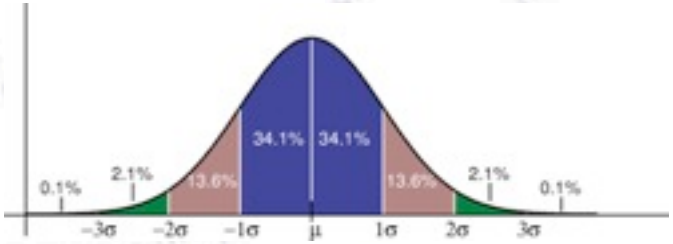
Applied Statistics

Problem Set - Solution & Discussion






Troels C. Petersen (NBI)



"Statistics is merely a quantization of common sense"

Problem 1.1

$$\int_0^C x^2\,\mathrm{d}x = 1 \Leftrightarrow \frac{1}{3}C^3 = 1 \Leftrightarrow C = \sqrt[3]{3} \approx \underline{\underline{1.44}} \tag{1}$$

The mean and the variance of the distribution is given by (Barlow, p. 35):

$$\mu = \int_{-\infty}^{+\infty} x f(x)\,\mathrm{d}x \tag{2}$$

$$\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\,\mathrm{d}x \tag{3}$$

Now we evaluate the PDF in the region from 0 to C, inserting the value of C obtained previously, and we get the mean of the distribution:

$$\mu = \int_0^C x^3\,\mathrm{d}x = \frac{1}{4}C^4 = \underline{\underline{1.082}} \tag{4}$$

We use eq. 3 to calculate variance:

$$\sigma^2 = \int_0^C (x-\mu)^2 f(x)\,\mathrm{d}x = \int_0^C (x^2+\mu^2-2x\mu) f(x)\,\mathrm{d}x = \frac{1}{5}C^5 + \mu^2\frac{1}{3}C^3 - 2\mu\frac{1}{4}C^4 = 0.078. \tag{5}$$

The width of x is just the square root of the variance: $\sigma = \sqrt{0.078} = \underline{\underline{0.279}}$.

Problem 1.2

This can be calculated but has a lot of terms. One way to simpilfy is by using that the chance of winning either 0, 1, 2 . . . or 100 times is 100%, and we can write:

$$\sum_{k=0}^{100} P(k) = \sum_{k=0}^{3} P(k) + \sum_{k=4}^{100} P(k) = 1 \tag{1}$$

$$\sum_{k=4}^{100} P(k) = 1 - \sum_{k=0}^{3} P(k) \tag{2}$$

This gives that the probability of Peter winning more than 3 times out of a 100 is $\sim 28.6\%$.

Since p is fairly small and n fairly large, this could also be represented by a Poisson distribution with mean expected number of successes $\lambda = pn = 100/36$. From Barlow (3.11) the chance of getting more than 3 successes would then be

$$P(r > 3) = \sum_{r=4}^{n} \frac{e^{-\lambda}\lambda^r}{r!} \approx 0.2865 = 28.65\%$$

Problem 1.3

The mean of the PDF $f(x) = 1/2\sin(x)$, $x \in [0, \pi]$ is

$$\mu = \langle x \rangle = \int_0^\pi x f(x) dx = \frac{1}{2}\int_0^\pi x \sin(x) dx = \frac{1}{2}\left[\sin(x) - x\cos(x)\right]_0^\pi = \frac{1}{2}\pi \approx 1.57 \quad (8)$$

while the width is given by

$$\langle x^2 \rangle = \int_0^\pi x^2 f(x) dx = \frac{1}{2}\int_0^\pi x^2 \sin(x) dx = \frac{1}{2}\left[2x\sin(x) + (2 - x^2)\cos(x)\right]_0^\pi = \frac{1}{2}\pi^2 - 2 \quad (9)$$

$$\sigma = \sqrt{\langle x^2 \rangle - \langle x \rangle^2} = \sqrt{\frac{1}{2}\pi^2 - 2 - \left(\frac{1}{2}\pi\right)^2} = \sqrt{\frac{1}{4}\pi - 2} \approx 0.68 \quad (10)$$

