3.3.1 Measures of Location (Keskiluvut)

Measures of Central Tendency or Location (keskilukuja), as the name says, are supposed to tell us where the center of the distribution is. We already know the first one of those, the 50'th percentile of the distribution:

The Median (mediaani)

As mentioned earlier, the median is the value below and above which there are equally many observations. The earlier used percentile formula for finding its value from the order statistics $x_{((n+1)p/100)}$ simplifies depending upon the numbers of observations n to:

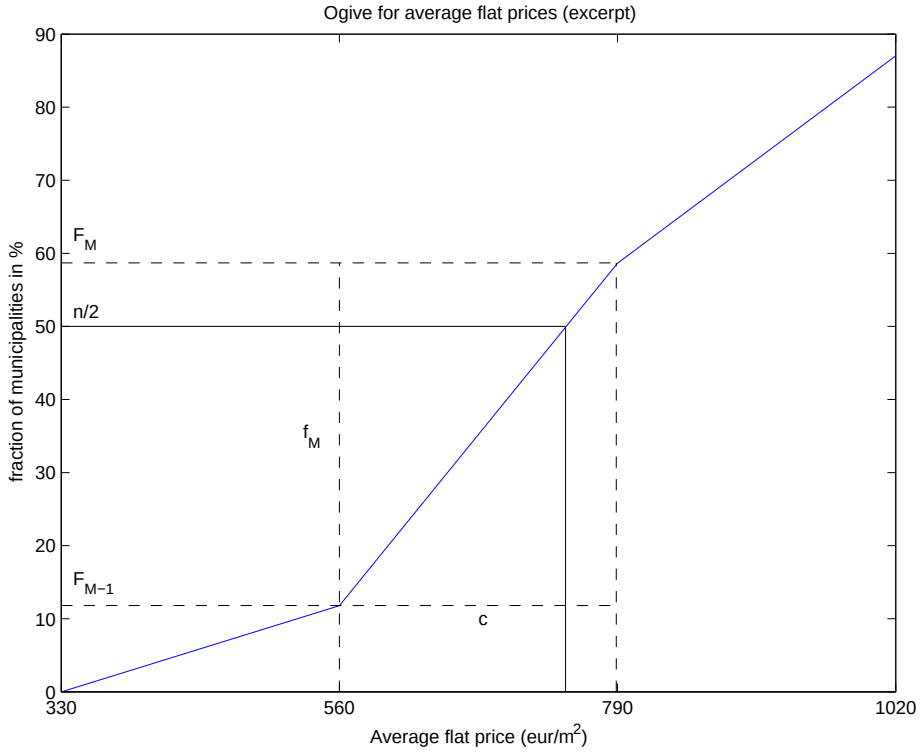
1. $M = x_{(k)}$ with $k = \frac{n+1}{2}$ for n odd, and
2. $M = \frac{x_{(k)}+x_{(k+1)}}{2}$ with $k = \frac{n}{2}$ for n even.

The median of a categorized continuous variable is estimated from

$$M = L_M + c_M \cdot \frac{n/2 - F_{M-1}}{f_M}, \quad \text{where}$$

$L_M =$ real lower limit of the median class,
$c_M =$ interval size of the median class,
$n =$ number of observations (not classes!)
$F_{M-1} =$ cumulative Frequency of the class preceeding the median class,
$f_M =$ frequency of the median class.

The median class is the first class satisfying $F_i \geq \frac{n}{2}$ ($\Leftrightarrow 100P_i \geq 50$).



Example. Recall our frequency table of average flat prices in 416 Finnish municipalities:

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

The median class is the second category (560–789€/m^2), because its cumulative freqency 244 exceeds 416/2=208 (or 58.7%≥50%.)

The median flat selling price may be therefore estimated as:

$$M = 559.5 + 230 \cdot \frac{208 - 49}{195} \approx 747€/\text{m}^2.$$

The same value may be obtained graphically by looking up a cumulative procentual frequency of 50% in an ogive and looking for the corresponding housing price.

