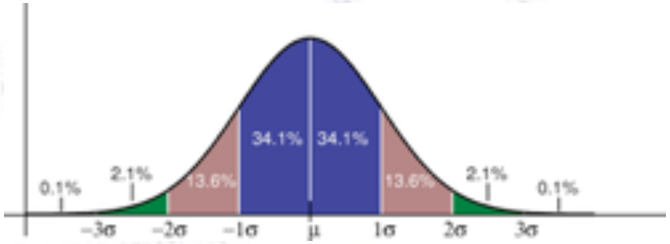
Applied Statistics

Exam - Solution & Discussion

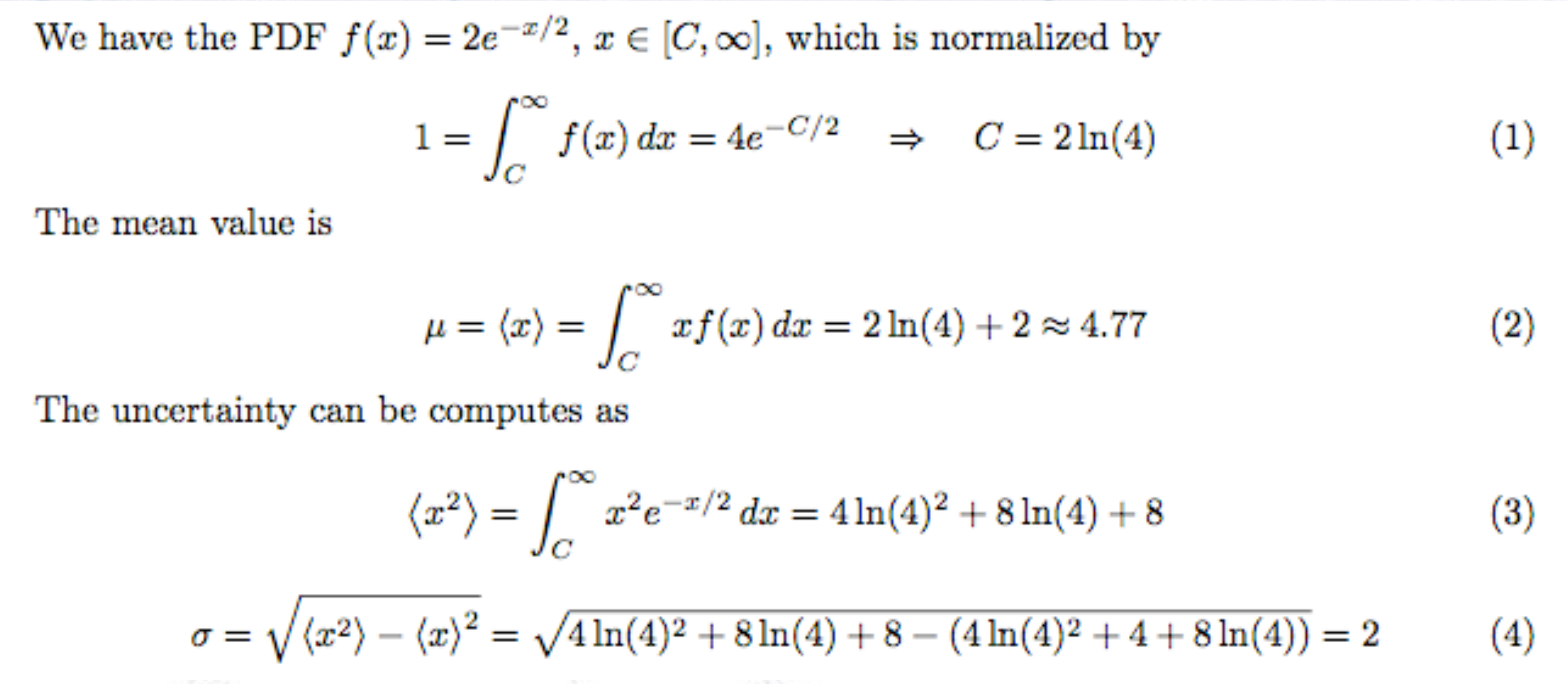
Troels C. Petersen (NBI)



"Statistics is merely a quantization of common sense"

Problem 1.1 (2+2 points)

Classic problem. Exponential with minimum (and thus mean) shifted by C, while width is left unchanged.



We have the PDF $f(x) = 2e^{-x/2}$, $x \in [C, \infty]$, which is normalized by

$$1 = \int_C^\infty f(x)\,dx = 4e^{-C/2} \quad \Rightarrow \quad C = 2\ln(4) \tag{1}$$

The mean value is

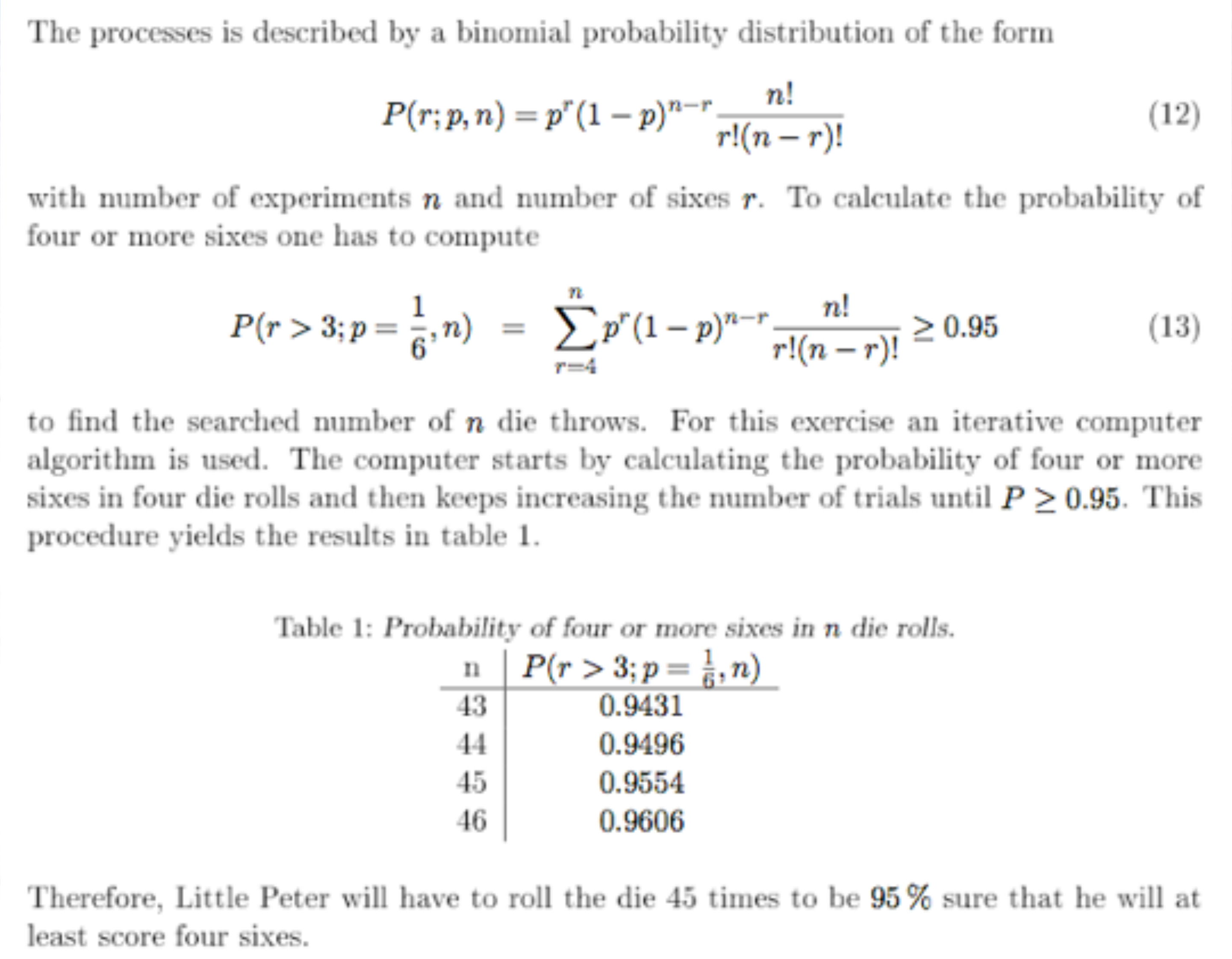
$$\mu = \langle x \rangle = \int_C^\infty x f(x)\,dx = 2\ln(4) + 2 \approx 4.77 \tag{2}$$

The uncertainty can be computes as

$$\langle x^2 \rangle = \int_C^\infty x^2 e^{-x/2}\,dx = 4\ln(4)^2 + 8\ln(4) + 8 \tag{3}$$

$$\sigma = \sqrt{\langle x^2 \rangle - \langle x \rangle^2} = \sqrt{4\ln(4)^2 + 8\ln(4) + 8 - (4\ln(4)^2 + 4 + 8\ln(4))} = 2 \tag{4}$$

Problem 1.2 (3 points)



The processes is described by a binomial probability distribution of the form

$$P(r; p, n) = p^r (1-p)^{n-r} \frac{n!}{r!(n-r)!} \tag{12}$$

with number of experiments n and number of sixes r. To calculate the probability of four or more sixes one has to compute

$$P(r > 3; p = \frac{1}{6}, n) \quad = \quad \sum_{r=4}^{n} p^r (1-p)^{n-r} \frac{n!}{r!(n-r)!} \geq 0.95 \tag{13}$$

to find the searched number of n die throws. For this exercise an iterative computer algorithm is used. The computer starts by calculating the probability of four or more sixes in four die rolls and then keeps increasing the number of trials until $P \geq 0.95$. This procedure yields the results in table 1.

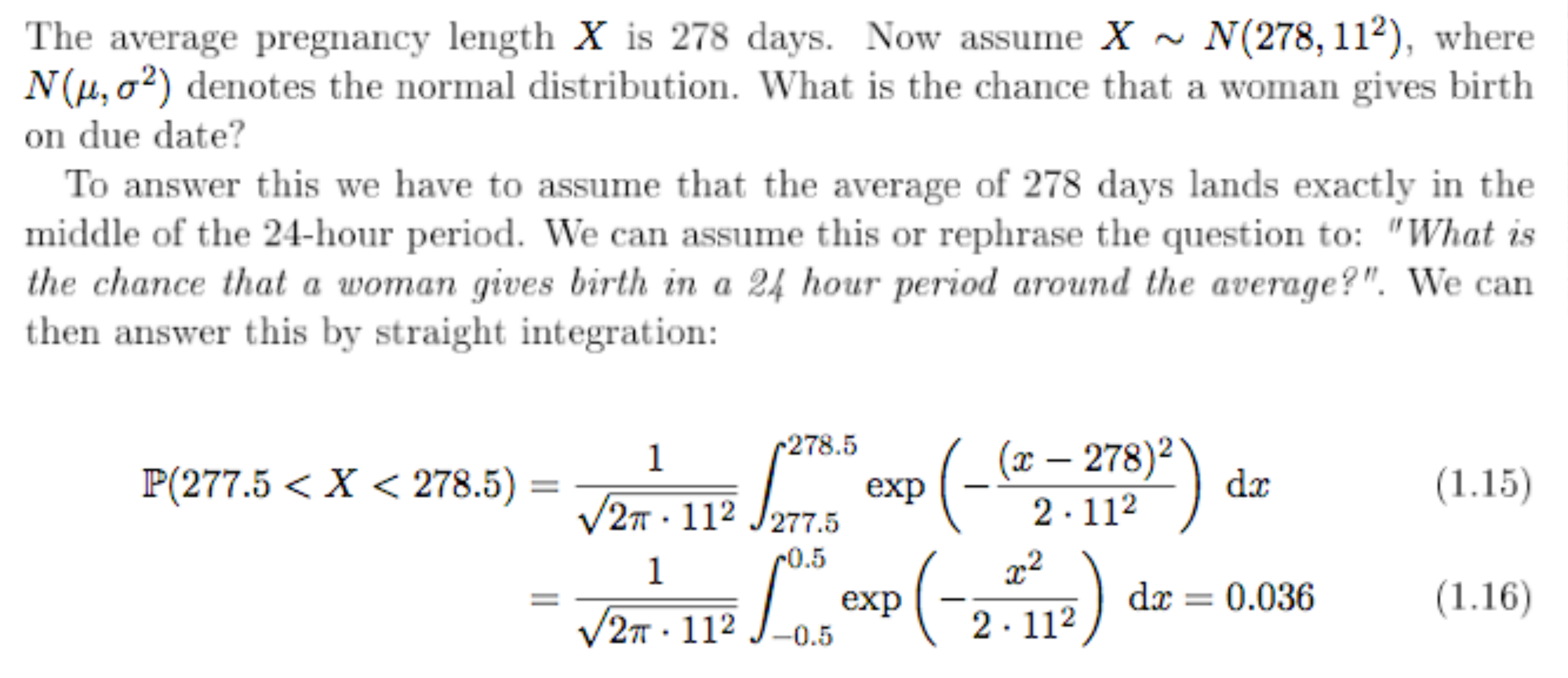
Table 1: *Probability of four or more sixes in n die rolls.*

n	$P(r > 3; p = \frac{1}{6}, n)$
43	0.9431
44	0.9496
45	0.9554
46	0.9606

Therefore, Little Peter will have to roll the die 45 times to be 95 % sure that he will at least score four sixes.

