Exploratory data analysis

Modelling in physical geography, 5cr, 30.10.-30.11.2017

Introduction

- Exploratory data analysis is an approach to analyze data in order to *summarize their main characteristics*
- **Get to know your data**
- *Formulate hypothesis* about causes of observed phenomena
- *Assess assumptions* on which statistical modelling will be based
- *Support the selection* of appropriate statistical tools and techniques
- Provide a basis for further data collection
- Approaches: graphical (histograms, boxplots, scatter plots, …), quantitative (e.g. ordination)

Exploratory analysis in R

- *summary()* –function returns five quantiles of each variables: minimum, lower quartile (25 %), median (50 % quartile), upper quartile (75 %), maximum and mean
- These statistics will provide a first idea of the distribution of the values

```
> summary(d$soil_moist)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  11.20   24.50   34.15   34.31   44.58   65.50
```

Mean and variance

- Arithmetic mean (μ) is the common *measure of central tendency*
- It's obtained by summing observed values and dividing the sum by the number of individual observations

> mean(soil_moist) # mean of soil_moist –variable

$$A = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

[1] 34.3

- Prefer *median* in the presence of outliers and/or with skewed data
- Variance (σ^2) is a *measure of deviation*, defined as the average of the squared differences from the mean $Var(X) = E[(X - \mu)^2]$

> var(soil_moist) # variance

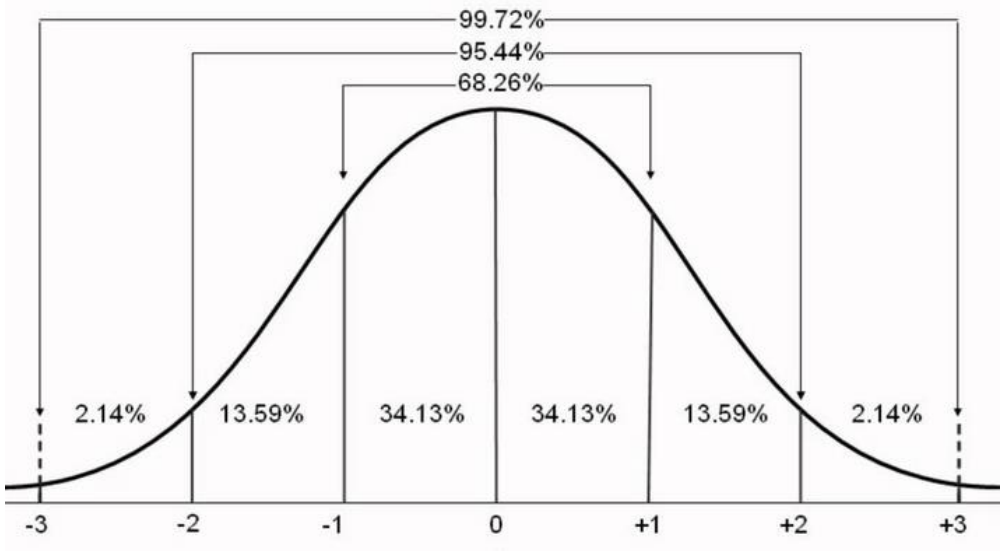
[1] 166.4

Standard deviation

- Most commonly used measure of deviation
- Standard deviation (σ) is the square-root of variance (σ^2)
- σ measures *the average spread* of observed values from their population mean

```
> sd(soil_moist) # standard deviation
[1] 12.9
> sqrt(var(soil_moist))
[1] 12.9
```

- Small σ indicates that values tend to be close to the mean and vice versa



σ is a useful measure of data spread, when random variable follows normal distribution

Range of variation

- *Range of variation* (range) is simply the minimum and maximum value of a variable
- Highly sensitive to extreme values

```
> range(soil_moist) # returns the lowest and the highest value of soil moisture
[1] 11.2 65.5
```

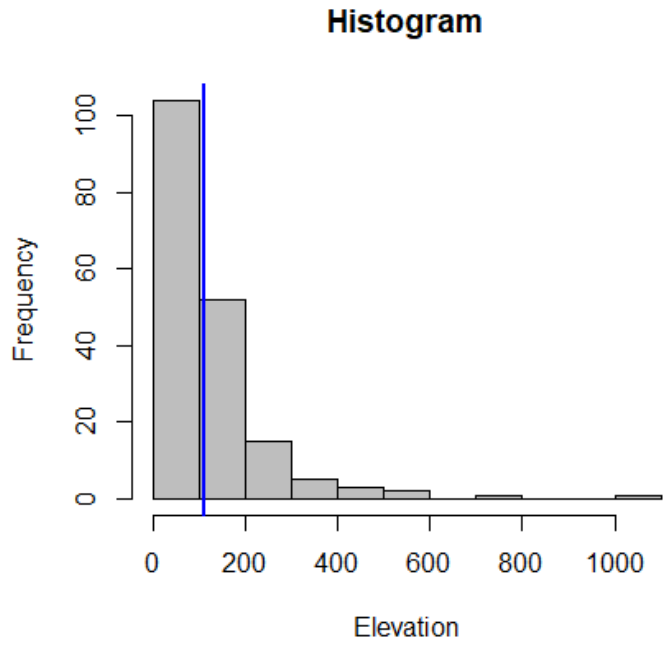
- *The length of a range* expresses the difference between maximum and minimum values, obtained in R using max() and min() -functions

```
> max(soil_moist) – min(soil_moist)
[1] 54.3
```

