Means, Correlation, and Regression

(StatLib Programs)

Population and Sample

- **Population** = the complete collection of elements to be studied (measured response times, memory utilizations, etc.)
- **Sample** = a subset of a population
- **Parameter** = a numerical characteristic of a population
- **Statistic** = numerical characteristic of a sample
- **Measured values:** $x_1, x_2, \ldots, x_n$

Using arithmetic, geometric and harmonic means

Mean (Arithmetic Mean)

$$\Sigma x = \sum_{i=1}^{n} x_i , \quad \Sigma(fx) = \sum_{i=1}^{n} f_i x_i$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{\Sigma(fx)}{\Sigma f} = \Sigma(px) , \quad p_i = \frac{f_i}{\Sigma f}$$

$$\overline{xy} = \frac{\Sigma(xy)}{n} , \ \overline{x^2} = \frac{\Sigma x^2}{n} , \ \bar{x}^2 = \left(\frac{\Sigma x}{n} \right)^2$$

Use and Abuse of Arithmetic Mean

Arithmetic mean can be use for averaging

- Memory use
- Response time

Arithmetic mean cannot be use for averaging

- Performance ratios
- Processing speeds

Performance ratios

$A, B = \text{benchmark programs}$

$S_1, S_2 = \text{competitive systems}$

$t_{A1}, t_{A2} = \text{run times for benchmark A}$

$t_{B1}, t_{B2} = \text{run times for benchmark B}$

$\text{Performance ratios of systems 1 and 2:}$

$R_A = t_{A1} / t_{A2}\ , \quad R_B = t_{B1} / t_{B2}$

$\text{What is the average performance ratio?}$

Average Performance Ratio

$$R_A = t_{A1} / t_{A2}\ , \quad R_B = t_{B1} / t_{B2}$$

Average performance ratio $R = f(R_A, R_B)$

Averaging function $\quad f = ?$

$$R = \sqrt{\frac{t_{A1}}{t_{A2}} \cdot \frac{t_{B1}}{t_{B2}}}\ , \quad \sqrt{\frac{1}{2} \cdot \frac{2}{1}} = 1$$

Important: $\quad R \neq \dfrac{R_A + R_B}{2} = \dfrac{2.5}{2} = 1.25$

Average Performance Ratio: Geometric Mean

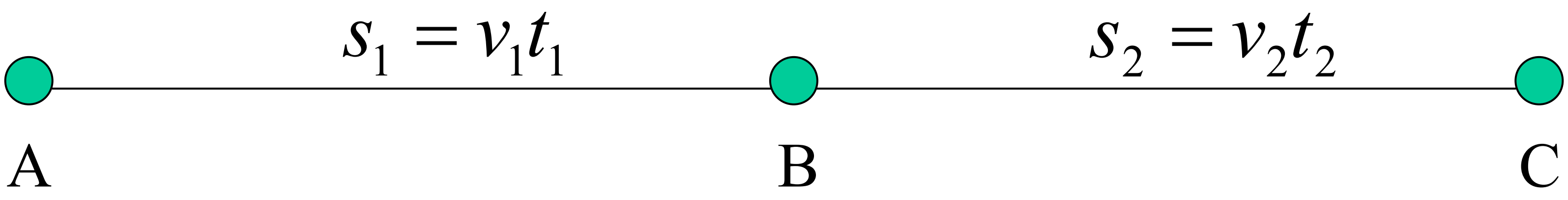
$$\text{The case of 3 programs: A, B, C}$$

$$R = \sqrt[3]{\frac{t_{A1}}{t_{A2}} \cdot \frac{t_{B1}}{t_{B2}} \cdot \frac{t_{C1}}{t_{C2}}}$$

$$\text{General case:}$$

$$R = \left(\frac{t_{11}}{t_{12}} \cdot \frac{t_{21}}{t_{22}} \cdots \frac{t_{n1}}{t_{n2}} \right)^{1/n}$$

Averaging Rates



$s_1 = v_1 t_1$ $s_2 = v_2 t_2$

A B C

We travel from point A to point C. The velocity between points A and B is v_1 and the traveling time is t_1 . The velocity between points B and C is v_2 and the traveling time is t_2 .

What is the average velocity between points A and C?

Could it be $v = (v_1 + v_2)/2$???

Averaging Rates: Harmonic Mean

$$s_1 = v_1 t_1 \ , \quad s_2 = v_2 t_2 \ , \quad s_1 + s_2 = s = vt$$

$$t = t_1 + t_2 \ , \quad \frac{s}{v} = \frac{s_1}{v_1} + \frac{s_2}{v_2}$$

$$v = \frac{1}{\dfrac{s_1 / s}{v_1} + \dfrac{s_2 / s}{v_2}} \qquad \frac{s_1}{s} + \frac{s_2}{s} = 1$$

Server Processing Rate Example

A web server processes two classes of transactions. The rate of processing short transactions is 80 per second. The rate of processing long transactions is 20 per second.

