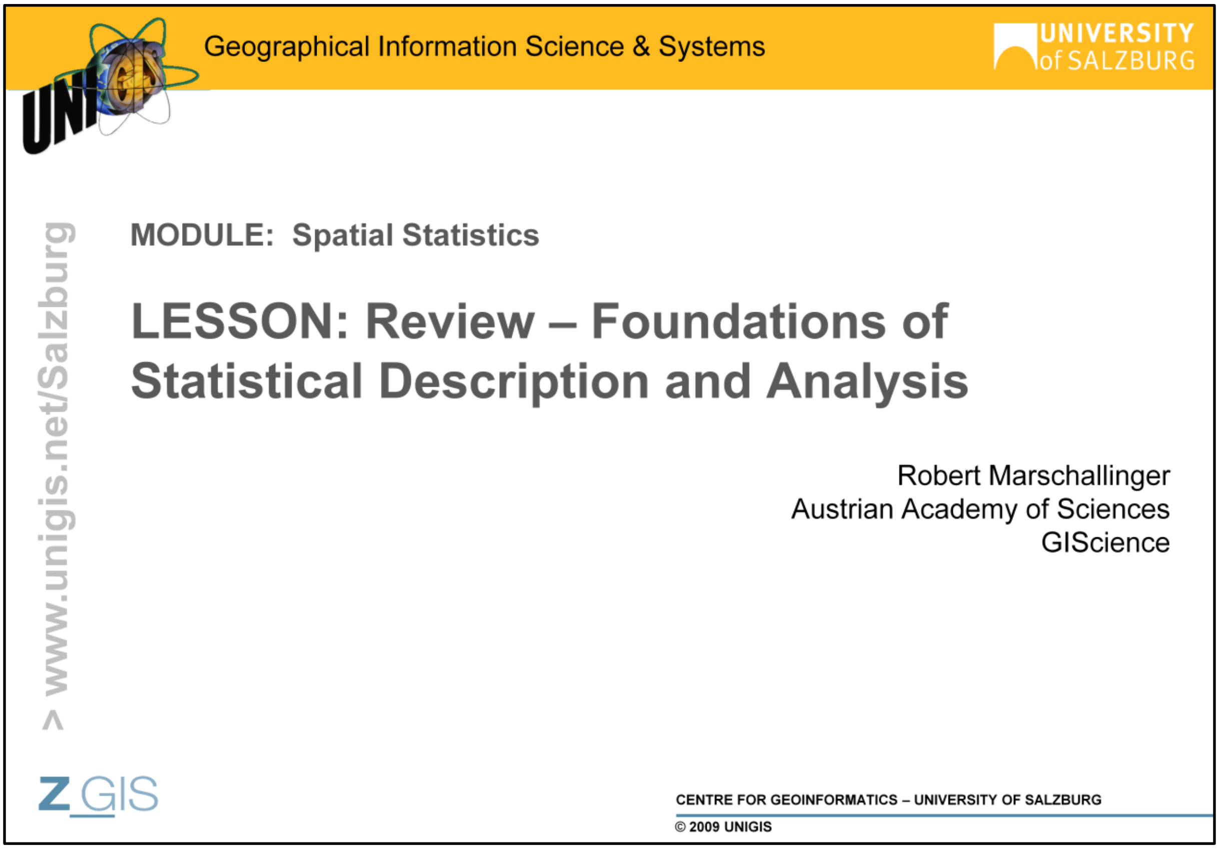


Geographical Information Science & Systems

UNIVERSITY of SALZBURG

> www.unigis.net/Salzburg

MODULE: Spatial Statistics

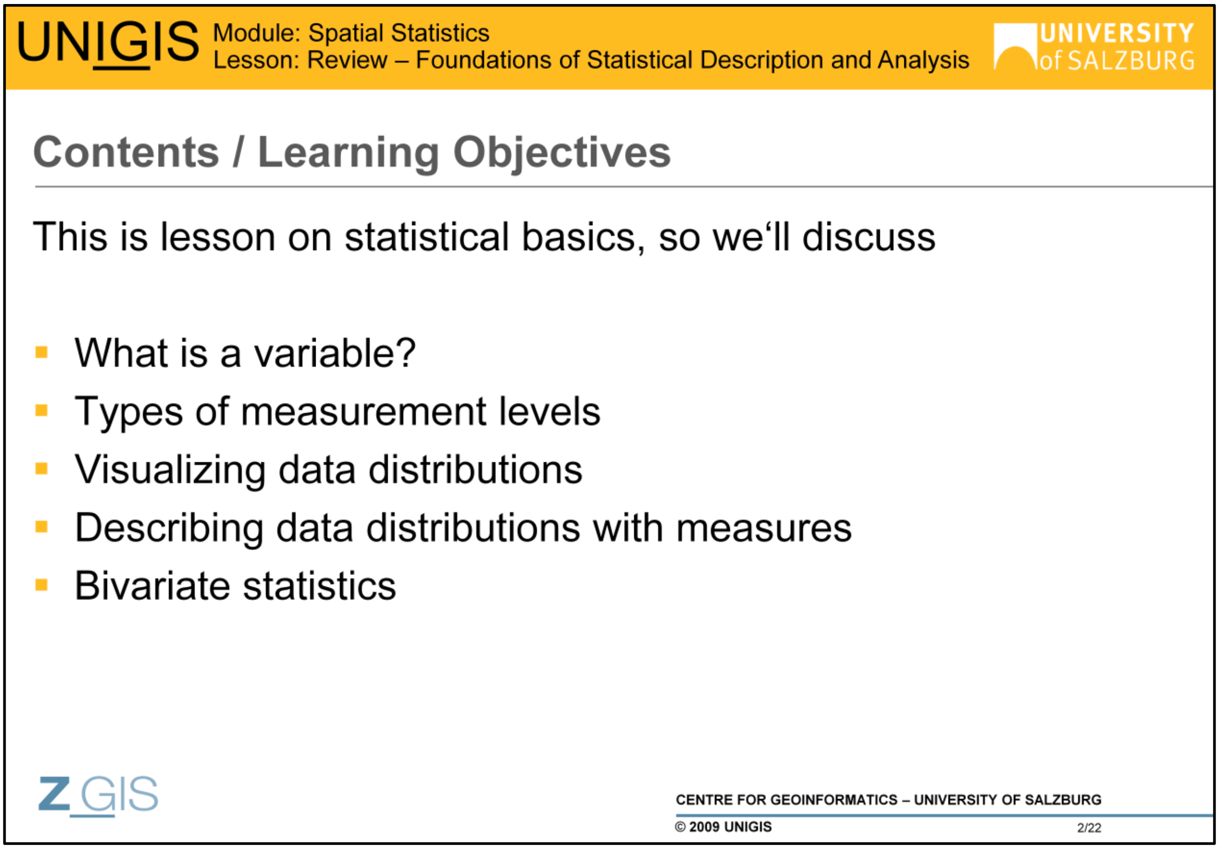
LESSON: Review – Foundations of Statistical Description and Analysis

Robert Marschallinger
Austrian Academy of Sciences
GIScience

Z_GIS

CENTRE FOR GEOINFORMATICS – UNIVERSITY OF SALZBURG

This is merely a review of uni- and bivariate descriptive statistics, along with using the R statistics package.



UNIGIS Module: Spatial Statistics
Lesson: Review – Foundations of Statistical Description and Analysis
UNIVERSITY of SALZBURG

Contents / Learning Objectives

This is lesson on statistical basics, so we'll discuss

- What is a variable?
- Types of measurement levels
- Visualizing data distributions
- Describing data distributions with measures
- Bivariate statistics

Z_GIS

CENTRE FOR GEOINFORMATICS – UNIVERSITY OF SALZBURG

© 2009 UNIGIS 2/22

R-Exercise (1)

