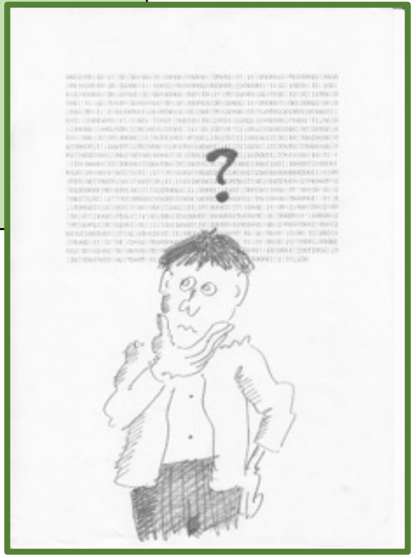
Correlations and their statistical significance - cont

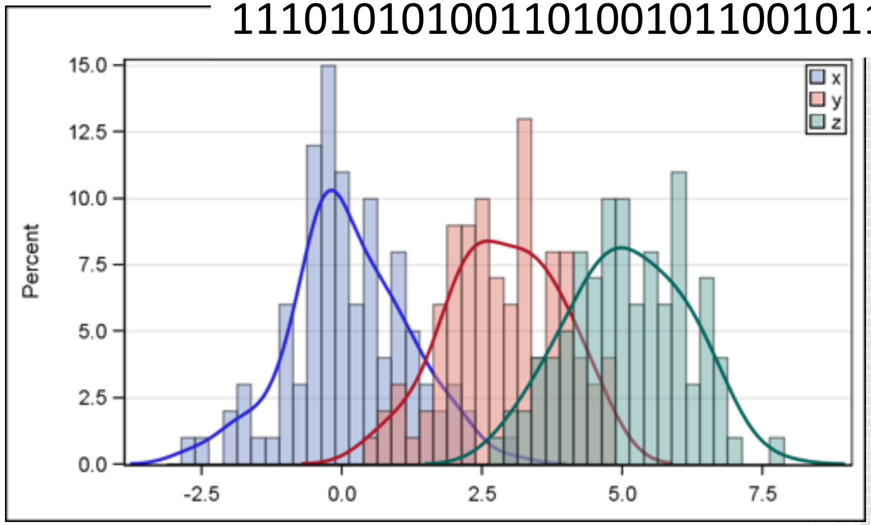
Statistics and data analysis

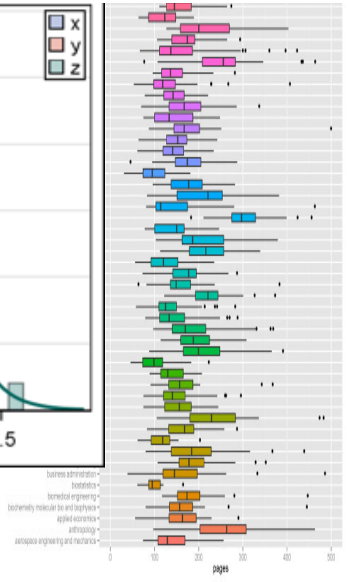
Zohar Yakhini

IDC, Herzeliya



001001110101010010101010010010001
010101000101011111010110100110010
01111010101001101001011001011011001
0





Outline

- Multinomial Correlations
- CI – an example
- Spurious correlations and intro to multiple testing

Multinomial Distribution

- Let $X \sim \mathrm{MNom}(N, P)$, $X = (X_1, X_2, \ldots, X_d)$.
- Example - roll a die N times, counts numbers on each face.
- What is d?
- What is P?
- What is $\mathrm{E}(X_i)$?
- What is $\mathrm{V}(X_i)$?
- Are these random variables collectively independent?
- Pairwise independent?

Multinomial Distribution - covariances

Let $X \sim \mathrm{MNom}(N, P)$, $X = (X_1, X_2, \dots, X_d)$.

$$Var(X_i + X_j) = V(X_i) + 2Cov(X_i, X_j) + V(X_j)$$

Now observe that $X_i + X_j \sim \mathrm{BiNom}(N, p_i + p_j)$ and therefore, from the above identity we get:

$$2Cov(X_i, X_j) = Var(X_i + X_j) - V(X_i) - V(X_j) =$$

$$= N\left[(p_i + p_j)(1 - p_i + p_j) - p_i(1 - p_i) - p_j(1 - p_j)\right]$$

$$= -2Np_ip_j$$

Multinomial Distribution - correlations

Let $X \sim \mathrm{MNom}(N, P)$, $X = (X_1, X_2 , \dots , X_d)$.
We saw that $Cov(X_i, X_j) = -Np_i p_j$
Using the fact that $\forall i \; Var(X_i) = Np_i(1 - p_i)$ we can conclude that

$$\rho(X_i, X_j) = \frac{-\sqrt{p_i p_j}}{\sqrt{(1 - p_i)(1 - p_j)}}$$

For the case of a fair die we get, for example, $\rho(X_1, X_2) = -\frac{1}{5}$.

