LEC3: DATA DESCRIPTION

1. UNIVARIATE DESCRIPTION (Chpt. 2)

- 1.1 DATA PRESENTATION
- 1.2 MEASURES OF CENTRAL TENDENCY
- 1.3 MEASURES OF VARIABILITY

2. BIVARIATE DESCRIPTION (Chpt. 3)

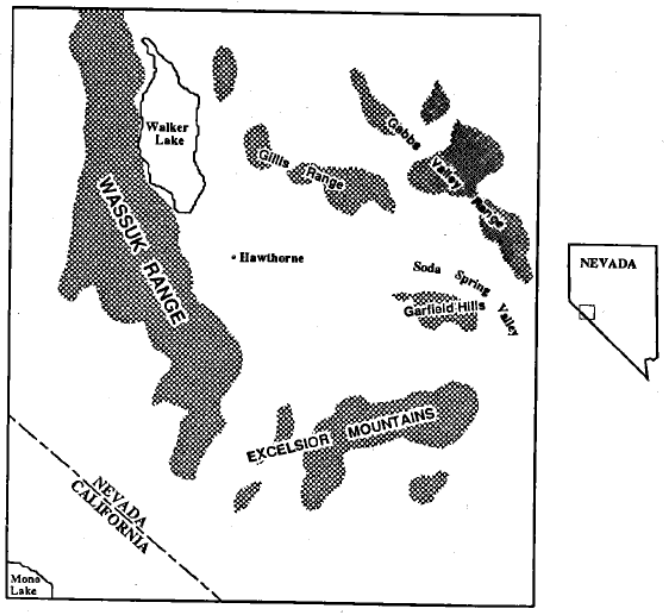
- 2.1 *q-q* PLOT AND SCATTER PLOT
- 2.2 CORRELATION COEFFICIENT
- 2.3 RNAK CORRELATION COEFFICIENT

Reading: This ppt and Chpt. 2 and 3 of the text

Let's review the datasets used in this book.

The Exhaustive Data Set

- The data set is derived from a **Digital Elevation Model (DEM)** from the **Walker Lake** area of Nevada in the western US (Figure 1.1);
- A data set consists of **3 variables** measured at each of **78,000 points on a 260m x 300m rectangular grid**;
- The **first two variables, V and U, are continuous** and their values range from 0 to several thousands and the **third variable, T, is discrete** and its value is either 1 or 2.



The continuous variables (*V* and *U*) could be:

- concentration of some pollutant
- thickness of a geological horizon
- soil moisture content
- aquifer permeability
- rainfall measurements
- air temperature
- wind speed
- the diameters of trees

Can you think another variable in your areas of interests?

The discrete variable (*T*) could be:

- color difference
- different species
- different rock types
- different soil lithology
- the presence or absence of a particular element in water, soil, or air sample

The Sample Data Set

The **sample data set** is a subset of the exhaustive data set. Using the sample data set we will address the following problems:

- The description of the important features of the data.
- The estimation of an average value over a large area.
- The estimation of an unknown value at a particular location.
- The estimation of an average value over small areas.
- The use of the available sampling to check the performance of an estimation methodology.
- The use of sample values of one variable to improve the estimation of another variable.
- The assessment of the uncertainty of our various estimates.

The Sampling History (Very important to know!)

1st campaign: sampled **V** at roughly 20 x 20 m^2 and got (13 * 15=) **195** data points for V;

2nd campaign: Each of the original 195 samples whose V value > 500 ppm was surrounded by 8 extra samples for **both V and U** located approximately at a 10 x 10 m^2 grid and got **150** data points for V and U;

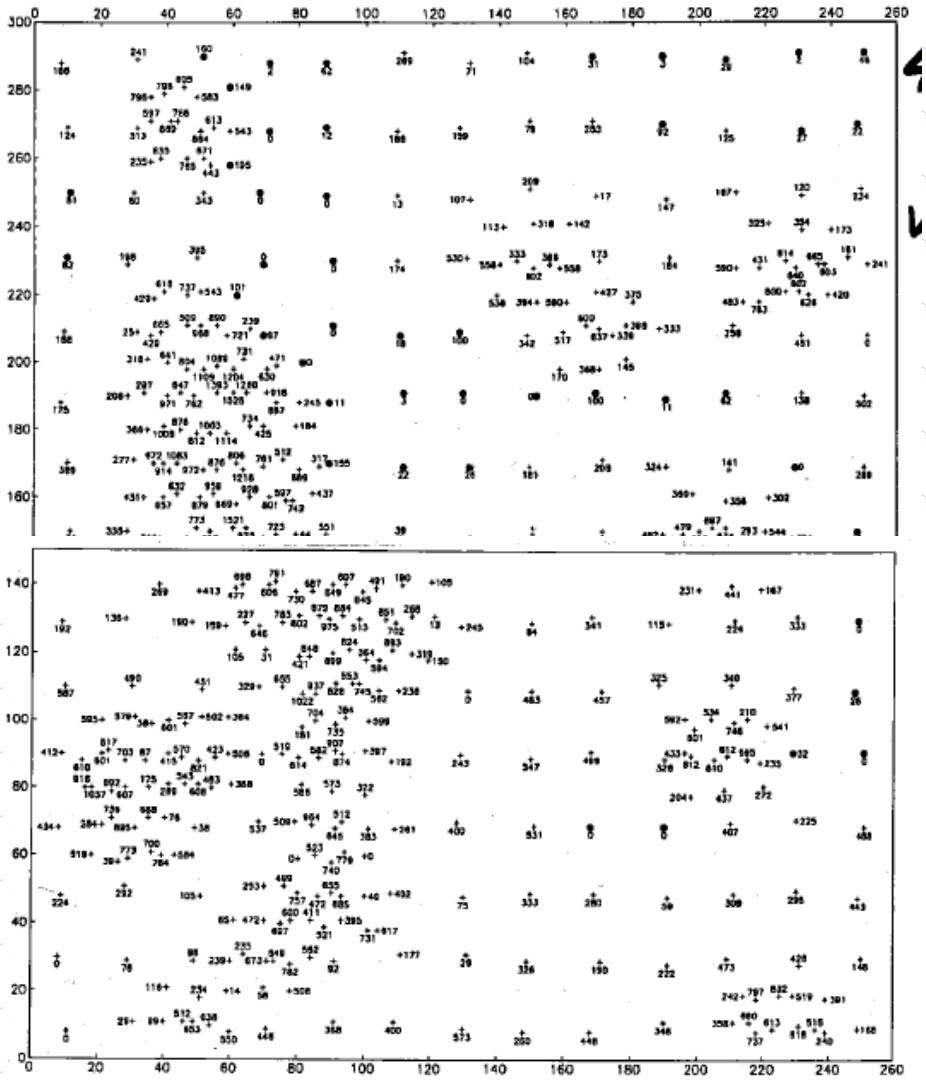
3rd campaign: added two more samples for V and U to those points with V > 500 ppm, one roughly 5m to the east and the other roughly 5 m to the west, and got **125** data points for V and U;

So, in the sample data set:

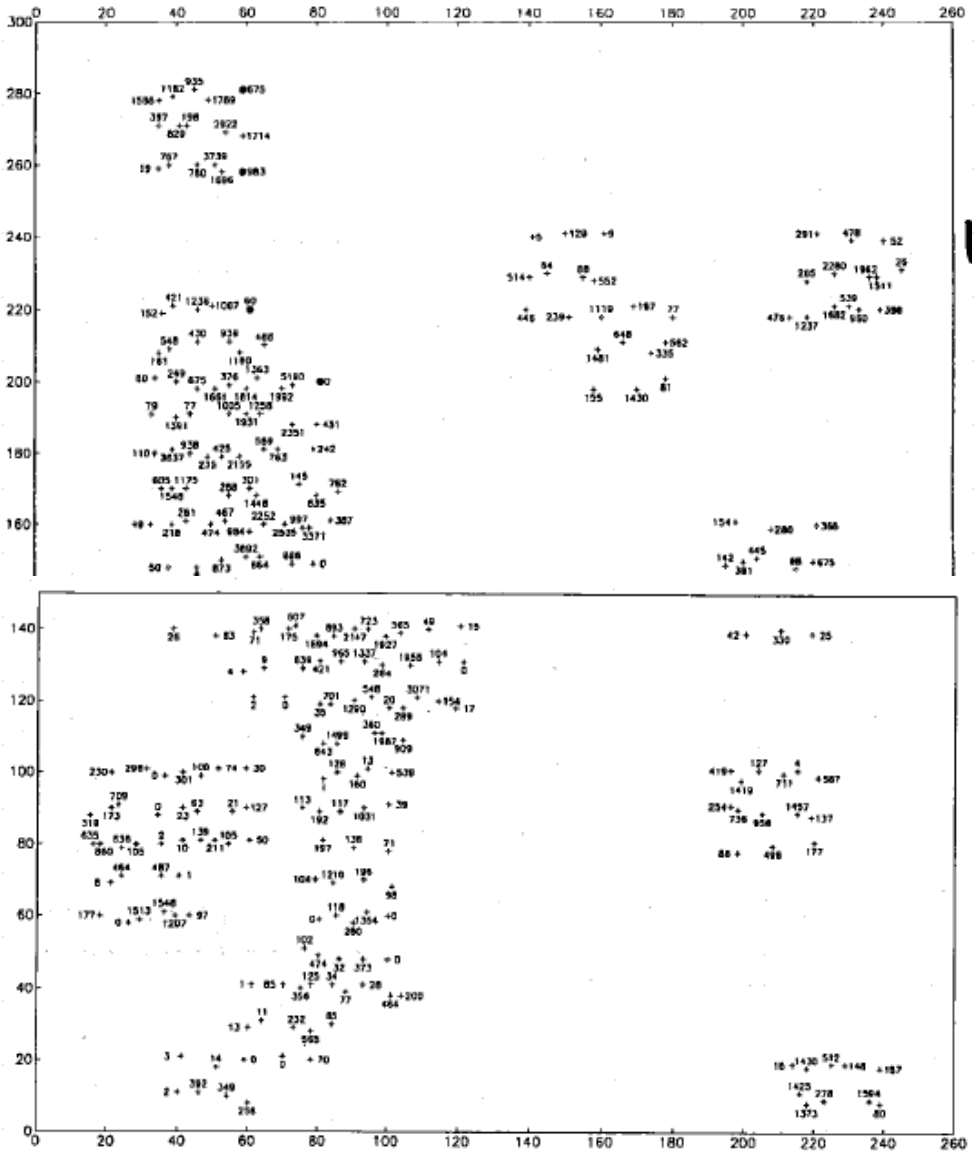
# of data points for V is: 195+150+125 = 470

# of data points for U is: 150+125 = 275

Distribution of 470 V values in the sample data set



Distribution of 275 U values in the sample data set



1. UNIVARIATE DESCRIPTION

1.1 Data presentation using tables and plots

- **Frequency Table and Histograms (*histogram(z)*)**
- **Cumulative Frequency Table and Histograms**
- **Normal and Lognormal Probability Plots (*normplot(z)*)**
- **Box-and-whisker plots (*boxplot(z)*)**

We selected 100 *V* values from the exhaustive data set on 10m x10m

How do you present your data using plots and graphs?

81	77	103	112	123	19	40	111	114	120
+	+	+	+	+	+	+	+	+	+
82	61	110	121	119	77	52	111	117	124
+	+	+	+	+	+	+	+	+	+
82	74	97	105	112	91	73	115	118	129
+	+	+	+	+	+	+	+	+	+
88	70	103	111	122	64	84	105	113	123
+	+	+	+	+	+	+	+	+	+
89	88	94	110	116	108	73	107	118	127
+	+	+	+	+	+	+	+	+	+
77	82	86	101	109	113	79	102	120	121
+	+	+	+	+	+	+	+	+	+
74	80	85	90	97	101	96	72	128	130
+	+	+	+	+	+	+	+	+	+
75	80	74	108	121	143	91	52	136	144
+	+	+	+	+	+	+	+	+	+
77	84	74	108	121	143	91	52	136	144
+	+	+	+	+	+	+	+	+	+
87	100	47	111	124	109	0	98	134	144
+	+	+	+	+	+	+	+	+	+

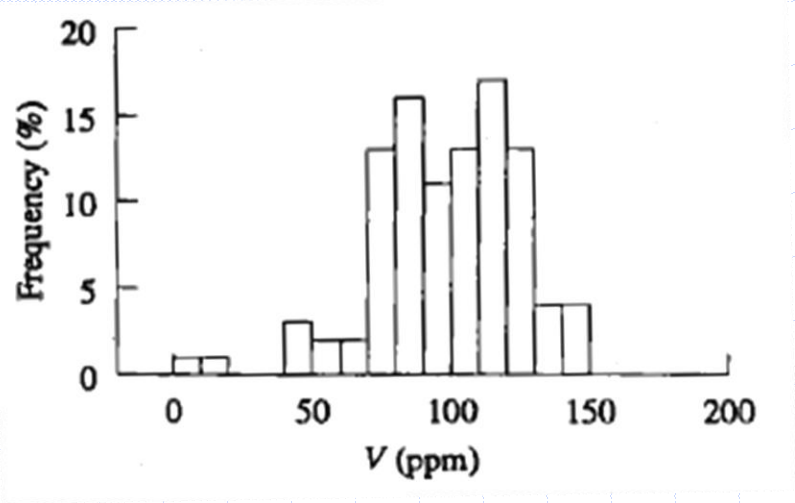
- ???
- ???
- ???
- ???
- ???

