Statistics Applied to Bioinformatics

Descriptive statistics

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Overview: descriptive statistics

- Data description
  - Enumeration
  - Frequency distribution
  - Class frequency distribution
- Graphical representations
  - Histogram
  - Frequency polygon
- Data reduction
  - Parameters of location (= central tendency)
  - Parameters of dispersion
  - Parameters of dissymmetry
  - Parameters of kurtosis
- Practical: descriptive statistics with R

Enumeration

- Example 1:
  - ORF lengths in the yeast genome
  - 3573 3531 987 648 1929 … (6217 values)
- Example 2:
  - Level of regulation at time point 2 during the diauxic shift
  - 1.19 1.23 1.32 1.33 0.88 … (6153 values)
- Not very convenient to read and interpret

Frequency distribution

- For each possible value (x_i), count its number of occurrences (n_i) in the enumeration

Occurrences n_i $\sum_{i=1}^{p} n_i = n$

Cumulative occurrences $N_i = \sum_{j=1}^{i} n_j$ $N_p = n$

- From these occurrences (also called absolute frequencies), one can also calculate

Frequencies $f_i = n_i / n$ $\sum_{i=1}^{p} f_i = 1$

Cumulative frequencies $F_i = \sum_{j=1}^{i} f_j$ $F_p = 1$

Frequency distribution example

x_i	n_i	N_i	f_i	F_i
1	0	0	0	0
2	0	0	0	0
…	…	…	…	…
77	0	0	0	0
78	3	3	0	0
…	…	…	…	…
327	26	327	0.004	0.053
328	0	327	0	0.053
329	0	327	0	0.053
330	24	351	0.004	0.056
331	0	351	0	0.056
…	…	…	…	…
14 732	0	6216	0	1
14 733	1	6217	0	1

- Still not very convenient when there are 15,000 possible distinct values

Class grouping

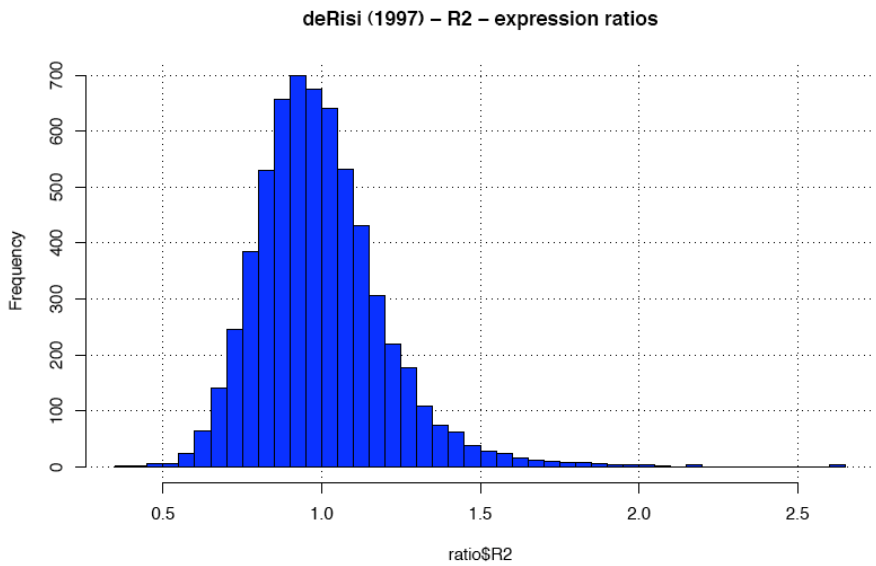
min	max	mid	occ	occ.cum	freq	freq.cum	intensity
0.0	0.2	0.1	0	0	0.0000	0.0000	0.0000
0.2	0.4	0.3	2	2	0.0003	0.0003	0.0016
0.4	0.6	0.5	43	45	0.0070	0.0073	0.0349
0.6	0.8	0.7	860	905	0.1398	0.1471	0.6988
0.8	1.0	0.9	2599	3504	0.4224	0.5695	2.1120
1.0	1.2	1.1	1895	5399	0.3080	0.8775	1.5399
1.2	1.4	1.3	523	5922	0.0850	0.9625	0.4250
1.4	1.6	1.5	154	6076	0.0250	0.9875	0.1251
1.6	1.8	1.7	45	6121	0.0073	0.9948	0.0366
1.8	2.0	1.9	20	6141	0.0033	0.9980	0.0163
2.0	2.2	2.1	9	6150	0.0015	0.9995	0.0073
2.2	2.4	2.3	0	6150	0.0000	0.9995	0.0000
2.4	2.6	2.5	0	6150	0.0000	0.9995	0.0000
2.6	2.8	2.7	3	6153	0.0005	1.0000	0.0024

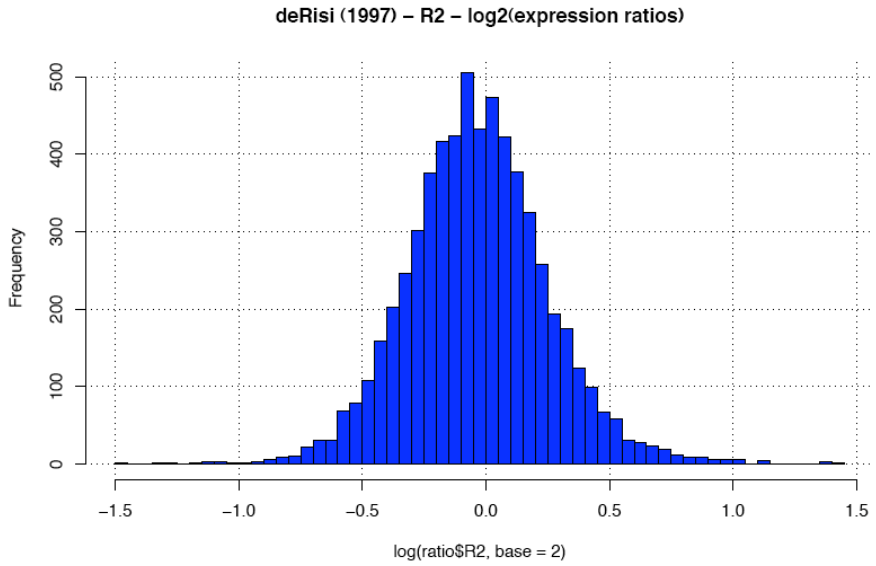
Class frequency distribution: level of gene regulation (red/green ratio) at time point 2 during the diauxic shift

