3.3.2 Measures of Variability (hajontaluvut)

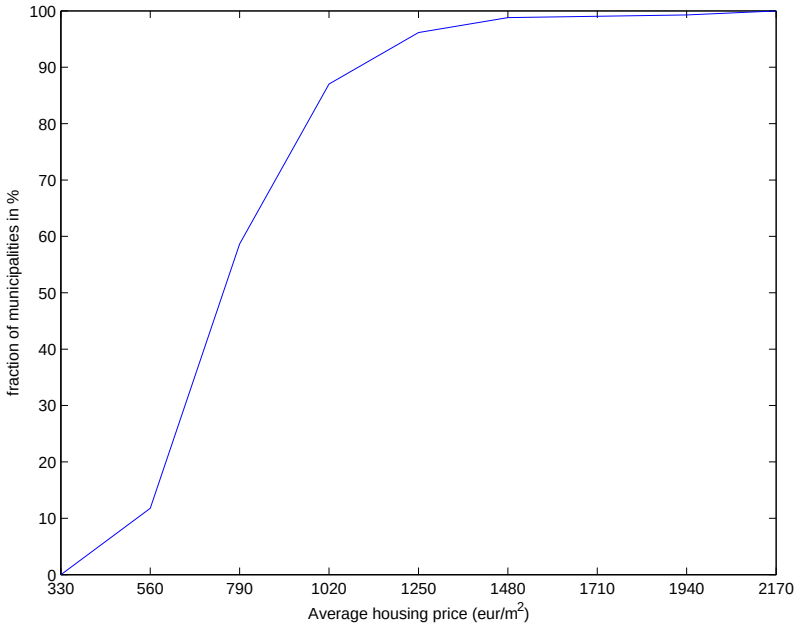
Range Intervals (vaihteluvälejä)

We know already the range interval (vaihteluväli) of statistical variables as the interval between the smallest and largest observations $(x_{(1)}, x_{(n)})$. It is only defined for variables measured on ordinal scale and above. If we know only the classes, we identify the lowest order statistics $x_{(1)}$ with the lower class limit of the first class, and the highest order statistics $x_{(n)}$ with the upper class limit of the highest class.

For quantitative variables measured on interval and ratio scale the range is defined as the difference between the highest and lowest order statistics $w = x_{(1)} - x_{(n)}$ (vaihteluvälin pituus). Again, if we know only the classes, it is estimated as the difference between the upper class limit of the highest category and the lower class limit of the first category.

Example. In our example of average housing prices in Finland the smallest price was 336€ /m^2 and the largest 2166€ /m^2. The range (vaihteluvälin pituus) of average flat prices was therefore w = 2166€ /m^2−336€ /m^2 = 1830€ /m^2.

Suppose we were given only the grouped data as presented in the ogive below:



Then we could not tell what the true range is, but would have estimated it from the upper limit of the highest class and the lower limit of the lowest class as

$$w = 2170€ /\text{m}^2 - 330€ /\text{m}^2 = 1840€ /\text{m}^2.$$

A disadvantage of the range as a measure of spread is that it takes only the most extreme observations into account. From our discussion of the five number summary and boxplots (section 3.3.0) we know already a better measure of spread, the interquartile range (kvartiilivälin pituus) defined as the difference between the upper and the lower quartile of the distribution, that is,

