Introductory Biostatistics for Biologists

IGC, September 11th - September 15th, 2017

Exercises III / Exercises IV

Distributions in R: probabilities, densities and simulation

Random numbers form a basic tool for any simulation study. Simulations require the ability to generate random numbers. On a computer, it is only possible to generate "pseudo-random" numbers which for practical purposes behave as if they were drawn randomly. All random number generators essentially work as follows:

(a) A seed number is needed as input for the process of generating a random number. This seed can be supplied by the user or the computer generates the seed e.g. as a function of the data.

(b) The seed number is put into mathematical functions that eventually return a random number and a new seed that will be used to generate the next random number.

In `R`, "set.seed" declares the seed for the random generator. If we use this command before a random number generating statement, we are able to retain the same number each time we provide the same seed.

`R` has the ability to sample with and without replacement. That is, choose at random from a collection of things such as the numbers 1 through 6 in the dice rolling example. The sampling can be done with replacement (like dice rolling) or without replacement (like a lottery). By default sample samples without replacement each object having equal chance of being picked. You need to specify `replace=TRUE` if you want to sample with replacement. Furthermore, you can specify separate probabilities for each if desired.

1. Suppose that 502 individuals were classified by blood group and sex as follows:

	sex		
blood goup	Male	Female	Total
0	113	113	226
A	103	103	206
B	25	25	50
AB	10	10	20
Total	251	251	502

   A person is picked at random from this group:

   (a) What is the probability of being a female?

   (b) Given that she is female, what is the probability of having blood group A or B?

2. Breast cancer is considered largely a hormonal disease. An important hormone in breast cancer research is estradiol. The following data on serum estradiol levels were obtained from 213 breast cancer cases and 432 age-matched controls. All women were age 50-59 years.

Serum estradiol (pg/mL)	Cases (n=213)	Controls (n=432)
1-4	28	72
5-9	96	233
10-14	53	86
15-19	17	26
20-24	10	6
25-29	3	5
≥ 30	6	4

   Suppose a serum estradiol level of ≥ 20 pg/mL is proposed as a screening criterion for identifying breast cancer cases.

   (a) What is the sensitivity of this test?

   (b) What is the specificity of this test?

(c) In the general population, the prevalence of breast cancer is about 2% among women 50-59 years of age. What is the probability of breast cancer among 50-59 years old women in the general population who have a serum estradiol level of ≥ 20 pg/mL? What is another name to this probability?

(d) Compute the negative predictive value.

3. The Binomial distribution is a possible probability model for the number of stormy days in a season.

   (a) Do you think that this is a realistic model? If there are 90 days in a winter and the probability of a stormy day in winter is 1/3, write down the parameters, n and p, of the Binomial model.

   (b) Compute the expectation and variance of the number of stormy days, and compute the probability of a winter having more than 30 stormy days.

   (c) Generate 100 Binomial variables to represent a sequence of 100 winters and plot the simulated data. What is the average number of stormy days simulated?

4. Suppose that for certain microRNA of size 20 the probability of a purine is binomially distributed with probability 0.7.
   a) What is the probability of 14 purines?
   b) What is the probability of less than or equal to 14 purines?
   c) What is the probability of the number of purines being between 10 and 15?
   d) How many purines do you expect?

5. The distribution of the expression values of the ALL (Acute Lymphoblastic Leukemia) patients on the Zyxin gene are distributed according to N(1.6, 0.42).
   a) Compute the probability of the expression values being smaller than 1.2?
   b) What is the probability of the expression values being between 1.2 and 2.0?
   c) What is the 15th percentile of that distribution? What does it mean?
   d) Use `rnorm` to draw a sample of size 1000 from the population and compare the sample mean and standard deviation with those of the population.

6. Given a sample 0.12, 0.24, 0.01, 0.16, 0.18, 0.55, 0.89, 1.00, 1.45 and 2.5 corresponding to intensity levels of one gene from 5 DNA chips:

   (a) Analyze the distribution considering normality, using `density` function in R and construct a normal QQ-plot to check for normality.

   (b) Apply a log_2 transformation and repeat the analysis.

   (c) Use the `qqnorm()` and `qqline()` functions to get the normal QQ-plot. Compare your results.