Histograms

- Fast and simple way to get an idea how your data is distributed
- Range of values are divided to series of intervals, and then count how many values fall into each interval
- Function *hist()* in R

```
> hist(d$elev, main = "Histogram", xlab="Elevation",
        col = "grey")
> abline(v=mean(d$elev), col = "blue", lwd=2)
```

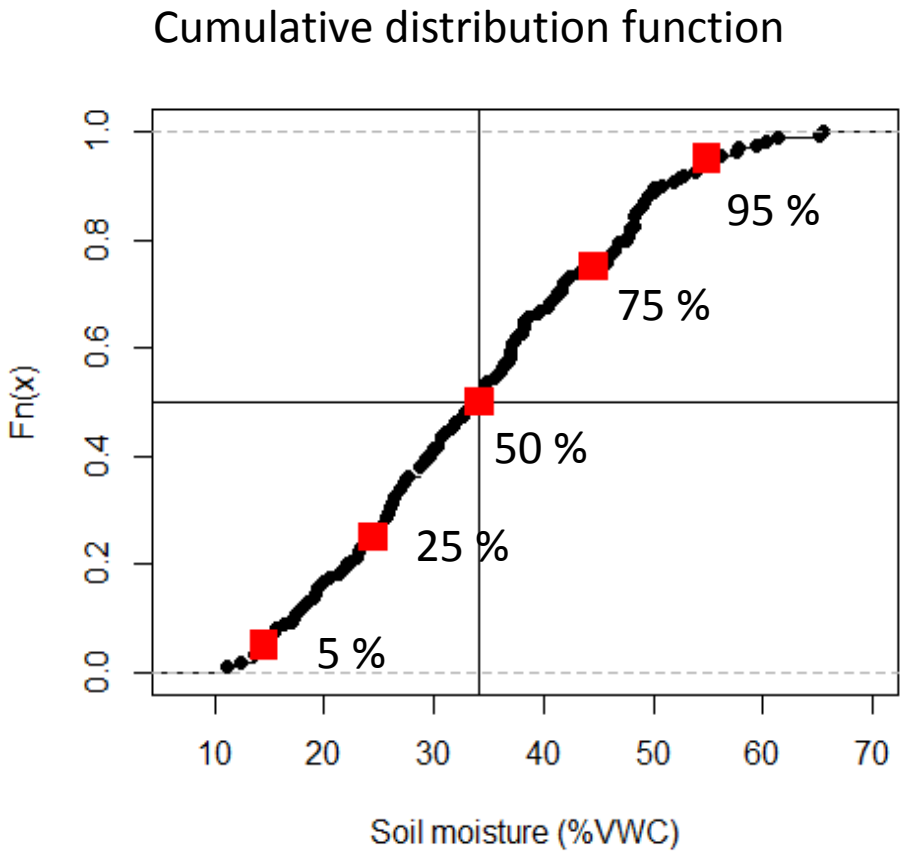


Quantiles

- Quantiles are percent points from a variable's *cumulative distribution function*
- Cutpoints for dividing data into contigous intervals with equal probability
- Some common terms:
  - percentiles (100-quantiles)
  - quantiles (10-quantiles)
  - quartiles (4-quantiles)

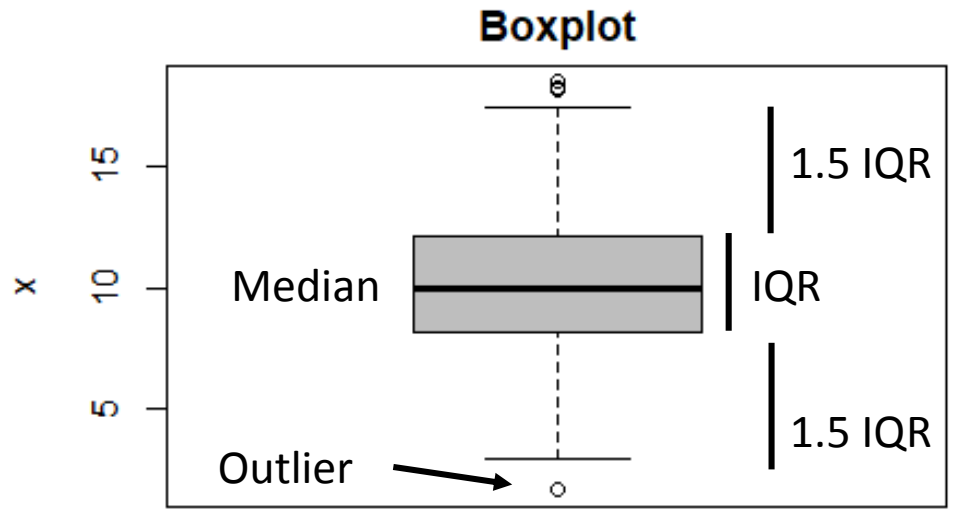
> quantiles(soil_moist, probs=seq(0, 1, 0.1))