(a) What is the average processing rate if 50% of transactions are long and 50% of transactions are short?

(b) What is the average processing rate if 20% of transactions are long and 80% of transactions are short?

Average Processing Rate - Case (a)

$$N = \text{number of precessed transactions}$$

$$N_1 = v_1 t_1 \;, \quad N_2 = v_2 t_2 \;, \quad N_1 + N_2 = N = vt$$

$$t = t_1 + t_2 \;, \quad \frac{N}{v} = \frac{N_1}{v_1} + \frac{N_2}{v_2}$$

$$v = \frac{1}{\dfrac{N_1 / N}{v_1} + \dfrac{N_2 / N}{v_2}} = \frac{1}{\dfrac{0.5}{80} + \dfrac{0.5}{20}} = 32 / \sec$$

Average Processing Rate - Case (b)

$$v = \frac{1}{\dfrac{N_1 / N}{v_1} + \dfrac{N_2 / N}{v_2}} = \frac{1}{\dfrac{0.8}{80} + \dfrac{0.2}{20}} = 50 / \sec$$

This model is called the weighted harmonic mean. Weights are N_1 / N and N_2 / N.

The sum of weights: $N_1 / N + N_2 / N = 1$

Weighted power means

Weighted Arithmetic Mean

$$\bar{x} = \sum_{i=1}^{n} W_i x_i$$

$$x_i \geq 0\,, \quad i = 1,\ldots,n$$

$$0 < W_i < 1\,, \quad i = 1,\ldots,n$$

$$\sum_{i=1}^{n} W_i = 1$$

Weighted Quadratic Mean

$$\bar{x} = \left(\sum_{i=1}^{n} W_i x_i^2 \right)^{1/2}$$

$$x_i \geq 0 \,, \quad i = 1, \ldots, n$$

$$0 < W_i < 1 \,, \quad i = 1, \ldots, n$$

$$\sum_{i=1}^{n} W_i = 1$$

Weighted Geometric Mean

$$\bar{x} = \prod_{i=1}^{n} x_i^{W_i} = \exp\left(\sum_{i=1}^{n} W_i \log x_i \right)$$

$$x_i \geq 0 , \quad i = 1,\ldots,n$$

$$0 < W_i < 1 , \quad i = 1,\ldots,n$$

$$\sum_{i=1}^{n} W_i = 1$$

Weighted Harmonic Mean

$$\bar{x} = \frac{1}{\sum_{i=1}^{n} \frac{W_i}{x_i}}$$

$$x_i \geq 0\,, \quad i = 1,\ldots,n$$

$$0 < W_i < 1\,, \quad i = 1,\ldots,n$$

$$\sum_{i=1}^{n} W_i = 1$$

Power Mean

$$\bar{x} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^r \right)^{1/r} \quad ,$$

$$-\infty \le r \le +\infty$$

$$x_i \ge 0 \,, \quad i = 1,\ldots,n$$

Weighted Power Mean

$$\bar{x} = \left(\sum_{i=1}^{n} W_i x_i^r \right)^{1/r} , \quad -\infty \le r \le +\infty$$

$$x_i \ge 0 , \quad i = 1,\ldots,n$$

$$0 < W_i < 1 , \quad i = 1,\ldots,n$$

$$\sum_{i=1}^{n} W_i = 1$$

Extreme Cases

If $r = -\infty$ then

$$\bar{x} = \left(\sum_{i=1}^{n} W_i x_i^r \right)^{1/r} = \min(x_1, \ldots, x_n)$$

If $r = +\infty$ then

$$\bar{x} = \left(\sum_{i=1}^{n} W_i x_i^r \right)^{1/r} = \max(x_1, \ldots, x_n)$$

Basic Properties of WPM (1)

If $x_1 = x_2 = \ldots = x_n = x$ then

$$\bar{x} = \left(\sum_{i=1}^{n} W_i x_i^r \right)^{1/r} = x \,, \quad \frac{\partial \bar{x}}{\partial r} = 0$$

In all other cases

$$\frac{\partial \bar{x}}{\partial r} = \frac{\partial}{\partial r} \left(\sum_{i=1}^{n} W_i x_i^r \right)^{1/r} > 0$$

Basic Properties of WPM (2)

$$\min(x_1, x_2, \ldots, x_n) \le \frac{1}{\sum_{i=1}^{n} W_i / x_i} \le$$

$$\le \prod_{i=1}^{n} x_i^{W_i} \le \sum_{i=1}^{n} W_i x_i \le$$

$$\le \left(\sum_{i=1}^{n} W_i x_i^2 \right)^{1/2} \le \max(x_1, x_2, \ldots, x_n)$$

$$\text{Min} \le \text{Har} \le \text{Geo} \le \text{Ari} \le \text{Quad} \le \text{Max}$$

