# Osamelci

http://goo.gl/forms/VjRKEX0eEY

# Osamelci

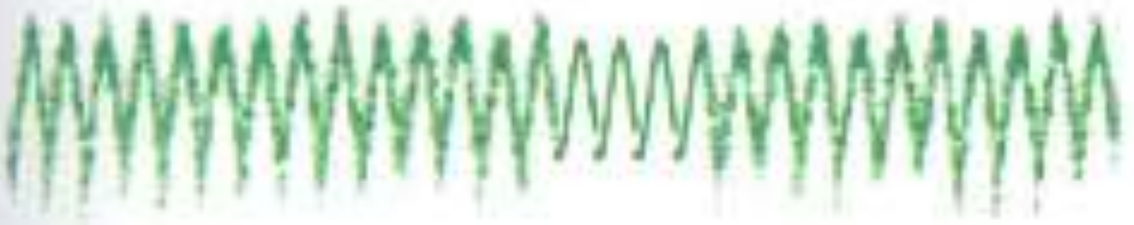
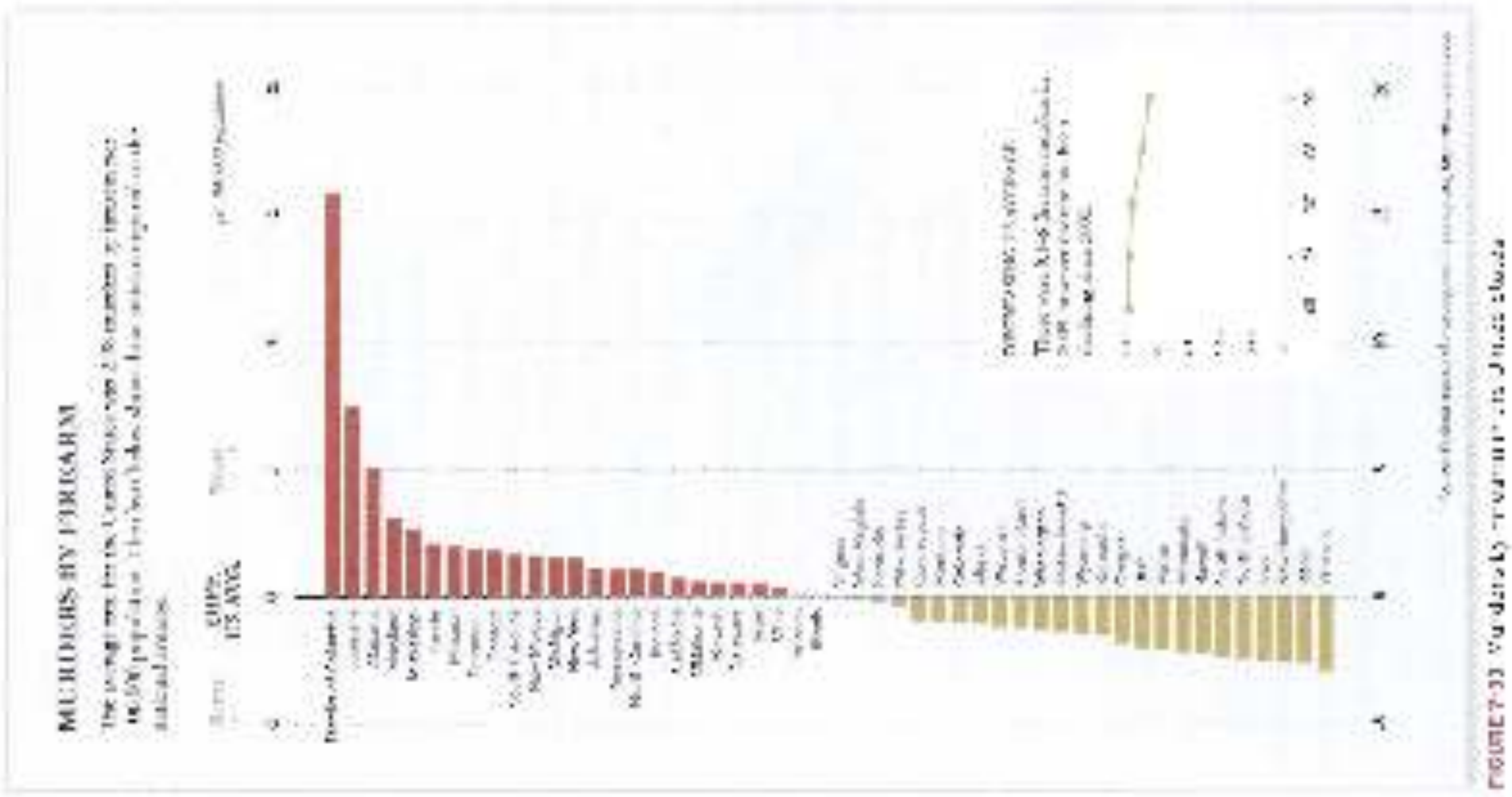


FIGURE 7-32 Estimated weather data from Weather Underground

[Yau: Visualize this.]

# Iskanje osamelcev



[Yau: Visualize this.]

# Iskanje osamelcev - primer

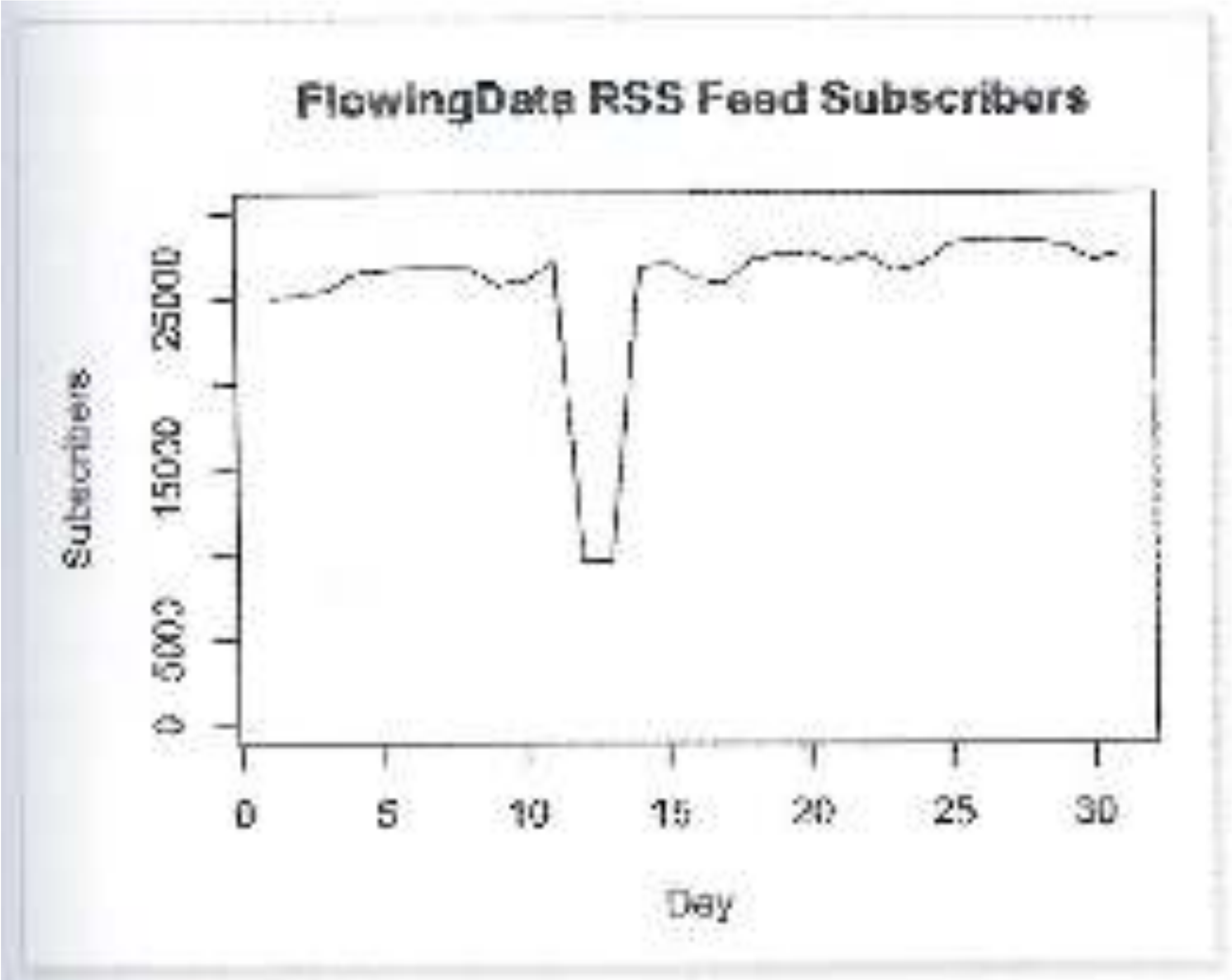
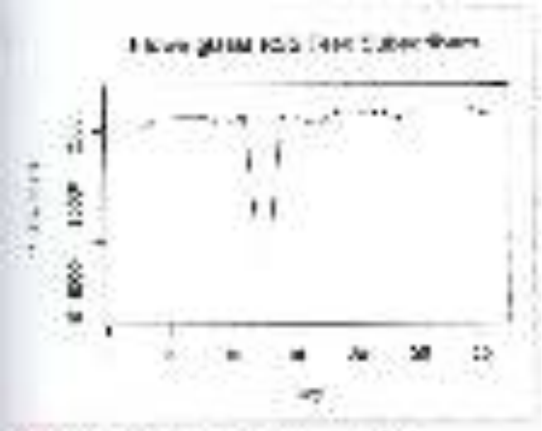


FIGURE 7-34 FlowingData subscriber counts over time

[Yau: Visualize this.]

