

# Chapter I: Introduction & Foundations

## ▶ 1.1 Introduction

- 1.1.1 Definitions & Motivations
- 1.1.2 Data to be Mined
- 1.1.3 Knowledge to be discovered
- 1.1.4 Techniques Utilized
- 1.1.5 Applications Adapted
- 1.1.6 Major Issues in Data Mining

## ▶ 1.2 Getting to Know Your Data

- 1.2.1 Data Objects and Attribute Types
- 1.2.2 Basic Statistical Descriptions of Data
- 1.2.3 Measuring Data Similarity and Dissimilarity

## ▶ 1.3 Basics from Probability Theory and Statistics

- 1.3.1 Probability Theory
- 1.3.2 Statistical Inference: Sampling and Estimation
- 1.3.3 Statistical Inference: Hypothesis Testing

mostly following “Statistical Inference” book by George Caselle et al. Plus other educational web sites

## 1.3.1 Probability Theory

- ▶ Statistics are build upon **probability theory**
- ▶ They provide means for modeling
  - Populations
  - Experiments
  - Anything that can be considered a random phenomenon
- ▶ Statistics are able to draw inferences about populations based on examination of only a part of the whole
  - Suitable for huge datasets
- ▶ The main objective of statisticians is to draw conclusions about a population by conducting an experiment
- ▶ The first step is to identify possible outcomes, in statistical terminology, the sample space.
- ▶ Here we start →

# Sample Space

- ▶ The set,  $S$ , of all possible outcomes of a particular experiment is called the **sample space ( $\Omega$ )** for the experiment

- **Example 1**

- Tossing a coin
- Two possible outcomes
- $\Omega = \{H, T\}$



Head



Tail



- **Example 2**

- Observing the scores of randomly selected students
- Scores are + integers between 200 and 800 that are multiples of 10
- $\Omega = \{200, 210, 220, \dots, 780, 790, 800\}$

- **Example 3**

- Reaction time to a certain stimulus
- $\Omega = \{0, \infty\}$

- ▶ A set  $\Omega$  is countable if its elements can be put into 1-1 correspondence with a subset of integers. Otherwise, it is uncountable

# Event

- ▶ An **event** is any collection of possible outcomes of an experiment, that is any subset of  $S$  (including  $S$  itself)
  - Let  $A$  be an event
  - An event  $A$  occurs if the outcome of the experiment is in the set  $A$
  - In general probabilities are of events rather than sets (the terms **sets** and **events** are both used)
  - **Example**
    - Select a card from a standard deck and note its suit
    - Clubs (**C**), Diamonds (**D**), Hearts (**H**), Spades (**S**)
    - **Sample set:**  $\Omega = \{C, D, H, S\}$       **Events:**  $A = \{C, D\}$  and  $B = \{D, H, S\}$
  - Given any two events (or sets)  $A$  and  $B$ , we have the following elementary operations:
    - $A \cup B = \{x: x \in A \text{ or } x \in B\}$
    - $A \cap B = \{x: x \in A \text{ and } x \in B\}$
    - The complement of  $A$  written  $\neg A$  is the set of all elements that are not in  $\neg A = \{x: x \notin A\}$
  - The set operations of events are commutative, associative, and distributive



# Probability Space

- ▶ A Probability Space is a triple  $(\Omega, \mathcal{A}, P)$  with
  - a sample space  $\Omega$
  - a family  $\mathcal{A}$  of events
  - a probability function (measure)  $P: \mathcal{A} \rightarrow [0,1]$  with
    - $P[\Omega]=1$
    - $P[\cup_i A_i] = \sum_i P[A_i]$  for pairwise disjoint
- ▶ *Notes*
  - Two events  $A$  and  $B$  are *disjoint* if  $A \cap B = \emptyset$
  - The events  $A_1, A_2, \dots$  are *pairwise disjoint* if  $A_i \cap A_j = \emptyset$  for all  $i \neq j$
- ▶ **Properties of P**
  - $P[A] = 1 - P[\neg A]$
  - $P[A \cup B] = P[A] + P[B] - P[A \cap B]$
  - $P[\emptyset] = 0$  (null/impossible event)
  - $P[\Omega] = 1$

# Probability Space

## ▶ Example

- Roll a dice and note the outcome number
- $\Omega = \{1, 2, 3, 4, 5, 6\}$  (all possible outcomes)
- We can define the following events:
  - the outcome of the experiment is **even** →  $A_{\text{even}} = \{2, 4, 6\}$
  - The outcome of the experiment is **odd** →  $A_{\text{odd}} = \{1, 3, 5\}$
- $A = \{A_{\text{even}}, A_{\text{odd}}\}$
- All outcomes has equal probability:  $1/6$
- $P(A_{\text{even}}) = P(2) + P(4) + P(6) = 1/6 + 1/6 + 1/6 = 3/6 = 1/2$
- $P(A_{\text{odd}}) = P(1) + P(3) + P(5) = 1/6 + 1/6 + 1/6 = 3/6 = 1/2$
- Complementary events
  - We consider the event  $A = \{1, 3\}$
  - Complementary event to  $A$  is  $\neg A = \{2, 4, 5, 6\}$
  - $P(A) = P(1) + P(3) = 1/6 + 1/6 = 2/6 = 1/3$
  - $P(\neg A) = 1 - P(A) = 2/3$
- Counting is a way for computing probabilities