For the selected 100 *V* values

Frequency

Class	Number	Percentage
$0 \leq V < 10$	1	1
$10 \leq V < 20$	1	1
$20 \leq V < 30$	0	0
$30 \leq V < 40$	0	0
$40 \leq V < 50$	3	3
$50 \leq V < 60$	2	2
$60 \leq V < 70$	2	2
$70 \leq V < 80$	13	13
$80 \leq V < 90$	16	16
$90 \leq V < 100$	11	11
$100 \leq V < 110$	13	13
$110 \leq V < 120$	17	17
$120 \leq V < 130$	13	13
$130 \leq V < 140$	4	4
$140 \leq V < \infty$	4	4

Histogram

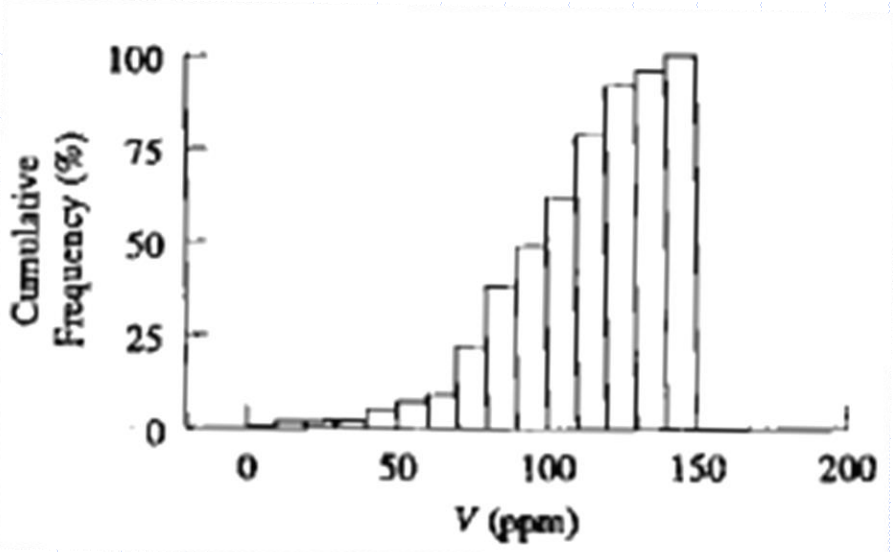


For the selected 100 *V* values

Cumulative Frequency

Class	Number	Percentage
$V < 10$	1	1
$V < 20$	2	2
$V < 30$	2	2
$V < 40$	2	2
$V < 50$	5	5
$V < 60$	7	7
$V < 70$	9	9
$V < 80$	22	22
$V < 90$	38	38
$V < 100$	49	49
$V < 110$	62	62
$V < 120$	79	79
$V < 130$	92	92
$V < 140$	96	96
$V < \infty$	100	100

Cumulative histogram



For the selected 100 *V* values

Normal Probability Plot

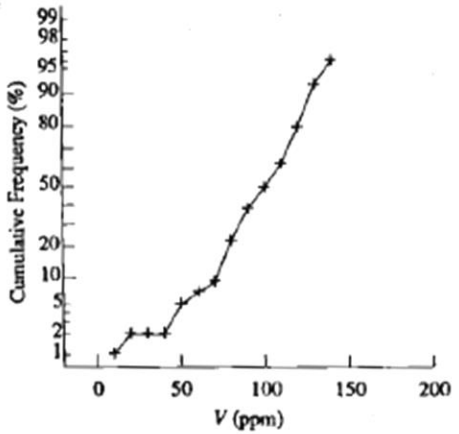


Figure 2.4 A normal probability plot of the 100 selected *V* data. The y-axis has been scaled in such a way that the cumulative frequencies will plot as a straight line if the distribution of *V* is Gaussian.

Log-normal Probability Plot

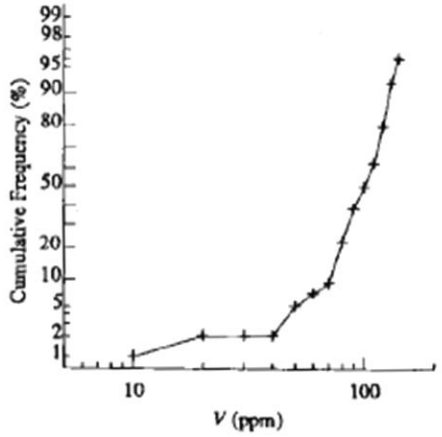


Figure 2.5 A lognormal probability plot of the 100 selected *V* data. The y-axis is scaled so that the cumulative frequencies will plot as a straight line if the distribution of the logarithm of *V* is Gaussian.

Normal Probability Plot

The plot The normal probability plot is *a graphical technique to identify substantive departures from normality.* This includes identifying outliers, skewness, kurtosis, a need for transformations, and mixtures. Normal probability plots are made of raw data, residuals from model fits, and estimated parameters.

In a normal probability plot (also called a "normal plot"), the sorted data are plotted vs. values selected to make the resulting image look close to a straight line if the data are approximately normally distributed. Deviations from a straight line suggest departures from normality. The plotting can be manually performed by using a special graph paper, called normal probability paper. With modern computers normal plots are commonly made with software.

The normal probability plot is ***a special case of the q–q probability plot for a normal distribution.*** The theoretical quantiles are generally chosen to approximate either the mean or the median of the corresponding order statistics.

Example

We have **118 Z values** along a cross section separated by **2m.** (Z can be elevation of ground surface, air temperature, precipitation, concentration of a chemical, thickness of a formation, water table, permeability, soil moisture, etc.)

1. How do you present your data?

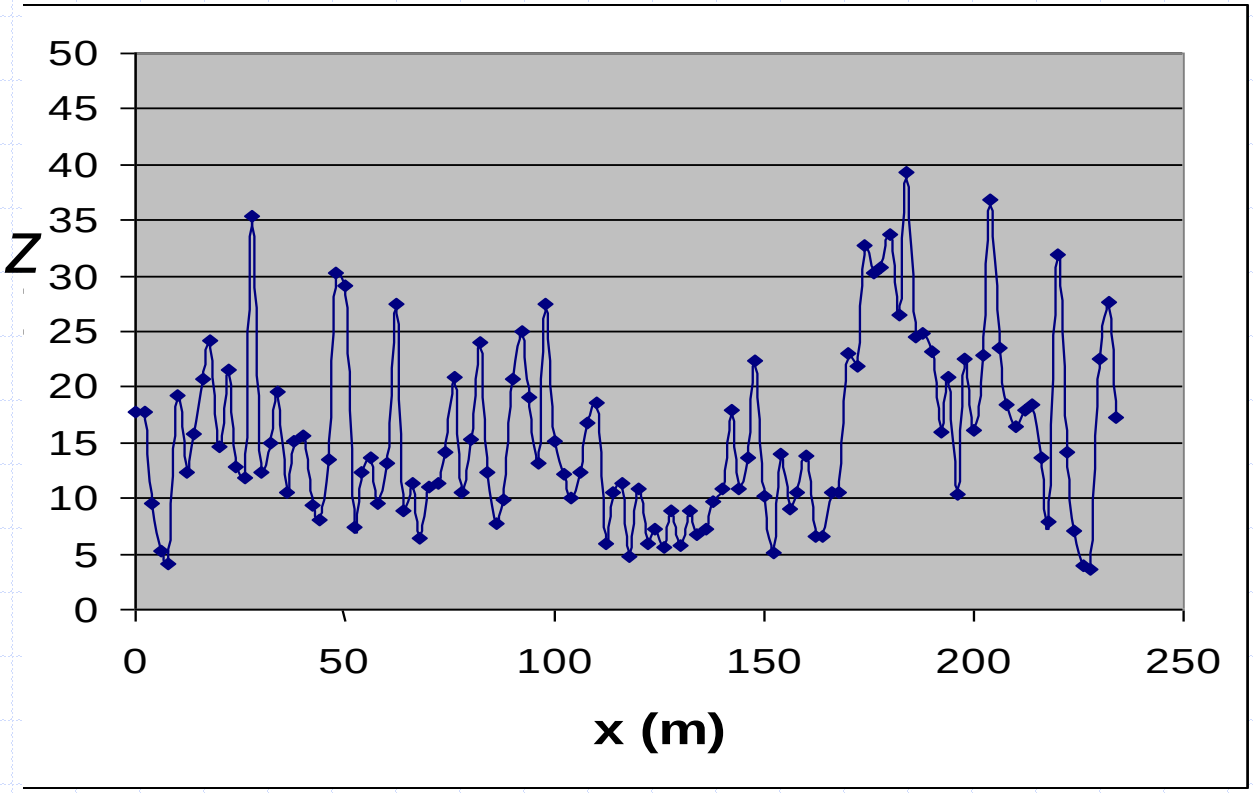
2. What information can you obtain from the data?

x (m)	Zi		x (m)	Zi		x (m)	Zi
0	17.7		82	24.0		164	6.5
2	17.8		84	12.3		166	10.6
4	9.5		86	7.8		168	10.6
6	5.2		88	9.9		170	23.0
8	4.1		90	20.7		172	21.8
10	19.2		92	25.0		174	32.8
12	12.4		94	19.1		176	30.2
14	15.8		96	13.1		178	30.8
16	20.8		98	27.4		180	33.7
18	24.1		100	15.2		182	26.5
20	14.7		102	12.2		184	39.3
22	21.6		104	10.1		186	24.5
24	12.8		106	12.3		188	24.9
26	11.9		108	16.7		190	23.2
28	35.4		110	18.6		192	16.0
30	12.3		112	6.0		194	20.9
32	14.9		114	10.6		196	10.3
34	19.6		116	11.3		198	22.6
36	10.6		118	4.7		200	16.2
38	15.1		120	10.9		202	22.9
40	15.6		122	6.0		204	36.9
42	9.3		124	7.2		206	23.5
44	8.1		126	5.6		208	18.5
46	13.5		128	8.9		210	16.4
48	30.2		130	5.8		212	17.9
50	29.1		132	8.9		214	18.5
52	7.4		134	6.7		216	13.6
54	12.3		136	7.2		218	7.9
56	13.6		138	9.7		220	31.9
58	9.5		140	10.8		222	14.1
60	13.1		142	17.9		224	7.1
62	27.4		144	10.9		226	3.9
64	8.8		146	13.7		228	3.7
66	11.4		148	22.3		230	22.5
68	6.4		150	10.2		232	27.6
70	11.0		152	5.1		234	17.3
72	11.4		154	13.9			
74	14.1		156	9.0			
76	20.9		158	10.6			
78	10.6		160	13.8			
80	15.3		162	6.5			

XY plot of 118 Z values

There are many ways to present your data. Here we only mention a few.

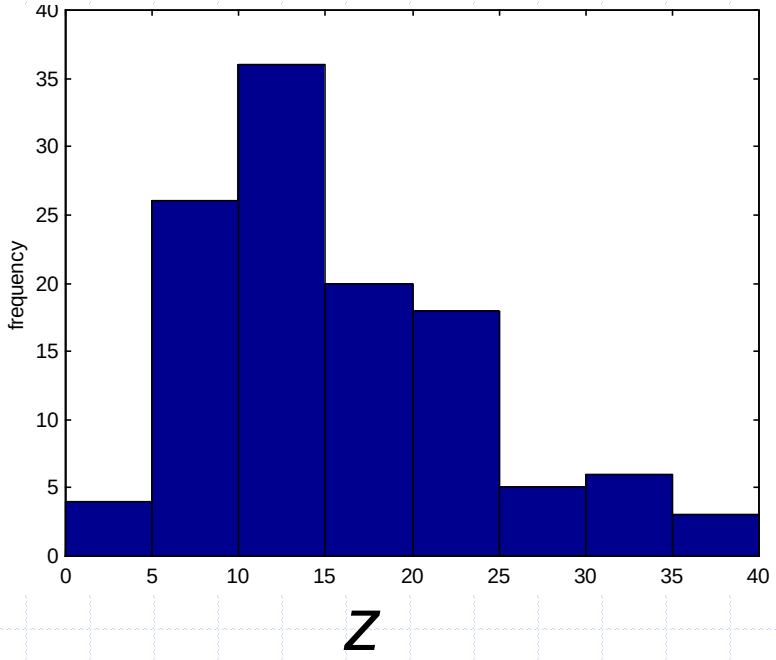
1. **XY plot** with X being the coordinate along x-axis and Y being the values of the variable measured



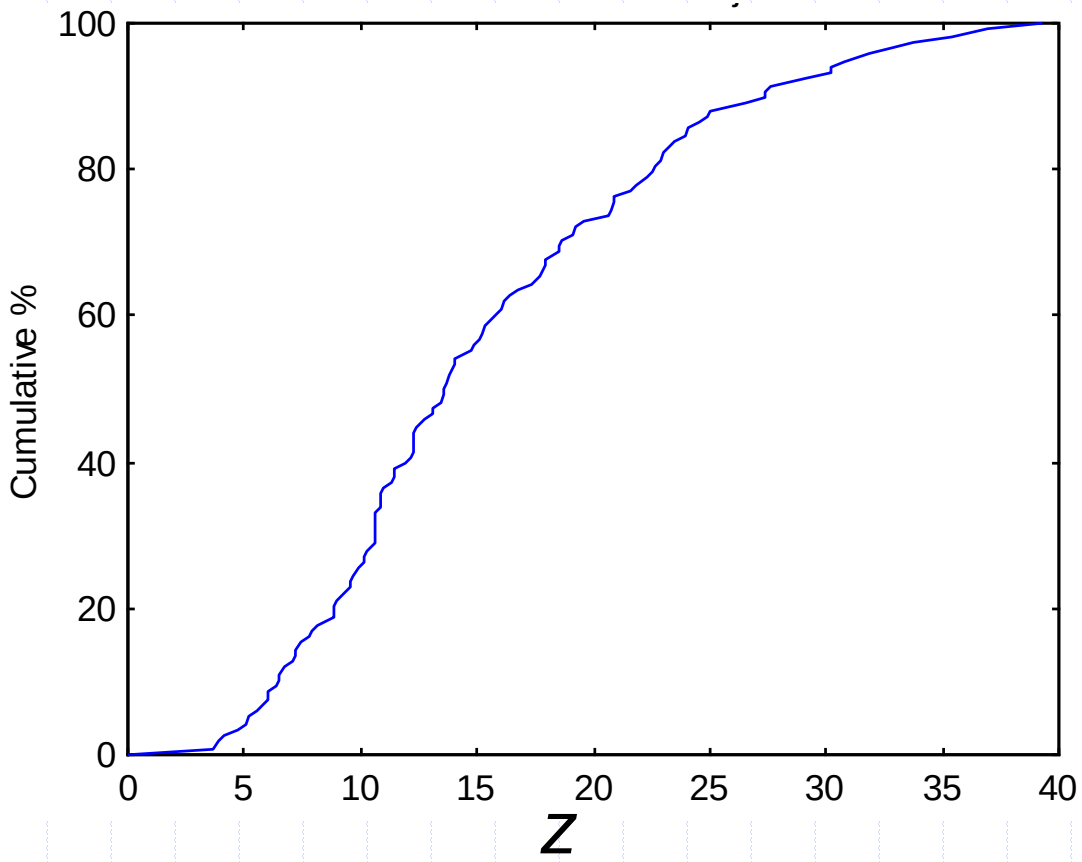
Histogram of 118 Z values

A bar chart in which a continuous variable is divided into discrete categories and the number or proportion of observations that fall into each category are represented by the area of the corresponding bars.

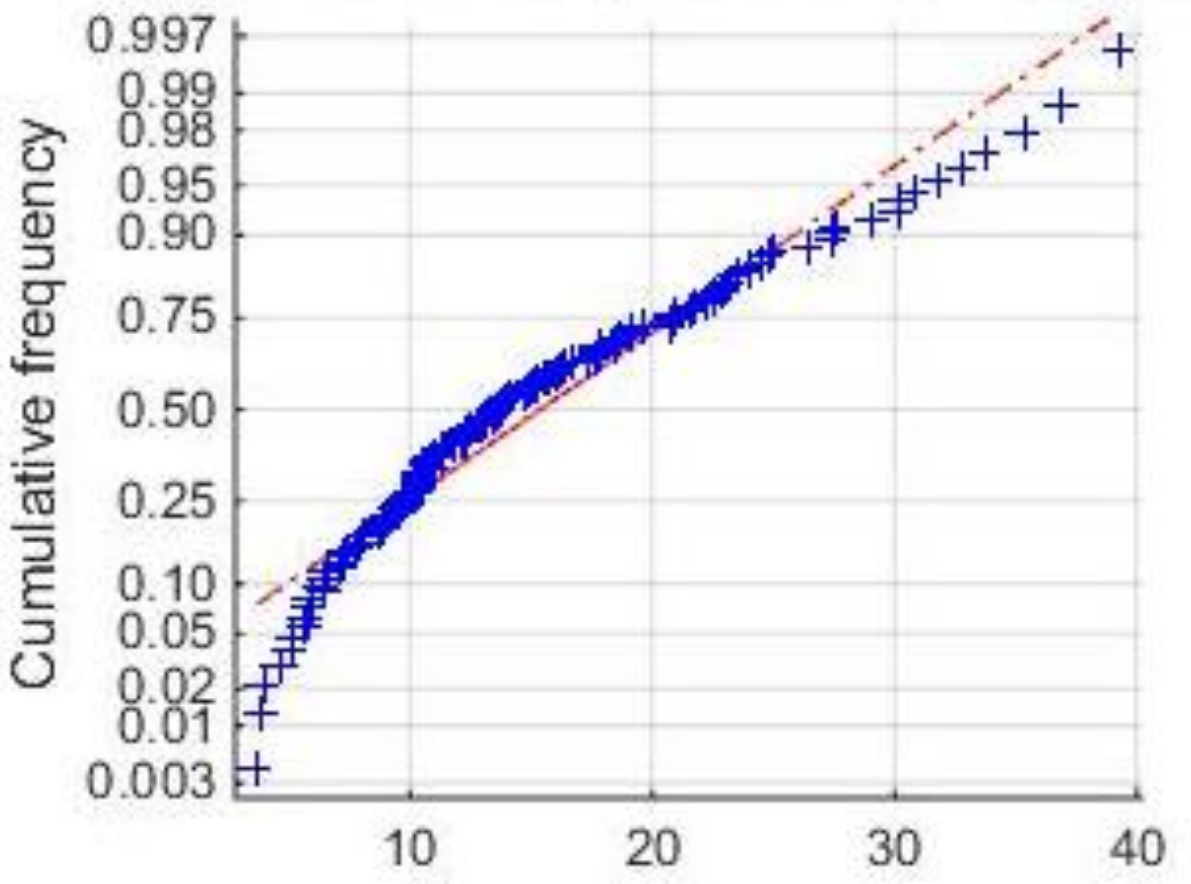
- Histogram is strongly affected by the number of categories.
- *Is it normally distributed? How do you check it?*



Cumulative Frequency Curve of 118 Z values

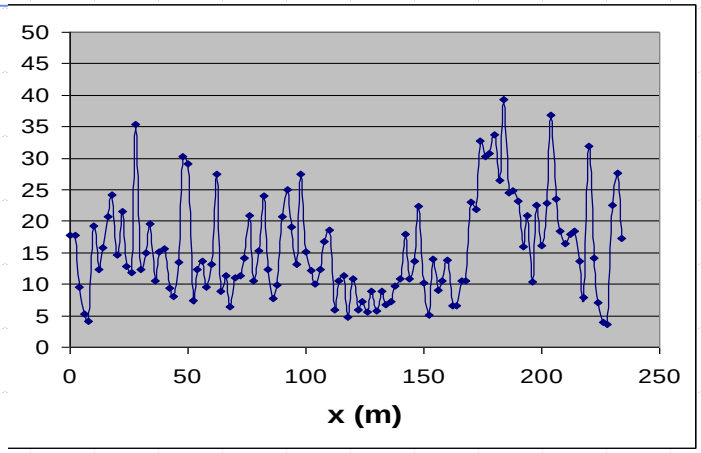


Normal Probability Plot of 118 Z values

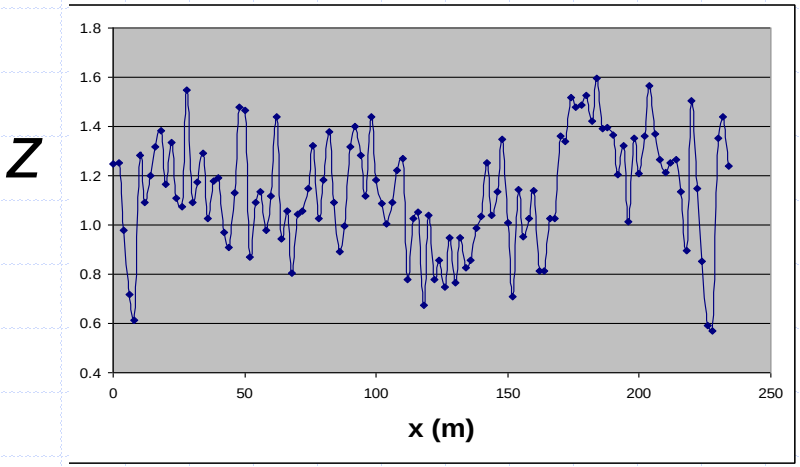


When a data is not normally distributed, what do you do?

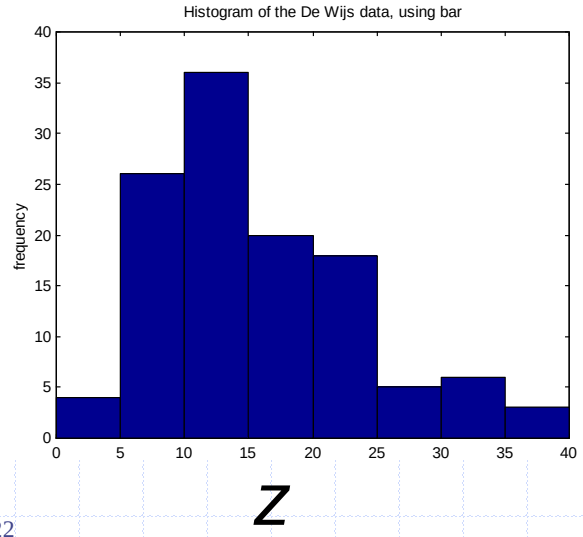
Original data



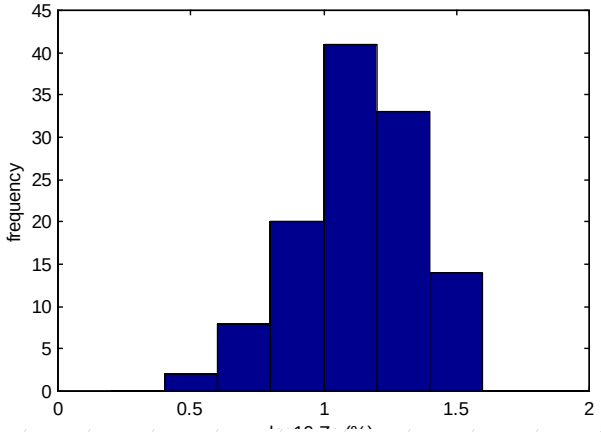
Log-Transformation



Histogram of original data



Histogram of Log-Transformation



1.2 MEASURES OF CENTRAL TENDENCY

***Central tendency is a typical or representative score*.** **If the mayor is asked to provide a single value which best describes the income level of the city, he or she would answer with a measure of central tendency.**

- **???**
- **???**
- **???**

1.2 MEASURES OF CENTRAL TENDENCY

***Central tendency is a typical or representative score*.** **If the mayor is asked to provide a single value which best describes the income level of the city, he or she would answer with a measure of central tendency. The three measures of central tendency that will be discussed here are**

- **Mode – the value that occurs with the greatest frequency or the largest number;**
- **Median – the value midway in the frequency distribution; *it is 50th percentile, 5th decile, or 2nd quartile.***
- **Mean – the arithmetic average**

Mode (M_0)

The mode, symbolized by M_0, is the **most frequently occurring score** value. If the scores for a given sample are:

32 32 35 36 37 38 38 39 39 39 40 40 42 45

then the mode would be **39** because a score of 39 occurs 3 times, more than any other score. The mode may be seen on a frequency distribution as the score value which corresponds to the highest point. For example, the following is a frequency polygon of the data presented above:

A distribution may have more than one mode if the two most frequently occurring scores occur the same number of times. For example, if the earlier score distribution were modified as follows:

32 32 32 36 37 38 38 39 39 39 40 40 42 45

then there would be two modes, **32** and **39**. Such distributions are called **bimodal.**

Mode (M_0)

The mode is not sensitive to extreme scores. Suppose the original distribution was modified by changing the first number 32 to **2** and the last number, 45 to **95** as follows:

2 32 35 36 37 38 38 **39** **39** **39** 40 40 42 95

The mode would still be **39**.

In any case, the mode is a quick and dirty measure of central tendency. Quick, because it is easily and quickly computed. Dirty because it is not very useful; that is, it does not give much information about the distribution.

Median (M_d)

The median, symbolized by M_d, is the score value which **cuts the distribution in half**, such that half the scores fall above the median and half fall below it.

Computation of the median is relatively straightforward.

- ✓ The first step is to rank order the scores from lowest to highest.
- ✓ The procedure branches at the next step: one way if there are an odd number of scores in the sample distribution, another if there are an even number of scores.

If there is an **odd** number of scores as in the distribution below:

32	32	35	36	36	37	38	38	39	39	39	40	40	45	46

then the median is simply the **middle number**. In the case above the median would be the number **38**, because there are 15 scores all together with 7 scores smaller and 7 larger.

