CZ.1.07/2.2.00/28.0041
Centrum interaktivních a multimediálních studijních opor pro inovaci výuky a efektivní učení

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Prologue

You should spent most of your time thinking about

what you should think about most of your time.

RANDOMIZED ALGORITHMS AND PROTOCOLS - 2020

RANDOMIZED ALGORITHMS AND PROTOCOLS - 2020

Prof. Jozef Gruska, DrSc
Wednesday, 10.00-11.40, B410

WEB PAGE of the LECTURE

http://www.fi.muni.cz/usr/gruska/random20

FINAL EXAM: You need to answer four questions out of five given to you.
CREDIT (ZAPOČET): You need to answer three questions out of five given to you.

EXERCISES/TUTORIALS

EXERCISES/TUTORIALS: Thursdays 14.00-15.40, C525

TEACHER: RNDr. Matej Pivoluška PhD

Language English

NOTE: Exercises/tutorials are not obligatory

CONTENTS - preliminary

LITERATURE

- R. Motwami, P. Raghavan: Randomized algorithms, Cambridge University Press, UK, 1995
- J. Gruska: Foundations of computing, International Thompson Computer Press, USA. 715 pages, 1997
- J. Hromkovič: Design and analysis of randomized algorithms, Springer, 275 pages, 2005
- N. Alon, J. H. Spencer: The probabilistic method, Willey-Interscience, 2008

Part I

Basics of Probability Theory

CHAPTER 3:
BASICS of PROBABILITY THEORY

PROBABILITY INTUITIVELY

Intuitively, **probability of an event** E is the ratio between the number of favorable elementary events involved in E to the number of all possible elementary events involved in E.

$$Pr(E) = \frac{\text{number of favorable elementary events involved in } E}{\text{number of all possible elementary events involved in } E}$$

Example: Probability that when tossing a perfect 6-sided dice we get a number divided by 3 is

$$2/6 = 1/3$$

.

Key fact: Any probabilistic statement must refer to a specific underlying probability space - a space of elements to which a probability is assigned.

PROBABILITY SPACES

A **probability space** is defined in terms of a **sample space** Ω (often with an algebraic structure – for example outcomes of some cube tossing) and a **probability measure** (*probability distribution*) defined on Ω.

Subsets of a sample space Ω are called **events**. Elements of Ω are referred to as **elementary events**.

Intuitively, the sample space represents the set of all possible outcomes of a **probabilistic experiment** – for example of a cube tossing. An event represents a collection (a subset) of possible outcomes.

Intuitively - again, probability of an event E is the ratio between the number of favorable elementary events involved in E and the number of all possible elementary events.

PROBABILITY THEORY

Probability theory took almost 300 years to develop

from intuitive ideas of Pascal, Fermat and Huygens, around 1650,

to the currently acceptable axiomatic definition of probability (due to A. N. Kolmogorov in 1933).

AXIOMATIC APPROACH - I.

Axiomatic approach: Probability distribution on a set Ω is every function $\Pr : 2^\Omega \to [0,1]$, satisfying the following axioms (of Kolmogorov):

1. $\Pr(\{x\}) \geq 0$ for any element (elementary event) $x \in \Omega$;
2. $\Pr(\Omega) = 1$
3. $\Pr(A \cup B) = \Pr(A) + \Pr(B)$ if $A, B \subseteq \Omega$, $A \cap B = \emptyset$.

Example: **Probabilistic experiment** – cube tossing; **elementary events** – outcomes of cube tossing; **probability distribution** – $\{p_1, p_2, p_3, p_4, p_5, p_6\}$, $\sum_{i=1}^{6} p_i = 1$, where p_i is probability that i is the outcome of a particular cube tossing.

In general, a sample space is an arbitrary set. However, often we need (wish) to consider only some (family) of all possible events of 2^Ω.

The fact that not all collections of events lead to well-defined probability spaces leads to the concepts presented on the next slide.

AXIOMATIC APPROACH - II.

Definition: A **σ-field** $(\Omega, \mathbf{F})$ consists of a sample space Ω and a collection $\mathbf{F}$ of subsets of Ω satisfying the following conditions:

1. $\emptyset \in \mathbf{F}$
2. $\varepsilon \in \mathbf{F} \Rightarrow \overline{\varepsilon} \in \mathbf{F}$
3. $\varepsilon_1, \varepsilon_2, \ldots \in \mathbf{F} \Rightarrow (\varepsilon_1 \cup \varepsilon_2 \cup \ldots) \in \mathbf{F}$

Consequence

A σ-field is closed under countable unions and intersections.

