Numerical Descriptive Measures

Jiří Neubauer

Department of Econometrics FVL UO Brno
office 69a, tel. 973 442029
email:Jiri.Neubauer@unob.cz

Numerical Descriptive Measures

- measures of location (center)
- measures of dispersion (variation)
- measures of concentration

Arithmetic mean

The most important aspect of studying the distribution of a sample of measurements is locating the position of a central value about which the measurements are distributed.

Definition

The **arithmetic mean (average)** of a set of n measurements $x_1, x_2 \ldots, x_n$ is given by the formula

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Arithmetic mean

If the data are organized in the frequency distribution table then we can calculate the mean by the formula

$$\overline{x} = \frac{1}{n} \sum_{j=1}^{k} n_j \cdot x_j,$$

where $n_1, n_2, \ldots, n_k$ are frequencies of variable varieties $x_1, x_2 \ldots, x_k$.

Arithmetic mean

Elementary properties of the arithmetic mean:

- the sum of deviations between the values and the mean is equal to zero

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0,$$

Arithmetic mean

Elementary properties of the arithmetic mean:

- the sum of deviations between the values and the mean is equal to zero

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0,$$

- if the variable is constant then the mean is equal to this constant

$$\frac{1}{n}\sum_{i=1}^{n} c = c,$$

Arithmetic mean

Elementary properties of the arithmetic mean:

- if we add a constant to the values of the variable, then

$$\frac{1}{n}\sum_{i=1}^{n}(x_i + c) = c + \overline{x},$$

Arithmetic mean

Elementary properties of the arithmetic mean:

- if we add a constant to the values of the variable, then

$$\frac{1}{n}\sum_{i=1}^{n}(x_i + c) = c + \overline{x},$$

- if we multiply the values of the variable by a constant c, then

$$\frac{1}{n}\sum_{i=1}^{n} c \cdot x_i = c \cdot \overline{x}.$$

Harmonic mean

Definition

The **harmonic mean** of a set of n measurements $x_1, x_2 \ldots, x_n$ is given by the formula

$$\overline{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

In certain situations, especially many situations involving rates and ratios, the harmonic mean provides the truest average.

Geometric mean

Definition

The **geometric mean** of a set of n measurements $x_1, x_2 \ldots, x_n$ is given by the formula

$$\overline{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}.$$

The geometric mean may be more appropriate than the arithmetic mean for describing percentage growth.
Suppose an orange tree yields 100 oranges one year, then 180, 210 and 300 the following years, so the growth is 80%, 16.7% and 42.9% for each of the years. Using the arithmetic mean, we can calculate an average growth as 46.5% (80% + 16.7% + 42.9 % divided by 3). However, if we start with 100 oranges and let it grow with 46.5% for three years, the result is 314 oranges, not 300.

Example

Calculate the arithmetic, harmonic and geometric mean of 1, 2, 5, 6, 7, 8, 8, 9.

- The arithmetic mean is

$$\overline{x} = \frac{1+2+5+6+7+8+8+9}{8} = 5.75.$$

- The harmonic mean is

$$\overline{x}_H = \frac{8}{\frac{1}{1}+\frac{1}{2}+\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}+\frac{1}{8}+\frac{1}{9}} \doteq 3.375.$$

- The geometric mean is

$$\overline{x}_G = \sqrt[8]{1 \cdot 2 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 8 \cdot 9} \doteq 4.709.$$

Notice that $\overline{x}_H \leq \overline{x}_G \leq \overline{x}$.

Quantile

Definition

The **quantile** x_p is the value of the variable which fulfils that $100p\,\%$ of values of ordered sample (or population) are smaller or equal to x_p and $100(1-p)\,\%$ of values of ordered sample (or population) are larger or equal to x_p.

The quantile is not uniquely defined.