Problem 1.3 (5 points)

Another standard problem…



The average pregnancy length X is 278 days. Now assume $X \sim N(278, 11^2)$, where $N(\mu, \sigma^2)$ denotes the normal distribution. What is the chance that a woman gives birth on due date?

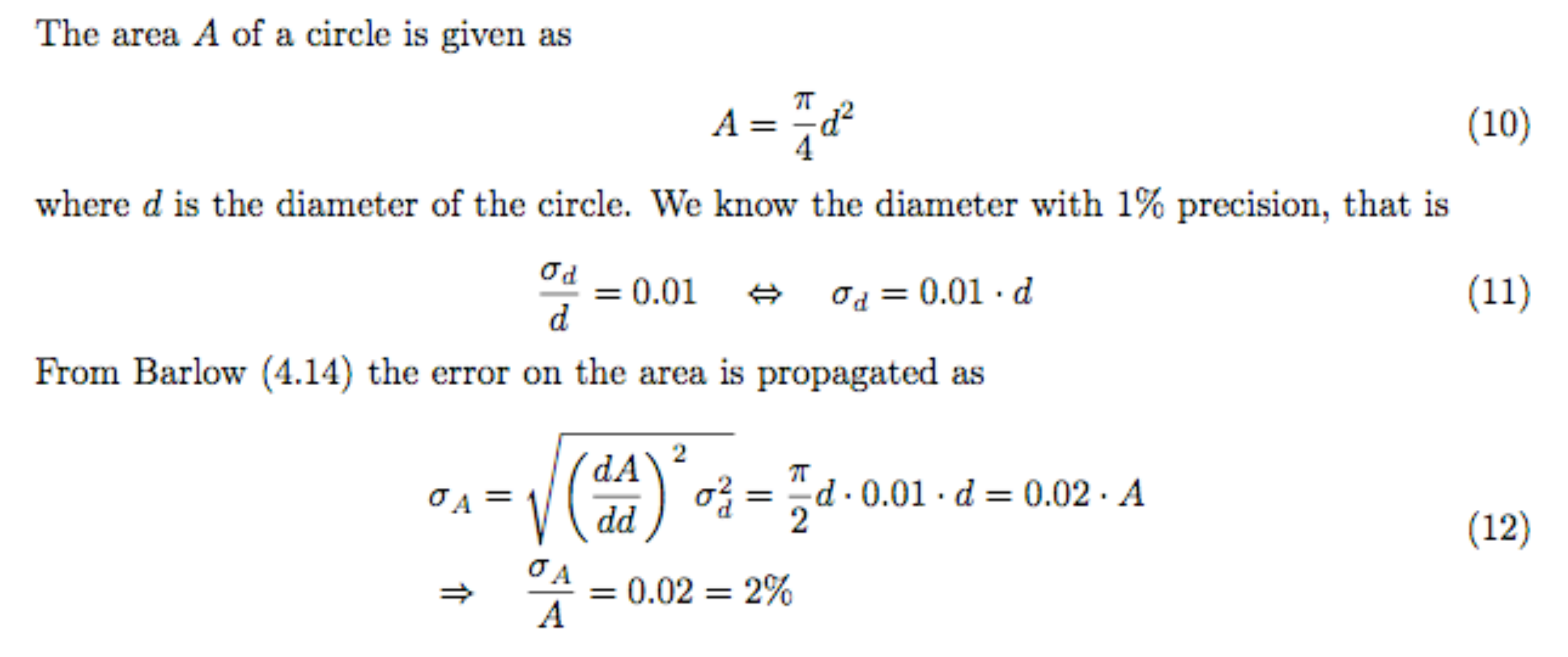
To answer this we have to assume that the average of 278 days lands exactly in the middle of the 24-hour period. We can assume this or rephrase the question to: *"What is the chance that a woman gives birth in a 24 hour period around the average?"*. We can then answer this by straight integration:

$$\mathbb{P}(277.5 < X < 278.5) = \frac{1}{\sqrt{2\pi \cdot 11^2}} \int_{277.5}^{278.5} \exp\left(-\frac{(x-278)^2}{2 \cdot 11^2}\right) \mathrm{d}x \qquad (1.15)$$

$$= \frac{1}{\sqrt{2\pi \cdot 11^2}} \int_{-0.5}^{0.5} \exp\left(-\frac{x^2}{2 \cdot 11^2}\right) \mathrm{d}x = 0.036 \qquad (1.16)$$

Problem 2.1 (2 points)

Simple... here well explained.



The area A of a circle is given as

$$A = \frac{\pi}{4} d^2 \tag{10}$$

where d is the diameter of the circle. We know the diameter with 1% precision, that is

$$\frac{\sigma_d}{d} = 0.01 \quad \Leftrightarrow \quad \sigma_d = 0.01 \cdot d \tag{11}$$

From Barlow (4.14) the error on the area is propagated as

$$\begin{aligned} \sigma_A &= \sqrt{\left(\frac{dA}{dd}\right)^2 \sigma_d^2} = \frac{\pi}{2} d \cdot 0.01 \cdot d = 0.02 \cdot A \\ \Rightarrow \quad & \frac{\sigma_A}{A} = 0.02 = 2\% \end{aligned} \tag{12}$$

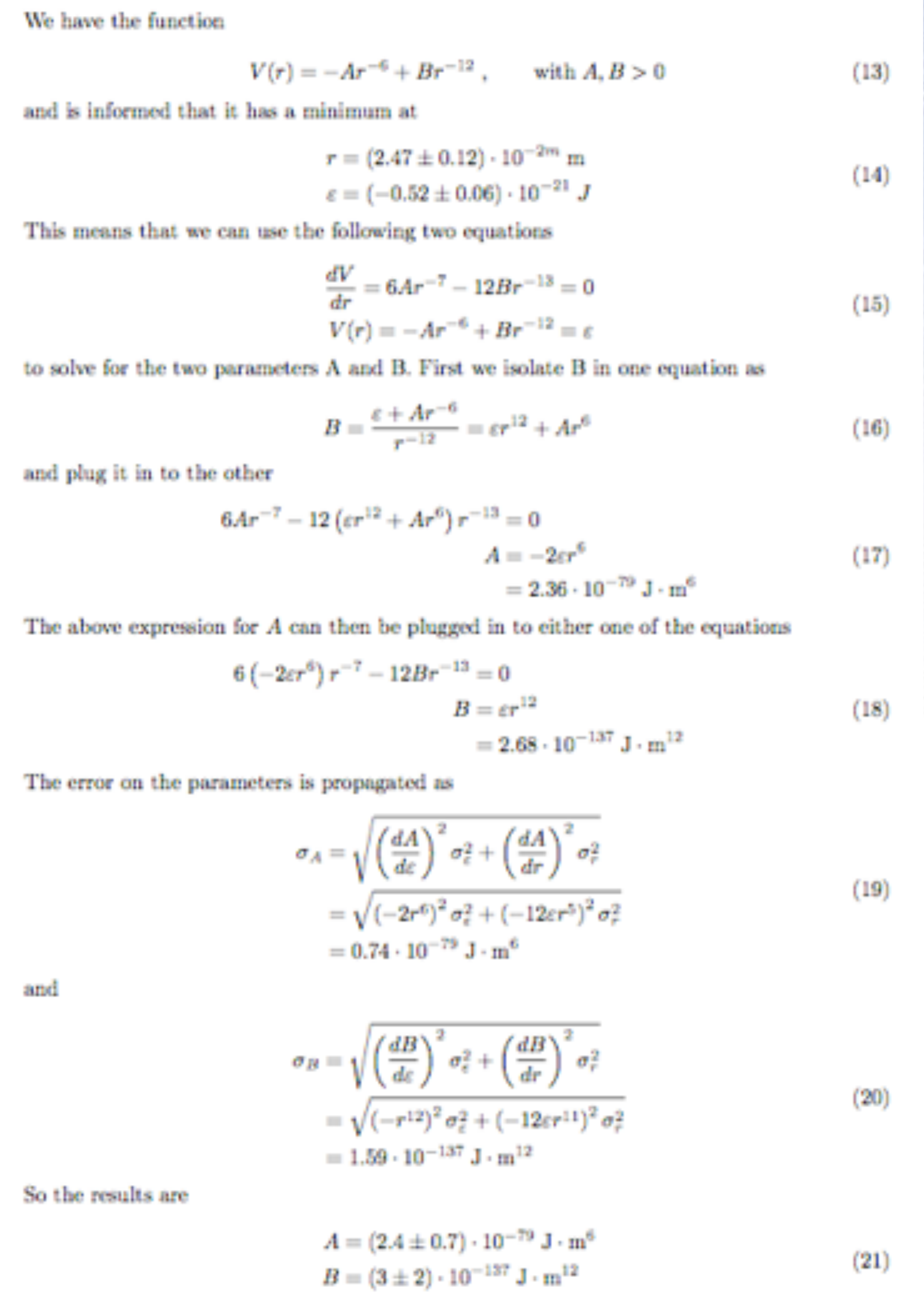
Problem 2.2

(8 points)

First real challenge. Two equations with two unknowns... but not too hard! The units are a bit crazy...

It is OK to give the resulting error on B with two significant digits as well (three is a bit too much!).

Extra points (beyond the 8) are awarded for noting the the errors are actually quite large, and so the error-propagation formula might not give a completely accurate result.
If an MonteCarlo is used to get a better result... that is fantastic!



We have the function

$$V(r) = -Ar^{-6} + Br^{-12}\,, \qquad \text{with } A, B > 0 \tag{13}$$

and is informed that it has a minimum at

$$\begin{aligned} r &= (2.47 \pm 0.12) \cdot 10^{-2m}\ \text{m} \\ \varepsilon &= (-0.52 \pm 0.06) \cdot 10^{-21}\ J \end{aligned} \tag{14}$$

This means that we can use the following two equations

$$\begin{aligned} \frac{dV}{dr} &= 6Ar^{-7} - 12Br^{-13} = 0 \\ V(r) &= -Ar^{-6} + Br^{-12} = \varepsilon \end{aligned} \tag{15}$$

to solve for the two parameters A and B. First we isolate B in one equation as

$$B = \frac{\varepsilon + Ar^{-6}}{r^{-12}} = \varepsilon r^{12} + Ar^{6} \tag{16}$$

and plug it in to the other

$$\begin{aligned} 6Ar^{-7} - 12\left(\varepsilon r^{12} + Ar^{6}\right) r^{-13} &= 0 \\ A &= -2\varepsilon r^{6} \\ &= 2.36 \cdot 10^{-79}\ \text{J} \cdot \text{m}^{6} \end{aligned} \tag{17}$$

The above expression for A can then be plugged in to either one of the equations

$$\begin{aligned} 6\left(-2\varepsilon r^{6}\right) r^{-7} - 12Br^{-13} &= 0 \\ B &= \varepsilon r^{12} \\ &= 2.68 \cdot 10^{-137}\ \text{J} \cdot \text{m}^{12} \end{aligned} \tag{18}$$

