Data Analysis and Statistical Thinking: An R Workbook

Fernando Racimo, ...

2021-03-26

Contents

| 1 | Introduction | 5 |
|----|---|-------------------------------------|
| 2 | Getting Started with Data Analysis | 7 |
| 3 | Probability Part 1 3.1 Today's programme | 9 9 9 11 12 13 17 |
| 4 | Probability Part 2 | 11 |
| 5 | Probability Catch-up15.1 Discrete distributions15.2 The Normal distribution and the Central Limit Theorem15.3 The exponential distribution1 | 114 |
| 6 | Linear Models16.1 Fitting a simple linear regression16.2 Interpreting a simple linear regression16.3 Simulating data from a linear model1 | 124 |
| 7 | Properties of Estimators and Inference17.1 Properties of point estimators17.2 Building confidence intervals17.3 ROC curve1 | 131 |
| 8 | Frequentist inference 1 | 33 |
| 9 | Bayesian Inference 1 | 35 |
| 10 | Classification | 37 |

4 CONTENTS

| 11 Model Assessment | 139 |
|--|-----|
| 12 Resampling | 141 |
| 12.1 The bootstrap | 141 |
| 12.2 Permutation test | |
| 12.3 Validation | 145 |
| 12.4 Cross-validation | 146 |
| 13 Mixed Models | 149 |
| 14 Ordination | 151 |
| 14.1 Libraries and Data | 151 |
| 14.2 Principal component analysis (PCA) | 152 |
| 14.3 PCA under the hood | 153 |
| 14.4 Principal components as explanatory variables | 155 |
| 14.5 NMDS | 156 |
| 15 Clustering | 161 |
| 15.1 Libraries and Data | 161 |
| 15.2 Distances | 161 |
| 15.3 K-means clustering | 162 |
| 16 REcoStats: Linear Models | 169 |
| 16.1 Fitting a simple linear regression | 169 |
| 16.2 Interpreting a simple linear regression | 173 |
| 16.3 Simulating data from a linear model | 175 |
| 16.4 Hypothesis testing and permutation testing | 177 |

Chapter 1

Introduction

Chapter 2

Getting Started with Data Analysis

Chapter 3

Probability Part 1

3.1 Today's programme

We will try out different R functions for simulating probability distributions and computing summary statistics from the resulting data.

We will cover the following topics

- Bernoulli distribution
- Binomial distribution
- Simulating discrete random variables
- Expected value
- Variance
- Probability mass function

3.2 The Bernoulli distribution: tossing a coin

We can "virtually" toss a coin in our R console, using the rbinom() function:

```
rbinom(1, 1, 0.5)
```

[1] 0

Try copying the above chunk to your R console and running it multiple times. Do you always get the same result?

This function has 3 required input parameters: n, size and prob. The first parameter (n) determines the number of trials we are telling R to perform, in other words, the number of coin tosses we want to generate:

```
rbinom(20, 1, 0.5)
```

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

Here, we generate 20 toin cosses, and the zeroes and ones represent whether we got a heads or a tails in each trial. For now, we will ignore the second parameter (size) and fix it at 1, but we'll revisit it in a moment. The third parameter (prob) dictates how biased the coin is. If we set it to 0.9, we'll get the outcomes of a biased coin toss, in particular biased towards heads:

```
rbinom(20, 1, 0.9)
```

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

Exercise: What happens when you set prob to 0.1? Or 0.999? Why?

What we are really doing here is simulating outcomes of a random variable that is governed by a particular probability distribution - in this case, the Bernoulli distribution. We can assign a name to this variable for storage and manipulation later on:

```
X <- rbinom(1, 1, 0.9)
```

If you type this in your console, X will now store the value of the outcome of a biased coin toss (either 0 or 1), which you can use later in your code.

How can we verify that R is really doing what we think it is doing? Well, if we think we have a fair coin and we throw it many times, then, on average, we should get the same number of heads and tails, right? This experiment should be more accurate the more trials we have. We can compute the average of our coin tosses by using the function sum(), which adds the elements of a vector, and then dividing by the total number of trials.

Let's create a new variable (n) that will determine how many trials we attempt, say 20.

```
nsims <- 20
sum(rbinom(nsims, 1, 0.5)) / nsims</pre>
```

```
## [1] 0.35
```

Exercise: Run the chunk of code above, in your own console. Do you get the same number as I do? Do you get exactly 0.5? If not, why not? Try the same exercise but with 100 trials, 1000 trials and 100000 trials. What happens as we increase the number of trials? This should illustrate how powerful R can be. We just threw 100 thousand coins into the air without even lifting our fingers! Try to repeat the exercise, but this time, set the Bernoulli prob parameter to be equal to a number of your choice (between 0 and 1). What is the average of all your coin tosses?

3.3 Adding up coin tosses

Let's say we are now not interested in any particular coin toss, but in the sum of several coin tosses. Each toss is represented by a 0 or a 1, so the sum of all our tosses cannot be smaller than 0 or larger than the total number of tosses we perform.

One way of doing this is by running 20 Bernoulli trials, and then adding them up using the function sum():

```
sum(rbinom(20, 1, 0.5))
```

[1] 12

Turns out there's a short-hand way of performing the same experiment, i.e. tossing a bunch of coins - each a Bernoulli random variable - observing their outcomes and adding them up, without using the sum() function at all. Here's where the second input parameter - size - of the rbinom() function comes into play. So far, we've always left it equal to 1 in all our command lines above, but we can set it to any positive integer:

```
rbinom(1, 20, 0.5)
```

[1] 7

The above code is equivalent to taking 20 Bernoulli trials, and then adding their outcomes up. The "experiment" we are running is now not a single coin toss, but 20 coin tosses together, so the first parameter is now 1 (note: if we had set it to 20, we would have performed 20 20-toss experiments). The outcome of this experiment is neither heads nor tails, but the sum of all the heads in all those coin tosses.

It turns out that this "experiment" is a probability distribution in its own right, and it is called the Binomial distribution. It has two parameters: the size of the experiment (how many tosses we perform) and the probability of heads for each toss (the prob parameter). The Bernoulli distribution is just a specific case of the Binomial distribution (the case in which we only toss 1 coin, i.e. size = 1). You can read more about this distribution if you go to the help menu for this function (type "?rbinom).

The Binomial and Bernoulli distributions are examples of distributions for discrete random variables, meaning random variables whose values can only take discrete values (0, 1, 2, 3, etc.). There are other types of distributions we'll study later, some of which can also take continuous values. For example, these could be any real number, or any real number between 2.4 and 8.3, or any positive number, etc. but we need not worry about these other distributions for now.

3.4 The expectation

We can compute the average of multiple Binomial trials. Let's try adding the results of 5 Binomial trials, each with size 20 (how many Bernoulli trials is this equivalent to?):

```
nsims <- 5
size <- 20
prob <- 0.5
X <- rbinom(nsims, size, prob)
X</pre>
```

```
## [1] 9 13 10 10 10
```

```
Xsum <- sum(X)
Xsum
```

```
## [1] 52
```

To get the average, we divide by the total number of trials. Remember here that the number of Binomial trials is 5:

```
Xave <- Xsum / nsims
Xave</pre>
```

```
## [1] 10.4
```

A shorthand for obtaining the mean is the function mean(). This should give you the same result:

```
Xave <- mean(X)
Xave</pre>
```

```
## [1] 10.4
```

Note that the mean need not be an integer, even though the outcome of each Binomial trial must be an integer.

Exercise: Try repeating the same exercise but using 100 Binomial trials, and then 100 thousand Binomial trials. What numbers do you get? What number do we expect to get as we increase the number of Binomial trials?

The number we "expect" our average to approach as we increase the number of trials is called the *Expectation* of a random variable. For discrete random variables, it is defined as follows:

$$E[X] = \sum_i x_i P[X = x_i]$$

Here the sum is over all possible values that the random variable X can take. In other words, it is equal to the sum of each of these values, weighted by the probability that the random variable takes that value.

In the case of a variable that follows the Binomial distribution, the expectation happens to be equal to the product of size and prob:

$$E[X] = np$$

Note that the n here refers to the size of a single Binomial trial. In the case of a variable Y that follows the Bernoulli distribution (which is just a Binomial with size 1), then:

$$E[Y] = p$$

This should make intuitive sense: if we throw a bunch of coins and add up their results, the number we expect to get should be approximately equal to the probability of heads times the number of tosses we perform. Note that this equality only holds approximately: for any given Binomial trial, we can get any number between 0 and the size of the Binomial experiment. If we take an average over many Binomial experiments, we'll approach this expectation ever more accurately. The average (also called "sample mean") over a series of n experiments is thus an approximation to the expectation, which is often unknown in real life. The sample mean is represented by a letter with a bar on top:

$$\bar{x} = \frac{\sum_{j=1}^{n} x_i}{n}$$

You can also think of the expectation as the mean over an infinite number of trials.

3.5 Our first probability mass function

Ok, all this talk of Bernoulli and Binomial is great. But what is the point of it? The nice thing about probability theory is that it allows us to better think about processes in nature, by codifying these processes into mathematical equations.

For example, going back to our coin tossing example, if someone asked you how many heads you'd expect among 20 tosses, your best bet would be to give the mean of a Binomial distribution with size 20 and probability of heads equal to 0.5: 0.5 * 20 = 10.

But this is a fairly intuitive answer. You didn't need probability theory to tell you that about half the coins would turn out heads. Plus, we all know that one might not get 10 heads: we might get 9, 13 or even 0 if we're very unlucky. What is then, the probability that we would get 10 heads? In other words, if we were to repeat our Binomial experiment of 20 tosses a large number of times,

how many of those experiments would yield exactly 10 heads? This is a much harder question to answer. Do you have a guess?

It turns out that probability theory can come to the rescue. The Binomial distribution has a very neat equation called its "Probability Mass Function" (or PMF, for short), which answers this question exactly:

$$P[X=k] = {n \choose k} p^k (1-p)^{n-k}$$

If we let k = 10, and plug in our values for the sample size and probability of heads, we get an exact answer:

$$P[X = 10] = {20 \choose 10} 0.5^{10} 0.5^{10} = 0.1762...$$

So in about 17% of all Binomial experiments of size 20 that we might perform, we should get that 10 out of the 20 tosses are heads.

Let's unpack this equation a bit. You can see that it has 3 terms, which are multiplied together. We'll ignore the first term for now. Let's focus on the second term: p^k . This is simply equivalent to multiplying our probability of heads k times. In other words, this means that we need k of the tosses to be heads, and the probability of this happening is just the product of the probability of heads in each of the total (n) tosses. In our case, k = 10, because we need 10 tosses, and n = 20 because we tossed the coin 20 times. So far, so good.

The third term is very similar. We not only need 10 heads, but also 10 tails (because we need exactly 10 of the tosses to be heads, no more, no less). The probability of this happening is the product of the probability of getting tails (1-p) multiplied n-k times. In our case, n-k happens to also be equal to 10.

But what about the first term: $\binom{n}{k}$? This is called a binomial coefficient. It is used to represent the ways in which one can choose an unordered subset of k elements from a fixed set of n elements. In our case, we need 10 of our 20 tosses to be heads, but we don't need to specify exactly which of the tosses will be heads. It could be that we get 10 heads followed by 10 tails, or 10 tails followed by 10 heads, or 1 head and 1 tail interspersed one after the other, or any other arbitrary combination of 10 heads and 10 tails. The binomial coefficient gives us the number of all these combinations. It is defined as:

$${n \choose k} = \frac{n!}{k! (n-k)!}$$

where

$$a! = a(a-1)(a-2)(a-3)...1$$

Ok, this is very neat, but how can we check this equation is correct? Well, we can use simulations! We can generate a large number of Binomial trials in R, and check how many of those are exactly equal to our choice of k. The fraction of all trials that are equal to k should approximate P[X=k]. Let's try this for k=10 and 500 trials.

```
nsims <- 500
size <- 20
prob <- 0.5
binomvec <- rbinom(nsims, size, prob)
binomvec</pre>
```

```
[1] 10 13 9 16 8 8 8 11 10 10 12 9 11 10 12 11 9 10 10 11 10 8 11 12 10
## [26] 10 12 8 11 12 12 5 9 9 7 8 10 11 10 10 12 12 8 12 7 12 11 11 5 14
## [51] 8 8 12 10 8 8 11 8 9 11 11 12 14 12 12 9 10 10 9 12 13 11 10 12 11
## [76] 9 9 15 13 12 10 6 9 14 9 5 6 11 11 7 7 7 13 13 10 12 11 11 8 7
## [101] 11 9 8 8 10 10 10 9 12 14 11 8 12 5 6 11 11 11 9 6 9 7 9 7 10
## [126] 9 11 8 12 11 13 8 7 9 11 9 10 13 11 8 9 8 10 11 10 8 9 10 10 12
## [151] 8 11 9 9 9 14 5 8 6 9 9 9 6 10 7 11 10 14 11 10 7 10 14 8 13
## [176] 13 11 8 9 6 10 10 9 8 7 10 13 11 11 14 9 9 6 8 11 13 9 9 11 9
## [201] 6 8 13 9 15 10 9 8 7 10 7 9 11 11 9 11 7 12 12 13 9 12 12 14 9
## [226] 11 7 9 15 6 7 11 7 9 11 10 8 13 10 13 8 11 10 9 15 10 10 13 11 13
## [251] 11 6 14 6 15 10 9 6 9 13 11 12 7 7 7 13 8 9 7 10 10 11 8 5 7
## [276] 11 9 8 11 9 15 12 10 9 11 8 10 9 8 12 8 12 8 8 14 9 6 14 15 8
## [301] 9 10 14 9 8 9 6 10 9 14 11 9 8 10 10 12 13 7 8 10 10 10 13 11 9
## [326] 7 9 10 9 12 13 11 11 11 10 12 9 6 14 8 5 12 7 14 13 9 11 7 11 6
## [351] 10 12 14 11 8 11 11 8 9 11 9 7 11 14 9 14 13 6 9 9 14 9 11 11 7
## [376] 7 14 9 12 7 11 6 13 12 12 4 13 13 10 9 10 8 8 11 8 9 11 13 14 7
## [401] 7 8 7 12 11 8 8 13 10 10 10 10 8 10 11 13 13 8 13 14 8 12 12 15 7
## [426] 6 9 13 8 10 8 9 7 9 11 11 11 10 4 6 9 12 12 11 11 10 12 7 13
## [451] 8 12 12 8 12 16 8 12 13 13 13 4 11 13 13 10 8 10 8 9 11 9 9 11 9
## [476] 10 10 7 12 8 10 9 12 9 11 8 9 11 10 11 9 11 11 11 10 11 11 10 11 12
```

We can determine which of these trials was equal to 10 using the "==" symbol:

```
k <- 10
verify <- (binomvec == k)
verify</pre>
```

```
## [1] TRUE FALSE FALSE FALSE FALSE FALSE FALSE TRUE TRUE FALSE FALSE FALSE
## [13] FALSE TRUE FALSE FALSE FALSE TRUE TRUE FALSE TRUE FALSE FALS
```

```
## [109] FALSE FALSE
## [121] FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## [133] FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE
## [145] TRUE FALSE FALSE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [157] FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE TRUE FALSE
## [169] FALSE TRUE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [181] TRUE TRUE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## [193] FALSE FALSE
## [205] FALSE TRUE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## [217] FALSE FALSE
## [229] FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE TRUE FALSE
## [241] FALSE FALSE TRUE FALSE FALSE TRUE TRUE FALSE FALSE FALSE FALSE
## [253] FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [265] FALSE FALSE FALSE FALSE TRUE TRUE FALSE FALSE FALSE FALSE FALSE
## [277] FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE TRUE FALSE
## [289] FALSE FALSE
## [301] FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [313] FALSE TRUE TRUE FALSE FALSE FALSE TRUE TRUE TRUE FALSE FALSE
## [325] FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE TRUE FALSE
## [337] FALSE FALSE
## [349] FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [361] FALSE FALSE
## [373] FALSE FALSE
## [385] FALSE FALSE FALSE TRUE FALSE TRUE FALSE FALSE FALSE FALSE FALSE
## [397] FALSE FALSE
## [409] TRUE TRUE TRUE FALSE TRUE FALSE FALSE FALSE FALSE FALSE
## [421] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE
## [433] FALSE FAL
## [445] FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [457] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE TRUE
## [469] FALSE FALSE FALSE FALSE FALSE FALSE TRUE TRUE FALSE FALSE FALSE
## [481] TRUE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE
## [493] FALSE FALSE TRUE FALSE FALSE TRUE FALSE FALSE
```

This returns a new vector in which each element is equal to TRUE if the corresponding element in "binomvec" is equal to 10, and FALSE otherwise. The nice thing is that R considers the value of TRUE to also be equal to 1, and the value of FALSE to also be equal to 0, so we can actually apply the function sum() to this vector!

```
how_many_tens <- sum(verify)
how_many_tens</pre>
```

[1] 74

Finally, to get at the fraction of all trials that were equal to 10, we simply divide by the number of trials:

proportion_of_tens <- how_many_tens / nsims
proportion_of_tens</pre>

[1] 0.148

You should have gotten a number pretty close to 17.62%. You can imagine that the more trials we perform, the more accurate this number will approximate the exact probability given by the PMF.

Exercise: Try repeating the above procedure but using a different value of k, between 0 and 20. Is your resulting probability lower or higher than P[X=10]?

Exercise: Plot a histogram of the vector "binomvec" using the function hist(). What do you observe?

3.6 The variance

There is another important property of a distribution: its *Variance*. This reflects how much variation we expect to get among different instances of an experiment:

$$Var[X] = E[(X - E[X])^2]$$

The variance is the expectation of $(X-E[X])^2$. This term represents the squared difference between the variable and it expectation, and so the variance is the expected value of this squared difference.

It turns out that the expectation of a function of a random variable is simply the sum of the function of each value the random variable can take, weighted by the probability that the random variable actually takes that value:

$$E[f(x)] = \sum_i f(x_i) P[X = x_i]$$

For a discrete random variable, we can thus write the variance as:

$$Var[X] = \sum_i (x_i - E[X])^2 P[X = x_i]$$

In the particular case of a discrete random variable that follows the Binomial distribution, the variance is a simple function of n and p:

$$Var[X] = np(1-p)$$

A measurable approximation to the *variance* is called the "sample variance" and can be computed from n samples of an experiment as follows:

$$s = \frac{\sum_{j=1}^{n} (x_j - \bar{x})^2}{n-1}$$

Just as we can compute the sample mean of a set of trials using the function mean(), we can easily compute the variance of a set of trials using the function var():

```
nsims <- 100000
size <- 10
prob <- 0.5
X <- rbinom(nsims, size, prob)</pre>
##
      [1] 4 5 5 6 7
                      3
                          5
                            7
                               6
                                 6
                                   6 6
                                          5
                                               7
                                                    4
                                                       6
                        6
                                        6
                                            5
                                                  3
                                                         3
                                                           4
                                                              5
                                                                3
                   6
                      5
                        5
                          3
                             7
                               8
                                               5
                                                  5
                                                    5
##
                                 4
                                    6
                                      8
                                        4
                                           4
                                             6
                                  5
                                           5
##
     [49]
                    6
                      4
                        4
                           6
                             3
                               4
                                    5
                                      4
                                         6
                                             6
                                                3
                                                  6
                                                    4
                                                       3
                                                         5
                                                              6
                                                                1
##
     [73]
          6
             3
               6
                 4
                    7
                      3
                        4
                           3
                             4
                               8
                                  5
                                    7
                                      7
                                         9
                                           5
                                             8
                                                4
                                                  6
                                                    5
                                                         7
##
     [97]
          8 3
               7
                 3
                   7
                      5
                        8 5
                             5
                               4
                                 7
                                    7
                                      2
                                         6
                                           5
                                             8
                                               7
                                                  6
                                                    4
                                                       1
             6
               5
                 3
                   5
                      6
                        5
                          4
                             7
                               5
                                  3
                                    3
                                      3
                                         6
                                           3
                                             4
                                                6
                                                    7
##
    [121]
          4
                    5
                      8
                        7
                           2
                               8
                                  6
                                    7
                                      7
                                         5
                                           2
                                             4
                                                5
                                                    7
##
    [145]
          4
             5
               4
                 4
                             6
                                                  5
                                                         5
                   5
                        7
                           4
                             6
                               7
                                  8
                                    2
                                      6
                                         6
                                           6
                                             5
                                                7
                                                    6
##
    [169]
          6
             6
               4
                 2
                      4
                                                  4
                                                         3
##
    [193] 7
             2 6 6 2 3 5 5 7
                               4 6 3 5
                                        5
                                          5
                                             2 6 4 3 6 5
                    8
                        2
                            4
                               7
                                  5
                                   5
                                      4
                                           4
                                               2
##
    [217] 7
             6 5
                 1
                      4
                          6
                                         7
                                             5
                                                  6
                                                    4 5
    [241]
          9
             6
               4
                 4
                    6
                      6
                        6
                           7
                             4
                               4
                                  2
                                    3
                                      5
                                         5
                                           7
                                             8
                                                8
                                                  5
                                                    5
                                                       7
                                                         5
##
             6 5 5
                    6
                        3 7
                             6
                               4
                                  5
                                    6
                                      4
                                           2
                                                    7
##
    [265] 7
                      4
                                         6
                                             6
                                                4
                                                  5
                                                       4
                                                         6
                                      3 7
    [289] 4 5 4 3 3 6 6 5
                            4 6 5 6
                                           3 6
                                               6 6
                                                    4 6 3 6
##
##
    [313] 4
            6
              7
                 7 3 5 7 6 5 2 4
                                    5
                                      7
                                         3 5 2 5 9
                                                    6 5 6
                             7
                               6
                                  8
                                             5
                                                6
                                                  6
                                                       7
##
    [337]
          7
             5
               6
                 5
                    5
                      4
                        5
                          5
                                    5
                                      4
                                         5
                                           4
                                                    5
                                                         3
    [361]
               9
                 4
                   3
                      4
                        3
                          1 5
                               6
                                 6
                                    2 7
                                         5
                                          8
                                             2 6
                                                  4
                                                    8
                                                      5
                                                         4
##
          4
             7
                                        2 4
                                                    6
##
                    4
                      6
                        4
                          6
                             8
                               1
                                 4 4
                                      6
                                             7
                                                7
             2
               5 5
                   6
                      4
                        4 6
                             6
                                  2 4
                                      6
                                        4
                                           3
                                             2 5 7 6 6 7
##
    [409]
          7
                               1
                      3
                        3
                             7
                               5
                                  7
                                                3
##
    [433]
             8
               3
                 4
                    6
                          3
                                    6
                                      6
                                         5
                                           6
                                             4
                                                  6
                                                    1
                                                       5
                                                         4
##
    [457]
          5
            5
               4
                 6
                   7
                      8 2 2
                             4
                               6
                                 6 8
                                      6
                                        5
                                           5
                                             2
                                                3
                                                  5
                                                    6
                                                      4
                                                         6
##
          5
               3 8
                   7
                      7
                        5 5
                             4
                               5
                                 3
                                    5
                                      7
                                         5
                                           5
                                             3
    [481]
##
    [505]
          6
             6
               3
                 4
                    4
                      5
                        5
                           2
                             7
                               3
                                 5
                                    5
                                      5
                                         4
                                           2
                                             7
                                                4
                                                  2
                                                    4
                                                       6
               8
                 6
                    7
                      3
                        6
                           6
                             7
                                6
                                 6
                                    3
                                      4
                                         4
                                           3
                                             6
                                               4
                                                  4
                                                    7
                                                       6
##
    [529]
               5
                 5
                        5
                          6
                            8
                               3 4
                                      4
                                         4
                                           6
                                             8 1 5 7
                                                       5 8
##
    [553]
                   4
                      7
                                    7
                                                                7
##
          6 6
               3 5
                   6
                      6
                        7
                           5
                             4
                               4
                                 1
                                    6
                                      3
                                        6
                                           5
                                             7 5
                                                  6 4 5
          3
               4
                    3
                        5
                             5
                               3
                                  3
                                    6
                                      5
                                         4
                                           2
                                             5
                                               5
                                                    8 5
                                                         6
##
    [601]
             4
                 7
                      5
                          4
                                                  3
               5
                 5
                    6
                      6
                        5
                           6
                             5
                               6
                                 2 5
                                      4
                                         4
                                           5
                                             5
                                                5
                                                  5
                                                    6 2
##
    [625]
          4
            5
                                                         7
##
          5 6 2 6 5 6 7 7
                             6 6 6 8
                                      8
                                        6
                                          4 2 6 4 3 5 3
    [649]
##
    [673] 7
            7 6 7 4 4 8 6 2 6
                                 7
                                    6
                                      3
                                        7
                                           6 4 4 4 2 5 6
##
    [697]
             4
               5
                 7
                    5
                      5
                        5
                          5
                             4
                               6
                                  6
                                    5
                                      3
                                         3
                                           8
                                             9
                                               5 3 3 6 2
                                                           7
##
    [721] 4
             3 4
                 4 5
                      4 5 5 6 6 7
                                    3 2 4 4 4 4 8 5 6 5 4 7
    [745] 7 4 4 4 3 5 4 7 4 4 5 7 6 4 5 6 2 5 5 3 5 6 4 5
##
    [769] 5 6 7 6 7 7 3 7 2 4 5 4 2 5 4 5 3 6 4 5 5 5 5 7
##
```

```
## [10729] 2 4 6 7 3 8 3 4 5 6 4 4 5 4 4 8 5 5 10 7 4 4 5 5
  [10753] 3 6 7 5 6 5 4 5 5 7 3 5 7 6 6 6 6 5 4 3 6 8 5 6
##
## [10777] 5 8 6 5 3 7 5 4 5 6 6 7 3 6 6 3 3 3 5 5 5 4 4 3
## [10801] 5 6 5 6 6 7 5 5 4 3 5 4 9 7 5 6 7 2 8 4 3 7 5 4
## [10825] 7 5 2 5 3 3 3 6 7 4 3 2 7 7 7 4 8 5 4 5 5 5 6 4
## [10849] 4 8 7 5 5 4 8 5 4 3 3 6 3 5 6 6 5 6 5 5 5 7 4 7
## [10873] 4 7 4 7 4 7 6 2 4 3 6 7 5 5 3 5 7 6 4 6 4 6 4 6
## [10897] 4 3 6 5 3 4 6 3 5 3 6 9 4 8 6 3 2 4 5 6 7 3 4 4
## [10921] 6 4 6 6 3 4 4 4 6 7 6 3 5 6 4 6 5 6 7 6 6 6 5 6
## [10945] 7 8 6 4 8 5 3 3 5 6 7 8 5 4 3 4 4 5 6 6 6 6 4 3
## [10969] 3 3 3 5 5 5 3 8 5 6 6 4 5 5 5 8 3 5 5 4 6 3 3 5
## [10993] 6 7 6 6 3 6 4 8 8 5 6 1 3 5 2 6 7 5 4 2 4 6 5 5
  [11017] 4 5 6 6 9 5 6 4 6 5 6 6 5 3 6 4 4 3 6 3 4 6 3 3
## [11041] 6 5 6 4 7 5 6 5 7 4 5 2 3 4 7 4 4 7 4 2 3 6 5 6
## [11065] 2 7 5 5 3 5 2 4 6 4 7 5 5 4 5 6 4 5 3 5 5 2 6 6
## [11089] 3 2 5 6 6 6 3 4 5 4 4 8 4 5 7 5 3 5 3 3 5 7 2 6
  [11113] 6 5 5 4 5 6 4 7 7 7 4 4 8 3 4 5 5 5 4 6 3 9 6 6
##
  [11137] 2 6 6 3 7 6 6 4 3 5 6 6 2 6 4 5 3 6 4 8 4 2 7 6
##
## [11161] 6 7 3 5 6 4 4 4 3 5 6 6 5 5 6 5 8 5 6 7 5 5 3 3
## [11185] 7 4 4 3 3 8 4 2 6 6 5 6 4 5 5 6 8 3 5 3 6 6 6 5
## [11209] 4 6 4 3 7 4 4 6 5 2 8 6 8 2 3 4 6 7 7 7 5 7 3 6
## [11233] 7 7 5 4 4 6 6 6 5 8 6 8 7 6 5 5 7 7 5 3 3 7 4 4
## [11257] 5 5 3 9 4 5 3 5 5 6 4 3 5 8 4 6 3 3 4 4 5 4 5 8
## [11281] 3 5 4 7 1 6 4 4 7 7 4 4 2 4 4 6 9 4 5 1 3 4 4 6
## [11305] 5 3 6 4 3 5 8 4 6 3 6 5 7 3 5 6 7 5 7 7 5 5 6 6
## [11329] 4 5 5 3 8 5 5 6 5 5 3 4 3 4 6 6 3 4 4 3 6 6 3 5
## [11353] 0 4 5 2 5 2 5 8 6 7 2 4 6 5 5 8 5 6 7 7 4 6 5 3
## [11377] 5 6 8 6 6 5 4 6 4 4 6 3 3 7 6 6 5 6 4 4 5 4 2 6
## [11401] 3 5 5 5 5 3 6 4 4 6 8 5 4 5 4 7 6 5 6 4 4 2 7 6
## [11425] 2 4 8 4 5 8 3 6 6 8 4 7 4 4 4 6 4 5 7 4 5 4 2 5
## [11449] 4 3 4 5 5 4 4 4 6 3 6 4 4 6 5 5 5 3 4 2 5 8 4 4
##
  [11473] 6 4 4 6 6 9 5 3 3 4 6 7 3 8 5 2 4 6 4 5 7 3 5 6