Quantile

Let us have the data set 2 5 7 10 12 13 18 21.
Possible methods of calculation

- Sort the data in ascending order. Find the sequential index i_p of the quantile x_p, which fulfils inequation

$$np < i_p < np + 1.$$

The quantile x_p is then equal to the value of variable with the sequential index i_p – $x_p = x_{(i_p)}$. If $np, np+1$ are integer, we calculate the quantile as an aritmetic mean of $x_{(np)}$ a $x_{(np+1)}$, $x_p = \frac{x_{(np)}+x_{(np+1)}}{2}$. Statistical software STATISTICA uses this method.

Quantile

- According to MATLAB
  We calculate
  $$\bar{i}_p = \frac{np + np + 1}{2} = \frac{2np + 1}{2}$$
  determining the location of the quantile. Using linear interpolation we get
  $$x_p = x_{([\bar{i}_p])} + (x_{([\bar{i}_p]+1)} - x_{([\bar{i}_p])})(\bar{i}_p - [\bar{i}_p]),$$
  where $[\cdot]$ denotes the integer part of the number. If $\bar{i}_p < 1$ then $x_p = x_{(1)}$, if $\bar{i}_p > n$ then $x_p = x_{(n)}$.

Quantile

- According to EXCEL
  We assign values $0, \frac{1}{n-1}, \frac{2}{n-1}, \ldots, \frac{n-2}{n-1}, 1$ to the data sorted in ascending order. If P is equal to the multiple of $\frac{1}{n-1}$, the quantile x_p is equal to the value corresponding to the given multiple. If P is not the multiple $\frac{1}{n-1}$, we use linear interpolation.

Quantile

x_p	0.10	0.25	0.50	0.75	0.90
STATISTICA	2	6	11	15.5	21
MATLAB	2.9	6	11	15.5	20.1
EXCEL	4.1	6.5	11	14.25	18.9

Example

Calculate the median, lower and upper quartile and lower and upper decile of 1, 2, 5, 6, 7, 8, 8, 9.

The range of the data set is $n = 8$. The median is the middle value of the data sorted in ascending order. There is not one middle value, but two (6 and 7). We calculate the median as

$$\tilde{x} = x_{0.50} = \frac{6+7}{2} = 6.5.$$

Interpretation: 50 % of ordered values are smaller or equal to 6.5, do not exceed value 6.5.

Example

Lower quartile $x_{0.25}$. Usign the formula

$$np < i_p < np + 1$$

we get $8 \cdot 0.25 < i_p < 8 \cdot 0.25 + 1 \Leftrightarrow 2 < i_p < 3$.

$$x_{0.25} = \frac{x_{(2)} + x_{(3)}}{2} = \frac{2+5}{2} = 3.5.$$

Analogously for upper decile: $x_{0.90}$,
$8 \cdot 0.90 < i_p < 8 \cdot 0.90 + 1 \Leftrightarrow 7.2 < i_p < 8.2$, we get $i_p = 8$ and

$$x_{0.90} = x_{(8)} = 9.$$

We say that 25 % of ordered values are smaller or equal to 3.5.
Analogously 90 % of values do not exceed 9.

Mode

Definition

The **mode** $\hat{x}$ is the value of variable with the highest frequency.

In the case of continuous variable (data) the mode is the value where the histogram reaches its peak.

