Introduction and descriptive statistics

Assoc. Prof. Dr. Sebastian Thöns
Department of Civil Engineering
Technical University of Denmark, Denmark

TU1402 Training School on the Value of Structural Health Monitoring Information

The objective of the COST TU1402 Training School is to provide the basic concepts for Value of Structural Health Monitoring (SHM) Information analyses in the framework of Bayesian decision theory.

A student who successfully has taken part in the course will be able to:

- Perform Value of SHM Information and decision analyses
- Select appropriate decision analyses types
- Develop and apply probabilistic models of the relevant uncertainties in decision analyses, in structural reliability analyses and for measurement information modelling

Day 1: Probabilistic fundamentals and modelling

Introduction and descriptive statistics

- Introduction of course and lecturers
- Numerical and graphical data description
- Data pair description

Probability theory fundamentals

- Definitions and axioms of probability
- Set theory, probability theory and Bayesian updating
- Uncertainty modelling with random variables

Probabilistic Modelling

- Maximum Likelihood Method
- Bayesian regression analysis
- Determination of model uncertainties

Day 2: Structural reliability and measurements

Structural reliability

- Introduction to reliability analysis
- First order reliability method (FORM)
- Monte Carlo simulation
- Importance sampling

Structural system reliability analysis

- Load and resistance modelling
- Logical systems, Daniels systems
- Target reliabilities

Adaptation of structural reliability with measurement information

- Probabilistic modelling of measurements (measurement uncertainty, probability of detection)
- Updating of structural reliability

Day 3: Decision analyses

Introduction and fundamentals of Bayesian decision analysis

- Prior, Pre-posterior and Posterior decision analysis
- Quantification of utilities

Value of Information analyses and decision analysis types

- Types of Value of Information analyses
- Decision analyses types
- Derivation of decision rules

Value of Information analyses cont.

- Introduction to Gaussian random field models
- Evaluation of Value of Information in Gaussian random fields
- Optimal sensors placement and inspection scheduling using Value of Information

Day 1: Probabilistic fundamentals and modelling

Introduction and descriptive statistics

- Introduction of course and lecturers
- Numerical and graphical data description
- Data pair description

Elementary data analysis

Data analysis provides the mathematical and graphical tools to analyse, document and communicate data.

Examples:

- Concrete compressive strength data
- Traffic counts (visual, cameras, weight in motion)
- Flood and sea level data
- Snow height data
- …

Elementary data analysis

39.2
34.6
35.8
33.3
37.1
27.6
27.8
32.8
27.9
24.4
34.1
33.5
30.3
28.5
35.9
39.7
31.7
36.8
30.1
32.2

Numerical data description

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Central measures

- Sample mean: The centre of gravity of a data set
- Sample median: Value corresponding to the 0.5 quantile
  - Average of the two middle numbers for even number of samples
  - Sample median number: The middle number
- Mode : Most frequent value

Numerical data description

$$\min(x_i), \max(x_i)$$

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

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\upsilon = \frac{s}{\overline{x}}$$

Dispersion measures

- Sample range
- Sample variance: Distribution around the sample mean
- Sample standard deviation: Distribution around the sample mean
- Sample coefficient of variation (CoV): The variability relative to the sample mean

Numerical data description

$$\eta = \frac{1}{n} \cdot \frac{\sum_{i=1}^{n} (x_i - \overline{x})^3}{s^3}$$

$$\kappa = \frac{1}{n} \cdot \frac{\sum_{i=1}^{n} (x_i - \overline{x})^4}{s^4}$$

Shape measures:

- Sample skewness: The skewness relative to the sample mean

- Sample kurtosis: The peakedness around the sample mean

Summary: Numerical data description

The numerical data analysis provides definitions for a numerical form of the statistical characteristics.

Central measures:

- Sample mean value: The centre of gravity of a data set
- Sample median: The mid value of a data set
- Sample mode: The most frequent value/range of a data set

Dispersion measures:

- Sample variance: The distribution around the sample mean
- Sample CoV: The variability relative to the sample mean

Shape measures:

- Sample skewness: The skewness relative to the sample mean
- Sample kurtosis: The peakedness around the sample mean

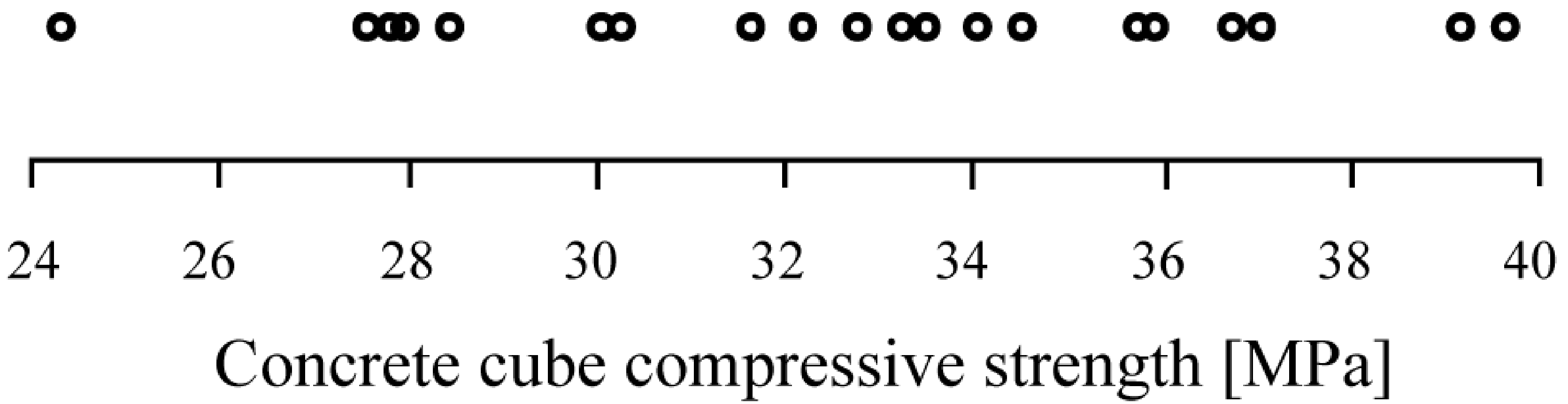
Graphical data description

Exemplary data: concrete compressive strengths

i	Unordered x_i	Ordered x_i^o
1	35.8	24.4
2	39.2	27.6
3	34.6	27.8
4	27.6	27.9
5	37.1	28.5
6	33.3	30.1
7	32.8	30.3
8	34.1	31.7
9	27.9	32.2
10	24.4	32.8
11	27.8	33.3
12	33.5	33.5
13	35.9	34.1
14	39.7	34.6
15	28.5	35.8
16	30.3	35.9
17	31.7	36.8
18	32.2	37.1
19	36.8	39.2
20	30.1	39.7