Definition: A **probability measure** (*distribution*) $Pr : \mathbf{F} \to \mathbf{R}^{\geq 0}$ on a σ-field $(\Omega,\mathbf{F})$ is a function satisfying conditions:

1. If $\varepsilon \in \mathbf{F}$, then $0 \leq Pr(\varepsilon) \leq 1$.
2. $Pr[\Omega] = 1$.
3. For mutually disjoint events $\varepsilon_1, \varepsilon_2, \ldots$ $Pr\left[\bigcup_i \varepsilon_i\right] = \sum_i Pr(\varepsilon_i)$

Definition: A **probability space** $(\Omega,\mathbf{F},Pr)$ consists of a σ-field $(\Omega,\mathbf{F})$ with a probability measure Pr defined on $(\Omega, \mathbf{F})$.

PROBABILITIES and their PROPERTIES - I.

Properties (for arbitrary events ε_i):

$$\begin{aligned} Pr(\overline{\varepsilon}) &= 1 - Pr(\varepsilon); \\ Pr(\varepsilon_1 \cup \varepsilon_2) &= Pr(\varepsilon_1) + Pr(\varepsilon_2) - Pr(\varepsilon_1 \cap \varepsilon_2); \\ Pr(\bigcup_{i \geq 1} \varepsilon_i) &\leq \sum_{i \geq 1} Pr(\varepsilon_i). \end{aligned}$$

Definition: Conditional probability of an event ε_1 given an event ε_2 is defined by

$$Pr[\varepsilon_1 | \varepsilon_2] = \frac{Pr[\varepsilon_1 \cap \varepsilon_2]}{Pr[\varepsilon_2]}$$

if $Pr[\varepsilon_2] > 0$.

Theorem: Law of the total probability Let $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k$ be a **partition** of a sample space Ω. Then for any event ε

$$Pr[\varepsilon] = \sum_{i=1}^{k} Pr[\varepsilon | \varepsilon_i] \cdot Pr[\varepsilon_i]$$

EXAMPLE:

Let us consider tossing of two perfect dices with sides labelled by 1, 2, 3, 4, 5, 6. Let

ε_1 be the event that the reminder at the division of the sum of the outcomes of both dices when divided by 4 is 3, and

ε_2 be the event that the outcome of the first cube is 4.
In such a case

$$Pr[\varepsilon_1 | \varepsilon_2] = \frac{Pr[\varepsilon_1 \cap \varepsilon_2]}{Pr[\varepsilon_2]} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}$$

PROBABILITIES and their PROPERTIES - II.

Theorem: (Bayes' Rule/Law)

(a) $Pr(\varepsilon_1) \cdot Pr(\varepsilon_2|\varepsilon_1) = Pr(\varepsilon_2) \cdot Pr(\varepsilon_1|\varepsilon_2)$ basic equality

(b) $Pr(\varepsilon_2|\varepsilon_1) = \frac{Pr(\varepsilon_2)\, Pr(\varepsilon_1|\varepsilon_2)}{Pr(\varepsilon_1)}$ simple version

(c) $Pr[\varepsilon_0|\varepsilon] = \frac{Pr[\varepsilon_0 \cap \varepsilon]}{Pr[\varepsilon]} = \frac{Pr[\varepsilon|\varepsilon_0] \cdot Pr[\varepsilon_0]}{\sum_{i=1}^{k} Pr[\varepsilon|\varepsilon_i] \cdot Pr[\varepsilon_i]}$. extended version

Definition: Independence

1. Two events $\varepsilon_1, \varepsilon_2$ are called **independent** if

$$Pr(\varepsilon_1 \cap \varepsilon_2) = Pr(\varepsilon_1) \cdot Pr(\varepsilon_2)$$

.

2. A collection of events $\{\varepsilon_i | i \in I\}$ is **independent** if for all subsets $S \subseteq I$

$$Pr\left[\bigcap_{i \in S} \varepsilon_i\right] = \prod_{i \in S} Pr[\varepsilon_i].$$

MODERN (BAYESIAN) INTERPRETATION of BAYES RULE

for the entire process of learning from evidence has the form

$$Pr[\varepsilon_1|\varepsilon] = \frac{Pr[\varepsilon_1 \cap \varepsilon]}{Pr[\varepsilon]} = \frac{Pr[\varepsilon|\varepsilon_1] \cdot Pr[\varepsilon_1]}{\sum_{i=1}^{k} Pr[\varepsilon|\varepsilon_i] \cdot Pr[\varepsilon_i]}.$$

In modern terms the last equation says that $Pr[\varepsilon_1|\varepsilon]$, the probability of a hypothesis ε_1 (given information ε), equals $\Pr(\varepsilon_1)$, our initial estimate of its probability, times $Pr[\varepsilon|\varepsilon_1]$, the probability of each new piece of information (under the hypothesis ε_1), divided by the sum of the probabilities of data in all possible hypothesis (ε_i).

