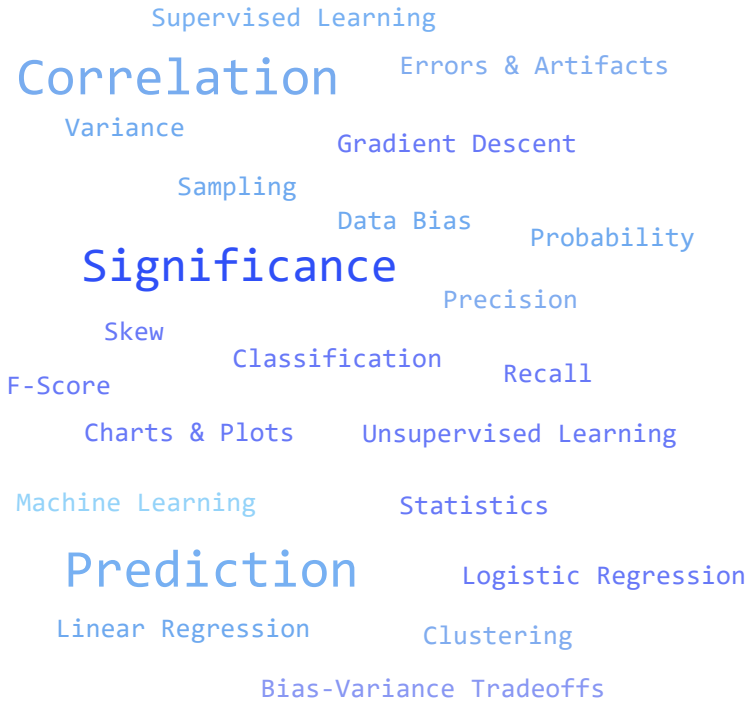


Data Science 1:
Introduction to Data Science

Probability, Statistics & Correlation

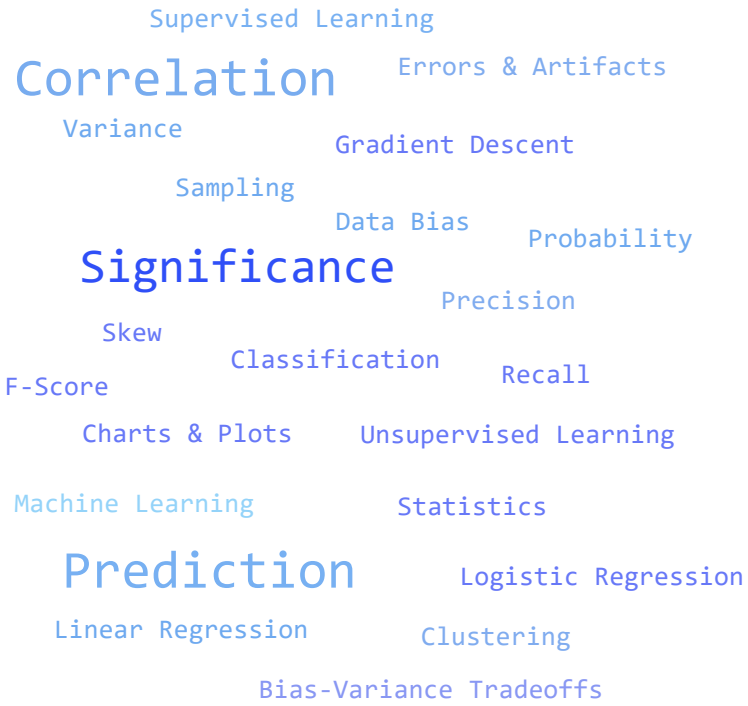
Winter 2025

Wolfram Wingerath, Jannik Schröder

Department for Computing Science
Data Science / Information Systems

Lecture slides based on content from “The Data Science Design Manual” (Steven Skiena, 2017) and associated course materials generously made available online by the author at https://www3.cs.stonybrook.edu/~skiena/data-manual/.

Special thanks to Professor Skiena for sharing these valuable teaching resources!



Data Science 1:
Introduction to Data Science

Probability, Statistics & Correlation

Winter 2025

Wolfram Wingerath, Jannik Schröder

Department for Computing Science
Data Science / Information Systems

Semester Schedule

CW 42	14. Oct	Lecture	1	Orga & Intro	1-26
CW 43	21. / 23. Oct	Lecture + Exercises	2	Probability, Statistics & Correlation	27-56
CW 44	28. Oct	Lecture	3	Data Munging, Cleaning & Bias	57-94 / “Invisible Women”
CW 45	04. / 06. Nov	Lecture + Exercises	4	Scores & Rankings	95-120
CW 46	11. Nov	Lecture	5	Statistical Distributions & Significance	121-154
CW 47	18. / 20. Nov	Lecture + Exercises	6	Building & Evaluating Models	201-236
CW 48	25. Nov	Guest Lecture	7	Data Visualization	155-200
CW 49	02. / 04. Dec	Lecture + Exercises	8	Intro to Machine Learning	351-390
CW 50	09. Dec	Lecture	9	Linear Algebra	237-266
CW 51	16. / 18. Dec	Lecture + Exercises	10	Linear Regression & Gradient Descent	267-288
CW 02	06. Jan	Lecture	11	Logistic Regression & Classification	289-302
CW 03	13. / 15. Jan	Lecture + Exercises	12	Nearest Neighbor Methods & Clustering	303-350
CW 04	20. Jan	Lecture	13	Data Science in the Wild	391-426
CW 05	27. / 29. Jan	Lecture + Exercises	14	Q&A / Feedback	
CW 06	03. / 04. Feb	Oral Exams (Block 1)		Preparation in our last session („Oral Exam Briefing“)	
CW 13	24. / 25. Mar	Oral Exams (Block 2)			

Probability

Probability theory provides a formal framework for reasoning about the likelihood of events.

The **probability** *p(s)* of an **outcome** *s* satisfies:

- *0 <= p(s) <= 1*
- $\sum_{s \in S} p(s) = 1$

These basic properties are often violated in casual use of “probability” in data science.

Probability vs. Statistics

- Probability deals with predicting the likelihood of **future** events, while statistics analyzes the frequency of **past** events.
- Probability is **theoretical** branch of mathematics on the consequences of definitions, while statistics is **applied** mathematics trying to make sense of real-world observations.

Compound Events and Independence

Suppose half my students are female (event A)
Half my students are above median (event B).

What is the probability a student is both A & B?

