3. 1-Dimensional Empirical Distributions

3.1. Frequency Distributions and Grouping

If the number of statistical units or observations n is very large, a data matrix does not always suffice to present the general properties of a statistical variable, as they may get easily lost in the mass of the details. In such cases we may wish to group the statistical units into categories or classes (luokat), based upon their observation values.

Let x denote the statistical variable in question and $E_i,\ i = 1, 2, \ldots, k$ the different categories, where k denotes the total number of categories (luokkien lukumäärä). The number f_i of observations belonging to class E_i are called the (absolute) frequency (or count) (frekvenssi) of that class. A table listing all classes E_i and their frequencies f_i is called a frequency distribution (frekvenssijakauma).

Often we are more interested in the relative rather than the absolute count of a class. We therefore define the relative frequency p_i (suhteellinen frekvenssi) of a class as:

$$p_i := \frac{\text{frequency in class } E_i}{\text{number of all observations}} = \frac{f_i}{n}.$$

$100 \times p_i$ is called the procentual frequency (prosentuaalinen frekvenssi).

For qualitative variables and discrete variables with only small sets of possible values the classes are naturally defined by their values.

Example. Frequency distribution of municipalities over Finlands counties in 2003:

County	f_i	p_i	$100p_i$
Etelä-Suomi	88	0.197	19.7
Länsi-Suomi	204	0.457	45.7
Itä-Suomi	66	0.148	14.8
Oulu	50	0.112	11.2
Lappi	22	0.049	4.9
Ahvenanmaa	16	0.036	3.6
Sum	446	1.000	100.0

Classes with small relative frequencies may be combined into larger classes if they fit logically together.

Grouping data of continuous variables is more difficult, because they may attain any value whatsoever within their range. Combining the values of continuous variables into classes always amounts to a loss of information, since the original observation values are replaced by the range of values for their particular class. Here comes a general *guideline for categorizing continuous variables:*

Denote the number of observations with n and the precision of measurements (mittaustarkkuus) with d (e.g. $d = 1$ for integer numbers; $d = 0.1$ if results have been measured with a precision of 1 decimal).

1. Find the smallest observation $x_{(1)}$ and the largest observation $x_{(n)}$ of the statistical variable x. The range interval (vaihteluväli) of x is the interval $(x_{(1)}, x_{(n)})$ with range/ width (pituus) $w = x_{(n)} - x_{(1)}$.

2. Decide whether you want to use a regular or an irregular classification (tasavälinen tai epätasavälinen luokitus). A regular classification means that all classes have the same width. Whenever possible one should use a regular classification, but in some cases the data is so unevenly distributed, that it becomes less meaningful.

3. Choose the number of categories k according to either $k \approx \sqrt[3]{n}$ or $k \approx \log_2 n$. For example, if $n = 125$, then $k \approx 5 - 7$, because $\sqrt[3]{125} = 5$ and $\log_2 125 \approx 7$. Usually there are 4–10 categories.

4. Determine the class interval size c (luokkavälin pituus) as $c \geq w/k$. Allowing c to be slightly larger than w/k is usually necessary in order to get somewhat more readable class intervals. On the other hand, c should not be chosen much larger than w/k, because then there will be artificially few observations in the lowest and highest category and overall less resolution.

5. Set up the categories in such a way that they cover the full range of x. The lower class limit (luokan alaraja) of the first category must be smaller or equal to $x_{(1)}$. The other classes are determined by successively adding up the class intervals. Alternatively one may start with the upper class limit (luokan yläraja) of the highest category exceeding $x_{(n)}$ and determine the other class limits by successively subtracting the class intervals.

6. Assign each observation to its corresponding class and find the count for each class. Each observation must be assigned to one and only one category.

Example:

Statistics Finland provided the average selling price for flats in 416 municipalities for 2002 in €/m^2. The numbers have been provided with a precision of 1€/m^2, such that d=1€/m^2. The smallest price was 336€/m^2 and the largest 2166€/m^2. The sellings price range is therefore 1830€/m^2. A suitable number of categories may be determined from

$$\sqrt[3]{416} \approx 7.5 \;\leq\; k \;\leq\; \log_2 416 \approx 8.7$$

as $k \approx 8$. Now the minimum class interval is $1830/8 = 228.75$, which suggests $c = 230$. Setting the lower class limit of the first category to 330, one obtains the following class limits (together with the frequency distribution, for later use):

Price(€/m^2)	f_i	$100p_i$
330 − 559	49	11.8
560 − 789	195	46.9
790 − 1019	118	28.4
1020 − 1249	38	9.1
1250 − 1479	11	2.6
1480 − 1709	1	0.2
1710 − 1939	1	0.2
1940 − 2169	3	0.7
Sum:	416	100.0

The rounding precision d is visible from the frequency table as the difference between the lower limit of class i and the upper limit of the preceeding class. As a matter of fact, average prices reported e.g. in the first class (330–559), were indeed in the range between 329,5 and 559,5 because of rounding. These are called the real class limits (todelliset luokkarajat), which can be found from the frequency table as the midpoints between the class limits shown and the reported class limits of the nearby classes. Denoting the lower (higher) real class limit of category i as L_i (U_i), the class interval size c_i of category i may be calculated as $c_i = U_i - L_i$.

The center of class i, $m_i = (L_i + U_i)/2$ is called its class mark (luokkakeskus). It is the best estimate of the average value in that class if nothing else is known.

If discrete quantitative variables may attain a lot of different values, we may categorize them as if they were continuous variables.