Median (M_d)

If there is an **even** number of scores, as in the distribution below:

32	35	36	36	37	38								
						38	39						
								39	39	40	40	42	45

then the median is the midpoint between the two middle scores: in this case the value **38.5**. It was found by adding the two middle scores together and dividing by two (38 + 39)/2 = 38.5. If the two middle scores are the same value then the median is that value.

Median (M_d)

The median, like the mode, ***is not effected by extreme scores,*** as the following distribution of scores indicates:

2	35	36	36	37	38								
						38	**39**						
								39	39	40	40	42	95

The median is still the value of **38.5**. The median is not as quick and dirty as the mode, but generally it is not the preferred measure of central tendency.

Mean ($\overline{X}$)

The mean, symbolized by $\overline{X}$, is *the sum of the scores divided by the number of scores*. The following formula both defines and describes the procedure for finding the mean:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where X is the sum of the scores and n is the number of scores. Application of this formula to the following data

32 35 36 36 37 38 38 39 39 39 40 40 42 45

yields the following results:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{536}{14} = 38.29$$

Mean ($\overline{X}$)

Use of means as a way of describing a set of scores is fairly common; batting average, bowling average, grade point average, and average points scored per game are all means. Note the use of the word "average" in all of the above terms. In most cases when the term "average" is used, it refers to the mean, although not necessarily. When a politician uses the term "average income", for example, he or she may be referring to the mean, median, or mode.

Note: **1)** Two desirable properties of the sample mean

- The sample mean is an unbiased estimate of the population mean;
- The sample mean is closer to the population mean than any other unbiased estimate (e.g., the median).

2) *The mean is sensitive to extreme scores.* For example, the mean of the following data is **42.86**, somewhat larger than the preceding example (**38.29**).

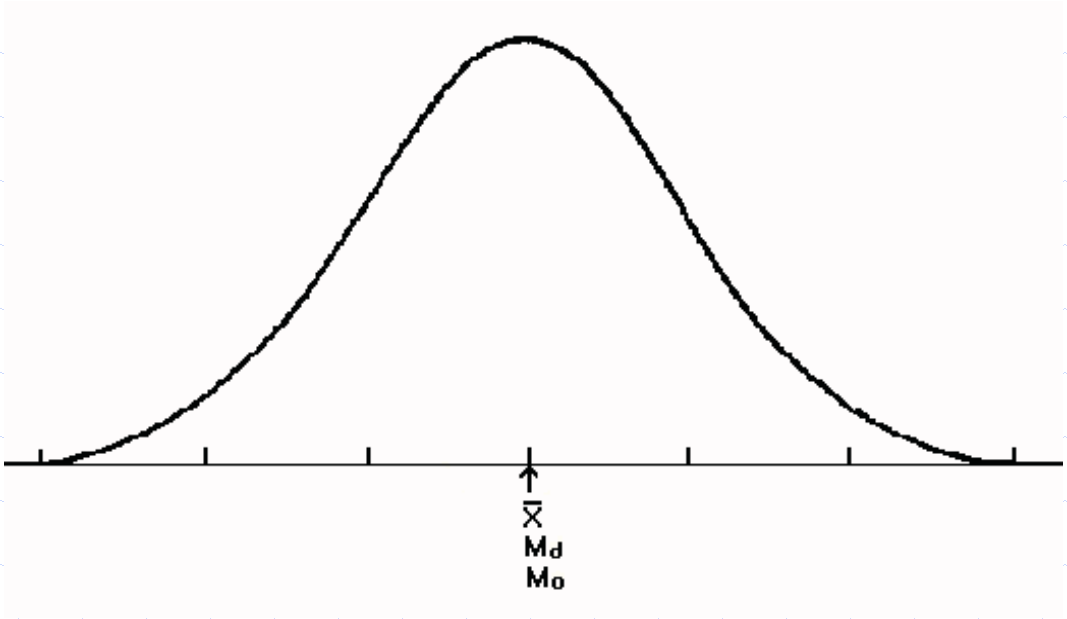
32 35 36 36 37 38 38 39 39 39 40 40 42 **95**

In most cases the **mean is the preferred measure of central tendency**, both as a description of the data and as an estimate of the parameter.

Symmetrical Distribution and Skewness

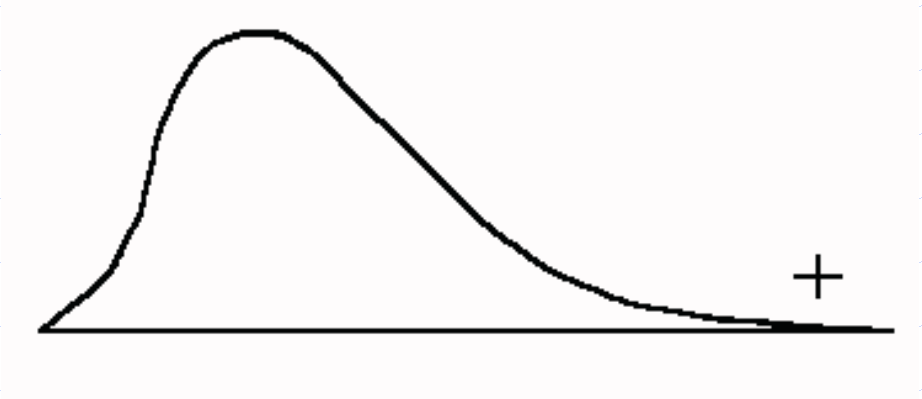
Skewness refers to the asymmetry of the distribution, such that a symmetrical distribution exhibits no skewness.

In a symmetrical distribution the mean, median, and mode all fall at the same point, as in the following distribution.



A positively skewed distribution

is asymmetrical and points in the positive direction. If a test was very difficult and almost everyone in the class did very poorly on it, the resulting distribution would most likely be positively skewed.



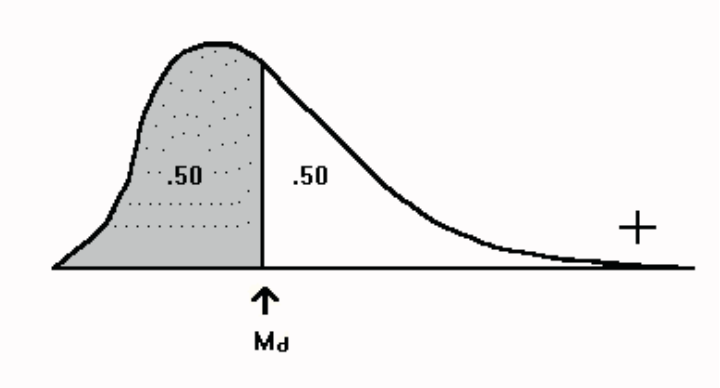
What is the relation among mode, medium, and mean for a positively skewed distribution?

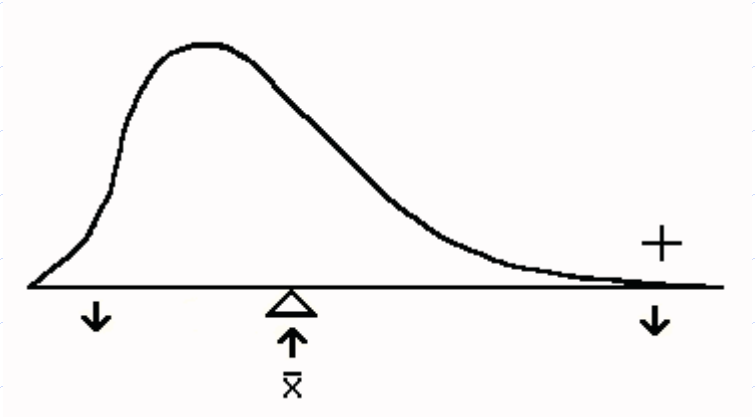
In the case of a positively skewed distribution

The mode is smaller than the median, which is smaller than the mean. This relationship exists because **the mode** is the point on the x-axis corresponding to the highest point, that is the score with greatest value, or frequency.

The median is the point on the x-axis that cuts the distribution in half, such that 50% of the area falls on each side.

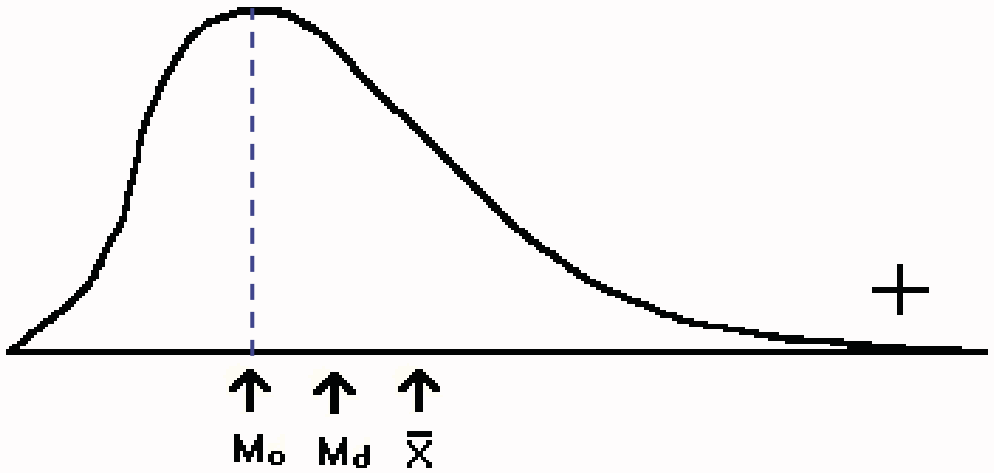
The mean is the balance point of the distribution. The mean is pulled in the direction the distribution is skewed. For the positively skewed distribution, the mean would be pulled toward larger numbers.





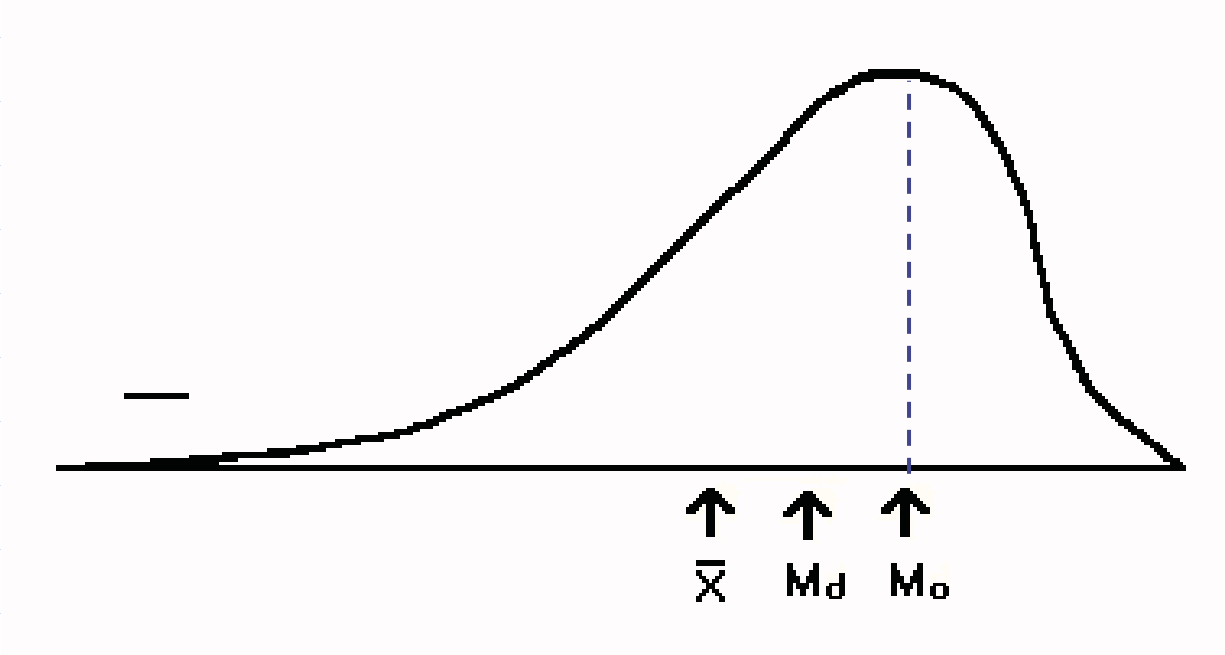
For a positively skewed distribution

One way to remember the order of the mean, median, and mode in a skewed distribution is to remember that the mean is pulled in the direction of the extreme scores. In a positively skewed distribution, the extreme scores are larger, thus the mean is larger than the median.



For a negatively skewed distribution

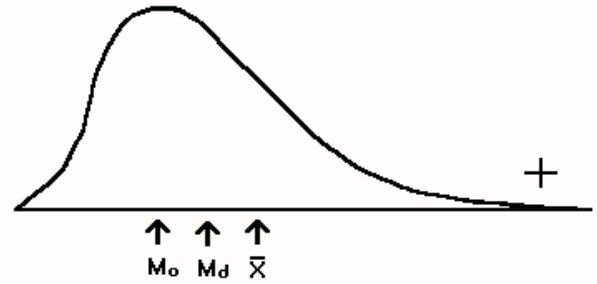
The order of the measures of central tendency would be the opposite of the positively skewed distribution, with the mean being smaller than the median, which is smaller than the mode.