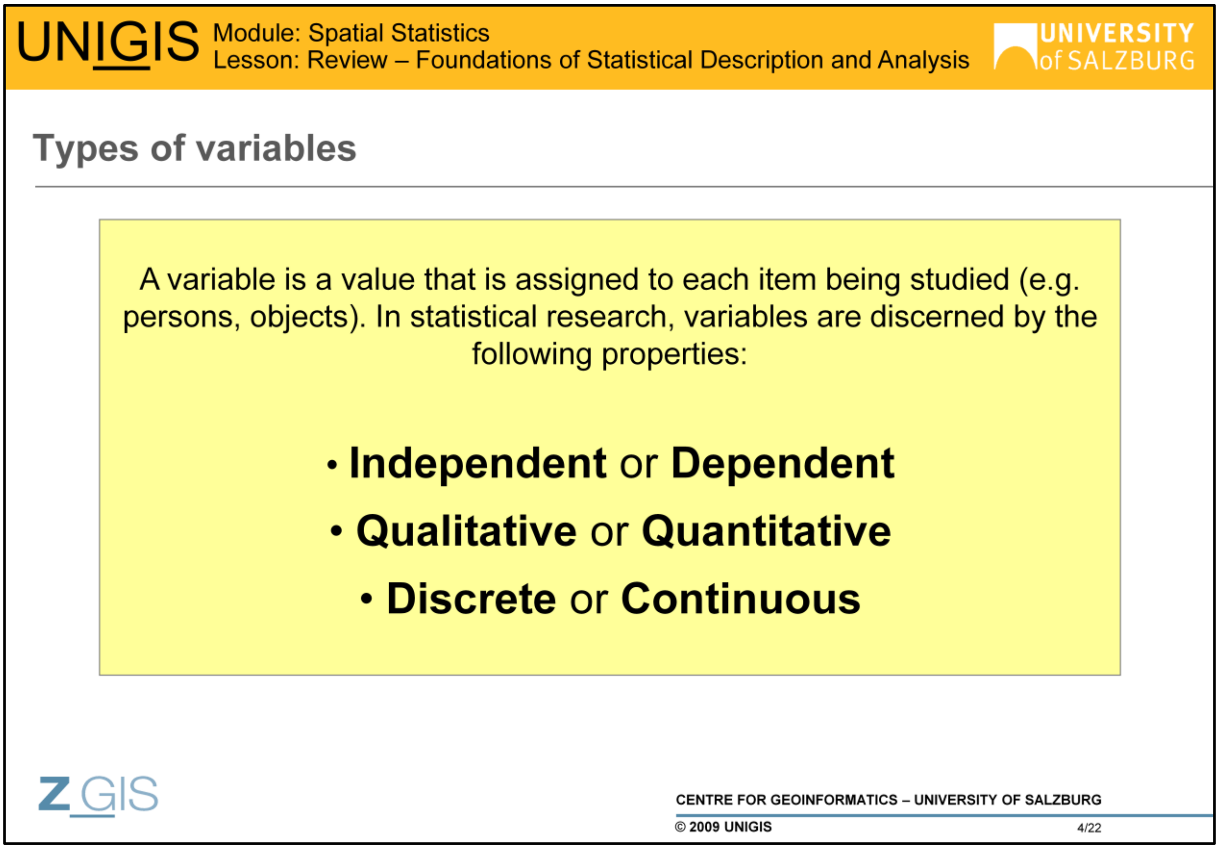
-----RRRRR-----

In this lesson, we'll use R along with the data set *limestone.dat* comprising 26 limestone samples, each with x,y coordinates and chemical analysis data for CaO, Mgo, Mno and SrO.

Start R. To read the contents of limestone.dat, type the following commands on the R console (R syntax in bold face); finish each line with CLRF - # are comments

```
>limedata = read.table(“c:/temp/M5/limestone.dat", header = TRUE)
                              # assuming limestone.dat is located in C:\temp\M5 –
                              obey the slash instead of backslash ;-)

>limedata                     # check data that are stored in array limedata
```

Leave R running or, when you'd like to interrupt, save the workspace before closing R. You can then continue from the last saved status.



UNIGIS Module: Spatial Statistics
Lesson: Review – Foundations of Statistical Description and Analysis
UNIVERSITY of SALZBURG

Types of variables

A variable is a value that is assigned to each item being studied (e.g. persons, objects). In statistical research, variables are discerned by the following properties:

- **Independent** or **Dependent**
- **Qualitative** or **Quantitative**
- **Discrete** or **Continuous**

Z_GIS

CENTRE FOR GEOINFORMATICS – UNIVERSITY OF SALZBURG
© 2009 UNIGIS 4/22

Statistics deals with variables that describe the objects being studied – there are different types of variables: the first distinction refers to the design of statistical experiments: **independent variables** are those that are manipulated whereas **dependent variables** are only measured or registered (e.g., organisms: age vs. height)

A **qualitative** variable is primarily non numeric. Examples of qualitative data include the colour of your hair, city names, and the type of car you drive. When qualitative data is presented, it is often summarized by totals or percentages. For example, 35% of the persons in the study have blond hair. Often qualitative data is assigned a number and entered into a database for reporting purposes. But if you were collecting data on hair colour, does it make sense to calculate an average colour? No! So you must be careful how you use and present qualitative data. Even if a number is assigned to qualitative data, it does not make sense to calculate an average colour. Instead you will likely report the number or percent of people who have black, brown, blond, etc. Your age or weight, your blood pressure and the temperature in your office are all **quantitative** variables. Furthermore, quantitative variables can be either **discrete** or **continuous**. A discrete variable has a specific or finite value. It can be counted. Examples of discrete variables include the number of desks in your office, the amount of memory (in bytes) in your computer, the number of pages in a script, or the number of people enrolled in a statistics course. A continuous variable is different. It can assume any value within a given range of precision. Your height is a good example of a continuous variable. Depending upon the precision of the measuring device, your height could be 1.8m, 1.82m, 1.824m etc. Continuous variables commonly occur when you measure something.

Read **chapter 1.F Introduction-Variables** of Online Statistics (http://onlinestatbook.com/chapter1/variables.html) and answer the questions at the end of the chapter. Then repeat these topics in Electronic Statistics Textbook StatSoft (http://www.statsoft.com/textbook/stathome.html): Go to Elementary Concepts in Statistics > **What are variables?**

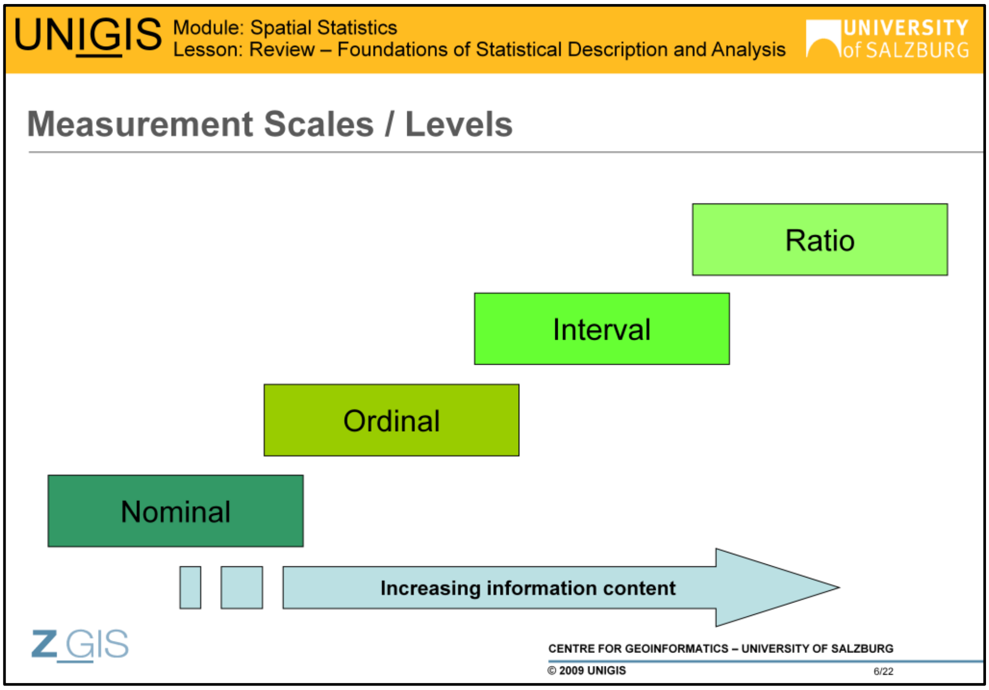
R-Exercise (2)

-----RRRRR-----

With R running, assign the contents of array limedata to variables. Type the following commands on the R console; finish each line with CLRF - # are comments

```
>id = limedata [, 1]                    # read array column 1 into variable id
>x = limedata [, 2]
                                        # read array column 2 into variable x
>y = limedata [, ‘y’]
                                        # another style of array column referencing
>CaO = limedata [, 4]
                                        # !!! oops – the R console is case
                                        sensitive!!!
>MgO = limedata [, 5]

>MnO = limedata [, 6]

>SrO = limedata [, 7]

>id
                                        # print variable id
>CaO
                                        # print variable CaO
```

Leave R running or, when you’d like to interrupt, save the workspace before closing R.



