Correlation and Regression Theory

1) Multivariate Statistics

- *What is a multivariate data set?*
- *How to statistically analyze this data set?*
- *Is there any kind of relationship between different variables in one data set? And how to present it? ➔ multiple correlation and multiple regression*
- *Is it better to use univariate or multivariate statistical analysis in dealing with such data sets?*

- Univariate for each individual variable
- Multivariate if there is a **relationship or dependencies** between different variables in a data set.

Correlation and Regression Theory

2) Bivariate Statistics

What is a bivariate statistics?

• It is a special case of multivariate statistics ➔Simple correlation and simple regression

Bivariate statistics deals with the organization, presentation, and summary of data of **TWO** variables. **Examples:**

•In geochemical analysis ➔ relationship between gold and tungsten?

•In hydrocarbon reservoirs characterization ➔ relationship between sonic wireline logs and porosity?

•In carbonate rocks ➔ relationship between porosity and magnesium content?

Compare the following distributions!

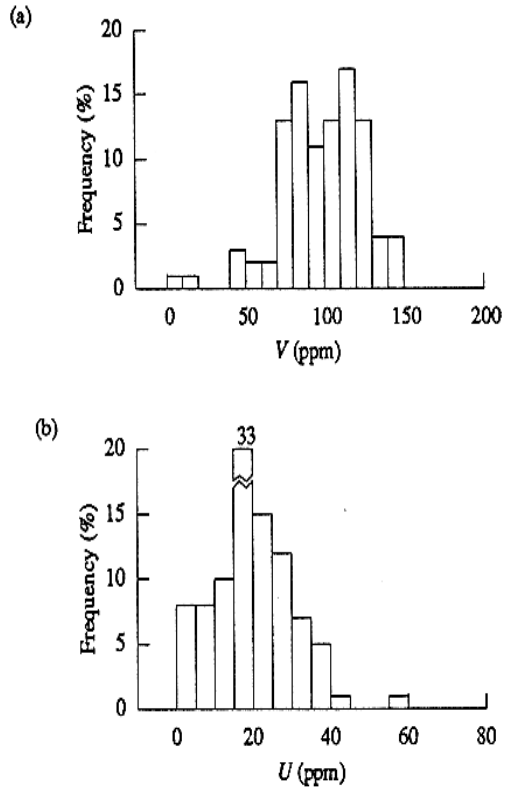


Figure 3.2 The histogram of the 100 V values in (a) and of the corresponding 100 U values in (b).

Table 3.1 Statistical summary of the V and U values shown in Figure 3.1.

	V	U
n	100	100
m	97.6	19.1
σ	26.2	9.81
CV	0.27	0.51
min	0.0	0.0
Q_1	81.3	14.0
M	100.5	18.0
Q_3	116.8	25.0
max	145.0	55.0

How to represent the relationship between both distributions?

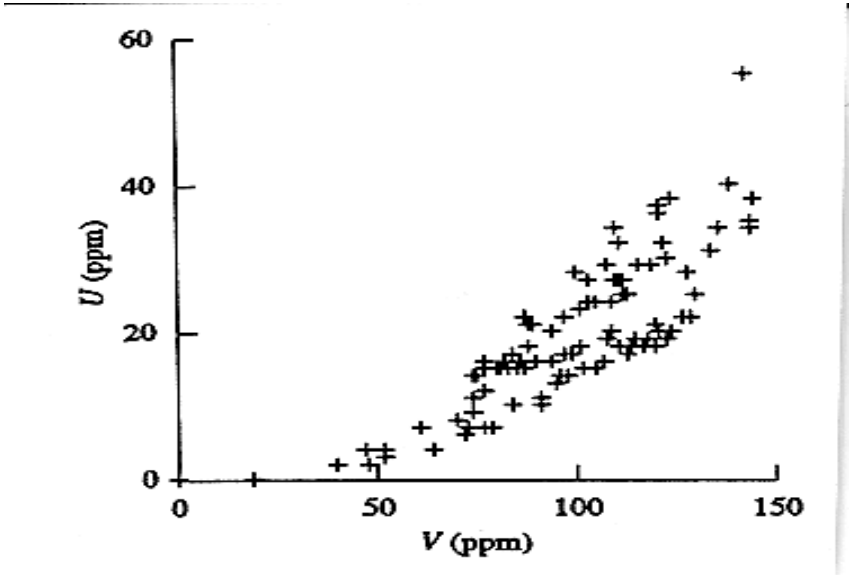
1) Scatterplots
2) Covariance and correlation
3) Regression analysis

Scatterplot

Def.: **Cloud** of points represented on a rectangular coordinate system.

