Introductory Statistics

Søren Lophaven

Topics

- Descriptive statistics
- Montgomery and Åsberg Depression Rating Scale (MADRS)
- Types of data
- Graphical presentation of data
- Tabular presentation of data

Descriptive statistics

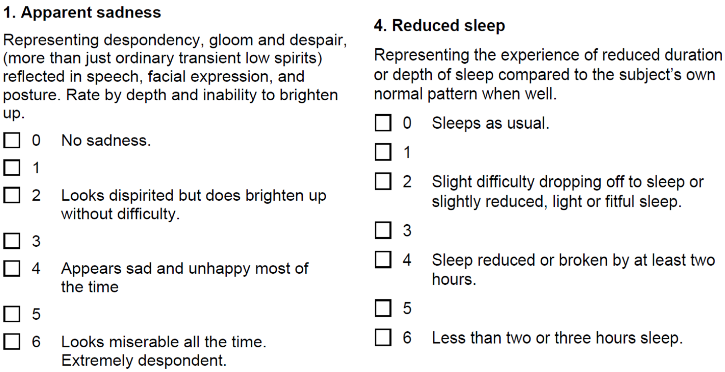
Often the primary statistical method used for:

- Phase I trials due to sparse data
- For demography data and other baseline characteristics
- For adverse event data in phase I, II and III trials
- For other types of safety data in phase I, II and III trials

Montgomery and Åsberg Depression Rating Scale (MADRS)

- MADRS: rating scale designed to assess the severity of depressive symptoms
- Based on a clinical interview
- Administered by trained psychiatrists
- MADRS total score is the sum of the score of the 10 individual items
- Symptoms rated on 7-point scales from 0 (no symptom) to 6 (severe symptom) with detailed anchor points
- MADRS total score goes from 0 to 60

Two MADRS items



1. Apparent sadness

Representing despondency, gloom and despair, (more than just ordinary transient low spirits) reflected in speech, facial expression, and posture. Rate by depth and inability to brighten up.

□ 0 No sadness.

□ 1

□ 2 Looks dispirited but does brighten up without difficulty.

□ 3

□ 4 Appears sad and unhappy most of the time

□ 5

□ 6 Looks miserable all the time. Extremely despondent.

4. Reduced sleep

Representing the experience of reduced duration or depth of sleep compared to the subject's own normal pattern when well.

□ 0 Sleeps as usual.

□ 1

□ 2 Slight difficulty dropping off to sleep or slightly reduced, light or fitful sleep.

□ 3

□ 4 Sleep reduced or broken by at least two hours.

□ 5

□ 6 Less than two or three hours sleep.

Types of data in clinical trials

- Demography and baseline characteristics
    - Age, gender, weight, height, BMI, medical history, smoking habits, FEV1 measurements, blood, pressure ...
- Safety data
    - Adverse events (seriousness, severity, causality, etc.), laboratory tests (blood and urine samples), vital signs (blood pressure, heart rate), ECG (QT, RR, PR, QRS interval etc.)
- Efficacy data
    - Depending on disease area

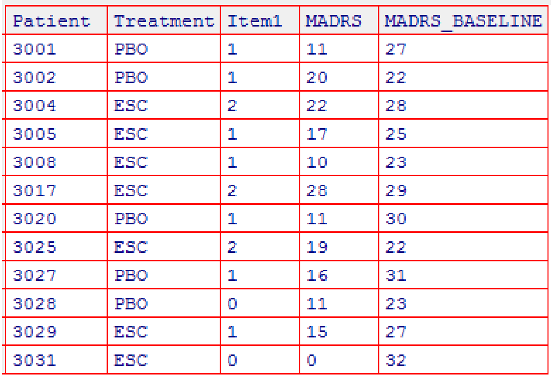
Types of data

- Continuous data
    - Body Mass Index, temperature, QTc interval, ...
- Ordered categorical data (ordinal)
    - MADRS total score, MADRS items, number of adverse events per patient, severity of adverse events (mild, moderate, severe), ...
- Unordered categorical data (nominal)
    - Political preference, sex, race, smoking history, ...