UNIGIS Module: Spatial Statistics
Lesson: Review – Foundations of Statistical Description and Analysis
UNIVERSITY of SALZBURG

Measurement Scales / Levels

Ratio

Interval

Ordinal

Nominal

Increasing information content

Z_GIS

CENTRE FOR GEOINFORMATICS – UNIVERSITY OF SALZBURG
© 2009 UNIGIS 6/22

Based upon its characteristics and how it is measured, data (and variables) is classified into four different **levels of measurement** or **measurement scales**. Different statistical analysis can be conducted on data, based upon the measurement scale it resides in. The four scales are: **nominal**, **ordinal**, **interval**, and **ratio**. Now let's look at each category.

Nominal scale data. Nominal data does not have any order to it. The data can only be counted and classified by categories or labels. Examples of nominal data include gender, the colour of your hair, or a landuse type (e.g., developing, industrial, ...). Survey answers of a YES / NO kind are another example of nominal data. Even if numbers are used to classify data, the numbers themselves have no meaning other than as a label or category.

Ordinal scale data. Ordinal data has some type of ranking or order to it. Ordinal data has the properties of nominal data, but the order or rank is meaningful. Many of us are acquainted with ordinal data which is found often on surveys. For example, if a survey asked for your opinion of this course, there could be five possible answers: excellent, very good, good, fair, or poor. The answers have a rank or value associated with them. Sometimes numbers are used to represent the possible answers: for example 1 = excellent, 2 = very good, etc. This is ordinal data as well.

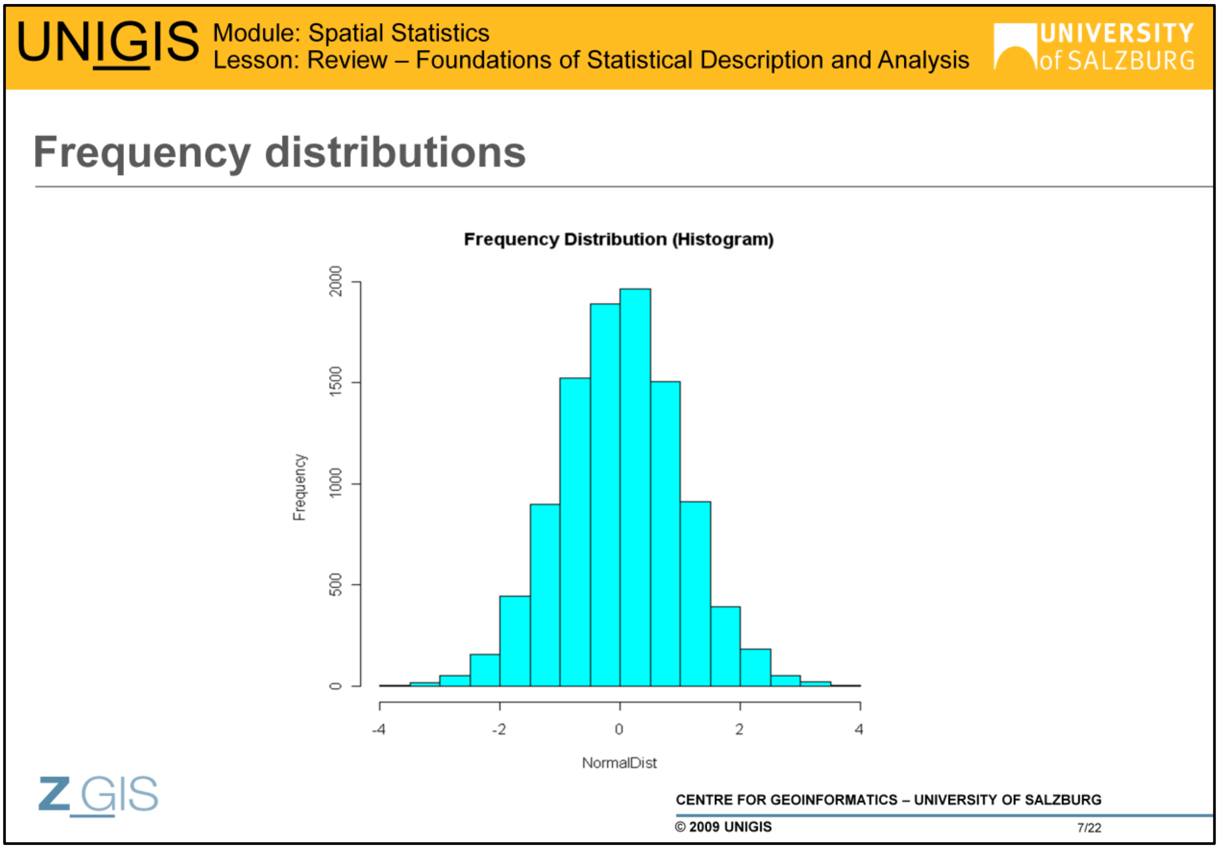
Interval scale data. Interval data has the properties of ordinal data and a fixed, measurable difference exists between the variables. The data contains an order that is based upon the amount of a characteristic it possesses. At interval scale the value of zero (0) does not have a meaningful value. Examples of interval data include the temperature using a thermometer or standardized test scores.

Ratio scale data. Ratio data has the highest possible level of measurement. It contains in addition to all of the characteristics of interval data the value of zero (0) which indicates that no value exists for a variable. Ratio data includes measuring distance, height, weight, and the cost of a good or service. For example, if you purchased a new computer today for €800, and the price of the TFT screen that you also purchased was €200; since this is ratio data, you could indicate that the price of the computer was 4 times the price of the TFT screen €800/€200). The most significant difference between ratio and interval data is that you can make comparisons like the computer and the TFT screen price with ratio data, but you cannot do the same with interval data. Is 40°C twice as hot as 20°C degrees? The numeric value is twice as much but it is difficult to determine if 40°C is twice as hot as 20°C (in fact, it is not!).

Notice that, via classification (binning), higher level measurement scales can be converted into lower scales (but not the other way round). As ratio data are by far the most common in geosciences, the focus will be on ratio data for the rest of this module.

For completeness, **cyclic data** and **circular data** are mentioned here. **Cyclic data** refer to processes (i.e., include time) will be discussed in the lesson on autocorrelation. **Circular data** (also known as 2D directional data), which are important in geography, will be considered in the lesson on explorative (spatial) data analysis.

Open the Electronic Statistics textbook Statsoft (http://www.statsoft.com/textbook/stathome.html) and then go to **Elementary Concept in Statistics > Measurement scales**. Read the explanations you find there on measurements scales and compare them with the **Levels of Measurement** from Online Statistics (http://onlinestatbook.com/chapter1/levels_of_measurement.html).



UNIGIS Module: Spatial Statistics
Lesson: Review – Foundations of Statistical Description and Analysis
UNIVERSITY of SALZBURG

Frequency distributions

Frequency Distribution (Histogram)

Z_GIS

CENTRE FOR GEOINFORMATICS – UNIVERSITY OF SALZBURG

© 2009 UNIGIS 7/22

A **frequency distribution** is used to organize data into classes, such that the number of observations within each class is shown. Classes within a frequency distribution are unique, i.e. each observation can be assigned to only one class (classes are mutually exclusive). Frequency distributions are plotted for presenting data in an organized way, to extract information and to analyze the data. **Often a frequency distribution is a first step in analyzing data**. The frequency per class can be displayed in several ways, the most common of which is the **histogram**. In a histogram, each bar represents one class; the bar height and the bar area are proportional to the class frequency, i.e. the absolute or relative number of observations per class. Frequency distributions can be applied to visualize data from all measurement scales; in the case of interval or ratio scale data, the continuous range of numbers has to be split up into classes, a process also known as classification (binning).

The example above was derived by drawing 10000 random samples from a standard normal distribution; each sample yields a number, which roughly is in the interval between -4 and +4, with a pronounced peak around a value of 0.

