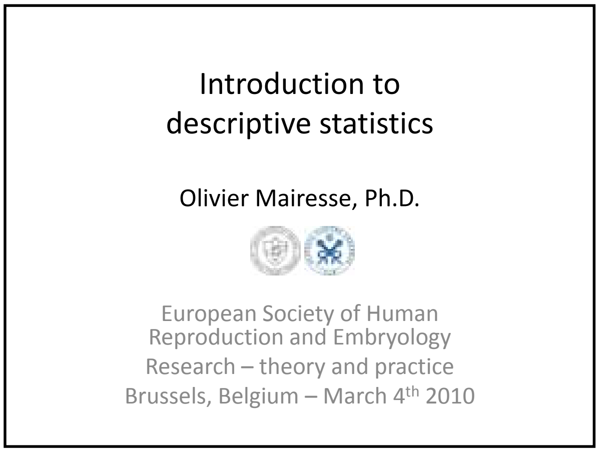
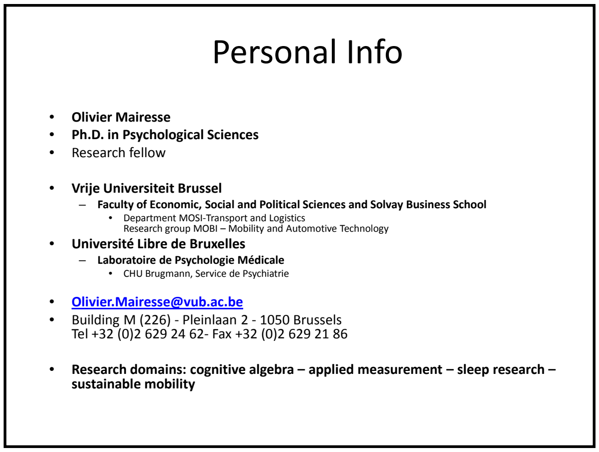


Introduction to descriptive statistics

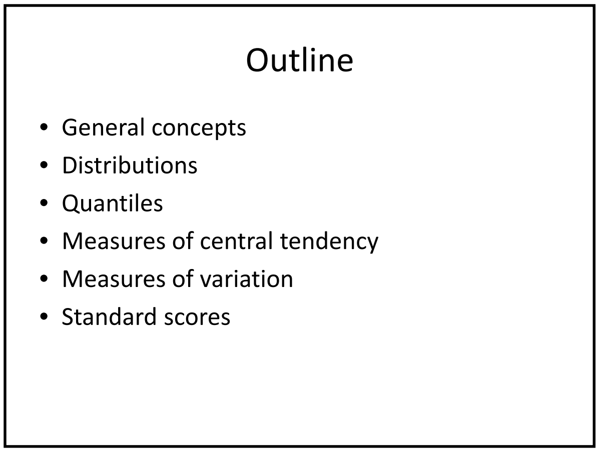
Olivier Mairesse, Ph.D.

European Society of Human Reproduction and Embryology
Research – theory and practice
Brussels, Belgium – March 4th 2010



Personal Info

- **Olivier Mairesse**
- **Ph.D. in Psychological Sciences**
- Research fellow
- **Vrije Universiteit Brussel**
  - **Faculty of Economic, Social and Political Sciences and Solvay Business School**
    - Department MOSI-Transport and Logistics
      Research group MOBI – Mobility and Automotive Technology
- **Université Libre de Bruxelles**
  - **Laboratoire de Psychologie Médicale**
    - CHU Brugmann, Service de Psychiatrie
- Olivier.Mairesse@vub.ac.be
- Building M (226) - Pleinlaan 2 - 1050 Brussels
  Tel +32 (0)2 629 24 62- Fax +32 (0)2 629 21 86
- **Research domains: cognitive algebra – applied measurement – sleep research – sustainable mobility**



Outline

- General concepts
- Distributions
- Quantiles
- Measures of central tendency
- Measures of variation
- Standard scores

Measurement

The process of obtaining the magnitude of a quantity (e.g. length, weight, …) relative to a unit of measurement (e.g. meter, kilogram).

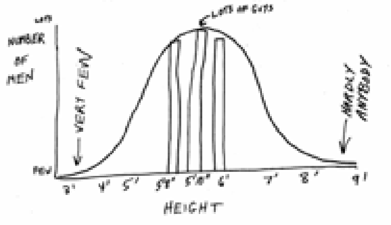
How much of what?

