Exercise week 1

Gherardo Varando, gherardo.varando@math.ku.dk

20/11/2019

1 Blood types genetics

In classical genetics, we recognize three alleles for the ABO gene that regulates ABO blood type: i, I^A and I^B. Any individual inherits a complete set of chromosomes from its two parents and thus the genotype of an individual is a pair of alleles (the order is not important).

Ex 1.1 Describe the sample space Ω related to the genotype of an individual.

Suppose now that the individuals inherit from each parent randomly one of the two alleles. That is if the two parents have alleles $\{x_1, x_2\}$ and $\{y_1, y_2\}$ respectively, then the offspring will receive x_1 or x_2 from the first parent and y_1 or y_2 from the second parents, with probabilities all equals to 0.5. Suppose moreover that the events related to different parents are independent. From now on we use the notation $ab = \{a, b\}$ for genotypes.

Ex 1.2 If both parents have genotype $iI^A = \{i, I^A\}$, compute the probability that the offspring will have genotype ii and the probability that at least one of the two alleles will be i.

Ex 1.3 Suppose the following two-generations scenario:

- Parents with genotypes iI^A and iI^B generate offspring A.
- Parents with equal genotypes I^AI^B generate offspring B.
- Parents A and B generate individual C

Compute the probability of the genotype of C being I^AI^B

We know that even if there are 6 genotypes (the sample space Ω), the possible phenotypes are only the following 4:

- **Type A** produced by I^AI^A and I^Ai.
- **Type B** produced by I^BI^B and I^Bi.
- **Type AB** produced by I^AI^B.

- **Type O** produced by ii.

Observe that the mapping from genotype to phenotype is an example of "extended" random variable.

Ex 1.4 Suppose the same two-generation scenario of Ex 1.3, which are the possible phenotypes for C? Can you describe the probability distribution of the phenotype of C?

Ex 1.5 What is the conditional probability of the phenotype of C being **Type B** given that the genotype of B is $I^A I^B$?

1.1 Some simulation in R

Lets now perform some simulations on the genotypes and phenotypes of blood types.

Ex 2.1 How we can simulate a random individual genotype ? (use the *sample* function).

Ex 2.2 Create an R function that given two genotypes return the (random) offspring genotype. Create a function that computes the phenotype from the genotype.

Ex 2.3 Approximate from simulations the probabilities in Ex 1.2. That is, simulate a large number of samples and approximate the probability with the empirical frequencies.

Ex 2.4 Can you code the two-generation scenario of Ex 1.3? And sample 1000 observation of the phenotype of C.

2 A case study of DNA sequence

Let Ω be the sample space of a single DNA nucleotide type,

$$\Omega = \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{G}\}.$$

Suppose that all four nucleotide types have equal probabilities in the genome and all locations in the DNA are independent.

A DNA motif is a specific sequence of nucleotides, for example AATG or CGGCC. The number of nucleotides in the motif is called the length of the nucleotides.

Ex 3.1 Describe the sample space of DNA sequences of length 2.

Ex 3.2 Consider a DNA sequence of length 5. Describe the event of observing the motif ACG inside the DNA sequence. What is the probability of this event?

Ex 3.3 Suppose we observe from a given position in the genome. We keep observing until we find the nucleotide G. Describe the sample space of the observed DNA sequence. What is the probability that the DNA sequence observed before observing G has length 10?

Ex 3.4 Now we remove the independence assumption. Suppose that the probability of observing A,T,C or G given the previous nucleotide being C, is respectively 0.2, 0.2, 0.5, 0.1. In general in different locations of the DNA sequence, this conditional probabilities will be different and this fact is important for biological sequence detecting.

1. Suppose we observe C at a given location, what is the probability of observing the motif CG immediately after?
2. Suppose now in a different region of the DNA sequence, the conditional probabilities of A, T, C, G given C become $0.1, 0.2, 0.3, 0.4$, then what is in that region the probability of observing CG given that we observed C in the previous location?

