# CSE 519: Data Science

# Steven Skiena
# Stony Brook University


Lecture 5: Correlation

# Correlation Analysis

Two factors are correlated when values of x has some predictive power on the value of y.

The correlation coefficient of X and Y measures the degree to which Y is a function of X (and visa versa).

Correlation ranges from -1 (anti-correlated) to 1 (fully correlated) through 0 (uncorrelated).

# The Pearson Correlation Coefficient

The numerator defines the covariance, which determines the sign but not the scale.

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

A point *(x,y)* makes a positive contribution to *r* when both are above or below their means.

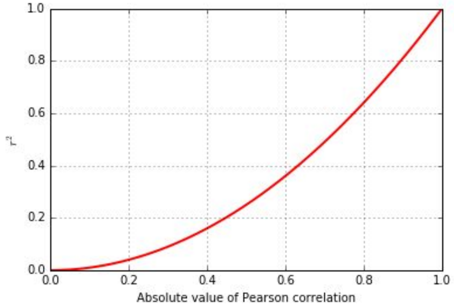
# Representative Pearson Correlations

- SAT scores and freshman GPA (r=0.47)
- SAT scores and economic status (r=0.42)
- Income and coronary disease (r=-0.717)
- Smoking and mortality rate (r=0.716)
- Video games and violent behavior (r=0.19)

# Interpreting Correlations: $r^2$

The square of the sample correlation coefficient $r^2$ estimates the fraction of the variance in *Y* explained by *X* in a simple linear regression.

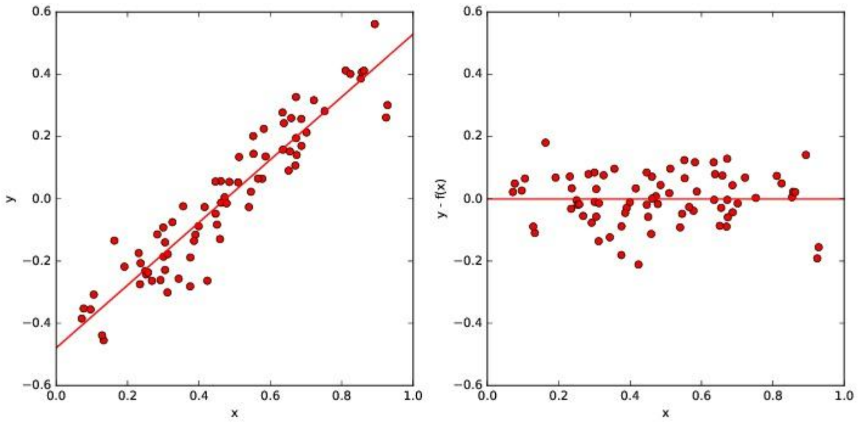
Thus the predictive value of a correlation decreases quadratically with r.
The correlation between height and weight is approximately 0.8, meaning it explains about ⅔ of the variance.



# Variance Reduction and R^2

If there is a good linear fit f(x), then the residuals y-f(x) will have lower variance than y.

Generally speaking,
1-r^2 = V(r)/ V(y)
Here r =0.94, explaining 88.4% of V(y).

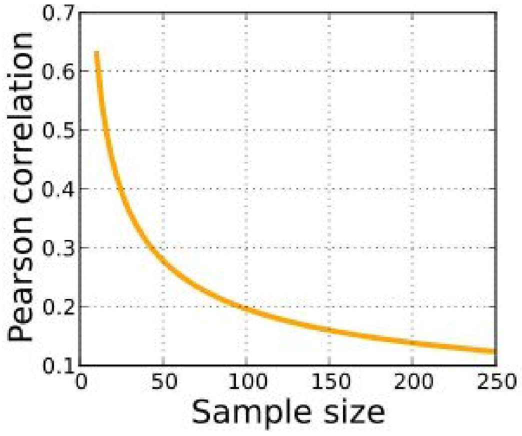


# Interpreting Correlation: Significance

The statistical significance of a correlation depends upon the sample size as well as r.

Even small correlations become significant (at the 0.05 level) with large-enough sample sizes.