For the PDF $f(x) = \ln(x)$, $x \in [1, e]$ the mean is

$$\mu = \langle x \rangle = \int_1^e x f(x) dx = \int_1^e x \ln(x) dx = \left[\frac{x^2}{2}\left(\ln(x) - \frac{1}{2}\right)\right]_1^e = \frac{e^2 + 1}{4} \approx 2.10 \quad (11)$$

and the width is determined as

$$\begin{aligned} \langle x^2 \rangle &= \int_1^e x^2 f(x) dx = \int_1^e x^2 \ln(x) dx = \left[\frac{x^3}{3}\left(\ln(x) - \frac{1}{3}\right)\right]_1^e \\ &= \frac{2e^3 + 1}{9} \end{aligned} \quad (12)$$

$$\sigma = \sqrt{\langle x^2 \rangle - \langle x \rangle^2} = \sqrt{\frac{2e^3 + 1}{9} - \left(\frac{e^2 + 1}{4}\right)^2} \approx 0.42 \quad (13)$$

Problem 2.1

For $\theta = 1.58 \pm 0.02$ the situation is the same for $\cos(\theta)$:

$$\sigma_{\cos} = \sqrt{(\sin(\theta)\sigma_\theta)^2} = \sqrt{(1.000 \cdot 0.02)^2} = 0.02 \tag{21}$$

For $\sin(\theta)$ the partial derivative is actually smaller than the error and the first order approximation used to derive the propagation formula is not a good one. But one could fx. go to second order in the taylor expansion. But a much easier way is of course to calculate the difference between $\sin(\theta)$ and $\sin(\theta \pm \sigma_\theta)$ (and it gives the same answer).

$$\sin(1.58) - \sin(1.56) = 0.999957646 - 0.99994172 = 1.59262688 \cdot 10^{-5} \tag{22}$$

$$\sin(1.58) - \sin(1.60) = 0.999957646 - 0.999573603 = 0.000384043457 \tag{23}$$

So the uncertainty on $\sin(\theta)$ would be -0.0004, which is only in one direction. No surprise since the value for $\sin(\theta) = 1.0000$. Note that this uncertainty is not the standard deviation and $\sin(\theta) = 1.0000$ is not the mean but the mode of the distribution. The uncertainty is actually overestimated since more than 68 % would be inside the interval. The 34% in the interval where $[\theta, \theta + \sigma_\theta]$ $\sin(\theta)$ will be in the interval $[1 - 0.0004, 1]$, but the 34 % just below the mean the $\sin(\theta)$ will be in $[1 - 1.59262688 \cdot 10^{-5}, 1]$ so measurements further out than one standard deviation in the distribution for θ will still be in the $[1 - 0.0004, 1]$ interval.

Problem 2.1

There is more than one person, who "discovered" this subtlety!

Here is a very nice reporting of the result:

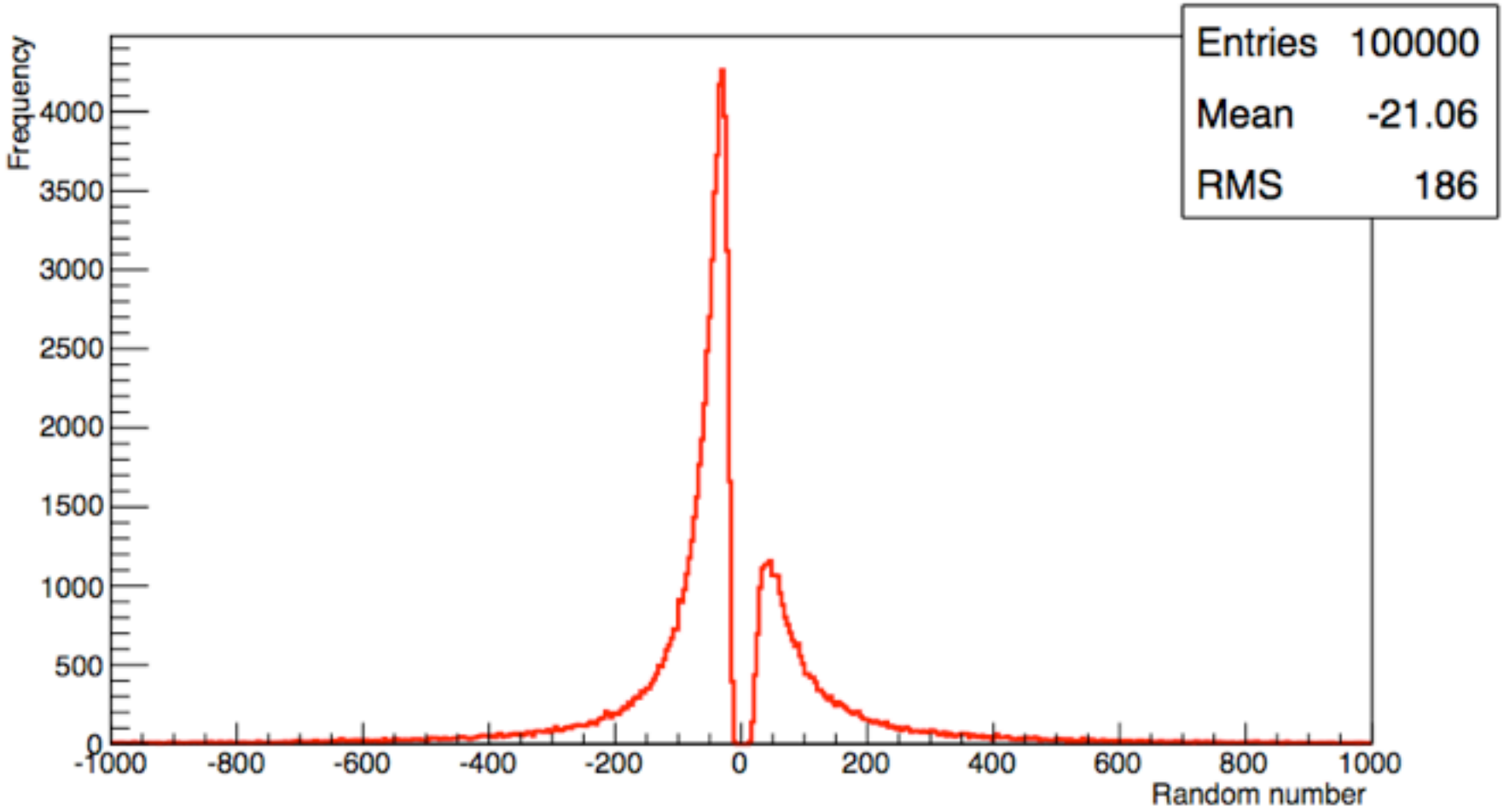
distribution. If we have to state an uncertainty, even though we now realize that this cannot be Gaussian, we can do a variation of error propagation without Taylor expanding. We have

$$\sin(1.58) - \sin(1.58 - 0.02) = 0.000002 \qquad \sin(1.58) - \sin(1.58 + 0.02) = 0.0004 \tag{2.9}$$

Going in the negative direction (-0.02) crosses $\pi/2$. The first of these are negligible compared to the other, so (if really we are forced to point estimation) we end up stating:

$$y = 1.0000^{+0.0000}_{-0.0004} \tag{2.10}$$

Problem 2.1



We draw N Gaussian random numbers from $N(1.58, 0.02^2)$ and simply evaluate tan on them, resulting in a histogram of $\tan(\theta)$. This is shown in Fig. 2.2. We see that is indeed not sensible to simply give a mean and a width. The full distribution is needed (or other measures), because of the bimodality. So we'll stop here.

Problem 2.2

Using the numbers provided in the assignment, we get $n_2 = 1.5$. The uncertainty can be calculated using propagation of errors as before

$$\sigma_{n_2} = \sqrt{\left(\frac{\partial n_2}{\partial \theta_1}\sigma_{\theta_1}\right)^2 + \left(\frac{\partial n_2}{\partial \theta_2}\sigma_{\theta_2}\right)^2} \tag{50}$$

$$= \sqrt{\left(\frac{n_1 \cdot \cos(\theta_1)}{\sin(\theta_2)}\sigma_{\theta_1}\right)^2 + \left(\frac{n_1 \cdot (-\sin(\theta_1)\cos(\theta_2))}{\sin(\theta_2)^2}\sigma_{\theta_2}\right)^2} \tag{51}$$

Using the numbers given in the assignment (it is especially important to remember that $\sigma_{\theta_1} = \sigma_{\theta_2} = \sin(0.2°)$), we get $\sigma_{n_2} = 0.02$. Since n_1 is given without uncertainties, we have omitted it from the propagation of errors formula. It is also given with so many digits that any error it might have is orders of magnitude lower than the ones for the angles.