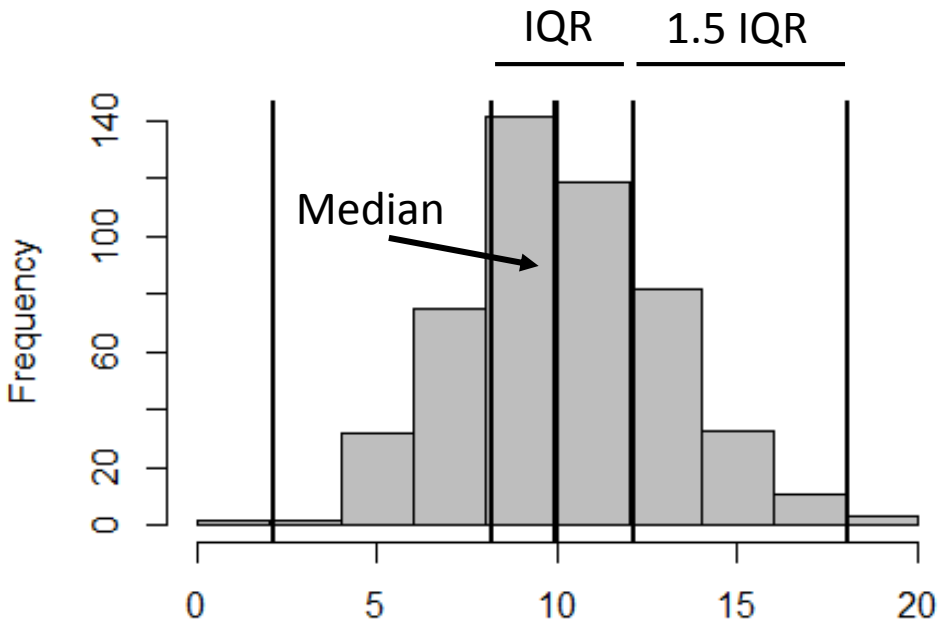
0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
11.20	17.38	22.42	26.00	29.72	34.15	37.20	41.39	47.56	50.91	65.50



Boxplots

- Powerful way to display numerical data through their quartiles
- The black line depicts median
- The *box* indicate interquartile range (*IQR*), stretching from first (25 %) to third quartile (75 %)
- In R default, the whiskers represent 1.5 IQR
- Data not falling within the whiskers are *outliers*
- In R boxplots are easy to do using *boxplot()* - function

