# MCMC Coffee | Season 1, Episode 2

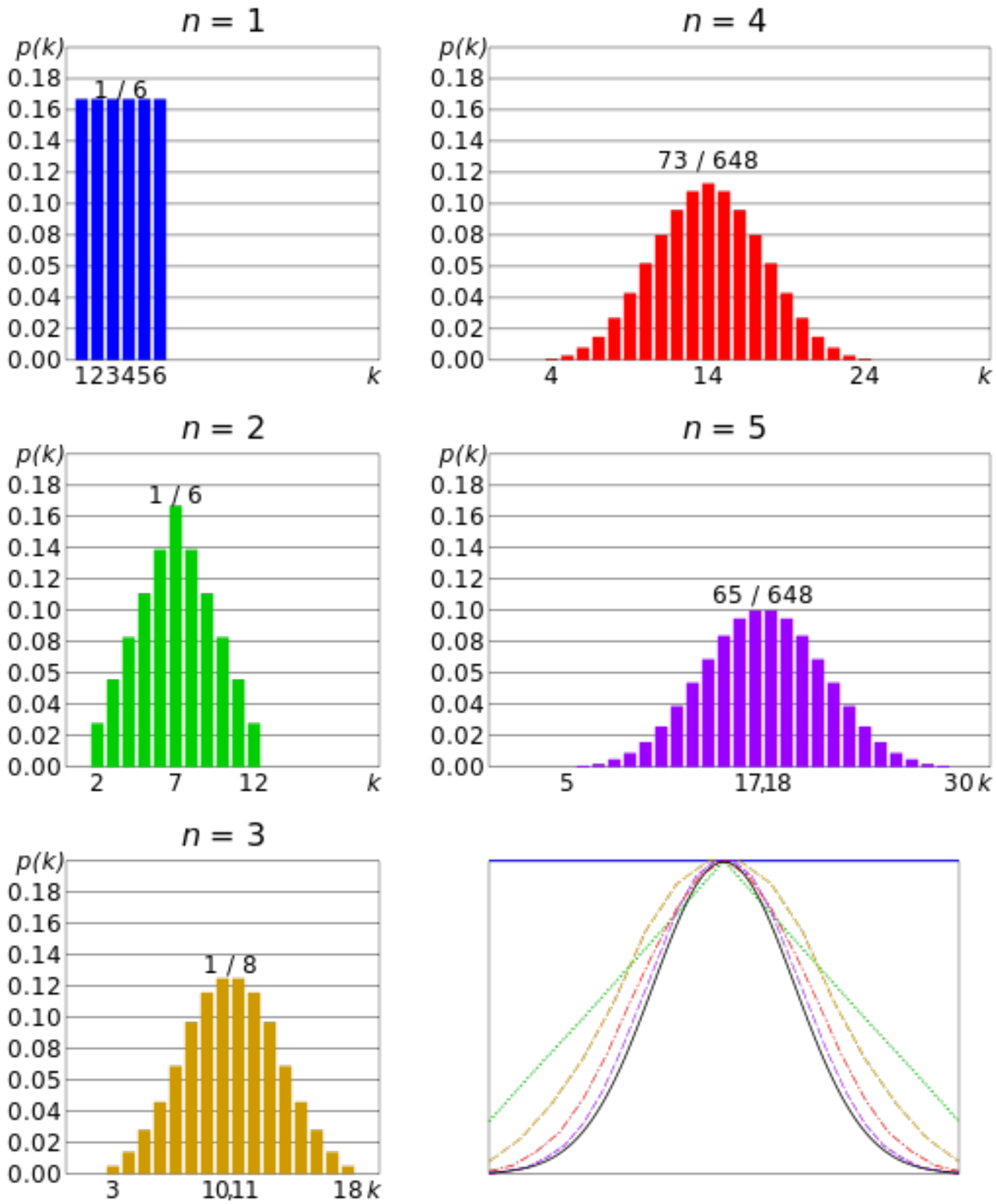
# Central Limit Theorem & Correlation Coefficients

Daniel Asmus

# The Central limit theorem

*Chocolate Hills in Bohol, Philippines*

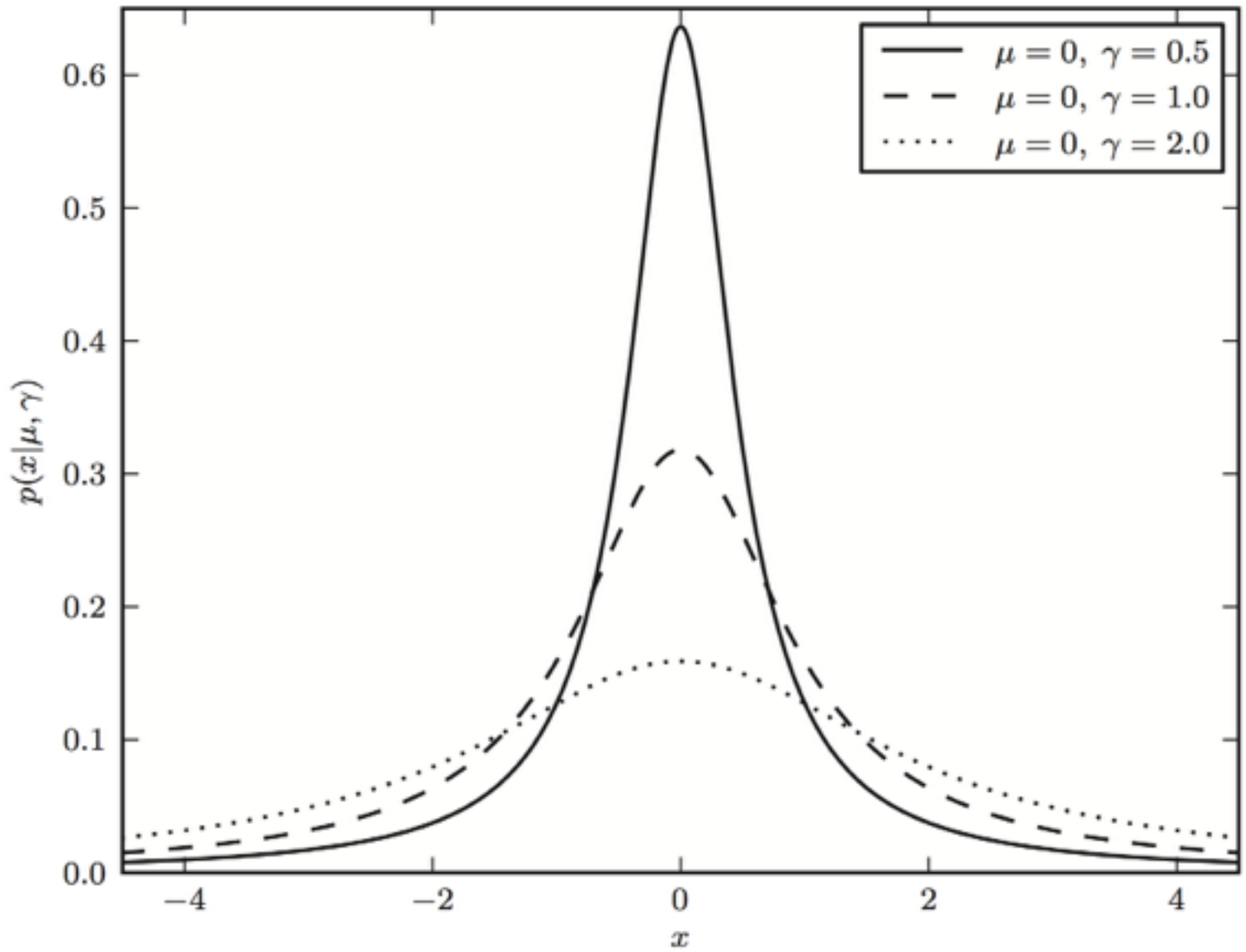
Given an *arbitrary* distribution $h(x)$, characterised by its mean $\mu$ and standard deviation $\sigma$, the central limit theorem says that the mean of $N$ values $x$ drawn from that distribution will approximately follow a Gaussian distribution $G(\mu,\sigma/\sqrt{N})$, with the approximation accuracy improving with $N$.