Graphical data description

One dimensional scatter plot



Graphical data description

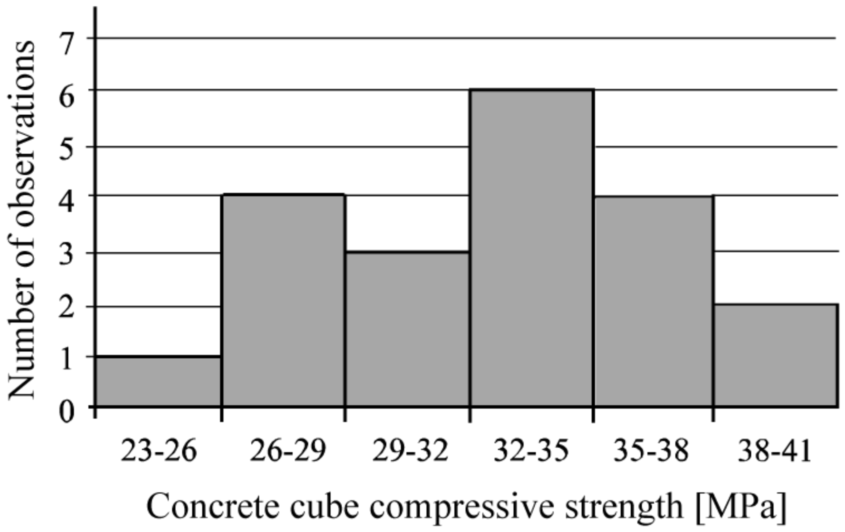
Histograms: Grouping the data into intervals

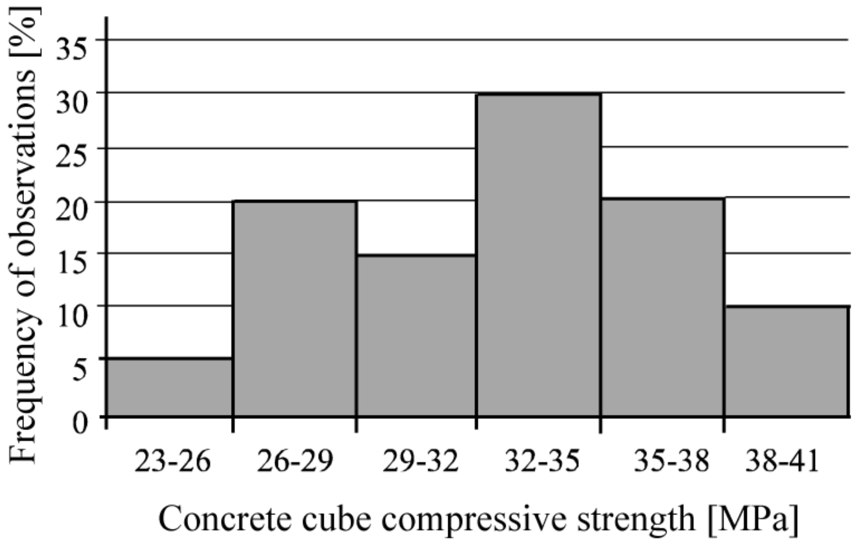
i	Unordered x_i	Ordered x_i^o
1	35.8	24.4
2	39.2	27.6
3	34.6	27.8
4	27.6	27.9
5	37.1	28.5
6	33.3	30.1
7	32.8	30.3
8	34.1	31.7
9	27.9	32.2
10	24.4	32.8
11	27.8	33.3
12	33.5	33.5
13	35.9	34.1
14	39.7	34.6
15	28.5	35.8
16	30.3	35.9
17	31.7	36.8
18	32.2	37.1
19	36.8	39.2
20	30.1	39.7

Interval	Midpoint	Number of observations	Frequency [%]	Cumulative frequency
23-26	24.5	1	5	0.05
26-29	27.5	4	20	0.25
29-32	30.5	3	15	0.40
32-35	33.5	6	30	0.70
35-38	36.5	4	20	0.90
38-41	39.5	2	10	1.00

