OUTLIER DETECTION
LECTURE: UNSUPERVISED LEARNING AND EVOLUTIONARY COMPUTATION USING R

Jakob Bossek

MALEO Group, Department of Computer Science, Paderborn University, Germany

11th Nov, 2024

Learning Goals

- Learn about the concept of *outliers*
- Recap Gaussian/Normal distribution and outlier detection of {uni,multi}-variate normally distributed data
- Visual methods for outlier detection
- Shapiro-Wilk normality test and Kolmogorov-Smirnow test
- Glimpse at some other methods

"Data values that are unusually large or small compared to the other values of the same construct." - (Aguinis, Gottfredson, and Joo 2013)

"An outlier is an observation which *deviates so much from the other observations* as to arouse suspicions that it was generated by a different mechanism." - (Hawkins 1980)

Causes of Outliers

Valid extreme values

Natural occurrence of extreme values which are unusual but are not related to any kind of errors. E.g.:

- the wealth of a very few individuals like Elon Musk
- unexpected and/or undiscovered areas of interest

Errors

Outliers are the product of some sort of erroneous behaviour or equipment. E.g.:

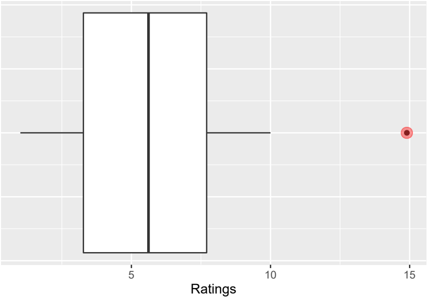
- measurement errors induced by some sort of equipment
- data entry mistakes, e.g., in a survey

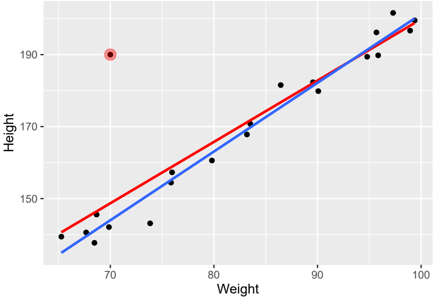
Importance of Outlier Detection

Dependent on the purpose or the goal of some endeavour, outlier detection might have a central or supportive role:

- Data cleaning (our focus)
- Anomaly detection, e.g., in financial transactions or network intrusion
- Discovery or research

Treatment of Outliers





- Treatment of outliers is highly context dependent.
- Outlier of the movie rating example is most likely an encoding error, i.e., can be removed or imputed.
- Outlier in the weight/height example might be a valid value, however, this outlier influences the coefficient of an otherwise good linear model.

Types of Outlier I

Recall:

> "An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism." - (Hawkins 1980)

Problem:

What constitutes a deviation that is larger than the natural pattern of variability present in the data? Is this related to a single and/or multiple features?

Types of Outlier II

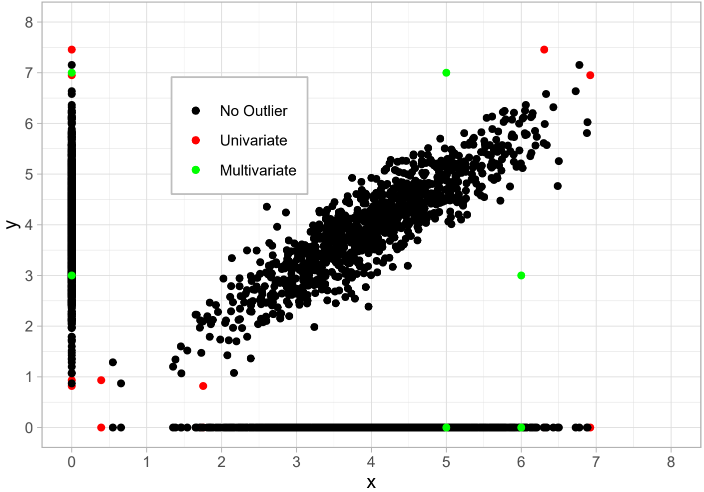
Univariate Outliers

An observation which exhibits an extreme value within a single feature.

Multivariate Outliers

Observations which deviate from the distribution of underlying data when looking at multiple features.

Types of Outlier III



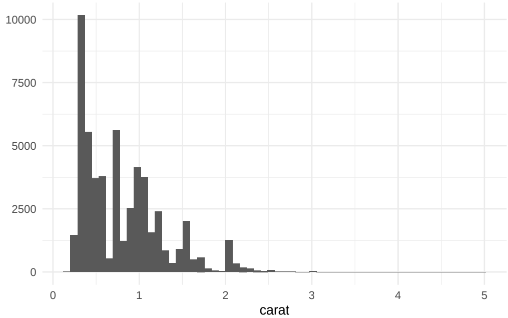
Methods to Detect Outliers

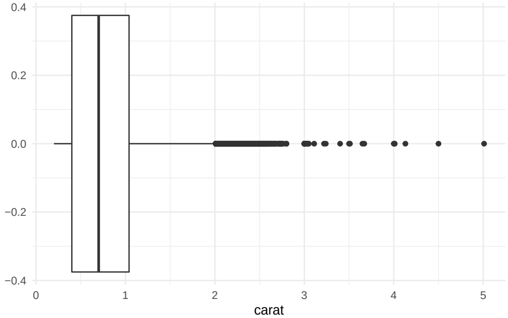
Upcoming methods to identify possible outliers. Some methods make assumption about the underlying distribution (parametric methods) whereas others are robust. Yet, each method comes with its own merit and pitfalls.