Summary: data description

- Class grouping is useful for graphical and tabular representations (summary reports)
- Whenever possible, avoid class grouping for calculation
  - using the class centre instead of the list values introduces a bias

Histogram

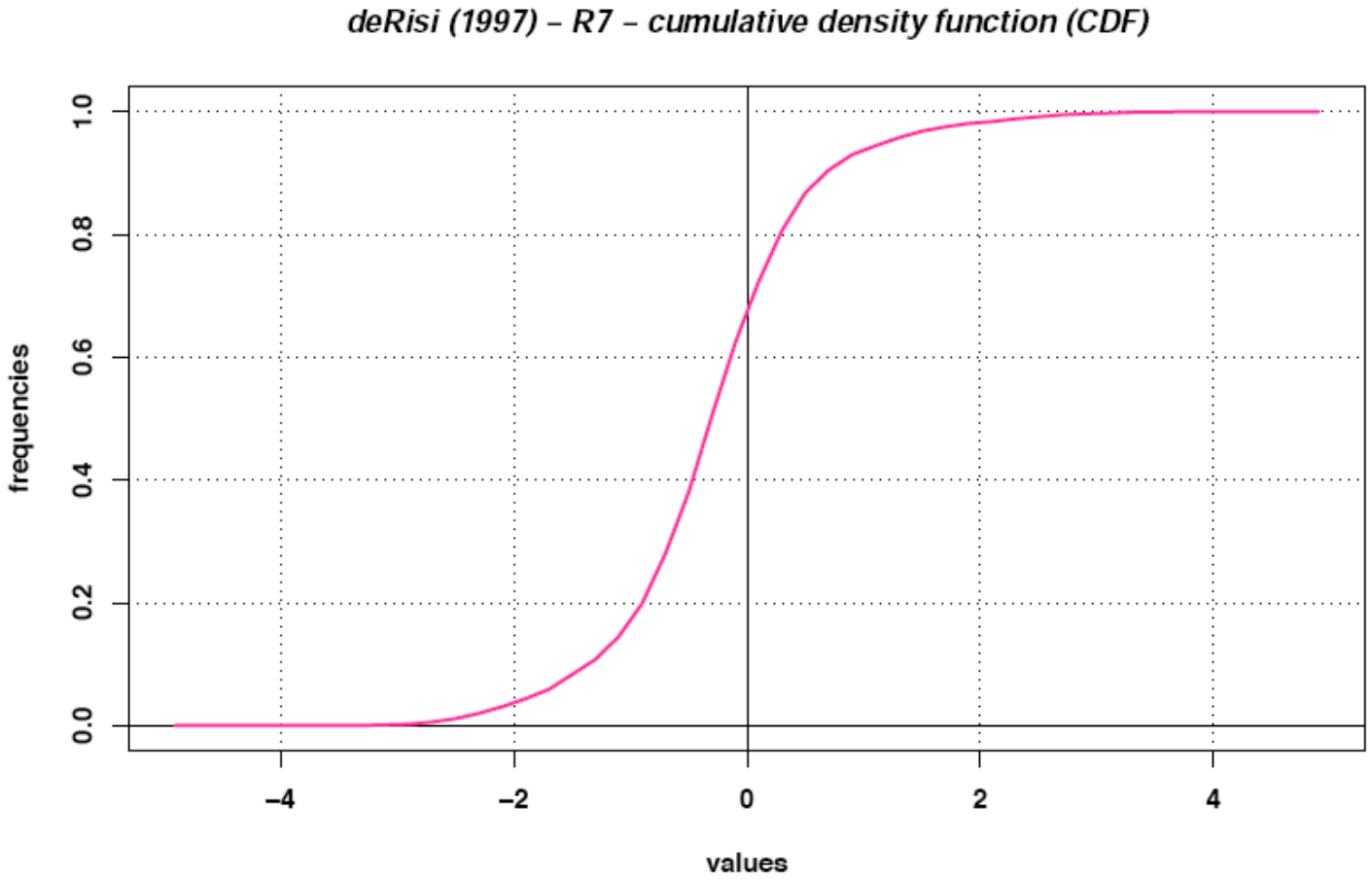




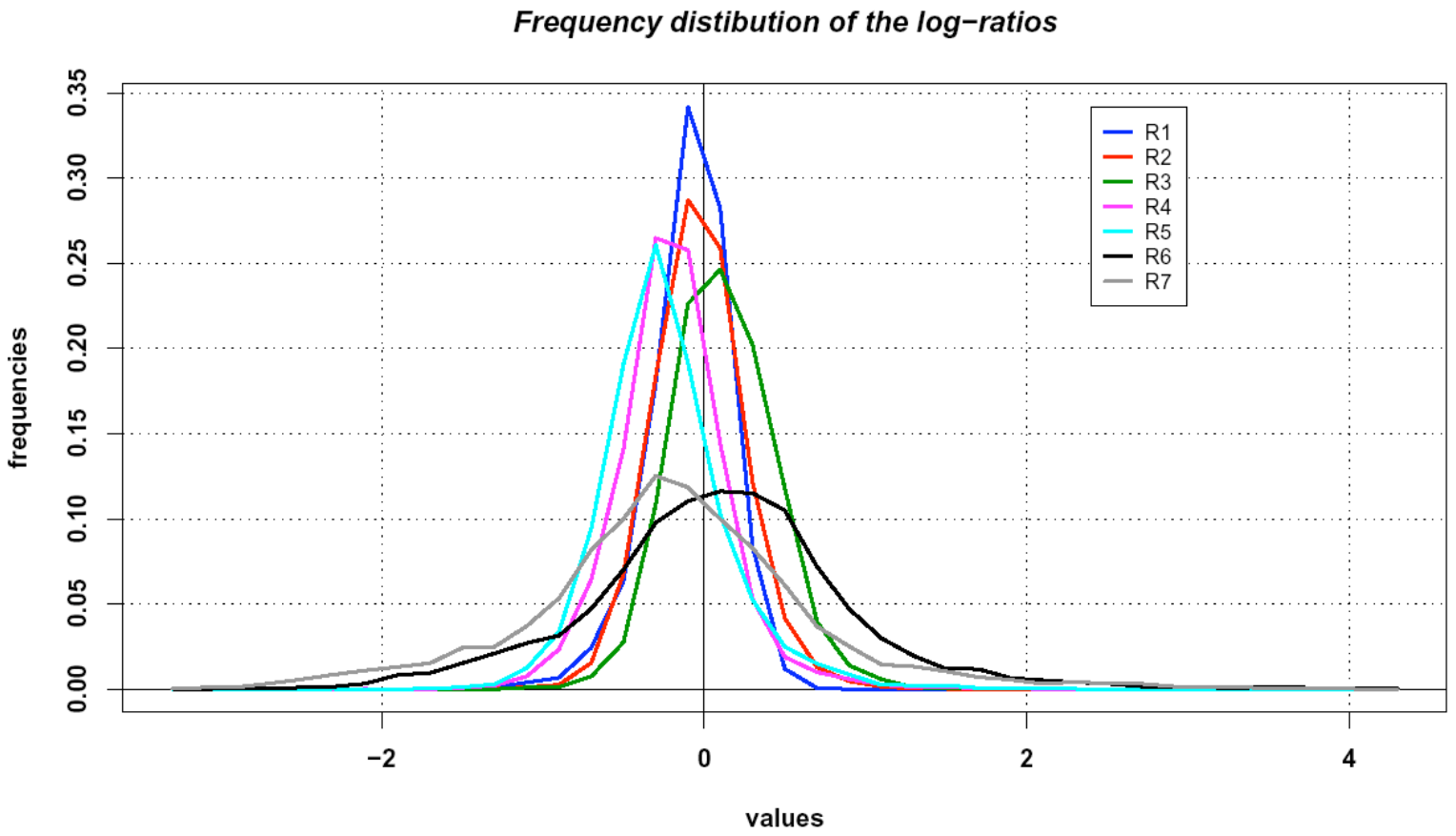
- The area above a given range is proportional to the frequency of this range
- Appropriate for absolute or relative frequencies