The Mode (moodi/ tyyppiarvo)

The mode is simply the value or class of a distribution with the highest count. Unlike the median, which like all other percentiles was only defined for variables measured on ordinal scale and above, the mode is also defined for variables measured on nominal scale.

Example.
Consider again our frequency distribution of municipalities over Finlands counties:

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

The mode of the distribution is the county of Länsi-Suomi, because it has the highest count.

For categorized variables measured on interval scale or higher, the classmark of the category with the highest count (moodiluokka) may be interpreted as the mode of the distribution.

Example.
Recall our frequency table for average flat prices in Finland:

Price(€/m^2)	f_i	m_i
330 – 559	49	444.5
560 – 789	195	674.5
790 – 1019	118	904.5
1020 – 1249	38	1134.5
1250 – 1479	11	1364.5
1480 – 1709	1	1594.5
1710 – 1939	1	1824.5
1940 – 2169	3	2054.5

The class with the highest count is the second category (560–789€/m^2). The classmark of this class (674.5€/m^2 $\approx$ 675€/m^2) may be interpreted as the mode of the distribution.

Note: The mode of a distribution is not necessarily unambiguously defined, as there may be several values or classes with equally highest frequency.

Example. Recall our 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

The mode observation is 21 years, because it is the age with the highest count. But suppose there would have been yet another student at the age of 20. Then there would have been two modes of the distribution, 20 and 21 years, since both would have had an equally highest count of 5 students.

Distributions with only one unambiguously defined mode are called unimodal (yksihuippuinen), otherwise they are called multimodal (monihuippuinen).

The Mean (keskiarvo)

The arithmetic mean (aritmeettinen keskiarvo) is defined as that value of a statistical variable, such that the distances of all observations from that value average out to zero. In other words, it is defined such that the sum of its distances from all observations smaller than the mean is just as large as the sum of its distances from all observations greater than the mean. The arithmetic mean is only defined for variables measured on interval and ratio scale.

Denoting the observations of a statistical variable x with x_i, $(i = 1, \ldots, n)$, the arithmetic mean $\bar{x}$ is given by:

$$\bar{x} := \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

This may be seen as follows. We start with the defining property that the sum of all distances from the mean should average out to zero:

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x}) = 0.$$

Therefore:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = \sum_{i=1}^{n} x_i - n\bar{x} \stackrel{!}{=} 0$$

$$\Leftrightarrow \quad n\bar{x} = \sum_{i=1}^{n} x_i \quad \Leftrightarrow \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Example. (Student's age continued)

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

$$\bar{x} = \frac{19 + 4 \cdot 20 + 5 \cdot 21 + \cdots + 42 + 46}{19} = \frac{462}{19} \approx 24\text{yrs.}$$

For categorized variables the arithmetic mean may be calculated as:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} f_i \, m_i, \quad \text{where}$$

$$\begin{aligned} n &= \text{ number of observations,} \\ k &= \text{ number of categories,} \\ f_i &= \text{ frequency of category } i, \\ m_i &= \text{ classmark of category } i. \end{aligned}$$

Example. (average flat prices continued)

Price(€/m^2)	f_i	m_i
330 – 559	49	444.5
560 – 789	195	674.5
790 – 1019	118	904.5
1020 – 1249	38	1134.5
1250 – 1479	11	1364.5
1480 – 1709	1	1594.5
1710 – 1939	1	1824.5
1940 – 2169	3	2054.5
Sum:	416	

$$\bar{x} = \frac{49 \cdot 444.5 + 195 \cdot 674.5 + \cdots + 3 \cdot 2054.5}{416} \approx 788 \frac{€}{\text{m}^2}.$$

Mathematical Properties of the Arithmetic Mean

1. If the observations on n statistical units are grouped into k goups of size $n_1, n_2, \ldots, n_k$ with respective arithmetic means $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$, then the overall arithmetic mean may be calculated from