TWO BASIC INTERPRETATIONS of PROBABILITY

In Frequentist interpretation, probability is defined with respect to a large number of trials, each producing one outcome from a set of possible outcomes - the probability of an event A , Pr(A), is a proportion of trials producing an outcome in A.

In Bayesian interpretation, probability measures a degree of belief. Bayes' theorem then links the degree of belief in a proposition before and after receiving an additional evidence that the proposition holds.

EXAMPLE 1

Let us toss a two regular cubes, one after another and let

ε_1 be the event that the sum of both tosses is ≥ 10

ε_2 be the event that the first toss provides 5

How much are: $Pr(\varepsilon_1), Pr(\varepsilon_2), Pr(\varepsilon_1|\varepsilon_2), Pr(\varepsilon_1 \cap \varepsilon_2)$?

$$Pr(\varepsilon_1) = \frac{6}{36}$$

$$Pr(\varepsilon_2) = \frac{1}{6}$$

$$Pr(\varepsilon_1|\varepsilon_2) = \frac{2}{6}$$

$$Pr(\varepsilon_1 \cap \varepsilon_2) = \frac{2}{36}$$

EXAMPLE 2

Three coins are given - two fair ones and in the third one heads land with probability $2/3$, but we do not know which one is not fair one.

When making an experiment and flipping all coins let the first two come up heads and the third one comes up tails. What is probability that the first coin is the biased one?

Let ε_i be the event that the ith coin is biased and B be the event that three coins flips came up heads, heads, tails.

Before flipping coins we have $Pr(\varepsilon_i) = \frac{1}{3}$ for all i. After flipping coins we have

$$Pr(B|\varepsilon_1) = Pr(B|\varepsilon_2) = \frac{2}{3}\frac{1}{2}\frac{1}{2} = \frac{1}{6} \qquad Pr(B|\varepsilon_3) = \frac{1}{2}\frac{1}{2}\frac{1}{3} = \frac{1}{12}$$

and using Bayes' law we have

$$Pr(\varepsilon_1|B) = \frac{Pr(B|\varepsilon_1)Pr(\varepsilon_1)}{\sum_{i=1}^{3} Pr(B|\varepsilon_i)Pr\,(\varepsilon_i)} = \frac{\frac{1}{6}\cdot\frac{1}{3}}{\frac{1}{6}\cdot\frac{1}{3}+\frac{1}{6}\cdot\frac{1}{3}+\frac{1}{12}\cdot\frac{1}{3}} = \frac{2}{5}$$

Therefore, the above outcome of the three coin flips increased the likelihood that the first coin is biased from 1/3 to 2/5

THEOREM

Let A and B be two events and let $\Pr(B) \neq 0$. Events A and B are independent if and only if

$$\Pr(A|B) = \Pr(A).$$

Proof

- Assume that A and B are independent and $\Pr(B) \neq 0$. By definition we have

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$$

  and therefore

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(A)\cdot\Pr(B)}{\Pr(B)} = \Pr(A).$$

- Assume that $\Pr(A|B) = \Pr(A)$ and $\Pr(B) \neq 0$. Then

$$\Pr(A) = \Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

  and multiplying by $\Pr(B)$ we get

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$$

  and so A and B are independent.

SUMMARY

- The notion of conditional probability, of A given B, was introduced in order to get an instrument for analyzing an experiment A when one has partial information B about the outcome of the experiment A before experiment has finished.
- **We say that two events A and B are independent if the probability of A is equal to the probability of A given B,**
- Other fundamental instruments for analysis of probabilistic experiments are **random variables** as functions from the sample space to $\mathbf{R}$, and **expectation** of random variables as the weighted averages of the values of random variables.

MONTY HALL PARADOX

Let us assume that you see three doors D1, D2 and D3 and you know that behind one door is a car and behind other two are goats.

Let us assume that you get a chance to choose one door and if you choose the door with car behind the car will be yours, and if you choose the door with a goat behind you will have to milk that goat for years.

Which door you will choose to open?

Let us now assume that you have chosen the door D1.

and let afterwords a moderator comes who knows where car is and opens one of the doors D_2 or D_3, say D2, and you see that the goat is in.

Let us assume that at that point you get a chance to change your choice of the door.

Should you do that?

Let C_1 denote the event that the car is behind the door D1.
Let C_3 denote the event that the car is behind the door D3.
Let M_2 denote the event that moderator opens the door D2.

