Univariate descriptives

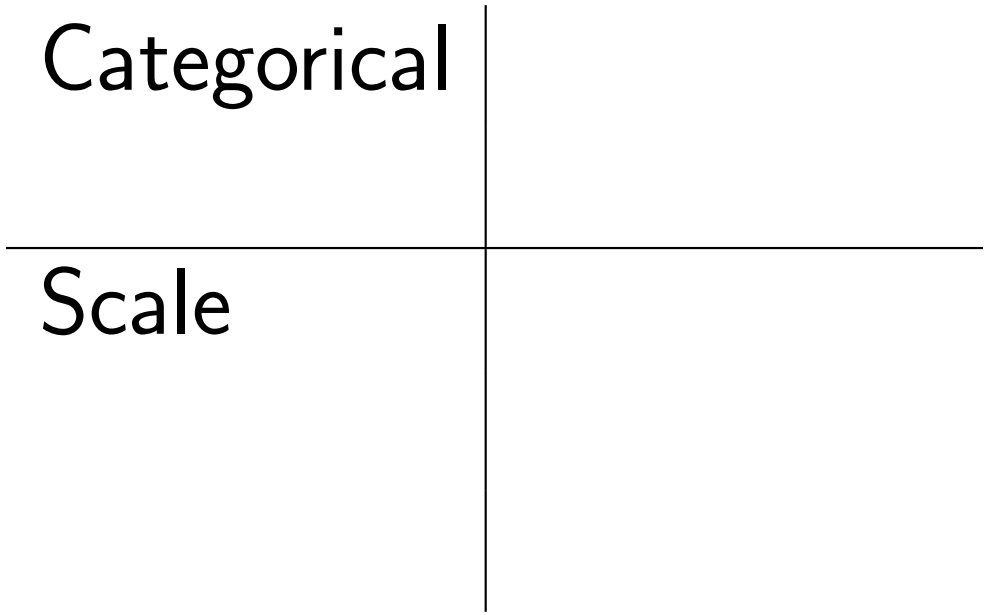
Johan A. Elkink

University College Dublin

16 September 2013

Introduction

Graphical presentations of single variables.



Introduction

Graphical presentations of single variables.

Categorical	pie-charts barplots
Scale	

Introduction

Graphical presentations of single variables.

Categorical	pie-charts barplots
Scale	histogram density plot boxplot

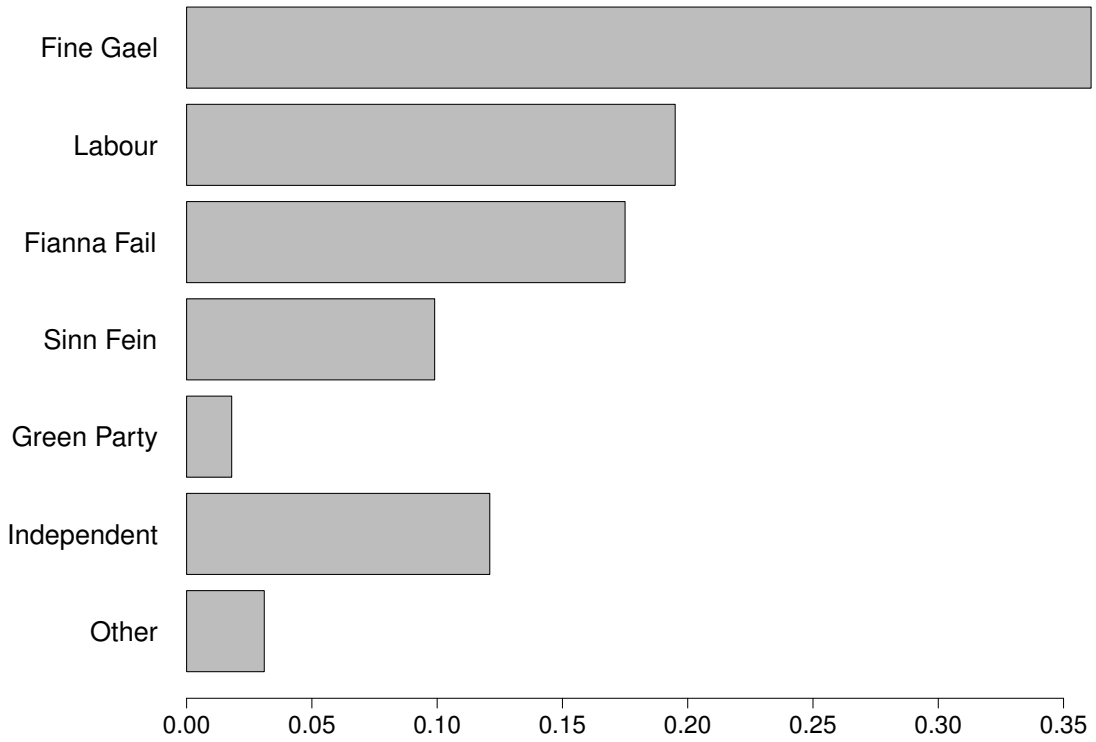
Outline

Categorical variables

For categorical variables, it is often useful to look at the number of cases or the proportion of cases in a particular category.