- Visual assessment
- Parametric methods
- Depth-based approaches
- many more exists (which are not covered in this lecture)

Visual Outlier Detection I

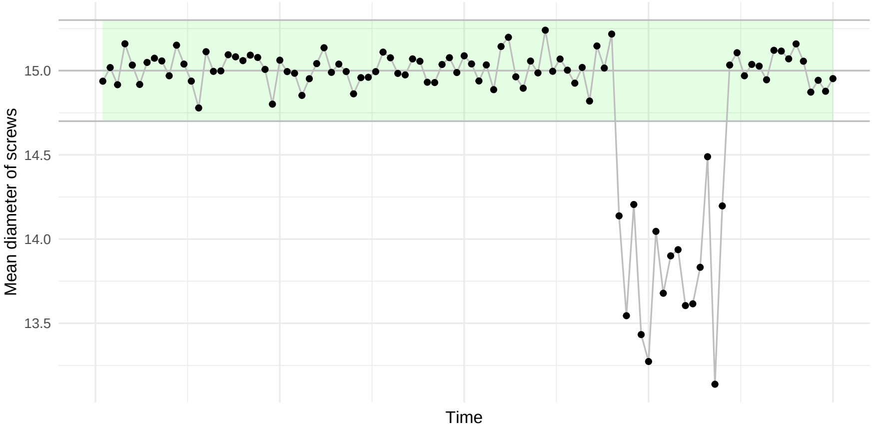
Example from the `carat` variable of the diamonds dataset:



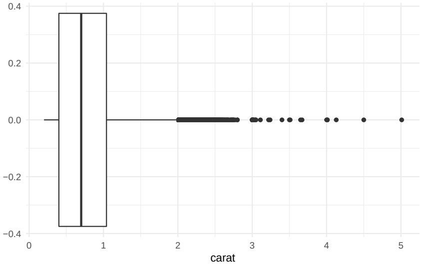


Visual Outlier Detection II

Example from fabrication of screws over time:



Visual Outlier Detection



- median ($q_{0.5}$) is a more robust than the arithmetic mean
- the box covers 50% of the data, starting from the quartile $q_{0.25}$ to $q_{0.75}$
- the *interquartile range* $IQR := q_{0.75} - q_{0.25}$ is used to draw the whiskers
- points outside of the interval $[q_{0.25} - 1.5 \cdot IQR; q_{0.75} + 1.5 \cdot IQR]$ are considered as outliers

Normality

Reasons for checking

- *Parametric* outlier detection
- Many statistical techniques assume data stemming from a normal distribution. E.g.,
    - One- or two-sample t-test
    - Distribution of residuals in linear regression
    - etc.
- Normal distribution violated?
    - Use transformation (e.g., *Box-Cox-transformation*) or
    - Use non-parametric alternatives

Normal/Gaussian Distribution
Importance

- **Natural and Social Phenomena:** Many real-world data, like heights and IQ scores, approximate a Normal distribution.
- **Foundation for Statistical Methods:** Assumed in methods like t-tests and regression, simplifying calculations and inference.
- **Symmetry and Predictability:** Symmetric and predictable within standard deviations, allowing easy probability estimates.
- **Central Limit Theorem (CLT):** Sums of large samples approach a Normal distribution, enabling inference even with non-normal data.
- **Maximizing Information:** Maximizes entropy, making it a preferred model for uncertain situations with given mean and variance.

Recap: Normal Distribution (e.g., Johnson and Wichern 2014)

Normal distribution

A continuous random variable X follows a Normal distribution[1] $\mathcal{N}(\mu, \sigma^2), \sigma > 0$ if X has the *density function*

$$f_X(x) = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} \cdot \exp\left(-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2\right).$$

It holds that $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$. The *cumulative distribiution function* is

$$P(x \leq X) = \phi(x) = \int_{-\infty}^{x} f_X(z)\, dz$$

[1] Also termed Gaussian distribution.

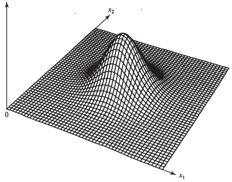
Recap: Multivariate Normal Distribution (e.g., Johnson and Wichern 2014)

A random vector $\mathbf{X}$ follows a multivariate Gaussian $\mathbf{X} \sim \mathcal{N}_p(\mu, \Sigma)$ with *covariance matrix*

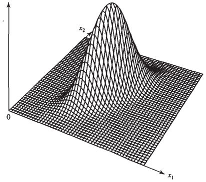
$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix} \sim (p, p).$$

and density function

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^p \cdot \det(\Sigma)}} \cdot \exp\left(-\frac{1}{2} \cdot (\mathbf{x} - \mu)^T \cdot \Sigma^{-1} \cdot (\mathbf{x} - \mu)\right)$$



$\sigma_{11} = \sigma_{22}, \rho_{12} = 0$



$\sigma_{11} = \sigma_{22}, \rho_{12} = 0.75$

Recap: Statistics

Covariance

Let X and Y be two random variables. Then

$$\mathrm{Cov}(X, Y) := E\left[(X - E(X)) \cdot (Y - E(Y))\right]$$

is the *covariance* of X and Y.