By Cmglee - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=18918612

# Notable exception: Cauchy (Lorentzian Distribution)

$$p(x|\mu,\gamma) = \frac{1}{\pi\gamma}\left(\frac{\gamma^2}{\gamma^2 + (x-\mu)^2}\right)$$

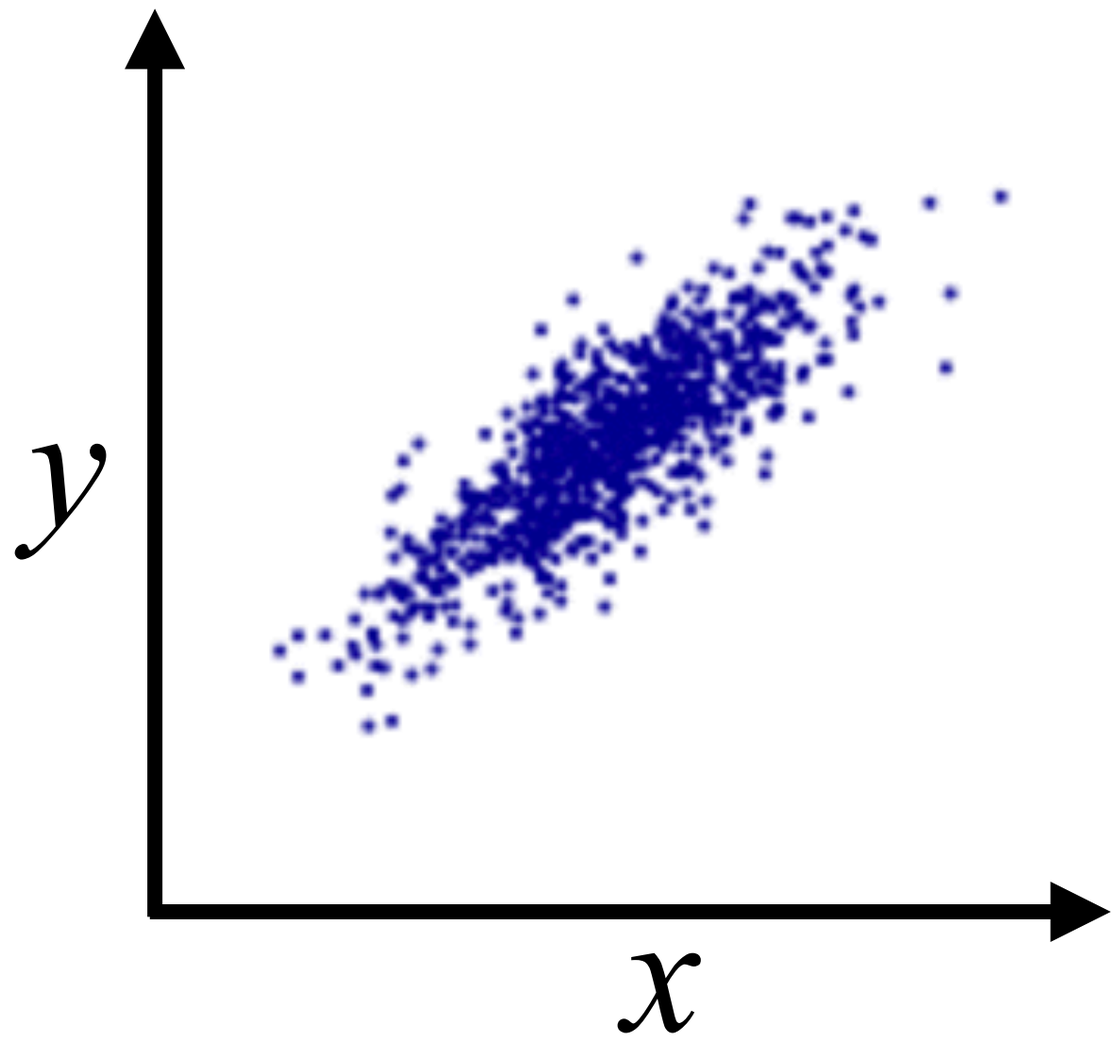


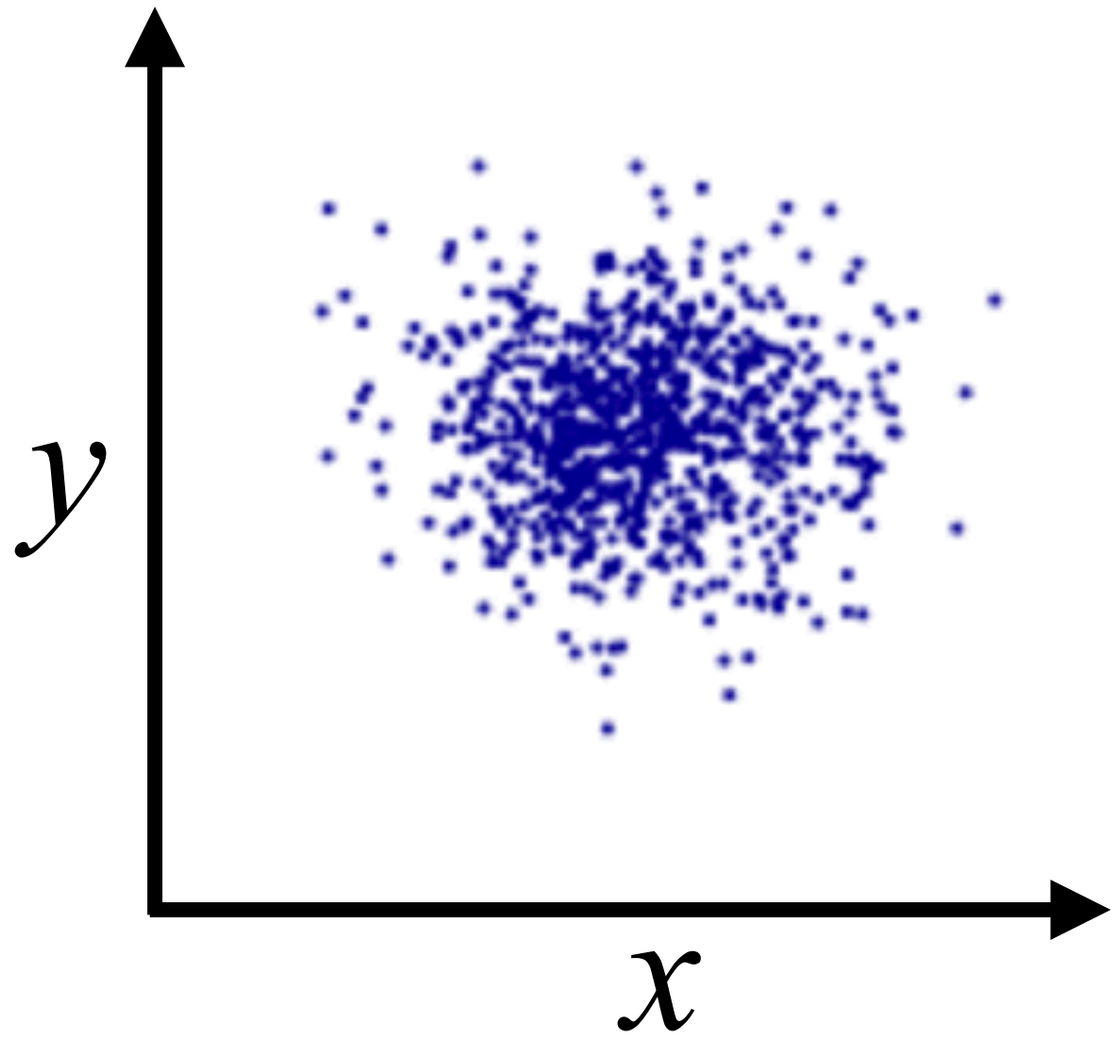

because mean & standard deviation ill-defined here
(tails drop only with $x^{-2}$)

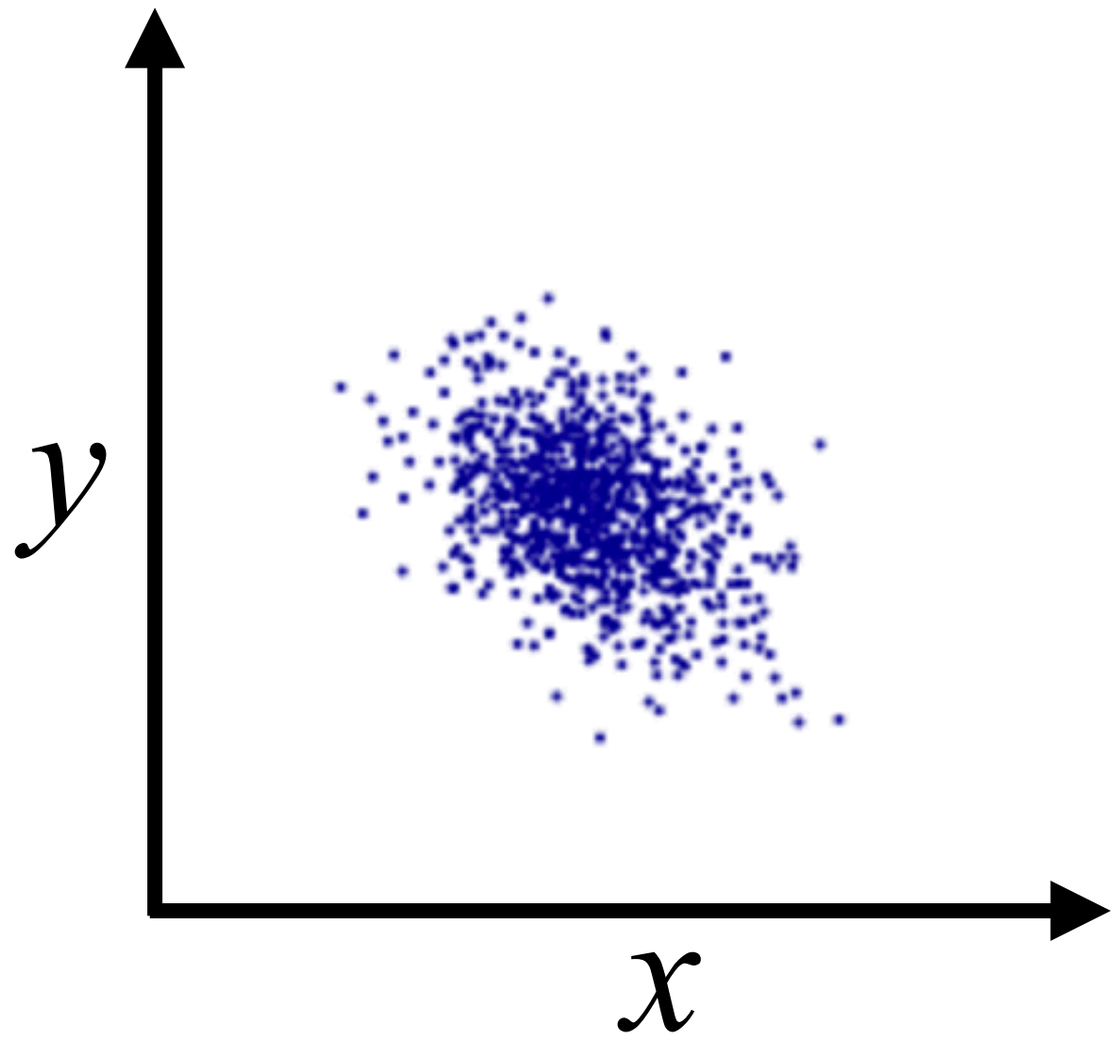
*Ivezic, Connolly, VanderPlas & Gray, 2014*