Types of data not always treated as you may think

- Continuous variables may be grouped
    - Remission (yes/no)
    - Response (yes/no or fast/partial/non-)
    - Age groups (21 - 30, 31 - 40, 41 - 50, ...)
    - QTcF>500 ms (yes/no)
- Ordered categorical variables may be presented/analysed as if they were continuous
    - MADRS total score and MADRS items
        - Generally more reasonable if many categories

MADRS dataset - Escitalopram versus placebo - first 12 observations



Patient	Treatment	Item1	MADRS	MADRS_BASELINE
3001	PBO	1	11	27
3002	PBO	1	20	22
3004	ESC	2	22	28
3005	ESC	1	17	25
3008	ESC	1	10	23
3017	ESC	2	28	29
3020	PBO	1	11	30
3025	ESC	2	19	22
3027	PBO	1	16	31
3028	PBO	0	11	23
3029	ESC	1	15	27
3031	ESC	0	0	32

Descriptive statistics - Mean

- Based on the first five observations from the MADRS dataset (MADRS total score at week 8 in the Escitalopram group): 22, 17, 10, 28 and 19
- The mean is a measure of the midpoint of the data distribution

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{1}{5} \cdot \left(22 + 17 + 10 + 28 + 19\right) = 19.2$$

Descriptive statistics - Median

- The median is a measure of the midpoint of the data distribution
- Order the observations by size:
- 10, 17, 19, 22, 28
- The median is then the observation in the middle: 19

- What if the dataset had been the first 6 observations of the MADRS total score at week 8 in the Escitalopram group?
- 22, 17, 10, 28, 19 and **15**

Descriptive statistics - Median

- Order the observations by size:
- 10, **15**, 17, 19, 22, 28
- The median is then the mean of the two middle observations: $\frac{1}{2} \cdot (17 + 19) = 18$

Mean versus median

- The mean and median are identical for a symmetrical distribution
- The median is more robust measure for the midpoint than the mean with respect to outliers
- For a skewed distribution the median is often a better measure of the midpoint
- The difference in means is directly related to a statistical test (t-test) while no test can be directly related to the difference in medians

Percentiles

- A percentile is a measure indicating the value below which a given percentage of observations in a dataset fall
- For example, the 20 percentile is the value below which 20 percent of the observations are found
- The median is the 50 percentile
- The 25 percentile is also called the lower or first quartile, the 75 percentile is called the upper or third quartile
- The 0 percentile s actually the minimum and the 100 percentile is the maximum
- If five observations are ordered the first is the minimum, the second is the 25 percentile, the third is the median etc.
- The range of the dataset if often presented as [min ; max]

Variance

- The variance is a measure of the variation is the dataset
- The variance measures the average squared 'distance' from the mean to the observations

$$
\begin{aligned}
s^2 &= \frac{1}{(n-1)} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \frac{1}{(5-1)} \Big((22-19.2)^2 + (17-19.2)^2 \\
&\quad + (10-19.2)^2 + (28-19.2)^2 + (19-19.2)^2 \Big) \\
&= 43.7
\end{aligned}
$$

Standard deviation (SD) and standard error (SE)

- The standard deviation measures the average 'distance' from the mean to the observations
- The standard error measures the variability of the mean

$$\begin{aligned} SD &= s = \sqrt{s^2} = \sqrt{43.7} = 6.61 \\ SE &= \frac{s}{\sqrt{n}} = \frac{6.61}{\sqrt{5}} = 2.96 \end{aligned}$$