- Appropriate for representing class frequencies

Frequency polygon – cumulative frequencies

- **Cumulative density function (CDF)**
  - the height (not the area) directly indicates the cumulative frequency of all values below x



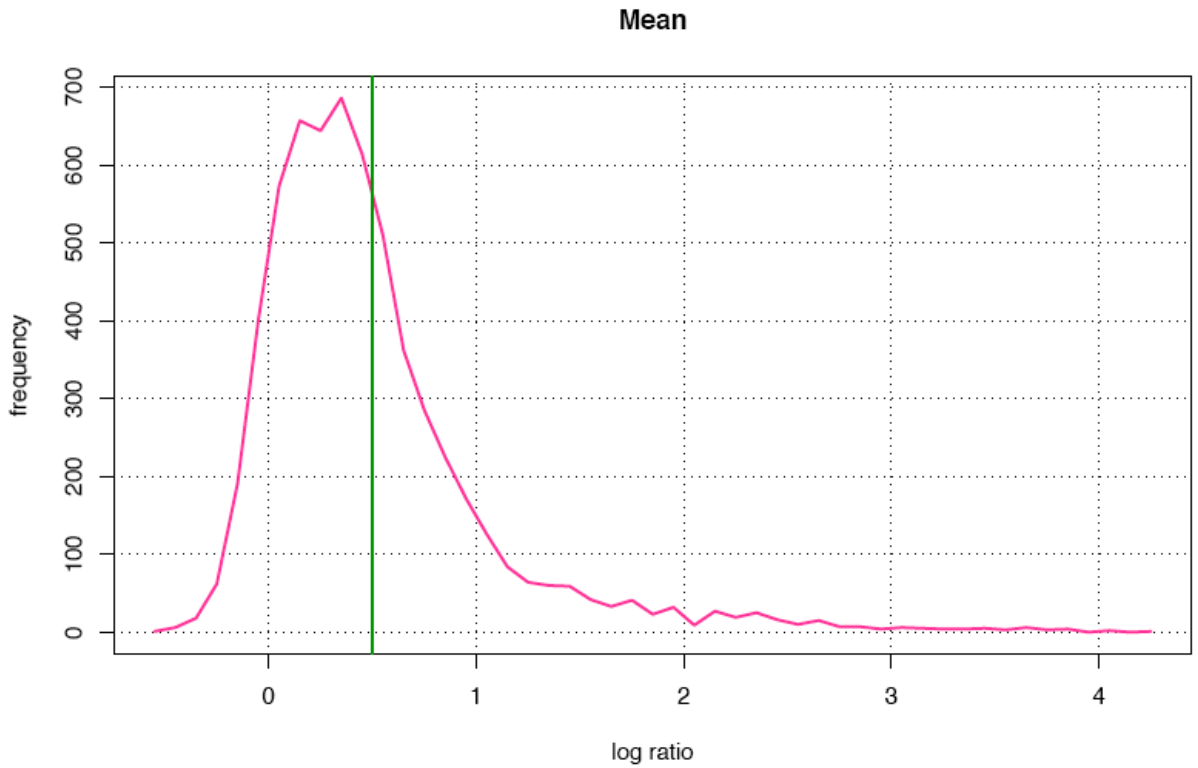
Frequency polygon – multiple curves



- Advantage: allows to visualise multiple curves on the same plot.
- Weakness: contrarily to histograms, the surface below the curve is not exactly proportional to the frequency.