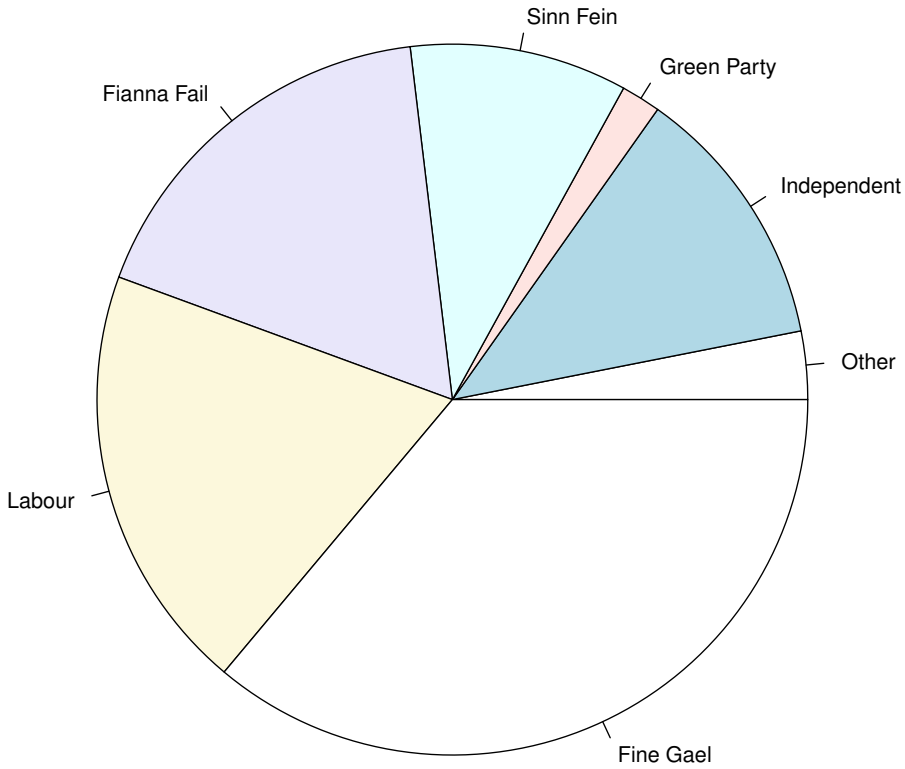
Barplots and pie charts are useful for this.

Barplot



(http://en.wikipedia.org/wiki/irish_general_election,_2011)

Pie chart


Sinn Fein
Green Party
Fianna Fail
Independent
Other
Labour
Fine Gael

Exercise

Using the data from the first lecture:

- Produce a pie chart for regime type
- Produce a bar chart for the cleavage variable

Outline

Histogram

For continuous (or scale) variables, we often want to get an idea of the distribution of values. How many low, medium, high values?

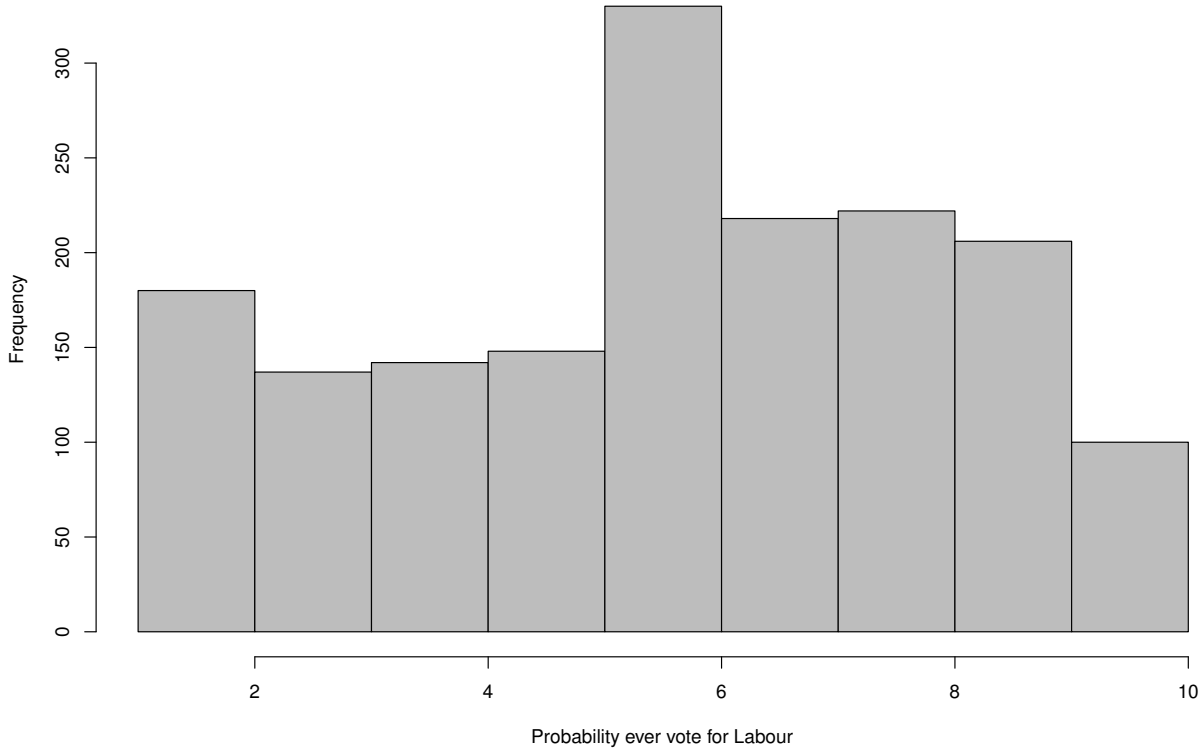
Histograms are useful to get an impression.