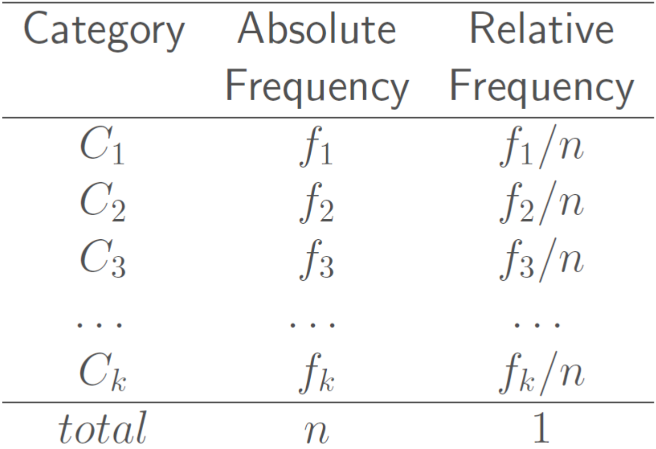
Descriptive statistics for the MADRS total score at week 8

	Placebo	Escitalopram
N	154	155
Mean	16.2	13.7
Variance	95.84	68.26
SD	9.79	8.26
SE	0.79	0.66
Minimum	0	0
25 percentile	9	8
Median	16	12
75 percentile	24	19.5
Maximum	45	36

First five observations versus full dataset for the Escitalopram group

	First five observations	Full dataset
Mean	19.2	13.7
Variance	43.7	68.26
SD	6.61	8.26
SE	2.96	0.66
Minimum	10	0
25 percentile	17	8
Median	19	12
75 percentile	22	19.5
Maximum	28	36

Descriptive statistics for categorical variables



Category	Absolute Frequency	Relative Frequency
C_1	f_1	f_1/n
C_2	f_2	f_2/n
C_3	f_3	f_3/n
$\dots$	$\dots$	$\dots$
C_k	f_k	f_k/n
total	n	1

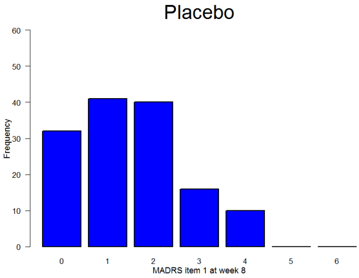
Descriptive statistics for MADRS item 1 at week 8

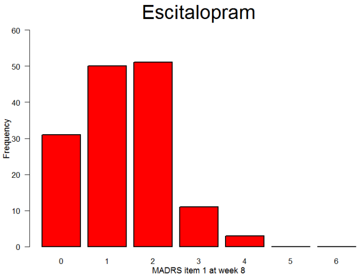
	Placebo		Escitalopram	
Item score	n	%	n	%
0	32	23.0	31	21.2
1	41	29.5	50	34.2
2	40	28.8	51	34.9
3	16	11.5	11	7.5
4	10	7.2	3	2.1
5	0	0	0	0
6	0	0	0	0

Graphical presentation of data

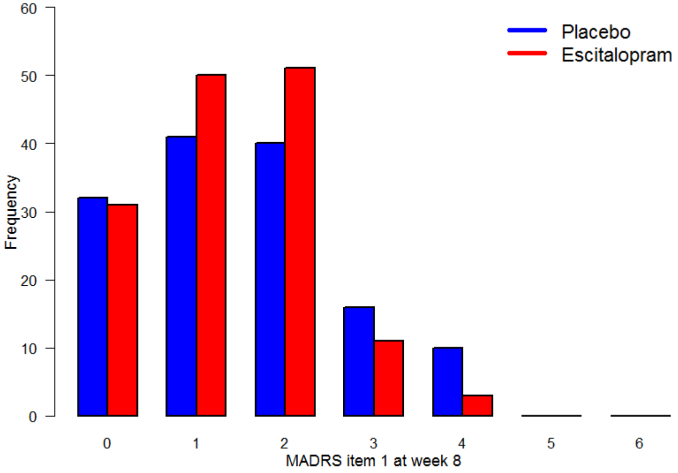
- Categorical variables
  - Bar chart
- Continuous variables
  - Histogram
  - Box plot
  - Cumulative distribution function
  - Scatterplot
- Data measured over time
  - Subject plot (spaghetti plot)
  - Mean plot

