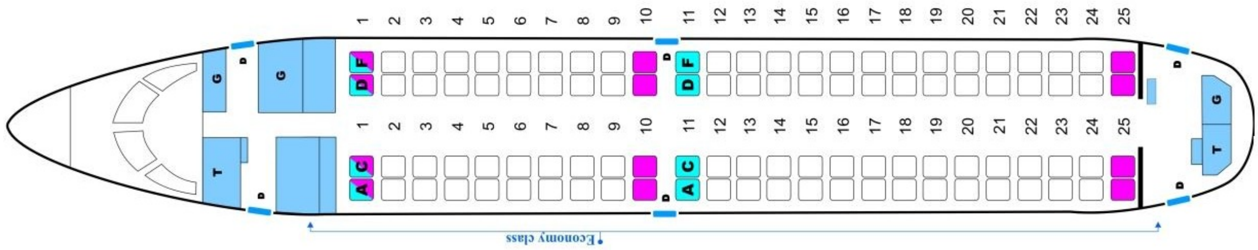
Home exam in “Statistical Methods I” 6–8.5.2014

(To be returned by 17.00 on Thu 8.5. either into locked box labelled “Home exam Statistical Methods I” at Physicums 2nd floor or by e-mail to kenneth.osterberg@helsinki.fi)

1. (i) The weights of the eggs produced by a farmer’s hens have a standard deviation of 10 g. He feeds some hens a vitamin supplement, which would be cost-effective if the weights of the eggs increase by at least 2 g. He measures 25 eggs from the vitamin-fed hens and the average weight has increased by 3 g. Is the increase significant? [1.5p]

   (ii) Utsjoki in Finnish Lapland is a good place to observe Northern lights. Assume they occur four out of five nights and the sky to be sufficiently cloudless for observation of them half of the nights. If you spend 5 nights in Utsjoki, what is the probability to observe Northern lights every night? What is the probability that you return home disappointed without having observed Northern lights at all? [1.5p]

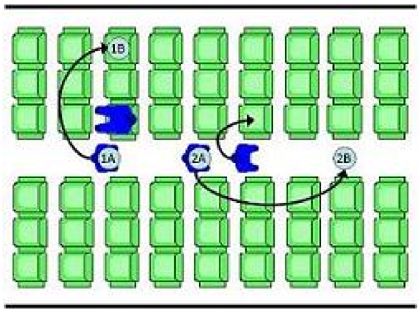
   (iii) A charged particle beam consists of 95 % pions and 5 % kaons (two types of elementary particles). We have a particle identification device which has the following performance: if a kaon traverses it, 90 % of the time the particle is identified as a kaon, 5 % of the time as a pion and the remainder of the time no decision is reported however if a pion traverses it, 80 % of the time the particle is identified as a pion, 5 % of the time as a kaon and the remainder of the time no decision is reported. Give the posterior probabilities that a particle is a pion and that a particle is a kaon (two probabilities per case below) given:
   a) The particle is identified as a pion
   b) The particle is identified as a kaon
   c) The particle identification device reports no decision [2p]



2. A classical Monte Carlo problem is simulation of the boarding of passengers on an airplane. Let’s look at the case of boarding an Embraer EMB 190 (see picture above). The seat rows are at 1 m distance and the distance from the gate to the first seat row is about 15 m. Assume that the plane is 100 % full and that people are randomly seated. Assume in addition that the average walking speed is 0.5 m/s. Include in your simulation ”aisle interference”, depicted by ”2A $\rightarrow$ 2B” in the figure below, i.e. passengers cannot get through to their seat because somebody is putting his/her luggage in the luggage storage above the seats and/or getting seated (assume it to take on average 20 s).

   (i) Simulate 100 ”outside-in” boardings i.e. that those with a window seat (”A” and ”F” in the picture above) board first and the those with aisle seat ("C" and "D" in the picture above) later. Assume random ordering in the boarding in terms of rows. Give as result the average and standard deviation of the total boarding time i.e. the time it takes for all passengers to be seated from your 100 simulations. [3.5p]

(ii) Another effect to take into account is "seat interference", depicted by "1A $\rightarrow$ 1B" in the figure below, i.e. that a passenger cannot get seated in his/her own window seat since somebody else is already seated in the aisle seat. Assume that such a swap takes on average 10 s. Simulate random boardings. Give as result the average and standard deviation of total boarding time. Is the "outside-in" boarding method of (i) significantly faster?[3.5p].



3. In an analysis, one needs to classify the measured events into two different categories: a and b. For each event two variables (x_1, x_2) are measured.

