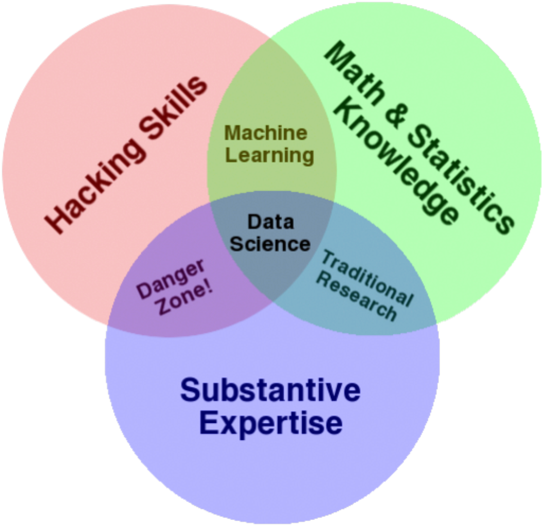


COMM054: Data Science Principles & Practices

UNIVERSITY OF SURREY

Introduction to Probability & Distributions

Dr. Manal Helal

31 October 2019

Outline

- Probability Definitions
- Descriptive Statistics Important Definitions
- Correlation Analysis
- Logarithms & Ratios

Probability Important Definitions

- An **experiment** is a procedure which yields one of a set of possible outcomes. As our ongoing example, consider the experiment of tossing two six-sided dice, one red and one blue, with each face baring a distinct integer {1,...,6}.

- A **sample space S** is the set of possible outcomes of an experiment. In our dice example, there are 36 possible outcomes, namely
  - S = {(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),(2,3),(2,4),(2,5),(2,6), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)}.

Probability Important Definitions- Cont'd

- An **event E** is a specified subset of the outcomes of an experiment. The event that the sum of the dice equals 7 or 11 is the subset
  - E = {(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1), (5, 6), (6, 5)}.

- The **probability of an outcome s**, denoted p(s) is a number with the two properties:
  - For each outcome s in sample space S, 0≤p(s)≤1.
  - The sum of probabilities of all outcomes adds to one: $\sum_{s \in S} \mathrm{p}(s) = 1$.

  - If we assume two distinct fair dice, the probability p(s) = (1/6) × (1/6) = 1/36 for all outcomes s ∈ S.

Probability Important Definitions - Cont’d

- The **probability of an event E** is the sum of the probabilities of the outcomes of the experiment that belong to E. Thus p(E) =$\sum_{s \in \mathrm{E}} \mathrm{p}(\mathrm{s})$
  - An alternate formulation is in terms of the complement of the event $\bar{E}$, the case when E does not occur. Then P (E) = 1 − P ($\bar{E}$) .
  - This is useful, because often it is easier to analyse P($\bar{E}$) than P(E) directly.
- A **random variable V** is a numerical function on the outcomes of a probability space.

Example 1:

- The function “sum the values of two dice” (V ((a, b)) = a + b) produces an integer result between 2 and 12. This implies a probability distribution of the values of the random variable. The probability P (V (s) = 7) = 1/6, as previously shown, while P (V (s) = 12) = 1/36.

- V (s) = 7 → E = {(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)}.
- P (V (s) = 7) = 6/36 = 1/6

- V (s) = 12 → E = {(6, 6)}.
- P (V (s) = 7) = 1/6

Probability Important Definitions - Cont'd

- The **expected value** of a random variable V defined on a sample space S, E(V) is defined as:
  - E(V) = $\sum_{s \in \mathrm{S}} \mathrm{p}(s). \mathrm{V}(s)$

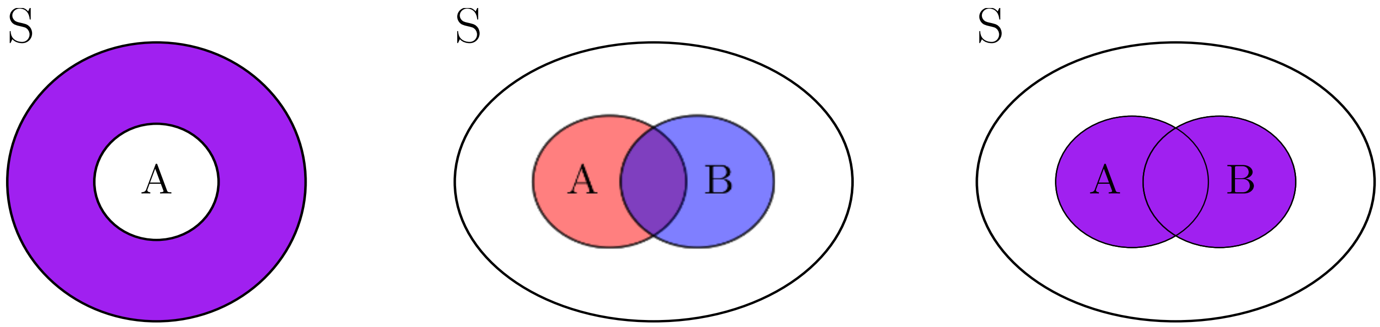
- **The complex events** are computed from simpler events A and B on the same set of outcomes.

