assignment

November 17, 2023

0.1 Assignment for BIO724P 2023-2024

0.1.1 Instructions

Perform the following tasks divided into four problems. Each problem is equally weighted for the final grade. Return this jupyter notebook with your solutions but with a cleaned output. If you can't return a cleaned jupyter notebook, submit a text file where you clearly indicate which lines refer to your solutions and comments. If you use libraries, make sure to indicate where to load them. You must use **R** to solve these problems. Other programming languages are not allowed. Make your code reproducible, interpretable, and efficient. We should be able to run your code as seamlessly as possible. You need to submit only one file with the solutions.

0.1.2 Problems

In the analysis of next-generation sequencing (NGS) data, one of the main challenges is to perform genotype calling. This operation consists of assigning genotypes to each sequenced individual at each genomic site. For instance, in humans and other diploid species with low mutation rates, at each site individuals are assigned either of three possible genotypes: homozygote for the reference allele, heterozygote, or homozygote for the alternate allele. The information on called genotypes is typically present in VCF files.

Assuming that the reference allele is coded as "0" and the alternate allele is coded as "1", the sample space S of all possible genotypes over two independent (unlinked) genomic sites is S={00-00, 00-01, 00-11, 01-00, 01-01, 01-11, 11-00, 11-01, 11-11}. In the following table, S is represented with the genotype for the first site on rows, and the genotype for the second site on columns.

S	00	01	11
00	00-00	00-01	00-11
01	01-00	01-01	01-11
11	11-00	11-01	11-11

In other words, assume that your experiment consists of drawing two genotypes for the same individual over two sites. S represents the possible outcomes of this experiment (e.g., the outcome 00-01 represents homozygote for allele "0" in one site {00} and heterozygote {01} in the second

site). Note that, as expected, we obtain $3^2 = 9$ possible outcomes. Finally, note that genotypes $\{01\}$ and $\{10\}$ are equivalent.

Problem 1

Task 1.1

Construct a random variable G representing the number of alternate alleles over two sites. In this case, we are not interested in the genotypes per se as previously defined, but rather at the count of alternate alleles "1" over two sites. For instance, a homozygote for the reference allele will not contribute to the count of the alternate alleles, while a heterozygote will contribute by adding one to the count. Note that we are interested in this metric over two sites. Assuming that each genotype has equal probability to occur, use this information to define the sample space of G and calculate its probability mass function.

Task 1.2

Plot the probability mass function and cumulative distribution function of the random variable G previously defined.

[ ]:

Problem 2

Task 2.1

Assume that we observed some data of variable G for several sequenced individuals. Specifically, the called genotypes for ten individuals over two sites are summarised in the following table:

individual	genotype at first site	genotype at second site
1	00	01
2	00	01
3	11	00
4	01	11
5	11	01
6	00	00
7	01	01
8	00	11
9	11	00
10	11	11

Produce an appropriate visualisation of the distribution of G based on these individuals. Also, calculate one metric of central tendency, one metric of scale, and one metric of skewness for the distribution of G.

[ ]:

Task 2.2

Test whether the mean of G is statistically significant different from the mean of the same random variable calculated on a different population *G2*. *G2* is defined as *G2 <- c(2, 4, 2, 3, 3, 3, 2, 3, 3, 4)*. Define your hypotheses and write a report statement.

[ ]:

Problem 3

Assume that you have the numerical variable `genotypes` of G values, as previously defined, for 200 individuals. For each individual you have also access to further variables, such as `ancestry`, `income`, and `risk`. Variable `ancestry` is categorical coding for arbitrary groups of genetic ancestry. The variable `ancestry` should be treated as a factor variable. Variable `income` is continuous and represents the household income in pounds (in thousand units). Variable `risk` is continuous and indicates the susceptibility to a certain disease, in arbitrary units. These variables are accessible from the file `assignment.csv`.

Test whether `genotypes`, `ancestry` and `income` are statistically significant explanatory variables for the response variable `risk` using a general linear model. Choose the variables that fit the model the best by reducing the original model. Write a report statement and produce a plot of the final model (i.e. significant explanatory variables against the response variable).

[ ]:

Problem 4

Assume that the variable `income` is distributed as a mixture of Normal distributions. Specifically, its probability density function is modelled as $\alpha\mathcal{N}(\mu_1, (\mu_1/10)^2) + (1-\alpha)\mathcal{N}(\mu_2, (\mu_2/10)^2)$ with parameters α, μ_1, and μ_2. In this mixture distribution, two Normal distributions with different means and variances are weighted by α and $1-\alpha$, respectively.

Calculate point estimates and confidence intervals (or other metrics of uncertainty) for each parameter. Choose a suitable statistical approach among the ones discussed in class.

[ ]: