EE550
Computational Biology

Week 11 Course Notes

Instructor: Bilge Karaçalı, PhD

Topics

- Bioinformatics
  - Preliminaries
    - Randomness in measurements
    - Probability distributions
    - Histograms and empirical cumulative distributions
    - Sample statistics
  - Hypothesis testing using t tests
  - Parametric and nonparametric classification

Motivation

- High throughput quantitative molecular biology data
  - Cannot be processed or analyzed manually
    - The data volume is well beyond the amount that can be handled manually
      - Sequence data from many thousands of genes and proteins
      - Signal transduction or gene transcription network maps
      - Gene expression data from microarrays
      - …
    - Manual analysis cannot provide any sense of statistical significance useful for making inferences regarding the biological problem at hand

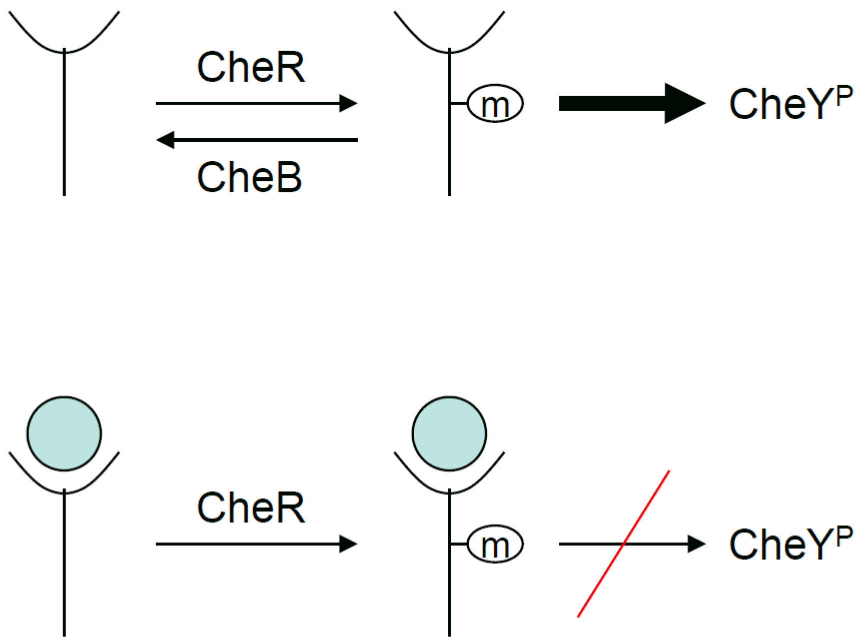
➔ Computer algorithms

Preliminaries

- Randomness in measurements
  - All measurements are subject to fluctuations
    - Fluctuations in the entity to be measured
    - Transient effects
    - Thermal noise in the measuring instrument
    - Quantization errors
  - Such fluctuations alter the measured value of a parameter of interest from its “true” value
  - In other instances, the parameter of interest fluctuates in and of itself from one instance to another
  - All these effects combine to produce deviations around some average

Preliminaries

- Example: Cell-to-cell variation of the amount of CheR in E. coli chemotaxis
  - When methylated, the receptor complex X phosphorylates CheY that in turn triggers direction change
  - The amount of CheR determining the steady state concentrations of the methylated receptor complex X changes from cell to cell
  - As a result, some cells are more nervous and change direction more often, while others are much more relaxed
  - All these effects combine to produce deviations around some average



Preliminaries

- Random variables
  - Technically:
    - A random variable is a mapping from a probability space (S, Ω, P) onto a measurable space (S, Ω)
      - S is the domain; also called the universal set of all possible outcomes/values
      - Ω is the sigma-algebra associated with the domain
      - $P: \Omega \rightarrow [0,1]$ is the probability measure such that $P(S) = 1$ and $P(\omega) \geq 0$ for all ω in Ω
  - Practically:
    - A random variable denotes the values of a parameter of interest measured under noisy or erroneous but generally stable conditions
      - The value of the random variable changes every time a measurement is made
      - Ranges of possible values that a random variable can take are associated with a probability between 0 and 1

Preliminaries

- Example:
  - Consider a fair die
    - A perfect cube with faces numbered from 1 to 6
    - When thrown, it has equal chance to land on its different faces
  - Throwing of this die corresponds to a random experiment
  - The measurement related to this random experiment is the reading of the number written on the face looking up
  - Each throw corresponds to a distinct realization of the random experiment
    - The measurement is simply the outcome of the experiment
  - Probabilities are assigned to collections (or sets) of events

- Q: Suppose a fair die is thrown. What are the chances that the outcome will be
  - Greater than or equal to 1?
  - Less than 10?
  - Less than 100?
  - 1 or 2 or 3?
  - 4 or 5 or 6?
  - 1 or 3 or 5?
  - 2 or 4 or 6?
  - 1 or 2?
  - 2 or 4?
  - 5 or 6?
  - 1?
  - 2?
  - 3?
  - …

Source: https://www.123learning.co.uk/pack-of-10-dice

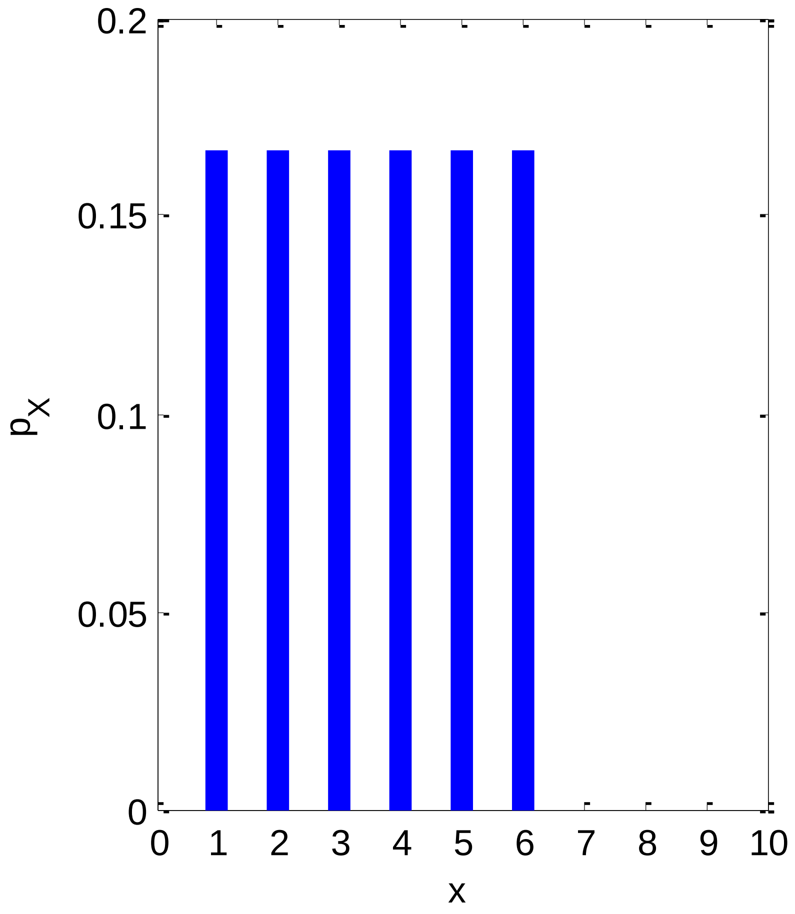
Preliminaries

- The odds of different possible outcomes are expressed in terms of probability distribution – mass or density – functions
  - Let X denote the random variable associated with the throwing of a fair die

$$\Pr\{X=1\}=1/6$$
$$\Pr\{X=2\}=1/6$$
$$\Pr\{X=3\}=1/6$$
$$\Pr\{X=4\}=1/6$$
$$\Pr\{X=5\}=1/6$$
$$\Pr\{X=6\}=1/6$$