Histogram

Histogram:

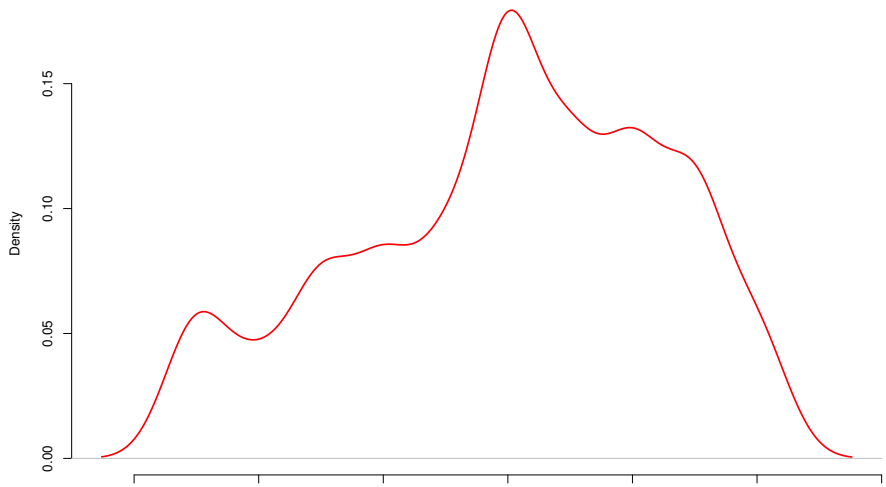
- bin the data using equal-distance cut-off points
- then produce a barplot of the number in each bin.

Histogram


Frequency
0
50
100
150
200
250
300
2
4
6
8
10
Probability ever vote for Labour

Density plot

A density plot is a smoothened version of a histogram, based on a non-parametric estimation of the shape of the distribution.



Distributions

For graphs of distributions (histogram, density plot, boxplot, etc.) you want to get an impression of:

- the shape of the distribution;
- the center and spread of the distribution;
- the presence of outliers.

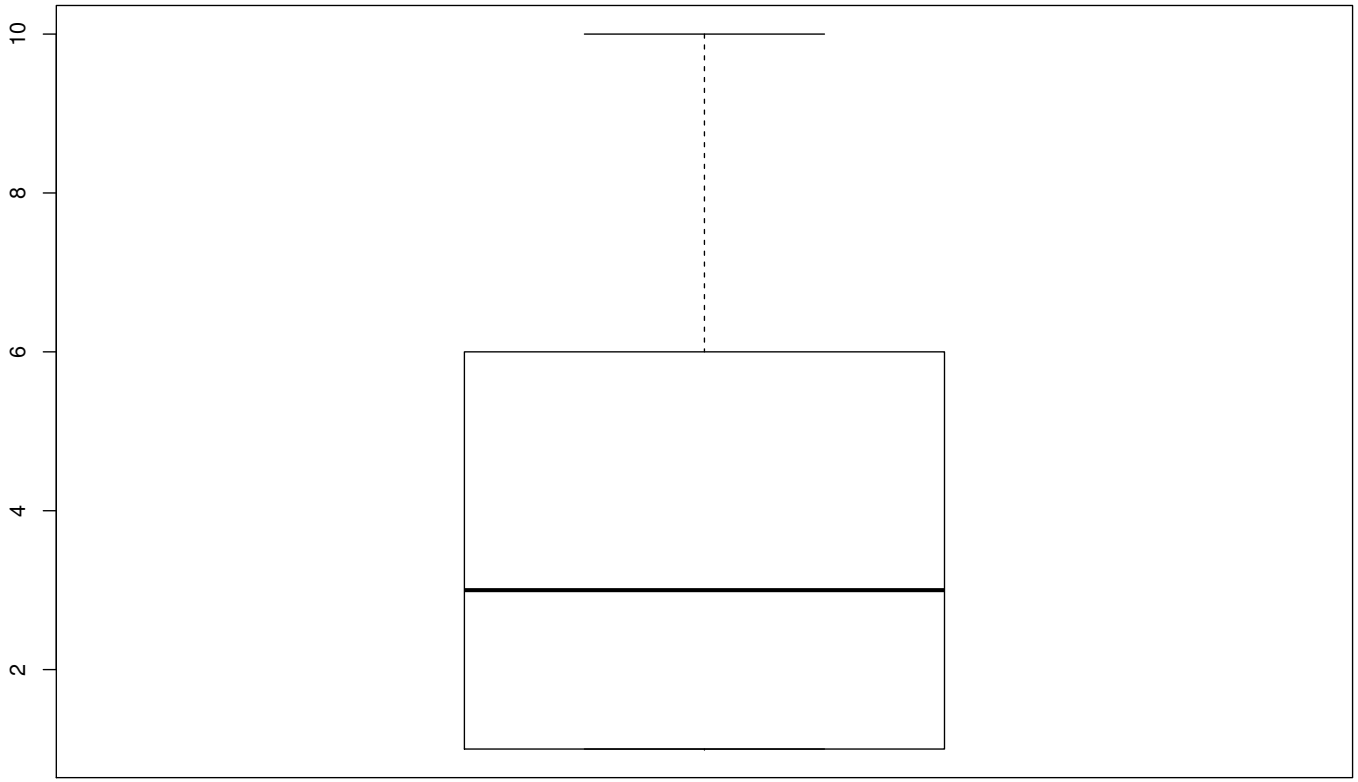
(Moore 2003: 12)

