Applied Statistics

Problems in fundamental concepts of statistics

The following exam has been handed out on Thursday the 1st of November 2012, and a written solution is to be handed in by Friday the 2nd of November at noon by Email or in person. Plots should be included were appropriate. Solving the exam in groups is **not** allowed.

Good luck and thank you for all your hard work, Troels & Sascha

---

Declare the past, diagnose the present, foretell the future. [Hippocrates, ca. 460-377 BC]

I – Distributions and probabilities:

1.1 Let x be distributed according to the PDF $f(x) = 2e^{-x/2}$ in the interval $[C, \infty]$.

- For which value of C is the PDF $f(x)$ normalized?
- What is the mean and width of x?

1.2 How many times will Little Peter have to roll a normal die ($p = 1/6$) to be 95% sure that he will have at least four sixes?

1.3 The average pregnancy is 278 days. Assume the time to follow a Gaussian distribution with a width of 11 days. What is the probability to give birth on the due date?

II – Error propagation:

2.1 If the diameter of a circle is known to 1% precision, how well is the area know?

2.2 The effective potential between uncharged atoms can be expressed as: $V(r) = -Ar^{-6} + Br^{-12}$, with $A, B > 0$. An experiment has measured the depth of the potential below zero to be $\epsilon = (0.52 \pm 0.06) \times 10^{-21}$ J and the position of this minimum of the potential to be $r = (2.47 \pm 0.12) \times 10^{-10}$ m. Determine the values and uncertainties of A and B.

2.3 A hospital ward has 10 beds. Each day an average of 6.8 patients arrive (all staying only one night). When the number of patients exceed the number of beds, they are send to another hospital.

- What is the chance of being send on as a patient?
- How many beds will be occupied on average?
- If the cost of transport is five times that of a bed, what is the (financially) optimal number of beds? And how many are transported a day then?

III – Monte Carlo:

3.1 Let $f(x, y) = e^{-x-y}$ be proportional to a two dimensional PDF for $x, y \in [0, \infty]$.

- Which method should be used to generate numbers according to $f(x, y)$? Explain?
- Make an algorithm, which from a uniform distribtion in the interval $[0, 1]$ generates values of x and y following the PDF $f(x, y)$. Plot the result of of this algorithm.
- Determine the size of the volume $\int_C f(x, y)dxdy$ and its uncertainty, where $C = \{x, y|(x-2)^2 + (y-2)^2 < 1\}$ by using 100.000 points.

3.2 On a floor made of parallel wooden strips of width L, you randomly drop a stick of length $l < L$.

- Show that the probability for the stick to lie across a line between two strips is $2l/\pi L$.
- Make a simulation that throws 1000 sticks, and from these give an estimate of π with uncertainty.
- For what value of l/L does this simulation give the most precise result?
- If a person using $l/L = 5/6$ got 113 crossings out of 213 throws, what value of π would he/she obtain? How close to the true value of π is this? How many throws would one normally have to do, to obtain such a high accuracy? Is the "113 out of 213" result realistic?

IV – Estimators:

4.1 Benford's ("first-digit") Law states that leading digits ($d \in 1, \ldots, 9$) occur with probability $P(d) = \log_{10}(1+1/d)$. Below is a table showing the first digit of countries population.

Digit	1	2	3	4	5	6	7	8	9
Frequency	64	43	24	31	15	20	17	11	12

- Test if country populations follow Benford's Law.

4.2 The table below shows the Sri Lankan population from 1871 to 1981.

Year (after 1900)	-29	-19	-9	1	11	21	31	46	53	63	71	81
Population (10^6)	2.3	2.6	3.0	3.5	4.1	4.4	5.3	6.6	8.1	10.6	12.7	14.9
Uncertainty (10^6)	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.4	0.4

- Does the Sri Lankan population follow an exponential distribution?
- Imagine a constant population, where the onset of better medical care and more efficient food production makes the population growth exponential. Does this hypothesis fit the data better?
- In 1802 a British administrator estimated that the Sri Lankan population was $(1.55 \pm 0.18) \times 10^6$. Given this additional information, what would be your best estimate of the Sri Lankan population in 1802?

V – Fitting data:

5.1 Below is a table of the goals scored in the 198 Danish Superliga games of the 2011-2012 season. There were never more than 5 goals scored by any team in a single match.

- What is the average number of goals scored home and away, respectively? Are the two numbers compatible? Are the two distributions compatible?
- Is the number of goals scored at home Poisson distributed? How about away goals?
- Is the fact that one team scores uncorrelated with the other team scoring (regardless of number of goals)?

Goals	Home					
Away	0	1	2	3	4	5
0	13	18	15	6	2	3
1	12	26	25	7	2	1
2	12	13	8	2	2	0
3	1	11	3	2	1	1
4	2	5	4	0	0	0
5	1	0	0	0	0	0

The table shows the number of matches with the score indicated, e.g. there were 15 matches, where the home team won 2-0.