Solution for Applied Statistics take-home exam 2016

Problem 1.1

This is a problem of probability calculation, though one can use a Binomial distribution for solving (and even a Poisson distribution for the second problem).
In Game 1 the probability of **NO** six in 4 rolls is $(5/6)^4 = \mathbf{0.4823}$, which means that (at equal odds) this game is advantages to play (player has more than 50% chance of winning).
In Game 2 the probability of **NO** double six in 24 rolls is $(35/36)^{24} = \mathbf{0.5086}$, which means that (at equal odds) this game is not advantages to play (player has less than 50% chance of winning).

- **Thus Game 1 is worth playing ($p_{winning}$ = 0.5177), while Game2 is not ($p_{winning}$ = 0.4914).**

One can also just consider the number of possibilities (pos):
In Game 1 there are 671 pos of winning and 625 pos of loosing out of 1296 in total.
In Game 2 there are approximately $11.033.126.465.280 \times 10^{24}$ pos of winning and $11.419.131.242.070 \times 10^{24}$ pos of loosing out of $22.452.257.707.350 \times 10^{24}$ in total.

° **A word on the problem**
This problem is known as Chevalier de Mere's problem, and is famous for having launched probability theory, as de Mere's (loosing large sums in the second game) wrote Blaise Pascal for help.

Notes on points for problem 1.1: 4, 4:
Essentially, one gets 4 points for each of the two probabilities calculated correctly.
There is 1 extra point for mentioning Binomial distribution (though not needed for calculation).
There is 1 minus point for making wrong interpretation of results.
The misunderstand that it is *exactly* one/two sixes also costs 1 point, unless argued well.

Problem 1.2

- The daily rate should follow a **Poisson distribution** with λ = 18.90.

The probability of 42 or more (extreme) events is:

$$p(42+\text{events}) = \sum_{i=42}^{i=\infty} Pois(i,\lambda) = 0.00000317 \quad (1)$$

Thus the probability of observing 42 events in a day is very low. However, the significance of a daily rate of 42 would a combination of probability and trial factor! The trial factor (i.e. number of days possible) is 1730.

- Thus the overall probability is:

$$p = 1-(1-p(42+\text{events}))^{Ntrials} = \mathbf{0.00547427} \quad (2)$$

This is still very low, but the significance is not quite as great, as some might think.

For comparison, Gaussian approximation gives a probability of $p(42+\text{events}) = 0.00000005$, which shows that for the tail, the Gaussian is not a good approximation at this low λ.

Notes on points for problem 1.2: 3, 4:
1.2.1: Minus 1 for arguing wrongly for the Poisson.
1.2.2: Minus 1 for forgetting Trial Factor of 1730, but "knowing" (i.e. writing interpretation), while minus 2 if not thinking of this at all. The global probability is low (i.e. no signal), but other conclusion OK if argued well.
Gaussian approximation is not accurate (but a typical thing to do), thus minus 1 point.

Problem 1.3

- The fraction of women taller than 1.85 (i.e. $(1.85-1.68)/0.06 = 2.83$ sigma, one sided) is:

$$f = \int_{1.85}^{\text{inf}} Gaus(1.68, 0.06) = \int_{2.83}^{\text{inf}} Gaus(0,1) = \mathbf{0.00230327} \quad (3)$$

- The cut to get top 20% tallest women is obtained from inverting the above integral to yield 20%, the result being 0.842σ, which corresponds to $1.68\,m + 0.06\,m * 0.842 = 1.730\,m$, though one can also "just" take the top 20% of randomly produced numbers. The average height of the 20 percent highest women is **1.764 m**. Note that if solved nummerically, the result should preferably have an uncertainty or a mention of this (the size of which of course depends on the number of points used).

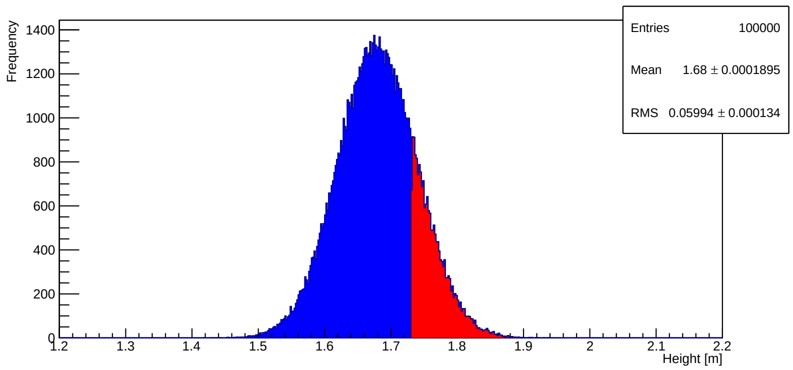

Figure 1: Distribution of heights, and illustration of the 20% tallest, for which the average height is 1.764 m.