Problem 2.3

Fejlpropageringsformlen for en funktion $x(y,z)$ af to variable *med korrelation* er

$$\sigma_x = \sqrt{\left(\frac{dx}{dy}\right)^2 \sigma_y^2 + \left(\frac{dx}{dz}\right)^2 \sigma_z^2 + 2\frac{dx}{dy}\frac{dx}{dz}\rho_{yz}\sigma_y\sigma_z},$$

som med funtionen

$$x = 2yz + z^2,$$

som har differentialkvotienterne

$$\frac{dx}{dy} = 2z \quad \text{og} \quad \frac{dx}{dz} = 2y + 2z,$$

giver

$$\sigma_x = 3,77.$$

Funktionsværdien er

$$x = 19,95,$$

så fejlen på x ligger procentvis i den samme størrelsesorden som fejlene på y og z. Det betyder ikke noget i sig selv, fejlen kunne i princippet sagtens have været noget andet, men det er beroligende.

Problem 2.4

If there are no correlation between N and τ, then we use Barlow (4.14) which gives

$$\sigma_N^2 = \left(\frac{dN}{dN_0}\right)^2 \sigma_{N_0}^2 + \left(\frac{dN}{d\tau}\right)^2 \sigma_\tau^2 = e^{-\frac{2t}{\tau}} \sigma_{N_0}^2 + N_0^2 e^{-\frac{2t}{\tau}} \frac{t^2}{\tau^4} \sigma_\tau^2$$

Since N_0 and τ are known with a relative uncertainty of $r = 1\%$, their uncertainties must be given by $\sigma_{N_0} = N_0 r$ and $\sigma_\tau = \tau r$, and hence

$$\sigma_N^2 = e^{-\frac{2t}{\tau}} {N_0}^2 r^2 + N_0^2 e^{-\frac{2t}{\tau}} \frac{t^2}{\tau^4} \tau^2 r^2 = {N_0}^2 r^2 e^{-\frac{2t}{\tau}} \left(1 + \frac{t^2}{\tau^2}\right)$$

So the uncertainty of N_0 contribute with a factor of 1, while the uncertainty of τ contribute with a factor of $\frac{t^2}{\tau^2}$ in the uncertainty of N. Hence, they contribute equally when $\frac{t}{\tau} = 1$

