# *Solution for Applied Statistics take-home exam 2018/19*

## *Problem 1.1*

**Problem 1.1.1:**

- The resulting distribution is **binomial**. The chance of winning exactly 25 times **0.110243**.

The chance of winning 26 or more times is **0.369458**, obtained by adding the individual binomial contributions.

**Problem 1.1.2:**

- In order to have 0.95 probability of winning at least 20 times, little Peter have to play $N$ times:

$$p_{20+}(N) = \sum_{i=20}^{N} C_N^i \; p^i (1-p)^{N-i} = 0.95 \qquad (1)$$

This gives **N = 53** games ($p_{20+}(52) = 0.947$ and $p_{20+}(53) = 0.959$).

**Notes on points for problem 1.1: (3 + 3 points)**
1.1.1: One should mention the Binomial distribution (though not strictly required). Using the Poisson gives at least -2 points, as this is a poor approximation here. There is -1 point for making wrong interpretation of the problem (i.e. not including 26 in the second problem) or getting the result wrong after arguing correctly.
1.1.2: The last problem is one of limit setting, and there is -1 point for rounding down to 52.

## *Problem 1.2*

- This is a non-analytical integral (but solvable with error function), which yields **p(1.2-2.5 $\sigma$) = 0.21772**.

It can be illustrated as follows:

**Notes on points for problem 1.2: (4 points)**
Basic understanding of the problem is a point in itself, and drawing it also gives a point. It is OK to do an estimate (typically based on a few iterations with evaluations of the Gaussian integral or numerically).

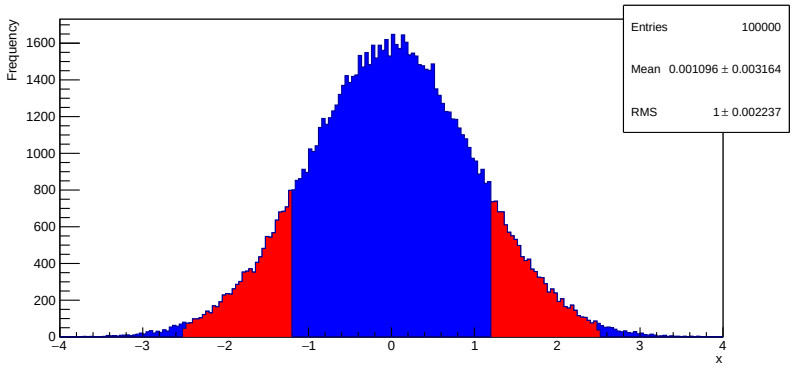


Figure 1: The chance to be between 1.2 and 2.5 $\sigma$ away from the mean is about 21.8% as illustrated in the figure

## *Problem 1.3*

- As there are many operations daily in Danish hospitals, and as the chance of a serious mistake is low, the underlying PDF is the **Poisson distribution**. Even if the chance of a serious mistake is not constant, it is still Poisson, as these are closed under addition.
- The chance of a critical day (CD) is $p_{CD} = 22/365 = 0.0603$. Thus one should find the Poisson distribution with an average number of daily mistakes $\lambda_{DM}$, for which the cumulative distribution from 8 and up is $p_{CD}$:

$$\sum_{8}^{\infty} Pois(N, \lambda_{DM}) = 1 - \sum_{0}^{7} Pois(N, \lambda_{DM}) = p_{CD} \qquad (2)$$

This can in principle be solved analytically using the upper incomplete gamma function or the regularized gamma function, though I doubt that (m)any will do this. The numerical approach is clearly the way forward, and yields the result $\lambda_{DM}$ **= 4.87560**.

**Notes on points for problem 1.3: (3 + 3 points)**
1.3.1: There is two points for identifying that this is a Poisson distribution, and two for getting the answer right. An extra point can maybe be gained by commenting (well) on the assumption, that the occurance of mistakes should be completely independent from another occurance elsewhere, and/or simulating the problem!
1.3.2: This problem is hard to solve analytically, so a numerical (or search like) solution is most likely. Missing if 8 mistakes was included or not costs 1 point. However, as it is possible to roughly estimate the answer, getting it far off (without commenting) costs at least two points, even if method is correct. Stating the solution outline correctly gives 3 out of 4 points, I think.