## [11497] 2 5 3 5 6 6 3 4 5 3 5 4 6 6 8 5 5 3 5 2 6 5 3 5
## [11521] 7 4 4 4 8 3 4 4 3 6 7 2 5 6 5 5 2 8 6 7 4 4 4 5
  [11545] 6 3 8 6 4 5 4 4 7 7 1 7 4 6 2 6 5 5 6 6 6 5 6 6
##
## [11569] 6 4 6 6 7 5 7 5 9 8 6 6 4 2 6 3 2 4 4 5 5 2 6 5
## [11593] 3 5 5 6 4 4 6 6 4 6 6 5 5 5 2 3 2 5 7 5 4 4 7 8
## [11617] 6 6 4 6 4 4 5 6 3 5 5 5 4 5 6 6 3 4 2 3 5 6 3 6
## [11641] 5 7 6 5 3 6 6 5 6 4 5 5 0 6 4 4 5 6 3 5 7 4 5 6
## [11665] 8 3 5 4 3 6 4 2 6 4 6 6 3 4 2 6 2 6 7 4 4 5 6 7
## [11689] 6 7 3 5 5 6 6 8 5 3 3 4 4 5 5 6 3 4 4 7 6 4 3 7
## [11713] 6 6 4 4 3 3 6 8 5 5 5 4 5 6 4 4 7 5 2 4 6 7 1 6
## [11737] 5 4 3 6 6 7 5 6 5 6 6 7 4 5 4 3 3 7 6 7 7 3 3 4
## [11761] 8 5 7 6 3 4 4 5 3 4 4 3 4 6 4 3 7 5 4 4 4 3 6 6
## [11785] 5 6 4 6 4 5 3 5 5 4 7 6 7 2 6 7 8 3 7 6 5 4 8 3
## [11809] 4 3 7 2 4 4 3 6 6 5 6 6 4 5 4 4 5 8 5 6 6 3 6 7
```

```
## [17353] 4 6 5 9 5 6 4 4 4 7 7 7 4 4 4 7 5 6 3 4 7 6 5 5
  [17377] 4 6 8 5 5 7 7 7 5 4 6 5 4 3 5 7 5 8 5 3 6 4 2 2
##
## [17401] 6 6 5 5 3 4 5 3 8 6 5 3 4 6 4 5 7 3 5 3 4 7
## [17425] 2 5 4 5 3 5 5 7 7 5 5 4 4 6 5 5 6 4 7 7 7 5 3 7
## [17449] 6 5 7 7 5 4 4 4 5 4 5 7 5 4 6 4 7 6 3 7 8 2 7 5
## [17473] 4 5 6 5 7 6 4 5 4 5 5 5 7 7 7 7 6 5 1 5 4 6 3 6
## [17497] 2 7 5 7 7 8 3 1 4 5 5 6 6 6 7 6 3 6 6 3 5 7 6 3
## [17521] 6 7 8 5 3 5 4 7 5 5 4 5 7 3 4 5 1 5 5 7 3 5 6 5
## [17545] 7 8 7 7 3 5 6 6 3 6 4 8 4 6 8 6 1 5 5 4 5 5 7 5
## [17569] 4 4 7 3 5 5 7 6 3 3 2 5 8 5 3 5 6 5 4 4 3 4 4 6
## [17593] 7 6 4 6 8 5 5 4 4 4 5 5 5 3 2 6 6 4 4 6 5 6 5 6
## [17617] 2 6 5 4 5 5 5 8 6 4 3 6 3 5 6 5 5 5 7 5 3 5 5 6
## [17641] 7 4 4 2 6 6 6 3 3 5 4 4 3 2 4 6 5 8 7 5 7 3 4 5
## [17665] 5 7 5 1 4 3 6 6 5 5 2 5 4 3 5 2 6 7 6 8 5 5 4 4
## [17689] 3 7 5 7 5 5 4 5 6 4 5 6 5 5 4 5 5 7 4 4 5 6 7 6
## [17713] 5 5 6 3 5 8 6 4 3 7 2 7 5 4 3 6 6 5 2 5 6 5 5 6
  [17737] 1 3 8 7 3 6 8 7 2 5 3 6 3 5 2 9 4 3 3 2 2 4 4 5
##
## [17761] 8 5 5 3 5 5 4 6 7 7 5 5 5 3 6 3 3 4 4 7 7 5 4 3
## [17785] 5 4 8 6 2 2 6 7 4 6 3 6 6 5 6 5 8 5 5 5 6 8 6 8
## [17809] 5 5 4 3 4 7 7 5 6 4 7 5 7 5 4 5 6 4 2 6 5 8 4 6
## [17833] 5 5 2 7 4 1 8 5 7 5 4 1 4 6 3 4 5 3 6 7 6 3 3 6
## [17857] 2 4 1 6 5 3 5 3 4 4 7 4 5 8 4 5 3 4 5 6 7 6 6 7
## [17881] 4 5 7 6 7 6 3 5 3 5 5 5 4 6 8 4 5 6 3 4 3 4 2 9
## [17905] 4 8 5 3 7 4 4 3 7 6 2 2 5 3 4 4 6 4 4 2 3 7 5 3
## [17929] 4 5 4 3 5 6 6 5 2 6 6 4 5 8 5 3 6 5 4 5 7 7 5 5
## [17953] 8 3 4 5 3 3 5 3 8 3 6 5 4 5 5 5 5 5 7 2 6 0 7 5
## [17977] 2 7 6 8 5 6 3 4 4 5 4 6 5 4 1 6 4 7 4 5 4 6 5 7
## [18001] 3 7 4 5 6 7 6 4 5 2 4 4 6 3 7 5 5 3 3 6 5 6 4 4
## [18025] 6 6 6 4 7 4 5 9 3 6 5 2 6 3 6 7 3 5 4 6 5 2 3 2
## [18049] 6 3 5 8 6 4 4 7 3 5 4 6 5 7 5 4 7 4 7 5 5 3 4 7
## [18073] 6 5 7 4 3 2 7 6 5 6 8 4 3 3 6 8 4 2 4 4 4 8 5 6
##
  [18097] 4 3 5 7 5 6 6 6 4 4 5 4 3 0 4 3 4 4 5 6 4 4 3 4
## [18121] 6 5 4 6 8 4 7 5 8 7 4 7 6 5 5 7 5 5 6 8 5 7 6 6
## [18145] 5 5 6 5 6 5 3 4 3 5 5 4 6 5 3 4 5 3 7 5 5 4 3 5
## [18169] 5 3 2 6 6 5 5 5 9 5 4 5 5 2 2 3 4 6 4 4 9 5 3 4
## [18193] 6 5 6 4 3 5 4 3 3 4 5 4 5 3 1 7 6 4 5 6 4 3 8 6
## [18217] 3 6 4 4 4 5 6 7 5 9 6 5 8 2 4 5 7 4 5 3 5 4 6 3
## [18241] 3 6 5 5 6 7 6 2 5 7 4 5 4 3 4 5 4 5 3 4 4 2 5 2
## [18265] 0 4 7 3 4 5 7 5 5 3 3 5 5 5 5 6 4 3 6 5 2 5 4 3
## [18289] 5 7 4 4 2 6 5 1 7 2 7 7 1 6 4 5 4 4 2 4 5 5 4 2
## [18313] 6 4 4 4 8 5 7 5 6 6 1 7 5 7 5 4 3 6 3 6 5 5 8 7
## [18337] 5 4 7 8 6 4 6 3 5 5 7 6 5 2 6 6 7 5 5 5 3 4 3 7
## [18361] 4 2 4 5 8 6 4 6 6 4 4 2 6 4 4 6 8 6 5 6 4 6 4 4
## [18385] 4 2 8 5 7 8 5 4 4 5 7 6 8 7 8 6 3 5 6 4 1 3 5 2
## [18409] 4 8 4 7 4 4 5 4 4 3 5 7 8 6 5 5 4 6 3 6 6 6 7 5
## [18433] 6 6 4 6 3 6 4 7 4 9 5 5 7 3 5 3 4 7 6 4 8 6 3 6
```

```
## [41641] 6 1 5 7 5 5 3 3 7 7 5 6 5 7 4 3 5 3 6 3 3 3 4 8
  [41665] 3 8 6 7 4 3 6 5 4 4 5 6 6 6 3 4 6 6 7 5 5 5 4 8
##
## [41689] 4 3 7 6 5 5 2 7 5 4 4 6 8 2 3 9 6 6 7 6 6 4 7 1
## [41713] 3 4 3 6 7 6 5 2 6 5 4 5 5 2 2 6 7 3 6 5 4 7 5
## [41737] 6 4 6 8 6 7 5 5 4 4 8 5 7 8 5 4 7 6 5 6 4 7 5 2
## [41761] 7 6 5 5 6 5 5 6 2 5 5 8 4 4 4 4 4 5 5 6 5 5 6 7
## [41785] 6 4 2 5 4 2 4 2 4 5 4 5 5 6 8 4 6 4 6 4 4 5 4 6
## [41809] 4 2 7 1 6 6 4 5 6 5 2 7 7 3 7 6 2 4 5 4 3 6 4 5
## [41833] 7 5 8 4 3 6 4 5 4 3 6 4 7 3 7 5 8 5 7 2 5 5 6 5
## [41857] 5 6 7 7 5 9 8 4 5 6 5 5 4 5 6 3 5 5 7 7 6 6 5 4
## [41881] 4 2 3 6 5 2 5 1 5 5 5 6 3 5 5 8 6 6 6 9 5 4 8 5
## [41905] 6 5 4 4 6 3 6 5 8 4 5 4 4 4 2 4 3 5 5 4 5 2 3 3
  [41929] 4 4 6 7 4 4 3 6 5 5 3 4 7 6 5 6 5 2 7 4 6 7 5 4
## [41953] 5 5 6 8 4 6 3 4 5 5 2 4 9 2 7 6 5 7 4 4 5 7 5 4
## [41977] 8 5 5 7 7 2 4 8 4 4 5 4 4 3 4 6 4 7 3 4 6 5 6 7
  [42001] 5 4 5 5 6 4 6 5 4 4 5 5 7 5 3 5 7 7 3 4 3 4 4 4
##
  [42025] 6 8 5 4 4 4 5 3 5 3 5 7 5 3 4 2 4 3 4 6 5 4 5 6
##
##
  [42049] 5 5 4 2 6 3 6 6 6 5 5 5 4 3 5 5 3 2 5 6 8 5 4 6
## [42073] 8 4 6 5 5 5 3 5 6 5 5 5 3 6 8 3 3 2 4 7 4 7 5 4
## [42097] 5 5 3 4 6 4 5 4 4 6 3 4 5 3 4 6 6 6 2 6 3 6 5 5
## [42121] 4 5 4 4 7 5 4 2 3 4 6 7 7 5 5 4 5 3 5 4 3 5 4 4
## [42145] 4 7 4 4 7 3 6 4 5 5 7 7 7 4 5 4 6 6 4 4 2 4 3 3
## [42169] 7 4 8 3 3 6 4 4 5 7 8 1 6 6 6 5 4 6 6 5 5 6 5 6
## [42193] 5 7 2 3 1 4 5 5 5 5 5 4 6 7 6 4 7 6 4 4 2 6 4 5
## [42217] 4 5 6 7 6 7 5 6 7 7 6 5 5 1 4 4 7 4 7 3 3 7 6 4
## [42241] 5 4 6 6 6 4 7 5 5 6 4 6 6 5 9 6 6 4 2 3 5 8 4 6
## [42265] 2 6 5 7 5 5 5 2 5 6 6 5 7 5 6 3 6 1 4 5 9 4 6 5
## [42289] 5 3 5 1 7 5 8 5 4 5 2 5 4 5 3 7 7 6 6 4 4 5 4 5
## [42313] 7 8 6 5 6 4 7 5 5 7 6 5 1 4 6 4 3 6 5 2 5 6 2 4
## [42337] 5 8 4 7 4 5 4 5 5 5 6 8 5 4 6 7 5 8 5 4 3 5 7 3
## [42361] 4 5 8 5 4 4 3 5 4 6 5 4 4 2 5 5 6 5 5 3 4 4 9 5
##
  [42385] 6 3 5 4 6 5 5 6 5 5 5 6 6 5 4 4 5 6 4 5 6 5 4 6
  [42409] 6 5 5 0 5 6 4 4 5 5 5 8 3 3 6 4 4 3 7 7 5 6 6 3
##
## [42433] 4 8 2 8 6 5 6 2 5 5 5 5 6 5 5 5 5 3 4 6 4 3 5 6
  [42457] 6 5 5 6 6 6 3 2 1 4 3 8 7 6 3 7 7 5 6 6 3 5 4 7
##
  [42481] 5 6 5 5 5 6 4 3 7 9 6 3 6 4 5 4 5 3 3 2 5 5 3 7
##
## [42505] 5 7 3 7 8 5 3 3 7 5 5 7 4 4 6 5 5 4 6 6 6 6 2 3
## [42529] 6 7 4 5 6 5 7 6 9 6 4 6 3 4 7 7 5 9 6 5 2 4 4 7
## [42553] 4 4 5 5 3 4 7 6 7 7 5 3 6 4 4 3 3 6 3 4 5 5 8 7
##
  [42577] 3 6 7 3 5 8 3 2 4 3 3 6 4 6 5 9 5 5 4 6 7 5 4 5
## [42601] 5 3 1 5 6 6 4 6 4 5 6 2 5 7 5 5 4 8 6 6 4 3 5 4
## [42625] 5 6 8 4 5 7 6 5 4 6 6 4 6 8 5 5 1 4 5 5 7 3 6 7
  [42649] 4 4 6 7 5 5 6 4 4 3 4 5 8 4 4 6 3 7 6 4 6 4 3 3
## [42673] 7 5 7 4 4 4 4 3 3 6 5 4 5 3 5 6 7 7 7 4 5 7 3 5
## [42697] 4 5 6 5 4 4 7 6 5 3 5 4 4 5 6 6 6 9 6 6 7 4 5 5
## [42721] 5 7 5 7 6 4 5 4 8 4 5 4 4 4 4 5 5 6 2 1 2 6 5 4
```

```
##
  [48265] 2 6 7 4 6 6 4 4 4 4 8 5 8 5 4 7 6 5 3 4 4 2 3 7
   [48289] 6 5 4 6 5 6 5 7 6 7 4 8 4 5 6 3 4 3 8 5 4 3 6 6
##
## [48313] 6 9 6 4 8 7 5 1 4 4 5 4 2 7 4 4 5 8 7 2 5 4 6 3
## [48337] 5 7 7 5 4 5 5 7 4 4 4 5 6 3 6 5 5 3 5 5 3 6 7 5
## [48361] 4 7 3 4 10 7 5 4 5 2 5 6 7 1 6 5 6 4 6 4 4 6 5 1
## [48385] 3 7 3 3 6 4 10 4 4 5 0 5 4 10 5 7 8 6 4 4 5 5 8 7
## [48409] 6 3 7 5 5 6 4 7 6 4 3 6 6 8 6 2 3 5 8 4 5 6 4 7
## [48433] 5 5 6 4 7 5 4 4 5 7 5 6 6 3 5 5 1 8 6 5 7 6 2 4
## [48457] 6 5 4 5 3 6 4 7 1 8 4 7 5 4 2 6 5 4 7 5 7 6 3 6
## [48481] 4 4 7 5 5 7 4 4 7 5 4 9 6 6 8 7 2 4 7 4 6 7 4 6
## [48505] 5 5 5 4 2 3 4 8 6 3 7 4 6 2 4 7 5 2 6 4 4 8 1 3
  [48529] 7 3 5 5 4 3 4 5 5 7 5 6 3 5 7 6 7 5 3 4 3 4 4 4
##
  [48553] 2 5 7 8 6 5 4 5 6 5 3 8 5 3 6 4 6 6 8 4 4 5 3 3
## [48577] 4 2 4 7 6 6 3 4 3 6 7 8 3 4 6 4 6 4 2 6 4 2 7 3
## [48601] 4 6 9 6 6 7 6 4 8 3 5 6 7 4 6 5 5 7 4 7 7 7 3 6
  [48625] 1 7 8 1 6 9 7 8 4 3 5 2 6 8 4 4 4 6 5 3 8 4 4 2
##
  [48649] 4 3 6 4 6 8 3 5 6 5 6 6 4 2 8 4 4 5 7 3 6 6 6 6
##
  [48673] 6 6 6 7 2 2 4 2 6 5 3 7 6 8 6 5 6 5 5 7 3 5 3 6
##
##
  [48697] 6 7 8 5 7 4 2 6 3 8 6 6 5 5 6 3 5 3 3 7 6 6 6 5
  [48721] 5 6 5 8 3 6 2 6 7 2 3 5 4 7 6 7 3 3 5 4 6 6 2 7
##
## [48745] 6 4 4 4 7 4 6 3 6 6 6 4 4 6 5 5 7 5 6 5 7 6 5 7
## [48769] 7 5 3 4 6 4 6 5 1 4 5 7 4 4 4 5 4 5 4 6 7 4 5 2
## [48793] 5 4 7 4 4 5 7 3 6 1 8 4 6 6 2 5 2 4 8 4 5 3 7 4
## [48817] 7 1 6 4 7 6 2 4 4 4 3 7 6 4 2 4 5 5 4 6 7 4 7 6
## [48841] 6 5 7 6 6 3 5 8 7 6 4 6 8 4 4 4 4 4 4 3 5 7 3 5
## [48865] 3 5 3 7 7 7 6 6 6 4 6 2 5 5 5 4 2 3 6 4 2 6 6 2
## [48889] 5 7 7 6 5 6 6 1 2 3 5 8 4 4 5 4 5 5 2 5 5 4 4 7
## [48913] 4 4 3 7 6 4 6 3 6 4 3 5 3 5 5 5 7 4 5 5 8 4 5 6
## [48937] 8 7 5 3 5 5 5 4 2 6 5 5 10 4 6 5 5 5 7 1 5 6 5 3
## [48961] 6 5 4 3 5 4 6 4 5 6 6 5 5 6 6 5 3 6 4 6 7 8 8 6
## [48985] 8 3 6 5 5 5 5 5 4 4 4 7 5 6 2 7 6 3 3 3 6 4 5 5
##
  [49009] 6 6 5 3 4 8 3 4 6 3 3 6 4 5 5 5 3 4 5 7 5 6 6 3
  [49033] 6 7 5 5 6 3 5 6 8 5 4 7 8 6 5 6 3 3 5 7 5 3 2 5
##
## [49057] 7 6 8 3 9 6 4 6 4 4 5 8 7 4 5 5 3 5 6 3 8 3 6 1
  [49081] 5 8 4 5 5 7 5 5 6 2 3 6 4 4 7 7 6 4 6 6 3 3 6 6
##
##
  [49105] 5 7 7 8 6 6 3 5 6 4 6 5 6 8 7 6 8 3 4 4 4 5 6 2
## [49129] 3 4 7 4 4 8 3 6 7 6 7 7 5 3 2 4 5 7 5 4 7 2 6 7
## [49153] 4 2 3 4 3 4 3 5 7 5 6 6 3 7 5 6 5 6 5 2 5 6 1 7
## [49177] 5 7 6 5 6 5 6 7 6 5 4 6 6 4 4 3 4 4 5 6 5 5 5 8
##
  [49201] 7 6 4 3 5 4 6 3 5 3 5 5 6 5 3 6 5 6 8 5 4 8 5 4
## [49225] 7 3 6 2 6 5 4 3 5 3 3 4 7 5 4 4 6 7 6 2 8 4 7 4
## [49249] 5 6 3 8 6 6 5 7 3 5 3 2 6 6 7 5 2 3 3 7 5 6 8 3
  [49273] 5 5 4 6 5 7 3 1 4 7 3 5 5 5 6 5 5 5 3 6 5 4 6 3
## [49297] 5 7 4 6 6 3 6 7 7 7 6 3 7 5 2 4 4 4 4 3 4 5 5 7
## [49321] 6 2 4 3 4 5 3 7 6 3 5 5 6 4 6 6 3 6 3 3 6 2 2 6
## [49345] 6 2 6 5 6 5 3 5 5 7 3 6 8 1 3 6 6 4 2 5 7 6 4 5
```

```
[50473] 5 5 4 4 6 5 5 5 2 6 4 5 4 4 4 4 6 4 4 5 3 5 7 7
##
   [50497] 4 6 4 5 6 5 4 4 5 3 6 6 2 8 1 5 4 2 5 5 4 4 5 5
##
  [50521] 5 1 7 4 4 6 6 5 6 7 6 2 5 5 4 5 9 7 5 4 8 3 7 7
##
## [50545] 4 3 5 4 5 3 6 8 5 7 7 3 4 8 6 5 5 4 4 2 5 6 6 5
##
  [50569] 6 3 3 8 2 7 5 7 7 9 3 5 6 4 4 5 6 6 7 3 6 5 6 5
## [50593] 6 5 7 5 3 5 5 7 6 2 5 5 3 2 6 8 7 4 5 3 5 6 6 3
## [50617] 7 4 5 7 7 4 5 8 4 6 4 5 6 4 3 8 5 6 6 3 5 3 3 9
## [50641] 5 3 6 5 3 4 3 5 6 0 4 4 5 7 4 3 6 3 3 4 5 3 7 6
## [50665] 5 6 6 6 6 5 6 6 4 7 5 6 4 7 4 6 4 5 5 6 6 3 6 5
## [50689] 5 5 3 5 6 5 7 8 5 3 6 5 5 4 6 4 7 3 5 3 2 5 3 4
  [50713] 4 1 7 6 1 5 4 5 5 4 5 8 4 4 7 3 6 7 5 3 3 6 4 4
  [50737] 3 5 6 6 7 3 5 7 7 8 4 4 4 3 5 5 6 5 5 6 5 1 3 6
##
  [50761] 7 5 5 4 5 4 6 5 3 3 4 4 6 2 4 5 5 7 3 6 5 6 5 4
## [50785] 4 7 5 3 4 4 7 7 5 6 5 3 4 3 6 6 5 4 5 3 2 5 4 6
  [50809] 5 2 7 8 3 5 7 4 2 7 7 6 1 5 6 7 5 5 3 6 5 3 6 3
##
  [50833] 8 6 6 5 4 4 6 4 5 8 6 8 5 7 4 4 7 4 4 5 5 4 6 4
##
   [50857] 6 8 5 3 4 6 5 2 6 6 8 7 4 4 6 4 3 3 7 4 5 2 4 6
##
##
  [50881] 7 6 5 3 7 6 4 3 3 4 6 2 4 5 2 6 4 6 5 7 3 4 6 4
##
  [50905] 5 5 5 4 6 4 4 6 5 8 3 7 6 8 6 3 6 5 5 3 8 3 7 5
  [50929] 6 4 3 5 5 6 5 4 4 6 8 3 4 5 3 2 2 4 6 5 4 4 3 5
##
##
  [50953] 8 5 3 5 5 3 5 4 4 5 4 3 3 3 6 4 7 5 4 1 5 6 5 3
## [50977] 3 3 4 6 6 7 5 3 5 6 5 6 5 4 2 5 5 4 6 5 4 7 3 6
## [51001] 6 4 7 5 7 8 5 4 5 7 7 4 3 5 5 8 4 5 4 2 6 7 6 6
## [51025] 6 5 5 5 9 3 5 4 3 6 4 4 4 3 5 3 7 2 7 5 7 4 3 5
## [51049] 5 3 6 8 5 4 7 5 7 6 5 7 7 5 5 5 7 7 6 4 3 8 5 4
## [51073] 3 2 5 6 4 6 5 4 5 6 3 3 5 6 4 6 7 6 4 5 6 5 5 6
## [51097] 4 5 6 4 3 7 7 6 4 4 7 5 4 3 6 9 5 7 6 7 6 5 7 3
## [51121] 4 5 5 7 5 5 6 5 5 3 4 5 6 4 3 8 8 3 4 5 4 4 5 3
## [51145] 6 4 4 6 4 5 5 5 7 5 3 5 3 6 6 4 8 4 4 4 4 7 4 3
## [51169] 3 3 5 5 4 3 4 4 4 6 4 5 4 4 6 2 6 5 6 2 6 3 3 6
## [51193] 4 6 4 4 7 6 5 5 5 6 2 5 3 3 7 3 5 3 4 4 6 8 4 9
##
  [51217] 4 4 2 6 3 4 5 7 3 4 5 6 4 2 6 6 4 4 5 6 3 1 4 6
  [51241] 6 5 4 6 4 2 7 5 4 7 7 6 4 5 5 3 7 3 4 7 4 7 7 4
##
  [51265] 6 5 5 6 4 5 5 4 4 2 6 9 5 5 5 7 3 3 4 6 5 5 5 6
  [51289] 4 6 5 5 5 3 6 6 3 5 4 7 4 5 6 7 7 6 4 3 5 6 5 8
##
  [51313] 5 3 4 5 4 6 2 4 3 6 7 5 5 8 4 9 5 4 4 5 3 5 4 9
##
## [51337] 6 2 5 7 5 7 4 7 6 6 4 7 4 5 3 5 5 4 5 4 5 6 4 3
## [51361] 3 7 4 6 5 3 7 5 7 5 5 5 5 6 4 7 8 5 6 3 1 5 4 6
  [51385] 5 4 4 5 7 4 6 6 4 8 5 8 6 7 5 6 6 5 6 3 5 7 4 6
##
##
  [51409] 4 6 2 5 6 6 8 5 8 3 3 3 6 4 4 6 2 7 5 7 2 3 3 6
  [51433] 6 5 4 5 6 4 6 4 4 4 6 6 8 5 4 6 5 3 4 5 5 5 5 4
  [51457] 5 3 5 9 4 6 4 8 6 5 4 5 5 3 5 3 4 6 2 3 3 7 4 5
##
##
  [51481] 5 5 6 2 4 3 5 4 4 5 7 7 5 4 6 5 4 5 8 6 5 3 8 7
## [51505] 4 3 5 3 7 4 5 5 4 4 2 6 5 5 4 8 3 7 9 6 5 3 4 7
## [51529] 4 4 5 5 3 6 1 6 3 2 9 5 8 4 6 3 6 6 2 3 5 6 8 3
## [51553] 5 5 5 5 5 7 9 3 6 6 8 3 4 5 7 6 5 5 5 6 6 7 6 5
```

```
[52681] 5 3 5 4 8 5 6 6 3 4 3 6 5 8 8 6 6 7 2 5 7 9 8 7
##
##
   [52705] 4 5 5 5 3 4 3 6 4 5 4 4 8 5 5 6 5 3 6 7 5 4 5 5
  [52729] 5 5 4 6 4 5 7 5 5 4 5 7 4 9 3 6 4 7 6 5 5 4
##
  [52753] 4 7 5 4 3 7 2 3 4 6 6 5 3 5 6 4 5 7 3 5 6 4 4 3
##
##
  [52777] 4 6 4 6 2 6 6 7 4 8 6 3 4 3 4 6 6 5 4 5 7 5 9 8
## [52801] 5 5 7 4 4 4 5 4 6 6 8 5 6 3 7 7 5 8 6 3 7 8 2 8
## [52825] 4 8 7 8 7 3 4 8 4 6 6 3 6 3 3 6 3 5 5 4 9 4 5 5
## [52849] 4 5 6 6 7 4 6 4 8 3 5 5 6 4 7 0 6 4 4 5 5 4 5 5
## [52873] 4 5 8 4 4 6 6 8 8 4 7 6 6 5 6 2 5 6 6 7 2 3 5 3
## [52897] 5 5 7 5 5 5 4 3 7 4 5 5 5 4 6 7 5 8 4 4 6 5 3 7
  [52921] 3 5 8 5 4 6 4 4 5 5 3 4 2 7 5 3 5 4 4 4 4 4 7 3
  [52945] 6 6 4 8 7 6 5 7 6 3 5 3 4 5 3 6 6 4 4 2 8 6 4 6
##
  [52969] 5 5 6 3 6 4 4 4 5 5 7 8 5 7 4 4 7 5 6 5 3 6 5 3
##
## [52993] 4 6 4 4 4 2 2 7 4 6 4 3 4 6 3 7 6 3 6 4 7 2 5 6
  [53017] 7 3 6 6 7 6 7 5 3 7 6 5 4 7 5 6 5 5 4 8 5 6 5 7
##
  [53041] 5 3 6 6 7 4 7 7 5 6 4 3 6 7 2 7 5 4 5 4 7 2 8 6
##
   [53065] 4 4 3 7 5 7 7 3 3 8 4 4 4 6 2 2 4 7 6 3 7 4 5 5
##
##
  [53089] 5 8 4 4 5 4 5 5 3 5 5 4 3 2 6 6 2 8 5 2 5 4 6 5
##
  [53113] 3 7 5 4 2 5 4 4 8 6 5 5 4 6 4 6 5 0 4 7 6 6 1 5
  [53137] 7 5 6 6 3 2 3 4 6 6 6 4 6 4 5 2 5 5 4 6 5 5 6 6
##
##
  [53161] 1 7 4 5 5 4 4 3 4 6 8 6 5 5 2 7 6 5 6 5 5 6 4 4
## [53185] 5 4 3 5 4 5 5 3 6 4 4 4 6 5 5 5 6 5 6 4 4 6 5 7
## [53209] 8 6 4 5 4 7 5 4 4 5 5 6 6 3 6 3 5 4 3 8 7 5 4 6
  [53233] 7 3 5 3 8 4 4 4 3 2 4 7 5 3 5 3 3 2 6 5 5 8 4 1
##
  [53257] 5 6 5 6 7 7 4 6 7 5 3 5 6 3 4 6 5 6 5 5 5 5 6 3
##
##
  [53281] 3 5 7 5 4 4 5 5 7 6 4 4 6 8 4 7 4 3 5 6 7 5 3 3
## [53305] 4 5 8 3 2 2 5 3 5 4 6 4 4 6 6 6 4 5 3 5 8 5 5 7
  [53329] 6 9 7 2 5 7 4 2 4 4 4 6 6 5 3 5 5 5 4 6 5 4 6 6
##
## [53353] 6 7 3 2 6 5 4 6 3 4 4 4 6 4 5 6 3 6 4 6 4 5 8 6
## [53377] 6 5 5 7 4 5 6 2 5 5 3 5 3 4 3 6 4 1 7 4 4 6 3 5
  [53401] 6 5 2 2 5 4 3 5 4 6 7 2 5 8 6 5 4 6 6 6 5 6 5 6
##
  [53425] 5 3 4 1 2 4 5 4 4 5 2 6 5 3 4 4 2 5 9 4 6 7
##
  [53449] 7 4 5 3 5 2 6 6 5 5 4 2 4 5 6 8 4 4 6 3 5 5 3 5
##
##
  [53473] 6 5 7 5 7 3 4 5 3 5 5 8 3 8 2 3 3 4 4 7 4 8 3 7
  [53497] 4 5 4 4 3 2 4 5 5 6 5 3 5 2 7 6 5 3 7 5 7 6 6 6
##
##
  [53521] 5 10 6 4 7 4 7 3 2 6 2 5 3 5 8 2 3 7 6 7 3 5 5 3
##
  [53545] 6 3 4 6 5 4 4 4 5 4 7 5 4 6 3 3 5 5 7 4 3 6 5 7
## [53569] 4 4 4 7 5 4 5 2 4 6 5 4 5 5 7 6 5 6 5 5 7 3 5 6
  [53593] 7 7 6 5 5 4 4 4 5 3 4 4 7 5 4 4 4 6 5 5 2 7 9 5
##
##
  [53617] 4 5 4 6 4 6 7 4 5 6 3 7 3 3 8 5 6 8 7 4 5 4 7 4
  [53641] 4 4 5 4 6 5 3 5 6 3 6 5 5 8 4 6 7 6 6 2 5 5 6 6
##
  [53665] 4 8 4 4 2 7 2 7 4 7 4 5 6 2 5 7 4 5 4 4 2 6 6 3
##
  [53689] 5 5 7 3 7 4 6 7 4 5 4 6 6 7 3 5 5 4 2 3 2 6 5 5
##
  [53713] 5 4 2 6 4 4 5 8 6 7 5 8 5 7 5 7 5 4 4 5 5 4 6 6
##
## [53737] 4 4 4 7 5 4 6 7 4 4 3 3 8 9 5 5 4 3 3 5 6 6 7 3
## [53761] 6 7 6 5 5 4 6 6 3 5 4 4 4 4 5 7 3 6 4 4 6 4 7 5
```

```
[54889] 6 4 3 8 5 5 8 8 8 2 4 2 6 4 6 7 5 5 3 6 5 5 3 4
##
   [54913] 6 4 7 2 6 5 7 4 8 5 9 6 1 3 6 5 3 3 6 6 4 6 8 4
##
  [54937] 7 3 5 8 5 3 2 5 5 5 6 7 6 6 4 7 6 5 3 8 6 6
##
## [54961] 5 6 5 7 5 6 8 2 6 4 8 3 5 8 4 4 5 8 3 6 8 4 7 5
##
  [54985] 7 5 5 4 5 7 3 5 4 6 4 4 7 6 3 7 5 5 5 4 4 4 7
## [55009] 3 5 6 7 6 2 4 5 4 7 8 5 7 4 6 7 3 8 1 5 3 6 4 4
## [55033] 3 6 5 5 3 5 8 5 6 6 6 7 6 6 6 5 4 3 6 3 4 4 8 7
## [55057] 4 4 6 6 7 4 6 4 4 5 6 4 5 2 5 3 4 9 4 5 7 4 6 8
  [55081] 2 7 5 7 4 5 3 5 7 7 6 3 9 3 3 7 8 5 4 7 4 4 6 9
##
  [55105] 7 6 7 4 9 8 5 4 4 4 2 4 3 5 5 6 2 2 6 7 7 5 4 8
##
  [55129] 7 5 4 7 5 5 2 4 5 4 6 5 5 5 8 6 4 7 7 4 4 7 5 5
  [55153] 1 6 8 8 6 6 2 5 6 2 5 3 5 6 4 7 6 7 4 8 3 4 7 5
##
  [55177] 2 5 4 6 5 2 4 5 4 4 6 4 2 6 6 3 5 4 4 3 5 3 4 6
##
  [55201] 5 7 5 4 7 1 7 4 5 4 4 7 5 4 6 3 3 6 4 5 5 6 6 5
  [55225] 5 6 4 7 6 6 4 6 5 5 3 6 6 5 6 8 4 6 4 3 4 3 8 4
##
  [55249] 4 5 5 4 5 1 7 7 6 2 6 5 2 2 6 5 4 5 4 4 6 9 5 4
##
  [55273] 4 6 5 4 3 7 7 7 6 6 4 4 5 4 7 6 4 6 3 5 4 3 7
##
  [55297] 4 3 5 6 7 6 6 3 5 4 6 6 6 5 2 5 4 7 3 6 2 5 4 3
##
##
  [55321] 4 4 5 7 3 5 5 5 6 5 5 7 6 2 8 5 4 5 4 5 5 5 4 3
  [55345] 4 6 3 4 8 2 5 3 4 7 5 5 7 4 4 5 5 5 6 3 5 4 3 1
##
##
  [55369] 7 5 4 6 5 4 5 3 3 7 4 7 3 2 7 3 4 3 6 5 5 3 5 5
## [55393] 1 3 4 5 5 7 6 5 8 2 4 9 7 6 6 4 4 5 6 3 6 3 5 3
## [55417] 4 4 6 5 8 3 6 4 7 5 3 8 4 6 6 3 4 3 7 8 4 7 6 5
  [55441] 2 4 2 5 3 3 4 6 7 5 4 6 6 6 5 2 8 7 6 3 4 7 4 7
##
  [55465] 4 8 7 3 6 7 5 4 3 4 7 7 7 4 6 5 5 6 6 5 6 7 5 3
##
## [55489] 3 9 1 4 5 5 4 5 4 6 7 2 8 7 6 3 5 6 7 4 9 6 3 3
## [55513] 4 8 4 3 5 7 6 7 4 5 5 2 2 4 2 4 5 4 5 4 5 5 5 4 5
  [55537] 2 3 2 5 5 5 5 6 4 7 9 4 3 6 5 6 3 8 4 7 5 4 6 3
##
## [55561] 3 3 3 5 5 5 4 5 3 5 3 4 6 5 4 4 9 7 3 4 7 5 4 9
## [55585] 7 7 4 6 4 7 6 0 5 7 3 7 4 4 4 5 6 3 6 5 7 4 9 6
  [55609] 6 7 5 1 6 3 4 5 3 2 8 5 6 6 2 4 3 6 4 8 4 6 3 4
##
##
   [55633] 4 3 3 7 5 6 4 6 3 6 4 6 5 5 6 3 4 7 4 5 5 7 4 5
  [55657] 6 3 6 8 5 8 1 4 8 5 3 4 2 8 4 7 4 4 3 6 5 7 3 6
##
##
  [55681] 4 7 3 6 5 6 5 3 7 6 6 6 6 6 4 6 8 3 3 7 7 5 4 7
  [55705] 3 6 4 4 5 5 6 5 4 2 5 6 5 5 4 6 5 7 6 5 4 3 7 5
##
##
  [55729] 5 6 5 8 4 4 6 3 4 5 4 6 2 6 5 7 4 5 6 5 6 3 5 4
##
  [55753] 7 3 6 5 7 2 4 5 7 6 4 4 3 6 5 2 5 9 4 6 6 3 4 6
## [55777] 6 7 1 6 5 4 3 4 5 7 5 4 5 5 5 5 5 4 2 2 4 6 9 3
  [55801] 7 4 7 1 5 6 3 7 4 4 3 4 5 5 6 6 5 6 6 4 3 6 2 6
##
##
  [55825] 4 5 5 4 8 7 4 5 3 7 8 7 3 4 8 5 6 6 4 3 5 2 4 5
##
  [55849] 8 4 7 4 4 5 8 6 4 4 6 7 6 3 4 6 4 6 2 2 5 3 2 6
  [55873] 1 6 5 5 5 6 2 6 4 7 7 6 5 4 4 5 6 3 7 5 5 8 6 6
##
  [55897] 7 4 6 7 7 5 3 6 5 4 4 5 4 6 3 7 8 6 7 5 6 5 5 6
##
  [55921] 6 5 7 5 5 6 8 4 5 2 8 6 3 6 4 4 5 4 4 6 6 4 5 2
##
  [55945] 6 4 4 7 4 3 3 5 8 7 4 5 7 3 5 6 9 7 4 3 3 3 4 6
##
## [55969] 3 4 6 4 6 3 7 6 6 4 5 4 7 3 5 5 6 7 5 7 2 7 6 6
```

```
[57097] 5 5 7 5 7 4 4 5 6 7 6 2 6 7 6 5 3 7 7 4 1 4 7 3
##
   [57121] 5 6 5 7 5 3 6 2 5 5 3 6 5 7 8 2 5 5 8 5 5 5 4 4
##
  [57145] 5 1 7 3 2 6 7 4 4 4 6 9 4 5 2 1 4 6 4 7 3 5 7 6
  [57169] 6 6 6 7 7 6 7 4 4 4 6 6 5 5 3 6 6 6 5 8 8 5 6 5
##
##
  [57193] 6 6 5 6 4 8 7 5 3 4 3 3 6 5 5 5 3 6 2 8 6 5 7 3
## [57217] 4 6 3 5 6 4 7 4 4 7 2 4 4 7 4 3 3 5 4 4 5 4 4 4
## [57241] 6 3 7 8 6 5 4 8 3 4 2 6 7 8 1 4 6 4 4 5 6 6 7 6
## [57265] 3 4 7 5 5 7 6 6 5 4 4 2 6 5 4 7 6 3 7 5 2 8 5 3
## [57289] 3 3 6 9 4 5 7 6 6 4 5 5 4 5 9 3 3 6 7 6 5 5 4 5
  [57313] 5 5 4 5 1 1 6 3 5 6 7 7 5 5 3 4 4 5 5 8 6 4 6 5
##
  [57337] 6 7 3 6 7 5 4 3 5 3 5 5 3 6 2 8 4 7 5 5 4 6 3 5
  [57361] 4 4 6 5 5 4 6 7 4 5 5 4 6 6 6 5 4 4 7 4 4 5 2 5
##
  [57385] 7 6 1 5 6 5 4 7 2 5 4 6 6 4 6 4 6 6 3 4 7 5 4 6
## [57409] 3 6 3 3 6 6 3 7 7 4 3 3 7 6 7 5 3 2 3 6 6 7 2 8
## [57433] 7 6 4 6 5 5 6 6 7 4 2 7 6 5 6 4 4 7 5 8 7 6 4 5
  [57457] 7 5 4 2 6 5 4 9 6 7 3 6 5 6 5 8 3 7 4 4 3 7 5 6
##
  [57481] 7 6 7 3 3 5 4 5 5 7 6 6 5 7 4 6 4 1 5 5 5 2 4 4
##
  [57505] 6 6 4 4 4 5 4 4 6 5 3 5 6 5 7 3 8 3 5 7 10 3 5 4
##
##
  [57529] 3 1 3 5 7 4 7 5 4 6 5 6 4 7 3 4 7 7 6 4 4 6 5 4
  [57553] 4 4 4 6 6 5 5 5 5 6 1 7 7 4 2 4 5 6 6 4 4 3 4 6
##
##
  [57577] 7 3 4 1 5 5 1 7 4 6 6 6 2 5 7 4 6 5 7 4 5 5 6 5
## [57601] 5 6 8 6 6 6 2 5 8 5 7 4 5 5 4 4 3 3 5 5 6 6 4 4
## [57625] 9 6 4 5 4 5 4 4 4 3 3 7 6 4 6 7 8 6 5 6 7 4 5 5
## [57649] 6 6 7 3 6 4 3 4 4 7 5 7 5 6 6 4 7 7 7 6 5 1 5 6
## [57673] 5 9 7 4 4 0 3 7 5 5 9 6 2 7 5 9 5 6 2 4 2 2 5 5
## [57697] 6 5 4 4 7 4 4 3 6 4 5 5 6 5 4 4 5 6 5 6 7 5 5 6
## [57721] 6 5 7 3 5 6 8 7 5 7 3 7 5 4 2 6 4 2 6 5 4 7 7 6
  [57745] 5 4 2 3 5 6 4 3 7 5 4 6 5 3 8 5 3 4 8 2 5 3 4 8
## [57769] 6 3 4 6 5 4 5 6 7 4 6 5 4 4 4 4 3 6 3 6 6 5 7 6
## [57793] 6 3 8 6 4 6 5 3 5 7 5 5 6 5 5 3 8 6 3 3 6 7 4 7
  [57817] 5 5 5 3 5 5 6 6 6 4 2 5 2 2 3 7 2 6 6 7 3 8 7 4
##
##
  [57841] 6 3 4 4 4 6 3 3 6 7 6 5 6 6 3 8 8 6 5 5 3 5 6 5
  [57865] 2 8 2 7 3 6 5 6 4 5 5 6 7 5 4 9 5 6 3 4 7 7 5 3
##
  [57889] 6 5 6 5 5 4 5 6 4 5 7 5 6 6 6 5 7 5 6 3 5 6 5 7
  [57913] 7 4 5 6 7 6 7 6 3 4 5 3 5 6 3 3 4 4 5 5 4 3 5 5
##
##
  [57937] 6 6 5 6 3 5 7 5 4 6 6 5 7 4 5 8 4 9 6 4 5 3 4 4
##
  [57961] 6 5 4 4 6 7 7 5 6 5 3 5 4 8 5 4 6 2 4 4 4 5 8 6
## [57985] 5 3 5 3 4 4 6 6 3 3 9 6 5 6 6 4 3 3 5 6 4 4 5 4
  [58009] 8 4 9 6 6 4 6 5 4 4 5 6 3 3 8 6 4 6 3 7 7 5 5 7
