Applied Statistics

Problem Set in Applied Statistics 2019/20

This is the problem set for Applied Statistics 2019/20. A solution in PDF format must be submitted on Absalon by 22:00 on Sunday the 5th of January 2020. Links to data files along with code to read the data can be found on the course webpage. Working in groups and discussing the problems with others is allowed. However, you should state your collaboration(s).

Happy solving, Troels, Sebastien, Etienne, Giulia, John & Nikki

---

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

I – Distributions and probabilities:

1.1 (6 points) Little Peter goes to the casino and puts money on black ($p = 18/37$).

- In 50 games, what are the chances that he will win exactly 25 times? 26 times or more?
- How many times does he have to play in order to be 95% sure of winning at least 20 times?

1.2 (4 points) What is the probability of a Gaussian value to lie between 1.25σ and 2.5σ away from the mean?

1.3 (6 points) The number of S-train delays is counted daily. Assume in the following, that delays are uncorrelated, and that the number of departures is the same every day.

- What distribution should the number of daily delays follow?
- Days with more than 7 delays are considered "delay days". If there were 19 "delay days" in a normal year, what is your estimate for the average number of daily delays?

II – Error propagation:

2.1 (13 points) A measurement of a tumor depth (in cm) was done using two methods. The first gave 4 measurements with uncertainty while the second gave 12 without, as shown in the table.

With unc.	2.05 ± 0.11			2.61 ± 0.10			2.46 ± 0.13			2.48 ± 0.12		
Without unc.	2.69	2.71	2.56	2.48	2.34	2.79	2.54	2.68	2.69	2.58	2.66	2.70

- Do the measurements with uncertainty agree with each other? Do those without?
- Which of the two methods provide the most precise positioning?
- What is your best estimate of the tumors position? And with what uncertainty?

2.2 (9 points) The spectral radiance B of a body is given by Planck's Law: $B(\nu, T) = \frac{2h\nu^3}{c^2} \frac{1}{e^{\frac{h\nu}{k_B T}} - 1}$ where ν is the frequency and T is the absolute temperature, while $h = 6.626 \times 10^{-34}\ Js$, $c = 299.7 \times 10^6\ m/s$, and $k_B = 1.381 \times 10^{-23}\ J/K$ are constants of nature.

- Given values of $\nu = (0.566 \pm 0.025) \times 10^{15}$ Hz and $T = (5.50 \pm 0.29) \times 10^3$ K (uncorrelated), what is the expected spectral radiance, B?
- How does the uncertainty change, if there is a correlation of $\rho(\nu, T) = 0.87$?

III – Monte Carlo:

3.1 (15 points) Let $f(x)$ be a PDF defined as $f(x) = C(1 - e^{-ax})$ for $x \in [0, 2]$ and $a = 2$.

- What is the mean and RMS of $f(x)$? Also, what is the value of C?
- What method(s) can be used to produce random numbers according to $f(x)$? Why?
- Produce 500 random numbers distributed according to $f(x)$ and plot these.
- Fit the numbers you produced above leaving a as a floating parameter.
- Let u be a sum of 5 random values from $f(x)$. Produce 1000 values of u and test if they are consistent with a Gaussian distribution?

IV – Statistical tests:

4.1 (15 points) The National UFO Reporting Center (NUFORC) has since 1974 catalogued reported UFO sightings. A subset of the data with 64719 entries containing date, time, place, shape, and duration of observation can be found at **www.nbi.dk/∼petersen/data_UfoSightings.txt**.

- Plot the distribution of duration of observation, and calculate both mean and median.
- Do these durations follow the same distribution on the East and West coast?
- What is the correlation between day in the year and time of the day of observation?
- Considering only the West Coast, is the distribution of number of observations uniform over the seven week days? How about when considering only Monday to Thursday?

4.2 (13 points) To test the fairness of dice, you roll 12 dices and count the number 5s and 6s (N_{56}). Repeating this many times yielded the following result:

Number of 5s & 6s	0	1	2	3	4	5	6	7	8	9	10	11	12
Observed frequency	185	1149	3265	5475	6114	5194	3067	1331	403	105	14	4	0

- What distribution should the number of 5s and 6s follow?
- Compare the data with the expected distribution. Does this hypothesis match the data well?
- Fit the data and test if alternative hypotheses match the data better. Also, determine the probability for a 5 or a 6, and decide if the dice are consistent with being fair.

V – Fitting data:

5.1 (19 points) Radioactive isotopes produce sharp peaks in gamma ray spectra by which they can be identified. The file **www.nbi.dk/∼petersen/data_GammaSpectrum.txt** contains 47173 measurements obtained from uranium ore. Each number refers to a *channel number* in the detector, which can be translated roughly linearly into an *energy* (in the 0-1000 keV range).

- The lead isotope ^{214}Pb produces three known, low energy gamma ray peaks at energies E (decay fractions f): 242 keV (7.4 %), 295 keV (19.3 %), and 352 keV (37.6 %). Fit these three ^{214}Pb peaks.
- For these three peaks, compare the *relative* distance $r = (E_3 - E_2)/(E_2 - E_1)$ (in channel number) with the corresponding tabular value (in energy). Does the relative distance match?
- The more energetic bismuth ^{214}Bi gamma rays produce peaks at 609 keV (46.1 %) and 1120 keV (15.1 %). From peaks of your choice, determine the energy scale (i.e. how to determine energy from channel number) and test if the relation is linear.
- Is the energy resolution (i.e. peak width) constant or does it change with energy?
- A theoretical calculation predicts a (small) peak in the spectrum in the range 700-800 keV. Does the data support that prediction? And if so, at what energy?
- Do you find any other peaks or features in the spectrum? If so, quantify your findings.

Don't worry too much about statistics! Just tell us what you do, and do what you tell us.
[Roger Barlow, ICHEP conference 2006, Moscow]