Quasi-arithmetic Means

$$\bar{x} = F^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} F(x_i)\right)$$

$$x_i \geq 0\,, \quad i = 1,\ldots,n$$

If $F(x) = x^r$ then we get the power mean

Weighted Quasi-arithmetic Means

$$\bar{x} = F^{-1}\left(\sum_{i=1}^{n} W_i F(x_i)\right)$$

$$x_i \ge 0\,, \quad i = 1,\ldots,n$$

If $F(x) = x^r$ then we get

the weighted power mean

Measures of central tendency

Averages

- Midrange
- Median
- Mode
- Mean (arithmetic)

These four indicators are also called “basic measures of central tendency”.

Midrange

- Defined as the mean value of the minimum and maximum element.
- If the population = 3, 2, 7, 5, 6, 5, 5 then min=2, max=7, and midrange = (7+2)/2=4.5
- Midrange is only affected by extreme values. It does not take into account every value. It is rarely used.

Median

- Defined as the value in the middle of the sorted list of values (for odd number of values) or as the mean of the two middle values (for populations with even number of values).
- If the population = 3, 2, 7, 5, 6, 5, 5 then the sorted list is 2, 3, 5, 5, 5, 6, 7 and the median is 5.
- Median does not take into account every value. It is not affected by extreme values. It is useful to avoid effects of questionable extreme values.

Mode

- Defined as the most frequent value in a population.
- It does not exist if the values are not repeated.
- If the population = 3, 2, 7, 5, 6, 5, 5 then the most frequent value (mode) is 5.
- It can be bimodal (e.g. 1,1,2,7,3,4,4), in the case of two modes, or multimodal for more than two modes.
- It is not affected by extreme values.
- Mode is the only average that can be used for populations with nominal (nonnumeric) data.

Mean

- Assumed to be the arithmetic mean, Σx/n .
- If the population = 3, 2, 7, 5, 6, 5, 5 then mean = 33/7=4.714
- Mean is the only average affected by all values. It is most frequently used.
- Sometimes is negatively affected by questionable extreme (outlier) values.

Measures of dispersion

Measures of Dispersion

- Range
- Variance
- Standard deviation
- Mean deviation
- Coefficient of variation

Range

$$X = (x_1, x_2, \ldots, x_n) = \text{population or sample}$$

$$x_{\min} = \min_{1 \le i \le n} x_i \ , \quad x_{\max} = \max_{1 \le i \le n} x_i$$

$$Range = [x_{\min}, \quad x_{\max}]$$

Variance (σ² or s²)

Whole population containing n elements $(\Sigma x = n\bar{x})$:

$$\sigma^2 = \frac{\Sigma(x-\bar{x})^2}{n} = \frac{\Sigma(x^2 - 2x\bar{x} + \bar{x}^2)}{n} =$$

$$= \frac{\Sigma(x^2) - 2\bar{x}\Sigma x + n\bar{x}^2}{n} = \frac{\Sigma(x^2) - n\bar{x}^2}{n} = \overline{x^2} - \bar{x}^2$$

Sample of n values (s is the best estimate of σ):

$$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{\Sigma(x^2) - n\bar{x}^2}{n-1} = \frac{n\Sigma(x^2) - (\Sigma x)^2}{n(n-1)}$$

Standard Deviation (σ or s)

Whole population containing n elements $(\Sigma x = n\bar{x})$:

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} = \sqrt{\frac{\Sigma(x^2) - n\bar{x}^2}{n}} = \sqrt{\overline{x^2} - \bar{x}^2}$$

Sample of n values (s is the best estimate of σ) :

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\frac{\Sigma(x^2) - n\bar{x}^2}{n-1}} = \sqrt{\frac{n\Sigma(x^2) - (\Sigma x)^2}{n(n-1)}}$$

Mean Deviation

$$\text{Mean deviation} = \frac{\sum \left| x - \bar{x} \right|}{n}$$

- A very reasonable indicator, equally affected by all components (the variance and standard deviation are predominantly affected by large values of the difference |x-xmean|).
- Problems with the derivative in the origin.
- Unsuitable for statistical inference methods.

