



# ***Module 1: Review of Basic Statistical Concepts***

## **1.2 Plotting Data, Measures of Central Tendency and Dispersion, and Correlation**

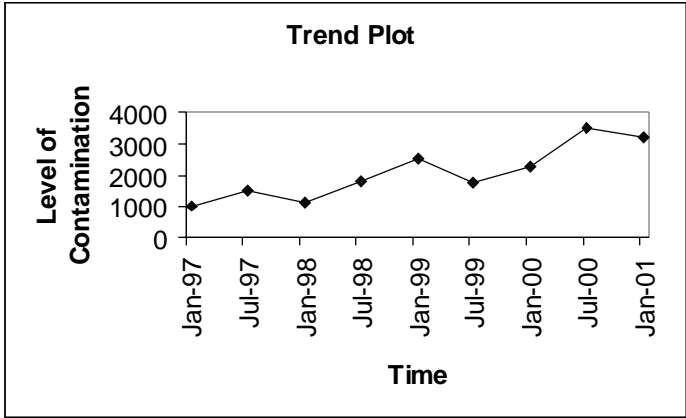


### ***Constructing a Trend Plot***

- A trend plot graphs the data against a variable of interest, often time or space. It is used to examine whether or not there is a relationship between the variable being examined and time/space.
  - Examples: Is the level of contamination in a well increasing over time? Is there a relationship between the measured concentration of a contaminant in a series of wells and their distance from a suspected source?



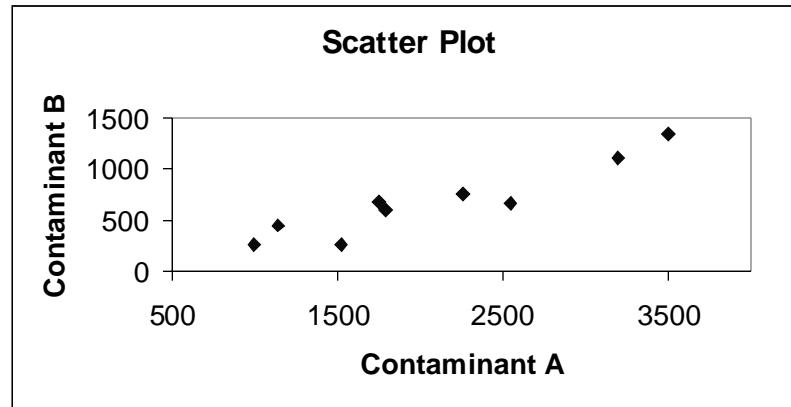
# Constructing a Trend Plot



# Constructing a Scatter Plot

- ◆ A scatter plot graphs values of one variable against the values of another variable. It is used to see if there is a relationship between the two variables.

## Constructing a Scatter Plot



4/12/2002

Module 1.2

5

## Constructing a Histogram

- A histogram is a graph of a sample pdf.
- It is constructed by:
  - Choose 5-10 non-overlapping intervals that cover the range of the data
  - Figure out how many data points fall into each interval
  - Divide the number of points in each interval by the total number of data points to get relative frequencies
  - Plot the data range on the X axis and the relative frequency on the Y
  - Draw a bar the width of each interval and the height of the relative frequency

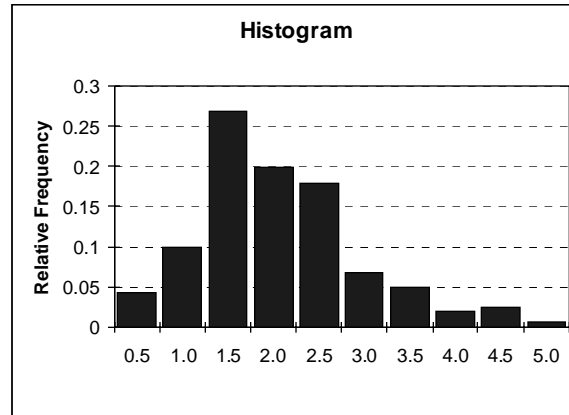
4/12/2002

Module 1.2

6



## Constructing a Histogram



4/12/2002

Module 1.2

7



## Measures of Central Tendency

- ♦ Mean – Same as Average

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

- ♦ Median – The middle value of a data set sorted from largest to smallest. If there are an even number of data points, average the two middle values.
- ♦ Mode – The most commonly occurring value

4/12/2002

Module 1.2

8



## *Measures of Central Tendency*

- ◆ Example: Heights to the nearest inch

60 64 65 67 67 67 69 70 72 72

- Mean =  
 $(60+64+65+67+67+67+69+70+72+72)/10$   
 $= 67.3$
- Median =  $(67 + 67)/2 = 67$
- Mode = 67

4/12/2002

Module 1.2

9



## *Measures of Central Tendency*

- ◆ Example: Salaries in a Start-up Dot Com company (in thousands)

27 27 33 35 85 150

- Mean = 59.5K
- Median = 34K
- Mode = 27K
- ◆ So, for symmetric distributions (like the normal) the mean is a good measure of central tendency but for skewed distributions (like income or environmental contamination) it is heavily influenced by a few unusual points.

