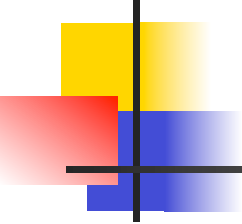


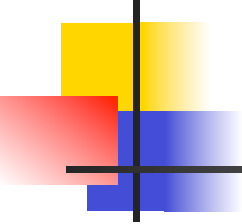
Data Modeling & Analysis Techniques

Probability & Statistics



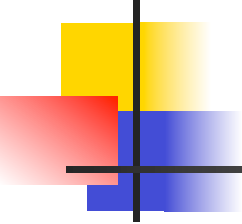
Probability and Statistics

- Probability and statistics are often used interchangeably but are different, related fields
 - Probability
 - Mainly a field of theoretical mathematics
 - Deals with predicting the likelihood of events given a set of assumptions (e.g. a distribution)
 - Statistics
 - Field of applied mathematics
 - Deals with the collection, analysis, and interpretation of data



Statistics

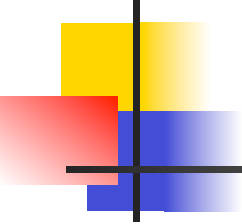
- Statistics deals with real world data
 - Data collection
 - How do we have to collect the data to get valid results
 - Analysis of data
 - What properties does the data have
 - What distribution does it come from
 - Interpretation of the data
 - Is it different from other data
 - What could cause issues in the data



Probability

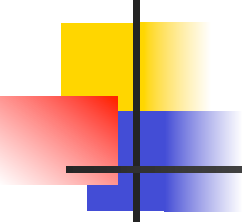
- Probability is a formal framework to model likelihoods mainly used to make predictions
 - There are two main (interchangeable) views of probability
 - Subjective / uncertainty view: Probabilities summarize the effects of uncertainty on the state of knowledge
 - In Bayesian probability all types of uncertainty are combined in one number
 - Frequency view: Probabilities represent relative frequencies of events

$$P(e) = (\# \text{ of times of event } e) / (\# \text{ of events})$$



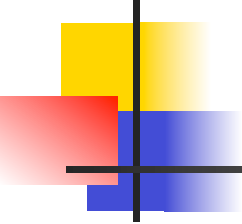
Data Modeling & Analysis Techniques

Probability Theory



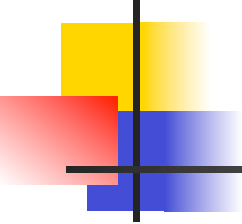
Probability

- Random variables define the entities of probability theory
 - Propositional random variables:
 - E.g.: IsRed, Earthquake
 - Multivalued random variables:
 - E.g.: Event, Color, Weather
 - Real-Valued random variables:
 - E.g.: Height, Weight



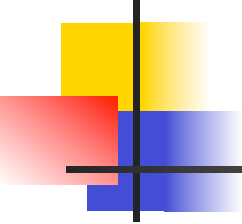
Axioms of Probability

- Probability follows a fixed set of rules
 - Propositional random variables:
 - $P(A) \in [0..1]$
 - $P(T) = 1, P(F) = 0$
 - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
 - $P(A \wedge B) = P(A)P(B | A)$
 - $\sum_{x \in \{T, F\}} P(X = x) = 1$



Axioms of Probability

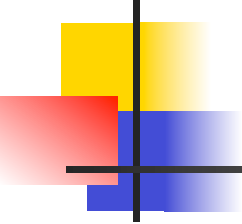
- The same axioms apply to multi-valued and continuous random variables
 - Multi-valued variables ($X \in S = \{x_1, \dots, x_N\}; Y, Z \subset S$):
 - $P(Y) \in [0..1]$
 - $P(Y = S) = 1, P(Y = \emptyset) = 0$
 - $P(Y \cup Z) = P(Y) + P(Z) - P(Y \cap Z)$
 - $P(Y \cap Z) = P(Y)P(Z|Y)$
 - $\sum_{x \in S} P(X = x) = 1$



Continuous Random Variables

- The probability of continuous random variables requires additional tools
 - The probability of a continuous random variable to take on a specific value is often 0 for all (or almost all) possible values
 - Probability has to be defined over ranges of values
 - Individual assignments to random variables have to be addressed using probability densities

$$p(X = x) \in [0.. \infty]$$



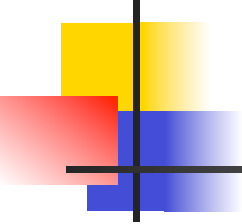
Continuous Random Variables

- Probability density is a measure of the increase in likelihood when adding the corresponding value to the range of values

$$P(a \leq X \leq b) = \int_a^b p(X = x) dx$$

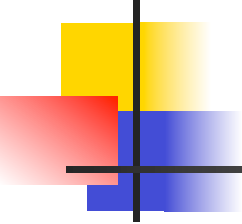
- The probability density is effectively the derivative of the cumulative probability distribution

$$p(X = x) = \frac{dF(x)}{dx}; F(x) = P(-\infty \leq X \leq x)$$



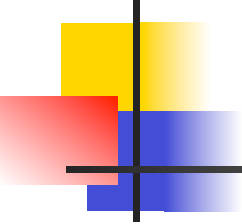
Probability Syntax

- Unconditional or prior probabilities represent the state of knowledge before new observations or evidence
 - $P(H)$
- A probability distribution gives values for all possible assignments to a random variable
- A joint probability distribution gives values for all possible assignments to all random variables