##
##
  [58033] 6 6 4 4 4 7 4 5 6 3 3 4 4 5 4 8 5 3 6 2 6 6 5 2
  [58057] 4 6 3 9 3 5 5 4 7 5 4 3 6 6 5 3 5 8 7 3 7 4 8 7
##
  [58081] 6 4 7 2 3 4 7 7 5 5 5 4 4 6 8 2 1 8 6 6 7 5 4 4
##
  [58105] 7 7 6 4 5 6 5 4 7 4 5 7 6 4 7 5 5 3 5 2 8 7 1 4
## [58129] 7 5 4 5 7 5 4 2 6 7 7 1 6 6 5 2 5 7 7 6 4 2 4 4
## [58153] 9 3 7 4 6 4 5 3 6 3 5 3 4 4 3 3 7 5 4 8 4 4 6 5
## [58177] 2 5 4 6 5 5 6 5 7 5 6 7 7 6 4 6 7 4 8 4 3 6 8 7
```

```
[61513] 5 4 8 3 7 5 7 6 6 6 6 7 6 4 7 7 3 7 6 6 5 6 5 4
##
   [61537] 5 7 3 5 6 8 8 5 5 5 6 6 8 4 8 8 5 5 3 5 3 5 5 7
##
  [61561] 2 1 4 6 6 6 4 3 3 6 6 7 6 5 8 4 6 3 5 3 3 4 8 3
##
  [61585] 5 5 5 7 7 4 6 5 4 4 5 4 5 5 5 3 4 5 4 5 3 4 4 5
##
  [61609] 4 8 3 5 4 4 6 4 1 6 5 5 5 8 5 3 6 7 3 5 5 1 1 6
##
## [61633] 7 4 5 5 5 5 5 5 7 5 4 6 5 5 6 6 5 8 5 8 2 3 5 5
## [61657] 3 7 3 5 7 4 7 6 5 7 7 6 6 5 7 4 4 4 5 3 4 5 7 5
## [61681] 7 7 6 5 5 6 3 3 5 5 6 5 7 3 3 3 5 3 3 5 5 4 5 6
##
  [61705] 7 4 2 5 5 3 5 5 5 5 6 2 5 6 5 6 4 6 8 5 5 3 5 6
  [61729] 3 3 5 6 4 4 5 6 4 6 5 4 4 3 6 5 6 8 2 3 7 4 8 4
##
  [61753] 6 5 5 5 7 3 4 9 5 7 2 7 6 7 3 3 3 7 4 5 4 2 6 5
  [61777] 4 4 2 6 7 7 7 3 6 6 7 3 5 5 6 5 3 6 7 6 6 4 4 4
##
  [61801] 4 6 7 4 7 6 4 5 1 5 6 5 7 7 6 6 6 7 3 3 6 6 2 4
##
  [61825] 5 4 4 5 7 2 3 3 5 4 3 3 5 3 4 5 5 3 5 4 4 6 6 6
  [61849] 4 5 6 7 3 6 8 3 6 5 4 3 5 8 8 4 3 7 7 3 3 7 7 4
##
  [61873] 2 7 3 6 6 3 4 5 4 4 5 7 5 6 4 7 5 5 7 6 7 3 4 4
##
  [61897] 6 7 3 7 6 3 5 6 4 4 5 1 4 4 6 4 1 6 6 5 4 4 6 4
##
##
  [61921] 6 5 3 8 6 4 3 2 6 7 7 3 5 3 4 5 8 5 7 5 8 3 3 6
##
  [61945] 6 8 6 7 4 6 5 3 3 4 4 5 5 6 4 2 3 5 6 3 4 5 7 3
  [61969] 6 5 5 5 5 6 8 7 7 5 3 5 7 7 2 5 5 3 7 8 1 6 5 5
##
##
  [61993] 8 4 3 6 4 7 4 3 4 6 5 7 3 4 4 6 7 6 6 5 3 7 7 6
## [62017] 8 6 4 4 6 3 7 7 4 6 4 5 4 5 5 5 3 7 9 4 4 8 8 5
## [62041] 2 5 3 7 3 5 5 5 3 7 8 5 6 2 7 4 6 6 5 4 2 5 4 6
  [62065] 7 6 5 5 3 2 5 4 6 7 7 6 5 3 4 6 7 5 8 7 4 2 1 5
##
##
  [62089] 6 3 6 2 5 2 6 5 6 6 2 6 3 3 5 5 4 6 8 4 4 3 4 5
## [62113] 6 7 4 5 2 4 4 5 4 6 3 7 4 3 6 6 4 6 4 5 8 5 4 4
## [62137] 2 6 4 4 3 3 4 6 6 4 4 5 8 5 7 6 4 6 3 6 8 5 4 5
  [62161] 4 2 3 2 3 6 6 4 5 4 6 6 7 5 3 7 5 3 8 1 5 5 6 7
##
## [62185] 3 6 5 8 6 6 2 6 2 8 7 4 6 2 5 6 4 4 5 4 4 6 4 3
## [62209] 4 6 7 6 4 5 8 4 6 6 6 5 4 5 4 5 8 3 6 6 9 5 2 6
  [62233] 4 4 8 6 6 2 6 2 2 4 7 7 2 5 6 6 4 6 1 6 8 7 1 7
##
##
  [62257] 3 8 3 6 5 6 5 4 6 7 3 5 6 7 6 5 5 6 4 6 3 4 4 5
  [62281] 8 6 5 5 6 6 4 5 3 3 3 9 4 5 2 8 4 4 6 5 3 8 4 4
##
  [62305] 2 4 8 5 7 6 3 8 6 3 5 5 4 5 6 3 6 6 8 5 5 7 5 4
  [62329] 6 5 5 5 5 3 5 2 5 6 6 4 4 7 6 4 5 3 3 6 8 5 5 5
##
##
  [62353] 6 5 5 7 4 4 6 6 5 6 6 4 7 6 5 8 5 3 6 3 5 5 6 3
##
  [62377] 8 7 5 7 5 5 6 4 6 4 5 6 5 7 6 8 3 2 6 6 5 2 3 4
## [62401] 3 8 4 3 5 6 9 3 4 9 6 6 4 7 3 7 5 6 4 7 5 4 2 3
  [62425] 4 6 5 4 5 5 5 6 5 3 5 8 4 4 2 3 3 5 3 7 5 8 4 4
##
##
  [62449] 4 3 6 5 2 3 6 3 5 4 5 5 8 4 5 6 5 4 3 4 6 5 3 5
  [62473] 7 6 6 4 5 6 3 6 7 8 3 4 5 5 4 3 5 7 7 6 2 8 3 8
  [62497] 3 7 7 2 4 7 5 3 5 4 4 5 5 3 4 4 7 7 4 8 5 4 8 2
##
##
  [62521] 7 3 7 7 4 4 4 5 4 4 7 5 4 4 3 4 7 4 4 7 5 5 9 6
  [62545] 7 2 7 5 5 5 5 5 3 4 4 3 5 4 6 5 7 7 7 4 4 8 6 6 5
##
  [62569] 5 5 4 2 7 6 7 8 2 3 5 3 3 3 3 6 7 7 1 5 6 3 6 4
##
## [62593] 6 6 4 3 5 6 4 2 4 7 7 5 4 4 2 4 8 8 2 5 3 7 5 3
```

```
[65929] 5 5 2 8 6 7 5 3 5 7 6 6 7 6 4 6 5 6 6 7 4 6 5 5
##
   [65953] 8 4 4 5 9 8 5 2 5 7 1 4 4 6 6 7 8 5 6 4 4 1 4 5
##
  [65977] 7 2 5 5 2 5 3 6 4 5 3 5 3 3 4 9 5 4 7 3 3 4 3 8
##
  [66001] 4 4 5 8 3 4 7 4 3 1 6 5 5 6 6 7 4 5 5 8 4 5 4 7
##
##
  [66025] 3 5 5 4 7 7 6 5 6 7 2 2 4 4 6 5 7 6 5 8 2 3 4 4
## [66049] 7 4 9 6 3 4 4 4 2 8 3 3 3 5 6 5 3 5 2 5 4 4 3 1
## [66073] 3 7 5 5 2 6 6 5 6 3 6 3 6 2 4 5 6 3 4 2 6 4 2 6
## [66097] 6 5 5 6 4 3 4 3 4 6 6 5 5 6 7 6 3 5 2 5 9 3 6 4
  [66121] 5 4 6 3 4 7 5 7 4 5 4 3 6 3 6 6 5 5 5 3 2 5 5 6
##
  [66145] 7 7 4 5 5 3 4 5 9 8 7 3 4 3 5 4 6 5 6 5 5 4 6 2
##
  [66169] 4 5 3 6 3 5 5 5 5 6 5 7 4 6 5 6 4 5 8 5 6 5 2 3
  [66193] 5 4 5 6 6 6 6 6 6 4 7 5 5 5 6 4 5 5 7 2 3 6 4 2 5
##
  [66217] 3 6 7 7 4 3 8 5 6 2 4 2 5 3 5 5 3 7 4 5 4 6 6 6
##
  [66241] 6 6 7 4 5 4 6 10 4 5 5 2 5 4 4 7 4 7 5 6 6 5 6 3
  [66265] 5 6 5 5 4 5 4 4 4 5 7 4 5 4 4 4 8 5 4 8 5 5 5 7
##
  [66289] 7 5 5 4 4 4 5 3 7 5 5 6 8 5 5 5 7 6 5 5 7 6 6 4
##
   [66313] 2 6 4 5 6 7 2 7 5 3 6 7 4 4 7 6 6 4 6 4 7 6 6 5
##
##
  [66337] 6 5 2 5 7 6 5 6 4 5 7 3 6 4 5 6 3 6 5 5 4 6 4 6
##
  [66361] 6 5 6 3 4 7 7 6 2 4 5 4 5 8 5 5 6 5 6 4 4 4 5 6
  [66385] 6 6 5 3 6 6 7 6 3 10 7 4 3 7 4 5 1 7 5 2 6 4 5 4
##
##
  [66409] 5 6 4 5 3 6 5 7 4 3 6 5 5 5 4 4 5 4 6 4 4 3 6 4
## [66433] 7 6 6 9 8 5 3 4 6 4 3 5 1 5 7 4 7 3 8 3 2 6 3 5
## [66457] 6 2 3 6 5 4 4 3 6 1 6 6 6 5 6 6 7 5 6 5 3 4 3 6
  [66481] 1 4 6 8 7 3 3 6 5 6 7 6 4 3 7 5 4 5 1 6 8 3 6 5
##
  [66505] 6 5 4 7 2 5 7 8 5 4 5 2 6 4 6 5 2 2 5 3 3 6 3 4
##
##
  [66529] 5 5 8 6 6 4 4 3 7 5 3 6 4 6 3 5 6 7 4 5 3 3 2 5
## [66553] 4 4 5 4 4 7 4 5 4 7 3 5 5 3 3 8 4 5 3 8 5 6 4 4
  [66577] 5 7 5 6 5 6 4 4 6 4 3 6 4 6 6 4 5 2 5 3 6 4 4 6
##
##
  [66601] 5 5 5 8 3 6 1 5 6 3 6 6 4 8 2 5 4 6 6 2 6 6 6 8
  [66625] 3 4 2 5 5 3 7 6 6 7 7 8 3 3 4 5 5 7 4 5 2 4 4 6
##
  [66649] 5 4 5 4 4 5 6 1 4 5 5 4 3 8 2 2 3 7 6 3 4 7 5 3
##
   [66673] 7 3 2 5 3 4 2 5 4 6 4 4 6 2 6 6 6 3 4 2 3 8 5 4
##
  [66697] 4 6 3 6 5 4 5 6 8 3 5 7 9 7 4 6 6 7 6 5 3 5 7 5
##
##
  [66721] 3 5 6 3 8 6 3 4 6 4 4 8 4 7 5 5 4 4 6 4 6 5 8 6
  [66745] 6 6 1 3 4 6 6 4 5 5 3 1 5 7 6 7 6 7 2 4 6 6 5 6
##
##
  [66769] 4 3 7 6 5 7 8 4 5 8 4 7 6 7 5 3 5 3 4 2 2 4 5 4
##
  [66793] 3 7 7 8 4 4 4 3 6 6 5 6 3 7 5 3 6 7 6 5 6 8 5 3
## [66817] 6 3 5 3 7 6 4 5 5 6 5 2 4 3 7 7 5 5 3 3 8 5 6 4
  [66841] 7 3 5 6 5 4 3 6 4 6 4 3 5 6 3 1 9 4 6 6 6 6 8 5
##
##
  [66865] 7 6 3 6 6 4 3 7 4 4 6 4 5 4 6 8 3 7 5 5 6 6 6 5
  [66889] 3 3 6 4 6 6 5 7 4 3 5 4 4 6 6 4 4 6 4 4 6 7 5 2
##
  [66913] 6 4 4 5 4 6 3 7 6 7 5 8 6 6 5 7 6 4 5 8 3 3 7 5
##
##
  [66937] 6 4 6 4 5 6 4 5 4 7 6 4 4 6 4 6 3 7 5 5 5 5 5 5
  [66961] 6 3 5 4 2 5 6 3 2 5 5 5 3 9 5 8 5 3 4 6 6 2 5 2
##
  [66985] 5 2 2 4 6 5 5 6 9 8 5 5 3 6 4 6 7 5 5 4 4 4 7 4
##
## [67009] 6 5 5 5 6 5 9 7 6 6 2 4 3 5 4 4 5 7 5 9 1 5 4 4
```

```
[72553] 4 6 5 2 5 5 5 6 3 4 5 5 4 6 6 4 5 4 5 5 5 8 7 6
##
   [72577] 1 5 4 7 6 5 5 6 4 6 4 4 5 4 9 7 5 5 3 5 9 6 6 4
##
  [72601] 5 5 8 6 6 6 6 5 7 3 4 3 3 6 8 3 5 6 7 5 6 4 7 4
## [72625] 4 3 4 8 4 5 6 5 6 4 2 4 5 5 5 4 7 7 5 6 3 4 6 3
## [72649] 4 4 7 4 8 5 4 5 7 6 6 5 5 5 7 3 5 6 7 7 6 2 4 7
## [72673] 6 5 5 4 4 3 5 3 4 7 8 6 8 2 6 4 5 4 4 6 2 5 5 3
## [72697] 5 3 4 5 5 5 4 7 4 6 6 8 5 7 5 8 7 7 5 7 2 6 5 4
## [72721] 7 7 5 5 4 5 6 6 6 6 5 7 3 2 4 5 7 5 4 6 7 5 6 4
## [72745] 4 4 7 5 4 6 7 8 3 6 5 4 4 7 3 6 8 3 2 1 7 7 5 6
## [72769] 6 6 5 4 7 4 5 7 6 5 4 2 5 4 5 5 3 1 6 6 4 7 3 6
## [72793] 6 4 6 6 5 4 5 6 5 5 3 7 4 5 5 6 4 4 5 4 5 7 4 5
## [72817] 6 7 4 6 6 5 4 7 7 6 4 5 7 1 3 6 3 5 6 5 7 7 3 5
  [72841] 5 3 6 6 4 6 3 6 6 4 7 6 3 7 6 5 2 2 6 5 5 4 7 6
## [72865] 2 5 2 4 4 4 6 7 5 8 7 6 6 6 7 8 4 7 4 4 5 6 5 7
## [72889] 5 7 6 6 7 7 4 2 6 5 5 3 5 3 1 5 5 6 4 5 5 6 4 9
## [72913] 5 5 6 6 4 5 3 4 4 2 6 6 8 4 1 5 7 4 5 4 6 7 6 3
  [72937] 6 7 4 6 7 7 4 6 6 5 4 7 7 8 5 4 7 5 5 5 6 4 5 5
##
##
  [72961] 6 6 3 3 5 6 5 3 4 5 5 6 5 6 5 6 5 5 5 4 6 4 6
## [72985] 7 9 7 4 5 4 6 6 3 8 4 5 2 3 5 5 2 2 6 7 9 5 6 4
  [73009] 4 1 6 7 7 6 7 6 4 2 5 6 3 7 5 5 7 7 5 5 6 5 3 2
##
## [73033] 6 4 8 5 6 6 3 7 6 1 5 4 2 5 5 10 3 6 2 5 3 1 5 3
## [73057] 9 5 6 7 6 5 6 5 4 6 6 6 6 6 5 3 4 4 6 3 4 5 4 3
## [73081] 4 5 4 5 5 6 1 5 5 7 7 3 6 6 5 5 5 2 4 5 9 6 6 6
## [73105] 4 5 2 4 6 3 5 8 4 3 5 7 6 6 6 4 6 6 6 6 6 7 5 6
## [73129] 5 5 4 4 5 6 4 2 6 6 4 7 5 5 4 8 7 2 9 7 5 3 6 6
## [73153] 6 4 3 7 2 6 6 4 3 4 3 6 6 6 2 9 1 6 6 4 6 4 6 8
## [73177] 6 6 4 3 7 5 7 7 6 2 5 3 6 5 7 4 8 9 6 2 4 4 4 6
## [73201] 7 7 5 4 3 7 2 2 3 4 3 1 6 5 6 5 7 6 4 5 6 3 5 3
## [73225] 4 5 3 4 5 3 5 6 4 5 5 6 3 5 3 5 5 2 6 5 6 6 8 6
## [73249] 5 5 6 6 6 4 4 4 4 6 6 6 2 3 3 6 4 4 6 5 8 5 4
## [73273] 4 5 4 5 5 4 5 4 9 6 5 2 6 5 4 6 1 6 3 4 7 4 5 4
  [73297] 7 6 4 3 4 6 4 5 5 3 5 3 3 5 6 8 5 5 3 6 1 6 4 7
##
  [73321] 3 2 4 8 8 6 5 5 5 3 6 8 4 4 7 6 7 7 4 5 5 6 8 4
##
  [73345] 7 4 6 8 7 6 6 5 4 5 7 3 5 4 4 5 7 6 4 5 5 5 7 2
  [73369] 5 2 5 3 5 7 7 4 3 3 3 6 4 5 4 5 5 3 9 8 8 5 7 4
##
##
  [73393] 6 7 5 4 3 5 2 8 5 5 5 3 7 4 4 6 4 4 8 5 6 6 6 6
## [73417] 4 4 6 5 4 8 5 5 7 6 3 5 6 7 6 6 2 5 3 4 5 2 4 5
## [73441] 8 5 5 2 6 6 7 5 4 4 6 6 6 2 5 3 7 5 6 4 6 8 6 3
## [73465] 7 4 6 4 3 2 5 6 5 6 5 4 5 3 6 8 7 6 5 5 5 4 4 8
##
  [73489] 6 7 3 5 5 8 3 7 5 6 7 5 9 5 7 5 6 6 6 5 5 5 7 4
## [73513] 6 4 7 8 5 9 3 3 7 5 4 7 2 9 3 6 5 6 6 2 7 3 9 4
## [73537] 5 5 3 4 1 5 6 1 6 3 2 4 7 5 6 7 4 5 6 3 7 5 8 5
  [73561] 1 5 5 3 5 6 5 6 6 8 6 5 2 5 4 2 3 6 7 8 8 4 4 6
##
## [73585] 7 5 5 7 6 5 6 4 6 3 6 4 5 5 7 7 5 5 8 5 4 7 7 6
## [73609] 3 4 7 5 4 3 4 6 6 4 5 3 4 5 9 2 1 6 3 4 6 6 4 5
## [73633] 5 4 7 6 4 2 5 5 3 6 5 5 7 3 7 4 5 4 3 7 6 5 6 5
```

```
[76969] 5 7 5 4 3 5 7 4 5 4 5 7 5 7 7 5 5 5 7 5 2 7 4 8
##
   [76993] 5 5 6 8 5 8 5 5 6 2 5 4 7 5 6 7 9 3 2 5 7 6 3 6
##
  [77017] 7 3 6 7 4 1 4 4 5 7 4 8 3 4 2 5 7 8 5 5 6 6 3 5
## [77041] 3 6 2 1 5 5 2 3 6 7 6 4 5 6 7 4 5 7 4 1 6 8 7 6
## [77065] 7 4 5 5 5 7 7 6 5 7 8 4 4 4 7 6 7 4 5 5 1 5 4 2
## [77089] 8 7 4 5 4 6 6 6 7 6 6 5 6 5 4 7 6 5 6 3 4 6 3 5
## [77113] 5 5 3 7 3 6 4 7 3 7 5 4 7 3 4 6 4 3 6 3 4 5 5 6
## [77137] 4 4 4 4 5 4 4 4 5 4 6 6 8 3 6 7 4 3 5 4 6 3 3 5
## [77161] 5 5 4 6 7 7 2 5 6 3 5 2 4 4 6 5 7 7 8 5 4 5 4 6
## [77185] 7 3 5 2 2 7 1 4 5 5 6 4 5 5 7 5 4 6 7 5 3 5 4 4
  [77209] 6 6 7 8 5 6 5 2 7 7 5 7 4 8 5 4 4 5 8 2 2 7 6 6
  [77233] 2 5 3 4 4 5 6 5 5 5 5 3 6 6 8 7 4 5 4 6 8 7 6 8
##
  [77257] 6 5 9 4 5 6 3 5 7 3 5 7 7 6 7 7 5 4 3 3 5 7 6 4
## [77281] 4 4 4 3 2 7 5 8 4 5 5 4 4 6 5 5 5 3 5 8 6 4 5 6
  [77305] 6 3 8 4 4 3 5 3 5 4 4 5 4 5 7 4 7 5 6 4 6 9 6 6
##
  [77329] 6 6 4 5 7 7 4 2 5 4 5 4 3 3 4 5 3 6 6 4 7 5 7 4
##
  [77353] 2 6 6 5 5 4 4 7 7 5 2 4 3 4 4 4 6 5 4 2 5 6 7 5
##
  [77377] 6 6 7 5 2 4 4 7 5 8 4 7 4 5 3 5 2 7 5 5 4 5 5 8
##
##
  [77401] 3 5 5 6 5 7 4 7 6 4 5 6 7 8 6 6 6 5 2 8 5 7 5 4
  [77425] 6 7 4 8 7 6 3 6 4 6 5 7 5 6 4 6 7 5 6 4 6 5 5 7
##
## [77449] 7 5 5 4 3 6 7 4 6 6 7 3 5 3 3 6 4 6 6 4 5 7 8 8
## [77473] 5 4 8 3 4 5 4 5 8 4 6 3 4 5 5 6 2 4 5 6 3 4 7 6
## [77497] 5 9 5 4 3 5 5 6 4 5 5 5 5 7 4 3 5 8 6 7 5 6 6 4
## [77521] 5 8 4 1 7 5 5 5 4 8 7 5 5 6 5 6 3 6 5 1 5 6 5 6
## [77545] 4 7 4 6 5 6 8 6 6 5 5 8 3 3 5 6 5 4 7 6 4 4 2 5
## [77569] 5 5 6 7 7 7 2 6 2 4 6 5 4 4 6 3 8 5 4 5 6 9 5 4
## [77593] 5 7 5 6 8 3 5 3 8 4 6 3 6 5 4 3 3 9 3 5 6 3 4 7
## [77617] 3 6 5 6 5 7 3 6 5 8 4 5 6 7 3 5 5 8 5 4 7 3 5 5
## [77641] 4 6 5 6 6 4 3 3 8 7 8 2 4 5 6 4 4 4 6 2 5 6 7 5
## [77665] 5 4 5 5 4 6 6 6 2 3 6 3 6 7 6 5 4 3 6 5 6 7 7 4
## [77689] 5 5 8 4 5 7 6 4 4 5 5 3 4 4 4 5 7 5 4 2 5 4 6 3
  [77713] 6 6 4 8 6 5 6 5 5 6 3 4 5 5 5 4 3 7 5 5 3 5 3 5
##
  [77737] 5 7 5 3 7 7 3 6 9 9 7 4 4 7 7 5 3 5 6 3 6 4 5 6
##
  [77761] 1 9 5 6 3 8 6 7 7 7 6 5 3 6 4 7 4 5 5 5 6 4 4 6
##
  [77785] 4 2 8 7 6 5 7 5 3 4 6 5 6 5 6 1 4 7 5 7 5 4 3 6
##
##
  [77809] 6 3 5 7 4 5 3 5 2 9 4 3 2 3 5 4 4 3 6 5 7 4 6 6
## [77833] 4 2 8 3 5 4 6 7 4 7 6 4 7 2 5 6 4 5 4 6 4 6 5 5
## [77857] 5 4 5 7 6 5 4 6 5 4 5 6 4 3 3 6 4 6 3 2 5 7 5 4
  [77881] 4 5 5 4 3 7 3 6 3 5 6 4 3 7 5 4 3 6 7 3 6 6 6 5
##
##
  [77905] 4 5 5 4 5 5 2 4 2 3 9 9 3 3 6 4 4 5 5 4 8 8 7 6
## [77929] 3 5 8 5 5 2 4 4 3 5 4 5 2 8 5 6 4 4 5 3 6 4 7 5
## [77953] 6 6 6 4 3 3 5 2 5 3 5 6 8 5 3 5 3 5 6 5 6 5 7 6
  [77977] 8 6 5 6 7 7 5 4 3 4 3 2 6 4 4 5 6 5 9 5 5 7 5 2
## [78001] 6 3 8 8 6 5 4 4 6 5 5 5 5 6 6 3 5 4 7 6 6 6 6 3
## [78025] 5 2 3 2 7 5 5 6 3 6 3 6 5 6 5 5 5 7 4 6 3 5 3
## [78049] 4 4 6 3 2 3 5 4 7 7 5 7 3 7 5 5 5 1 6 6 6 7 8 5
```

```
[85801] 5 2 5 5 7 5 3 6 5 6 6 4 5 3 7 6 7 4 5 4 7 2 3 4
##
   [85825] 4 4 7 2 6 6 4 4 5 3 9 2 5 6 3 6 4 4 3 4 3 4 5 3
##
  [85849] 5 5 7 1 5 4 4 3 5 3 7 7 5 6 3 5 5 8 6 8 6 4
##
  [85873] 3 8 5 4 6 5 6 6 6 5 6 3 3 0 5 5 4 4 5 5 4 6 5 6
##
##
  [85897] 3 5 7 6 6 4 3 8 9 4 8 4 5 7 5 4 7 6 3 4 5 3 3 6
## [85921] 5 6 7 6 3 4 6 7 3 5 6 3 4 4 7 5 4 2 3 6 5 7
                                                       2 4
## [85945] 6 3 5 6 3 4 4 6 8 7 6 6 7 4 6 4 8 3 5 7 6 6 4 6
## [85969] 9 4 7 5 6 4 6 3 5 4 3 3 3 4 7 6 7 8 5 2 4 7 6 4
  [85993] 6 3 6 3 5 7 8 6 5 4 4 3 4 6 3 4 5 5 4 4 3 2 7 6
##
  [86017] 5 2 3 7 4 2 3 2 5 5 0 4 7 4 5 5 4 3 5 7 9 5 5 4
##
  [86041] 6 5 4 6 4 6 5 3 6 5 6 3 4 3 7 4 3 6 4 4 3 5 5 6
  [86065] 6 4 3 5 4 4 4 5 4 6 5 3 5 6 2 5 5 6 7 6 4 7 7 4
##
  [86089] 5 4 4 5 4 5 5 3 2 4 3 3 6 7 6 3 5 4 6 3 4 5 4 5
##
  [86113] 7 3 5 3 4 3 4 4 5 6 6 4 3 7 6 6 4 5 4 5 3 4 5 4
  [86137] 6 5 7 5 4 7 4 7 4 4 8 5 5 5 2 8 9 5 6 5 6 6 4 5
##
  [86161] 5 6 7 6 4 4 7 3 5 4 5 3 4 4 4 8 5 7 6 4 4 5 4 5
##
   [86185] 8 6 6 4 6 5 4 6 1 6 4 5 6 6 3 7 7 4 4 3 4 6 5 4
##
##
  [86209] 4 5 4 5 2 6 4 9 3 6 5 6 3 5 7 6 4 5 5 7 4 5 5 5
##
  [86233] 5 5 6 3 4 7 3 7 5 5 5 4 3 5 6 6 5 5 3 3 3 4 3 4
  [86257] 7 6 6 7 6 3 6 7 5 8 6 4 6 3 3 6 6 6 4 7 3 5 4 5
##
##
  [86281] 7 3 6 6 6 5 3 6 5 5 4 6 6 7 3 3 8 5 5 3 8 4 2 5
  [86305] 5 5 7 1 4 2 4 5 6 4 5 5 5 5 5 4 5 3 4 6 5 3 5 6
##
## [86329] 2 5 6 5 7 5 3 6 5 2 5 8 4 9 4 6 4 6 4 4 5 4 3 2
  [86353] 5 4 4 5 6 6 7 6 6 6 5 4 7 3 5 4 6 5 8 5 7 6 6 4
##
  [86377] 7 1 5 7 4 5 7 5 7 6 8 6 9 5 1 8 8 3 7 6 3 2 7 3
##
##
  [86401] 3 6 4 5 6 6 2 4 5 4 7 5 6 6 7 4 5 9 4 5 3 5 2 4
## [86425] 7 3 4 6 4 5 6 6 5 4 6 6 5 3 4 4 6 4 3 3 4 5 4 6
  [86449] 4 6 7 2 5 3 5 5 3 4 4 3 5 6 8 6 5 5 5 4 4 4 5 2
##
## [86473] 2 6 8 4 7 4 3 3 6 5 4 6 3 3 7 6 4 7 2 7 3 6 5 3
  [86497] 7 5 5 4 6 3 4 7 3 5 8 5 5 3 6 4 3 4 3 7 4 5 7 6
  [86521] 4 7 5 7 5 5 3 4 3 8 7 4 7 4 3 5 6 3 5 7 4 3 3 4
##
##
   [86545] 5 7 5 5 5 8 4 3 4 4 4 6 7 8 5 5 6 8 5 6 2 3 2 3
  [86569] 3 4 5 6 5 3 3 5 5 1 6 4 4 7 6 5 4 6 4 6 4 5 6 3
##
##
  [86593] 3 5 8 4 3 4 4 6 5 8 5 5 4 5 2 4 4 3 4 7 4 4 6 3
  [86617] 4 2 4 6 7 1 6 5 6 8 6 6 6 3 2 3 4 1 2 6 5 3 6 8
##
##
   [86641] 2 5 6 8 5 5 5 5 5 5 4 8 3 5 6 3 6 3 4 5 3 5 5 5 6
##
  [86665] 2 7 5 7 5 5 5 4 3 5 2 3 7 4 4 6 2 4 4 5 8 5 4 6
##
  [86689] 5 4 3 6 7 5 4 6 4 4 8 8 5 5 5 5 5 6 4 5 4 4 6 7
  [86713] 5 4 4 4 7 5 4 6 8 6 6 3 4 6 3 3 6 6 6 5 6 5 5 4
##
##
  [86737] 5 6 6 5 5 2 3 6 6 7 6 5 4 6 5 6 5 8 7 4 3 5 6 6
  [86761] 4 7 7 5 0 4 4 9 7 6 6 7 4 4 4 5 5 5 5 4 7 5 6 6
##
  [86785] 3 6 3 9 4 4 4 6 6 6 3 5 4 5 6 7 3 6 6 5 4 5 5 6
##
##
  [86809] 5 6 4 8 6 8 4 2 2 5 6 5 2 6 7 4 3 5 7 6 4 4 3 6
  [86833] 4 6 7 4 3 5 6 5 6 8 5 8 4 6 5 7 5 6 4 5 4 5 5 4
##
  [86857] 6 3 1 4 6 3 3 8 4 5 3 5 3 3 8 3 5 3 6 6 2 5 5 6
##
## [86881] 3 4 5 5 3 6 4 7 5 3 4 4 2 4 5 4 3 3 4 5 5 5 4 6
```

```
[88009] 7 6 5 3 5 7 4 6 4 4 3 8 3 4 8 8 4 6 3 3 7 4 6 4
##
   [88033] 4 4 6 5 5 4 6 2 5 7 3 7 4 4 4 5 6 7 7 5 2 8 7
##
                                                          5
  [88057] 5 5 6 8 4 7 4 8 7 6 4 6 7 5 5 4 7 7 3 6 4 5 7
##
  [88081] 5 6 5 4 8 7 1 8 5 7 4 3 4 3 7 5 9 5 7 6 5 6 6 5
##
##
  [88105] 6 7 7 4 6 6 3 6 5 5 6 4 5 5 5 5 4 10 5 6 5 7 3 8
##
  [88129] 3 5 6 6 6 2 6 3 5 3 3 7 2 7 5 6 6 5 5 3 5 9 3 4
## [88153] 6 7 5 5 4 5 6 5 5 5 4 3 4 5 4 5 4 8 5 5 2 4 4 7
## [88177] 4 1 5 3 6 7 5 6 6 2 7 7 4 5 7 6 5 6 6 4 3 5 3 4
##
  [88201] 2 6 3 6 7 3 3 3 3 5 5 4 6 6 4 4 6 4 5 6 7 8 5 4
  [88225] 5 5 4 6 3 7 3 7 4 5 5 4 2 7 6 3 6 5 3 6 6 4 4 7
##
  [88249] 6 5 6 1 5 6 3 8 5 3 6 4 5 5 3 2 6 8 5 5 4 8 4 6
  [88273] 6 3 5 4 4 5 3 4 3 5 4 8 4 4 5 7 3 4 4 7 5 6 4 7
##
  [88297] 8 4 4 4 5 7 5 3 1 4 3 2 5 5 5 6 5 4 4 5 7 6 5
##
##
  [88321] 5 4 5 7 3 3 5 3 2 7 9 6 5 7 5 7 3 6 7 6 4 5 6 4
  [88345] 6 4 6 7 4 4 5 2 5 6 5 2 3 4 9 8 5 5 7 5 6 4 5 3
##
  [88369] 3 5 4 4 4 8 5 5 2 4 3 4 3 8 7 5 2 7 4 6 4 5 5 10
##
   [88393] 5 3 4 4 3 5 2 5 5 6 2 5 3 2 5 4 3 4 7 3 4 5 5 4
##
##
  [88417] 5 5 5 4 4 7 4 4 6 7 6 4 6 7 6 3 3 4 6 5 6 6 5 6
##
  [88441] 8 3 4 3 4 4 5 5 3 5 6 4 2 4 9 4 6 7 4 4 4 4 6 4
  [88465] 3 5 5 2 3 2 6 5 7 8 5 5 5 2 6 5 6 5 6 3 4 3 4
##
                                                          2
##
  [88489] 4 3 4 4 5 3 4 4 7 6 7 6 6 9 6 3 7 5 5 6 5 7 5 4
## [88513] 7 3 6 5 2 4 6 5 2 8 6 4 6 5 5 3 7 6 7 5 5 6 3 4
## [88537] 2 4 5 3 4 7 5 6 8 6 4 5 6 5 4 6 3 2 6 6 6 6 3 5
  [88561] 9 5 7 5 4 6 6 6 4 6 4 4 2 6 6 6 6 5 4 4 3 5 7 6
##
##
  [88585] 5 7 4 7 4 5 6 8 3 8 4 8 4 4 3 5 5 4 5 5 1 5 7 6
##
  [88609] 4 4 5 6 5 5 3 8 9 3 5 3 9 3 1 6 4 5 5 4 6 5 8 5
  [88633] 6 8 6 5 5 2 5 5 4 7 5 1 7 5 4 6 5 7 4 2 5 2 5 3
##
  [88657] 5 3 2 4 7 5 4 5 6 3 7 6 7 5 6 5 2 8 7 5 5 4 3 4
##
##
  [88681] 3 8 4 5 4 6 6 6 3 6 5 2 5 8 9 7 4 5 5 6 5 9 9 4
  [88705] 5 6 1 7 4 7 5 4 4 5 7 6 5 4 6 6 5 4 8 5 4 4 4 5
  [88729] 4 8 6 5 8 6 5 4 5 5 3 6 5 7 6 7 9 4 6 3 4 4 7 4
##
##
   [88753] 4 5 4 4 3 4 5 4 7 5 5 5 7 5 4 6 4 6 6 4 1 6
  [88777] 6 7 6 5 4 6 5 5 5 5 4 5 7 5 4 4 4 1 5 7 5 7 5 8
##
  [88801] 7 8 5 7 6 9 2 4 6 7 4 4 3 8 4 5 3 4 7 1 6 4 6 5
##
  [88825] 3 5 4 7 6 8 2 5 5 1 5 3 4 6 5 3 5 3 5 4 3 6 5 3
##
##
   [88849] 5 5 3 7 5 4 3 4 4 6 9 5 5 5 7 7 6 7 7 7 3 4 6 3
##
  [88873] 5 5 3 7 6 4 5 7 6 6 7 6 4 3 5 6 6 3 5 7 6 5 5 6
##
  [88897] 5 3 6 6 5 4 5 2 2 3 7 4 6 4 5 6 7 5 4 3 8 8 5 7
  [88921] 5 7 6 5 3 5 4 5 7 6 4 5 5 7 5 5 4 4 3 3 5 5 7 6
##
##
  [88945] 8 6 5 3 7 4 6 2 6 4 4 4 3 5 5 4 3 6 4 7 8 6 6 7
  [88969] 6 5 4 6 7 4 4 6 5 2 5 4 5 5 6 7 4 7 7 5 3 4 1 5
##
  [88993] 5 3 3 8 5 4 5 5 2 6 5 4 6 4 4 5 7 7 4 5 3 5 4 6
##
##
  [89017] 5 2 8 4 5 4 5 3 5 3 6 5 5 6 5 5 3 2 3 1 2 3 2 6
  [89041] 5 3 3 7 3 2 2 4 4 4 5 7 5 4 2 4 4 5 4 3 6 6 6 8
##
  [89065] 6 3 4 5 3 8 6 7 5 7 5 7 4 4 5 4 6 7 9 5 3 8 3 5
##
## [89089] 4 4 4 5 4 6 5 8 3 6 4 5 3 5 5 4 5 5 4 4 6 5 6 2
```

```
##
  [99049] 3 3 5 5 1 4 5 6 6 3 4 7 5 3 7 3 4 2 4 6 5 4 4 3
   [99073] 2 2 4 7 3 7 5 6 5 3 7 5 4 4 5 6 5 3 5 5 1 5 7 5
##
  [99097] 5 5 5 3 4 6 7 6 4 5 5 5 5 6 8 6 6 2 6 8 2 6 5 5
## [99121] 4 6 6 6 7 3 6 5 6 2 5 5 4 4 9 4 6 4 2 3 5 4 4 5
##
  [99145] 7 6 5 5 5 6 8 6 5 3 6 6 3 3 4 4 7 3 6 6 7 3 5 6
## [99169] 6 7 6 2 3 5 4 3 3 6 5 7 5 6 6 4 5 4 7 4 4 6 8 1
## [99193] 5 5 6 4 4 6 5 4 5 6 7 5 4 4 4 6 3 7 4 7 4 6 4 6
## [99217] 7 7 6 2 6 4 6 7 8 5 5 3 4 7 7 7 2 6 6 6 5 5 9 6
  [99241] 5 3 7 7 5 4 2 5 4 2 3 0 7 8 4 8 5 8 5 5 6 7 5 6
##
  [99265] 4 3 5 3 4 3 4 5 6 4 5 6 6 7 5 5 3 5 7 2 4 4 6 4
##
  [99289] 3 7 3 5 6 5 4 6 6 7 7 7 3 4 3 5 4 5 8 6 5 6 6 5
  [99313] 4 5 6 4 6 3 5 6 6 4 4 8 6 6 8 4 5 6 4 6 5 3 7 5
##
##
   [99337] 5 3 5 3 4 3 4 5 4 5 6 6 6 6 5 5 7 6 5 7 5 7 4 6
##
  [99361] 7 7 6 6 5 6 6 4 5 6 4 7 4 4 5 7 4 6 8 4 5 6 5 1
  [99385] 6 5 6 4 4 5 6 4 4 3 5 4 4 8 6 8 6 5 5 6 3 4 5 3
##
  [99409] 3 4 5 5 7 1 6 5 3 4 5 5 7 3 5 2 5 4 4 2 5 5 6 6
##
   [99433] 5 5 7 7 5 6 7 5 3 5 6 5 6 5 7 4 4 5 4 5 4 5
##
                                                       7 4
##
  [99457] 4 4 5 6 3 3 5 6 5 2 6 4 7 6 5 6 5 7 6 3 3 4 4 7
##
  [99481] 5 4 5 6 6 7 5 6 3 4 3 3 5 4 5 5 5 4 5 5 3 6 4 5
          2 4 6 4 4 5 4 4 1
                            3 3 6 4 6 4 6 6 4 4 5 5 4
##
  [99505]
##
   [99529] 7 7 4 4 1 5 5 5 3 5 6 4 6 5 4 4 6 4 5 6 8 4 5
                                                         6
## [99553] 5 8 7 3 5 4 7 8 4 4 9 8 3 3 3 4 6 3 1 3 5 4 4 8
## [99577] 5 3 3 4 4 6 4 6 1 3 7 4 6 3 1 7 6 7 7 4 6 6 4 6
## [99601] 8 7 9 6 3 5 2 4 5 4 2 5 7 3 7 6 5 7 4 3 5 5 7 6
## [99625] 5 5 7 4 4 4 4 5 5 4 6 4 4 5 6 5 2 5 4 7 5 7 7 4
## [99649] 1 3 6 8 4 7 5 4 5 5 7 4 6 6 5 3 3 4 6 3 8 5 6 8
## [99673] 6 5 5 6 6 2 5 6 8 4 6 3 4 4 3 6 5 4 5 7 7 5 5 5
## [99697] 3 5 4 9 5 3 4 5 5 4 5 5 5 4 5 3 6 7 4 4 6 8 5 7
## [99721] 3 7 6 6 7 5 1 4 5 3 4 5 7 8 2 2 4 4 7 2 4 4 6 7
## [99745] 8 4 4 3 5 3 4 3 7 5 4 8 7 7 6 7 6 4 5 3 8 5 7 5
  [99769] 5 8 5 4 7 7 6 5 5 6 5 4 5 5 5 7 2 4 5 4 2 6 5 7
##
##
   [99793] 1 5 6 5 6 5 6 6 6 3 9 8 4 6 4 2 6 5 7 3 6 7 7 2 5
##
  [99817] 7 5 6 4 3 5 5 6 3 4 6 3 5 4 4 6 5 4 7 5 3 1 6 8
  [99841] 5 5 6 7 5 4 5 4 7 6 7 4 4 4 3 3 5 4 7 9 5 3 4 5
   [99865] 7 2 5 6 5 7
                      3 5 5 5 8 5 4 6 8 3 5 3 2 5 5 5 6 5
##
   [99889] 2 4 4 3 5 3 3 5 5 7 6 5 6 4 4 5 4 5 7 6 5 5 4 4
##
##
  [99913] 7 8 5 5 3 7 6 5 8 3 5 9 6 6 4 7 5 5 3 4 8 6 4 4
## [99937] 3 2 5 3 5 5 3 6 6 5 4 6 5 3 4 5 3 8 5 3 6 4 4 5
## [99961] 5 5 5 5 7 7 4 6 6 4 2 4 3 5 6 3 7 3 6 6 7 5 5 3
##
   [99985] 5 6 4 5 5 3 6 4 6 6 5 4 7 6 3 5
mean(X)
```