- Empirical analogue is the *empirical covariance* $s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})$.
- The covariance is symmetric: $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
- $\mathrm{Cov}(X, Y) = 0$ means no linear relationship.
- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
- For $a, b, c, d \in \mathbb{R}$ bilinearity holds, i.e.,

$$\mathrm{Cov}(aX + b, cY + d) = ac\,\mathrm{Cov}(X, Y).$$

Example: empirical covariance

Consider $n = 5$ observations of two RVs X and Y:

i	1	2	3	4	5
x_i	8	6	2	1	3
y_i	10	7	2	4	2

Then we get $\bar{x} = 4, \bar{y} = 5$ and

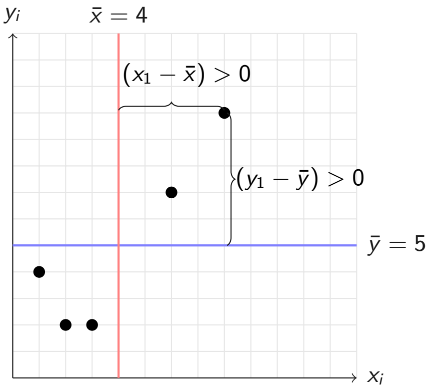
$$\begin{aligned} s_{xy} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y}) \\ &= \frac{1}{5}((8-4)\cdot(10-5) + (6-4)\cdot(7-5) + (2-4)\cdot(2-5) \\ &\quad + (1-4)\cdot(4-5) + (3-4)\cdot(2-5)) \\ &= 7.2. \end{aligned}$$

Recap: Statistics

Empirical covariance illustration

Recall: $s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})$

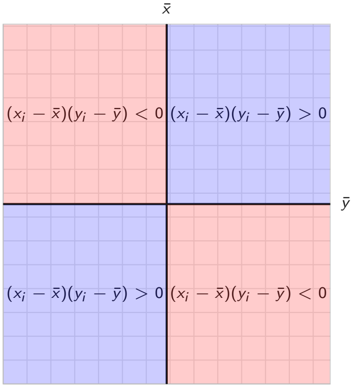
i	1	2	3	4	5
x_i	8	6	2	1	3
y_i	10	7	2	4	2



Recap: Statistics

Covariance illustration

For all points in the red regions the contribution to the covariance is negative while it is positive for all points in the blue regions.



Recap: Statistics
Covariance: interpretation

Let X and Y be two random variables with covariance

$$\mathrm{Cov}(X, Y) := E\left[(X - E(X)) \cdot (Y - E(Y))\right].$$

- If $\mathrm{Cov}(X, Y) > 0 \rightsquigarrow$ positive linear relationship: high values of X go hand in hand with high values of Y and low values of X with low values of Y.
- If $\mathrm{Cov}(X, Y) < 0 \rightsquigarrow$ negative linear relationship: high values of X go hand in hand with low values of Y and low values of X with high values of Y.
- If $\mathrm{Cov}(X, Y) = 0$, there is no linear relationship.[2]
- Strengh of the linear relationship not measurable with the covariance.

---

[2] Attention: there might be a non-linear relationship though!

Recap: Statistics

Correlation

Let X and Y be two random variables. Then

$$\mathrm{Cor}(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)} \cdot \sqrt{\mathrm{Var}(Y)}} \in [-1, 1].$$

is the *correlation* between X and Y.

- Kind of "normalized" covariance.
- Nice interpretation: correlation 1 $\rightsquigarrow$ perfect linear relationship.
- Empirical analogue is the *empirical covariance*

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2} \cdot \sqrt{s_y^2}} = \frac{s_{xy}}{s_x \cdot s_y}.$$

Recap: Statistics
Covariance matrix

Let $X = (X_1, \ldots, X_p)^T$ be a vector of random variables each with finite expected value and variance. The the *covariance matrix* $\mathrm{Cov}(X)$ of X is a square $(p \times p)$ matrix where the components describe the covariance between pairs of variables. I.e. for $1 \leq i, j \leq p$:

$$\mathrm{Cov}(X)_{ij} = \mathrm{Cov}(X_i, X_j) = E\left[(X_i - E(X_i)) \cdot (X_j - E(X_j))\right].$$

- The covariance matrix is symmetric.
- The diagonal entries are the individual variances since:

$$\mathrm{Cov}(X_i, X_i) = E\left[(X_i - E(X_i))^2\right] = \mathrm{Var}(X_i).$$

- Correlation matrix is defined analogeously: $\mathrm{Cor}(X)_{ij} = \mathrm{Cor}(X_i, X_j)$.

Recap: Statistics

Diagonal covariance matrix

If all variables p RVs of a random vector $X = (X_1, \dots, X_p)^T$ are pairwise uncorrelated, i.e., $\operatorname{Cov}(X_i, X_j) = 0$ for $1 \leq i \neq j \leq p$, the covariance matrix has diagonal form

$$\begin{aligned}\operatorname{Cov}(X) &= \begin{pmatrix} \operatorname{Var}(X_1) & 0 & 0 & \dots & 0 \\ 0 & \operatorname{Var}(X_2) & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & \operatorname{Var}(X_{p-1}) & 0 \\ 0 & \dots & 0 & 0 & \operatorname{Var}(X_p) \end{pmatrix} \\ &= \operatorname{diag}(\operatorname{Var}(X_1), \operatorname{Var}(X_2), \dots, \operatorname{Var}(X_p)) \\ &= \operatorname{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2)\end{aligned}$$

Exercises

Let $X \sim \mathcal{N}(0, 1)$. Determine

1. an $x \in \mathbb{R}$ such that $P(X \geq x) = 0.5$ and
2. an $y \in \mathbb{R}$ such that $P(|X| \geq y) = 0.5$.

