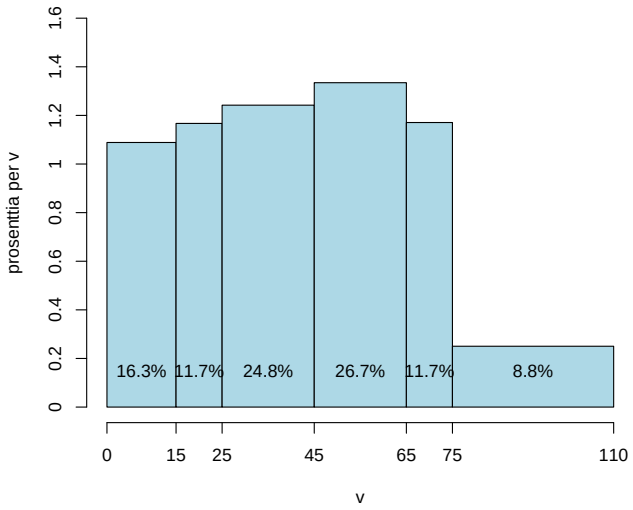
4A Graphs and statistics of data sets

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?

Solution.

(a) In the 15-year-long interval $[0, 15)$ we have 59735 persons per year, and in the 10-year-long interval $[65, 75)$ we have 64243 persons per year. Assuming each year in an interval contains the same number of persons, we have more 66-year-olds than 1-year-olds. (Instead of absolute numbers, one could also work with percentages of population). Observe that the density "persons per year" (or "% of population per year") is higher in the latter interval, even though it contains fewer persons. It is a shorter interval. The higher density is appropriately seen as a higher bar in the histogram.

(b) The first two bars cover 28% of the population. To collect the lowest half of the population, we need another 22% of the population. Since the third bar covers 24.8%, we only need $22/24.8 \approx 0.8871$ of the people in the third bar. Since the third age group is 20 years long, we take the lowest $0.8871 \cdot 20 = 17.7$ years of the third group, that is, ages from 25 to $25 + 17.7 = 42.7$.

Our estimate is that the median is 42.7 years.

(c) Again, assuming that the age distribution is uniform in each group, the average age within the first group is $(0{+}15)/2 = 7.5$ years, within the second group $(15{+}25)/2 = 20$ years, and so on. To find the average age of all Finns, we need to find the *sum* of their ages, and then divide by the total count. Adding up the ages in each group (by our uniformity assumption), we get the average

$$\frac{896023 \cdot 7.5 + 640387 \cdot 20 + 1363155 \cdot 35 + 1464640 \cdot 55 + 642428 \cdot 70 + 480675 \cdot 92.5}{5487308}$$

which is about 43.2 years. This is the *weighted average* of the group averages, where weights are the frequencies of the groups. Alternatively, we could have used the relative frequencies as weights.

(d) The median is certainly somewhere in $[25, 45]$, because the first two bars contain less than half, but the first three bars contain already more than half of the population. Nothing more can be said, because we don't know how the people are distributed over $[25, 45)$. Perhaps all of them are at the low end, or all are at the high end.

The average is as small as possible if everybody is at the lower endpoint of their age interval. Then the average is

$$\frac{896023 \cdot 0 + 640387 \cdot 15 + 1363155 \cdot 25 + 1464640 \cdot 45 + 642428 \cdot 65 + 480675 \cdot 75}{5487308}$$

or about 34.2 years. But if everybody is (almost) at the higher endpoint of their age group, the average is about 52.3 years. So we can say that the average is somewhere within $[34.2, 52.3]$.

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.