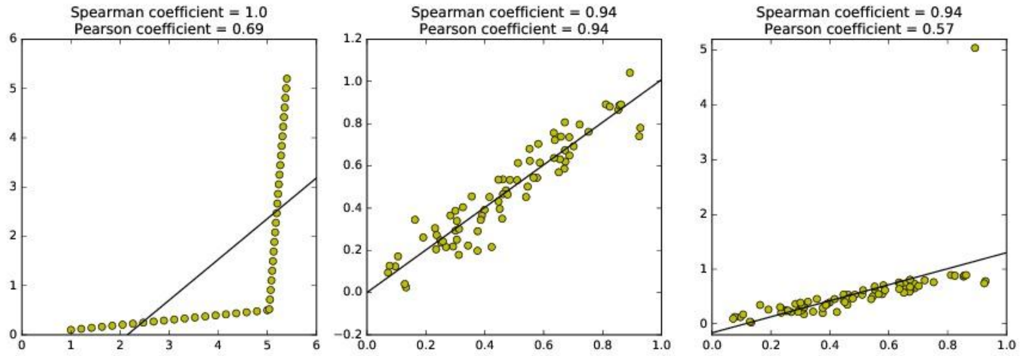
This motivates “big data” multiple parameter models: each single correlation may explain/predict only small effects, but large numbers of weak but *independent* correlations may together have strong predictive power.



# Spearman Rank Correlation

Counts the number of disordered pairs, not how well the data fits a line.

Thus better with non-linear relationships and outliers.



# Computing Spearman Correlation

Let *rank(x_i)* be the rank position of *x_i* in sorted order, from 1 to n. Then:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

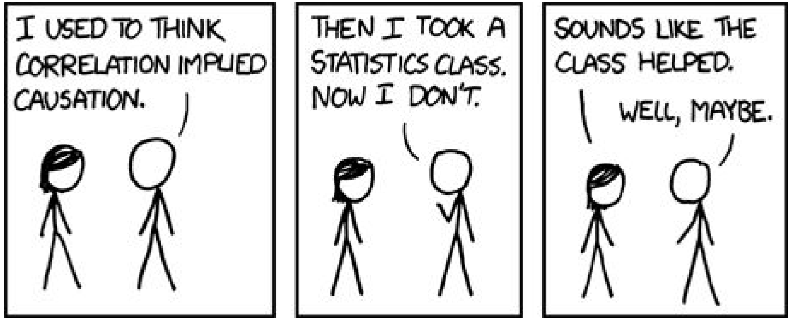
where *d_i = rank(x_i) - rank(y_i).*

It is the Pearson correlation of the X and Y value ranks, so it ranges from -1 to 1.

# Correlation vs. Causation

Correlation does not mean causation.

The number of police active in a precinct correlated strongly with the local crime rate, but the police do not cause the crime.




# Autocorrelation and Periodicity

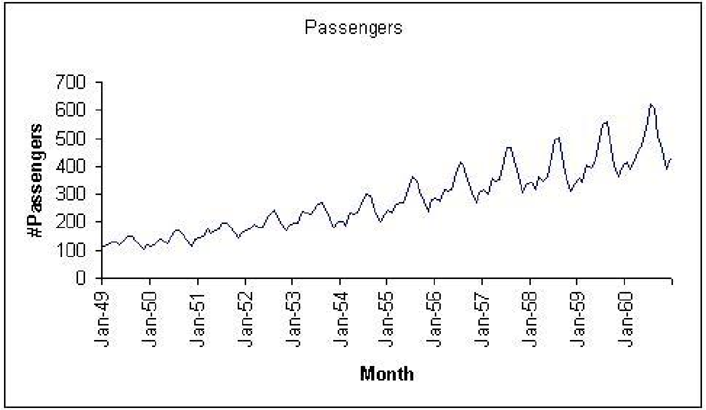
Time-series data often exhibits cycles which affect its interpretation.