Sample Solutions

Statistical Significance Tests
Brief reminder

- The first thing any statistical hypothesis does is to calculate a so-called *test-statistic* W based on theoretical assumptions and the given data
- The distribution of W <u>under H_0</u> is known!
- Given data, we can calculate the *p-value* p^*.
  Interpretation: p^* is the probability to get this sample data given H_0 is true
  - $p^* = 1 \rightsquigarrow$ sample is very lickely under H_0
  - $p^* < \alpha \in \{0.05, 0.01, 0.001\} \rightsquigarrow$ sample is very unlickly given H_0

$\rightsquigarrow$ **Decision:** reject H_0 if $p^* < \alpha$

Statistical Significance Tests

Brief reminder

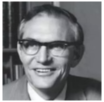
Overview of the four possible outcomes of a statistical hypothesis test at significance level α:

		Zero-hypothesis H_0 is ...	
		True	**False**
Test result is ...	**rejected** ($p^* < \alpha$)	false positive ☹ **Type I error / α-error** Probability α	true positive ☺ probability $(1-\beta)$
	not rejected ($p^* \geq \alpha$)	true negative ☺ probability $(1-\alpha)$	false negative ☹ **Type II error / β-error** probability β

Parametric Methods: Prerequisites

- Parametric methods assume that the data is normal distributed.
- If the data is not normal distributed, you will get meaningless results!
- Never apply parametric methods to detect outliers without checking for normality first!
- Hence, in the following we learn how to check for normality.

Shapiro-Wilk Test
Working principle




Shapiro-Wilk hypothesis test checks the hypothesis pair

$$H_0 : \text{data normally distr. versus } H_1 : \text{data is } \underline{\text{not}} \text{ normally distr.}$$

where F_0 is a Gaussian distribution. The *test-statistic* is

$$W = \frac{b^2}{(n-1)s^2} \in (0,1)$$

where

- b^2 is an estimator for the variance if it would stem from a normal distribution
- s^2 is the ordinary unbiased estimator $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

$\rightsquigarrow$ W close to 1 indicates a Gaussian, values $W < W^*$ with critical value W^* indicate a non-Gaussian distribution

Shapiro-Wilk Test
Recipe

Given a sample of numerical observations $x_1, \dots, x_n$ with $3 \leq 3 \leq 5\,000$

1. Calculate the *test statistic*

$$W = \frac{b^2}{(n-1)s^2} = \frac{b^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

   where

$$b = a_{(1)}(x_{(n)} - x_{(1)}) + a_{(2)}(x_{(n-1)} - x_{(2)}) + \dots$$

   and the $a_{(i)}$ result from normality assumption

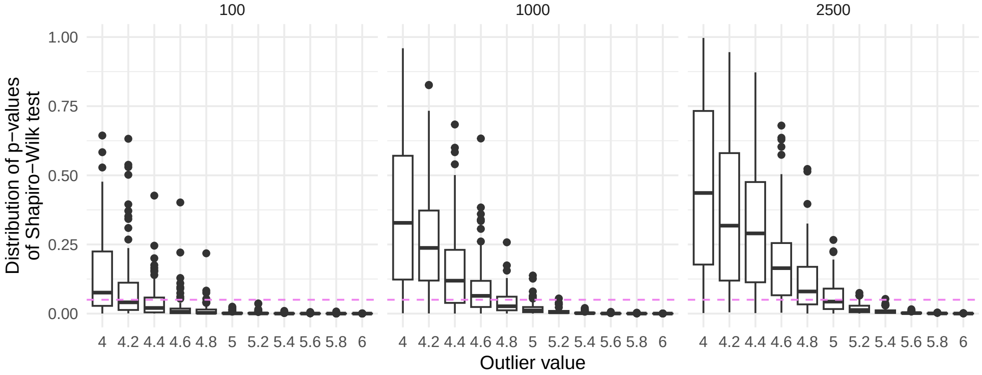
2. Compare W with the *critical value* W^* and reject H_0 if $W < W^*$

3. Nowadays: calculate the p-value and reject H_0 if p-value is below *significance level* $\alpha \in [0, 1]$[3]

[3] In R: `shapiro.test(x)$p.value`

Shapiro-Wilk Test
Sensitivity to Outliers

Data from a $\mathcal{N}(0,1)$ with <u>a single</u> outlier (x-axis):



Shapiro-Wilk Test
Properties

Advantages ☺

- Objective measure (in comparison to subjective graphical methods, e.g., QQ-plots or histograms)
- Mean value and variance of the Gaussian is not necessary
- Available in most software libs (e.g., R, SAS, SPSS)

Disadvantages ☹

- Applicable for $3 \leq n \leq 5\,000$
- Sensitive to outliers
- Quite sensitive to duplicates/ties
- SW is a so-called *omnibus test*: can tell there is a deviation, but does not give an explanation (e.g., skewness)
- No general adapdation test $\leadsto$ specialised normality test

Box-Cox Transformation

Transforming into normal

Box-Cox-transformation is a statistical means to stabalise the variance and make the data follow a normal distribution.