[o] **Note for censors**
We have worked with data, that should be combined with or without uncertainties, but in separate problems. The rejection of data with uncertainties should be clear.

## ***Problem 2.1***

NOTE: Final values have to be updated for the weighted mean, as I changed the first value!!!

The weighted mean of the data with uncertainties gives:

$$\begin{aligned} \mu &= 2.429 \pm 0.058 \\ \text{chi2} &= 11.5 \\ \text{P}(\text{chi2}, N_{\text{dof}} = 3) &= 0.009 \end{aligned} \tag{3}$$

The unweighted mean of the data without uncertainties gives:

$$\mu = 2.618 \pm 0.036 \tag{4}$$

The chi2 and the probability of the weighted mean of the data with uncertainties suggest that the data are not consistent with each other. This may be due to the outlier ($2.09 \pm 0.12$). The unweighted mean of the data without uncertainties is, by construction, consistent with the data. In addition, the uncertainty on the mean, i.e. the positioning, is lower than those of the weighted mean.

Without considering the outlier in the weighted mean of the data with uncertainties, the weighted mean gives:

$$\begin{aligned} \mu &= 2.532 \pm 0.066 \\ \text{chi2} &= 1.10 \\ \text{P}(\text{chi2}, N_{\text{dof}} = 2) &= 0.58 \end{aligned} \tag{5}$$

with slightly higher uncertainty on the mean, but a much better agreement in the chi2 and the probability.

We can also exclude the lowest and highest data of the unweighted for the data without uncertainties, which gives a more accurate positioning:

$$\mu = 2.629 \pm 0.026 \tag{6}$$

However, we might exclude important data point without knowing the uncertainties behind the measurements. The best solution is still to get those dame missing uncertainties.

NOTE: The combination is missing!

**Notes on points for problem 2.1: (5, 4, 4 points)**
2.1.1: The first problem is the longest part, as one needs to evaluate both sets of data.
-1 point for not doing weighted mean.
-1 point for not doing Chi2.
-1 point for not rejecting the first point.
-1 point for not getting weighted mean uncertainty right.
+1 point for discussing Chauvenet's Criterion regarding largest outlier (2.34).
2.1.2: This forces students to provide mean and uncertainty for each of the two classes of measurements. Miscalculations may lead to wrong conclusion, but any earlier mistake(s) should not "ruin" later correct reasoning. Note: The working should have been "precise" instead of "accurate".
-1 point for each uncertainty calculated wrongly.
2.1.3: Unless they conclude, that the measurements without uncertainty far dominates, they should combine the two measurements with a weighted mean, and CHECK with a Chi2, that they agree.
-1 for not doing Chi2. -1 for not commenting on Chi2 (i.e. that they can be combined).

## *Problem 2.2*

This is a classic error propagation problem! The uncertainty of the radiance according to the Planck's law is given by the error propagation:

$$\begin{aligned} \sigma_{\mathrm{B}}^2 &= (\frac{\partial \mathrm{B}}{\partial \nu}\sigma_\nu)^2 + (\frac{\partial \mathrm{B}}{\partial \mathrm{T}}\sigma_{\mathrm{T}})^2 \\ \frac{\partial \mathrm{B}}{\partial \mathrm{T}} &= \frac{2\mathrm{h}\nu^3}{\mathrm{c}^2}\frac{\frac{\mathrm{h}\nu}{\mathrm{k}T^2}\exp(\mathrm{h}\nu/\mathrm{k}T)}{(\exp(\mathrm{h}\nu/\mathrm{k}T)-1)^2} \\ \frac{\partial \mathrm{B}}{\partial \nu} &= \frac{6\mathrm{h}\nu^2}{\mathrm{c}^2}\frac{1}{\exp(\mathrm{h}\nu/\mathrm{k}T)-1} - \frac{2\mathrm{h}\nu^3}{\mathrm{c}^2}\frac{\frac{\mathrm{h}}{\mathrm{k}T}\exp(\mathrm{h}\nu/\mathrm{k}T)}{(\exp(\mathrm{h}\nu/\mathrm{k}T)-1)^2} \end{aligned} \tag{7}$$