Graphical data description

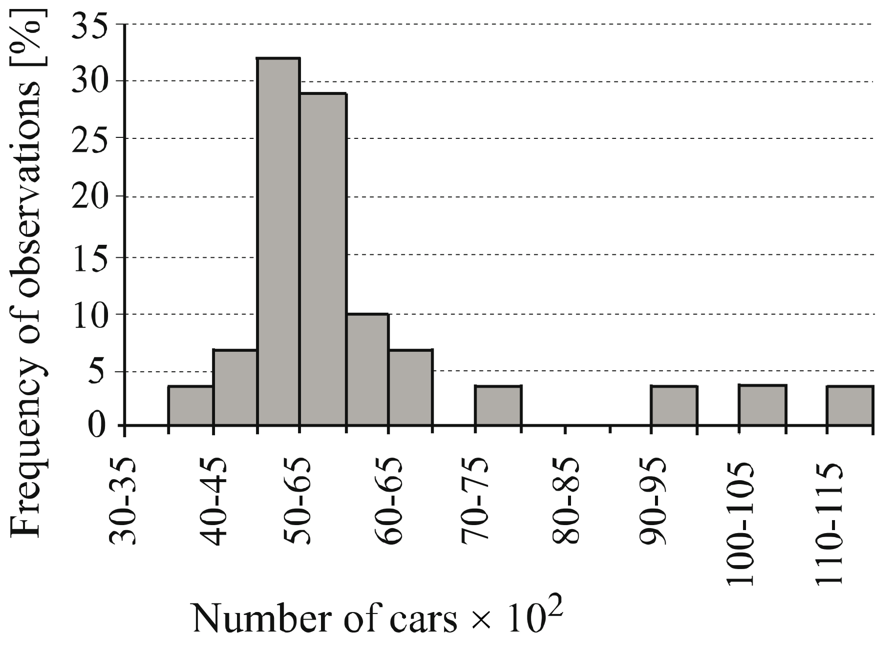
Histogram plot

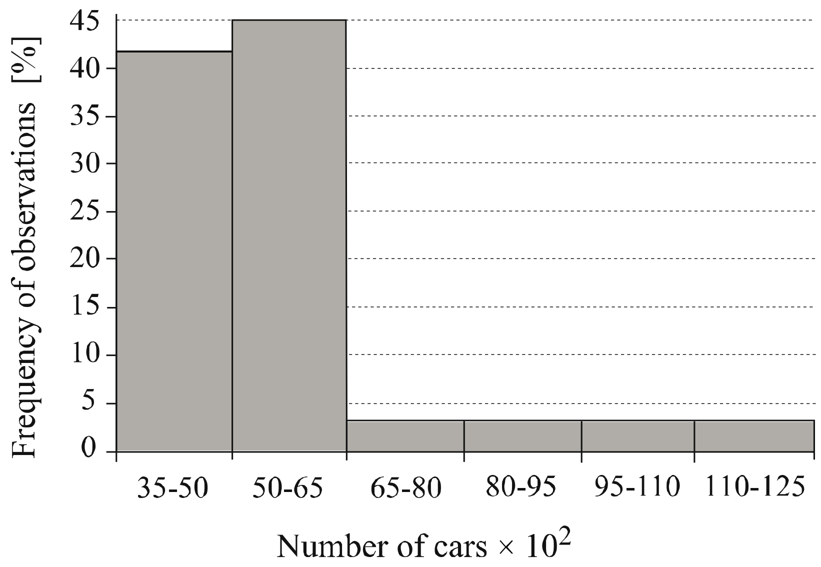




Graphical data description

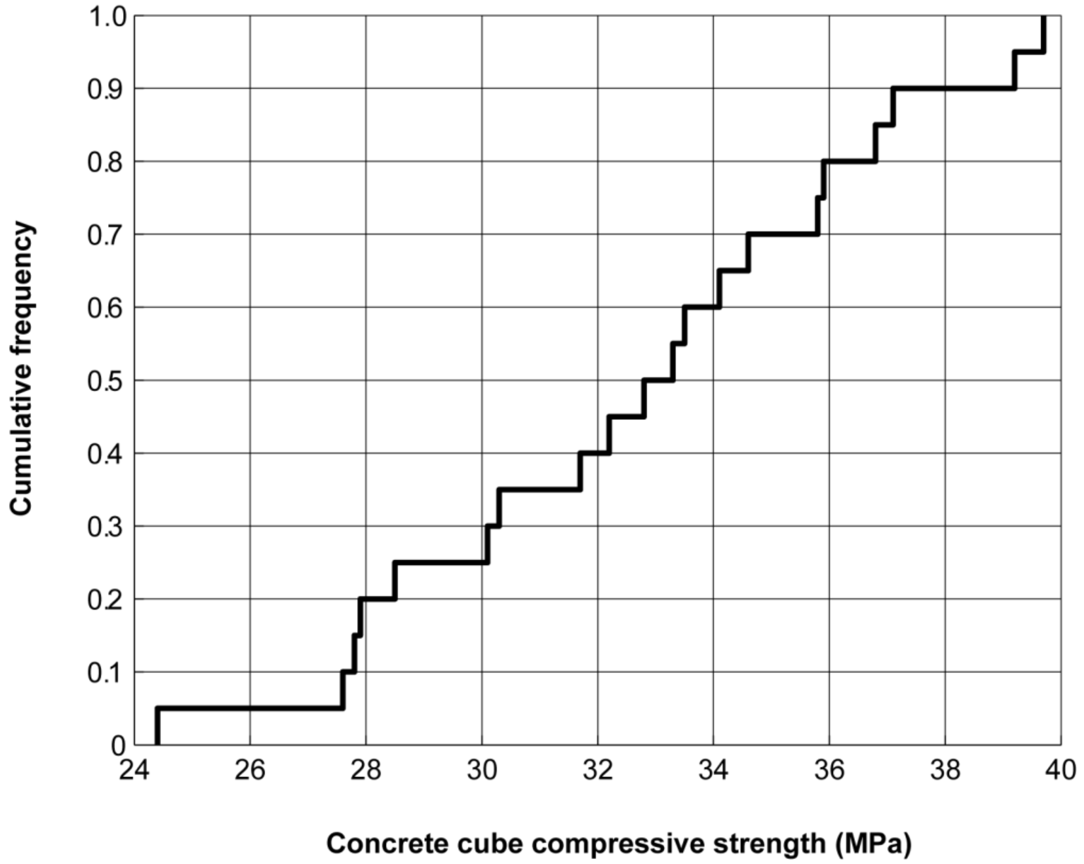
Histogram plots with a different number of intervals