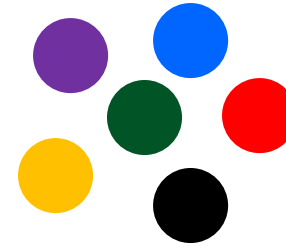
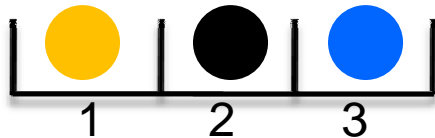


# Counting & Enumerating Outcomes

- ▶ Probabilities of events can be calculated by simply **counting** the number of outcomes in the event
- ▶ Suppose a sample space  $\Omega = \{s_1, \dots, s_N\}$
- ▶ If all outcomes are equally likely  $\rightarrow P(s_i) = 1/N$
- ▶ For any event A
  - $\rightarrow P(A) = \sum_{s_i \in A} P(s_i) = \sum_{s_i \in A} 1/N = (\# \text{ elements in } A) / (\# \text{ elements in } \Omega)$
- ▶ An easy way to count is to break the problem to small tasks
- ▶ There are 4 different ways of counting
  - $\rightarrow$  Ordered, without replacement
  - $\rightarrow$  Ordered, with replacement
  - $\rightarrow$  Unordered, without replacement
  - $\rightarrow$  Unordered, with replacement
- ▶ We will analyze these possibilities by example

# Counting & Enumerating Outcomes

## ▶ Ordered, without replacement



## ▶ In **how many** different **ordered ways** can we put these **6** colors in the **3** places shown above?

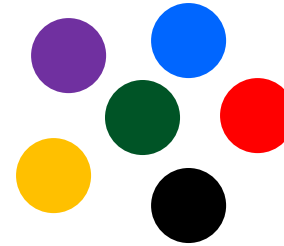
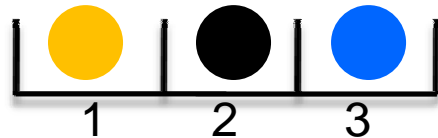
- The first color can be chosen from **6** different colors
- The second color can be chosen from **5** different colors
- The third color can be chosen from **4** different colors
- Possible arrangements =  $6 \times 5 \times 4 = 120$   
 $= (6 \times 5 \times 4 \times 3 \times 2 \times 1) / (3 \times 2 \times 1)$   
 $= 6! / 3!$

## ▶ **Generalization**

- **n** the number of objects
- **r** the arrangement size
- # possible arrangements =  $(n!) / (n-r)!$

# Counting & Enumerating Outcomes

## ▶ Ordered, with replacement



## ▶ In **how many** different **ordered ways** can we put these **6** colors in the **3** places shown above?

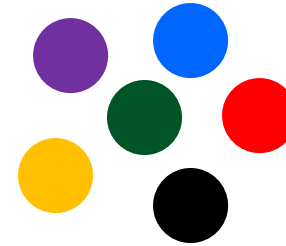
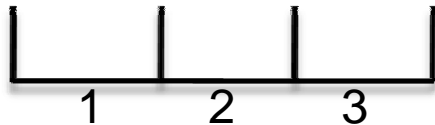
- The first color can be chosen from **6** different colors
- The second color can be chosen from **6** different colors
- The third color can be chosen from **6** different colors
- Possible arrangements =  $6 \times 6 \times 6 = 216$

## ▶ **Generalization**

- **n** the number of objects
- **r** the arrangement size
- # possible arrangements =  $n^r$

# Counting & Enumerating Outcomes

## ▶ Unordered, without replacement



▶ In **how many** different **unordered ways** can we put these **6** colors in the **3** places shown above?

→ Order is not important → remove redundant ordering

→ From the *Fundamental Theorem of counting* **3** numbers can be arranged in **3!** ways.

→ We know the number of arrangement in the ordered case **6!/3!**

→ We divide by the number of redundant ordering: **6!/(3! 3!) = 20**

## ▶ Generalization

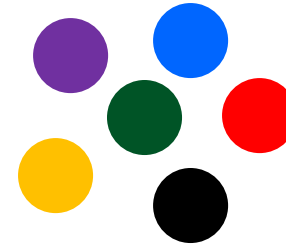
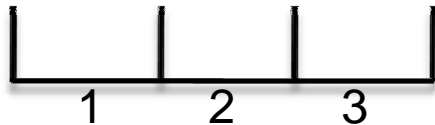
→ **n** the number of objects

→ **r** the arrangement size

→ # possible arrangements =  $n!/r!(n-r)! = \binom{n}{r}$

# Counting & Enumerating Outcomes

## ▶ Unordered, with replacement



▶ In **how many** different **unordered ways** can we put these **6** colors in the **3** places shown above?