# Iskanje osamelcev - primer



[illegible]

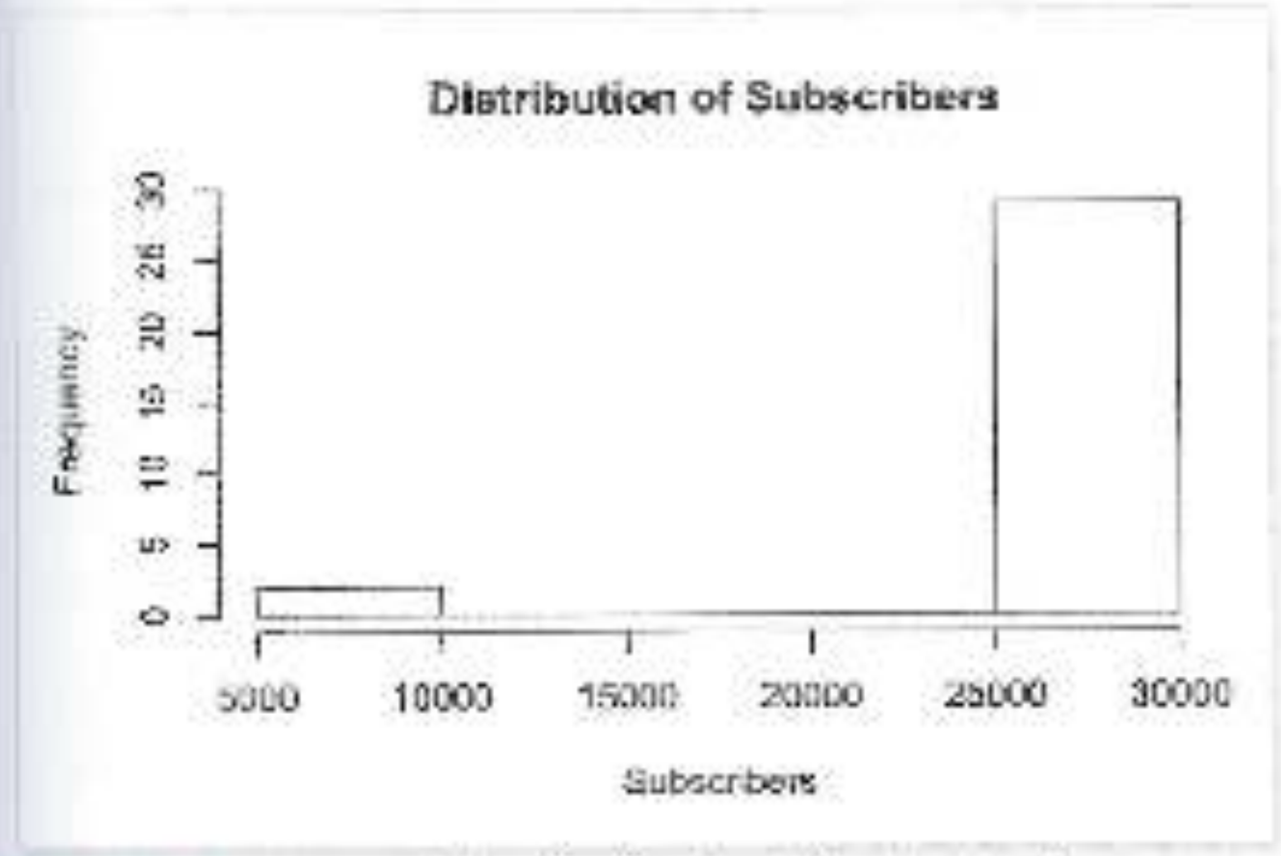
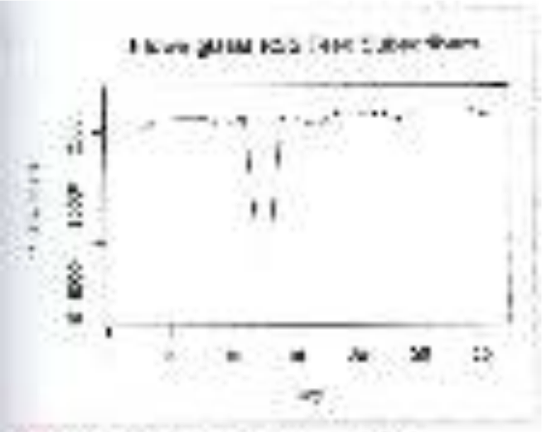


FIGURE 7-35 Histogram showing distribution of subscriber counts

[Yau: Visualize this.]

# Iskanje osamelcev - primer



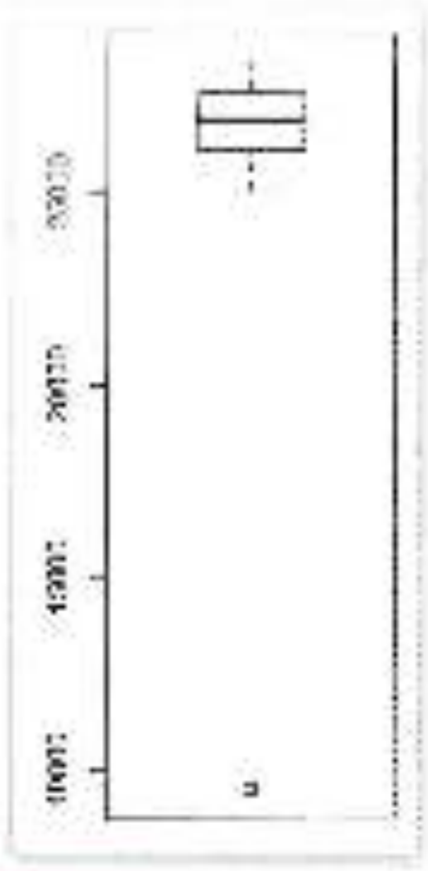
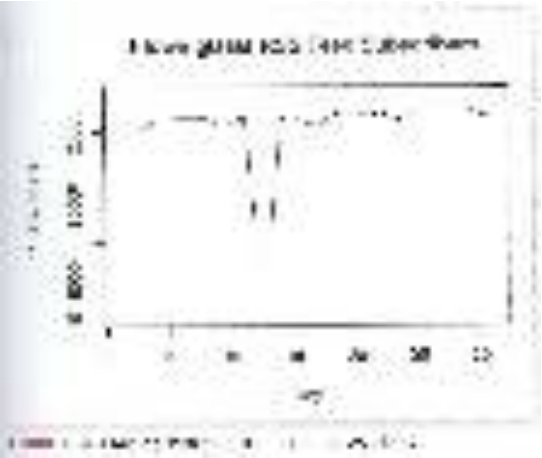
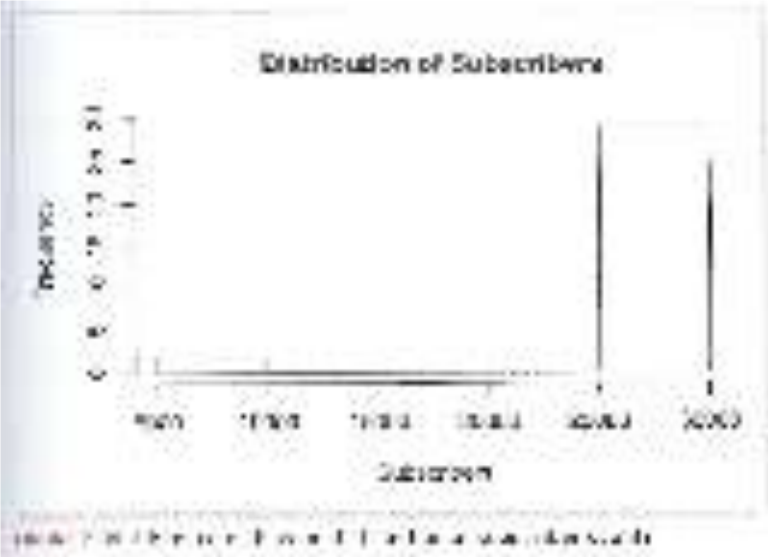


FIGURE 7-36 Boxplot showing distribution of subscriber counts

[Yau: Visualize this.]

# Iskanje osamelcev - primer





[Yau: Visualize this.]

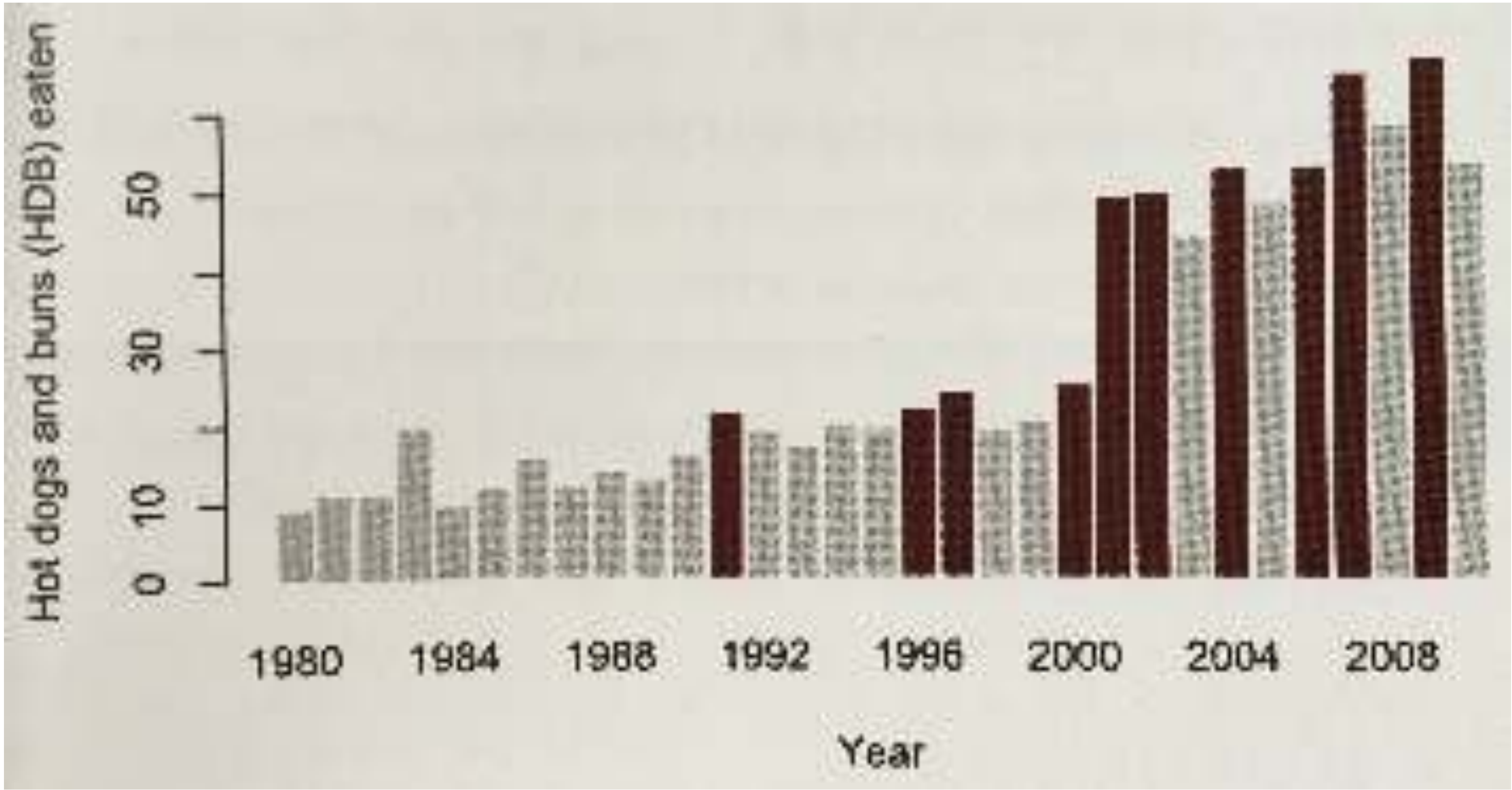
# Iskanje osamelcev – primer II

Tekmovanje (10-12 min) “Nathan’s Hot Dog Eating Contest” v Coney Island, New York, ob prazniku ”Independence day”, vsako leto od 1972.

[Yau: Visualize this.]

# Iskanje osamelcev – primer II

Tekmovanje (10-12 min) "Nathan's Hot Dog Eating Contest" v Coney Island, New York, ob prazniku "Independence day", vsako leto od 1972.