Problem 3.1

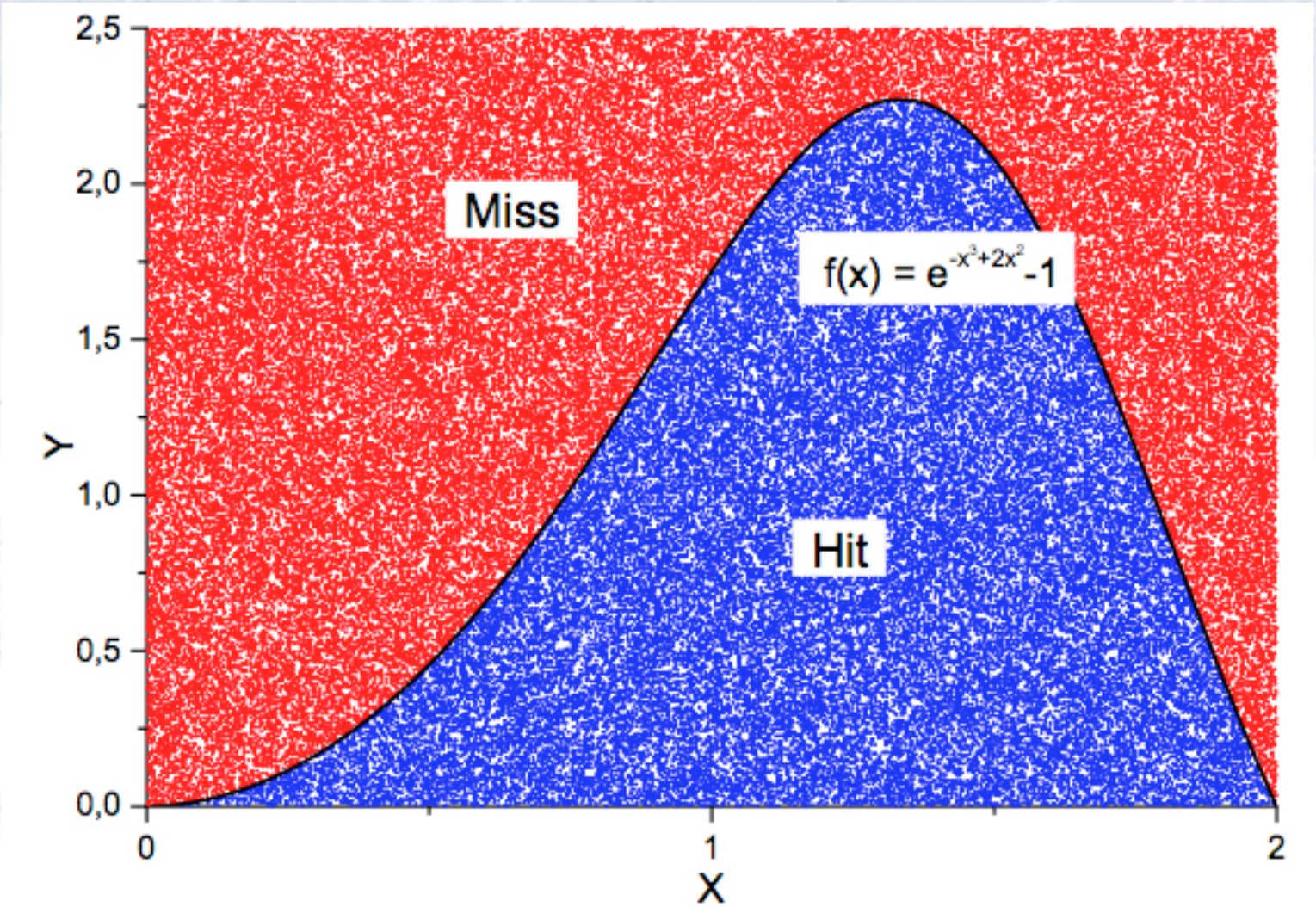


Figure 3.1: "Hit and miss" run, showing the good (blue) and the bad (red) trials.

Problem 3.1

In total 10^7 points were distributed in a box as shown in figure 1. 4422469 points fell in the area included by f(x) from this the included area is calculated

$$\int_0^2 f(x) \approx \frac{N_{hit}}{N_{total}} \cdot A_{box} = \frac{4422469}{10^7} \cdot 2.5 \cdot 2.0 \Longrightarrow \underline{\underline{\mathbf{A = 2.2112}}} \, . \tag{11}$$

The error on the enclosed area is given by the fractional error.

$$\sigma_A = A * \sqrt{\frac{f(1-f)}{N_{total}}} \quad \text{where} \quad f = \frac{N_{hit}}{N_{total}} \tag{12}$$

$$\Longrightarrow \underline{\underline{\boldsymbol{\sigma_A = 0.0008}}} \, . \tag{13}$$

$f(x)$ is normalized by dividing by the enclosed area A:

$$f(x)_{norm} = \frac{1}{2.2112} \cdot \left(e^{-x^3+2x^2} - 1 \right) \, . \tag{14}$$

Problem 4.1

a) To calculate the mean and spread using Barlow (2.1) and (2.8d) which gives

$$\bar{x}_n = \frac{1}{10}\sum_{i=1}^{10} x_i = 1{,}94\ \mu s \quad \sigma_n = \frac{1}{\sqrt{10}}\sqrt{\frac{1}{9}\sum_{i=1}^{10}(x_i - \bar{x}_n)^2} = 0{,}08\ \mu s$$