- The most difficult case
- Possible arrangements =  $8!/3!5! = 56$
- See the reasoning behind in (reference)

## ▶ Generalization

- **n** the number of objects
- **r** the arrangement size
- # possible arrangements =  $(n + r - 1)!/r!(n-1)! = \binom{n + r - 1}{r}$

# Counting & Enumerating Outcomes

## ▶ Example of using counting and enumeration for computing probabilities

→ Data set  $S = \{2, 4, 9, 12\}$

### → Experiment

- Select all possible averages of four numbers from  $S$
- Numbers are drawn with replacement
- Assume that the order is important
- Total number of ordered outcomes :  $n^r = 4^4 = 256$

→ **Question:** what is the probability of the average **4.25** to occurs?

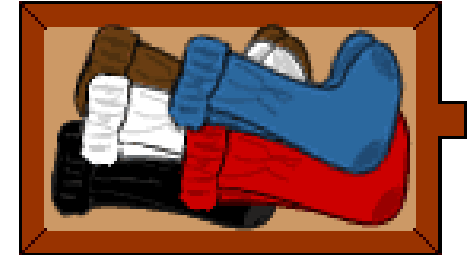
### → Answer

- The average **4.25** occurs when the outcome contains one **2**, two **4** and one **9**
- Compute the probability of drawing the ordered outcome  **$\{2, 4, 4, 9\}$**
- **12 Possible outcomes**  
(2,4,4,9), (2,4,9,4), (2,9,4,4), (4,2,4,9), (4,2,9,4), (4,4,2,9)  
(4,4,9,2), (4,9,2,4), (4,9,4,2), (9,2,4,4), (9,4,2,4), (9,4,4,2)

→ **The probability of drawing the unordered sample  $\{2, 4, 4, 9\}$  is  $12/256$**

# Independence

- ▶ Two events **A**, **B** in a probability space are **independent** if the fact that **A** occurs does not affect the probability of **B** occurring
  - choose a pair of socks without looking
  - The first pair you pull out is red --the wrong color
  - You replace this pair
  - Choose another pair of socks
  - What is the probability of choosing the red pair twice?
  - **Important**
    - Since the first pair was replaced, choosing a red pair on the first try has no effect on the probability of choosing a red pair on the second try
    - These two events are independent



## ▶ Formally

- Two events **A**, **B** in a probability space are **independent** if

$$P(A \cap B) = P(A) P(B)$$

- A finite set of events  $A = \{A_1, \dots, A_n\}$  is independent if for every subset  $S \subseteq A$ , the equation  $P(\bigcap_{A_i \in S} A_i) = \prod_{A_i \in S} P(A_i)$  holds

# Independence

## ▶ Example

→ Rolling two dice

→  $\Omega = \{ (1,1), (1,2), \dots, (6,5), (6,6) \}$  36 ordered pairs

→ Assume the following events

- $A = \{ \text{doubles appear} \} = \{ (1,1), (2,2), (3,3), (4,4), (5,5), (6,6) \}$  (6 possibilities)
- $B = \{ \text{the sum is between 7 and 10} \}$  (18 possibilities)
- $C = \{ \text{the sum is 2 or 7 or 8} \}$  (12 possibilities)

→ By counting among the 36 cases we obtain

$$P(A) = 1/6 \quad P(B) = 1/2 \quad P(C) = 1/3$$

→  $P(A \cap B \cap C) = P(\text{the sum is 8, composed of double 4s})$  [1 possibility]

$$= 1/36$$

$$= 1/6 \times 1/2 \times 1/3$$

$$= P(A)P(B)P(C) \Rightarrow A, B, \text{ and } C \text{ are independent}$$



# Conditional Probability

## ▶ Conditional Probability

- The probability of some event **A (hypothesis)**, given the occurrence of some other event **B (evidence)**
- Written  $P(A | B)$

## ▶ Example

- Rolling two dice
- **Hypothesis A**: the sum is greater than 8
- **Evidence B**: first dice is a 6
- What is the probability of **A** given **B** ?
- **With Evidence B** there are **6** outcomes for which the first die is a 6  
 $\{(6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\}$
- **Total > 8 and Dice 1 = 6**:  $\{(6,3), (6,4), (6,5), (6,6)\}$  4 pairs

$$P(A | B) = 4/6 = 2/3$$

## ▶ Formally

- $P(A | B) = P(A \cap B) / P(B)$
- Event **A** is **conditionally independent** from **B** if  $P(A | BC) = P(A | C)$



# Bayes' Theorem

## Intuition

- ▶ Provide a mathematical rule explaining how you should change your existing beliefs in the light of new evidence
- ▶ **Example**
  - Without looking at the sky we ask you what you think about it.
  - You believe that it is completely cloudy
  - Let's call your belief **hypothesis A**
    - **A**: The sky is completely cloudy
    - **P(A)**: the probability of the sky being completely cloudy without having any observation. It is called **prior** probability
  - You observe that it is raining. We call this observation **evidence B**
  - Now you would like to compute the probability of **hypothesis A** given evidence **B**

# Bayes' Theorem

$P(\text{sky is completely cloudy} \mid \text{it rains})$  can be computed using the **likelihood** and the **prior** probability.