[Yau: Visualize this.]

# Iskanje osamelcev – primer II

Tekmovanje (10-12 min) “Nathan’s Hot Dog Eating Contest” v Coney Island, New York, ob prazniku ”Independence day”, vsako leto od 1972.

Leta 2001 prvič tekmuje Takeru Kobayashi (l.r. 1978), ki se zelo resno loti priprav in uspe podvojiti rekord Iz 25.25 na 50 hot dog-ov.

[Yau: Visualize this.]

# Iskanje osamelcev – primer III

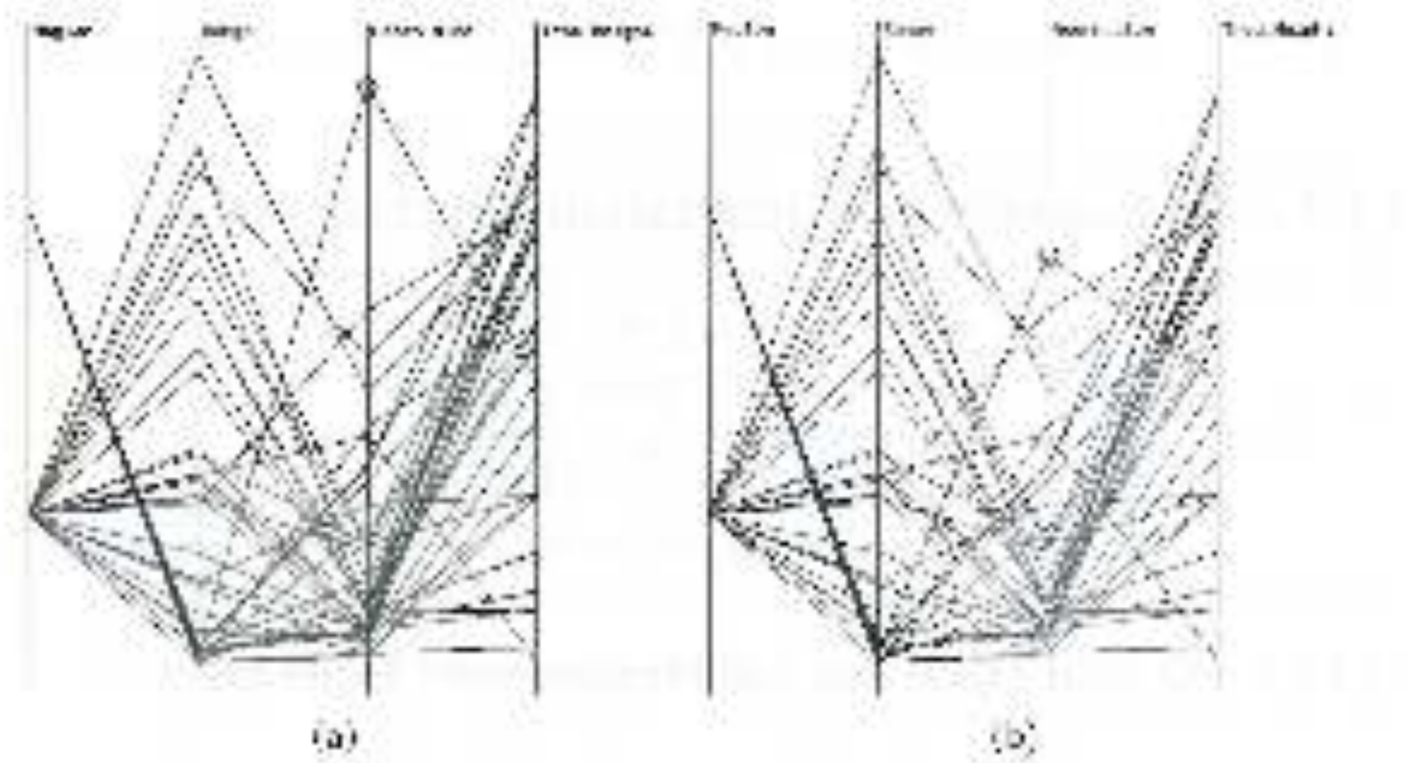


Figure 13.1 Parallel coordinates view of data set describing acorn attributes with a single outlier (circled, in the acorn size dimension) (a) in its original position and (b) with the distance artificially shortened [378]. (Image © 1997 IEEE.)

[Ward et al.: Interactive Data Visualization.]

# Iskanje osamelcev – primer III

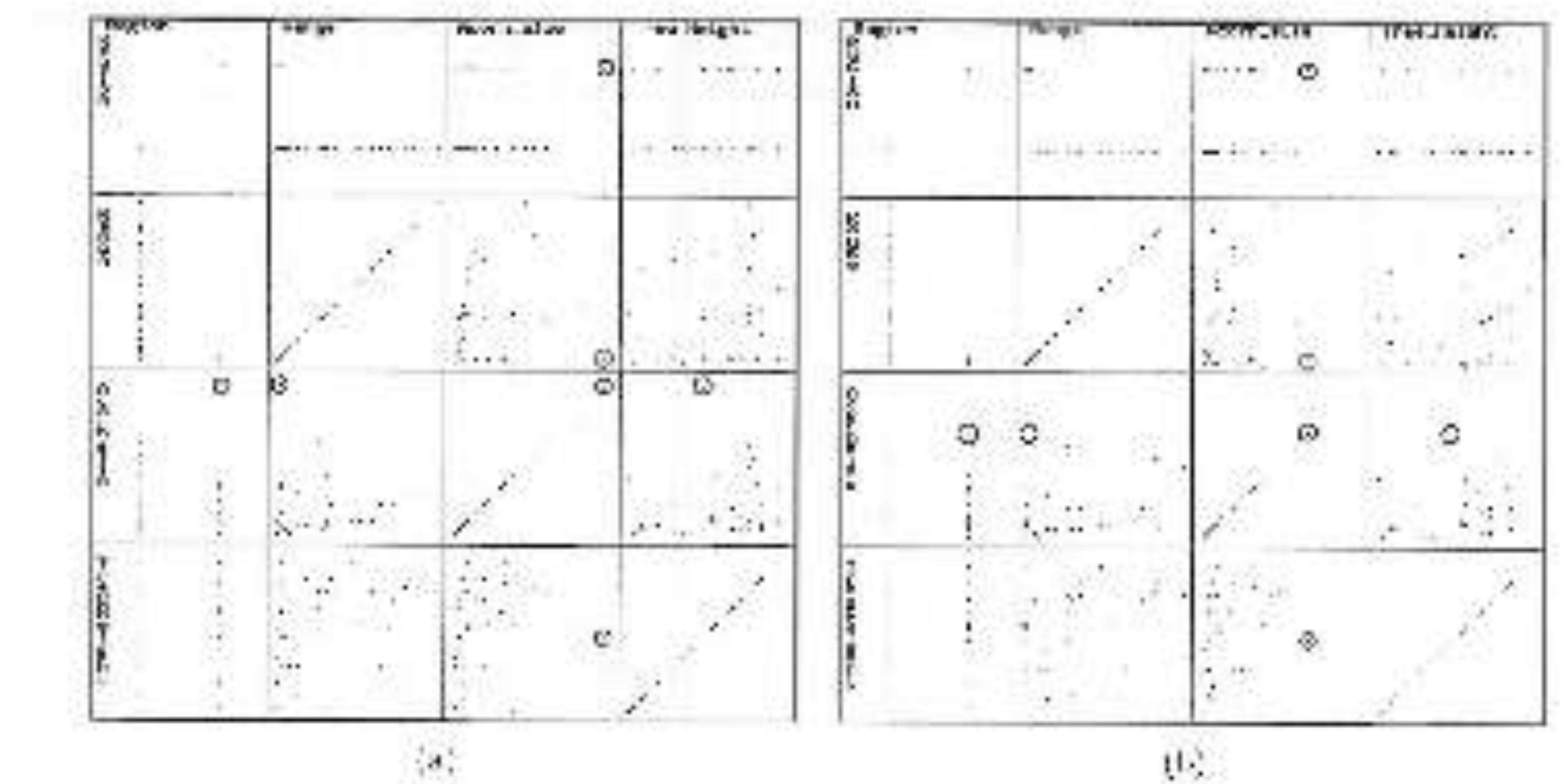
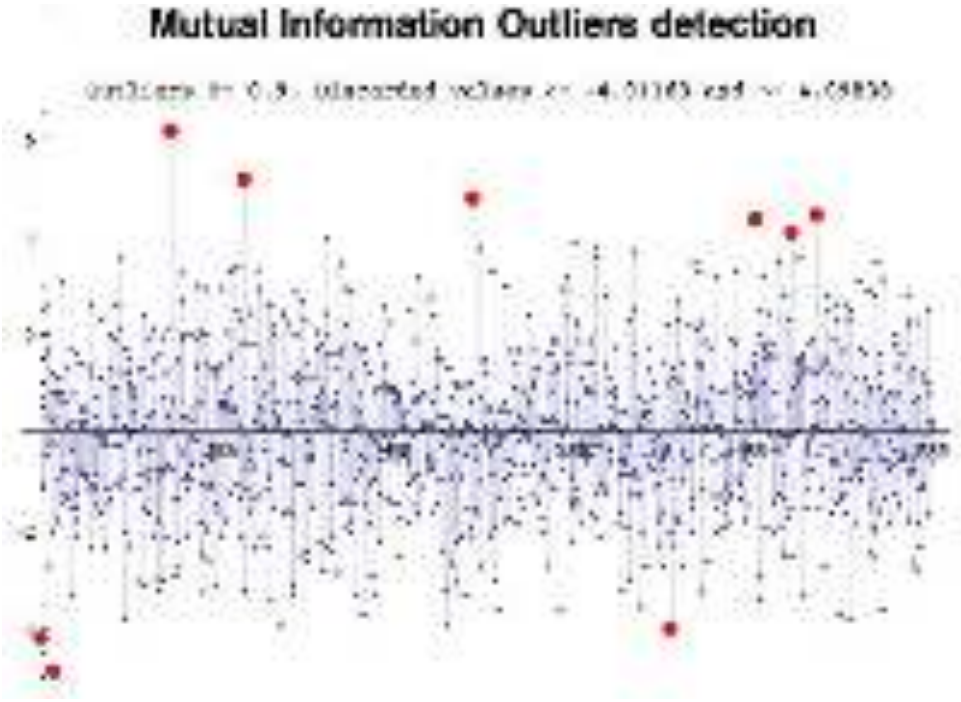


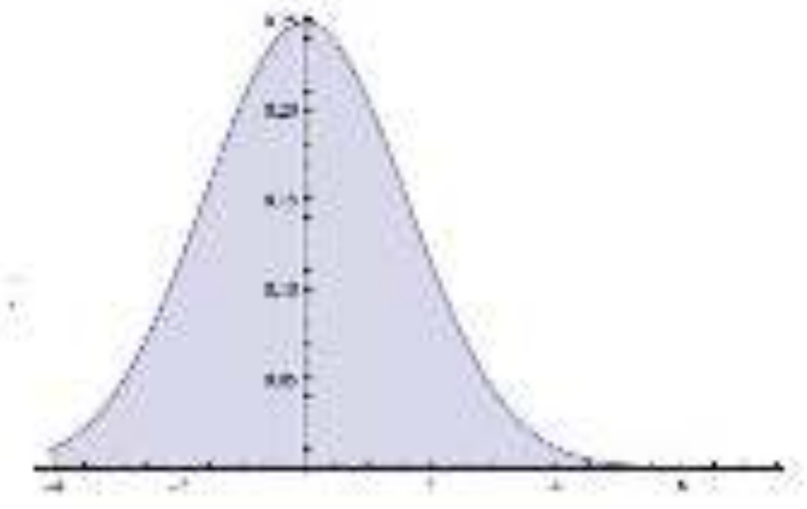
Figure 13.2. Identifying outliers with scatterplot matrices: same data as previous figure, using scatterplot matrices [875]. (Image © 1987 IEEE.)