Example 2:

- event A is that at least one of two dice be an even number, while event B denotes rolling a total of either 7 or 11. Note that there exist certain outcomes of A which are not outcomes of B, specifically:
  - B = {(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1), (5, 6), (6, 5)}.
  - A − B = {(1, 2), (1, 4), (2, 1), (2, 2), (2, 3), (2, 4), (2, 6), (3, 2), (3, 6), (4, 1), (4, 2), (4, 4), (4, 5), (4, 6), (5, 4), (6, 2), (6, 3), (6, 4), (6, 6)}.
- This is the set **difference** operation. Observe that here B − A = {}, because every pair adding to 7 or 11 must contain one odd and one even number.

Probability Important Definitions - Cont’d

- The common outcomes in both events A and B are called the intersection, denoted A ∩ B. This can be written as
  - A ∩ B = A − (S − B).
- Outcomes which appear in either A or B are called the union, denoted A ∪ B. With the complement operation A̅ = S−A, we get a rich language for combining events, shown in the Figure.



$\overline{A}$ $A \cap B$ $A \cup B$

Independent and dependent Events

- The events A and B are **independent** if and only if P (A ∩ B) = P (A) × P (B).
  - This means that there is no special structure of outcomes shared between events A and B. Assuming that half of the students in my class are female, and half the students in my class are above average, we would expect that a quarter of my students are both female and above average if the events are independent.
- Events A and B are **dependent** if there is a correlation that makes the occurrence of one lead to the other.

Example 3: Conditional Probability

- Suppose a person always uses an umbrella if and only if it is raining. Assume that the probability it is raining (event B) is, say, p = 1/5. This implies the probability that this person is carrying an umbrella (event A) is q = 1/5. But even more, if you know the state of the rain you know exactly whether this person is carrying an umbrella. These two events are perfectly correlated.
  - $P(A|B) = \frac{\mathrm{P\ (A \cap B)}}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$

- The **conditional probability** of A given B, P(A|B) is defined:
  - $P(A|B) = \frac{\mathrm{P\ (A \cap B)}}{P(B)}$

Conditional Probability

- From **Example 2**, Event A is that at least one of two dice be an even number. Event B is the sum of the two dice is either a 7 or an 11.
- Observe that P(A|B) = 1, because any roll summing to an odd value must consist of one even and one odd number. Thus A ∩ B = B, analogous to the umbrella case above. For P (B|A), note that P (A ∩ B) = 9/36 and P (A) = 25/36, so P (B|A) = 9/25.
- $P(B|A) = \frac{\mathrm{P\ (B\ \cap A)}}{P(A)} = \frac{9/36}{25/36} = \frac{9}{25}$

Bayes theorem

- **Bayes theorem** reverses the direction of the **dependencies**:

Likelihood: Probability of collecting this data when our hypothesis is true

Prior: The Probability of hypothesis being true before collecting the data

Posterior: The Probability of hypothesis being true given the data collected

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

Marginal: What is the probability of collecting this data under all possible hypothesis

For **Example 2** by Bayes theorem:

P(B|A) = (1·9/36)/(25/36) = 9/25, exactly what we got before.

https://gerardnico.com/data_mining/bayes

Random Variables

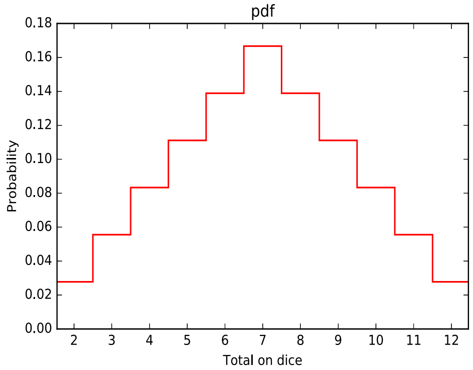
- **Random variables** are numerical functions where the values are associated with **probabilities of occurrence**.

