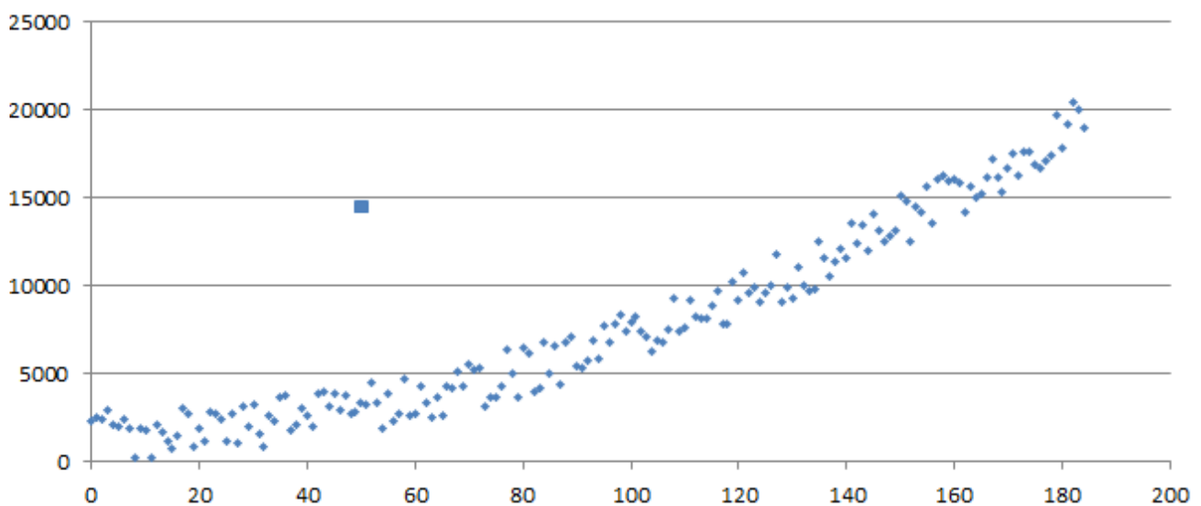
Anomaly Detection
in Exploratory Data Analysis

Definitions

Def:

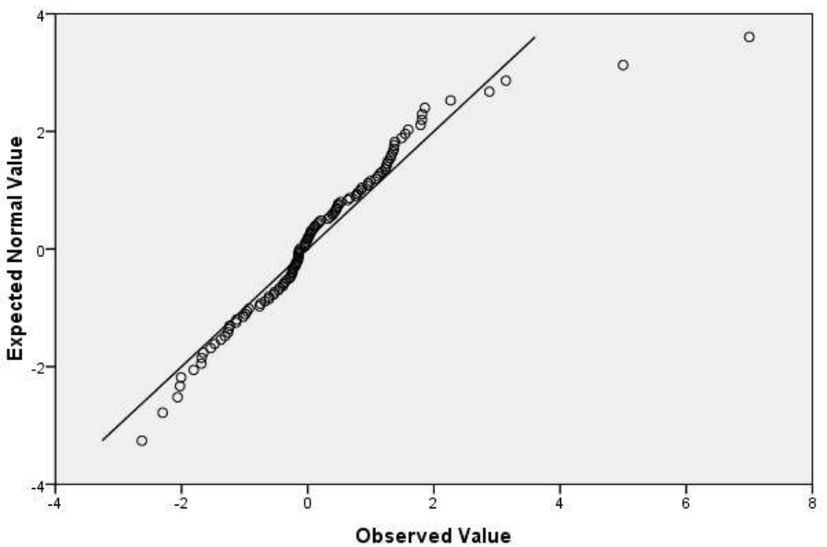
Suppose we have a sample $S = \{x_1, \dots, x_n\}$ of realizations of some random variable ξ with cumulative density function F. Then observation x_j is called **anomaly** (also known as **outlier**) if its value is not consistent with the distribution of ξ.