[Ward et al.: Interactive Data Visualization.]

# Iskanje osamelcev – različne distribucije



MI

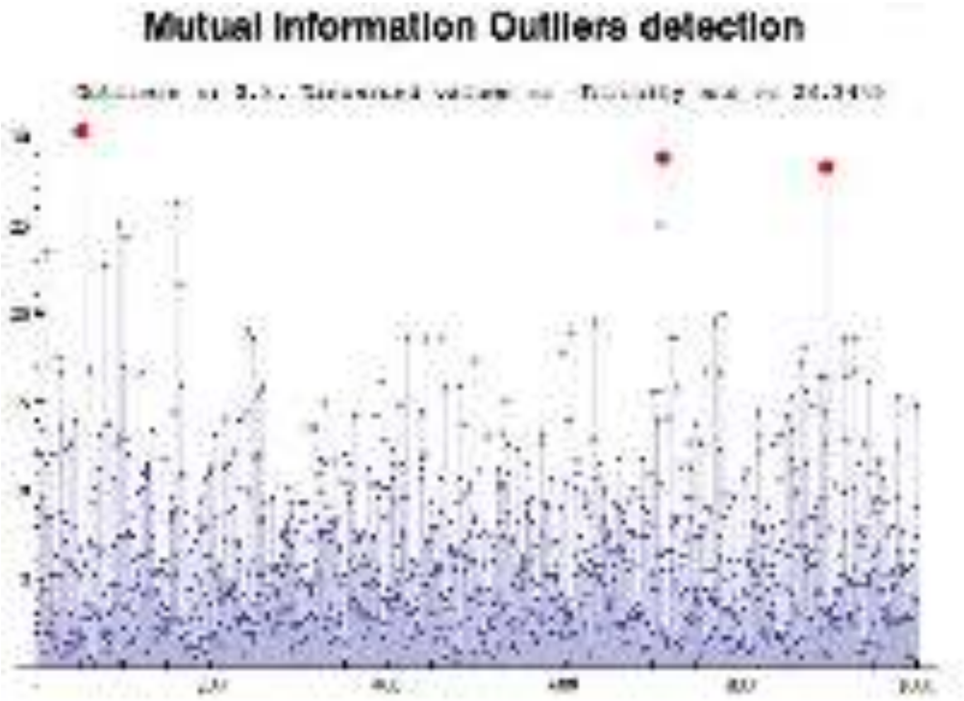


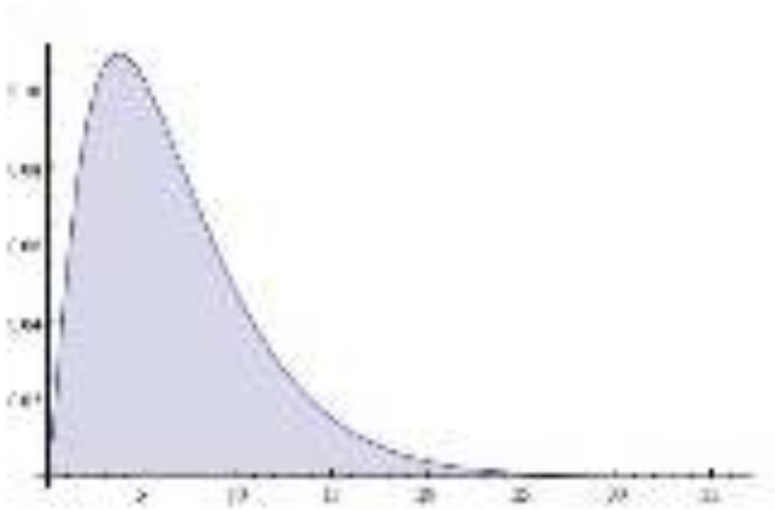
MI

[http://www.analyticbridge.com]

# Iskanje osamelcev – različne distribucije

MI

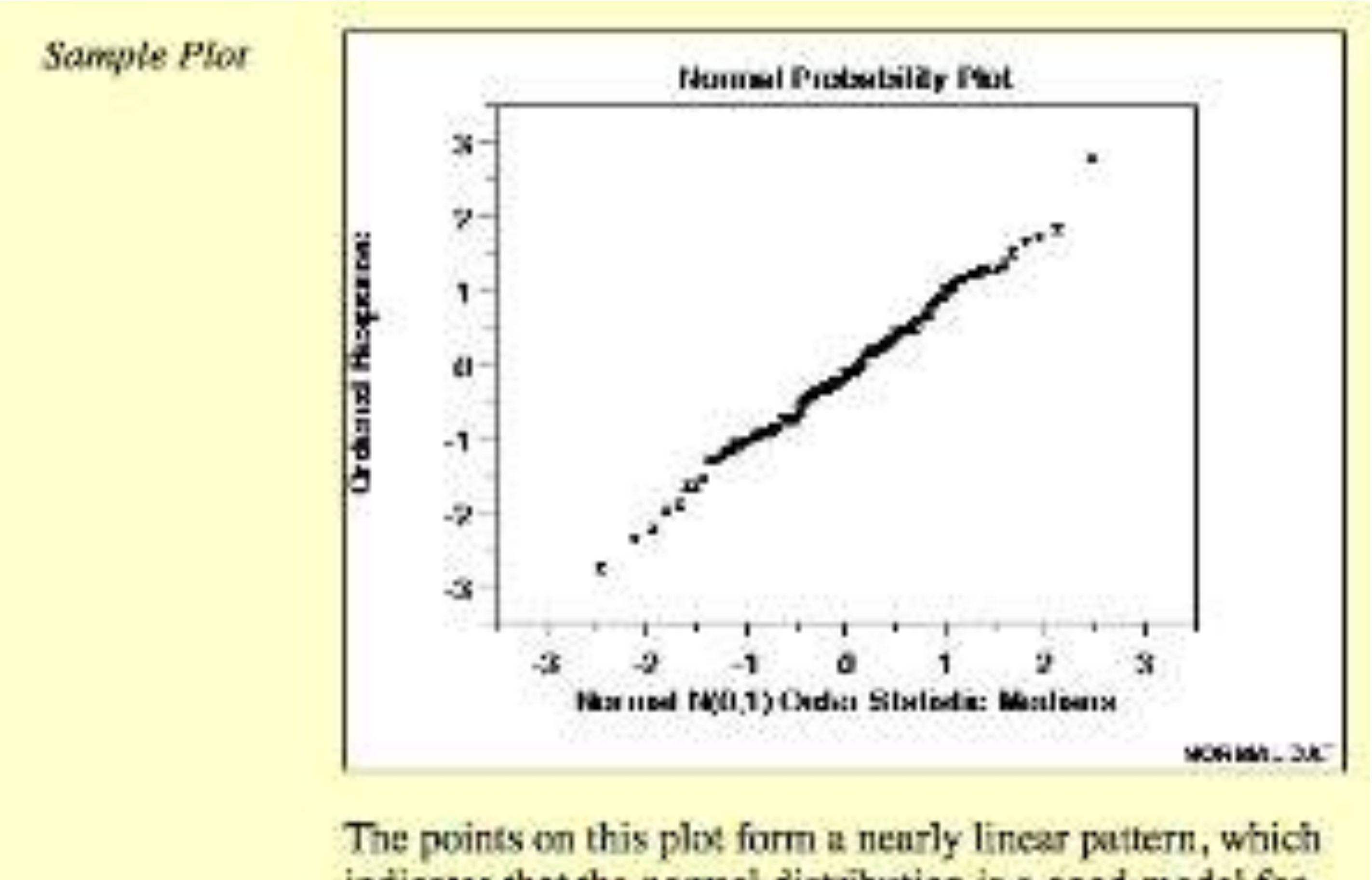




MI

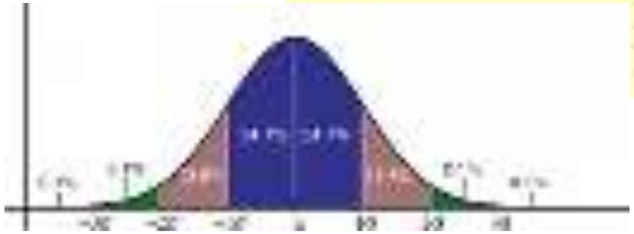
[http://www.analyticbridge.com]