Location parameters - Arithmetic mean

$$\bar{x} = a_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$



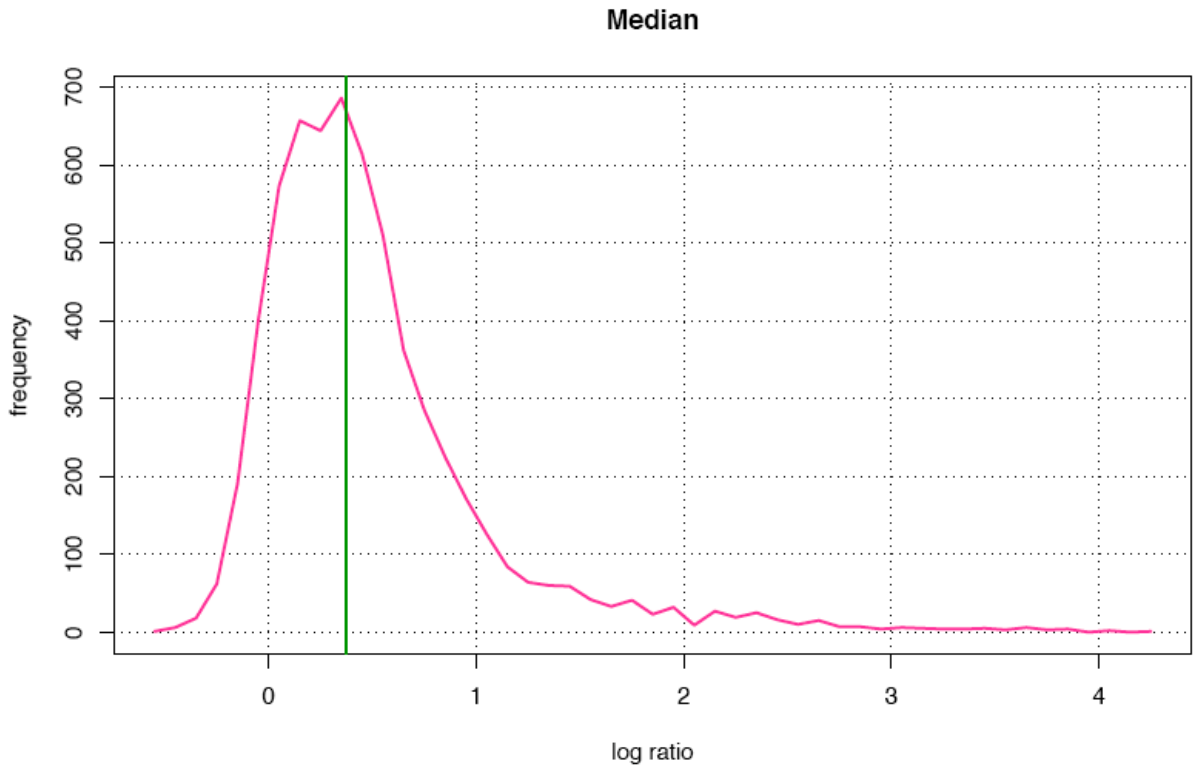
- The mean is the gravity center of the distribution
- Beware: the mean is strongly influenced by outliers.
- Statistical "outliers" are generally biologically relevant objects (e.g. regulated genes).

Location parameters - Median

$$\tilde{m} = x_{(n+1)/2}$$ *if n is odd*

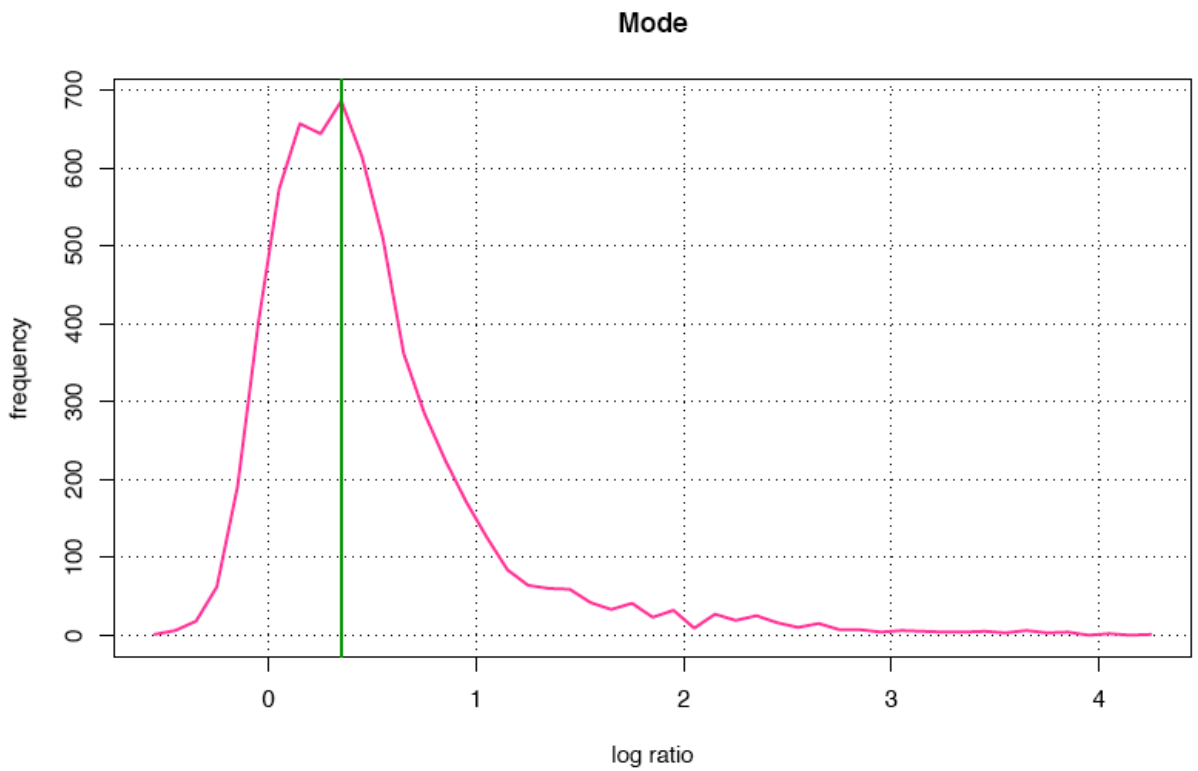
$$\tilde{m} = \frac{x_{n/2} + x_{n/2+1}}{2}$$ *if n is even*

- Left area = right area
- The median is robust to the presence of outliers because it does not take into account the values themselves, but the ranks.