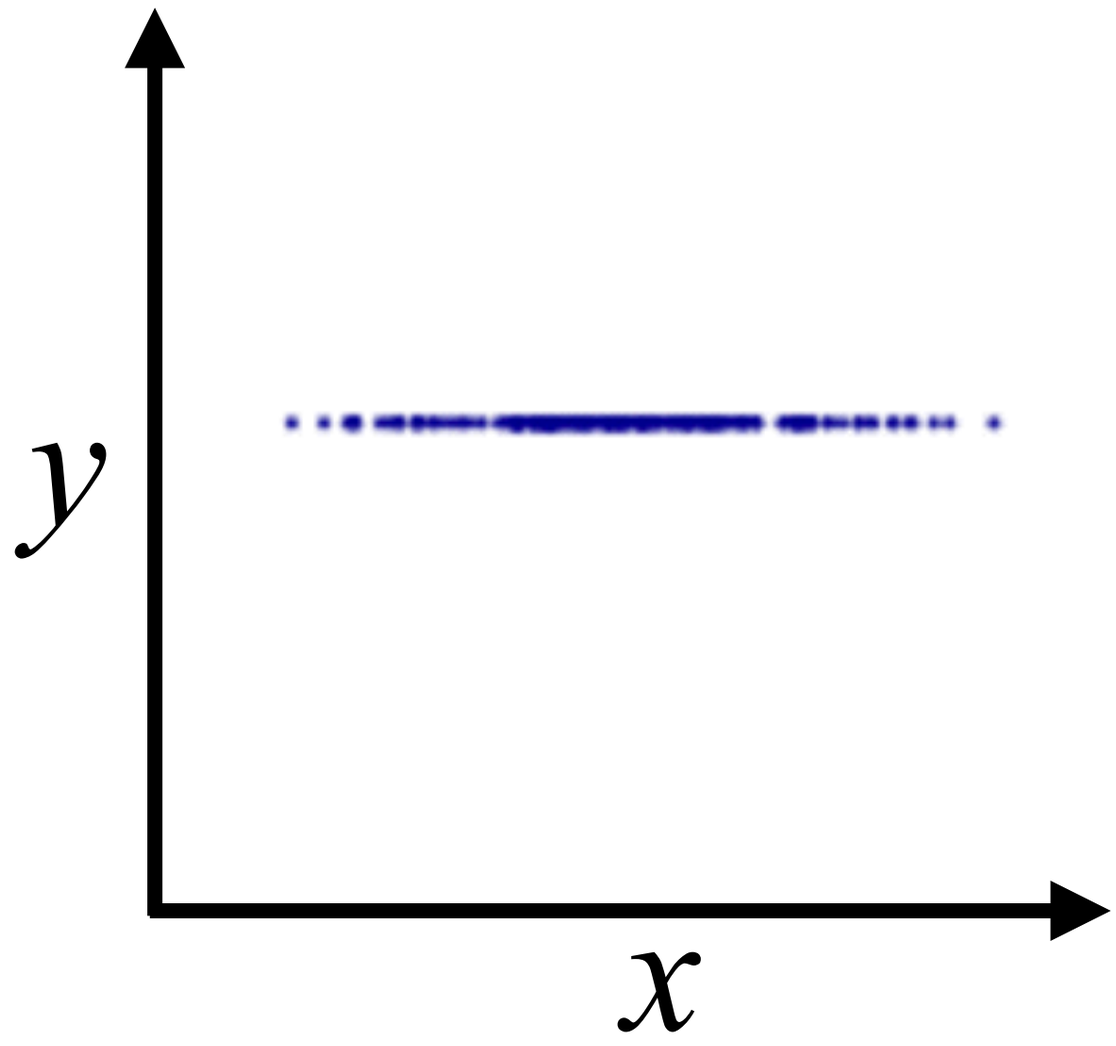
# Correlation

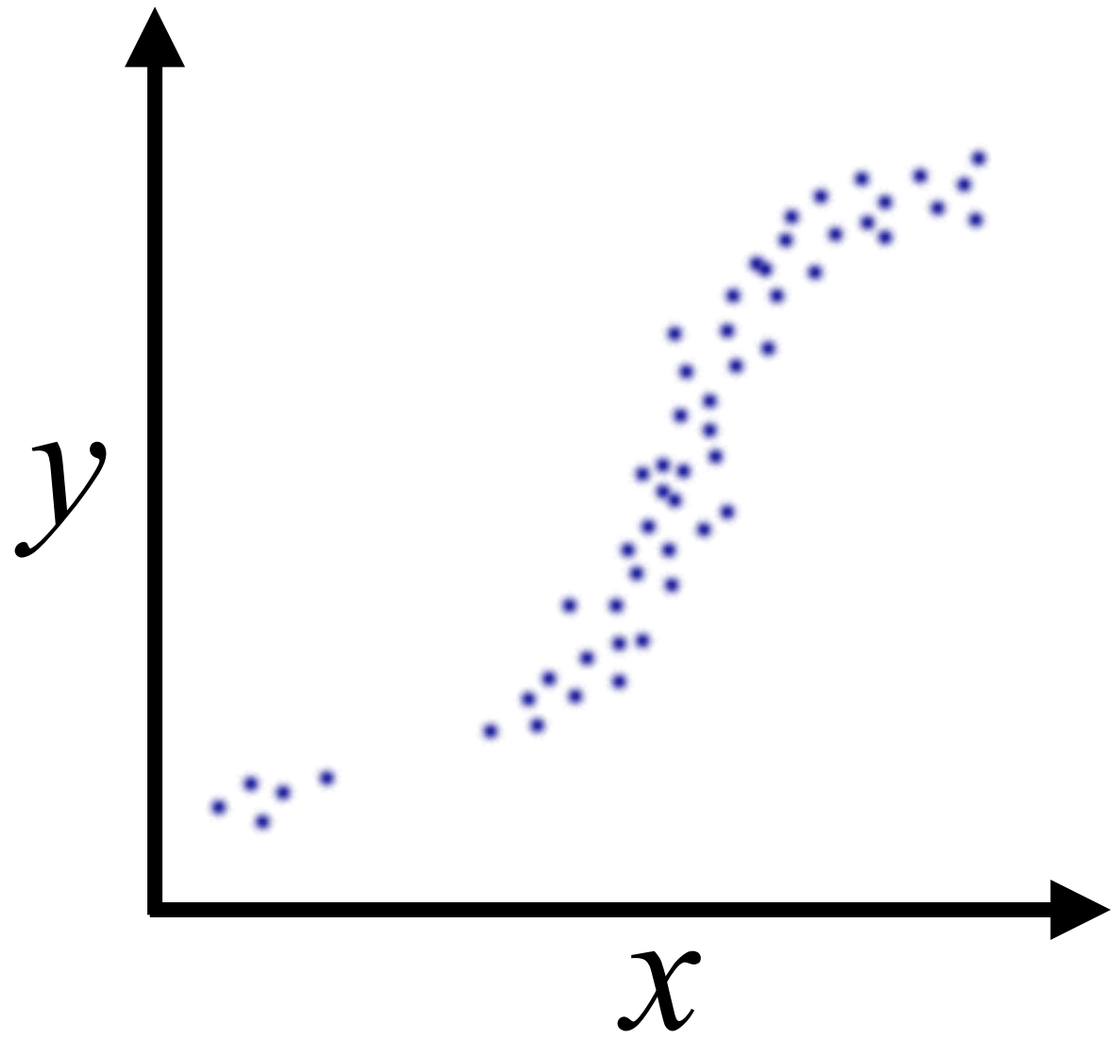
Are $x$ and $y$ correlated?