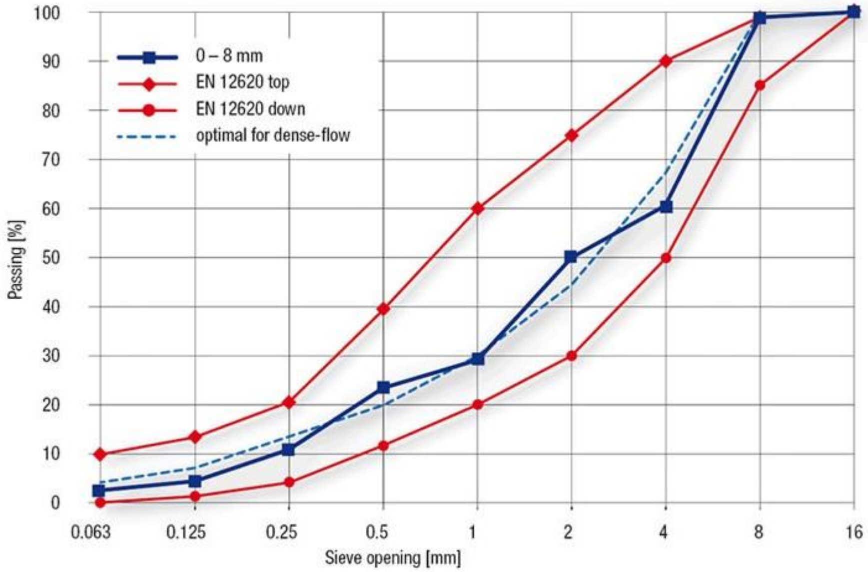
Graphical data description

Histogram plot, cumulative frequency



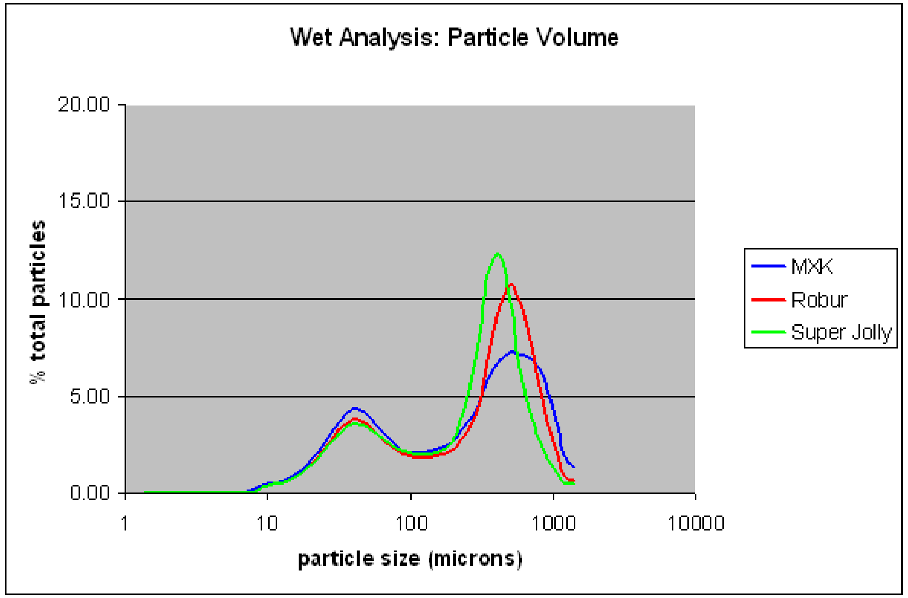
Graphical data description

Further examples: Histogram plot of concrete aggregates



Graphical data description

Further examples: Histogram plot of grinded coffee for espresso



Graphical data description

i	Ordered x_i^o	Q_i
1	24.4	0.048
2	27.6	0.095
3	27.8	0.143
4	27.9	0.190
5	28.5	0.238
6	30.1	0.286
7	30.3	0.333
8	31.7	0.381
9	32.2	0.429
10	32.8	0.476
11	33.3	0.524
12	33.5	0.571
13	34.1	0.619
14	34.6	0.667
15	35.8	0.714
16	35.9	0.762
17	36.8	0.810
18	37.1	0.857
19	39.2	0.905
20	39.7	0.952

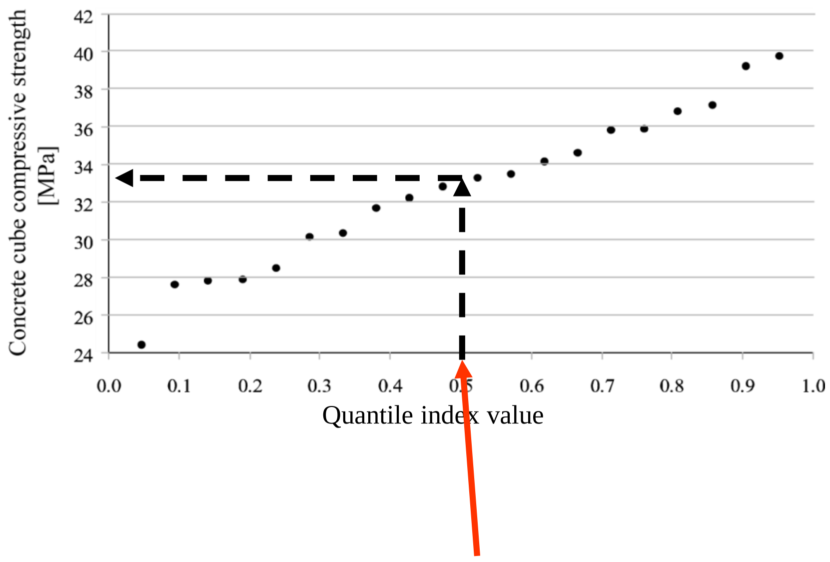
$$v = \frac{i}{1+n}$$

Quantile Plots: A quantile is related to a percentage.

- Example: The 0.65 quantile of a given data set of observations is the observation for which 65% of all observations in the data set have smaller values.

Quantile plots are generated by plotting the ordered data against the respective quantile index values.