Events A and B are independent iff

$$P(A \cap B) = P(A) \times P(B)$$

Independence (zero correlation) is good to simplify calculations, but bad for prediction.

Conditional Probability

The conditional probability P(A|B) is defined:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Conditional probabilities get interesting only when events are **not independent**, otherwise:

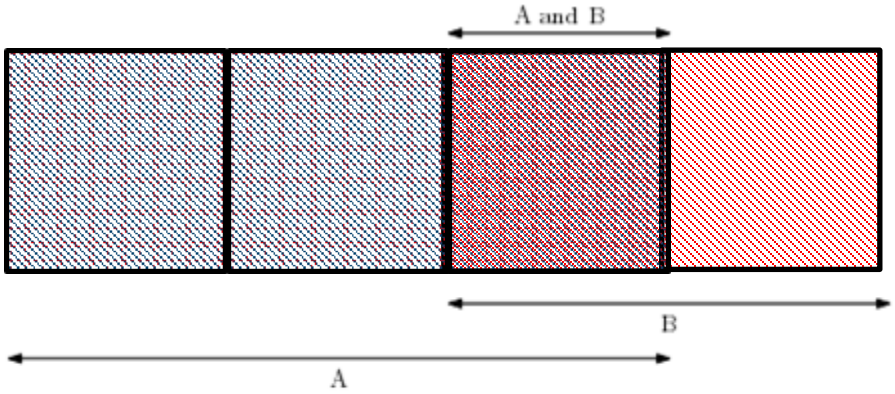
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \times P(B)}{P(B)} = P(A)$$

Bayes Theorem

Bayes theorem is an important tool which reverses the direction of the dependences:

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A)}$$

$$= \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{3}{4}} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{4}{3} = \frac{1}{3}$$



Proof of Bayes Theorem

The probability of two events A and B happening, P(A∩B), is the probability of A, P(A), times the probability of B given that A has occurred, P(B|A).

$$P(A \cap B) = P(A)P(B|A) \tag{1}$$

On the other hand, the probability of A and B is also equal to the probability of B times the probability of A given B.

$$P(A \cap B) = P(B)P(A|B) \tag{2}$$

Equating the two yields:

$$P(B)P(A|B) = P(A)P(B|A) \tag{3}$$

and thus

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)} \tag{4}$$

(q.e.d.) 

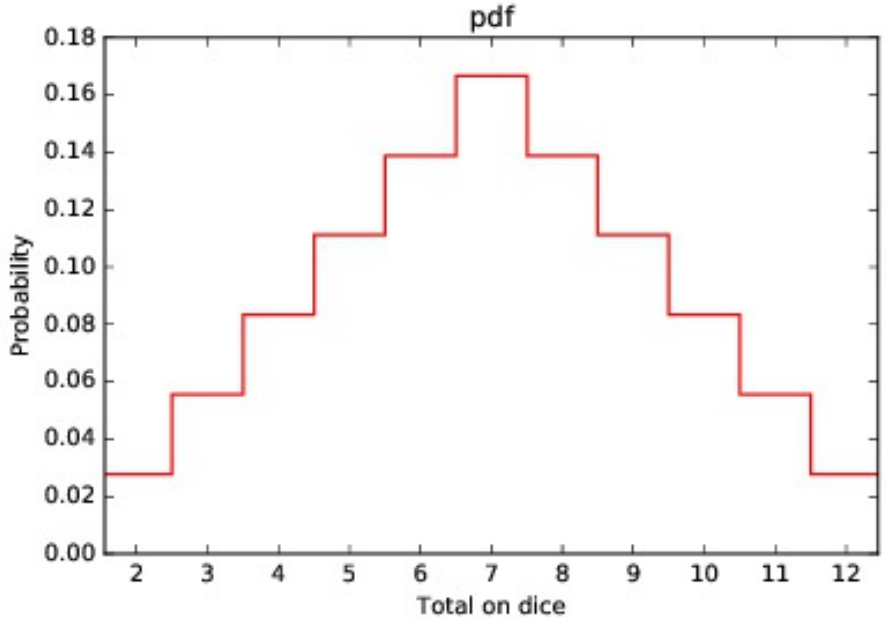
Distributions of Random Variables

Random variables (RVs) are numerical functions where values come with probabilities.

Probability density functions (pdfs) represent RVs, essentially as histograms.

Distributions of Random Variables

Example: the sum of two dice throws.

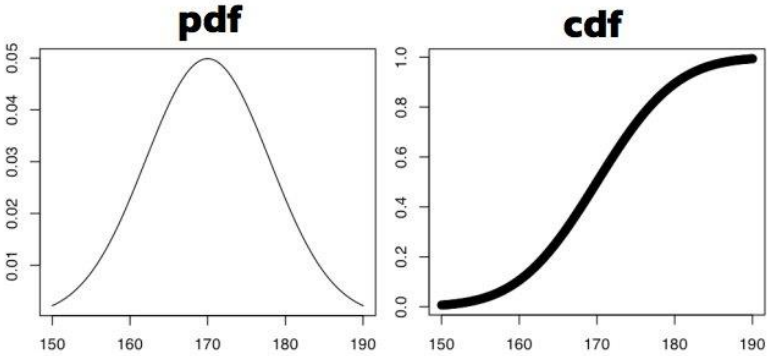


Probability/Cumulative Distributions

The cdf is the running sum of the pdf:

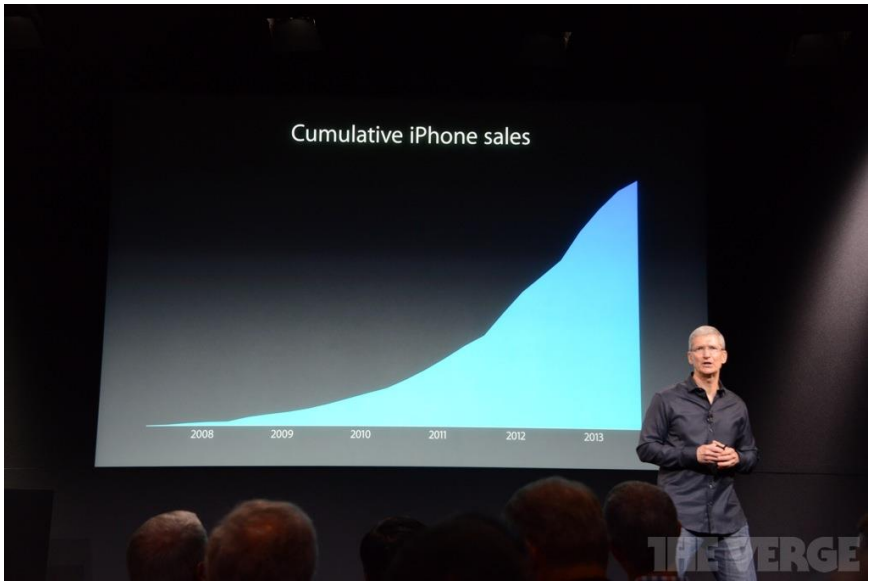
$$C(X \leq k) = \sum_{x \leq k} P(X = x)$$

The pdf and cdf contain exactly the same information, one being the integral / derivative of the other.