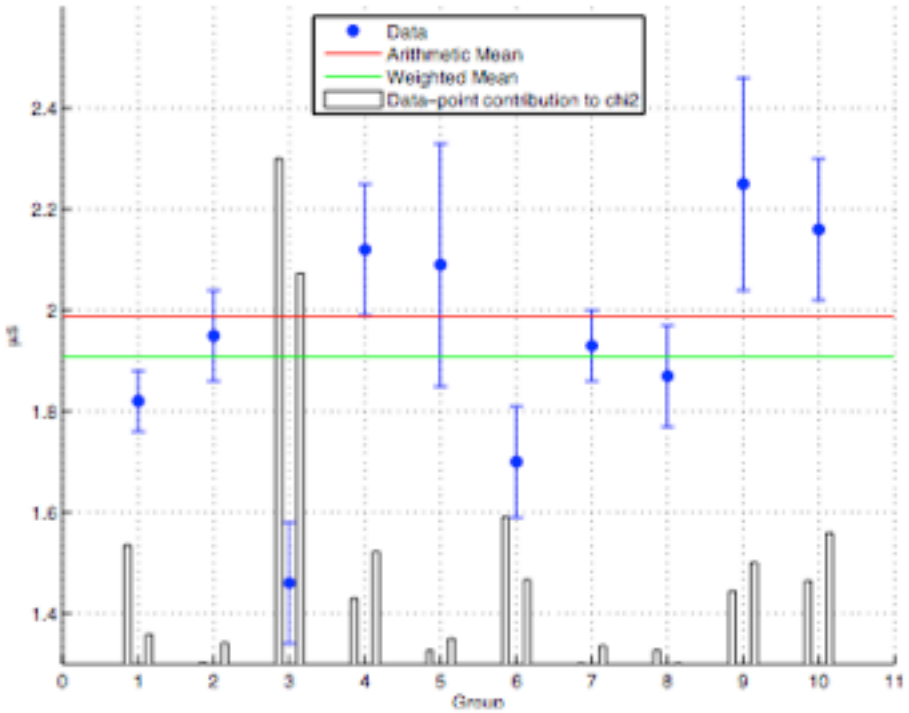
Instead of doing this, can we calculate a weighted mean and spread using Barlow (4.6) and (4.7)

$$\bar{x}_w = \frac{\sum_{i=1}^{10}\frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{10}\frac{1}{\sigma_i^2}} = 1{,}88\ \mu s \quad \sigma_w = \sqrt{\frac{1}{\sum_{i=1}^{10}\frac{1}{\sigma_i^2}}} = 0{,}03 \mu s$$

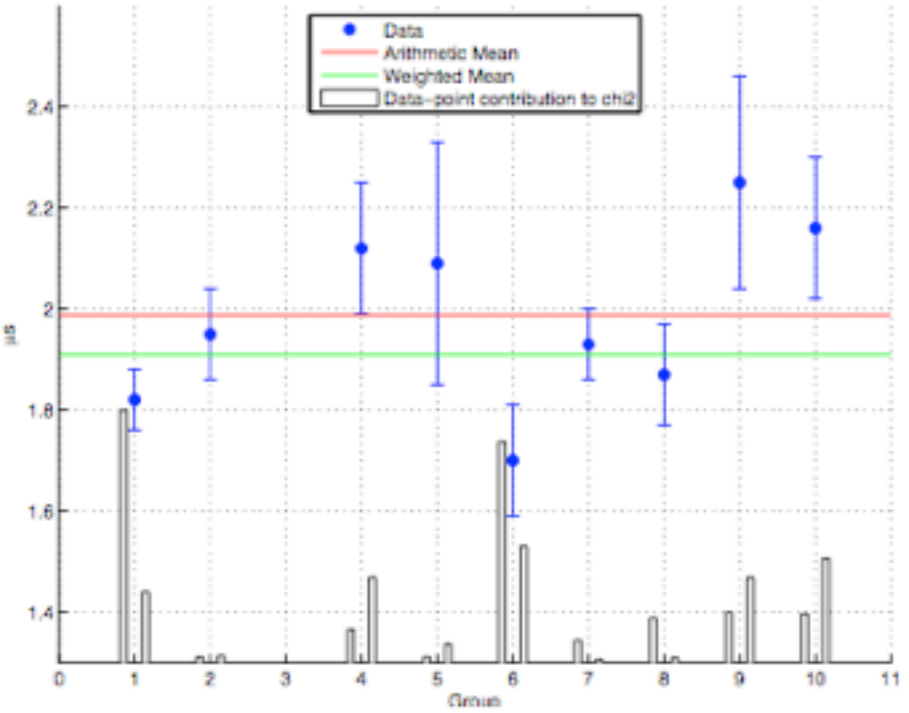
The χ^2 value for these two results are found from Barlow (6.1) giving