Boxplot

Another way of looking at the distribution of a continuous variable is to find out where the lowest 25% are located, where the lowest 50% are located, and where the top 25% are located.

A plot that shows this is the boxplot.

Boxplot



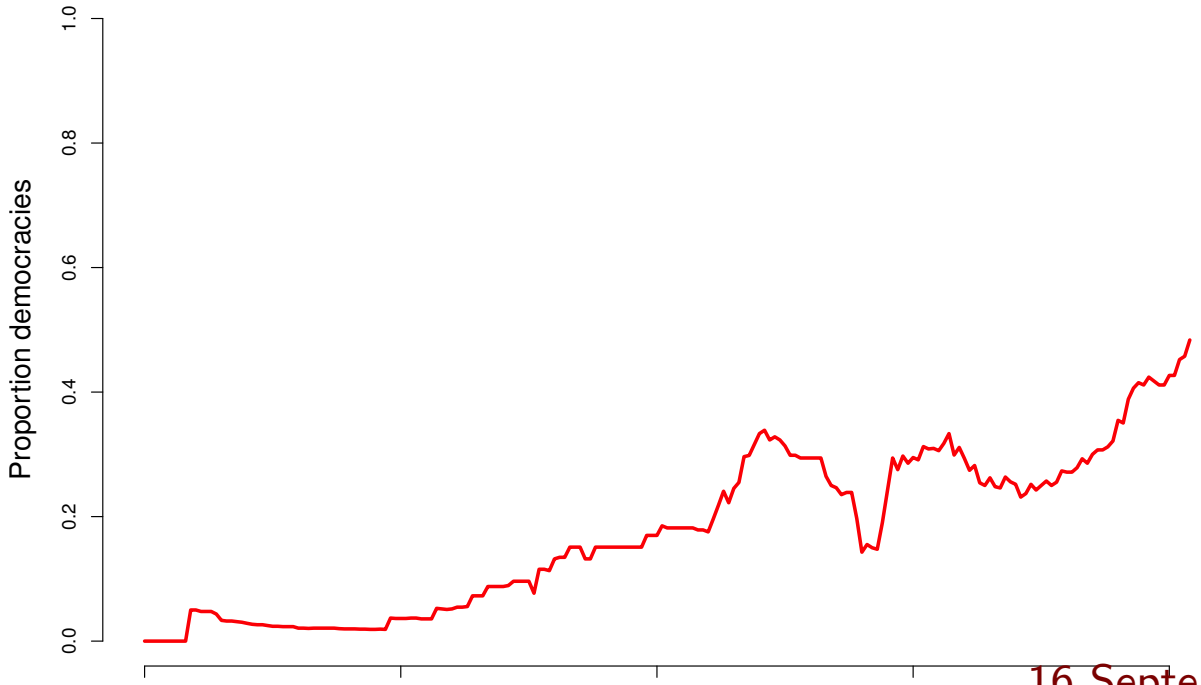
Probability to ever vote Fianna Fail

Time plot

When data is measured over time, another useful plot is a time plot, to see trends over time.

Time plot

When data is measured over time, another useful plot is a time plot, to see trends over time.



Exercise

Using the data from the first lecture:

- Produce a histogram for the Polity IV democracy score
- Produce a boxplot for the number of battle deaths

Exercise

Open the `demdev.dta` data file in SPSS. Select only cases where the year is 1990.

Produce the following plots of GDP per capita (`laggdppc`):

1. Histogram
2. Boxplot
3. Repeat for the logarithm of GDP
4. Repeat for energy usage (`energy2`)

Describe the distribution in words.