R-Exercise (3)

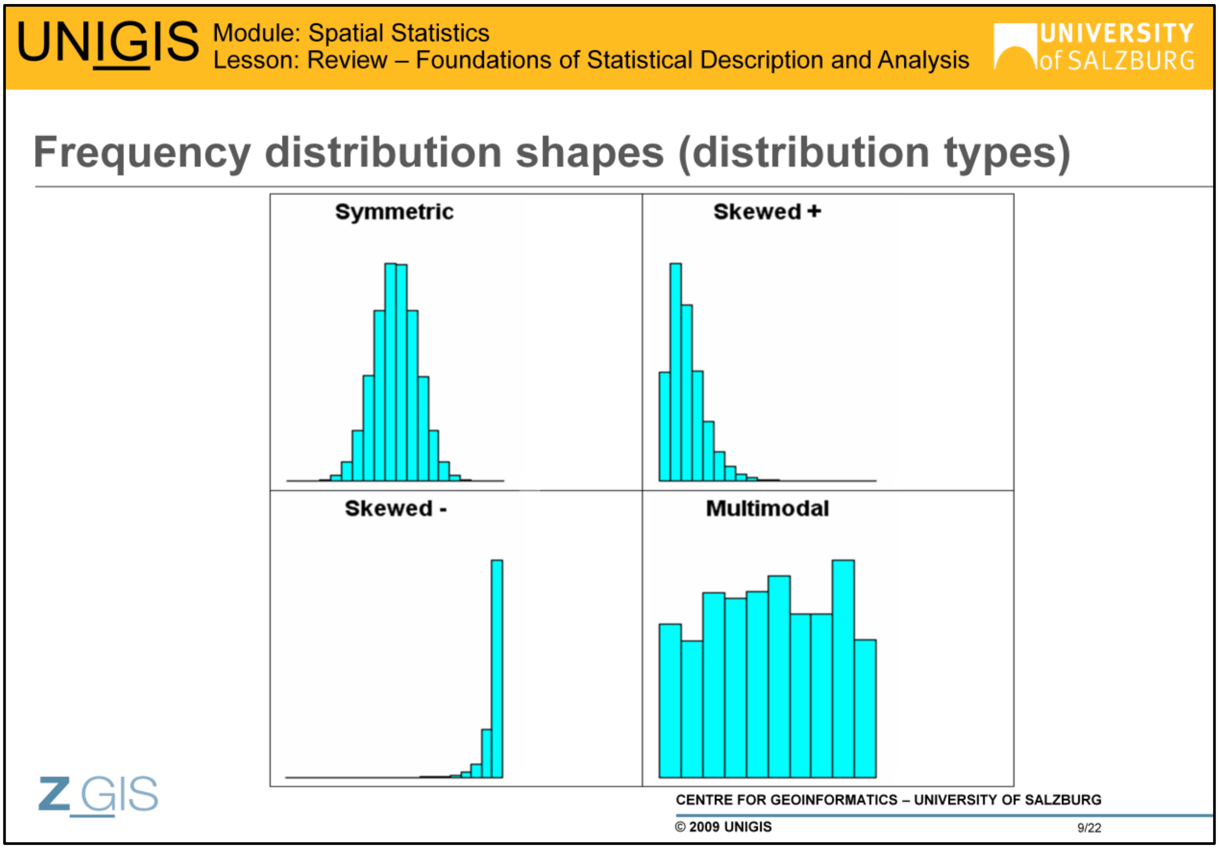
-----RRRRR-----

With R running, create a sample of the normal distribution. Type the following commands on the R console; finish each line with CLRF - # are comments

```
>hist (rnorm (100000))          # just for fun: visualize a random sample
                                (n=100000) of the normal distribution
```

Experiment with various sample sizes, e.g., rnorm (30), rnorm(1000) & compare the shape of the resulting frequency distribution.

Leave R running or, when you'd like to interrupt, save the workspace before closing R.



UNIGIS Module: Spatial Statistics
Lesson: Review – Foundations of Statistical Description and Analysis

UNIVERSITY of SALZBURG

Frequency distribution shapes (distribution types)

ZGIS

CENTRE FOR GEOINFORMATICS – UNIVERSITY OF SALZBURG

© 2009 UNIGIS

9/22

According to their shape, frequency distributions are commonly classified as follows:

- **Symmetric distributions** have one peak with a symmetric falloff in frequency to both sides.
- **Skewed distributions** have one peak, but an asymmetric falloff.
According to the relative position of the peak, we're talking of
    - **positive skewness** (most data to the right of the peak);
    positively skewed distributions have outliers with high values, i.e., a tail to the right
    - **negative skewness**
    the contrary is due for negatively skewed distributions, i.e., a tail to the right
- **Multimodal distribution** is the one where more than one peak is discernible

In the frequency distribution, both symmetric and skewed distributions do only have one peak.

Open the Electronic Statistics Textbook StatSoft (http://www.statsoft.com/textbook/stathome.html) go to **Contents >Basic Statistics > Shape of the Distribution, Normality** and read about the Shape of the Distribution

R-Exercise (4)

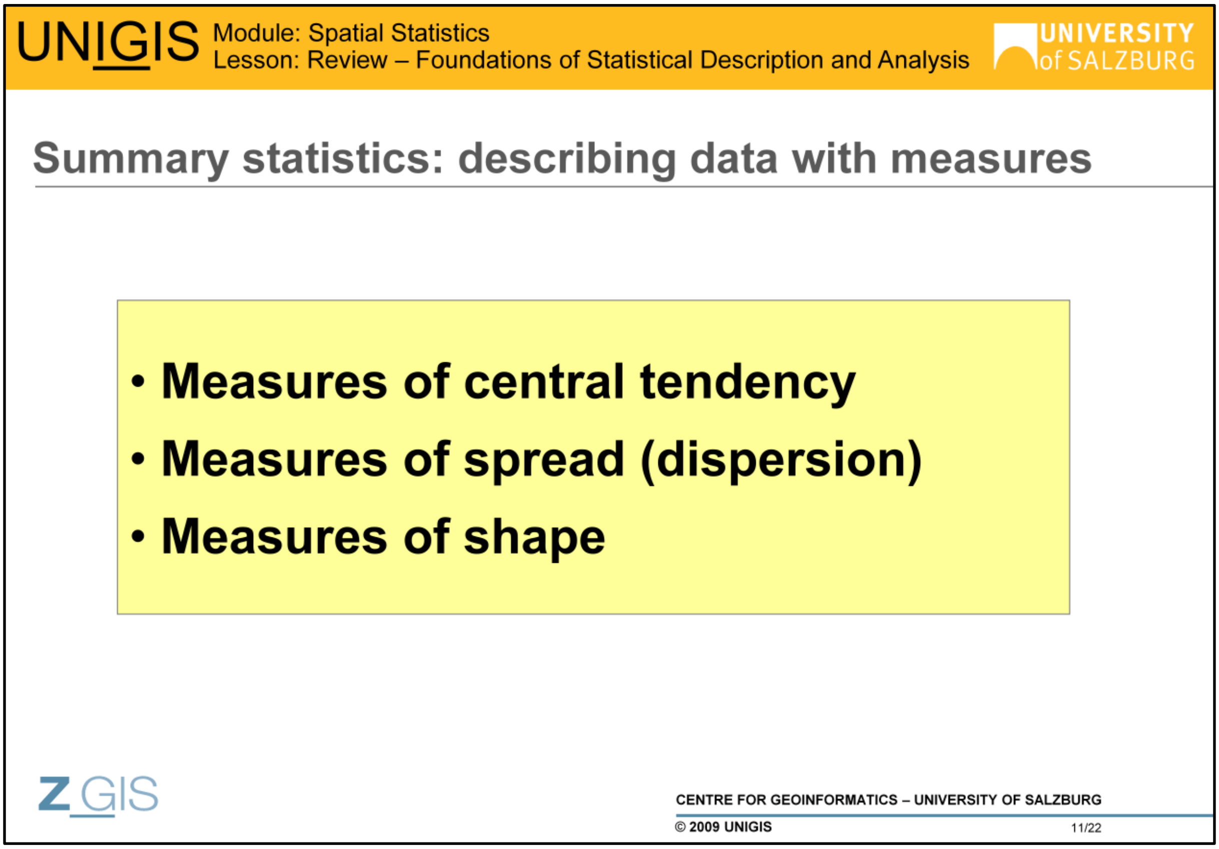
-----RRRRR-----

With R running, create histograms. Type the following commands on the R console; finish each line with CLRF - # are comments

```
>hist (CaO, col = 2)      # visualize the frequency distribution of CaO as red
                          histogram

>rug (CaO, col = 3)       # visualize the frequency distribution as line densities
```

Repeat for MgO, MnO, SrO, use different colors. Compare the shape of the resulting frequency distribution, try to describe the shapes.

Leave R running or, when you’d like to interrupt, save the workspace before closing R.



