Data analysis with R final exam 2022

2022-07-04

1.

```
tmp <- c(4, 6, 3)
```

Create the vectors

(a) $(4, 6, 3, 4, 6, 3, \ldots, 4, 6, 3)$ where there are 10 occurrences of 4.

(b) $(4, 4, \ldots, 4, 6, 6, \ldots, 6, 3, 3, \ldots, 3)$ where there are 10 occurrences of 4, 20 occurrences of 6 and 30 occurrences of 3.

2.

Execute the following lines which create two vectors of random integers which are chosen with replacement from the integers $0, 1, \ldots, 999$. Both vectors have length 250.

```
xVec <- sample(0:999, 250, replace=T)
yVec <- sample(0:999, 250, replace=T)
```

(a) Create the vector $(y_2 - x_1, \cdots, y_n - x_{n-1})$.

(b) Pick out the values in yVec which are > 600.

(c) What are the index positions in yVec of the values which are > 600?

(d) Sort the numbers in the vector xVec in the order of increasing values in yVec.

(e) Pick out the elements in yVec at index positions $1, 4, 7, 10, 13, \cdots$

3.

By using the function cumprod and other functions to calculate:

$$1 + \frac{2}{3} + \left(\frac{2}{3}\frac{4}{5}\right) + \left(\frac{2}{3}\frac{4}{5}\frac{6}{7}\right) + \cdots + \left(\frac{2}{3}\frac{4}{5}\cdots\frac{38}{39}\right)$$

4.

For this problem we'll use the (built-in) dataset state.x77.

```
data(state)
state.x77 <- as.data.frame(state.x77)
```

a. Find out how many states have an income of less than 4300.

b. Find out which is the state with the highest income.

c. Add a variable to the data frame which should categorize the level of illiteracy: $[0, 1)$ is low, $[1, 2)$ is some, $[2, \infty)$ is high.

d. Find out which state with low illiteracy, has the highest income, and what that income is.

5.

Simulate 1,000 observations from (X_1, X_2) which follow the uniform distribution over the square $[0, 1] \times [0, 1]$.

a. Get an approximation of the probability that the distance between (X_1, X_2) and the nearest edge is less than 0.25.

b. The same question for the distance to the nearest vertex.

6.

A discrete random variable X has probability mass function

x	0	1	2	3	4
$p(x)$	0.1	0.2	0.2	0.2	0.3

Generate a random sample of size 1000 from the distribution of X using the R `sample()` function. Construct a relative frequency table and compare the empirical with the theoretical probabilities.

7.

Mortality rates per 100,000 from male suicides for a number of age groups and a number of countries are given in the following data frame.

```
suicrates <- tibble(Country = c('Canada', 'Israel', 'Japan', 'Austria', 'France', 'Germany',
'Hungary', 'Italy', 'Netherlands', 'Poland', 'Spain', 'Sweden', 'Switzerland', 'UK', 'USA'),
Age25.34 = c(22,  9, 22, 29, 16, 28, 48,  7,  8, 26,  4, 28, 22, 10, 20),
Age35.44 = c(27, 19, 19, 40, 25, 35, 65,  8, 11, 29,  7, 41, 34, 13, 22),
Age45.54 = c(31, 10, 21, 52, 36, 41, 84, 11, 18, 36, 10, 46, 41, 15, 28),
Age55.64 = c(34, 14, 31, 53, 47, 49, 81, 18, 20, 32, 16, 51, 50, 17, 33),
Age65.74 = c(24, 27, 49, 69, 56, 52, 107, 27, 28, 28, 22, 35, 51, 22, 37))
```

a. Transform `suicrates` into *long* form.

b. Construct side-by-side box plots for the data from different age groups, and comment on what the graphic tells us about the data.

8.

The steam data in the `MASS` package has a nonlinear regression model,

$$P = \alpha \exp\left\{\frac{\beta t}{\gamma + t}\right\} + \varepsilon$$

Fit the model with `nls()` function and find the fitted values, using initial value $a = 5, b = 20, g = 200$. Plot them with the data points on the original scale.

```
library(MASS)
data(steam)
```