Visualizing Cumulative Distributions

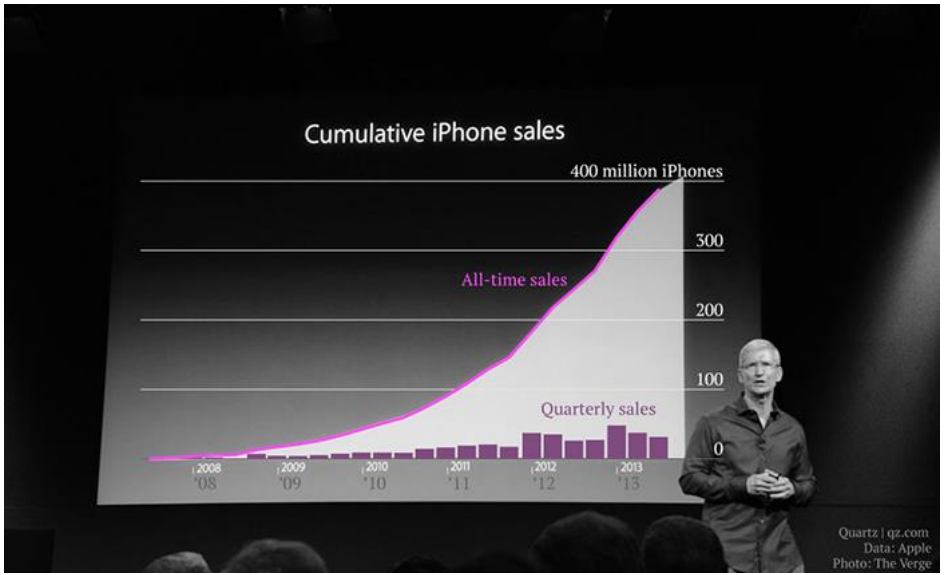
Apple iPhone sales have been exploding, right?



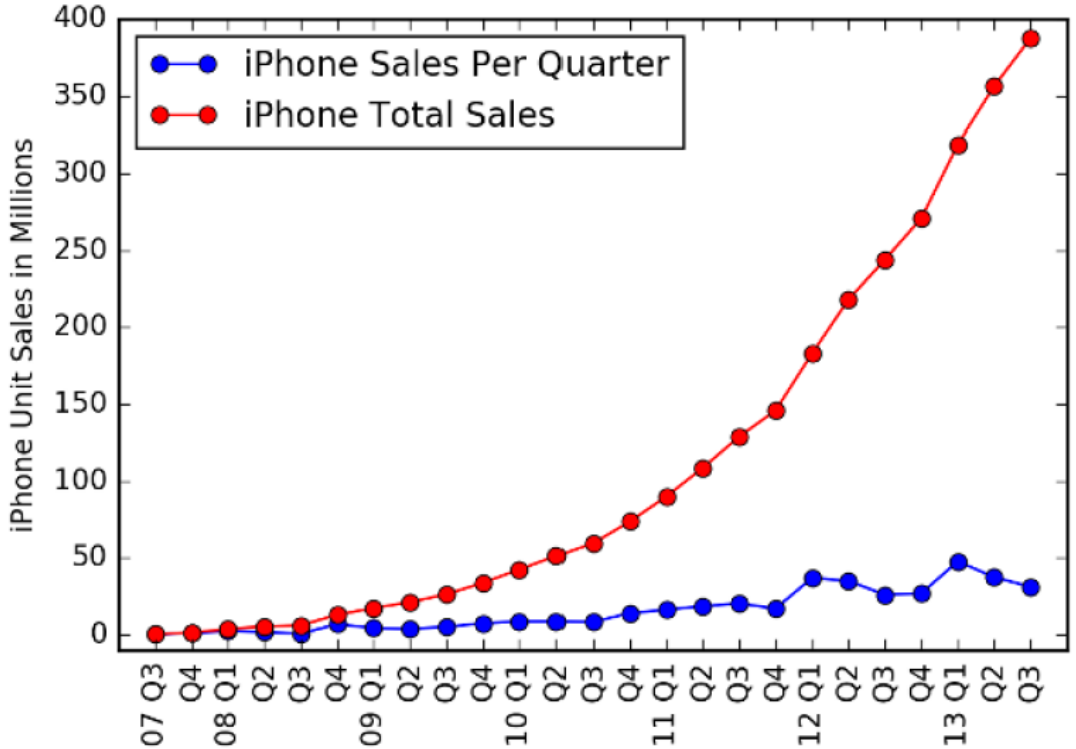
How explosive is that growth, really?

Cumulative distributions present a misleading view of growth rate.

The incremental change is the derivative of this function, which is hard to visualize



How explosive is that growth, really?



Descriptive Statistics

Descriptive statistics provides ways to capture the properties of a given data set / sample.

- Central tendency measures describe the center around the data is distributed.
- Variation or variability measures describe data spread, i.e. how far the measurements lie from the center.

Centrality Measure: Mean

To calculate the mean, sum values and divide by number of observations: $\mu_X = \frac{1}{n}\sum_{i=1}^{n} x_i$

Mean is meaningful for symmetric distributions without outliers.

Other Centrality Measures

The median represents the middle value.

The geometric mean is the nth root of the product of n values: $\left(\prod_{i=1}^{n} a_i\right)^{1/n} = \sqrt[n]{a_1 a_2 \cdots a_n}.$

The geometric mean is always <= arithmetic mean, and more sensitive to values near zero.

Geometric means make sense with ratios:

1/2 and 2/1 *should* average to 1.

Which Measure is Best?

Mean is meaningful for symmetric distributions without outliers: e.g. height and weight.

Median is better for skewed distributions or data with outliers: e.g. wealth and income.