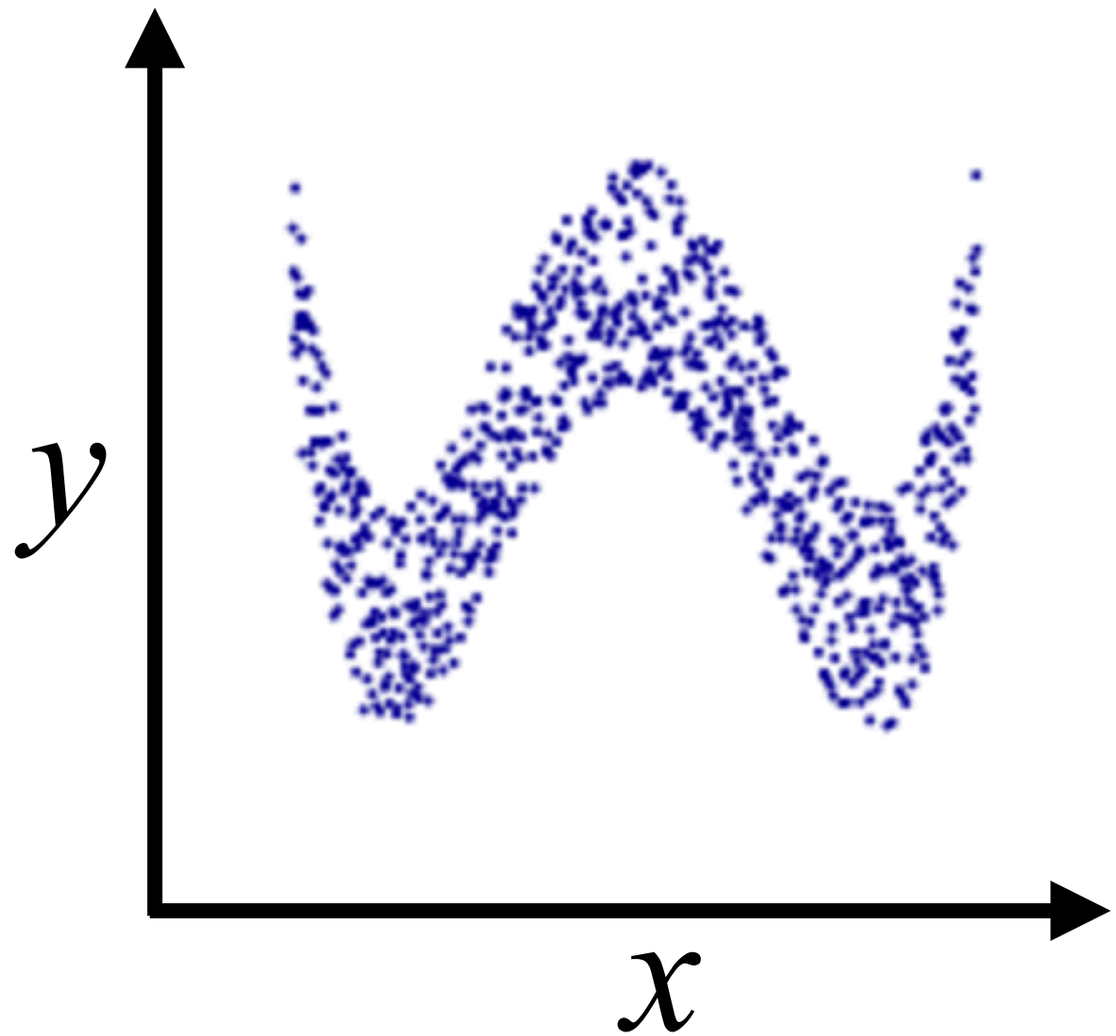

y
x

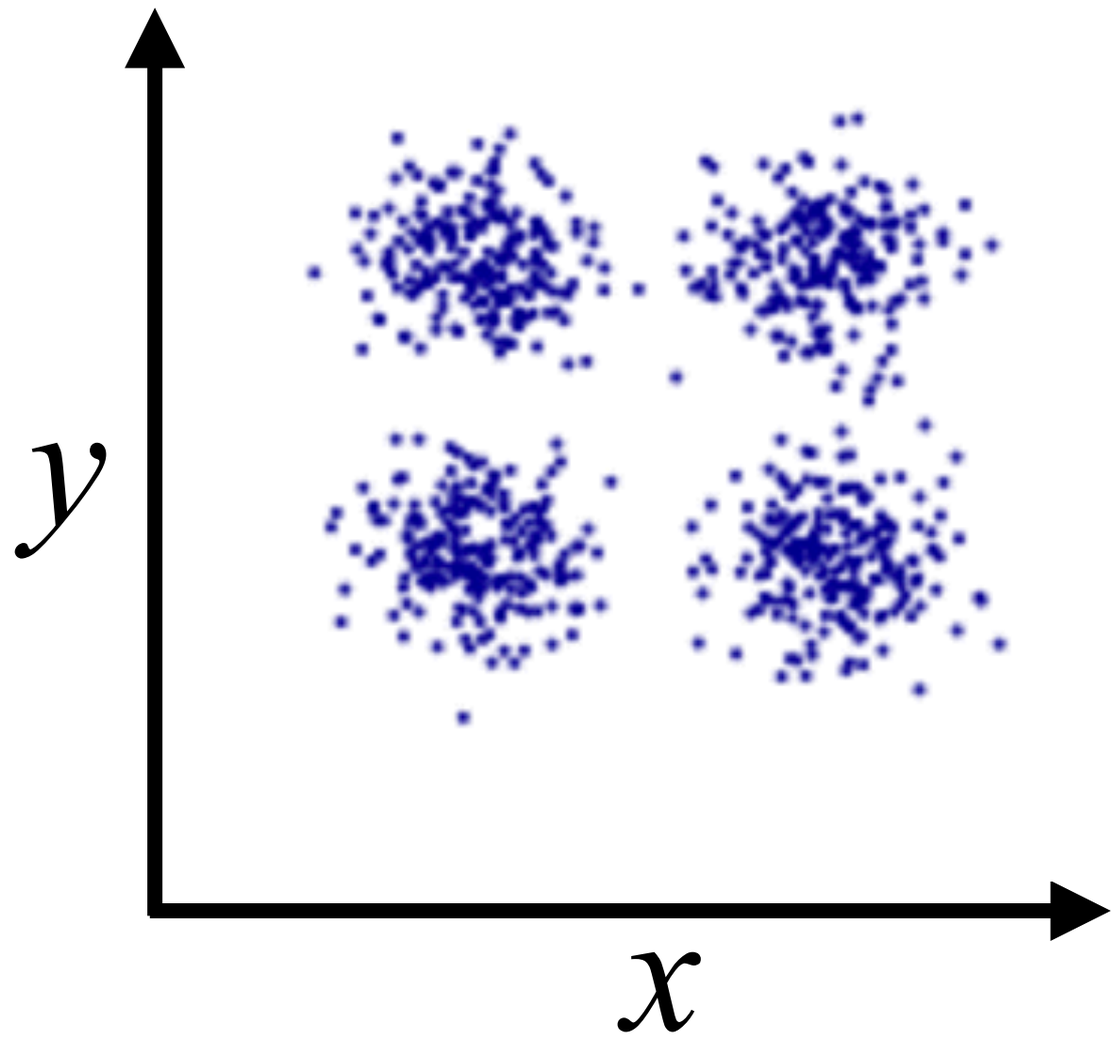

y
x


y
x


y
x

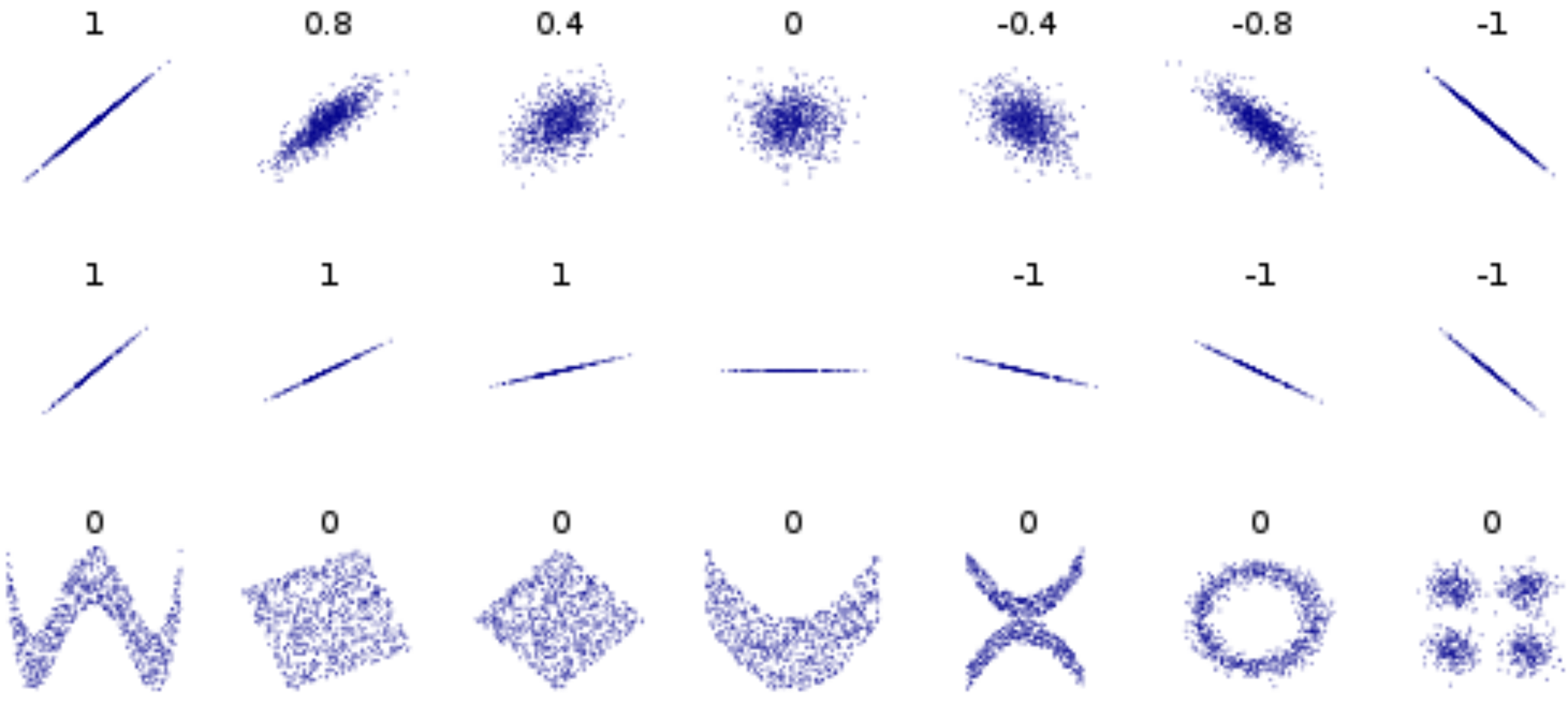

y
x


y
x


y
x

# The straightforward approach: Pearson's sample correlation coefficient

$$r = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}},$$




By DenisBoigelot, original uploader was Imagecreator - Own work, original uploader was Imagecreator, CC0, https://commons.wikimedia.org/w/index.php?curid=15165296

# No correlation without a probability

The p-value is defined as the probability of obtaining a result equal to or "more extreme" than what was actually observed, when the null hypothesis is true.

Different ways to calculate the p-value:

- permutation test
