Chapter 2

October 22, 2019

R by Example

Alex Chi

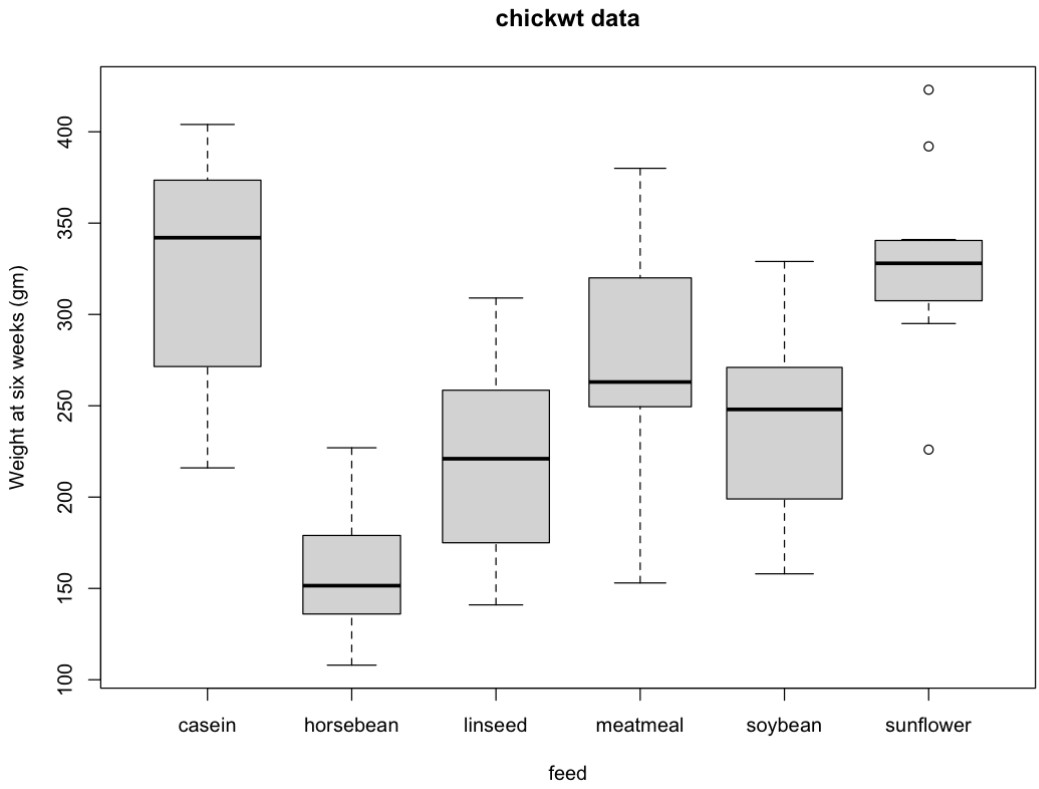
```
In [67]: library(repr)
         options(repr.plot.width=9, repr.plot.height=7)
```

Chapter 2

Exercises

2.1 (chickwts data). The chickwts data are collected from an experiment to compare the effectiveness of various feed supplements on the growth rate of chickens (see ?chickwts). The variables are weight gained by the chicks, and type of feed, a factor. Display side-by-side boxplots of the weights for the six different types of feeds, and interpret.

```
In [68]: boxplot(weight ~ feed, data = chickwts, col = "lightgray",
                 varwidth = TRUE, notch = FALSE, main = "chickwt data",
                 ylab = "Weight at six weeks (gm)")
```



2.2 (iris data). The iris data gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of three species of iris. There are four numeric variables corresponding to the sepal and petal measurements and one factor, Species. Display a table of means by Species (means should be computed separately for each of the three Species).

Note: 在新版 R 语言中，`mean` 不能用于 Dataframe。经过 Google & StackOverflow 查询，发现需要用 `colMeans` 函数替代。