UNIGIS Module: Spatial Statistics
Lesson: Review – Foundations of Statistical Description and Analysis

UNIVERSITY of SALZBURG

Summary statistics: describing data with measures

- Measures of central tendency
- Measures of spread (dispersion)
- Measures of shape

CENTRE FOR GEOINFORMATICS – UNIVERSITY OF SALZBURG

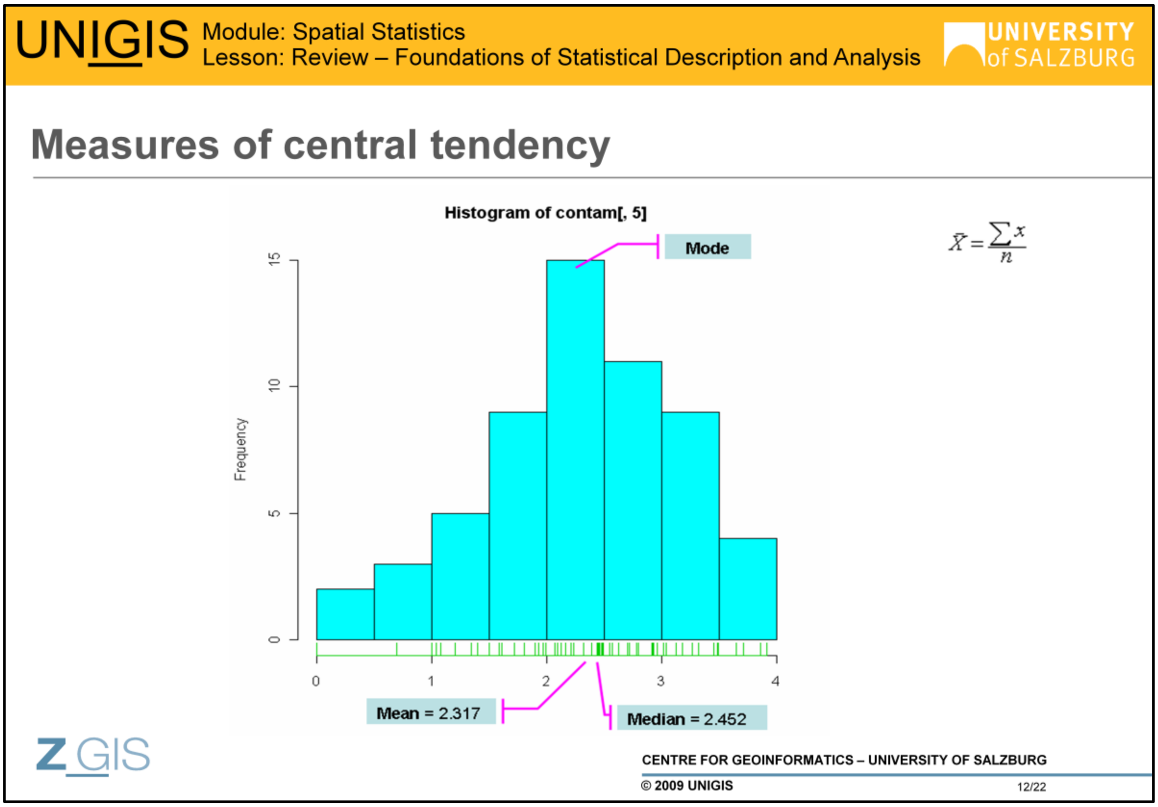
© 2009 UNIGIS 11/22

More quantitative than just plotting frequency distributions and verbally characterizing their shape, **summary statistics** aims at conveying information about the distribution of sample data by means of **descriptive measures**; this distinguishes summary statistics from **Inferential Statistics**, which comprises the use of statistics to make inferences concerning some unknown aspect of a population – we will not deal with inferential statistics in this lesson.

An important use of descriptive statistics is to summarize a collection of data in a clear and understandable way. The most popular measures are concerning **central tendency**, **spread**, and **shape** of a distribution.

-----RRRRR-----

With R running, start the online help system and read: **An Introduction to R -> Manuals -> Probability Distributions -> Examining the distribution of a set of data**



UNIGIS Module: Spatial Statistics
Lesson: Review – Foundations of Statistical Description and Analysis
UNIVERSITY of SALZBURG

Measures of central tendency

$\bar{X} = \frac{\sum x}{n}$

Z_GIS
CENTRE FOR GEOINFORMATICS – UNIVERSITY OF SALZBURG
© 2009 UNIGIS
12/22

The ***mode*** is the value that occurs most frequently in a set of data. In nominal or categorical data, it is the category with the most observations. Recall that data distributions can be uni- or multimodal.

In sparsely sampled data, the mode strongly depends on the class width.

The ***median*** is the middle number or value from a set of data.

- It is the middle value or number in a group of values that are sorted or ranked from lowest to highest.
- If there is an even set of data values, the median is found by taking the middle two values and adding them together and then dividing by two.
- The median is not influenced by outliers or extreme values. Outliers, whether they are large or small, are considered just one value in a series of values. All of the values have equal weight since each value is counted as one.

The ***mean*** is a widespread measurement of central tendency for numeric data. Simply stated, it delivers a central or average value for a set of values. The mean is calculated by summing all observations and dividing the sum by the number of observations (simple average). The symbol $\mathbf{X}_{\mathbf{bar}}$ represents the sample mean. The symbol ∑ indicates that all of the following items should be added together. ***x*** is the value of an observation or one data value in a sample. The symbol n is the total number of data values or observations in a sample. The mean is also known as **first moment statistics.**

Characteristics of the mean:

- It is the average or central value of a set of numbers.
- The mean is often referred to as a balance point for a set of numeric values. This indicates that all of the values added together above the mean are at the same distance as the values added together below the mean.
- The mean is influenced by large or small values. Extreme values in a set of data are called *outliers*. Outliers affect the mean and cause it to be larger or smaller. As a result if outliers are present, the mean may not accurately represent a set of data.

Read the introductory text about **Central Tendency** (http://onlinestatbook.com/chapter3/central_tendency.html). Then, experiment with **Balance scale Simulation** (http://onlinestatbook.com/chapter3/balance.html) and **Mean and Median Simulation** (http://onlinestatbook.com/chapter3/mean_median_sim.html). This helps in getting a feeling for the usefulness of measures of central tendency in the case of multimodal distributions.

R-Exercise (5)

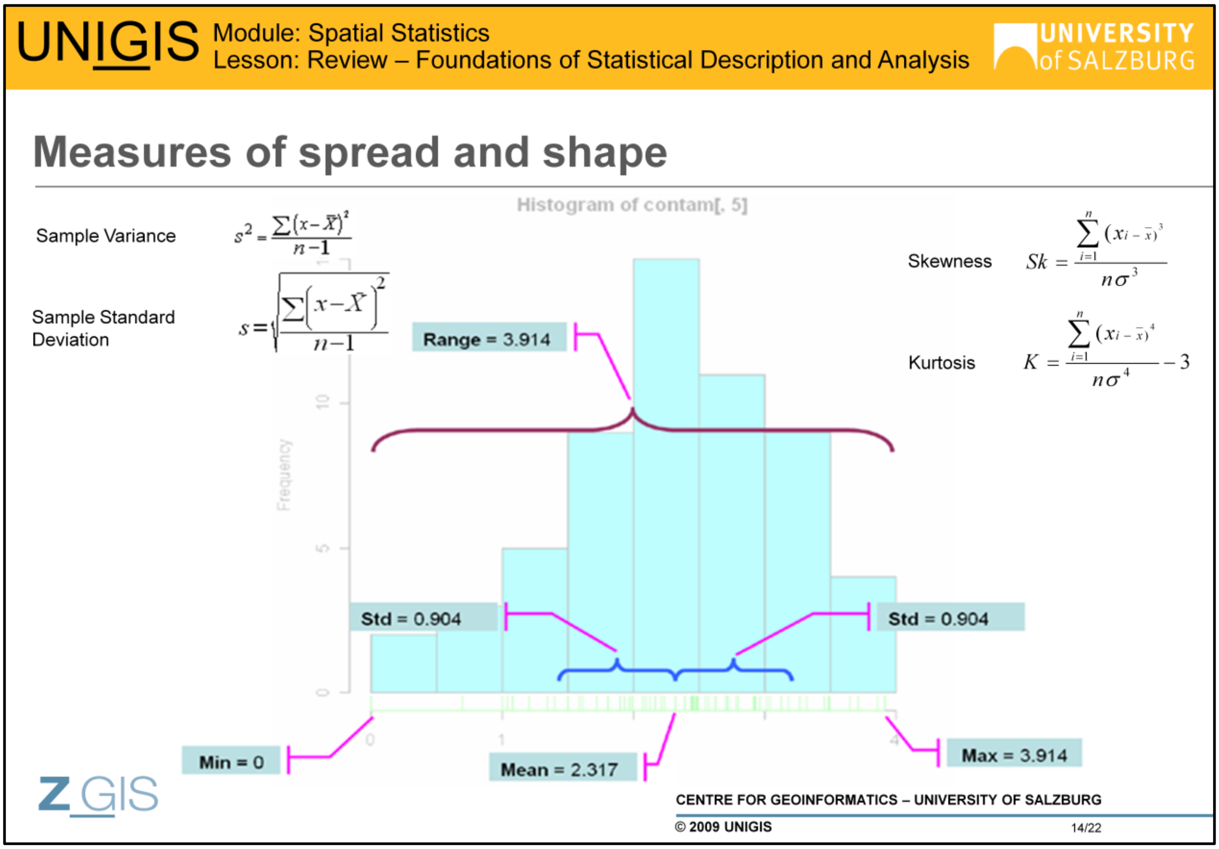
-----RRRRR-----

With R running, compute measures of central tendency for oxides in limedata. Type the following commands on the R console; finish each line with CLRF - # are comments

```
>mean (CaO)              # calculate the mean of the variable CaO

>median (CaO)            # calculate the median of the variable CaO
```

Repeat for MgO, MnO, SrO. Compare the values of mean/median pairs.

