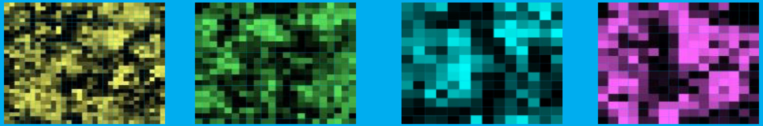
Gene Set Enrichment Analysis (GSEA) Part I



Network Analysis in Systems Biology

Neil Clark, PhD

Postdoctoral Fellow, Ma’ayan Lab

Department of Pharmacology and Systems Therapeutics

Icahn School of Medicine at Mount Sinai, New York, NY 10029

Introduction to GSEA

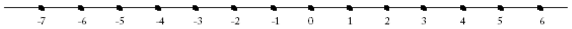
- Microarray experiments give the expression level of many genes.
- We would like to evaluate the difference in expression between two conditions, e.g., diseased and normal
- Traditionally methods look for individual genes which are differentially expressed between the two conditions but this has problems:
  - No single gene may stand out in the noise
  - Many genes may stand out but without any unifying biological theme
- GSEA looks for sets of genes which are differentially expressed, the advantages being:
  - A set of genes may be more likely to stand out (larger signal-noise ratio)
  - Biological theme is integral, and aids understanding and further investigation.

Elements of GSEA

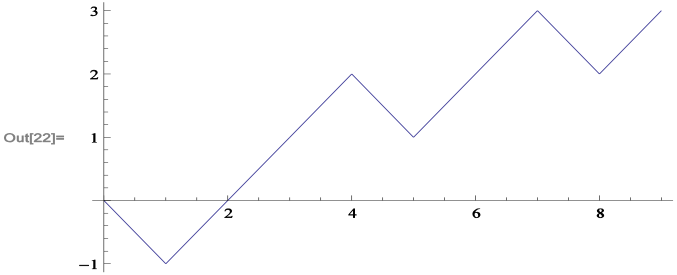
- One dimensional random walks
    - The Weiner process
    - The Brownian bridge
- The Kolmogorov-Smirnov test
    - Probability distributions
    - Statistical test of ‘goodness of fit’
- The GSEA test
    - The statistical test and evaluation of significance
    - An example from the literature

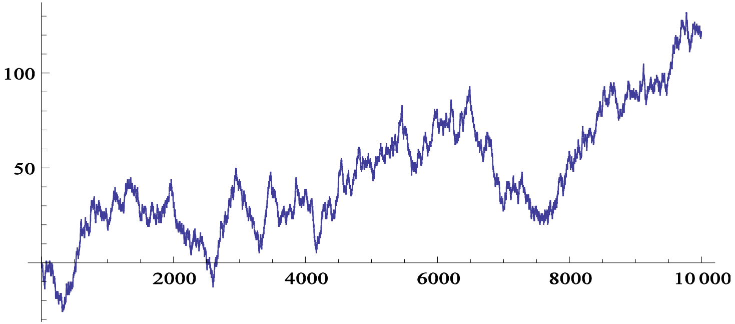
One-Dimensional Random Walks

Random walk on a one dimensional lattice



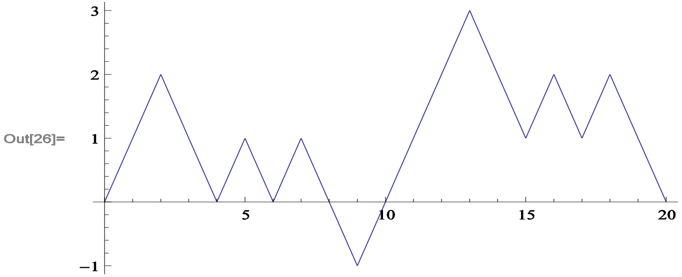
- Start at 0 and take discrete steps, left or right with equal probability.
- Can plot a random walk on a graph (upper right)
- It is possible to show that after *n* steps the mean distance from the starting point is proportional to $\sqrt{n}$
- With more steps comes more fluctuations (lower right)
- As the number of steps tends to infinity while the length of each step tends to zero the random walk tends to a ‘Weiner process’.