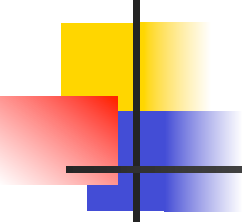
Conditional Probability

- Conditional probabilities represent the probability after certain observations or facts have been considered
 - $P(H/E)$ is the posterior probability of H after evidence E is taken into account
 - Bayes rule allows to derive posterior probabilities from prior probabilities
 - $P(H | E) = P(E | H) P(H)/P(E)$



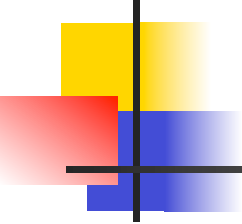
Conditional Probability

- Probability calculations can be conditioned by conditioning all terms
 - Often it is easier to find conditional probabilities
- Conditions can be removed by marginalization
 - $P(H) = \sum_E P(H | E)P(E)$



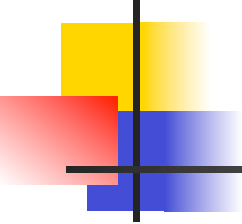
Joint Distributions

- A joint distribution defines the probability values for all possible assignments to all random variables
 - Exponential in the number of random variables
 - Conditional probabilities can be computed from a joint probability distribution
 - $P(A | B) = P(A \cap B) / P(B)$



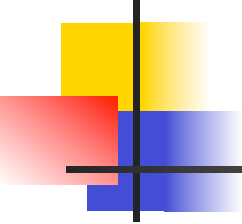
Inference

- Inference in probabilistic representation involves the computation of (conditional) probabilities from the available information
 - Most frequently the computation of a posterior probability $P(H/E)$ from a prior probability $P(H)$ and new evidence E



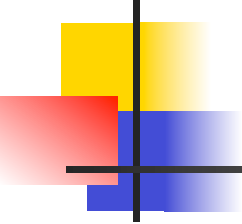
Statistics

- Statistics attempt to represent the important characteristics of a set of data items (or of a probability distribution) and the uncertainty contained in the set (or the distribution).
 - Statistics represent different attributes of the probability distribution represented by the data
 - Statistics are aimed at making it possible to analyze the data based on its important characteristics



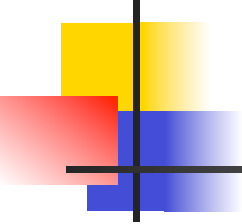
Experiment and Sample Space

- A (random) experiment is a procedure that has a number of possible outcomes and it is not certain which one will occur
- The sample space is the set of all possible outcomes of an experiment (often denoted by S).
 - Examples:
 - Coin : $S = \{H, T\}$
 - Two coins: $S = \{HH, HT, TH, TT\}$
 - Lifetime of a system: $S = \{0..∞\}$



Statistics

- A number of important statistics can be used to characterize a data set (or a population from which the data items are drawn)
 - Mean
 - Median
 - Mode
 - Variance
 - Standard deviation



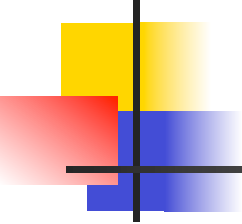
Mean

- The arithmetic mean μ represents the average value of data set $\{X_i\}$

$$\mu = \frac{\sum_{i=1}^N X_i}{N}$$

- The arithmetic mean is the expected value of a random variable, i.e. the expected value of a data item drawn at random from a population

$$\mu = E[X]$$

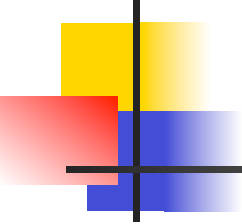


Median and Mode

- The median m is the middle of a distribution

$$|\{X_i \mid X_i \leq m\}| = |\{X_i \mid X_i \geq m\}|$$

- The mode of a distribution is the most frequently (i.e. most likely) value



Variance and Standard Deviation

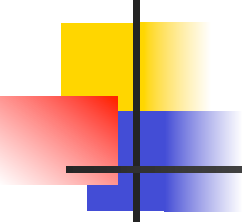
- The variance σ^2 represents the spread of a distribution

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

- In a data set $\{X_i\}$ an unbiased estimate s^2 for the variance can be calculated as

$$s^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N - 1}$$

- $N-1$ is often called the number of degrees of freedom of the data set
- The standard deviation σ is the square root of the variance
 - In the case of a sample set, s is often referred to as standard error



Moments

- Moments are important to characterize distributions
 - r^{th} moment: $E \left[(x - a)^r \right]$
 - Important moments:
 - Mean: $E \left[(x - 0)^1 \right]$
 - Variance: $E \left[(x - \mu)^2 \right]$
 - Skewness: $E \left[(x - \mu)^3 \right]$