Notes on points for problem 1.3: 3, 5:
1.3.1: All or "nothing". Of course, right integral but wrong value gives minus 1 point.
1.3.2: All or "nothing". Again, the right integral alone (easy) gives only minus 1 point.
Also, if solved numerically, there should be either a source or an uncertainty!

Problem 2.1

- As the radius r appears squared, while the length L appears linearly, the radius r has to be determined with **twice the relative precision** compared to the length.

$$2 \times \left(\frac{\sigma(r)}{r}\right) = \left(\frac{\sigma(L)}{L}\right) \tag{4}$$

Notes on points for problem 2.1: 5:
2.1: Right use of error propagation formula, but wrong result gives minus 1-2 points.

Problem 2.2

- The mean velocity is $310.4 \pm 28.1\mathrm{m/s}$, but should be given without decimals as 310 ± 28**m/s** or actually $(0.31 \pm 0.03) \times 10^3$**m/s**.
- The kinetic energy of a bullet is then on average: $E_{kin} = 404.7 \pm 24.1(\text{mass}) \pm 73.3\text{velocity} = \mathbf{404.7 \pm 77.2}$
- For the two uncertainties to be equal, the uncertainty on velocity should drop by a factor $73.3/24.1 = 3.04$, requiring a factor $3.04^2 = 9.25$ more experiments, thus a total number of experiments of $10 \times 9.25 = 92.5 \simeq \mathbf{93}$.

Notes on points for problem 2.2: 3, 3, 3:
2.2.1: Missing the sqrt(N) from RMS to error on mean costs 2 points (and all my respect!).
2.2.2: Right formula, wrong result gives minus 1 point.
2.2.3: They should know that the uncertainty goes like 1/sqrt(N)...

Problem 3.1

• We consider the function $f(x) = Cx^{-0.9}$ defined in the range $x \in [0.005; 1.0]$. The normalisation constant C should equal $C = 1/(10 \times (\sqrt[10]{1} - \sqrt[10]{0.005})) = 0.2431$.

• In order to generate random numbers according to this, one can use both the **Transformation method** and the **Accept-Rejection method**. The transformation method is preferable, as the efficiency of the Accept-Rejection method is very low. If the range is enlarged down to zero, then **only the transformation method works**, as the function is then no longer bounded (in y).

[o] **Note for censors**
We have discussed the efficiency of the Accept-Rejection method, but I said that as long as it is still fast, there is no need to be alarmed by a low efficiency. "Fast computers breed lazy programmers".

• Then generated random numbers are shown in Figure 2.

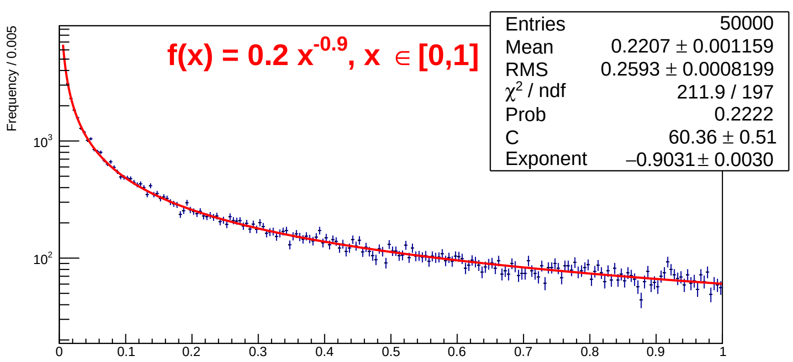

Figure 2: Distribution of random numbers x according to $f(x)$ with an overlaying fit.

• The distribution of t can be seen in Figure 3. It looks rather Gaussian, but that is in fact not really the case (I get $p = 0.99$, but it depends on the random numbers generated!). However, the mean (of course) matches the analytical expectation (which is $\mu(t) = 50 \times C/1.1 \times (1 - 0.005^{1.1}) = 11.01901...$), but the student should consider the error on the mean, and quantify this (I get 11.035 ± 0.058).