- **Likelihood** that it rains given that the sky is completely cloudy
- **Prior** probability that the sky is completely cloudy
- You need to normalize considering the likelihood that it rains given all other situations of the sky.
- Consider that the sky can be:
  - completely cloudy
  - partially cloudy
  - bleu
- **Marginal probability**: is the unconditional probability of the rain, regardless of the situation of the sky:

Likelihood that it rains given that it is completely cloudy + Likelihood that it rains given that the sky is partially cloudy + Likelihood that it rains given that the sky is bleu.

$$P(A \mid B) = \text{prior} \times \text{likelihood} / \text{marginal probability}$$

# Bayes' Theorem

## Formally

- ▶ **A** is the **hypothesis** that we are interested in testing
- ▶ **B** represents a new piece of **evidence** that seems to confirm or disconfirm the hypothesis

$$P(A | B) = \underbrace{P(B | A)}_{\text{likelihood}} \underbrace{P(A)}_{\text{prior}} / \underbrace{P(B)}_{\text{marginal probability}}$$

- ▶  $P(A | B)$  is called **posterior probability**
- ▶  $P(A)$  is called **prior probability**
- ▶ **Total probability theorem**
  - For a partitioning of  $\Omega$  into events  $B_1, \dots, B_n$ :

$$P(A) = \sum_{i=1, n} P(A | B_i) P(B_i)$$

# Bayes' Theorem

- ▶ **Example (from wikipedia)**

- ▶ Observe a student
- ▶ All an observer can see is that this student is wearing trousers.

	Girls	Boys	Total
Trousers	20	60	80
Skirts	20	0	20
Total	40	60	100

- ▶ What is the probability the observed student is a girl?
- ▶ **Hypothesis A:** the student observed is a girl
- ▶ **Evidence B:** The student observed is wearing trousers
  - $P(A)=0.4$  (40% of students are girls)
  - $P(B | A)=0.5$  (50% wearing skirts and 50% wearing trousers)
  - Using the total probability theorem we compute

$$P(B) = P(B | A)P(A) + P(B | \neg A)P(\neg A),$$
$$= 0.5 \times 0.4 + 1 \times 0.6 = 0.8$$

- ▶  $P(A | B) = \frac{P(B | A) P(A)}{P(B)} = \frac{0.5 \times 0.4}{0.8}$   
 $= 0.25$

# Random Variables

## Why do we need random variables?

- ▶ In many experiments, it is easier to deal with a summary of variable than with the original probability structure
- ▶ **Example**
  - Ask **50** people about an issue
  - Record "**1**" for agree and "**0**" for disagree
  - The sample space has  $2^{50}$  elements, each an ordered string of 1s and 0s of length 50
  - **Goal**: reduce the space size
  - **Solution**: look for the quantity of interest that captures the essence of the problem
    - In this example, what is important is the number of people who agree (equivalently, disagree)
    - Define a variable  **$X = \text{number of 1s}$**
    - The sample space for  $X$  is  $\{0, 1, \dots, 50\}$
    - It is much easier to deal with this space than the original one

# Random Variables

- ▶ A random variable on the probability space  $(\Omega, \mathcal{A}, \mathbb{P})$  is a function that maps  $\Omega$  into real numbers:

$$X: \Omega \rightarrow \mathbb{M} \text{ with } \mathbb{M} \subseteq \mathbb{R}$$

- ▶ **Examples**

Experiment

Toss two dice

Toss a coin 25 times

Random variable

$X =$  sum of the numbers

$X =$  number of heads in 25 tosses

- ▶ The **cumulative distribution function (cdf)** of a random variable  $X$  is defined by

$$F_X(x) = P(X < x), \text{ for all } x$$

- ▶ The **probability density function (pdf)** of a random variable  $X$  is given by

