1.9 Reports and Reproducible Research

Most data analysis will be summarized in some type of report or article. The process of "copy and paste" for commands and output can lead to errors and omissions. Reproducible research refers to methods of reporting that combine the data analysis, output, graphics, and written report together in such a way that the entire analysis and report can be reproduced by others. Various formats for reports may include word processing documents, LaTeX, or HTML.

The `Sweave` function in R facilitates generating this type of report. There is a LaTeX package (*Sweave*) that generates a .tex file from the Sweave output. Various other packages such as *R2wd* (R to Word), *R2PPT* (R to PowerPoint), *odfWeave* (open document format), *R2HTML* (HTML), can be installed. Commercial packages are also available (e.g. *RTFGen* and *Inference for R*). For more details see the Task View "Reproducible Research" on CRAN.[8]

Exercises

1.1 (Normal percentiles). The `qnorm` function returns the percentiles (quantiles) of a normal distribution. Use the `qnorm` function to find the 95^{th} percentile of the standard normal distribution. Then, use the `qnorm` function to find the quartiles of the standard normal distribution (the quartiles are the 25^{th}, 50^{th}, and 75^{th} percentiles). Hint: Use `c(.25, .5, .75)` as the first argument to `qnorm`.

1.2 (Chi-square density curve). Use the `curve` function to display the graph of the $\chi^2(1)$ density. The chi-square density function is `dchisq`.

1.3 (Gamma densities). Use the `curve` function to display the graph of the gamma density with shape parameter 1 and rate parameter 1. Then use the `curve` function with `add=TRUE` to display the graphs of the gamma density with shape parameter k and rate 1 for $2, 3$, all in the same graphics window. The gamma density function is `dgamma`. Consult the help file `?dgamma` to see how to specify the parameters.

1.4 (Binomial probabilities). Let X be the number of "ones" obtained in 12 rolls of a fair die. Then X has a Binomial($n = 12, p = 1/3$) distribution. Compute a table of binomial probabilities for $x = 0, 1, \ldots, 12$ by two methods:

a. Use the probability density formula

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

[8] http://cran.at.r-project.org/web/views/ReproducibleResearch.html.

and vectorized arithmetic in R. Use `0:12` for the sequence of x values and the `choose` function to compute the binomial coefficients $\binom{n}{k}$.

b. Use the `dbinom` function provided in R and compare your results using both methods.

1.5 (Binomial CDF). Let X be the number of "ones" obtained in 12 rolls of a fair die. Then X has a Binomial($n = 12, p = 1/3$) distribution. Compute a table of cumulative binomial probabilities (the CDF) for $x = 0, 1, \ldots, 12$ by two methods: (1) using `cumsum` and the result of Exercise 1.4, and (2) using the `pbinom` function. What is $P(X > 7)$?

1.6 (Presidents' heights). Refer to Example 1.2 where the heights of the United States Presidents are compared with their main opponent in the presidential election. Create a scatterplot of the loser's height vs the winner's height using the `plot` function. Compare the plot to the more detailed plot shown in the Wikipedia article "Heights of Presidents of the United States and presidential candidates" [54].

1.7 (Simulated "horsekicks" data). The `rpois` function generates random observations from a Poisson distribution. In Example 1.3, we compared the deaths due to horsekicks to a Poisson distribution with mean $\lambda = 0.61$, and in Example 1.4 we simulated random Poisson($\lambda = 0.61$) data. Use the `rpois` function to simulate very large ($n = 1000$ and $n = 10000$) Poisson($\lambda = 0.61$) random samples. Find the frequency distribution, mean and variance for the sample. Compare the theoretical Poisson density with the sample proportions (see Example 1.4).

1.8 (*horsekicks*, continued). Refer to Example 1.3. Using the `ppois` function, compute the cumulative distribution function (CDF) for the Poisson distribution with mean $\lambda = 0.61$, for the values 0 to 4. Compare these probabilities with the empirical CDF. The empirical CDF is the cumulative sum of the sample proportions `p`, which is easily computed using the `cumsum` function. Combine the values of `0:4`, the CDF, and the empirical CDF in a matrix to display these results in a single table.

1.9 (Custom standard deviation function). Write a function `sd.n` similar to the function `var.n` in Example 1.5 that will return the estimate $\hat{\sigma}$ (the square root of $\hat{\sigma}^2$). Try this function on the temperature data of Example 1.1.

1.10 (Euclidean norm function). Write a function `norm` that will compute the Euclidean norm of a numeric vector. The Euclidean norm of a vector $x = (x_1, \ldots, x_n)$ is

$$\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Use vectorized operations to compute the sum. Try this function on the vectors $(0, 0, 0, 1)$ and $(2, 5, 2, 4)$ to check that your function result is correct.

1.11 (Numerical integration). Use the `curve` function to display the graph of the function $f(x) = e^{-x^2}/(1+x^2)$ on the interval $0 \le x \le 10$. Then use the `integrate` function to compute the value of the integral

$$\int_0^\infty \frac{e^{-x^2}}{1+x^2}\,dx.$$

The upper limit at infinity is specified by `upper=Inf` in the `integrate` function.

1.12 (Bivariate normal). Construct a matrix with 10 rows and 2 columns, containing random standard normal data:

```
x = matrix(rnorm(20), 10, 2)
```

This is a random sample of 10 observations from a standard bivariate normal distribution. Use the `apply` function and your `norm` function from Exercise 1.10 to compute the Euclidean norms for each of these 10 observations.

1.13 (*lunatics* data). Obtain a five-number summary for the numeric variables in the *lunatics* data set (see Example 1.12). From the summary we can get an idea about the skewness of variables by comparing the median and the mean population. Which of the distributions are skewed, and in which direction?

1.14 (Tearing factor of paper). The following data describe the tearing factor of paper manufactured under different pressures during pressing. The data is given in Hand et al. [21, Page 4]. Four sheets of paper were selected and tested from each of the five batches manufactured.

Pressure	Tear factor			
35.0	112	119	117	113
49.5	108	99	112	118
70.0	120	106	102	109
99.0	110	101	99	104
140.0	100	102	96	101

Enter this data into an R data frame with two variables: tear factor and pressure. Hint: it may be easiest to enter it into a spreadsheet, and then save it as a tab or comma delimited file (.txt or .csv). There should be 20 observations after a successful import.

1.15 (Vectorized operations). We have seen two examples of vectorized arithmetic in Example 1.1. In the conversion to Celsius, the operations involved one vector `temps` of length four and scalars (32 and 5/9). When we computed differences, the operation involved two vectors `temps` and `CT` of length four. In both cases, the arithmetic operations were applied element by element. What would happen if two vectors of different lengths appear together in an arithmetic expression? Try the following examples using the colon operator `:` to generate a sequence of consecutive integers.

a.

```
x = 1:8
n = 1:2
x + n
```

b.

```
n = 1:3
x + n
```

Explain how the elements of the shorter vector were "recycled" in each case.