ST203: R for Data Science and Statistics

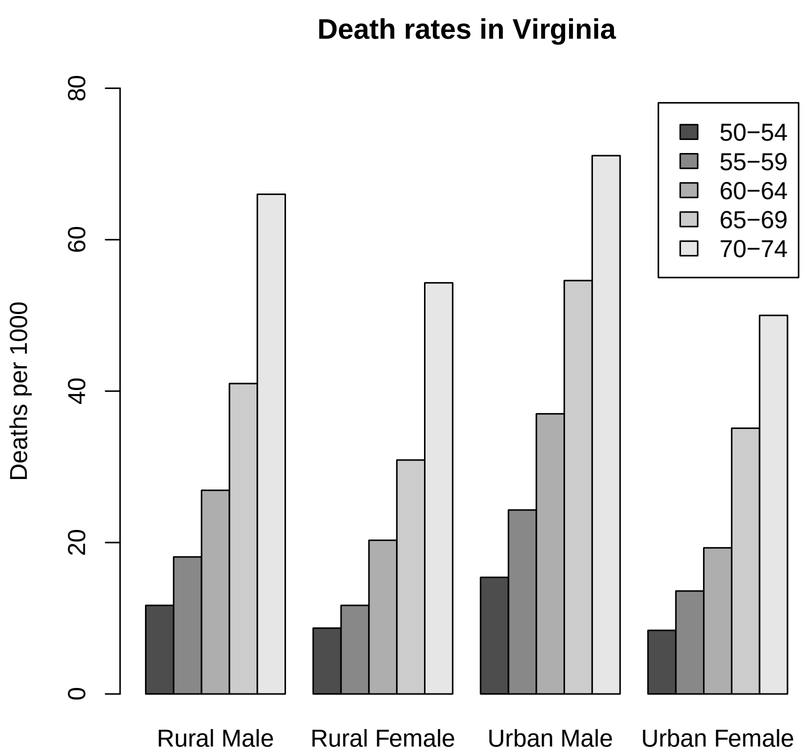
Rafael Moral

Assignment 2 – 2021

- Do all questions: only one randomly chosen question will be marked.
- Upload your script file via Moodle before 23:59 on Friday 12 November.
- You may include your code and your *commented* answers in the same script file.
- You may submit either an R script (‘.R‘) or an R Markdown file (‘.Rmd‘).
- Place your name and student number on the first line of your R script or in the YAML header in your R Markdown file.

Question 1

Look at the help file for the built in data set `VADeaths`. Then use the `barplot` function to produce the following graph.



Question 2

a) Read in the `eupop` dataset, available on Moodle. (You may need to use `setwd` to set the working directory to the folder where you stored the data, and then `read.table` to read in the data.)

b) Draw barcharts, one beside the other, comparing the population breakdowns of Ireland and the UK.

c) Draw back to back barcharts comparing the population breakdowns of Ireland and the UK. Make the Irish bars green and the UK bars blue.

d) Draw divided (stacked) barcharts comparing the population breakdowns of all countries.

Question 3

a) Generate a sample of size 100 from the standard normal distribution. Now generate a sample of size 100 from the binomial distribution with $n =$ `size` $= 20$ and `prob = .25` (look at `?rbinom`).

b) Use `qqplot` to compare the samples. Use `qqnorm` to plot the sample from the binomial distribution. Comment on your results.

c) Generate a sample of size 100 from the binomial distribution with $n =$ `size` $= 20$ and `prob = .5` and repeat part (b).

Question 4

a) Define a function `fbinom` which when given inputs k, n, and p calculates the sum

$$\sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$$

(Hint: use the built-in function `choose`.)

b) Try the function for $p = 0.3$, $k = 5$, and $n = 10$. Then try the function for $p = .3$, $k = 30$, and $n = 100$.

c) Write a second function which compares the result of `fbinom` to the result of the built-in function `pbinom`.

d) Try the function for $p = .3$, $k = 5$ and $n = 10$. Then try the function for $p = .3$, $k = 30$ and $n = 100$.

Question 5

Write a function which simulates arrivals in a supermarket. Suppose that in each 5-minute period the number of arrivals is a Poisson random variable with a mean of 2. (Hint: see `rpois`)

a) Simulate the number of customers in the supermarket over a 3-hour period. You may assume that no one ever leaves this magic supermarket so that the process is completely driven by the number of arrivals.

b) Plot the simulation results for 3 different 3-hour periods. Your `ylim` argument should account for the different ranges and your simulation results should begin from a count of zero at time 0. (Hint: use `max` and `cumsum`).

c) Add appropriate titles and labels, and a legend, to your plot.

Question 6

The geometric mean of a set of positive data is defined as

$$\tilde{x} = \left(\prod_{i=1}^{n} x_i\right)^{1/n} = (x_1 x_2 \ldots x_n)^{1/n}$$

a) Write an R function that calculates the geometric mean, using the built-in R function `prod`.

b) Re-express $\tilde{x}$ using log and exp. Implement this with another R function that calculates the geometric mean.

c) Generate 1000 samples from an exponential distribution with rate $= 1/100$ by executing the following code:

```
# set.seed(your_student_number)
x <- rexp(1000, rate = 1/100)
```

Compare the results from both versions of your function to compute the geometric mean. Can you account for this difference?

d) Using the `replicate` function, write code to collect the differences between the arithmetic mean and the geometric mean (second version) for 1000 random samples of size 100 from the exponential distribution with mean $= 1$.

e) Draw a histogram of the differences and calculate the mean difference.