[1] 4.99712

var(X)

[1] 2.498577

Exercise: Compute the variance of a set of 100,000 Binomial trials, each of size 10, for different values of the probability of heads, ranging from 0 to 1 (in steps of 0.1). This is equivalent to performing one hundred thousand 5-toss experiments, with different types of biased coins in each experiment. For what value of the binomial probability is the variance maximized? Afterwards, try overlaying the theoretical variance of the binomial distribuiton (size*p*(1-p)) on top of your plot.

Probability Part 2

Probability Catch-up

5.1 Discrete distributions

Take a look at the help pages for rpois (and dpois and ppois), and for rgeom (and dgeom and sample). There is a whole world of R functions we can use to sample random values from probability distributions or obtain probabilities from them! The _pois family of functions allow us to work with Poisson distributions, while the _geom family of functions allow us to work with Geometric distributions.

Be careful with which type of function you use! For example, rpois serves to generate random samples from the Poisson distribution. In turn, dpois allow us to calculate the Poisson PMF (in other words, P[X=k]) for any value of k. Finally, ppois allow us to calculate the "distribution function", which is just a fancy way of saying $P[X \leq k]$ for any value of k.

After having navigated these help pages, let's take a stab at answering some discrete probability questions:

Excercise: Let's assume the number of meteorites hitting a planet follows a Poisson distribution. If meteorites hit the planet at a rate of 5 per year, what is the probability that 3 or less meteorites will hit on a given year?

Exercise: Assuming the same rate as above, what is the probability that more than 10 meteorites hit on a given year?

Exercise: Each hour, 3 buses arrive at a station on average. Simulate a thousand 1-hour instances of this random process, assuming the number of buses per hour follows a Poisson distribution. Compute their sample mean and variance.

Exercise: If I'm tossing a series of unbiased coins, what is the probability I get exactly 5 tails before I get a head?

Exercise: If I'm tossing a series of unbiased coins, what is the probability I get at least 5 tails before I get a head?

Finally, we can use the function sample() to simulate from a discrete uniform distribution. As input, this requires a vector containing the possible values that we can sample from (for example, the numbers 1, 2, 3, 4, 5 and 6 in a six-sided dice), and the number of trials we want to perform. Make sure to set the option replace to TRUE, so that we can sample with replacement from the specified vector.

Exercise: Using the sample() function, throw a fair 16-sided dice a thousand times and compute the sample mean of your results.

5.2 The Normal distribution and the Central Limit Theorem

The Central Limit Theorem states that if $X_1, X_2, ..., X_m$ are random samples from a distribution (any distribution) with mean μ and finite variance σ^2 , and the sample mean is \bar{X} , then, as m grows large, \bar{X} converges in distribution to a Normal distribution with mean μ and variance σ^2/m .

In other words, for large values of m (large sample sizes), we can treat the average of our sample as being drawn from $N(\mu, \sigma^2/\sqrt{m})$.

How can we verify this to be true? Well, one way to do this is to pretend (simulate) we have many sample sets, each composed of a large number of samples, compute their sample mean, and look at the histogram over all the sample means. If the CLT holds, that histogram should look very similar to a Normal distribution. Let's try that here!

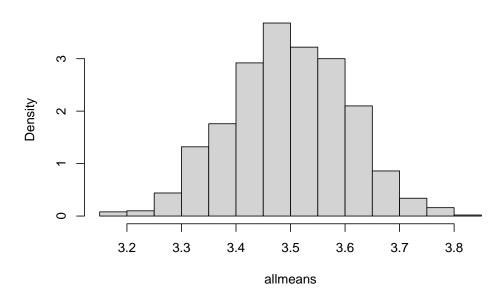
We'll begin with a binomial distribution as our initial "sampling" distribution. We can, for example, draw m = 100 values from a binomial distribution with parameters [p = 0.7, n = 5], and then compute their mean:

```
m <- 100 # sample size
p <- 0.7 # binomial success parameter
n <- 5 # binomial size parameter
samp1 <- rbinom(m,n,p) # simulation
mean1 <- mean(samp1) # sample average</pre>
```

The CLT theorem is a statement about multiple means from multiple samples, so let's repeat the above exercise 1,000 times:

```
allmeans <- sapply(seq(1,1000), function(i){mean(rbinom(m,n,p))})
hist(allmeans,freq=FALSE)</pre>
```

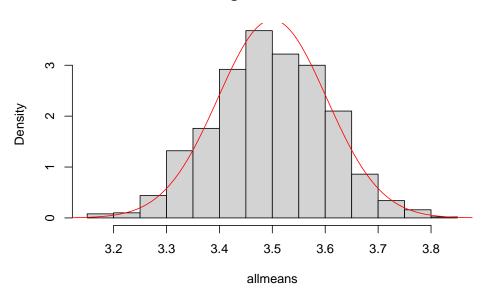
Histogram of allmeans



The CLT states that this distribution should be very close to a Normal distribution with mean $\mu=np=5*0.7=3.5$ and $\sigma^2=(np(1-p))/m=5*0.7*0.3/100=0.0105$. Let's verify that:

```
hist(allmeans, freq=FALSE) # Histogram of sample means
mu <- n*p # mean of Normal distribution under the CLT
var <- n*p*(1-p)/m # variance of Normal distribution under the CLT
sd <- sqrt(var) # standard deviation of Normal distribution under the CLT
curve(dnorm(x, mean=mu, sd=sd), from=-5, to=5, n=5000, add=TRUE, col="red") # Normal distribution
```





Exercise: Repeat the exercise above but instead of a binomial distribution, use a Poisson distribution with parameter $\lambda = 5$ (considering what the mean and variance of the corresponding Normal distribution should be).

Exercise: Repeat the exercise above but instead of a binomial distribution, use a Geometric distribution with parameter p=0.8 (considering what the mean and variance of the corresponding Normal distribution should be).

5.3 The exponential distribution

Previously, we simulated the *quantity* of buses that would randomly arrive at a station over a particular period of time (an hour), given that the *rate* of this process is constant. For this, we used the Poisson distribution with parameter λ equal to this rate. Using the same rate assumption, we can also model the *waiting time* until the next bus arrives, using the exponential distribution. Take a look at the help page for ?rexp.

Exercise: Buses arrive at a station at an average rate of 3 per hour. Using the rexp function, simulate 1,000 waiting times (in hours) of this random process, assuming the waiting time follows an exponential distribution. Create a histogram of these waiting times. Then, calculate their sample mean and compare it to the expected value of an exponential distribution with rate 3.

Exercise: I've arrived at a station. Assuming the same rate as above, use the pexp function to obtain the probability that I will have to wait less than 15 minutes till the next bus arrives.

Exercise: Assuming the same rate as above, use the pexp function to obtain the probability that I will have to wait more than 30 minutes till the next bus arrives.

Exercise: Assuming the same rate as above, use the pexp function to obtain the probability that I will have to wait between 33 and 48 minutes for the next bus to arrive.

Linear Models

We describe linear models in this chapter. First we need to load some libraries (and install them if necessary).

```
if (!require("tidyverse")) install.packages("tidyverse") # Library for data analysis
## Loading required package: tidyverse
## -- Attaching packages ------ tidyverse 1.3.0 --
## v ggplot2 3.3.2 v purrr 0.3.4
## v tibble 3.0.4 v dplyr 1.0.2
## v tidyr 1.1.2 v stringr 1.4.0
## v readr 1.4.0 v forcats 0.5.0
## -- Conflicts ------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()
                  masks stats::lag()
if (!require("stargazer")) install.packages("stargazer") # Library for producing pretty tables of
## Loading required package: stargazer
##
## Please cite as:
## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
if (!require("devtools")) install.packages("devtools")
## Loading required package: devtools
## Loading required package: usethis
```

```
if (!require("report")) devtools::install_github("easystats/report") # Library for pro-
## Loading required package: report
```

report is in alpha - help us improve by reporting bugs on github.com/easystats/report/is

6.1 Fitting a simple linear regression

We'll use a dataset published by Allison and Cicchetti (1976). In this study, the authors studied the relationship between sleep and various ecological and morphological variables across a set of mammalian species: https://science.sciencemag.org/content/194/4266/732

Let's start by loading the data into a table:

```
allisontab <- read.csv("Data_allison.csv")</pre>
```

This dataset contains several variables related to various body measurements and measures of sleep in different species. Note that some of these are continuous, while others are discrete and ordinal.