  - Therefore, the probability mass function of X, denoted by p_X, is

$$p_X(x)=\begin{cases}1/6 & \text{if } x\in\{1,2,3,4,5,6\}\\ 0 & \text{otherwise}\end{cases}$$



Preliminaries

- The **probability distribution** of a random variable governs the odds of observing some specific values in a chance event
- In case the exact form of this probability is not known, it can be estimated
  - using many realizations of the corresponding chance event
- A most common way of estimating underlying probability distributions is by way of **histograms**
  - The more realizations, the better the estimate
  - Still, ambiguities abound

Preliminaries

- Consider estimating the underlying probability distribution of a fair die experiment from 100 independent realizations
  - The die is thrown $N = 100$ times
  - The numbers that come up each time are recorded
  - Let N_1 be the number of times the face with the number 1 comes up, and similarly for N_2, N_3, N_4, N_5, and N_6
    - Or, simply, N_x for $x = 1, \dots, 6$
    - Clearly,

$$N_1 + N_2 + N_3 + N_4 + N_5 + N_6 = 100$$

  - Define h by

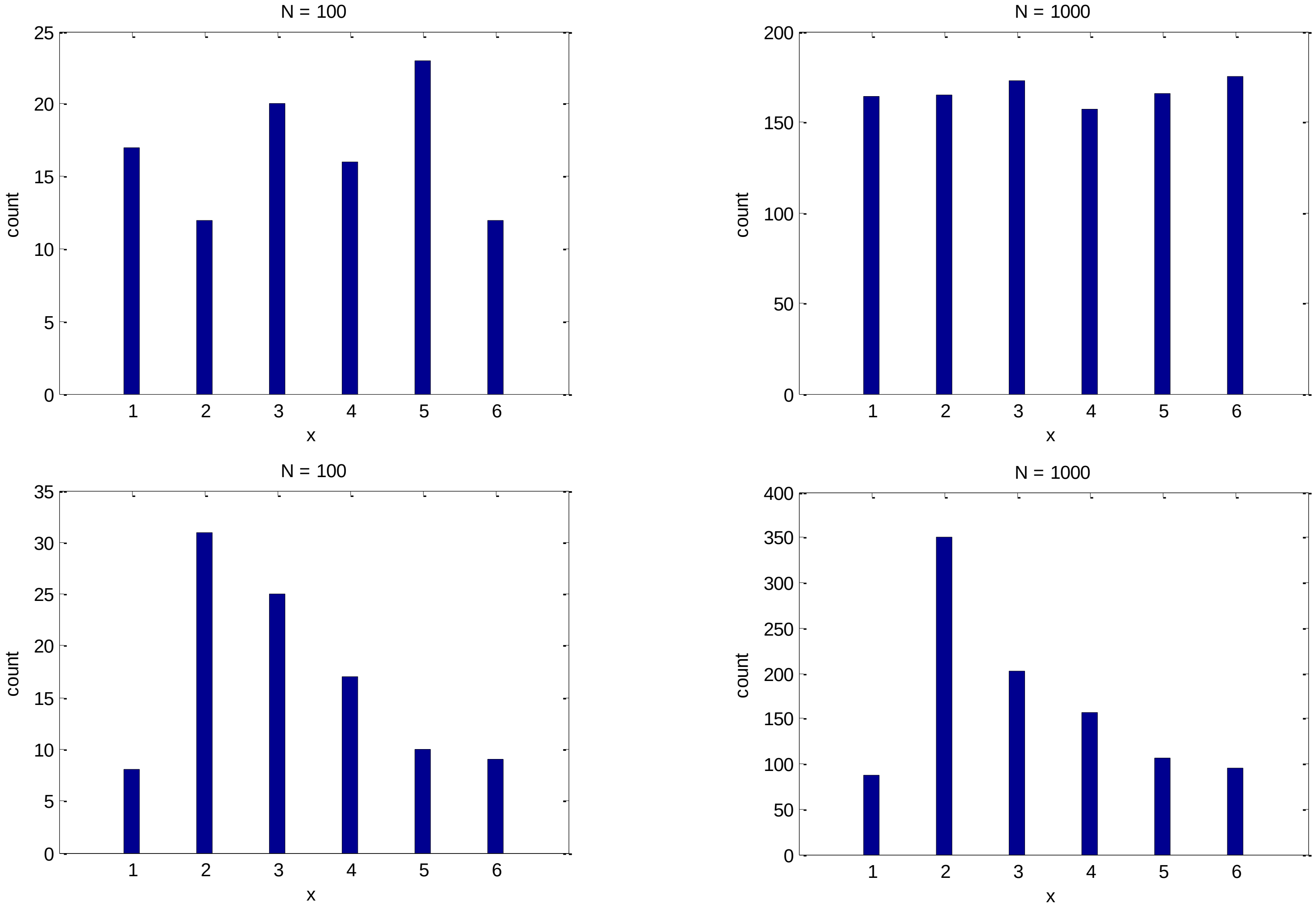
$$h(x) = \frac{N_x}{100}, x = 1,2, \dots, 6$$

  - Then, h is a histogram of the 100 realizations of the random variable X, and an estimate of p_X

Preliminaries

fair die

unfair die



Preliminaries

- The die throwing experiment describes a **discrete** random variable
  - The outcomes are elements in a finite set $\{1,2,3,4,5,6\}$
- More interesting examples tend to assume values from a continuum
- The random variables associated with such parameters are called **continuous** random variables
  - Continuous random variables possess similar definitions as the discrete random variables
    - Probability measures, chance events, …
  - But they differ in certain crucial ways, especially in how the probability distributions are defined
    - Let X denote the height of a freshman at IYTE in meters
    - Q: What is the probability that a freshman at IYTE will be 1.70m tall, i.e., $\Pr\{X = 1.70\} = ?$
    - A: ZERO!!!
    - But, but, but… A freshman does have a certain height; if it's not 1.70 EXACTLY, it is somewhere near…
    - So what?