Levels of measurement

Measurement level	characteristic	example
Ratio	*possess an absolute zero*	length
Interval	*distance is meaningful*	temperature in °C
Ordinal	*attributes can be ordered*	patient admission rankings
Nominal	*attributes are only named*	types of medication

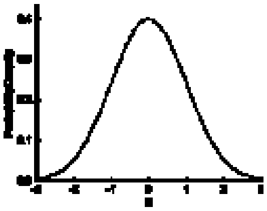
Frequency distribution

- The number of times a value appears in a sample



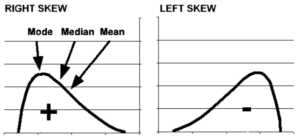
Probability distribution

- Normal (Gaussian) distribution → The Bell Curve
- Symmetric; M = 0; SD= 1



Skewed distributions

- Tilt in the normal distribution



The skewed part (the tail) is to the right.
Most cases are to the left.
There is more count in the tail than expected in a normal distribution.
"Positive skew"

The skewed part is to the left.
Most cases are to the right.

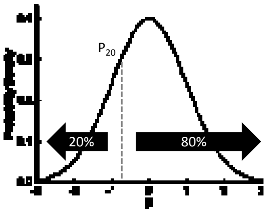
There is more count in the tail than expected in a normal distribution.
"Negative skew"

Describing variables

- Have a variable with observations on a (possibly the largest) number of cases
- Produce a number of summary measures that **meaningfully** characterize those data
- Focus here is on
  - Distribution
  - Central tendency
  - Variation

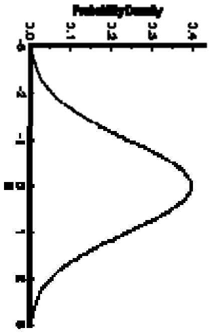
Quantiles:
Percentiles, deciles and quartiles

- Percentile: value of a variable below which a certain percent of observations fall (e.g. P_{20})
- 1/100 “jumps”



Relationship between percentiles, deciles and quartiles

- $P_{10} = D_1$
- $P_{20} = D_2$
- $P_{25} = Q_1$
- $P_{30} = D_3$
- $P_{40} = D_4$
- **$P_{50} = D_5 = Q_2$**
- $P_{60} = D_6$
- $P_{70} = D_7$
- $P_{75} = Q_3$
- $P_{80} = D_8$
- $P_{90} = D_9$
- $P_{100} = D_{10} = Q_4$

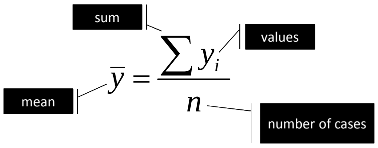


Measures of central tendency

- Mean
- Median
- Mode

Mean

- Sum of the values divided by the number of cases



$$\bar{y} = \frac{\sum y_i}{n}$$

Calculating mean temperatures

patient	temperature
Liam	36.6
Luka	35.8
Noah	37.2
Mohammed	36.8
Yasmine	36.9
Otis	37
John	40.5
Peter	36.4
Lily	36.6
Milo	36.2
SUM	**370**