Mode

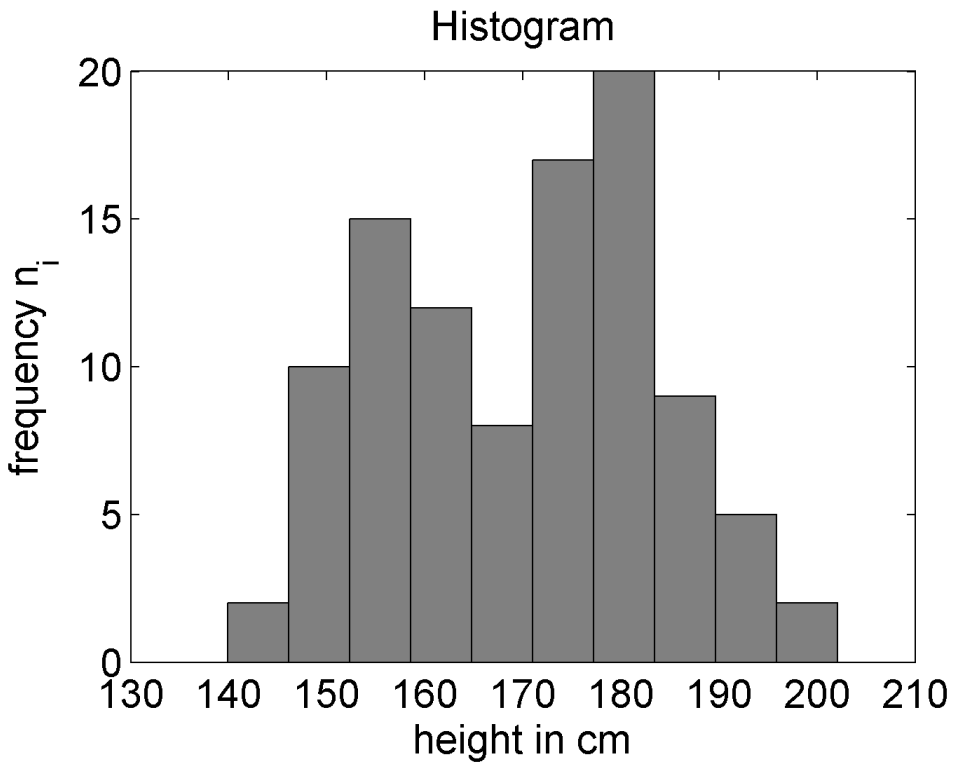

Figure: Non-homogeneous sample

Measures of Dispersion

Means, quantiles and a mode – measures of location – describe one property of frequency distribution – location.
Another important property is dispersion (variation) which we describe by several measures of variation

Measures of Dispersion

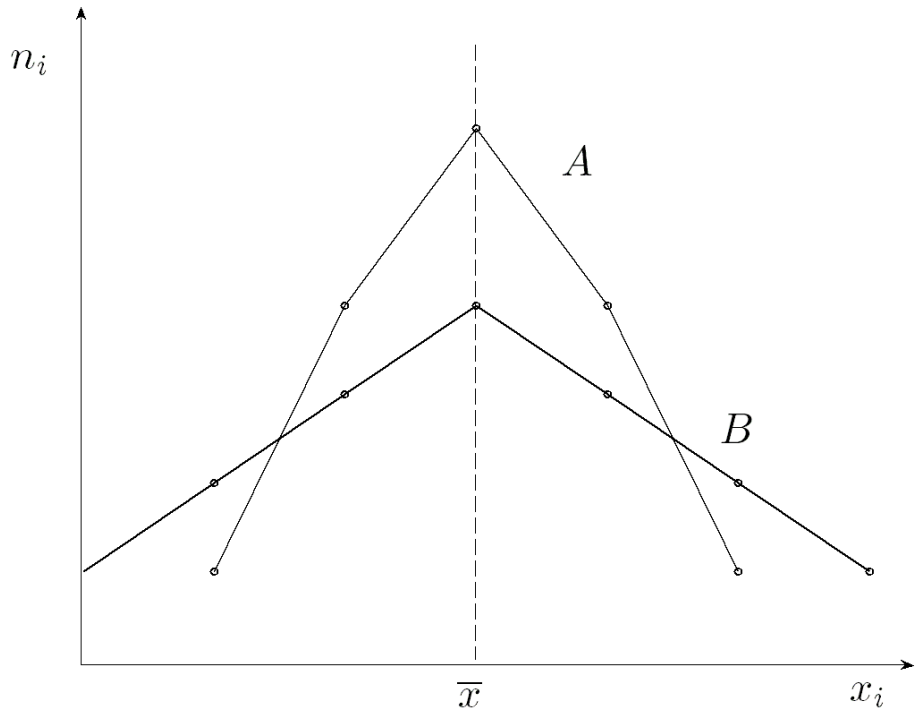

Figure: Two samples with different variation

Range of Variation

Definition

The **range of variation** R is defined as difference between the largest and the smallest value of the variable

$$R = x_{\max} - x_{\min}.$$

It is the simplest but the rawest measure of variation. It indicates the width of the interval where all values are included.