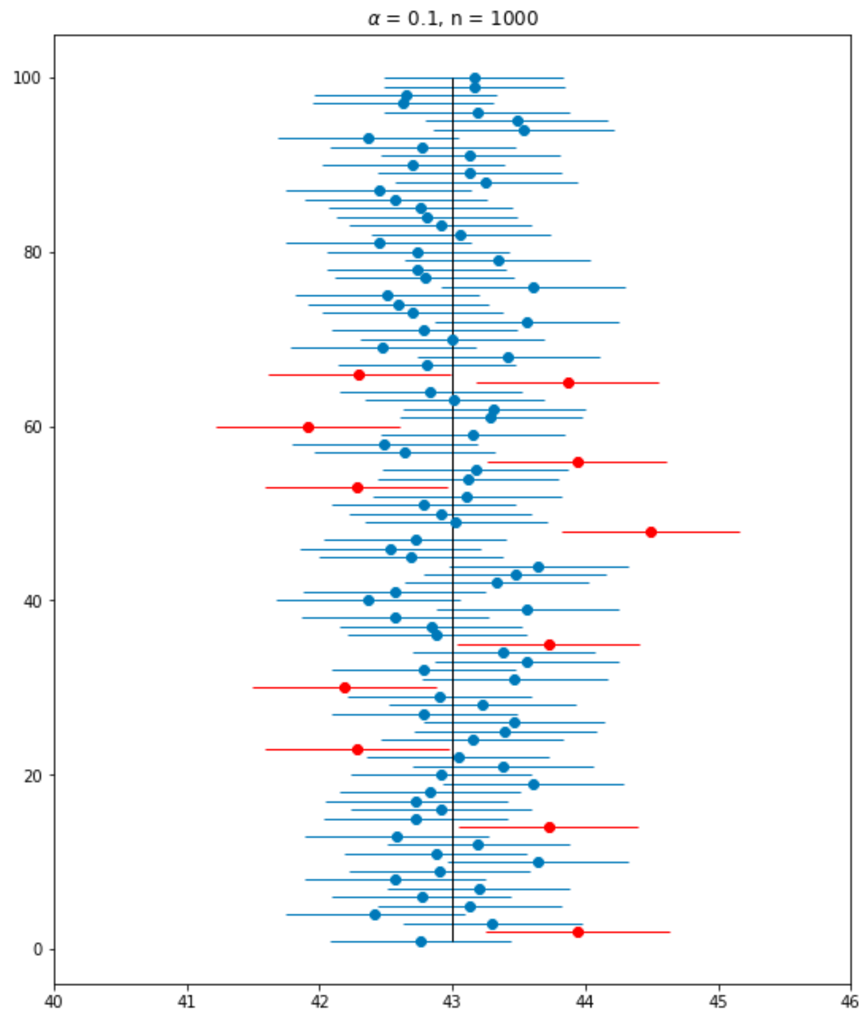
Confidence intervals

$$P\left(\hat{p}(A) - \gamma \cdot \hat{\sigma}(A) \le p \ \le \hat{p}(A) + \gamma \cdot \hat{\sigma}(A)\right) \approx 2\Phi(\gamma) \ - 1$$

where $\hat{p}(A)$ is the empirical proportion

and $\hat{\sigma}(A) = \dfrac{\sqrt{\hat{p}(A)(1-\hat{p}(A))}}{\sqrt{n}}$

$$P\left(p \in \hat{p}(A) \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \hat{\sigma}(A)\right) \approx 1 - \alpha$$



CIs for correlations?
based on the Fisher Transform

$$\Sigma = \begin{pmatrix} V(X) & Cov(X,Y) \\ Cov(X,Y) & V(Y) \end{pmatrix}$$

$$Cov(X,Y) = \rho(X,Y)\sigma(X)\sigma(Y)$$

When drawing n samples from a bivariate normal distribution with correlation coefficient ρ and computing the empirical Pearson correlation on the sample, $\hat{\rho}_n$, we have:

$$P\left(F(\rho) \in F(\hat{\rho}_n) \pm \frac{1}{\sqrt{n-3}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha$$

And also:

$$P\left(F(\rho) \geq F(\hat{\rho}_n) - \frac{1}{\sqrt{n-3}} \Phi^{-1}(1 - \alpha)\right) = 1 - \alpha$$

How do we use this?

Suppose that we obtained $\hat{\rho} = 0.8$ when calculated for grades in two exams in a class of 20 students.

What can you say about the correlation between the grades in these two exams, in the entire world population?
Assume that the grades in these two exams follow a bivariate normal joint distribution.
Also, assume, of course, that our 20 students were sampled from the relevant (entire world) population.

How do we use this?

we obtained $\hat{\rho} = 0.8$ when calculated for grades in two exams in a class of 20 students.