Outline

Frequency tables

Democracy is preferable to any other kind of government	1924	80.2%
In some circumstances, a non-democratic government can be preferable	182	7.6%
For someone like me, it doesn't matter what kind of government we have	124	5.2%
Missing / don't know	168	7.0%
Total	2398	100.0%

Example

6	2	3	0	7	9	6	7	5	3	7	0

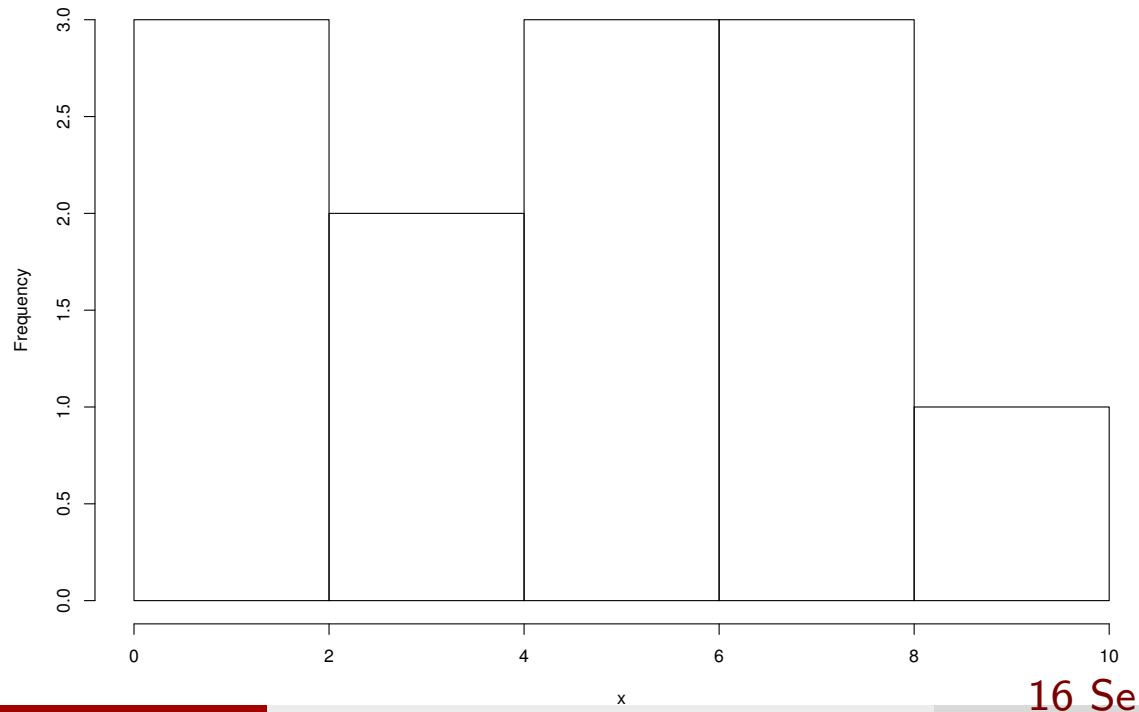
Example frequency table

6	2	3	0	7	9	6	7	5	3	7	0

value	count	percentage
0	2	17
2	1	8
3	2	17
5	1	8
6	2	17
7	3	25
9	1	8
total	12	100%

Example summary

6	2	3	0	7	9	6	7	5	3	7	0



Outline

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?"

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

The mode is the category with the highest frequency.

Example

6	2	3	0	7	9	6	7	5	3	7	0

Example mode

6	2	3	0	7	9	6	7	5	3	7	0

The mode is the value that most often occurs, i.e. 7.

Median

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

Median

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.

Example

6	2	3	0	7	9	6	7	5	3	7	0

Example median

6	2	3	0	7	9	6	7	5	3	7	0

First, sort the data:

0	0	2	3	3	5	6	6	7	7	7	9

Example median

6	2	3	0	7	9	6	7	5	3	7	0

First, sort the data:

0	0	2	3	3	5	6	6	7	7	7	9

Then, find the centre:

0	0	2	3	3	5	6	6	7	7	7	9

Example median

6	2	3	0	7	9	6	7	5	3	7	0

First, sort the data:

0	0	2	3	3	5	6	6	7	7	7	9

Then, find the centre:

0	0	2	3	3	5	6	6	7	7	7	9

The median is $\frac{5+6}{2} = 5.5$.

Mean

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

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

Example

6	2	3	0	7	9	6	7	5	3	7	0

Example mean

6	2	3	0	7	9	6	7	5	3	7	0

$$\bar{x} = \frac{6+2+3+0+7+9+6+7+5+3+7+0}{12} = \frac{55}{12} = 4.6$$

Exercise

Calculate appropriate measures of central tendency for each of the following variables related to party membership:

gender	class	years in party	gender	class	years in party
M	High	32	M	Medium	17
M	Low	32	M	Low	50
M	Low	25	M	Medium	25
F	High	12	F	High	10
F	Medium	21	F	Medium	33
M	Low	37	F	Low	15
F	Low	31			

(Healey 1997: 81)

Outliers

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

Mean or median?

- For a symmetric distribution, median and mean are the same.

Mean or median?

- For a symmetric distribution, median and mean are the same.
- The more skewed the distribution, the more mean and median differ.

Mean or median?

- 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 or median?

- 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

Measures of dispersion

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

Range

The range is the highest value minus the lowest value.

Range

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%.

Range

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.

Example

6	2	3	0	7	9	6	7	5	3	7	0

Example range

6	2	3	0	7	9	6	7	5	3	7	0

Range: $9 - 0 = 9$.

