Universiteit Utrecht

Stochastic Hydrology

# Descriptive statistics

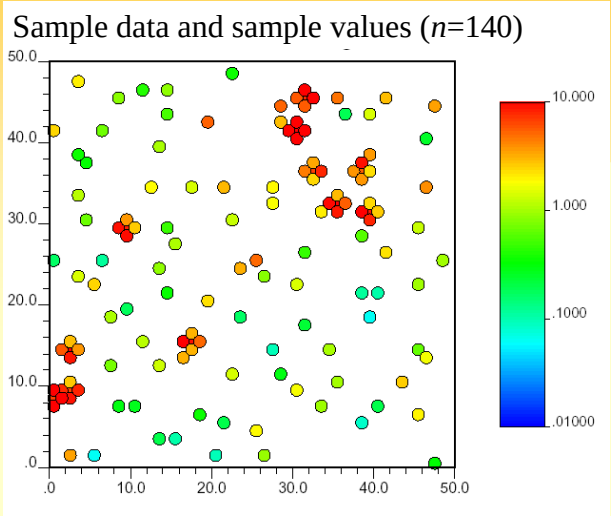
Marc F.P. Bierkens

*Professor of Hydrology*
*Faculty of Geosciences*

Universiteit Utrecht

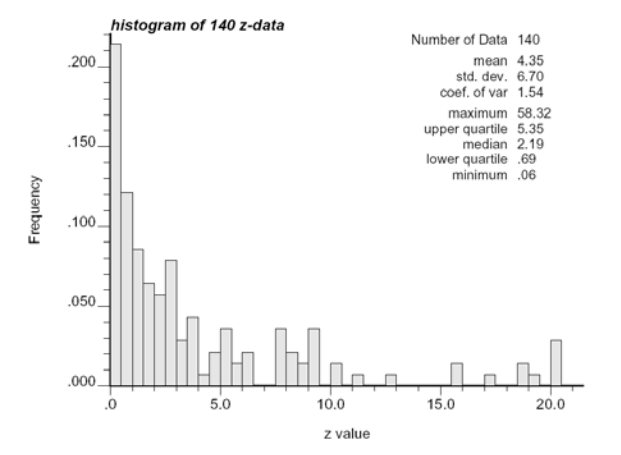
## Univariate statistics

Sample data and sample values ($n$=140)



Universiteit Utrecht

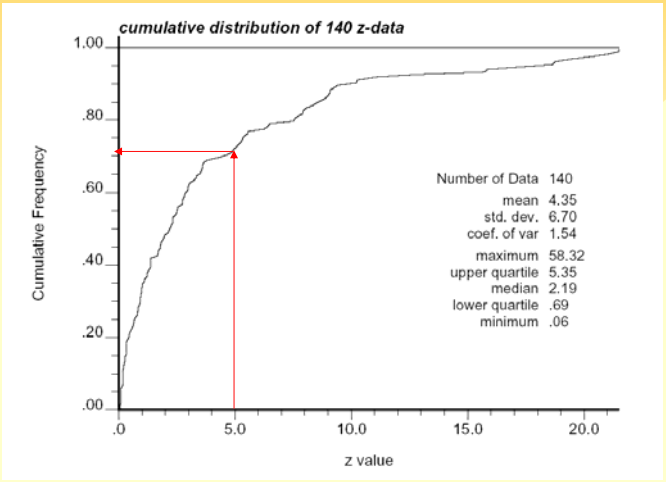
## Histogram or frequency distribution




Universiteit Utrecht

## Cumulative frequency distribution

$z_i$ versus $\frac{i}{n+1}$ (with $z_i$ in ascending order)

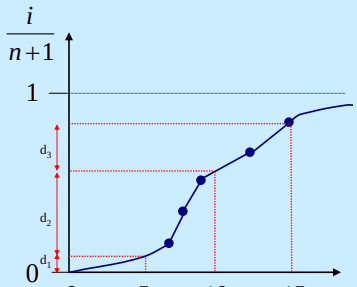


Universiteit Utrecht

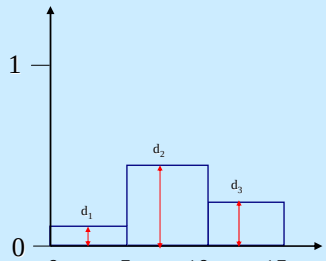
## Relation between cfd and histogram

10 7 9 8 15 : $n = 5$

i : 4 1 3 2 5




Relation with histogram




Universiteit Utrecht

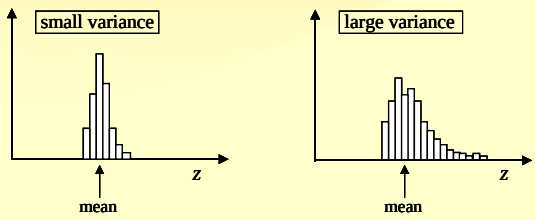
## Statistical measures

*Mean*: measure of locality

$$m_z = \frac{1}{n}\sum_{i=1}^{n} z_i$$

*Variance*: measure of spread

$$s_z^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - m_x)^2 = \frac{1}{n}\sum_{i=1}^{n} z_i^{\,2} - m_z^2$$