Location parameters - Mode

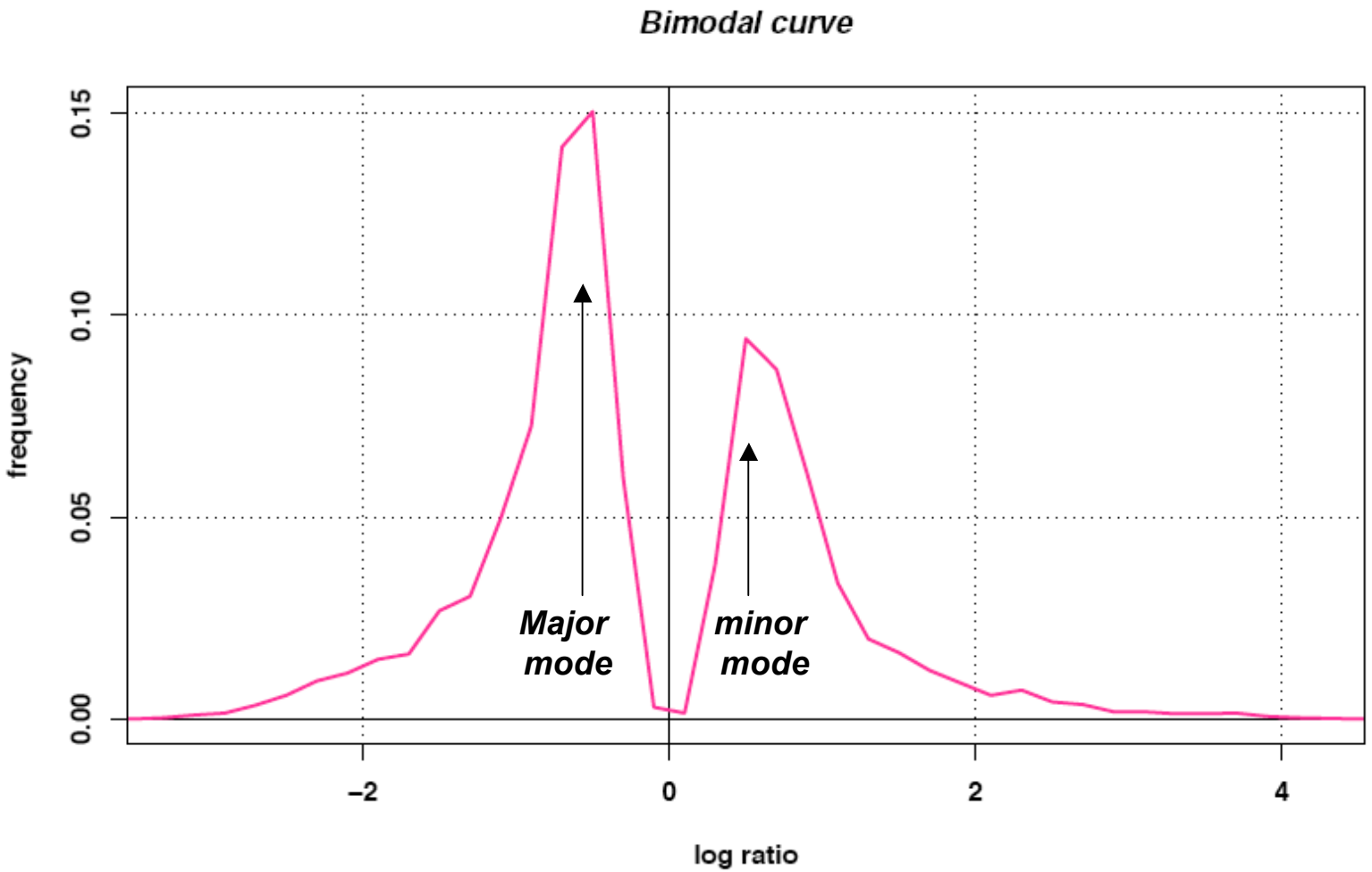
$$M'' = \arg\max_x \left(F(x) \right)$$



- The mode is the value associated to the maximal frequency
- Not a very robust statistics:
  - for small samples, the distribution can be irregular
  - the precise location of the mode is depends on the choice of class boundaries.