Graphical data description



Median = 0.5 quantile value

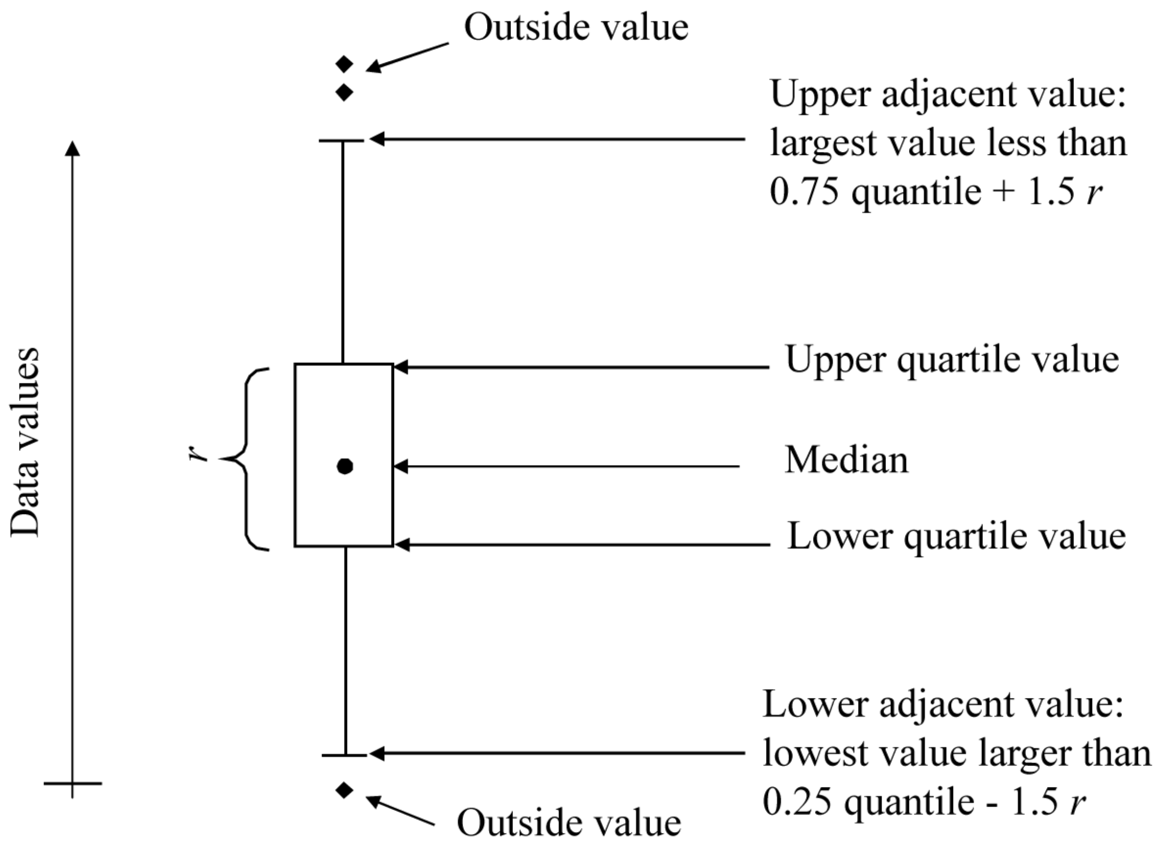
Quantile Plots: A quantile is related to a percentage.

- Example: The 0.65 quantile of a given data set of observations is the observation for which 65% of all observations in the data set have smaller values.

Quantile plots are generated by plotting the ordered data against the respective quantile index values.

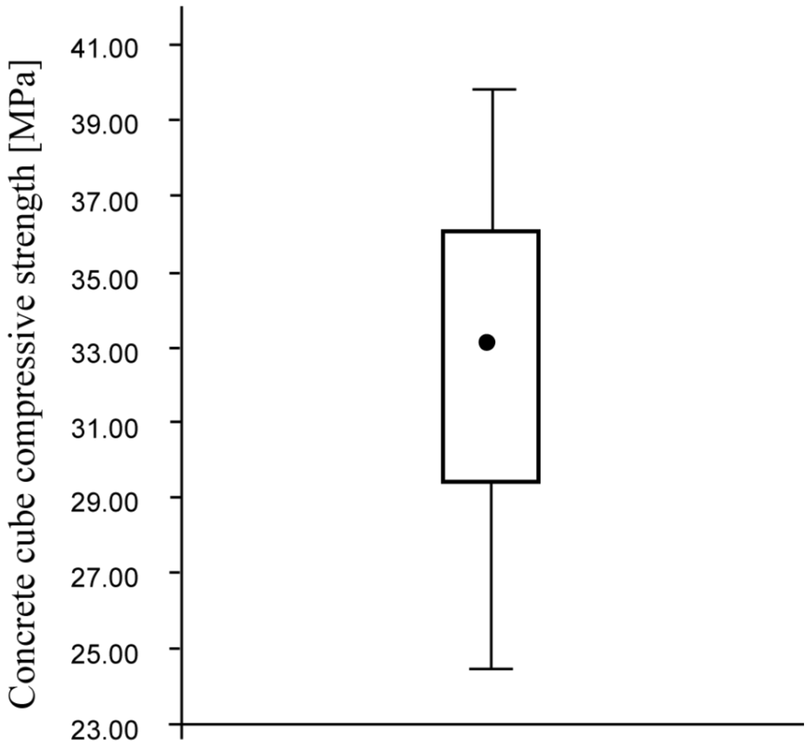
Graphical data description

Tukey Box plots



r : Inter-quartile range (50% of data)

Graphical data description



Tukey Box plot of the concrete compressive strength data

Statistic	Value
Lower quartile	29.30
Lower adjacent value	24.40
Median	33.05
Upper adjacent value	39.70
Upper quartile	35.85

Summary: Graphical data description

A graphical analysis delivers definition to describe the data visually.

One-dimensional scatter plots

- Range and distribution of a dataset along one axis; indicate symmetry and dispersion.

Histograms

- Data distribution over a range; indicate mode, symmetry and dispersion.