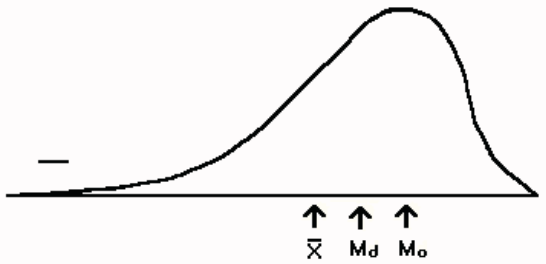


Skewness Coefficient

$$k_s = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^3 \Big/ s^3$$

This is a dimensionless number. A symmetric distribution has k_s zero; if the data contain many values slightly smaller than the mean and a few values much larger than the mean, k_s is positive; if there are many values slightly larger and a few values much smaller than the mean, k_s is negative.





Quantiles –successive divisions of a distribution

- **Percentile** – each category is a percentile if we rank all observations in a sample and then divide the rank into 100 equal-sized categories,

  Percentile of x_i = 100(rank of x_i/n)*

  where n is the number of observations.

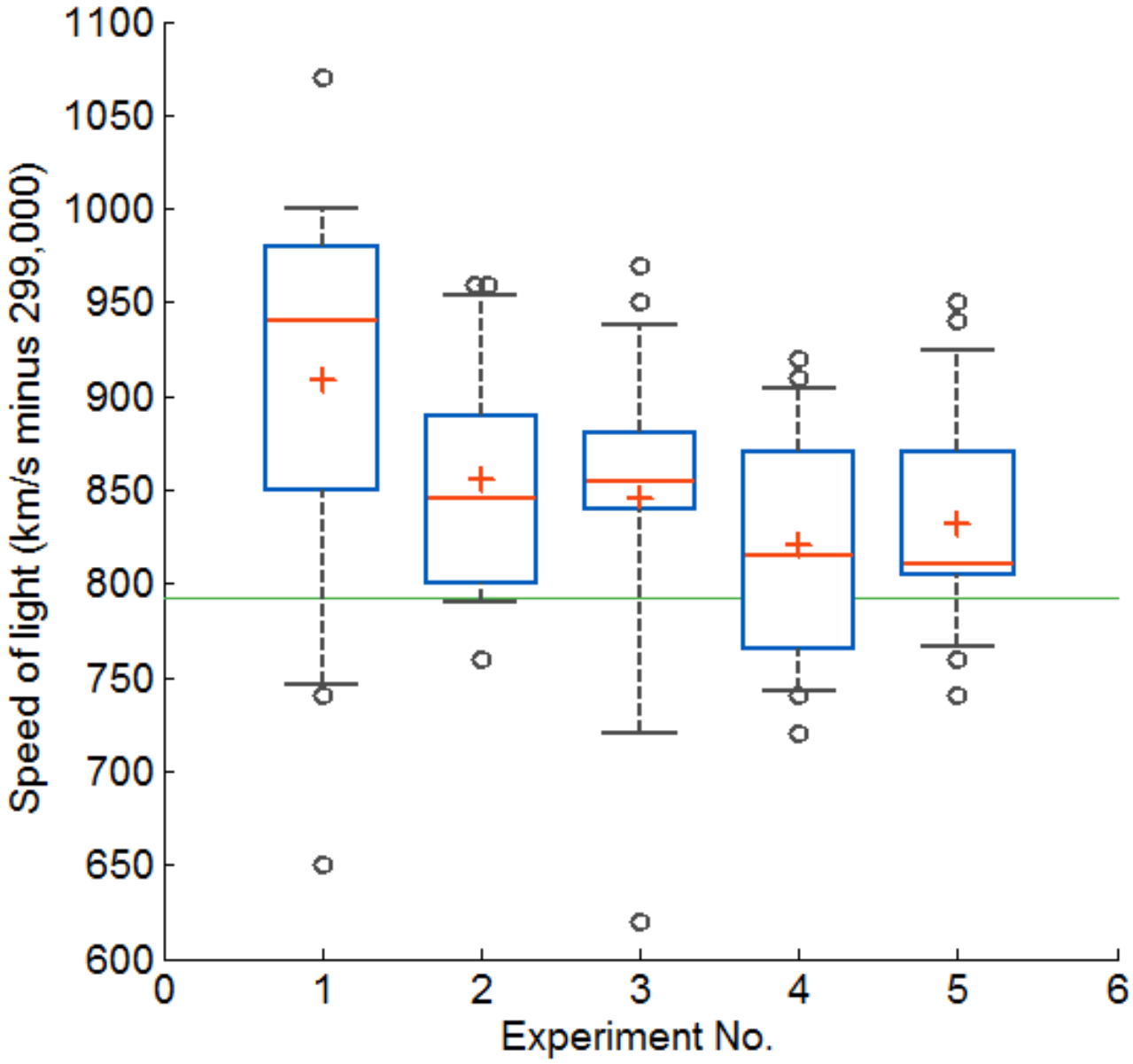
- **Decile** - each category is a decile if the rank is divided into 10 equal-sized categories

- **Quartile** - each category is a quartile the rank is divided into 4 equal-sized categories.

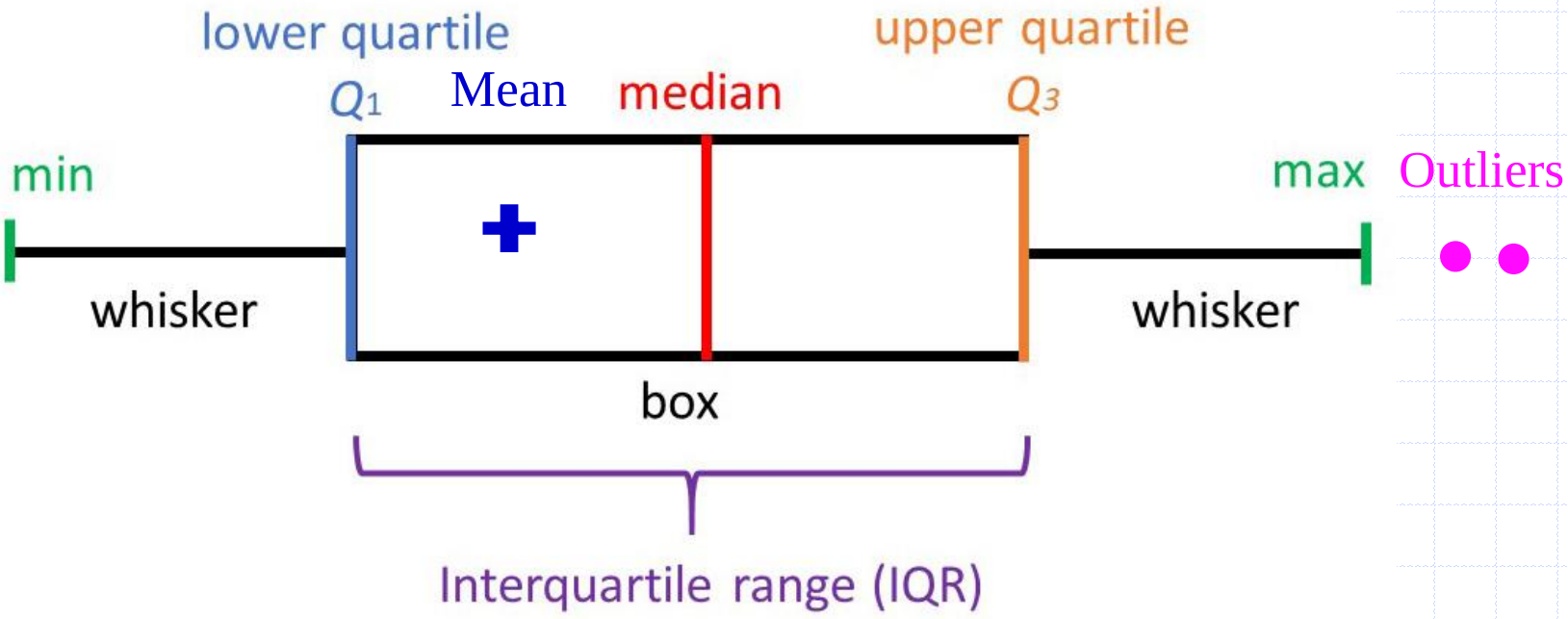
  5th and 95th percentile

  25th and 75th percentile (or 1st and 3rd quartiles or lower and upper quartiles)

  50th percentile is 5th decile, the 2nd quartile, or the median.


Speed of light (km/s minus 299,000)
1100
1050
1000
950
900
850
800
750
700
650
600
0
1
2
3
4
5
6
Experiment No.

Box-and-whisker plots



$IQR=Q_3-Q_1$

四分距

1.3 MEASURES OF VARIABILITY

Variability refers to the spread or dispersion of scores. A distribution of scores is said to be highly variable if the scores differ widely from one another.

Three statistics will be discussed which measure variability:

What are they ?

1.3 MEASURES OF VARIABILITY

Three statistics will be discussed which measure variability:

- **the range,**
- **the variance or standard deviation**
- **coefficient of variation**

The latter two are very closely related and will be discussed in the same section.

Maximum is the largest value in the data set.

Minimum is the smallest value in the data set.

Range

The Range is the maximum score minus the minimum score. It is a quick and dirty measure of variability, although when a test is given back to students they very often wish to know the range of scores. Because the range is greatly affected by extreme scores, it may give a distorted picture of the scores. The following two distributions have the same range, 13, yet appear to differ greatly in the amount of variability.

Distribution 1:	32	35	36	36	37	38	40	42	42	43	45
Distribution 2:	32	32	33	33	33	34	34	34	34	34	45

For this reason, among others, the range is **not** the most important measure of variability.

The Variance and Standard Deviation

Variance is a measure of variability. It is easier to define the variance with an algebraic expression than words, thus the following formula:

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

Standard Deviation are root of the variance, i.e., s and σ for a population or sample, respectively. It measures variability in units of measurement, while the variance does so in units of measurement squared. For this reason, the standard deviation is usually the preferred measure when describing the variability of distributions.

A formula suitable for computation with a calculator:

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

Coefficient of Variation (C_v)

is a dimensionless measure of variability expressed as a fraction of the mean

$$C_v = \frac{s}{\overline{X}}$$

where *s* is the standard deviation and $\overline{X}$ is the mean.

Why do we need coefficient of Variation?

Coefficient of Variation (C_v)

Example:

Dataset A: 1, 3, 5 ; Dataset B: 10, 15, 20

Which dataset has larger variation? A or B?

Coefficient of Variation (C_v)

Example:

Data set A: 1, 3, 5 ; Data set B: 10, 15, 20

For A: $\overline{X}$ = (1+3+5)/3 = 3 ;

s^2 = [(1-3)^2+(3-3)^2 +(5-3)^2]/3 = 8/3 = 2.67

C_v = (8/3)^1/2/3 = **0.54**

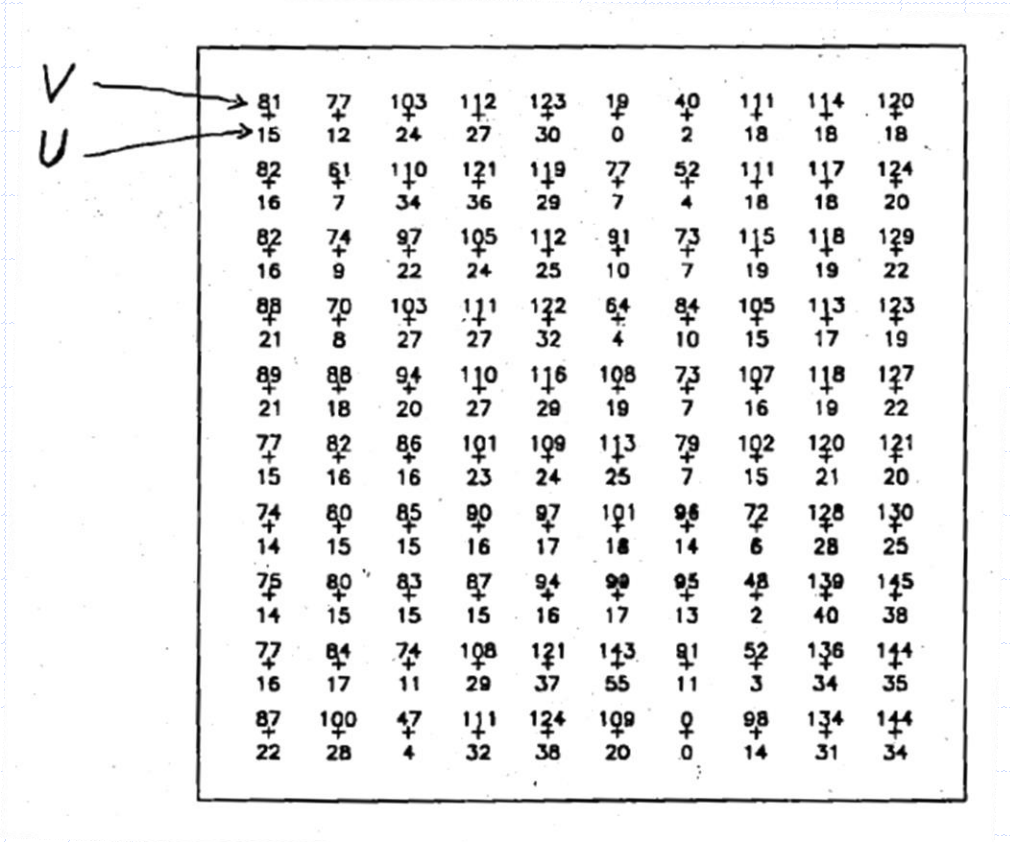
For B: $\overline{X}$ = (10+15+20)/3 = 15 ;

s^2 = [(10-15)^2+(15-15)^2 +(20-15)^2]/3 = 50/3 = 16.67

C_v = (50/3)^^1/2/15 = **0.27**

2. BIVARIATE DESCRIPTION

Location map of the 100 Vs and Us. Vs are above and U are below each measurement location (the plus)

