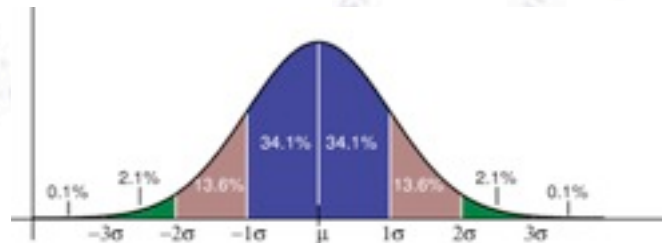# Applied Statistics

Correlations

Troels C. Petersen (NBI)

"Statistics is merely a quantization of common sense"

Correlation

Recall the definition of the Variance, V:

$$V = \sigma^2 = \frac{1}{N} \sum_i^n (x_i - \mu)^2 = E[(x - \mu)^2] = E[x^2] - \mu^2$$
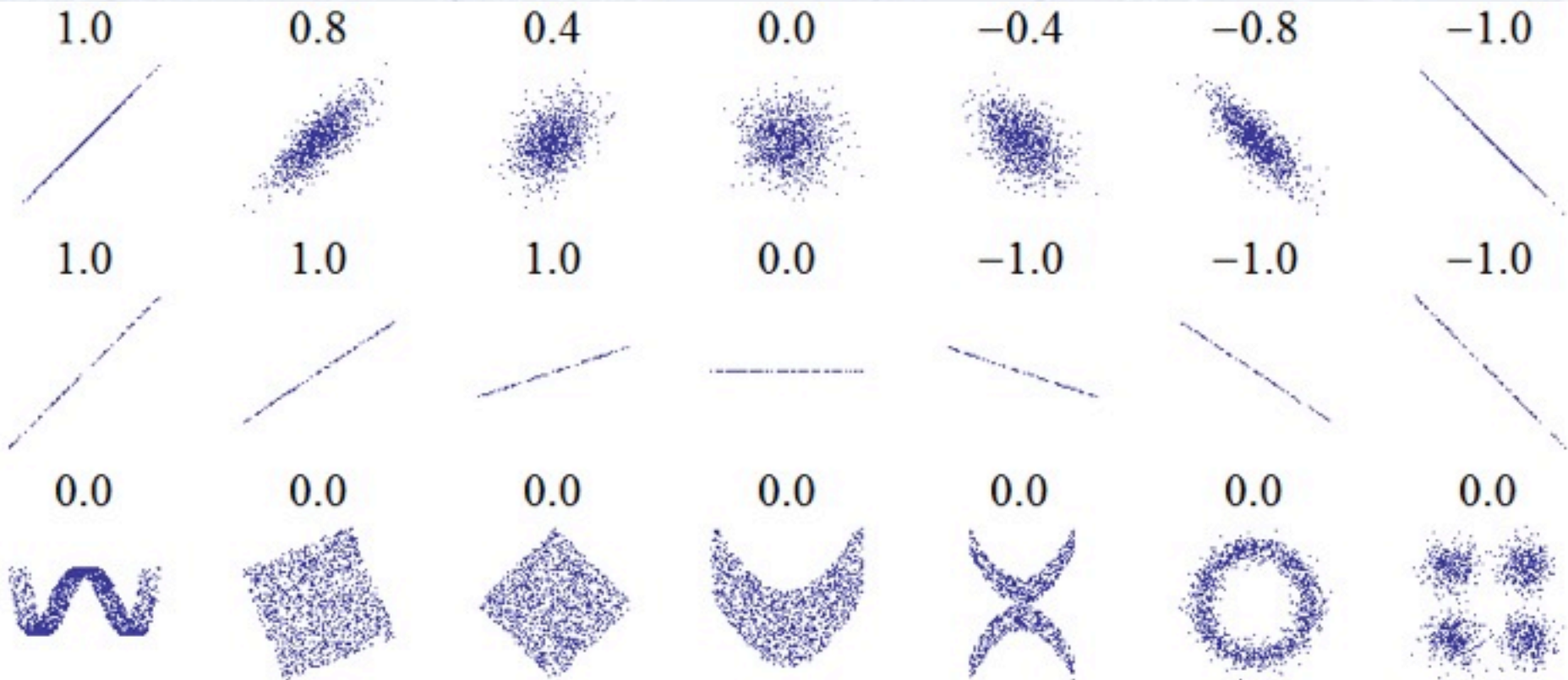
Likewise, one defines the **Covariance, V_{xy}:**

$$V_{xy} = \frac{1}{N} \sum_i^n (x_i - \mu_x)(y_i - \mu_y) = E[(x_i - \mu_x)(y_i - \mu_y)]$$

"Normalizing" by the widths, gives the (linear) correlation:

$$\rho_{xy} = \frac{V_{xy}}{\sigma_x \sigma_y}$$

$$-1 < \rho_{xy} < 1$$

$$\sigma(\rho) \simeq \sqrt{\frac{1}{n}(1 - \rho^2)^2 + O(n^{-2})}$$

Correlation

Correlations in 2D are in the gaussian case the "degree of ovalness"!