The error on the parameters is propagated as

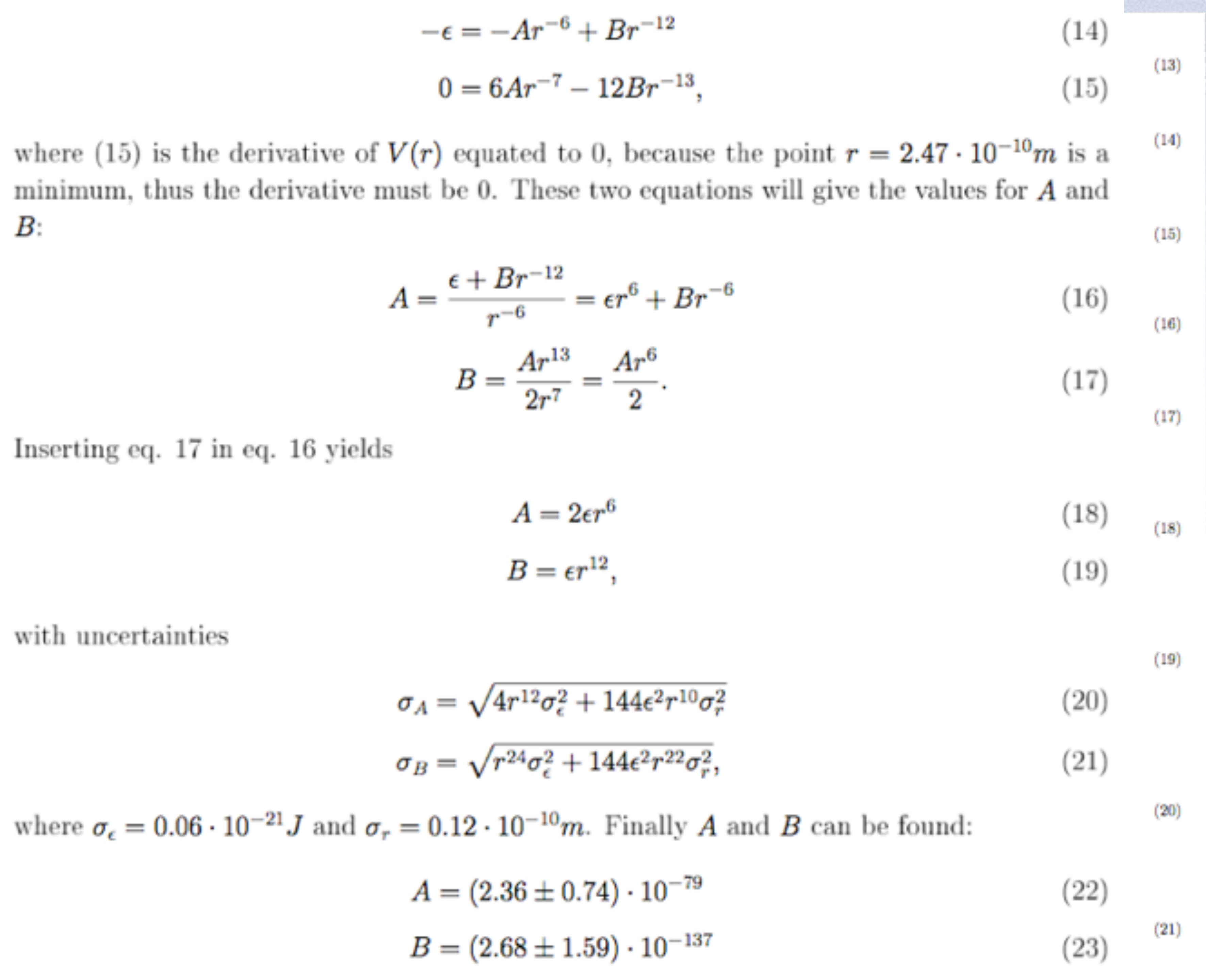
$$\begin{aligned} \sigma_A &= \sqrt{\left(\frac{dA}{d\varepsilon}\right)^2 \sigma_\varepsilon^2 + \left(\frac{dA}{dr}\right)^2 \sigma_r^2} \\ &= \sqrt{(-2r^{6})^2 \sigma_\varepsilon^2 + (-12\varepsilon r^{5})^2 \sigma_r^2} \\ &= 0.74 \cdot 10^{-79}\ \text{J} \cdot \text{m}^{6} \end{aligned} \tag{19}$$

and

$$\begin{aligned} \sigma_B &= \sqrt{\left(\frac{dB}{d\varepsilon}\right)^2 \sigma_\varepsilon^2 + \left(\frac{dB}{dr}\right)^2 \sigma_r^2} \\ &= \sqrt{(-r^{12})^2 \sigma_\varepsilon^2 + (-12\varepsilon r^{11})^2 \sigma_r^2} \\ &= 1.59 \cdot 10^{-137}\ \text{J} \cdot \text{m}^{12} \end{aligned} \tag{20}$$

So the results are

$$\begin{aligned} A &= (2.4 \pm 0.7) \cdot 10^{-79}\ \text{J} \cdot \text{m}^{6} \\ B &= (3 \pm 2) \cdot 10^{-137}\ \text{J} \cdot \text{m}^{12} \end{aligned} \tag{21}$$



$$-\epsilon = -Ar^{-6} + Br^{-12} \tag{14}$$

$$0 = 6Ar^{-7} - 12Br^{-13}, \tag{15}$$

where (15) is the derivative of $V(r)$ equated to 0, because the point $r = 2.47 \cdot 10^{-10}m$ is a minimum, thus the derivative must be 0. These two equations will give the values for A and B:

$$A = \frac{\epsilon + Br^{-12}}{r^{-6}} = \epsilon r^{6} + Br^{-6} \tag{16}$$

$$B = \frac{Ar^{13}}{2r^{7}} = \frac{Ar^{6}}{2}. \tag{17}$$

Inserting eq. 17 in eq. 16 yields

$$A = 2\epsilon r^{6} \tag{18}$$

$$B = \epsilon r^{12}, \tag{19}$$

with uncertainties

$$\sigma_A = \sqrt{4r^{12}\sigma_\epsilon^2 + 144\epsilon^2 r^{10}\sigma_r^2} \tag{20}$$

$$\sigma_B = \sqrt{r^{24}\sigma_\epsilon^2 + 144\epsilon^2 r^{22}\sigma_r^2}, \tag{21}$$

where $\sigma_\epsilon = 0.06 \cdot 10^{-21}J$ and $\sigma_r = 0.12 \cdot 10^{-10}m$. Finally A and B can be found:

$$A = (2.36 \pm 0.74) \cdot 10^{-79} \tag{22}$$

$$B = (2.68 \pm 1.59) \cdot 10^{-137} \tag{23}$$

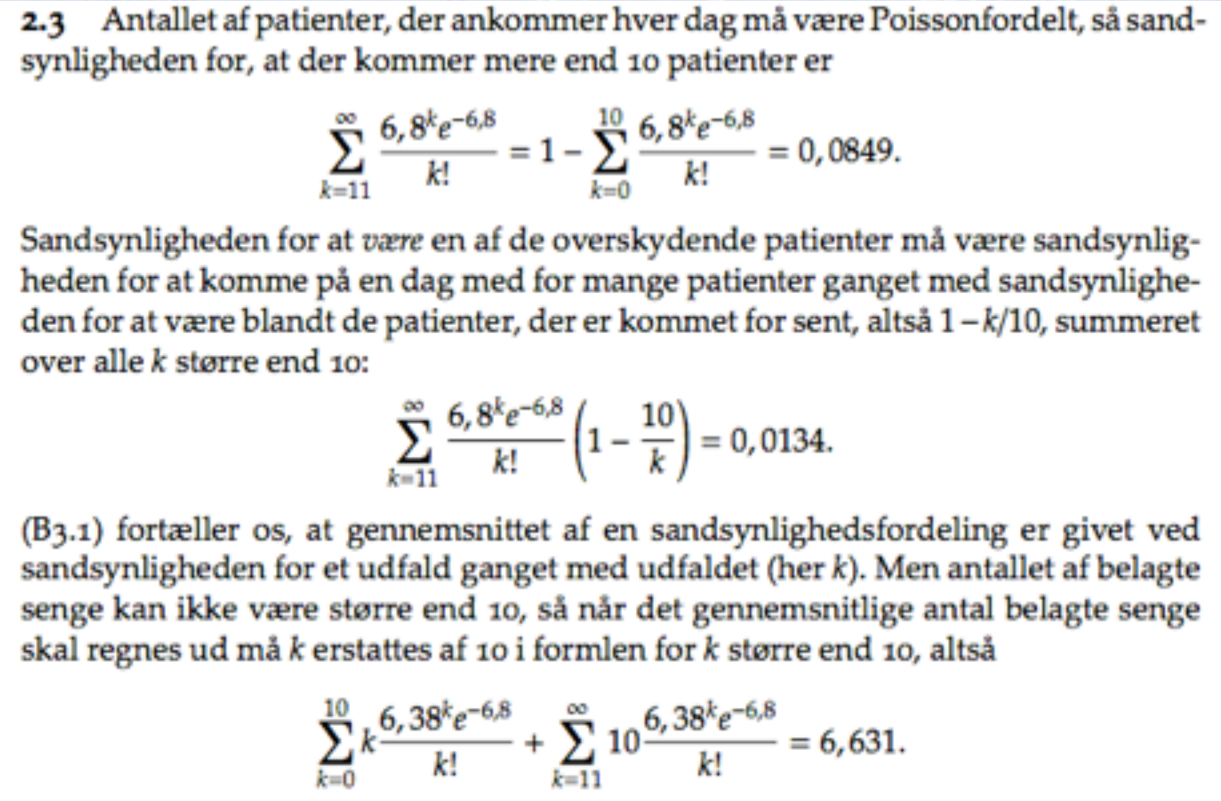
Problem 2.3 (3+4+4 points)

This was a harder problem than I first imagined.

First question has two acceptable solutions, depending interpretation.

Second question one has to remember to include the 10+ cases (at weight 10).

Third question was really hard! Simulation required.



2.3 Antallet af patienter, der ankommer hver dag må være Poissonfordelt, så sandsynligheden for, at der kommer mere end 10 patienter er

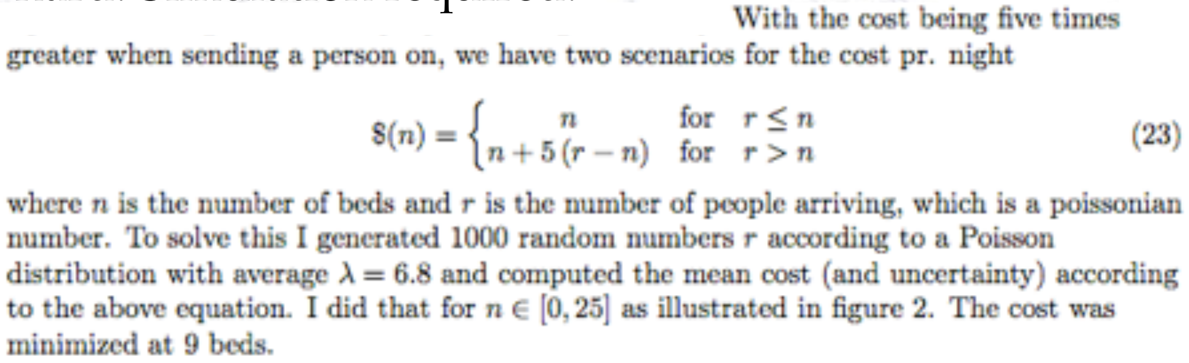
$$\sum_{k=11}^{\infty} \frac{6,8^k e^{-6,8}}{k!} = 1 - \sum_{k=0}^{10} \frac{6,8^k e^{-6,8}}{k!} = 0,0849.$$

Sandsynligheden for at *være* en af de overskydende patienter må være sandsynligheden for at komme på en dag med for mange patienter ganget med sandsynligheden for at være blandt de patienter, der er kommet for sent, altså $1 - k/10$, summeret over alle k større end 10:

$$\sum_{k=11}^{\infty} \frac{6,8^k e^{-6,8}}{k!} \left(1 - \frac{10}{k}\right) = 0,0134.$$

(B3.1) fortæller os, at gennemsnittet af en sandsynlighedsfordeling er givet ved sandsynligheden for et udfald ganget med udfaldet (her k). Men antallet af belagte senge kan ikke være større end 10, så når det gennemsnitlige antal belagte senge skal regnes ud må k erstattes af 10 i formlen for k større end 10, altså

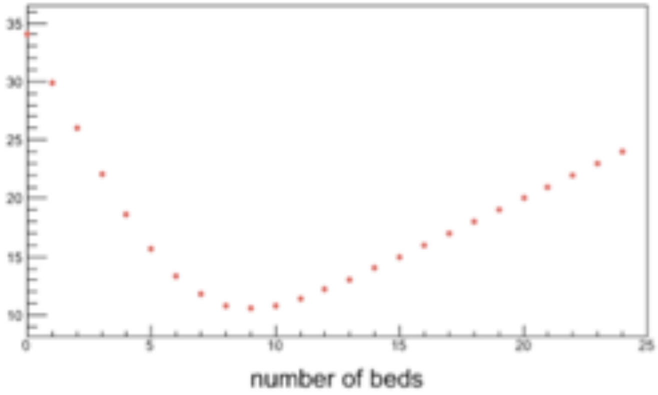
$$\sum_{k=0}^{10} k \frac{6,38^k e^{-6,8}}{k!} + \sum_{k=11}^{\infty} 10 \frac{6,38^k e^{-6,8}}{k!} = 6,631.$$