# Iskanje osamelcev: Q-Q plot



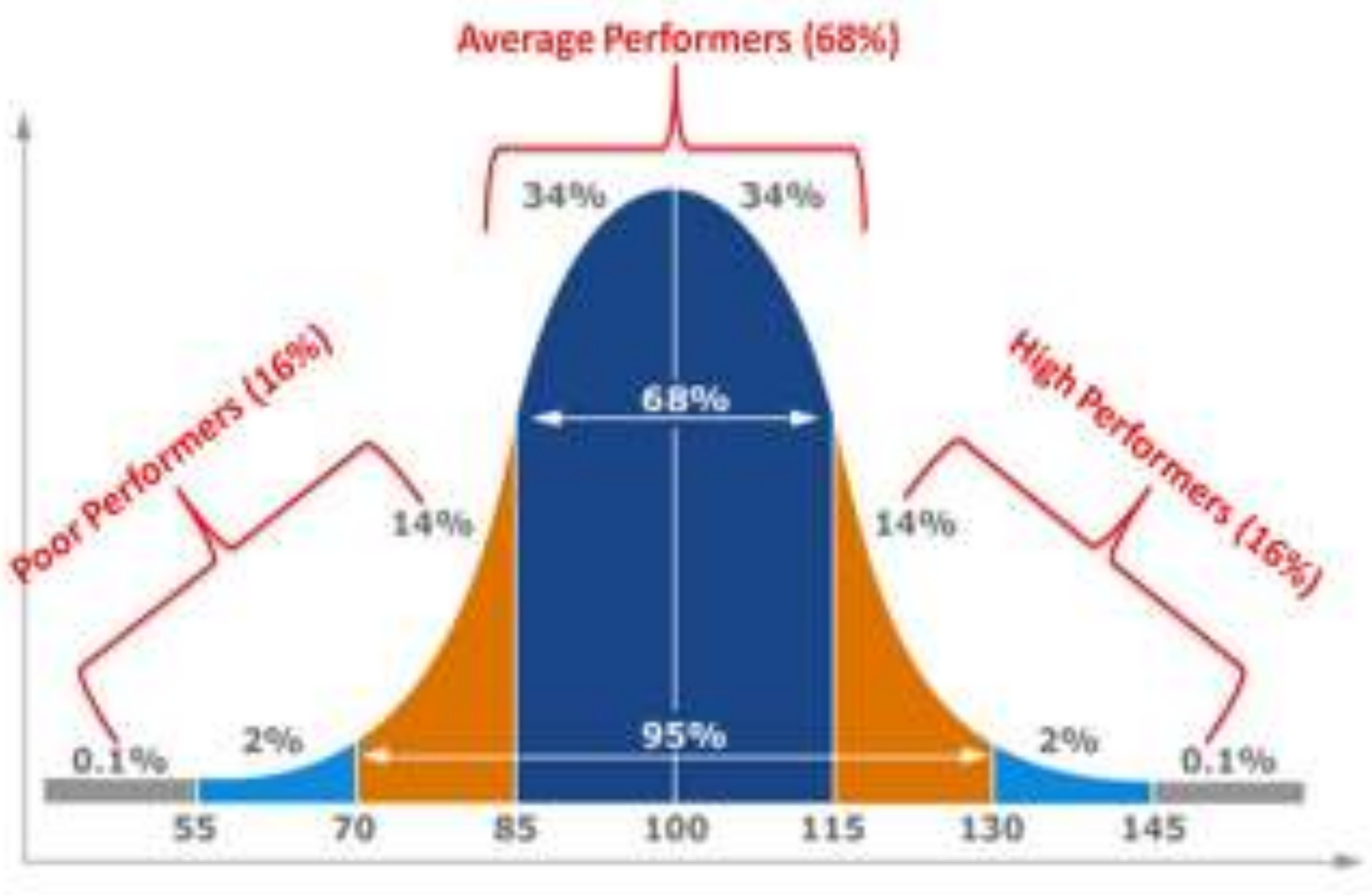
*Sample Plot*

The points on this plot form a nearly linear pattern, which indicates that the normal distribution is a good model for this data set.



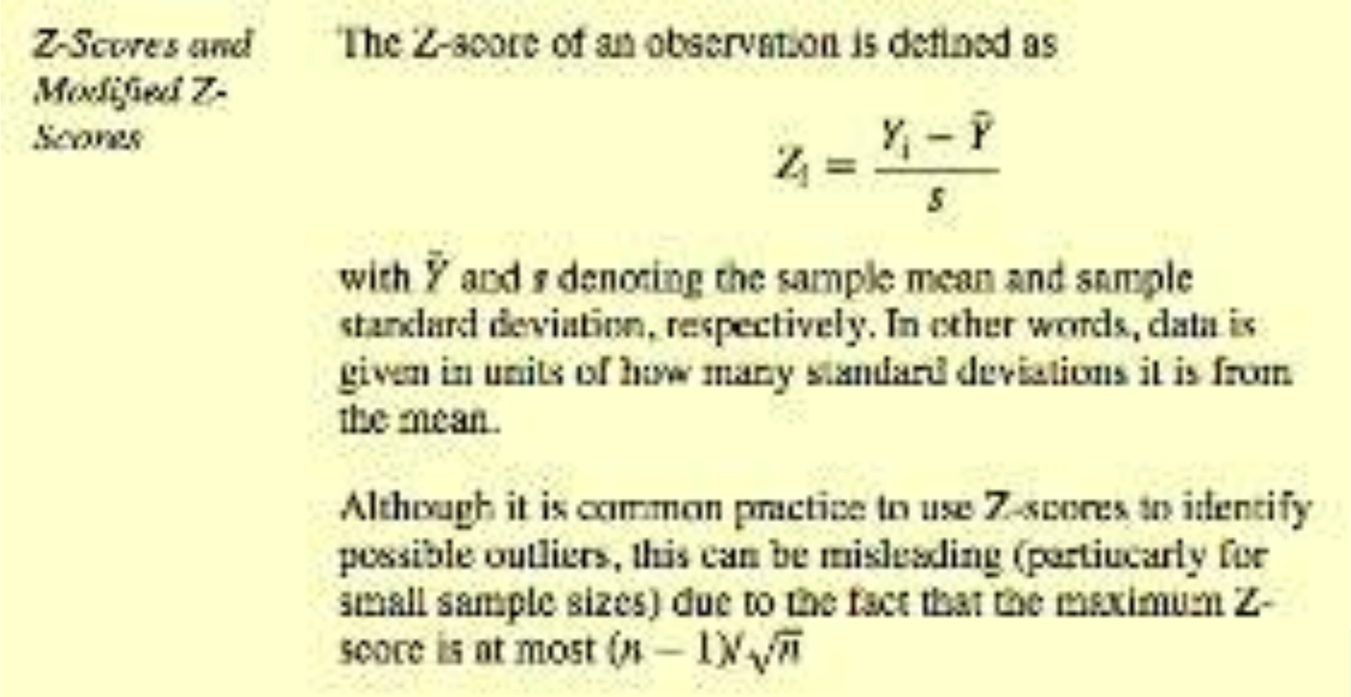
[http://www.itl.nist.gov/div898/handbook/eda/section3/normprpl.htm]

# Iskanje osamelcev: Z-score



[http://www.compensationcafe.com/2014/04/ding-dong-the-wicked-bell-curve-is-dead.html]

# Iskanje osamelcev: Z-score



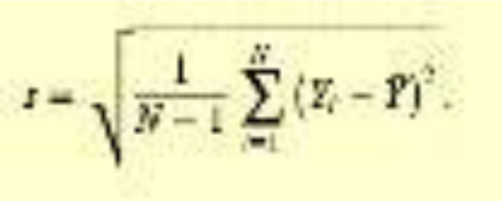
*Z-Scores and Modified Z-Scores*

The Z-score of an observation is defined as

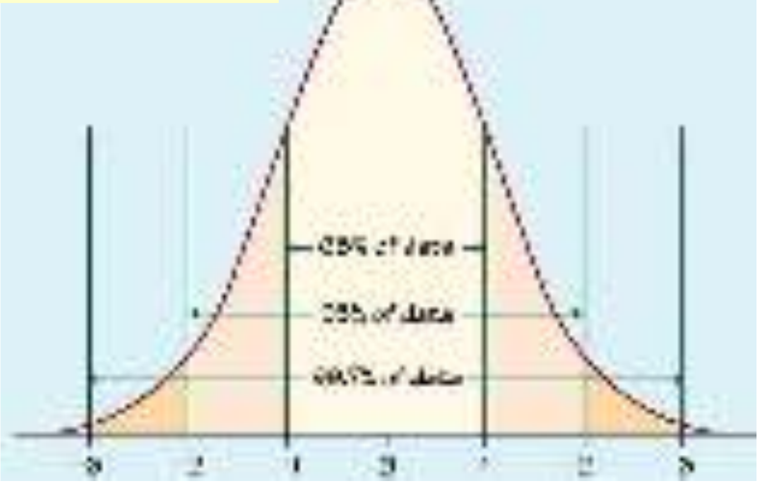
$$Z_i = \frac{Y_i - \bar{Y}}{s}$$

with $\bar{Y}$ and $s$ denoting the sample mean and sample standard deviation, respectively. In other words, data is given in units of how many standard deviations it is from the mean.

Although it is common practice to use Z-scores to identify possible outliers, this can be misleading (partiucarly for small sample sizes) due to the fact that the maximum Z-score is at most $(n-1)/\sqrt{n}$



$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2}.$$

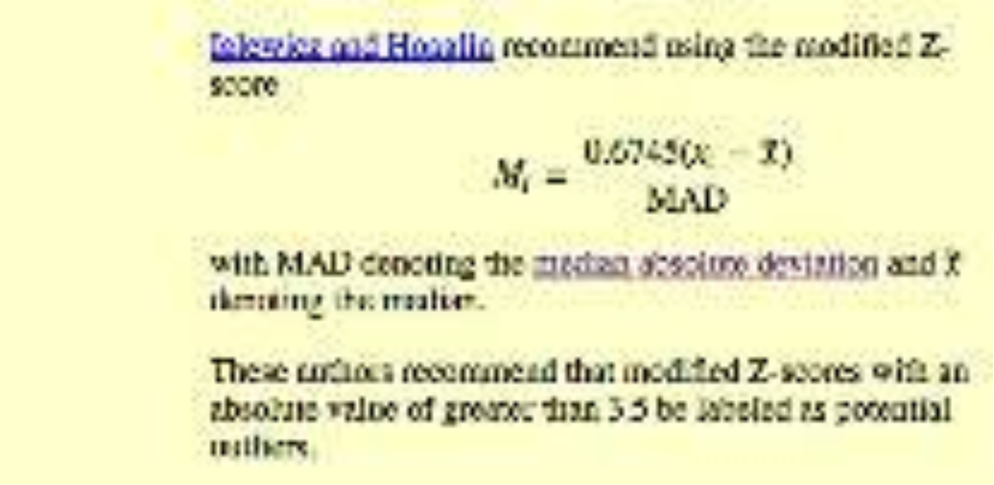


probability

Z-score [s]

[http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm]

# Iskanje osamelcev: Z-score

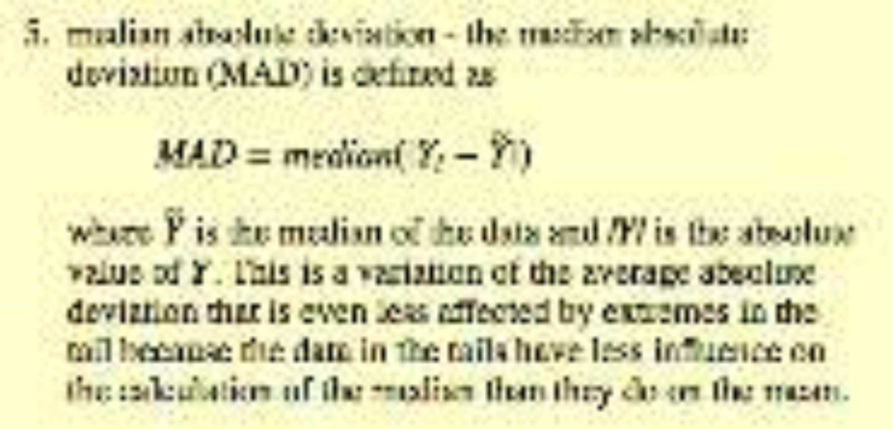


Iglewicz and Hoaglin recommend using the modified Z-score

$$M_i = \frac{0.6745(x_i - \tilde{x})}{\text{MAD}}$$

with MAD denoting the median absolute deviation and $\tilde{x}$ denoting the median.

These authors recommend that modified Z-scores with an absolute value of greater than 3.5 be labeled as potential outliers.



