Applied Statistics

Exam in applied statistics 2015/2016

The following problem set is the take-home exam for the course applied statistics. It will be distributed Thursday the 21st of January 2016, and a solution in PDF format sent by email must be handed in by noon Friday the 22nd of January 2016, along with all code used to work out your solution. Working in groups or discussing the problems with others is **not** allowed. The use of computers is both allowed and recommended.

Good luck and thanks for all your hard work, Troels, Mathias & Niccolo.

Statistics may be defined as "a body of methods for making wise decisions in the face of uncertainty".
[W.A. Wallis, US statistician 1912-1998]

I – Distributions and probabilities:

1.1 - 2015 (7 points) You pick 20 M&M's from a (large) bag, where 5% are blue (you are told).

- What distribution does the number of blue M&M's you picked follow? Why?
- If you get six blue M&M's, do you trust the fraction of these to be 5%?

1.2 - 2015 (6 points) Two bags contain red and blue M&M's. Bag 1 has 50% red M&M's, while bag 2 has 75% red M&M's. You pick a random M&M from a random bag, and it is red.

- What is the probability that you picked an M&M from bag 1?

II – Error propagation:

2.1 (8 points) An experiment to measure resistance R using Ohm's Law $V = RI$ yields the measurements $V = 5.01 \pm 0.07V$ and $I = 1.178 \pm 0.013A$.

- Assuming no correlations, what is the resistance and its uncertainty?

The experiment was designed to have a linear correlation between V and I of -0.95.

- What is the uncertainty on R then?

2.2 (8 points) Given $x = 0.58 \pm 0.02$, what is the value and uncertainty of $\exp(x)$, $\sin(x)$ and $\tan(x)$? And what if $x = 1.58 \pm 0.02$? Also, comment on the degree of Gaussianity of the uncertainties.

2.3 (10 points) The age of the Earth can be estimated from the current amounts of U238 and U235 and their ratio, $f = U235/U238$. Assume equal amounts at Earth's creation (i.e. $f(t=0) = 1$) and half-lifes $\tau_{U238} = (4.5 \pm 0.5) \times 10^9 y$ and $\tau_{U235} = (0.704 \pm 0.032) \times 10^9 y$.

- Given a current ratio $f = 0.0072 \pm 0.0003$, what is your estimate of the age of the Earth?
- How much does the result change if $f(t=0)$ is rather uncertain, e.g. 0.5 ± 0.1?

III – Monte Carlo:

3.1 (12 points) Let $f(x) = C \sin^2(\pi/x)/\sqrt{x}$ be a PDF for $x \in [0.05, 1.0]$.

- 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.
- In order for this PDF to be normalized, what value should C have?
- Perform a fit to the produced data points. Do you manage to get a good χ^2?

IV – Fitting data:

4.1 (12 points) The data **www.nbi.dk/~petersen/data_MuonLifetime.txt** contains the results of an experiment that measures muon decay times. The exponential signal has a background at very short times resulting from random noise around $t = 0$, and one that is constant in time.

- Determine the constant background by fitting a suitable range at high times.
- Extract the lifetime of the decay along with its uncertainty.
- Fit the entire distribution with a suitable function. Do you obtain a good fit?

4.2 (12 points) You are measuring the gravitational acceleration g by letting a magnet drop a ball, which you then measure the position (d) of at certain times (t). The data can be found in the file **www.nbi.dk/~petersen/data_FreeFall.txt**. The Gaussian uncertainties are $\sigma_t = 0.001$s and $\sigma_d = 5$mm. To begin with, consider only the first 8 data points.

- Assuming that $d(t{=}0) = 0$ and $v(t{=}0) = 0$, determine g and its uncertainty.
- Is the ball released at $t = 0$? Repeat the fit, and measure the possible offset in time, Δt.
- Now consider all 20 points and test which of the following three hypothesis matches the data best, where τ (characteristic time) and v_∞ (terminal velocity) are parameters:
  ⋆ No air drag: $d(t) \sim \frac{1}{2}gt^2$
  ⋆ Linear drag: $d(t) \sim g\tau(t - \tau(1 - e^{-t/\tau}))$
  ⋆ Quadratic drag: $d(t) \sim v_\infty^2 \ln(\cosh(gt/v_\infty))/g$

V – Statistical tests:

5.1 (15 points) The file **www.nbi.dk/~petersen/data_KentuckyDerby.txt** contains the winning times 1874-2015 of the famous Kentucky Derby. First, consider the data from 1950 and onwards.

- Have racing horses gotten faster? As always, please quantify your answer.

The course length was 1.5 mile from 1874-1895, and 1.25 mile after that. The variations in winning times are mainly due to racing conditions. Assume an uncertainy of 2.5s before 1950, and 1.5s from then on.

- Fit the data from 1874-2015, possibly with a step at 1895-1896.
- Do any time(s) seem extreme? Quantify this, and possibly exclude point(s).
- Test the hypothesis that the size of the step 1895-1896 corresponds to the distance ratio?

5.2 (10 points) In 1929 Edwin Hubble investigated the relationship between distance (D) and radial velocity (v) of extra-galactic nebulae. His original data from 1929 is listed below.

D (Mpc)	0.032	0.034	0.214	0.263	0.275	0.275	0.45	0.5	0.5	0.63	0.8	0.9
v (km/s)	170	290	-130	-70	-185	-220	200	290	270	200	300	-30
D (Mpc)	0.9	0.9	0.9	1.0	1.1	1.1	1.4	1.7	2.0	2.0	2.0	2.0
v (km/s)	650	150	500	920	450	500	500	960	500	850	800	1090

- Assuming uncertainties of 12% on the distance and 170 km/s on the radial velocity, fit the data to extract the constant of linear proportionality H_0 (Hubble constant), $v = H_0 D$.
- Hubble's distance measurement was biased by a factor 5.3 ± 0.3 due to the existance of two types of Cepheids. Given this correction, how good is the agreement with the modern value of the Hubble constant $H_0 = 67.80 \pm 0.77$.

Numbers are like people; torture them enough and they'll tell you anything. [Unknown]