Coefficient of Variation

$$\text{Whole population:}$$

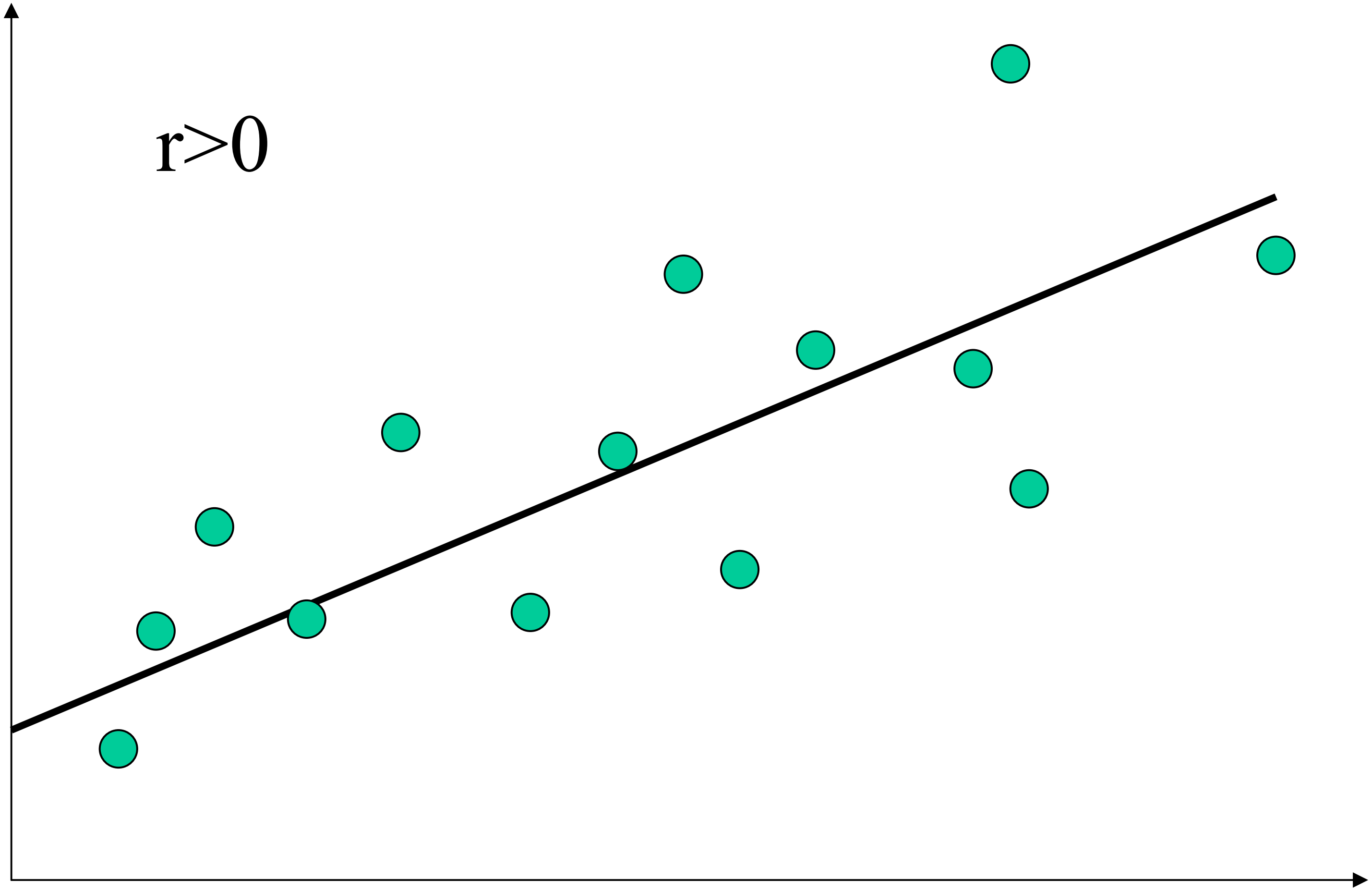
$$v = 100\sigma / \bar{x}\ [\%], \quad \bar{x} > 0$$