   (i) Build a new variable, x_3, as a linear (alternatively non-linear) combination of x_1 and x_2 for linear (non-linear) **optimal** discrimination between the 2 event categories. For this you have 2 training samples, `exam_sample_cata.txt` and `exam_sample_catb.txt`, found in the same folder as the exam paper that contains the x_1 (1. column) and x_2 (2. column) values of events from only category a and only category b, respectively. Plot the x_3 value for the events of category a and b. [2p]

   (ii) What is the requirement on x_3 to have a 95 % rejection of events from category a? What is the efficiency for accepting events of category b with this requirement? Plot the data points of the two training samples in the x_1x_2-plane using different colours. Add the x_3 requirement giving 95 % rejection of category a into the same plot. [1.5p]

   (iii) Assume observation of an event at $x_1 = 0.50$ and $x_2 = 0.25$. Calculate the probability that it belongs to category a and the probability that it belongs to category b. Assume the category-dependent probability distributions for both variables (x_1 and x_2) to be gaussian and that x_1 and x_2 are independent from each other for each category. [1.5p]

4. The specific heat capacity (C) of a substance is the amount of heat energy (E) required to change its temperature by a given amount per unit of mass m (see `http://en.wikipedia.org/wiki/Specific_`

heat_capacity). C can be expressed as $C = E/m\Delta T$, where ΔT is the change in temperature.

A way to determine C for a liquid is to heat it in a thermally isolated container with an electric heater while measuring the temperature of the system at different times. The heater produces as much heat energy in the time t as it consumes electrical energy E; the energy produced is $E = U \cdot I \cdot t$ (U is the voltage and I the current). A perfect transfer of the heat from the heater to the substance can be assumed.

A chemist uses an electric heater rated at $U = 12$ V and $I = 10$ A to heat $m = 1000$ g of an unknown liquid. He obtains the following results for the change in temperature at different times:

t (s)	352 ± 5	701 ± 9	1048 ± 9	1398 ± 9
ΔT ($°C$)	10.0 ± 0.1	19.7 ± 0.2	30.2 ± 0.2	40.4 ± 0.2
t (s)	1751 ± 9	2099 ± 15	2446 ± 15	2805 ± 15
ΔT ($°C$)	49.9 ± 0.3	60.5 ± 0.3	70.4 ± 0.4	80.0 ± 0.4

(i) Determine the specific heat capacity C of the liquid and its uncertainty, neglecting the uncertainties of the temperature measurements. You may assume that C is constant over this temperature range. What are the χ^2 and P-values of your solution? [2p]

(ii) Check if the variance of C obtained in (i) is at the RCF-bound.[1.5p]

(iii) Determine C and its uncertainty now taking into account both the uncertainties of the times and the temperatures. What is the change in χ^2 and P-values? [2.5p]

(iv) Are the chemist's error estimates reasonable? Is your result compatible with theory? [1p]

5. Group of particle physicists are looking for beyond Standard Model (SM) particles i.e. particles not contained in the SM, our best current model of elementary particles and their interactions. Before anybody can pick up his/her Nobel price for the discovery, somebody needs to validate that the real data is well modelled by the simulation.

   The file `real_mass.dat` contains mass measurements of real data events: first two columns are the boundaries of the bins (lower and upper), and third column the number of entries in the corresponding bin $n_i, i = 1, ..., 20$. The files `MC1_mass.dat` and `MC2_mass.dat` contains simulated events of two types of possible backgrounds to the non-SM particle production process of interest. The files mentioned above can be found in the same folder as the exam paper.

(i) Test whether the real data can be described only by the first background (`MC1_mass.dat`). Test also whether the real data can be described only by the second background (`MC2_mass.dat`). In both cases quantify your test with a relevant variable. What is the conclusion?[2p]

(ii) More realistic tests are obtained with pseudo-experiments based on expected background distributions. Generate pseudo data sample $data_{\text{pseudo}}$ with N_{data} events. Make the same tests as in (i). Are the quantitative results obtained in (i) and (ii) identical? If not, why? [2p]

(v) Find constant a, so that the combination "a*MC1 + $(1-a)$*MC2" best models data. Give optimal a and quantitive variable describing the similarity of the distributions for both direct and pseudo-experimental methods. Is the real data fully described by this optimal background combination? [2p]

6. The decay rate of orthopositronium (a bound state made up of an electron–positron pair with parallel J_z's and angular momentum $L = 0$ that due to conservation of the CP quantum number always decays into 3 photons) has been measured. The number of candidate decays measured in 10 ns time windows spaced by 50 ns starting from 300 ns after the creation of the bound positronium states are given in Table 1.

Table 1: Number of decays in 10 ns windows as function of time t.

Time (t/ns)	N	Time (t/ns)	N
300	803 000	550	188 487
350	581 083	600	150 737
400	429 666	650	124 103
450	320 016	700	105 397
500	242 783	750	92 748

The measured rates contain both background as well as signal. Assume that the decay rate, R, can be parametrized by following functional form:

$$R(t) \quad = \quad A\exp(-\lambda t) + B,$$

where decay constant $\lambda \equiv \ln 2/\tau$, τ is the lifetime and A & B constants proportional to the number of signal and background events.

(i) Determine λ, A and B. Estimate also their uncertainties. [4p]

(ii) Estimate the lifetime of orthopositronium and its uncertainty. [2p]