Quantile plots

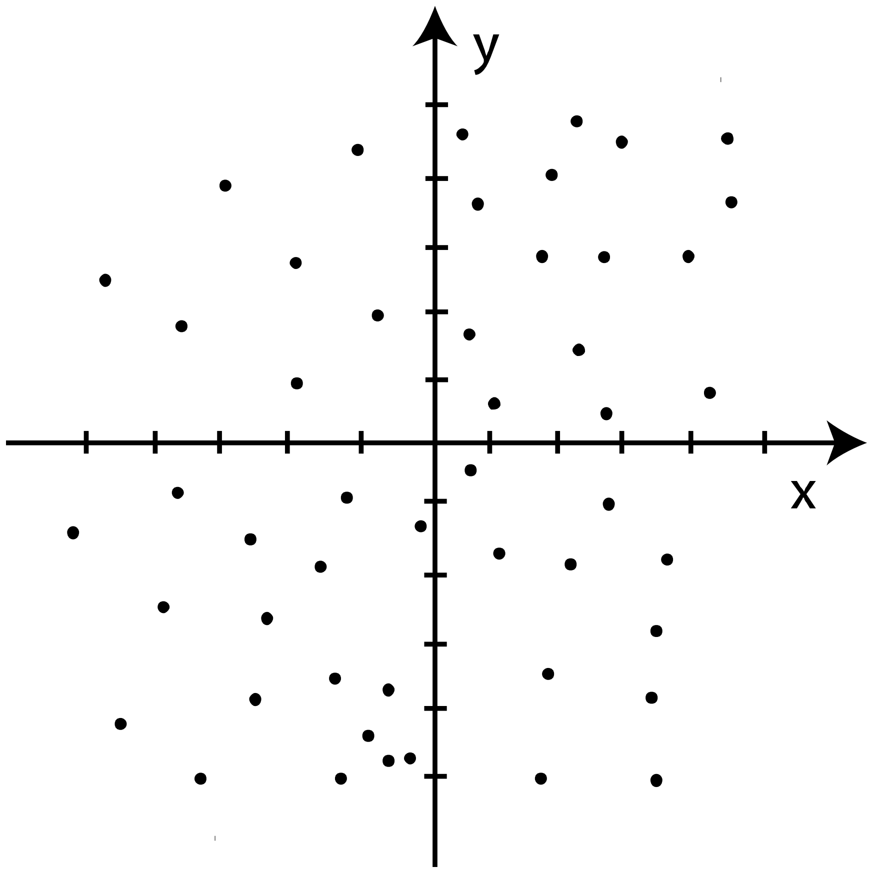
- Median, distribution and symmetry.

Tukey – Box plots

- Median, upper/lower quartiles, symmetry and distribution.

Data pair description

Scatter diagram for visualisation of two properties



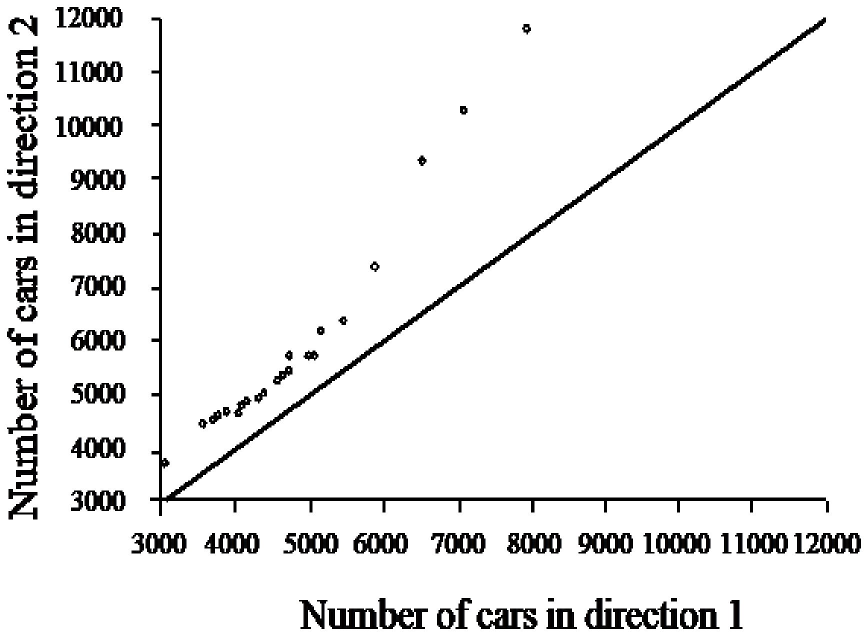
Data pair description

i	Direction 1 ordered $\hat{x}_i^o$	Direction 2 ordered $\hat{x}_i^o$	υ
1	3087	3677	0.0313
2	3578	4453	0.0625
3	3710	4480	0.0938
4	3737	4560	0.1250
5	3906	4635	0.1563
6	4029	4648	0.1875
7	4041	4672	0.2188
8	4085	4757	0.2500
9	4103	4791	0.2813
10	4164	4815	0.3125
11	4323	4880	0.3438
12	4359	4928	0.3750
13	4366	4946	0.4063
14	4368	5005	0.4375
15	4371	5013	0.4688
16	4419	5100	0.5000
17	4563	5220	0.5313
18	4588	5235	0.5625
19	4664	5281	0.5938
20	4667	5318	0.6250
21	4727	5398	0.6563
22	4739	5401	0.6875
23	4741	5679	0.7188
24	5001	5688	0.7500
25	5098	5729	0.7813
26	5193	6183	0.8125
27	5457	6308	0.8438
28	5892	7357	0.8750
29	6551	9323	0.9063
30	7118	10256	0.9375
31	7974	11748	0.9688

Q-Q plot: Data points of the two data sets with the same quantile values are plotted
against each other

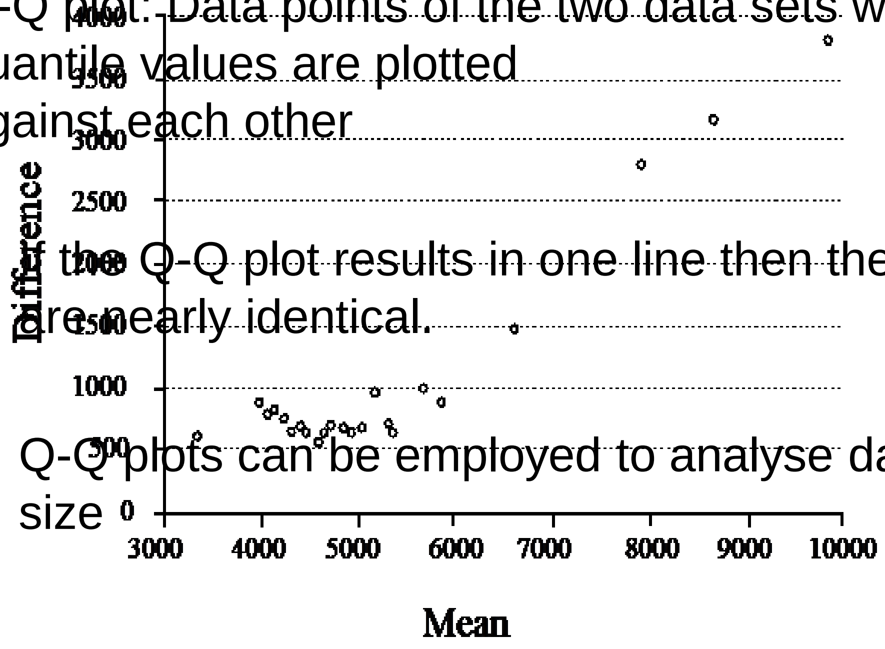
- If the Q-Q plot results in one line then the distributions are nearly identical.
- Q-Q plots may be employed to analyse data with different sample sizes