$$\text{Sample:}$$

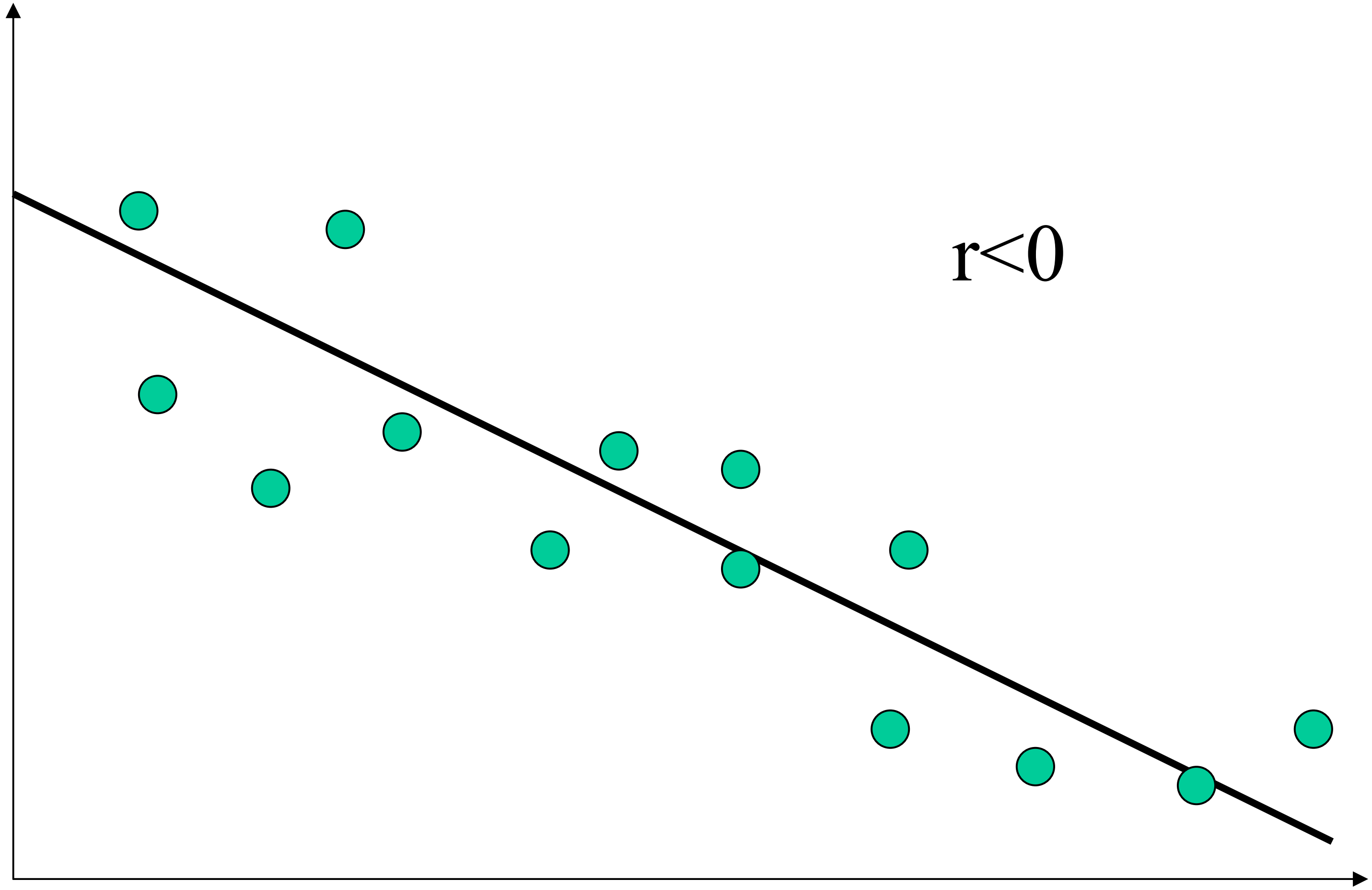
$$v = 100s / \bar{x}\ [\%], \quad \bar{x} > 0$$