Variance

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

Variance

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

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

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

Example mean

6	2	3	0	7	9	6	7	5	3	7	0

$$\bar{x} = \frac{6+2+3+0+7+9+6+7+5+3+7+0}{12} = \frac{55}{12} = 4.6$$

Example variance

6	2	3	0	7	9	6	7	5	3	7	0

$$\bar{x} = \frac{6+2+3+0+7+9+6+7+5+3+7+0}{12} = \frac{55}{12} = 4.6$$

$$s_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x - \bar{x})^2 = \frac{1}{12}\sum_{i=1}^{N}(x - 4.6)^2$$

Example variance

6	2	3	0	7	9	6	7	5	3	7	0

$$s_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x-\bar{x})^2 = \frac{1}{12}\sum_{i=1}^{N}(x-4.6)^2$$

$$= \frac{1}{12}((1.4)^2 + (-2.6)^2 + (-1.6)^2 + (-4.6)^2 + (2.4)^2 + (4.4)^2$$
$$+ (1.4)^2 + (2.4)^2 + (0.4)^2 + (-1.6)^2 + (2.4)^2 + (-4.6)^2)$$

Example variance

6	2	3	0	7	9	6	7	5	3	7	0

$$s_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x-\bar{x})^2 = \frac{1}{12}\sum_{i=1}^{N}(x-4.6)^2$$

$$= \frac{1}{12}((1.4)^2 + (-2.6)^2 + (-1.6)^2 + (-4.6)^2 + (2.4)^2 + (4.4)^2$$
$$+ (1.4)^2 + (2.4)^2 + (0.4)^2 + (-1.6)^2 + (2.4)^2 + (-4.6)^2)$$

$$= \frac{1}{12}(2.0 + 6.7 + 2.5 + 21.0 + 5.8 + 19.5 + 2.0 + 5.8 + 0.2 + 2.5$$
$$+ 5.8 + 21.0) = \frac{94.9}{12} = 7.9$$

Variance: sample data

Country	Asylum seekers	$(x - \bar{x})$	$(x - \bar{x})^2$
Denmark	2.3		
Finland	3.6		
Ireland	4.3		
Norway	5.4		
Netherlands	12.4		
Belgium	16.0		
Sweden	17.5		
Germany	28.9		
United Kingdom	30.5		
France	50.1		
Total	170.8		

$$s^2 = \frac{\sum_{i=1}^{N}(x-\bar{x})^2}{N} \quad .$$

.

Variance: sample data

Country	Asylum seekers	$(x - \bar{x})$	$(x - \bar{x})^2$
Denmark	2.3	$2.3 - 17.1 = -14.8$	
Finland	3.6	$3.6 - 17.1 = -13.5$	
Ireland	4.3	$4.3 - 17.1 = -12.8$	
Norway	5.4	$5.4 - 17.1 = -11.7$	
Netherlands	12.4	$12.4 - 17.1 = -4.7$	
Belgium	16.0	$16.0 - 17.1 = -1.1$	
Sweden	17.5	$17.5 - 17.1 = 0.4$	
Germany	28.9	$28.9 - 17.1 = 11.8$	
United Kingdom	30.5	$30.5 - 17.1 = 13.4$	
France	50.1	$50.1 - 17.1 = 33$	
Total	170.8		

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

.

.

Variance: sample data

Country	Asylum seekers	$(x - \bar{x})$	$(x - \bar{x})^2$
Denmark	2.3	$2.3 - 17.1 = -14.8$	
Finland	3.6	$3.6 - 17.1 = -13.5$	
Ireland	4.3	$4.3 - 17.1 = -12.8$	
Norway	5.4	$5.4 - 17.1 = -11.7$	
Netherlands	12.4	$12.4 - 17.1 = -4.7$	
Belgium	16.0	$16.0 - 17.1 = -1.1$	
Sweden	17.5	$17.5 - 17.1 = 0.4$	
Germany	28.9	$28.9 - 17.1 = 11.8$	
United Kingdom	30.5	$30.5 - 17.1 = 13.4$	
France	50.1	$50.1 - 17.1 = 33$	
Total	170.8	0	

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

.

.

Variance: sample data

Country	Asylum seekers	$(x - \bar{x})$	$(x - \bar{x})^2$
Denmark	2.3	$2.3 - 17.1 = -14.8$	$(-14.8)^2 = 219.0$
Finland	3.6	$3.6 - 17.1 = -13.5$	$(-13.5)^2 = 182.3$
Ireland	4.3	$4.3 - 17.1 = -12.8$	$(-12.8)^2 = 163.8$
Norway	5.4	$5.4 - 17.1 = -11.7$	$(-11.7)^2 = 136.9$
Netherlands	12.4	$12.4 - 17.1 = -4.7$	$(-4.7)^2 = 22.1$
Belgium	16.0	$16.0 - 17.1 = -1.1$	$(-1.1)^2 = 1.2$
Sweden	17.5	$17.5 - 17.1 = 0.4$	$(0.4)^2 = 0.2$
Germany	28.9	$28.9 - 17.1 = 11.8$	$(11.8)^2 = 139.2$
United Kingdom	30.5	$30.5 - 17.1 = 13.4$	$(13.4)^2 = 179.6$
France	50.1	$50.1 - 17.1 = 33$	$(33)^2 = 1089$
Total	170.8	0	2133.3

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

.

