Examination: Statistical Computing with R, Part I

January 25, 2013

Introduction

The following figure is taken from a paper by social psychologist Dirk Smeesters "The effect of color (red versus blue) on assimilation versus contrast in prime-to-behavior effects". It shows the average scores on a simple multiple choice general knowledge test alleged to have been taken by 169 subjects (psychology students) who were split into 12 experimental groups according to three different experimental conditions. The subsequent table contains the summary statistics of this experiment: average score, sample standard deviation, and size of group. The numerical part of this table can be downloaded from `http://www.math.leidenuniv.nl/~gill/summary.csv`.

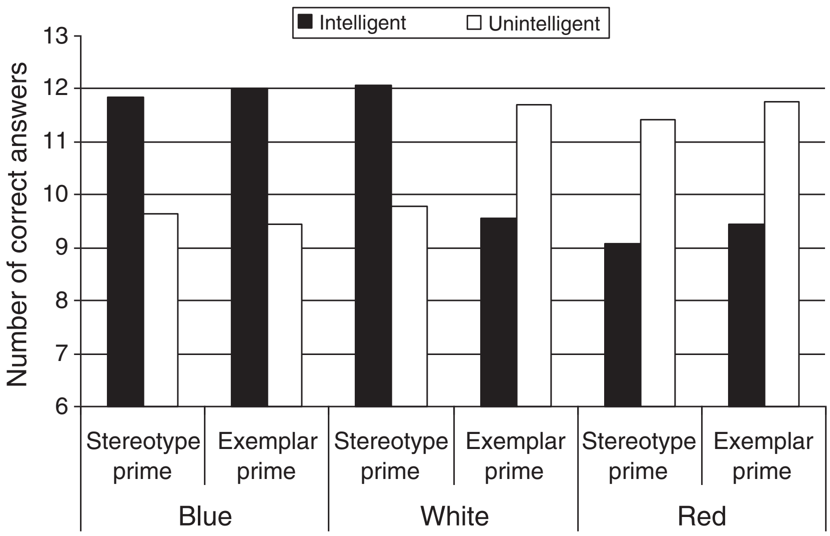

Fig. 1. Number of correct answers as a function of color, prime, and dimension.

The data in this table, and later the raw data from the experiment, has been analysed by fraud buster Uri Simonsohn who was thereby able to obtain strong evidence that the data was actually completely fabricated.

group	means	sds	numbers	colour	prime	dimension
1	11.85	2.66	14	blue	stereotype	intelligent
2	9.64	3.03	14	blue	stereotype	unintelligent
3	12.00	3.37	14	blue	exemplar	intelligent
4	9.43	2.82	14	blue	exemplar	unintelligent
5	12.07	2.78	14	white	stereotype	intelligent
6	9.78	2.66	14	white	stereotype	unintelligent
7	9.56	2.83	16	white	exemplar	intelligent
8	11.71	2.87	14	white	exemplar	unintelligent
9	9.07	2.55	14	red	stereotype	intelligent
10	11.43	2.79	14	red	stereotype	unintelligent
11	9.43	3.06	14	red	exemplar	intelligent
12	11.77	3.03	13	red	exemplar	unintelligent

In this examination project you will construct a complete data-set matching these summary statistics, and use a permutation approach to investigate one of Simonsohn's claims: *the data is too good to be true.* Notice that there are six groups with similar, large average; and six groups with similar, small average score. This was actually the prediction of Smeester's psychological theory and the experiment certainly appears to confirm that. You will see that the variation of the group averages of groups predicted to be similar is much too small, compared to the variation within the groups.

The data concerns 12 experimental groups, defined according to three factors: *colour* (blue, white, red), *prime* (stereotype, exemplar), and *dimension* (intelligent, unintelligent). The sizes of the groups vary (see column "number"). There are 10 groups of size 14, one of size 16, one of size 13, so altogether there were 10*14+16+13=169 subjects in the experiment. For each group we know the sample average ("mean") and sample standard deviation ("sd") on the score of a certain test carried out on each of the subjects in each of the groups.

1 Step 1 (data manipulation).

Read the summary statistics into R and use it to generate a data frame with 12 rows corresponding to the 12 groups, and having, as well as the three numerical variables reported in the table (mean, sd and number), also the three factors *colour*, *prime*, *dimension* with the just mentioned levels.

2 Step 2 (graphics).

Reproduce the barplot Figure 1 within R as well as you are able (but don't spend more than 30 minutes on this part of the exam). The heights of the bars are the values of the sample means in the table, minus a “baseline” value of 6.

Don’t bother to reproduce the annotation *below* the bar plot. But do try to get the labelling of the vertical axis correct.

3 Step 3 (more data manipulation).

Create a data frame with 169 rows corresponding to 169 *subjects* with variables: *colour*, *prime*, *dimension*, *group*, *id*.

The first three variables should be factors, with the levels just defined. The variable *group* is to be numerical and is just the group number, 1 through 12. The variable *id* is to be numerical and is to represent the within-group subject identity, so it runs from 1 up to the number of subjects in the group. Add to this table a new variable consisting of any numbers you like chosen to have, per group, sample average equal to the group mean from the original (summary) table and sample standard deviation equal to the group standard deviation from the table. Hint: suppose `X` is a numerical vector and `a`, `b` are any two numbers (numerical vectors of length 1). Define `Y = a + b*(X-mean(X))/sd(X)`. What is the sample mean and sample standard deviation of the numbers in the vector `Y`?

4 Step 4 (statistical analysis and a permutation test).

Notice from Figure 1 that six of the groups have large and roughly equal averages, and the other six have small and roughly equal averages. Use analysis of variance to test the hypothesis that the true means of the six “small” groups are all equal to one another and that the true means of the six “large” groups are all equal to one another. Extract the p-value of the F-statistic. [Reminder from your course on linear models: suppose group is a factor with 12 levels and megagroup is a factor with 2 levels corresponding to combining the original groups into two. The ANOVA you need could be done by something like `anova(aov(Y~megagroup),aov(Y~group))`.]

If the two groups of six were indeed statistically indistinguishable, we could pool all subjects in each of the two combined “mega-groups” (of sizes 84 and 86) and then redistribute the scores at random again to groups of the original sizes (14, 14, 14, 14, 14, and 13 for the megagroup of the large averages, 14, 14, 14, 16, 14, and 14 for the megagroup of the small averages).

1000 times, create a new data set with the subjects of the large average groups randomly permuted, and those of the small average groups randomly permuted, perform the analysis of variance, and extract the p-value of the F-statistic. Compare the p-value of the F-statistic for the original (unpermuted) data with the empirical distribution of p-values for the randomly permuted data sets (i.e., determine what proportion of the 1000 permutation p-values are smaller and what proportion are larger than the original p-value).

Note: a random permutation of a finite number of distinct elements is just a sample without replacement from the set of elements of the same size as the original set.