- Considering $\nu = (0.566 \pm 0.025) \times 10^{15}$ Hz and T = $(5.50 \pm 0.29) \times 10^3$ K, the analytical solution gives: $\mathbf{B(\nu, T) = (1.93 \pm 0.53) \times 10^{-8}\ Jm^{-2}s^{-1}Hz^{-1}}$.

If the frequency $\nu$ and the brightness temperature $T$ are partially correlated, with a correlation factor $\rho = 0.87$ the uncertainty depends on the non diagonal element of the correlation matrix. Therefore, we have to use the general formulation of the error propagation. This leads to:

$$\sigma_{\mathrm{B}}^2 = \mathbf{J}^{\mathrm{T}}\mathbf{V}\mathbf{J}, \text{ with}$$
$$\mathbf{J} = \begin{pmatrix} \partial \mathrm{B}/\partial \nu \\ \partial \mathrm{B}/\partial \mathrm{T} \end{pmatrix} \quad \text{and} \quad \mathbf{V} = \begin{pmatrix} \sigma_\nu^2 & \rho\sigma_\nu\sigma_{\mathrm{T}} \\ \rho\sigma_\nu\sigma_{\mathrm{T}} & \sigma_{\mathrm{T}}^2 \end{pmatrix} \tag{8}$$

The partially correlated uncertainty found (using simulation) is $\sigma_{\mathbf{B}}$ $= \mathbf{0.37 \times 10^{-8}\ Jm^{-2}s^{-1}Hz^{-1}}$.

**Notes on points for problem 2.2: (4 + 5 points)**
One should do this with the error propagation formula, though a simulation can do as well. It is a bit of a hard problem, as both the function and the numbers are complicated.
$-1$ point for showing the error propagation equation but getting wrong result.
$-1$ for getting the error propagation equation wrong.
$-1$ for getting the correlated value wrong.
$+1$ points for doing both analytical and simulation solution. $+1$ points for noting (through simulation) that the distribution is not entirely Gaussian.

## *Problem 3.1*

- The mean = 1.17 and the RMS = 0.51. $C = 1/(2 + (\exp(-2a) - 1)/a) = 0.6626$.
- The function $f(x) = C(1 - \exp(-ax))$ defined in the range $x \in [0, 2]$, where $a = 2$ can be integrated, but not inverted analytically (at least not easily - only through special functions defined seemingly only in Wolfram-Alpha!), thus the **Transformation Method is NOT possible**. Since the range is limited and there are no asymptotes, **the Accept-Reject method can be used**.
- •
-

**Notes on points for problem 3.1: (3, 3, 3, 3, 3 points)**
3.1.1: -1 point for each value missing or wrong (well, give some points if method is there).
3.1.2: -1 point for thinking that the transformation method will work (unless really well argued).

: -1 point for not arguing why the Accept-Reject-method works (bounded in $x$ and $y$).
3.1.3: Almost any plot is enough, as long as the number of events and range looks OK.
3.1.4: -1 point, if fit done with a $\chi^2$ without comments on the low(ish) statistics.

+1 point, if both $\chi^2$ and LLH fits are done and commented on.
3.1.5: -1 point, if there is no quantification ($\chi^2$ and/or p-value) of likeness to Gaussian.

+1 point, if doing anything beyond commented $\chi^2$ fit with Gaussian.

## *Problem 4.1*

- The **mean is 471.3s**, while the **median is 180s** (3 minutes!).

There is no requirement of commenting on the large difference. The plot is tricky, because some observations extend very far (e.g. 18 hours!), which is hard to get in a linear plot. If using a linear scale, one should possibly provide two plots with different scales. Also, the data is "very discontinuous", as specific lengths (full minutes) are much more abundant that values just around.