For variables measured at least on ordinal scale we may be interested in how many units belong to class E_i or lower. This information is called cumulative frequency (summafrekvenssi/ kumulatiivinen frekvenssi) F_i and simply obtained by summing up all frequencies from class E_1 to E_i, that is:

$$
\begin{aligned}
F_1 &= f_1 \\
F_2 &= f_1 + f_2 = F_1 + f_2 \\
F_3 &= f_1 + f_2 + f_3 = F_2 + f_3 \\
&\vdots \\
F_i &= \sum_{j=1}^{i} f_j = f_1 + \cdots + f_i = F_{i-1} + f_i \\
&\vdots \\
F_k &= F_{k-1} + f_k = \sum_{j=1}^{k} f_j = n.
\end{aligned}
$$

The relative cumulative frequency (suhteellinen summafrekvenssi) P_i is similar to the relative frequency p_i defined as $P_i := F_i/n$, that is:

$$P_i = \frac{F_i}{n} = \frac{\sum_{j=1}^{i} f_j}{n} = \sum_{j=1}^{i} \frac{f_j}{n} = \sum_{j=1}^{i} p_j.$$

The procentual cumulative frequency $100P_i$ (prosentuaalinen summafrekvenssi) is defined analogous to the procentual frequency p_i as:

$$100P_i := 100 \cdot \sum_{j=1}^{i} p_j.$$

Example. Below is the preceeding table suplemented with cumulative frequencies (absolute + procentual), upper and lower real class limits, and class marks:

Price(€/m^2)	f_i	$100p_i$	F_i	$100P_i$	L_i	U_i	m_i
330 – 559	49	11.8	49	11.8	329.5	559.5	444.5
560 – 789	195	46.9	244	58.7	559.5	789.5	674.5
790 – 1019	118	28.4	362	87.0	789.5	1019.5	904.5
1020 – 1249	38	9.1	400	96.2	1019.5	1249.5	1134.5
1250 – 1479	11	2.6	411	98.8	1249.5	1479.5	1364.5
1480 – 1709	1	0.2	412	99.0	1479.5	1709.5	1594.5
1710 – 1939	1	0.2	413	99.3	1709.5	1939.5	1824.5
1940 – 2169	3	0.7	416	100	1939.5	2169.5	2054.5
Sum:	416	100.0					

Note: The number (fraction) of units belonging to class E_{i+1} and above may be determined from $\bar{F}_i := n - F_i$ and $\bar{P}_i := 1 - P_i$, respectively.

3.2. Graphical Presentation

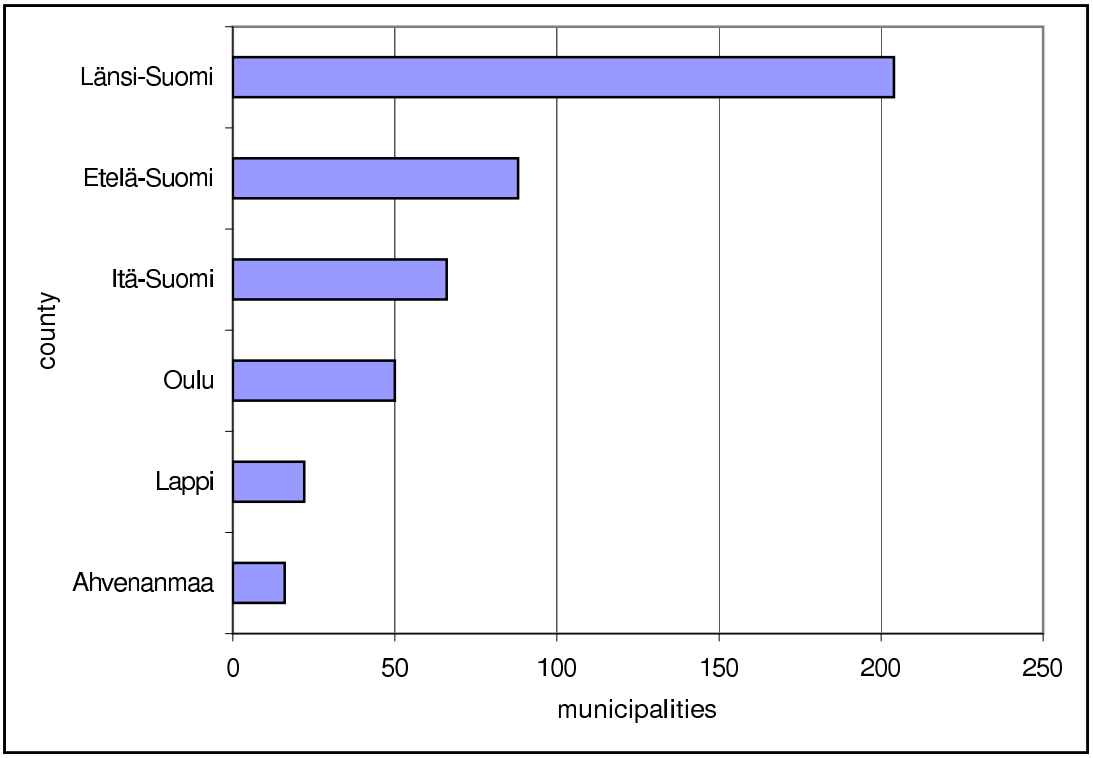
Frequency distributions may also be presented graphically. An often used representation method are so called bar charts (pylväskuviot). There is one bar for each class E_i and its area (or length, if all bars have the same width) indicates its (absolute, relative or procentual) frequency.

Horizontal bar charts (vaakapylvskuviot) are used for qualitative variables. If the variable is measured on nominal scale, bars are ordered according to their frequency (largest frequency first). Variables measured on ordinal scale are presented in the order of their ranking. It is customs that the bars in a horizontal bar chart do not touch each other.