Note how ALL of the bottom distributions have $\varrho = 0$, despite obvious correlations!

Correlation

The correlation matrix Vxy explicitly looks as:

$$V_{xy} = \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 & \cdots & \sigma_{1N}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2N}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_N^2 & \sigma_{N2}^2 & \cdots & \sigma_{NN}^2 \end{bmatrix}$$

Very specifically, the calculations behind are:

$$V = \begin{bmatrix} \mathrm{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \mathrm{E}[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_1 - \mu_1)(X_n - \mu_n)] \\ \mathrm{E}[(X_2 - \mu_2)(X_1 - \mu_1)] & \mathrm{E}[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n - \mu_n)(X_1 - \mu_1)] & \mathrm{E}[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}.$$

Correlation and Information

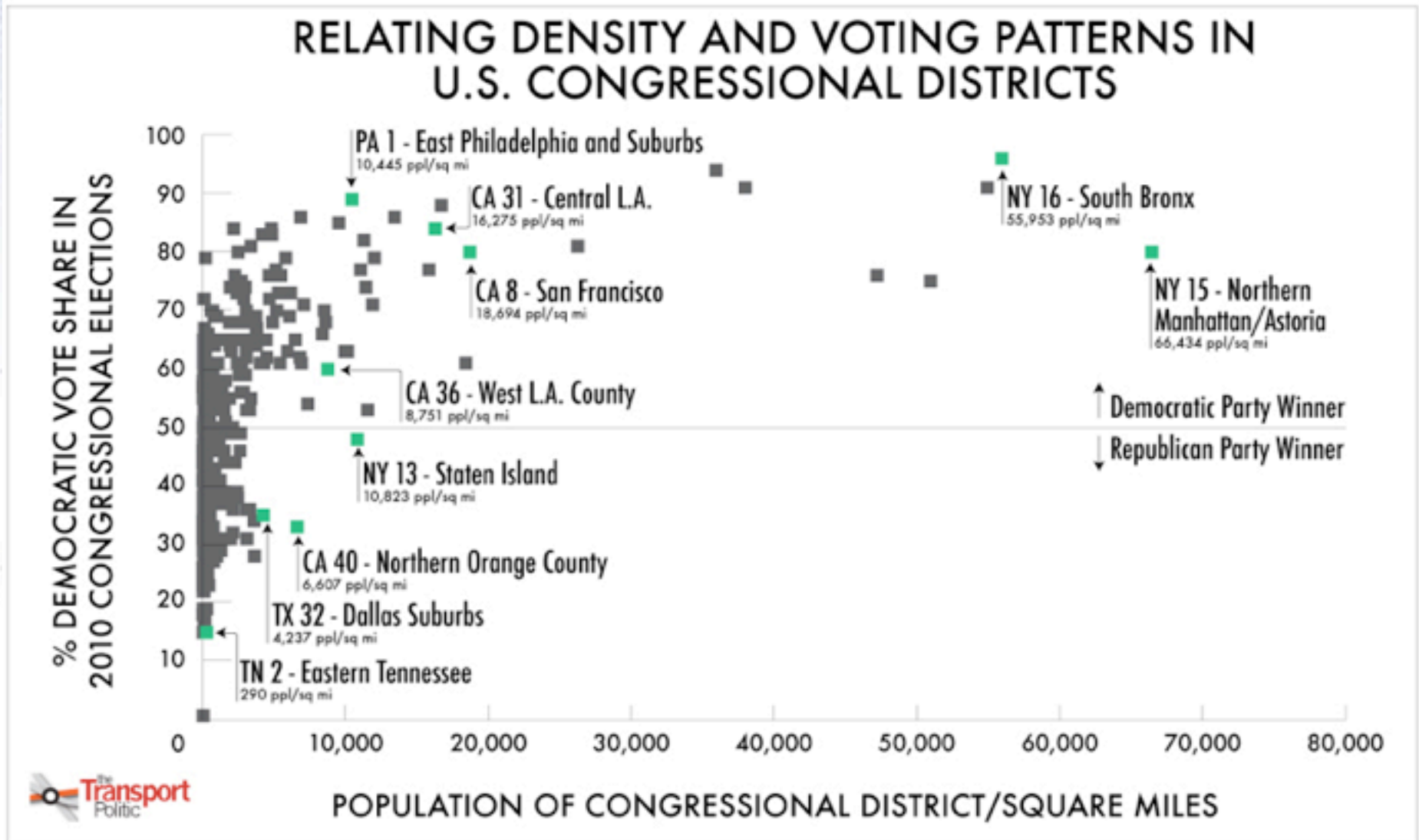Correlations influence results in complex ways!