[o] **A word on the data**
This strange/fun data set was used by several projects in the past years, though in other contexts.

- The likeness between East and West coast should be done with a Kolmogorov-Smirnov (K-S) test, though a $\chi^2$ is also acceptable. But it is not enough to just comment, that they look alike in a plot. It needs some form of quantification. **The result is that they are compatible**.
- The **linear correlation coefficient is 0.024**, which is actually slightly significant (the error is roughly $\pm 0.004$, though this is not required to calculate), but very low. However, if one plots the two against each other, it become visibly clear, that there are more patterns in the relation, that this reveals. Though there are many effects playing in (and one needs to divide the data for example at noon and midnight), the correlation should roughly follow a sinus function of a yearly period with the maximum in summer and minimum in winter.
- This should be tested with a simple $\chi^2$, as the uncertainties on the counts/frequencies are from high statistics (i.e. highly Gaussian). A K-S test can be used, though it is not easy to get right, nor

reason why the order of the days play a significant role (except for the weekend, which clearly stands out).

**Notes on points for problem 4.1: (3, 4, 4, 4 points)**
4.1.1 $-1$ point for getting mean and median wrong.

$-1$ point for plots that exclude much of the data (without comments) or are impossible to see.

$+1$ point for commenting on the difference between mean and median (due to the skewness).
4.1.2 $+1$ point for noting that the K-S test requires continuous variables, which this is a borderline case of.
4.1.3 $-1$ point for just calculating the linear correlation coefficient. A plot of some form (or more elaborate measures) are needed for full points.

$+1$ point for commenting on the pattern seen, and how especially dusk, but also dawn seems to be the most populated areas, and that they vary with the year, as height of sun and hour of sunset/sunrise changes.
4.1.4 $-1$ A simple $\chi^2$ test (applied twice) gives full points. Using a K-S test is hard to get to work, and since the distribution is not continuous, this gives -1 point (-2 if the solution is also far off). A plot with (correct) errors on should give 2 points, maybe even 3, if commented correctly on (i.e. showing understanding but lacking quantification).

[o] **A word on the data**
The data is from Raphael Weldon, who famously did this experiment, and wrote about it in a letter to Francis Galton (which is why it is known).

## *Problem 4.2*

- The distribution should clearly follow a **binomial distribution**, as there is a fixed number of trials (12) and a fixed probability of succes (around $1/3$). The Poisson is not a good (enough) approximation, as $N$ is not large nor is $p$ small.
- The "naive" $p = 1/3$ gives a $\chi^2 = 35.7$, and with 13 degrees of freedom (there are no fit parameters), this yields a p-value of 0.000656. Thus, given this much data, the "naive" null hypothesis is rejected, almost no matter what significance level is chosen.
- A fit with a binomial gives a $\chi^2 = 9.8$, and with 12 degrees of freedom (there are now one fit parameter), this yields a p-value of 0.636. Thus, this is a good fit, and the binomial hypothesis is accepted with $p_{56} = 0.33764 \pm 0.00084$. If one (also) performs a (binned – unbinned doesn't make sense here) likelihood fit, then one obtains the result $p_{56} = 0.33770 \pm 0.00084$.

**Notes on points for problem 4.2: (4, 4, 5 points)**
4.2.1 $-2$ points for suggesting the Poisson. All other things would be strange and max. 1-2 points.
4.2.2 $-1$ point for getting the number of degrees of freedom wrong.
4.2.3 $+1$ point for also doing a likelihood fit. +1 additional, if followed up by (good) comments about the slight change in value, while the uncertainty is also smaller, but only on the third decimal, because the Gaussian assumption is ensured in most bins, given the high statistics.

