Applied Statistics

Problem set in applied statistics

The following is a problem set on the curriculum lectured on so far. It will be handed out Friday the 30th of September 2011, and a solution in writing should be handed in by Monday the 10th of October at the latest. Working in groups is allowed, but separate solutions are required. The use of computers is both allowed and recommended along with modifications of the programs you've worked with.

Good Luck, Troels

Those who are good at archery learnt from the bow and not from Yi the Archer. Those who know how to manage boats learnt from boats and not from Wo [the legendary boatman]. Those who can think learnt for themselves and not from the sages.

[Guan Yin, 8th Century]

I – Distributions and probabilities:

1.1 Little Peter rolls a normal die 13 times and obtains 7 sixes. What is the probability of obtaining this result, or something more extreme? Did he cheat?

1.2 A beam of particles contains 10^{-4} parts electrons the rest being photons. The particles pass through a detector with two layers, which gives the signals 0, 1 or 2 depending on how many layers the particle is detected in. The probability for these signals are for electrons (e) and photons (γ) as follows:

$$\begin{array}{lll} P(0|e) = 0.001 & \text{and} & P(0|\gamma) = 0.99899 \\ P(1|e) = 0.01 & & P(1|\gamma) = 0.001 \\ P(2|e) = 0.989 & & P(2|\gamma) = 0.00001 \end{array}$$

- What is the probability for a particle to be a photon, if it is detected in one layer?
- What is the probability for a particle to be an electron, given signal in two layers?

1.3 Let x be a uniformly distributed PDF in the interval $[\alpha, \beta]$, where $0 < \alpha < \beta$.

- What is the expectation value and the variance of $1/x$?
- Compare the expectation value $E[1/x]$ with $1/E[x]$.

II – Error propagation:

2.1 Tennis balls are officially required to weight between 56.0 and 59.4 grams. For which mean and uncertainty will 90% of tennis balls lie within these requirements?
If a tennis serve has a speed of (73 ± 3) m/s, what is then the kinetic energy of the (official) ball, if there are no correlations? And with a correlation of $\rho_{mv} = -0.7$?

2.2 If $\theta = 0.54 \pm 0.02$, what is then the uncertainty on $\cos\theta$, $\sin\theta$, and $\tan\theta$? What if $\theta = 1.54 \pm 0.02$?

2.3 Snell's Law states that $n_1 \sin\theta_1 = n_2 \sin\theta_2$. Find n_2 and its error from the following measurements:

$$\theta_1 = (22.03 \pm 0.2)^\circ \qquad \theta_2 = (14.45 \pm 0.2)^\circ \qquad n_1 = 1.0000$$

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

3.1 Let $f(x) = \frac{1}{2\sqrt{x}}$ be a PDF for $x \in [0, 1]$.

- Which method should be used to generate this distribution? Why?
- Make an algorithm, which from a uniform distribution of random numbers in the interval $[0, 1]$, generates 1000 numbers following the PDF $f(x)$. From these calculate the average and the uncertainty on the average of $f(x)$.
- Compare this value with the analytically calculated average.

IV – Estimators:

4.1 Consider the dataset on Tibetan skull sizes, which can be found at:

`http://www.nbi.dk/~petersen/Teaching/Stat2011/Data_TibetanSkulls.txt`

- What is the mean and width of skulls of type A and type B for each of the five variables? And how large is the separation (i.e. $(\mu_A - \mu_B)/\sqrt{\sigma_A^2 + \sigma_B^2}$) between skulls of type A and type B for each of the variables?
- What are the linear correlations between the five variables for each type of skull?
- (**Bonus question**) How well can you separate the two types of skulls for example using a Fisher discriminant? Is there a new human race in Tibet?

4.2 Consider the classic 1910 dataset on Polonium 210 decays by Rutherford and Geiger, showing the number of decays in a 72s period for 2608 periods:

Counts	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Frequency	57	203	383	525	532	408	273	139	45	27	10	4	0	1	1

- Argue what distribution these counts should follow and test if they do indeed. What is the average level of activity?
- Calculate the χ^2 and $-2\ln(\mathit{likelihood})$ for a range of activity levels around the average value. Plot these and determine the (possibly asymmetric) statistical uncertainty on the average level both using the χ^2 and the likelihood.
- Given the lifetime of Polonium 210 and the time used for the experiment, which systematic uncertainty would you ascribe this result? Should Rutherford and Geiger actually have conducted a shorter experiment? Or reported the data differently?

V – Fitting data:

5.1 An experiment gave the following binned (counting) result, where the uncertainty on y, σ_y, has been estimated to be the square root of the number of events plus a systematic uncertainty of 1.5 to be added in quadrature:

x	y	x	y	x	y	x	y	x	y
0.05	14	0.45	9	0.85	13	1.25	12	1.65	5
0.15	7	0.55	17	0.95	9	1.35	9	1.75	10
0.25	6	0.65	23	1.05	8	1.45	3	1.85	9
0.35	9	0.75	19	1.15	14	1.55	11	1.95	6

- Assume a linear relation between x og y, and make a χ^2-fit to data. Is the fit good?
- If one expects a Gaussian signal somewhere in the region 0.5 to 0.8, what would you then fit with? Discuss the validity of the fit and the significance of the signal.