JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Probabilities

Outline

- Probabilities
- Distributions
- Moments of Distributions
- Covariance Matrix
- Random Variables
- Hypothesis Testing

Why Study Probabilities

- Understanding the theoretical meaning of probabilities provides a frame work for the analysis of real world data sets.

- Having a mathematical framework to do an initial analysis of data allows a baseline to work on.

- Understanding probabilities allow for the development of more sophisticated methods such as
  - Random number generators
  - Data prediction
  - Data classification
  - Data mapping

Probabilities

- When data is not represented in numerical form there is some preprocessing involved either prior to using an algorithm or within the algorithm itself to convert it to a numerical form.

- For the study of algorithm analysis assume the data is represented by real numbers $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ where $\mathbf{x} \in \mathbb{R}^d$
  - d represents the dimension of the data or the number of random variables in the data (features).

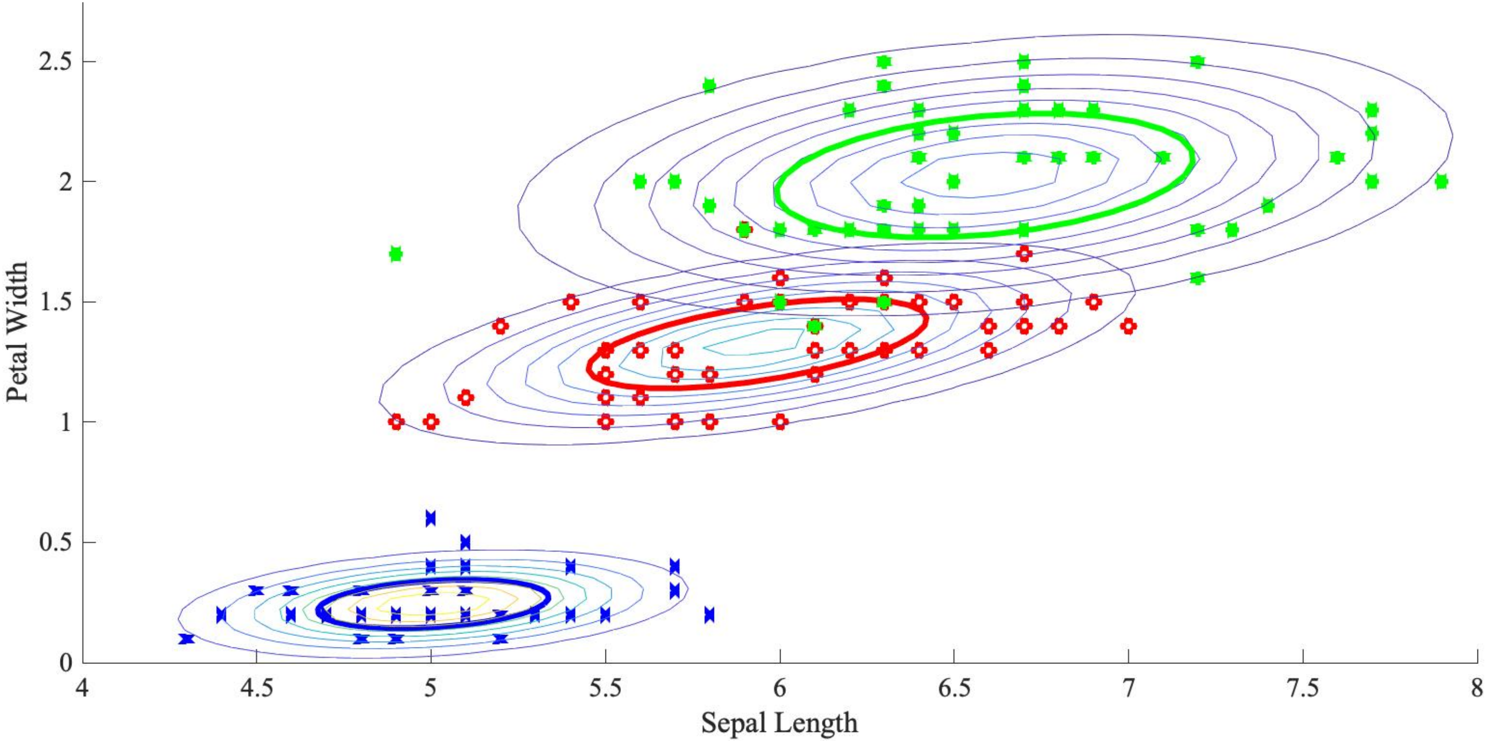
- When data is analyzed using probability theory it is possible to use pattern recognition techniques as well as machine learning to analyze the data in order to determine what the data represents.