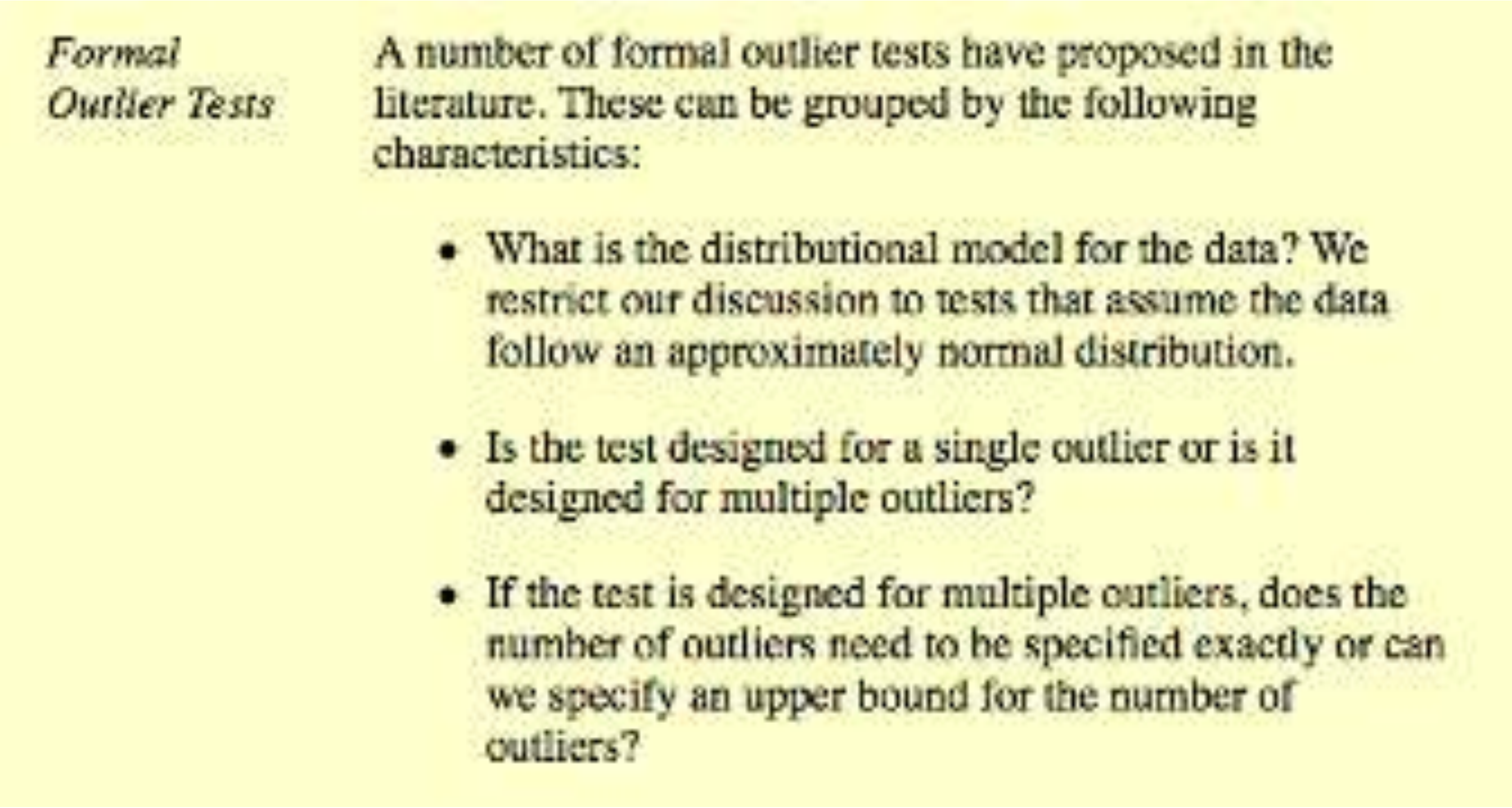
5. median absolute deviation - the median absolute deviation (MAD) is defined as

$$MAD = median(|Y_i - \tilde{Y}|)$$

where $\tilde{Y}$ is the median of the data and $|Y|$ is the absolute value of $Y$. This is a variation of the average absolute deviation that is even less affected by extremes in the tail because the data in the tails have less influence on the calculation of the median than they do on the mean.

http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm

# Iskanje osamelcev: Z-score
# Koliko osamelcev?



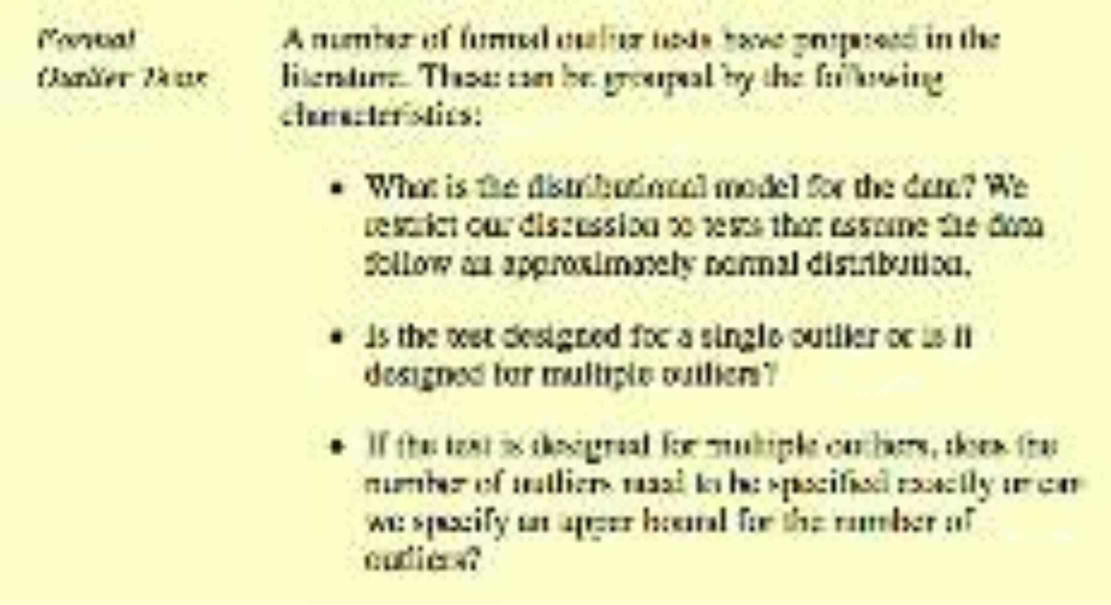
*Formal Outlier Tests*

A number of formal outlier tests have proposed in the literature. These can be grouped by the following characteristics:

- What is the distributional model for the data? We restrict our discussion to tests that assume the data follow an approximately normal distribution.
- Is the test designed for a single outlier or is it designed for multiple outliers?
- If the test is designed for multiple outliers, does the number of outliers need to be specified exactly or can we specify an upper bound for the number of outliers?

http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm

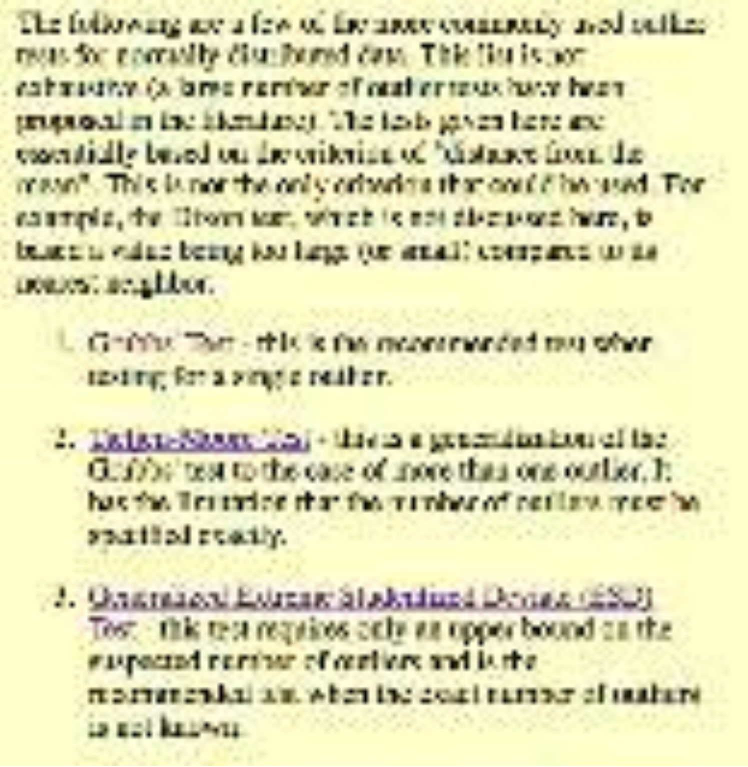
# Iskanje osamelcev: Z-score
# Koliko osamelcev?



*Formal Outlier Tests*

A number of formal outlier tests have proposed in the literature. These can be grouped by the following characteristics:

- What is the distributional model for the data? We restrict our discussion to tests that assume the data follow an approximately normal distribution.
- Is the test designed for a single outlier or is it designed for multiple outliers?
- If the test is designed for multiple outliers, does the number of outliers need to be specified exactly or can we specify an upper bound for the number of outliers?