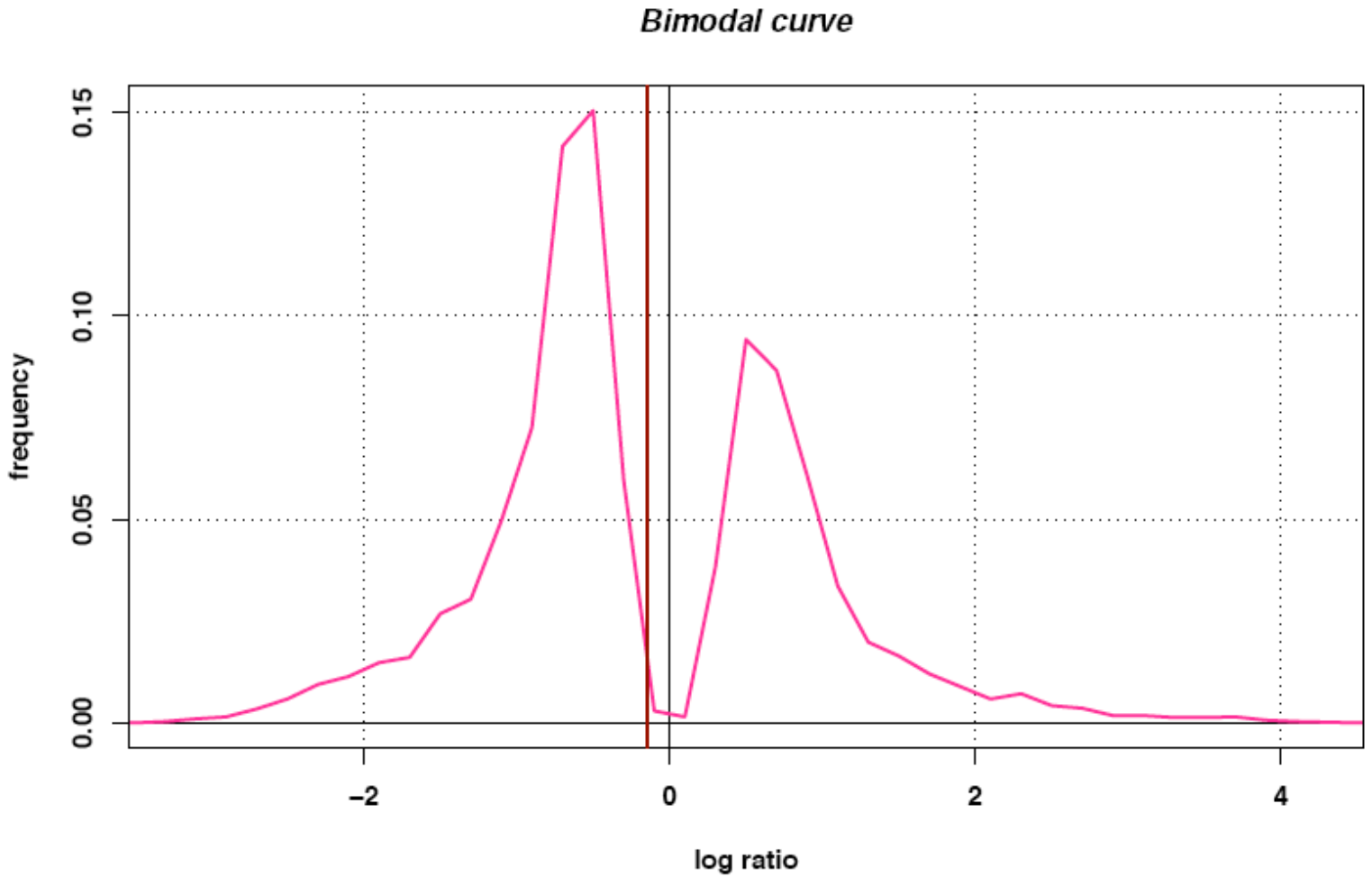
Multimodal curves

- E.g.: Extreme values in the gene expression data



Mean and bimodal curves

- For bimodal curves, the mean and the median poorly reflect the tendency of the population (almost no point has the mean value)



Comparison of location parameters

- Symmetric distributions
  →mean=median
- Unimodal and symmetric
  → mode=mean

Dispersion parameters - Range

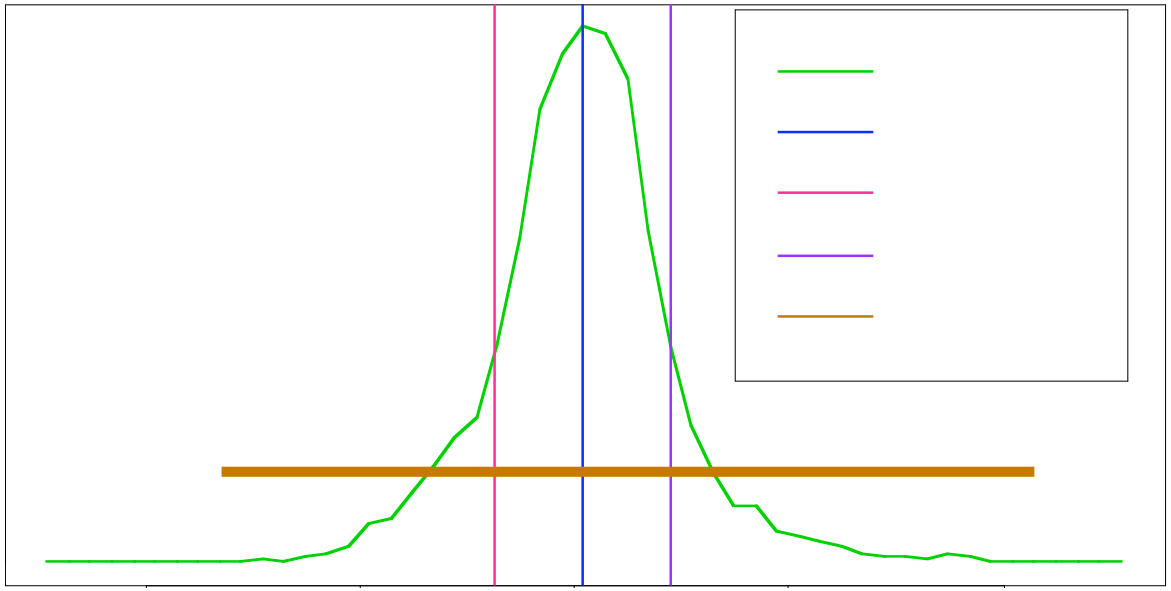
- $Range = max - min$
- The range only reflects 2 values: the min and max
- Strongly affected by outliers →
  poor representation of the general characteristics of the sample

Dispersion parameters - Variance

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

- The variance is strongly affected by exceptional values

Dispersion parameters - Standard deviation



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

- Same units as the mean

Dispersion parameters – Variation coefficient

- $V = s/m$
- Has no unit
- Makes only sense if the data is measured on a scale with a real 0 (e.g. Kelvin degrees)
- Counter-example
  - for a sample of mean=0 (with negative and positive values), V is infinite (it is thus absolutely not appropriate)

Dispersion parameters - interquartile range (IQR)

- The quartiles are an extension of the median
  - The first quartile ($Q1$) leaves 1/4 of the observations on its left.
  - The second quartile is the median.
  - The third quartile ($Q3$) leaves 3/4 of the observations on its left.
- The inter-quartile range (IQR=$Q3$-$Q1$) indicates the spread of the 50% central values.
- The inter-quartile range is robust to outliers, since it is is based on the ranks rather than the values themselves.