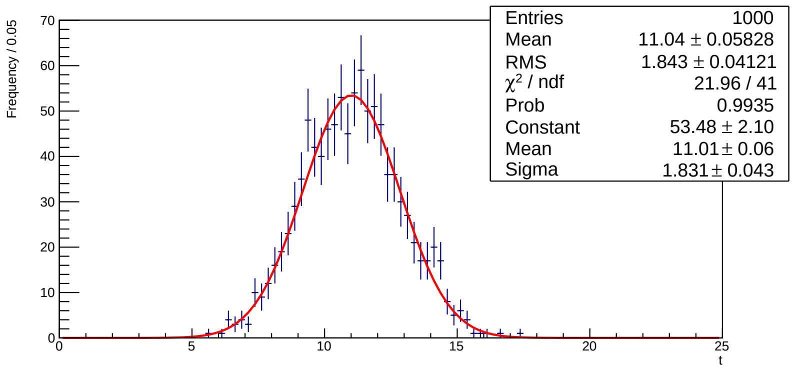

Figure 3: Distribution of random numbers t according to $\sum_{20} f(x)$ with an overlaying Gaussian fit.

Notes on points for problem 1.3: 3, 4, 4, 4:
3.1.1: They typically get C from Matematica...
3.1.2: One method for the first case gives full points. Then 1 extra for saying that both works.
3.1.3: Here we just want to see it work.
3.1.4: There is statistics enough for a Chi2 fit, but commenting on it gives 1 extra point.

Problem 4.1

- The distribution is very consistent with being a Gaussian, despite the one "jumpy" point around 15.5-16.0. The p-value is 0.66.

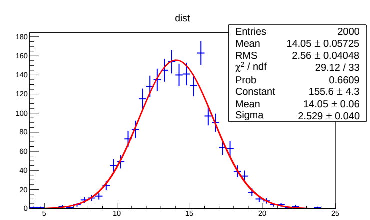

Figure 4: Fit with Gaussian, which shows that it fits well.

- The linear correlation between B and C for ill people is: $\rho_{B,C} = -0.39630$

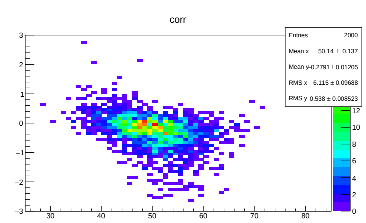

Figure 5: There is correlation, but it is not linear.

- Each of the three variables have some degree of separation, while their combination (fx. through a Fisher Discriminant) is much stronger. Among the three single variables, C is the strongest, and choosing a cut at 0.2, the error rates are:

  Type I error: α = 223 / 3000 = 0.074
  Type II error: β = 263 / 2000 = 0.132

If the variables are combined in a Fisher Discriminant, then the combined covariance matrix becomes:

$$\begin{bmatrix} 15.558 & 31.586 & -0.595 \\ 31.586 & 86.184 & -1.147 \\ -0.595 & -1.147 & 0.362 \end{bmatrix} \tag{5}$$

Inverted combined covariance matrix yields:

$$\begin{bmatrix} 0.2568 & -0.0924 & 0.1291 \\ -0.0924 & 0.0454 & -0.0081 \\ 0.1291 & -0.0081 & 2.9473 \end{bmatrix} \tag{6}$$

In the end, the Fisher coefficients come out to be:

$$\begin{pmatrix} -1.326 \\ 0.627 \\ 2.100 \end{pmatrix} \tag{7}$$

The total separation between the samples using the Fisher is 3.24σ, and in terms of error rates using a selection cut of 18, one gets:
Type I error: α = 29 / 3000 = 0.010
Type II error: β = 30 / 2000 = 0.015

Notes on points for problem 4.1: 4, 4, 9:
4.1.1 Requires Gaussian Chi2 fit (enough stat.). Subtract 1 point for not commenting on p-value.
4.1.2 Simple calculation from data. Give 1 extra point for showing plot and commenting on non-linearity of correlation.
4.1.3 Can be solved very simply. Subtract 3 points for not choosing best variable (C). Subtract 1 point for "only" choosing C without commments. Give 1 extra point for Fisher and 2 extra for other MVA methods. Judge yourself how well the problem has been solved!

