Mathematical methods and biological applications

SS 2012

Institut für Genetik und Mathematisches Institut

Administration and Schedule

Preliminary schedule:

- Lectures: Wednesday 8.15-9.45
- Exercises: Friday 8.15-9.45

Date	Lecture topic	Name	Date	Exercises?
04.04.2012	Fundamentals of probability	Wiehe	06.04.2012	holiday
11.04.2012	Stochastic processes: Wright-Fisher model	Wiehe	13.04.2012	yes, C-Pool Ger
18.04.2012	Stochastic processes: Coalescent	Wiehe	20.04.2012	cancelled
25.04.2012	Fundamentals of Combinatorics	Disanto	27.04.2012	yes
02.05.2012	Counting trees, Permutations and Grey code	Disanto	04.05.2012	yes
09.05.2012	Introduction to Number theory	Bringmann	11.05.2012	yes?
16.05.2012	Number theory	Bringmann	18.05.2012	yes?
23.05.2012	Ranked trees and generating functions	Disanto	25.05.2012	yes, C-Pool Ger
06.06.2012	ODEs, PDEs and dynamical systems	Sweers	08.06.2012	yes?
13.06.2012	ODEs, PDEs and dynamical systems	Sweers	15.06.2012	yes?
20.06.2012	Introduction to Numerics	Tischendorf	22.06.2012	yes?, C-Pool Ge
27.06.2012	Numerics, NOTE: Computerpool	Tischendorf	29.06.2012	yes?, C-Pool Ge
04.07.2012	Summary, NOTE: Computerpool	all	06.07.2012	exam prep, C-P
13.07.2012	Written exam (1.5 hours)			

- Credits: 4 ECTS
- Exam regulation: 1.5 hours written exam and exercises/homework problems
- Pass grade: pass in written exam ($\geq 50\%$) and $\geq 50\%$ of points in exercise/homework problems

04.04.2012

1 Fundamentals of Probability theory

1.1 Biological applications

- Hypothesis testing
  - H_0: The proportion of males and females among newborns is identical
- Patterns
  - Arrangement of branches around a tree trunk
  - Coat patterning
- Diffusion
  - Dispersal of seeds
  - Infectious diseases
  - Spatial concentration of molecules in cells during early development
- Modeling of stochastic processes
  - **Evolutionary process**
- Modeling of dynamical systems
  - Predator prey systems

1.2 Sample space. Random Variables. Distributions

Definition
The *sample space* is the set of all possible outcomes of a *random* experiment.
Example:
Throwing a fair die. Sample space $\Omega_1 = \{1, ..., 6\}$.
Example:
Number of rabbits in a colony. Sample space $\Omega_2 = \{0, 1, 2, ...\}$.
Example:
GC-content in a DNA sequence. Sample space $\Omega_3 = [0, 1]$.
Example:
Three traffic lights in Zülpicher Straße. Sample space $\Omega_4 = \{RRR, RRG, RGR, GRR, ...GGG\}$.

Definition

An *event* E is a subset of Ω.
Example:
$E_4 = \{RGG, GRG, GGR\} \subset \Omega_4$ "encounter exactly one red light in Zülpicher Straße".

The usual set operations (union, intersection, complement, ...) can be performed on events.

Definition
A *probability measure* is a function P from the power set $\mathcal{P}$ (or, if Ω is not countable, a σ-algebra $\mathfrak{A}$) of Ω to the real unit interval which satisfies

1. $P(\Omega) = 1$
2. if $E \subset \Omega$, then $P(E) \geq 0$
3. if E_1 and E_2 are disjoint, then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$.

The triplet $(\Omega, \mathcal{A}, \mathcal{P})$ is called a *probability space*.
Note, that it follows that

$$P(E_1 \cup E_2) \leq P(E_1) + P(E_2)$$

for arbitrary events $E_1, E_2 \subset \Omega$.

Example:
Assume that all outcomes in the traffic lights experiment are equally likely. Then $P(E_4) = 3/8 = 0.375$.

In case of finite Ω: may determine probabilities of events by counting. In contrast, consider Ω_2: there is no uniform probability measure!

Consider experiment 4 as a repetition of three simpler experiments: look at traffic light 1, 2 and 3 as if they were "independent": $\Omega_{4'} = \{R, G\}^3$ with identical probability measures $P(R) = 3/4$ and $P(G) = 1/4$ on each projection.
Example:
What is the probability of "exactly one red light among three independent traffic lights"? $P(E_{4'}) = 33/64 = 0.141$. In practice, often they are not independent. If you find G at the first light, you may have higher chances to find G also on the second and third light.