They need to be taken into account, for example in **Error Propagation!**

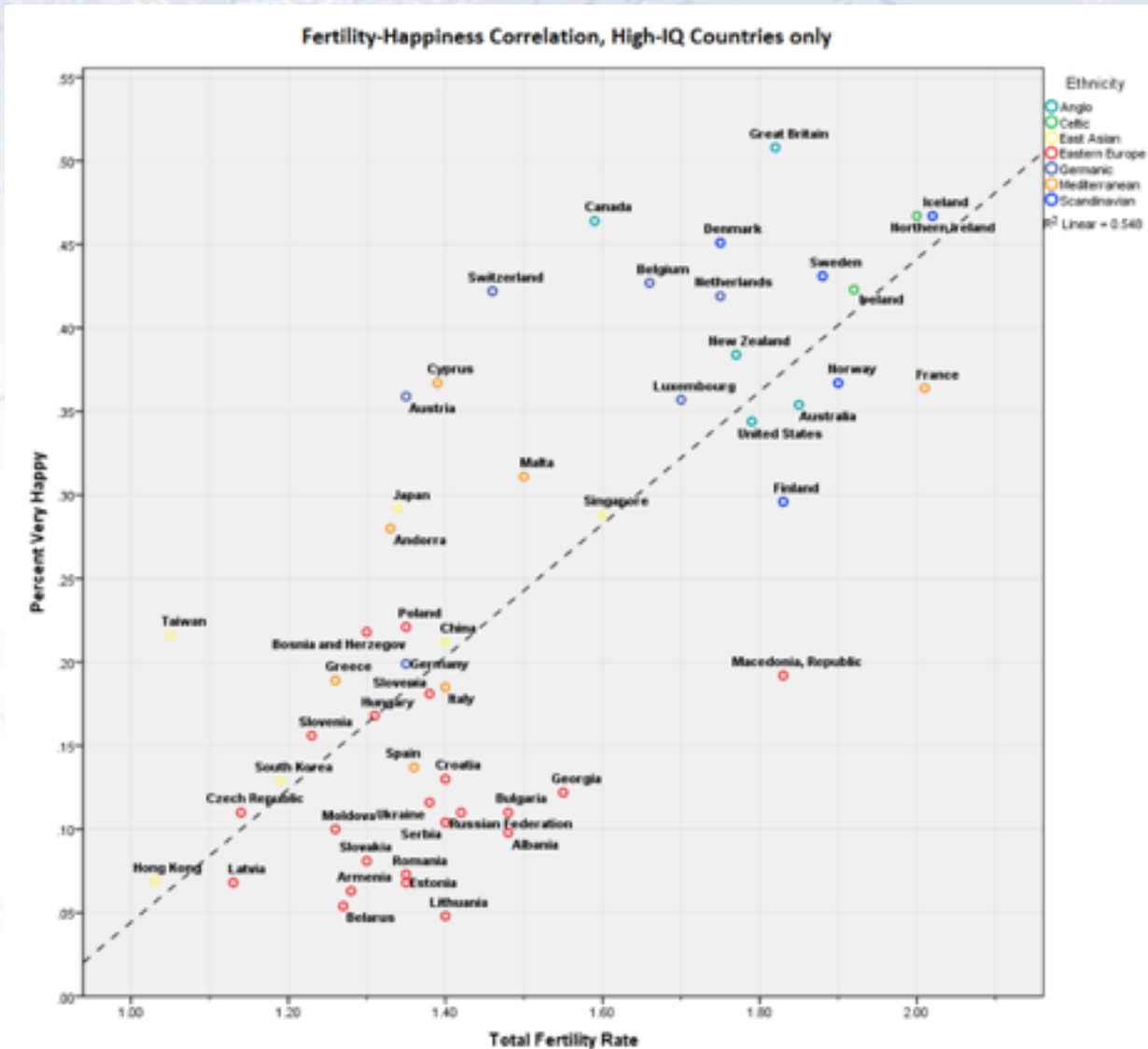Correlations may contain a significant amount of information.

We will consider this more when we play with multivariate analysis.