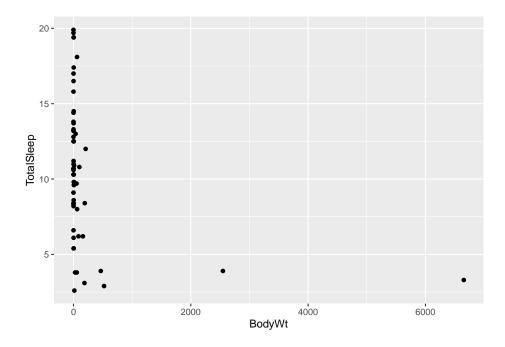
```
summary(allisontab)
```

```
##
     Species
                          BodyWt
                                            BrainWt
                                                           NonDreaming
   Length:62
                     Min.
                                0.005
                                        Min.
                                                  0.14
                                                         Min.
                                                                : 2.100
                                 0.600
                                        1st Qu.:
                                                   4.25
                                                          1st Qu.: 6.250
   Class:character
                     1st Qu.:
   Mode :character
                     Median:
                                 3.342
                                        Median: 17.25
                                                          Median: 8.350
##
                           : 198.790
                                       Mean
                                              : 283.13
                                                         Mean
                                                                : 8.673
##
                   3rd Qu.: 48.202
                                      3rd Qu.: 166.00
                                                        3rd Qu.:11.000
##
                           :6654.000
                   Max.
                                       Max.
                                              :5712.00
                                                         Max.
                                                                :17.900
##
                                                             NA's
##
                        TotalSleep
                                          LifeSpan
                                                             Gestation
       Dreaming
##
    Min.
           :0.000
                     Min.
                            : 2.60
                                     Min.
                                             : 2.000
                                                        Min.
                                                                : 12.00
    1st Qu.:0.900
                     1st Qu.: 8.05
##
                                     1st Qu.: 6.625
                                                        1st Qu.: 35.75
    Median :1.800
                     Median :10.45
                                     Median: 15.100
                                                        Median: 79.00
##
           :1.972
                            :10.53
                                             : 19.878
                                                                :142.35
    Mean
                     Mean
                                     Mean
                                                        Mean
##
    3rd Qu.:2.550
                     3rd Qu.:13.20
                                     3rd Qu.: 27.750
                                                        3rd Qu.:207.50
##
    Max.
           :6.600
                     Max.
                            :19.90
                                     Max.
                                             :100.000
                                                        Max.
                                                                :645.00
##
    NA's
            :12
                     NA's
                             :4
                                       NA's
                                               :4
                                                           NA's
                                                                   :4
##
      Predation
                         Exposure
                                           Danger
##
    Min.
            :1.000
                             :1.000
                                       Min.
                                               :1.000
                     \mathtt{Min}.
##
    1st Qu.:2.000
                     1st Qu.:1.000
                                       1st Qu.:1.000
##
    Median :3.000
                     Median :2.000
                                       Median :2.000
##
    Mean
            :2.871
                     Mean
                             :2.419
                                       Mean
                                               :2.613
    3rd Qu.:4.000
                     3rd Qu.:4.000
##
                                       3rd Qu.:4.000
##
    Max.
            :5.000
                     Max.
                             :5.000
                                       Max.
                                               :5.000
##
```

We'll begin by focusing on the relationship between two of the continuous variables: body size (in kg) and total amount of sleep (in hours). Let's plot these to see what they look like:

```
ggplot(allisontab) + geom_point(aes(x=BodyWt,y=TotalSleep))
```

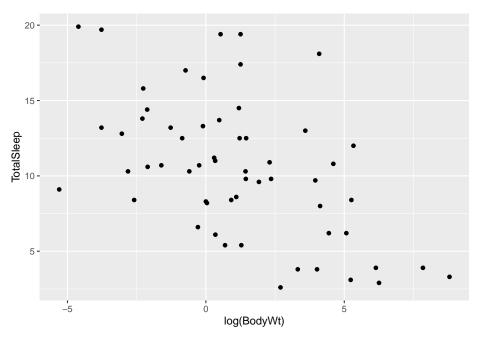
Warning: Removed 4 rows containing missing values (geom_point).



Hmmm this looks weird. We have many measurements of body weight around 0 (small values) and a few very large values of thousands of kilograms. This is not surprising: given that this dataset spans several different species, the measurements spans several orders of magnitude (from elephants to molerats). To account for this, variables involving body measurements (like weight or length) are traditionally converted into a log-scale when fitted into a linear model. Let's see what happens when we log-scale the body weight variable:

```
ggplot(allisontab) + geom_point(aes(x=log(BodyWt),y=TotalSleep))
```

Warning: Removed 4 rows containing missing values (geom_point).



A pattern appears to emerge now. There seems to be a negative correlation between the log of body weight and the amount of sleep a species has. Indeed, we can measure this correlation using the cor() function:

cor(log(allisontab\$BodyWt), allisontab\$TotalSleep, use="complete.obs")

[1] -0.5328345

Let's build a simple linear model to explain total sleep, as a function of body weight. In R, the standard way to fit a linear model is using the function lm(). We do so by following the following formula:

fit <- lm(formula, data)</pre>

The formula within an lm() function for a simple linear regression is:

$$y \sim x_1$$

Where y is the response variable and x_1 is the predictor variable. This formula is a shorthand way that R uses for writing the linear regression formula:

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

In other words, R implicitly knows that each predictor variable will have an associated β coefficient that we're trying to estimate. Note that here y, x_1, ϵ , etc. represent lists (vectors) of variables. We don't need to specify additional terms for the β_0 (intercept) and ϵ (error) terms. The lm() function automatically accounts for the fact that a regression should have an intercept, and that there

will necessarily exist errors (residuals) between our fit and the the observed value of Y.

We can also write this exact same equation by focusing on a single (example) variable, say y_i :

$$y_i = \beta_0 + \beta_1 x_{1,i} + \epsilon_i$$

In general, when we talk about vectors of variables, we'll use boldface, unlike when referring to a single variable.

In our case, we'll attempt to fit total sleep as a function of the log of body weight, plus some noise:

```
myfirstmodel <- lm(TotalSleep ~ log(BodyWt), data=allisontab)
myfirstmodel</pre>
```

```
##
## Call:
## lm(formula = TotalSleep ~ log(BodyWt), data = allisontab)
##
## Coefficients:
## (Intercept) log(BodyWt)
## 11.4377 -0.7931
```

This way, we are fitting the following model:

$$TotalSleep = \beta_0 + \beta_1 log(BodyWt) + \epsilon$$

Remember that the β_0 coefficient is implicitly assumed by the lm() function. We can be more explicit and incorporate it into our equation, by simply adding a value of 1 (a constant). This will result in exactly the same output as before:

```
myfirstmodel <- lm(TotalSleep ~ 1 + log(BodyWt), data=allisontab)
myfirstmodel</pre>
```

```
##
## Call:
## lm(formula = TotalSleep ~ 1 + log(BodyWt), data = allisontab)
##
## Coefficients:
## (Intercept) log(BodyWt)
## 11.4377 -0.7931
```

Exercise: the function attributes() allows us to unpack all the components of the object outputted by the function lm() (and many other objects in R). Try inputting your model output into this function. We can observe that one of the attributes of the object is called coefficients. If we type myfirstmodel\$coefficients, we obtain a vector with the value of our two fitted coefficients (β_0 and β_1). Using the values from this vector, try plotting the

line of best fit on top of the data. Hint: use the geom_abline() function from the ggplot2 library.

6.2 Interpreting a simple linear regression

We can obtain information about our model's fit using the function summary(): summary(myfirstmodel)

```
##
## Call:
## lm(formula = TotalSleep ~ 1 + log(BodyWt), data = allisontab)
## Residuals:
      Min
                10 Median
                                3Q
                                       Max
## -6.6990 -2.6264 -0.2441
                           2.1700 9.9095
##
## Coefficients:
##
               Estimate Std. Error t value Pr(>|t|)
## (Intercept)
               11.4377
                            0.5510 20.759 < 2e-16 ***
## log(BodyWt)
                -0.7931
                            0.1683 -4.712 1.66e-05 ***
## ---
                  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Signif. codes:
## Residual standard error: 3.933 on 56 degrees of freedom
     (4 observations deleted due to missingness)
## Multiple R-squared: 0.2839, Adjusted R-squared: 0.2711
## F-statistic: 22.2 on 1 and 56 DF, p-value: 1.664e-05
```

The summary() function provides a summary of the output of lm() after it's been given some data and a model to fit. Let's pause and analyze the output here. The first line just re-states the formula we have provided to fit our model. Below that, we get a summary (min, max, median, etc.) of all the residuals (error terms) between our linear fit and the observed values of TotalSleep.

Below that, we can see a table with point estimates, standard errors, and a few other properties of our estimated coefficients: the intercept $(\beta_0$, first line) and the slope $(\beta_1$, second line). The standard error is a measure of how confident we are about our point estimate (we'll revisit this in later lectures). The "t value" corresponds to the statistic for a "t-test" which serves to determine whether the estimate can be considered as significantly different from zero. The last column is the P-value from this test. We can see that both estimates are quite significantly different from zero (P < 0.001), meaning we can reject the hypothesis that these estimates are equivalent to zero.

Finally, the last few lines are overall measures of the fit of the model. 'Multiple R-squared' is the fraction of the variance in TotalSleep explained by the fitted

model. Generally, we want this number to be high, but it is possible to have very complex models with very high R-squared but lots of parameters, and therefore we run the risk of "over-fitting" our data. 'Adjusted R-squared' is a modified version of R-squared that attempts to penalize very complex models. The 'residual standard error' is the sum of the squares of the residuals (errors) over all observed data points, scaled by the degrees of freedom of the linear model, which is equal to n - k - 1 where n = total observations and k = total model parameters. Finally, the F-statistic is a test for whether any of the explanatory variables included in the model have a relationship to the outcome. In this case, we only have a single explanatory variable (log(BodyWt)), and so the P-value of this test is simply equal to the P-value of the t-test for the slope of log(BodyWt).

We can use the function report() from the library easystats (https://github.com/easystats/report) to get a more verbose report than the summary() function provides.

```
report(myfirstmodel)
```

```
## Formula contains log- or sqrt-terms. See help("standardize") for how such terms are standardized.
## Formula contains log- or sqrt-terms. See help("standardize") for how such terms are standardized.
## We fitted a linear model (estimated using OLS) to predict TotalSleep with BodyWt (formula: TotalSleep)
```

##
- The effect of BodyWt [log] is significantly negative (beta = -0.79, 95% CI [-1.13, -0.46], t(56)
##

Standardized parameters were obtained by fitting the model on a standardized version of the dataset Note that this function "standardizes" the input variables before providing a

Note that this function "standardizes" the input variables before providing a summary of the output, which makes the estimates' value to be slightly different than those stored in the output of lm(). This makes interpretation of the coefficients easier, as they are now expressed in terms of standard deviations from the mean.

Another way to summarize our output is via a summary table in , which can be easily constructed using the function stargazer() from the library stargazer (https://cran.r-project.org/web/packages/stargazer/index.html).

stargazer(myfirstmodel, type="text")

```
## Constant
                          11.438***
                           (0.551)
##
##
## Observations
                             58
## R2
                            0.284
## Adjusted R2
                            0.271
## Residual Std. Error
                        3.933 (df = 56)
## F Statistic
                    22.203*** (df = 1; 56)
## Note:
                   *p<0.1; **p<0.05; ***p<0.01
```

This package also supports LaTeX and HTML/CSS format (see the type option in ?stargazer), which makes it very handy when copying the output of a regression from R into a working document.

Exercise: try fitting a linear model for TotalSleep as a function of brain weight (BrainWt). Keep in mind that this is a size measurement that might span multiple orders of magnitude, just like body weight. What are the estimated slope and intercept coefficients? Which coefficients are significantly different from zero? What is the proportion of explained variance? How does this compare to our previous model including BodyWt?

Exercise: Plot the linear regression line of the above exercise on top of your data.

6.3 Simulating data from a linear model

It is often useful to simulate data from a model to understand how its parameters relate to features of the data, and to see what happens when we change those parameters. We will now create a function that can simulate data from a simple linear model. We will then feed this function different values of the parameters, and see what the data simulated under a given model looks like.

Let's start by first creating the simulation function. We'll simulate data from a linear model. The model simulation function needs to be told: 1) The number (n) of data points we will simulate 1) How the explanatory variables are distributed: we'll use a normal distribution to specify this. 2) What the intercept (β_0) and slope (β_1) for the linear relationship between the explanatory and response variables are 3) How departures (errors) from linearity for the response variables will be modeled: we'll use another normal distribution for that as well, and control the amount of error using a variable called sigma.res. We'll assume errors are homoscedastic (have the same variance) in this exercise.

```
linearmodsim <- function(n=2, beta_0=0, beta_1=1, sigma.res=1, mu.explan=5, sigma.expl
# Simulate explanatory variables
explan <- r_explan(n,mu.explan,sigma.explan)</pre>
```

```
# Sort the simulated explanatory values from smallest to largest
explan <- sort(explan)

# Standardize the response variables so that they are mean-centered and scaled by their standed
explan.scaled <- scale(explan)

# OPTIONAL: If errors are heteroscedastic (hetero does not equal 0), then their standard deviate
sdev.err <- sapply(sigma.res + explan.scaled*hetero,max,0)

# Simulate the error values using the above-specified standard deviation
error <- rerror(n,0,sdev.err)

# Simulate response variables via the linear model
response <- beta_0 + beta_1 * explan + error

# Output a table containing the explanatory values and their corresponding response values
cbind(data.frame(explan,response))
}
```

Exercise:

- a) Carefully read the code for the function above. Make sure you understand every step in the function.
- b) Plot the output of a simulated linear model with 40 data points, an intercept of 1.5 and a slope of 3. Simulate from the same model one more time, and plot the output again.
- c) Now, fit the data from your latest simulation using the lm() function. Does your fit match your simulations?
- d) Try increasing the sample size (say, to 200 data points), and repeat the lm() fitting. How does this influence the accuracy of your fitted model? Try simulating and fitting multiple times to get an idea of how well you can estimate the parameters.
- e) Try changing the standard deviation of the simulated residual errors (make sigma.res smaller or larger), and repeat the lm() fitting. How does this influence the accuracy of your fitted model?

Properties of Estimators and Inference

7.1 Properties of point estimators

This exercise is partly based on Exercise 6.1 of the book Statistical Rethinking by M.D. Edge. We'll draw many simulated samples from a known distribution with known parameters. We will then pretend this are instances of real datasets and estimate some parameters of the distributions using different estimators. Our goal will be to analyze properties of these parameter estimators, namely their bias and variance. Recall the definitions of these two properties:

Bias:

$$B(\hat{\theta}_n) = E[\hat{\theta}_n(D) - \theta]$$

The bias reflects the average difference between our estimator and the true parameter.

Variance:

$$Var(\hat{\boldsymbol{\theta}}_n) = E[(\hat{\boldsymbol{\theta}}_n(D) - E[\hat{\boldsymbol{\theta}}_n(D)])^2]$$

The variance reflects the variation of the estimator around its own mean (regardless of what the true parameter is).

Note: in real life, we will generally not know the distribution from which our data comes from, or the value of the parameters of that distribution (otherwise we wouldn't be performing statistical inference). When developing or applying a particular estimator to our data, it is thus important to test the properties of that estimator under a variety of simulations and parameter choices. These simulations should aim to, as much as possible, mimic the possible range of processes and parameters that could be generating our data.

We will first create a function that draws a user-specified number (nsim) of independent data samples of size sampsize from a distribution of choice. By default this distribution is the Normal distribution (rnorm), but it can be changed by the user:

```
simsamps <- function(sampsize, nsim = 10000, rsim = rnorm, ...){
   sims <- rsim(sampsize*nsim,...) # generate draws from a specified distribution
   simmat <- matrix(sample(sims), nrow=nsim, ncol=sampsize) # organize those draws into
   return(simmat) # provide the matrix as output
}</pre>
```

The output of this function is a matrix, with row number equal to the number of generated simulations, and column number equal to the size of each simulation. For example, to create 10,000 simulations, each of size 30, from a normal distribution, we would write:

```
testmat <- simsamps(30,10000,rnorm) # generate matrix of 10,000 Normal data sets, each dim(testmat) # dimensions of matrix
```

testmat <- simsamps(30,100,rnorm, mean=3,sd=5) # qenerate matrix of 1000 Normal(3,5) da

```
## [1] 10000 30
```

What type of Normal distribution are we sampling from though? By default, the function rnorm samples from a standard Normal distribution with mean equal to 0 and standard deviation equal to 1. In our function simsamps, we are only feeding one argument to the normal distribution (the number of draws we want to obtain from it). The three dots placed in both the argument of the simsamps function and in its internal call to rsim allows us to feed more parameters to rsim:

We can obtain the sample mean and sample median from each of these datasets, using the apply function:

```
testmean <- apply(testmat,1,mean)
testmedian <- apply(testmat,1,median)</pre>
```

Exercise: Create 20,000 simulated data sets of size 5, each drawn from a Normal distribution with expected value equal to 4 and standard deviation equal to 10. Create a histogram of the sample means from all the data sets. Draw a red line where the expected value of the original source distribution is located. Next, do the same for all the sample medians.

Now, try answering these questions: Is the sample mean a biased estimator of the expected value of a normal distribution? Is the sample median a biased estimator of the expected value of a normal distribution? Which estimator has the highest variance? Which estimator would you use if given a dataset like one of the ones we simulated, in order to estimate the expected value of this Normal distribution?

Exercise: Create 20,000 simulated data sets of size 10, each drawn from an

Exponential distribution with rate equal to 2. Create a histogram of the sample means from all the data sets. Draw a red line where the expected value of the original source distribution is located. Recall that the expected value of this distribution is 1/rate = 0.5.

Now, try answering these questions: Is the sample mean a biased estimator of the expected value of an exponential distribution? Is the sample median a biased estimator of the expected value of an exponential distribution? Which of these two estimators would you use if you were trying to estimate the rate of an exponential process (e.g. the rate at which buses arrive at a station), from a dataset like one of the ones we just simulated?

7.2 Building confidence intervals

This exercise is partly based on Exercise 7.2 of the book Statistical Rethinking by M.D. Edge.

7.3 ROC curve

Frequentist inference

Bayesian Inference

Classification

Model Assessment

Resampling

```
library("tidyverse")
```

12.1 The bootstrap

We'll work with a subset of the Allison et al. data. We'll start by using the body and brain weight measurements from all the species, after log-scaling them. Later on, we'll also use the TotalSleep variable as well, so let's remove any rows that have missing data for any of these 3 variables.

```
# Load table
allisontab <- tibble(read.csv("Data_allison.csv"))
# Remove rows with missing data in columns of interest
allisontab <- filter(allisontab,!is.na(BrainWt) & !is.na(BodyWt) & !is.na(TotalSleep))
# Log-scale body and brain weight
allisontab <- mutate(allisontab,logBody=log10(BodyWt), logBrain=log10(BrainWt))</pre>
```

Below is a function to obtain a single bootstrapped sample from an input data. Take a close look at each step.

```
bootstrap <- function(tab){
    # Preliminary check: if the table is a vector with a single variable, turn it into a matrix
    if(is.null(dim(tab))){tab <- matrix(tab,ncol=1)}

# Count the number of elements in our data
    numelem <- nrow(tab)

# Sample indexes with replacement
    bootsidx <- sample(1:numelem, replace=TRUE)

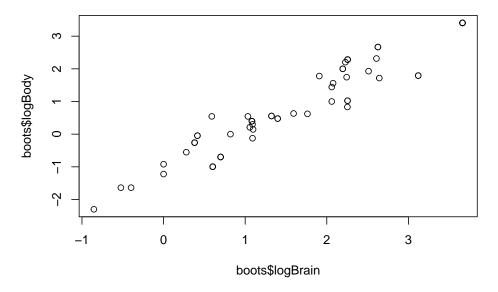
# Obtain a boostrapped sample by selecting the bootstrapped indexes from the original sample
    final <- tab[bootsidx,]

# Produce bootstrapped sample as output</pre>
```

```
return(final)
}
```

Let's see what happens when we run this function on our data.

```
boots <- bootstrap(allisontab)
plot(boots$logBrain,boots$logBody)</pre>
```



Repeat the above command lines multiple times. What happens?

Let's estimate a parameter: the slope coefficient in a linear regression of log brain weight on log body weight:

```
estimate <- lm(logBrain ~ logBody, data=allisontab)$coeff[2]
estimate</pre>
```

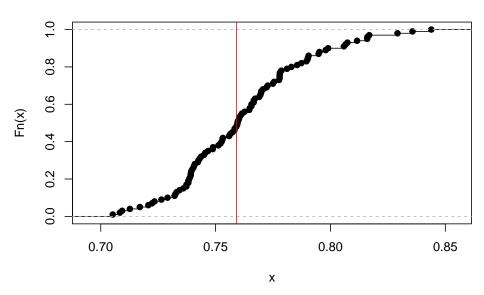
```
## logBody
## 0.7591064
```

Exercise: try estimating the same parameter from a series of 100 bootstrapped samples of our original data, and collecting each of the bootstrapped parameters into a vector called "bootsvec". Hint: you might want to use a for loop or a vectorized sapply() function.

Let's plot the ecdf of all our estimates, using the function ecdf().

```
plot(ecdf(bootsvec))
abline(v=estimate,col="red")
```





We are now ready to obtain confidence intervals (CIs) of our original parameter estimate, using our bootstrapped distribution. There are multiple ways to obtain CIs from a bootstrapped distribution. Some of these assume that the ECDF has particular properties, while others are more generally applicable:

- a) Standard error approach assumes ECDF is normal
- b) Percentile approach assumes ECDF is symmetric and median-unbiased
- c) Pivotal approach most general, makes very few assumptions.

These three approaches generally result in very similar CIs, but they differ (slightly) in methodology. The most widely used method is the pivotal approach, though the motivation for its construction is a bit long-winded. In the interest of time, we'll demonstrate how to run the first two approaches in R. We'll leave the third approach as an exercise you can do at home (read Box 8-1 in the Edge book for an explanation of it, and a code example).

12.2 Permutation test

Let's evaluate the relationship that there is no relationship between logBrain and logBody. Recall that one way to do it would be by using a linear model, and testing whether the value of the fitted slope is significantly different from zero, using a t-test:

```
summary(lm(logBrain ~ logBody, data=allisontab))
##
## Call:
```

lm(formula = logBrain ~ logBody, data = allisontab)

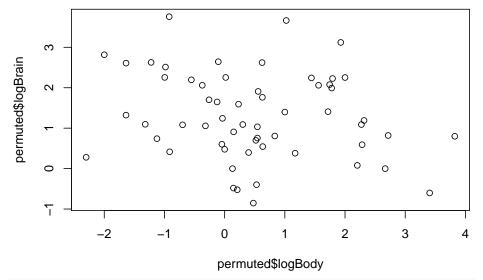
```
##
## Residuals:
##
       Min
                  1Q
                      Median
                                    3Q
                                           Max
## -0.75701 -0.21266 -0.03618 0.19059
                                       0.82489
##
## Coefficients:
##
              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.93507
                          0.04302
                                    21.73
                                             <2e-16 ***
## logBody
               0.75911
                           0.03026
                                    25.09
                                             <2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3071 on 56 degrees of freedom
## Multiple R-squared: 0.9183, Adjusted R-squared: 0.9168
## F-statistic: 629.2 on 1 and 56 DF, p-value: < 2.2e-16
```

This test, however, makes assumptions on our data that sometimes may not be warranted, like large sample sizes and homogeneity of variance. We can perform a more general test that makes less a priori assumptions on our data - a permutation test - as long as we are careful in permuting the appropriate variables for the relationship we are trying to test. In this case, we only have two variables, and we are trying to test whether there is a significant relationship between them. If we randomly shuffle one variable with respect to the other, we should obtain a randomized sample of our data. We can use the following function, which takes in a tibble and a variable of interest, and returns a new tibble in which that particular variable's values are randomly shuffled.

```
permute <- function(tab,vartoshuffle){
    # Extract column we wish to shuffle as a vector
    toshuffle <- unlist(tab[,vartoshuffle],use.names=FALSE)
    # The function sample() serves to randomize the order of elements in a vector
    shuffled <- sample(toshuffle)
    # Replace vector in new table (use !! to refer to a dynamic variable name)
    newtab <- mutate(tab, !!vartoshuffle := shuffled )
    return(newtab)
}</pre>
```

Now we can obtain a permuted version of our original data, and compute the slope estimate on this dataset instead:

```
permuted <- permute(allisontab, "logBrain")
plot(permuted$logBody,permuted$logBrain)</pre>
```



permest <- lm(logBrain ~ logBody, data=permuted)\$coeff[2]
permest</pre>

logBody ## -0.1553871

Exercise: try estimating the same parameter from a series of 100 permuted versions of our original data, and collecting each of the permuted parameters into a vector called "permvec".