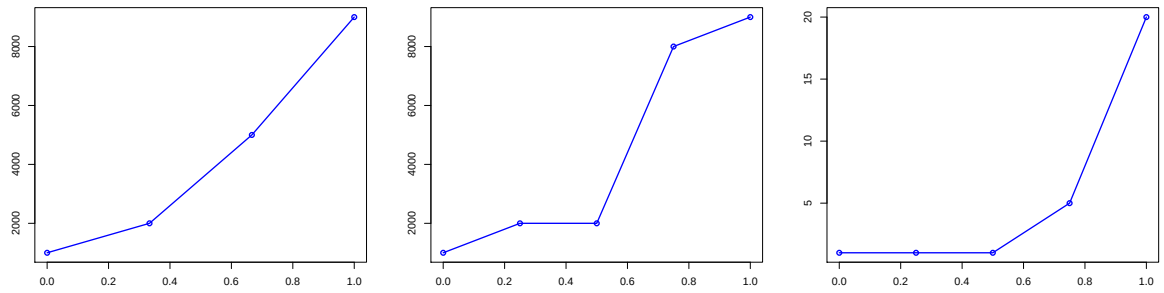
(f) The lower quartile is always smaller or equal to the mean.

Solution.

(a) Graph below.
Lower quartile $Q(0.25) = 1750$, median $Q(0.50) = 3500$, upper quartile $Q(0.75) = 6000$.

(b) Graph below.
Lower quartile $Q(0.25) = 2000$, median $Q(0.50) = 2000$, upper quartile $Q(0.75) = 8000$.

(c) Graph below.
Lower quartile $Q(0.25) = 1$, median $Q(0.50) = 1$, upper quartile $Q(0.75) = 5$.



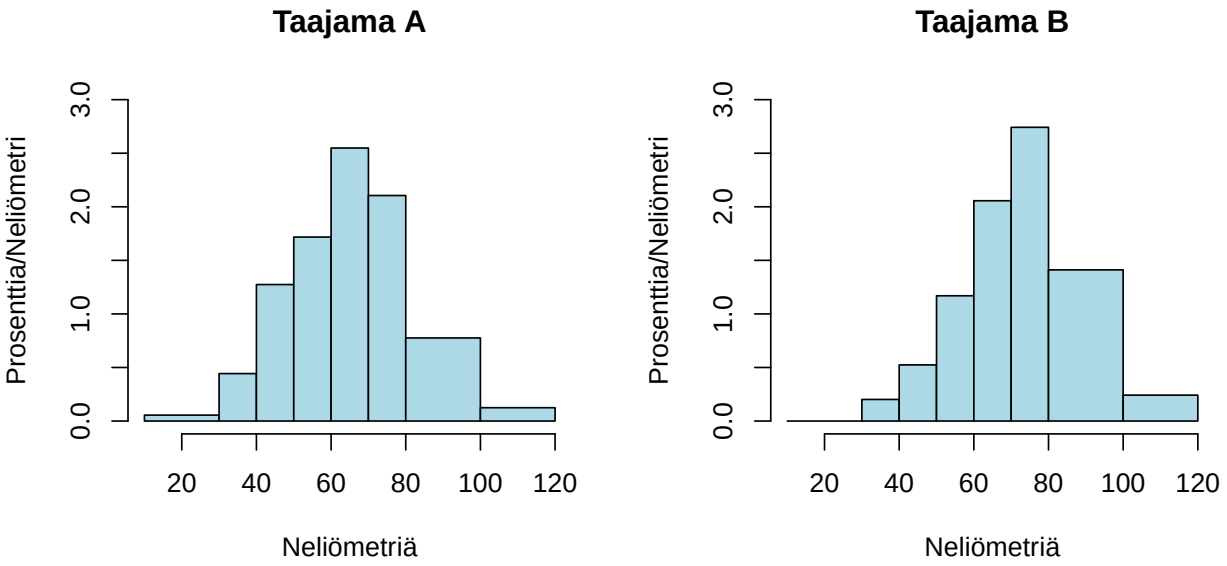
(d) False. Any of (a)–(c) is a counterexample.

(e) True, because the quantile function is nondecreasing. (The kth point cannot be smaller than the $(k-1)$th point, because the data was sorted in increasing order.)

(f) False. One very small value is enough to make the average very small, below the lower quartile. Consider, for example, the data $x = (-1000, 1, 1, 5, 20)$. It has the same quartiles as (c), but its average -194.6 is way below the lower quartile.

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?

Grading. (a) 1 p. for correct frequency with reasonable precision (within 82 ± 8). Otherwise 0 p.
(b) 1 p. for correct answer with a somewhat reasonable argument. Bar areas must be taken into account (heights not enough). Correct answer to the last question ("additional assumptions") not required for the point.

Solution.