Distributions

- Distributions have various shapes depending on the data being analyzed
  - Normal or Gaussian
  - Chi Square
  - Binomial
  - Poisson

- In this class we revert to the normal distribution so that the focus is on the analysis of algorithms

Normal Distribution Example



Moments of Distributions

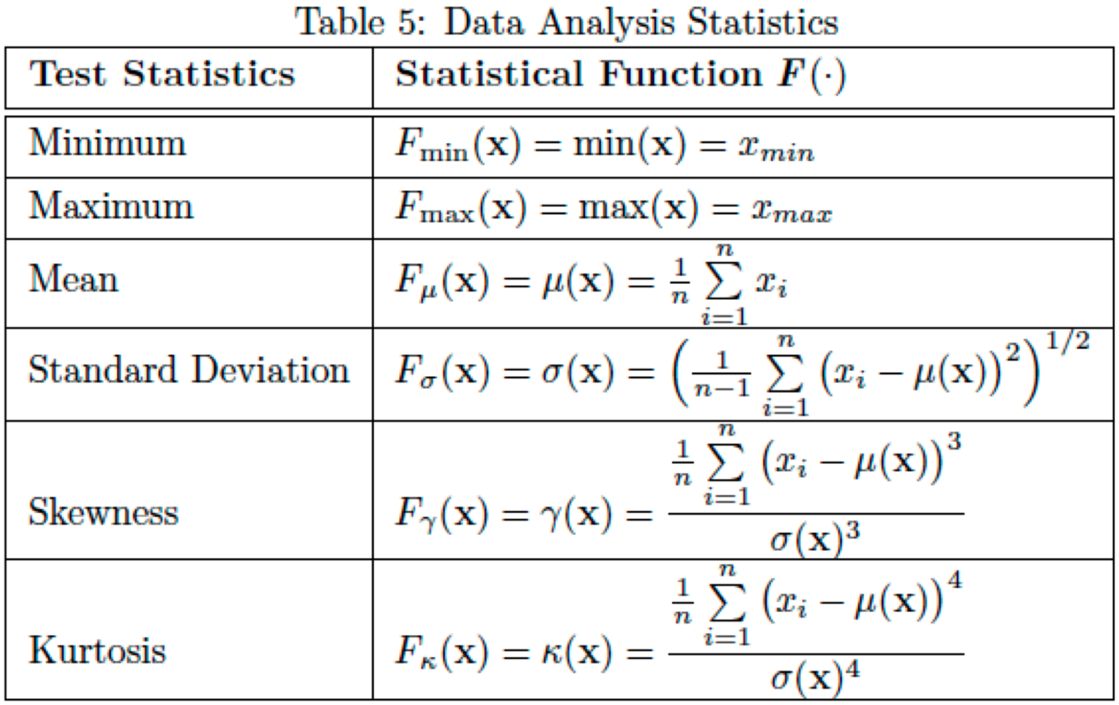


Table 5: Data Analysis Statistics

Test Statistics	Statistical Function $F(\cdot)$
Minimum	$F_{\min}(\mathbf{x}) = \min(\mathbf{x}) = x_{min}$
Maximum	$F_{\max}(\mathbf{x}) = \max(\mathbf{x}) = x_{max}$
Mean	$F_{\mu}(\mathbf{x}) = \mu(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$
Standard Deviation	$F_{\sigma}(\mathbf{x}) = \sigma(\mathbf{x}) = \left(\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \mu(\mathbf{x})\right)^2\right)^{1/2}$
Skewness	$F_{\gamma}(\mathbf{x}) = \gamma(\mathbf{x}) = \dfrac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu(\mathbf{x})\right)^3}{\sigma(\mathbf{x})^3}$
Kurtosis	$F_{\kappa}(\mathbf{x}) = \kappa(\mathbf{x}) = \dfrac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu(\mathbf{x})\right)^4}{\sigma(\mathbf{x})^4}$

The moments are

- Mean
- Standard deviation
- Skewness
- Kurtosis

In this class we will call these mathematical measurements “test statistics"

Covariance Matrix

$$C = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})\,(X_i - \bar{X})^T \; where\; C \;\in \mathbb{R}^{d\times d}$$

- The covariance matrix give the variance between each pair of random vectors (features)
  - The diagonal values are the variances of each of the vectors.

- The importance in understanding the covariance matrix is the the use of linear transformation
  - PCA
  - Rotation of random data to represent the data the covariance was calculated from