Definition
The *conditional probability* of an event E_1 *given* an event E_2 is found by considering the occurrences of E_1 under the condition that also E_2 has occurred. Formally,

$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}.$$

Example:
Traffic light problem. Simple conditional probabilities may be given in terms of a "transition matrix":

$$\begin{array}{ccc} \begin{array}{c}\mathrm{light}_1 \rightarrow \\ \mathrm{light}_2 \downarrow\end{array} & R & G \\ R & 0.5 & 0.4 \\ G & 0.5 & 0.6 \end{array}$$

For instance, the probability for $P(\text{"second light is green"}|\text{"first light is green"}) = 0.6$. What is the probability of "exactly one red light among three non-independent traffic lights"? To answer this, one needs to know the *initial* probability measure for the first light. Let this be $P(R) = 3/4$ and $P(G) = 1/4$.

$P(RGG, GRG, GGR) = 0.75\times0.5\times0.6+0.25\times0.4\times0.5+0.25\times0.6\times0.4 = 0.335$

The probability $P(E_2)$ can be calculated as so-called *total* probability:

$$P(E_2) = P(E_2|E_1)P(E_1) + P(E_2|E_1^c)P(E_1^c)\,,$$

where E^c is the *complement* of E.

For instance, $P(\text{"second light is green"}) = 0.6 \times 0.25 + 0.5 \times 0.75 = 0.525$.

This is formalized as "Bayes' rule":

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

More generally, for disjoint events $B_1, B_2, B_3, ...B_n$ with $\cup_{i=1}^n B_i = \Omega$

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)}$$

where $P(A)$ is computed as total probability $P(A) = \sum_{i=1}^n P(A|B_i)P(B_i)$.

The possibility of "exchanging conditions" is important in Bayesian inference and testing theory (the probability P("data have property j given observation A)" can be calculated by means of Bayes' rule)

Two events E_1 and E_2 are independent, if

$$P(E_1|E_2) = P(E_1).$$

In words: "Knowledge of occurrence of E_2 does not alter the probability for E_1".

Definition
A (real) *random variable* X is a mapping from the sample space to the real numbers (or a subset thereof)

$$X : \Omega \rightarrow X(\Omega) \subset R$$

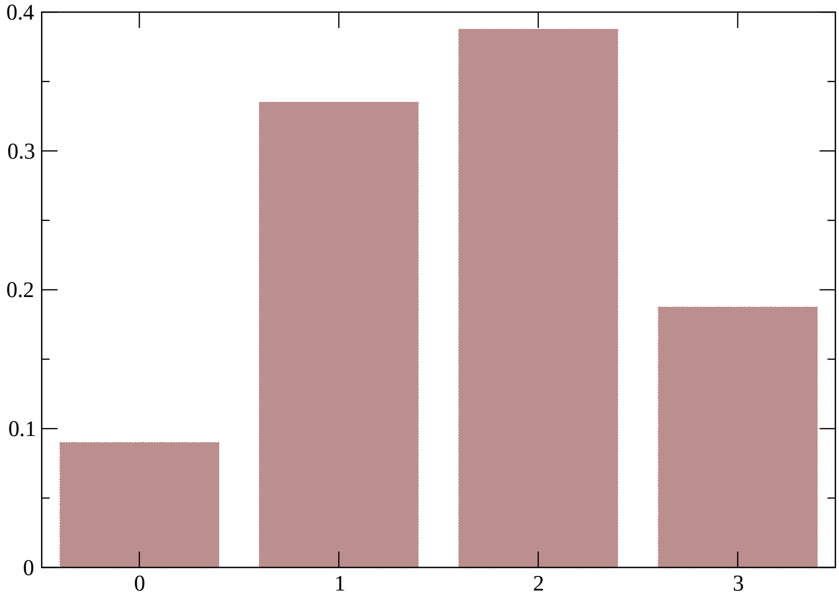

Figure 1: Distribution of red traffic lights, given the initial distribution and transition probabilities mentioned in the text

Since an outcome of an random experiment in sample space Ω is random, also the number produced by the function $X = X(\omega)$ is random.

Often, only the properties of X, but not the details of Ω are of interest.

Example:
The "number of red traffic lights in Zülpicher Straße" is a random variable and can take values $0, 1, 2, 3$. It is a mapping $X : \Omega_4 \to \{0, 1, 2, 3\}$.

What are the probabilities? For instance, $P(X = 0) = P(X^{-1}(0)) = P(GGG) = 0.25 \times 0.6 \times 0.6 = 0.09$.

Similarly, one can determine $P(X = x)$ for $x = 1, 2, 3$. Together, these numbers constitute the *distribution* of X. Here,

x	$X^{-1}(x)$	$P(X = x)$
0	GGG	0.0900
1	RGG, GRG, GGR	0.3350
2	RRG, RGR, GRR	0.3875
3	RRR	0.1875

Definition
The function

$$p : X(\Omega) \to R$$

with $p(x) = P(X^{-1}(x))$, where X is a *discrete* random variable, is called the *probability mass function* of X.