Preliminaries

- The laws governing the chance structure associated with the values of continuous random variables are given in terms of set probabilities
  - The probability of interest is not $\Pr\{X = 1.70\}$, but $\Pr\{X \leq 1.70\}$
- The **cumulative distribution function** of X, denoted by $F_X(x)$, is defined as
$$F_X(x) = \Pr\{X \leq 1.70\}$$
- Note that
  - $\lim_{x \to -\infty} F_X(x) = 0$
  - $\lim_{x \to \infty} F_X(x) = 1$
- In turn, the **probability density function** $f_X(x)$ is defined as the derivative of $F_X(x)$ as
$$f_X(x) = \frac{dF_X(x)}{dx}$$

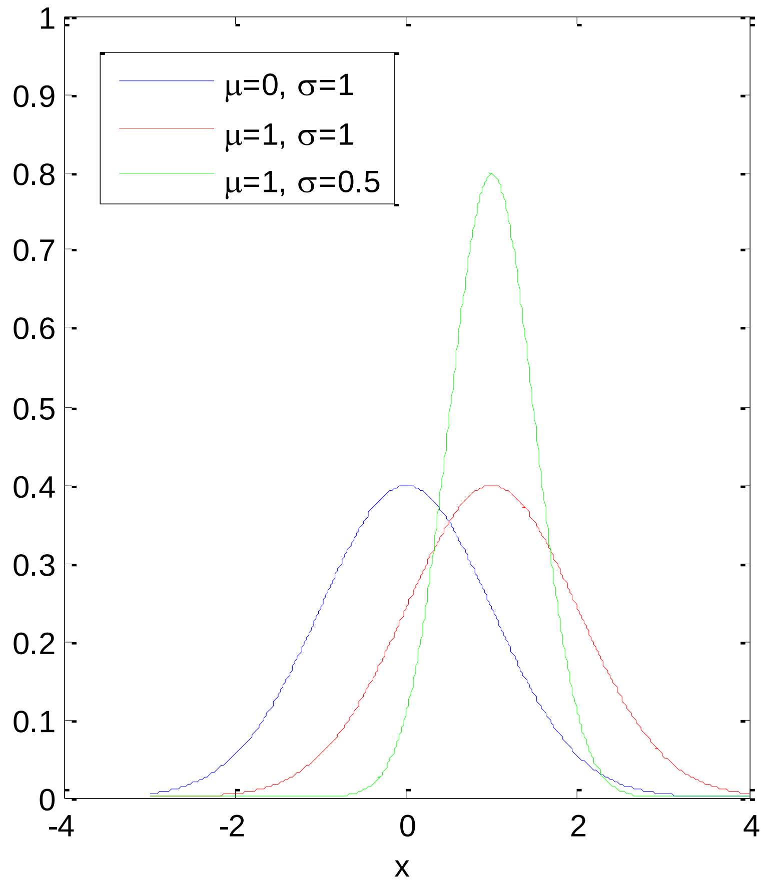
Preliminaries

- There are certain key **probability distribution families** that have been found very useful in describing the chance structures associated with real life random events
  - Gaussian probability distribution function
    - Bell curve
  - Exponential probability distribution function
    - Time-to-event
  - Binary probability distribution function
    - Heads or tails?
  - Binomial probability distribution function
    - How many heads or tails in so many repeats?
  - Poisson probability distribution function
    - How many heads or tails so far?

Preliminaries

- Gaussian probability distribution
  - A continuous function with two parameters
    - Mean μ
    - Variance σ^2

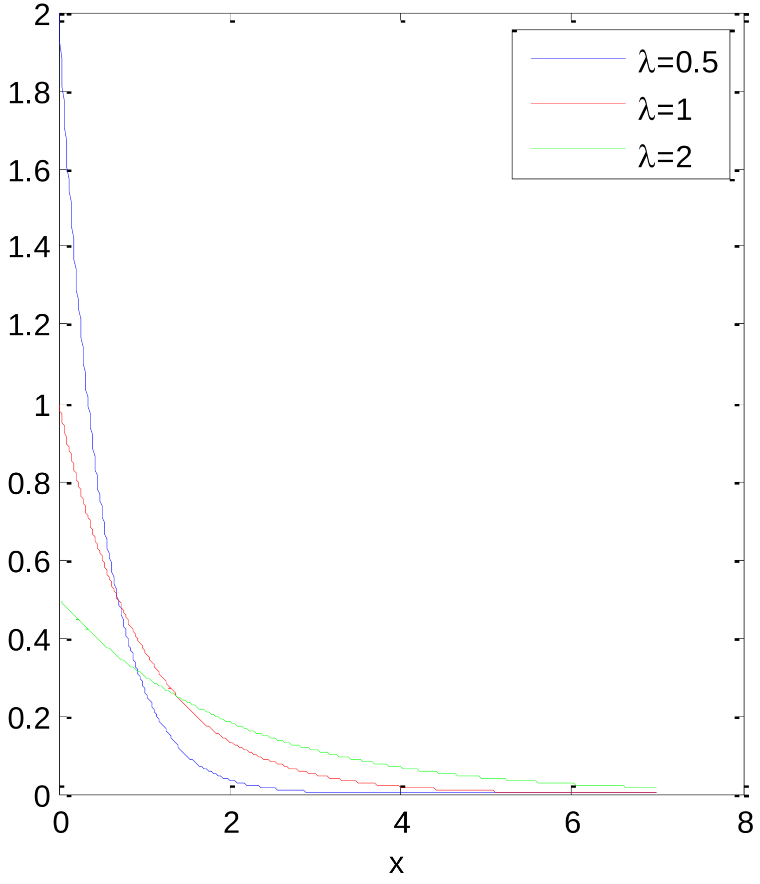
$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$



Preliminaries

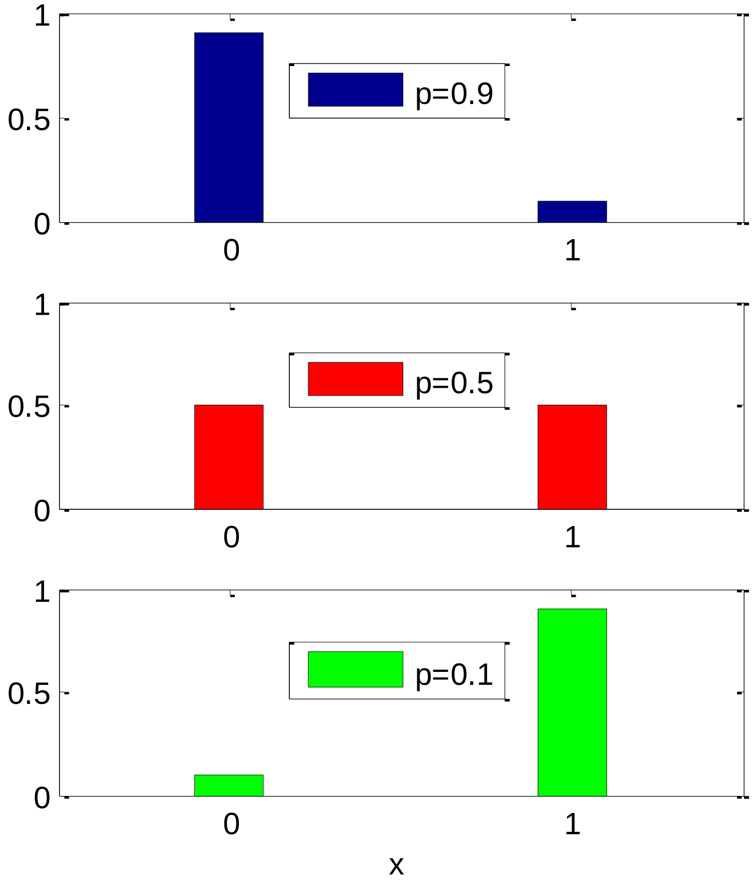
- Exponential probability distribution
  - Another continuous distribution, this time with one parameter
    - The rate of change λ
  - This is the only memoryless continuous distribution