Therefore, working with $\alpha = 0.05$, say, we have

$$F(\rho) \in F(0.8) \pm \frac{1}{\sqrt{17}} \Phi^{-1}(0.975) = 1.099 \pm 0.243 \cdot 1.96 = [0.624 , 1.574]$$

To be useful, we now want to convert this to a CI for ρ .
Since the Fisher Transform, F, is monotone and invertible we get:

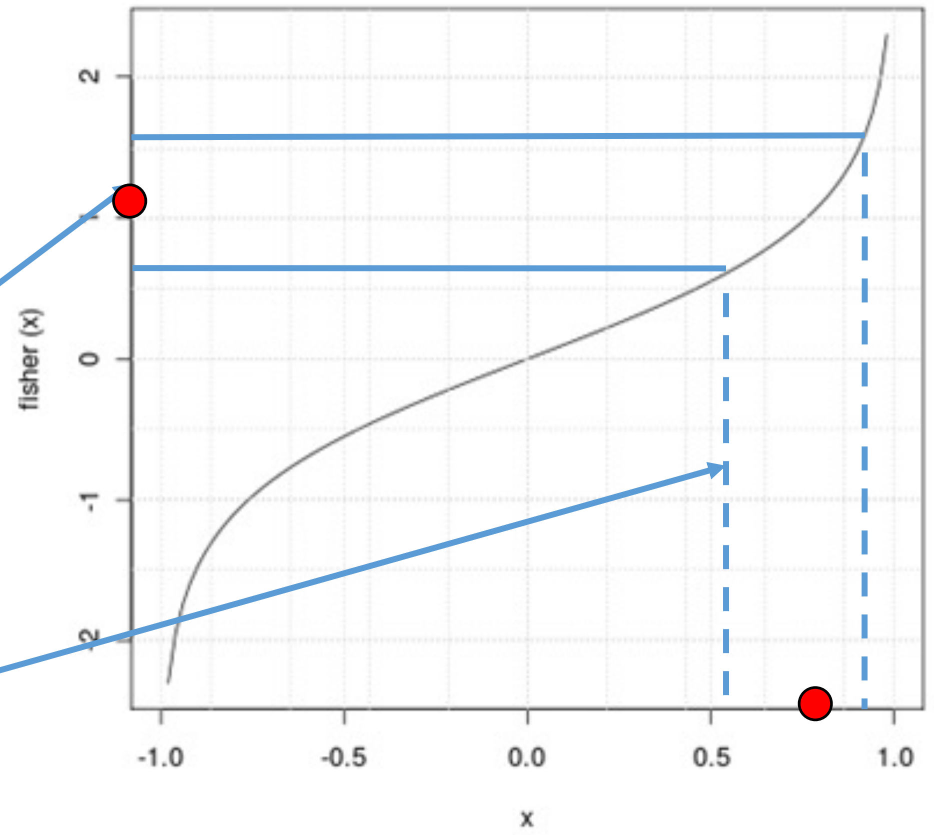
$$\rho \in F^{-1}([0.624 , 1.574]) = [\, F^{-1}(0.624) , F^{-1}(1.574) \,] = [0.554 , 0.918] \ .$$

And so we can state that $0.554 \leq \rho \leq 0.918$ with 95% confidence.