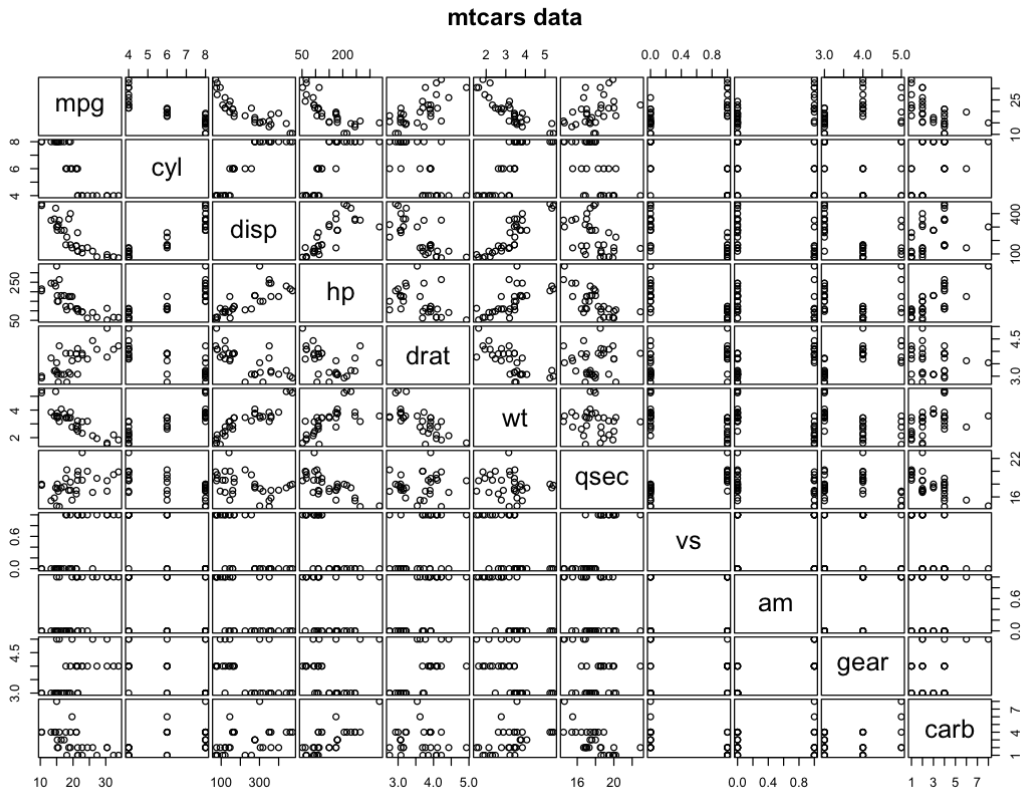
```
In [69]: by(data=iris[1:4], INDICES=iris$Species, FUN=colMeans)
```

```
iris$Species: setosa
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
       5.006        3.428        1.462        0.246
------------------------------------------------------------
iris$Species: versicolor
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
       5.936        2.770        4.260        1.326
------------------------------------------------------------
iris$Species: virginica
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
       6.588        2.974        5.552        2.026
```

2.3 (mtcars data). Display the mtcars data included with R and read the documentation using ?mtcars. Display parallel boxplots of the quantitative variables. Display a pairs plot of the quantitative variables. Does the pairs plot reveal any possible relations between the variables

Answer: mpg 和 disp 似乎有反比关系?

```
In [70]: pairs(mtcars, main = "mtcars data", gap = 1/4)
```



2.4 (mammals data). Refer to Example 2.7. Create a new variable r equal to the ratio of brain size over body size. Using the full mammals data set, order the mammals data by the ratio r. Which mammals have the largest ratios of brain size to body size? Which mammals have the smallest ratios? (Hint: use head and tail on the ordered data.)

```
In [71]: library(MASS)
         mammals$r = mammals$brain / mammals$body
         max_id = which.max(mammals$r)
         mammals[max_id,]
```