Universiteit Utrecht

## Statistical measures

*Standard deviation*

$$s_z = \sqrt{s_z^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_i - m_x)^2}$$

*Coefficient of variation*

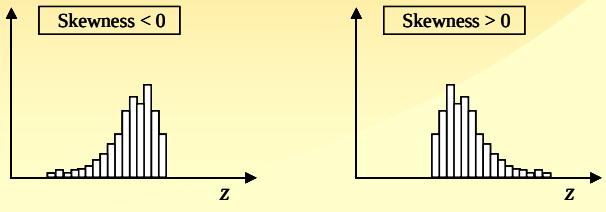
$$CV_z = \frac{s_z}{m_z}$$

Universiteit Utrecht

## Statistical measures

*Skewness* (measure of form)

$$CS_z = \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i - m_z)^3}{s_z^3}$$

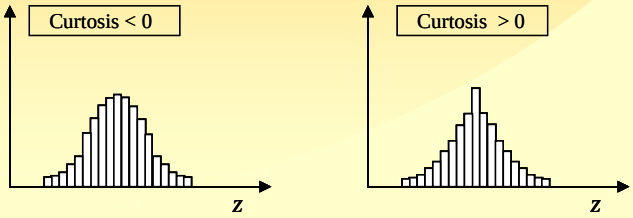


Universiteit Utrecht

## Statistical measures

*Curtosis* (measure of form)

$$CC_z = \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i - m_z)^4}{s_z^4} - 3$$

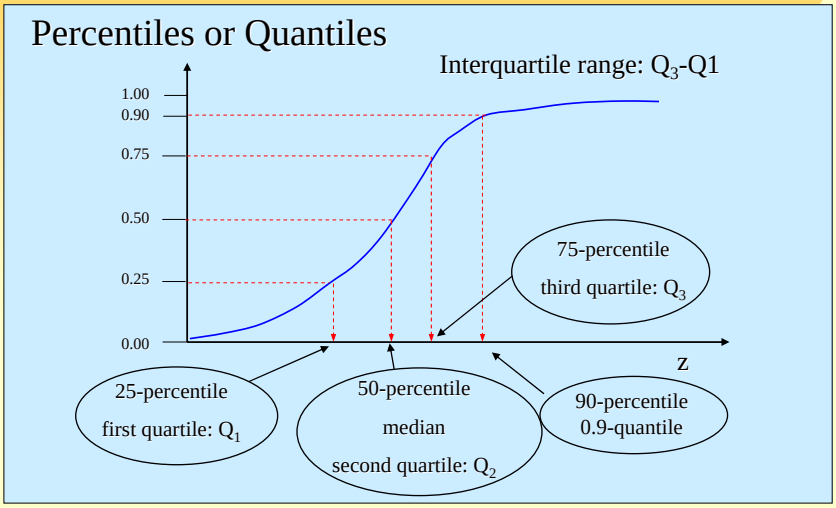



Universiteit Utrecht

## Statistical measures

Other measures of locality and spread

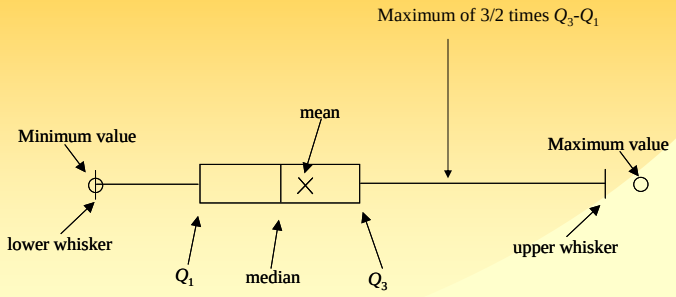
### Percentiles or Quantiles



Universiteit Utrecht

## Statistical measures

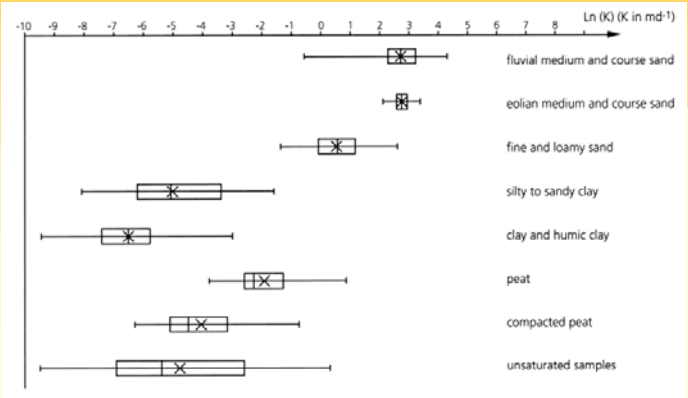
### Box-and-whisker plot




Universiteit Utrecht

## Statistical measures

### Box-and-whisker plot: example




Bierkens (WRR 32(8), 1996)