$$f_X(x) = P_X(X = x), \text{ for all } x$$

# Cdf and Pdf

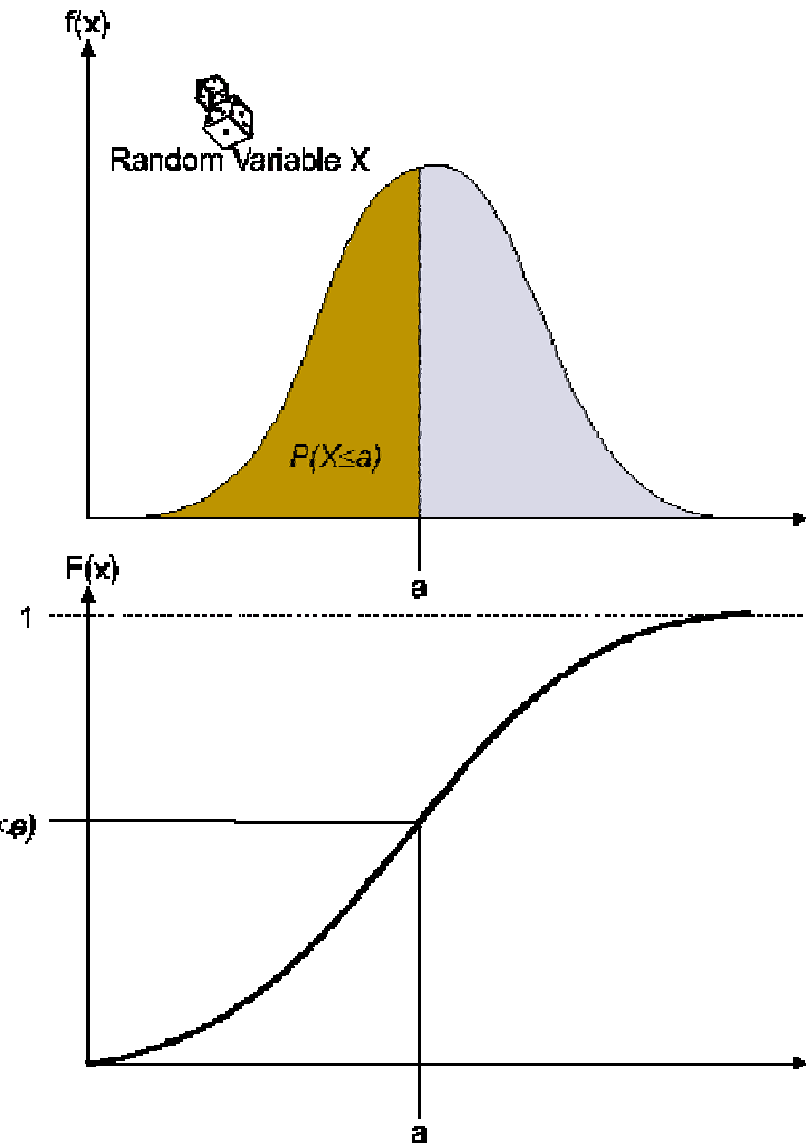
- ▶ The probability density function (**pdf**) is concerned with *"point probabilities"*
- ▶ The cumulative distribution function (**cdf**) can be obtained by

→ Summing up the values of the pdf (in the discrete case)

$$F_X(x) = \sum_{x_i \leq x} f(x_i)$$

→ Computing the integral of the pdf in the continuous case

$$F_X(X) = \int_{-\infty}^x f_X(t) dt$$



# Important Discrete Distributions

- ▶ **Bernoulli** distribution with parameter  $p$ : (tossing a coin)

$$P(X = k) = p^k (1-p)^{1-k}, \text{ for } k \in \{0,1\}$$

- ▶ **Uniform** distribution over  $\{1,2,\dots,m\}$ :

$$P(X = k) = f_X(k) = \frac{1}{m}, \text{ for } 1 \leq k \leq m$$

- ▶ **Binomial** distribution (coin toss  $n$  times repeated;  $X$ :# heads)

$$P(X = k) = f_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- ▶ **Poisson** distribution (with rate  $\lambda$ )

$$P(X = k) = f_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

- ▶ **Geometric** distribution (# coin tosses until first head)

$$P(X = k) = f_X(k) = (1-p)^k p$$

- ▶ **2-Poisson mixture** (with  $a_1 + a_2 = 1$ )

$$P(X = k) = f_X(k) = a_1 e^{-\lambda_1} \frac{\lambda_1^k}{k!} + a_2 e^{-\lambda_2} \frac{\lambda_2^k}{k!}$$

# Important Continuous Distributions

- ▶ **Uniform** distribution in the interval  $[a,b]$

$$f_X(x) = \frac{1}{b-a}, \text{ for } a \leq x \leq b \text{ (0 otherwise)}$$

- ▶ **Exponential** distribution with rate  $\lambda$

$$f_X(x) = \lambda e^{-\lambda x}, \text{ for } x \geq 0 \text{ (0 otherwise)}$$

- ▶ **Hyper-exponential** distribution

$$f_X(x) = p\lambda_1 e^{-\lambda_1 x} + (1-p)\lambda_2 e^{-\lambda_2 x}, \text{ for } x \geq 0 \text{ (0 otherwise)}$$

- ▶ **Pareto** distribution

$$f_X(x) \rightarrow = \frac{a}{b} \left( \frac{b}{x} \right)^{a+1} \text{ for } x > b, 0 \text{ otherwise}$$

- ▶ **Logistic** distribution

$$F_X(x) = \frac{1}{1-e^{-x}}$$

# Normal Distributions

## ▶ Normal distributions are

- a family of distributions that have the same general shape
- are symmetric with scores more concentrated in the middle than in the tails

