

WEKA DATA MINING

Some statistics calculations, Creating box plot, scatter plot and a q-q plot

DATA PREPROCESSING PROBLEMS FOR LAB

- ▶ Suppose that the data for analysis includes the attribute *age*. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
- ▶ (a) What is the *mean* of the data? What is the *median*?
- ▶ (b) What is the *mode* of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).
- ▶ (c) What is the *midrange* of the data?
- ▶ (d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?
- ▶ (e) Give the *five-number summary* of the data.
- ▶ (f) Show a *boxplot* of the data.
- ▶ (g) How is a *quantile-quantile plot* different from a *quantile plot*?

DATA PREPROCESSING PROBLEMS FOR LAB

- (a) What is the *mean* of the data? What is the *median*?

Answer:

The (arithmetic) mean of the data is: $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i = 809/27 = 30$ (Equation 2.1).

The median (middle value of the ordered set, as the number of values in the set is odd) of the data is: 25.

- (b) What is the *mode* of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).

Answer:

This data set has two values that occur with the same highest frequency and is, therefore, bimodal.

The modes (values occurring with the greatest frequency) of the data are 25 and 35.

- (c) What is the *midrange* of the data?

Answer:

The midrange (average of the largest and smallest values in the data set) of the data is: $(70+13)/2 = 41.5$

DATA PREPROCESSING PROBLEMS FOR LAB

(d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

Answer:

The first quartile (corresponding to the 25th percentile) of the data is: 20. The third quartile (corresponding to the 75th percentile) of the data is: 35.

(e) Give the *five-number summary* of the data.

Answer:

The five number summary of a distribution consists of the minimum value, first quartile, median value, third quartile, and maximum value. It provides a good summary of the shape of the distribution and

for this data is: 13, 20, 25, 35, 70.

(f) Show a *boxplot* of the data.

Answer:

DATA PREPROCESSING PROBLEMS FOR LAB

(g) How is a *quantile-quantile plot* different from a *quantile plot*?

Answer:

A quantile plot is a graphical method used to show the approximate percentage of values below or equal to the independent variable in a univariate distribution. Thus, it displays quantile information for all the data, where the values measured for the independent variable are plotted against their corresponding quantile.

A quantile-quantile plot however, graphs the quantiles of one univariate distribution against the corresponding quantiles of another univariate distribution.

DATA PREPROCESSING PROBLEMS FOR LAB

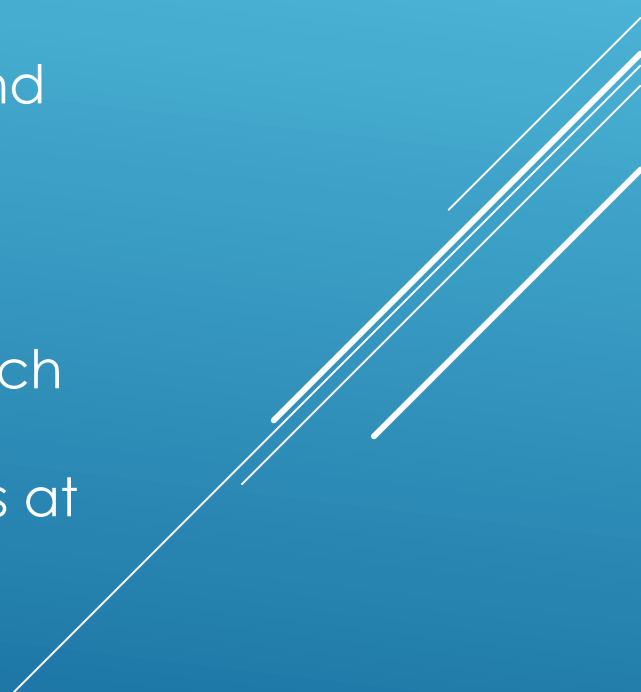
(g) How is a *quantile-quantile plot* different from a *quantile plot*?

Answer: continued from previous slide

Both axes display the range of values measured for their corresponding distribution, and points are plotted that correspond to the quantile values of the two distributions.

A line ($y = x$) can be added to the graph along with points representing where the first, second and third quantiles lie to increase the graph's informational value. Points that lie above such a line indicate a correspondingly higher value for the distribution plotted on the y-axis than for the distribution plotted on the x-axis at the same quantile.

The opposite effect is true for points lying below this line.