- bootstraping
- r follows Student's t distribution for no correlation
- Fisher transformation

Result reliable only for larger samples (500+ or so)

Measurement errors can be included with Monte Carlo resampling

Disadvantages of Pearson's coefficient:

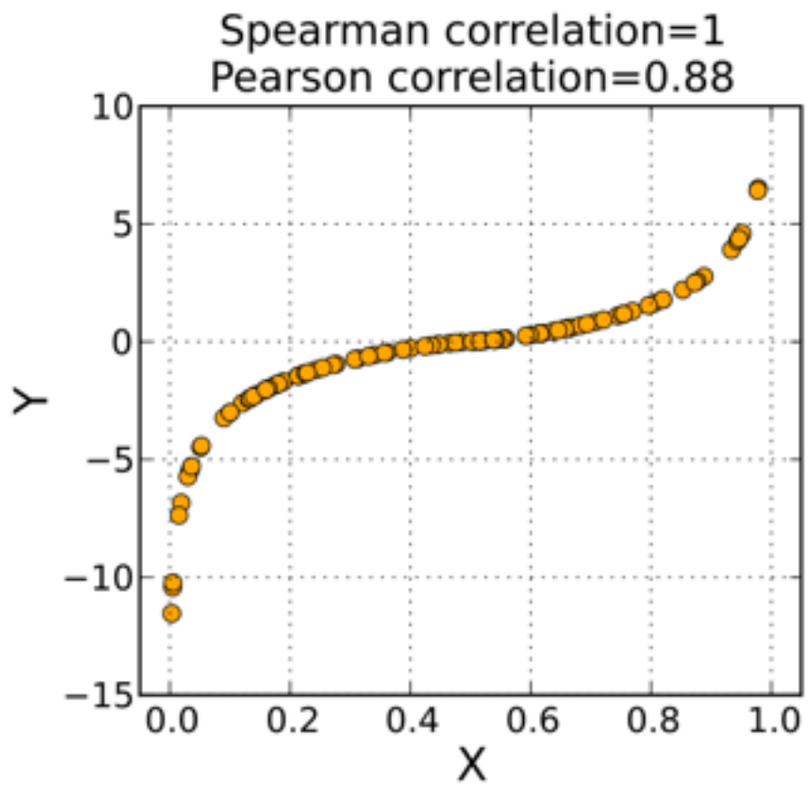
- measures only linear correlations
- assume normally distributed variables
- sensitive to outliers
- …

# Spearman's correlation coefficient

Using of ranks: sort data set $x_i$ in ascending order. The index $i$ of the sorted data is its rank, $R_i^x$

$$r_S = \frac{\sum_{i=1}^{N}(R_i^x - \overline{R^x})(R_i^y - \overline{R^y})}{\sqrt{\sum_{i=1}^{N}(R_i^x - \overline{R^x})^2}\sqrt{\sum_{i=1}^{N}(R_i^y - \overline{R^y})^2}},$$