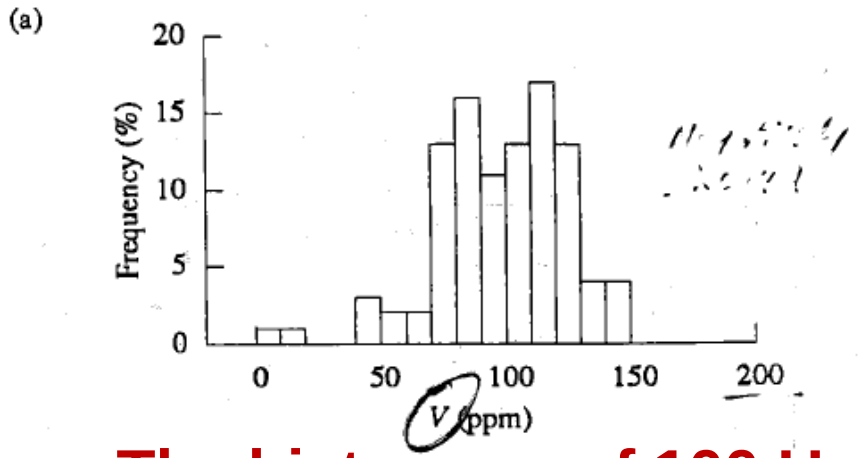


- ***q-q plot***
- ***Scatterplot***

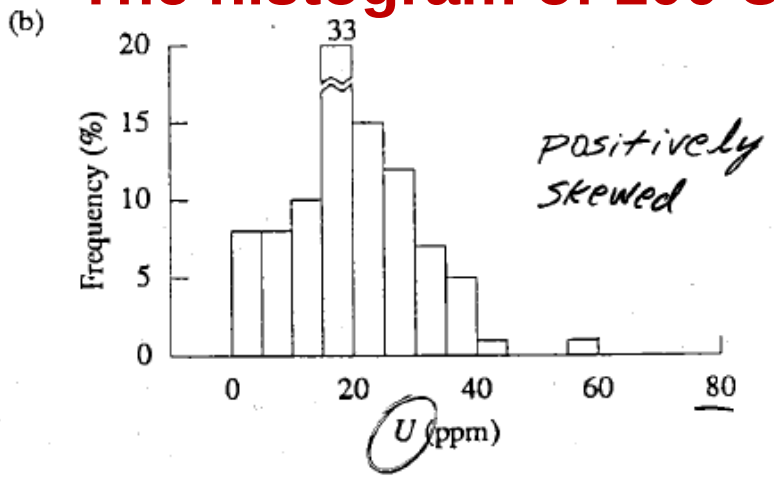
- ***Correlation Coefficient (r)***
- ***Coefficient of Determination (R^2)***
- ***Rank Correlation Coefficient (r_{rank})***

2.1 Comparison of Distributions of 100 *V* and *U*

The histogram of 100 V



The histogram of 100 U



Their statistics

	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

Comparison of the *V* and *U* quantiles

Cumulative Frequency	Quantile		Cumulative Frequency	Quantile	
	V	U		V	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

q – q plot is a graph on which the **quantiles** from two distributions are plotted versus one another. It is a good visual comparison of two distributions.

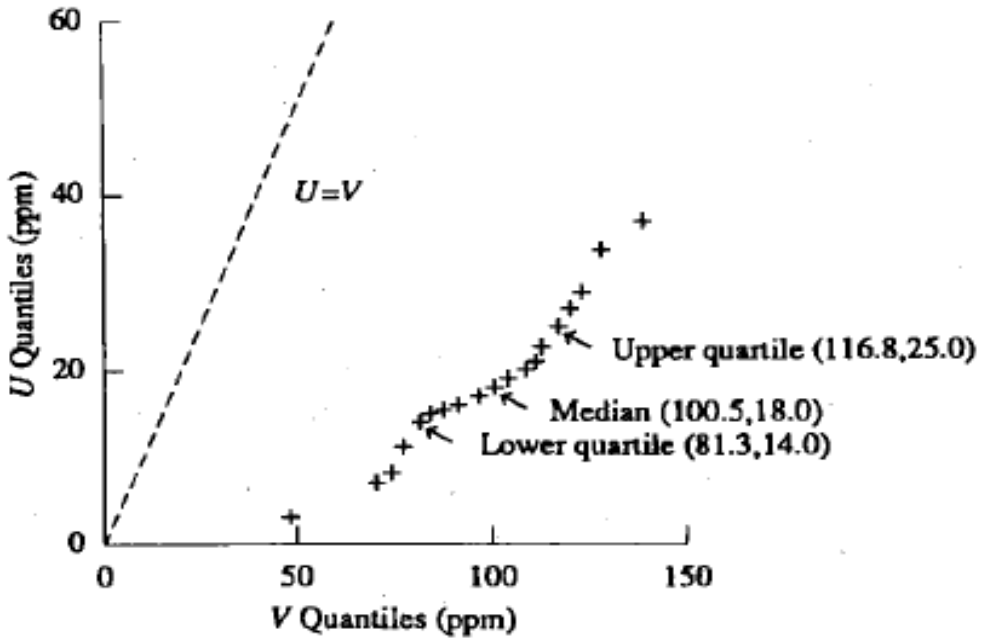


Figure 3.3 A q-q plot of the distribution of the 100 special U values versus the 100 V values. Note the different scales on the axes.

Note the different scales on the axes!

2.2 Scatterplot

Scatterplot is a x-y graph of the data on which the x-coordinate corresponding to the value of one variable and the y-coordinate to the value of the other variable.

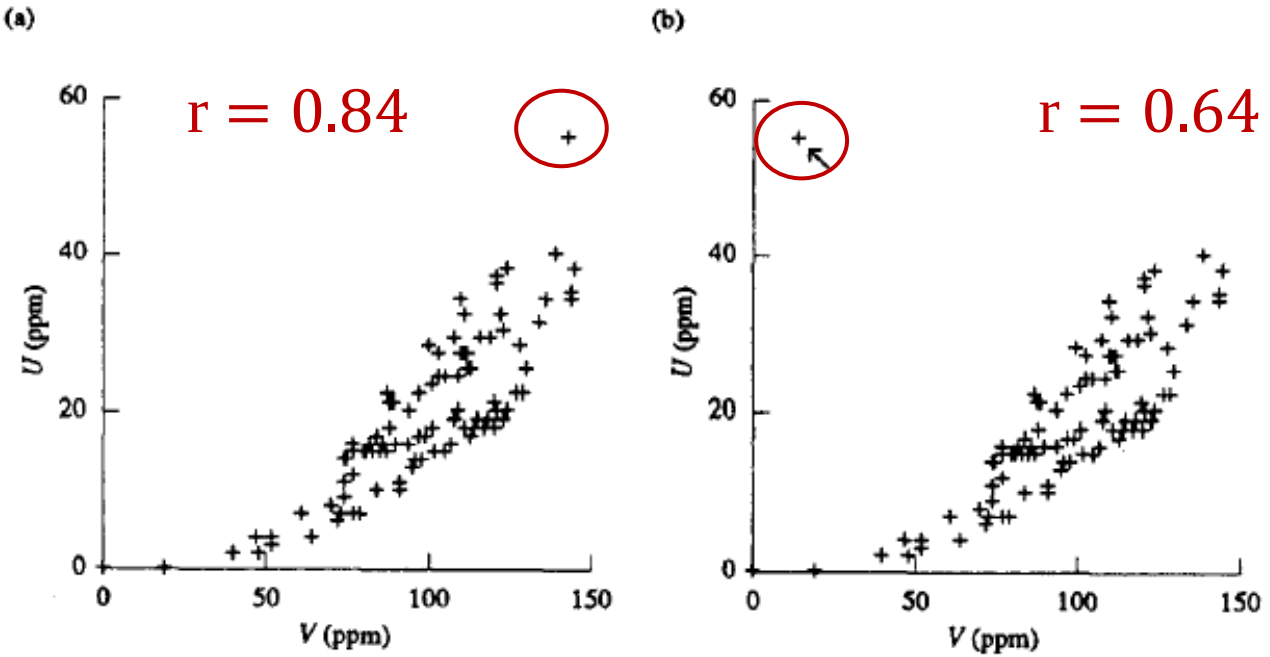


Figure 3.4 Scatterplot of 100 U versus V values. The actual 100 data pairs are plotted in (a). In (b) the V value indicated by the arrow has been "accidentally" plotted as 14 ppm rather than 143 ppm to illustrate the usefulness of the scatterplot in detecting errors in the data.

2.3 Correlation

The **correlation coefficient** for short is a measure of the degree of **linear** relationship between two variables, usually labeled X and Y.

$$r = \frac{COV(X,Y)}{s_X s_Y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{s_X s_Y}$$

where **COV(X,Y)** is the covariance of X and Y, s_X and s_Y are the standard deviation of X and Y, respectively.

While in regression the emphasis is on predicting one variable from the other, in correlation the emphasis is on the degree to which a linear model may describe the relationship between two variables. In regression the interest is directional, one variable is predicted and the other is the predictor; in correlation the interest is non-directional, the relationship is the critical aspect.

Correlation Coefficient

- The correlation coefficient may take on any value between plus and minus one. $-1 \leq r \leq 1$
- The sign of the correlation coefficient (+ , -) defines the direction of the relationship, either positive or negative. A **positive** correlation coefficient means that as the value of one variable increases, the value of the other variable increases; as one decreases the other decreases. A **negative** correlation coefficient indicates that as one variable increases, the other decreases, and vice-versa.
- Taking the absolute value of the correlation coefficient measures the strength of the relationship.
- A correlation coefficient of r = .50 indicates a stronger degree of linear relationship than one of r = .40. Likewise a correlation coefficient of r = -.50 shows a greater degree of relationship than one of r = -.40. Thus a correlation coefficient of zero (r = 0.0) indicates the absence of a linear relationship and correlation coefficients of r= + 1.0 and r = -1.0 indicate a perfect linear relationship.

Covariance

$$COV = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})$$

Example: Calculate the covariance of the heights and weights of 6 students

	X(Height)	Y(Weight)	xi-Xmean	yi-Ymean	(xi-Xmean)(yi-Ymean)
	73.00	223.00	11.67	35.17	410.28
	57.00	166.00	-4.33	-21.83	94.61
	84.00	305.00	22.67	117.17	2655.78
	56.00	157.00	-5.33	-30.83	164.44
	65.00	188.00	3.67	0.17	0.61
	33.00	88.00	-28.33	-99.83	2828.61
mean	61.33	187.83		**COV**	1025.72
variance	302.67	5271.77		r	0.81
stdev	17.40	72.61			

Understanding and interpreting the correlation coefficient

- **Scatterplots**

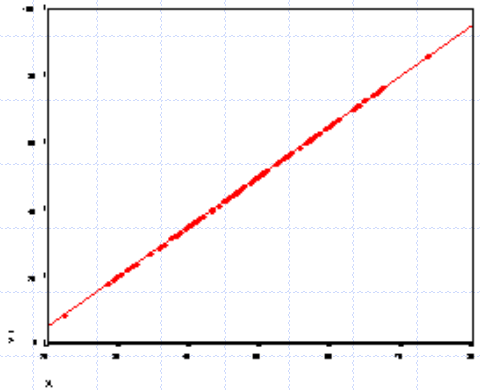
The scatterplots presented before perhaps best illustrate how the correlation coefficient changes as the linear relationship between the two variables is altered. When r = 0.0 the points scatter widely about the plot, the majority falling roughly in the shape of a circle. As the linear relationship increases, the circle becomes more and more elliptical in shape until the limiting case is reached (r =1.0 or r = -1.0) and all the points fall on a straight line.

A number of scatterplots and their associated correlation coefficients are presented below in order that the student may better estimate the value of the correlation coefficient based on a scatterplot in the associated computer exercise.