Dispersion parameters - MAD

$$MAD = k * median\left(\left|x - median(x)\right|\right)$$

- The median of the sample is used as a robust estimator of the central tendency.
- The **median absolute deviation** (MAD) is the median of the absolute difference between each value and the median.
- The constant k ensures consistency
- With a value of $k=1.4826$, for normal population, the expected MAD is the standard deviation.
  - E[MAD]=σ
- The MAD is robust to outliers.

Moments

- k-order moment about c

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

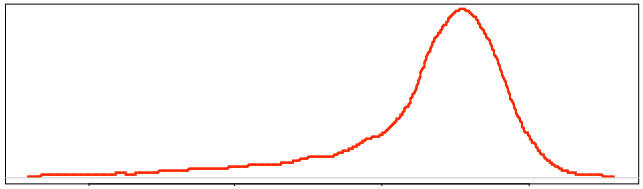
$c = center$

$k = order$

- In particular
  - a_k Moment about the origin ($c=0$)
  - a_1 = arithmetic mean
  - m_k Central moment
    = moment about the mean ($c=m=a_1$)
  - m_1 always = 0
  - m_2 = variance

Dissymmetry parameters – g1

Row.min ; g1 = -0.4



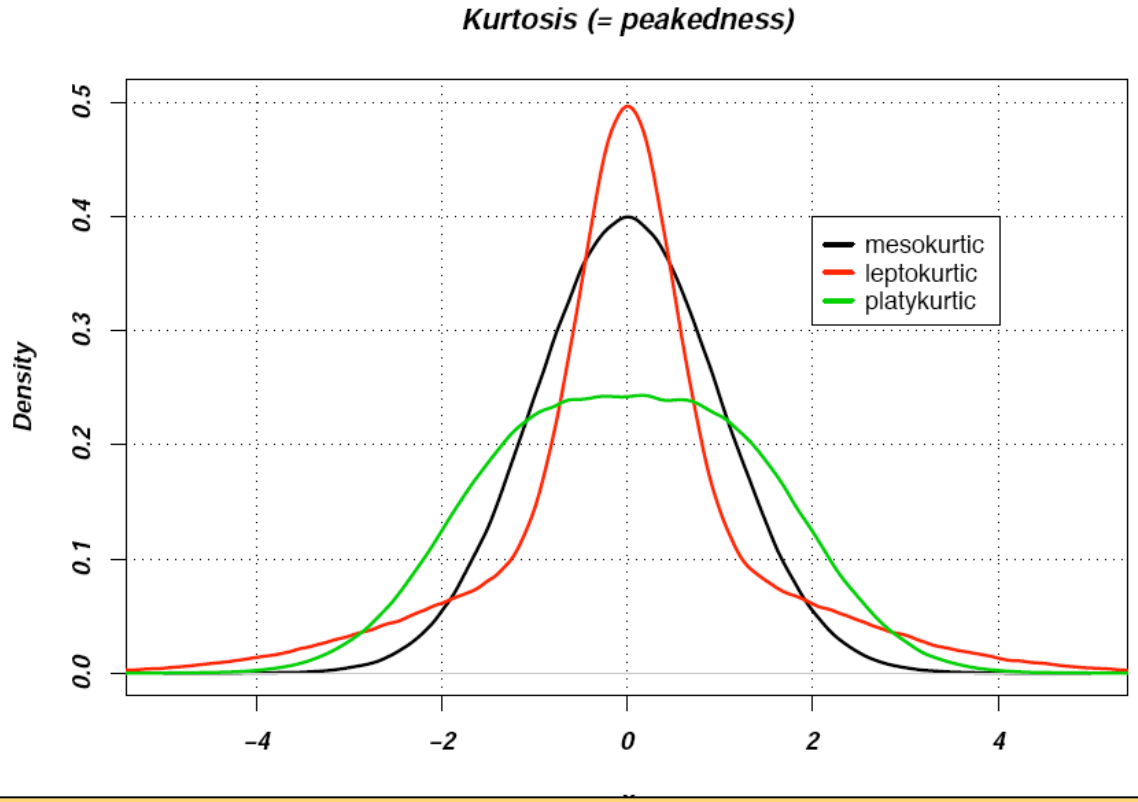
$$g_1 = m_3 \Big/ \left(m_2\right)^{3/2} = m_3 / s_3$$