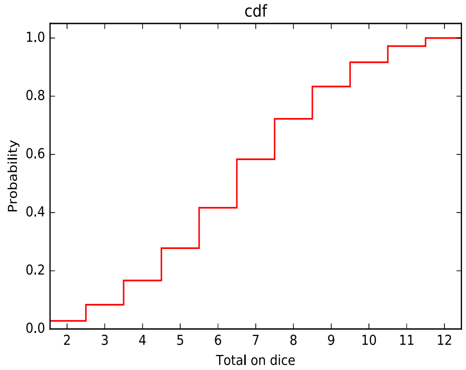
- From **Example 1**, the distribution of V(S) as the sum of two tossed dice, the function produces an integer between 2 and 12. The probability of a particular value V (s) = X is the sum of the probabilities of all the outcomes which add up to X.

Probability Density Function & Cumulative DF

- The probability density function, or pdf of a random variable, is a graph where the x-axis represents the range of values the random variable can take on, and the y-axis denotes the probability of that given value.
- The pdf of V(s) in example 1, is illustrated in the Left Figure. Observe that the peak at X = 7 corresponds to the most frequent dice total, with a probability of 1/6, 2 and 12 are the least frequent dice total with a probability of 1/36.

- **The cumulative density function or cdf** is the running sum of the probabilities in the pdf; as a function of k, it reflects the probability that X ≤ k instead of the probability that X = k. Right Figure shows the cdf of the dice sum distribution.



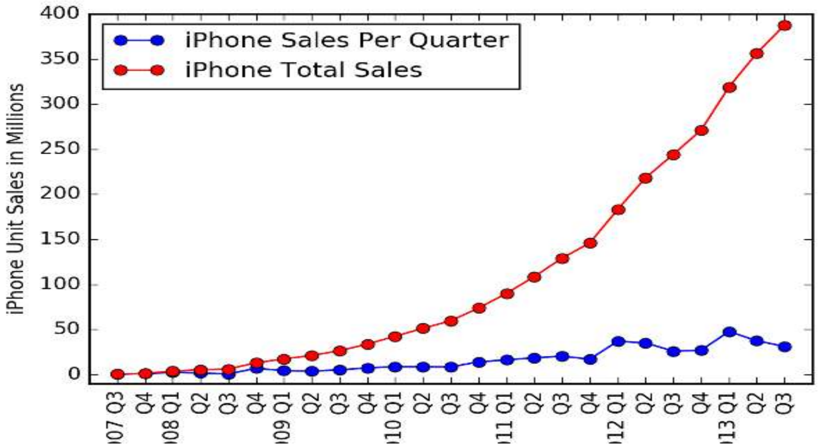


Python Notebook Exercises

- **Section 1** imports the required libraries.
- **Section 2** shows how the pdf is created from the histogram function. This is because pdf plots have a strong relationship to histograms of data frequency, where the x-axis again represents the range of value, but y now represents the observed frequency of exactly how many event occurrences were seen for each given value X. Converting a histogram to a pdf can be done by dividing each bucket by the total frequency over all buckets. The sum of the entries then becomes 1, so we get a probability distribution.
- **Section 3** in the code illustrated how to generate cdf plots, by using the attribute “cumulative=True” in the histogram function.

Python Notebook Exercises – Cont’d

- **Section 4** in the code illustrates the iphone sales data in cdf. The cumulative is continuously increasing until it reaches 1 (total probability). The cumulative distribution of the iphone sales (red) shows that sales are exploding, right? But it presents a misleading view of growth rate, because incremental change is the derivative of this function, and hard to visualize. Indeed, the sales-per-quarter plot (blue) shows that the rate of iPhone sales actually had declined for the last two periods before the presentation.