$$f_X(x) = \lambda e^{-\lambda x}$$



Preliminaries

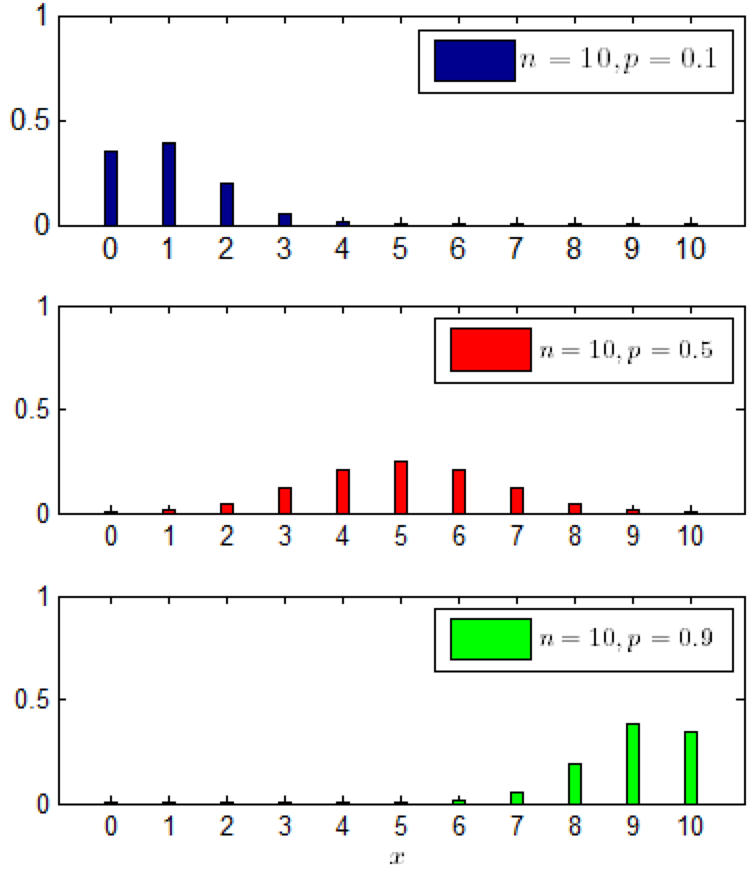
- Binary probability distribution
  - A discrete distribution with only two possible outcomes
    $p_X(\text{"first outcome"}) = p$
    $p_X(\text{"second outcome"}) = 1 - p$
  - The set of outcomes can be varied
    - $\{0,1\}$
    - $\{-1,1\}$
    - $\{A, B\}$
    - $\ldots$



Preliminaries

- Binomial probability distribution
  - A discrete distribution counting two possible outcomes in so many independent repeats with
    $p_X(\text{"first outcome"}) = p$
    $p_X(\text{"second outcome"}) = 1 - p$
  - The probabilities are then given by

$$\Pr\{\text{"}k\text{ first outcome in }n\text{ repeats"}\} = \binom{n}{k} p^k (1-p)^{n-k}$$

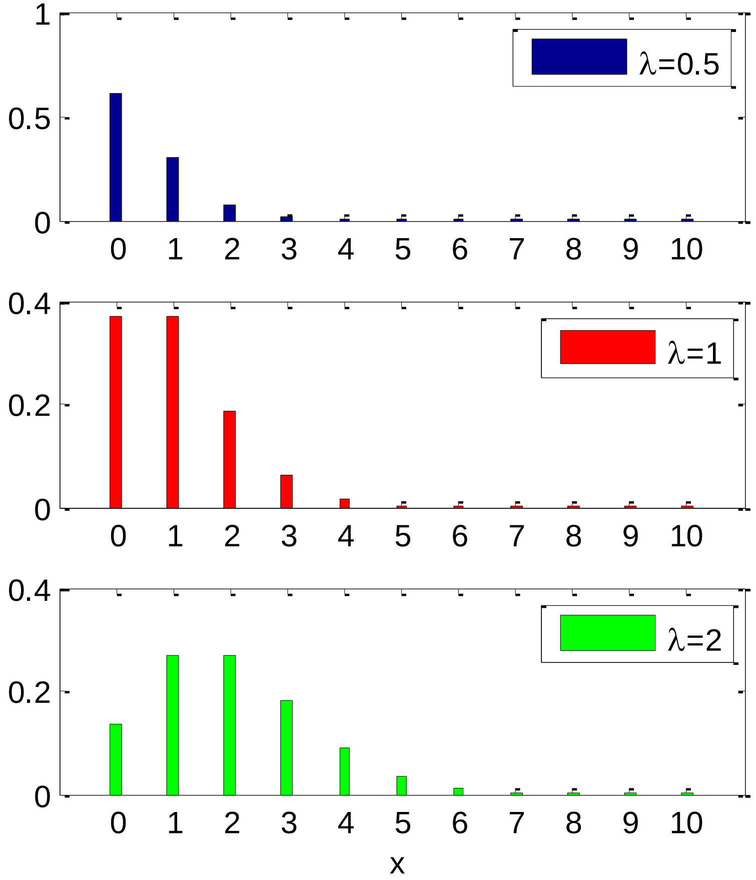


Preliminaries

- Poisson probability distribution
  - Another discrete distribution with one parameter
    - Rate of change λ
  - Counts the number of times an event of interest occurs in a fixed period of time

$$p_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

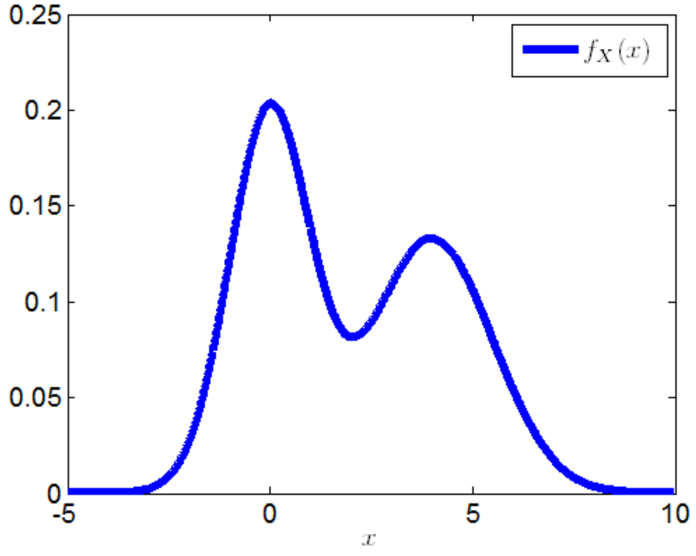
  - Interestingly, the time separation between successive events is exponentially distributed