Problem 5.1

The first two answers depends on which criteria one rejects hypothesis by, but in this (non-controversial) case, I would say that 5% is reasonable, with 1-2% as alternatives.

- The constant income (well, deficit) can hardly be upheld for the first year, given a **p-value of 0.025**.

- The linear relation for the first 12 months is likely (p-value 0.068), and can be extended to cover 14 months (p-value 0.057) but not 15 months (p-value 0.016).

- From a full fit (see Figure 6) the size of the "jump" is estimated to be $\Delta = 0.708 \pm 0.079$, but could take different values for different fits.

- The full fit is a challenge, and the fits needs to be build up! The initial questions are leading up to this, and the last step is to add some sort of an "onset" function, here a sigmoid, but many other similar functions will also work (atan, Gompertz, piecewise linear, etc.). Finally, a constant offset is introduced at 31.5.

$$\begin{aligned} f(t) &= c_0 + c_1 t + (c_2 - c_0)/(1 + \exp(-(t - c_3)/c_4)) && \text{for } t < 31.5 \\ f(t) &= c_0 + c_1 t + (c_2 - c_0)/(1 + \exp(-(t - c_3)/c_4)) + c_5 && \text{for } t > 31.5 \end{aligned} \tag{8}$$

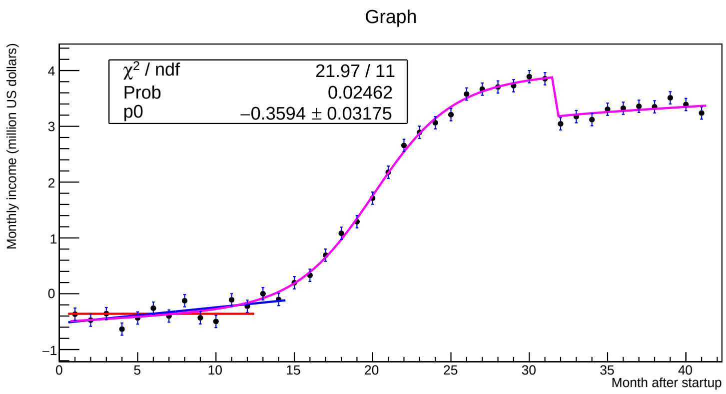

Figure 6: Fit to the income vs. months.

Notes on points for problem 5.1: 3, 4, 4, 4:
5.1.1 It is OK to accept, if criterium (p-critical) is stated.
5.1.2 It is OK to expand further/less, if criterium (p-critical) is stated.
5.1.3 It is perfectly alright to do linear fit (or similar) in a small range around jump to get it. 5.1.4 A high-degree polynomial will **not** do the trick and gives only 1-2 points, depending on discussion! Other bad fits followed by (correct) comments gives only 1 minus point.

Problem 5.2

- Given an RMS of 0.0878 ± 0.0014 sec, this would be considered the typical timing uncertainty. However, the RMS is affected by a few outliers, and a Chi2 fit (which is not too sensitive to outliers) yields $\sigma(t) = 0.066$ sec, which could also be an answer, if described. The mean is 0.0007 ± 0.0021, and thus in perfect agreement with zero, as it should be.

- Using the RMS and data size, one would expect to see about one event with 5% probability at a residual of $t_{cut} = 4\sigma \times 0.0878 = \pm 0.351$ (obtained by solving $1 - (1 - P(|t| > t_{cut}))^{1726} \simeq 0.05$). Using Chauvenet's criterion with $p_{reject} = 0.01$ (i.e. excluding events with a global probability less than this), three data points are excluded:

Measurement	Residual (s)	Nσ	p_{global}	Conclusion
537	0.606	6.90	0.00000000	Rejected
946	0.482	5.56	0.00002266	Rejected
428	0.463	5.38	0.00006271	Rejected
42	0.354	4.15	0.02786367	Accepted

After excluding these three points, the RMS is 0.0852s.

- The single Gaussian does not yield a satisfactory fit. If one performs a Chi2 fit, the p-value is about 3×10^{-31}, and it is very clear from the plot, that it does not fit. If anyone tests it with a (high stat.) Kolmogorov-Smirnov-test, because of the somewhat lower statistics, that would be fantastic (+2 points).