We now have a distribution of the parameter estimate under the assumption that there is no relationship between these two variables:

Exercise: obtain an empirical one-tailed P-value from this distribution by counting how many of the permuted samples are as large as our original estimate, and dividing by the total number of permuted samples we have. Note: you should add a 1 to both the denominator and the numerator of this ratio, in case there are no permuted samples that are as large as the original estimate, so as not to get an infinite number.

12.3 Validation

We'll perform a validation exercise to evaluate the error of various models on the data. In this case, we'll create a predictor for TotalSleep as a function of logBody, using a linear model, and then test how well it does. We'll first divide our data into a "training" partition - which we'll use to fit our model - and a separate "test" partition - which we'll use to test how well our model is doing, and avoid over-fitting. Each partition will be one half of our original data.

```
# Obtain the number of data points we have
numdat <- dim(allisontab)[1]
# For the training set, randomly sample 50% of the data indexes
trainset <- sample(numdat, round(numdat*0.5))
# For the test set, obtain all indexes that are not in training set
testset <- seq(1,numdat)[-trainset]</pre>
```

Let's begin by calculating the mean squared error (MSE) between our observations and our predictions in our test partition, after fitting a linear model to our training partition:

```
# Fit the linear model to the training subset of the data
fit1 <- lm(TotalSleep ~ logBody, data=allisontab, subset=trainset)
# Predict all observations using the fitted linear model
predall <- predict(fit1,allisontab)
# Compute mean squared differences between observations and predictions
sqdiff <- (allisontab$logBrain - predall)^2
# Extract the differences for the test partition
sqdiff.test <- sqdiff[testset]
# Compute the mean squared error
mse1 <- mean(sqdiff.test)</pre>
```

Now, we'll try to fit our data to two more complex models: a quadratic model and a cubic model, using the function poly:

```
fit2 <- lm(TotalSleep ~ poly(logBody,2), data=allisontab,subset=trainset)
mse2 <- mean(((allisontab$logBrain - predict(fit2,allisontab))^2)[testset])
fit3 <- lm(TotalSleep ~ poly(logBody,3), data=allisontab,subset=trainset)
mse3 <- mean(((allisontab$logBrain - predict(fit3,allisontab))^2)[testset])</pre>
```

We can see that the MSE appears to increase for the more complex models. This suggests a simple linear fit performs better at predicting values that were not included in the training set.

Exercise: compute the MSE on the training partition. Compare the resulting values to the MSE on the test partition. What do you observe? Is the difference in errors between the three models as large as when computing the MSE on the test partition? Why do you think this is?

12.4 Cross-validation

We'll now perform a cross-validation exercise. If you haven't installed it, you'll need to install the library boot before loading it.

```
if(require("boot") == FALSE){install.packages("boot")}
```

```
## Loading required package: boot
```

```
library("boot")
```

The function cv.glm() from the library boot can be used to compute a cross-validation error. This function is designed to work with the glm() function for fitting generalized linear models in R, but we can compute a simple linear regression using glm() as well, and then feed the result into cv.glm():

```
fit1=glm( TotalSleep ~ logBody, data=allisontab )
# The LOOCV error is computed using the function cv.glm()
cv.err=cv.glm(allisontab,fit1)
```

The value of the cross-validation error is stored in the second element of the attribute delta of the output of cv.glm. By default, this is a "leave-one-out" cross-validation (LOOCV) error, meaning it computes error by leaving 1 data point out of the fitting and evaluating the error at that data point. The process is iterated over all data points, and the errors are then averaged together. We can obtain the value of the LOOCV error by writing:

```
cv.err$delta[1]
```

[1] 15.97798

Now, let's compute the LOOCV error for increasingly complex polynomial models (linear, quadratic, cubic, etc.):

```
CVerr=rep(0,5)
for(m in 1:5){
  fit=glm(TotalSleep ~ poly(logBody,m), data=allisontab)
  CVerr[m]=cv.glm(allisontab,fit)$delta[1]
}
```

Exercise: Plot the results of this cross-validation exercise. Which model has the lowest LOOCV error?

Exercise: Take a look at the help function for cv.glm. Which argument would you modify to be able to compute the 10-fold cross-validation error, instead of the LOOCV error. Can you do this for the five models we tested above?

Chapter 13

Mixed Models

Chapter 14

Ordination

14.1 Libraries and Data

Today, we will work with the package vegan (useful for ordination techniques) and the packages ggplot2 and ggbiplot (useful for fancy plotting). Make sure all these libraries are installed before you begin.

Let's begin by installing and loading the necessary libraries:

```
if (!require("vegan")) install.packages("vegan")
## Loading required package: vegan
## Loading required package: permute
## Attaching package: 'permute'
## The following object is masked _by_ '.GlobalEnv':
##
##
       permute
## The following object is masked from 'package:recipes':
##
##
       check
## The following object is masked from 'package:devtools':
##
##
       check
## Loading required package: lattice
## Attaching package: 'lattice'
```

```
## The following object is masked from 'package:boot':
##
       melanoma
## Registered S3 method overwritten by 'vegan':
                     from
##
     method
##
     print.nullmodel parsnip
## This is vegan 2.5-6
## Attaching package: 'vegan'
## The following object is masked from 'package:parsnip':
##
##
       nullmodel
if (!require("devtools")) install.packages("devtools")
if (!require("ggplot2")) install.packages("ggplot2")
```

We will use a dataset on measurements of particular parts of the iris plant, across individuals from three different species.

```
data(iris)
```

Exercise: Take a look at the iris data matrix. How many samples does it have? How many variables? What happens when you run the function plot() on this matrix? Which variables seem to be strongly correlated? (you can use the function cor() to compute the strength of correlations). Speculate as to why some of these variables could be strongly correlated.

14.2 Principal component analysis (PCA)

We'll perform a PCA of the data. The function prcomp() performs the PCA, and we can assign the result of this function to a new variable (let's call it "fit"). We must first remove the last column to whatever we give as input to prcomp, as the species names are a non-linear (categorical) variable and we don't have (for now) any natural measures of distance for species. The option scale=T standardizes the variables to the same relative scale, so that some variables do not become dominant just because of their large measurement unit. We use

```
irisnumvar <- iris[-5] # Remove the categorical variable
fit<-prcomp(irisnumvar, scale=TRUE) # Perform PCA</pre>
```

Exercise: Try using the summary() and plot() functions to obtain a summary of the resulting PCA object How many principal components were created? (note that the number of PCs always equals the number of original variables). How much variance does the first principal component serve to explain in our

data? How much variance does the second component explain? How many PCs would we need to be able to explain at least 95% of the variation in our data?

The "Rotation" matrix is included inside the fit object we just constructed. You can retrieve it by typing fit[2] \$rotation. This matrix contains the "loadings" of each of the original variables on the newly created PCs.

Exercise: Take a look at the rotation matrix. The larger the absolute value of a variable in each PC, the more that variable contributes to that PC. For each component, use the function barplot() to plot the loadings (contributions) of each variable into that component. Which variables contribute most to each component?

Exercise: Use the function "biplot" to plot the first two PCs of our data. The plotted arrows provide a graphical rendition of the loadings of each of the original variables on the two PCs. Across this reduced dimensional space, we can see that particular variables tend to co-vary quite strongly. Which ones? We can also see a separation into two groups on PC1. Based on the previous exercise (looking at the rotation matrix), which variables do you think would be most different between samples in one group and in the other?

14.3 PCA under the hood

Rather than just using a ready-made function to compute a PCA, let's take a longer route to understand exactly what's happening under the hood of the prcomp() function.

First, let's standardize each column of our data so that each column has mean 0 and variance 1

```
irisdat <- iris[-5]
irisstandard <- apply(irisdat,2,function(x){(x-mean(x))/sd(x)})</pre>
```

Now, calculate the covariance matrix. Because the data has been standardized, this is equivalent to calculating the correlation matrix of the pre-standardized data.

```
cormat <- cov(irisstandard)</pre>
```

Then, extract the eigenvalues and eigenvectors of correlation matrix:

```
myEig <- eigen(cormat)</pre>
```

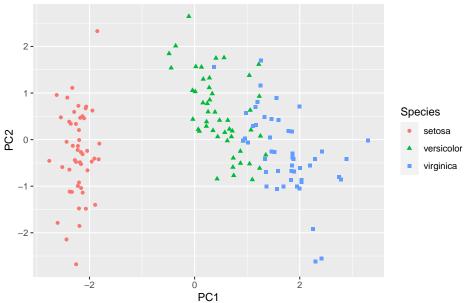
Now, we'll manually obtain certain values that were automatically computed by the prcomp function when we ran it earlier. We'll calculate the singular values (square root of eigenvalues) and also obtain the eigenvectors, also called loadings.

```
sdLONG <- sqrt(myEig$values)
loadingsLONG <- myEig$vectors
rownames(loadingsLONG) <- colnames(irisstandard)</pre>
```

Using the loadings, we can plot our original (standardized) data matrix into the new PC-space, by multiplying the data matrix by the matrix of loadings. Plotting the first two rows of the resulting product should reveal the location of our data points in the first two principal components (like we had before):

```
scoresLONG <- irisstandard %*% loadingsLONG
iristoplot <- data.frame(scoresLONG,iris$Species)
colnames(iristoplot) <- c("PC1","PC2","PC3","PC4","Species")
ggplot(iristoplot, aes(PC1, PC2)) +
  geom_point(aes(color=Species, shape = Species)) +
  xlab("PC1") +
  ylab("PC2") +
  ggtitle("Iris PCA")</pre>
```

Iris PCA



You can compare the results from the first section (using the ready-made function prcomp) and this section (taking a longer road), to check that the results are equivalent. The function range() returns a vector containing the minimum and maximum of a given vector. Using this function, we can observe that the minimum and maximum differences in values for the loadings, the scores and the standard deviations of the PCs are all infinitesimally small (effectively zero).

```
range(fit$sdev - sdLONG)
## [1] -6.661338e-16 2.220446e-16
range(fit$rotation - loadingsLONG)
## [1] -6.661338e-16 7.771561e-16
range(fit$x - scoresLONG)
## [1] -2.359224e-15 3.108624e-15
```

Principal components as explanatory variables

We can use principal components as explanatory variables to any linear model. In this case, we'll use the first two principal components of the PCA we performed above, to perform a logistic regression on the probability that an individual belongs to the species 'virginica'. First, let's create a new variable that is equal to 1 if an individual belongs to this species, and is 0 otherwise. We'll use this variable as the response variable

```
isvirginica <- as.numeric(iris[,5] == "virginica")</pre>
```

We now collate the principal components from the exercise above into a new dataframe that also includes the response variable we just created.

```
# The PC scores are stored in the fifth element of fit. Here, we could have also used the object
PC.scores <- fit[5]
newiris <- data.frame(PC.scores,isvirginica)</pre>
colnames(newiris) <- c("PC1","PC2","PC3","PC4","isvirginica")</pre>
head(newiris)
```

```
##
           PC1
                      PC2
                                  PC3
                                               PC4 isvirginica
## 1 -2.257141 -0.4784238
                           0.12727962
                                       0.024087508
## 2 -2.074013 0.6718827
                           0.23382552
                                       0.102662845
                                                              0
## 3 -2.356335  0.3407664 -0.04405390  0.028282305
                                                              0
## 4 -2.291707 0.5953999 -0.09098530 -0.065735340
                                                              0
## 5 -2.381863 -0.6446757 -0.01568565 -0.035802870
                                                              0
## 6 -2.068701 -1.4842053 -0.02687825 0.006586116
```

Exercise: use the glm() function on the newly created newiris data-frame, to perform a logistic regression for the probability that an individual belongs to the species virginica, using the first principal component (PC1 and PC2) as explanatory variables. Do both components have fitted effects that are significantly different from 0? Do these results make sense in light of the PCA biplots created in the sections above?

4

5

Exercise: Compare the logistic model to another logistic model, this time using only PC1 as the explanatory variable. Which model has the highest AIC score?

14.5 NMDS

We'll now perform non-metric multidimensional scaling. Let's first take a look at the raw data we will use. This is a data matrix containing information about dune meadow vegetation. There are 30 species and 20 sites. Each cell corresponds to the number of specimens of a particular species that has been observed at a particular site (Jongman et al. 1987). As one can see, there are many sites where some species are completely absent (the cell value equals 0):

| data(dune) head(dune) | | | | | | | | | |
|-----------------------|----------|----------|----------|----------|----------|----------|----------|----------|--|
| ## | Achimill | Agrostol | Airaprae | Alopgeni | Anthodor | Bellpere | Bromhord | Chenalbu | |
| ## | 1 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| ## | 2 3 | 0 | 0 | 2 | 0 | 3 | 4 | 0 | |
| ## | 3 0 | 4 | 0 | 7 | 0 | 2 | 0 | 0 | |

| ## | 6 | 2 | 0 | 0 | 0 | 3 | 0 | 0 | 0 |
|----|---|---------|------------------|----------|----------|----------|----------|----------|----------|
| ## | C | irsarve | ${\tt Comapalu}$ | Eleopalu | Elymrepe | Empenigr | Hyporadi | Juncarti | Juncbufo |
| ## | 1 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 |
| ## | 2 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 |
| ## | 3 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 |
| ## | 4 | 2 | 0 | 0 | 4 | 0 | 0 | 0 | 0 |
| ## | 5 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 |

| ## | | Lolipere | Planlanc | Poaprat | Poatriv | Ranuflam | Rumeacet | Sagiproc | Salire |
|----|---|----------|----------|---------|---------|----------|----------|----------|--------|
| ## | 1 | 7 | 0 | 4 | 2 | 0 | 0 | 0 | 0 |
| ## | 2 | 5 | 0 | 4 | 7 | 0 | 0 | 0 | 0 |
| ## | 3 | 6 | 0 | 5 | 6 | 0 | 0 | 0 | 0 |
| ## | 4 | . 5 | 0 | 4 | 5 | 0 | 0 | 5 | 0 |
| ## | 5 | 2 | 5 | 2 | 6 | 0 | 5 | 0 | 0 |
| ## | 6 | 6 | 5 | 3 | Λ | 0 | 6 | Λ | Ο |

| ## | | Scorautu | Trifprat | Trifrepe | ${\tt Vicilath}$ | ${\tt Bracruta}$ | Callcusp |
|----|---|----------|----------|----------|------------------|------------------|----------|
| ## | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| ## | 2 | 5 | 0 | 5 | 0 | 0 | 0 |
| ## | 3 | 2 | 0 | 2 | 0 | 2 | 0 |
| ## | 4 | 2 | 0 | 1 | 0 | 2 | 0 |
| ## | 5 | 3 | 2 | 2 | 0 | 2 | 0 |
| ## | 6 | 3 | 5 | 5 | 0 | 6 | 0 |

Note that here linearity is not a good assumption to make for our ordination: a difference between a site containing 0 specimens (absence) and a site containing

14.5. NMDS 157

1 specimen is conceptually larger than a difference between a site containing 1 specimen and another site containing 2. In other words, the difference between presence and absence is more important than a difference in quantity of specimens. Thus, our first instinct should not be to perform PCA on it. Because NMDS relies on "distances", we need to specify a distance metric that we'll use. The function for performing NMDS in the package 'vegan' is called metaMDS() and its default distance metric is "bray", which corresponds to the Bray-Curtis dissimilarity: a statistic used to quantify the compositional dissimilarity between two different sites, based on counts at each site

Let's perform NMDS ordination using the Bray-Curtis dissimilarity. Remember that, unlike PCA, NMDS requires us to specify the number of dimensions (k) a priori (the default in vegan is 2). It also performs a series of transformations on the data that are appropriate for ecological data (default: autotransform=TRUE). The trymax option ensures that the algorithm is started from different points (in our case, 50) to avoid local minima.

```
ord <- metaMDS(dune, k=2, autotransform = TRUE, trymax=50, distance="bray")
## Run 0 stress 0.1192678
## Run 1 stress 0.1192679
## ... Procrustes: rmse 0.0001966789 max resid 0.0006000853
## ... Similar to previous best
## Run 2 stress 0.1183186
## ... New best solution
## ... Procrustes: rmse 0.0202763 max resid 0.06498442
## Run 3 stress 0.1192678
## Run 4 stress 0.2087025
## Run 5 stress 0.1183186
## ... Procrustes: rmse 0.0001579953 max resid 0.0004779912
## ... Similar to previous best
## Run 6 stress 0.2853324
## Run 7 stress 0.119268
## Run 8 stress 0.1183186
## ... New best solution
## ... Procrustes: rmse 4.360698e-05 max resid 0.0001393975
## ... Similar to previous best
## Run 9 stress 0.1808916
## Run 10 stress 0.1183186
## ... Procrustes: rmse 7.003877e-05 max resid 0.000188974
## ... Similar to previous best
## Run 11 stress 0.1183186
## ... Procrustes: rmse 6.23413e-05 max resid 0.0001900353
## ... Similar to previous best
## Run 12 stress 0.2035424
## Run 13 stress 0.1192688
## Run 14 stress 0.1183186
```

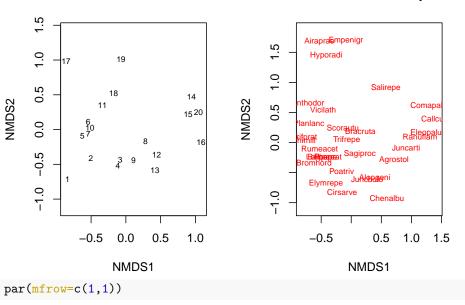
```
## ... New best solution
## ... Procrustes: rmse 1.19404e-05 max resid 2.4161e-05
## ... Similar to previous best
## Run 15 stress 0.1183186
## ... New best solution
## ... Procrustes: rmse 9.399469e-06 max resid 3.209052e-05
## ... Similar to previous best
## Run 16 stress 0.1192683
## Run 17 stress 0.1922248
## Run 18 stress 0.1886532
## Run 19 stress 0.1192679
## Run 20 stress 0.1183186
## ... Procrustes: rmse 6.031738e-05 max resid 0.0001860327
## ... Similar to previous best
## *** Solution reached
```

As you can see, the function goes through a series of steps until convergence is reached. Let's plot the results:

```
par(mfrow=c(1,2))
plot(ord,choices=c(1,2),display="sites",main="NMDS ordination of sites",type="t")
plot(ord,choices=c(1,2),display="species",main="NMDS ordination of species",type="t")
```

NMDS ordination of sites

NMDS ordination of species



Here, the option choices determines which NMDS dimensions are being plotted. We only have two dimensions, so there's only two dimensions that we can plot here. In turn, the option display allows us to plot an ordination of the sites (assuming species are properties of each site), or of the species (assuming sites

14.5. NMDS 159

are properties of each species).

Exercise: Take a look at the plots. Which species tend to co-occcur with each other? Which sites tend to have similar species compositions?

Exercise: Change the number of dimensions and re-run the ordination. Note that you'll have to create multiple plots to observe all the dimensions if there are more than 2 of them. How do the results change?

Exercise: Change the distance metric used and re-run the ordination with k=2 (e.g. try using the Euclidean distance instead). You can take a look at the list of possible distances and their definitions using ?vegdist. Do the results change? Why?

Chapter 15

Clustering

Today, we'll perform a clustering of the iris dataset, using the K-means clustering method.

15.1 Libraries and Data

```
We'll use the following libraries for clustering:
```

```
if (!require("cluster")) install.packages("cluster")

## Loading required package: cluster
if (!require("NbClust")) install.packages("NbClust")

## Loading required package: NbClust
if (!require("tidyverse")) install.packages("tidyverse")

Now, load the iris dataset using the data() function.
data(iris)
```

15.2 Distances

The variables we will use to cluster our data are the four flower measurements in the iris dataset. They all represent the measured length of a segment between two points (e.g. sepal length, petal width), the Euclidean distance is an obvious choice of distance for clustering our observation. The clustering method we will apply to our data (and the ordination methods we applied before) implicitly use the function dist() to calculate distances between observations.

```
d <- dist(iris[,1:4], method="euclidean")
d <- as.matrix(d)</pre>
```

Exercise: Take a look at the first row of this matrix (d[1,]). Under the chosen distance metric, what is the most disparate observation from specimen 1? What is the distance between this specimen and specimen 1? Do these specimens belong to different species? (Hint: the species ID is stored in the origin iris data frame)

15.3 K-means clustering

Ok, we're almost ready to cluster our data. There are just a few preliminary details to keep in mind.

The first of these details is that our variables lie on different scales: some span a wider range of variation than others. A common practice in clustering is to scale all variables in our data such that they are mean-centered and have a standard deviation of 1, before performing the clustering. This ensures that all variables have the same "vote" in the clustering. If we don't scale, then variables that have large amounts of variation (large standard deviations) will disproportionately affect the Euclidean distance, and therefore our clustering will be highly influenced by those variables to the detriment of other variables. This is not inherently wrong, but it is important to keep in mind that unscaled data might result in different clusters than scaled data. To scale our data, we use the function scale().

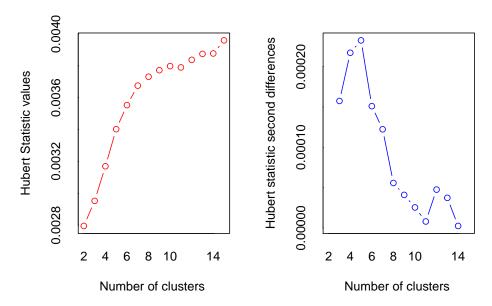
```
df <- scale(iris[,1:4])</pre>
```

Exercise: what's the mean and standard deviation of each of the four variables in the iris dataset before scaling? what is their mean and standard deviation after scaling?

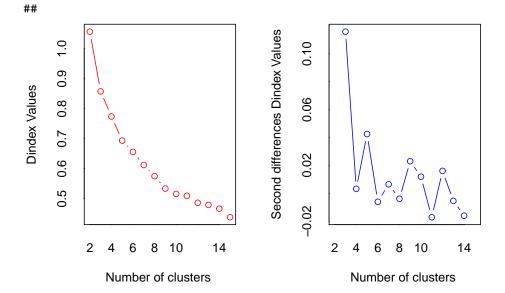
Ok, now on to the second preliminary detail. Clustering methods require the specification of the number of clusters that we a priori choose to fit the data to. Deciding on what is the "best" number of clusters depends on a number of criteria (e.g. minimizes the total within-cluster variance, homogenizing percluster variance, etc), and there are many methods with different criteria. We'll use the function NbClust which runs 26 of these different methods on our data for assessing the "best" number of clusters. We'll then choose to use the number of clusters that is recommended by the largest number of methods.

```
numclust <- NbClust(df, min.nc=2, max.nc=15, distance="euclidean", method="kmeans")</pre>
```

##



***: The Hubert index is a graphical method of determining the number of clusters.
In the plot of Hubert index, we seek a significant knee that corresponds to a
significant increase of the value of the measure i.e the significant peak in Hubert
index second differences plot.



***: The D index is a graphical method of determining the number of clusters.
In the plot of D index, we seek a significant knee (the significant peak in Dindex
second differences plot) that corresponds to a significant increase of the value of
the measure.