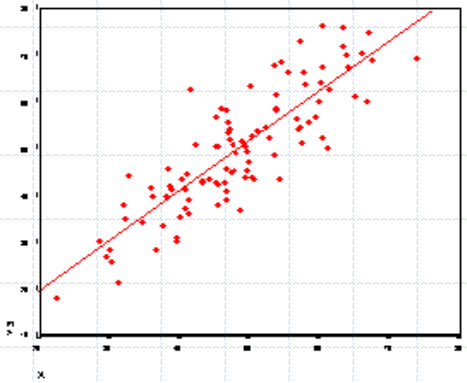
In the following plots:

- x-axis is the variable X, e.g., height or concentration of nitrate;
- y-axis is the variable Y, e.g., weight or concentration of chloride;

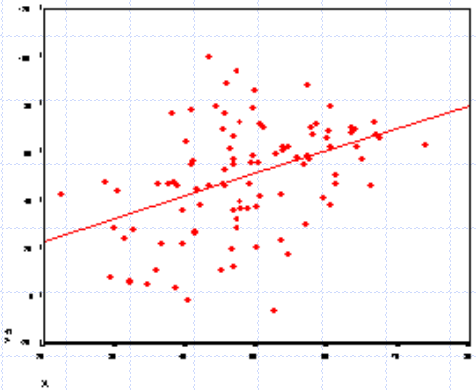
r = 1.00



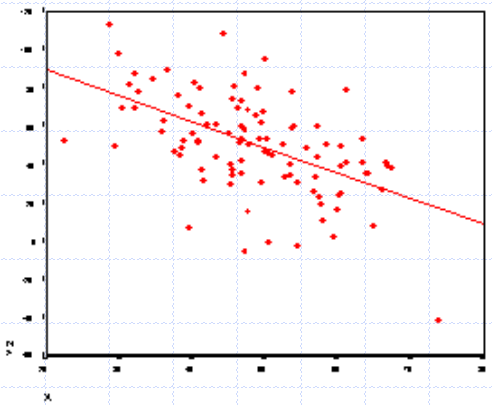
r = 0.85



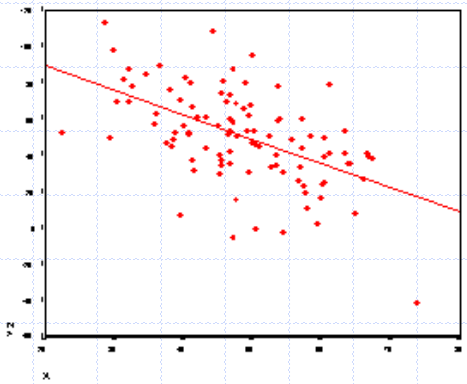
r = 0.42



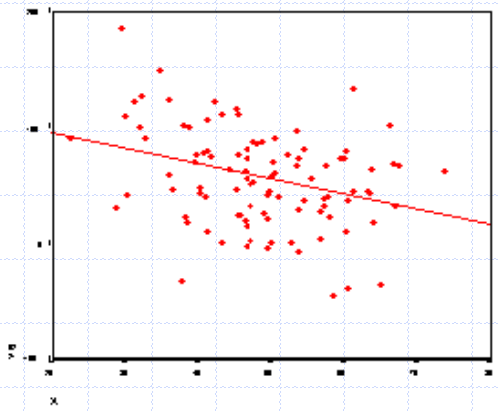
r = -0.94



r = -0.54



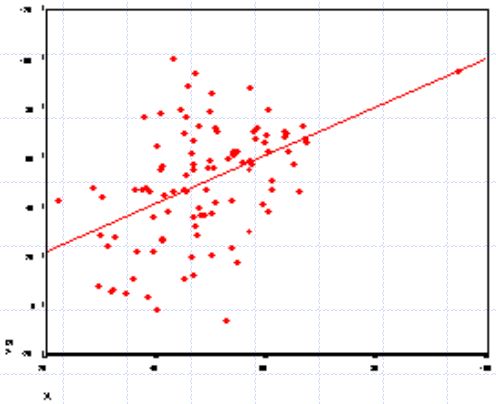
r = -0.33



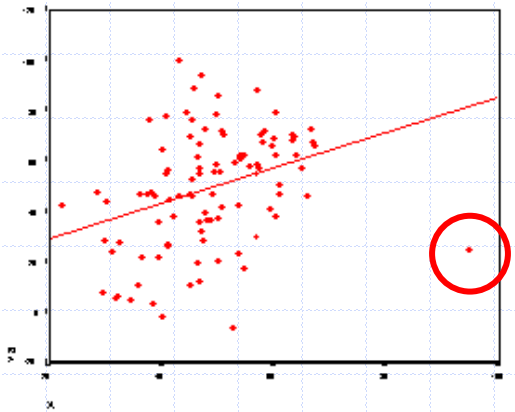
Effect of Outliers

An outlier is a score that falls outside the range of the rest of the scores on the scatterplot. For example, if age is a variable and the sample is a statistics class, an outlier would be a retired individual. Depending upon where the outlier falls, the correlation coefficient may be increased or decreased.

r = .457



r = .336



An outlier that falls some distance away from the original regression line would decrease the size of the correlation coefficient, as seen above

About outliers

The effect of the outliers on the above examples is somewhat muted because the sample size is fairly large (N=100). The smaller the sample size, **the greater the effect of the outlier**. At some point the outlier will have little or no effect on the size of the correlation coefficient.

When a researcher encounters an outlier, a decision must be made whether to include it in the data set. It may be that the respondent was deliberately malingering, giving wrong answers, or simply did not understand the question on the questionnaire. On the other hand, it may be that the outlier is real and simply different. The decision whether to include or not include an outlier remains with the researcher; **he or she must justify deleting any data** to the reader of a technical report, however.

It is suggested that the correlation coefficient be computed and reported both with and without the outlier if there is any doubt about whether or not it is real data. In any case, the best way of spotting an outlier is by drawing the scatterplot.

2.4 Linear Regression

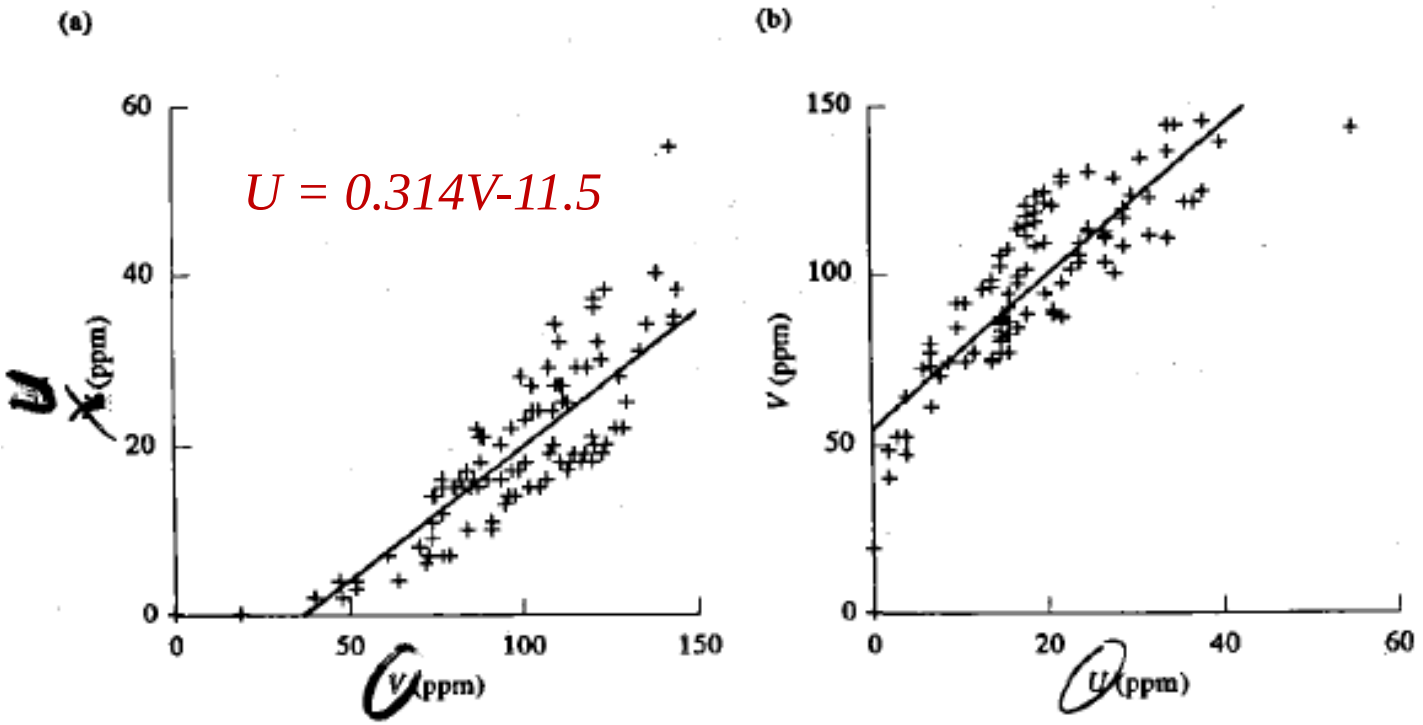
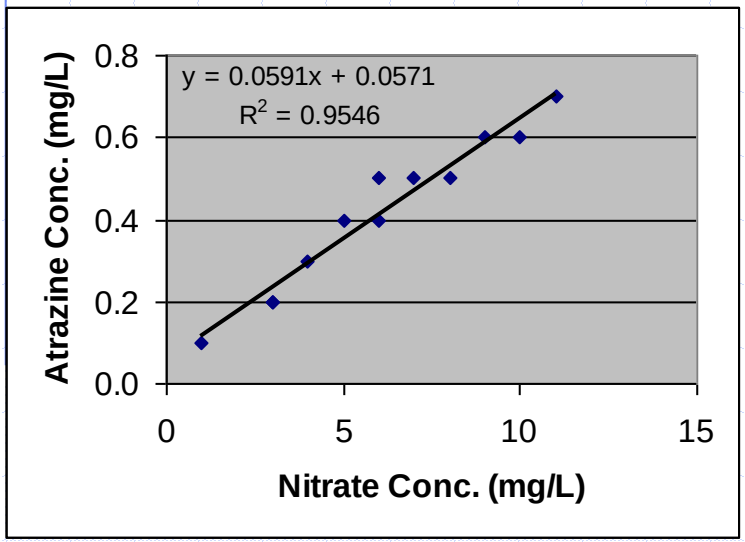


Figure 3.5 Linear regression lines superimposed on the scatterplot. The regression line of U given V is shown in (a), and of V given U in (b).

R^2 is Coefficient of Determination



You may be wondering about the meaning of R^2. The simple explanation is that this parameter, **coefficient of determination,** is a measure of how well your data fits a linear equation.

The closer R^2 is to unity, the better the fit.

In our example,

$R^2 = 0.9546$

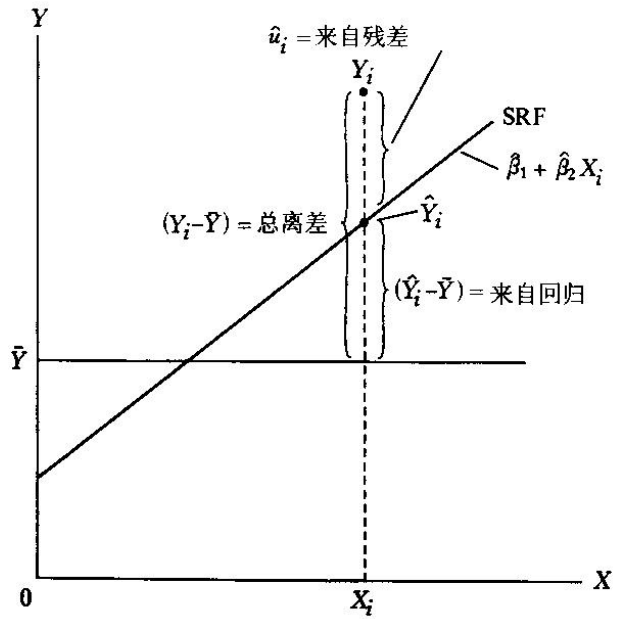
which is very good.