Preliminaries

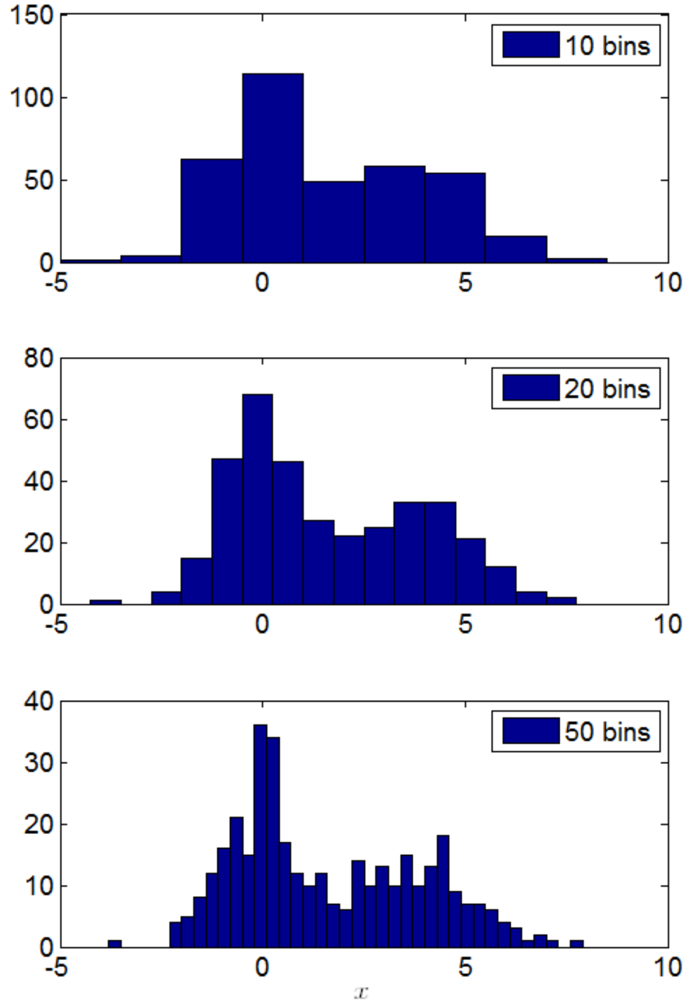
- A **sample set** represents a collection $\{x_j\}$, $i = 1,2,\dots,N$ of values observed from a given random variable
  - A collection of freshman heights from a randomly selected group of 10 first year students
  - Sequence lengths of 10000 randomly selected human proteins
  - Ages (in years) of 120 Alzheimer's Disease patients
  - …
- The distribution of values in the sample set can be characterized using
  - Histograms
    - N_k represents the number of samples in an interval $(x'_{k-1}, x'_k]$ with
      $$x'_0 < x'_1 < \cdots < x'_{K-1} < x'_K$$
    - K represents the number of bins
  - Sample statistics
    - Sample mean $m = \frac{1}{N}\sum_i x_i$
    - Sample variance $s^2 = \frac{1}{N-1}\sum_i (x_i - m)^2$

Preliminaries



Above: The probability density function of some random variable X

Right: Histograms of 1000 realizations of X with different bin sizes



Preliminaries

- Remarks:
  - Histograms are informative only when the bins are located and sized appropriately
    - There is no sense in placing the bins on regions of zero occurrence
    - If the bins are too small, the resolution will be high, but they will cover only a few samples producing large errors
    - Larger bins will possess many samples providing a smaller error, but the resolution will be poor
  - Sample mean m and variance s^2 (standard deviation s too) describe where the samples are centered and how wide they are dispersed
    - This is usually fine for unimodal distributions with a single peak
    - On the other hand, this is terribly inadequate to represent multimodal distributions
      - The samples may be clustered around a handful of values with little or no dispersion
      - The mean will not capture this localization, and the standard deviation will indicate large dispersion when there is only very little

Hypothesis Testing

- Suppose we are given two sample sets

$$\{x_i\}, i = 1,2, \ldots, N_x$$

and

$$\{y_j\}, j = 1,2, \ldots, N_y$$

  - The heights of freshman students in EE and MB&G
  - The sequence lengths of human and yeast proteins
  - …
- The task is to decide if these two sample sets represent events with different characteristics
  - These sample sets represent events with different characteristics if and only if the underlying probability distributions are different
- One option it to generate histograms for the two sets and see if they look different
  - Feasible first-attempt, but difficult to infer a statistical significance measure
    - Requires a measure of distance between histograms and permutation tests
- Another option is to assume these sample sets originate from distributions of the Gaussian family with potentially different parameters, and test to see if their parameters might be the same

Hypothesis Testing

- Presumptions about the statistical nature of the observed data are tested against empirical evidence presented by the data
- Formally:
  - A **null hypothesis** H_0 is formulated postulating a statement
    - the uninteresting explanation for the observed data
  - A **complementary hypothesis** H_c is automatically formulated postulating the invalidity of the statement
    - the interesting/desired/hoped-for explanation
  - A **probability** P is computed as the probability of observing the actual observed sample statistic under the null hypothesis
  - If the probability is smaller than a prescribed **significance threshold**, the **null hypothesis is rejected** at the benefit of the complementary hypothesis
    - Small P values indicate that the sample statistic is unlikely to be observed if null hypothesis were true
    - Typical P value thresholds are 5% or 0.1%
- Note that this strategy requires **a statistic** to be computed from the data with a **known distribution** under the null hypothesis
  - Any statistic can be used as long as its distribution can be *guessed* well

Hypothesis Testing Using a Two-Sample t-Test

- Consider the following problem:
  - Two sample sets $\{x_i\}$ and $\{y_j\}$ are provided representing the value observed for a parameter of interest from two different groups
    - $\{x_i\}$ are the realizations of a random variable X
    - $\{y_j\}$ are the realizations of a random variable Y
  - Let μ_X and μ_Y represent the unknown means of the random variables X and Y
  - The task is to test the validity of the null hypothesis

$$H_0\text{: } \mu_X = \mu_Y$$

with a significance threshold $\alpha \ll 1$

➔ two-sided two-sample t-test

Hypothesis Testing Using a Two-Sample t-Test

- A t-test is a statistical comparison test that computes a probability for the null hypothesis given the data
- If the probability is smaller than the prescribed significance α, the null hypothesis is rejected in favor of the complementary hypothesis
- A few variants exist for the t-test
  - Equal sample sizes, equal variances
  - Unequal sample sizes, equal variances
  - Unequal sample sizes, unequal variances
  - Paired vs. unpaired
- The test calculates a T statistic for each case, and computes its probability when the null hypothesis is true as the P value