p-value can be computed similar to Pearson's r



# Kendall's correlation coefficient

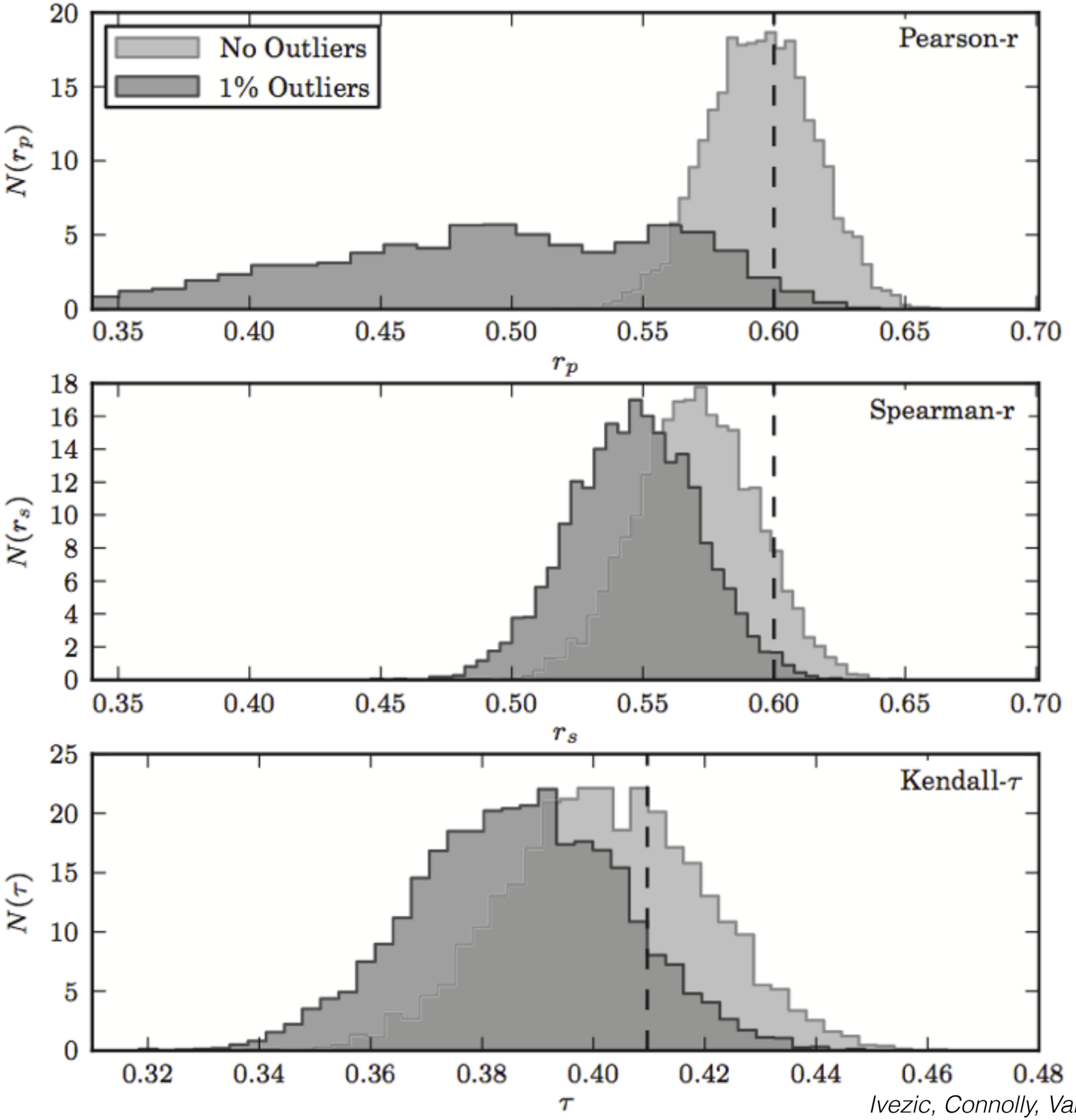
Like Spearman's coeff but instead of using the actual differences, $R_i^x - R_i^y$, count the numbers of concordant, $(x_j - x_k)(y_j - y_k) > 0$, and discordant pairs $(x_j - x_k)(y_j - y_k) < 0$ for $R_j^x = R_j^y$ and $R_k^x = R_k^y$

$$\tau = 2\frac{N_c - N_d}{N(N-1)}$$

p-value can be computed using a Gaussian with $\mu = 0$ and

$$\sigma_\tau = \left[\frac{2(2N+5)}{9N(N-1)}\right]^{1/2}$$

approximating the distribution of Kendall's $\boldsymbol{\tau}$




*Ivezic, Connolly, VanderPlas & Gray, 2014*

## Pearson's:

- parametric test
- tests for linear relationship
- variables should be normally distributed
- very sensitive to outliers

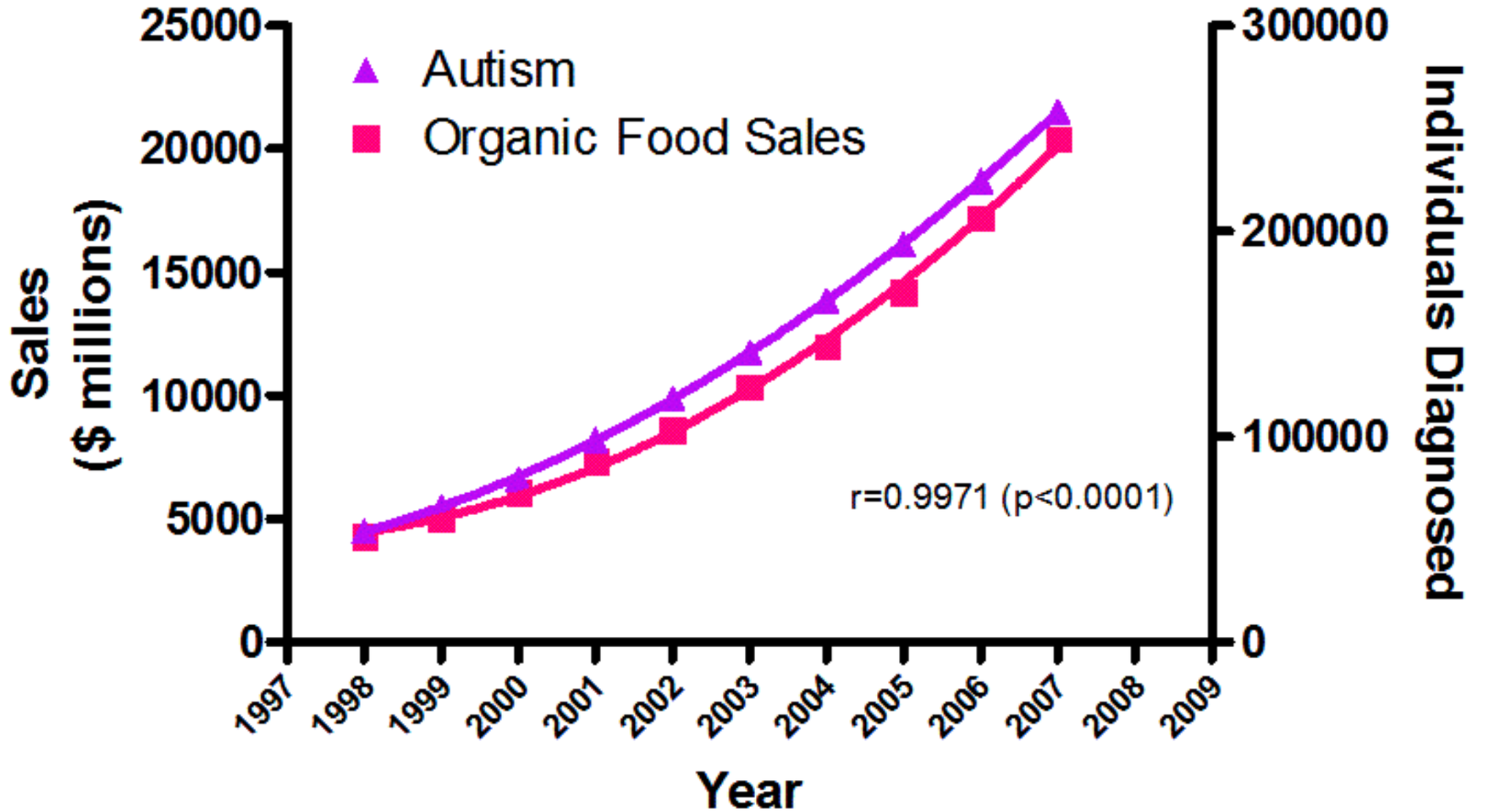
## Spearman's:

- non-parametric
- tests for monotonic relationship
- any variable distribution
- somewhat sensitive to outliers