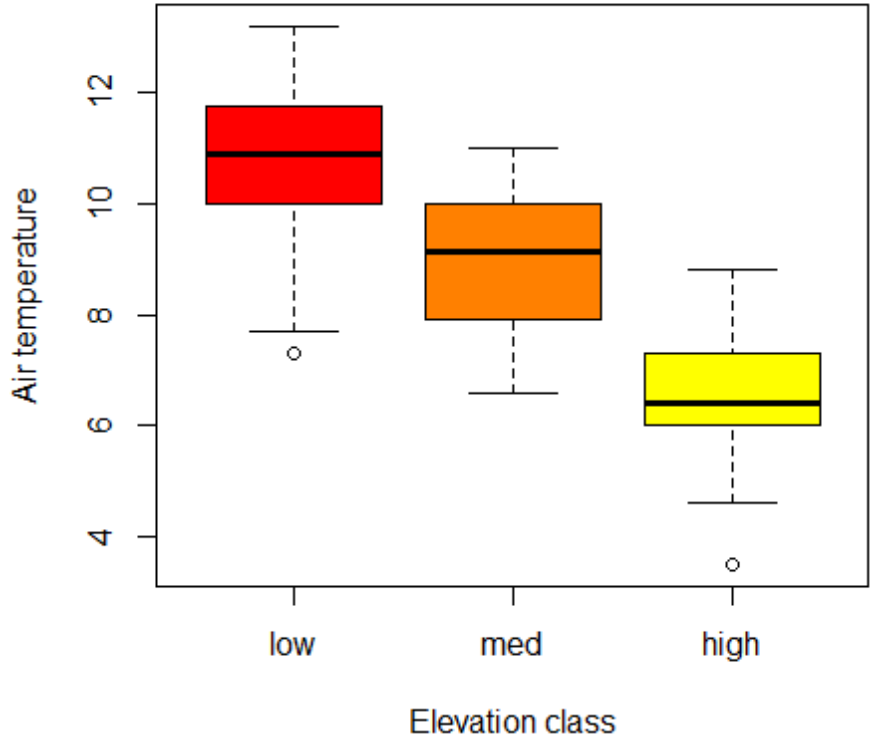




Group statistics

- Use *tapply()* –function to produce statistics for different groups
- In this example, let's summarize the variation in air temperature over three elevation classes (low=0-100m, med=101-200, high=201 -)

```
> tapply(d$temp, d$elev_class, mean)
[1] low  med high
    10.8  9.0   6.6
> boxplot(d$temp~d$elev_class, ylab="Air
          temperature", xlab="Elevation class",
          col=heat.colors(3))
```

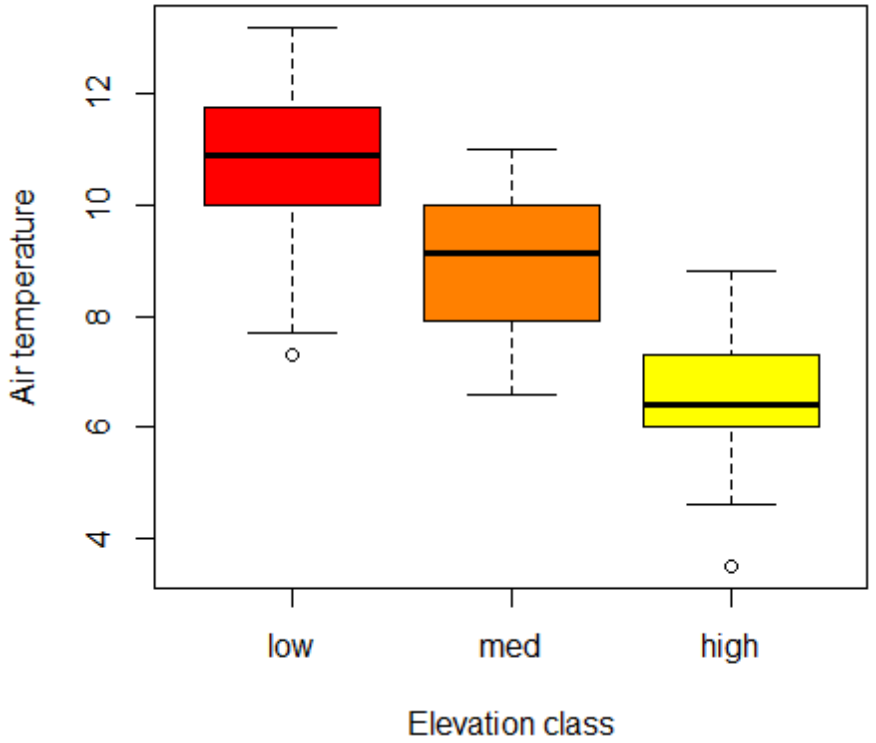


Comparing groups

- Test for differences between two groups of same variable
- Can the difference between the means of the groups emerge due chance alone?
- Two-sample t-test agains *null hypothesis*; there is no difference between the means of ”low” and ”med”

```
> t.test(d$temp[d$elev_class=="low"],
         d$temp[d$elev_class=="med"])$p.value
[1] 3.903991e-13
```

- Mann-Whitney’s U-test is a non-parametric equivalent for t-test, function *wilcox.test()*



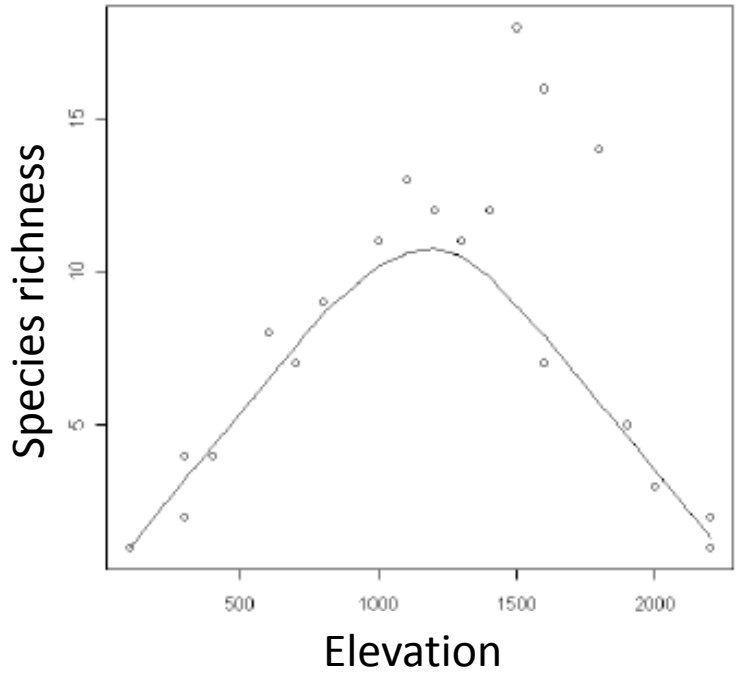
$p < 0.001$, difference is highly significant, the difference in means between ”low” and ”med” is *not likely to be a result of chance*