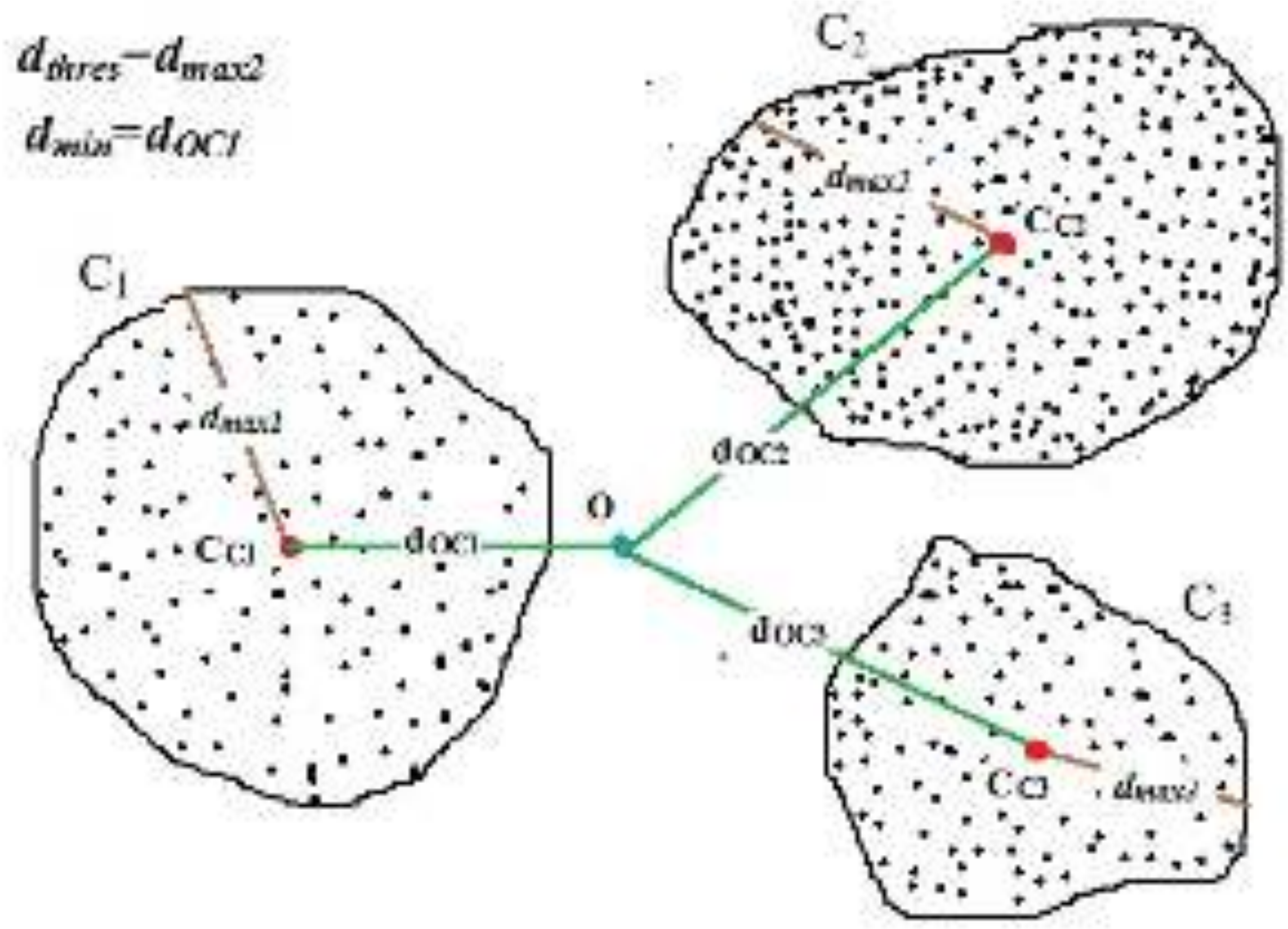


The following are a few of the more commonly used outlier tests for normally distributed data. This list is not exhaustive (a large number of outlier tests have been proposed in the literature). The tests given here are essentially based on the criterion of "distance from the mean". This is not the only criterion that could be used. For example, the Dixon test, which is not discussed here, is based a value being too large (or small) compared to its nearest neighbor.

1. *Grubbs' Test* - this is the recommended test when testing for a single outlier.
2. Tietjen-Moore Test - this is a generalization of the *Grubbs'* test to the case of more than one outlier. It has the limitation that the number of outliers must be specified exactly.
3. Generalized Extreme Studentized Deviate (ESD) Test - this test requires only an upper bound on the suspected number of outliers and is the recommended test when the exact number of outliers is not known.

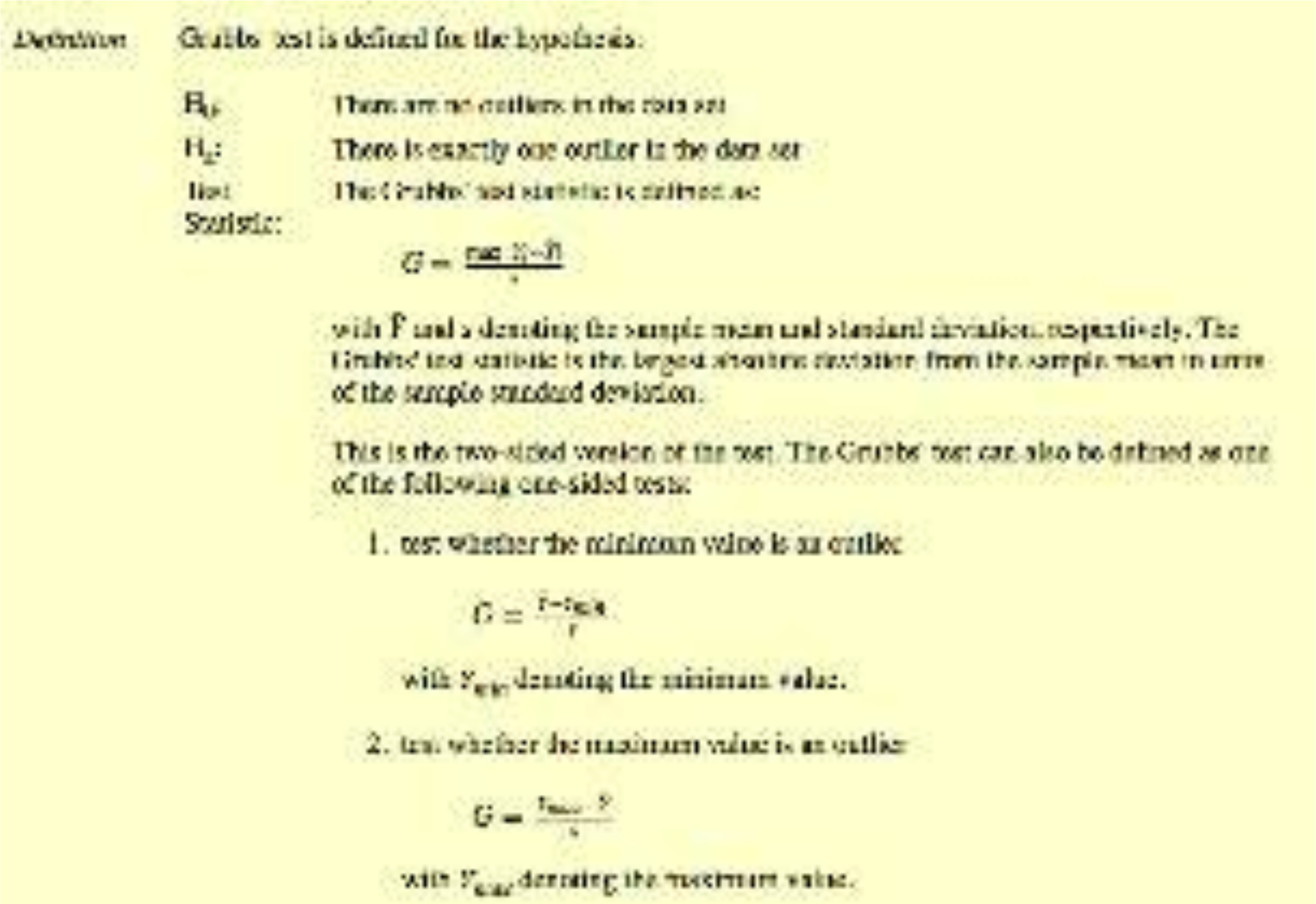
http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm

# Iskanje osamelcev – več skupin



[Mira et la. (2012) RODHA: Robust Outlier Detection using Hybrid Approach]

# Iskanje osamelcev: Z-score
# Samo en osamelec?



Definition Grubbs' test is defined for the hypothesis:

$H_0$: There are no outliers in the data set

$H_a$: There is exactly one outlier in the data set

Test Statistic: The Grubbs' test statistic is defined as:

$$G = \frac{\max |Y_i - \bar{Y}|}{s}$$

with $\bar{Y}$ and $s$ denoting the sample mean and standard deviation, respectively. The Grubbs' test statistic is the largest absolute deviation from the sample mean in units of the sample standard deviation.

This is the two-sided version of the test. The Grubbs' test can also be defined as one of the following one-sided tests:

1. test whether the minimum value is an outlier

$$G = \frac{\bar{Y} - Y_{min}}{s}$$

with $Y_{min}$ denoting the minimum value.

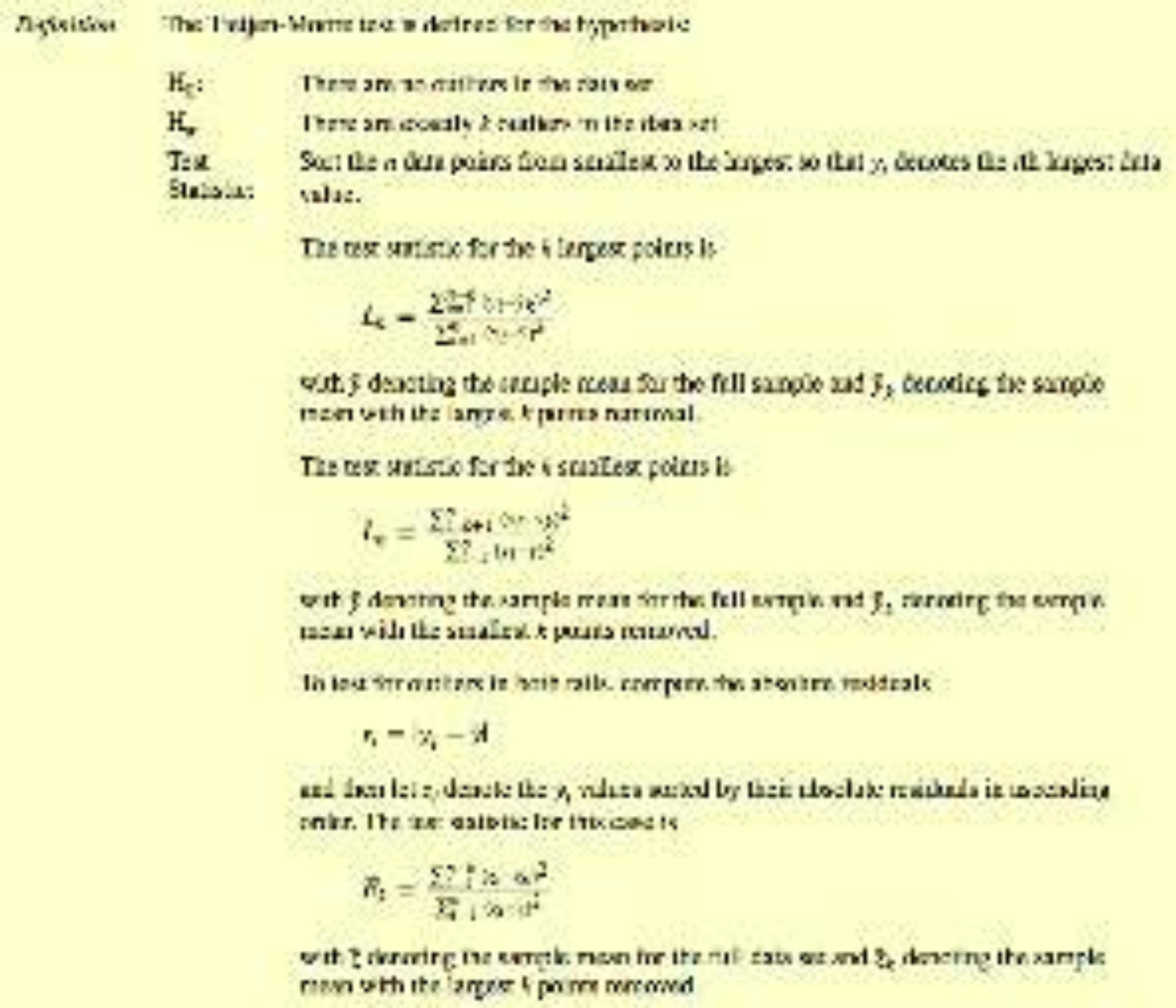
2. test whether the maximum value is an outlier

$$G = \frac{Y_{max} - \bar{Y}}{s}$$

with $Y_{max}$ denoting the maximum value.

http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h1.htm

# Iskanje osamelcev: Z-score
# Točno k osamelcev?



*Definition* The Tietjen-Moore test is defined for the hypothesis:

$H_0$: There are no outliers in the data set

$H_a$: There are exactly $k$ outliers in the data set

Test Statistic: Sort the $n$ data points from smallest to the largest so that $y_i$ denotes the $i$th largest data value.