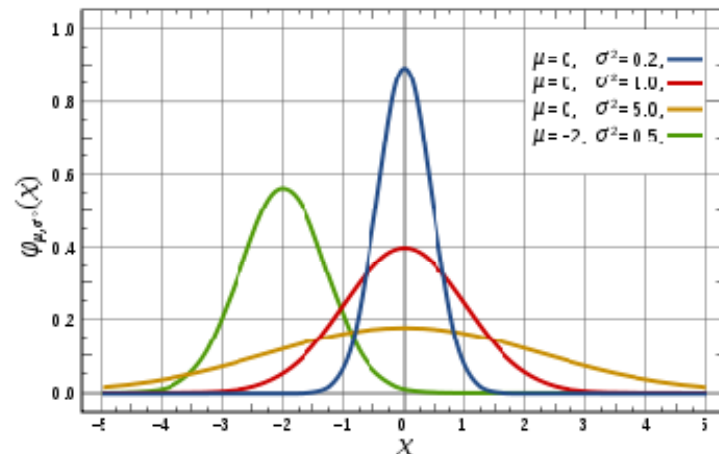
## ▶ Why they are important?

- many psychological and educational variables are distributed approximately normally
- easy for mathematical statisticians to work with

## ▶ Formally

- Normal distribution  $N(\mu, \sigma^2)$  approximates sums of independent identically distributed random variables

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$$



# Moments

- ▶ **Moments** describe the nature (and the shape) of a distribution
- ▶ The **n<sup>th</sup>** moment about a point **a** is given by:

$$\mu_n(a) = \sum_{x \in M} (x - a)^n f_X(x)$$

in the discrete case

$$\mu_n(a) = \int_{-\infty}^{+\infty} (x - a)^n f_X(x) dx$$

in the continuous case

- ▶ The first moment about zero is the **Mean (a=0, n=1)**  
(Also called Expectation value E[X])

$$\text{Mean} = \mu = \sum_{x \in M} x f_X(x)$$

$$\text{Mean} = \mu = \int_{-\infty}^{+\infty} x f_X(x) dx$$

- ▶ The second moment about the mean is the **Variance (a=μ, n=2)**

$$\text{Variance} = \sigma^2 = \sum_{x \in M} (x - \mu)^2 f_X(x)$$

$$\text{Variance} = \sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f_X(x) dx$$

# Moments

- ▶ The third moment about the mean is used to measure the symmetry of the shape of a distribution: **Skewness** ( $\mu_3/\sigma^3$ )

$$\text{Skewness} = \frac{\sum_{x \in M} (x - \mu)^3 f_X(x)}{\sigma^3}$$

$$\text{Skewness} = \frac{\int_{-\infty}^{+\infty} (x - \mu)^3 f_X(x) dx}{\sigma^3}$$

→ Skewness=0  $\Rightarrow$  symmetric, Skewness > 0  $\Rightarrow$  positively skewed,  
Skewness < 0  $\Rightarrow$  negatively skewed

- ▶ The fourth moment about the mean is used to describe the **Kurtosis** ( $\mu_4/\sigma^4$ )
  - measure of the flatness or peakedness of a distribution

$$\text{Kurtosis} = \frac{\sum_{x \in M} (x - \mu)^4 f_X(x)}{\sigma^4}$$

$$\text{Kurtosis} = \frac{\int_{-\infty}^{+\infty} (x - \mu)^4 f_X(x) dx}{\sigma^4}$$

# Tail bounds

- ▶ When a variable deviates far from its mean we need to bound its probability.
- ▶ **Tail bounds** are used for untraceable distributions

→ **Markov inequality**

$$P(X \geq t) \leq \frac{E(X)}{t}, \text{ for } t > 0 \text{ and non negative } X$$

→ **Chebyshev inequality**

$$P(|X - E[X]| \geq t) \leq \frac{\sigma^2}{t^2}, \text{ for } t > 0 \text{ and non negative } X$$

→ **Mill's inequality**

$$P(|Z| > t) \leq \frac{\sqrt{2}}{\pi} \frac{e^{-t^2/2}}{t}, \text{ for } Z \sim N(0,1) \text{ and } t > 0$$

→ **Cauchy-Schwarz inequality**

$$E[XY] \leq \sqrt{E[X^2]E[Y^2]}$$

## Tail bounds (Example)

- ▶ Consider
  - 3 n-faceted dice (1/n probability for each outcome)
  - 1 m-faceted dice (1/m probability for each outcome)
- ▶ Derive an upper bound for the probability that the sum of all four dice exceed  $t$  ( $4 < t < 3n+m$ ) using the Markov inequalities
- ▶ **Solution**
  - N**= the sum of k tosses of 3 n-faceted dice
  - M**= the sum of k tosses of one m-faceted dice
  - S=N+M**

$$E(S) = E(M + N) = \frac{3}{2}(n+1) + \frac{1}{2}(m+1)$$

$$P(S \geq t) \leq \frac{E(S)}{t} = \frac{\frac{3}{2}(n+1) + \frac{1}{2}(m+1)}{t}$$

## Summary of Section 1.3.1

- ▶ Probabilities and statistics provide means for modeling populations and experiments
- ▶ Bayes' Theorem very simple and very powerful
- ▶ Rich variety of well studied distribution functions
- ▶ Moments capture distributions
- ▶ Tail bounds useful for intractable distributions

