Assignment 1 - Statistics for molecular biomedicine

March 16, 2021

- Please read the assignment instructions in the assignment module on Absalon.
- In this assignment only use base R - dont use other packages i.e. dont use library().
- In the following questions use a significance threshold of 0.05 unless otherwise stated.
- This is an individual assignment. You cannot share code and answers with someone else.
- Use the discussion board for question about the assignment - however DO NOT post your suggested solution (see pinned post in the discussion board)

1 Part1

The GC-content (or GC-ratio) is the fraction of nucleotides in a specific DNA region that are either guanine or cytosine (from a possibility of four different nucleotides). We have calculated the GC-content of 4187 human introns (`intron_gc.txt`) and 4187 exons (`exon_gc.txt`). Each line in these files corresponds to a gene who's GC content in the intronic regions and exonic regions have been calculated. Thus the first line in both files corresponds to the same gene and the GC content for this gene's intronic regions is found in the first line of the file `intron_gc.txt`, while the GC content for this same gene's exonic regions is found in the first line of the file `exon_gc.txt`.

1A Test the null hypothesis that there is no difference in the mean of the GC content in exons and introns using a parametric test. Report the chosen test, the p-value and what you can conclude based on the test?

1B Test the null hypothesis that there is no difference in the GC content in exons and introns using a non-parametric test. Which test did you use? Also report the p-value and what you can conclude based on the test?

1C Very briefly name the main advantage and main disadvantage of non-parametric test

1D For each of the genes there are 15 exonic ~~and intronic~~ regions. Each of these regions either contain a GC repeat or not, and the number of regions that do contain them has been counted for each gene. The result is summarized in the following table:

Number of **exonic** regions with GC repeats	0	1	2	3	4	5	≥ 6
Number of genes	390	1000	1149	872	481	205	90

What is the mean number of **exonic** regions with GC repeats in a gene if you assume there are on avg. 6.25 repeats in the ≥ 6 category?

1E Assume that the number of **exonic** regions with GC repeats for each gene is binomial distributed where each exomic region has a chance to contain a ~~single~~ GC repeat. Based on you estimated mean what is the probability that a single exonic region has a GC repeat and what is the probability of observing more than **exonic** 6 GC regions in a gene?

1F We want to evaluate our assumption of the data following the previously described binomial distribution. Construct and print a table of expected and observed number of genes with $0, 1, 2, \ldots, 5, \geq 6$ **exonic** GC regions that you can use for a goodness of fit test.

1G Apply the goodness of fit test. Report the test statistic, the p-value and the number of degrees of freedom of the χ^2 distribution.

1H What is your conclusion on the goodness of fit test for the binomial distribution?

2 Part 2

In a study with 90 test participants, 30 from each of 3 age categories (< 25, $25 - 40$ and > 40) were assigned to each of 3 exercise schemes; low (l), moderate(m) and high(h). Then the change in their physical condition after 3 months, summarized in a single number, was recorded. We wish to test if the effect of exercise intensity on physical condition depends on age. The data is found in the file `conditions.txt`.

2A First we want to establish whether the exercise intensity has an effect on the change of physical condition. Choose an appropriate test. Report your test choice, your null hypothesis and you alternative hypothesis.

2B Report the p-value of your test, and the conclusion.

2C Plot the group mean of physical condition for each of the nine combinations of age and exercise level.

2D Fit a model for the additive effect of both age and exercise intensity on change of physical condition. Report the Multiple R-squared (use Adjusted).

2E Based on you fitted model, what is the predicted change of physical condition for a 36 year old after moderate exercise intensity?

2F We want to test if the effect of exercise intensity depends on age. Perform the appropriate test and report the relevant test statistic and p-value, that tells us if the effect of exercise intensity on physical condition depends on age.

2G State you conclusion based on this test.

2H Check if the assumptions of this test are satisfied. Please keep the answer brief.