Let $x_1, \ldots, x_n$ be the sample data. Then the transformation is defined as

$$y_i = \begin{cases} \frac{x_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(x_i) & \text{if } \lambda = 0 \end{cases}, i = 1, \ldots, n$$

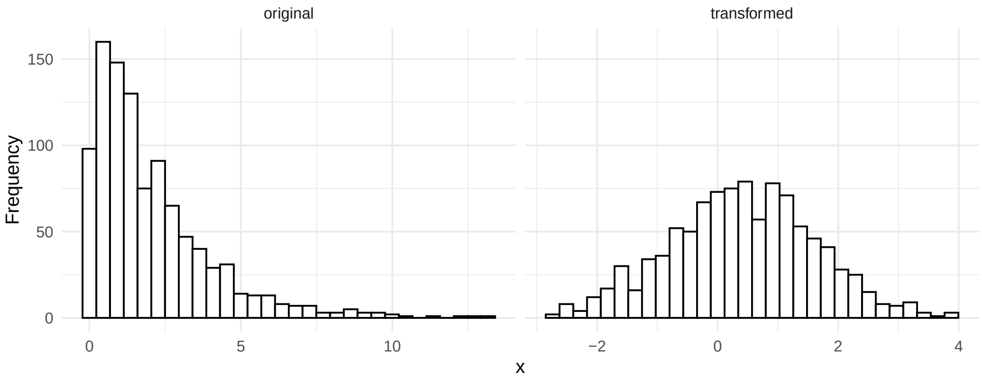
where λ is a *transformation parameter*[4]. E.g.,

- $\lambda = 0 \rightsquigarrow$ logarithmic transformation
- $\lambda = -1 \rightsquigarrow$ reciprocal transformation

[4] Parameter λ is chosen to maximise the likelihood function, effectively finding the value that makes the transformed data as close to normally distributed as possible.

Box-Cox-Transformation
Example

Left: data from an Exp(2)-distribution. **Right**: Box-Cox transformed data ($\lambda = 0.3$)



Box-Cox-Transformation
Properties

Advantages ☺

- Simple and straight-forward approach
- Possibility to apply sophisticated statistical methods that are based on normality assumption

Disadvantages ☹

- Only applicable to strictly positive data
- For data with extreme skewness normality might not be achieve normality
- Certainly make interpretation harder

Kolmogorov-Smirnow Test
Working principle


Kolmogorov-Smirnov test checks the hypothesis pair

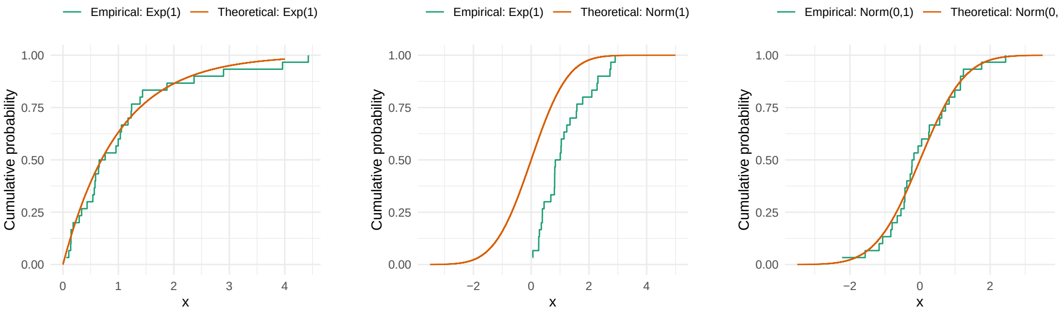
$$H_0 : F = F_0 \text{ versus } H_1 : F \neq F_0$$

where F_0 is any theoretical distribution and F is the data distribution. The *test-statistic* is

$$W = \sup_{x \in \mathbb{R}} |\tilde{F}_n(x) - F_0(x)| \text{ with } \tilde{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(-\infty,x]}(X_i), x \in \mathbb{R}.$$

Interpretation: Maximum deviation between the empirical cumulative distribution function $\tilde{F}_n(x)$ and its theoretical counterpart $F_0(x)$

Kolmogorov-Smirnow Test
Visualisation



Kolmogorov-Smirnow Test
Properties

Advantages ☺

- Simple and intuitive
- Nonparametric test[5]
- General distribution test
- Versatility:
  - One-sample K-S test: test if a sample comes from a specific distribution
  - Two-sample K-S test: test if two samples are drawn from the same distribution

Disadvantages ☹

- Requires continuous data
- Assumes that there are no ties (identical values) in the data. Ties can affect the test statistic, leading to potential inaccuracies

[5] However, if testing against a specific distribution, the parameter need to be known.

Excurse: Probability Plots

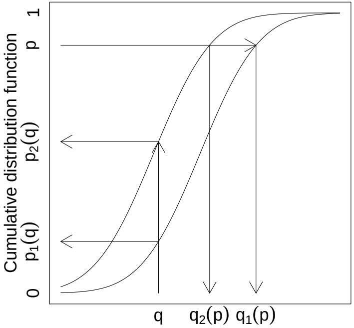
Observations are ordered and plotted against an *assumed* cumulative distribution function. There are different kinds of probability plots:

- Probability-probability plot whose coordinates are

$$p_1(q) = P(X_1 \leq q)$$
$$p_2(q) = P(X_2 \leq q)$$

- Quantile-quantile (QQ) plot whose coordinates are

$$q_1(p) = p_1^{-1}(p)$$
$$q_2(p) = p_2^{-1}(p)$$



Excurse: QQ-Plots

QQ-plots can be used for checking the **univariate** normality assumption:

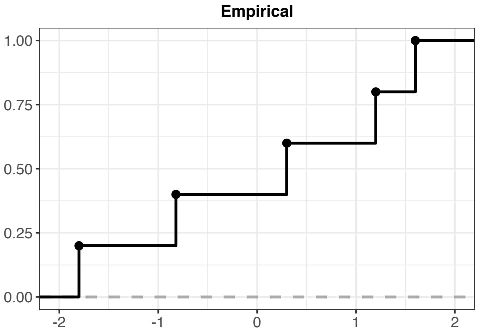
1. Given a sample of X, i.e., realisations $x_1, x_2, \ldots, x_n$
2. Order the sample to $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$

   $\rightsquigarrow$ the (ordered) $x_{(i)}$, $i = 1, \ldots, n$ are the *(empirical) quantiles* of the sample.
3. As the sample is ordered, exactly j observations are less than or equal to $x_{(j)}$.
4. For analytical convenience, the proportion j/n to the left of $x_{(j)}$ is approximated by

$$j^* = \frac{j - 1/2}{n} = \frac{j}{n} - \frac{1}{2} \cdot \frac{1}{n} \quad \text{(continuity correction).}$$

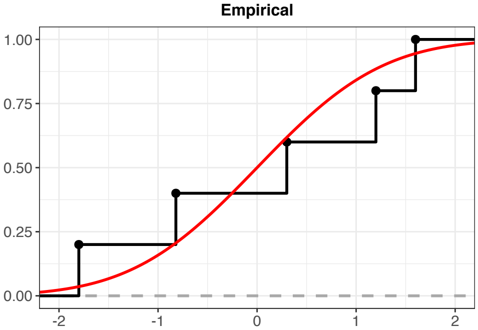
Excurse: QQ-Plots

$x_{(j)}$	ranks (j)	j^*	$q_{(j^*)}$ (normal)
-1.80	1	0.1	
-0.82	2	0.3	
0.30	3	0.5	
1.20	4	0.7	
1.60	5	0.9	



Excurse: QQ-Plots

$x_{(j)}$	ranks (j)	j^*	$q_{(j^*)}$ (normal)
-1.80	1	0.1	
-0.82	2	0.3	
0.30	3	0.5	
1.20	4	0.7	
1.60	5	0.9	

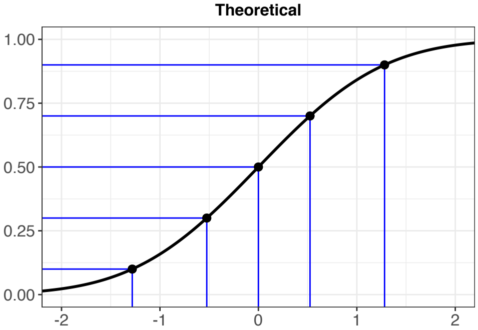


For a standard normal distributed variable $X \sim \mathcal{N}(0,1)$, the probability for observing a value, which is smaller or equal to $q(j)$ is

$$p_{(j)} := P\left(X \leq q_{(j)}\right) = \int_{-\infty}^{q_{(j)}} \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2} \cdot z^2\right) dz.$$

Excurse: QQ-Plots

$x_{(j)}$	ranks (j)	j^*	$q_{(j^*)}$ (normal)
-1.80	1	0.1	-1.28
-0.82	2	0.3	-0.52
0.30	3	0.5	0.00
1.20	4	0.7	0.52
1.60	5	0.9	1.28

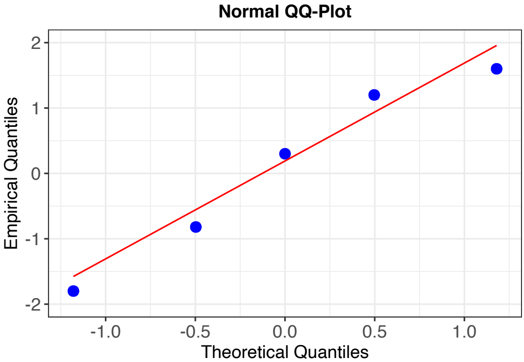


Given that $j^* = \frac{j-1/2}{n}$ represents the (corrected) proportion on the left of (or equal to) $q_{(j^*)}$ the corresponding quantiles $q_{(j^*)}$ of X are

$$q_{(j^*)} = \Phi^{-1}(j^*) = \Phi^{-1}\left(\frac{j-1/2}{n}\right).$$

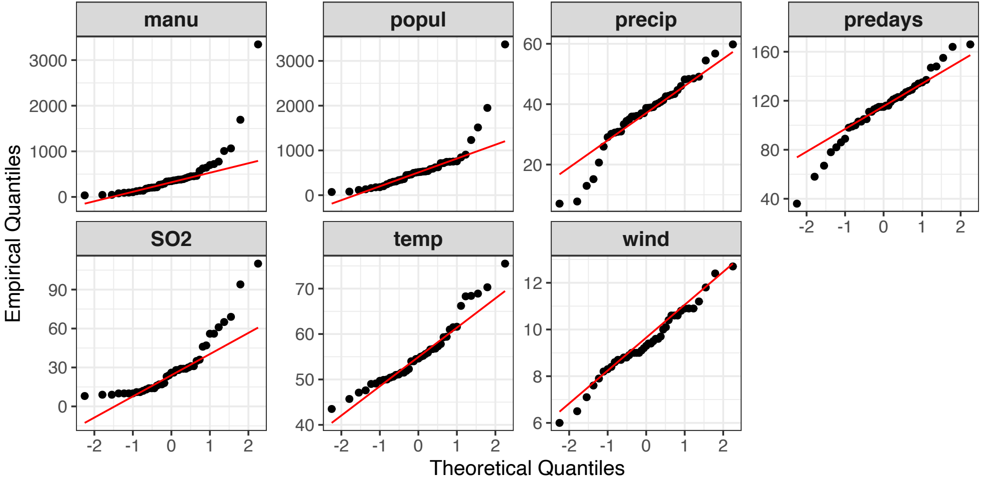
Excurse: QQ-Plots

$x_{(j)}$	ranks (j)	j^*	$q_{(j^*)}$ (normal)
-1.80	1	0.1	-1.28
-0.82	2	0.3	-0.52
0.30	3	0.5	0.00
1.20	4	0.7	0.52
1.60	5	0.9	1.28