Inverting the Fisher Transform

Note: the CI around ρ is NOT symmetric

$$r = F^{-1}(u) = \frac{e^{2u} - 1}{e^{2u} + 1}$$



CI for $F(\rho)$, obtained from a normal cdf

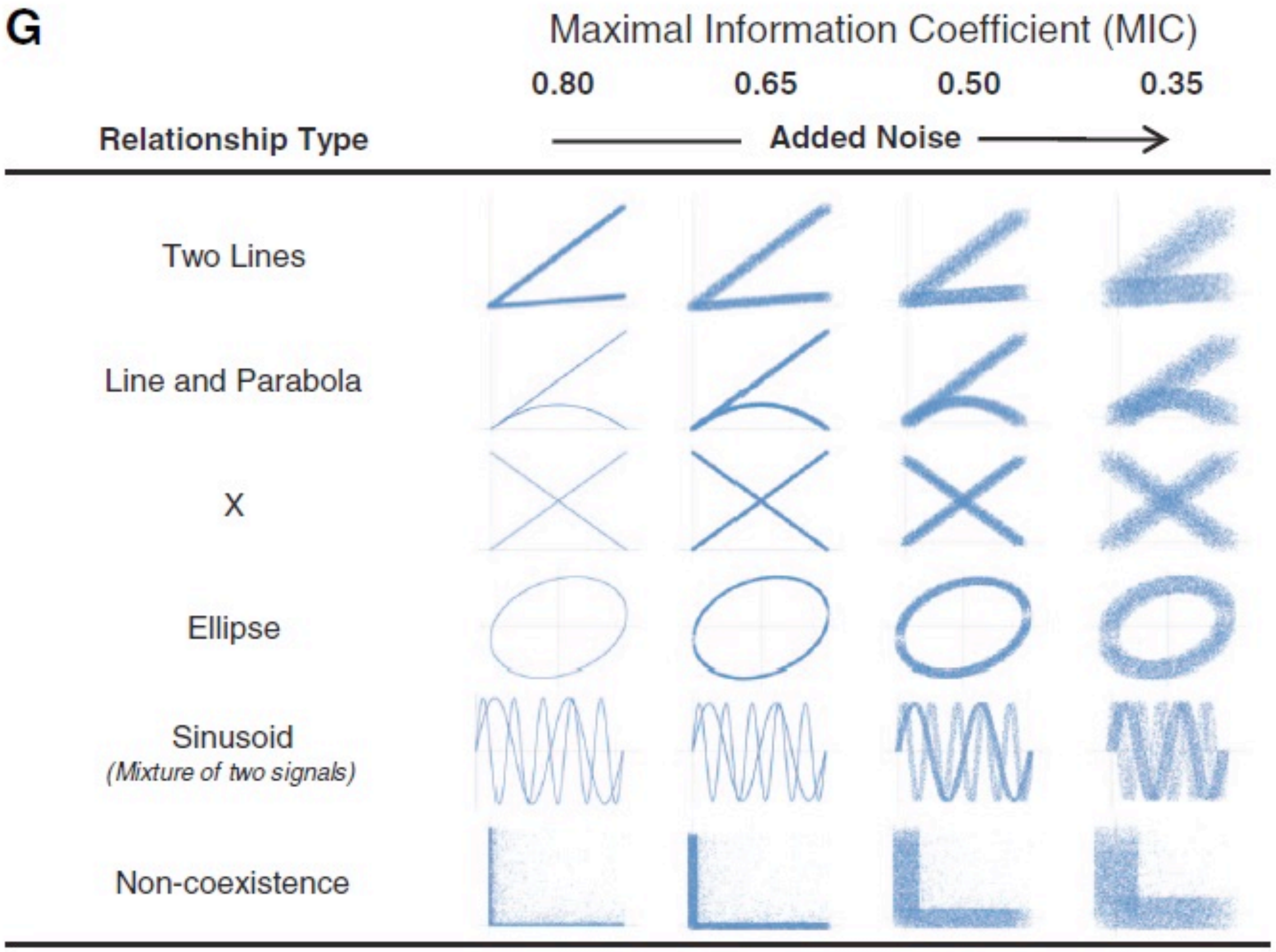
CI for ρ , obtained from the above by taking the inverse image, under F.

