Applied Statistics

Final exam in applied statistics

The following problem set is the take-home exam for the course applied statistics. It will be distributed Thursday morning the 28th of October 2010, and a written or electronic solution should be handed in by noon Friday the 29th. Working in groups is **not** allowed.
The use of computers is both allowed and recommended along with modifications of the programs you've worked with. For some of the problems, the use of computers will be necessary.

Good luck, Troels

I – Distributions and probabilities:

1.1 Little Peter has dropped gambling and is sitting on Blegdamsvej counting red cars. He has been told, that 30% of the cars passing are red. He observes 50 cars out of which 6 are red. Which distribution should the number of red cars follow, and what is the probability of this or less given the expected rate?

1.2 The members of a larger collaboration running the gigantic neutrino observatory Kamiokande in Japan, observed 11 neutrinos in their apparature on Monday the 23rd of February 1987. This coincided with the Supernova 1987A.

- What is the probability of observing 11 or more neutrinos in one day, if the expected average rate is 2.1 neutrinos per day?
- If the expected rate of 2.1 neutrinos per day was measured over 100 days, what is then the uncertainty on this number?

(Actually, the 11 neutrinos arrived with a period of 13 seconds, leaving no doubt if they were signal or not!)

II – Error propagation:

2.1 Using seven independent data samples (different directions) from the WMAP satellite, the age of the Universe has been determined. The results (in 10^9 years) were:

14.4 ± 0.7	14.3 ± 0.8	13.6 ± 0.4	13.2 ± 1.0	13.1 ± 0.5	15.9 ± 0.6	14.2 ± 0.6

- What is the mean age and the uncertainty on that mean? And what is the χ^2 and probability of the data matching a common mean?
- Do you find one of the measurements unlikely? Argue, if this is the case and remove the measurement, if you find reason for it.
- The star HE 1523-0901 has been measured to be $(13.2 \pm 0.3) \times 10^9$ years old by ESO's Very Large Telescope. What is the probability that the combined measurement from WMAP lies below the age of this star?

2.2 The distance d a ball will fly given an initial velocity of v_0 at an angle θ is given by $d = 2v_0^2 \cos(\theta)\sin(\theta)/g$, where $g = 9.80\text{m/s}^2$ is the gravitational acceleration. In an experiment the initial velocity is measured to be $v_0 = (3.53 \pm 0.15)\text{m/s}$ and the angle $\theta = (0.42 \pm 0.06)\text{rad}$.

- Assuming no correlations between v_0 and θ, what is the expected result and uncertainty on the distance d?
- What is the result, if v_0 and θ are 70% linearly correlated (i.e. $\rho_{v_0,\theta} = 0.7$)?

III – Monte Carlo: (For this part the use of computers is adviced. Plots can be enclosed in the solution).

3.1 Let a Monte Carlo algorithm generate 1000 squares with side lengths $a = 1.03 \pm 0.24$ and $b = 1.07 \pm 0.19$ where the uncertainties are Gaussian and uncorrelated.

- Plot the distribution of square areas A $(= ab)$, and determine the mean and the width.
- Compare the width with the result of analytically propagating the uncertainties.
- What is the probability of getting a negative area?

3.2 Let $f(x) = \frac{1}{\pi}(1 - \sin(2x))$ be a PDF for $x \in [0, \pi]$.

- By which method would you generate random numbers according to this PDF?
- Produce an algorithm, which from a uniform distribution in the interval $[0, 1]$ generates 1000 randon numbers following the PDF $f(x)$. Calculate the average of these numbers and the uncertainty on the average. Compare this to the analytical value for the average.

IV – Statistical tests:

4.1 In searching for ancient DNA, a test statistic t is developped to distinguish between DNA and non-DNA samples. For non-DNA the test statistic t follows a unit exponential distribution, while the DNA is Gaussianly distributed with a mean of 2.7 and a width of 0.4. A test is constructed which selects DNA samples by requiring that $t > 1.9$.

- What is the probability for the test of selecting a DNA sample and rejecting a non-DNA sample?
- Assume that non-DNA samples are 100 times more frequent than DNA samples. What is the purity (i.e. fraction out the total) of DNA samples selected by the criteria $t > 2.4$?
- What is the maximum purity, that can be obtained, considering any (small) interval in t?

V – Fitting data:

5.1 An experiment in the basement of NBI is measuring the lifetime of the muon. The number of decays in each 1μs time interval was measured as follows:

Time (s)	0.50	1.50	2.50	3.50	4.50	5.50	6.50	7.50	8.50	9.50
Counts	110	66	50	47	34	16	25	28	23	25

- Fit this data with an exponential function. What uncertainty do you ascribe each measurement? And is the fit good?
- It turns out that there is a constant background. Include this in your fit, and repeat it. Does this improve your fit significantly?
- A subsequent measurement shows that the background is 19.5. Given this knowledge, how much (in relative precision) does this improve your lifetime measurement?

Coincidences, in general, are great stumblingblocks in the way of that class of thinkers who have been educated to know nothing of the theory of probabilities – that theory to which the most glorious objects of human research are endebted for the most glorious of illustration.

[Edgar Allan Poe (1809-1849), The murders in the Rue Morgue]