Plot the pairs $(q_{(j^*)}, x_{(j)})$. If the sample belongs to a normal distribution, the pairs *should* possess an approximately linear relationship.

Example: US Air Pollution Data



Example: US Air Pollution Data

Findings

- Precipitation and SO_2 deviate considerably from normality.
- There is evidence for outliers in plots of manufacturing, predays and population.
- We can only check each variable separately.

Excurse: QQ-Plots

So far:

- only compared to $\mathcal{N}(0,1)$
- n point pairs: $\left(\Phi^{-1}\left(\frac{j-0.5}{n}\right), x_{(j)}\right)$

Now:

- generalize to all normal distributions $\mathcal{N}(\mu,\sigma^2)$
- standardize the data, i.e., $z_i = \frac{x_i-\mu}{\sigma}$
- afterwards compare it to $\mathcal{N}(0,1)$
- n point pairs: $\left(\Phi^{-1}\left(\frac{j-0.5}{n}\right), z_{(j)}\right)$

Univariate Parametric Method for Outlier Detection

Given a standard normal distributed random variable $Z \sim \mathcal{N}(0, 1)$, the general idea of this method is to classify observations as outlier which is far away from the mean. Drawbacks:

- This is restricted to univariate outliers only.
- A standard normal distribution is assumed.
- The reliable testing for normality can only be done with a moderate to large sample.

A transformation from a normal distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ to $Z \sim \mathcal{N}(0, 1)$ is called standardization.

$$Z = \frac{X - \mu}{\sigma}.$$

As a rule of thumb, values which are outside the interval $[-3.5, 3.5]$ can be considered as critical.

Exercises

1. Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Show that the so-called z-transformed random variable

$$Z := \frac{X - \mu}{\sigma}$$

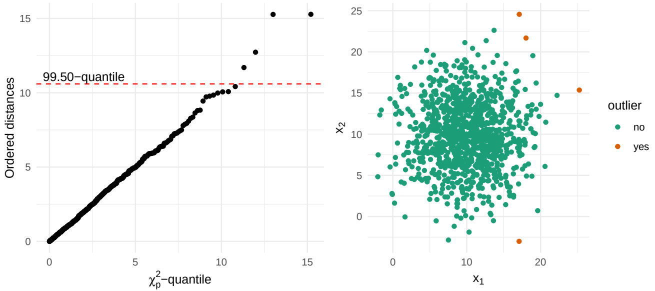
follows a standard normal distribution, i.e., $Z \sim \mathcal{N}(0, 1)$.

2. Show that for $X \sim \mathcal{N}(\mu, \sigma^2)$ it holds that $P(X \in [\mu - k\sigma, \mu + k\sigma]) = 2\phi(k) - 1$.

Sample solutions I

Sample solutions II

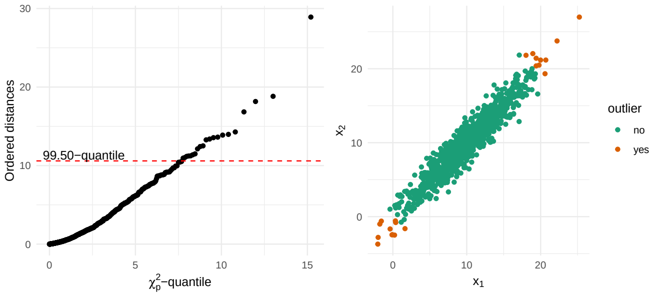
Multivariate Parametric Methods for Outlier Detection



The calculation of the sum of squared distances would be as follows:

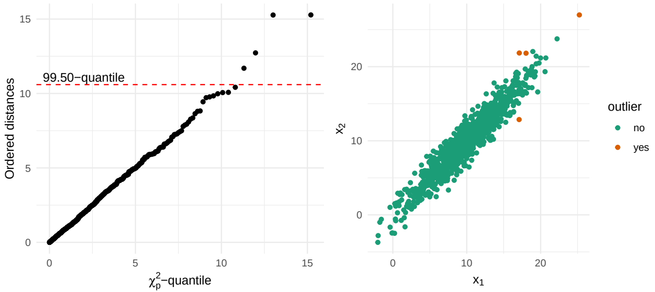
$$\hat{d}^2 = (X - E(X))^T \cdot (X - E(X)) = \sum_{i=1}^{p} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$$

Multivariate Parametric Methods for Outlier Detection



- The problem with $\hat{d}^2$ is that this approach assumes that the observations are spherically scattered around the center of mass.
- If the data is ellipsoidal (i.e. correlated), the output yields unexpected results