Example. Consider again the frequency distribution of municipalities over Finlands counties presented earlier:

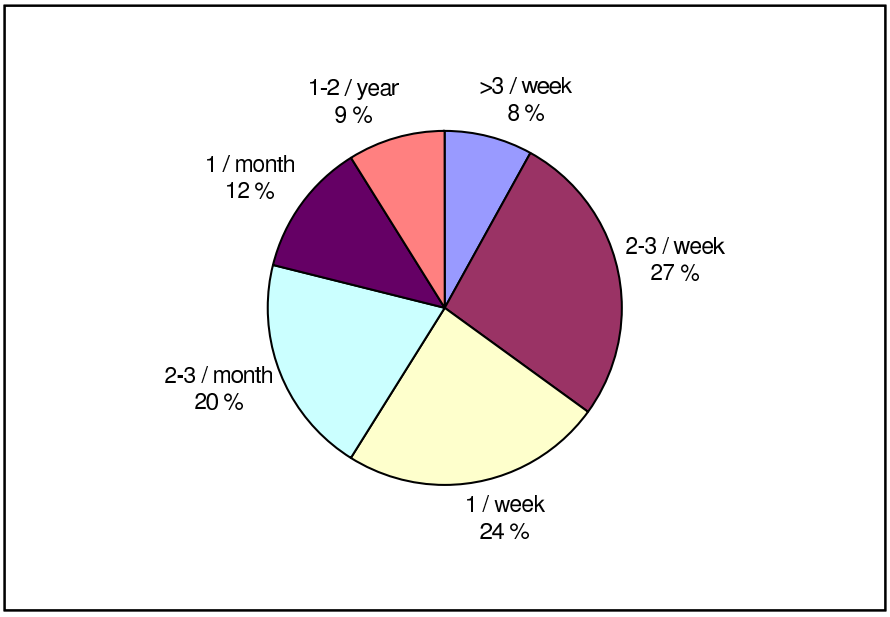
County	f_i	p_i	$100p_i$
Etelä-Suomi	88	0.197	19.7
Länsi-Suomi	204	0.457	45.7
Itä-Suomi	66	0.148	14.8
Oulu	50	0.112	11.2
Lappi	22	0.049	4.9
Ahvenanmaa	16	0.036	3.6
Sum	446	1.000	100.0

Below is the corresponding bar chart:



Pie charts (sektoridiagrammi, ympyräkuvio, piirakkakuvio) are mostly used for procentual frequencies, as they stress the relative count of a class as a part of the whole. However, small differences between frequencies are easier spotted in bar charts.

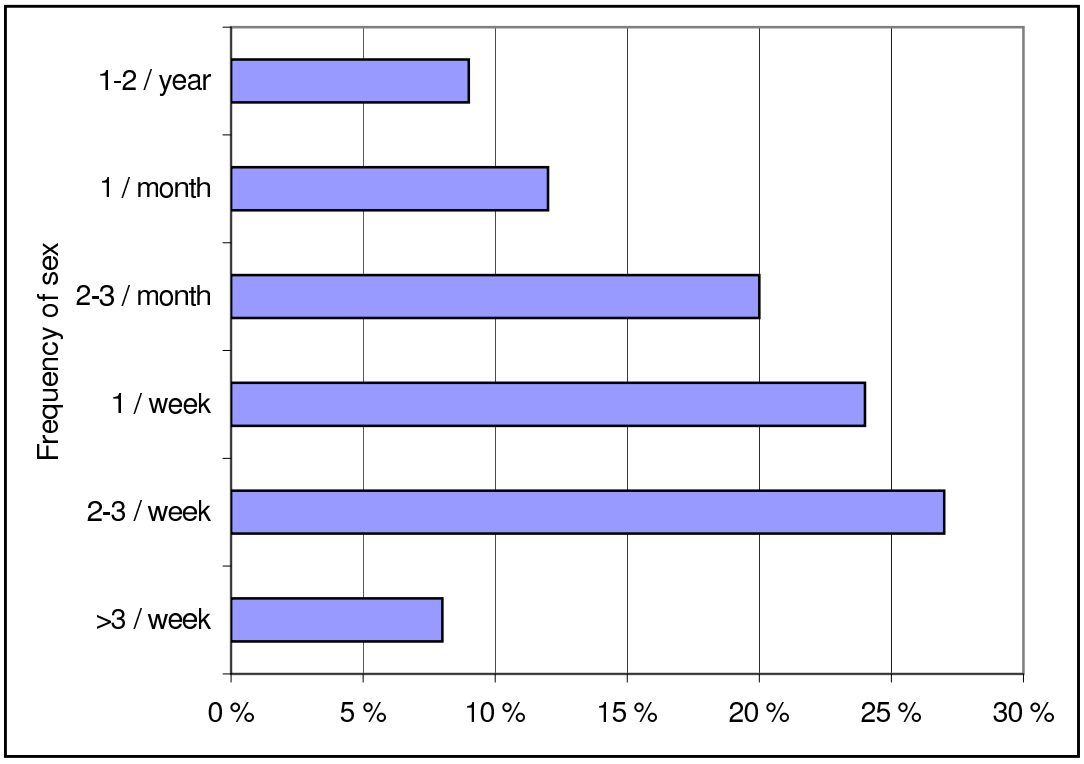
Example. Frequency of sex for sexually active heterosexual US citizens according to Blanchflower and Oswald*:



Note: The area of each segment is proportional to the frequency of the corresponding class.