Descriptive Statistics Important Definitions

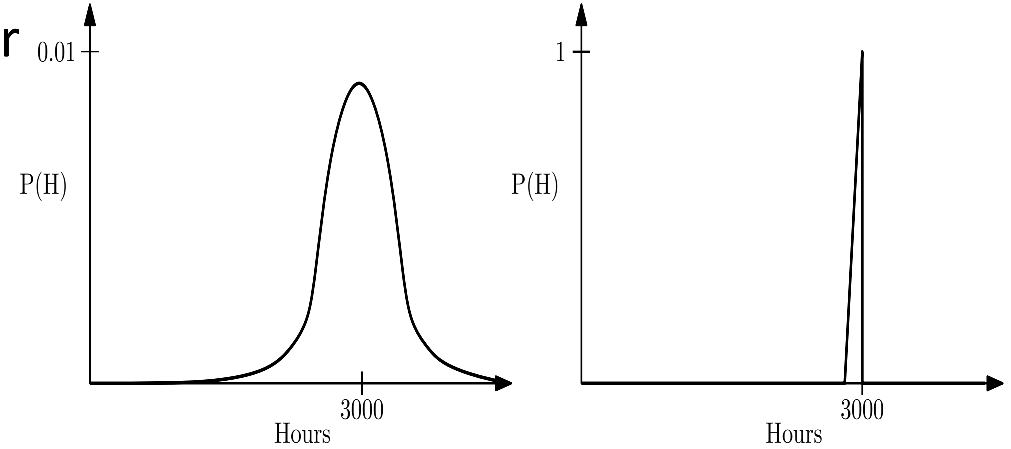
- **Central tendency measures**, such as:
  - **arithmetic mean**: $\mu x = \frac{1}{n}\sum_{i=1}^{n} x_i$, in Example 1, it is = 7
  - **geometric mean**: $(\prod_{i=1}^{n} a_i)^{\frac{1}{n}} = \sqrt[n]{a_1 a_2 \dots a_n}$, in Example 1, it is = 6.5201
  - **median**: the exact middle value, in Example 1, it is = 7
  - **mode**: the most frequent element in the data set, in Example 1, it is = 7, because it occurs 6 times out of 36.

Variation or Variability Measures

- Measures how data is spread away from the mean, such as:
    - **standard deviation**: sum of square differences between the elements and the mean: σ $= \sqrt{\frac{\sum_{i=1}^{n}(a_i - a)^2}{n-1}}$. The squaring makes one outlier value d units from the mean contributes as much to the variance as d^2 points each one unit from the mean.
    - **variance** is 8 characters shorter than standard deviation, making it very sensitive to outliers, V = σ^2,

Example 4

- A light bulb comes with an expected working life, say μ = 3000 hours, derived from some underlying distribution shown in the Figure.
- In a conventional bulb, the chance of it lasting longer than μ is presumably about the same as that of it burning out quicker, and this degree of uncertainty is measured by σ.
- An evil manufacturer builds very robust bulbs, but includes a counter to prevent it from ever glowing after 3000 hours of use. Here μ = 3000 and σ = 0. Both distributions have the same mean, but substantially different variance.



Correlation Analysis Important Definitions

- Given two variables x and y, represented by a sample of n points of the form (x_i,y_i), for 1 ≤ i ≤ n, the **correlation coefficient r(X,Y)** is a statistic that measures the degree to which Y is a function of X, and vice versa.
- x and y are correlated when the value of x has some predictive power on the value of y.
- The value of the correlation coefficient ranges from −1 to 1, where 1 means fully correlated and 0 implies no relation, or independent variables. Negative correlations imply that the variables are anti-correlated, meaning that when X goes up, Y goes down.

- **Covariance** is measures as: $Cov(X,Y) = \sum_{i=1}^{n}(X_i - \bar{X})(Y - \bar{Y})$.

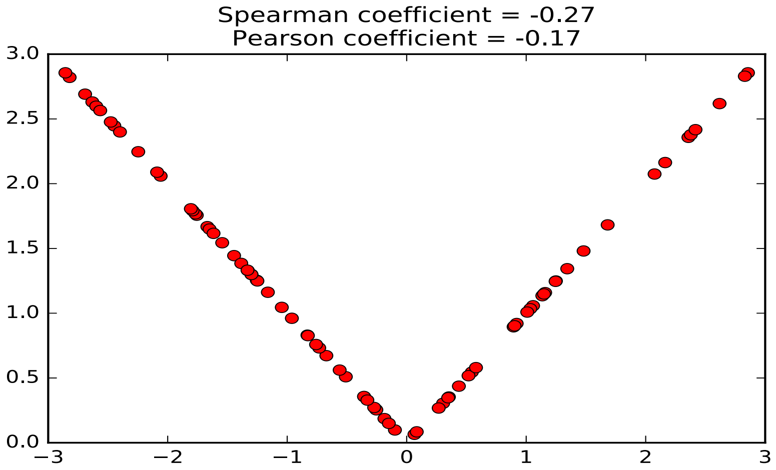
- $\bar{X}\ is$ the mean of X , and $\bar{Y}$ is the mean of Y.