- g1 < 0 → left skewed
- g1 = 0 → symmetric
- g1 > 0 → right skewed

Random normal ; g1 = 0.08

Row.max ; g1 = 0.93

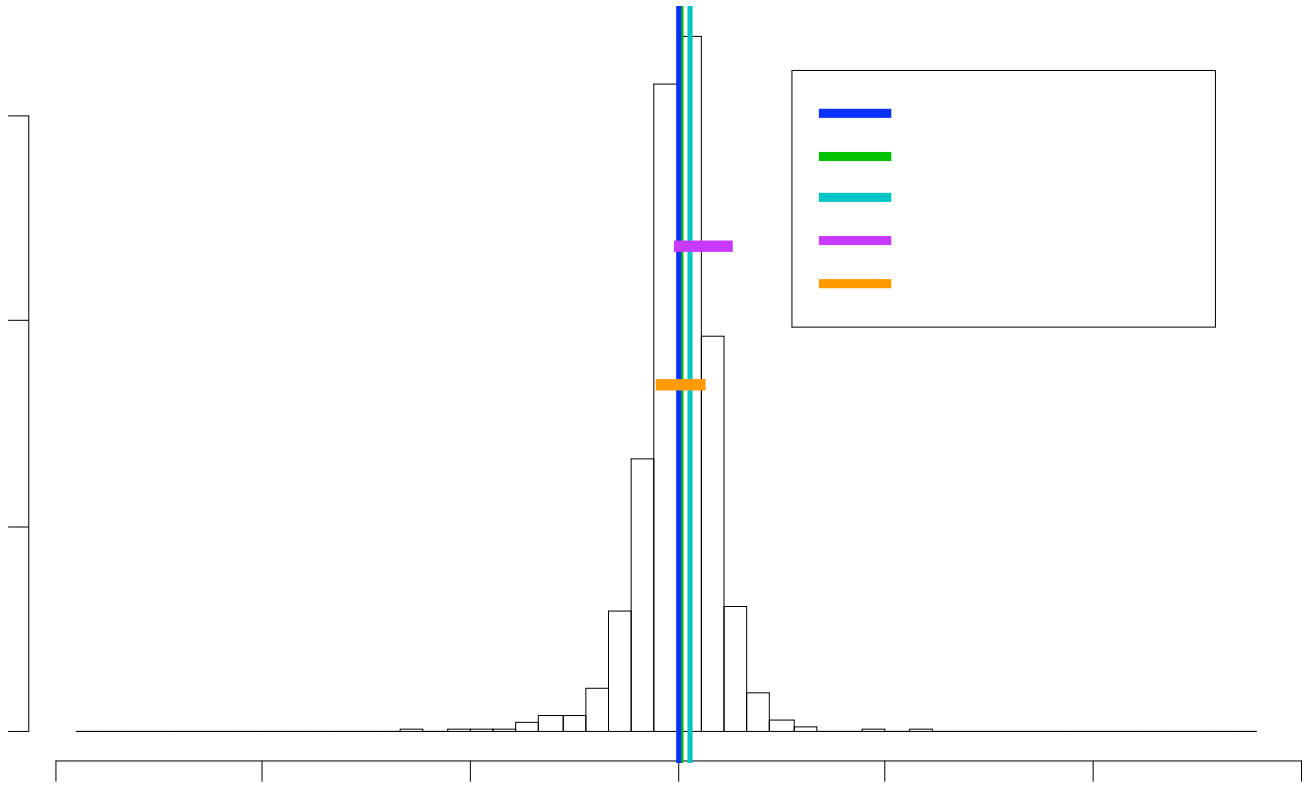
Kurtosis (flatness) parameters – g2



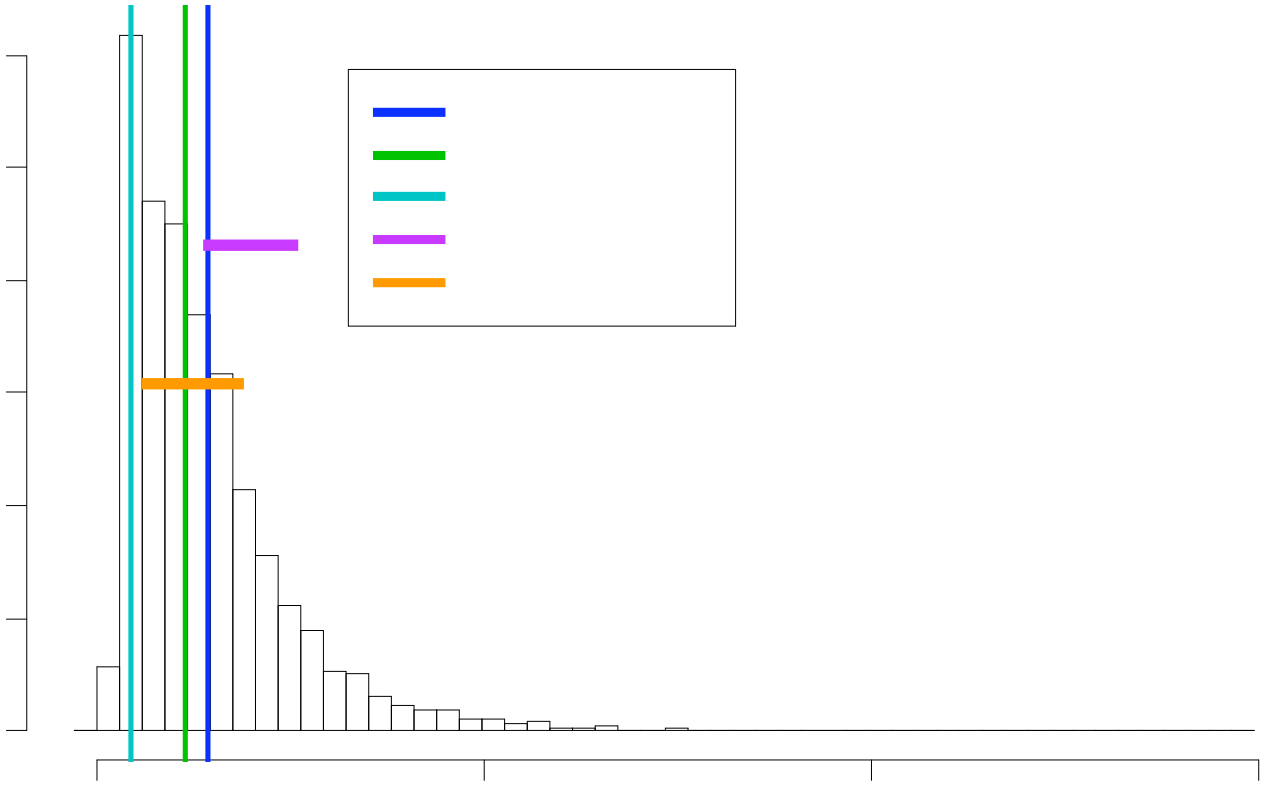
$$g_2 = \left(m_4 / m_2^2\right) - 3 = m_4 / s^4 - 3 = b_2 - 3$$

- g = 0 → **mesokurtic**
- g > 0 → **leptokurtic** (peaked)
- g < 0 → **platykurtic** (flat)

Descriptive parameters - DNA chip sample



Descriptive parameters - yeast ORF lengths



Statistics Applied to Bioinformatics

Descriptive statistics - exercises

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Descriptive statistics - Exercises

- Explain why the median is a more robust estimator of central tendency than the mean ?
- Which kind of problem can be indicated by
  - a platykurtic distribution ?
  - a mesokurtic distribution ?