Bill Gates adds $250 to the mean per capita wealth in the US, but nothing to the median.

Aggregation as Data Reduction

Representing a group of elements by a new derived element, like mean, min, count, sum reduces a large dataset to a small **summary statistic**.

Such statistics can become features when taken over natural groups or clusters in the full data set.

Variance Metric: Standard Deviation

The variance is the square of the standard deviation (SD) sigma.

Do we divide by n or n-1?

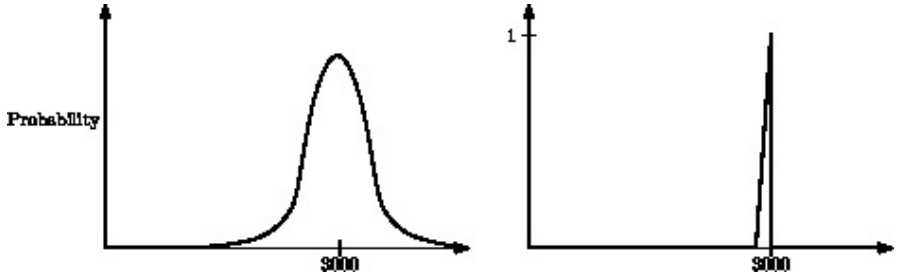
$$\hat{\sigma} = \sqrt{\frac{\sum_{i}^{n} (x_i - \bar{x})^2}{n-1}}$$

The population SD divides by n, the sample SD by n-1, but for large n, n ~ (n-1) so it doesn't really matter.

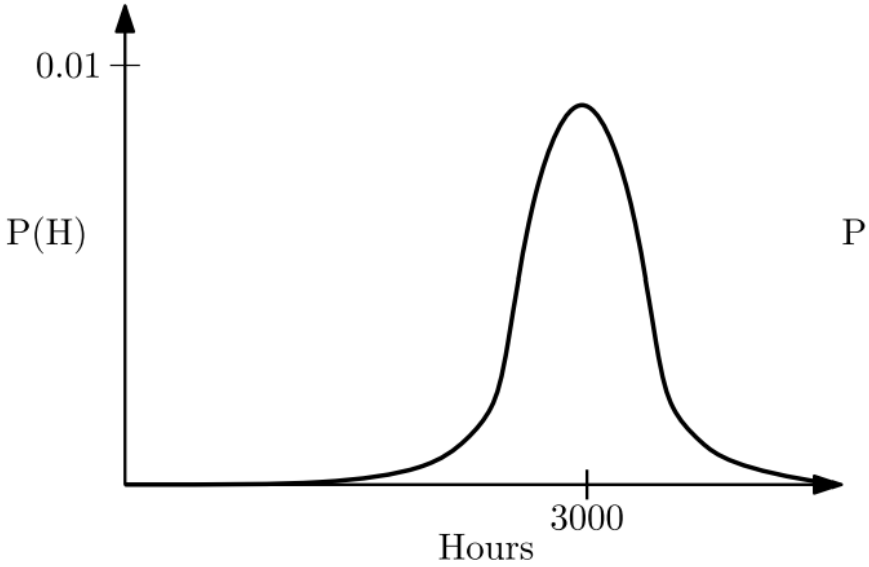
The Light Bulb Life Distribution

Distributions with the same mean can look very different.

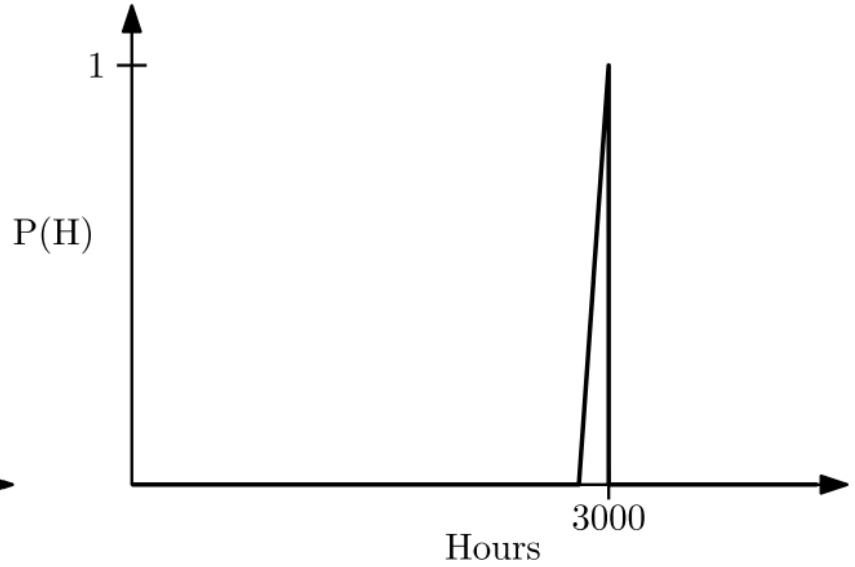
But together, the mean and standard deviation fairly well characterize any distribution.



The Light Bulb Life Distribution



Normal light bulb



Super-reliable light bulb with built-in end-of-warrantee killswitch

Parameterizing Distributions

Regardless of how data is distributed, at least $(1 - \frac{1}{k^2})th$ of the points must lie within *k* sigma of the mean (Chebyshev's inequality).

Thus at least 75% must lie within two sigma of the mean.

Even tighter bounds apply for normal distributions.

Interpreting Variance (Stock Market)

It is hard to measure **“signal to noise” ratio**, because much of what you see is just variance.

Consider measuring the relative “skill” of different stock market investors.

Annual fluctuation in performance among funds is such that investor performance is random, meaning there is little real difference in skill.

Interpreting Variance (Batting Avg)

In baseball, 0.300 hitters (30% success rate) represent consistency over 500 at-bats/season.

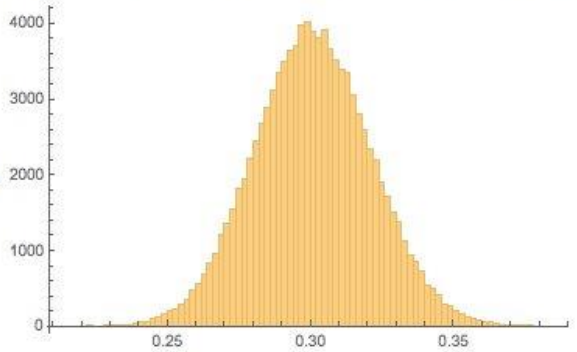
But simulations show a real 0.300 hitter has a 10% chance of hitting 0.275 or below.