$$IQR = Q_3 - Q_1,$$

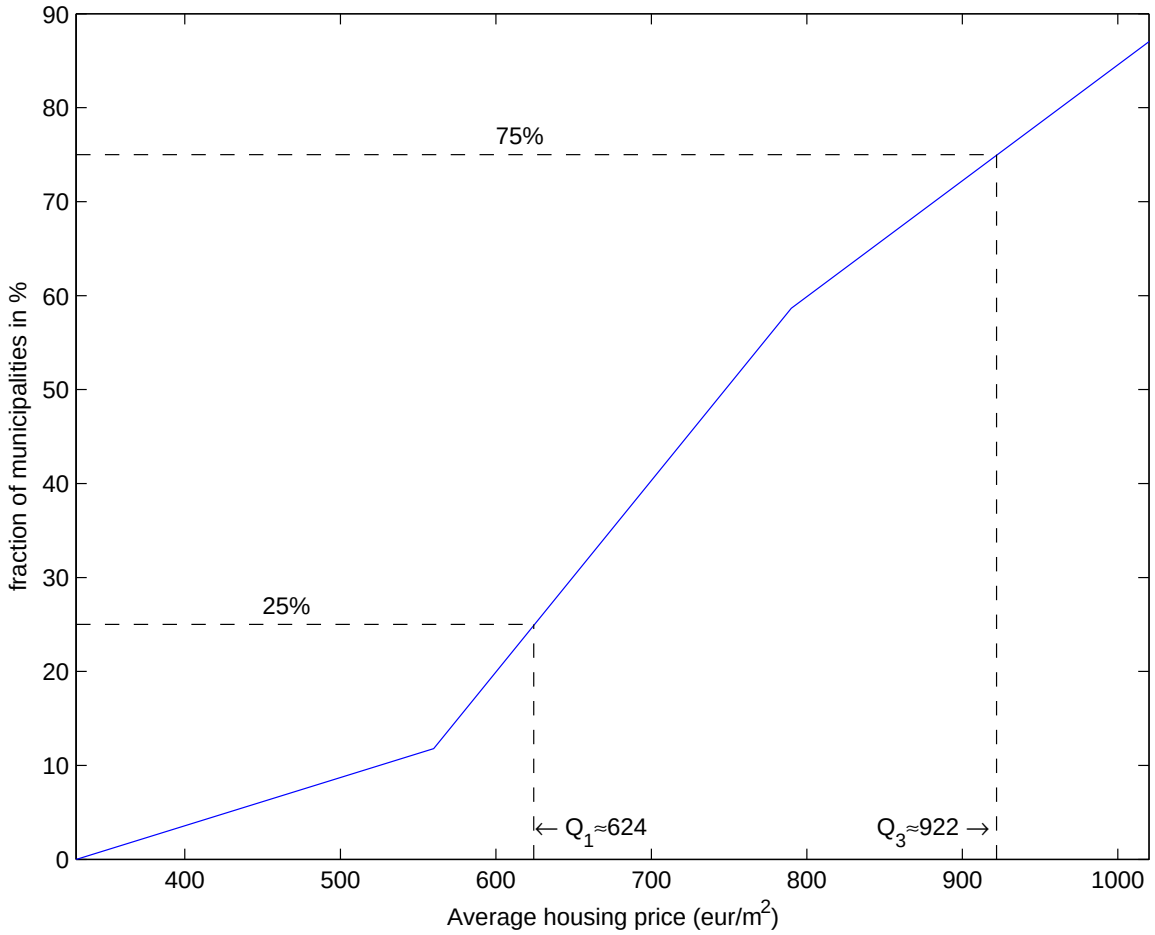
containing the 50% most central observations around the median.

The quartile deviation (kvartiilipoikkeama) Q is defined as half the interquartile range, that is,

$$Q = \frac{1}{2}(Q_3 - Q_1).$$

Note that while the quartiles are defined for variables measured on all scales except nominal scale, the interquartile range and the quartile deviation are defined only for variables measured on interval and ratio scale.

Example. Average housing prices continued.



The interquartile range and quartile deviation may be determined from the ogive of average selling prices as:

$$IQR \approx 922€/\mathrm{m}^2 - 624€/\mathrm{m}^2 = 298€/\mathrm{m}^2,$$
$$Q \approx (298€/\mathrm{m}^2)/2 = 149€/\mathrm{m}^2.$$

Variance and Standard Deviation

Suppose you want a measure of spread around the mean, which like the mean itself takes all observations into account. A natural candidate for this is the sum of all distances from the mean $\sum(x_i - \bar{x})$. However, we just defined the arithmetic mean $\bar{x}$ as the number from which all distances add up to zero. That is $\sum(x_i - \bar{x}) = 0$, no matter how large or small the spread of the distribution is, so it is not a good measure of spread. Instead one defines the sum of squared deviations (poikkeamien neliöiden summa) SSD from the mean as

$$SSD := \sum_{i=1}^{n} (x_i - \bar{x})^2 = (x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2,$$

where n denotes the number of observations. Squaring all observations from the mean has the advantage that differences from the left and to the right of the mean do no longer cancel out and one obtains a non-negative measure of spread, which becomes zero only if all observations equal the arithmetic mean $\bar{x}$.