Bar chart - MADRS item 1 at week 8

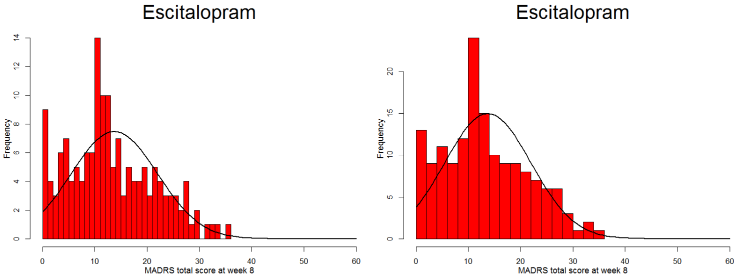




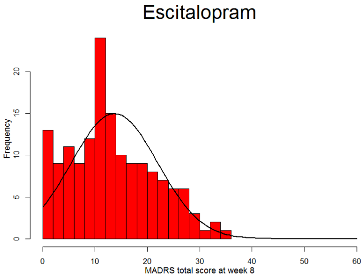
Bar chart - MADRS item 1 at week 8

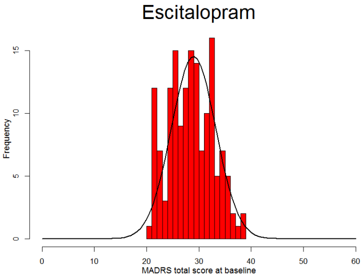


Histogram - MADRS total score at week 8

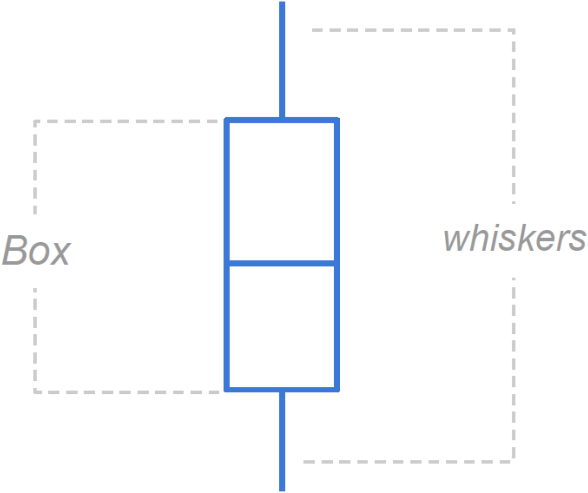


Histogram - MADRS total score week 8 versus baseline

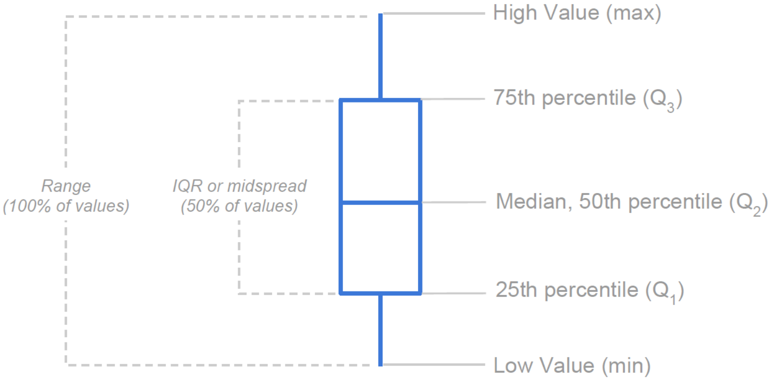




Boxplot Basics



Boxplot Basics

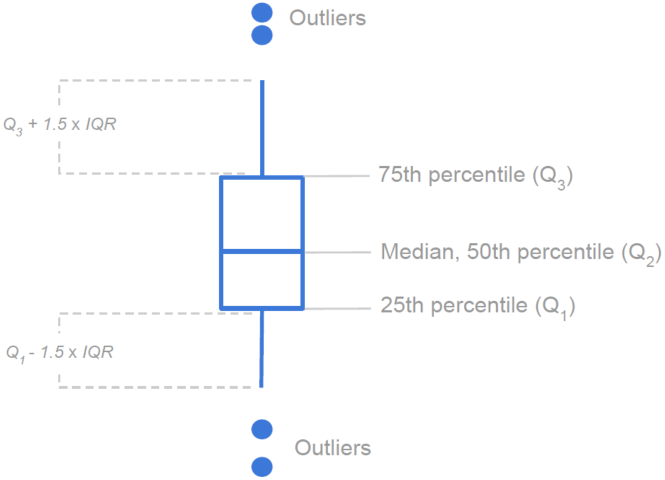


Boxplot with Outliers