Let us assume that the moderator chosen a door at random if goats were behind both doors he could open. In such a case we have

$$Pr[C_1] = \frac{1}{3} = Pr[C_3], \quad Pr[M_2|C1] = \frac{1}{2}, \quad Pr[M_2|C_3] = 1$$

Then it holds

$$Pr[C_1|M_2] = \frac{Pr[M_2|C_1]Pr[C_1]}{Pr[M_2]} = \frac{Pr[M_2|C_1]Pr[C_1]}{Pr[M_2|C_1]Pr[C_1] + Pr[M_2|C_3]Pr[C_3]} = \frac{1/6}{1/6+1/3} = \frac{1}{3}$$

Similarly

$$Pr[C_3|M_2] = \frac{Pr[M_2|C_3]Pr[C_3]}{Pr[M_2]} = \frac{Pr[M_2|C_3]Pr[C_3]}{Pr[M_2|C_1]Pr[C_1] + Pr[M_2|C_3]Pr[C_3]} = \frac{1/3}{1/6+1/3} = \frac{2}{3}$$

RANDOM VARIABLES - INFORMAL APPROACH

A random variable is a function defined on the elementary events of a probability space and having as values real numbers.

Example: In case of two tosses of a fair six-sided dice, the value of a random variable V can be the sum of the numbers on te two top spots on the dice rolls.

The value of V can therefore be an integer from the interval $[2, 12]$.

A random variable V with n potential values $v_1, v_2, \ldots, v_n$ is characterized by a probability distribution $p = (p_1, p_2, \ldots, p_n)$, where p_i is probability that V takes the value v_i.

The concept of random variable is one of the most important of modern science and technology.

INDEPENDENCE of RANDOM VARIABLES

Definition Two random variables X, Y are called **independent random variables** if

$$x, y \in \mathbf{R} \Rightarrow Pr_{X,Y}(x, y) = Pr[X = x] \cdot Pr[Y = y]$$

EXPECTATION – MEAN of RANDOM VARIABLES

Definition: The **expectation (mean or expected value) E**$[X]$ of a random variable X is defined as

$$\mathbf{E}[X] = \sum_{\omega \in \Omega} X(\omega) Pr_X(\omega).$$

Properties of he mean for random variabkes X and Y and a constant c:

$$\begin{aligned} \mathbf{E}[X+Y] &= \mathbf{E}[X] + \mathbf{E}[Y]. \\ \mathbf{E}[c \cdot X] &= c \cdot \mathbf{E}[X]. \\ \mathbf{E}[X \cdot Y] &= \mathbf{E}[X] \cdot \mathbf{E}[Y], \quad \text{if } X, Y \text{ are independent} \end{aligned}$$

The first of the above equalities is known as **linearity of expectations**. It can be extended to a finite number of random variables $X_1, \ldots, X_n$ to hold

$$\mathbf{E}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathbf{E}[X_i]$$

and also to any countable set of random variables $X_1, X_2, \ldots$ to hold: If $\sum_{i=1}^{\infty} \mathbf{E}[|X_i|] < \infty$, then $\sum_{i=1}^{\infty} |X_i| < \infty$ and

$$\mathbf{E}[\sum_{i=1}^{\infty} X_i] = \sum_{i=1}^{\infty} \mathbf{E}[X_i].$$

EXPECTATION VALUES

For any random variable X let $\mathbf{R}_X$ be the set of values of X. Using $\mathbf{R}_X$ one can show that

$$E[X] = \sum_{x \in \mathbf{R}_X} x \cdot Pr(X = x).$$

Using that one can show that for any real a, b it holds

$$\begin{aligned} \mathbf{E}[aX + b] &= \sum_{x \in \mathbf{R}_X} (ax + b) \Pr(X = x) \\ &= a \sum_{x \in \mathbf{R}_X} x \cdot \Pr(X = x) + b \sum_{x \in \mathbf{R}_X} \Pr(X = x) \\ &= a \cdot \mathbf{E}[X] + b \end{aligned}$$

The above relation is called **weak linearity of expectation**.

INDICATOR VARIABLES

A random variable X is said to be an indicator variable if X takes on only values 1 and 0.

For any set $A \subset S$, one can define an indicator variable X_A that takes value 1 on A and 0 on $S - A$, if $(S, \Pr)$ is the underlying probability space.

It holds:

$$\begin{aligned} \mathbf{E}_{\Pr}[X_A] &= \sum_{s \in S} X_A(s) \cdot \Pr(\{s\}) \\ &= \sum_{s \in A} X_A(s) \cdot \Pr(\{s\}) + \sum_{s \in S - A} X_A(s) \cdot \Pr(\{s\}) \\ &= \sum_{s \in A} 1 \cdot \Pr(\{s\}) + \sum_{s \in S - A} 0 \cdot \Pr(\{s\}) \\ &= \sum_{s \in A} \Pr(\{s\}) \\ &= \Pr(A) \end{aligned}$$