They also have a 10% chance of hitting 0.325 or above.

```
In[28]:= Season[p_Real, n_Integer] :=
           Count[Table[If[RandomReal[1] ≤ p, 1, 0], {n}], 1] / (1.0 * n)

In[29]:= Histogram[d = Table[Season[0.300, 500], {100 000}], 100]
```

Out[29]=



Good or bad season, or lucky/unlucky?

→ It's really easy to interpret something as signal that is actually just noise

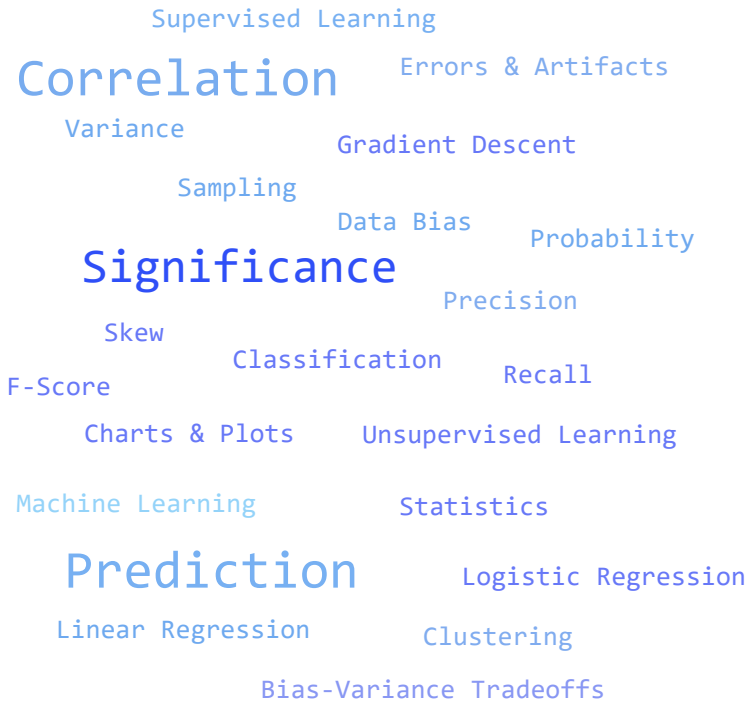
→ This is the kind of problem where wisdom helps (arguably)

Interpreting Variance (Many Models)

We will typically develop several models for each challenge, from very simple to complex.

Some difference in performance will be explained by simple variance: which training/evaluation pairs were selected, how well parameters were optimized, etc.

Small performance wins argue for simpler models.



Data Science 1:
Introduction to Data Science

Probability, Statistics &

Correlation

Winter 2025

Wolfram Wingerath, Jannik Schröder

Department for Computing Science
Data Science / Information Systems

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.

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

The Pearson Correlation Coefficient

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

Covariance

Std. Dev. of X

Std. Dev. of Y

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

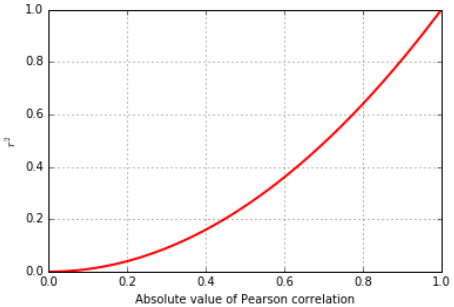
Representative Pearson Correlations

- 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.



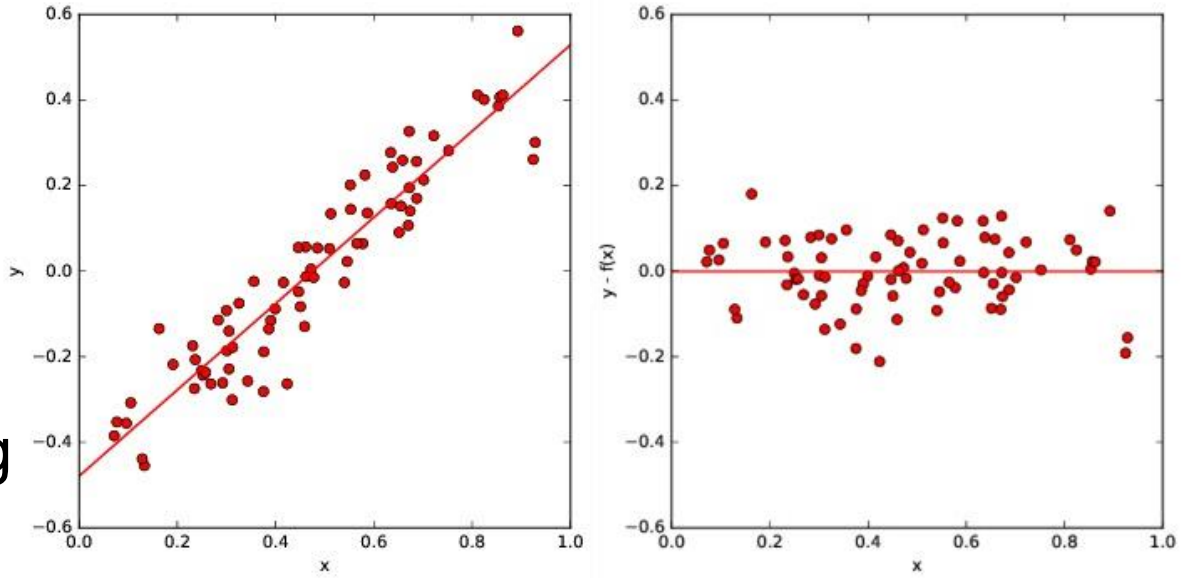
Variance Reduction and r^2

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

Generally speaking,

$$1 - r^2 = \frac{V(residuals)}{V(y)}$$

Here r = 0.94, explaining 88.4% of V(y).