Linear correlation

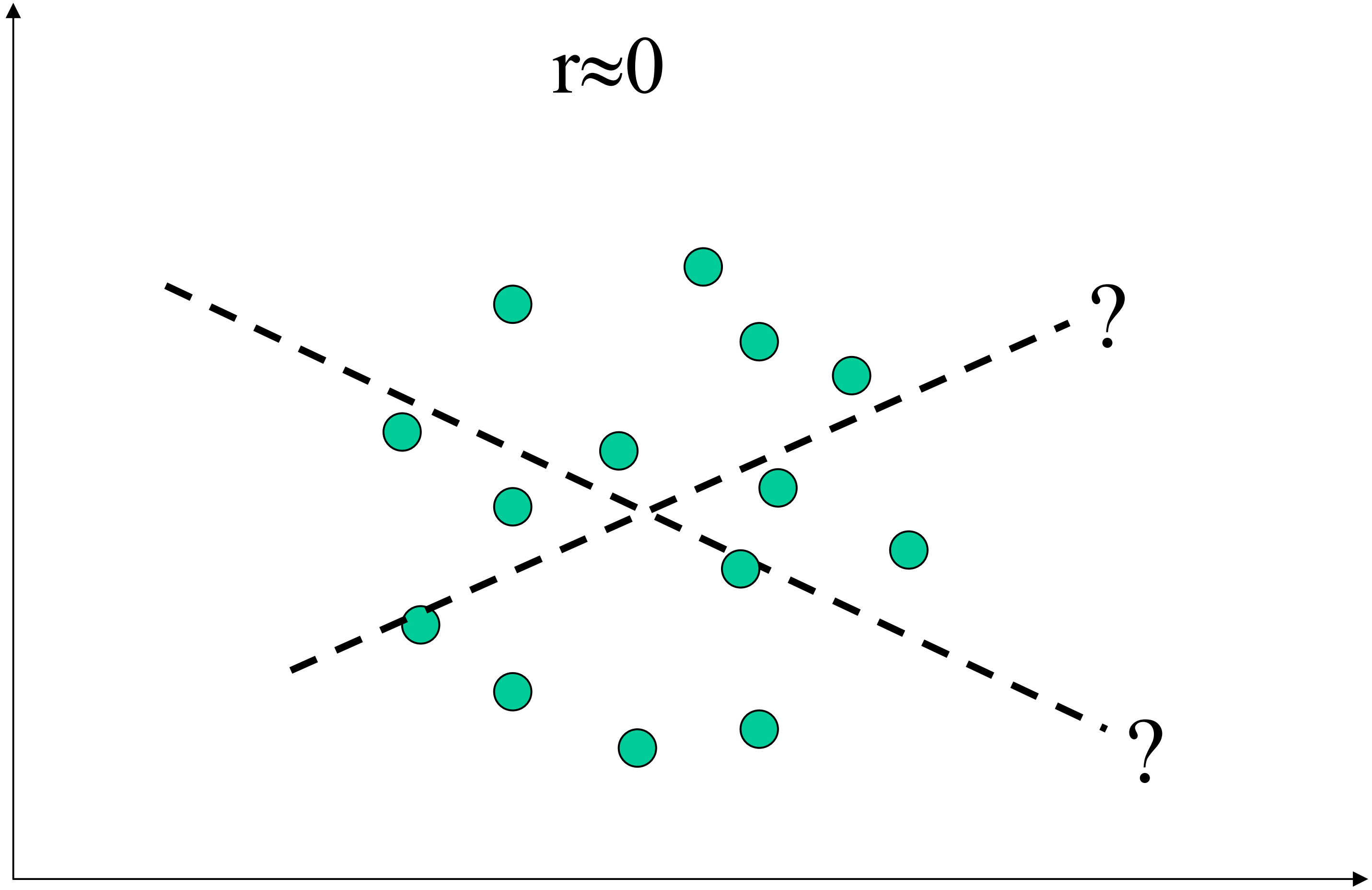
Positive Correlation



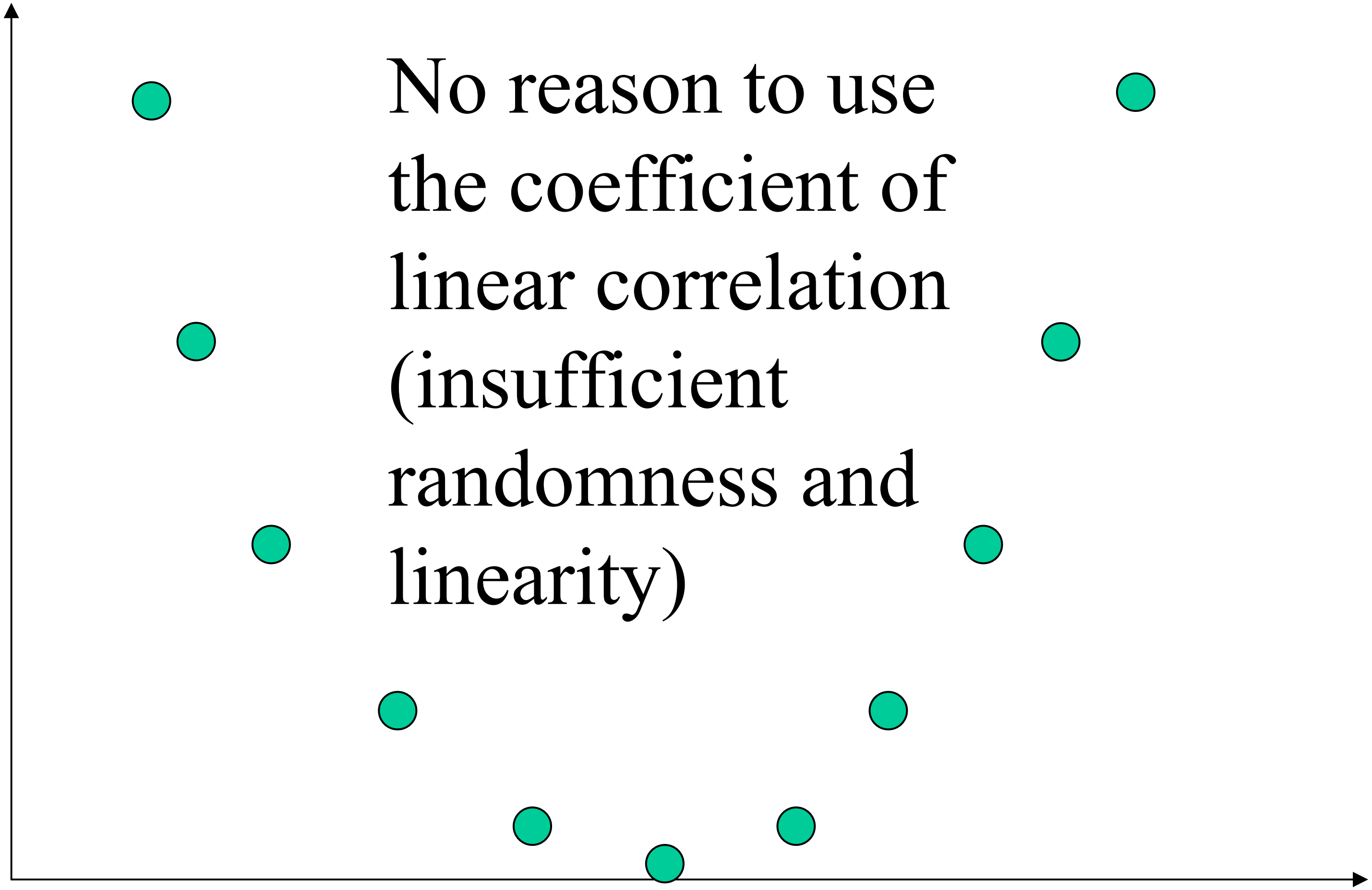
Negative Correlation



No Correlation



Nonlinear Functional Relationship



No reason to use the coefficient of linear correlation (insufficient randomness and linearity)

Linear Correlation

- Measure of the strength of linear relationship between the paired random values x and y.
- Example: x = size of program, y= program development time (or program compile time).
- Quantitative measure for a given sample (or a population) is the coefficient of correlation:

$$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{(n-1)s_x s_y}, \quad \rho = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y}, \quad \begin{cases} -1 \le r \le +1 \\ -1 \le \rho \le +1 \end{cases}$$

Other Forms of the Correlation Formula for Samples

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{(n-1)s_x s_y}$$

$$= \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

Correlation Formula for Population

$$\frac{[n\Sigma xy-(\Sigma x)(\Sigma y)]/n^2}{\sqrt{n(\Sigma x^2)-(\Sigma x)^2}\sqrt{n(\Sigma y^2)-(\Sigma y)^2}\;/\,n^2}=$$

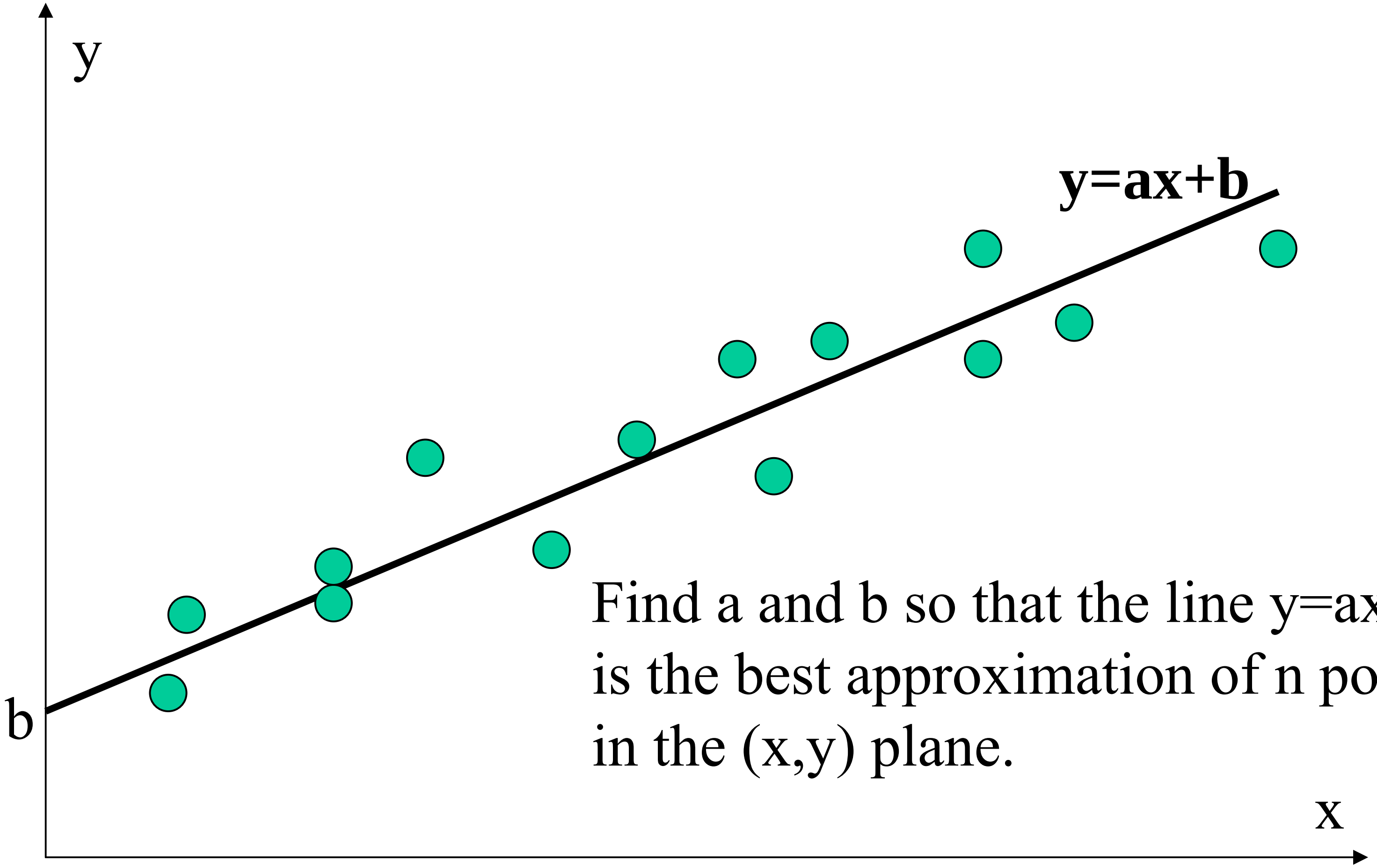
$$=\frac{\overline{xy}-\bar{x}\;\bar{y}}{\sigma_x\sigma_y}=\rho$$

