Introductory Biostatistics for Biologists

IGC, September 11th - September 15th, 2017

Exercises I / Exercises II

Basic exploratory statistics with `R`

Create your own folder for your R session. All your analysis will happen in this directory. At any point during this session you may save the workspace so that you can come back to the same session later.

1. The famous (Fisher's or Anderson's) `iris` data set from `R`, gives the measurements in centimeters of the variables *sepal length*, *sepal width*, *petal length* and *petal width*, respectively, for 50 flowers from each of 3 species of iris. The species are *setosa*, *versicolor* and *virginica*.

   (a) To describe a distribution we often want to know where is it centered and what is the spread. These are typically measured with mean and variance (or standard deviation), or the median and more generally the five-number summary. The `R` commands for these are `mean`, `var`, `sd`, `median`, `fivenum` and `summary`. Calculate the descriptive statistics for the variables of the `iris` data.

   (b) Compute the skewness using the function `skewness` of the `moments` package of the variable *sepal width* for the three species.

   (c) Compare the four groups using box plots and save it into a pdf file.

   (d) Construct histograms and box plots for the four groups on the same plot.

   (e) Plot a pie chart for the variable *species*.

   (f) Considering the four groups, create an error plot using SD and another with 95 % CI. Plot them into the same window and compare. Install the `psych` package and use the function `error.bars()`.

   (g) Plot the kernel density estimate using the function `density()` for the four groups on the same plot. Characterize the groups concerning skewness.

2. Consider a sample of measurements (in centimeters) of butterfly wing lengths:

3.3	3.5	3.6	3.6	3.7	3.8	3.8	3.8	3.9	3.9	3.9	4.0
4.0	4.0	4.0	4.1	4.1	4.1	4.2	4.2	4.3	4.3	4.4	4.5

   (a) Classify the type of variable.

   (b) Compute descriptive statistics for this variable.

   (c) Represent the data graphically.

3. The data in the file "Tumor" are a sample of weights, in ounces, of malignant tumors extracted from the abdomen of 57 individuals. Compute descriptive statistics and interpret the results.

4. The following table shows the measurements (in centimeters) of wing and tail lengths among birds of a particular species:

Bird	Wing length	Tail length	Bird	Wing length	Tail length
1	10.4	7.4	7	10.7	7.4
2	10.8	7.6	8	10.5	7.2
3	11.1	7.9	9	10.8	7.8
4	10.2	7.2	10	11.2	7.7
5	10.3	7.4	11	10.6	7.8
6	10.2	7.1	12	11.4	8.3

   (a) Compute descriptive statistics for both variables and represent the data graphically.

   (b) Compute the correlation coefficient and interpret the results.

5. The Apgar score (Apgsc) was developed in 1952 as a measure of the physical condition of an infant at 1 and 5 minutes after birth. The score is obtained by summing five components, each of which is rated as 0, 1 or 2 and represents different aspects of the condition of an infant at birth (Table 1). The score is routinely calculated for most newborn infants. The results obtained from a sample of 24 infants are described at the following table:

Infant	Apgsc-1min	Apgsc-5min	Infant	Apgsc-1min	Apgsc-5min
1	10	10	13	6	9
2	3	6	14	8	10
3	8	9	15	9	10
4	9	10	16	9	10
5	8	9	17	9	10
6	9	10	18	9	9
7	8	9	19	8	10
8	8	9	20	9	9
9	8	9	21	3	3
10	8	9	22	9	9
11	7	9	23	7	10
12	8	9	24	10	10

Study the association between the Apgar scores at 1 and 5 minutes.

Table 1: Apgar score

The five criteria of the Apgar score:

	Score of 0	Score of 1	Score of 2	Component of Acronym
Appearance/Complexion	blue or pale all over	blue at extremities body pink (acrocyanosis)	no cyanosis body and extremities pink	Appearance
Pulse rate	Absent	<100	>100	Pulse
Reflex irritability	no response to stimulation	grimace/feeble cry when stimulated	cry or pull away when stimulated	Grimace
Activity	none	some flexion	flexed arms and legs that resist extension	Activity
Respiratory Effort	absent	weak, irregular, gasping	strong, lusty cry	Respiration