EXAM Basic Skills - Part 2: Statistics and Programming

Johnny van Doorn & Abe Hofman
UvA

March 2022

Instructions

- **Before you start**: change the file name to the specified format: `Surname_FirstName_StudentNumber_R_Programming.rmd`
- You can work in a RMarkdown of .R file
- For each question, provide any code you use to answer the question
- Submit your answers through Canvas before 12:30.
- You can use all materials you like, but any form of communication is forbidden
- Time for this part of the exam: 90 min
- The number in front of the question indicates how many points can be obtained (14 total)

Statistics (30 min)

1. **Confidence Intervals.** Means or Medians? Make some R code in which you take a sample of $n = 20$ from a normal density of with $\mu = 0.3$ en $\sigma = 1$. Do this a 1000 times. In each run calculate the mean and the median of the 20 values.

   (a) (1 pt) Use your simulated values to calculate the average and the SD of the two vectors (that contain your simulated means and of the medians). Based on these values calculate the two 95% confidence intervals. What statistic results in the smallest interval?

   (b) (1 pt) Based on these numbers and your statistical intuition, which statistic would you pick if you would want to reject $H_0 : \mu = 0$? The means or the medians?

2. **Regression.** Inspect the code below:

```r
set.seed(10)
p <- numeric(100)
for(i in 1:100) {
  x <- rnorm(20, 2, 2)
  y <- 3 + .15 * x + rnorm(20)
  mod <- lm(y ~ x)
  p[i] <- coefficients(summary(mod))[2,4]
}
table(p < .05)

##
## FALSE  TRUE
##    74    26
```

(a) (1 pt) Carefully study the code. Explain why 'table(p $<$ .05)' represents the power of the regression test.

(b) (1 pt) The p-values refer to the significance level of the slope parameter in the regression model ($b_1 = .15$). What is the null-hypothesis (H0) for the slope parameter? And, is the H0 in this case true? Explain.

(c) (1 pt) Show, using the R-code above, how we can increase the probability of rejecting the H0, if the H0 is false.

Programming in R (60 min)

1. **Matrix.**

   (a) Make the following matrix without typing it out:

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,]    5    6    7    8    9   10   11   12   13    14
##  [2,]   15   16   17   18   19   20   21   22   23    24
##  [3,]   25   26   27   28   29   30   31   32   33    34
##  [4,]   35   36   37   38   39   40   41   42   43    44
##  [5,]   45   46   47   48   49   50   51   52   53    54
##  [6,]   55   56   57   58   59   60   61   62   63    64
##  [7,]   65   66   67   68   69   70   71   72   73    74
##  [8,]   75   76   77   78   79   80   81   82   83    84
##  [9,]   85   86   87   88   89   90   91   92   93    94
## [10,]   95   96   97   98   99  100  101  102  103   104
```

   (b) (1 pt) Select all odd rows, without simply writing out the odd numbers.

   (c) (1 pt) Calculate the median of each column of the whole matrix using an **explicit** loop.

   (d) (1 pt) Do the same with an **implicit** loop.

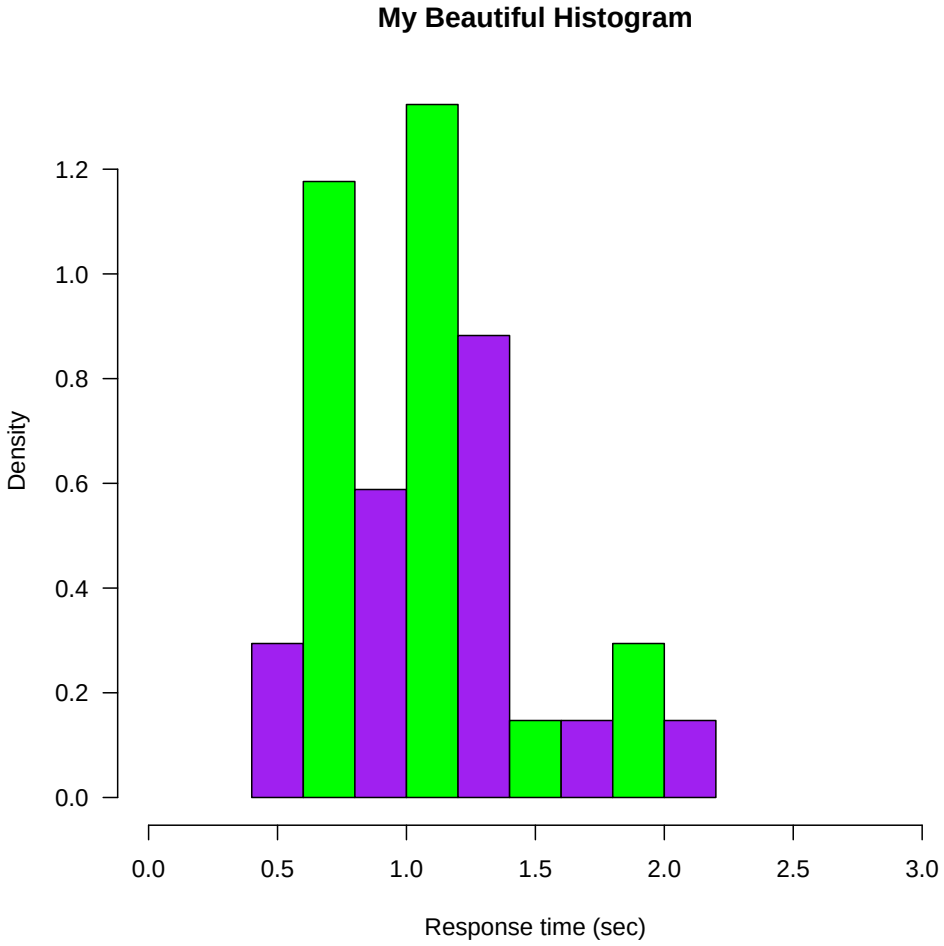
2. **Response Time Analysis.**

   (a) (1 pt) Read in the data file **BasicSkillsExam2022.txt** and store the data in an R object called `dat`.

   (b) (1 pt) What is the mean response time (RT) for each of the conditions? If you did not succeed in the previous question, you can use the code below to generate the data set:

```r
set.seed(167)
dat <- data.frame(pp = rep(1:40),
    condition = rep(letters[1:5], each = 8),
    age = rep(sample(18:40, 10, T), each=4),
    RT  = rlnorm(40, .15, .4),
    ACC = sample(0:1, 40, T))
naPP <- sample(1:40, 6, FALSE)
dat$ACC[naPP] <- dat$RT[naPP] <- NA
```

(c) (1 pt) Reproduce the histogram of the response times below, and write the plot to a `.pdf` (you do not need to include the pdf when you hand in your answers). Note that your results might differ due to sampling variance – be sure to have the same axis limits and graphical parameters. N.B. the histogram contains colors, please see the exam pdf for the color version.



3. **Functions.**

(a) (1 pt) Write a function that takes in a data frame with a single variable and returns a list with the mean and the median of that variable.

(b) (1 pt) Adjust your function such that, if the data frame has more than 1 variable, the function computes all correlations between the variables, and returns the highest correlation.

(c) (1 pt) Make two data frames: (1) a data frame with 1 numeric variable (2) a data frame with 5 numeric variables. You can simulate these variables from the standard normal distribution. Apply your function to both of these data frames.