$$\chi_n{}^2 = \sum_{i=1}^{10}\frac{(x_i - \bar{x}_n)^2}{\sigma_i^2} = 31{,}64 \quad \chi_w{}^2 = \sum_{i=1}^{10}\frac{(x_i - \bar{x}_w)^2}{\sigma_i^2} = 28{,}33$$

Problem 4.1



(a) *All data*

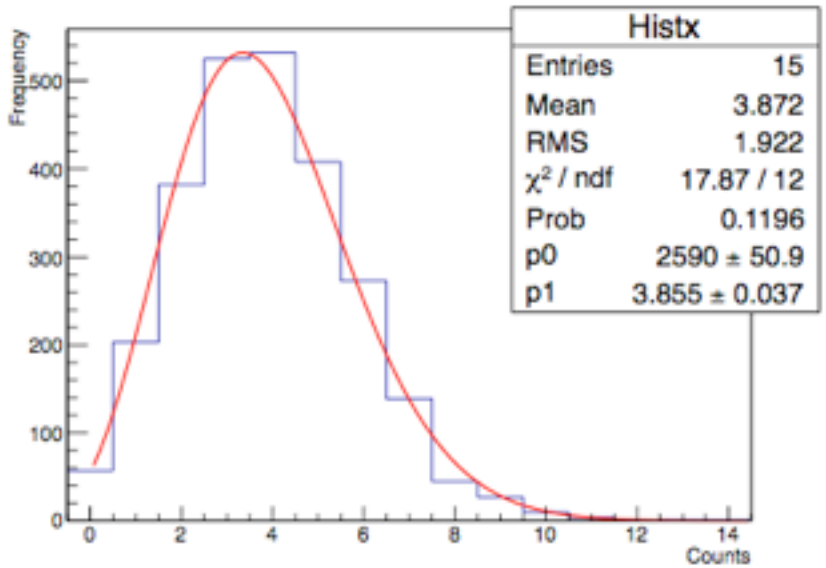


(b) *"Outlier" excluded*

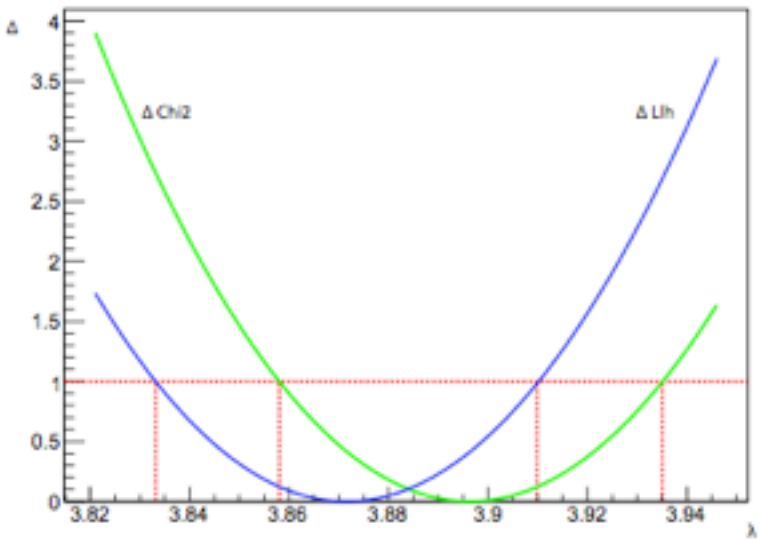
Figure 3: ***Left:*** *Data with error bars plotted along with the arithmetic and weighted mean. The bars represent the contribution from each point to the* χ^2 *(left bars are the* χ^2 *found with the arithmetic mean)*

Problem 4.2

$$\lambda_{\chi^2} = 3.896\,^{+0.039}_{-0.038} \qquad \lambda_{-2\ln(L)} = 3.872\,^{+0.0384}_{-0.0387}\,.$$



(a). Histogram over the Dataset in table 4 with a fitted gaussian function.