Interquantile Ranges

Definition

- the **interquartile range**

$$R_Q = x_{0.75} - x_{0.25}$$

Interquantile Ranges

Definition

- the **interquartile range**

$$R_Q = x_{0.75} - x_{0.25}$$

- the **interdecile range**

$$R_D = x_{0.90} - x_{0.10}$$

Interquantile Ranges

Definition

- the **interquartile range**

$$R_Q = x_{0.75} - x_{0.25}$$

- the **interdecile range**

$$R_D = x_{0.90} - x_{0.10}$$

- the **interpercentile range**

$$R_C = x_{0.99} - x_{0.01}$$

Interquantile Ranges

The interquartile range indicates the width of the interval which includes 50 % of middle values of ordered sample. By analogy the interdecile or the interpercentile range indicatethe width of the interval which includes 80 % or 98 % of middle values of ordered sample.

Example

We have calculated quantiles of the data 2, 5, 7, 10, 12, 13, 18 and 21. According to STATISTICA: $x_{0.10} = 2$, $x_{0.25} = 6$, $x_{0.50} = 11$, $x_{0.75} = 15.5$, $x_{0.90} = 21$.

The range of variation is $R = x_{\max} - x_{\min} = 21 - 2 = 19$.
The interquartile range is $R_Q = x_{0.75} - x_{0.25} = 15.5 - 6 = 9.5$.
The interdecile range is $R_D = x_{0.90} - x_{0.10} = 21 - 2 = 19$.

Quantile Deviations

Definition

- the **quartile deviation**

$$Q = R_Q/2$$

Quantile Deviations

Definition

- the **quartile deviation**

$$Q = R_Q/2$$

- the **decile deviation**

$$D = R_D/8$$

Quantile Deviations

Definition

- the **quartile deviation**

$$Q = R_Q/2$$

- the **decile deviation**

$$D = R_D/8$$

- the **percentile deviation**

$$C = R_C/98$$

Example

Calculate the quartile and the decile deviation of 2, 5, 7, 10, 12, 13, 18 and 21.

The quartile deviation is $Q = R_Q/2 = 9{,}5/2 = 4{,}75$.
The decile deviation is $D = R_D/8 = 19/8 = 2{,}375$.
It means that the average width of two (eight) middle quartile (decile) intervals is 4.75 (2.375).

Average Deviation

Definition

The **average deviation** is defined as the arithmetic mean of the absolute deviations

$$d_{\overline{x}} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \overline{x}|.$$

Example

Find the average deviation of a data set 1, 2, 5, 6, 7, 8, 8 and 9.

The arithmetic mean is $\overline{x} = 5.75$. We obtain

$$\overline{d}_{\overline{x}} = \frac{|1 - 5.75| + |2 - 5.75| + |5 - 5.75| + |6 - 5.75|}{8} +$$
$$+ \frac{|7 - 5.75| + |8 - 5.75| + |8 - 5.75| + |9 - 5.75|}{8} = 2.3125.$$

Variance

Definition

The **variance** s_n^2 is defined as the arithmetic mean of squares of deviations

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Variance

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\overline{x}^2\right)$$
$$= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - 2n\overline{x}^2 - n\overline{x}^2\right) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 = \overline{x^2} - \overline{x}^2.$$

Variance

Elementary properties of the variance:

- if the variable is constant and is equal to c, then the variance is zero

$$\frac{1}{n}\sum_{i=1}^{n}(c-c)^2 = 0,$$

Variance

Elementary properties of the variance:

- if the variable is constant and is equal to c, then the variance is zero

$$\frac{1}{n}\sum_{i=1}^{n}(c-c)^2 = 0,$$

- if we add a constant to the values of the variable, then

$$\frac{1}{n}\sum_{i=1}^{n}[(x_i+c)-(\overline{x}+c)]^2 = s_n^2,$$

Variance

Elementary properties of the variance:

- if the variable is constant and is equal to c, then the variance is zero

$$\frac{1}{n}\sum_{i=1}^{n}(c-c)^2 = 0,$$

- if we add a constant to the values of the variable, then

$$\frac{1}{n}\sum_{i=1}^{n}[(x_i+c)-(\overline{x}+c)]^2 = s_n^2,$$

- if we multiply the values of the variable by a constant c, then

$$\frac{1}{n}\sum_{i=1}^{n}(c\cdot x_i - c\cdot\overline{x})^2 = c^2\cdot s_n^2.$$

Standard Deviation

Definition

The square root of the variance is called the **standard deviation**

$$s_n = \sqrt{s_n^2}$$

Sample Variance and Standard Deviation

Definition

The **sample variance** s^2 if defined by the formula

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

the square root of the sample variance is called the **sample standard deviation**

$$s = \sqrt{s^2}.$$

It is obvious that

$$s_n^2 = \frac{n-1}{n}s^2.$$

Example

Calculate the variance, the standard deviation, the sample variance and the sample standard deviation of the data set 1, 2, 5, 6, 7, 8, 8 and 9.

The arithmetic mean is $\overline{x} = 5.75$.

$$s_n^2 = \frac{(1-5.75)^2 + (2-5.75)^2 + (5-5.75)^2 + (6-5.75)^2}{8} +$$
$$+ \frac{(7-5.75)^2 + (8-5.75)^2 + (8-5.75)^2 + (9-5.75)^2}{8} = 7.4375.$$

Example

The variance can be also calculated by the formula $s_n^2 = \overline{x^2} - \overline{x}^2$.

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1^2 + 2^2 + 5^2 + 6^2 + 7^2 + 8^2 + 8^2 + 9^2}{8} = 40.5,$$

$$s_n^2 = \overline{x^2} - \overline{x}^2 = 40.5 - 5.75^2 = 7.4375.$$

The standard deviation is

$$s_n = \sqrt{s_n^2} = \sqrt{7.4375} \doteq 2.72718.$$

Example

To get the sample variation we apply the formula

$$s^2 = \frac{n}{n-1} s_n^2 = \frac{8}{7} \cdot 7.4375 = 8.5.$$

The sample standard deviation is

$$s = \sqrt{s^2} = \sqrt{8.5} \doteq 2.91548.$$

Moments

Definition

The r^{th} **moment** is defined by the formula

$$m'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r,$$

The r^{th} **central moment** is defined by the formula

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^r.$$

Sample Skewness

Definition

The **sample skewness** is defined by the formula

$$a_3 = \frac{m_3}{m_2^{3/2}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{ns_n^3} = \frac{m_3}{s_n^3}$$