Statistical depency and correlation

- Statistical depency means that variables co-vary; for example as a tree is getting taller, it's diameter increases
- Pairwise covariance implies that variables values vary similarly in the data

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

- **Statistical depency does not mean causality!**
- The stronger the depency, the *more likely* is that the variables have a causal relationship
- In the example, species richness and elevation are weakly correlated, but strongly associated. However, they are hardly causally linked to each other



Correlation coefficient

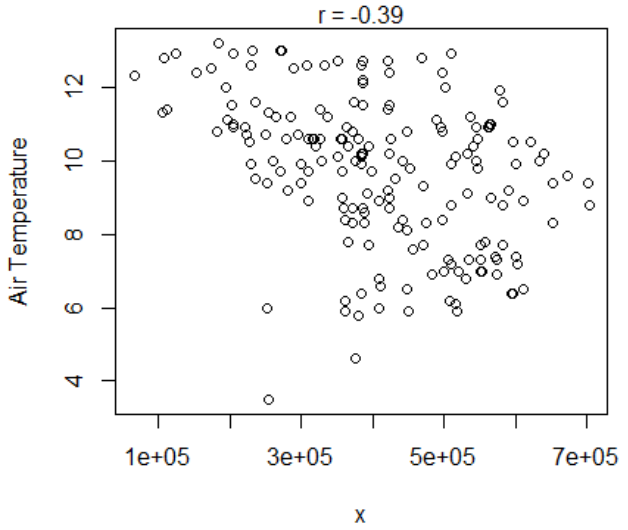
- Examine potential for pairwise dependencies using a scatter plot (*plot* -function)
- Correlation coefficient (r) is normalized covariance: they are always between -1 and 1

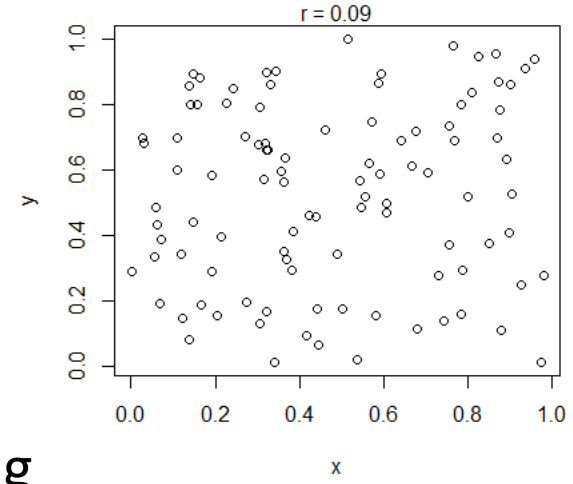
> cov(d$temp, d$elev)

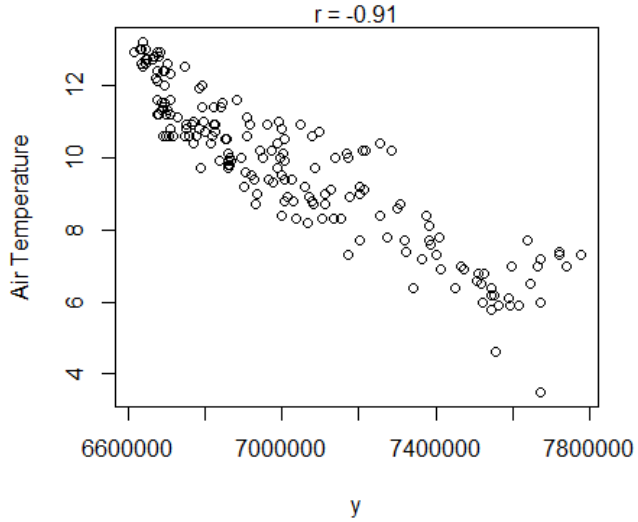
[1] -206.7

> cor(d$temp, d$elev)