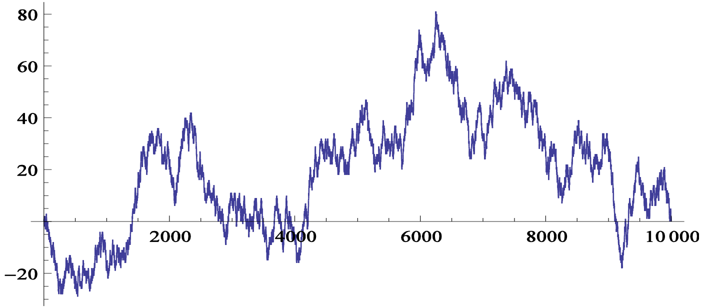




The Brownian Bridge

- A random walk with the end points fixed at zero is shown in the upper right figure
- As before we can plot the same kind of walk but with more steps, and see a broader range of fluctuations
- Let the number of steps go to infinity while letting the step size tend to zero, and this becomes the 'Brownian Bridge', *B(t)*
- May ask how far from the fixed end points is the walk likely to travel in the course of the Brownian Bridge.
- This maximum displacement is called the 'Supremum' of the Bridge, and has a cumulative probability distribution:





$$cdf(x) = 1 - 2\sum_{i}^{\infty}(-1)^{i-1}e^{-2i^2x^2}$$

Probability Distributions

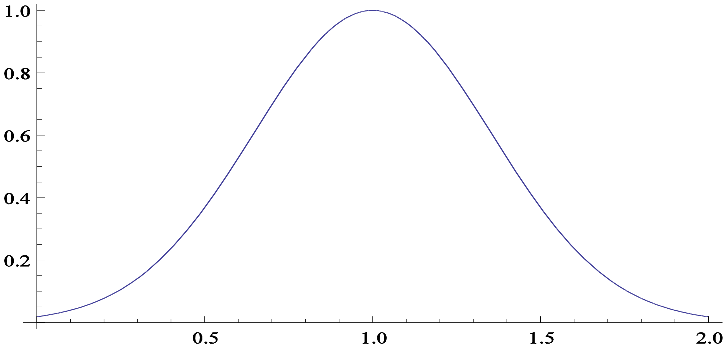
- The probability density function of the random variable *X*, gives the probability of measuring *X* in a given range by integration,

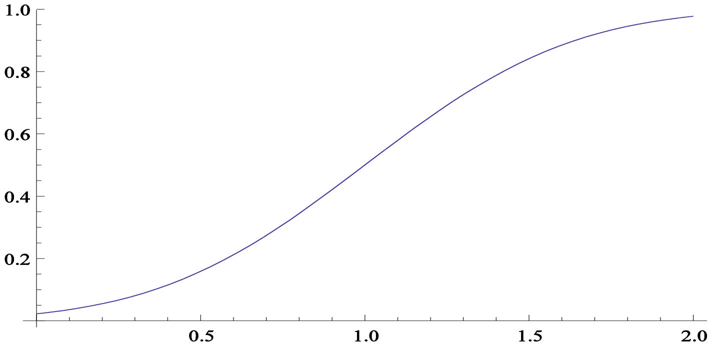
$$\int_a^b p(x)dx$$

- The upper right figure shows the probability density function of a Gaussian variable with mean 1, and variance 0.5
- The cumulative distribution function cdf(x), gives the probability of measuring *X* to have the value of *x* or lower,

$$cdf(x) = \int_{-\infty}^{x} p(x')dx$$

- The lower right figure shows the cumulative distribution function of the Gaussian variable described above





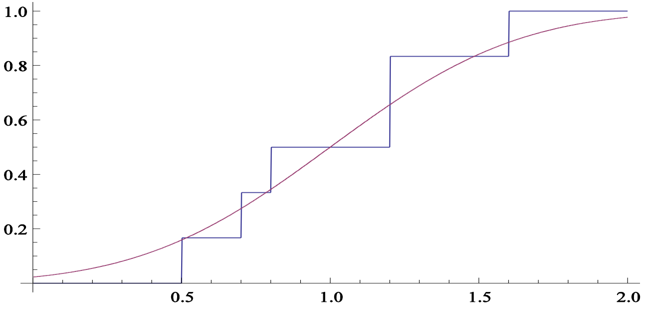
The Kolmogorov-Smirnov Test

- A statistical test of ‘Goodness of fit’
- Tests whether the data is consistent with a hypothesized cumulative distribution function
- Good when the sample size is small as there is no binning of the data