DATA PREPROCESSING FILTER OPTION EXAMPLE

Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following

result

age 23 23 27 27 39 41 47 49 50

%fat 9.5 26.5 7.8 17.8 31.4 25.9 27.4 27.2 31.2

age 52 54 54 56 57 58 58 60 61

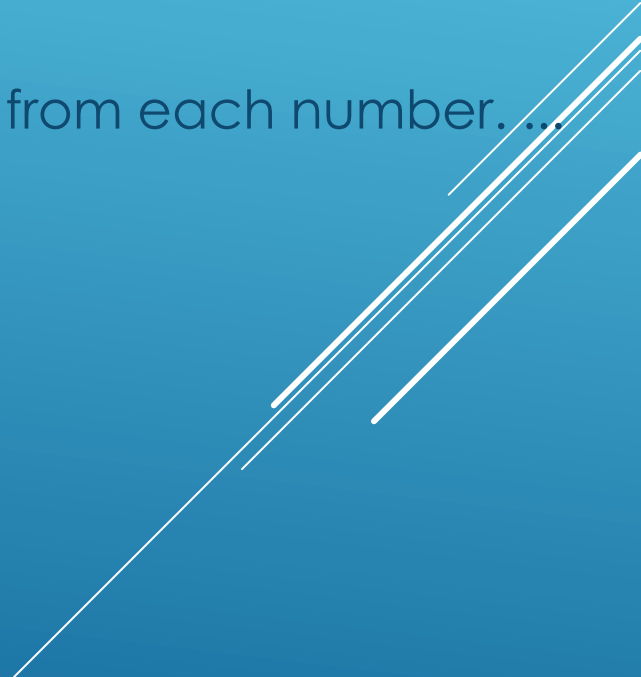
%fat 34.6 42.5 28.8 33.4 30.2 34.1 32.9 41.2 35.7

(a) Calculate the mean, median and standard deviation of *age* and *%fat*.

(b) Draw the boxplots for *age* and *%fat*.

(c) Draw a *scatter plot* and a *q-q plot* based on these two variables.

DATA PREPROCESSING FILTER OPTION EXAMPLE

- ▶ **Calculate** the mean or average of each data set. ...
 - ▶ Subtract the deviance of each piece of data by subtracting the mean from each number. ...
 - ▶ Square each of the **deviations**.
 - ▶ Add up all of the squared **deviations**.
 - ▶ Divide this value by the number of items in the data set.
- 
- Several white lines of varying lengths and slopes are positioned in the bottom right corner of the slide, creating a modern, abstract design element.

DATA PREPROCESSING FILTER OPTION EXAMPLE

Ans :

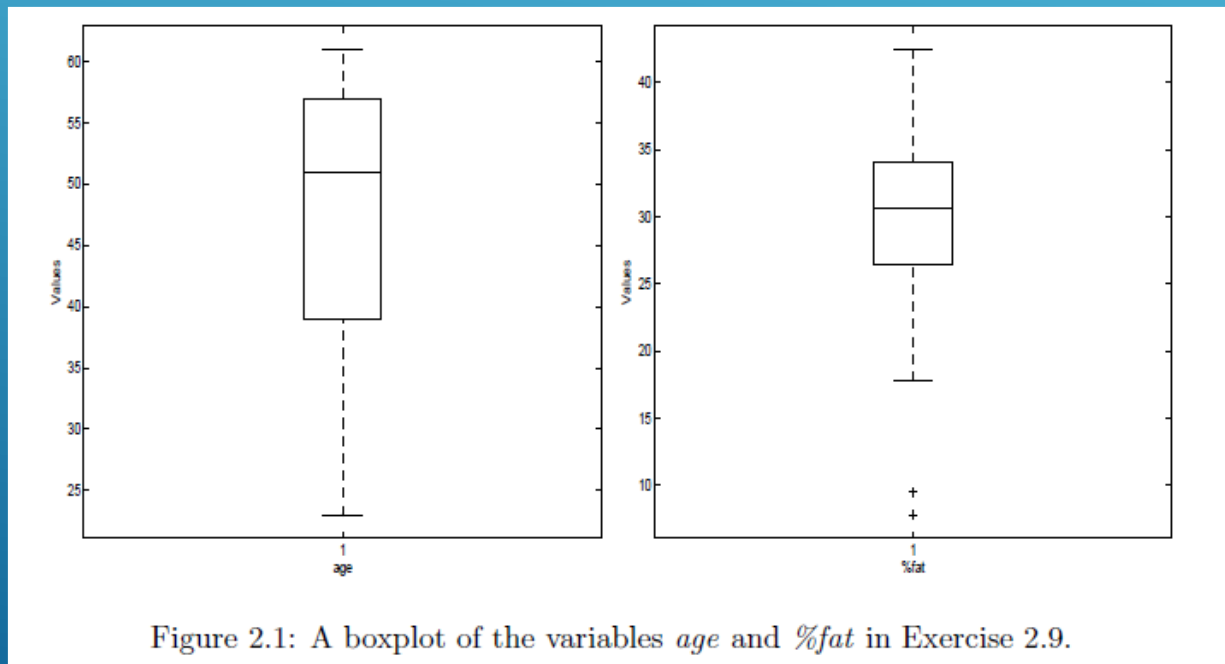
(a) Calculate the mean, median and standard deviation of *age* and *%fat*.