Coefficient of Determination (R^2)
决定系数

It is named R^2 which is the proportion of ESS (explained sum of squares) in TSS (total sum of squares). 决定系数(又称R方)。该系数为回归平方和（ESS-explained sum of squares）在总变差（TSS-total sum of squares）中所占的比重。

R2 can be used to measure how well your data fits a linear equation. 决定系数可作为综合度量回归模型对样本观测值拟合优度的度量指标。

The larger the value of R2, the larger ESS in TSS and the better your data fits a linear model and vice versa. 决定系数越大，说明在总变差中由模型作出了解释的部分占的比重越大，模型拟合优度越好。反之，决定系数小，说明模型对样本观测值的拟合程度越差。



Y_i 的变异分解为两个成分

Coefficient of determination (R^2) vs. Correlation Coefficient

Coefficient of Determination	Correlation Coefficient of
About the model 就模型而言	About two variables 就两个变量而言
Explain how well your data fits a linear equation. 说明自变量对因变量的解释程度	Quantify linear correlation between two variable 度量两个变量的线性相关程度
Range of its value： (0, 1) 取值范围	Range of its value： (-1, 1) 取值范围

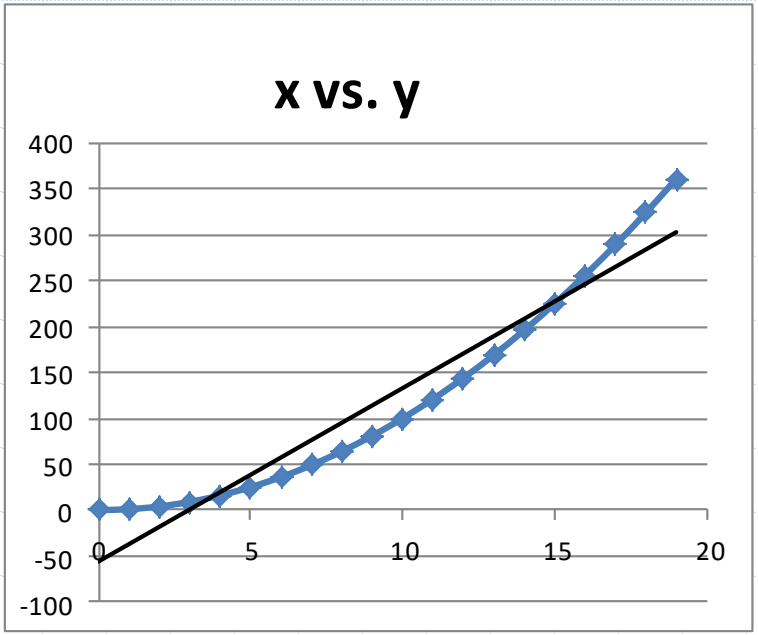
What is the value of correlation coefficient for

$$y = x^2$$

We know they are perfectly correlated but

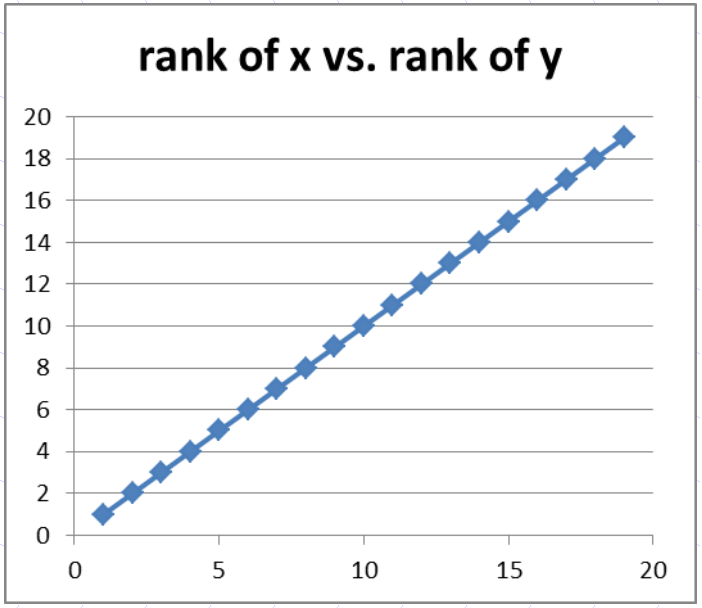
r < 1.0

Why?



2.5 Rank Correlation Coefficient (r_{rank})

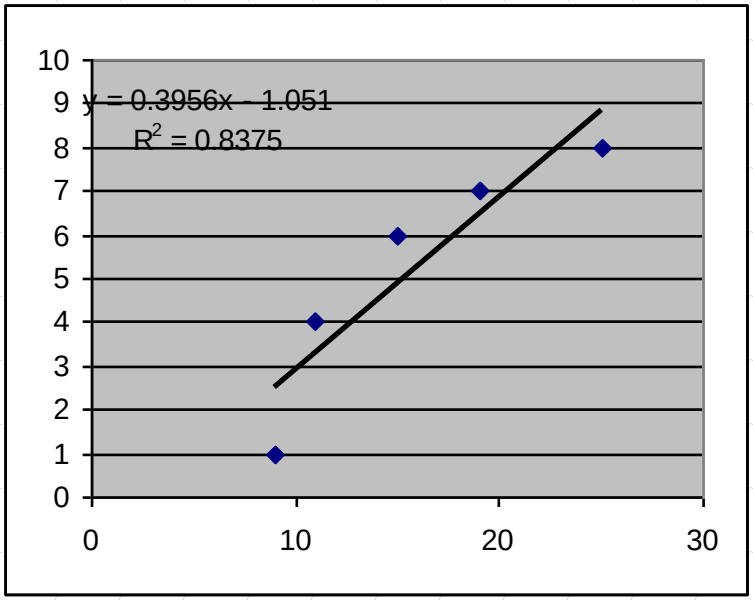
$$r_{rank} = \frac{\frac{1}{n}\sum_{i=1}^{n}(R_{x_i} - m_{R_x})(R_{y_i} - m_{R_y})}{\sigma_{R_x}\sigma_{R_Y}}$$



$r_{rank} = 1.0$

Example

i	x_i	y_i	R_{xi}	R_{yi}
1	11	4		
2	19	7		
3	25	8		
4	9	1		
5	15	6		



$$m_{R_x} = m_{R_y} = \frac{1+2+3+4+5}{5} = 3$$

$$\sigma^2_{R_x} = \sigma^2_{R_y} = \frac{1}{5}\sum_{i=1}^{5}(i-3)^2 = 2$$

$$r_{rank} = \frac{1}{2}\left\{\frac{1}{5}\left[(2-3)(2-3)+(4-3)(4-3)+\ldots+(3-3)(3-3)\right]\right\}$$

$$= \frac{1}{10}\{1+1+4+4+0\} = 1$$

Note:

- r_{rank} is not strongly influenced by extreme pairs;
- r_{rank} = 1 means (a) larger x, larger y, (b) x and y don't have to be linear

CORRELATION AND CAUSATION

No discussion of correlation would be complete without a discussion of causation. It is possible for two variables to be related (correlated), but not have one variable cause another.

For example, suppose there exists a high correlation between the number of popsicles sold and the number of drowning deaths. Does that mean that one should not eat popsicles before one swims? Not necessarily. Both of the above variables are related to a common variable, the heat of the day. The hotter the temperature, the more popsicles sold and also the more people swimming, thus the more drowning deaths. This is an example of correlation without causation.

Much of the early evidence that cigarette smoking causes cancer was correlational. It may be that people who smoke are more nervous and nervous people are more susceptible to cancer. It may also be that smoking does indeed cause cancer. The cigarette companies made the former argument, while some doctors made the latter. In this case I believe the relationship is causal and therefore do not smoke.

CORRELATION AND CAUSATION (cont.)

Sociologists are very much concerned with the question of correlation and causation because much of their data is correlational. Sociologists have developed a branch of correlational analysis, called path analysis, precisely to determine causation from correlations (Blalock, 1971). Before a correlation may imply causation, certain requirements must be met. These requirements include:

(1) the causal variable must temporally precede the variable it causes, and

(2) certain relationships between the causal variable and other variables must be met.

If a high positive correlation was found between the age of the teacher and the students' grades, it does not necessarily mean that older teachers are more experienced, teach better, and give higher grades. Neither does it necessarily imply that older teachers are soft touches, don't care, and give higher grades. Some other explanation might also explain the results. The correlation means that older teachers give higher grades; younger teachers give lower grades. It does not explain why it is the case.

CAUTIONS ABOUT INTERPRETING CORRELATION COEFFICIENTS

- **Appropriate Data Type**

Correct interpretation of a correlation coefficient requires the assumption that both variables, X and Y, meet the interval property requirements of their respective measurement systems. Calculators and computers will produce a correlation coefficient regardless of whether or not the numbers are "meaningful" in a measurement sense.

The interval property is rarely, if ever, fully satisfied in real applications. There is some difference of opinion among statisticians about when it is appropriate to assume the interval property is met. My personal opinion is that as long as a larger number means that the object has more of something or another, then application of the correlation coefficient is useful, although the potentially greater deviations from the interval property must be interpreted with greater caution. When the data is clearly nominal categorical with more than two levels (1=Protestant, 2=Catholic, 3=Jewish, 4=Other), application of the correlation coefficient is clearly inappropriate.

Summary

- The data set is derived from a DEM. There are three variables, V, U, and T. The exhaustive data set consists of 78,000 points on a 260m x 300m rectangular grid. The sample data set is a subset of the exhaustive data set which was collected in three campaigns:

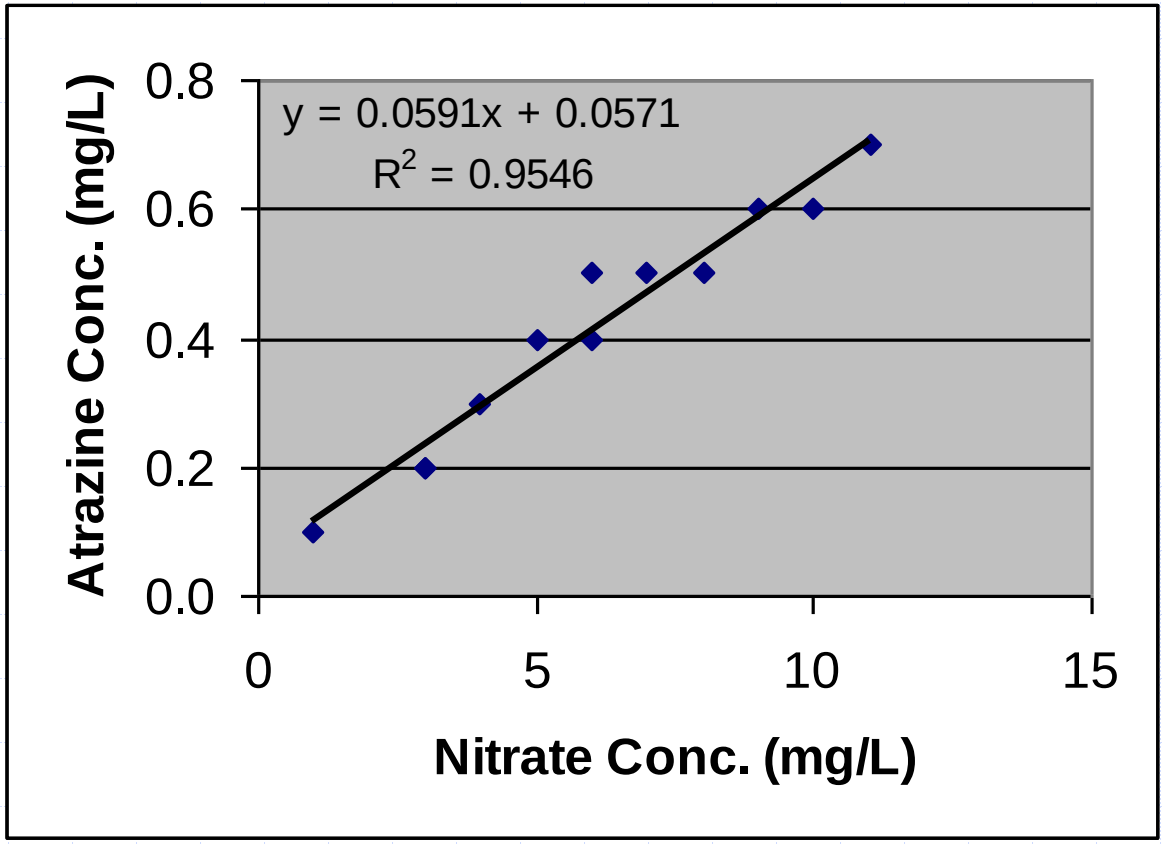
  # of data points for V is: 195+150+125 = 470
  # of data points for U is: 150+125 = 275

- One variable can be presented using *Frequency Table and Histograms, Cumulative Frequency Table and Histograms, Normal and Lognormal Probability Plots, Box-and-whisker plots*;
- Measures of central tendency are *mean, median, mode, and skewness coefficient; measures of variability are range, variance, standard deviation, coefficient of variation,*
- Two variables can be presented using *q-q plot and scatterplot* and described with *Correlation Coefficient (r), Coefficient of Determination (R^2), Rank Correlation Coefficient (r_{rank}).*

Thanks!

Regression Equation

$$Y = a + bx = 0.0571 + 0.0591x$$



Regression with MATLAB

regression calculates the linear regression between each element of the network response and the corresponding target.

[R,M,B] = regression(T,Y) takes cell array or matrix targets T and output Y, each with total matrix rows of N, and returns the linear regression for each of the N rows: the regression values R, slopes M, and y-intercepts B.

Here a feed forward network is trained and regression performed on its targets and outputs.

```
[x, t] = simplefit_dataset;
net = feedforwardnet(10);
net = train(net,x,t);
y = net(x);
[r,m,b] = regression(t,y)
```

Regression and Curve Fitting with EXCEL

Step 1 Open an Excel file

Step 2 Type in your data

Step 3 Plot the data

Step 4 Right click on any data point and select *Insert Trendline* from the resulting menu. Select the *Linear type*

Step 5 Select Options and put √ in the boxes labeled Display Equation on Chart and Display R-squared Value on Chart.

Regression and Curve Fitting with EXCEL