VARIANCE and STANDARD DEVIATION

Definition For a random variable X **variance** VX and **standard deviation** σX are defined by

$$\mathbf{V}X = \mathbf{E}((X - \mathbf{E}X)^2)$$

$$\sigma X = \sqrt{\mathbf{V}X}$$

Since

$$\begin{aligned} \mathbf{E}((X - \mathbf{E}X)^2) &= \mathbf{E}(X^2 - 2X\mathbf{E}X + (\mathbf{E}X)^2) = \\ &= \mathbf{E}(X^2) - 2(\mathbf{E}X)^2 + (\mathbf{E}X)^2 = \\ &= \mathbf{E}(X^2) - (\mathbf{E}X)^2, \end{aligned}$$

it holds

$$\mathbf{V}X = \mathbf{E}(X^2) - (\mathbf{E}X)^2$$

Example: Let $\Omega = \{1, 2, \ldots, 10\}$, $Pr(i) = \frac{1}{10}$, $X(i) = i$; $Y(i) = i - 1$ if $i \leq 5$ and $Y(i) = i + 1$ otherwise.
$\mathbf{E}X = \mathbf{E}Y = 5.5$, $\mathbf{E}(X^2) = \frac{1}{10} \sum_{i=1}^{10} i^2 = 38.5$, $\mathbf{E}(Y^2) = 44.5$; $\mathbf{V}X = 8.25$, $\mathbf{V}Y = 14.25$

TWO RULES

For independent random variables X and Y and a real number c it holds

- $\mathbf{V}(cX) = c^2\mathbf{V}(X)$; $\sigma(cX) = c\sigma(X)$
- $\mathbf{V}(X+Y) = \mathbf{V}(X) + \mathbf{V}(Y)$. $\sigma(X+Y) = \sqrt{V(X)+V(Y)}$.

MOMENTS

Definition

For $k \in \mathbf{N}$ the k-th moment m_X^k and the k-th central moment μ_X^k of a random variable X are defined as follows

$$\begin{aligned} m_X^k &= \mathbf{E}X^k \\ \mu_X^k &= \mathbf{E}((X - \mathbf{E}X)^k) \end{aligned}$$

The **mean** of a random variable X is sometimes denoted by $\mu_X = m_X^1$ and its variance by μ_X^2.

EXAMPLE I

Each week there is a lottery that always sells 100 tickets. One of the tickets wins 100 millions, all other tickets win nothing.

What is better: to buy in one week two tickets (Strategy I) or two tickets in two different weeks (Strategy II)?

Or none of these two strategies is better than the second one?

EXAMPLE II

With Strategy I we win (in millions)

$$0 \text{ with probability } 0.98$$

$$100 \text{ with probability } 0.02$$

With Strategy II we win (in millions)

$$0 \text{ with probability } 0.9801 = 0.99 \cdot 0.99$$

$$100 \text{ with probability } 0.0198 = 2 \cdot 0.01 \cdot 0.99$$

$$200 \text{ with probability } 0.0001 = 0.01 \cdot 0.01$$

Variance at Strategy I is 196

Variance at Strategy II is 198

PROBABILITY GENERATING FUNCTION

The **probability density function** of a random variable X whose values are natural numbers **can be represented** by the following **probability generating function** (PGF):

$$G_X(z) = \sum_{k\geq 0} Pr(X=k)\cdot z^k.$$

Main properties

$$G_X(1) = 1$$

$$\mathbf{EX} = \sum_{k\geq 0} k\cdot Pr(X=k) = \sum_{k\geq 0} Pr(X=k)\cdot(k\cdot 1^{k-1}) = \mathbf{G'_X(1)}.$$

Since it holds

$$\begin{aligned}\mathbf{E(X^2)} &= \sum_{k\geq 0} k^2\cdot Pr(X=k)\\ &= \sum_{k\geq 0} Pr(X=k)\cdot(k\cdot(k-1)\cdot 1^{k-2} + k\cdot 1^{k-1})\\ &= \mathbf{G''_X(1)} + \mathbf{G'_X}(1)\end{aligned}$$

we have

$$\mathbf{V}X = G''_X(1) + G'_X(1) - (G'_X(1))^2.$$

AN INTERPRETATION

- Sometimes one can think of the expectation $\mathbf{E}[Y]$ of a random variable Y as the "best guess" or the "best prediction" of the value of Y.
- It is the "best guess" in the sense that among all constants m the expectation $\mathbf{E}[(Y-m)^2]$ is minimal when $m = \mathbf{E}[\mathbf{Y}]$.

WHY ARE PGF USEFUL?

Main reason: For many important probability distributions their PGF are very simple and easy to work with.

For example, for the **uniform distribution** on the set $\{0,1,\ldots,n-1\}$ the PGF has form

$$U_n(z) = \frac{1}{n}(1+z+\ldots+z^{n-1}) = \frac{1}{n}\cdot\frac{1-z^n}{1-z}.$$

Problem is with the case $z=1$.