.

Variance: sample data

Country	Asylum seekers	$(x - \bar{x})$	$(x - \bar{x})^2$
Denmark	2.3	$2.3 - 17.1 = -14.8$	$(-14.8)^2 = 219.0$
Finland	3.6	$3.6 - 17.1 = -13.5$	$(-13.5)^2 = 182.3$
Ireland	4.3	$4.3 - 17.1 = -12.8$	$(-12.8)^2 = 163.8$
Norway	5.4	$5.4 - 17.1 = -11.7$	$(-11.7)^2 = 136.9$
Netherlands	12.4	$12.4 - 17.1 = -4.7$	$(-4.7)^2 = 22.1$
Belgium	16.0	$16.0 - 17.1 = -1.1$	$(-1.1)^2 = 1.2$
Sweden	17.5	$17.5 - 17.1 = 0.4$	$(0.4)^2 = 0.2$
Germany	28.9	$28.9 - 17.1 = 11.8$	$(11.8)^2 = 139.2$
United Kingdom	30.5	$30.5 - 17.1 = 13.4$	$(13.4)^2 = 179.6$
France	50.1	$50.1 - 17.1 = 33$	$(33)^2 = 1089$
Total	170.8	0	2133.3

$$s^2 = \frac{\sum_{i=1}^{N}(x-\bar{x})^2}{N} = \frac{2133.3}{10} = 213.3.$$

.

Standard deviation

Standard deviation:

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

Example variance

6	2	3	0	7	9	6	7	5	3	7	0

$$s_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x-\bar{x})^2 = \frac{1}{12}\sum_{i=1}^{N}(x-4.6)^2$$

$$= \frac{1}{12}((1.4)^2 + (-2.6)^2 + (-1.6)^2 + (-4.6)^2 + (2.4)^2 + (4.4)^2$$
$$+ (1.4)^2 + (2.4)^2 + (0.4)^2 + (-1.6)^2 + (2.4)^2 + (-4.6)^2)$$

$$= \frac{1}{12}(2.0 + 6.7 + 2.5 + 21.0 + 5.8 + 19.5 + 2.0 + 5.8 + 0.2 + 2.5$$
$$+ 5.8 + 21.0) = \frac{94.9}{12} = 7.9$$

Example standard deviation

6	2	3	0	7	9	6	7	5	3	7	0

Standard deviation: $s = \sqrt{s^2} = \sqrt{7.9} = 2.8$.

Variance: sample data

Country	Asylum seekers	$(x - \bar{x})$	$(x - \bar{x})^2$
Denmark	2.3	$2.3 - 17.1 = -14.8$	$(-14.8)^2 = 219.0$
Finland	3.6	$3.6 - 17.1 = -13.5$	$(-13.5)^2 = 182.3$
Ireland	4.3	$4.3 - 17.1 = -12.8$	$(-12.8)^2 = 163.8$
Norway	5.4	$5.4 - 17.1 = -11.7$	$(-11.7)^2 = 136.9$
Netherlands	12.4	$12.4 - 17.1 = -4.7$	$(-4.7)^2 = 22.1$
Belgium	16.0	$16.0 - 17.1 = -1.1$	$(-1.1)^2 = 1.2$
Sweden	17.5	$17.5 - 17.1 = 0.4$	$(0.4)^2 = 0.2$
Germany	28.9	$28.9 - 17.1 = 11.8$	$(11.8)^2 = 139.2$
United Kingdom	30.5	$30.5 - 17.1 = 13.4$	$(13.4)^2 = 179.6$
France	50.1	$50.1 - 17.1 = 33$	$(33)^2 = 1089$
Total	170.8	0	2133.3