- For unequal sample sizes, equal variances:

$$T = \frac{m_X - m_Y}{s\sqrt{\frac{1}{N_X} + \frac{1}{N_Y}}}$$

where

$$m_X = \frac{1}{N_X}\sum_i x_i\,, m_Y = \frac{1}{N_Y}\sum_j y_j$$

$$s_X^2 = \frac{1}{N_X - 1}\sum_i (x_i - m_X)^2$$

$$s_Y^2 = \frac{1}{N_Y - 1}\sum_j \left(y_j - m_Y\right)^2$$

$$s = \sqrt{\frac{(N_X - 1)s_X^2 + (N_Y - 1)s_Y^2}{N_X + N_Y - 2}}$$

and

$$DF = N_X + N_Y - 2$$

Hypothesis Testing Using a Two-Sample t-Test

- Procedure for testing for the equality of means:
  1. Given the sample sets $\{x_i\}$ and $\{y_j\}$
  2. Calculate the sample means and variances
  3. Calculate the T statistic
  4. Compare the absolute value of the T statistic to the critical value T_c for which

$$F_t(T_c) = 1 - \alpha/2$$

where F_t denotes the cumulative distribution function of a t random variable with the corresponding degrees of freedom under the null hypothesis

OR

Calculate the P value via

$$P = 2 \cdot \left(1 - F_t(|T|)\right)$$

and see if it is smaller than α

Hypothesis Testing Using a Two-Sample t-Test

- Equality of the means example:
  - Let the sample sets be given as
  x's: {0.83, -0.09, -0.46, 0.05, -1.36, -0.21}
  y's: {-0.08, 0.13, 1.94, 0.57, 0.27, 0.99, 0.41, 0.87, 0.35}
  - Compute
    - The sample means
    $m_x = -0.2067$ and $m_y = 0.6056$
    - The sample variances
    $s_x^2 = 0.5097$ and $s_y^2 = 0.3640$
    - The T statistic

$$T = -2.3778$$

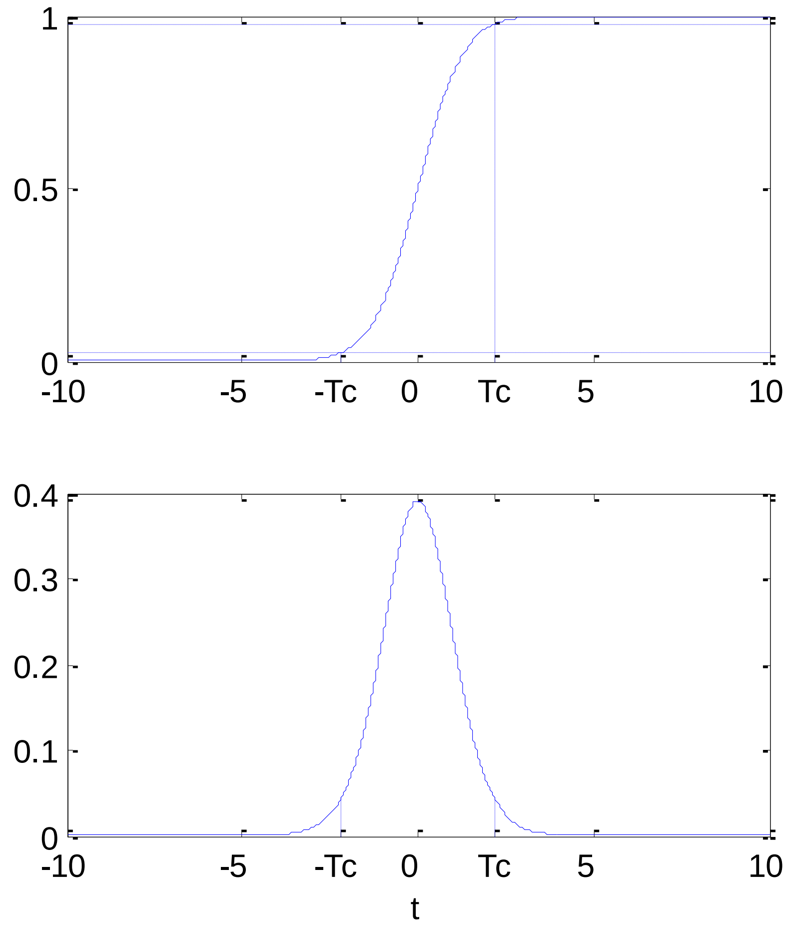
  - The P value associated with this T statistic is

$$P = 2 \cdot (1 - P_t(|T|)) = 2 \cdot 0.0167 = 0.0334$$

  - The critical value T_c is

$$T_c = 2.1448$$

  - **The null hypothesis is rejected!!**



Hypothesis Testing Using a Two-Sample t-Test

- Example (continued):
  - Now, let the sample sets be given as

    x's: {0.83, -0.09, -0.46, 0.05, -1.36, -0.21}

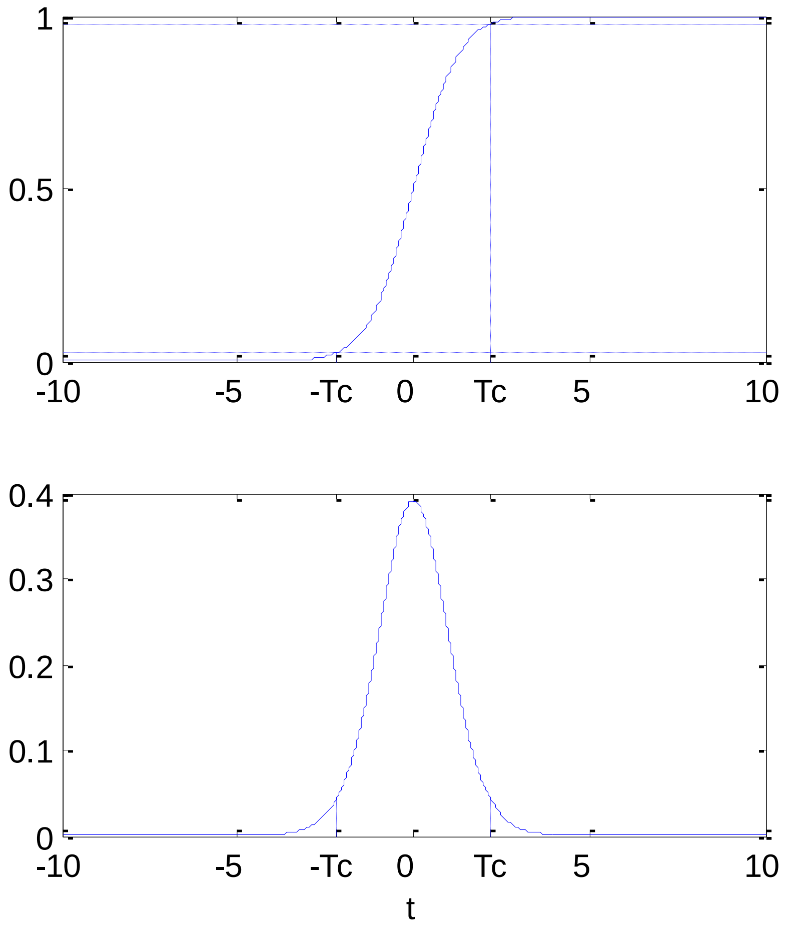
    y's: {-0.08, 0.13, 4.94, 0.57, 0.27, 0.99, 0.41, 0.87, 0.35}
  - Compute
    - The sample means