p-Values for correlations

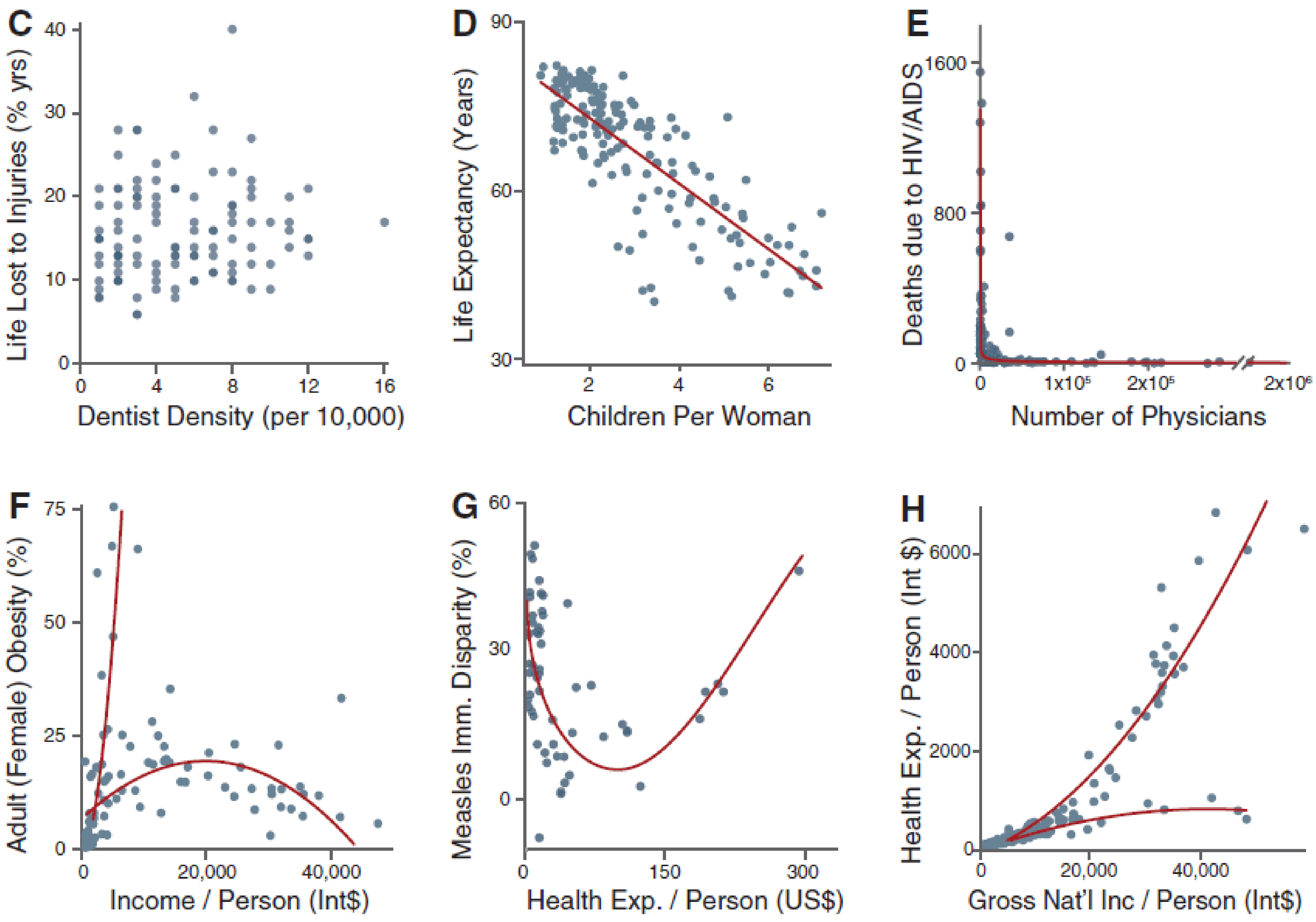
Built in functions typically compute p-values for Pearson's correlation using normal approximations (+Fisher), under the null model of X and Y following an independent standard bivariate normal distribution.