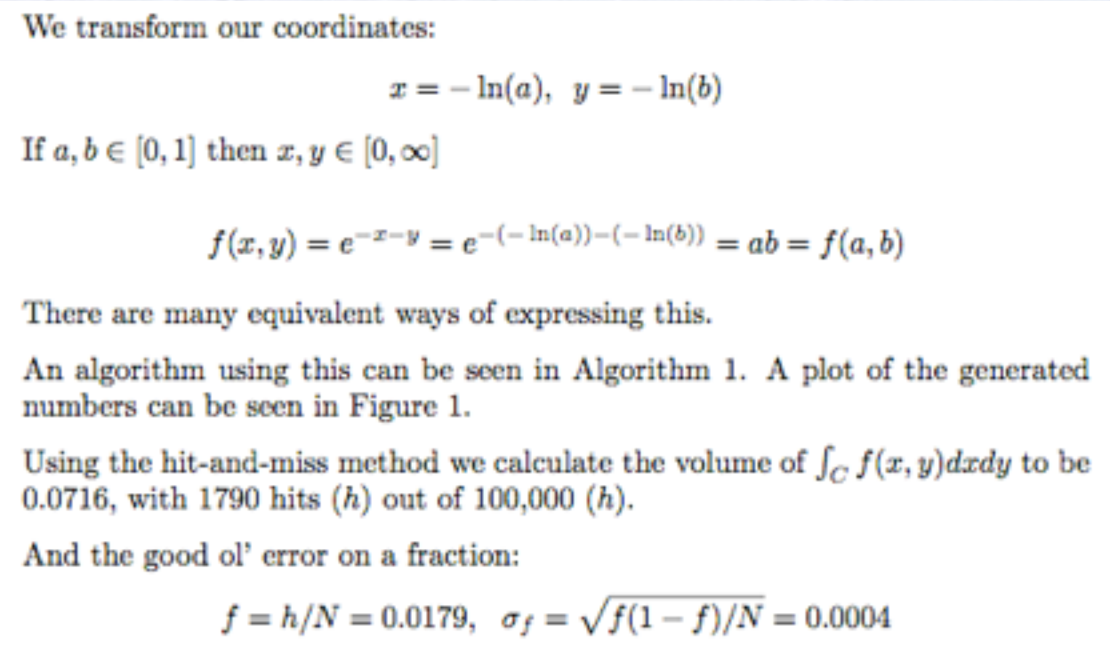
With the cost being five times greater when sending a person on, we have two scenarios for the cost pr. night

$$\mathcal{S}(n) = \begin{cases} n & \text{for} \quad r \leq n \\ n + 5\,(r - n) & \text{for} \quad r > n \end{cases} \tag{23}$$

where n is the number of beds and r is the number of people arriving, which is a poissonian number. To solve this I generated 1000 random numbers r according to a Poisson distribution with average $\lambda = 6.8$ and computed the mean cost (and uncertainty) according to the above equation. I did that for $n \in [0, 25]$ as illustrated in figure 2. The cost was minimized at 9 beds.



Problem 3.1 (2+5+5 points)



We transform our coordinates:

$$x = -\ln(a), \quad y = -\ln(b)$$

If $a, b \in [0,1]$ then $x, y \in [0, \infty]$

$$f(x,y) = e^{-x-y} = e^{-(-\ln(a))-(-\ln(b))} = ab = f(a,b)$$

There are many equivalent ways of expressing this.

An algorithm using this can be seen in Algorithm 1. A plot of the generated numbers can be seen in Figure 1.

Using the hit-and-miss method we calculate the volume of $\int_C f(x,y)dxdy$ to be 0.0716, with 1790 hits (h) out of 100,000 (h).

And the good ol' error on a fraction:

$$f = h/N = 0.0179, \quad \sigma_f = \sqrt{f(1-f)/N} = 0.0004$$

The volume is:

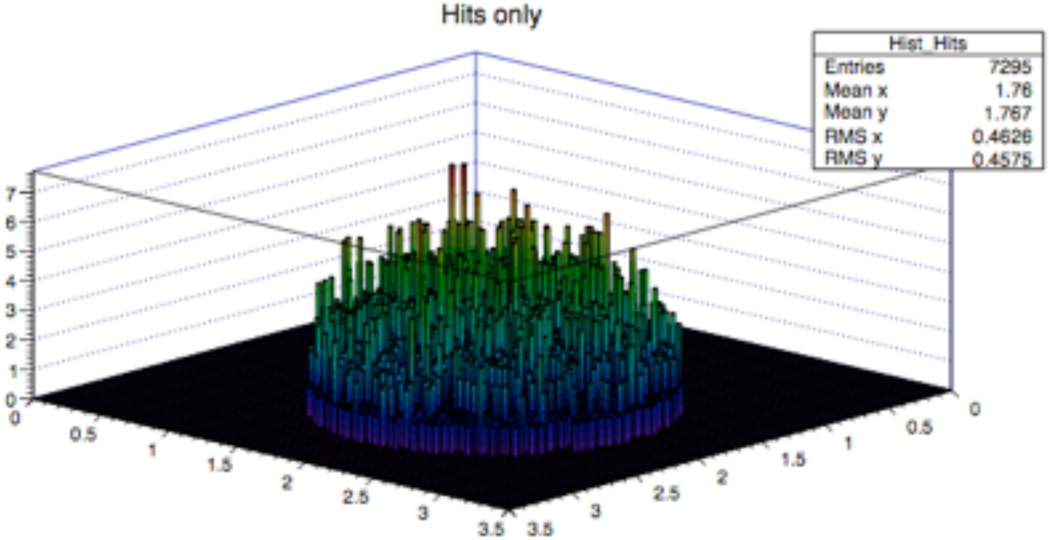
$$V = f \cdot V_{box} = 4f, \quad \sigma_V = \sigma_f \cdot V_{box}$$

And so the final result $V = \underline{0.0716 \pm 0.0017}$.

The transformation method is to be used for generating f(x,y), and either that or the Hit&Miss for estimating the integral.

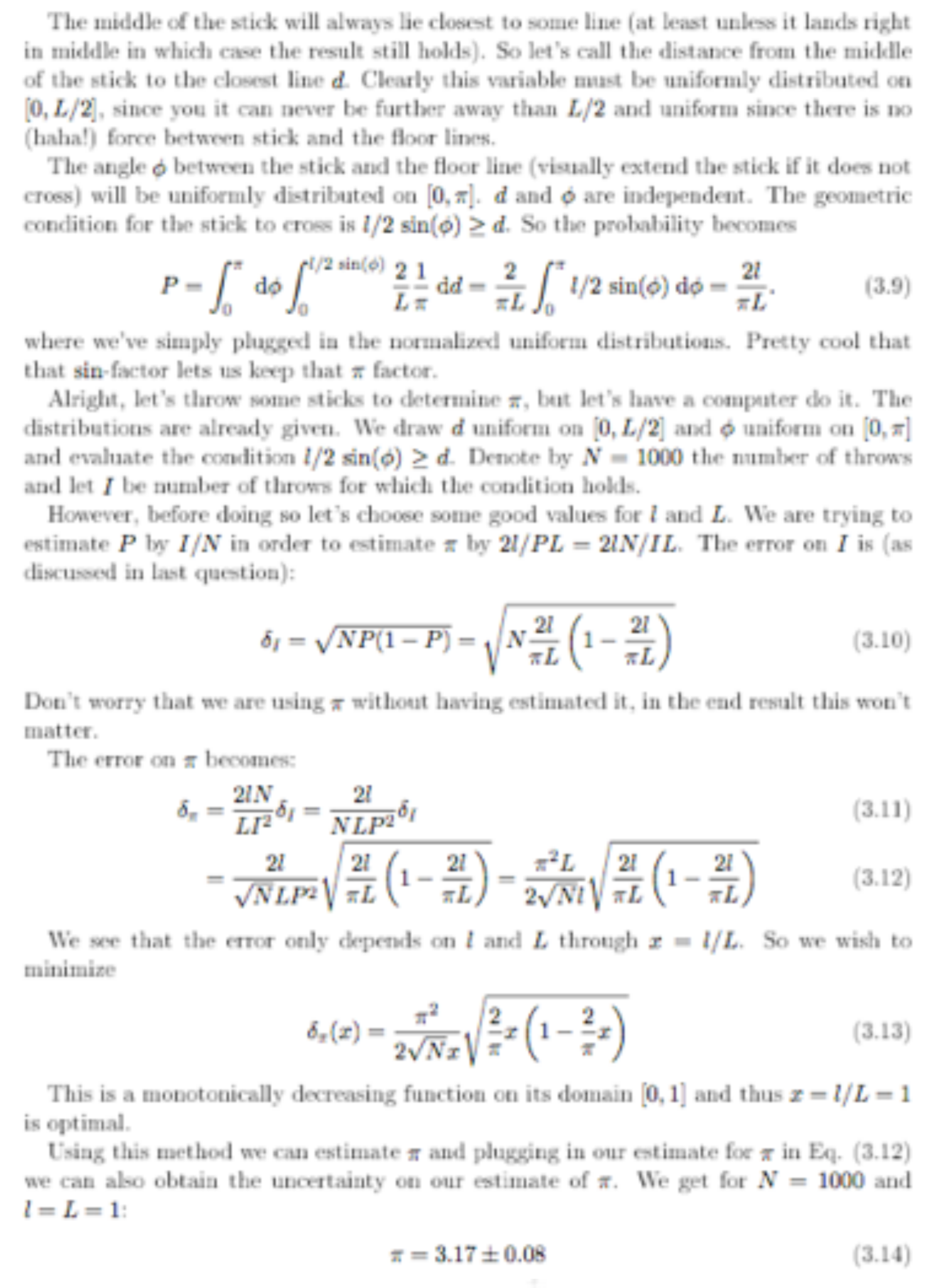
A plot is enough to show that they got the first two problems right.

The estimate of the integral should have an uncertainty roughtly in the range 0.0007-0.0018 (depending on interpretation of how to produce points).



Problem 3.2

(3+4+4+4 points)



The middle of the stick will always lie closest to some line (at least unless it lands right in middle in which case the result still holds). So let's call the distance from the middle of the stick to the closest line d. Clearly this variable must be uniformly distributed on $[0, L/2]$, since you it can never be further away than $L/2$ and uniform since there is no (haha!) force between stick and the floor lines.

The angle ϕ between the stick and the floor line (visually extend the stick if it does not cross) will be uniformly distributed on $[0, \pi]$. d and ϕ are independent. The geometric condition for the stick to cross is $l/2 \sin(\phi) \geq d$. So the probability becomes

$$P = \int_0^\pi \mathrm{d}\phi \int_0^{l/2 \sin(\phi)} \frac{2}{L}\frac{1}{\pi}\, \mathrm{d}d = \frac{2}{\pi L}\int_0^\pi l/2 \sin(\phi)\, \mathrm{d}\phi = \frac{2l}{\pi L}. \tag{3.9}$$

where we've simply plugged in the normalized uniform distributions. Pretty cool that that sin-factor lets us keep that π factor.

Alright, let's throw some sticks to determine π, but let's have a computer do it. The distributions are already given. We draw d uniform on $[0, L/2]$ and ϕ uniform on $[0, \pi]$ and evaluate the condition $l/2 \sin(\phi) \geq d$. Denote by $N = 1000$ the number of throws and let I be number of throws for which the condition holds.

However, before doing so let's choose some good values for l and L. We are trying to estimate P by I/N in order to estimate π by $2l/PL = 2lN/IL$. The error on I is (as discussed in last question):

$$\delta_I = \sqrt{NP(1-P)} = \sqrt{N\frac{2l}{\pi L}\left(1 - \frac{2l}{\pi L}\right)} \tag{3.10}$$

Don't worry that we are using π without having estimated it, in the end result this won't matter.

The error on π becomes:

$$\delta_\pi = \frac{2lN}{LI^2}\delta_I = \frac{2l}{NLP^2}\delta_I \tag{3.11}$$

$$= \frac{2l}{\sqrt{N}LP^2}\sqrt{\frac{2l}{\pi L}\left(1 - \frac{2l}{\pi L}\right)} = \frac{\pi^2 L}{2\sqrt{N}l}\sqrt{\frac{2l}{\pi L}\left(1 - \frac{2l}{\pi L}\right)} \tag{3.12}$$

We see that the error only depends on l and L through $x = l/L$. So we wish to minimize

$$\delta_\pi(x) = \frac{\pi^2}{2\sqrt{N}x}\sqrt{\frac{2}{\pi}x\left(1 - \frac{2}{\pi}x\right)} \tag{3.13}$$

This is a monotonically decreasing function on its domain $[0, 1]$ and thus $x = l/L = 1$ is optimal.

Using this method we can estimate π and plugging in our estimate for π in Eq. (3.12) we can also obtain the uncertainty on our estimate of π. We get for $N = 1000$ and $l = L = 1$:

$$\pi = 3.17 \pm 0.08 \tag{3.14}$$

The first problem should either have a relevant integral (as shown) or a ratio argument.

The value should have an uncertainty of about 0.08 give and take a little (if people used 1000).