- Sum of values

$$\sum y_i = 370$$

- Number of cases

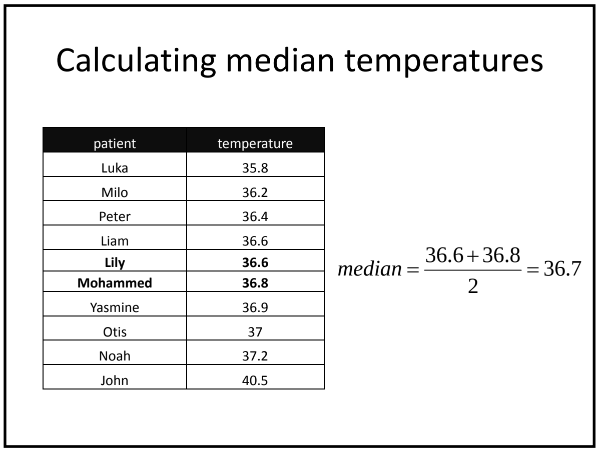
$$n = 10$$

- Calculate mean

$$\bar{y} = \frac{\sum y_i}{n} = \frac{370}{10} = 37$$

Median

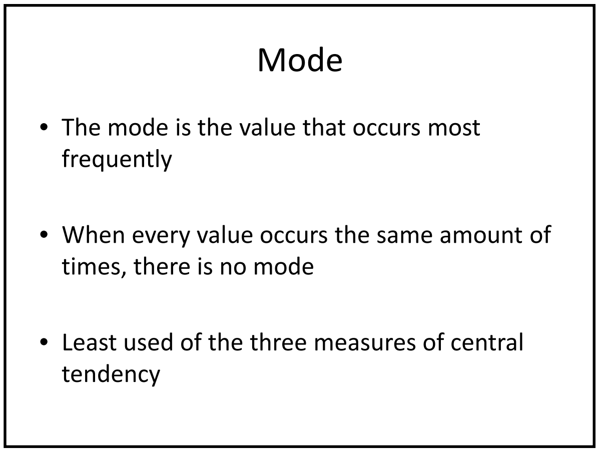
- The median represents the middle of the ordered sample data
- When the sample size is odd, the median is the middle value
- When the sample size is even, the median is the midpoint/mean of the two middle values
- $P_{50} = D_5 = Q_2$ = median!



Calculating median temperatures

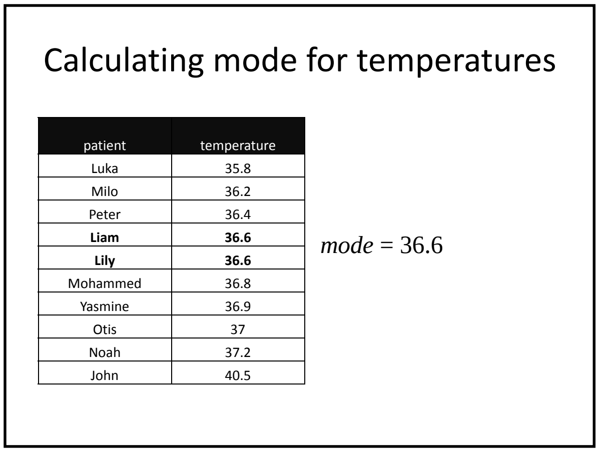
patient	temperature
Luka	35.8
Milo	36.2
Peter	36.4
Liam	36.6
Lily	**36.6**
Mohammed	**36.8**
Yasmine	36.9
Otis	37
Noah	37.2
John	40.5

$$median = \frac{36.6 + 36.8}{2} = 36.7$$



Mode

- The mode is the value that occurs most frequently
- When every value occurs the same amount of times, there is no mode
- Least used of the three measures of central tendency



Calculating mode for temperatures

patient	temperature
Luka	35.8
Milo	36.2
Peter	36.4
Liam	**36.6**
Lily	**36.6**
Mohammed	36.8
Yasmine	36.9
Otis	37
Noah	37.2
John	40.5

$$mode = 36.6$$

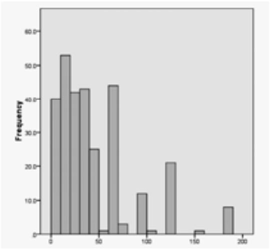
Measures of central tendency and levels of measurement

- Mean assumes numerical values and requires interval data for meaningful descriptions
- Median requires ordering of values and is used with both interval and ordinal data
- Mode only involves determination of most common value and is used with interval, ordinal, and nominal data

The mean and median and the distribution of the data

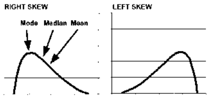
- For symmetric distributions, the mean and the median are the same
- For skewed distributions, the mean lies in the direction of the skew (the longer tail) relative to the median

Comparison of mean and median



Yearly amount of succesful IVFs/hospital

n = 294
Mean = 41.72
Median = 30



The skewed part (the tail) is to the right. Most cases are to the left. There is more count in the tail than expected in a normal distribution. "Positive skew"

The skewed part is to the left. Most cases are to the right.

There is more count in the tail than expected in a normal distribution. "Negative skew"

Comparison of mean and median

- Mean
  - Uses all of the data
  - Has desirable statistical properties
  - Affected by extreme high or low values (**outliers**)
  - May not best characterize skewed distributions
- Median
  - Not affected by outliers
  - May better characterize skewed distributions

Measures of variation

- Range
- Variance and standard deviation
- Interquartile range

Range

- Range is the difference between the minimum and maximum values

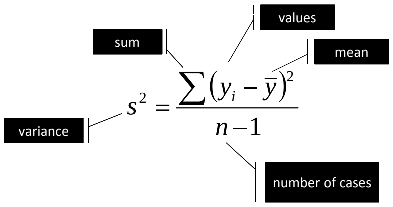
Calculating the range for temperatures

patient	temperature
Luka	**35.8**
Milo	36.2
Peter	36.4
Liam	36.6
Lily	36.6
Mohammed	36.8
Yasmine	36.9
Otis	37
Noah	37.2
John	**40.5**

range = 40.5 - 35.8 = 4.7

Variance and standard deviation

- The variance s^2 is the sum of the squared deviations from the mean divided by the number of cases minus 1



$$s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

sum

values

mean

variance

number of cases

Why squared? Why *n-1*?

