4A Graphs and statistics of data sets

Class problems

4A1 (Grouped data) This table represents the age distribution of Finland on 31.12.2015. The table lists age groups in whole years, but in this exercise we treat ages as real numbers, so someone in the age group 0–14 might be 14.7 years old.

Age (years)	Frequency
0–14	896 023
15–24	640 387
25–44	1 363 155
45–64	1 464 640
65–74	642 428
75–	480 675

(Source: Tilastokeskus)

(a) Draw a histogram of the grouped data. The units of the horizontal and vertical axes should be years and %/year, respectively. You can assume the last bin ends at 110.

Try to answer the following questions by using the grouped data.

(b) Which are more common in the population, 1-year-olds or 66-year-olds?

(c) What is the median age of the population?

(d) What is the average age of the population?

In (b)–(d), did you have to make additional assumptions? If yes, how valid do you think they are?

4A2 (Quantiles) The R software defines the quantile function of data $x = (x_1, \ldots, x_n)$ as follows. Let $x_{(1)}$ = the smallest number in the data, $x_{(2)}$ = second smallest, etc. Thus we have ordered data $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Then the horizontal unit interval $[0, 1]$ is divided into $n-1$ equal parts, at points $p_k = (k-1)/(n-1)$, $k = 1, \ldots, n$. The quantile function is defined by drawing points $(p_k, x_{(k)})$ and connecting them with straight line segments.

Draw (on paper by hand) the quantile functions of the following data sets, and for each data set, determine the lower quartile $Q(0.25)$, median $Q(0.50)$ and upper quartile $Q(0.75)$:

(a) $x = (1000, 2000, 5000, 9000)$,

(b) $x = (1000, 2000, 2000, 8000, 9000)$,

(c) $x = (1, 20, 1, 5, 1)$.

Then consider the following claims. For each claim, either argue why the claim is true (for all data sets), or show it false by a counterexample.

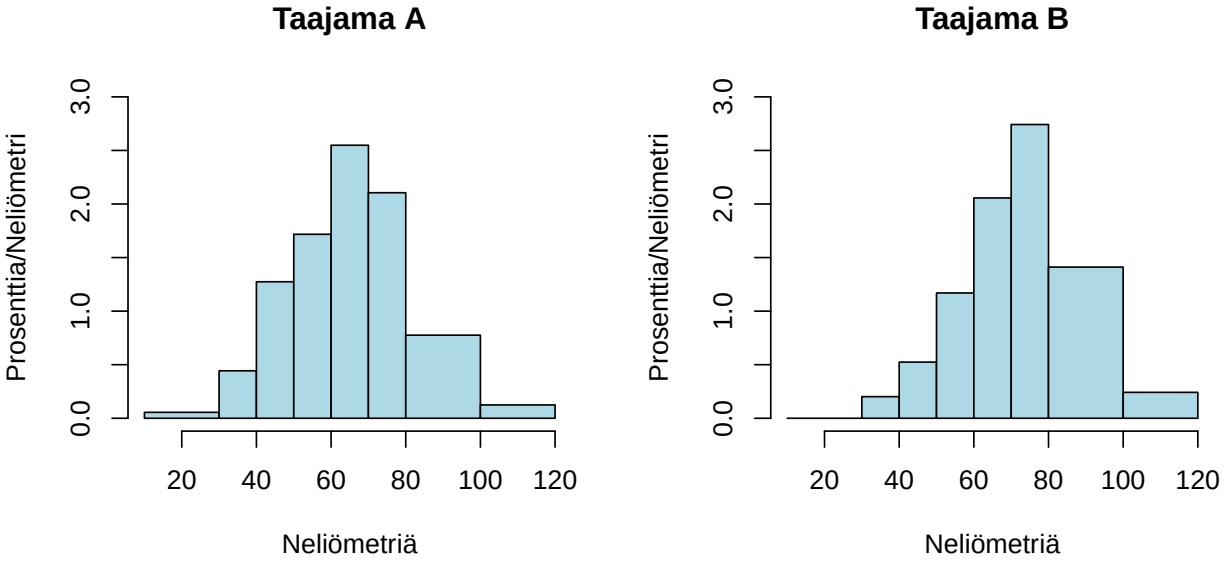
(d) The mean and median of a data set are always equal.

(e) The lower quartile is always smaller or equal to the median.

(f) The lower quartile is always smaller or equal to the mean.

Home problems

4A3 (Apartment sizes.) In town A there are 361 apartments, and in town B 248 apartments. The following histograms describe the size distributions (in square meters, "neliömetriä").



Answer the following questions by using the histograms. Assume, for simplicity, that no apartment has area exactly at a bin boundary.

(a) How many apartments in town B have area at least 80 m^2?

(b) In which town is the median area larger? Did you have to make additional assumptions about the distribution to answer this question?

4A4 (Two dice) The lecturer rolled two dice, red and yellow, each $n = 18$ times. The results are a bivariate data set $((r_1, y_1), \ldots, (r_n, y_n))$. The following contingency table shows the counts of each possible pair or values (r, y) in the data set.

		r					
		1	2	3	4	5	6
	1	0	0	0	0	0	0
	2	0	0	1	2	0	0
y	3	0	0	0	1	0	0
	4	1	0	1	0	0	0
	5	1	0	0	0	0	0
	6	2	4	1	1	2	1

(a) Find the empirical distributions of the red die and the yellow die (two different distributions). Calculate the average yellow result and the average red result.

(b) Calculate the standard deviations of each empirical distribution.

(c) Calculate the correlation coefficient between red and yellow results, from the empirical joint distribution. Hint: First calculate $\mathrm{E}(RY)$, where R and Y are random variables following the empirical distribution. Then use $\mathrm{Cov}(R, Y) = \mathrm{E}(RY) - \mathrm{E}(R)\,\mathrm{E}(Y)$. Finally calculate the correlation coefficient from the covariance.

(d) Is the empirical correlation coefficient negative, zero or positive? Explain in words what this tells about the data set.

(e) All of the above concerns the empirical distribution. Let us now think about the generating distribution (stochastic process) of the two dice. Let R and Y be random variables expressing the process of rolling the red die and the yellow die. Based on your physical understanding (and possibly on the empirical observations), do you think R and Y are independent or dependent? What is their correlation coefficient in the *generating* distribution?

(f) Based on the empirical observations, give your opinion on whether the red die is fair, and on whether the yellow die is fair. (So far we do not have mathematical tools for this, thus you have to rely on your opinion.)

Hint: See Lecture 3B about empirical distributions and contingency tables.