A data.frame: 1 × 3

	body	brain	r
	<dbl>	<dbl>	<dbl>
Ground squirrel	0.101	4	39.60396

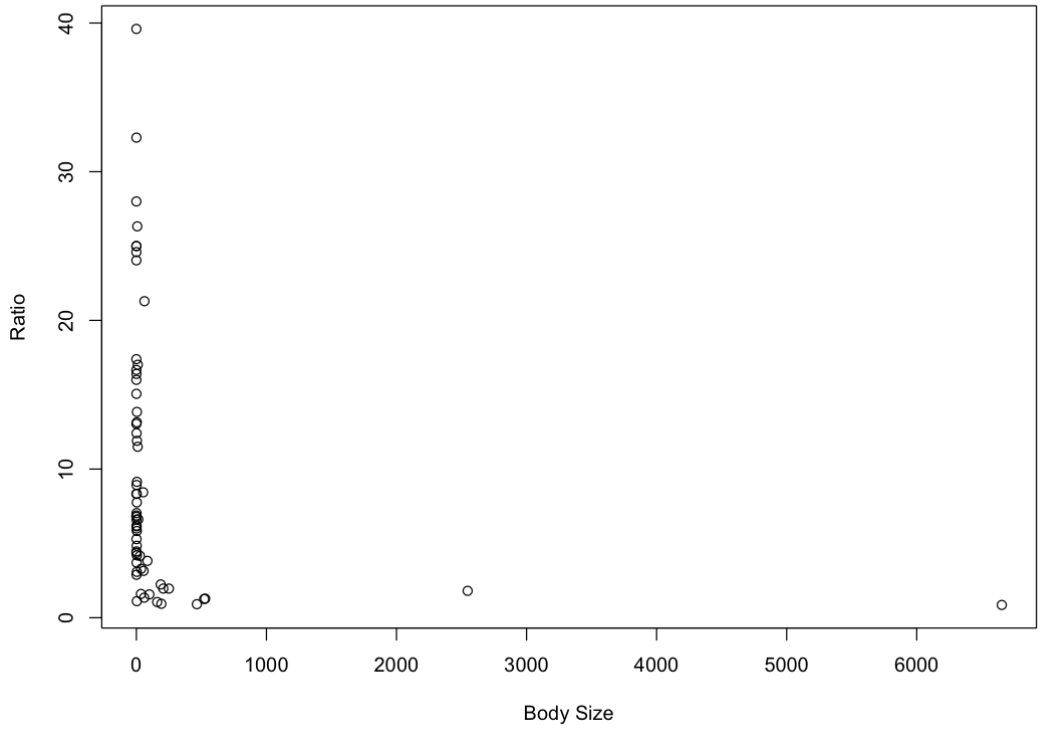
```
In [72]: min_id = which.min(mammals$r)
         mammals[min_id,]
```

A data.frame: 1 × 3

	body	brain	r
	<dbl>	<dbl>	<dbl>
African elephant	6654	5712	0.858431

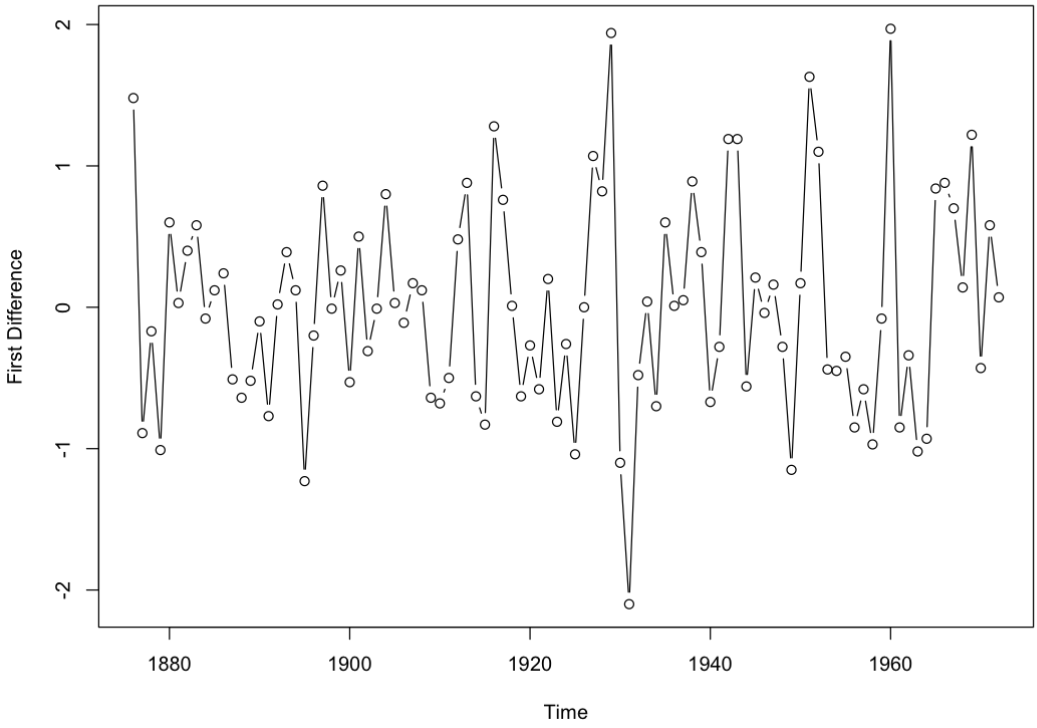
2.5 (mammals data, continued). Refer to Exercise 2.5. Construct a scatterplot of the ratio r = brain/body vs body size for the full mammals data set

```
In [73]: plot(mammals$body, mammals$r, xlab="Body Size", ylab="Ratio")
```



