University of Copenhagen Niels Bohr Institute, 6. december 2018

Applied Statistics

Problem Set in applied statistics 2018/19

This problem set was distributed Friday the 7th of December 2018, and a solution in PDF format must be submitted via the course webpage on Absalon by Sunday the 6th of January 2019 at 22:00. Links to data files can also be found on the course webpage. Working in groups or discussing the problems with others is (unlike at the exam!) allowed, but you should state your collaboration(s).

Thanks for all your hard work so far, Troels

Statistics like veal pies, are good if you know the person that made them, and are sure of the ingredients.
[Harvard President Lawrence Lowell, 1909]

I – Distributions and probabilities:

1.1 (5 points) How many times do you have to roll a normal die ($p_{six} = 1/6$) to be 99% sure of having rolled at least one six?

1.2 (5 points) A high jumper experiences that she can clear 2.01m 13% of her jumps and 1.93m 87% of her jumps. Estimate the μ and σ of her (assumed Gaussian and consistent) jumping performance?

1.3 (5 points) In Palm Springs it rains 14 days a year. On rainy days, the forecast predicted it in 80% of the cases, while on sunny days the forecast is (wrongly) for rain in 10% of the cases. Given a forecast for rain, what is the chance that it will actually rain?

1.4 (8 points) The mean number of occurances of a species (A) of a plant is found to be 7.1/km^2.

- What distribution should the number of plants in a fixed area follow? What is the chance of finding four or more of this plant in 0.3 km^2?
- For another species (B) the occurance is 12.6/km^2. In an area of size 0.2 km^2, what is the chance of finding exactly two of each species? And finding more than four plants in total?

II – Error propagation:

2.1 (6 points) Given Snell's Law ($n_1 \sin\theta_1 = n_2 \sin\theta_2$) and the measurements $\theta_1 = 1.282 \pm 0.007$, $\theta_2 = 0.671 \pm 0.004$ and $n_1 = 1.0003 \pm 0.0001$, determine n_2.
Plate, Crown, and Flint glass have index of refraction n_2 of 1.52, 1.54, and 1.60, respectively. Which of these materials could the above measurement come from? Quantify!

2.2 (9 points) The challeging measurement of the W boson mass has been done by seven experiments:

Measurement	ALEPH	Delphi	Opal	L3	CDF	D0	ATLAS
Result (GeV)	80.440	80.336	80.415	80.270	80.387	80.367	80.370
Uncertainty (GeV)	0.051	0.067	0.052	0.055	0.019	0.026	0.019

- Assuming independent measurements, what is the W mass and its uncertainty?
- The Electro-Weak fit with (without) the Higgs boson included, predicts a W mass of 80.358 ± 0.008 GeV (80.249 ± 0.008 GeV). How consistent are these with the measured W mass?
- If the measurements all have a common systematic uncertanty of 0.011 GeV (i.e. this part of the uncertainties is 100% correlated between measurements), what average is then obtained?

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

III – Monte Carlo:

3.1 (18 points) Let $f(x) = C\exp(-p_0 x)\ln(1+\sin^2(p_1 x))$ be a PDF for $x \in [0,2]$, where $(p_0, p_1) = (1, 3\pi)$. Let $g(x)$ have the same formula as $f(x)$, but defined for the range $x \in [0, \infty]$.

- What method would you use to produce random numbers according to $f(x)$? Why?
- Produce 1000 random numbers distributed according to $f(x)$ and plot these.
- Fit these numbers with $f(x)$, where p_0 and p_1 are left floating. Do they match the input?
- Let u be a sum of 10 random values from $f(x)$. Produce 1000 values of u and test if they follow a Gaussian distribution.
- How would you produce numbers according to $g(x)$?
- Let v and w be a sum of 75 and 50 random values from $f(x)$ and $g(x)$, respectively. Generate 100 values of v and w. Can you tell the difference between these two distributions?

IV – Statistical tests:

4.1 (14 points) The data file **www.nbi.dk/~petersen/data_WomenAndMen.txt** contains the height, auditory ability, loudness tolerance, eye-sight score, IQ and gender (0: Female, 1: Male) for 1923 women and 1372 men, respectively.

- Calculate the mean height and its uncertainty for these 3295 persons.
- Calculate the mean height for women and men separately, and combine the results in a weighted mean. Do you get the same result as above?
- What is the auditory ability and loudness tolerance correlation for women? Men? Both?
- Based on these variables, how well are you able to distinguish between men and women?

4.2 (12 points) Below is a table of the results from the 3420 English Premier League games from 2008-2017. There were never more than 9 goals scored by any team in a single match.

- What is the average number of goals scored per match by the home and away teams, respectively? Is there a significant home advantage?
- Is the number of goals scored by the home team Poisson distributed? How about away team goals?
- Is the fact that one team scores uncorrelated with the other team scoring (possibly regardless of number of goals)?

Goals	Away										
Home	0	1	2	3	4	5	6	7	8	9	Sum:
0	278	234	137	82	30	12	3	0	0	0	776
1	346	355	229	102	33	8	5	1	0	0	1079
2	279	299	192	68	15	4	0	0	0	0	857
3	162	161	76	36	8	2	2	0	0	0	447
4	70	51	29	14	5	1	0	0	0	0	170
5	25	19	6	4	1	1	0	0	0	0	56
6	9	9	3	2	0	0	0	0	0	0	23
7	2	3	1	1	0	0	0	0	0	0	7
8	3	0	1	0	0	0	0	0	0	0	4
9	0	1	0	0	0	0	0	0	0	0	1
Sum:	1174	1132	674	309	92	28	10	1	0	0	3420

The table shows the number of matches with the score indicated, e.g. there were 279 matches, where the home team won 2-0.

V – Fitting data:

5.1 (18 points) The Arecibo Observatory was in 1974 used for detecting the Hulse-Taylor pulsar. Amplitudes recorded every 0.01s for three observation periods of a few seconds (several seconds apart) can be found at **www.nbi.dk/~petersen/data_HulseTaylor.txt**.

- Consider the first **calibration** run (about 1-6s) consisting of noise. What is the distribution of the amplitudes? What uncertainty would you, based on this, ascribe single measurements?
- In the **first observation run** (about 8-12s), how consistent is the data with being constant? Also, fit the distribution with an oscillating function. Is this fit hypothesis reasonable?
- Fit the **second observation run** (about 20-24s) with the same oscillating function. Do the fitted period and noise level match the ones obtained from first observation run fit?
- Now try to fit both observation runs simultaneously in one fit spanning both observation runs. Do you get a good fit? In any case, see if you can improve the fit by incorporating other effects.