Built in functions compute p-values for Kendall's tau and Spearman's rho using either the exact permutation distributions (for small sample sizes), or large-sample normal approximations.
(what is the null model?)

Correlations don’t cover all association types

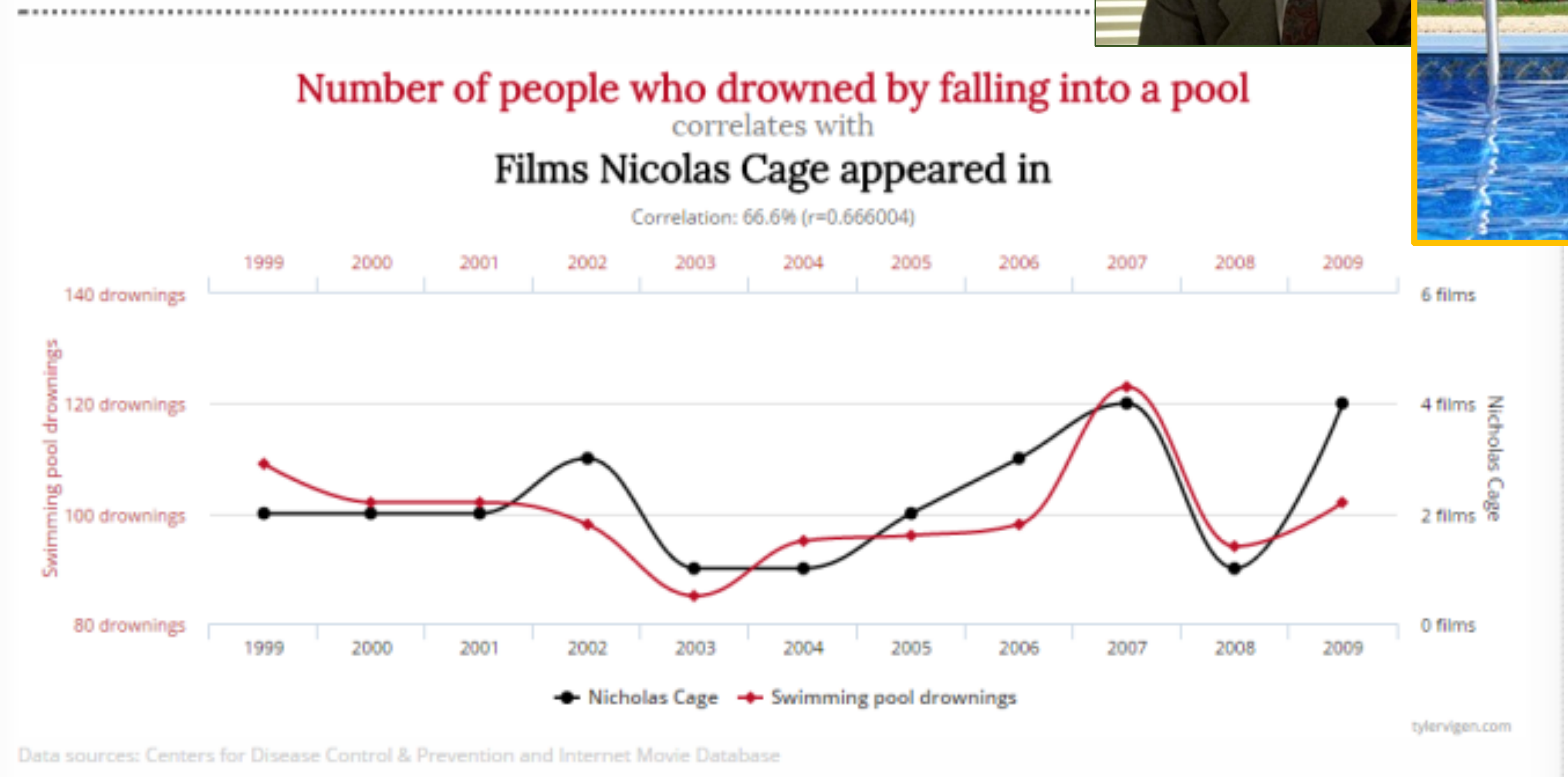


Reshef et al,
Detecting Novel Associations in Large Data Sets,
Science 2011