Pearson Correlation Coefficient

- The **Pearson Correlation Coefficient**: $\frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$.
- It defines the degree to which a linear predictor of the form y = m·x+b can fit the observed data.
- The denominator reflects the amount of variance in the two variables, as measured by their standard deviations. The covariance between X and Y potentially increases with the variance of these variables, and this denominator is the magic amount to divide it by to bring correlation to a −1 to 1 scale.

The linearity of Pearson Correlation Coefficient

- Consider points of the form (x, |x|), where x is uniformly (or symmetrically) sampled from the interval [−1, 1] as shown in Figure 5. The correlation will be zero because for every point (x, x) there will be an offsetting point (−x, x), yet y = |x| is a perfect predictor. Pearson correlation measures how well the best linear predictors can work but says nothing about weirder functions like absolute value. This is illustrated in The Figure.



Spearman Rank Correlation Coefficient

- The **Spearman rank correlation coefficient** essentially counts the number of pairs of input points which are out of order. Suppose that our data set contains points (x_1,y_1) and (x_2,y_2) where $x_1 < x_2$ and $y_1 < y_2$. This is a vote that the values are positively correlated, whereas the vote would be for a negative correlation if $y_2 < y_1$.
- Summing up over all pairs of points and normalizing properly gives us Spearman rank correlation.

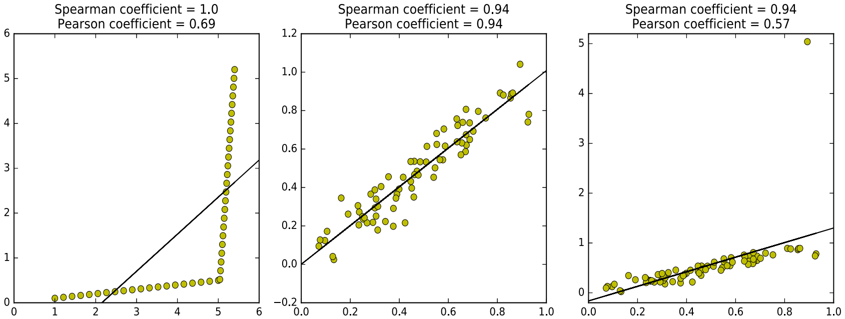
- Let **rank(x_i)** be the rank position of x_i in sorted order among all x_i, so the rank of the smallest value is 1 and the largest value n.

- **Spearman rank correlation coefficient** $\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$, where d_i = rank(x_i) − rank(y_i).

Pearson vs. Spearman

- The relationship between the two coefficients is illustrated in Figure 6. In addition to giving high scores to non-linear but monotonic functions, Spearman correlation is less sensitive to extreme outlier elements than Pearson. Let p = (x_1, y_{max}) be the data point with largest value of y in a given data set. Suppose we replace p with p′ = (x_1, ∞). The Pearson correlation will go crazy, since the best fit now becomes the vertical line x = x_1. But the Spearman correlation will be unchanged, since all the points were under p, just as they are now under p′.

- Section 5 in the accompanying code illustrates the code to calculate both Pearson and spearman coefficients for a dataset that is randomly selected with a uniform distribution, showing why Pearson fails for absolute function as shown in the Figure.



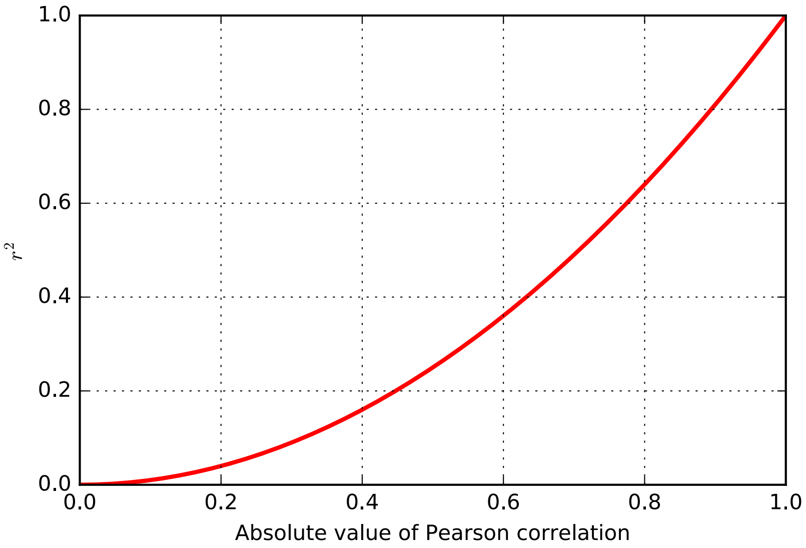
Spearman coefficient = 1.0
Pearson coefficient = 0.69