## *Problem 5.1*

- The three peaks are identifiable by their low energy, ratio in distance and ratio in height They can be fitted with a $\chi^2$ fit using a Gaussian for the signal, and a linear background (constant over a short range or high polynomials can also be used). The fits yield:

Pb1: $\mu = 113.50 \pm 0.19$ $\sigma = 1.33 \pm 0.22$ Prob($\chi^2$) = 0.7366
Pb2: $\mu = 137.72 \pm 0.07$ $\sigma = 1.20 \pm 0.06$ Prob($\chi^2$) = 0.4809
Pb3: $\mu = 163.86 \pm 0.03$ $\sigma = 1.05 \pm 0.03$ Prob($\chi^2$) = 0.5026

- The ratio of the distances should be the same between channel number (from problem above) and energy (tabular values). This is thus a problem of error propagation and single sample test. **The ratio comes out to be** $r = 1.079 \pm 0.010$ **compared to table value** $r = 1.075$**, thus less than one sigma away, and in good agreement**.
- The Bismuth peaks comes out to be:

Bi1: $\mu = 281.37 \pm 0.05$ $\sigma = 0.99 \pm 0.05$ Prob($\chi^2$) = 0.3474
Bi2: $\mu = 514.71 \pm 0.10$ $\sigma = 0.94 \pm 0.11$ Prob($\chi^2$) = 0.8922

Using these five peak values, and plotting their known energies on one ($x$) axis (since they don't have errors), and the fitted channel number on the other ($y$) axis gives five points on a line, which can be fitted. The result is: $CN = (2.997 \pm 0.061) + E \times (0.45698 \pm 0.00012)$, and a $\chi^2 = 4.6$ with Ndof = 5-2 = 3, yield Prob=0.197. This can be inverted to yield the conversion of channel number to energy. Note that "linear" allows for a constant offset (as there is), and that the problem would have stated "proportional" otherwise. But since the distinction is small, several may fit without the offset, and get low p-values (about 0.0001 or so).

- Plotting the width of the peaks (again as many as you like!), one can fit these with a constant. This gives $\chi^2 = 10.2$ with Ndof = 5-1 = 4, yielding Prob=0.0373. That is a bit low, but not necessarily enough to reject the hypothesis. The student should decide, but argue.
- There is clearly a bump around 340 in channel number, and fitting it gives:

$N = 75.7 \pm 13.1$, $\mu = 342.92 \pm 0.16$, $\sigma = 0.75 \pm 0.14$, and Prob = 0.486. Given that the normalisation is more than 5 sigma away from zero, this is very significant. It can be solved even more elegantly using the likelihood ratio, but that is not required. Since one is searching for a peak in a 100 keV range, and the width is typically around 3.3 keV, there is a trial factor of about 30, but this is again not required at all. Should be converted to energy!!!

- In addition to the step at 840 keV, there is a structure around 920 keV (422 in channel number), which is a double peak. Fitting it with two hypothesis yields:

Single Gaussian: $N = 152.6 \pm 29.0$, $\mu = 423.30 \pm 0.63$, $\sigma = 2.74 \pm 0.38$, and Prob = 0.0102, thus not a very good fit.

Double Gaussian: $N1 = 74.4 \pm 15.3$, $\mu 1 = 421.15 \pm 0.17$, $\sigma 1 = 0.72 \pm 0.15$, $N2 = 109.6 \pm 16.8$, $\mu 2 = 425.42 \pm 0.13$, $\sigma 2 = 0.76 \pm 0.12$, and Prob = 0.7612, thus giving a very good fit. The latter hypothesis is accepted, and clearly prefered over the single Gaussian.

**Notes on points for problem 5.1: (3, 3, 3, 3, 3, 4 points)**
5.1.1 -1 point for fitting the peaks with no background or with a flat for a larger range. Comments can save points!
5.1.2 -1 point for getting some other ratio (i.e. misunderstanding the problem).

-1 point for not propagating the error correctly.
5.1.3 -1 point for not getting the linearity but a proportionality.
5.1.4 -1 point for not getting the errors right on peak widths (should be taken from fits).
5.1.5 -1 point for not converting to energy.

+1 point for including uncertainty in linearity determination.
5.1.6 This is a really hard problem, and getting either the double peak right or the step fitted should give full points or close to. Getting both features and doing it correct should give at least +2 points.