PROPERTIES of GENERATING FUNCTIONS

Property 1 If $X_1,\ldots,X_k$ are independent random variables with PGFs $G_1(z),\ldots,G_k(z)$, then the random variable $Y=\sum_{i=1}^k X_i$ has as its PGF the function

$$G(z) = \prod_{i=1}^{k} G_i(z).$$

Property 2 Let $X_1,\ldots,X_k$ be a sequence of independent random variables with the same PGF $G_X(z)$. If Y is a random variable with PGF $G_Y(z)$ and Y is independent of all X_i, then the random variable $S = X_1+\ldots+X_Y$ has as PGF the function

$$G_S(z) = G_Y(G_X(z)).$$

IMPORTANT DISTRIBUTIONS

Two important distributions are connected with experiments, called **Bernoulli trials**, that have two possible outcomes:

- **success** with probability p
- **failure** with probability $q = 1 - p$

Coin tossing is an example of a Bernoulli trial.

1. Let values of a random variable X be the number of trials needed to obtain a success. Then

$$Pr(X = k) = q^{k-1}p$$

Such a probability distribution is called the **geometric distribution** and such a variable geometric random variable. It holds

$$\mathbf{E}X = \frac{1}{p} \qquad \mathbf{V}X = \frac{q}{p^2} \qquad G(z) = \frac{pz}{1 - qz}$$

2. Let values of a random variable Y be the number of successes in n trials. Then

$$Pr(Y = k) = \binom{n}{k} p^k q^{n-k}$$

Such a probability distribution is called the **binomial distribution** and it holds

$$\mathbf{E}Y = np \qquad \mathbf{V}Y = npq \qquad G(z) = (q + pz)^n$$

and also

$$\mathbf{E}Y^2 = n(n-1)p^2 + np$$

BERNOULLI DISTRIBUTION

Let X be a binary random variable (called usually Bernoulli or indicator random variable) that takes value 1 with probability p and 0 with probability $q = 1 - p$, then it holds

$$\mathbf{E}[X] = p \qquad \mathbf{V}X = pq \qquad G[z] = q + pz.$$

BINOMIAL DISTRIBUTION revisited

Let $X_1, \ldots, X_n$ be random variables having Bernoulli distribution with the common parameter p.

The random variable

$$X = X_1 + X_2 + \ldots + X_n$$

has so called binomial distribution denoted $B(n, p)$ with the density function denoted

$$B(k, n, p) = Pr(X = k) = \binom{n}{k} p^k q^{(n-k)}$$

POISSON DISTRIBUTION

Poisson distribution

Let $\lambda \in \mathbf{R}^{>0}$. The Poisson distribution with the parameter λ is the probability distribution with the density function

$$p(x) = \begin{cases} \lambda^x \frac{e^{-\lambda}}{x!}, & \text{for } x = 0,1,2,\dots \\ 0, & \text{otherwise} \end{cases}$$

For large n the Poisson distribution is a good approximation to the Binomial distribution $B(n, \frac{\lambda}{n})$

Property of a Poisson random variable X:

$$\mathbf{E}[X] = \lambda \qquad \mathbf{V}X = \lambda \qquad G[z] = e^{\lambda(z-1)}$$

EXPECTATION+VARIANCE OF SUMS OF RANDOM VARIABLES

Let

$$S_n = \sum_{i=1}^{n} X_i$$

where each X_i is a random variable which takes on value 1 (0) with probability p $(1-p=q)$.
It clearly holds

$$\begin{aligned}
\mathbf{E}(X_i) &= p \\
\mathbf{E}(X_i^2) &= p \\
\mathbf{E}(S_n) &= \mathbf{E}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathbf{E}(X_i) = np \\
\mathbf{E}(S_n^2) &= \mathbf{E}((\sum_{i=1}^{n} X_i)^2) = \mathbf{E}(\sum_{i=1}^{n} X_i^2 + \sum_{i \neq j} X_i X_j) = \\
&= \sum_{i=1}^{n} \mathbf{E}(X_i^2) + \sum_{i \neq j} \mathbf{E}(X_i X_j)
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbf{E}(S_n^2) &= \mathbf{E}((\sum_{i=1}^{n} X_i)^2) = \mathbf{E}(\sum_{i=1}^{n} X_i^2 + \sum_{i \neq j} X_i X_j) = \\
&= \sum_{i=1}^{n} \mathbf{E}(X_i^2) + \sum_{i \neq j} \mathbf{E}(X_i X_j)
\end{aligned}$$

and therefore, if X_i, X_j are pairwise independent, as in this case, $\mathbf{E}(X_i X_j) =$ $= \mathbf{E}(X_i)\mathbf{E}(X_j)$ Hence

$$\begin{aligned}
\mathbf{E}(S_n^2) &= np + 2\binom{n}{2} p^2 \\
&= np + n(n-1)p^2 \\
&= np(1-p) + n^2 p^2 \\
&= n^2 p^2 + npq \\
VAR[S_n] &= \mathbf{E}(S_n^2) - (\mathbf{E}(S_n))^2 = n^2 p^2 + npq - n^2 p^2 = npq
\end{aligned}$$

MOMENT INEQUALITIES

The following inequality, and several of its special cases, play very important role in the analysis of randomized computations:

Let X be a random variable that takes on values x with probability $p(x)$.