Dividing SSD by the number of observations yields the population variance (populaation varianssi), which is the average squared deviation (poikkeamien neliöiden keskiarvo) from the mean. The population variance is not much used as a measure of spread though, for the following reason:

Often we do not have time and energy enough to collect all observations from the population we are interested in. Then we just collect observations from a part of it, called a sample (otos). To fix notation, let's assume we have a population of N units, but measurements only on a sample of $n < N$ units. Denote the sample mean $\sum_{i=1}^{n} x_i/n$ with $\bar{x}$ and the population mean $\sum_{i=1}^{N} x_i/N$ with μ. In this notation, the population variance σ^2 (populaatiovarianssi) becomes

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}.$$

Now suppose we want to estimate the variance of the full population σ^2 with the sum of squared deviations SSD_n we have from our sample with $n < N$. A natural strategy would be to calculate SSD_n/n, assuming that a typical deviation $(x_i - \bar{x})$ in our sample is representative for a typical deviation $(x_i - \mu)$ in the full population. This is however not the case, unless our sample mean $\bar{x}$ happens to coincide exactly with the population mean μ. In general $(x_i - \bar{x})$ will be smaller than $(x_i - \mu)$, because we constructed the arithmetic mean such that $\bar{x}$ it is the center of all observations in our sample, and not μ, which from the viewpoint of our sample is just some arbitrary number. So SSD_n/n will underestimate the population variance!

As it turns out, dividing SSD_n by $n-1$ instead of n will correct for this estimation error, and this is exactly how the sample variance (otosvarianssi) is defined. That is, the sample variance is our best estimate of the population variance if we have only $n < N$ observations.

The sample variance (otosvarianssi) s^2 is given by

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2/n}{n-1},$$

where n denotes the number of sample observations. The left hand side is equal to the right hand side because, recalling that $\bar{x} = (\sum_{i=1}^n x_i)/n$:

$$\begin{aligned}
\sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \sum_{i=1}^n x_i^2 - 2\bar{x}\sum_{i=1}^n x_i + n\bar{x}^2 \\
&= \sum_{i=1}^n x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2 \\
&= \sum_{i=1}^n x_i^2 - n\bar{x}^2 \\
&= \sum_{i=1}^n x_i^2 - n\frac{\left(\sum_{i=1}^n x_i\right)^2}{n^2}.
\end{aligned}$$

Note that the right hand side, despite its more complicated look, is faster to evaluate than the left hand side and is not affected by possible rounding errors in $\bar{x}$.

The sample standard deviation (otoskeskihajonta) s is defined as

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

Similarly, one defines the population standard deviation (populaatiokeskihajonta) σ as

$$\sigma := \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}.$$

The advantage of the standard deviation as compared to the variance is, that *the standard deviation has the same unit as the statistical variable under investigation*, whereas the variance is measured in units squared, which is often counterintuitive.

In the future, when we speak of variance (varianssi) and standard deviation (keskihajonta), we mean the sample variance and standard deviation, and not their population counterparts (unless stated otherwise).

Note: The variance and standard deviation are only defined for variables measured on interval and ratio scale.

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

We calculate the (sample) variance according to

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \underbrace{\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}_{n\bar{x}^2}\right).$$

Now:

$$\sum_{i=1}^{n} x_i = 19 + 4\cdot 20 + 5\cdot 21 + \cdots + 46 = 462,$$

$$\sum_{i=1}^{n} x_i^2 = 19^2 + 4\cdot 20^2 + \cdots + 46^2 = 12214.$$

$$\Rightarrow \quad s^2 = \frac{1}{19-1}\left(12214 - \frac{462^2}{19}\right) \approx 54.45\text{years}^2$$

$$\Rightarrow \quad s = \sqrt{s^2} \approx 7.4\text{years}.$$

Note that by summarizing students of the same age in the example above we actually calculated the standard deviation as

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k} f_i x_i^2 - \frac{\left(\sum_{i=1}^{k} f_i x_i\right)^2}{n}\right)$$

rather than

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right),$$

where k denotes the number of ages, and f_i was each age classes count.

Generally, for variables classified into k categories with frequencies f_i $(i = 1, \ldots, k)$, we may calculate the variance according to the same scheme with the individual observations x_i replaced by their classmarks m_i:

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k} f_i m_i^2 - \frac{\left(\sum_{i=1}^{k} f_i m_i\right)^2}{n}\right)$$

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
Sum:	416	

We calculate the variance according to

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k} f_i m_i^2 - \frac{\left(\sum_{i=1}^{k} f_i m_i\right)^2}{n}\right):$$

$$s^2 = \frac{1}{416-1}\left[\left(49\cdot 444.5^2 + \cdots + 3\cdot 2054.5^2\right) \underbrace{-\frac{\overbrace{327742^2}^{(n\bar{x})^2}}{416}}_{-n\bar{x}^2}\right]$$

$$= \frac{1}{415}\left(282858954 - \frac{327742^2}{416}\right) \approx 59398.2\,€^2/\mathrm{m}^4.$$

$$\Rightarrow \quad s \approx \sqrt{59398.2\,€^2/\mathrm{m}^4} \approx 244€/\mathrm{m}^2.$$

Dependence on Unit of Measurement

Both the variance and the standard deviation, like the mean, depend upon the unit in which the statistical variable is measured. Changing the scale of the variable x to a new scale according to the linear transformation $y = a + bx$ will change the variance s_x^2 and standard deviation s_x of the variable x into

$$s_y^2 = b^2 s_x^2 \quad \text{and} \quad s_y = |b| s_x$$

for the transformed variable y.