- Why square differences between data values and mean?
  - Gives positive values
  - Gives more weight to larger differences
- Why *n - 1* for sample variance?
  - Dividing by *n* underestimates population variance
  - Dividing by *n-1* gives an ‘unbiased’ estimate of population variance

Variance and standard deviation

- The standard deviation *s* is the square root of the variance
- Easier to interpret because the unit of measurement remains the same
- Measure of absolute deviation

$$s = \sqrt{\frac{\sum\left(y_i - \bar{y}\right)^2}{n-1}}$$

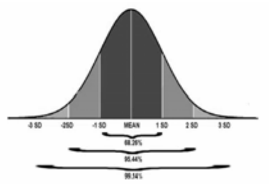
Calculating the variance and standard deviation for temperatures

patient	temperature	difference (value-mean)	squared difference
Liam	36.6	-0.4	0.16
Luka	35.8	-1.2	1.44
Noah	37.2	0.2	0.04
Mohammed	36.8	-0.2	0.04
Yasmine	36.9	-0.1	0.01
Otis	37	0	0
John	40.5	3.5	12.25
Peter	36.4	-0.6	0.36
Lily	36.6	-0.4	0.16
Milo	36.2	-0.8	0.64
SUM	**370**		**15.1**
n	10		
Mean	**37**		

$$s^2 = \frac{\sum\left(y_i - \bar{y}\right)^2}{n-1} = \frac{15.1}{10-1} = 1.68 \qquad s = \sqrt{\frac{\sum\left(y_i - \bar{y}\right)^2}{n-1}} = \sqrt{1.68} = 1.30$$

Interpretation of standard deviation

- If distribution of data approximately bell shaped, then
  - About 68% of the data fall within one standard deviation of the mean
  - About 95% of the data fall within two standard deviations of the mean
  - Nearly all of the data fall within three standard deviations of the mean (99%)



Interquartile range

- Difference between upper (third) and lower (first) quartiles
- Quartiles divide data into four equal groups
  - Lower (first) quartile is 25th percentile
  - Middle (second) quartile is 50th percentile and is the median
  - Upper (third) quartile is 75th percentile

Calculating the interquartile range for temperatures

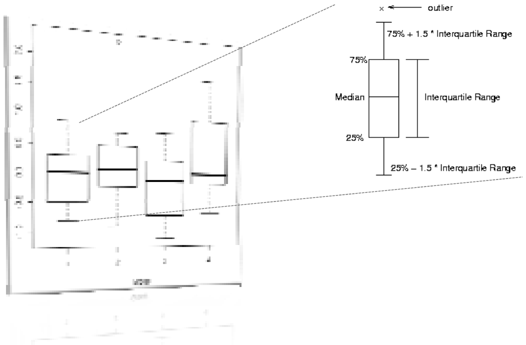
patient	temperature
Luka	35.8
Milo	36.2
Peter	*36.4*
Liam	36.6
Lily	**36.6**
Mohammed	**36.8**
Yasmine	36.9
Otis	*37*
Noah	37.2
John	40.5

interquartile range = 37-36.4 = 0.6

Interquartile range and outliers

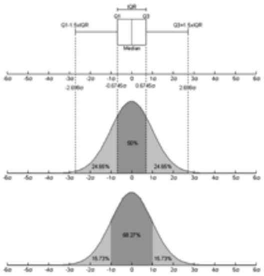
- Value can be considered to be an **outlier** if it falls more than 1.5 times the interquartile range above the upper quartile or more than 1.5 times the range below the lower quarter
- Example for temperatures
  - Interquartile range is .6
  - 1.5 times interquartile range is 0.9
  - Outliers would be values
    - Above UQ → 37 + 0.9 = 37.9 (**John**)
    - Below LQ → 36.4 – 0.9 = 35.5 (none)

Interquartile range and Boxplots


outlier
75% + 1.5 * Interquartile Range
75%
Median
Interquartile Range
25%
25% − 1.5 * Interquartile Range

