

STATISTICS FOR EES — EXERCISE SHEET 6

1. A plant of a certain species reaches on average a height of 123 cm after four months. The height, however, also depends in the amount of sun (S) and the amount of rain (R) that the plant will get during this time. S and R (which are measured in certain units) vary from year to year, and their standard deviations are $\sigma_S = 3.6$ and $\sigma_R = 4.7$. Furthermore, sunny years tend to be less rainy, which can be quantified by $\text{Cor}(R, S) = -0.1$. Each unit of sun increases the plant height (after four months) by 1.2 cm, and each unit of rain increases the height by 4.2 cm. Additional genetical and environmental factors, which are not correlated to R and S , adds a random increase or decrease F to plant height, with $\sigma_F = 5.7$. Calculate the total standard deviation of plant height after four months.

2. In a series of experiments with a male bird you played four times a recording of the song of a female of the same species and measured how long it took until the male replied with his song. The measured “reply times” in seconds were 1.52, 2.31, 1.24, 0.91.

- Calculate expectation value, variance and standard deviation of the from the data (as estimations of these parameters of the reply times of this male bird).
- Apply the transformation $g(x) = 1/x$ to the data to obtain the “reply rate” and calculate the expectation value, variance and standard deviation also for the transformed data.
- Apply another transformation f of the data, which consists in subtracting the minimum possible reaction time of 0.3 seconds and measuring the rest in milliseconds. Calculate expectation value, variance and standard deviation of the transformed data.
- To what extent could you also calculate expectation value, variance and standard deviation by applying the transformations g and f or modifications of them to the expectation value, variance and standard deviation calculated for the untransformed data?

3. In a published study you find a comparison between the genomes of two closely related species. In this study, 100 different alignable genomic regions of 1 kb each have been sampled, sequenced and compared. The average number of substitutions observed in a region was 7.8, where 4.2 and 3.6 were the average numbers of transitions and transversions, respectively. Of course, the actual numbers differed between the genomic regions, and the standard deviation was 2.4 for the transitions, 3.4 for the transversion and 4.9 for the total number of substitutions.

For your own research project the correlation of the number of transitions and the number of transversions in 1 kb regions is of interest. Can you calculate (more precisely, reasonably estimate) this correlation from the statistics given above or would you need to contact the authors of the study (or download the raw data)? If it is possible to calculate the correlation, what is the result? If not, why not?

4. Let X_1, X_2, \dots, X_n be independent random variables with the same distribution with finite variance σ^2 . Proof that the corrected sample variance is unbiased, that is:

$$\mathbb{E} \left(\frac{1}{n-1} \sum_{i=1}^n \left(X_i - \frac{1}{n} \sum_{j=1}^n X_j \right)^2 \right) = \sigma^2.$$

5. A breeder crossed 200 pairs of plants of some crop. The yield of the father plant, the mother plant and the offspring (F1) is given in table yield.csv (in some measuring unit). For the next generation, the breeder crosses plants from the F1 generation, but selects for this only plants of a yield of more than 20.

- (a) Visualize the distributions of the yield in the parent population, in the offspring population before selection and in the offspring population after selection.
- (b) Add lines indicating the mean values to your plot.
- (c) Predict the average yield of the plants of the F2 generation and show it in your plot.