[1] -0.76



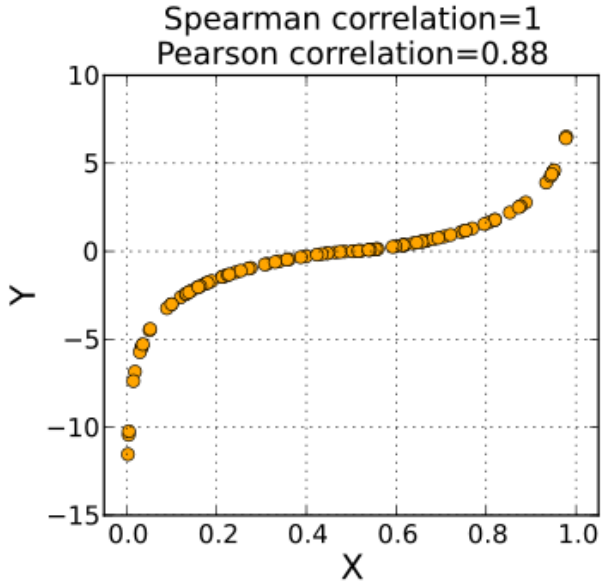


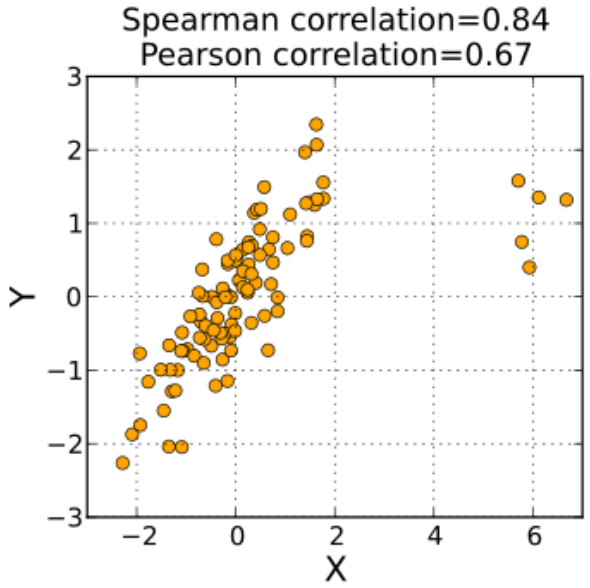


- Coefficient close to [1] means strong correlation, close to zero weak correlation

Correlation coefficient

- Pearson's correlation coefficient r_p for testing *linear* relationships with variable following normal distribution (*parametric*)
- Spearman's (r_s; and Kendall's tau, r_k) measures *monotonic* association when variables are not normally distributed (*non-parametric*)
- r_s uses *pairwise ranks* to quantify statistical association and it's not restricted to linear relationships
- r_s is less sensitive to outliers in the data than r_p





Correlation in R – functions

Pearson (normal distribution)
> cor.test(d$temp, d$lake)

Spearman (non-normal distribution)
> cor.test(d$temp, d$lake, method="spearman")

```
        Pearson's product-moment correlation

data:  d$temp and d$lake
t = -0.085165, df = 181, p-value = 0.9322
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1512479  0.1388540
sample estimates:
         cor
-0.006330139
```

```
        Spearman's rank correlation rho

data:  d$temp and d$lake
S = 1334600, p-value = 2.418e-05
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.3066677
```

Spearman's correlation coefficient implies highly significant ($p < 0.001$) relationship between air temperatures and proximity to lakes