The higher l/L the better (till 1), so l/L = 1 is optimal.

The last problem is not very precise, but people should realize that one gets a value of pi with 6 digits correct - and that this normally requires 10^{12} - 10^{14} throws. But not necessarily “unrealistic”, when N=213 and l/L = 5/6.

Problem 3.2

(3+4+4+4 points)

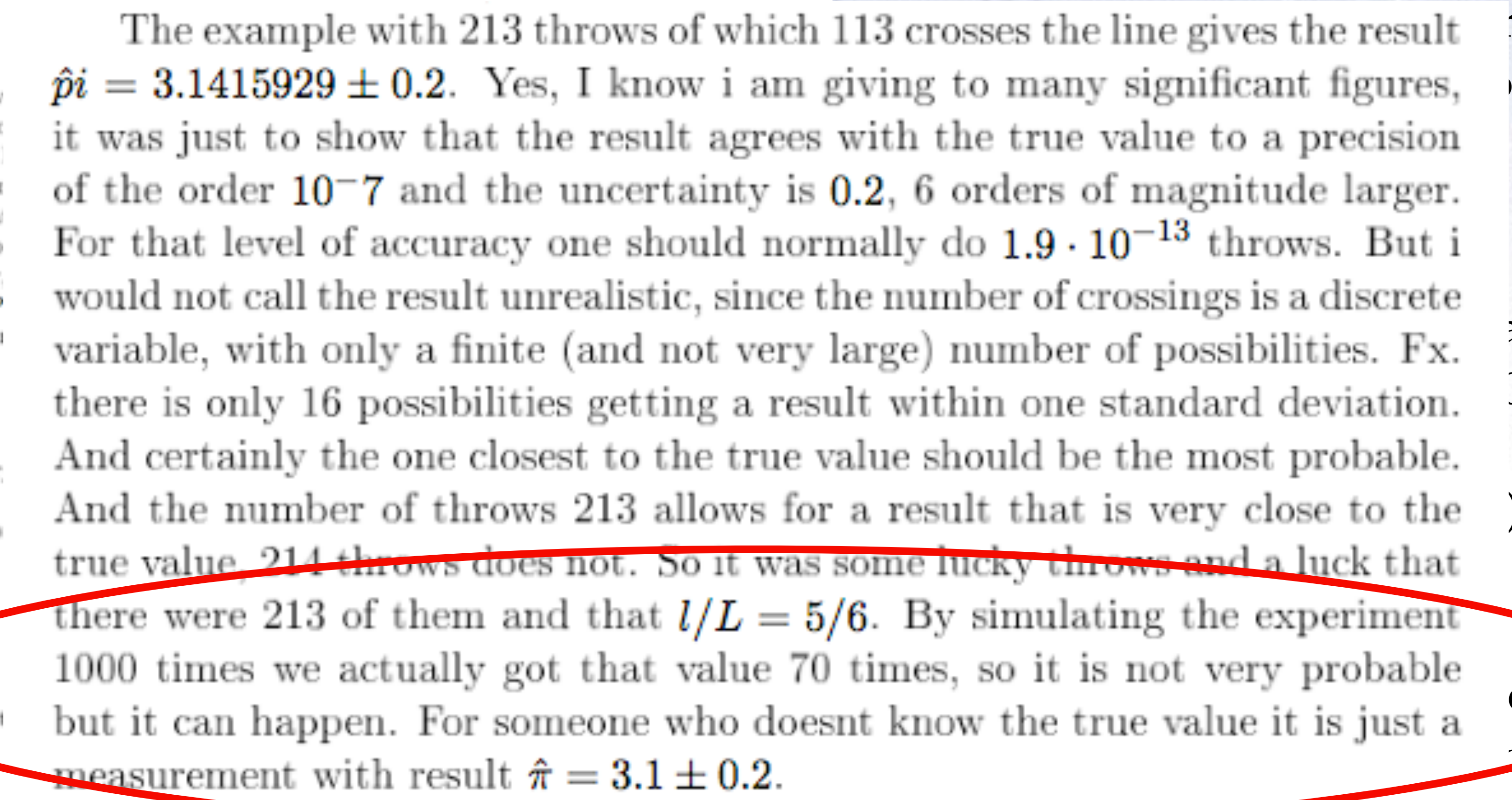
The middle of the stick will always lie closest to some line (at least unless it lands right in middle in which case the result still holds). So let's call the distance from the middle of the stick to the closest line d. Clearly this variable must be uniformly distributed on $[0, L/2]$, since you it can never be further away than $L/2$ and uniform since there is no (haha!) force between stick and the floor lines.

The angle ϕ between the stick and the floor line (visually extend the stick if it does not cross) will be uniformly distributed on $[0, \pi]$. d and ϕ are independent. The geometric condition for the stick to cross is $l/2 \sin(\phi) \geq d$. So the probability becomes

where we'v
that sin-fac
Alright, l
distributio
and evalua
and let I b
However,
estimate P
discussed i

Don't worr
matter.
The erro

We see t
minimize



The example with 213 throws of which 113 crosses the line gives the result $\hat{pi} = 3.1415929 \pm 0.2$. Yes, I know i am giving to many significant figures, it was just to show that the result agrees with the true value to a precision of the order $10^{-}7$ and the uncertainty is 0.2, 6 orders of magnitude larger. For that level of accuracy one should normally do $1.9 \cdot 10^{-13}$ throws. But i would not call the result unrealistic, since the number of crossings is a discrete variable, with only a finite (and not very large) number of possibilities. Fx. there is only 16 possibilities getting a result within one standard deviation. And certainly the one closest to the true value should be the most probable. And the number of throws 213 allows for a result that is very close to the true value, 214 throws does not. So it was some lucky throws and a luck that there were 213 of them and that $l/L = 5/6$. By simulating the experiment 1000 times we actually got that value 70 times, so it is not very probable but it can happen. For someone who doesnt know the true value it is just a measurement with result $\hat{\pi} = 3.1 \pm 0.2$.

This is a monotonically decreasing function on its domain $[0,1]$ and thus $x = l/L = 1$ is optimal.

Using this method we can estimate π and plugging in our estimate for π in Eq. (3.12) we can also obtain the uncertainty on our estimate of π. We get for $N = 1000$ and $l = L = 1$:

$$\pi = 3.17 \pm 0.08 \qquad (3.14)$$

f
own)

and
0)).

), so

ecise,
one
gets a value of pi with 6 digits correct - and that this normally requires 10^{12} - 10^{14} throws. But not necessarily “unrealistic”, when N=213 and l/L = 5/6.

Problem 4.1 (9 points)

There are several ways to test this:

- ChiSquare test:
  Gives p = 57% (Chi2)
  Gives p = 54% (LLH)

Note that since the number of contries is known, one can actually have Ndof=9. Also, one can choose to use Binomial errors.

- Kolmogorov test:
  Gives p = 94%
- Runs test is not enough (but fun!).

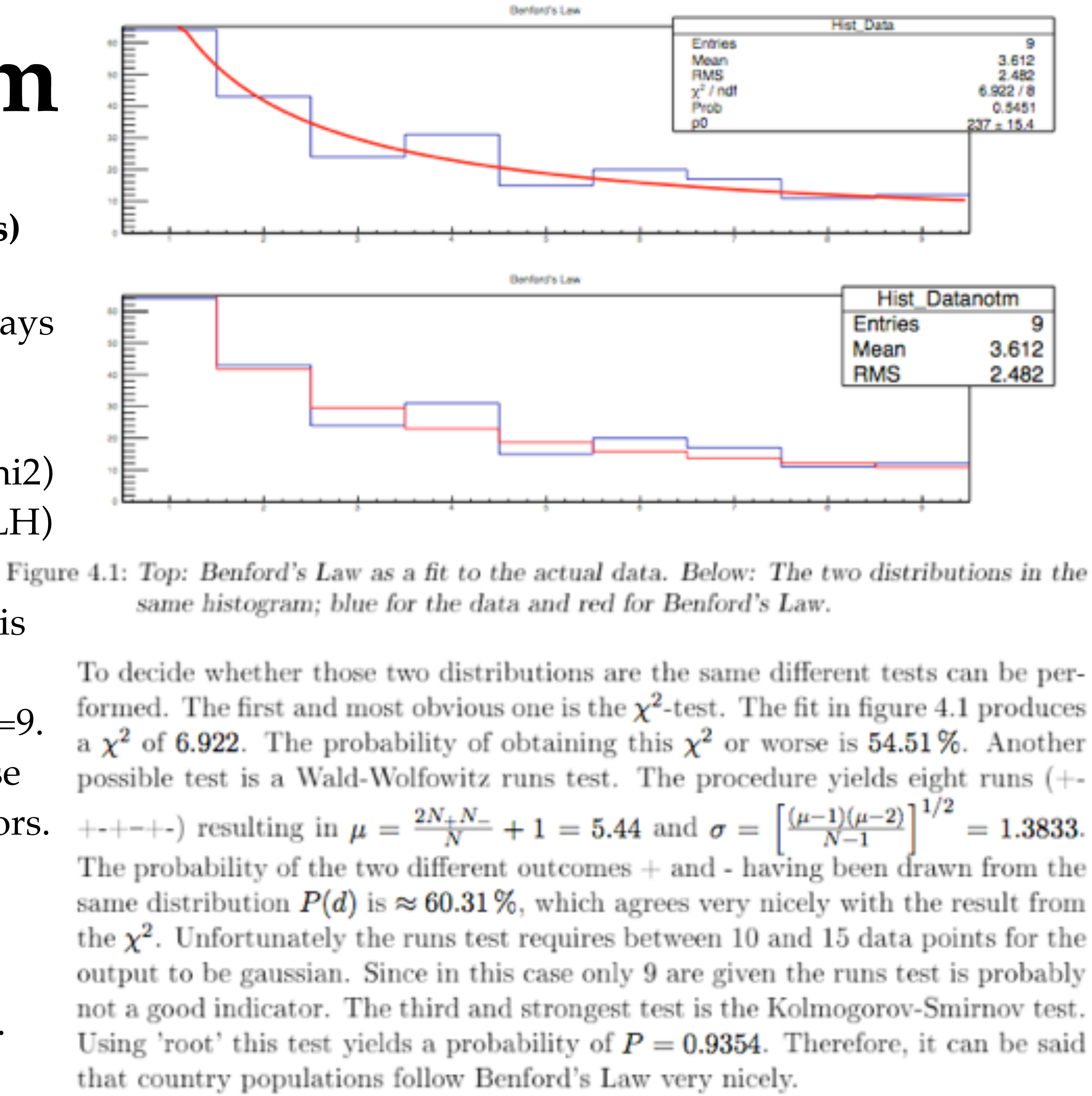
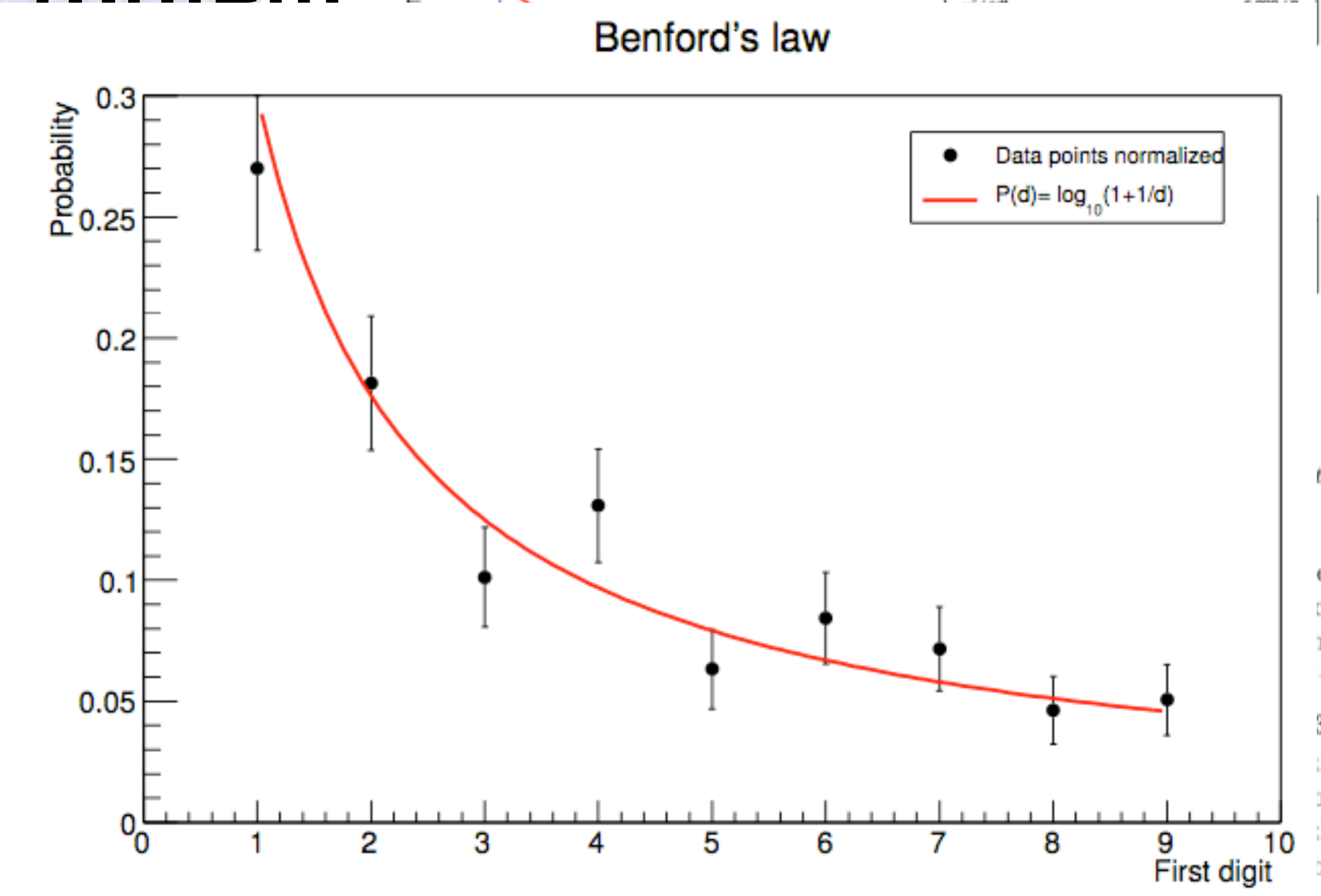