Sample Skewness

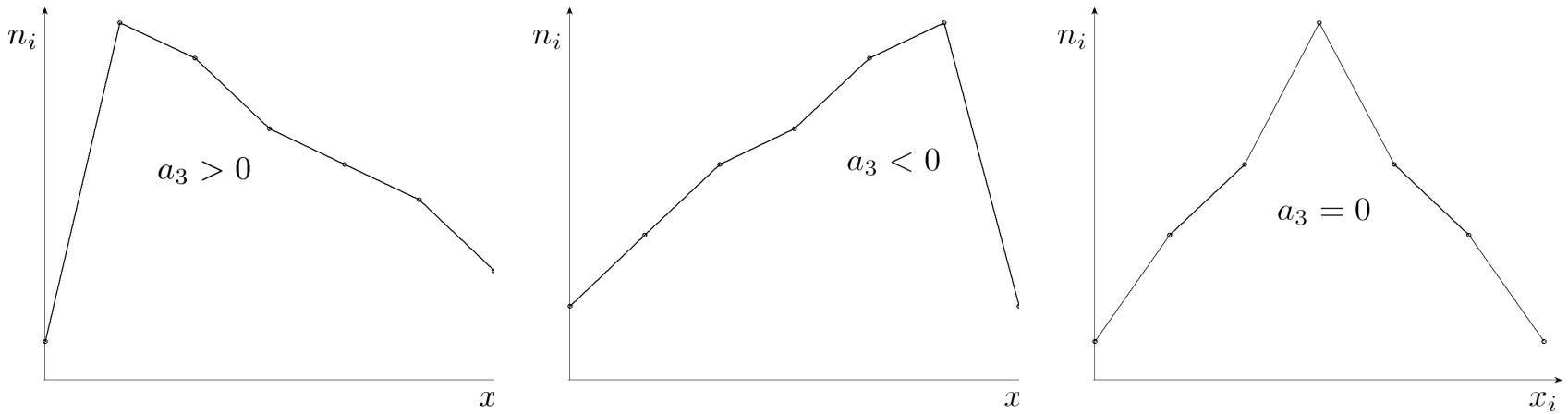

Figure: Frequency distribution with the different sample skewness

Sample Kurtosis

Definition

The **sample kurtosis** is defined by the formula

$$a_4 = \frac{m_4}{m_2^2} - 3 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{ns_n^4} - 3$$

Sample Kurtosis

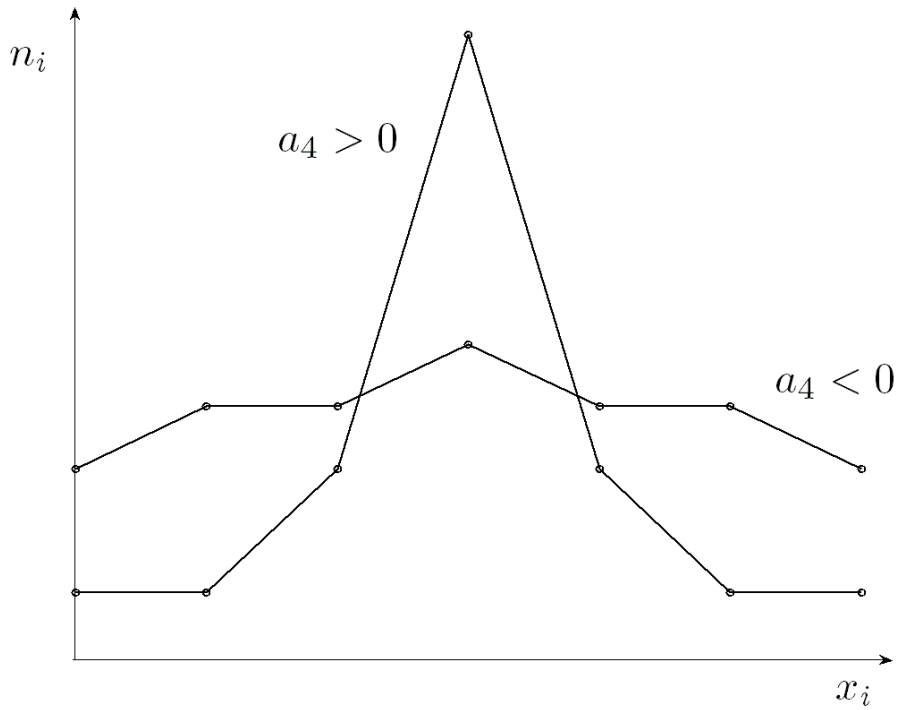

Figure: Frequency distribution with the different sample kurtosis

Note

Excel functions SKEW and KURT calculate skewness and kurtosis by formulas

$$a_3^* = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s} \right)^3,$$

$$a_4^* = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}.$$

It can be derived that

$$a_3 = \frac{n-2}{\sqrt{n(n-1)}} \cdot a_3^*,$$

$$a_4 = \frac{(n-2)(n-3)}{n^2-1} \cdot a_4^* - \frac{6}{n+1}.$$