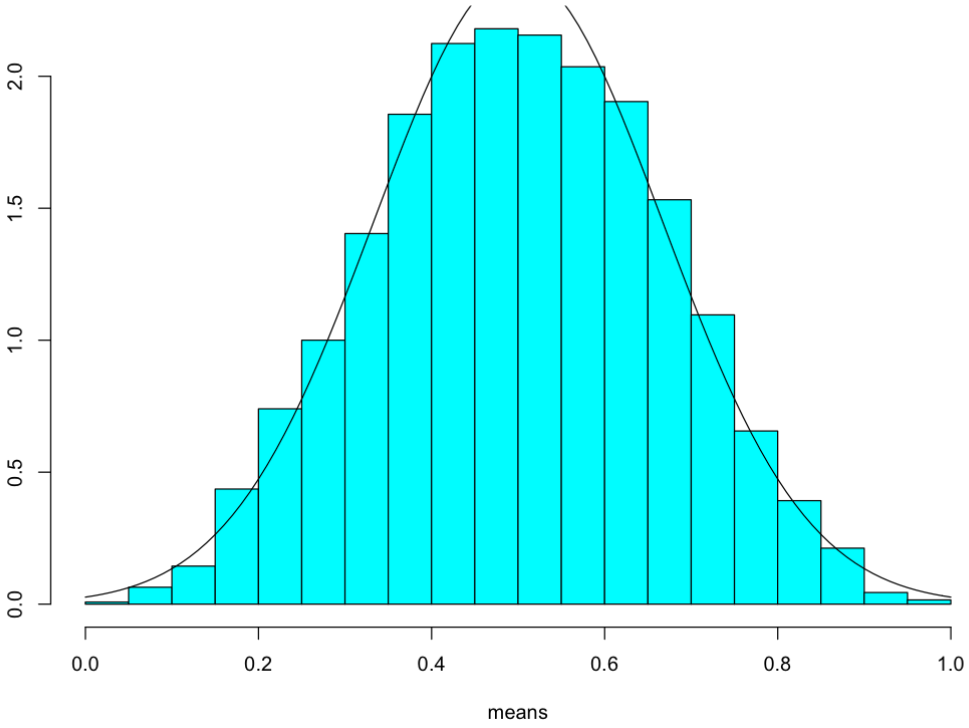
2.6 (LakeHuron data). The LakeHuron data contains annual measure- ments of the level, in feet, of Lake Huron from 1875 through 1972. Display a time plot of the data. Does the average level of the lake appear to be stable or changing with respect to time? Refer to Example 2.4 for a possible method of transforming this series so that the mean is stable, and plot the resulting series. Does the transformation help to stabilize the mean?

```
In [74]: data(LakeHuron)
         plot(diff(LakeHuron), type="b", ylab="First Difference")
```



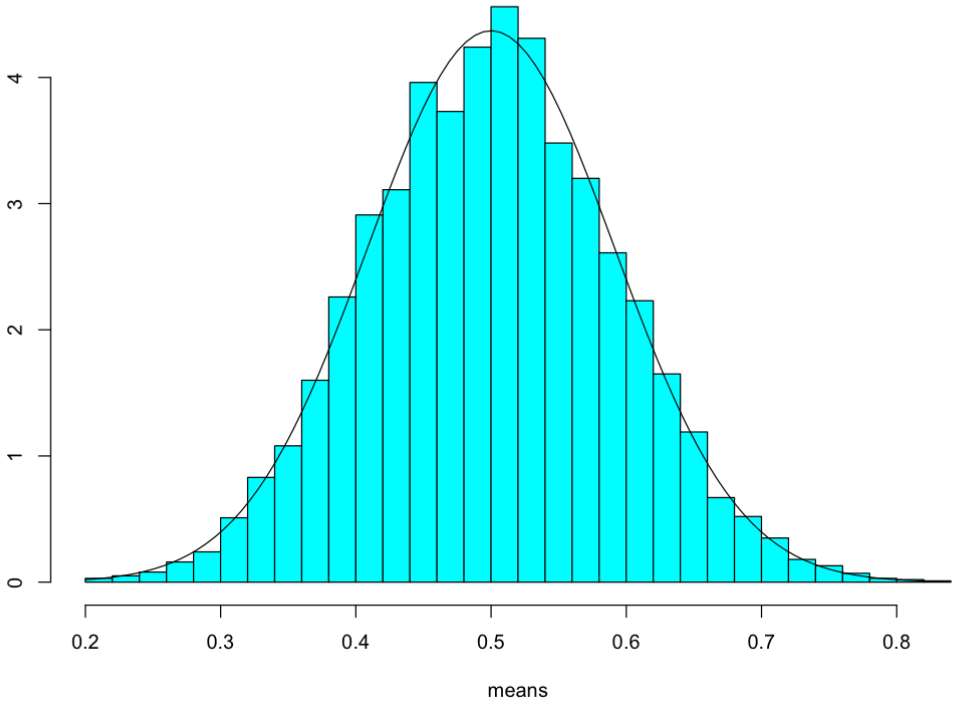
2.7 (Central Limit Theorem with simulated data). Refer to Example 2.6, where we computed sample means for each row of the randu data frame. Repeat the analysis, but instead of randu, create a matrix of random numbers using runif.

```
In [75]: N = 5000
         x = runif(N)
         y = runif(N)
         z = runif(N)
         m = cbind(x, y, z)
         means = apply(m, MARGIN=1, FUN=mean)
         truehist(means, prob=TRUE)
         curve(dnorm(x, 1/2, sd=sqrt(1/36)), add=TRUE)
```