- Clearly, there is a mixture of resolution, and a two Gaussian fit does much better (p-value 0.041, and 0.045 if cleaned). As the mean was consistent with zero, this parameter should be eliminated, yielding a four parameter fit. Also, to minimise correlations, there should be one common normalisation constant and a fraction of each of the two normalised Gaussians.

$$f(t) = N(f_{core} \times G_{core}(0, \sigma_{core}) + (1 - f_{core}) \times G_{tail}(0, \sigma_{tail})) \quad (9)$$

The Voigtian (Lorentz distributed folded with a Gaussian) and the Student's t distribution are even better descriptions than the double Gaussian. All the fits can be seen in Figures 7 and 8. One can venture beyond this to include a e.g. third Gaussian, but that is probably speculation, though we will of course award the courages students with extra points, if they dare tread this path.

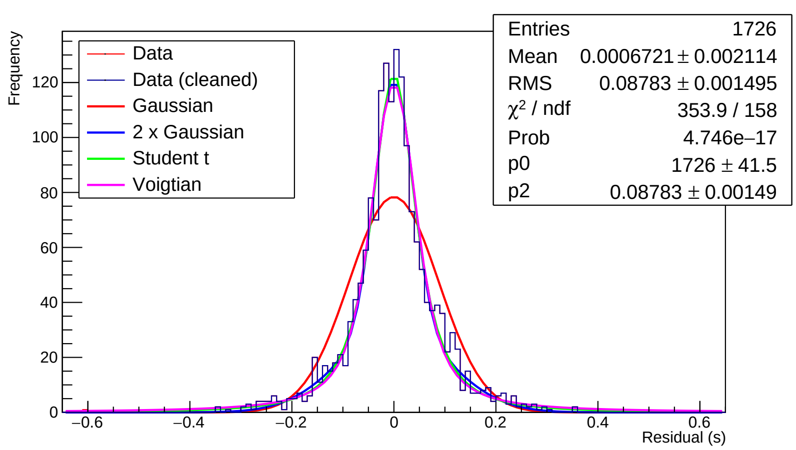

Figure 7: Fit to the time residuals. The single Gaussian is clearly not satisfactory, while the double Gaussian, Student's t, and Voigtian is.

Distribution	Npar	Prob(χ^2)	Comment
Gaussian ($\mu = 0$)	2	4.7×10^{-17}	Very poor model
2 x Gaussian ($\mu = 0$)	4	0.045	Reasonable and interpretable model
Student's t ($\mu = 0$)	3	0.305	Best model, also matching outliers
Voigtian ($\mu = 0$)	3	0.104	Good model, though large tails

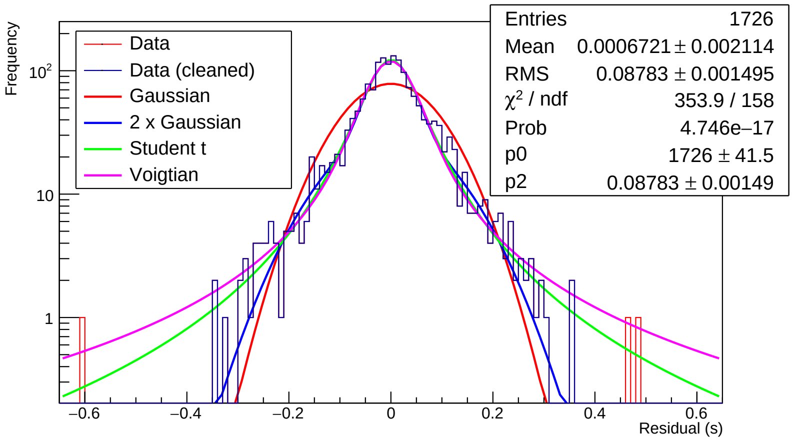

Figure 8: Log version of the above plot.

Notes on points for problem 5.2: 4, 4, 3, 4:
5.2.1 This should be standard and done on the data itself or possibly the histogram. 5.2.2 1 extra point for applying Chauvenet's criterion or mentioning it. Rough exclusion gives near full points (as this is hard!), however the cut should not be lower than 3σ.
5.2.3 1 extra point for fitting both with Chi2 and likelihood (or mentioning it). 5.2.4 Fit with any double Gaussian gives full points, but 1 extra point for fixing mean to zero, and 1 extra point for using optimal parametrisation with fractions and normalisation in place. Check of this: Norm is the total number of entries (requires binwidth to also be included).