Benford's Law
Hist_Data
Entries 9
Mean 3.612
RMS 2.482
χ² / ndf 6.922 / 8
Prob 0.5451
p0 237 ± 15.4
Benford's Law
Hist_Datanotm
Entries 9
Mean 3.612
RMS 2.482

Figure 4.1: *Top: Benford's Law as a fit to the actual data. Below: The two distributions in the same histogram; blue for the data and red for Benford's Law.*

To decide whether those two distributions are the same different tests can be performed. The first and most obvious one is the χ^2-test. The fit in figure 4.1 produces a χ^2 of 6.922. The probability of obtaining this χ^2 or worse is 54.51 %. Another possible test is a Wald-Wolfowitz runs test. The procedure yields eight runs (+-+-+-+-) resulting in $\mu = \frac{2N_+N_-}{N} + 1 = 5.44$ and $\sigma = \left[\frac{(\mu-1)(\mu-2)}{N-1}\right]^{1/2} = 1.3833$. The probability of the two different outcomes + and - having been drawn from the same distribution $P(d)$ is $\approx$ 60.31 %, which agrees very nicely with the result from the χ^2. Unfortunately the runs test requires between 10 and 15 data points for the output to be gaussian. Since in this case only 9 are given the runs test is probably not a good indicator. The third and strongest test is the Kolmogorov-Smirnov test. Using 'root' this test yields a probability of $P = 0.9354$. Therefore, it can be said that country populations follow Benford's Law very nicely.

Problem

Th
to
-



N
nu
kn
ac
A
to
-
-
enough (but fun!).

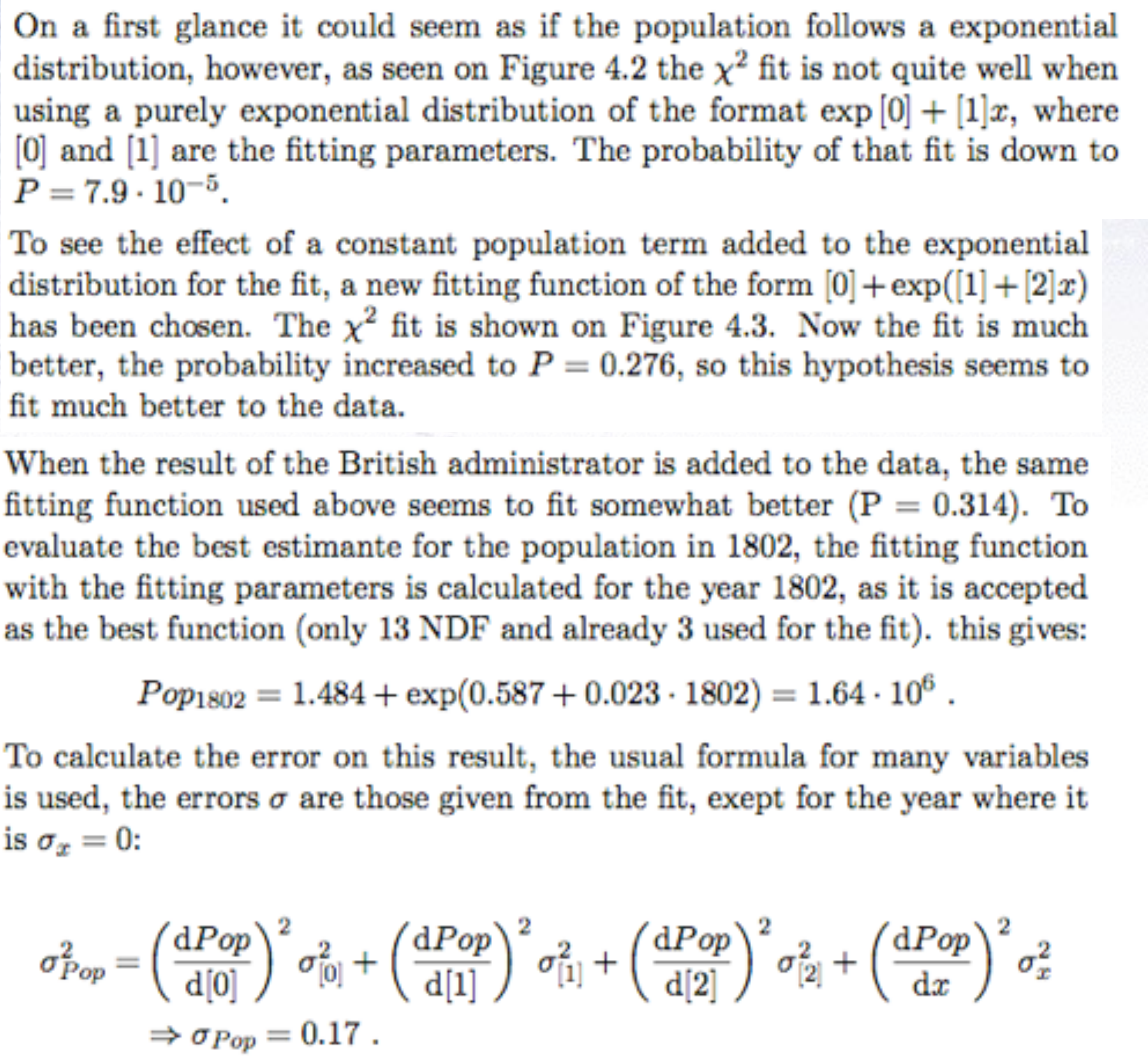
the
er-
ces
her
+-
33.
he
om
he
bly
st.

Using 'root' this test yields a probability of $P = 0.9354$. Therefore, it can be said that country populations follow Benford's Law very nicely.

Problem 4.2 (3+4+7 points)

The exponential fit does not very work, while including a constant makes the fit good (p=28%).

For the last problem, one can either estimate f(-98) ~ 1.65 ± 0.20 and combine this with the estimate, OR include the estimate as an additional point.
The final estimate should have errors, but including correlations between fit parameters is not required (bonus).



On a first glance it could seem as if the population follows a exponential distribution, however, as seen on Figure 4.2 the χ^2 fit is not quite well when using a purely exponential distribution of the format $\exp[0] + [1]x$, where [0] and [1] are the fitting parameters. The probability of that fit is down to $P = 7.9 \cdot 10^{-5}$.

To see the effect of a constant population term added to the exponential distribution for the fit, a new fitting function of the form $[0] + \exp([1] + [2]x)$ has been chosen. The χ^2 fit is shown on Figure 4.3. Now the fit is much better, the probability increased to $P = 0.276$, so this hypothesis seems to fit much better to the data.

When the result of the British administrator is added to the data, the same fitting function used above seems to fit somewhat better (P = 0.314). To evaluate the best estimante for the population in 1802, the fitting function with the fitting parameters is calculated for the year 1802, as it is accepted as the best function (only 13 NDF and already 3 used for the fit). this gives:

$$Pop_{1802} = 1.484 + \exp(0.587 + 0.023 \cdot 1802) = 1.64 \cdot 10^6 \ .$$

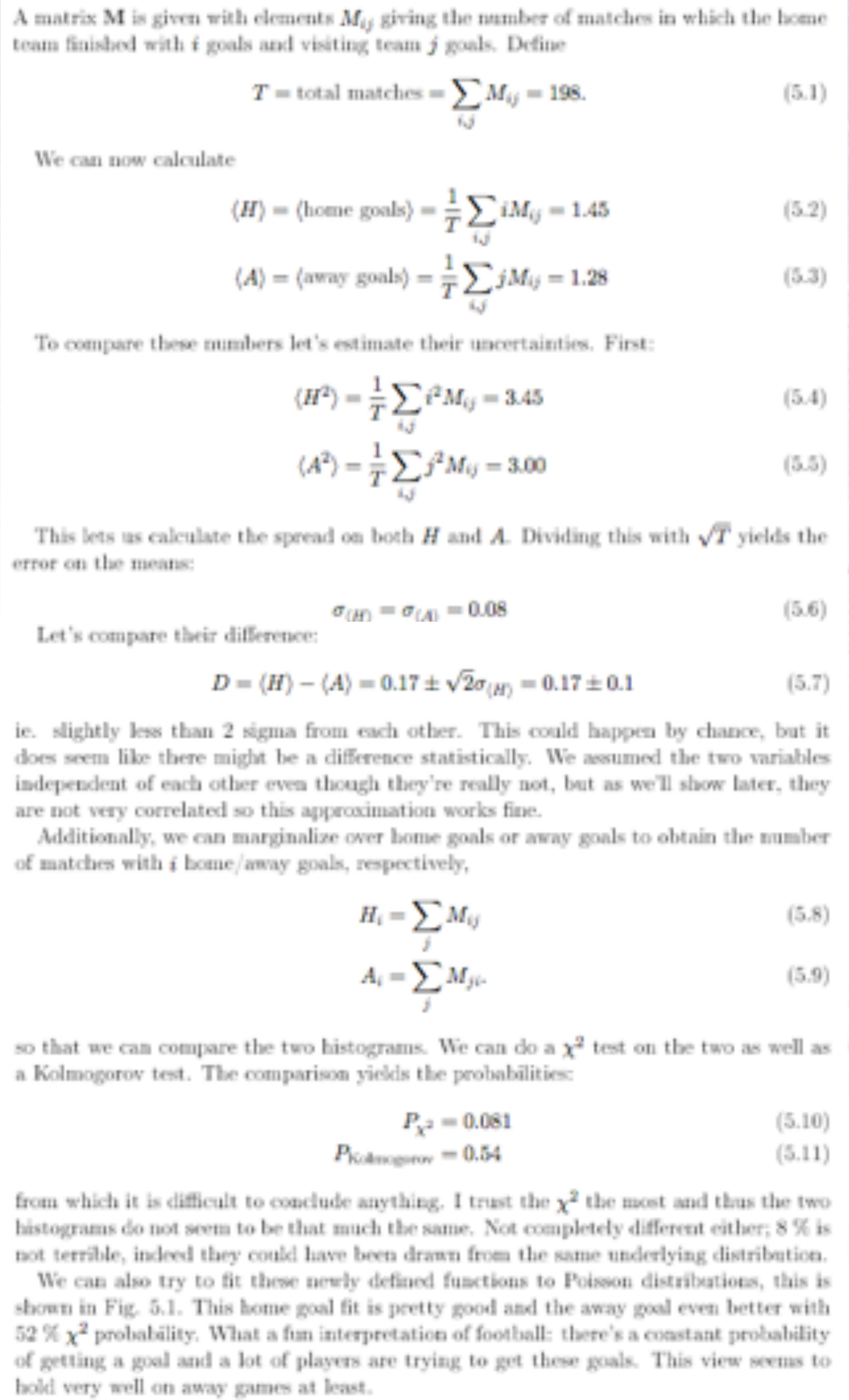
To calculate the error on this result, the usual formula for many variables is used, the errors σ are those given from the fit, exept for the year where it is $\sigma_x = 0$:

$$\sigma^2_{Pop} = \left(\frac{\mathrm{d}Pop}{\mathrm{d}[0]}\right)^2 \sigma^2_{[0]} + \left(\frac{\mathrm{d}Pop}{\mathrm{d}[1]}\right)^2 \sigma^2_{[1]} + \left(\frac{\mathrm{d}Pop}{\mathrm{d}[2]}\right)^2 \sigma^2_{[2]} + \left(\frac{\mathrm{d}Pop}{\mathrm{d}x}\right)^2 \sigma^2_x$$

$$\Rightarrow \sigma_{Pop} = 0.17 \ .$$

Problem 5.1

(5+6+6 points)