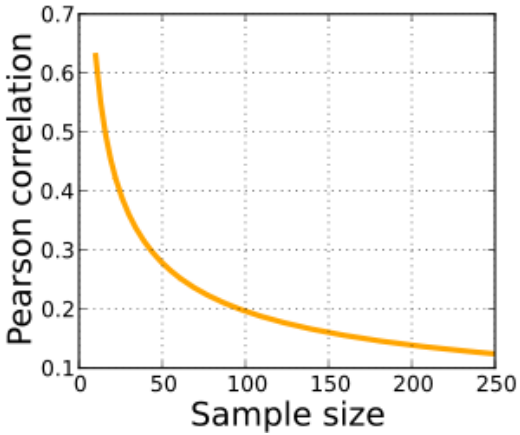


Interpreting Correlation: Significance

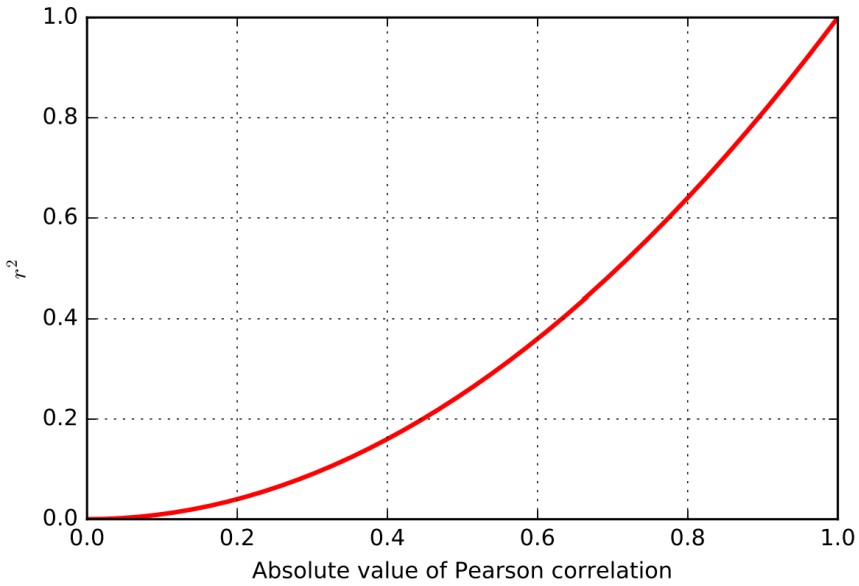
The statistical significance of a correlation depends upon the sample size as well as r.

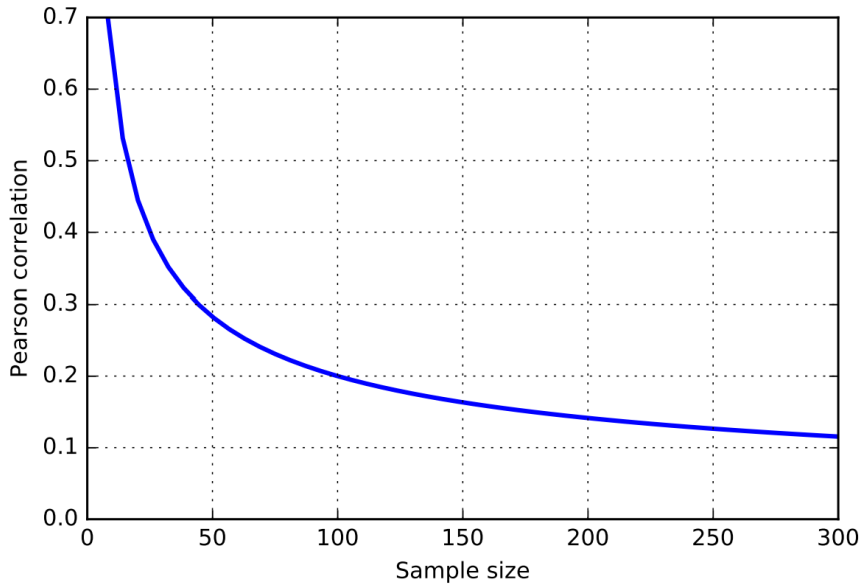
Even small correlations become significant (at the 0.05 level) with large-enough sample sizes.

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.



Interpreting Correlations: r^2





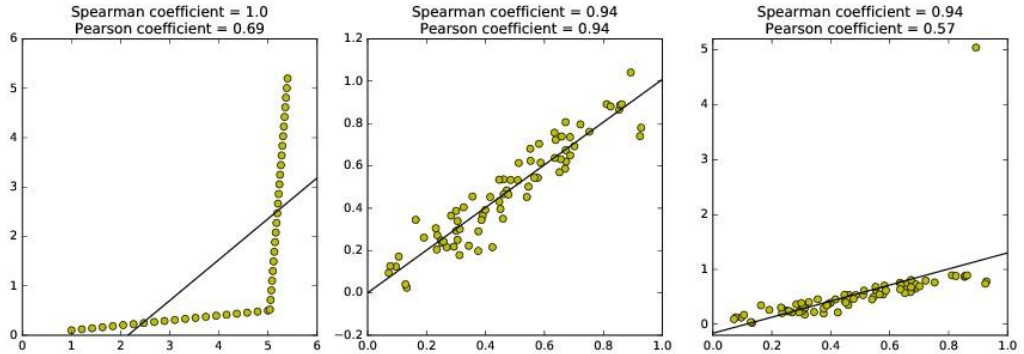
Weak correlations only explain a small fraction of the variance.

With more samples, even weak correlations become significant.

Spearman Rank Correlation

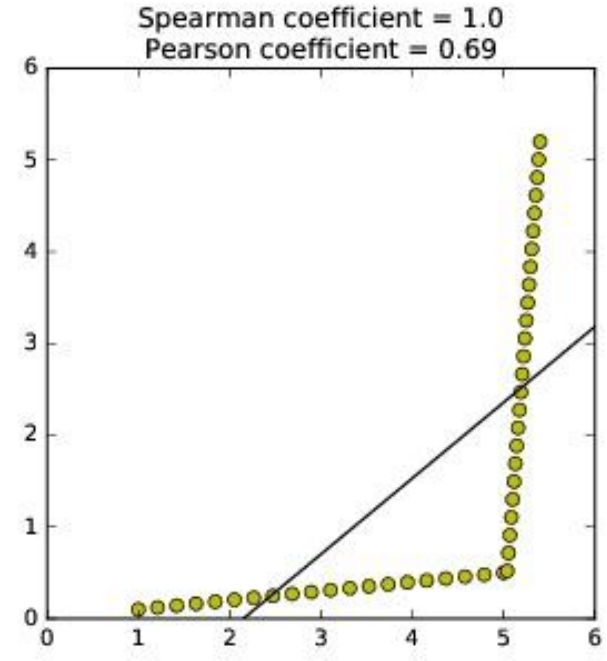
Counts the number of disordered pairs, not how well the data fits a line.