      $m_x = -0.2067$ and $m_y = 0.9389$
    - The sample variances

      $s_x^2 = 0.5097$ and $s_y^2 = 2.3648$
    - The T statistic

      $$T = -1.6914$$
  - The P value associated with this T statistic is

    $$P = 2(1 - P_t(|T|)) = 2 \cdot 0.0573 = 0.11$$
  - This time, **the null hypothesis is not rejected!!**
  - What is going on??



Hypothesis Testing Using a Two-Sample t-Test

- Remarks:
  - The t test is susceptible to deviations from the presumptions
    - Gaussianity of the underlying distributions
    - Presence of outliers
  - In addition, it determines whether there is reason to believe that the unknown means are different, but says little about how different they are
    - Given sufficient number of samples, the statistical power may suffice to detect even the tiniest differences between the means
    - Conversely, not detecting a difference of the means in a significant manner may simply be because the available data does not provide sufficient statistical power to detect a small difference
  - Finally, it is helpless when the random variables are multivariate
    - Hotelling's T^2 test can be used but is problematic

Parametric and Nonparametric Classification

- Often, several parameters are measured jointly and recorded in experiments
  - Heights, ages, and grade point averages of college freshmen
  - Lengths and amino acid compositions of amino acid sequences of human proteins
  - Gene expression of 40K genes in microarray experiments
  - …
- Such multivariate data sets require multivariate data analysis methods
- A common task when multivariate data from two or more sample sets are present is whether classification rules that distinguish these sets from one another can be constructed
  - If such a rule can be constructed, one can then determine
    - to which group a novel sample should belong
    - which parameter values are critical to distinguish the samples of different groups and in what conditions
  - Both these possibilities are absolutely vital to understand the biological problems in consideration

Parametric and Nonparametric Classification

- Given training data, classifier construction strategies are studied under two general categories
  - Parametric classification rules
    - A parametric model is assumed for the underlying multivariate probability distributions of the different groups
    - The parameters for these distribution models are estimated from available data
    - An optimal decision boundary is deduced from the estimated probability distributions
  - Nonparametric classification rules
    - No parameter-based model is assumed
    - Classification rules are constructed based on the similarity and distance structure between the available –manually annotated– “training” samples

Parametric and Nonparametric Classification

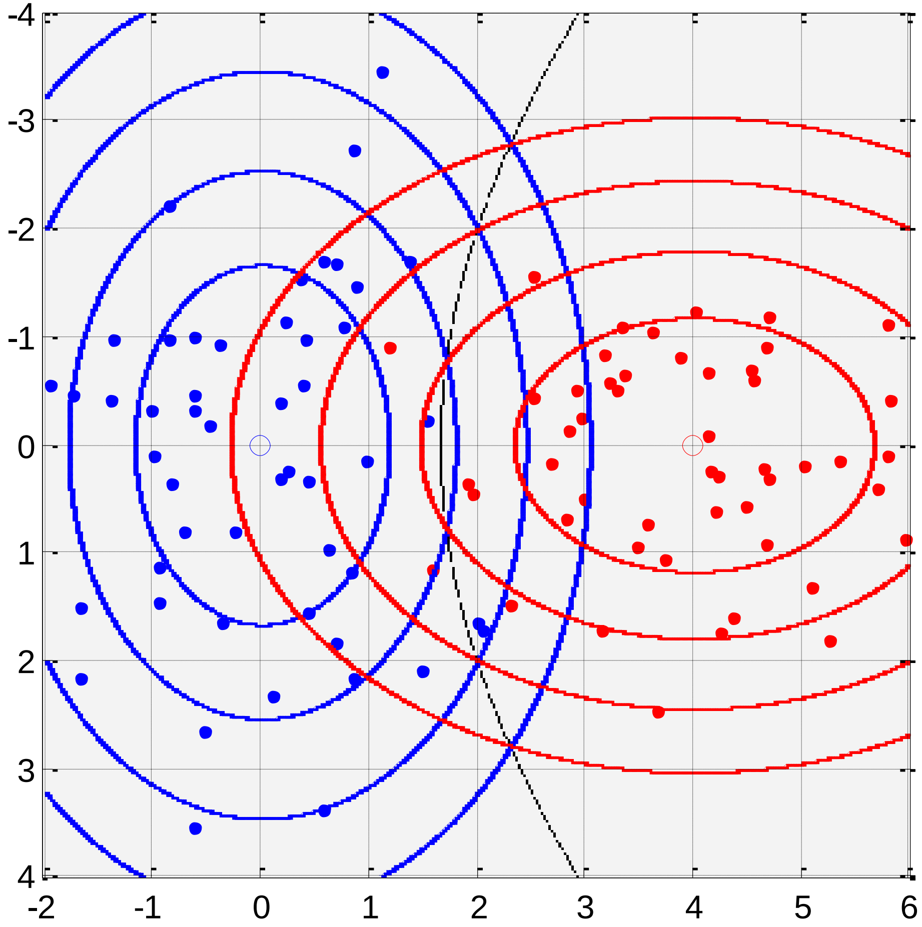
- Maximum likelihood classification
  - Given the multivariate training data
  - Estimate the means and the covariance matrices for all sample sets
    - The estimated sample distributions then become multivariate Gaussian distributions with the corresponding mean vectors μ_i and the covariance matrices Σ_i as

$$f_i(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n |\det(\Sigma_i)|}} e^{-\frac{1}{2}(\boldsymbol{x}-\mu_i)^T \Sigma_i^{-1} (\boldsymbol{x}-\mu_i)}$$