Theorem For any $\lambda > 0$ the so called k^{th} **moment inequality** holds:

$$Pr[|X| > \lambda] \leq \frac{\mathbf{E}(|X|^k)}{\lambda^k}$$

Proof of the above inequality;

$$\begin{aligned}
\mathbf{E}(|X|^k) &= \sum_{x \in X} |x|^k p(x) \geq \sum_{|x| > \lambda} |x|^k p(x) \geq \\
&\geq \lambda^k \sum_{|x| > \lambda} p(x) = \lambda^k Pr[|X| > \lambda]
\end{aligned}$$

Two important special cases - I.1

of the moment inequality;

$$Pr[|X| > \lambda] \leq \frac{\mathbf{E}(|X|^k)}{\lambda^k}$$

Case 1 $\quad k \to 1 \quad \lambda \to \lambda \mathbf{E}(|X|)$

$$Pr[|X| \geq \lambda \mathbf{E}(|X|)] \leq \frac{1}{\lambda} \qquad \textbf{Markov's inequality}$$

Case 2 $\quad k \to 2 \quad X \to X - \mathbf{E}(X), \lambda \to \lambda\sqrt{V(X)}$

$$Pr\left[|X - \mathbf{E}(X)| \geq \lambda\sqrt{V(X)}\right] \leq \frac{\mathbf{E}((X - \mathbf{E}(X))^2)}{\lambda^2 V(X)} =$$

$$= \frac{V(X)}{\lambda^2 V(X)} = \frac{1}{\lambda^2} \qquad \textbf{Chebyshev's inequality}$$

Another variant of Chebyshev's inequality:

$$\Pr[|X - \mathbf{E}(X)| \geq \lambda] \leq \frac{V(X)}{\lambda^2}$$

and this is one of the main reasons why variance is used.

Two important special cases - I.2

The following generalization of the moment inequality is also of importance:

Theorem

If $g(x)$ is non-decreasing on $[0, \infty)$, then

$$Pr[|X| > \lambda] \leq \frac{\mathbf{E}(g(X))}{g(\lambda)}$$

As a special case, namely if $g(x) = e^{tx}$, we get:

$$Pr[|X| > \lambda] \leq \frac{\mathbf{E}(e^{tX})}{e^{t\lambda}} \qquad \textbf{basic Chernoff's inequality}$$

Chebyshev's inequalities are used to show that values of a random variable lie close to its average with high probability. The bounds they provide are called also **concentration bounds**. Better bounds can usually be obtained using Chernoff bounds discussed in Chapter 5.

FLIPPING COINS EXAMPLES on CHEBYSHEV INEQUALITIES

Let X be a sum of n independent fair coins and let X_i be an indicator variable for the event that the i-th coin comes up heads. Then $\mathbf{E}(X_i) = \frac{1}{2}$, $\mathbf{E}(X) = \frac{n}{2}$, $\mathrm{Var}[X_i] = \frac{1}{4}$ and

$\mathrm{Var}[X] = \sum \mathrm{Var}[X_i] = \frac{n}{4}$.

Chebyshev's inequality

$$\Pr[|X - \mathbf{E}(X)| \geq \lambda] \leq \frac{V(X)}{\lambda^2}$$

for $\lambda = \frac{n}{2}$ gives

$$Pr[X = n] \leq Pr[|X - n/2| \geq n/2] \leq \frac{n/4}{(n/2)^2} = \frac{1}{n}$$

THE INCLUSION-EXCLUSION PRINCIPLE

Let $A_1, A_2, \ldots, A_n$ be events – not necessarily disjoint. The **Inclusion-Exclusion principle**, that has also a variety of applications, states that

$$\begin{aligned} Pr\left[\bigcup_{i=1}^n A_i\right] &= \sum_{i=1}^n Pr(A_i) - \sum_{i<j} Pr(A_i \cap A_j) + \sum_{i<j<k} Pr(A_i \cap A_j \cap A_k) - \\ &- \ldots + (-1)^{k+1} \sum_{i_1 < i_2 < \ldots < i_k} Pr\left[\bigcap_{j=1}^k A_{i_j}\right] \ldots + \\ &+ (-1)^{n+1} Pr\left[\bigcap_{i=1}^n A_i\right] \end{aligned}$$