Leave R running or, when you’d like to interrupt, save the workspace before closing R.



UNIGIS Module: Spatial Statistics
Lesson: Review – Foundations of Statistical Description and Analysis
UNIVERSITY of SALZBURG

Measures of spread and shape

Sample Variance $s^2 = \frac{\sum(x-\bar{X})^2}{n-1}$

Sample Standard Deviation $s=\sqrt{\frac{\sum\left(x-\bar{X}\right)^2}{n-1}}$

Skewness $Sk = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{n\sigma^3}$

Kurtosis $K = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{n\sigma^4} - 3$

Range = 3.914

Std = 0.904

Std = 0.904

Min = 0

Mean = 2.317

Max = 3.914

ZGIS

CENTRE FOR GEOINFORMATICS – UNIVERSITY OF SALZBURG

© 2009 UNIGIS 14/22

Measures of spread and shape of a distribution

A measure of spread (also known as dispersion or variation) is used to determine the variability of data, i.e. to determine how well the mean represents the data. If there is little variability in our data, then the mean can represent the data accurately enough. However if there is a lot of variability in the data, if data that are scattered and distant from the mean, then the mean should be used to represent the data values only with caution. Complementing the mean, dispersion tells us about the spreading of data around the mean.

The following are some popular measures of variation.

The **Range** is *one* of the easiest measurements of data dispersion. It is a difference between the highest and lowest values in a set of data. The range is often referred to as the simplest or crudest measure of dispersion.

The **Variance** measures the mean of the squared deviations from the mean on a set of data. As squared deviations from the mean are used in its calculation, variance is also known as a **second moment statistic**.

The variance will have a value of zero when all the data values are the same, the greater the differences in the data values the larger the variance will be.

The symbol **s^2** is the symbol used to represent the **sample variance** and is called *s squared*. Once again ***x*** represents the value of an observation or one data value, this time in a **sample**. The symbol X_{bar} represents the **sample** mean. The symbol n represents the total number of data values or observations in a sample; using **n-1** provides a better estimate of population variance when you calculate the sample variance.

The **Standard Deviation** is the square root of the variance. Standard deviation is also a **second moment statistic**.

- The steps for calculating the standard deviation for a sample are identical to those for calculating sample variance, except that you take the square root of the variance.
- The small value of the standard deviation indicates that the data values are close to the mean value and are not scattered.
- The standard deviation is expressed in the original data units.

The Empirical Rule states that if a distribution is normal, then:

- About 68% of all data values lie within one standard deviation of the mean (positive or negative)
- About 95% of all data values lie within two standard deviations of the mean (positive or negative)
- About 99.7% of all data values lie within three standard deviations of the mean (positive or negative)

The **Coefficient of Variation**: dividing the standard deviation by the mean results in the coefficient of variation, which is frequently used to compare the dispersion of sets of data that use different measurement values or have extreme differences between the means.

Percentiles are more robust estimates of the spread of a distribution (i.e., without the assumption of an underlying "near-normal" distribution). Like the median, they are portions of a ordered sequence of values. Well-known percentiles are the *lower (25%)* and *upper (75%) quartiles*. Percentiles can be calculated for all portions.

Skewness measures the directional bias of a distribution – the extent to which data in a distribution cluster to higher or lower values than the mean. When most values are less than the mean, it is a positively skewed distribution and vice versa. Skewness is based on third powers of mean deviation an therefore it is a **third moment statistics** (compare formula above; this is just one of several formulas to calculate skewness).

The **Kurtosis** expresses whether values of a distribution are concentrated around one value; high kurtosis values indicate a sharp peak (mode) in the distribution. Kurtosis is a **fourth moment statistics**.

R-Exercise (6)

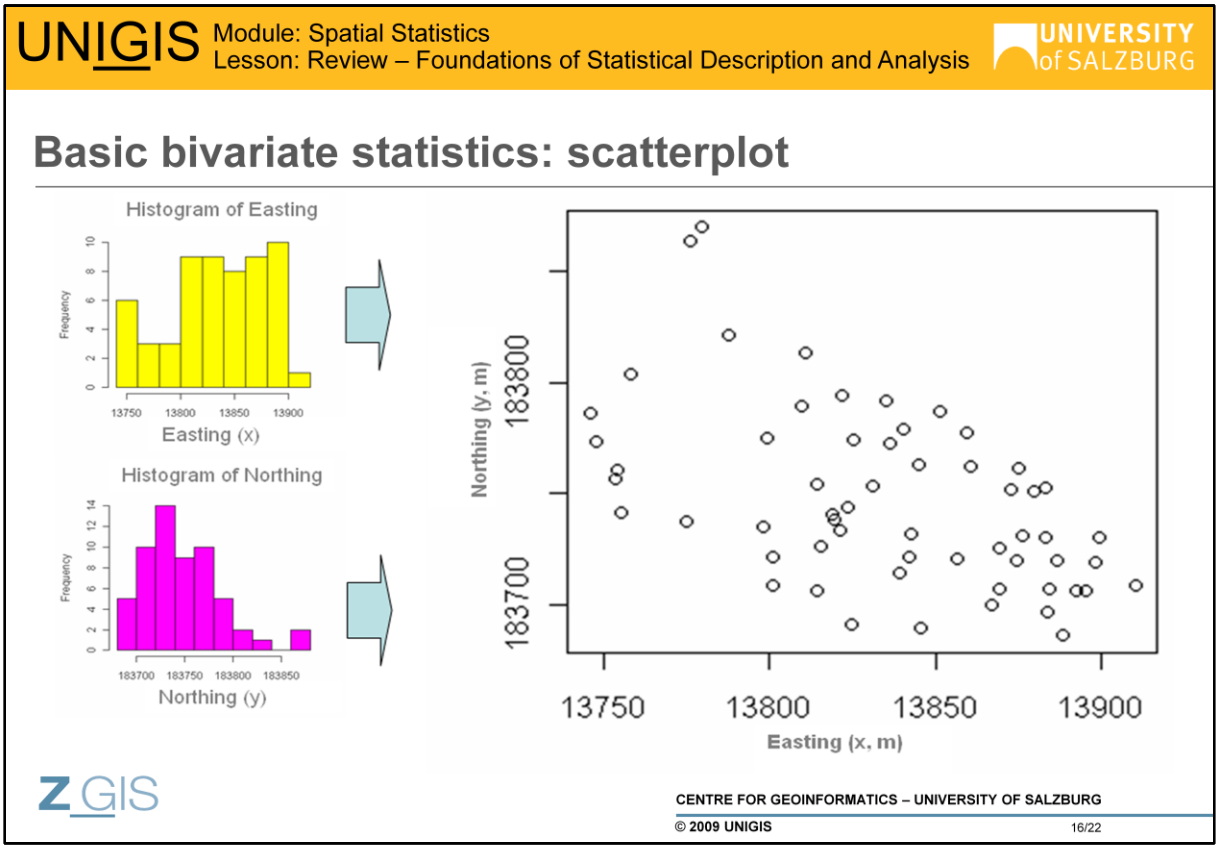
-----RRRRR-----

With R running, compute measures of central tendency for oxides in limedata. Type the following commands on the R console; finish each line with CLRF; # are comments.

```
>min (CaO)            # calculate the minimum of the variable CaO

>max (CaO)            # calculate the maximum of the variable CaO

>range (CaO)          # calculate the range of the variable CaO

>var (CaO)            # calculate the variance of the variable CaO

>std (CaO)            # calculate the standard deviation of CaO

>summary (CaO)        # short version of summary statistics of CaO
```

Repeat the procedure for MgO, MnO, SrO.

Leave R running or, when you'd like to interrupt, save the workspace before closing R.