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} n_i \, \bar{x}_i.$$

Example.
The average hourly wage in a company with 400 women and 500 men is 26.58€ for women and 34.59€ for men. The overall average hourly wage is therefore:

$$\bar{x} = \frac{400 \cdot 26.58 + 500 \cdot 34.59}{400 + 500} = 31.03€\,.$$

2. The arithmetic mean of a linear function y of a statistical variable x, $y = a + b \cdot x$ say, where are a and b are known constants, that is, $y_i = a + b \cdot x_i$ with the same a and b for all $i = 1, \ldots, n$, may be directly calculated from $\bar{y} = a + b \cdot \bar{x}$.

Note.
The property $\overline{f(x)} = f(\bar{x})$ holds only for linear transformations of x, that is functions of the form $f(x) = a + b \cdot x$. In general: $\overline{f(x)} \neq f(\bar{x})$.

Sensitivity to Outliers

The arithmetic mean is the most complete description of a distributions center because it takes all observations into account. (Recall that the mode accounted only for the observation(s) with the highest count, and the median only for the central order statististics.) This strength may be also considered a weakness if one is not sure about ones most extreme observations, since these have a disproportionate large effect upon the mean as compared to the mode and the median.

Example. (Student's age continued)

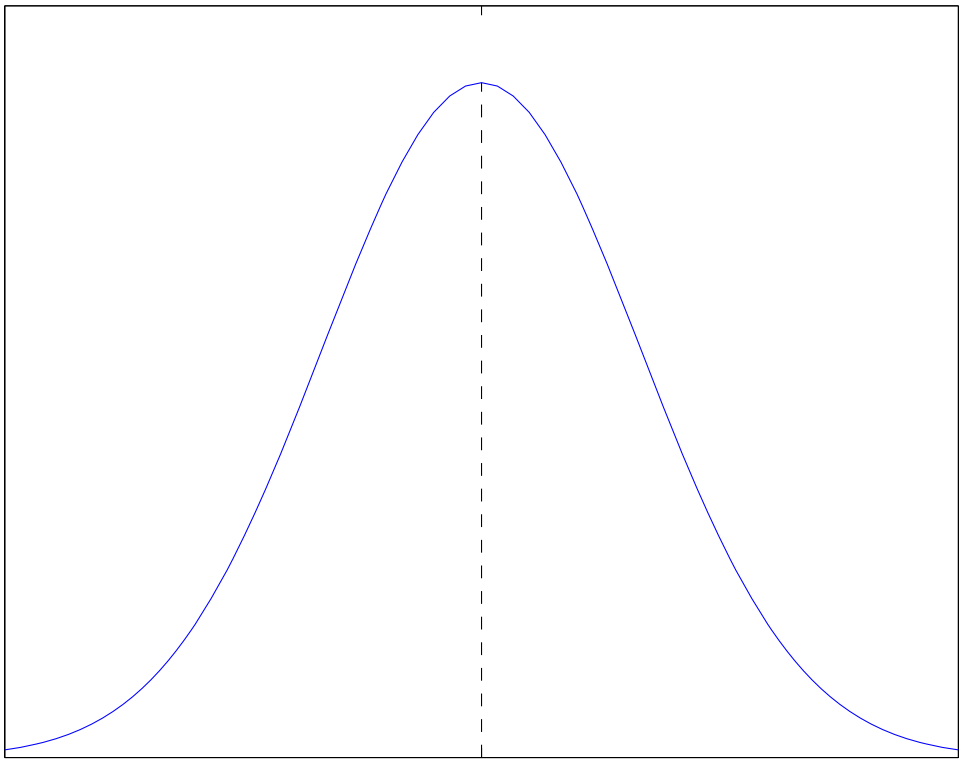
Age	19	20	21	22	23	25	26	29	42	46
f	1	4	5	2	2	1	1	1	1	1

Suppose we want to remove the two oldest students, because for some reason we don't find them representative for our statistics classes in general. Then our new measures of central tendency become:

Mode:	$x_{\max} = 21$	(original: 21)
Median:	$x_{(9)} = 21$	(original: 21)
Mean:	$374/17 = 22$	(original: 24.3)

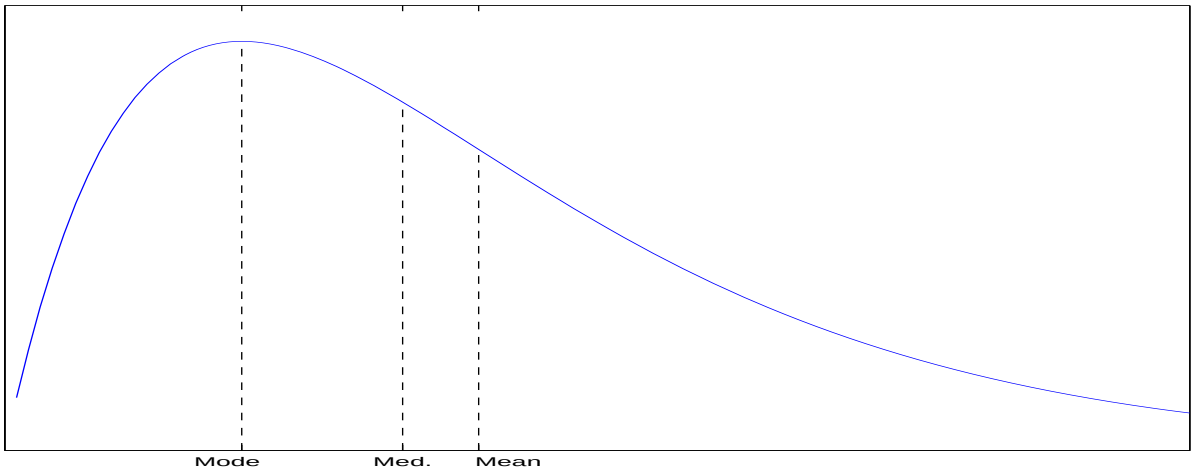
Location of Center in Unimodal Distributions

The fact that the different measures of location have differerent sensitivities towards extreme observations implies that they need not be the same even for unimodal distributions. However, for symmetric (symmetrinen) unimodal distributions, that is when for each observation on the left of the center there is exactly one balancing observation to the right of the center, then the mode, median and mean will always coincide:



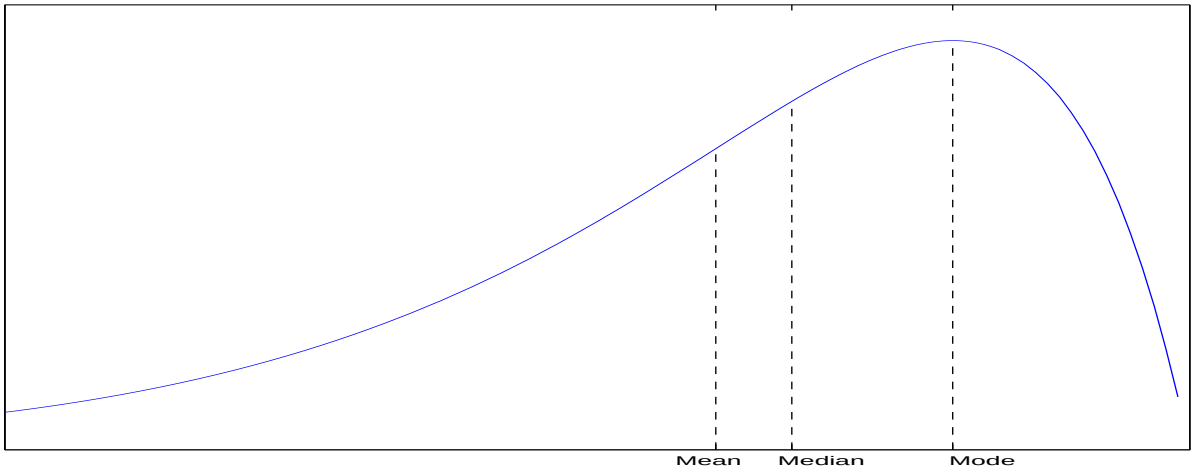
Distributions with a long right tail are called right-skewed (oikealle viinoutta). Usually:

$$\text{Mode} \leq \text{Median} \leq \text{Mean}.$$



Distributions with a long left tail are called left-skewed (vasemmalle viinoutta). Usually:

$$\text{Mean} \leq \text{Median} \leq \text{Mode}.$$



The Geometric Mean (geometrinen keskiarvo)

The geometric mean is used to express the average change per time unit of a variable measured on ratio scale. It can be only calculated for time series of purely positive observations and is given by

$$G_x = \sqrt[n]{x_1 \cdot x_2 \cdots x_n},$$

where G_x denotes the geometric mean, and $x_1, \ldots, x_n$ denote the observed values in a time series of n observations.

Example. The price of a product rises by 50% in the first year, by a factor of 5 in the second year, and by a factor of 4 in the third year. On average the price rises then by a factor of $G = \sqrt[3]{1.5 \cdot 5 \cdot 4} \approx 3.1$ per year.

Note: (logarithmic transformation)
Taking the logarithm of a geometric mean is the same as taking the arithmetic mean of the logarithm of the individual observations, that is:

$$\overline{\log x_i} = \log G_x \quad \Rightarrow \quad G_x = \exp(\overline{\log x_i}).$$

This may be seen as follows:

$$\begin{aligned}
\log G_x &= \log \sqrt[n]{x_1 \cdot x_2 \cdots x_n} \\
&= \log(x_1 \cdot x_2 \cdots x_n)^{1/n} \\
&= \log\left(x_1^{1/n} \cdots x_n^{1/n}\right) \\
&= \log x_1^{1/n} + \cdots + \log x_n^{1/n} \\
&= \frac{1}{n}\log x_1 + \cdots + \frac{1}{n}\log x_n \\
&= \frac{1}{n}(\log x_1 + \cdots + \log x_n) \\
&= \overline{\log x_i}.
\end{aligned}$$

Example. (continued)

$$\begin{aligned}
\overline{\log x_i} &= \frac{1}{3}(\log 1.5 + \log 5 + \log 4) \\
&= \frac{1}{3}(0.405 + 1.609 + 1.386) \\
&\approx 1.13.
\end{aligned}$$

$$G_x = \exp(\overline{\log x_i}) = e^{1.13} \approx 3.1.$$

The Harmonic Mean (harmoninen keskiarvo)

The harmonic mean is often appropriate when an average of rates is desired. Like the geometric mean, it is defined only for time series of purely positive observations of variables measured on ratio scale. The harmonic mean H_x of a series of n observations $x_1, \ldots, x_n$ is given by

$$H_x = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i} \right)^{-1}.$$

Example. A car travels half of the distance at 40 kilometres per hour and the other half at 60 kilometres per hour. The average speed is then:

$$H = \frac{2}{\frac{1}{40} + \frac{1}{60}} = 48\text{km/h}.$$

Note. In the example above, the need for calculating the harmonic instead of the arithmetic mean arose because the distance, over which the average was to be taken, is the reciprocal of the unit in which speed is measured: We measure speed in distance traveled per time unit used (e.g. km/h), not in time used per distance traveled (e.g. h/km), as it could equally well be defined. Now, because we were asked to average over the "wrong" unit *distance*, the harmonic mean was the right average to be used. If instead, the car would have traveled half of the *time* at 40km/h and the other half at 60km/h, the arithmetic mean $\bar{x} = 50$km would have provided the right answer.

Note. For the same set of observations, the harmonic mean H_x, the geometric mean G_x, and the arithmetic mean $\bar{x}$, are always related by

$$H_x \leq G_x \leq \bar{x},$$

where the equal signs apply only if all observations $x_1, \ldots, x_n$ are identical.