Measurements of Disperion

Example [2]

There is the following data set: 22, 82, 27, 43, 19, 47, 41, 34, 34, 42, 35 (the data from the previous example).

Determine:

a) All quartiles

b) Inter-Quartile Range

c) MAD

d) Draw the Empirical Distribution Function

Solution:

a) You need to determine Lower Quartile $x_{0,25}$; Median $x_{0,5}$ and Upper Quartile $x_{0,75}$.

First, you order the data by size and assign a sequence number to each value.

Original data	Ordered data	Sequence
22	19	1
82	22	2
27	**27**	**3**
43	34	4
19	34	5
47	**35**	**6**
41	41	7
34	42	8
34	**43**	**9**
42	47	10
35	82	11

Now you can divide the data set into quartiles and mark their variable values accordingly:

Lower Quartile $x_{0,25}$: $p = 0.25; n = 11 \Rightarrow z_p = 11 \text{ x } 0.25 + 0.5 = 3.25 \cong 3 \Rightarrow x_{0.25} = 27$
i.e. 25% of musicians are under 27 (75% of them are 27 years old or older).

Median $x_{0,5}$: $p = 0.5; n = 11 \Rightarrow z_p = 11 \text{ x } 0.5 + 0.5 = 6 \Rightarrow x_{0.5} = 35$
i.e. a half of the musician are under 35 (50% of them are 35 years old or older).

Upper Quartile $x_{0,75}$: $p = 0.75; n = 11 \Rightarrow z_p = 11 \text{ x } 0.75 + 0.5 = 8.75 \cong 9 \Rightarrow x_{0.75} = 43$

i.e. 75% musicians are under 43 (25% of them are 43 years old or older).

b) Inter-Quartile Range IQR:

$$\textbf{IQR} = x_{0.75} - x_{0.25} = 43 - 27 = 16$$

c) MAD

If you want to determine this characteristic you must follow its definition (the median of absolute deviations from the median).

$$x_{0.5} = 35$$

Original data x_i	Ordered data y_i	Absolute values of deviations of the ordered data from their median $\lvert y_i - x_{0.5} \rvert$	Ordered absolute values M_i
22	19	$16 = \lvert 19 - 35 \rvert$	0
82	22	$13 = \lvert 22 - 35 \rvert$	1
27	27	$8 = \lvert 27 - 35 \rvert$	1
43	34	$1 = \lvert 34 - 35 \rvert$	6
19	34	$1 = \lvert 34 - 35 \rvert$	7
47	35	$0 = \lvert 35 - 35 \rvert$	**8**
41	41	$6 = \lvert 41 - 35 \rvert$	8
34	42	$7 = \lvert 42 - 35 \rvert$	12
34	43	$8 = \lvert 43 - 35 \rvert$	13
42	47	$12 = \lvert 47 - 35 \rvert$	16
35	82	$47 = \lvert 82 - 35 \rvert$	47

$$MAD = M_{0.5}$$
$$p = 0.5;\ n = 11 \Rightarrow z_p = 11 \text{ x } 0.5 + 0.5 = 6 \Rightarrow x_{0.5} = 8$$

(MAD is a median absolute deviation from the median i.e. 6th value of ordered absolute deviations from the median)

MAD = 8.

d) The last task was to draw the Empirical Distribution Function. Here is its definition:

$$F(x) = \begin{cases} 0 & for x \le x_1 \\ \sum_{i=1}^{j} p(x_i) & for x_j < x \le x_{j+1}, 1 \le j \le n-1 \\ 1 & for x_n < x \end{cases}$$

Arrange the variable values as well as their frequencies and relative frequencies in ascending order and write them down in the table. Then derive the empirical distribution function from them:

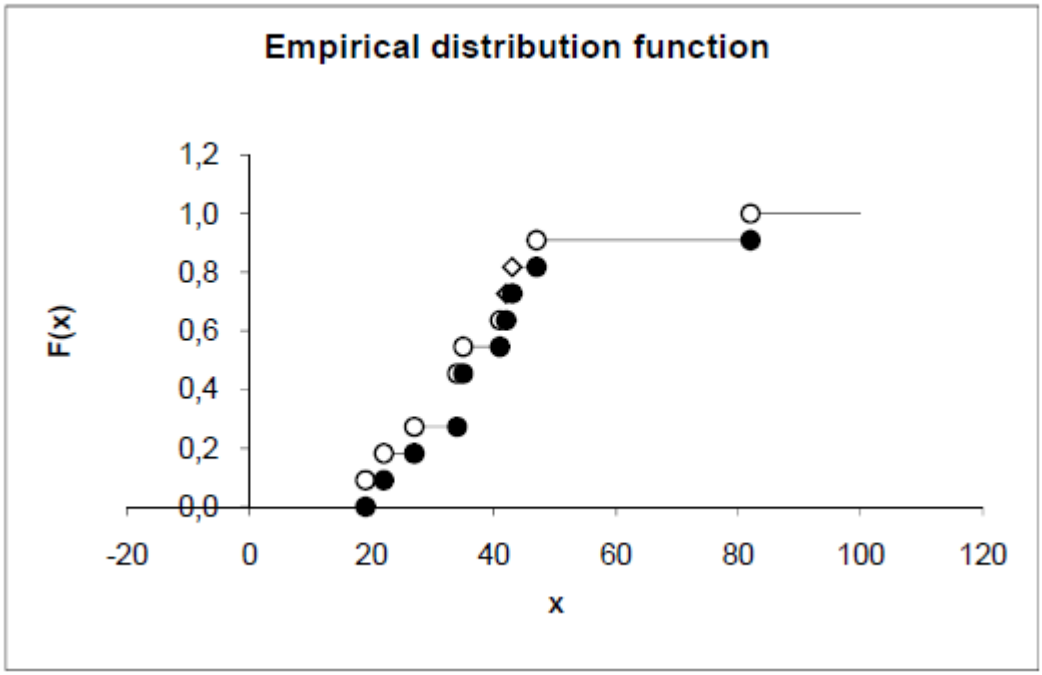
Original data x_i	Ordered data a_i	Absolute frequencies of the ordered values n_i	Relative frequencies of the ordered values p_i	Empirical distribution function $F(a_i)$
22	19	1	1/11	0
82	22	1	1/11	1/11
27	27	1	1/11	2/11
43	34	2	2/11	3/11
19	35	1	1/11	5/11
47	41	1	1/11	6/11
41	42	1	1/11	7/11
34	43	1	1/11	8/11
34	47	1	1/11	9/11
42	82	1	1/11	10/11
35				

As by its definition - the empirical distribution function F(x) - equals 0 for each x <19; F(x) equals 1/11 for all $22 \geq x > 19$; F(x) equals 1/11 + 1/11 for all $27 \geq x > 22$; and so it goes on.

X	$(-\infty;19\rangle$	$(19;22\rangle$	$(22;27\rangle$	$(27;34\rangle$	$(34;35\rangle$
F(x)	0	1/11	2/11	3/11	5/11

X	$(35;41\rangle$	$(41;42\rangle$	$(42;43\rangle$	$(43;47\rangle$	$(47;82\rangle$	$(82;\infty)$
F(x)	6/11	7/11	8/11	9/11	10/11	11/11

- **Sample Variance s^2**



Sample Variance is the most common measure of variability

The sample variance is given by:

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

Where

x_i is the variables

$\bar{x}$ is the variables mean

n number of variables.

- Sample Variance is the sum of all squared deviations from their mean divided by one less than the sample size

General properties of the sample variance are for example:

- The sample variance of a constant number is zero

In other words: if all variable values are the same, the sampling has zero diffuseness

- $$\forall a \in \Re : \left[\left(s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \right) \wedge (y_i = a + x_i) \right] \Rightarrow \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} = s^2$$

In other words: if you add the same constant number to all variable values, the sample variance doesn't change

- $$\forall b \in \Re : \left[\left(s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \right) \wedge (y_i = bx_i) \right] \Rightarrow \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} = b^2 s^2$$

In other words: if you multiply all variable values by an arbitrary constant number (b) the sample variance increases by square of this constant number (b^2).

Disadvantage of using the sample variance

Disadvantage of using the sample variance as a measure of variability is that it employs squared values of the variable. For example: if the variable represents cash denominated in EUR, then the sample variation of this variable will be in EUR^2. That is why we use another measure of variability called standard deviation.