*Money, Sex and Happiness: An Empirical Study, Scandinavian Journal of Economics (2004)

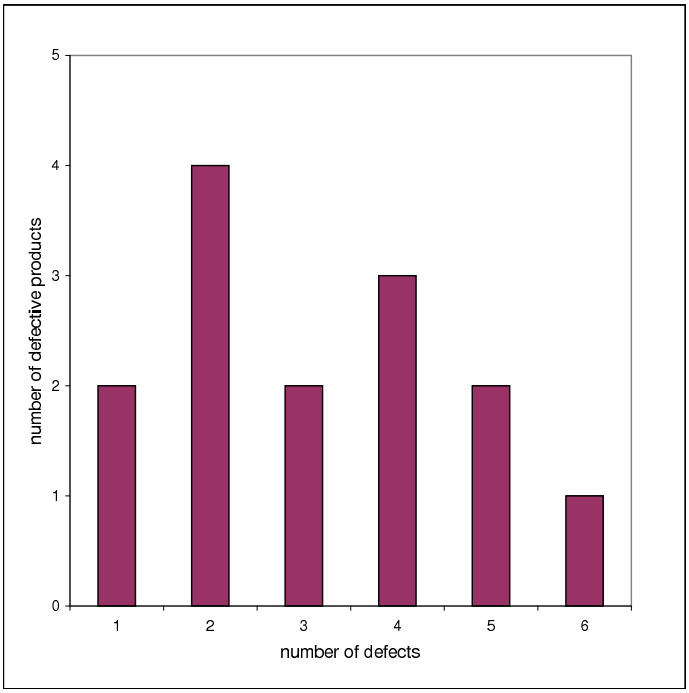
Example. (continued)



Frequency distributions of discrete variables may be displayed in vertical bar charts (janakuviot/ pystypylväskuviot) with the variables' values below the bars representing their corresponding frequencies.

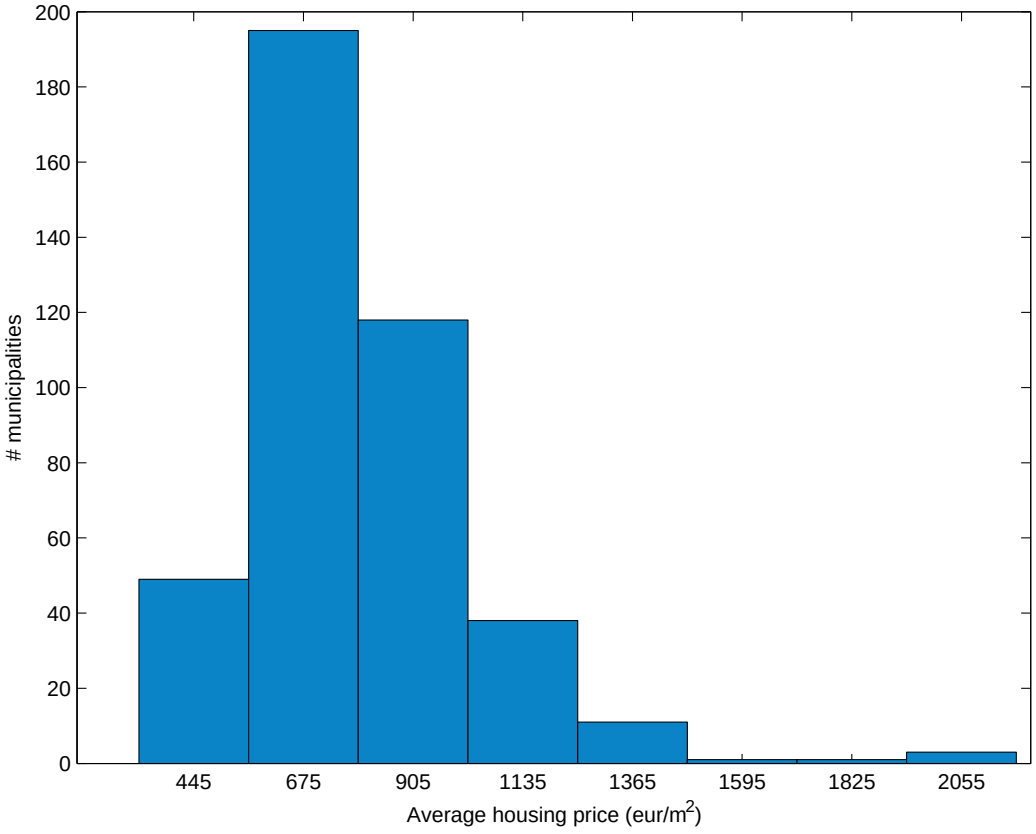
Example. Frequency table and bar chart of the number of defects in defective products.

# defects	f_i
1	2
2	4
3	2
4	3
5	2
6	1



Frequency distributions of continuous variables are displayed in histograms (frekvenssihistogramma). These are vertical bar charts, where the width of the bars are given by the lengths of the class intervals c_i and their height by their corresponding frequency f_i. The edges of the left and right boundaries of the bars are given by their lower and upper real class limits. That is, in contrast to bar charts of discrete and qualitative variables, the bars do touch each other.

Example. Histogram of average housing prices:



Note: The class marks are shown below the bars, rounded to some readable number.

Recall that the *area* in each bar is proportional to each classes count, which coincides with a bars *height* only if all bars have the same width. This implies that we may not include any frequencies on the y-axis of histograms based on an irregular categorisation, since for diverging class interval widths c_i, the same height does not necessarily correspond to the same frequency.