## Kendall's:

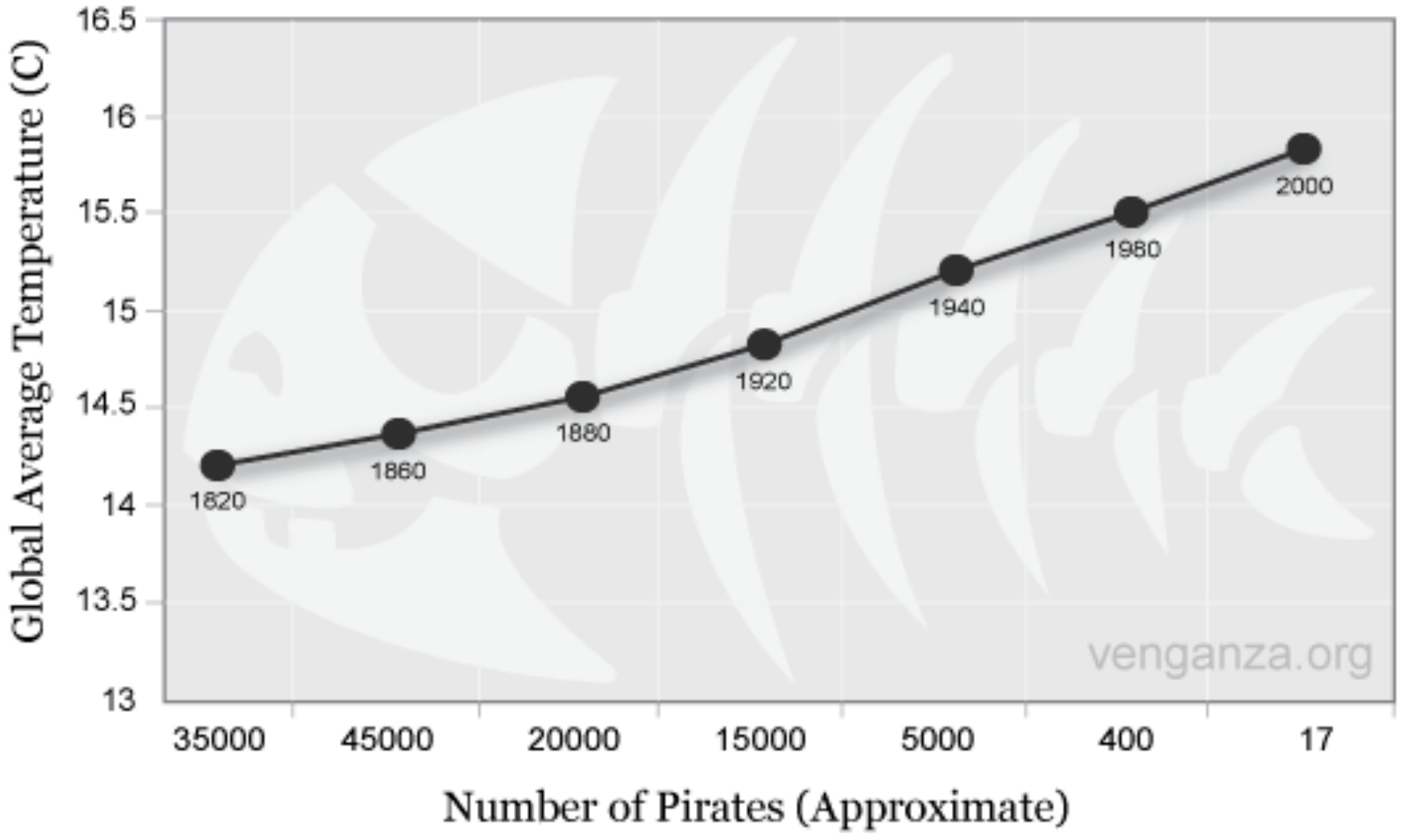
- non-parametric
- tests for monotonic relationship
- defined only for discrete variables
- not sensitive to outliers
- p-values more accurate for smaller samples

*but see: https://arxiv.org/pdf/1011.2009*




Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals
with Disabilities Education Act

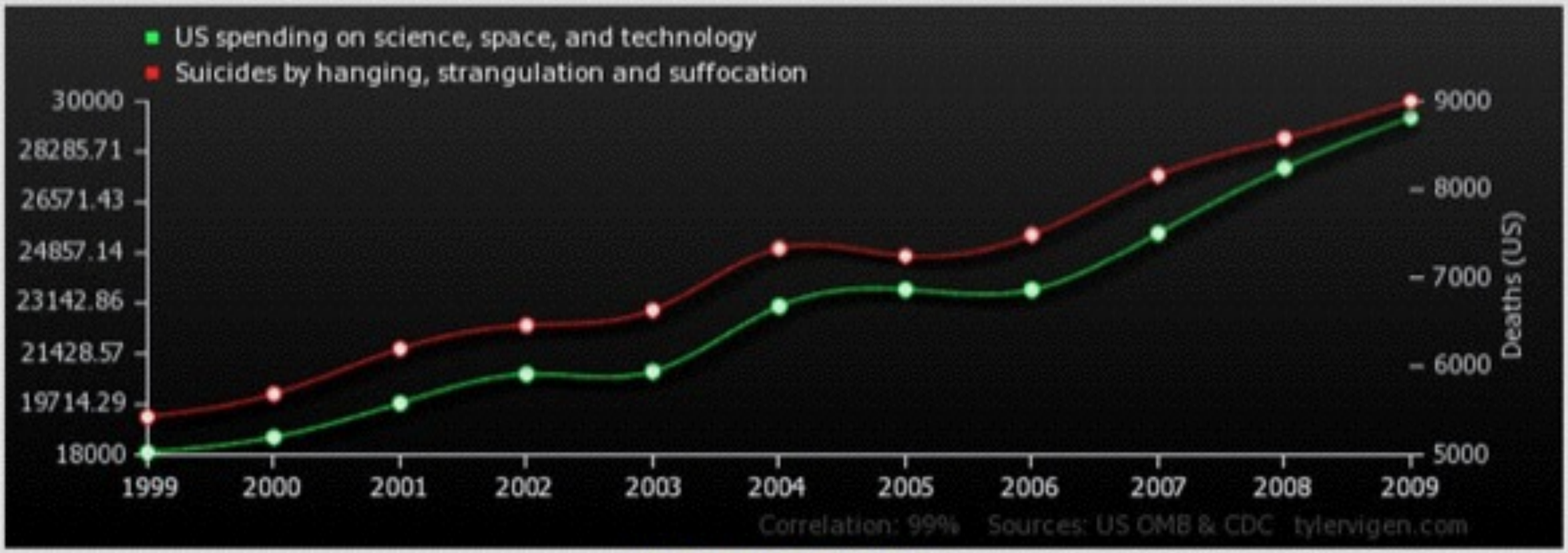
# Global Average Temperature Vs. Number of Pirates



# US spending on science, space, and technology

correlates with

# Suicides by hanging, strangulation and suffocation




Upload this chart to Imgur

MISLEADING STATISTICS

# OVER 2 MILLION AMERICANS EXPOSED TO DRINKING WATER WILL DIE THIS YEAR

Correlation ≠ Causation

CONTENT SHOULD BE USEFUL, NOT JUST PRETTY

*vert.ms/Baddata*