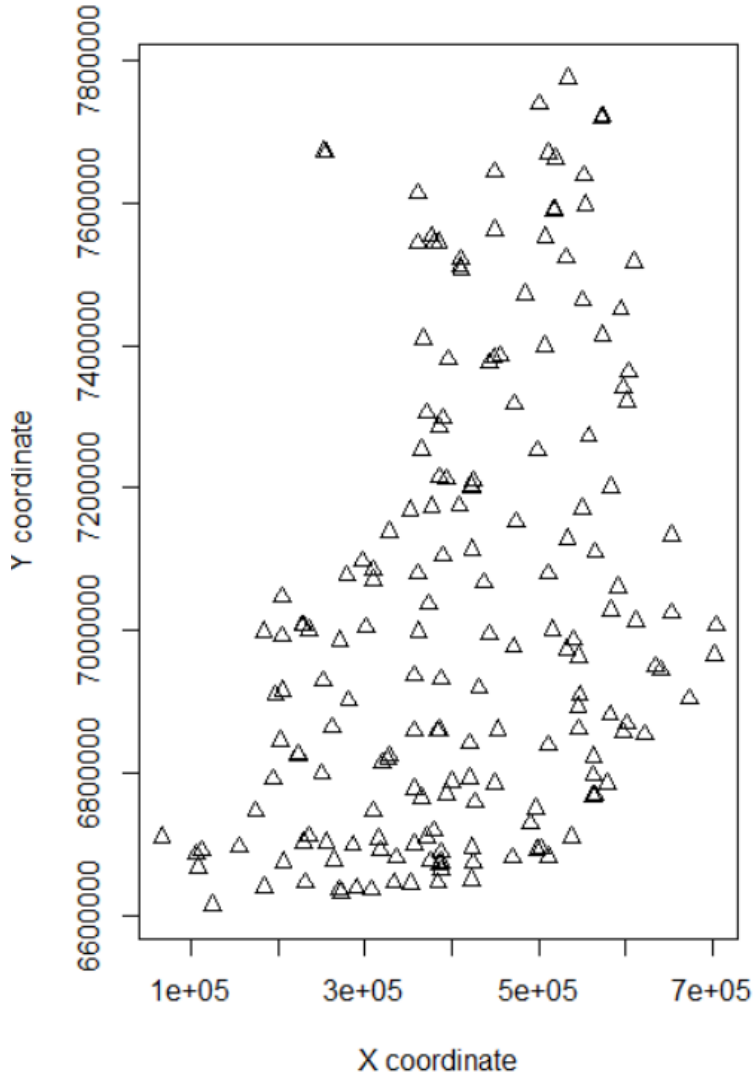
Missing values in R

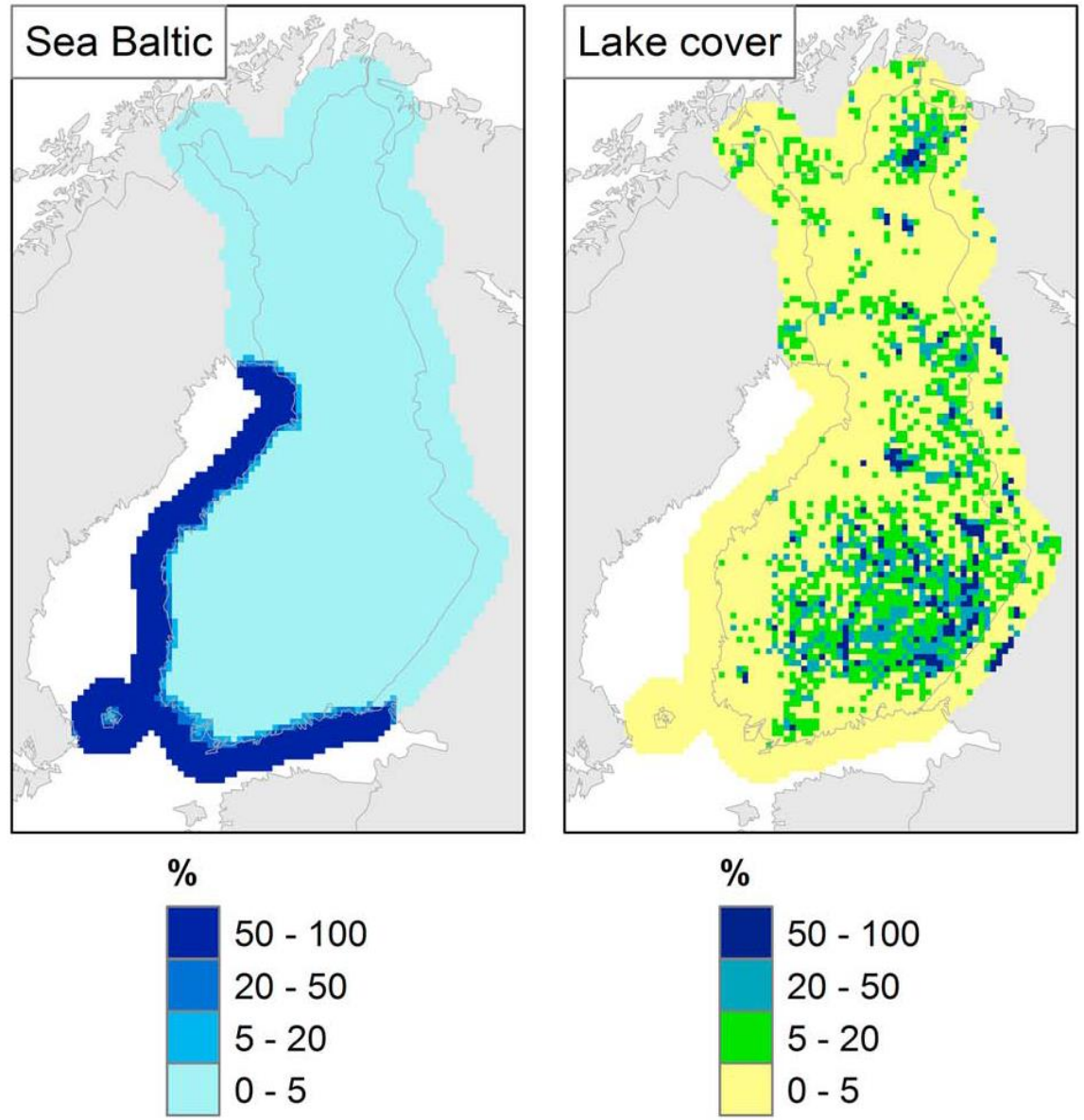
- In R missing valus are represented by the symbol *NA* (not available)
- **It’s not the same as zero!!**

```
> x <- c(1, 2, NA, 3)
> mean(x)
[1] NA
> mean(x, na.rm = TRUE)
[1] 2
> y <- c(2, 4 ,5 , 6)
> cor(x, y) # returns NA
> cor(x,y, use = ”complete.obs”)
[1] 1
```