  - Construct the classification rule that assigns a new sample to the sample set with the greatest value of the probability density function at the new sample

$$f^{\mathrm{ML}}(\boldsymbol{x}) = \arg\max_j f_j(\boldsymbol{x})$$

Parametric and Nonparametric Classification



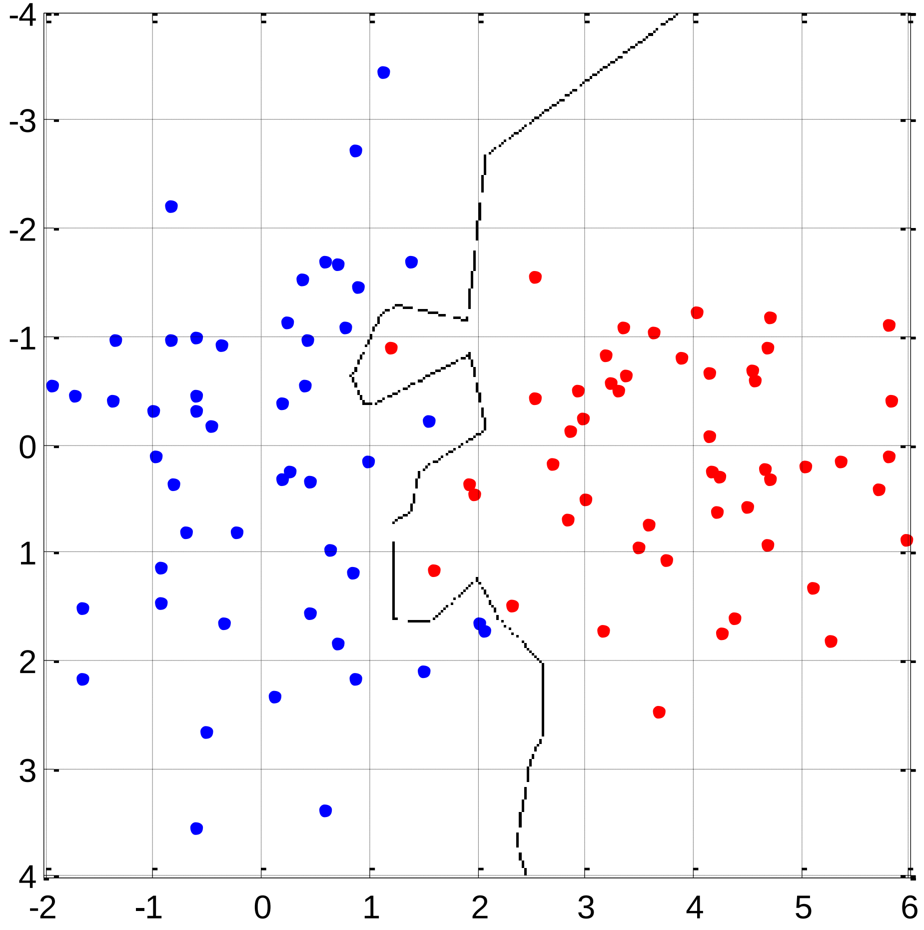
Parametric and Nonparametric Classification

- Nearest neighbor classification:
  - Store all the available data for training in a reference set $\{\boldsymbol{x}_j, y_j\}$, $\boldsymbol{x}_j \in IR^n$, $y_j \in \{1,2\}$ with $j = 1,2,\dots,\ell$
  - Assign the newly observed sample to the class with most similar samples
    - Similarity computed in terms of a defined measure, or as inverse distance, or a weighted combination, …
  - The classification rule is given by

$$f^{\mathrm{NN}}(\boldsymbol{x}) = y_{j_0}$$

where $j_0 = \arg\min_j \rho(\boldsymbol{x}, \boldsymbol{x}_j)$, with $\rho(\boldsymbol{x}, \boldsymbol{x}_j)$ calculating the distance between samples $\boldsymbol{x}$ and $\boldsymbol{x}_j$

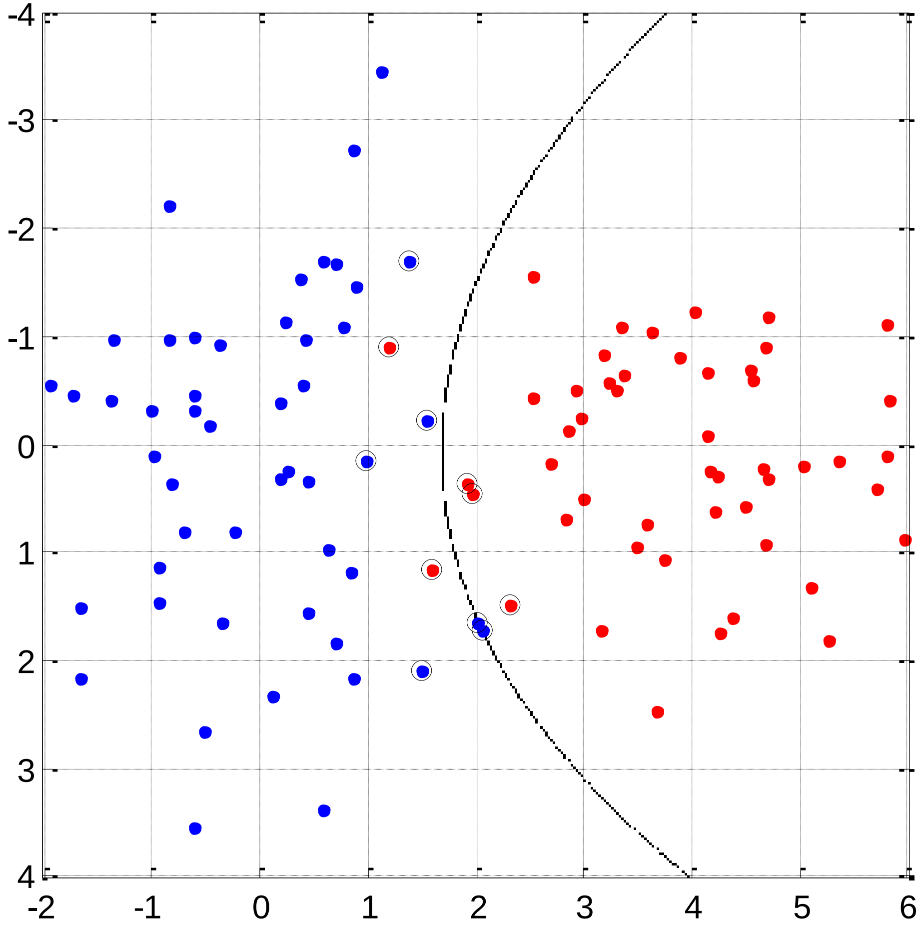
Parametric and Nonparametric Classification



Parametric and Nonparametric Classification

- Support vector machine classification:
  - A maximum margin linear classifier is constructed to separate the samples of two different classes
  - Nonlinear solutions are obtained by employing an inner product kernel to replace the original inner product between the samples
    - polynomial, Radial Basis Function, sigmoid, …
  - Linear maximum-margin solution in the transform space corresponds to a nonlinear solution in the observation space
  - For more details, see the literature
    - Maximization of the margin using the method of Lagrange multipliers
    - Karush-Kuhn-Tucker optimality conditions that produce the support vectors
    - Generalization to multiple class problems

Parametric and Nonparametric Classification



Summary

- Bioinformatics uses statistical analysis techniques to address molecular biology questions emanating from quantitative data in large volumes
  - The data collected from high throughput experiments can only be handled using computational methods
  - These methods use different strategies to answer a variety of questions
    - Whether the nature of measured parameters change from one group to another
    - Whether it is possible to derive classification rules to distinguish the different groups based on the measured data