Spearman coefficient = 0.94
Pearson coefficient = 0.94

Spearman coefficient = 0.94
Pearson coefficient = 0.57

Strength of correlation: R^2

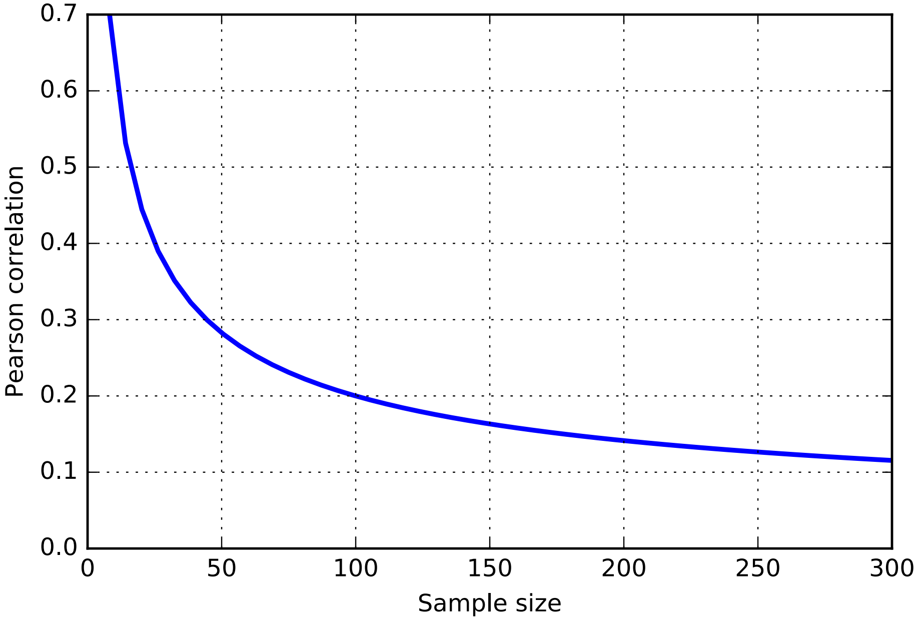
- The square of the sample correlation coefficient r^2 estimates the fraction of the variance in Y explained by X in a simple linear regression. For Example: The correlation between height and weight is approximately 0.8, meaning it explains about two thirds of the variance.
- The left Figure shows how rapidly r^2 decreases with r. There is a profound limit to how excited we should get about establishing a weak correlation. A correlation of 0.5 possesses only 25% of the maximum predictive power, and a correlation of r = 0.1 only 1%. Thus,
  the predictive value of correlations
  decreases rapidly with r.



- This is implemented in section 7 of the accompanying code. It generated a linearly spaced r values and plot them against their squared values. Left Image explains sample size effects on Correlation significance.

Statistical Significance

- The statistical significance of a correlation depends upon its sample size n as well as r. By tradition, we say that a correlation of n points is significant if there is an α ≤ 1/20 = 0.05 chance that we would observe a correlation as strong as r in any random set of n points.
- This is not a particularly strong standard. Even small correlations become significant at the 0.05 level with large enough sample sizes, as shown in the Figure. A correlation of r = 0.1 becomes significant at α = 0.05 around n = 300, even though such a factor explains only 1% of the variance.
- This is implemented in Section 8 in the code.