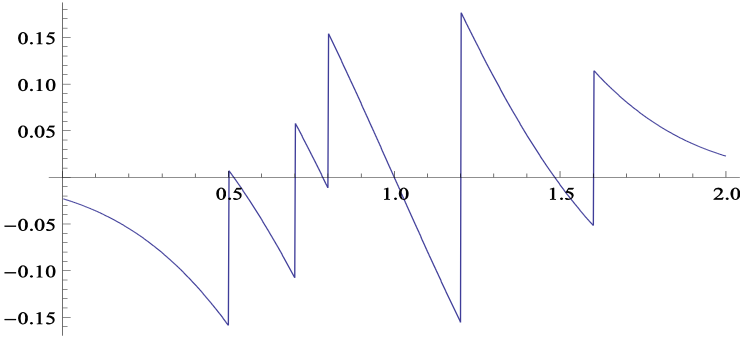
An example of the Kolmogorov-Smirnov test

- Take the data set {0.5, 0.7, 0.8, 1.2, 1.2, 1.6} and ask whether it is consistent with a Gaussian distribution of mean 1 and variance 0.5?
- The premise of the Kolmogorv-Smirnov test is that, if the data is consistent with the cdf, then the difference between them should be a random walk.
- The supremum of the difference should not be an outlier in the distribution for a Brownian Bridge.
- Use the cdf for the supremum of a Brownian bridge to estimate the significance of the test.

cdf of data (blue), and hypothesized distribution(red):



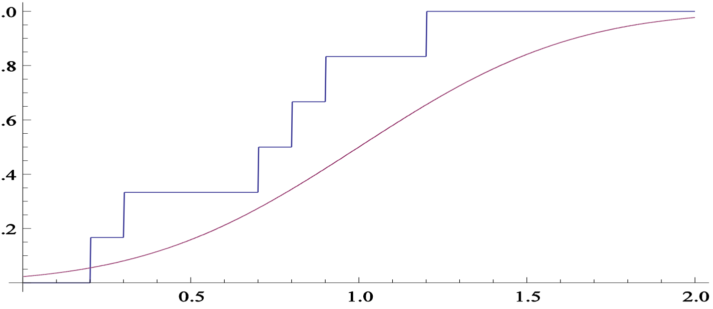
The difference between the two :



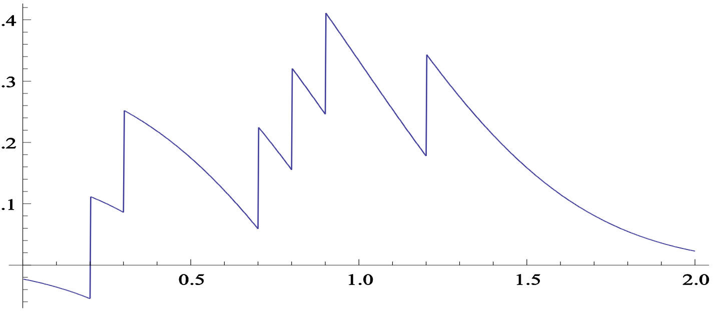
Another example of the Kolmogorov-Smirnov test

- Take the data set {0.2, 0.3, 0.7, 0.8, 0.9, 1.2} and repeat the test
- In this case we see the walk shown on the lower right: there is a clear bias.
- Comparing to the statistical tables, this data does not fit the hypothesized distribution so any significant degree.

cdf of data (blue), and hypothesized distribution(red):



Difference between the two



Overview of GSEA

- Take gene expression data from two different conditions and rank according to the differential expression across the conditions
- Take a test set of genes and determine whether they are collectively differentially expressed
- Randomly swap the class labels of the data and repeat the test many times as a gauge of significance