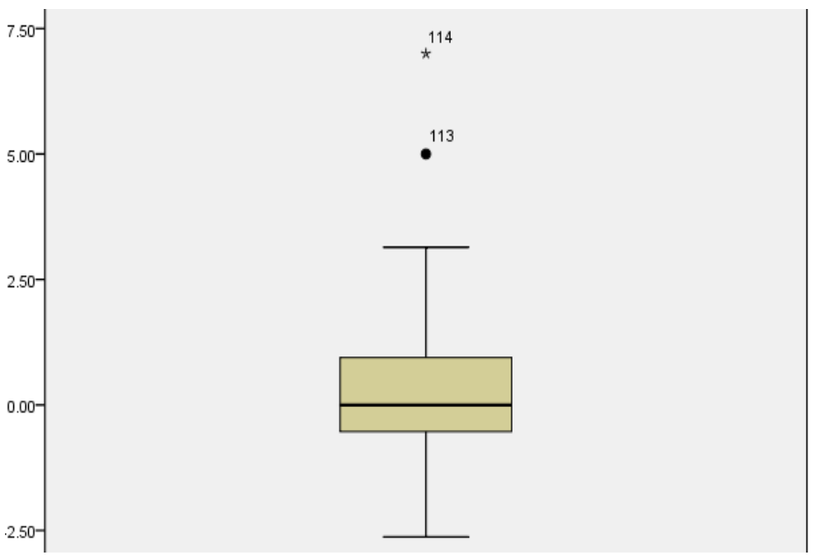
Note that criteria given below can be applied only for normal samples.



Scalar Measurement Processing

- Grubbs' Test
- Thompson Tau Test
- Tietjen-Moore Test
- Graphical Methods (histogram, probability plot, stem-and-leaf plot, box plot, etc.)





Grubbs' Test

1) $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,\ s = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right]}$
2) $z_i = |x_i - \bar{x}|, i = \overline{1,n}$
3) $z_{(1)}, z_{(2)}, \ldots, z_{(n)}$ where $z_{(n)}$ corresponds to the element of the sample $x_{i(n)}$ suspected to be anomaly;
4) $G = \frac{x_{i(n)} - \bar{x}}{s}$
5) In case $|G| > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\frac{\alpha}{2n}, n-2}}{n-2+t^2_{\frac{\alpha}{2n}, n-2}}}$ we say that $x_{i(n)}$ is outlier, delete it from the sample and repeat all the steps.

Thompson Tau Test

1) Repeat steps 1-4 from the previous slide (calculate $\bar{x}, s, z_i$, sort z_i in ascending order, calculate statistic G);
2) Calculate statistic $t = \frac{G\sqrt{n-2}}{\sqrt{n-1-G^2}}$
3) In case $|t| > t_{\frac{\alpha}{2},n-2}$ (where α is significance level) delete $x_{i(n)}$ and repeat all previous steps again;
4) Otherwise there are no outliers in the sample and stop.

Tietjen-Moore Test

1) Calculate $\bar{x}$ and $z_i = |x_i - \bar{x}|, i = \overline{1,n}$;
2) Sort z_i in ascending order and obtain $z_{(1)}, \dots, z_{(n)}$ – sorted list;
3) Take k last elements of sorted list $\{z_{(i)}\}$; they correspond to $x_{i(n-k+1)}, \dots, x_{i(n)}$ which are suspected to be outliers;
4) Calculate $E(n,k)$:

$$E(n,k) = \frac{\sum_{i=1}^{n-k} \left(z_{(i)} - \bar{z}(n-k)\right)^2}{\sum_{i=1}^{n} \left(z_{(i)} - \bar{z}(n)\right)^2}$$

where $\bar{z}(m) = \frac{1}{m}\sum_{i=1}^{m} z_{(i)}$;

5) In case $E(n,k) < E_{1-\alpha}(n,k)$ we say that $x_{i(n-k+1)}, \dots, x_{i(n)}$ are outliers.

Multivariate Outlier Detection

Let $S = \{x_1, \dots, x_n\}$ be a sample, $x_i \in \mathbb{R}^q, i = \overline{1,n}$.

1) $\bar{x}_i = \frac{1}{n-1}\sum_{j\neq i} x_j$, $\hat{\Sigma}_i = \frac{1}{n-2}\sum_{j\neq i}\left(x_j - \bar{x}_i\right)\left(x_j - \bar{x}_i\right)^T$, $i = \overline{1,n}$
2) Calculate Mahalanobis distances:

$$D_i^2 = (x_i - \bar{x}_i)^T \hat{\Sigma}_i^{-1}(x_i - \bar{x}_i)$$

3) $F_i = \frac{(n-1)(n-1-q)}{n(n-2)q} D_i^2, i = 1, \dots, n$
4) $i_0 = \arg\max_i F_i$
5) In case $F_{i_0} > F_\alpha(q, n-1-q)$ we delete x_{i_0} and repeat the procedure from the beginning.

Have any questions?

Website: http://assignment4student.com

Email: info@assignment4student.com

Facebook: http://www.facebook.com/Assignment4Student

Twitter: http://www.twitter.com/assign4student