

Exercises for SPAß: Statistics, Probability and Algorithms in Bioinformatics

Freie Universität Berlin, WS 2003/04
Dipl.-Math. Sven Rahmann

Problem set 1 · Handed out on 22.10.2003

Please hand in solutions by 29.10.2003 · Late solutions will not be accepted

Problem 1 (Countability). Using the following definition from the lecture,

A set I is countable if there exists an injective mapping $I \rightarrow \mathbb{N}$,

show that the set $\mathbb{Z}^2 := \mathbb{Z} \times \mathbb{Z}$ is countable.

Solution. There are two steps: (1) \mathbb{Z} is countable. This is easily seen by mapping $z \in \mathbb{Z} \mapsto 2z + 1$ if $z \geq 0$, and $z \mapsto 2|z|$ if $z < 0$ (nonnegative numbers map to odd numbers, negative numbers map to even numbers). (2) Now it is sufficient to show that $\mathbb{N} \times \mathbb{N}$ is countable. Then the result follows by composition. The idea is to use enumeration by diagonal, as shown below. The element (a, b) belongs to diagonal number $d = a + b - 1$ (so $(1, 1)$ makes up diagonal number 1). Within this diagonal, it is the a -th element. The first $d - 1$ diagonals contain $1 + 2 + \dots + (d - 1) = (d - 1) \cdot d/2$ elements. Therefore (a, b) is enumerated as the $[(d - 1) \cdot d/2 + a]$ -th element. In other words, the map $(a, b) \mapsto [(a + b - 2) \cdot (a + b - 1) + 2a]/2$ does the job.

b / a	1	2	3	...
1	1	3	6	...
2	2	5	9	...
3	4	8	13	
⋮	⋮	⋮		

Problem 2 (Sum of independent random variables). Let X_1, X_2, X_3 be independent random variables that have the uniform distribution on $\{1, \dots, 10\}$. What is the distribution of $X_1 + X_2 + X_3$? Try to solve this intelligently!

Solution. While it is possible to enumerate all 10^3 combinations of (x_1, x_2, x_3) and count how many of them give which sum, this is *not* the most intelligent solution. If there were 10 instead of 3 variables, we would be in serious trouble. It is much better to proceed step by step. We first construct the distribution of $S_2 := X_1 + X_2$ and then proceed inductively to construct the distribution of $S_n = S_{n-1} + X_n$. In each step, we only have to look at pairs of random variables (with bounded range) and hence avoid the curse of dimensionality. This is easy to program with two nested for loops, and one can even do this by hand: For $k = n..(10n)$, we have

$$\mathbb{P}(S_n = k) = \sum_{i=1}^{10} \mathbb{P}(X_n = i) \cdot \mathbb{P}(S_{n-1} = k - i).$$

We can even exploit that $\mathbb{P}(X_n = i)$ is constant $1/10$. We only need to pay attention to the boundary conditions because $\mathbb{P}(S_{n-1} = k - i)$ might be zero for some i . The solution is given in the table below for $n = 3$ and $k = 3$.16, the other side is symmetric.

k	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
$\mathbb{P}(S_3 = k)$	1	3	6	10	15	21	28	36	45	5	63	69	73	75	/1000

Problem 3 (Discretization of a normal distribution). Compute the $(-10, 1, 10)$ -discretization of the standard normal distribution. Which function is essential for this task?

Solution. We need the cdf Φ of the Normal distribution. While there is no closed form for it, we may nevertheless assume that it is available numerically as an “elementary” function, e.g., via MATLAB or R. You can also look at “Numerical Recipes”, Section 6.2. In MATLAB, we can conveniently write

```
d = diff([0 normcdf([-10.5:1:10.5]) 1]);
```

to obtain the discretization, including the values for R_- and R_+ . Note that evaluating the pdf φ at the midpoints of the interval would be a less accurate approximation. We can also exploit the symmetry. The solution is given in the table below. Make sure to get a feeling for the order of magnitude.