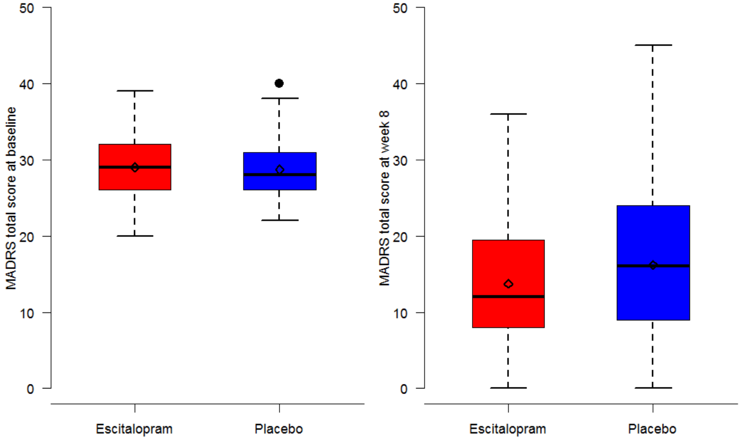
The 1.5 x IQR rule for outliers

- Call an observation a suspected outlier if it falls more than 1.5 x IQR above the third quartile or below the first quartile

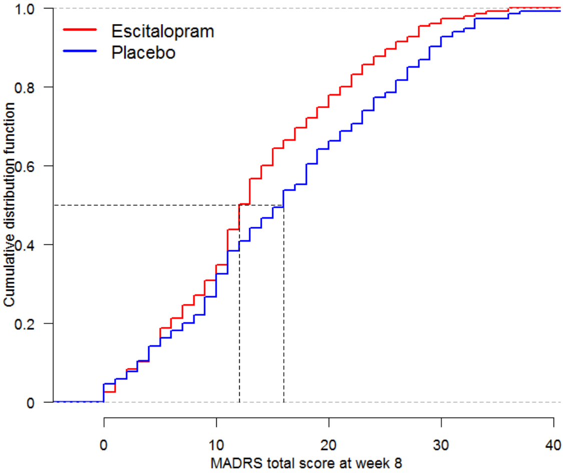
Boxplot with Outliers



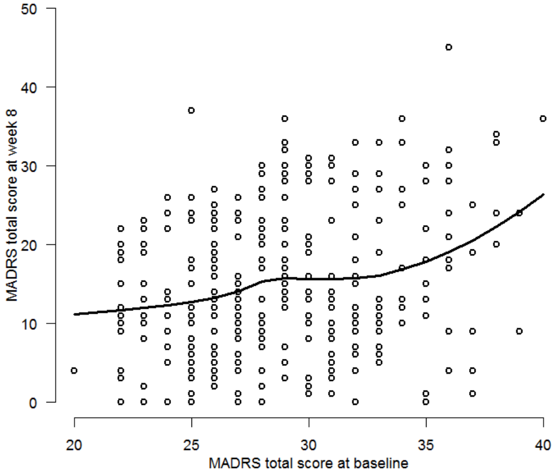
Box plot - MADRS total score at week 8 versus baseline



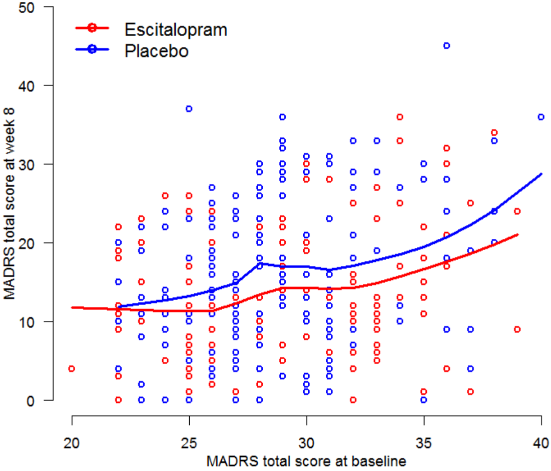
Cumulative distribution function - MADRS total score at week 8