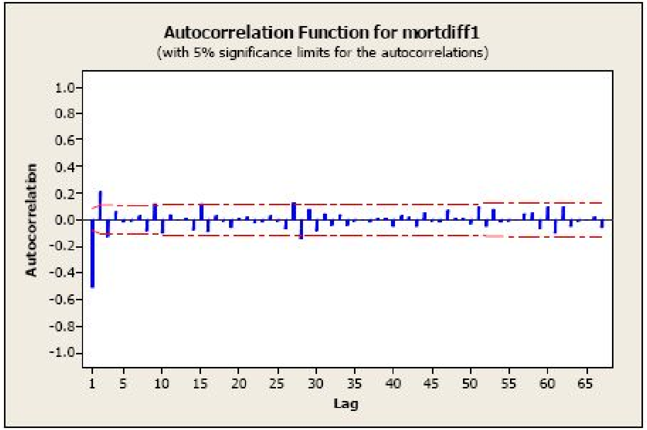
Sales in different businesses may well have 7 day, 30 day, 365 day, and 4*365 day cycles.

A cycle of length $k$ can be identified by unexpectedly large autocorrelation between $S[t]$ and $S[t+k]$ for all $0 < t < n-k$.

# The Autocorrelation Function

Computing the lag-k autocorrelation takes O(n), but the full set can be computed in O(n log n) via the Fast Fourier Transform (FFT).



# Logarithms

The logarithm is the inverse exponential function, i.e. $y = \log_b x \implies b^y = x$

We will use them here for reasons different than in algorithms courses:

Summing logs of probabilities is more numerically stable than multiplying them:

$$\prod_{i=1}^{n} p_i = b^P \text{ where } P = \sum_{i=1}^{n} \log_b(p_i)$$

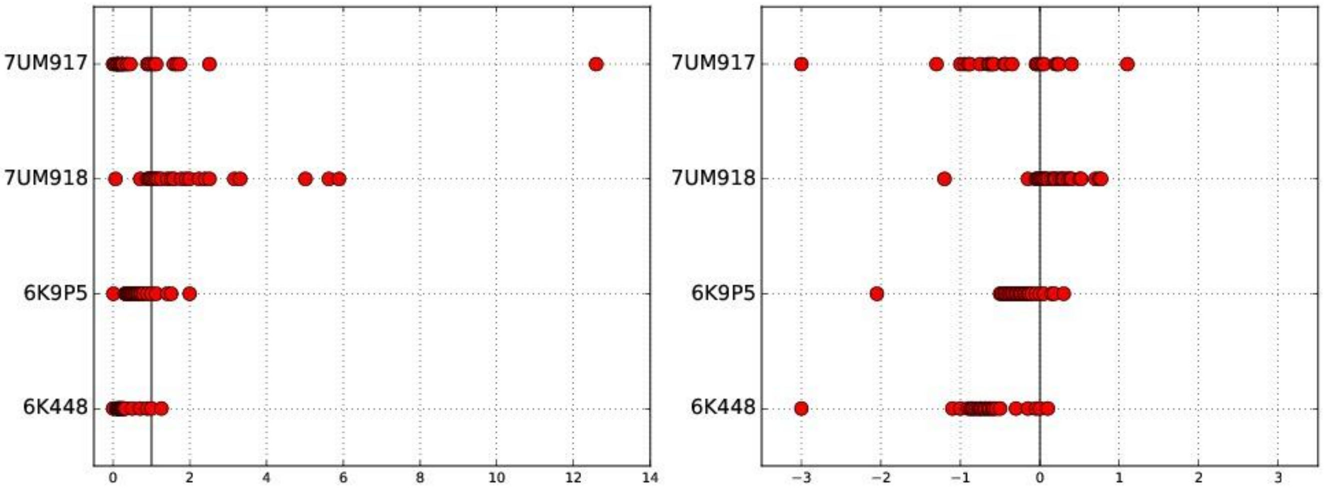
# Logarithms and Ratios

Ratios of two similar quantities (e.g new_price / old_price) behave differently when reflecting increases vs. decreases.

200/100 is 200% above baseline, but 100/200 is 50% below despite being similar changes!

Taking the log of the ratios yield equal displacement: 1.0 and -1.0 (for base-2 logs)

# Always Plot Logarithms of Ratios!



# Logarithms and Power Laws

Taking the logarithm of variables with a power law distribution brings them more in line with traditional distributions.

My wealth is roughly the same number of logs from typical students as I am from Bill Gates!

# Normalizing Skewed Distributions

Taking the logarithm of a value before analysis is useful for power laws and ratios.