$$s^2 = \frac{\sum_{i=1}^{N}(x-\bar{x})^2}{N} = \frac{2133.3}{10} = 213.3.$$

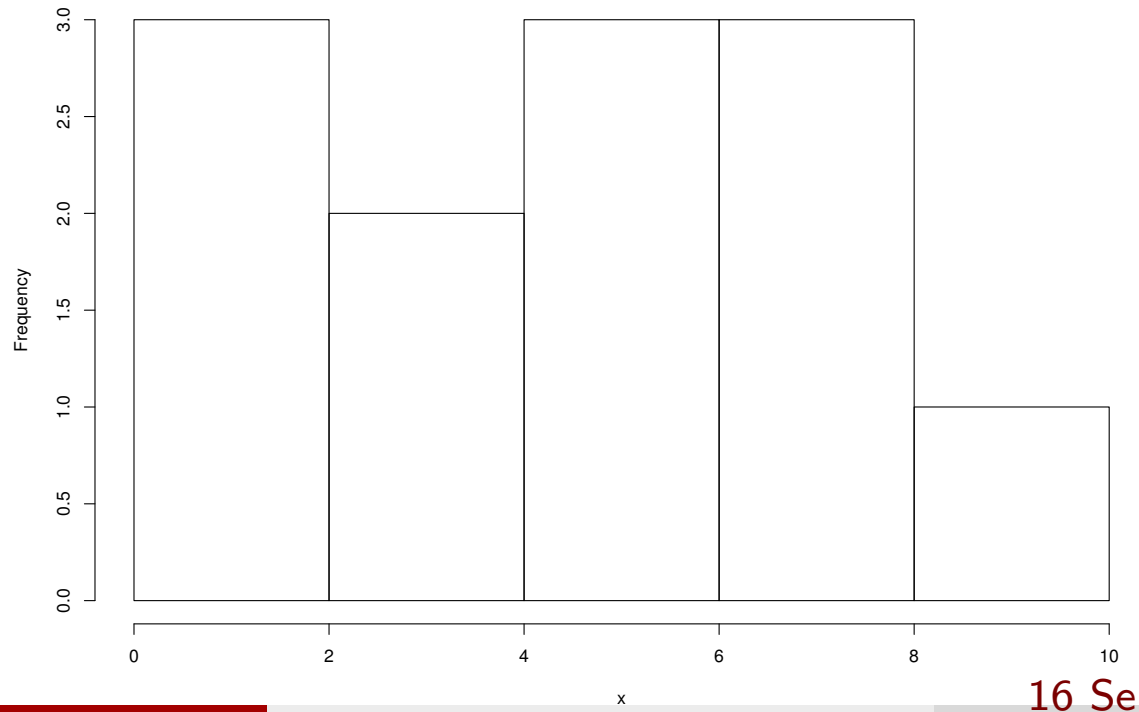
$$s = \sqrt{s^2} = \sqrt{213.3} = 14.6.$$

Variance and outliers

The variance refers to the variation in the data around the mean. It is similarly sensitive to outliers - a few extreme values in the data can significantly increase the variance.

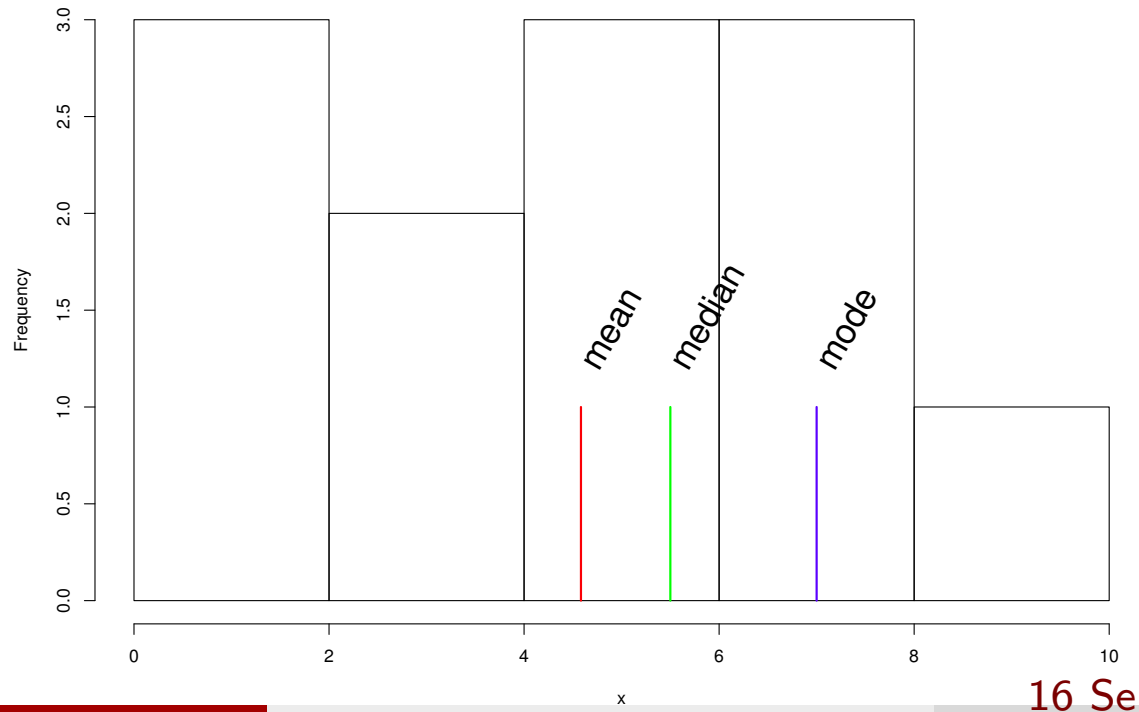
Example summary

6	2	3	0	7	9	6	7	5	3	7	0



Example summary

6	2	3	0	7	9	6	7	5	3	7	0



Exercise

Open the `demdev.dta` data file in SPSS. Select only year 1980.

1. Take a glance at the data.
2. Using SPSS, produce a histogram of GDP per capita (`laggdppc`).
3. Calculate the median, mean, range, variance, and standard deviation of GDP per capita.
4. Create a new variable that is the log of GDP per capita.
5. Repeat steps 2 and 3 on the logged variable.
6. Are there any extreme cases? What happens if you remove this?