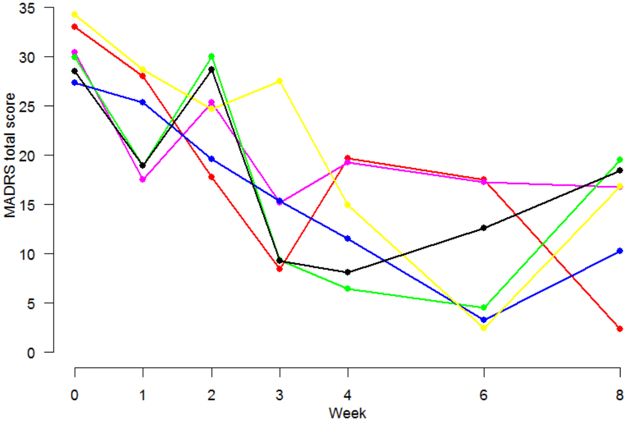
Scatterplot - MADRS total score at week 8 versus baseline



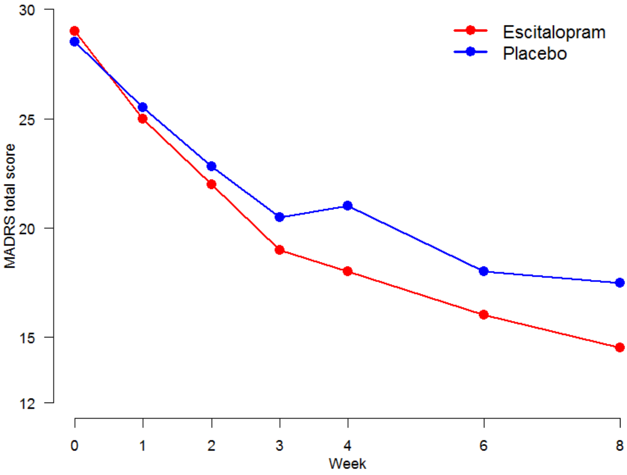
Scatterplot - MADRS total score at week 8 versus baseline



Subject plot (spaghetti plot) - MADRS total score



Mean plot - MADRS total score



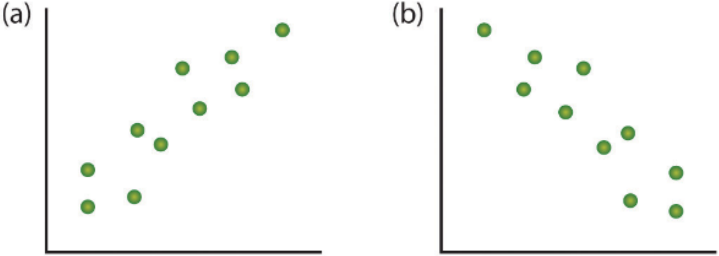
Association between variables

Pearson correlation coefficient:

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{(n-1)\sigma_x\sigma_y}$$

- A measure of linear association between two variables
- Not a measure of causality

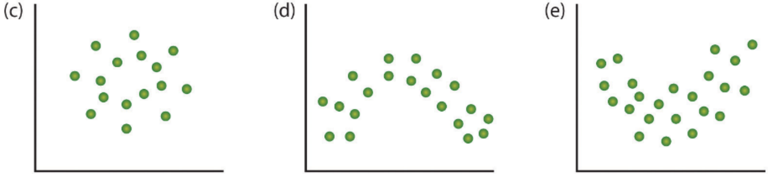
Correlation - some examples



Positive linear
$r = +.82$

Negative linear
$r = -.70$

Correlation - some examples



Independent
$r = 0.00$

Curvilinear
$r = 0.00$

Curvilinear
$r = 0.00$