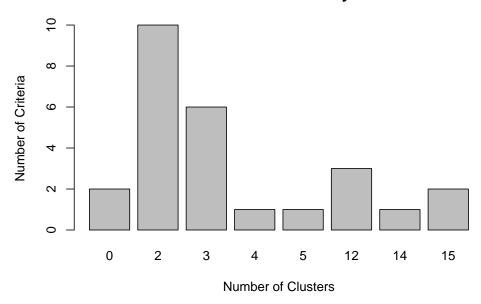
```
## * Among all indices:
## * 10 proposed 2 as the best number of clusters
## * 6 proposed 3 as the best number of clusters
## * 1 proposed 4 as the best number of clusters
## * 1 proposed 5 as the best number of clusters
\#\# * 3 proposed 12 as the best number of clusters
\#\# * 1 proposed 14 as the best number of clusters
## * 2 proposed 15 as the best number of clusters
##
##
                 **** Conclusion ****
##
## * According to the majority rule, the best number of clusters is 2
##
table(numclust$Best.n[1,])
```

```
##
## 0 2 3 4 5 12 14 15
## 2 10 6 1 1 3 1 2
```

It seems like ten of the methods suggest that the best number of clusters should be 2. There is also a considerable (but smaller) number of methods (six), that suggest it should be 3.

```
barplot(table(numclust$Best.n[1,]),
xlab="Number of Clusters", ylab="Number of Criteria", main="Number of Clusters Chosen")
```





Let's go with 2 clusters then. We are now finally ready to perform a K-means clustering. We do so as follows:

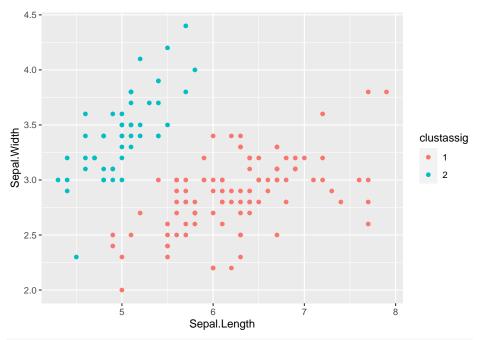
```
fit.kmeans <- kmeans(df, 2)
```

The cluster assignments for all observations are stored in the "cluster" attribute of the resulting object:

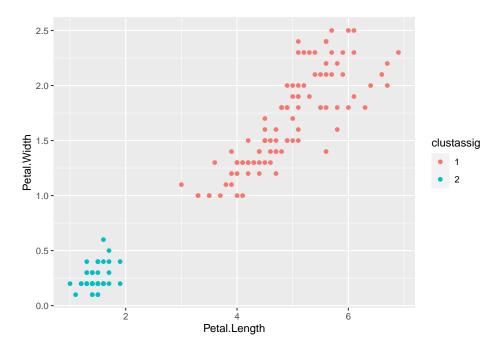
```
clustassig <- fit.kmeans$cluster
clustassig</pre>
```

We can use this vector to plot our data, colored by the resulting clusters.

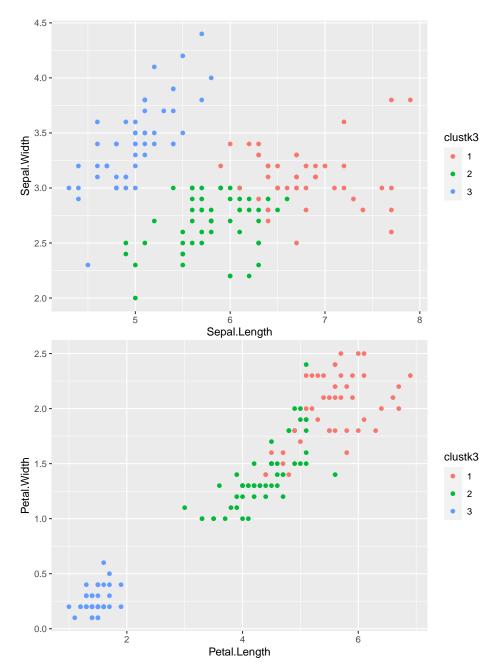
```
clustassig <- as.factor(clustassig) # Treat the cluster assignments as discrete factors
irisclust <- data.frame(iris,clustassig) # Combine data with cluster assignments
ggplot(data=iris) + geom_point(aes(x=Sepal.Length,y=Sepal.Width,color=clustassig)) # Plot sepal assignments</pre>
```



ggplot(data=iris) + geom_point(aes(x=Petal.Length,y=Petal.Width,color=clustassig)) # P



Exercise: Now try computing a K-means clustering on the data yourself, this time using 3 clusters. Plot the result using a different color for each cluster.



Exercise: How well do the clusters from the previous exercise correspond to the 3 iris species? Are they perfectly matched? Why? Why not?

Chapter 16

REcoStats: Linear Models

We describe linear models in this chapter. First we need to load some libraries (and install them if necessary).

```
if (!require("tidyverse")) install.packages("tidyverse") # Library for data analysis
if (!require("stargazer")) install.packages("stargazer") # Library for producing pretty tables of
if (!require("devtools")) install.packages("devtools")
if (!require("report")) devtools::install_github("easystats/report") # Library for producing nice
```

16.1 Fitting a simple linear regression

We'll use a dataset published by Allison and Cicchetti (1976). In this study, the authors studied the relationship between sleep and various ecological and morphological variables across a set of mammalian species: https://science.sciencemag.org/content/194/4266/732

Let's start by loading the data into a table:

```
allisontab <- read.csv("Data_allison.csv")</pre>
```

This dataset contains several variables related to various body measurements and measures of sleep in different species. Note that some of these are continuous, while others are discrete and ordinal.

```
summary(allisontab)
```

```
## Species BodyWt BrainWt NonDreaming
## Length:62 Min.: 0.005 Min.: 0.14 Min.: 2.100
## Class:character 1st Qu.: 0.600 1st Qu.: 4.25 1st Qu.: 6.250
## Mode:character Median: 3.342 Median: 17.25 Median: 8.350
## Mean: 198.790 Mean: 283.13 Mean: 8.673
## 3rd Qu.: 48.202 3rd Qu.: 166.00 3rd Qu.:11.000
```

```
##
                   Max.
                          :6654.000
                                      Max.
                                             :5712.00
                                                        Max.
                                                               :17.900
##
                                                            NA's
                                                                    :14
##
       Dreaming
                       TotalSleep
                                          LifeSpan
                                                            Gestation
                                            : 2.000
           :0.000
                            : 2.60
##
    Min.
                    Min.
                                     Min.
                                                               : 12.00
                                                        Min.
    1st Qu.:0.900
                    1st Qu.: 8.05
##
                                     1st Qu.: 6.625
                                                        1st Qu.: 35.75
##
    Median :1.800
                    Median :10.45
                                     Median: 15.100
                                                        Median: 79.00
##
    Mean
           :1.972
                            :10.53
                                            : 19.878
                                                        Mean
                                                               :142.35
                    Mean
                                     Mean
##
    3rd Qu.:2.550
                    3rd Qu.:13.20
                                     3rd Qu.: 27.750
                                                        3rd Qu.:207.50
           :6.600
                            :19.90
                                                                :645.00
##
    Max.
                    Max.
                                     Max.
                                            :100.000
                                                        Max.
##
    NA's
           :12
                     NA's
                             :4
                                       NA's
                                              :4
                                                          NA's
                                                                  :4
##
      Predation
                        Exposure
                                           Danger
##
    Min.
           :1.000
                     Min.
                             :1.000
                                      Min.
                                              :1.000
##
    1st Qu.:2.000
                     1st Qu.:1.000
                                       1st Qu.:1.000
##
    Median :3.000
                     Median :2.000
                                       Median :2.000
##
    Mean
           :2.871
                     Mean
                             :2.419
                                       Mean
                                              :2.613
                     3rd Qu.:4.000
##
    3rd Qu.:4.000
                                       3rd Qu.:4.000
##
    Max.
            :5.000
                     Max.
                             :5.000
                                       Max.
                                              :5.000
##
```

We'll begin by focusing on the relationship between two of the continuous variables: body size (in kg) and total amount of sleep (in hours). Let's plot these to see what they look like:

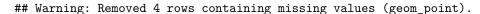
```
ggplot(allisontab) + geom_point(aes(x=BodyWt,y=TotalSleep))
```

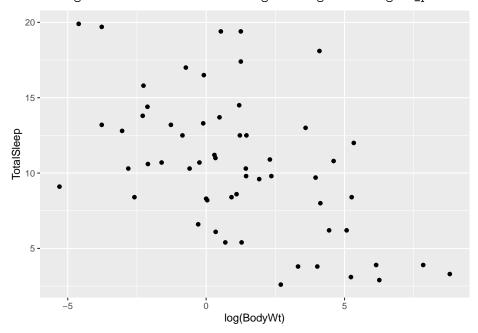
Warning: Removed 4 rows containing missing values (geom_point).



Hmmm this looks weird. We have many measurements of body weight around 0 (small values) and a few very large values of thousands of kilograms. This is not surprising: given that this dataset spans several different species, the measurements spans several orders of magnitude (from elephants to molerats). To account for this, variables involving body measurements (like weight or length) are traditionally converted into a log-scale when fitted into a linear model. Let's see what happens when we log-scale the body weight variable:

ggplot(allisontab) + geom_point(aes(x=log(BodyWt),y=TotalSleep))





A pattern appears to emerge now. There seems to be a negative correlation between the log of body weight and the amount of sleep a species has. Indeed, we can measure this correlation using the cor() function:

cor(log(allisontab\$BodyWt), allisontab\$TotalSleep, use="complete.obs")

[1] -0.5328345

Let's build a simple linear model to explain total sleep, as a function of body weight. In R, the standard way to fit a linear model is using the function lm(). We do so by following the following formula:

fit <- lm(formula, data)</pre>

The formula within an ${\tt lm}()$ function for a simple linear regression is:

Where y is the response variable and x_1 is the predictor variable. This formula is a shorthand way that R uses for writing the linear regression formula:

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

In other words, R implicitly knows that each predictor variable will have an associated β coefficient that we're trying to estimate. Note that here y, x_1, ϵ , etc. represent lists (vectors) of variables. We don't need to specify additional terms for the β_0 (intercept) and ϵ (error) terms. The lm() function automatically accounts for the fact that a regression should have an intercept, and that there will necessarily exist errors (residuals) between our fit and the the observed value of Y.

We can also write this exact same equation by focusing on a single (example) variable, say y_i :

$$y_i = \beta_0 + \beta_1 x_{1,i} + \epsilon_i$$

In general, when we talk about vectors of variables, we'll use boldface, unlike when referring to a single variable.

In our case, we'll attempt to fit total sleep as a function of the log of body weight, plus some noise:

```
myfirstmodel <- lm(TotalSleep ~ log(BodyWt), data=allisontab)
myfirstmodel</pre>
```

```
##
## Call:
## lm(formula = TotalSleep ~ log(BodyWt), data = allisontab)
##
## Coefficients:
## (Intercept) log(BodyWt)
## 11.4377 -0.7931
```

This way, we are fitting the following model:

$$TotalSleep = \beta_0 + \beta_1 log(BodyWt) + \epsilon$$

Remember that the β_0 coefficient is implicitly assumed by the lm() function. We can be more explicit and incorporate it into our equation, by simply adding a value of 1 (a constant). This will result in exactly the same output as before:

```
myfirstmodel <- lm(TotalSleep ~ 1 + log(BodyWt), data=allisontab)
myfirstmodel</pre>
```

```
##
## Call:
## lm(formula = TotalSleep ~ 1 + log(BodyWt), data = allisontab)
```

```
##
## Coefficients:
## (Intercept) log(BodyWt)
## 11.4377 -0.7931
```

Exercise: the function attributes() allows us to unpack all the components of the object outputted by the function lm() (and many other objects in R). Try inputting your model output into this function. We can observe that one of the attributes of the object is called coefficients. If we type myfirstmodel\$coefficients, we obtain a vector with the value of our two fitted coefficients (β_0 and β_1). Using the values from this vector, try plotting the line of best fit on top of the data. Hint: use the geom_abline() function from the ggplot2 library.

16.2 Interpreting a simple linear regression

We can obtain information about our model's fit using the function summary(): summary(myfirstmodel)

```
##
## Call:
## lm(formula = TotalSleep ~ 1 + log(BodyWt), data = allisontab)
##
## Residuals:
##
      Min
               1Q Median
                                30
                                       Max
## -6.6990 -2.6264 -0.2441 2.1700
                                   9.9095
##
## Coefficients:
##
              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.4377
                            0.5510 20.759 < 2e-16 ***
## log(BodyWt) -0.7931
                            0.1683 -4.712 1.66e-05 ***
                  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Signif. codes:
## Residual standard error: 3.933 on 56 degrees of freedom
     (4 observations deleted due to missingness)
## Multiple R-squared: 0.2839, Adjusted R-squared: 0.2711
## F-statistic: 22.2 on 1 and 56 DF, p-value: 1.664e-05
```

The summary() function provides a summary of the output of lm() after it's been given some data and a model to fit. Let's pause and analyze the output here. The first line just re-states the formula we have provided to fit our model. Below that, we get a summary (min, max, median, etc.) of all the residuals (error terms) between our linear fit and the observed values of TotalSleep.

Below that, we can see a table with point estimates, standard errors, and a few

other properties of our estimated coefficients: the intercept (β_0 , first line) and the slope (β_1 , second line). The standard error is a measure of how confident we are about our point estimate (we'll revisit this in later lectures). The "t value" corresponds to the statistic for a "t-test" which serves to determine whether the estimate can be considered as significantly different from zero. The last column is the P-value from this test. We can see that both estimates are quite significantly different from zero (P < 0.001), meaning we can reject the hypothesis that these estimates are equivalent to zero.

Finally, the last few lines are overall measures of the fit of the model. 'Multiple R-squared' is the fraction of the variance in TotalSleep explained by the fitted model. Generally, we want this number to be high, but it is possible to have very complex models with very high R-squared but lots of parameters, and therefore we run the risk of "over-fitting" our data. 'Adjusted R-squared' is a modified version of R-squared that attempts to penalize very complex models. The 'residual standard error' is the sum of the squares of the residuals (errors) over all observed data points, scaled by the degrees of freedom of the linear model, which is equal to n - k - 1 where n = total observations and k = total model parameters. Finally, the F-statistic is a test for whether any of the explanatory variables included in the model have a relationship to the outcome. In this case, we only have a single explanatory variable (log(BodyWt)), and so the P-value of this test is simply equal to the P-value of the t-test for the slope of log(BodyWt).

We can use the function report() from the library easystats (https://github.com/easystats/report) to get a more verbose report than the summary() function provides.

report(myfirstmodel)

Formula contains log- or sqrt-terms. See help("standardize") for how such terms are star ## Formula contains log- or sqrt-terms. See help("standardize") for how such terms are star

We fitted a linear model (estimated using OLS) to predict TotalSleep with BodyWt (formul

- The effect of BodyWt [log] is significantly negative (beta = -0.79, 95% CI [-1.13, -0.44

Standardized parameters were obtained by fitting the model on a standardized version of

Note that this function "standardizes" the input variables before providing a summary of the output, which makes the estimates' value to be slightly different than those stored in the output of lm(). This makes interpretation of the coefficients easier, as they are now expressed in terms of standard deviations from the mean.

Another way to summarize our output is via a summary table in , which can be easily constructed using the function stargazer() from the library stargazer (https://cran.r-project.org/web/packages/stargazer/index.html).

```
stargazer(myfirstmodel, type="text")
##
##
                     Dependent variable:
##
##
                        TotalSleep
## -----
## log(BodyWt)
                         -0.793***
##
                          (0.168)
##
## Constant
                         11.438***
##
                          (0.551)
## Observations
                           58
## R2
                          0.284
## Adjusted R2
                          0.271
                   3.933 (df = 56)
## Residual Std. Error
## F Statistic
                   22.203*** (df = 1; 56)
*p<0.1; **p<0.05; ***p<0.01
## Note:
```

This package also supports LaTeX and HTML/CSS format (see the type option in ?stargazer), which makes it very handy when copying the output of a regression from R into a working document.

Exercise: try fitting a linear model for TotalSleep as a function of brain weight (BrainWt). Keep in mind that this is a size measurement that might span multiple orders of magnitude, just like body weight. What are the estimated slope and intercept coefficients? Which coefficients are significantly different from zero? What is the proportion of explained variance? How does this compare to our previous model including BodyWt?

Exercise: Plot the linear regression line of the above exercise on top of your data.

16.3 Simulating data from a linear model

It is often useful to simulate data from a model to understand how its parameters relate to features of the data, and to see what happens when we change those parameters. We will now create a function that can simulate data from a simple linear model. We will then feed this function different values of the parameters, and see what the data simulated under a given model looks like.

Let's start by first creating the simulation function. We'll simulate data from a

linear model. The model simulation function needs to be told: 1) The number (n) of data points we will simulate 1) How the explanatory variables are distributed: we'll use a normal distribution to specify this. 2) What the intercept (β_0) and slope (β_1) for the linear relationship between the explanatory and response variables are 3) How departures (errors) from linearity for the response variables will be modeled: we'll use another normal distribution for that as well, and control the amount of error using a variable called sigma.res. We'll assume errors are homoscedastic (have the same variance) in this exercise.

```
linearmodsim <- function(n=2, beta_0=0, beta_1=1, sigma.res=1, mu.explan=5, sigma.expl.
# Simulate explanatory variables
explan <- r_explan(n,mu.explan,sigma.explan)
# Sort the simulated explanatory values from smallest to largest
explan <- sort(explan)
# Simulate the error values using the specified standard deviation for the residuals
error <- rerror(n,0,sigma.res)
# Simulate response variables via the linear model
response <- beta_0 + beta_1 * explan + error
# Output a table containing the explanatory values and their corresponding response
cbind(data.frame(explan,response))</pre>
```

Exercise:

- a) Carefully read the code for the function above. Make sure you understand every step in the function.
- b) Plot the output of a simulated linear model with 40 data points, an intercept of 1.5 and a slope of 3. Simulate from the same model one more time, and plot the output again.
- c) Now, fit the data from your latest simulation using the lm() function. Does your fit match your simulations?
- d) Try increasing the sample size (say, to 200 data points), and repeat the lm() fitting. How does this influence the accuracy of your fitted model? Try simulating and fitting multiple times to get an idea of how well you can estimate the parameters.
- e) Try changing the standard deviation of the simulated residual errors (make sigma.res smaller or larger), and repeat the lm() fitting. How does this influence the accuracy of your fitted model?

16.4 Hypothesis testing and permutation testing

Let's evaluate again the hypothesis that there is no relationship between Total-Sleep and log(BodyWt). Recall that one way to do it would be by using a linear model, and testing whether the value of the fitted slope is significantly different from zero, using a t-test:

```
summary(lm(TotalSleep ~ log(BodyWt), data=allisontab))
##
## Call:
```

```
## Call:
## lm(formula = TotalSleep ~ log(BodyWt), data = allisontab)
## Residuals:
##
      Min
               1Q Median
                               3Q
                                      Max
  -6.6990 -2.6264 -0.2441 2.1700 9.9095
##
## Coefficients:
##
              Estimate Std. Error t value Pr(>|t|)
## (Intercept)
               11.4377
                           0.5510 20.759 < 2e-16 ***
## log(BodyWt)
               -0.7931
                           0.1683 -4.712 1.66e-05 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.933 on 56 degrees of freedom
     (4 observations deleted due to missingness)
## Multiple R-squared: 0.2839, Adjusted R-squared: 0.2711
## F-statistic: 22.2 on 1 and 56 DF, p-value: 1.664e-05
```

Take a look at the P-values for the intercept and the slope. If you look at the help page ?summary.lm, you can see that the P-values from these values come from a two-sided t-test. t-tests are usually deployed to compare the means of two populations, or to assess whether the mean of a population has a value specified by a hypothesis. In the case of the slope, for example, we're assessing whether our parameter estimate for the slope has a value specified by the null hypothesis, which in our case is zero. In other words, we're testing whether the value of the slope is consistent with there being no relationship between the two variables (such that if we had an infinite number of data points, their estimated slope would be zero)

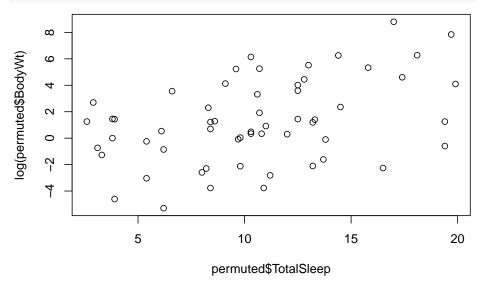
The above t-test makes assumptions on our data that sometimes may not be warranted. Most importantly, the t-test assumes we have a large number of samples, which might not always be the case. We can perform a more robust test that makes less a priori assumptions on our data - a permutation test. To do so, we need to be careful to permute the appropriate variables relevant to the relationship we are trying to test. In this case, we only have two variables

(TotalSleep and log(BodyWt)), and we are trying to test whether there is a significant relationship between them. If we randomly shuffle one variable with respect to the other, we should obtain a randomized sample of our data. We can use the following function, which takes in a tibble and a variable of interest, and returns a new tibble in which that particular variable's values are randomly shuffled.

```
permute <- function(tab,vartoshuffle){
    # Extract column we wish to shuffle as a vector
    toshuffle <- unlist(tab[,vartoshuffle],use.names=FALSE)
    # The function sample() serves to randomize the order of elements in a vector
    shuffled <- sample(toshuffle)
    # Replace vector in new table (use !! to refer to a dynamic variable name)
    newtab <- mutate(tab, !!vartoshuffle := shuffled )
    return(newtab)
}</pre>
```

Now we can obtain a permuted version of our original data, and compute the slope estimate on this dataset instead:

```
permuted <- permute(allisontab, "TotalSleep")
plot(permuted$TotalSleep,log(permuted$BodyWt))</pre>
```



```
permest <- lm(TotalSleep ~ log(BodyWt), data = permuted)$coeff[2]
permest</pre>
```

```
## log(BodyWt)
## 0.6247139
```

Exercise: try estimating the same parameter from a series of 100 permuted versions of our original data, and collecting each of the permuted parameters

into a vector called "permvec".

##

We now have a distribution of the parameter estimate under the assumption that there is no relationship between these two variables:

Exercise: obtain an empirical one-tailed P-value from this distribution by counting how many of the permuted samples are as extreme or more extreme (in the negative or positive direction, than our original estimate, and dividing by the total number of permuted samples we have. Note: you should add a 1 to both the denominator and the numerator of this ratio, in case there are no permuted samples that are as large as the original estimate, so as not to get an infinite number.

The R package coin provides a handy way to apply permutation tests to a wide variety of problems.

```
if (!require("coin")) install.packages("coin")
## Loading required package: coin
## Loading required package: survival
##
## Attaching package: 'survival'
## The following object is masked from 'package:boot':
##
##
       aml
##
## Attaching package: 'coin'
## The following object is masked from 'package:infer':
##
##
       chisq_test
## The following object is masked from 'package:scales':
##
##
       pvalue
library("coin") # Library with pre-written permutation tests
```

The spearman_test() function runs a permutation test of independence between two numeric variables, like the one in the permute() function we coded above. The advantage is that we don't need to actually code the function, we can just run the pre-made function in the coin package directly, as long as we know what type of dependency we're testing. In this case, we perform a test using 1000 permutations (the more permutations, the more exact the test):

Approximative Spearman Correlation Test

```
spearman_test(TotalSleep ~ log(BodyWt), data=allisontab, distribution=approximate(nresample=1000)
##
```

```
##
## data: TotalSleep by log(BodyWt)
## Z = -3.8188, p-value < 0.001
## alternative hypothesis: true rho is not equal to 0</pre>
```

Exercise: Perform a permutation test to assess whether there is a significant relationship between log(BrainWt) and TotalSleep. Compare this to the p-value form the t-statistic testing the same relationship in your fitted linear model.

```
##
##
    Approximative Spearman Correlation Test
##
## data: TotalSleep by log(BrainWt)
## Z = -4.2069, p-value < 0.001
## alternative hypothesis: true rho is not equal to 0
##
## Call:
## lm(formula = TotalSleep ~ log(BrainWt), data = allisontab)
## Residuals:
##
      Min
                1Q Median
                                3Q
                                       Max
## -6.7029 -3.2406 0.0498 2.5162 9.0210
##
## Coefficients:
##
                Estimate Std. Error t value Pr(>|t|)
                 13.7245
                             0.8046 17.058 < 2e-16 ***
## (Intercept)
## log(BrainWt)
                -1.0571
                             0.2076 -5.092 4.3e-06 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.842 on 56 degrees of freedom
     (4 observations deleted due to missingness)
## Multiple R-squared: 0.3165, Adjusted R-squared: 0.3043
## F-statistic: 25.93 on 1 and 56 DF, p-value: 4.297e-06
```

Let's perform a different type of permutation test. In this case, we'll test whether the mean scores of two categories (for example, the math exam scores from two classrooms) are equal to each other.

```
mathscore <- c(80, 114, 90, 110, 116, 114, 128, 110, 124, 130)
classroom <- factor(c(rep("X",5), rep("Y",5)))
scoretab <- data.frame(classroom, mathscore)
```

The standard way to test this is using a t-test, which assumes we have many observations from the two classrooms (do we?) and that these observations come from distributions that have the same variance:

```
t.test(mathscore~classroom, data=scoretab, var.equal=TRUE)

##

## Two Sample t-test

##

## data: mathscore by classroom

## t = -2.345, df = 8, p-value = 0.04705

## alternative hypothesis: true difference in means is not equal to 0

## 95 percent confidence interval:

## -38.081091 -0.318909

## sample estimates:

## mean in group X mean in group Y
```

Exercise: look at the help menu for the oneway_test in the coin package and find a way to carry out the same type of statistical test as above, but using a permutation procedure. Apply it to the scoretab data defined above. Do you see any difference between the P-values from the t-test and the permutation-based test. Why?

121.2

##

102.0