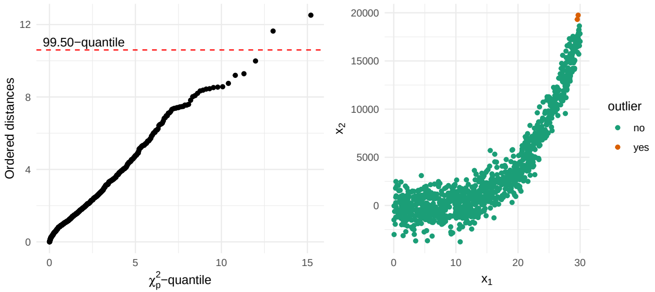
Multivariate Parametric Methods for Outlier Detection



- We require a distance metric which takes correlations into account, i.e., the generalized squared distances

$$d^2 = (X - E(X))^T \cdot \Sigma^{-1} \cdot (X - E(X))$$

Multivariate Parametric Methods for Outlier Detection



Drawbacks:

- Method works only if the data follows a multivariate normal distribution.
- Normality checks can only be done if the sample is large enough.
- Generalized squared distances become meaningless. This phenomenon is known as 'Curse of Dimensionality'.

Recap: χ^2-distribution

χ^2-distribution

Let $X_1, \ldots, X_p$ be *independent, standard normal* random variables. The the sum of their squares is χ^2*-distributed* with p *degree of freedom*, i.e.,

$$Q = \sum_{i=1}^{p} X_i^2 \sim \chi_p^2.$$

- Less often used for modelling nature phenomena
- Very common in hypothesis testing (due to relation to normal-, t- and F-distribitions)
- Since by definition Q is the sum of independent random variables with finite mean and variance, it converges to a normal distribution for large p by the central limit theorem

Chi-Square-Plots

A QQ-Plot for the χ^2-distribution

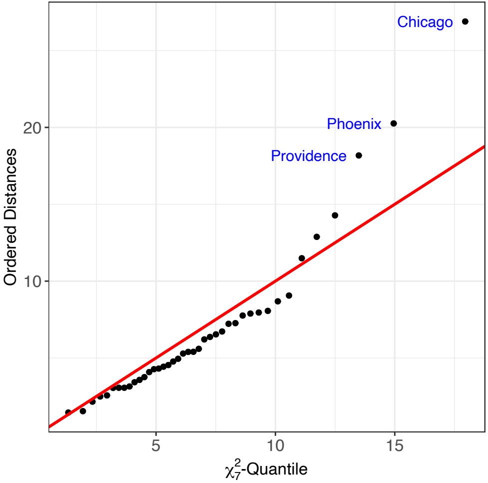
Based on generalized distances:

$$d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})^T \cdot \mathbf{S}^{-1} \cdot (\mathbf{x}_j - \bar{\mathbf{x}}), \quad j = 1, \dots, n \qquad (\text{Note: } d_j^2 \sim \chi_p^2)$$

Steps:

1. Order the squared distances: $d_{(1)}^2 \leq d_{(2)}^2 \leq \cdots \leq d_{(n)}^2$.
2. Plot the pairs $(q_{c,p}((j-1/2)/n), d_{(j)}^2)$ with $q_{c,p}(\alpha)$ being the α-quantile of the χ^2-distribution with p degrees of freedom
3. In case of multivariate normal distributed data, the plot should resemble a straight line through the origin $(0,0)$ with slope 1.

Example: Us Air Pollution Data



Depth-Based Approach

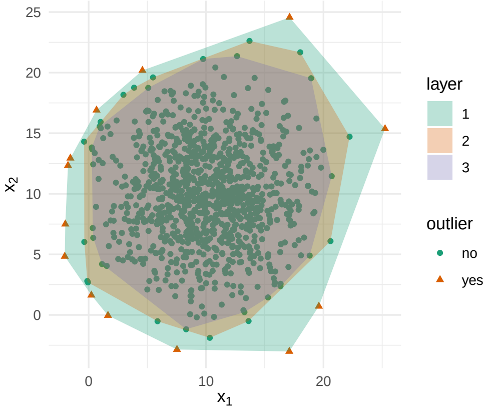
In the *depth-based-approach* (Graham 1972) outliers are located at the border of the space spanned by the data:

1. Initialize a layer counter $L = 1$.
2. Assign all points on the convex hull to layer L.
3. Remove all layer L points from the data.
4. Set $L = L + 1$.
5. Repeat step 2-4 until no points remain.
6. Define how many $k \geq 1$ outer layers L should be labeled as outliers.

Depth-Based Approach

Drawbacks

- Major drawback is the computational complexity
- No differentiation between points within a layer



What we learned today

- The concept of *outliers* or *unusual observations*
- Recap Gaussian/Normal distribution and outlier detection of {uni,multi}-variate normally distributed data
- Visual methods for outlier detection (box-plots, QQ-plots)
- Testing for normality (QQ-plots, Shapiro-Wilk test, KS-test)
- Glimpse at some other methods (e.g., depth-based)

References I

Hawkins, D. M. (1980). *Identification of outliers.* Monographs on applied probability and statistics. Chapman and Hall. ISBN: 041221900X.

Johnson, Richard Arnold and Dean W. Wichern (2014). *Applied Multivariate Statistical Analysis.* 6th. Pearson, Prentice-Hall. URL: https://www1.udel.edu/oiss/pdf/617.pdf.

Graham, R.L. (1972). "An efficient algorithm for determining the convex hull of a finite planar set". In: *Information Processing Letters* 1.4, pp. 132–133. ISSN: 0020-0190. DOI: https://doi.org/10.1016/0020-0190(72)90045-2.