(a) In town B, the bar heights on intervals 80–100 and 100–120 are approximately 1.4 and 0.25. Multiplying by the interval lengths, we get the bar areas. Adding them up we get the *relative frequency* of apartments of 80 m^2 or more as

$$1.4 \times 20 + 0.25 \times 20 \;=\; 33\%.$$

Because town B contains 248 apartments in total, the *frequency* of 80 m^2 or more is approximately $248 \times 0.33 \approx 82$.

(b) We can start from either end of the distribution, and try to find how many bars are needed to cover 50%. In this case we start from the right end because the bars seem bigger. (This is just a matter of convenience.) When we exceed 50%, we know that the median is within the bar where that happened.

In town B, starting from the right, we find bars of heights approximately $0.25, 1.4, 2.75$, and widths $20, 20, 10$. By multiplication, we get relative frequencies $5\%, 28\%, 27.5\%$. The two rightmost bars totalled only 33%, but with the third bar we have over 50%, so the median is certainly within $[70, 80]$.

In town A, starting from the right, we find approximate heights $0.1, 0.75, 2.1$ and thus areas $2\%, 15\%, 21\%$. Their total is $38\% < 50\%$, so the median must be lower than 70%.

According to this calculation, the median is larger in town B. In this case, we did *not* need to assume uniform distribution within the bars. Even if (and when) the distributions are non-uniform, we know that the median of town B is *somewhere* within $[70, 80]$.

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 empirical distribution of the observed pairs (r_i, y_i).

		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.

Grading. 0.5 points per item.

Solution.

(a) Empirical distribution of the red die (take column sums of the contingency table, and divide by 18):

r	1	2	3	4	5	6
$f_{\vec{r}}(r)$	4/18	4/18	3/18	4/18	2/18	1/18

The average is $m(\vec{r}) = \mathbf{2.9444}$.

Empirical distribution of the yellow die (take row sums and divide by 18):

y	1	2	3	4	5	6
$f_{\vec{y}}(y)$	0	3/18	1/18	2/18	1/18	11/18

The average is $m(\vec{y}) = \mathbf{4.8889}$.

(b)

$$\mathrm{sd}(\vec{r}) = \sqrt{\sum_{r=1}^{6} (r - m(\vec{r}))^2 f_{\vec{r}}(r)} = \mathbf{1.5082}$$

$$\mathrm{sd}(\vec{y}) = \sqrt{\sum_{y=1}^{6} (y - m(\vec{y}))^2 f_{\vec{y}}(y)} = \mathbf{1.5595}$$

Alternatively one could compute (in the empirical distribution) $\mathrm{E}(R^2)$ and then use the formula $\mathrm{Var}(R) = \mathrm{E}(R^2) - (\mathrm{E}(R))^2$.

(c) Following the hint, calculate

$$\mathrm{E}(RY) = \sum_{r=1}^{6} \sum_{y=1}^{6} ry f_{\vec{r}\vec{y}}(r, y) = 14.0556.$$

(Although there are 36 possible values of (r, y), only 12 were observed so the sum has only 12 nonzero terms.)

Then

$$\mathrm{Cov}(R, Y) = \mathrm{E}(RY) - \mathrm{E}(R)\,\mathrm{E}(Y) = 14.0556 - 2.9444 \cdot 4.8889 = -0.3393$$

(we are using the probabilistic notation, but remember that all this is in the empirical distribution). Finally

$$\mathrm{Cor}(R, Y) = \frac{\mathrm{Cov}(R, Y)}{\mathrm{SD}(R)\,\mathrm{SD}(Y)} = \mathbf{-0.1442}.$$

(d) The dataset is small, so we cannot say much. The empirical distribution of the red die is fairly close to uniform, so perhaps its generating distribution might be uniform. The empirical distribution of the yellow die is very far from uniform, which creates suspicions. In the latter half of the course we will learn tools for evaluating these suspicions mathematically.

The empirical correlation coefficient was nonzero. We know that nonzero correlation indicates dependence, because independent random variables have zero correlation.

But the empirical distribution is only an approximation of the generating distribution. In the generating distribution the correlation might well be exactly zero. If the rolls are physically separate events, it would be quite plausible that R and Y are independent.