The test statistic for the $k$ largest points is

$$L_k = \frac{\sum_{i=1}^{n-k}(y_i-\bar{y}_k)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

with $\bar{y}$ denoting the sample mean for the full sample and $\bar{y}_k$ denoting the sample mean with the largest $k$ points removed.

The test statistic for the $k$ smallest points is

$$L_k = \frac{\sum_{i=k+1}^{n}(y_i-\bar{y}_k)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

with $\bar{y}$ denoting the sample mean for the full sample and $\bar{y}_k$ denoting the sample mean with the smallest $k$ points removed.

To test for outliers in both tails, compute the absolute residuals

$$r_i = |y_i - \bar{y}|$$

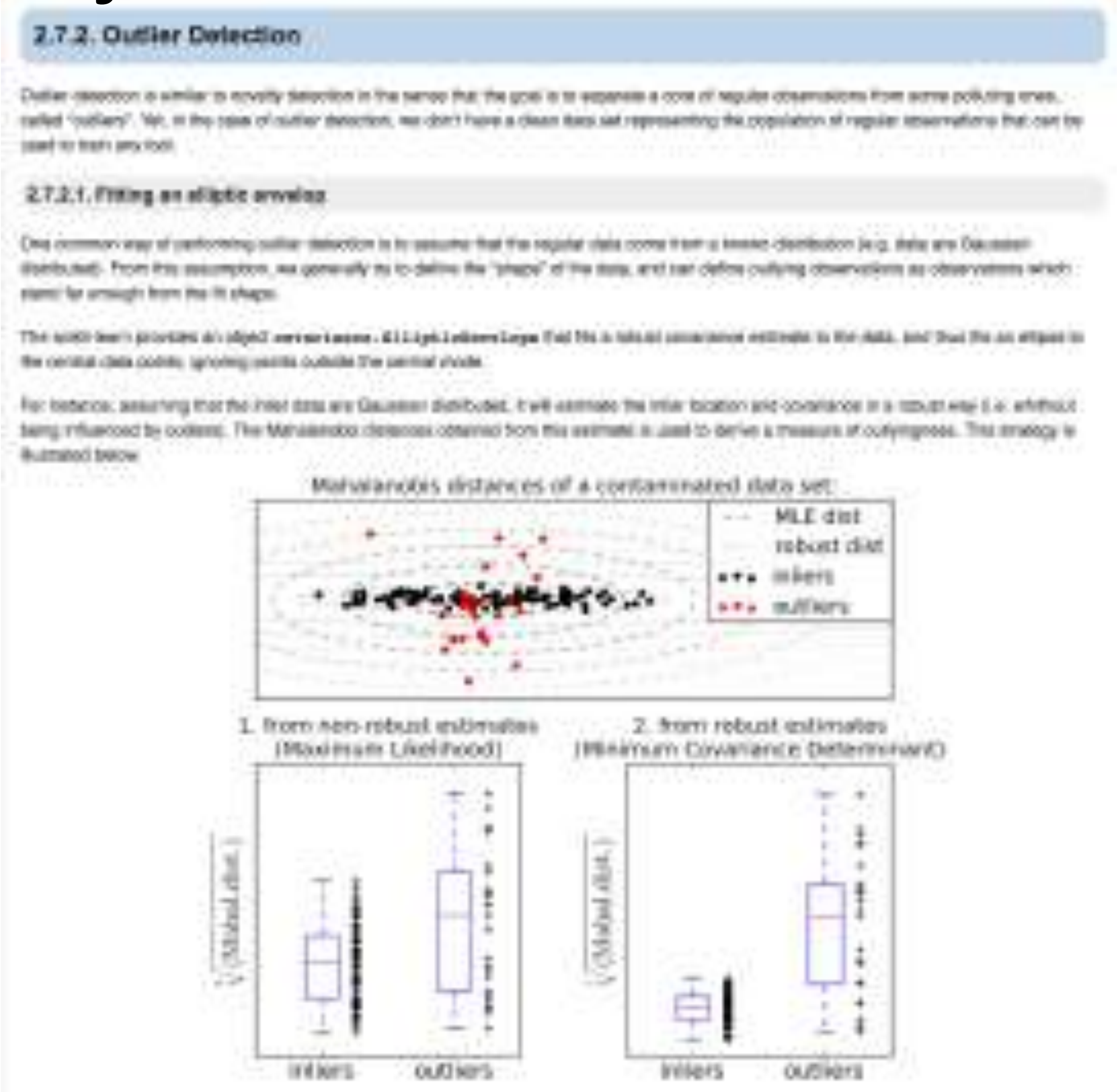
and then let $z_i$ denote the $y_i$ values sorted by their absolute residuals in ascending order. The test statistic for this case is

$$E_k = \frac{\sum_{i=1}^{n-k}(z_i-\bar{z}_k)^2}{\sum_{i=1}^{n}(z_i-\bar{z})^2}$$

with $\bar{z}$ denoting the sample mean for the full data set and $\bar{z}_k$ denoting the sample mean with the largest $k$ points removed.

http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h1.htm

# Iskanje osamelcev: model-based



## 2.7.2. Outlier Detection

Outlier detection is similar to novelty detection in the sense that the goal is to separate a core of regular observations from some polluting ones, called "outliers". Yet, in the case of outlier detection, we don't have a clean data set representing the population of regular observations that can be used to train any tool.

### 2.7.2.1. Fitting an elliptic envelop

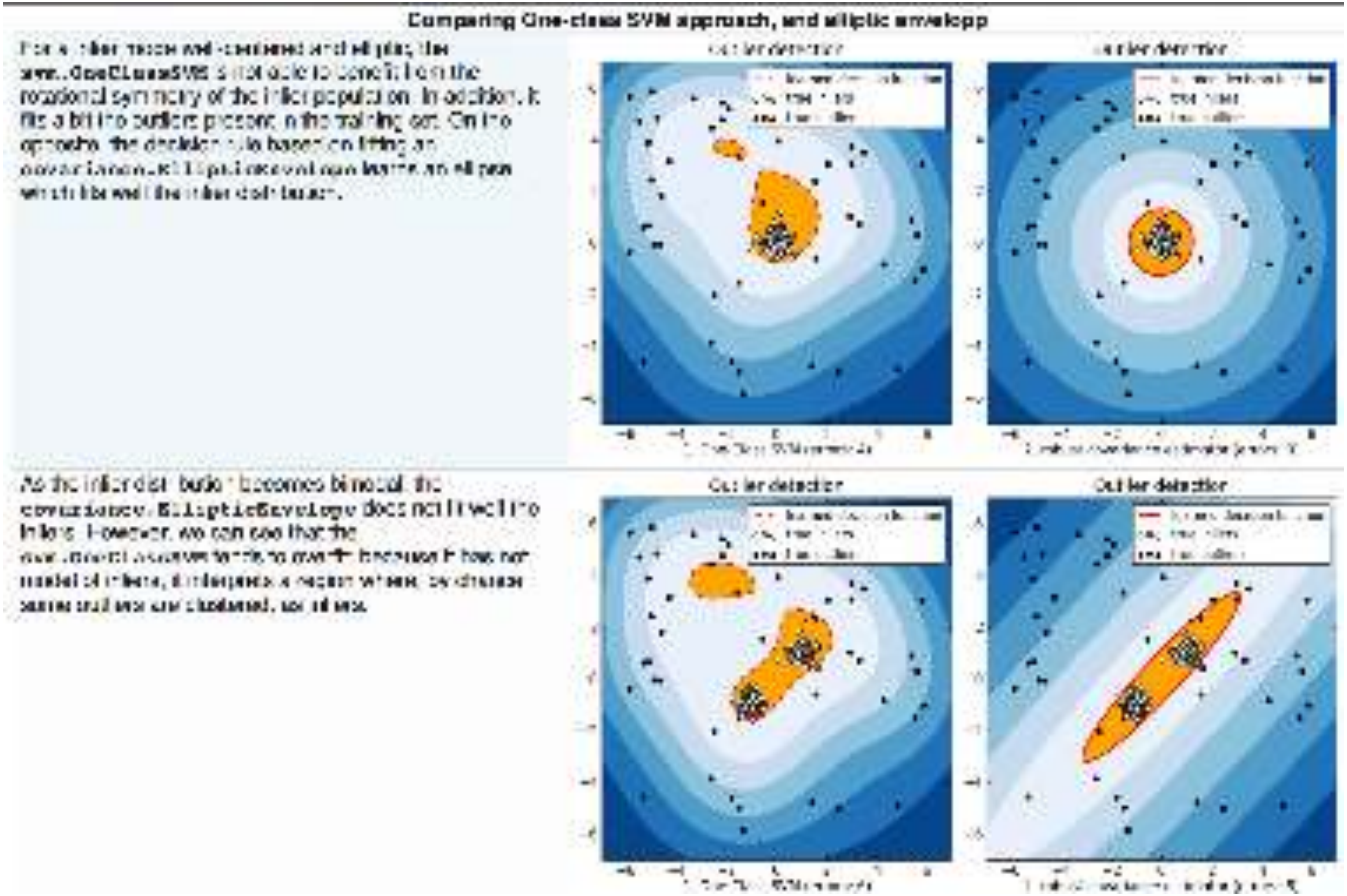
One common way of performing outlier detection is to assume that the regular data come from a known distribution (e.g. data are Gaussian distributed). From this assumption, we generally try to define the "shape" of the data, and can define outlying observations as observations which stand far enough from the fit shape.

The scikit-learn provides an object `covariance.EllipticEnvelope` that fits a robust covariance estimate to the data, and thus fits an ellipse to the central data points, ignoring points outside the central mode.

For instance, assuming that the inlier data are Gaussian distributed, it will estimate the inlier location and covariance in a robust way (i.e. without being influenced by outliers). The Mahalanobis distances obtained from this estimate is used to derive a measure of outlyingness. This strategy is illustrated below.

http://scikit-learn.org/stable/modules/outlier_detection.html

# Iskanje osamelcev: model-based



**Comparing One-class SVM approach, and elliptic envelopp**

For a inlier mode well-centered and elliptic, the svm.OneClassSVM is not able to benefit from the rotational symmetry of the inlier population. In addition, it fits a bit the outliers present in the training set. On the opposite, the decision rule based on fitting an covariance.EllipticEnvelope learns an ellipse, which fits well the inlier distribution.

As the inlier distribution becomes bimodal, the covariance.EllipticEnvelope does not fit well the inliers. However, we can see that the svm.OneClassSVM tends to overfit: because it has not model of inliers, it interprets a region where, by chance some outliers are clustered, as inliers.

http://scikit-learn.org/stable/modules/outlier_detection.html

# Feedback – predavanje 3

http://goo.gl/forms/NC0tQUEA5b