Data pair description



Q-Q plot: Data points of the two data sets with the same quantile values are plotted against each other

- If the Q-Q plot results in one line then the distributions are nearly identical.
- Q-Q plots can be employed to analyse data of different size



Data pair description

Numerical data set description:

- Sample covariance

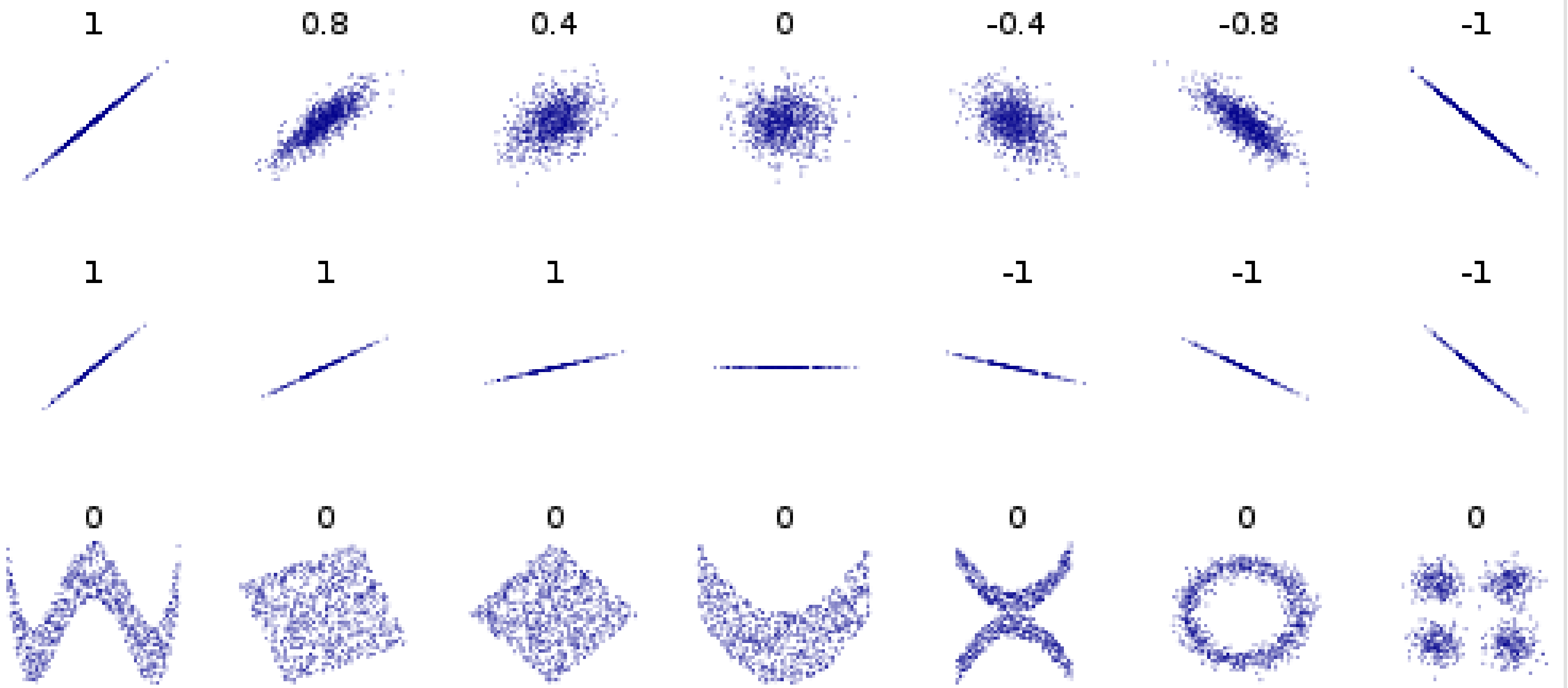
$$c_{XY} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$$