Example. Suppose that the variance of some length x measured in inches is $s_x^2 = 5(\text{inches})^2$. Now, one inch is 2.54cm and therefore the variance s_y^2 measured in cm^2 is

$$s_y^2 = 2.54^2 \cdot 5 = 32.258\text{cm}^2.$$

When comparing different statistical variables with each other, it is often desirable to express their values in a way that does not depend upon their particular units of measurement. This may be done by calculating standardized observations (standardoitu havainto) or z-scores z_i according to

$$z_i = \frac{x_i - \bar{x}}{s},$$

where x_i denotes observation i of the statistical variable x, $\bar{x}$ its arithmetic mean, and s its standard deviation.

A z-score measures the number of standard deviations s that a data value x_i is from the mean $\bar{x}$. Standardized observations have the following important properties:

1. Zero Mean: $\bar{z} = 0$,
2. Unit Variance: $s_z = s_z^2 = 1$,
3. z is a pure number (no units involved).

Example. A student attends a statistics exam and gets 32 points. She also attends an exam in business mathematics and obtains 30 points. The average number of points attained in the statistics exam by all students was 29 with a standard deviation of 6, while in the business mathematics exam the corresponding numbers were 22 and 8. In which exam did the student do better relative to the other students?

The standardized exam results are:

$$z_{\text{Stat}} = \frac{32 - 29}{6} = 0.5, \quad z_{\text{BMat}} = \frac{30 - 22}{8} = 1.$$

Therefore, the student did relatively better in the Business Mathematics than in the Statistics exam, even though she obtained more points in Statistics than in Business Mathematics.

If the statistical variable x is measured on ratio scale, we may define a dimensionless measure of spread as the so called coefficient of variance (variaatiokerroin) V:

$$V = \frac{s}{\bar{x}}.$$

The coefficient of variance of x tells how large his spread is relative to its center and it is usually reported as a percentage, that is $100V\%$. Again, the fact that it is dimensionless implies its usefulness in comparing spreads from different statistical variables.

Example.

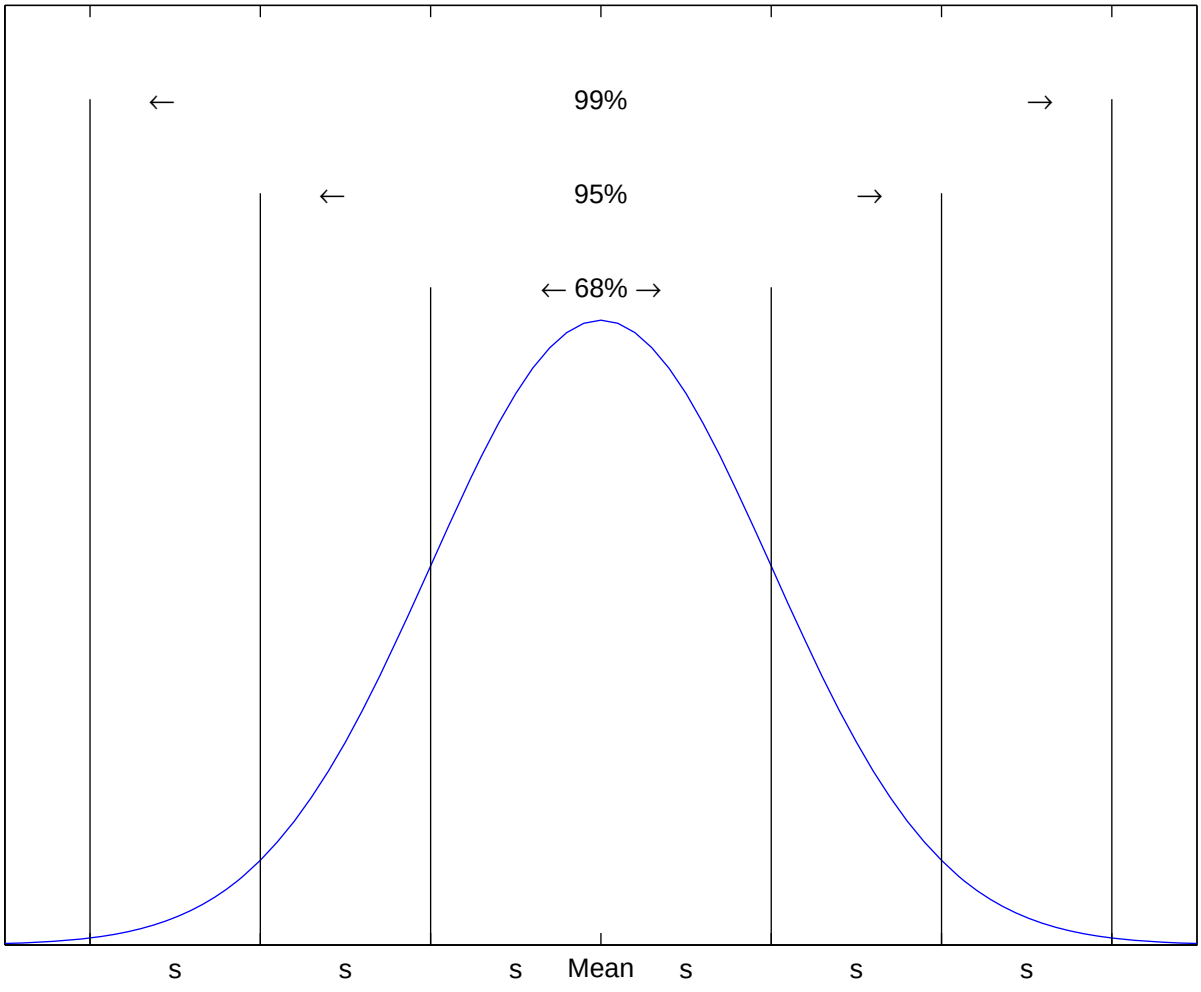
In the preceeding example the coefficients of variance were

