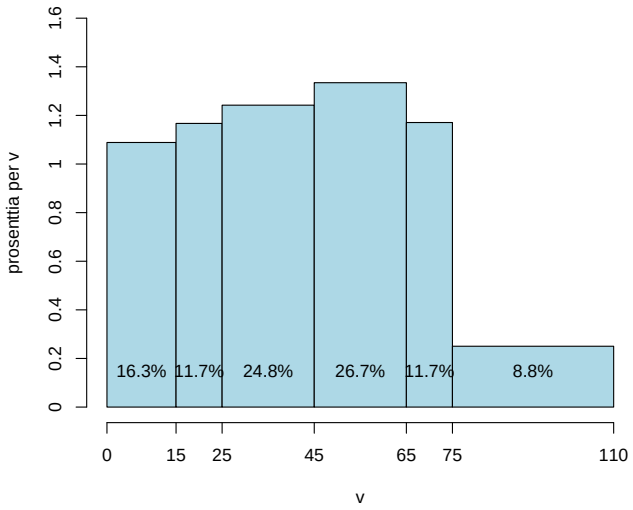
4A Graphs and statistics of data sets

Class problems

4A1 (Grouped data) The table and the histogram represent the age distribution of Finland on 31.12.2015. Here we treat ages as real numbers; a person who is 14.9 years old belongs to the interval $[0, 15)$. That's a half-open interval, it contains the point 0 but does not contain 15.

Age (years)	Frequency
$[0, 15)$	896 023
$[15, 25)$	640 387
$[25, 45)$	1 363 155
$[45, 65)$	1 464 640
$[65, 75)$	642 428
$[75, 110)$	480 675
Total	5 487 308

(Source: Tilastokeskus)



Answer the following questions by using the grouped data. In (a)–(c), assume that within each group, the ages are distributed uniformly.

(a) Which are more common in the population, 1-year-olds or 66-year-olds? (By a "1-year-old" we mean a person whose age, as a real number, is in the interval $[1, 2)$.)

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

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

(d) What can we say about median and average age, if we cannot assume uniform age distribution within groups? Can we know them exactly? If not, how small and how big could they be?

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.

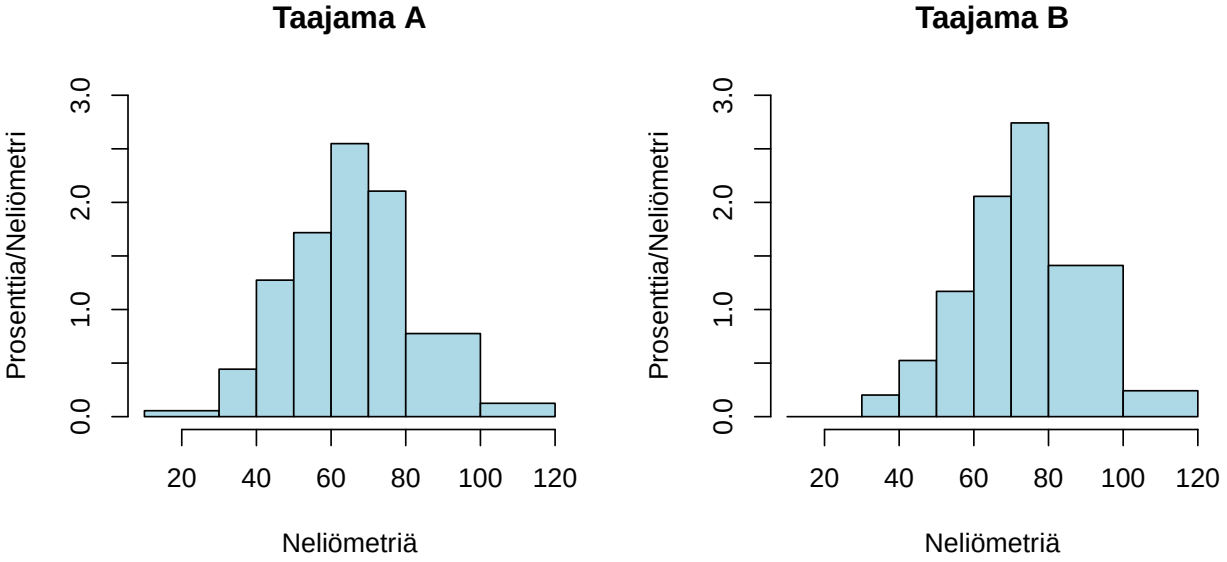
(d) The mean and median of a data set are always equal.

(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 performed 18 times the experiment of rolling a red die (R) and a yellow die (Y). The dice might be fair or not. The following contingency table shows the observed counts of pairs of values (r, y).

		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) Calculate the empirical distributions of each die (red and yellow) separately. Calculate their averages.

(b) Calculate their standard deviations.

(c) Calculate the empirical correlation coefficient. Hint: First calculate $\mathrm{E}(RY)$, in the empirical joint distribution, by considering all observed values of the pair (R, Y) and their relative frequencies. Then use the formula $\mathrm{Cov}(R, Y) = \mathrm{E}(RY) - \mathrm{E}(R)\,\mathrm{E}(Y)$. Finally recall how correlation coefficient is obtained from covariance.

(d) All of the above concerns the empirical distribution. Now think of the *generating distribution* of this random experiment (rolling these two dice, with results R and Y). Based on your observations, do you think the generating distributions of R and Y are uniform over the set $\{1, 2, \ldots, 6\}$? Do you think R and Y are dependent or independent?

Hint: Recall empirical distributions and contingency tables from lecture 3B. Note that you can use all of the probability calculus rules and formulas also with empirical distributions.