Correlation example

RELATING DENSITY AND VOTING PATTERNS IN
U.S. CONGRESSIONAL DISTRICTS

% DEMOCRATIC VOTE SHARE IN 2010 CONGRESSIONAL ELECTIONS

PA 1 - East Philadelphia and Suburbs
10,445 ppl/sq mi

CA 31 - Central L.A.
16,275 ppl/sq mi

NY 16 - South Bronx
55,953 ppl/sq mi

CA 8 - San Francisco
18,694 ppl/sq mi

NY 15 - Northern Manhattan/Astoria
66,434 ppl/sq mi

CA 36 - West L.A. County
8,751 ppl/sq mi

↑ Democratic Party Winner
↓ Republican Party Winner

NY 13 - Staten Island
10,823 ppl/sq mi

CA 40 - Northern Orange County
6,607 ppl/sq mi

TX 32 - Dallas Suburbs
4,237 ppl/sq mi

TN 2 - Eastern Tennessee
290 ppl/sq mi

The Transport Politic

POPULATION OF CONGRESSIONAL DISTRICT/SQUARE MILES

Correlation example

Fertility-Happiness Correlation, High-IQ Countries only

Correlation Vs. Causation

"Com hoc ergo propter hoc"

(with this, therefore because of this)

Fig. 1
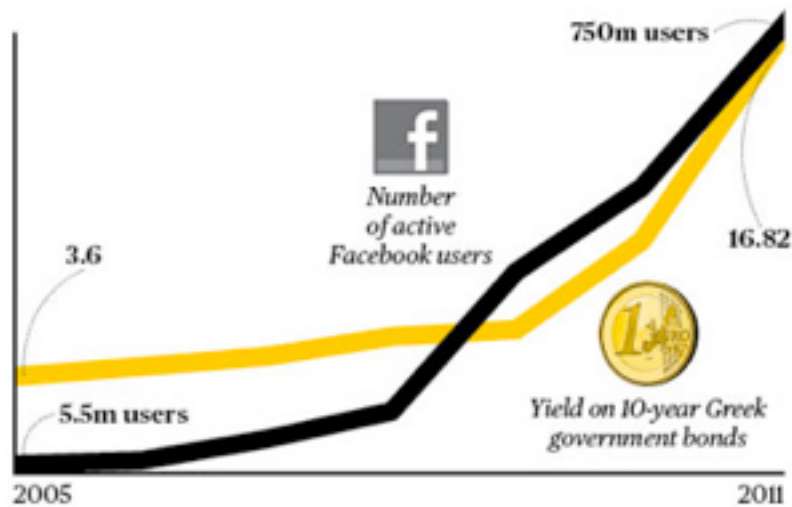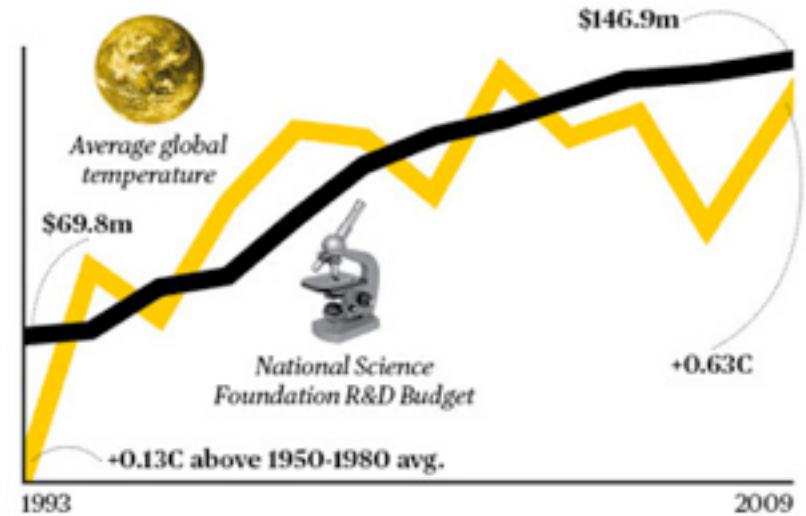**IS FACEBOOK DRIVING
THE GREEK DEBT CRISIS?**

750m users

Number
of active
Facebook users

3.6

16.82

5.5m users

Yield on 10-year Greek
government bonds

2005   2011

Fig. 2
**IS GLOBAL WARMING A HOAX
PROPAGATED BY SCIENTISTS?**

$146.9m

Average global
temperature

$69.8m

National Science
Foundation R&D Budget

+0.63C

+0.13C above 1950-1980 avg.

1993   2009

Correlation Vs. Causation

"Com hoc ergo propter hoc"

(with this, therefore because of this)