## 1.3.2 S.Inf.: Sampling and Estimation

- ▶ A **statistical model** is a set of distributions (or regression functions)
- ▶ **Statistical inference**: given a sample  $X_1, \dots, X_n$  how do we infer the distribution or its parameters within a given model
- ▶ Statistical inference includes:
  - **Estimation**
    - Non parametric estimation
    - Parametric estimation
  - **Hypothesis Testing**
  - **Prediction** or **regression**

# Non-Parametric Estimation

- ▶ In **non-parametric** models
  - The distribution is not specified a **priori** but it is determined from the data
  - Non-parametric **does not refer to the absence of parameters**. It refers to the fact that the parameters are flexible and not fixed in advance
- ▶ **Some Methods**
  - A **histogram** is a non parametric estimate of a probability distribution
  - **Empirical distribution function** is a step function such that the height of each step is  $1/n$  (for the  $n$  numbers of the sample)
  - A **statistical functional**  $T(F)$  is any function of  $F$  (mean, variance, skewness, median, quantiles, etc.)
  - **Kernel density estimation** makes it possible to extrapolate the data to the entire population

# Parametric Estimation

- ▶ A **parametric** model
  - is a set of distributions that is completely described by a finite number of parameters (e.g., the family of Normal distributions)
  - assume data come from a type of probability distribution and makes inferences about the parameters of the distribution
- ▶ **Example**
  - Assume the data follows a normal distribution  $N(\mu, \sigma^2)$
  - Estimate the parameters  $\mu$  and  $\sigma^2$  given a sample  $X_1, \dots, X_n$
- ▶ **Some Methods**
  - Method of **moments**
  - **Maximum likelihood** estimator
  - **Expectation Maximization** method

# Statistical Estimators

- ▶ The estimate of a parameter can be
  - a single number: **Point Estimate**
  - a range of scores: **Interval estimate**
    - use **confidence intervals**
- ▶ Example of point estimate
  - you would like to estimate the average age of students in Bozen
  - you take a random sample of students
  - the sample mean would be a **point estimate** of the population mean
- ▶ Point estimates are **important** because many statistical formula are based on them
  - interval confidence
  - Significance testing
- ▶ **Formally**: A **point estimator** for a parameter  $\theta$  of a probability distribution is a random variable  $X$  derived from a random sample  $X_1, \dots, X_n$

# Characteristics of Estimators

- ▶ An estimator  $T$  for parameter  $\theta$  is **unbiased**
  - if  $E[T] = \theta$
  - Otherwise the estimator has a bias  $E[T] - \theta$
- ▶ An estimator on a sample of size  $n$  is **consistent**
  - if the estimator tends to get closer to the parameter it is estimating as the sample size increases
  - **Formally**: if  $\lim_{n \rightarrow \infty} P[|T - \theta| < \varepsilon] = 1$  for each  $\varepsilon > 0$

## ▶ Examples

→ Assume a random sample  $X_1, \dots, X_n$

→ Sample mean  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

→ Sample variance  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

Sample mean and sample variance are **unbiased**, **consistent** estimators with **minimal variance**

# Estimator Error

- ▶ Let  $\hat{\theta}_n = T(\theta)$  be an estimator for parameter  $\theta$  over sample  $X_1, \dots, X_n$   
The distribution of  $\hat{\theta}_n$  is called the **sampling distribution**

- ▶ The **standard error for  $\hat{\theta}_n$**  is:

$$se(\hat{\theta}) = \sqrt{\text{Var}[\hat{\theta}]}$$

- ▶ The **mean squared error (MSE)** for  $\hat{\theta}_n$  is

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= E[(\hat{\theta}_n - \theta)^2] \\ &= \text{bias}^2(\hat{\theta}_n) + \text{Var}[\hat{\theta}_n] \end{aligned}$$



If  $\text{bias} \rightarrow 0$  and  $se \rightarrow 0$  then the estimator is consistent

- ▶ The estimator  $\hat{\theta}_n$  is **asymptotically Normal** if  
 $(\hat{\theta}_n - \theta)/se$  converges in distribution to standard Normal(0,1)

# Method of Moments

- ▶ Consider a distribution  $f(x | \theta_1, \dots, \theta_k)$  that depends on  $k$  parameters
- ▶ We have the first  $k$  moments  $m_1, m_2, \dots, m_k$  of the sample

$$m_i = \frac{1}{n} \sum_{j=1}^n (x_j)^i$$

- ▶ We have the first  $k$  moments  $\mu_1, \mu_2, \dots, \mu_k$  of the distribution

$$u_i = E[X^i]$$

Note that these moments are functions of the parameters  $(\theta_1, \dots, \theta_k)$  of the distribution

- ▶ Solve a set of  $k$  equations  $\{m_i = u_i\}$  to estimate parameters  $\{\theta_i\}$
- ▶ Method-of-moment estimators are usually **consistent** and **asymptotically Normal**, but may be **biased**