Practical – exploratory data analysis in R

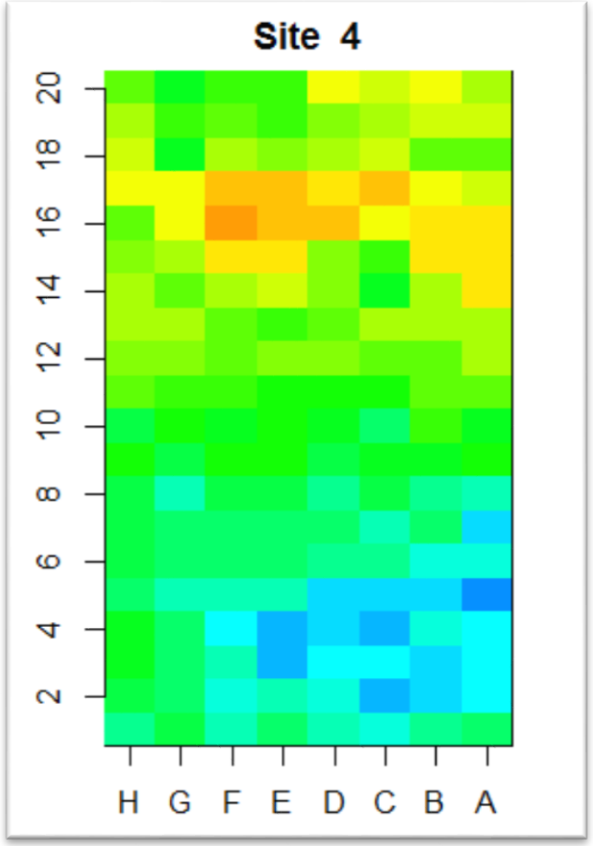
- R-script ”*Exploratory_Analysis_Practical.pdf*”
- ”*AirTemperatureData.csv*”, average air temperatures
- pvm = year and month
- station = FMI station id
- **Temp = air temperature (°C)**
- x and y = geographical location (Euref fin)
- elev = elevation above sea level (m)
- lake = lake percentage (%), see next slide!
- sea = sea percentage (%), see next slide!




Sea Baltic
%
50 - 100
20 - 50
5 - 20
0 - 5
Lake cover
%
50 - 100
20 - 50
5 - 20
0 - 5

Questions – data description

- ”*SaanaSoilMoisture.csv*”
- soil_moist = soil moisture (% VWC) inside a 1m2 study-plot measured in Mt. Saana Kilpisjärvi
- mesotopo = measures of local topography (1 = valley bottom, 5 = mid slope, 10 = ridge top)
- veg_height = maximum vegetation height (cm) at each plot
- veg_cover = total vegetation cover (%) at each plot



Questions

1. Summarize the variation in (i) soil moisture and (ii) vegetation cover (mean, median, standard deviation, range of variation). Plot distributions of both variables using histograms in the same figure and add lines indicating mean values.
2. Let’s assume that 2.5 % and 97.5 % percentiles of soil moisture characterize extremely dry and wet soil conditions, respectively. What are the corresponding soil moisture values? Calculate the length of 95 % range of variation.
3. Create two groups of topography (”valley” and ”ridge”) based on ”mesotopo” –variable, using mesotopo value of 5 as a cut-off. Plot soil moisture variation in both ”valley” and ”ridge”. Calculate mean and standard deviation of soil moisture over the two groups. Is the difference between the means statistically significant, according to t-test (at 0.05 significance level)? What about for vegetation height?
4. Examine pairwise relationships by calculating correlation matrices for variables ”mesotopo”, ”soil_moist”, ”veg_height” and ”veg_cover”, based (i) Pearson’s and (ii) Spearman’s correlation coefficients.
5. Plot soil moisture againts all other numerical variables organized as a 2 x 2 figure matrix. Add Spearman’s correlation (**if significant, $p < 0.05$**) to each of the plot at two decimal precision, using functions such as *text()* or *mtext().*