# Bivariate statistics

*Covariance*: measure of co-variation

$$C_{zy} = \frac{1}{n}\sum_{i=1}^{n}(z_i - m_z)(y_i - m_y) = \frac{1}{n}\sum_{i=1}^{n} z_i y_i - m_z m_y$$

*Correlation coefficient*: standardized measure of co-variation

$$r_{zy} = \frac{C_{zy}}{s_z s_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i - m_z)(y_i - m_y)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_i - m_z)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - m_y)^2}}$$

Universiteit Utrecht

Statistical measures

*Correlation coefficient*: easy way of calculation

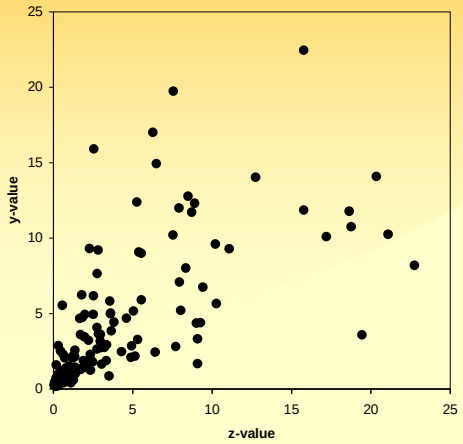
$$r_{zy} = \frac{n\sum_{i=1}^{n} z_i y_i - \sum_{i=1}^{n} z_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} z_i^2 - \left(\sum_{i=1}^{n} z_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

calculate $\Sigma z_i$, $\Sigma y_i$, $\Sigma z_i^2$, $\Sigma y_i^2$ and $\Sigma z_i y_i$

Universiteit Utrecht

## Statistical measures

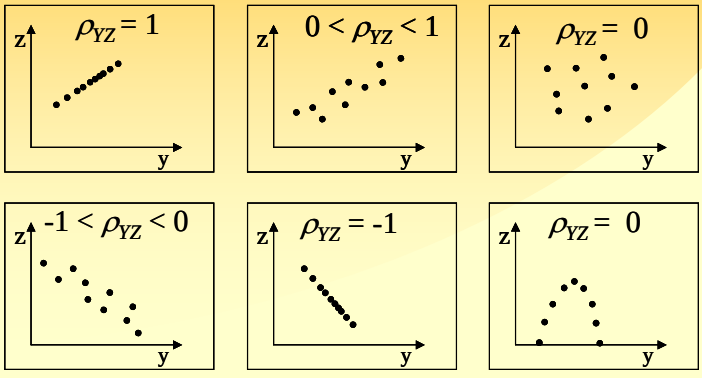
*Correlation coefficient:* 0.57




Universiteit Utrecht

## Statistical measures

*Correlation coefficient* = <u>linear</u> dependence



Universiteit Utrecht

# Descriptive statistics: exercises

Consider the following data set:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *n* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| *z* | 1.7 | 6.26 | 7.56 | 7.92 | 0.96 | 2.47 | 2.55 | 0.28 | 1.34 | 0.71 |
| *y* | 1.3 | 17.02 | 19.74 | 12.01 | 0.66 | 1.8 | 15.91 | 0.62 | 2.15 | 2.07 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *n* | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| *z* | 1.66 | 2.99 | 8.71 | 0.09 | 0.62 | 0.99 | 10.27 | 2.96 | 5.54 | 3.61 |
| *y* | 4.68 | 2.74 | 11.72 | 0.24 | 2.3 | 0.52 | 5.67 | 3.17 | 5.92 | 5.03 |

1. Make a histogram of *z* with class-widths of 5 units. What fraction of the data has values between 5 and 10?
2. Construct the cumulative distribution of *z* and *y*
3. Calculate the mean, the variance, the skewnes, the quantiles, the medium and the interquantile range of *z* and *y*.
4. Draw a box-and-whisker plot of the z- and y-values. Are there any possible outliers?
5. Suppose that *z* is the concentration of some pollutant in the soil (mg/kg). Suppose that the samples have been taken randomly from the site of interest. If the critical concentration is 5 mg/kg and the site is 8000 $m^2$. Approximately what area of the site has be cleaned up?
6. Calculate the correlation coefficient between *z* and *y*?
7. What fraction of the data has a z-value smaller than 5 *and* a y-value smaller than 10?
8. What fraction of the data has a z-value smaller than 5 *or* a y-value smaller than 10?