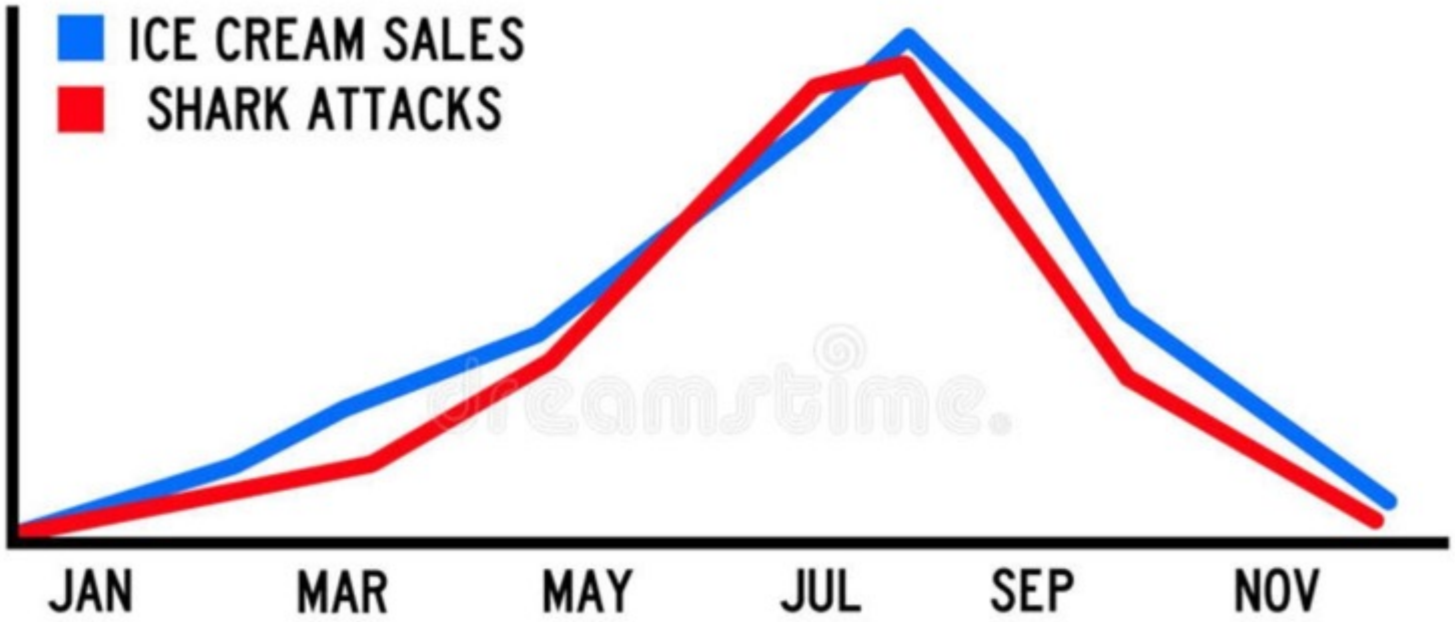


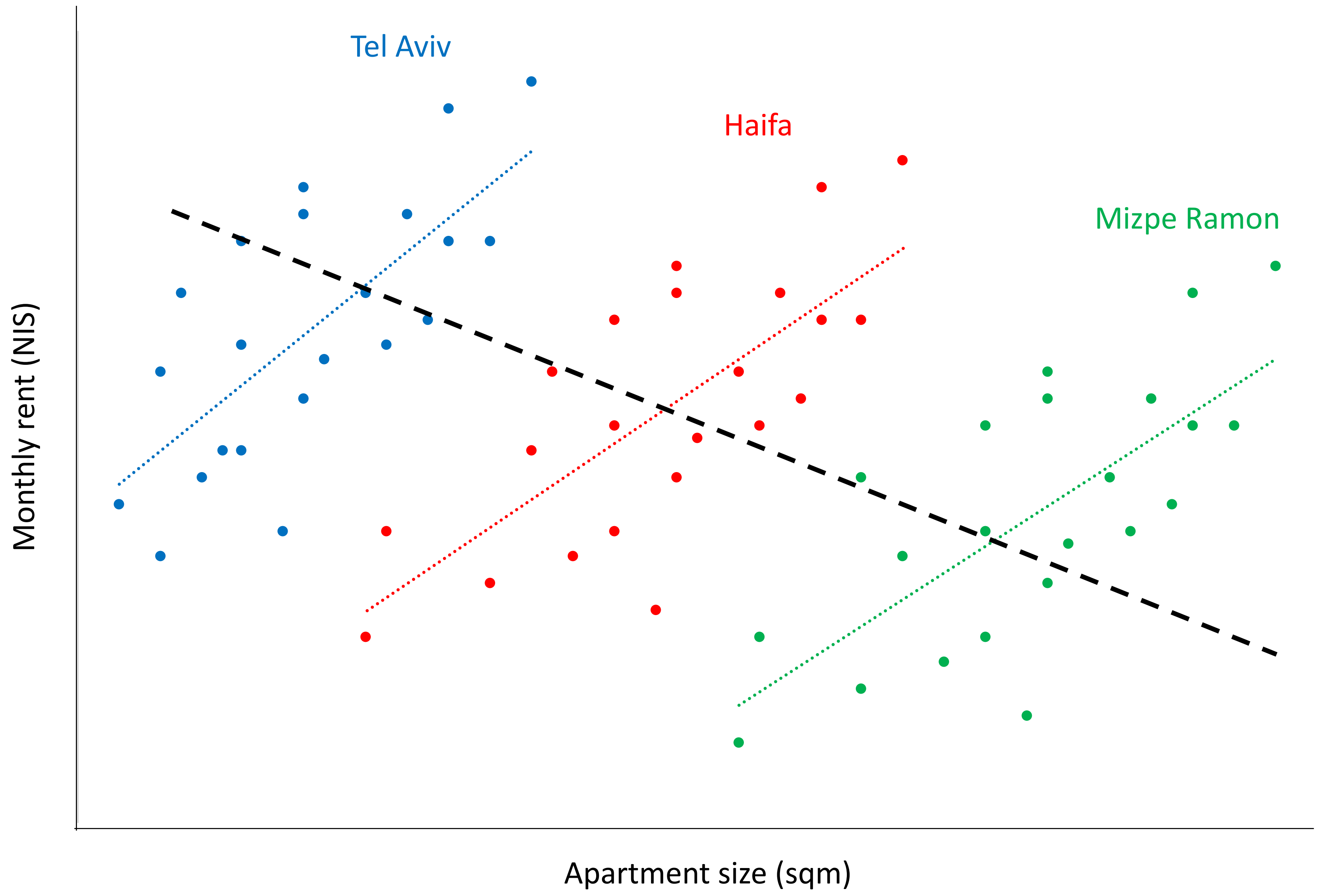
C
Life Lost to Injuries (% yrs)
Dentist Density (per 10,000)
D
Life Expectancy (Years)
Children Per Woman
E
Deaths due to HIV/AIDS
Number of Physicians
F
Adult (Female) Obesity (%)
Income / Person (Int$)
G
Measles Imm. Disparity (%)
Health Exp. / Person (US$)
H
Health Exp. / Person (Int $)
Gross Nat'l Inc / Person (Int$)