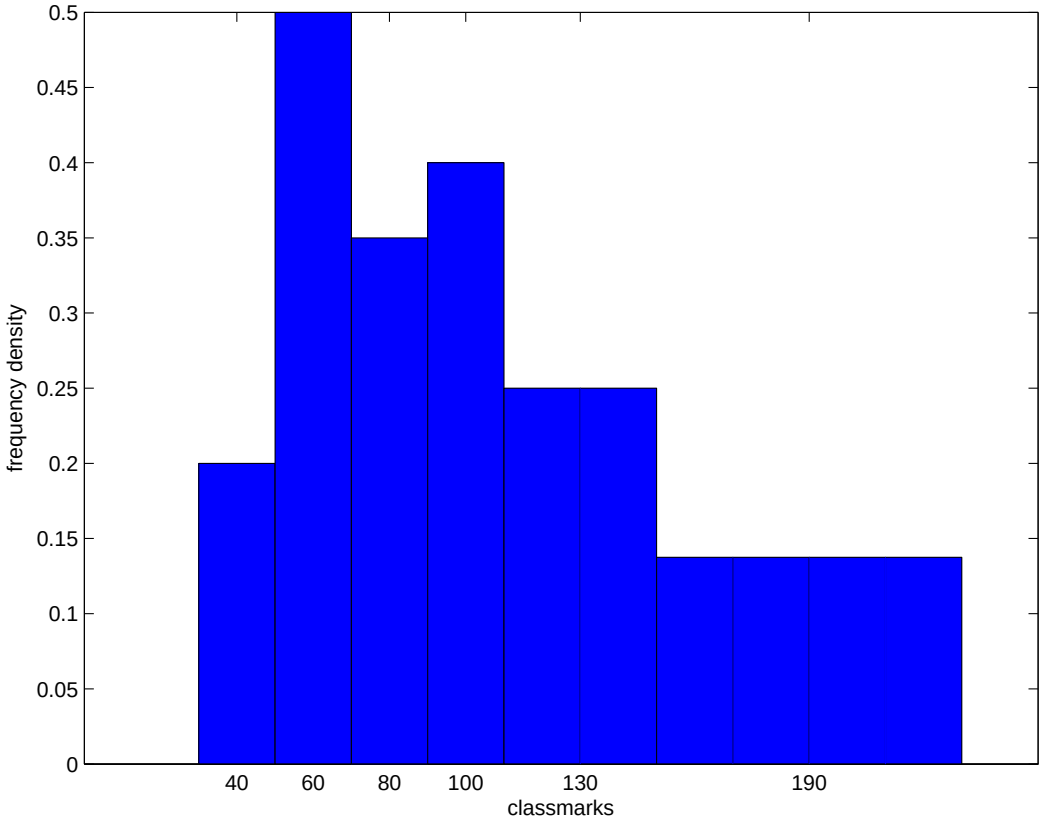
For an irregularly categorized variable it is therefore advisable to use a relative histogram (suhteellinen histogramma) instead, where the absolute frequencies f_i are replaced by so called frequency densities (havaintotiheys), which is each classes frequency divided by its class interval size (f_i/c_i).

In relative histograms the height of each bar represents its frequency density, while each bars area is equal to its classes count, regardless of categorisation. The total area under relative histograms is therefore equal to the total number of observations.

Consider the following example:

Classes	m_i	f_i	c_i	f_i/c_i
30 – 49	39.5	4	20	0.2
50 – 69	59.5	10	20	0.5
70 – 89	79.5	7	20	0.35
90 – 109	99.5	8	20	0.4
110 – 149	129.5	10	40	0.25
150 – 229	189.5	11	80	0.1375
Sum:		50		

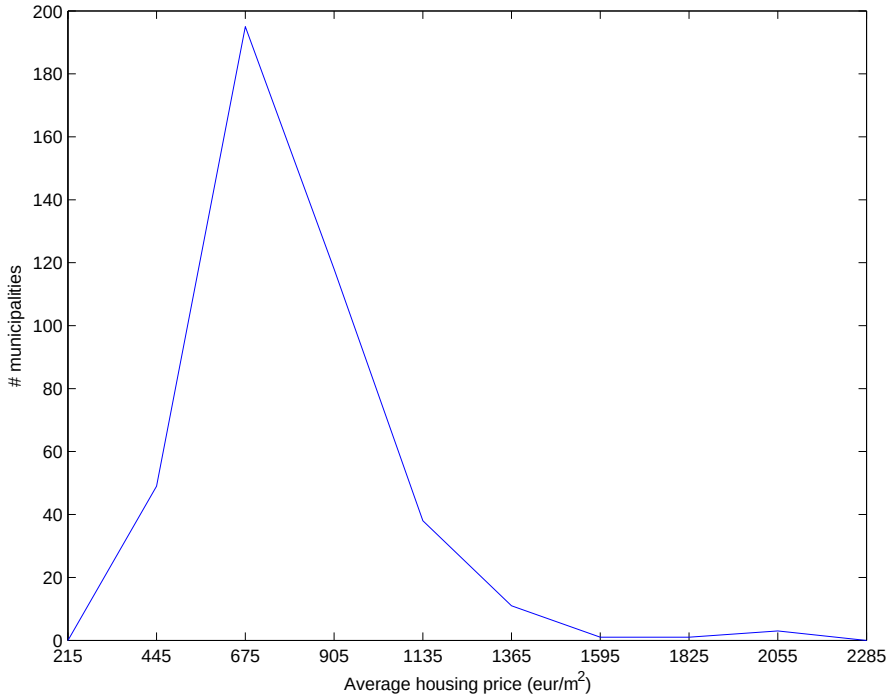
The relative histogram looks like this:



Distributions of continuous variables may also be displayed in so called frequency polygons (frekvenssimonikulmio/ frekvenssimurtoviiva). For each classmark there is a dot at the height of the classes frequency, which are connected by straight lines. Frequency polygons always start and end at the x-axis. For that purpose, the class marks of one additional class below the lowest category and another additional class above the highest catgory are also displayed in the plot.