Interval	Probability
$] -\infty, -10.5] = R_-$	4.319006317809258e-26
$] -10.5, -9.5] = R_{-10}$	1.049408317473094e-21
R_{-9}	9.478485370695816e-18
R_{-8}	3.189943719428700e-14
R_{-7}	4.012809692186214e-11
R_{-6}	1.894940246004914e-08
R_{-5}	3.378683562264174e-06
R_{-4}	2.292314059107950e-04
R_{-3}	0.00597703624674
R_{-2}	0.06059753594308
R_{-1}	0.24173033745713
R_{-0}	0.38292492254803

Problem 4 (IEEE floating point format). For the following two questions, give both the bit representation and the number: What is the smallest positive representable IEEE double precision number? What is the smallest representable IEEE double precision number larger than 1.0? Finally, give the IEEE single precision bit representation of $1/100 = 0.01$.

Solution. Smallest positive double: S=0, E=0, F=(0...01) gives a value of $2^{-1022} \cdot 2^{-52} = 2^{-1074}$, which is approximately $10^{-323.3}$.

Smallest positive double > 1 : S=0, E=1023=(011...1), F=(0...01) gives a value of $2^0 \cdot (1 + 2^{-52}) = 1 + 2^{-52}$, which is approximately $1 + 10^{-15.65356}$ or $1 + 2.22 \cdot 10^{-16}$.

Now we must represent $1/100$ with powers of 2. Clearly $2^{-7} = 1/128 < 1/100 < 1/64$, so the we will have an exponent of -7 , so $E=120$. It quickly becomes messy to compute the fraction, though. We'd better write a short C program to read out the bits. We need to pay close attention to the byte-order in memory for this to work. You can download `printieee.c` from the webpage; it seems to work on Linux i386 and Compaq Tru64 machines. The solution is

(0 | 01111000 | 01000111101011100001010).

Project 1 (Hadamard matrices). An $n \times n$ matrix $H = (H_{ij})$ is called a *Hadamard matrix* of order n if and only if

- $H_{ij} \in \{+1, -1\}$ for all i, j ,
- $H^T \cdot H = n \cdot I_n$, where $(\cdot)^T$ denotes transposition, and I_n is the $n \times n$ identity matrix.

To warm up, note the following:

1. $H^T \cdot H = n \cdot I_n$ is equivalent to $H \cdot H^T = n \cdot I_n$.
2. Any two rows (columns) of a Hadamard matrix are orthogonal.
3. Exchanging two rows (columns) of a Hadamard matrix preserves the Hadamard property.
4. Multiplying a row (column) of a Hadamard matrix by -1 preserves the Hadamard property.
5. If a Hadamard matrix of order n exists, then there exists one whose first row and column consist entirely of 1s. Such a matrix is said to be *normalized*.

We are interested in the existence of Hadamard matrices of order n .

- Obviously there exist Hadamard matrices of order 1, 2, and 4. Write them down.
- Show that if there exists a Hadamard matrix of order $n > 2$, then n is divisible by 4.
- Given Hadamard matrices of orders p and q , show how to construct a Hadamard matrix of order pq .
- It is conjectured that there exists a Hadamard matrix for *every* order $n = 4m$, but this is currently unproven. The smallest undecided case is $n = 428 = 4 \cdot 107$. Your task is to find such a matrix, using whatever methods you can think of! (Alternatively, find a proof that no such matrix exists.)

More information is available at the following URLs and in the references given there:

- <http://www.research.att.com/~njas/hadamard/>

- <http://mathworld.wolfram.com/HadamardMatrix.html>
- <http://www.math.ntua.gr/people/ckoukouv/openhad.htm>

The undecided cases for $n < 2000$ are 428, 668, 716, 764, 892, 956, 1004, 1132, 1244, 1388, 1436, 1676, 1772, 1852, 1912, 1916, 1948, 1964. Good luck!

Note: Projects are voluntary. If you decide do to one, do it well! Solutions for projects can be handed in at any time during the semester.