UNIGIS Module: Spatial Statistics
Lesson: Review – Foundations of Statistical Description and Analysis
UNIVERSITY of SALZBURG

Basic bivariate statistics: scatterplot

Z GIS

CENTRE FOR GEOINFORMATICS – UNIVERSITY OF SALZBURG

© 2009 UNIGIS 16/22

As outlined before, summary statistics are useful for understanding how values of one variable are distributed within one dataset. These descriptive **univariate** measures (central tendency, spread, etc.) cannot, however, measure **multivariate** relationships among different distributions. Frequently, more than one variable is measured per case – consider a simple set of points with the x,y coordinates recorded per point (top). In such situations, not only the x,y variables themselves but merely the inter-variable relationships are used to gain insight into the data set. One way to portray these relationships would be to visually compare the frequency distributions of each variable (the histograms top left); however, this does not maintain the inter-pair relationships. Indeed, we can learn much more by displaying the bivariate data in a graphical form that maintains the pairing – in the form of a **scatterplot**, *which portrays the geographical location of the samples* (top right).

Going beyond simply visualizing bivariate data, we‘re often interested in quantifying the relationships between two variables.

As two examples of basic bivariate statistics, we‘ll consider interval and ratio scale **correlation**, which *measures the dependence of one variable on another* and **regression**, which *statistically measures the direction and strength of relationship between two variables.*

R-Exercise (7)

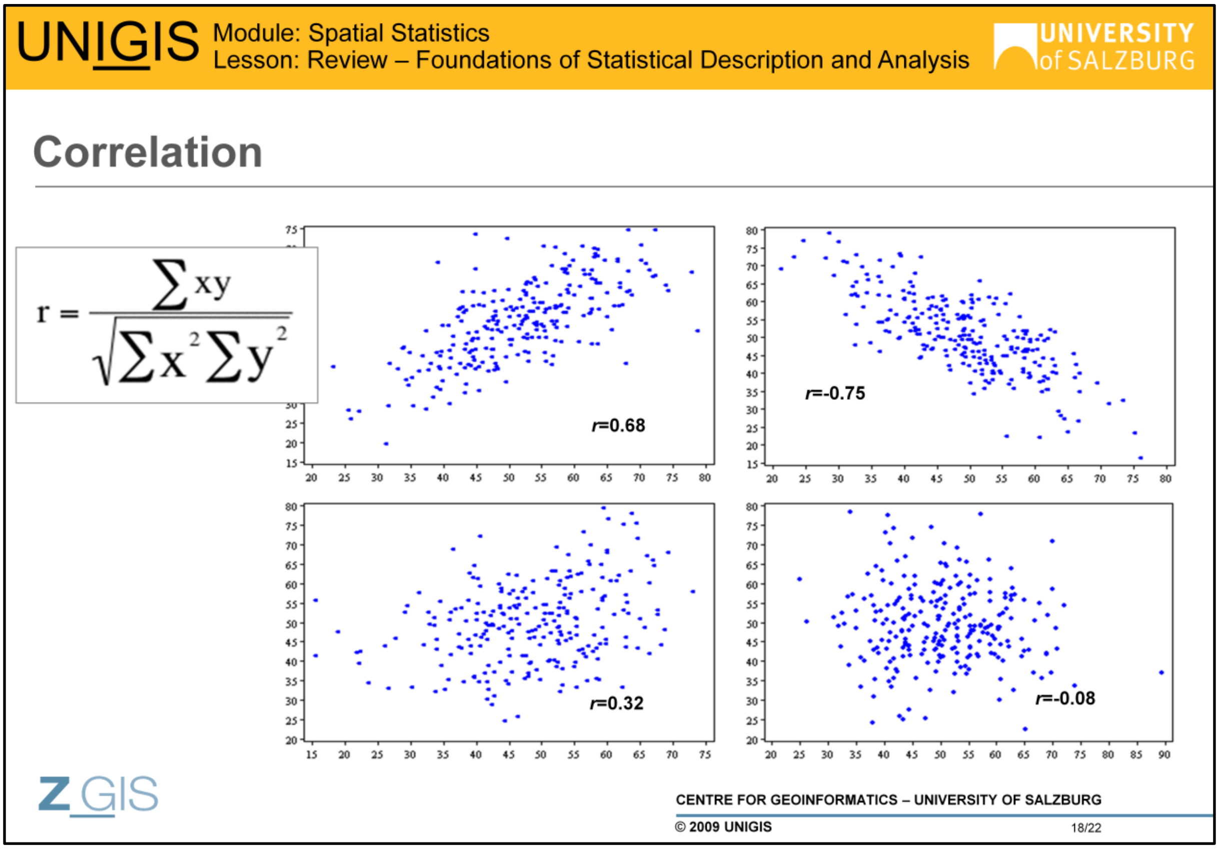
-----RRRRR-----

With R running, visualize a scatterplot of xy coordinate pairs in limedata. Type the following commands on the R console; finish each line with CLRF - # are comments

```
>plot (x,y)                              # visualize the scatterplot
```

Do you have any idea how to change the color, size, and stlye of xy symbols?

Leave R running or, when you'd like to interrupt, save the workspace before closing R.



UNIGIS Module: Spatial Statistics
Lesson: Review – Foundations of Statistical Description and Analysis
UNIVERSITY of SALZBURG

Correlation

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

ZGIS

CENTRE FOR GEOINFORMATICS – UNIVERSITY OF SALZBURG
© 2009 UNIGIS 18/22

Correlation *statistically measures the direction and strength of the relationship between two variables.*

Generally, a **positive correlation** between two variables X and Y means increasing X will cause Y to increase its value and vice versa. Similarly, a **negative correlation** means increasing X will decrease Y. If you want to measure the magnitude and direction of correlation between two interval or ratio scale variables, the most popular statistics is **Pearson's product moment correlation coefficient, *r***. It indicates **a linear relationship** between two measurement variables. In a **scatter diagram**, two perfectly linearly correlated variables share a line while non correlated points are randomly distributed in a scatter diagram (there are exceptions of course – a circular arrangement is „uncorrelated" in that sense).

The range of ***r*** is -1 to +1, with

- **-1** indicating *a perfectly negative correlation,*
- **0** means *no correlation at all,* and
- **+1** is a *perfectly positive correlation.*

In essence, Pearson's ***r*** gives you a first indication whether a variable can be considered a linear predictor of another variable; compare the shape of the scatterplots above and the associated ***r*** values.

From the chapter **Describing Bivariate Data** (http://onlinestatbook.com/chapter4/bivariate.html) read the **introductory text**. Then, experiment with **guessing r** (http://onlinestatbook.com/chapter4/pearson_demo.html).

R-Exercise (8)