Spurious Correlations



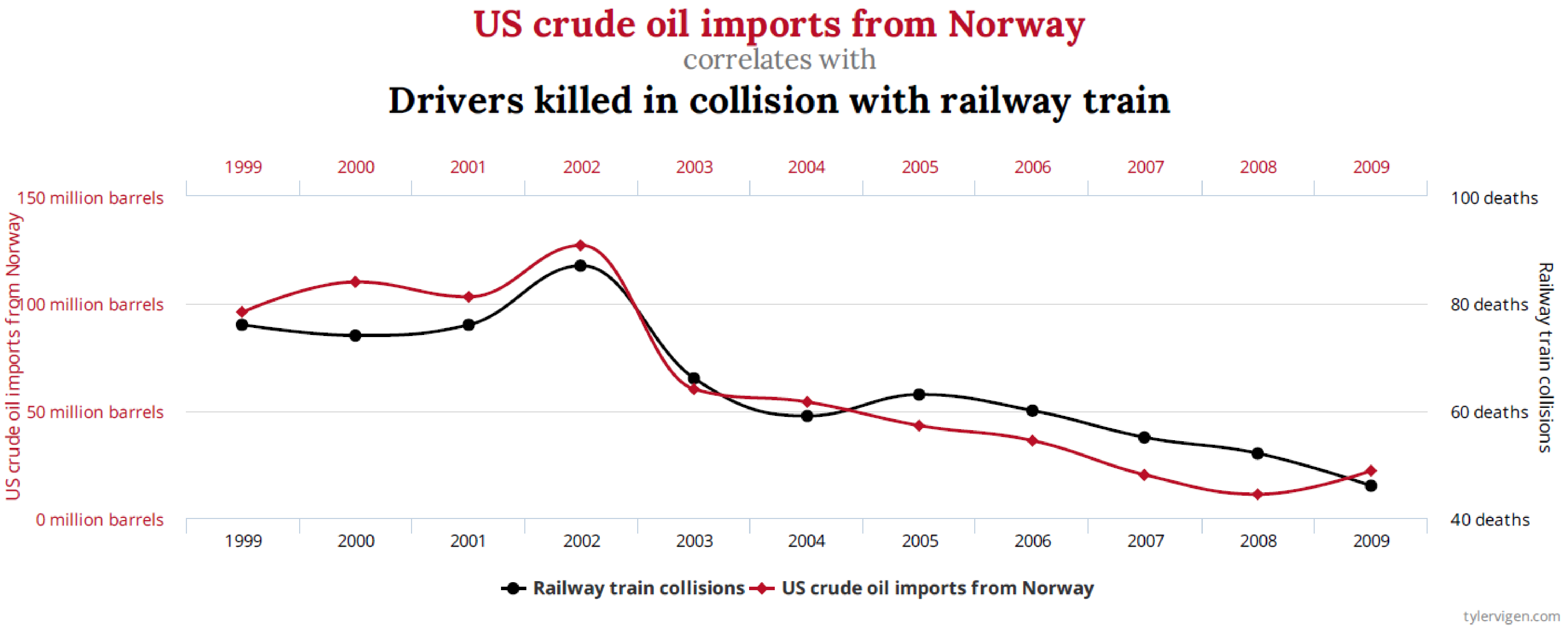
Correlation and causality

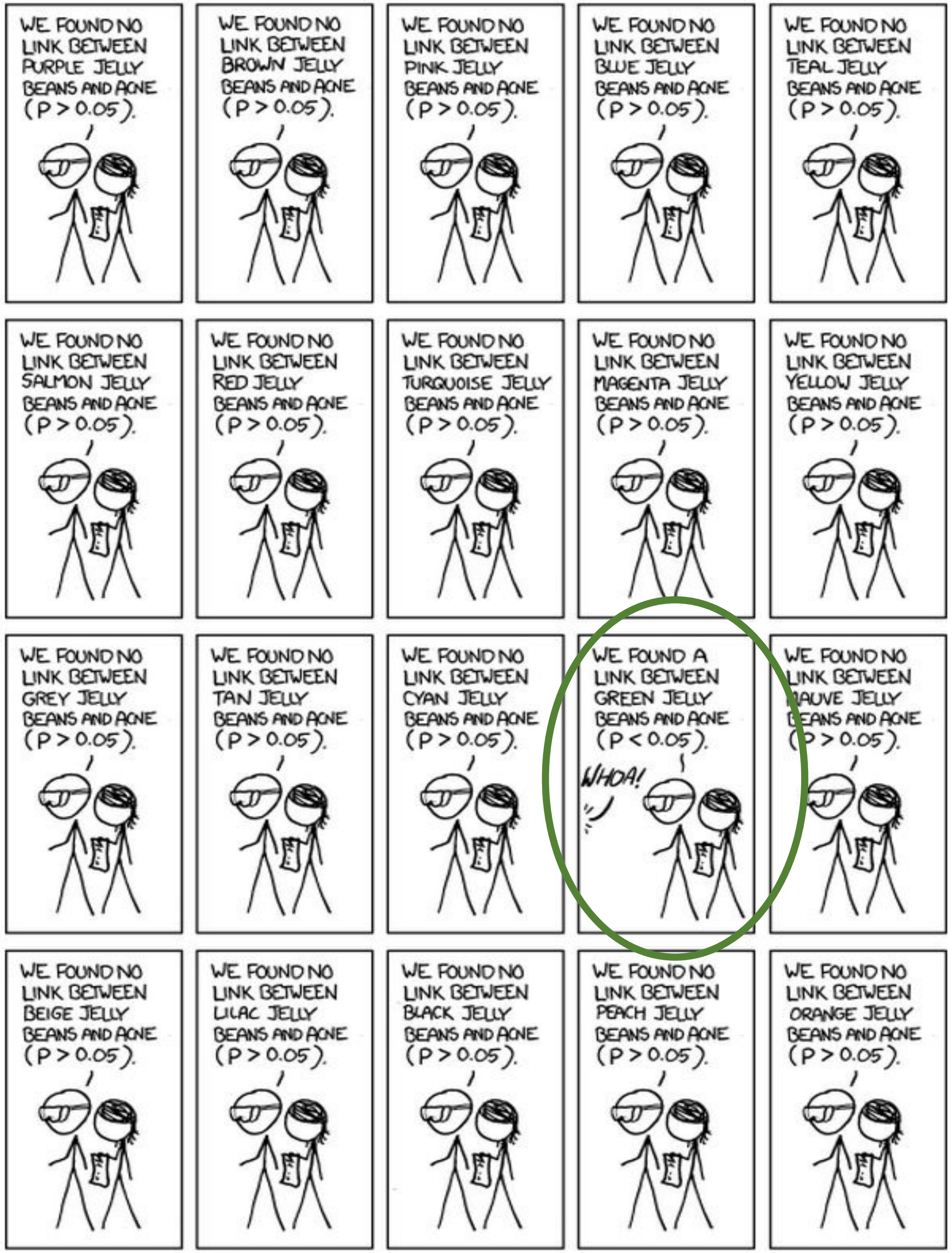




Tel Aviv
Haifa
Mizpe Ramon
Monthly rent (NIS)
Apartment size (sqm)

Spurious correlations




WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

Summary

- Correlation measures serve to quantify relationships between different aspects of observed data
- Statistical assessment under a null model
- Pearson correlation is the classical and most popular correlation coefficients. It’s a sample version of the population covariance.
- Pearson correlation affords confidence interval calculations.
- How can we capture more complex relationships between two variables/features/quantities?
- Between more than two variables/features/quantities?
- Finally – again, always assess presented correlations with a good measure of skepticism and attention to multiple testing