BONFERRONI'S INEQUALITIES

the following Bonferroni's inequalities follow from the Inclusion-exclusion principle:
For every odd $k \leq n$

$$Pr\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{j=1}^{k} (-1)^{j+1} \sum_{i_1 < \ldots < i_j \leq n} Pr\left(\bigcap_{l=1}^{j} A_{i_l}\right)$$

For every even $k \leq n$

$$Pr\left(\bigcup_{i=1}^{n} A_i\right) \geq \sum_{j=1}^{k} (-1)^{j+1} \sum_{i_1 < \ldots < i_j \leq n} Pr\left(\bigcap_{l=1}^{j} A_{i_l}\right)$$

SPECIAL CASES of THE INCLUSION-EXCLUSION PRINCIPLE

"Markov"-type inequality - Boole's inequality or Union bound

$$Pr\left(\bigcup_i A_i\right) \leq \sum_i Pr(A_i)$$

"Chebyshev"-type inequality

$$Pr\left(\bigcup_i A_i\right) \geq \sum_i Pr(A_i) - \sum_{i<j} Pr(A_i \cap A_j)$$

Another proof of Boole's inequality:

Let us define $B_i = A_i - \bigcup_{j=1}^{i-1} A_j$. Then $\bigcup A_i = \bigcup B_i$. Since B_i are disjoint and for each i we have $B_i \subset A_i$ we get

$$Pr[\bigcup A_i] = Pr[\bigcup B_i] = \sum Pr[B_i] \leq \sum Pr[A_i]$$

APPENDIX

APPENDIX

PUZZLE - HOMEWORK

Puzzle 1 Given a biased coin, how to use it to simulate an unbiased coin?

Puzzle 2 n people sit in a circle. Each person wears either red hat or a blue hat, chosen independently and uniformly at random. Each person can see the hats of all the other people, but not his/her hat. Based only upon what they see, each person votes on whether or not the total number of red hats is odd. Is there a scheme by which the outcome of the vote is correct with probability greater than $1/2$.

MODERN (BAYESIAN) INTERPRETATION of BAYES RULE

Bayes rule for the process of learning from evidence has the form:

$$Pr[\varepsilon_1|\varepsilon] = \frac{Pr[\varepsilon_1 \cap \varepsilon]}{Pr[\varepsilon]} = \frac{Pr[\varepsilon|\varepsilon_1] \cdot Pr[\varepsilon_1]}{\sum_{i=1}^{k} Pr[\varepsilon|\varepsilon_i] \cdot Pr[\varepsilon_i]}.$$

In modern terms the last equation says that $Pr[\varepsilon_1|\varepsilon]$, the probability of a hypothesis ε_1 (given information ε), equals $\Pr(\varepsilon_1)$, our initial estimate of its probability, times $Pr[\varepsilon|\varepsilon_1]$, the probability of each new piece of information (under the hypothesis ε_1), divided by the sum of the probabilities of data in all possible hypothesis (ε_i).

EXAMPLE - DRUG TESTING

Suppose that a drug test will produce 99% true positive and 99% true negative results.

Suppose that 0.5% of people are drug users.

If the test of a user is positive, what is probability that such a user is a drug user?

SOLUTION

$$\Pr(\text{drg-us}|+) = \frac{\Pr(+|\text{drg-us})\Pr(\text{drg-us})}{\Pr(+|\text{drg-us})\Pr(\text{drg-us}) + \Pr(+|\text{no-drg-us})\Pr(\text{no-drg-us})}$$

$$\Pr(drg - us|+) = \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995} =\approx 33.2\%$$

BAYES' RULE INFORMALLY

Basically, Bayes' rule concerns of a broad and fundamental issue: how we analyze evidence and change our mind as we get new information, and make rational decision in the face of uncertainty.

Bayes' rule as one line theorem: by updating our initial belief about something with new objective information, we get a new and improved belief

BAYES' RULE STORY

- Reverend Thomas Bayes from England discovered the initial version of the "Bayes's law" around 1974, but soon stopped to believe in it.
- In behind were two philosophical questions
  - Can an effect determine its cause?
  - Can we determine the existence of God by observing nature?
- Bayes law was not written for long time as formula, only as the statement: **By updating our initial belief about something with objective new information, we can get a new and improved belief.**
- Bayes used a tricky thought experiment to demonstrate his law.
- Bayes' rule was later invented independently by Pierre Simon Laplace, perhaps the greatest scientist of 18th century, but at the end he also abounded it.
- Till the 20 century theoreticians considered Bayes rule as unscientific. Bayes rule had for centuries several proponents and many opponents in spite that it has turned out to be very useful in practice.
- Bayes rule was used to help to create rules of insurance industries, to develop strategy for artillery during the first and even Second World War (and also a great Russian mathematician Kolmogorov helped to develop it for this purpose).

- It was used much to decrypt ENIGMA codes during 2WW, due to Turing, and also to locate German submarines.