A matrix **M** is given with elements M_{ij} giving the number of matches in which the home team finished with i goals and visiting team j goals. Define

$$T = \text{total matches} = \sum_{i,j} M_{ij} = 198. \tag{5.1}$$

We can now calculate

$$\langle H \rangle = \langle \text{home goals} \rangle = \frac{1}{T} \sum_{i,j} i M_{ij} = 1.45 \tag{5.2}$$

$$\langle A \rangle = \langle \text{away goals} \rangle = \frac{1}{T} \sum_{i,j} j M_{ij} = 1.28 \tag{5.3}$$

To compare these numbers let's estimate their uncertainties. First:

$$\langle H^2 \rangle = \frac{1}{T} \sum_{i,j} i^2 M_{ij} = 3.45 \tag{5.4}$$

$$\langle A^2 \rangle = \frac{1}{T} \sum_{i,j} j^2 M_{ij} = 3.00 \tag{5.5}$$

This lets us calculate the spread on both H and A. Dividing this with $\sqrt{T}$ yields the error on the means:

$$\sigma_{\langle H \rangle} = \sigma_{\langle A \rangle} = 0.08 \tag{5.6}$$

Let's compare their difference:

$$D = \langle H \rangle - \langle A \rangle = 0.17 \pm \sqrt{2}\sigma_{\langle H \rangle} = 0.17 \pm 0.1 \tag{5.7}$$

ie. slightly less than 2 sigma from each other. This could happen by chance, but it does seem like there might be a difference statistically. We assumed the two variables independent of each other even though they're really not, but as we'll show later, they are not very correlated so this approximation works fine.

Additionally, we can marginalize over home goals or away goals to obtain the number of matches with i home/away goals, respectively,

$$H_i = \sum_j M_{ij} \tag{5.8}$$

$$A_i = \sum_j M_{ji}. \tag{5.9}$$

so that we can compare the two histograms. We can do a χ^2 test on the two as well as a Kolmogorov test. The comparison yields the probabilities:

$$P_{\chi^2} = 0.081 \tag{5.10}$$

$$P_{\text{Kolmogorov}} = 0.54 \tag{5.11}$$

from which it is difficult to conclude anything. I trust the χ^2 the most and thus the two histograms do not seem to be that much the same. Not completely different either; 8 % is not terrible, indeed they could have been drawn from the same underlying distribution.

We can also try to fit these newly defined functions to Poisson distributions, this is shown in Fig. 5.1. This home goal fit is pretty good and the away goal even better with 52 % χ^2 probability. What a fun interpretation of football: there's a constant probability of getting a goal and a lot of players are trying to get these goals. This view seems to hold very well on away games at least.

The two distributions of goals somewhat have different averages. However, they are OK within about 1.6 σ.
The Kolmogorov test gives p ~ 50%.

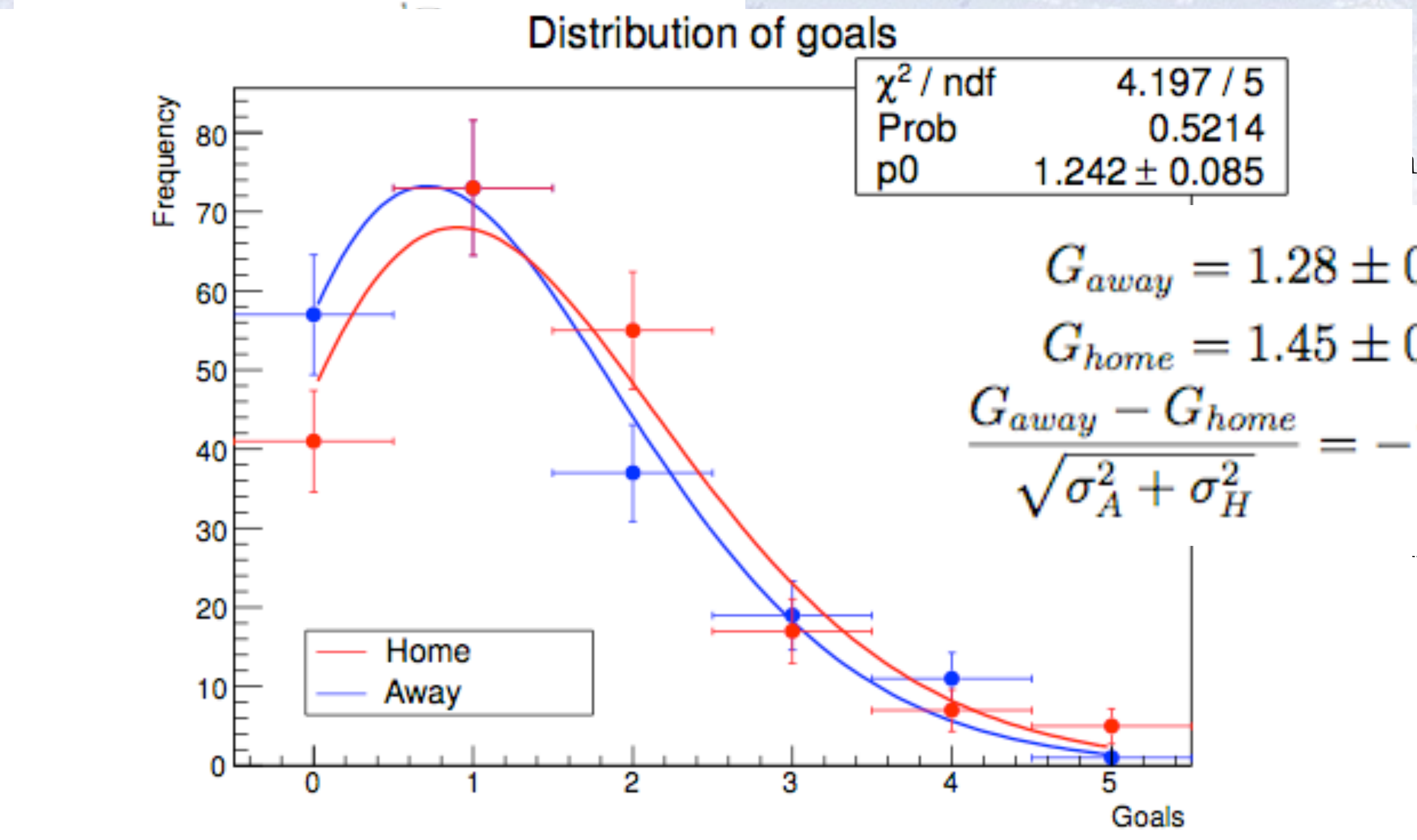
They are really Poisson distributed. Note that this fit should ideally only have one parameter, namely λ, but a normalization is also OK.

Problem 5.1

A matrix **M** is given with elements M_{ij} giving the number of matches in which the home team finished with i goals and visiting team j goals. Define

$$T = \text{total matches} = \sum_{ij} M_{ij} = 198. \qquad (5.1)$$

We can now calculate



ewhat

$$G_{away} = 1.28 \pm 0.08$$

$$G_{home} = 1.45 \pm 0.09$$

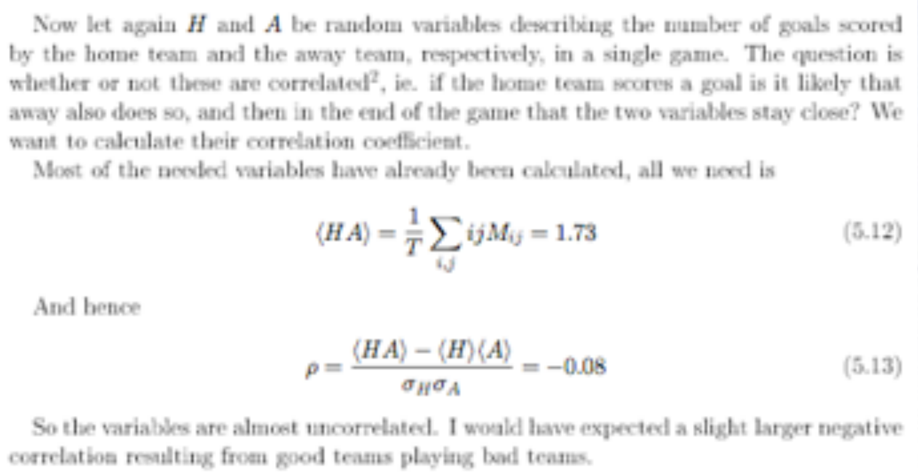
$$\frac{G_{away} - G_{home}}{\sqrt{\sigma_A^2 + \sigma_H^2}} = -1.5,$$

ization

Figure 8: The data for the home and away team with fits. The fit results for the fit to the away team is shown. The values for the home team is $p0 = 1.427 \pm 0.081$, χ^2 /ndf= 5.952/5 and Prob = 0.3114.

Problem 5.1

(5+6+6 points)



Now let again H and A be random variables describing the number of goals scored by the home team and the away team, respectively, in a single game. The question is whether or not these are correlated[2], ie. if the home team scores a goal is it likely that away also does so, and then in the end of the game that the two variables stay close? We want to calculate their correlation coefficient.

Most of the needed variables have already been calculated, all we need is

$$\langle HA \rangle = \frac{1}{T}\sum_{i,j} ijM_{ij} = 1.73 \qquad (5.12)$$

And hence

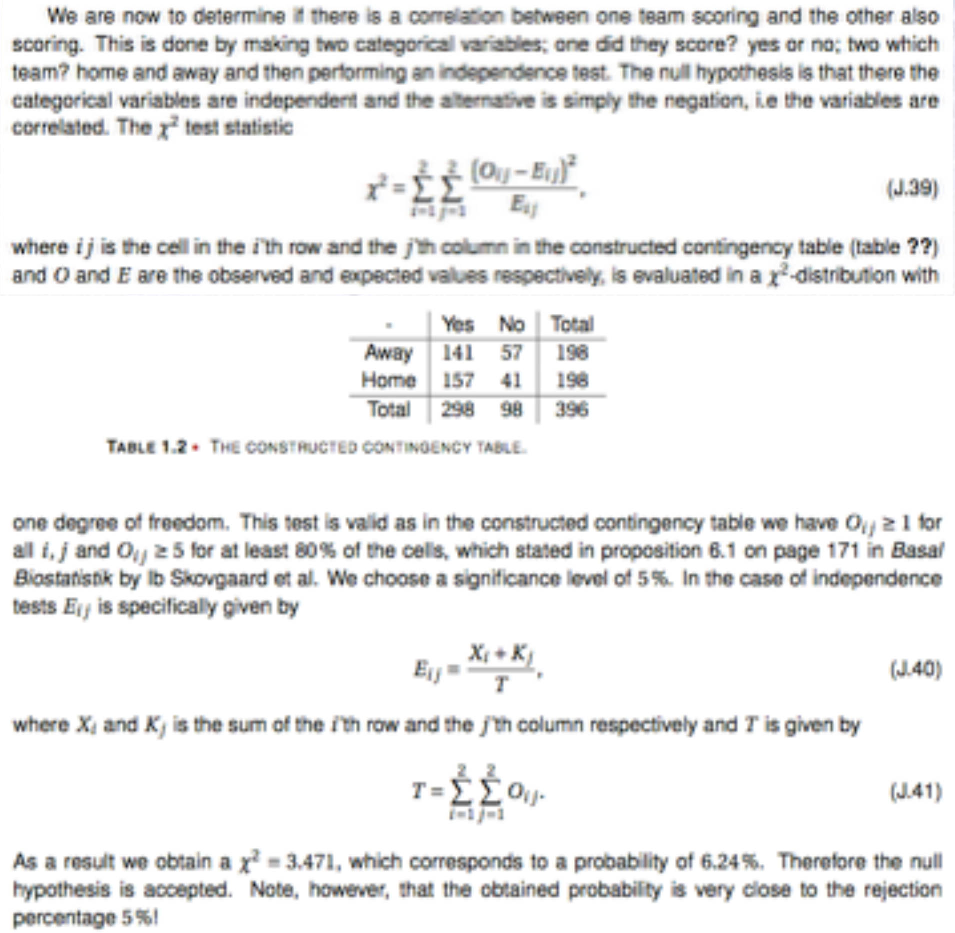
$$\rho = \frac{\langle HA \rangle - \langle H \rangle\langle A \rangle}{\sigma_H \sigma_A} = -0.08 \qquad (5.13)$$

So the variables are almost uncorrelated. I would have expected a slight larger negative correlation resulting from good teams playing bad teams.