The function

$$F : X(\Omega) \to R$$

with $F(x) = P(X \leq x)$ is called the *(cumulative) distribution function* of X.

Note:

- For *continuous* random variables (for instance, the outside temperature at 8.00 am) one may define a cumulative distribution function F in an analogous way.
- Often, but not always, continuous random variables have a *probability density* function f. If F can be differentiated, then $F' = f$.

1.3 Moments

The *expectation*, or mean or first moment, of a R.V. is the sum

$$\mathcal{E}(X) = \sum_i x_i p(x_i)$$

in case of a discrete R.V. and the integral

$$\mathcal{E}(X) = \int x f(x) dx$$

in case of a continuous R.V. with a density function f.

The k*-th moment*s are

$$\mathcal{E}(X^k) = \sum_i x_i^k p(x_i)$$

and

$$\mathcal{E}(X^k) = \int x^k f(x) dx \, ,$$

respectively. The *variance* of a R.V. X is

$$V(X) = E(X^2) - E^2(X)$$

The expectation is linear. Given two R.V. X and Y with finite expectation and constants a and b, then

$$\mathcal{E}(aX + bY) = a\mathcal{E}(X) + b\mathcal{E}(Y)$$

But

$$V(aX) = a^2 V(X)$$

The variance is not linear!

1.4 Special random variables and their distributions

The binomial distribution (discrete, finite)

A random variable X on the integers 0,...n is *binomially distributed* with parameters n and p if its distribution function is given by

$$F(X \leq x) = \sum_{i=0}^{x} P(X = i)$$

where

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

are the "binomial" probabilities.
Example:
Consider a population of N individuals with the relative frequency p of an allele (=gene variant) A, i.e. $\mathrm{freq}(A) = p$. Given random mating and that population size remains constant, the frequency of A in the next generation is binomially distributed with parameters N and p. The random change in frequency from generation to generation is called "genetic drift".

The Poisson distribution (discrete, infinite)

A random variable X on the positive integers is *Poisson distributed* with parameter μ if its distribution function is given by

$$F(X \leq x) = \sum_{i=0}^{x} P(X = i)$$

where

$$P(X^{-1}(i) = \exp(-\mu)\mu^i / i!$$

are the Poisson probabilities.
Example:
The number of mutations in a DNA sequence, accumulating in k generations, is Poisson distributed with parameter $k\mu$.

The exponential distribution (continuous, unbounded)

A random variable X on the positive real numbers is exponentially distributed with parameter λ if its distribution function is given by

$$F(X \leq x) = \int_{i=0}^{x} f_\lambda(x) dx$$

where $f(x)$ is the *density* function of the exponential distribution and given by

$$f_\lambda(x) = \lambda \exp(-\lambda x)\,.$$

Example:
The coalescent time of two genealogical lineages in a population of constant size N is (approximately) exponentially distributed with parameter $\lambda = 1/N$.

Example:
The number of genetic differences between two individuals drawn randomly from a diploid population of constant size N is Poisson distributed with parameter $\theta = 2 \times 2N\mu$.

The uniform distribution (discrete or continuous, bounded)

A random variable X on the real unit interval is uniformly distributed if its distribution function is given by

$$F(X \leq x) = x\,.$$

A random variable X on the integers $\{1, ...n\}$ is uniformly distributed if its distribution function is given by

$$F(X \leq x) = x/n\,, 1 \leq x \leq n\,.$$

Example:
Consider a binary tree with n leafs (n odd) generated under the coalescent process (i.e. random bifurcation). The number of "left-leafs" (w.l.o.g.) is uniformly distributed on $\{1, ...\lfloor n/2 \rfloor\}$.

1.5 Stochastic processes

Definition
A *stochastic process* is a family of (real) random variables $(X_t)_t$, $t \in T$, where T is a totally ordered, finite or infinite, index set ("time").
Example:
Temperature at Wednesdays, 8.00 am, during this course.

1.6 Exercises 1. 04.04.2012

1. Determine the distribution of red traffic lights (among 3), when they are independent and when $P(G) = 1/4$ and $P(R) = 3/4$. Draw the probability mass function and the distribution function. (5P)

2. Given $P(G) = 1/4$ and traffic lights are independent. What is the probability to find a "run" of 0,1,2,3 green ones? (5P)

3. * What is the expected run length? (extra 5P)

4. (Easter egg problem) Easter bunny tells you that behind exactly one of three doors you will find Easter eggs. You can select one of the three doors, however without opening it yet. Now, Easter bunny opens one of the two remaining doors and shows you that nothing is hidden behind it. Do you reconsider your first choice? If yes, why? (10P)