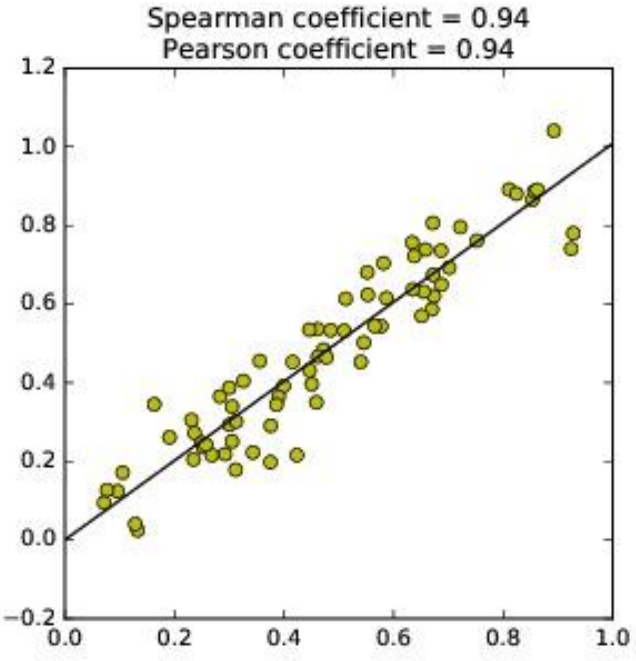
Thus better with non-linear relationships and outliers.

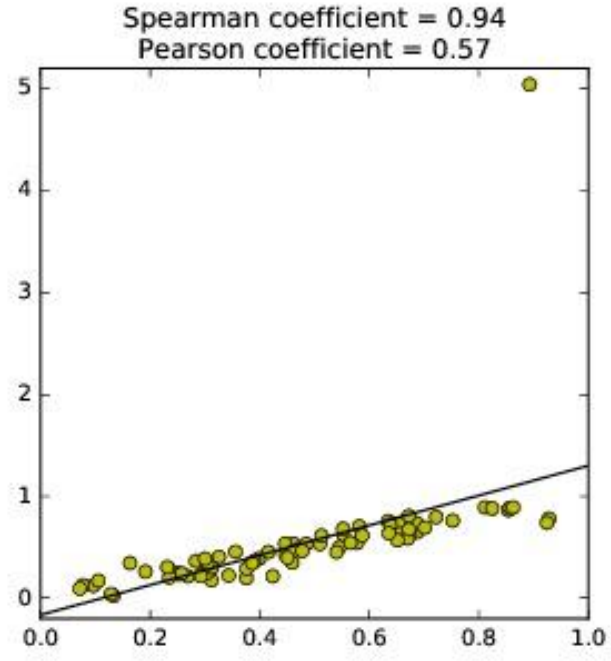


Spearman Rank Correlation

Thus better with non-linear relationships & 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.

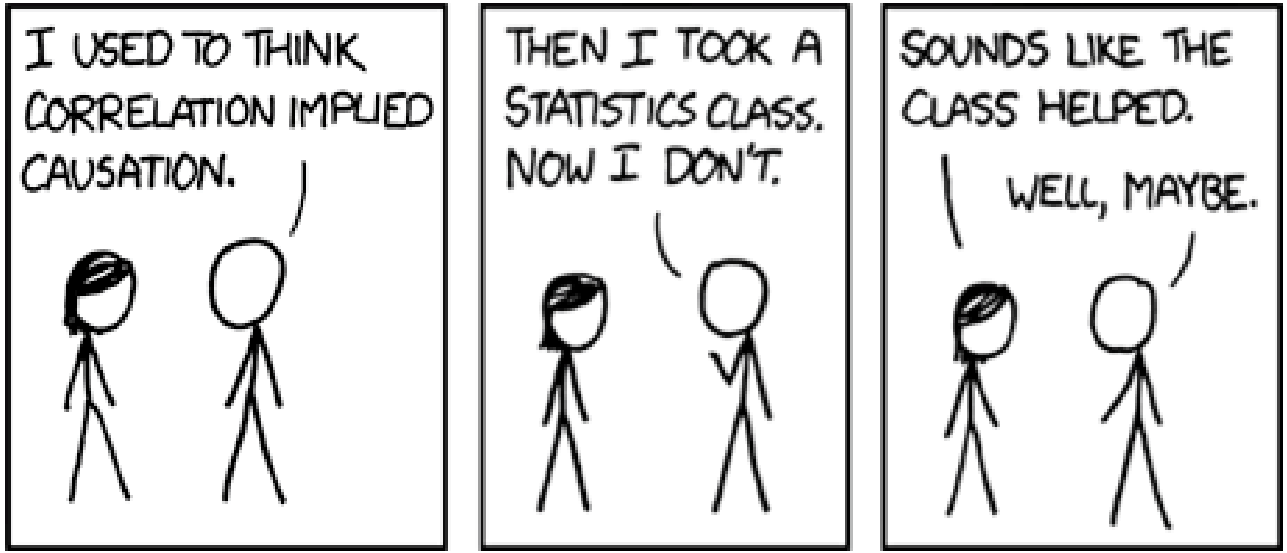
Correlation vs. Causation

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.

The amount of medicine people take is correlated strongly with their probability to get sick, but medicine is typically not causing the sickness.

Correlation vs. Causation



“Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there’.”

XKCD: Correlation

Autocorrelation and Periodicity

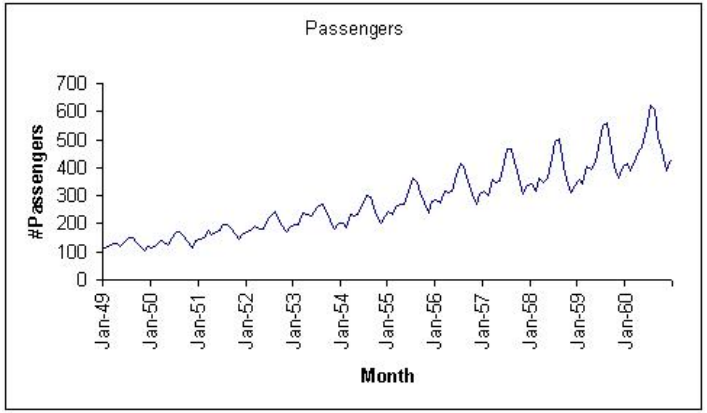
Time-series data often exhibits cycles which affect its interpretation.