(b). The distribution of $\delta\chi^2$ and $\delta\left(-2\ln(L)\right)$ around in the range $\lambda = [3.821, 3.946]$.

Figure 2. Analysis of the classic 1910 dataset on Polonium 210 decays by Rutherford and Geiger.

Problem 4.2

- Polonium-210's halveringstid er 138,4 dage [Wikipedia], så i de godt to dage, som eksperimentet har taget, hvis målingerne er lavet i et stræk, hvilket man må formode, er polonium-prøvens aktivitet faldet med

$$1 - e^{-\frac{t_{\text{exp}}}{\tau_{1/2}} \ln 2} = 1,1\%.$$

De målte frekvenser må så fordele sig ifølge hver sin Poisson-fordeling, hvor middelværdien har flyttet sig med de 1,1% af hvad den var da forsøget startede. Vi får stadig en Poisson-fordeling ud, fordi summen af Poisson-fordelinger stadig er en Poisson-fordeling, med en middelværdi, der ligger et sted i midten af feltet, og med en spredning circa på 0,05%. Den systematiske fejl fra denne effekt lander på ±0,02, circa samme størrelse som den statistiske fejl. Denne usikkerhed ville have været mindre med et kortere eksperiment, men når den statistiske usikkerhed går som

$$\frac{1}{\sqrt{N}},$$

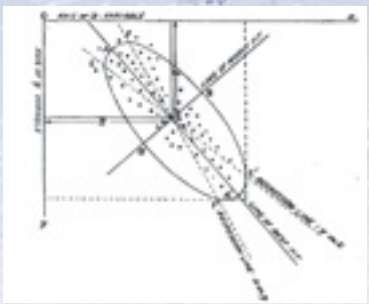
mens den systematiske usikkerhed tilnærmelsesvis stiger lineært for så små værdier af $t_{\text{exp}}/\tau_{1/2}$, bliver den statistiske usikkerhed større hurtigere end den systematiske usikkerhed bliver mindre, når N bliver mindre. Det kunne til gengæld have hjulpet at rapportere tiden for hver måling med, så de løbende kan korrigeres for faldet i aktivitet, og på den måde reducere den systematiske usikkerhed uden at hæve den statistiske.

Problem 5.1

	hypothesis	χ^2	ndf	Probability
1	$a+bx$	48.20	9	5.69×10^{-7}
2	$a+bx+cx^2$	21.24	8	1.164×10^{-2}
3	$a+bx+c\sqrt{kx}$	35.1	8	2.57×10^{-5}
4	$a+bx+c\sqrt{k(x+2)}$	37.66	7	8.70×10^{-6}
5	$a+bx+cx^2+kx^3$	17.03	7	2.981×10^{-2}
6	$a+bx+c\sin(k(x-\pi/2))$	14.53	7	6.889×10^{-2}
7	$a+bsin(c(x+2-\pi/2))$	17.55	8	4.075×10^{-2}

X

The



Top 10

most important things in applied statistics

1. Errors decrease with the **square root of N**
2. The **ChiSquare** is simple, powerful, and robust and provides a fit quality measure
3. Binomial distribution → Poisson distribution → Gaussian distribution
4. The error is sqrt(N) on a (Poisson) number and sqrt(f (1-f)/N) on a fraction
5. Correlations are important and needs consideration
6. The likelihood (ratio) is generally the optimal estimator (test)
7. Low statistics is terrible – needs special attention
8. Error propagation is craftsmanship/fitting is an art
9. Prior probabilities needs attention, i.e. Bayes' Theorem
10. Hypothesis testing is done with a test statistic t (e.g. Likelihood ratio, Fisher, etc.)