Regression

Linear Regression (1)

- In the x-y plane find a linear regression line y=ax+b that is the best approximation of n points in the plane.
- Method:
  - – Create the mean square error function E(a,b) as a sum of squares of differences between the points and the line.
  - – Compute the values of *a* and *b* that minimize the error E(a,b).

Linear Regression (2)



Find a and b so that the line y=ax+b is the best approximation of n points in the (x,y) plane.

Linear Regression (3)

Criterion function (mean square error):

$$E(a,b) = \Sigma(ax+b-y)^2$$

Find *a* and *b* that minimize *E(a,b)*:

$$\left.\begin{array}{l}\dfrac{\partial E}{\partial a} = 2\Sigma(ax+b-y)x = 0 \\[2ex] \dfrac{\partial E}{\partial b} = 2\Sigma(ax+b-y) = 0\end{array}\right\}$$

$$\left.\begin{array}{l}a\Sigma x^2 + b\Sigma x = \Sigma xy \\ a\Sigma x + bn = \Sigma y\end{array}\right\}\text{Solve for } a \text{ and } b$$

Linear Regression (4)

Solution :

$$a = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2},$$

$$b = \frac{(\Sigma y - a\Sigma x)}{n} = \bar{y} - a\bar{x} = \frac{\Sigma y \Sigma x^2 - \Sigma x \Sigma xy}{n\Sigma x^2 - (\Sigma x)^2}$$

Linear Regression (5)

$$\text{Regression line}:$$

$$\frac{y-\bar{y}}{\sigma_y} = \rho \frac{x-\bar{x}}{\sigma_x}$$

$$y = ax + b = \bar{y} + \rho \frac{\sigma_y}{\sigma_x}(x-\bar{x})$$

Nonlinear Regression

$\text{Measured values}: (t_1, n_1), (t_2, n_2), \ldots, (t_m, n_m)$

$t(n) = an^b \quad | \quad \log$

$\log t = \log a + b \log n$

$T = \log t, \quad A = \log a, \quad N = \log n$

$T = A + bN$. Now, we can apply the linear regression formulas and compute A and b.

Then, $t(n) = \exp(A) n^b$

Computing parameters in 2 points

$$\text{Measured values}: (t_1, n_1), (t_2, n_2), \ldots, (t_m, n_m)$$

$$\text{Model}: \quad t(n) = an^b$$

$$\left.\begin{array}{l} t_j = an_j^b \\ t_k = an_k^b \end{array}\right\} \text{Zero error in points } j \text{ and } k$$

$$\frac{t_j}{t_k} = \left(\frac{n_j}{n_k}\right)^b \quad , \quad \log\left(\frac{t_j}{t_k}\right) = b\log\left(\frac{n_j}{n_k}\right)$$

$$b = \frac{\log t_j - \log t_k}{\log n_j - \log n_k} \quad , \quad a = t_j / n_j^b$$