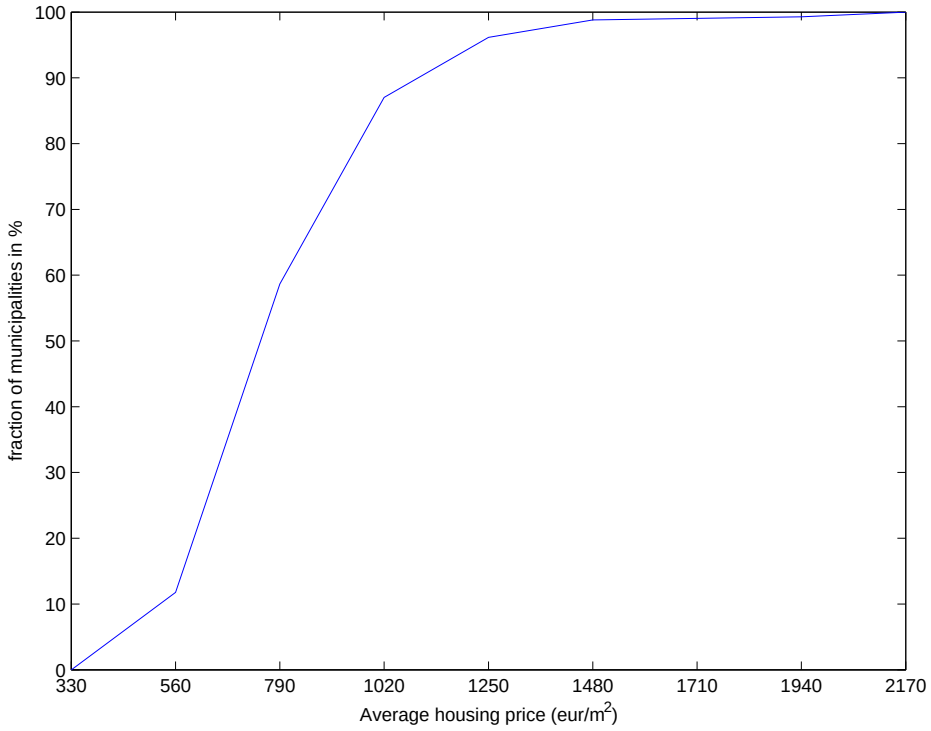
Note: Like for histograms, we replace frequencies by frequency densities in the case of irregular classifications.

Example. Average housing prices again:



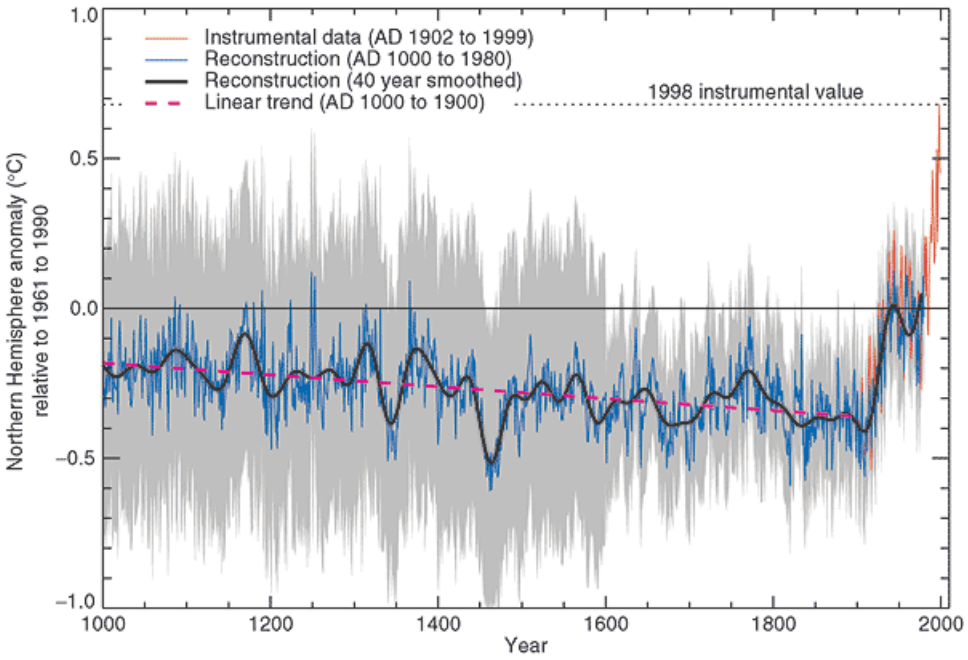
The cumulative (relative, procentual) frequencies of continuous variables may be displayed in so called ogives (suhtellinen, prosentuaalinen summakäyrä). Each classes cumulative frequency is plotted at its upper real class limit, and the dots are again connected with straight lines. Ogives start always at the x-axis (at the lowest categories lower real class limit) and finnish with the number of observations n (or 1 for relative resp. 100% for procentual frequencies).

Example. Average housing prices again:



Line graphs (viivakuviot) are mainly used to plot time series (aikasarjat), that is the evolution of a variables value through time. Time is plotted on the x-axis and the variables value on the y-axis. Again, the observation values are connected by straight lines.