Comparison of range, standard deviation, and interquartile range

- Sensitivity to extreme values
  - Range – extremely sensitive
  - Standard deviation – very sensitive
  - Interquartile range – not sensitive
- Standard deviation
  - Has desirable statistical properties (units!)
  - Suggests numbers of cases in different intervals for bell-shaped distributions

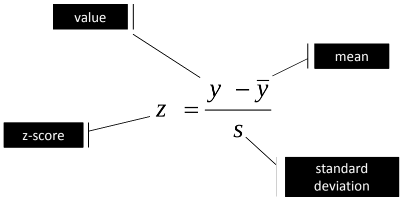


Standard scores

- *z*-score
- *t*-score and other deviation scores

z-score

- Expresses the distance between the value and the mean in number of standard deviations



$$z = \frac{y - \bar{y}}{s}$$

value

mean

z-score

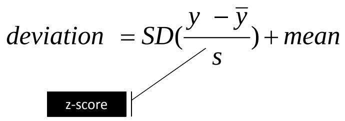
standard deviation

Z-scores for temperatures

patient	temperature	difference (value-mean)	*z*-scores
Liam	36.6	-0.4	-0.31
Luka	35.8	-1.2	-0.92
Noah	37.2	0.2	0.15
Mohammed	36.8	-0.2	-0.15
Yasmine	36.9	-0.1	-0.08
Otis	**37**	**0**	**0.00**
John	*40.5*	*3.5*	*2.69*
Peter	36.4	-0.6	-0.46
Lily	36.6	-0.4	-0.31
Milo	36.2	-0.8	-0.62
SUM	**370**		
n	10		
Mean	**37**		
SD	**1.3**		

t-scores and other deviation scores

- Analogous to *z*-scores, adapted for relevant distributions
- e.g. *t*-distribution: M= 50; SD= 10
- e.g. IQ: M=100; SD= 15



$$deviation = SD(\frac{y - \bar{y}}{s}) + mean$$

z-score

t-score

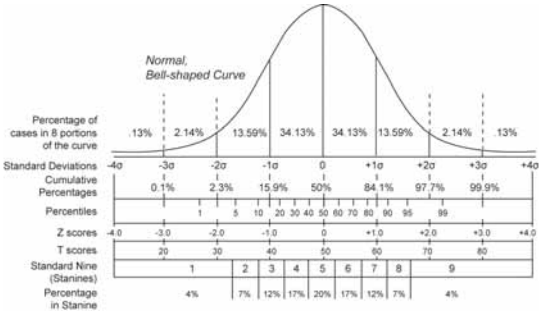
patient	temperature	*z*-score	*t*-scores
Liam	36.6	-0.31	46.92
Luka	35.8	-0.92	40.77
Noah	37.2	0.15	51.54
Mohammed	36.8	-0.15	48.46
Yasmine	36.9	-0.08	49.23
Otis	**37**	**0.00**	**50.00**
John	*40.5*	*2.69*	**76.92**
Peter	36.4	-0.46	45.38
Lily	36.6	-0.31	46.92
Milo	36.2	-0.62	43.85
SUM	**370**		
n	10		
Mean	**37**		
SD	**1.3**		

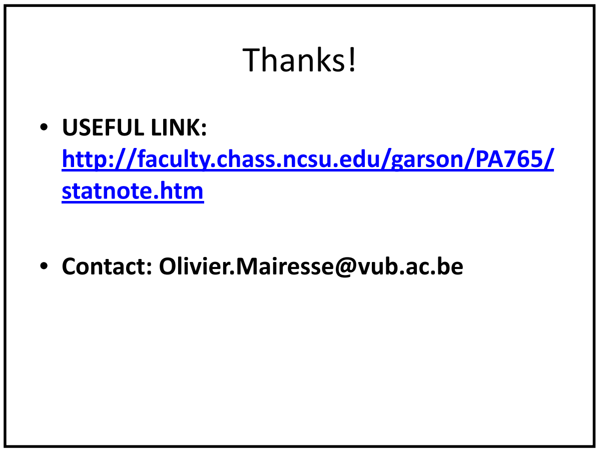
$$t = 10\frac{y - \bar{y}}{s} + 50$$

Standard scores

- Useful to compare values with different units (need for standardization)
- Useful to detect outliers
  - Generally, a value can be considered to be an **outlier** if it falls more than 2 standard deviations times above or below the mean, or in other words if the *z*-score is above or below 2
  - Example for temperatures
    - John: *z*-score = 2.69

Visual summary





Thanks!

- **USEFUL LINK:** **http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm**
- **Contact: Olivier.Mairesse@vub.ac.be**