# Maximum Likelihood Estimators (MLE)

- ▶ Estimate parameter  $\theta$  of  $f(\theta, x)$  such that the probability that the data of the sample are generated by this distribution is maximized

- ▶ **Maximum Likelihood estimation**

Maximize  $L(x_1, \dots, x_n, \theta) = P[x_1, \dots, x_n \text{ originates from } f(\theta, x)]$   
or maximize  $\log L$

often written as:

$$L(\theta \mid x_1, \dots, x_n) = f(x_1, \dots, x_n \mid \theta)$$

- ▶ Maximum Likelihood estimators are **consistent** and **asymptotically Normal**

# MLE Example

▶ Given

- coin with Binomial distribution with unknown parameter  $p$  for head and  $1-p$  for tail
- Sample (data):  $k$  times head with  $n$  coin tosses needed: maximum likelihood estimation of  $p$

Let  $L(k, n, p) = P[\text{sample is generated from distr. with parameter } p]$

$$= \binom{n}{k} p^k (1-p)^{n-k}$$

Maximize log - likelihood function  $\log L(k, n, p)$

$$\log L = \log \binom{n}{k} + k \log p + (n-k) \log(1-p)$$

$$\frac{\partial \log L}{\partial p} = \frac{k}{p} - \frac{n-k}{1-p} = 0 \Rightarrow p = \frac{k}{n}$$

## Summary of Section 1.3.2

- ▶ **Samples** and **Estimators** are Random Variables
- ▶ Estimators should be **unbiased**
- ▶ **MLE** is canonical estimator for parameters

## 1.3.3 S.Inf: Hypothesis Testing

### James Bond Tasting Martini

- ▶ We gave to Mr. Bond a series of 16 taste tests
- ▶ At each test, we flipped a fair coin to determine whether to stir or shake the martini
- ▶ Mr. Bond was correct on 13/16 tests:
  - Is Mr. Bond able to distinguish between stirred and shaken martini?
  - Was he just lucky?



- ▶ Using a binomial distribution ( $N=16$ ,  $k=13$ ,  $p=0.5$ )
  - $P(\text{someone who is lucky would be correct } 13/16 \text{ or more}) = 0.0106$
- ▶ The hypothesis that Mr. Bond was lucky is not proven false (but considerable doubt is cast on it)
- ▶ There is a strong evidence that Mr. Bond can tell whether a drink was shaken or stirred

# Probability Value (p-value)

- ▶ In the James Bond example, the computed probability of **0.0106** is the probability he would be correct on 13 or more taste tests (out of 16) if he was just guessing
- ▶ **Important**
  - This is not the probability that he cannot tell the difference
  - The probability of **0.016** is the probability of a **certain outcome** (13 or more out of 16) assuming a **certain state of the world** (James Bond was only guessing.)
- ▶ Using statistical terminology, the **probability value** is the probability of an outcome given the hypothesis. It is not the probability of the hypothesis given the outcome.

# Null Hypothesis & Statistical Significance

- ▶ In the previous example, the hypothesis that an effect is due to chance is called the **null hypothesis**
- ▶ The null hypothesis is typically the opposite of the researcher's hypothesis
- ▶ The null hypothesis is rejected when the probability value is lower than a specific threshold (0.05 or 0.01) called **test level**
- ▶ When the null hypothesis is rejected, the effect is said to be **statistically significant**

# Statistical Hypothesis Testing

- ▶ A hypothesis test determines a probability  $1-\alpha$  (**test level  $\alpha$ , significance level**) that a sample  $X_1, \dots, X_n$  from some unknown probability distribution has a certain property
- ▶ Examples
  - under the assumption of a normal distribution
  - Two random variables are independent

## ▶ General Form

### Null hypothesis $H_0$ vs, alternative hypothesis $H_1$

Needs **test variable  $X$**  (derived from  $X_1, \dots, X_n, H_0, H_1$ ) and

Test region  $R$  with

$X \in R$  for rejecting  $H_0$  and

$X \notin R$  for retaining  $H_0$

	Retain $H_0$	Reject $H_0$
$H_0$ true	✓	Type I error
$H_1$ true	Type II error	✓

# Hypothesis and p-Values

## ▶ Hypothesis

A hypothesis of the form  $\theta = \theta_0$  is called a **simple hypothesis**

A hypothesis of the form  $\theta > \theta_0$  or  $\theta < \theta_0$  is called a **composite hypothesis**

## ▶ Tests

A test of the form  $H_0: \theta = \theta_0$  vs.  $H_1: \theta \neq \theta_0$  is called a **two-sided test**

A test of the form  $H_0: \theta \leq \theta_0$  vs.  $H_1: \theta > \theta_0$  or  $H_0: \theta \geq \theta_0$  vs.  $H_1: \theta < \theta_0$  is called a **one-sided test**

## ▶ P-value

Small p-value means strong evidence against  $H_0$

## Summary of Section 1.3.3

- ▶ **Hypothesis testing:** reject or retain  $H_0$  at level  $\alpha$
- ▶ **P-value:** smallest level  $\alpha$  for rejecting  $H_0$