Example. Temperatures on the northern hemisphere during the last thousand years:*



*Intergovernmental Panel on Climate Change (2001): Climate Change 2001: The Scientific Basis, available at: http://www.grida.no/climate/ipcc_tar/wg1/index.htm

3.3. Statistics of 1-Dimensional Distributions

3.3.0 The 5-number summary and boxplots

Consider the following frequency distribution of the age of 19 students in a statistics class:

Age	19	20	21	22	23	25	26	29	42	46
f	1	4	5	2	2	1	1	1	1	1

Suppose we want to know the age such that about p percent of the students are below that age. One way to do this is to calculate the cumulative procentual frequencies of each age, plot the ogive, look up the percentage p at the vertical axis and find the corresponding age at the horizontal axes of the ogive. But there is an easier way to do it. We may also just write down the observations in increasing order and take the number right to the left p percent of the numbers. A percentile (prosenttipiste, prosentin fraktiili) is almost defined like that. However, in order to have a measure that also works where the discrete numbers are just classmarks of some continuous variable (as is actually the case with students ages), we move only p percent of an observation, rather than a full observation to the right of the left p percent of the numbers.

To sum up: The p'th percentile (prosenttipiste) of a group of numbers is that value below which lie about p% of the observations. The position of the p'th percentile in ordered data is $(n+1)p/100$, where n is the number of observations.

Example. 25'th percentile of the students age: First we have to write down the students ages in increasing order, that is,

19, 20, 20, 20, 20, 21, 21, 21, 21, 21, 22, 22, 23, 23, 25, 26, 29, 42, 46.

The observation in position k is called the k'th order statistics (järjestystunnusluku) $x_{(k)}$. We want to find the $p = 25$'th percentile in a population of $n = 19$ observations, that is we are looking for position number:

$(19+1) \cdot 25/100 = 5$ (5'th order statistics).

Now the entry at position number 5 is 20 ($x_{(5)} = 20$), that is, the 25'th percentile of the students age is 20, meaning that about 25% of the students are 20 years old or less.

Example. (continued) Suppose there would have been yet another student at the age of 46, such that the order statistics would be:

19, 20, 20, 20, 20, 21, 21, 21, 21, 21, 22, 22, 23, 23, 25, 26, 29, 42, 46, 46.

Then we would have been looking for position number $(20+1)\cdot 25/100 = 5.25$, meaning $x_{(5)}$ plus 0.25 times the difference between $x_{(6)}$ and $x_{(5)}$, that is,

$$20 + 0.25 \cdot (21 - 20) = 20.25.$$

This is how the 25'th percentile is defined, but of course, on a discrete scale it still means that about 25% of the students are 20 years old or below.

The 25'th percentile is also known as the first quartile Q_1 (alakvartiili), because one quarter of the observations are below it.

Other important percentiles include:

The median M (mediaani), which is the 50'th percentile, that is, 50% lie below and 50% lie above that value.

The third quartile Q_3 (yläkvartiili), which is the 75'th percentile, that is, three quarters of observations lie below and one quarter above that value.

Example. (continued)
Consider again our original order statistics: 19, 20, 20, 20, 20, 21, 21, 21, 21, 21, 22, 22, 23, 23, 25, 26, 29, 42, 46.

Then:

$$k_M = (19+1)\frac{50}{100} = 10 \;\Rightarrow\; M = x_{(10)} = 21,$$

$$k_{Q_3} = (19+1)\frac{75}{100} = 15 \;\Rightarrow\; Q_3 = x_{(15)} = 25,$$

meaning that about one half of the students are 21 years or less, and about one quarter of the students are more than 25 years old.

Note: All percentiles (that is, in particular the median, and the lower and upper quartile) are only defined for variables measured on ordinal scale and above, but not for variables measured on nominal scale.

Note: Percentage measures based upon percentiles are not precise whenever their is more than one count in any one class.

The median gives a good idea about the center of the distribution. In order to get an idea about its spread, one may apply the interquartile range (kvartiilivälin pituus) defined as the difference between the upper and the lower quartile of the distribution, that is, $Q_3 - Q_1$. It tells the range within which the 50% most central observations are located. The quartile deviation (kvartiilipoikeama) is defined as $\frac{1}{2}(Q_3 - Q_1)$.

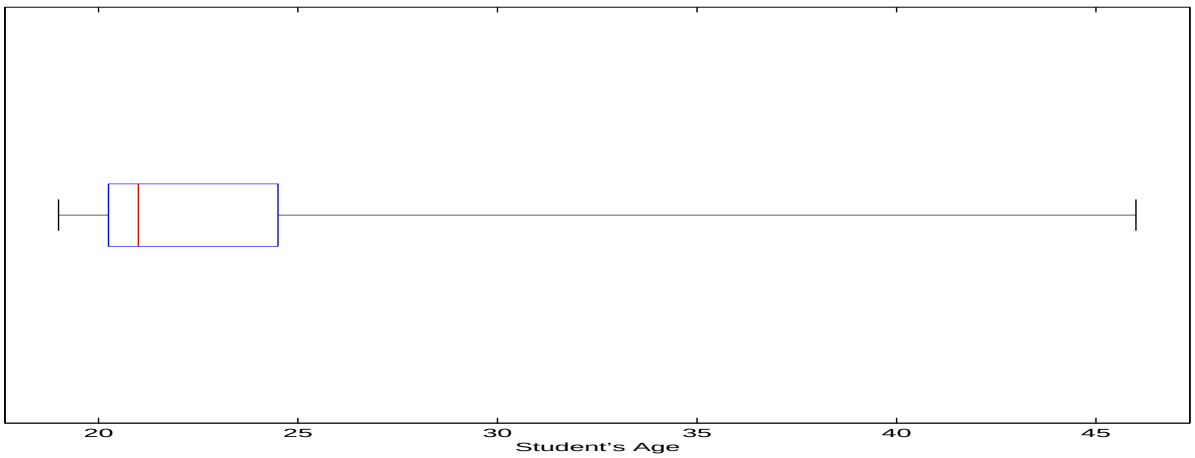
Example. (continued)
The interquartile range in our distribution of student ages is $Q_3 - Q_1 = 25 - 20 = 5$ years. The quartile deviation is $\frac{1}{2}(Q_3 - Q_1) = 2.5$ years.