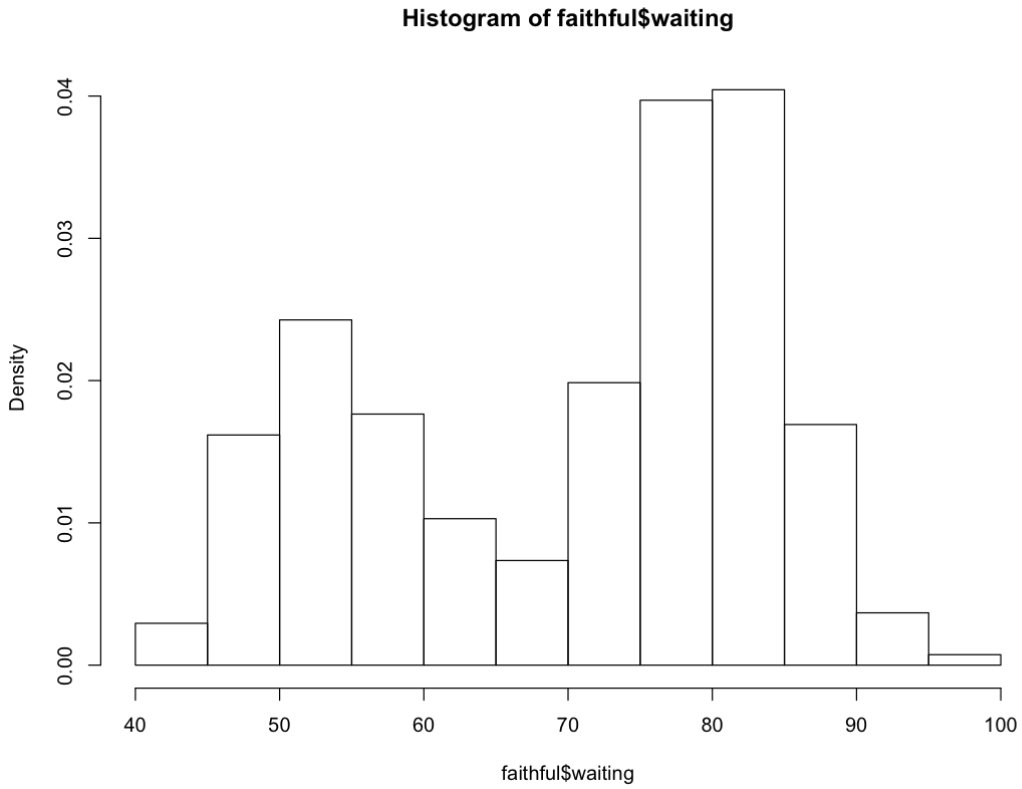
2.8 (Central Limit Theorem, continued). Refer to Example 2.6 and Exercise 2.7, where we computed sample means for each row of the data frame. Repeat the analysis in Exercise 2.7, but instead of sample size 3 generate a matrix that is 400 by 10 (sample size 10). Compare the histogram for sample size 3 and sample size 10. What does the Central Limit Theorem tell us about the distribution of the mean as sample size increases?

```
In [76]: N = 5000
         x = runif(N * 10)
         m = matrix(x, N, 10)
         means = apply(m, MARGIN=1, FUN=mean)
         truehist(means, prob=TRUE)
         curve(dnorm(x, 1/2, sd=sqrt(1/12/10)), add=TRUE)
```