2.1 Some simulations in R

We now conduct some simulations in R for the DNA case study.

Ex 4.1 Run a simulation of exercise 3.2 and check the probability of observing the motif ACG in a sequence of length 5. Idea: generate a sequence of length 5 and check if it contains the motif ACG. Repeat the above sampling and checking precess 1000, 10000 or 100000 times. Obtain for each case the proportion of sequences containing the motif ACG, we can use such proportions as an approximation of the probability.

Ex 4.2 Run an R simulation of exercise 3.3 and calculate the probability of observing 10 nucleotides before G, using 10000 repetitions.

Ex 4.3 Repeat exercise 4.2 for different lengths of the sequence $\{0, 1, 2, \ldots, 20\}$. Show the result in a scatter plot, where the x-axis is $\{0, 1, \ldots, 20\}$ and the y-axis is approximated probabilities. Plot then in the same plot the real probabilities.

3 A case study of neuronal data

Neurons work by generating and propagating, action potentials, called “spikes”. The time interval between two adjacent spikes, or inter-spike interval (ISI), is often used in computational neuroscience. In the file `neurospikes.txt` you can find some ISI measurements.

To load the data in R, you can use the following line,

```
isidata <- read.table("neuronspikes.txt", col.names = "isi")
```

Ex 5.1 Plot the histogram of the ISI data using the function `hist`, try with different values for the `breaks` parameter. We can use then `breaks = 50`. Moreover use the argument `prob = TRUE` to plot probabilities.

Ex 5.2 As a first approximation we can use the exponential distribution, that is often used to describe intervals between events, e.g. neural spikes. The density of the exponential distribution is $f(x) = \lambda e^{-\lambda x}$, and in R we can use the `dexp` function where the λ parameter is called `rate`. Plot the densities of the exponential distribution on top of the histogram of the ISI data for different values of the parameter $\lambda = 0.5, 0.6, \ldots, 1.4, 1.5$. Which values of λ do you think fits better the data?

Ex 5.3 Using the value λ^* that you found visually in the previous exercise, find the value of $P(X \leq 1)$ and the value of $P(X \in [0.5, 1.5])$ (where X is the random variable that records the ISI value and we assume $X \sim exp(\lambda^8)$). Remember that the CDF of a random can be computed in R for an exponential distribution with the function `pexp`. The CDF F_X is defined as,

$$F_X(x) = P(X \leq x).$$

4 Brain cell database

The brain cell database is a survey of biological features measured from single cells (neurons), in human and mouse. It is available in the Allen Institute webpage, `http://celltypes.brain-map.org/data` but you can also directly download the file from absalon course page.

The dataset consists of both human and mouse cells and includes morphological and electrophysiological values.

Download `cell_types.csv` file and load the dataset in R using the command: `cells <- read.csv("cell_types.csv", na.strings = "")`

Ex 6.1 Explain why we need the `na.strings = ""` argument in the function call of `read.csv`, have a look to the function documentation.

Ex 6.2 Find which observations in the dataset are from human donors and which observations are from mice. What is the proportion of human cells in the database?

Ex 6.3 Plot an histogram of the probabilities of the variable named `ef__peak_t_ramp` that record the *ramp spike time* in seconds, which is the time to the first spike when the cell is stimulated with a ramp type stimulus (you can read more details at `http://celltypes.brain-map.org/data`). Plots then histograms for the human and the mouse cells separately.

Ex 6.4 Plot the density of a log-normal distribution (`dlnorm`) with parameters `sdlog = 0.6` and `meanlog` varying from 1 to 3.

Ex 6.5 Which value of the parameter `meanlog` (fixing `sdlog = 0.6`) fits better the histogram of ramp spike time observations?

Ex 6.6 Plot now the density of the log-normal distribution with the parameters found in ex 6.5 on top of the histogram for the human cells only. Does the density fit the data?

Ex 6.6 Compute how many human cells in the dataset are from male donors and how many are from female donors.