Applied Statistics

Problem Set in applied statistics 2015

The following problem set is for the course applied statistics. It will be distributed 4th of December 2015, and an electronic solution in writing sent by email to petersennbi.dk must be handed in by 22:00 Thursday the 17th of December. Working in groups is allowed. The use of computers is both allowed and recommended.

Good luck and thanks for all your hard work so far, Troels.

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

I – Distributions and probabilities:

1.1 (7 points) Two chess players play 50 non-draw games to determine who is the best.

- What distribution will the number of victories for the first player follow?
- If the two players are equally good, what is the chance that the first player wins 30 games or more? And with 30 wins, can he then claim to be the better player?

1.2 (7 points) Let x be distributed according to the PDF $f(x) = a(4x - x^3)$, $x \in [0, 2]$.

- For what value of a is $f(x)$ normalized?
- What is mean and width of x?

II – Error propagation:

2.1 (10 points) The energy of a damped harmonic oscillator is $E = Cme^{-bt/m}$, where $C = 1J/g$. At $t = 1s$ the following measurements were made: $m = 12.5 \pm 1.5$ g and $b = 0.91 \pm 0.15$ g/s.

- Assuming no correlations, what is the energy and its uncertainty?
- If the mass m and damping constant b are linearly correlated by -90%, what is the answer then?

2.2 (10 points) Consider the classic 1910 dataset on Polonium 210 decays by Rutherford and Geiger, showing the number of decays in a 7.5s period for 2608 periods:

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

- Find the mean number of decays per period and its uncertainty.
- If the source was 1.24 ± 0.05 years old, and given an exponential decay with half-life 138.4 days, what was its initial mean number of decays per period and its uncertainty?

Statistics ... is the most important science in the whole world, for upon it depends the practical application of every other science and of every art.

[Florence Nightengale (1820-1910)]

III – Monte Carlo:

3.1 (15 points) Let $f(x) = \frac{c}{\sqrt{x}}$ be proportional to a PDF for $x \in [0, 1]$.

- For what value of c is $f(x)$ normalized, and what is the mean and width of $f(x)$?
- What method would you use to produce random numbers according to $f(x)$. Why?
- Produce an algorithm that generates random numbers according to $f(x)$, and use 10000 such numbers to numerically determine the mean and width of $f(x)$.
- Fit the distribution of 10000 random numbers with $f(x)$ in the range $x \in [0.01, 1]$, possibly with a floating exponent of x. Share your thoughts on choice of binning!

3.2 (12 points) Let $f(x) = ae^{-x}cos^2(x)$ be a PDF for $x \in [0, \infty]$.

- What method would you use to produce random numbers according to $f(x)$? Why?
- Produce 100000 random numbers according to $f(x)$ and plot these elegantly.
- From the algorithm calculate the normalization a and compare to analytic value.

IV – Fitting data and optimization:

4.1 (12 points) The probability of obtaining a High Threshold hit (pHT) in the ATLAS Transition Radiation Tracker depends on the logarithm of the γ-factor of the particle traversing. In the file **www.nbi.dk/~petersen/data_problem41.txt**, 60 measurements of pHT including uncertainties for various values of $\log(\gamma)$ can be found.

- At low values of $\log(\gamma)$, pHT is constant. Up to what value of $\log(\gamma)$ is it consistent with being constant? And what (constant) value of pHT do you find?
- Fit the entire distribution with suitable function(s), and discuss which one best describes this distribution.

V – Statistical tests:

5.1 (15 points) From independent measurements of diameter (d) and (cross sectional) area (A), 180 students have measured the volume of the dwarf planet Orcus, assuming that Orcus is spherically shaped. The file **www.nbi.dk/~petersen/data_problem51.txt** contains their results, where the units are km and km^2.

- Using all measurements, what are the means and the uncertainty on the means of the diameter and area measurements, respectively?
- Expecting good measurements of both d and A to follow a Gaussian distribution, would you consider discarding data points, and if so why and how many?
- Estimate the volume of Orcus based on the diameter and area measurements separately. Are the two results consistent?
- Given a mass of $(6.1 \pm 0.2) \times 10^{20}$ kg, what is the mean density?

5.2 (12 points) 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 frequency of first digits of countries size measured in km^2 and miles^2 (km^2/miles^2).

Digit	1	2	3	4	5	6	7	8	9
Frequency	58/56	34/37	22/20	21/17	10/14	14/14	11/14	7/12	10/4

- Test if country sizes in km^2 and miles^2 follow Benford's Law.
- Are the two distributions consistent with being from the same underlying distribution?