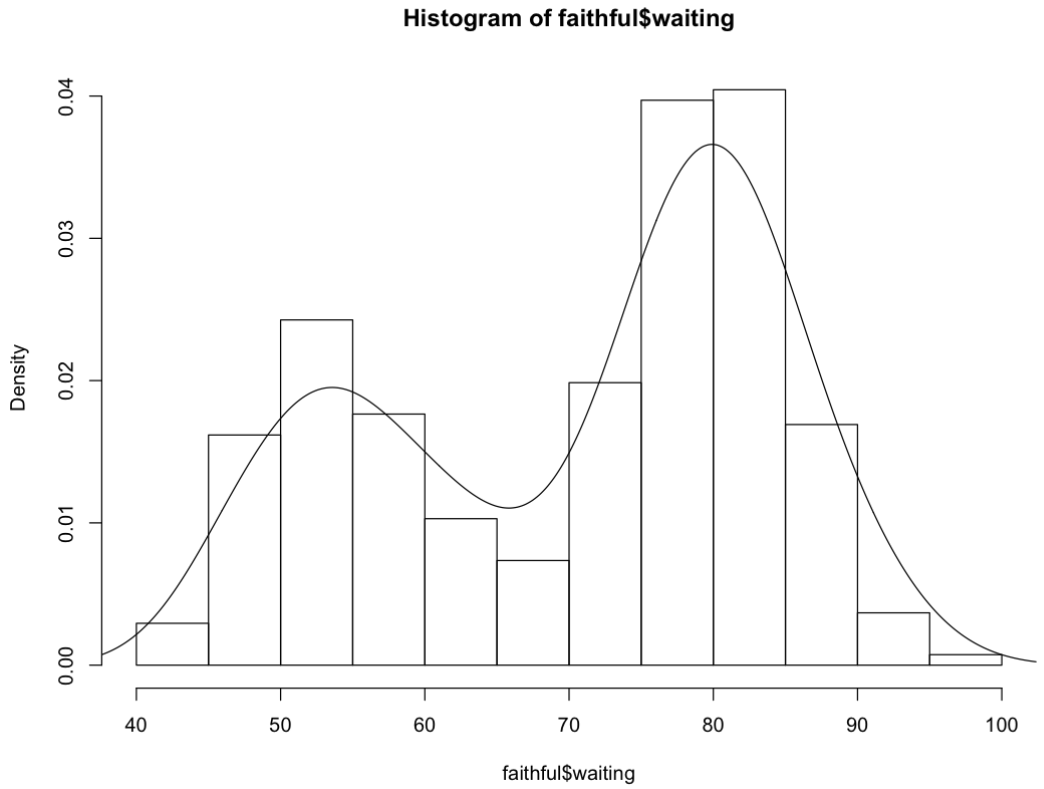


2.10 (“Old Faithful” histogram). Use hist to display a probability his- togram of the waiting times for the Old Faithful geyser in the faithful data set (see Example A.3). (Use the argument prob=TRUE or freq=FALSE.)

```
In [77]: hist(faithful$waiting, prob=TRUE)
```




```
In [78]: hist(faithful$waiting, prob=TRUE)
         lines(density(faithful$waiting))
```



2.12 (Ordering the mammals data by brain size). Refer to Example 2.1. Using the full mammals data set, order the data by brain size. Which mammals have the largest brain sizes? Which mammals have the smallest brain sizes?

```
In [79]: o = order(mammals$brain)
         sorted_mammals = mammals[o, ]
         head(sorted_mammals, 1)
         tail(sorted_mammals, 1)
```

A data.frame: 1 × 3

	body	brain	r
	<dbl>	<dbl>	<dbl>
Lesser short-tailed shrew	0.005	0.14	28

A data.frame: 1 × 3

	body	brain	r
	<dbl>	<dbl>	<dbl>
African elephant	6654	5712	0.858431

2.14 (mammals cluster analysis). Refer to Example 2.10. Repeat the cluster analysis using Ward's minimum variance method instead of near- est neighbor (complete) linkage. Ward's method is implemented in hclust with method="ward" when the first argument is the squared distance matrix. Display a dendrogram and compare the result with the dendrogram for the nearest neighbor method.

```
In [80]: d = dist(log(mammals))
         h = hclust(d, method="ward.D")
         plot(h)
```