- Coefficient of correlation

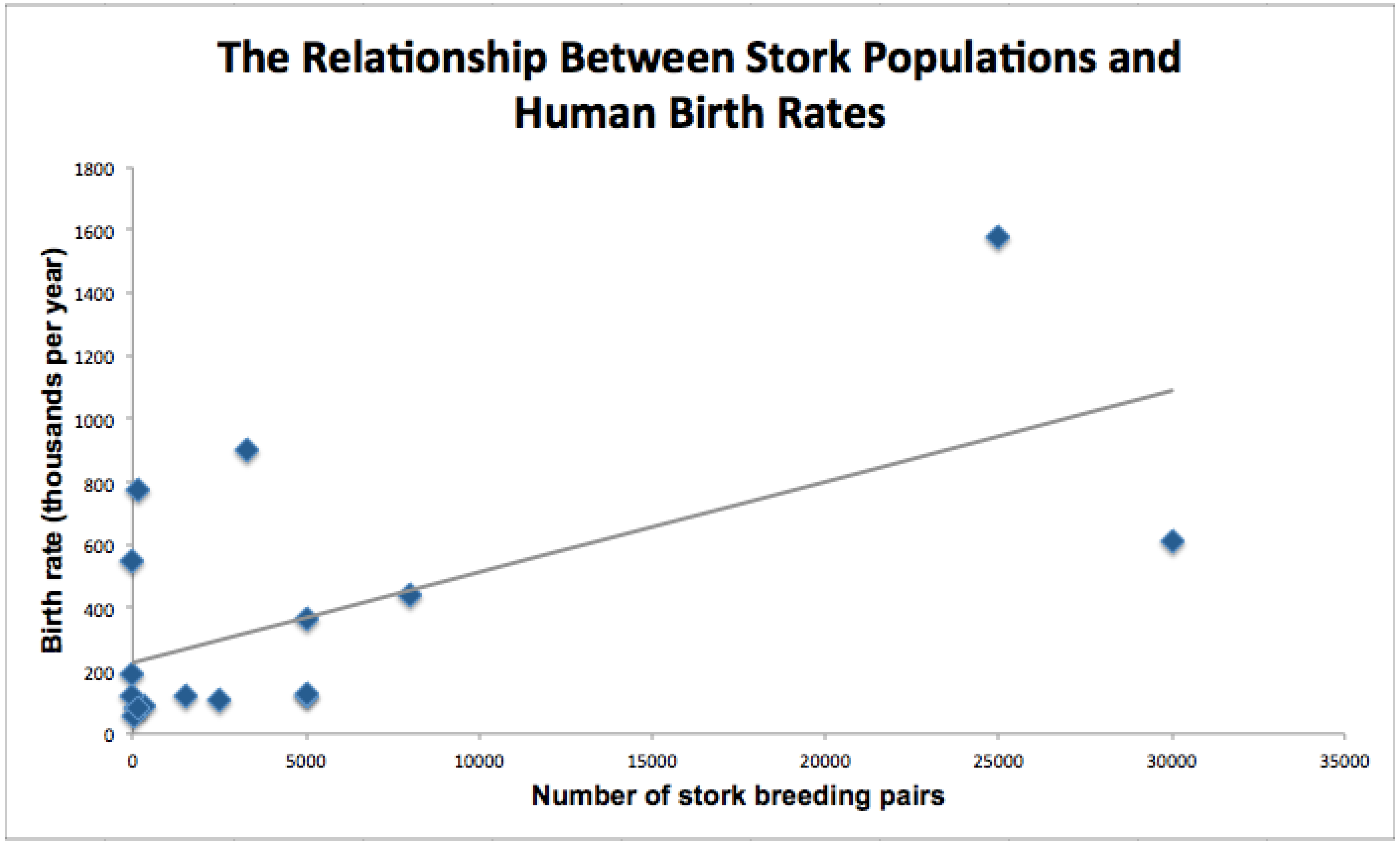
$$r_{XY} = \frac{c_{XY}}{s_X s_Y}$$

Data pair description

Coefficient of correlation and scatter plots

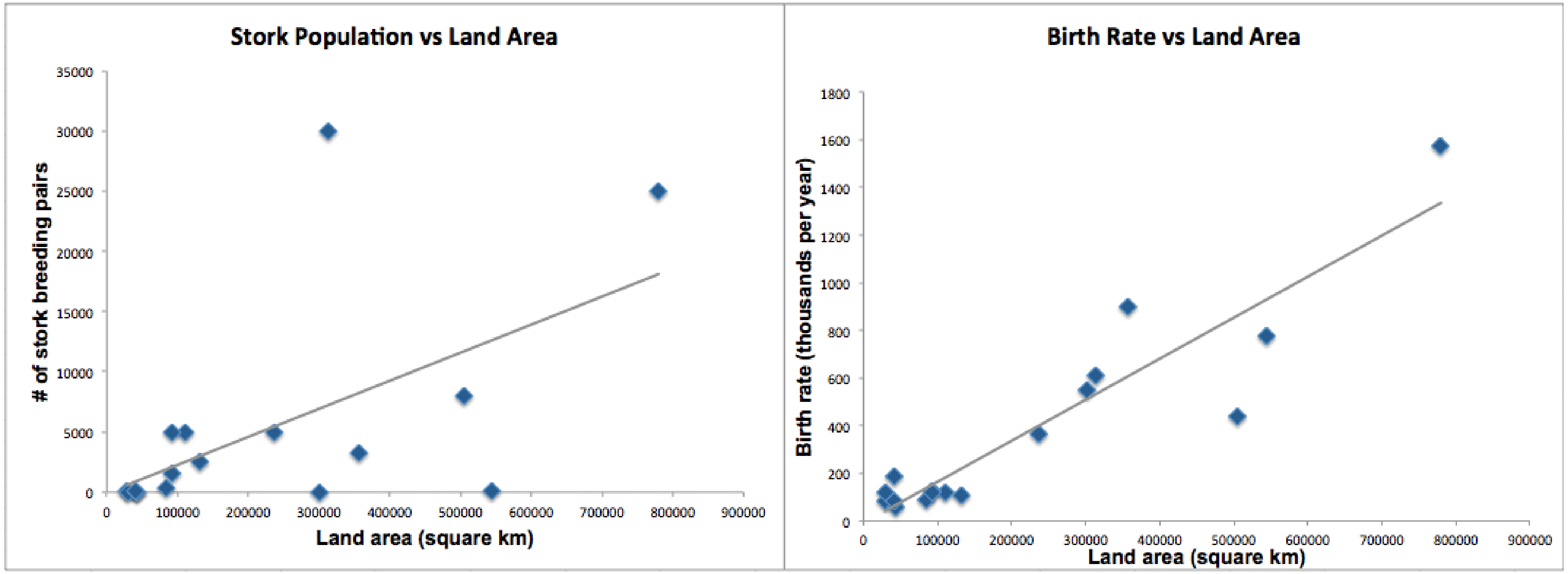


Data pair description



Matthews, R. (2000). Storks Deliver Babies (p= 0.008). Teaching Statistics 22(2): 36-38. DOI: 10.1111/1467-9639.00013.

Data pair description



Summary: Data pair description

Numerical

- Sample covariance
- Coefficient of correlation
  - Does not account for non-linear dependencies!

Graphical

- Scatter plot: To visualize common data distribution
- QQ- Plot to compare distributions

Summary: Elementary data analysis

Elementary data analysis plays an important role in engineering risk and decision analysis.

- Elementary data analysis provides a defined basis to analyse and to document data for understanding and describing uncertainties

- Elementary data analysis formulates the data characteristics in a suitable form
  - For further analyses
  - For communication in a defined format to other professionals

Elementary data analysis: Task

Self Assessment Questions of Chapter 3: Descriptive Statistics.

Faber, M. H. (2012). Statistics and Probability Theory in Pursuit of Engineering Decision Support, Springer. ISBN: 978-94-007-4055-6.

Thank you for your attention.

COST TU1402: Quantifying the Value of Structural Health Monitoring
www.cost-tu1402.eu

DTU

ERDA
Engineering Risk and
Decision Analysis