Midterm Review

Review of previous weeks

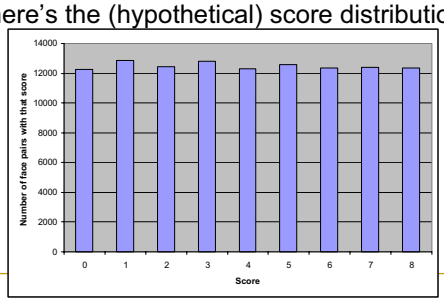
- Pairwise sequence alignment
  - Scoring matrices
    - PAM, BLOSUM,
  - Dynamic programming
    - Needleman-Wunsch (Global)
    - Semi-global (no end-gap penalties)
    - Smith-Waterman (Local)

Review of previous weeks

- Statistical significance of alignments
  - p-value, E-value, z-score?
  - Computing significance using random samples.
  - Example:
    - Suppose we have a face matching algorithm and we assign scores based on match of gender, complexion, eye-color, hair-color, use of glasses, use of face jewellery, existence of mustache, beard.

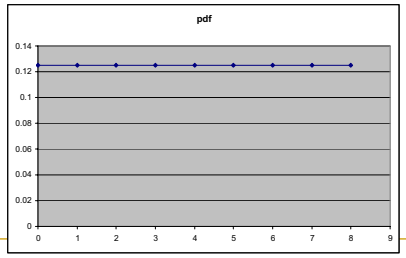
Statistical significance

- We match 100K pairs of faces randomly and here's the (hypothetical) score distribution:



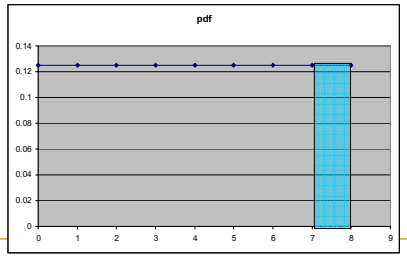
Statistical significance

- What is the p-value of a match with score 7? score 1?



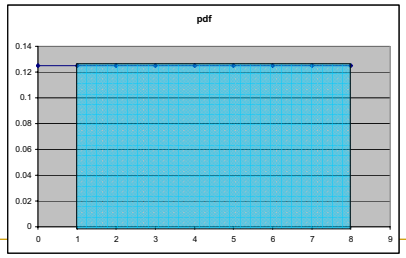
Statistical significance

- p-value(x>=7) = 0.125



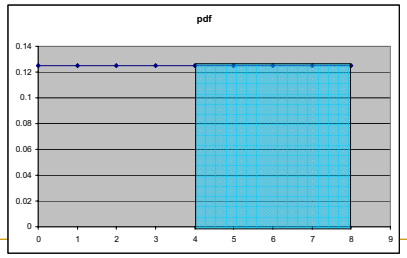
Statistical significance

- p-value(x>=1) = 0.875



Statistical significance

- p-value(x>=4) = 0.5



Suffix Trees and Suffix Arrays

- Construction of suffix trees and suffix arrays
- Pattern search in suffix trees/arrays
- Other applications
  - E.g.,
    - Finding the most occurring pattern of length 2 in a string (Solution: count the # of leaves below the nodes that have string depth>=2)