Provides a **QUALITATIVE** "**feel**" of how the 2 variables are related.

Draws attention of aberrant "**diverging**" data points or outliers!



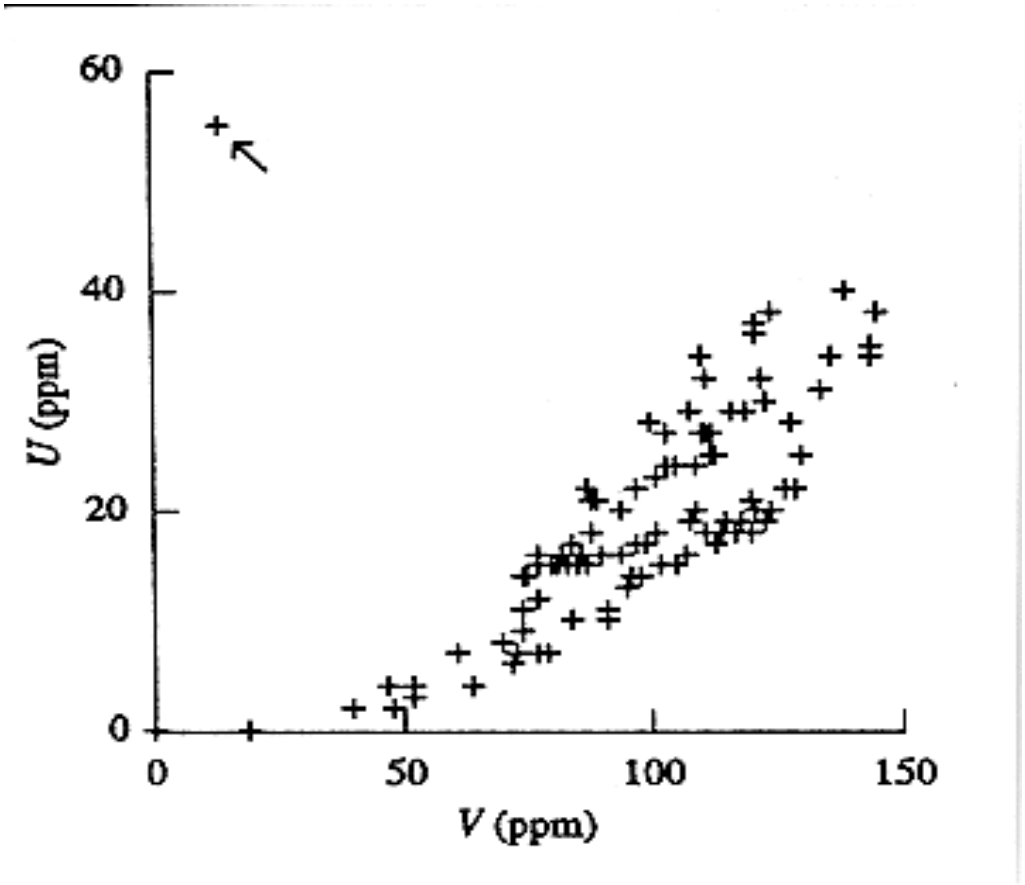
Scatterplot and Outliers

- Is it a real data point?

- Is it an outlier?
  - Rosner's test.
  - Control charts

- How to treat it?
  - Power or logarithmic **transformation**! This might be a more **robust** statistical technique.
  - Cancellation! This decision **should not be taken for guarantee** especially if the data set contains **multiple "outilers"!**

Scatterplot and Outliers



Covariance and Correlation

Quantitative measures of the **relationship** between two variables.