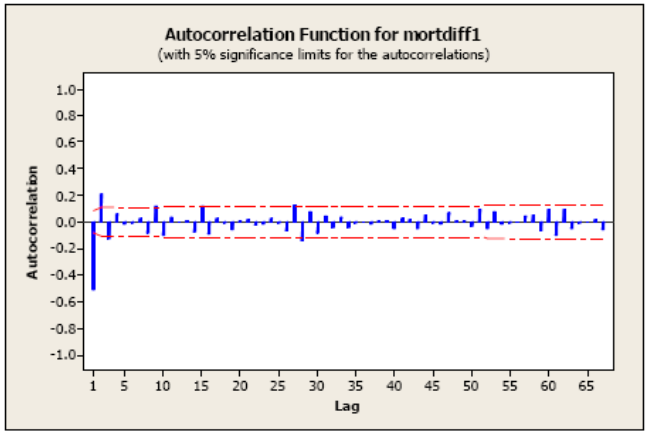
Sales in different businesses may well have 7 day, 30 day, 365 day, and 4*365 day cycles.

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. $y = \log_b x \Rightarrow b^y = x$

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

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

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

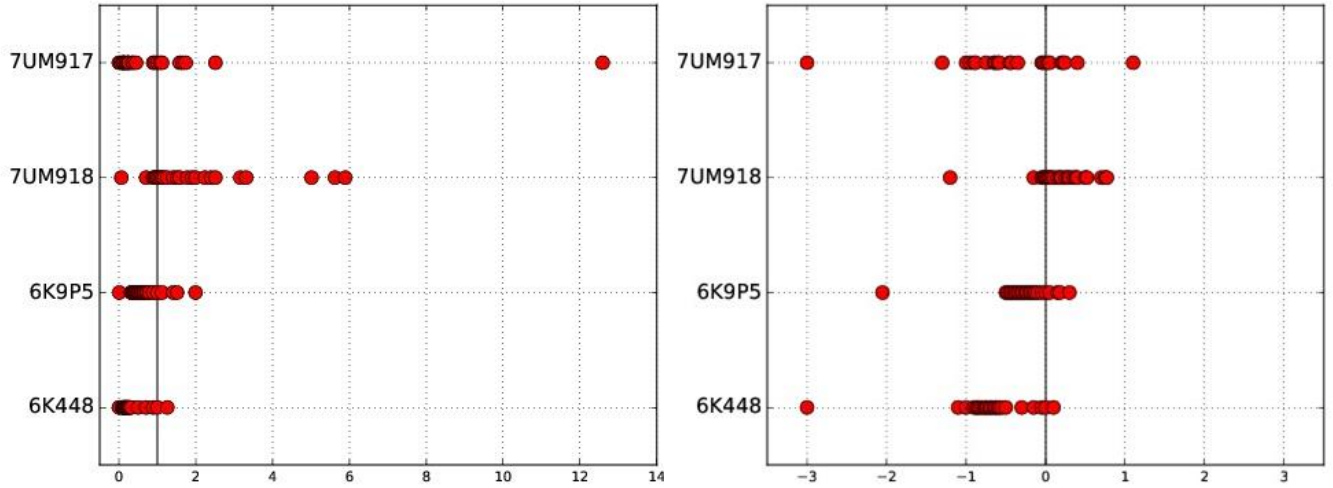
Logarithms and Ratios

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

200/100 is 200% of the baseline, but 100/200 is only 50% despite both 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!



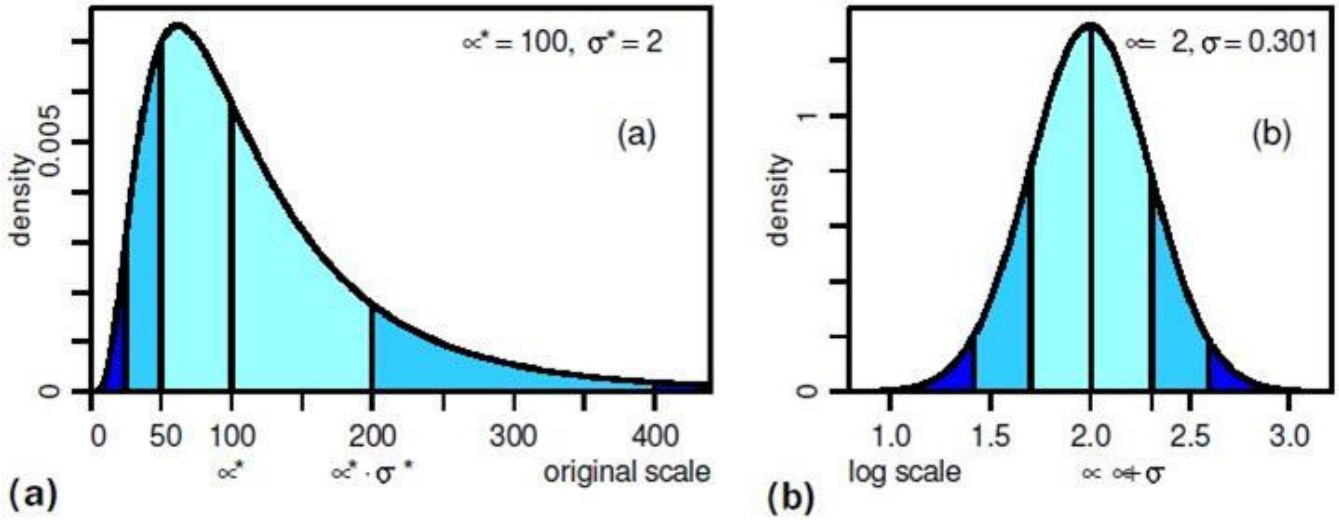
Logarithms and Power Laws

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

Example: John Romero's wealth is reportedly about the same number of logs from typical students' as his from Bill Gates'!

Normalizing Skewed Distributions

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

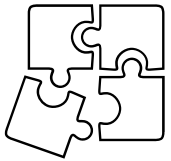


3 Use Cases for Logarithms

1. Higher precision for **probability multiplication**: sum up logarithms, don't multiply probabilities!
2. Representation of increase/decrease of **ratios**: plot ratio logarithms rather than actual ratios!
3. Visualize **distributions with skew or outliers**: put X-axis on a logarithmic scale when you are looking at a power law variable!

Wrapup:

Probability, Statistics & Correlation



- Probability & statistics are fundamental for making predictions and summarizing data
- Correlation & significance can help understand the relationship between variables in data sets
- Logarithms can be used to normalize skewed distributions and to make power law variables easier to interpret