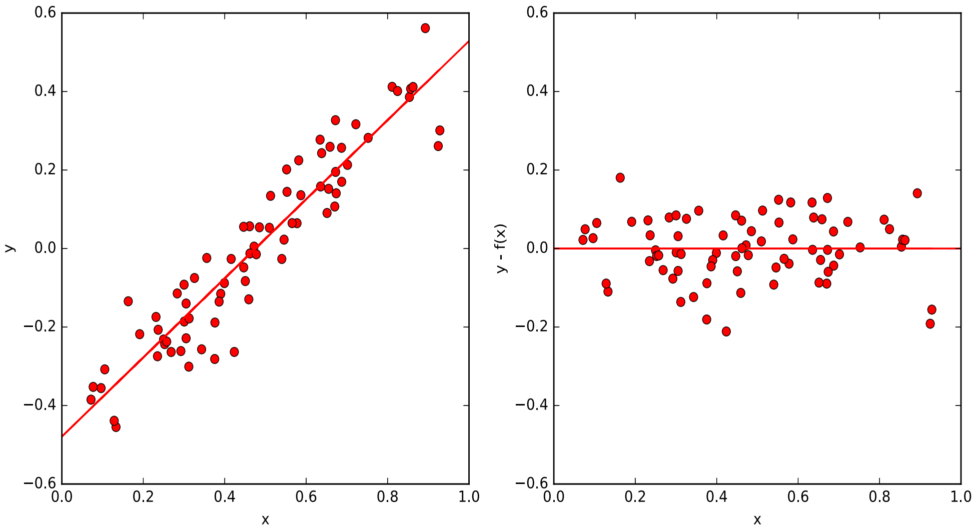
Explaining the Variance

- Let f(x) = mx+c be the predictive value of y from x, with the parameters m and c corresponding to the best possible fit. The residual values $r_i = y_i - f(x_i)$ will have mean zero, as shown in the Figure.
- Further, the variance of the full data set V (Y) should be much larger than V (r) if there is a good linear fit f (x). If x and y are perfectly correlated, there should be no residual error, and V (r) = 0. If x and y are totally uncorrelated, the fit should contribute nothing, and V (y) ≈ V (r). Generally speaking, $1 - r^2 = V(r)/V(y)$.

This is implemented in **section 9** in the code.

- Consider the left Figure 8, showing a set of points admitting a good linear fit, with correlation r = 0.94. The corresponding residuals $r_i = y_i - f(x_i)$ are plotted on the right. The variance of the y values on the left V (y) = 0.056, substantially greater than the variance V (r) = 0.0065 on the right. Indeed,

$$1-r^2 = 0.116 \quad \leftarrow\rightarrow \quad V(r)/V(y) = 0.116.$$

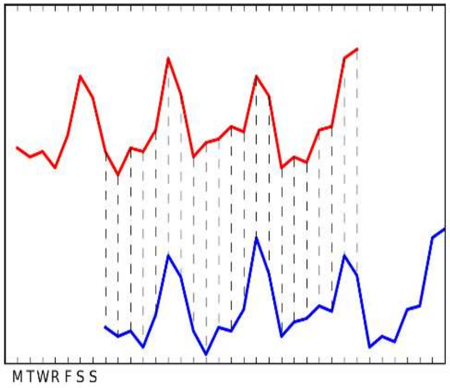


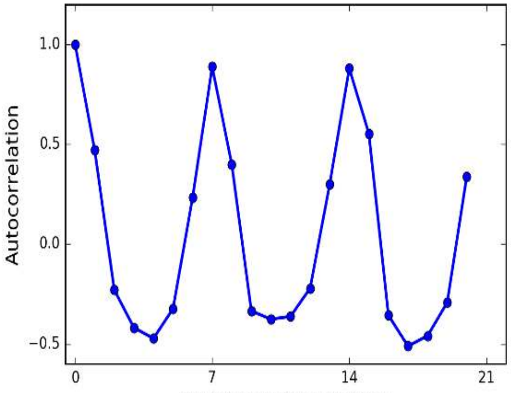
Some More Definitions

- Correlation Does Not Imply **Causation**.

- **Autocorrelation**: Comparing a sequence to itself.
- **Autocorrelation function:** the series of correlations for all $1 \le k \le n - 1$.
- **Periodicities by Autocorrelation:** to recognize a cyclic pattern in a sequence S, we correlate the values of S_i with S_{i+p}, for all $1 \le i \le n-p$. If the values are in sync for a particular period length p, then this correlation with itself will be unusually high relative to other possible lag values.