4/12/2002

Module 1.2

10



## ***Measures of Dispersion***

- ◆ Measures of dispersion measure how spread out the data are.
- ◆ Range = largest value – smallest value
- ◆ The problem with the range is that it tells you nothing about all of the rest of the data

4/12/2002

Module 1.2

11



## ***Measures of Dispersion***

- ◆ Intuitively, you can think of the Standard Deviation as the average difference between the data points and the mean. Unlike the range, it's a function of all of the data points.
- ◆ A deviation is a difference between two values. We can easily calculate the deviations of each data point from the mean. If we summed these, we would get zero. So, we must either square them or take their absolute value.
- ◆ Absolute values are difficult to work with mathematically so we'll square the deviations. Then we "average" them to get the variance. Then, since we squared the deviations, the units of the variance are the square of the data points so we take the square root to get back to original units.

4/12/2002

Module 1.2

12



## Measures of Dispersion

- Because of some mathematical properties of the statistic, we use  $n-1$  rather than  $n$  in taking the “average” of the deviations.

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

4/12/2002

Module 1.2

13



## Measures of Dispersion

Example: Heights to the nearest inch

$X_i$        $X_i - \bar{X}$        $(X_i - \bar{X})^2$

60	-7.3	53.29
64	-3.3	10.89
65	-2.3	5.29
67	-0.3	0.09
67	-0.3	0.09
67	-0.3	0.09
69	1.7	2.89
70	2.7	7.29
72	4.7	22.09
72	4.7	22.09

124.10

Variance is  $s^2 = (1/(10-1)) * 124.10 = (1/9) * 124.10 = 13.79$

Standard deviation is the square root of  $13.79 = 3.71$

4/12/2002

Module 1.2

14



## *Percentiles of a Distribution*

- ♦ The population median is the point that has 50% of the distribution above it and 50% below. The sample median has 50% of the data above and 50% below.
- ♦ The percentiles of the distribution (or sample) are similar. The Xth percentile has X percent of the distribution (or data) below it and 1-X percent above it.
- ♦ For example, a 95 percentile has 95% of the distribution below it and 5% above it.

4/12/2002

Module 1.2

15



## *Correlation*

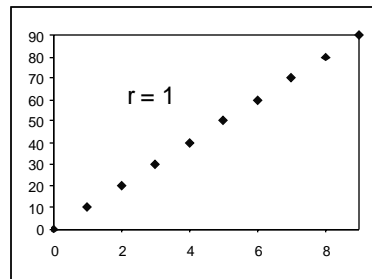
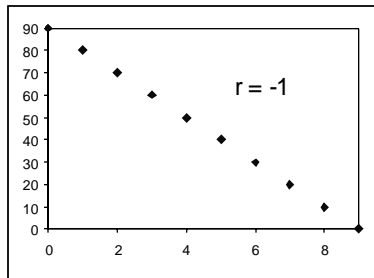
- ♦ The correlation coefficient measures the degree of linear association between two variables. It ranges between -1 and 1.
- ♦ A perfect association gives points that plot on a straight line. No association gives points that plot as a cloud.
- ♦ A positive association means that high values of one variable are associated with high values of the other. A negative association means that high values of one variable are associated with low values of the other.

4/12/2002

Module 1.2

16

# Examples of the Correlation Coefficient

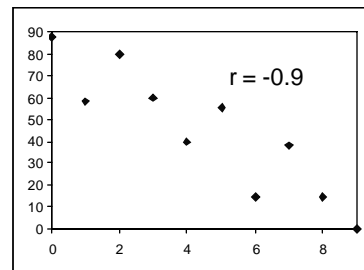
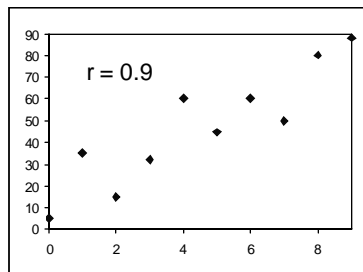


4/12/2002

Module 1.2

17

# Examples of the Correlation Coefficient

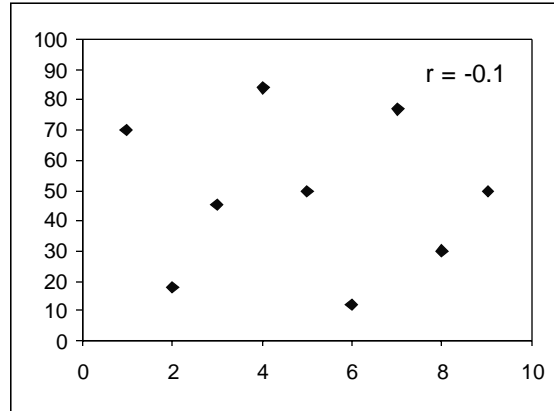


4/12/2002

Module 1.2

18

# Examples of the Correlation Coefficient



4/12/2002

Module 1.2

19