How to calculate and what does it mean?

Covariance

$$COV(u,v) = S_{uv} = \frac{1}{n-1}\sum_{i=1}^{n}(u_i - m_u)(v_i - m_v)$$

Correlation Coefficient

$$r = \rho = \frac{S_{uv}}{S_u S_v}$$

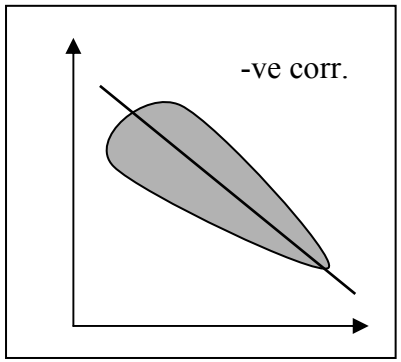
Def.: This coefficient measures the strength of the **linear relationship** between two variables

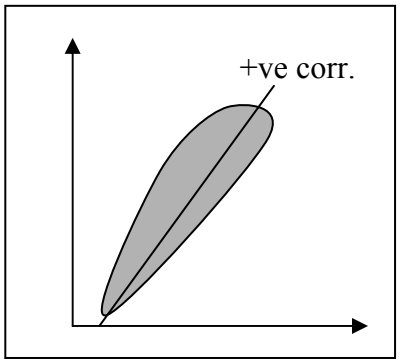
• Its value range is between **–1** (completely negative ➔ one variable increases at the expense of the other!) and **+1** (completely positive ➔ both variables increase and decrease in the same way!)

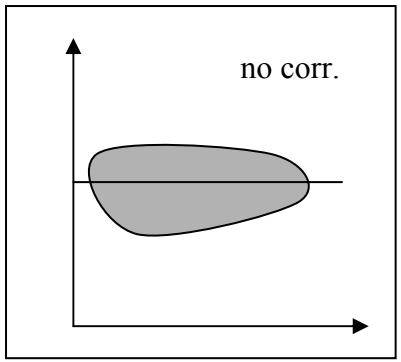
• It is **sensitive** to outliers

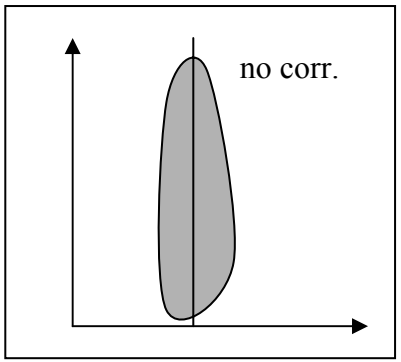
• The absence of linear relationship does not mean that there is no relationship between the two variables ➔ **Non-linear relationship** might exist!

Positively or Negatively correlated, or Uncorrelated Variables









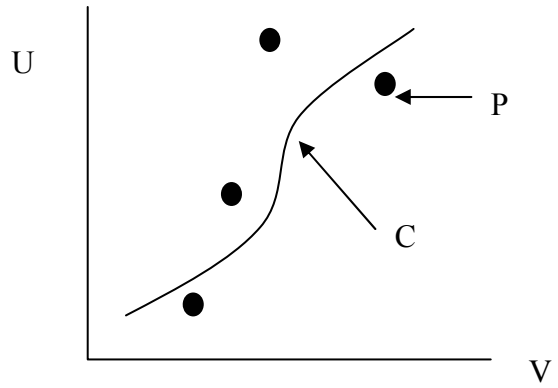
Regression Analysis

Def.: The process of **fitting a curve** through the cloud of point of a **scatterplot.**

The Process ➔ The Method of Least Squares

Of all curves approximating a given set of data points, the curve having the property that $D_1^2 + D_2^2 + \ldots\ldots\ldots + D_n^2$ is a minimum is called a **Best-Fitting Curve** or **Least-Squares Curve**.

D = The deviation or residual error between curve C and each point P.



Why regression analysis?
It helps in predicting one variable from the other (Is it always the case?)