$$V_{\text{Stat}} = \frac{6}{29} \approx 20.7\%, \quad V_{\text{BMat}} = \frac{8}{22} \approx 36.4\%,$$

meaning that the spread of the statistics exam was about 20%, and the spread of the business mathematics exam about 36% of their respective means.

The Empirical Rule

For *reasonably symmetric unimodal distributions* (and only for those) there are about 68% of the observations within one standard deviation around the mean, 95% within two standard deviations around the mean, and 99% within three standard deviations around the mean:



Chebyshev's Theorem

Chebyshev's Theorem asserts that regardless of the distributions shape:

1. At least three-quarters of all observations will lie within 2 standard deviations of the arithmetic mean.
2. At least eight-ninths of all observations will lie within 3 standard deviations of the arithmetic mean.

In general, the theorem states that at least $1 - 1/k^2$ of all observations will lie within k standard deviations of the arithmetic mean (where k does not necessarily have to be an integer number).

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

According to Chebychev, no more than one quarter, that is 25% of all students should be older than $\bar{x} + 2s \approx 24.3 + 2 \cdot 7.4 \approx 39$. Indeed there are two students older than that, that is $2/19 \approx 10,5\% < 25\%$ (but more than 2.5%, which the empirical rule would suggest).

3.3.3 Measures of Shape

Skewness (Vinous)

Recall that non-symmetric unimodal distribution are skewed to the right (oikealle viinoutta) if the observations concentrate upon the lower values or classes, such that it has a long tail to the right, and skewed to the left (vasemmalle viinoutta), if the observations concentrate upon the higher values or classes, such that the distribution has a long tail to the left. The (coefficient of) skewness (vinous eli vinouskerroin)

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

is supposed to measure this asymmetry. Usually a distribution is skewed to the left (right) if g_1 is smaller (larger) than zero. Unimodal distributions with $g_1 \approx 0$ ($-0.5 < g_1 < 0.5$) are regarded as fairly symmetric.

Kurtosis (Huipukkuus)

The (coefficient of) Kurtosis (huipukkuus eli huipukkuuskerroin), defined as

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - 3,$$

as the finnish name says, is a measure of peakedness (at least for unimodal distributions). That is, unimodal distributions with low kurtosis ($g_2 < 0$), called platykurtic (platykurtinen), are rather evenly spread across all possible values or classes, and unimodal distributions with high kurtosis ($g_2 > 0$), called leptokurtic (leptokurtinen), have a sharp peak at their mode. Distributions with $g_2 \approx 0$ are called mesokurtic (mesokurtinen).

The kurtosis of the well known normal distribution (to be discussed later) is exactly zero. Therefore, the sign of g_2 tells for unimodal distributions whether they are more ($g_2 > 0$) or less ($g_2 < 0$) sharp peaked than the normal distribution.