A reasonable summary of a frequency distribution may then be given by the set of the following five numbers:

1. Smallest Observation (Minimum)
2. First Quartile (Q_1)
3. Median (M)
4. Third Quartile (Q_3)
5. Largest Obervation (Maximum)

Example. (Student's age continued)
The five-number summary of our Student's age distribution reads 19 20 21 25 46.

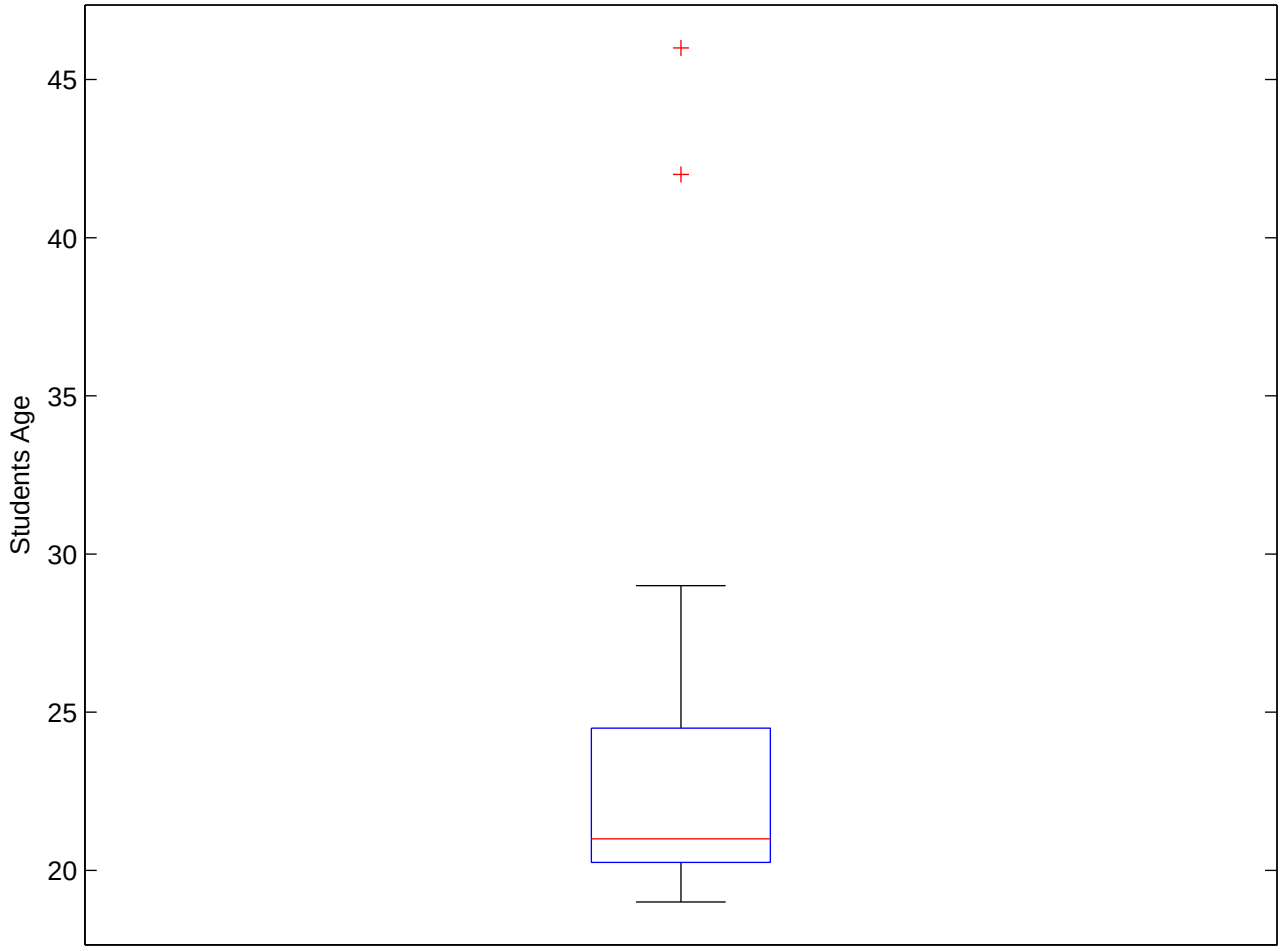
This may be displayed in a so called boxplot:



Even more information of the original distribution is preserved in modified boxplots where the so called whiskers (the lines above and below the box indicating the interquartile range) do not necessarily extend to the most extreme values, but only to the largest and smallest values not further than 1.5 times the interquartile range away from the lower resp. upper quartile. All observations outside that range are called (suspected) outliers (vieras/poikkeava havainto) and plotted separately.

Example. (Student's age continued)
Recall that the interquartile range (IQR) was 5 years, $Q_1 = 20$, and $Q_3 = 25$. The whiskers of the modified boxplot may therefore not exceed $20 - 1.5 \cdot 5 = 12.5$ from below and $25 + 1.5 \cdot 5 = 32.5$ from above. The smallest observation above 12.5 is 19, and the largest observation below 32.5 is 29. The students of age 42 and 46 are therefore regarded as outliers, and we obtain the following modified boxplot:

Example. (...continued)



Note: The range how far the whiskers of the boxplot may extend is sometimes defined differently. 1.5×IQR is the standard, but it pays to check how it is defined in the particular chart you're looking at.