Some Types and Equations of Regression Curves

1) Linear Regression (Straight line equation)

A mathematical expression that models the linear relationship between two variables (for example u and v).

$$u = a + bv$$

u = The **dependent** or regressed or predicted variable
v = The **independent** or regressor or predictor variable
a = The **intercept** of the line on the u axis

$$a = m_u - bm_v$$

b = The **slope** of the line

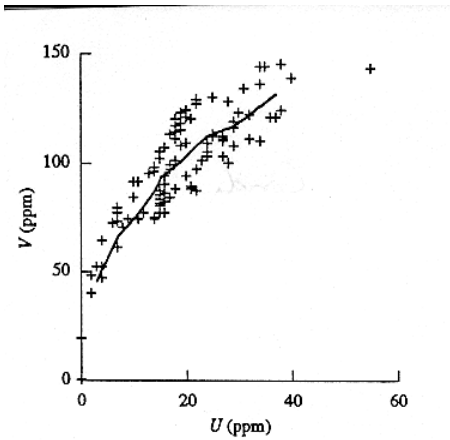
$$b = \rho \frac{S_u}{S_v}$$

Some Types and Equations of Regression Curves

2) Curvilinear Regression

a) Transform the data (i.e power, reciprocal, log, … etc.)
b) Use polynomials

1- Parabola (quadratic) curve
2- Cubic curve
3- n^{th}-degree curve
4- Hyperbola curve
5- Geometric curve
6- Exponential curve



q-q Plot

A method to compare two distributions

A graph on which quantiles from two distributions are plotted versus one another.

- **Identical distribution**

1.q-q plot shows a straight line relationship (i.e. x=y)
2.small departures from the x=y line reveals where the distributions differ.

- **Different distributions**

1.q-q plot does not come close to a straight line x=y.
2.If q-q plot shows a straight line other than x=y, the two distributions have the same shape but different location and spread statistical parameters.

q-q Plot

A method to compare two distributions

Table 3.2 Comparison of the V and U quantiles.

Cumulative Frequency	Quantile V	Quantile U	Cumulative Frequency	Quantile V	Quantile U
0.05	48.1	3.1	0.55	104.1	19.0
0.10	70.2	7.0	0.60	108.6	20.0
0.15	74.0	8.1	0.65	111.0	21.0
0.20	77.0	11.2	0.70	112.7	22.7
0.25	81.3	14.0	0.75	116.8	25.0
0.30	84.0	15.0	0.80	120.0	27.0
0.35	87.4	15.4	0.85	122.9	29.0
0.40	91.0	16.0	0.90	127.9	33.8
0.45	96.5	17.0	0.95	138.9	37.0
0.50	100.5	18.0			

q-q Plot

A method to compare two distributions

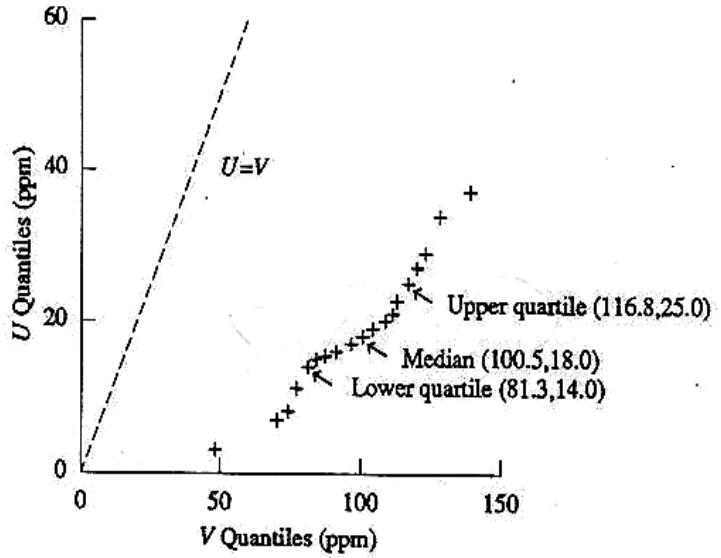


Figure 3.3 A q-q plot of the distribution of the 100 special U values versus the 100 V values. Note the different scales on the axes.