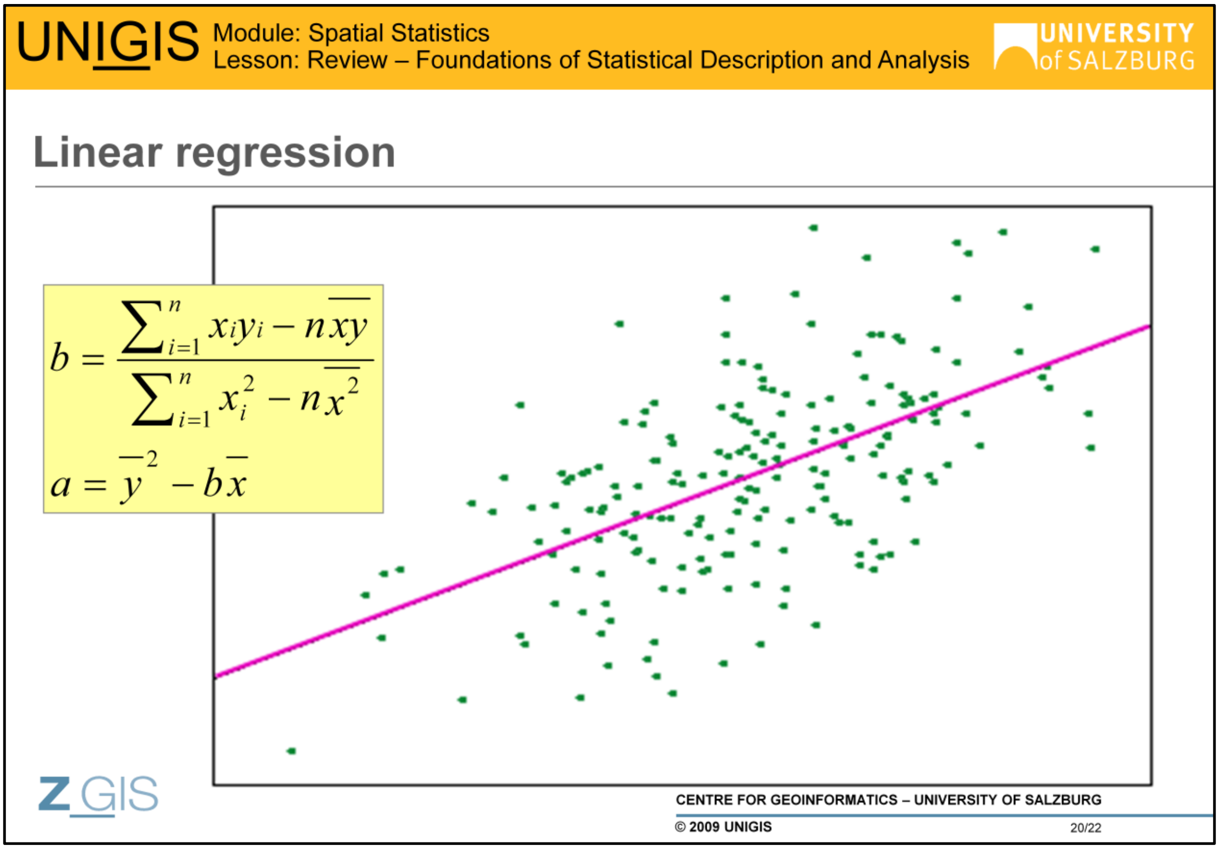
-----RRRRR-----

With R running, first do a scatterplot of oxides CaO and MgO in limedata, then calculate the Pearson's r for these variables. Type the following commands on the R console; finish each line with CLRF - # are comments

```
>plot (CaO,MgO)        # scatterplot of CaO and MgO to view correlation

>cor (CaO,MgO)         # calculate Pearson's r for CaO and MgO
```

Repeat the procedure for CaO-MnO and CaO-SrO. Try to synthesize graphics and values of r.

Leave R running or, when you'd like to interrupt, save the workspace before closing R.



UNIGIS Module: Spatial Statistics
Lesson: Review – Foundations of Statistical Description and Analysis
UNIVERSITY of SALZBURG

Linear regression

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

$$a = \overline{y}^{2} - b\overline{x}$$

Z GIS

CENTRE FOR GEOINFORMATICS – UNIVERSITY OF SALZBURG

© 2009 UNIGIS 20/22

In this section, we discuss **trend analysis** basics, i.e. *measuring the linear dependence of one variable on another*. As discussed before, r measures strength and direction of a correlation.

Regression allows us to *calculate the value of one variable based on the value of another variable*. **Linear regression** attempts to model the relationship between two variables by **fitting a linear equation** to observed data. One variable is considered to be an *explanatory variable*, and the other is considered to be a *dependent variable* (for example, a modeller might want to relate the weights of individuals to their heights using a linear regression model). The regression is expressed as a linear equation with parameters ***b*** (slope) and ***a*** (y intercept); these depend on the variables ***x*** and ***y*** only (see regression equations top left). One of the most important characteristics of the **regression line** is that it is the ***best fit line***, meaning it **minimizes the distances between observations and predictions**. In the above scatterplot, the observations are the green points, through which the regression line is fit (magenta).

Before attempting to fit a linear model to observed data, a modeller should first determine whether or not there is a relationship between the variables of interest. A scatterplot can be a helpful tool in determining the strength of the relationship between two variables.

Whenever a linear regression model is fit to a group of data, the range of the data should be carefully observed. Attempting to use a regression equation to predict values outside of this range, i.e. to ***extrapolate***, is often inappropriate, and may yield meaningless answers. In contrast, given the assumptions of a linear dependence, ***interpolation*** is generally safe. Once a regression model has been fit to a group of data, examination of the residuals (the deviations from the fitted line to the observed values) allows the modeller to investigate the validity of his or her assumption that a linear relationship exists.

Read the introduction to **Simple Linear Regression** (http://onlinestatbook.com/chapter12/intro.html), then experiment with **the Linear Fit** (http://onlinestatbook.com/chapter12/linear_fit_demo.html) and **Prediction Line** (http://onlinestatbook.com/chapter12/prediction_line_demo.html) **simulations**.

R-Exercise (9)

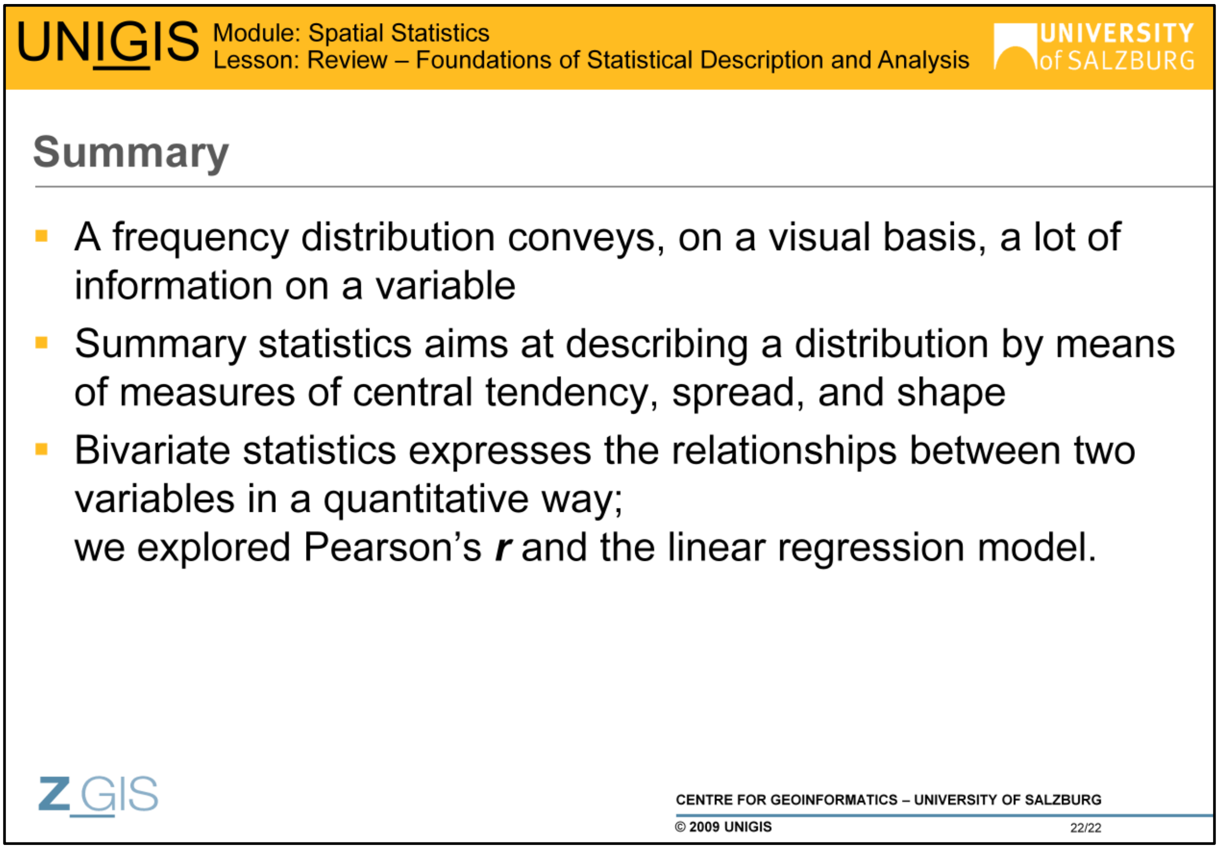
-----RRRRR-----

With R running, first do a scatterplot of oxides CaO and MgO in limedata, then calculate the linear regression on these variables. Type the following commands on the R console; finish each line with CLRF - # are comments

```
>plot (CaO,MgO)                # scatterplot of CaO and MgO to view
                                 correlation

>abline (lm(MgO ~ CaO))        # visualize the regression line

>lm(MgO ~ CaO)                 # print linear regression coefficients
```

Repeat the procedure for CaO-MnO and CaO-SrO. Try to synthesize graphics and coefficients of the regression line.

Leave R running or, when you’d like to interrupt, save the workspace before closing R.



UNIGIS Module: Spatial Statistics
Lesson: Review – Foundations of Statistical Description and Analysis

UNIVERSITY of SALZBURG

Summary

- A frequency distribution conveys, on a visual basis, a lot of information on a variable
- Summary statistics aims at describing a distribution by means of measures of central tendency, spread, and shape
- Bivariate statistics expresses the relationships between two variables in a quantitative way;
  we explored Pearson's ***r*** and the linear regression model.

CENTRE FOR GEOINFORMATICS – UNIVERSITY OF SALZBURG

© 2009 UNIGIS 22/22