- The Figure shows a peak at a shift of seven days (and every multiple of seven days) establishes that there is a weekly periodicity in sales: more stuff gets sold on weekends.
  This is implemented in section 10 and 11 of the code





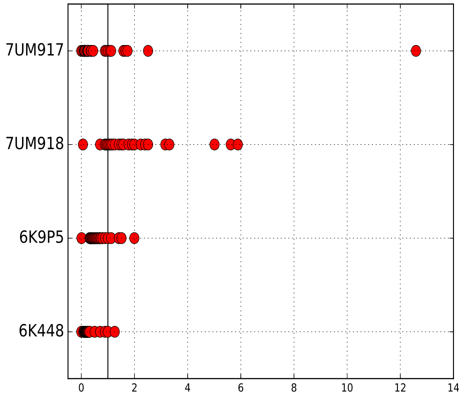
Logarithms Important Definitions

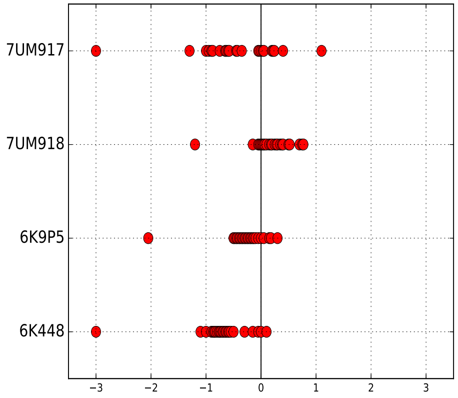
- The logarithm is the inverse exponential function $y = b^x$, an equation that can be rewritten as $x = \log_b y$.
- This definition is the same as saying that $b^{\log_b y} = y$.

- **Exponential functions** grow at a very fast rate: consider b = $\{2^1, 2^2, 2^3, 2^4, . . .\}$.
- **Logarithms functions** grow at a very slow rate: these are just the exponents of the previous series {1, 2, 3, 4, . . .}.
- Summing the logarithms of probabilities is much more numerically stable than multiplying them, but yields an equivalent result

Ratios

- **Ratios** are quantities of the form a/b appear in dataset features values, or comparing pair of features, or normalising data over weights. Ratios behave differently when reflecting increases than decreases.
- The ratio 200/100 is 200% above baseline, but 100/200 is only 50% below despite being a similar magnitude change.
- Thus, averaging ratios is committing a statistical sin. Do you really want a doubling followed by a halving to average out as an increase, as opposed to a neutral change? You can use geometric mean or the logarithms of these ratios.
- The left Figure is a graph from a student paper, showing the ratio of new score over old score on data over 24 hours (each red dot is the measurement for one hour) on four different data sets (each given a row).
- The solid black line shows the ratio of one, where both scores give the same result. try to read it?
- The right Figure plots the logarithms of the ratios. The space devoted to left and right of the black line can now be equal.





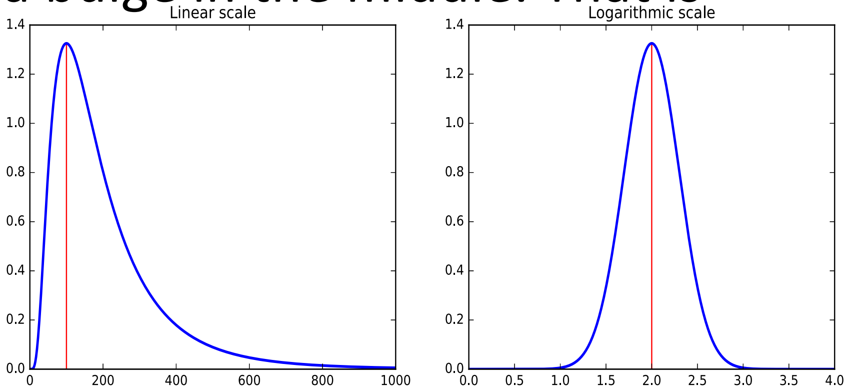
- Both algorithms are implemented in **sections 12** in the code.

Skewed Distributions

- Skewed distributions happens often and is not as a nice feature as normal bell distribution. For Example the human wealth, poor people at the zeros, average people at the thousands, and Bill Gates is pushing $100 billion as of this writing. Normalization to convert distributions of the shape in the left Figure to the log normal distribution shape in the right by taking the logarithm of variables with a power law distribution brings them more in line with traditional distributions.

- The acid test is to plot a frequency distribution of the transformed values and see if it looks bell-shaped: grossly-symmetric, with a bulge in the middle. That is when you know you have the right function.

- This is implemented in **sections 13** in the code.


Linear scale
Logarithmic scale