



Data Analytics for Social Science

Distributions and descriptives

Johan A. Elkink

School of Politics & International Relations
University College Dublin

26 September 2017



Outline

Central
tendency

Variation

References

1 Central tendency

2 Variation



Measures of central tendency

Measures of central tendency provide information about the centre of a distribution, roughly put: “what is a typical value for this variable?”

Different measures are available for different levels of measurement:

	mode	median	mean
nominal	x		
ordinal	x	x	
scale	(x)	x	x



Mode

Central
tendency

Variation

References

The **mode** is the category with the highest frequency.

For a continuous variable, the highest point in an estimated density curve can be taken as the mode.



Median

Central
tendency

Variation

References

The **median** is the value where 50% of the cases has a lower value on this variable and 50% a higher value.

If N is uneven: the middle value after sorting in ascending order.

If N is even: the average of the two middle values after sorting in ascending order.

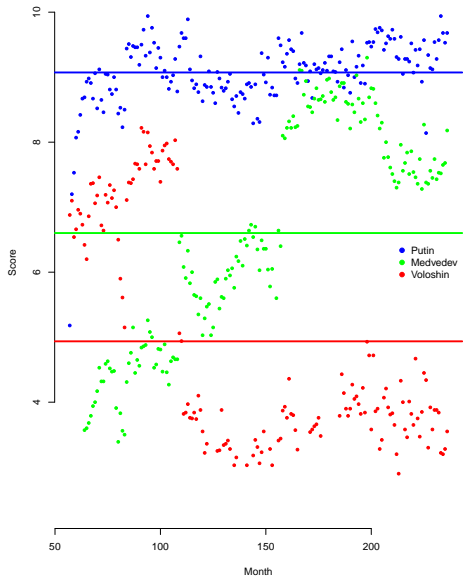


The **mean** is the sum of all values, divided by the number of values.

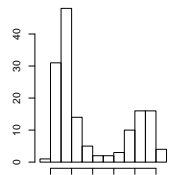
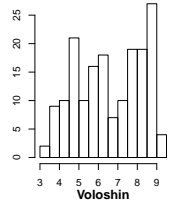
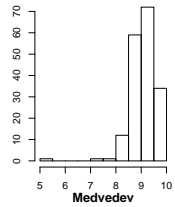
$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N}$$



Influence scores



Putin





Mean or median?

Outliers is a general term of values that are very far from the main values of the distribution.

- For a symmetric distribution, median and mean are the same.
- The more **skewed** the distribution, the more mean and median differ.
- Mean is sensitive to outliers, while the median is not.
- Mean has better understood mathematical properties.



Outline

Central
tendency

Variation

References

1 Central tendency

2 Variation



Measures of dispersion

Measures of dispersion provide an indication of the amount of variation or heterogeneity in a variable.

The **range** is the highest value minus the lowest value.

The **interquartile range** (IQR) is the range between the lowest 25% and the top 25%. A boxplot typically provides the median and the IQR, with some indication of outliers.



Variance

$$\text{Var}(x) = s_x^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}$$

Note that many software packages do not calculate this sample variance, but the unbiased estimator of the population variance:

$$\hat{\sigma}_x^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1}$$

Standard deviation:

$$s_x = \sqrt{\text{Var}(x)} = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}}$$



Baturo, Alexander and Johan A. Elkind. 2014. "Office or Officeholder? Regime Deinstitutionalisation and Sources of Individual Political Influence." *Journal of Politics* 76(3).

Baturo, Alexander and Johan A. Elkind. 2015. "Dynamics of Regime Personalization and Patron-Client Networks in Russia, 1999–2014." *Post-Soviet Affairs* 32(1):75–98.

Moore, David S. 2003. *The basic practice of statistics*. 3rd ed. New York: W.H. Freeman.