For the variable *age* the mean is 46:44, the median is 51, and the standard deviation is 12:85. For the

variable *%fat* the mean is 28:78, the median is 30:7, and the standard deviation is 8:99.

DATA PREPROCESSING FILTER OPTION EXAMPLE

(b) Draw the boxplots for *age* and *%fat* independently as in the previous example.



DATA PREPROCESSING FILTER OPTION EXAMPLE

(c) Draw a *scatter plot* and a *q-q plot* based on these two variables.

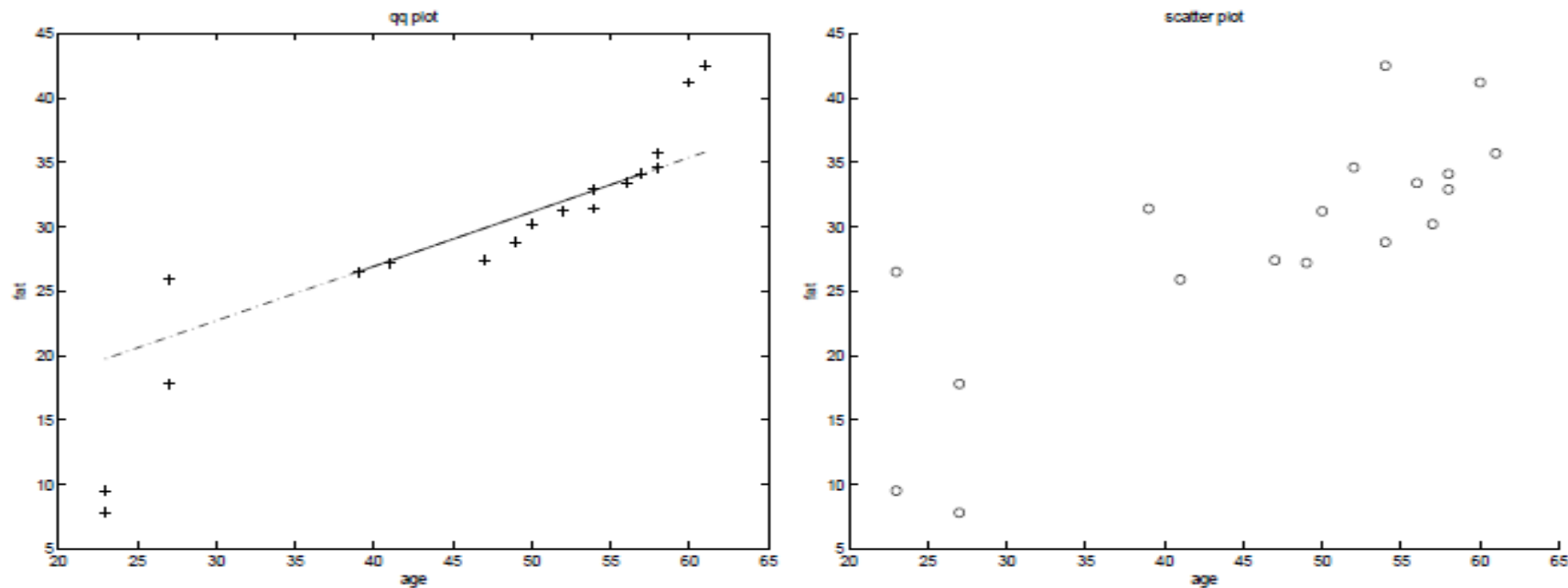


Figure 2.2: A *q-q plot* and a *scatter plot* of the variables *age* and *%fat* in Exercise 2.9.