As always there are several solutions. The above lacks the uncertainty on ϱ, but is otherwise nice.

The contingency table on the right is fully correct. Anything like it works.

Finally, Fisher's exact method gives the answer (there is no obvious correlation).



We are now to determine if there is a correlation between one team scoring and the other also scoring. This is done by making two categorical variables; one did they score? yes or no; two which team? home and away and then performing an independence test. The null hypothesis is that there the categorical variables are independent and the alternative is simply the negation, i.e the variables are correlated. The χ^2 test statistic

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \qquad (J.39)$$

where ij is the cell in the i'th row and the j'th column in the constructed contingency table (table **??**) and O and E are the observed and expected values respectively, is evaluated in a χ^2-distribution with

·	Yes	No	Total
Away	141	57	198
Home	157	41	198
Total	298	98	396

TABLE 1.2 • THE CONSTRUCTED CONTINGENCY TABLE.

one degree of freedom. This test is valid as in the constructed contingency table we have $O_{ij} \geq 1$ for all i, j and $O_{ij} \geq 5$ for at least 80% of the cells, which stated in proposition 6.1 on page 171 in *Basal Biostatistik* by Ib Skovgaard et al. We choose a significance level of 5%. In the case of independence tests E_{ij} is specifically given by

$$E_{ij} = \frac{X_i + K_j}{T}, \qquad (J.40)$$

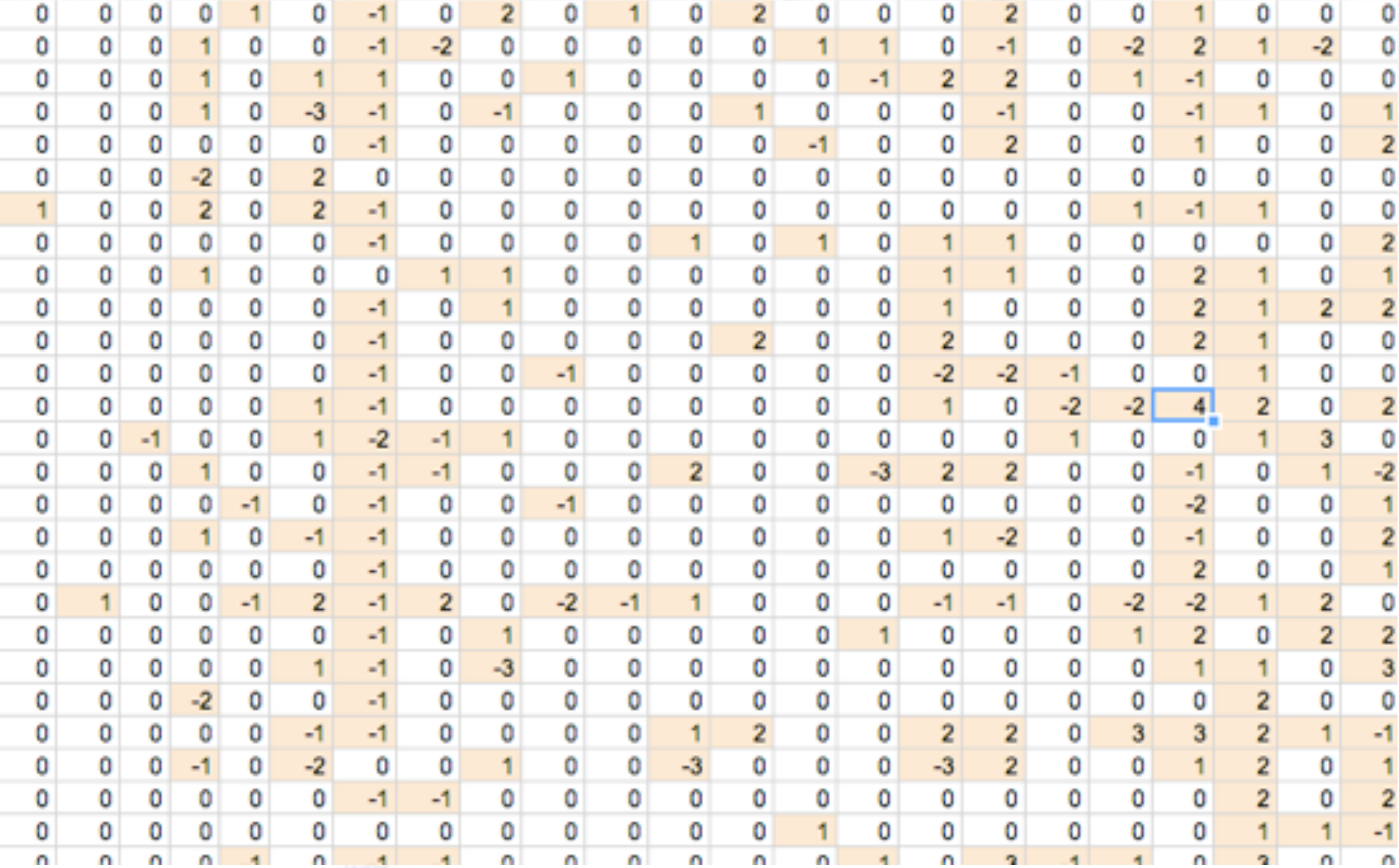
where X_i and K_j is the sum of the i'th row and the j'th column respectively and T is given by

$$T = \sum_{i=1}^{2}\sum_{j=1}^{2} O_{ij}. \qquad (J.41)$$

As a result we obtain a $\chi^2 = 3.471$, which corresponds to a probability of 6.24%. Therefore the null hypothesis is accepted. Note, however, that the obtained probability is very close to the rejection percentage 5%!

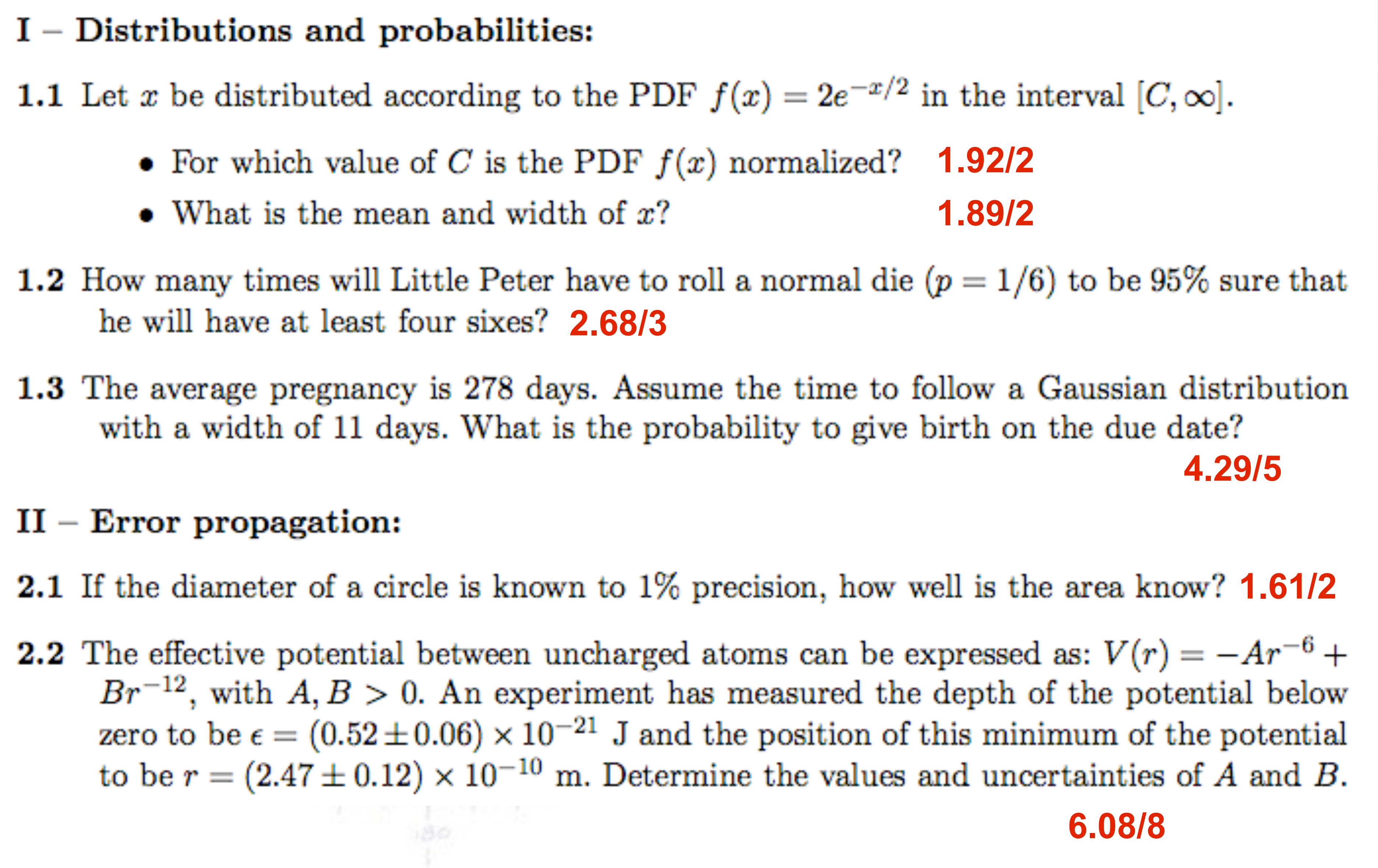
General comments

There seems to be a tendency to loose more points towards the end of the exam. Perhaps the problems are harder, you were more tired, or there were too many problems. Anyhow, well done.



0	0	0	0	1	0	-1	0	2	0	1	0	2	0	0	0	2	0	0	1	0	0	0
0	0	0	1	0	0	-1	-2	0	0	0	0	0	1	1	0	-1	0	-2	2	1	-2	0
0	0	0	1	0	1	1	0	0	1	0	0	0	0	-1	2	2	0	1	-1	0	0	0
0	0	0	1	0	-3	-1	0	-1	0	0	0	1	0	0	0	-1	0	0	-1	1	0	1
0	0	0	0	0	0	-1	0	0	0	0	0	0	-1	0	0	2	0	0	1	0	0	2
0	0	0	-2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	2	0	2	-1	0	0	0	0	0	0	0	0	0	0	0	1	-1	1	0	0
0	0	0	0	0	0	-1	0	0	0	0	1	0	1	0	1	1	0	0	0	0	0	2
0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	1	1	0	0	2	1	0	1
0	0	0	0	0	0	-1	0	1	0	0	0	0	0	0	1	0	0	0	2	1	2	2
0	0	0	0	0	0	-1	0	0	0	0	0	2	0	0	2	0	0	0	2	1	0	0
0	0	0	0	0	0	-1	0	0	-1	0	0	0	0	0	-2	-2	-1	0	0	1	0	0
0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	1	0	-2	-2	4	2	0	2
0	0	-1	0	0	1	-2	-1	1	0	0	0	0	0	0	0	0	1	0	0	1	3	0
0	0	0	1	0	0	-1	-1	0	0	0	2	0	0	-3	2	2	0	0	-1	0	1	-2
0	0	0	0	-1	0	-1	0	0	-1	0	0	0	0	0	0	0	0	0	-2	0	0	1
0	0	0	1	0	-1	-1	0	0	0	0	0	0	0	0	1	-2	0	0	-1	0	0	2
0	0	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1
0	1	0	0	-1	2	-1	2	0	-2	-1	1	0	0	0	-1	-1	0	-2	-2	1	2	0
0	0	0	0	0	0	-1	0	1	0	0	0	0	0	1	0	0	0	1	2	0	2	2
0	0	0	0	0	1	-1	0	-3	0	0	0	0	0	0	0	0	0	0	1	1	0	3
0	0	0	-2	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
0	0	0	0	0	-1	-1	0	0	0	0	1	2	0	0	2	2	0	3	3	2	1	-1
0	0	0	-1	0	-2	0	0	1	0	0	-3	0	0	0	-3	2	0	0	1	2	0	1
0	0	0	0	0	0	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	-1
0	0	0	0	-1	0	-1	-1	0	0	0	0	0	0	1	0	3	-1	1	0	3	0	0

Solving fraction



I – Distributions and probabilities:

1.1 Let x be distributed according to the PDF $f(x) = 2e^{-x/2}$ in the interval $[C, \infty]$.

- For which value of C is the PDF $f(x)$ normalized? 1.92/2
- What is the mean and width of x? 1.89/2

1.2 How many times will Little Peter have to roll a normal die ($p = 1/6$) to be 95% sure that he will have at least four sixes? 2.68/3

1.3 The average pregnancy is 278 days. Assume the time to follow a Gaussian distribution with a width of 11 days. What is the probability to give birth on the due date?

4.29/5

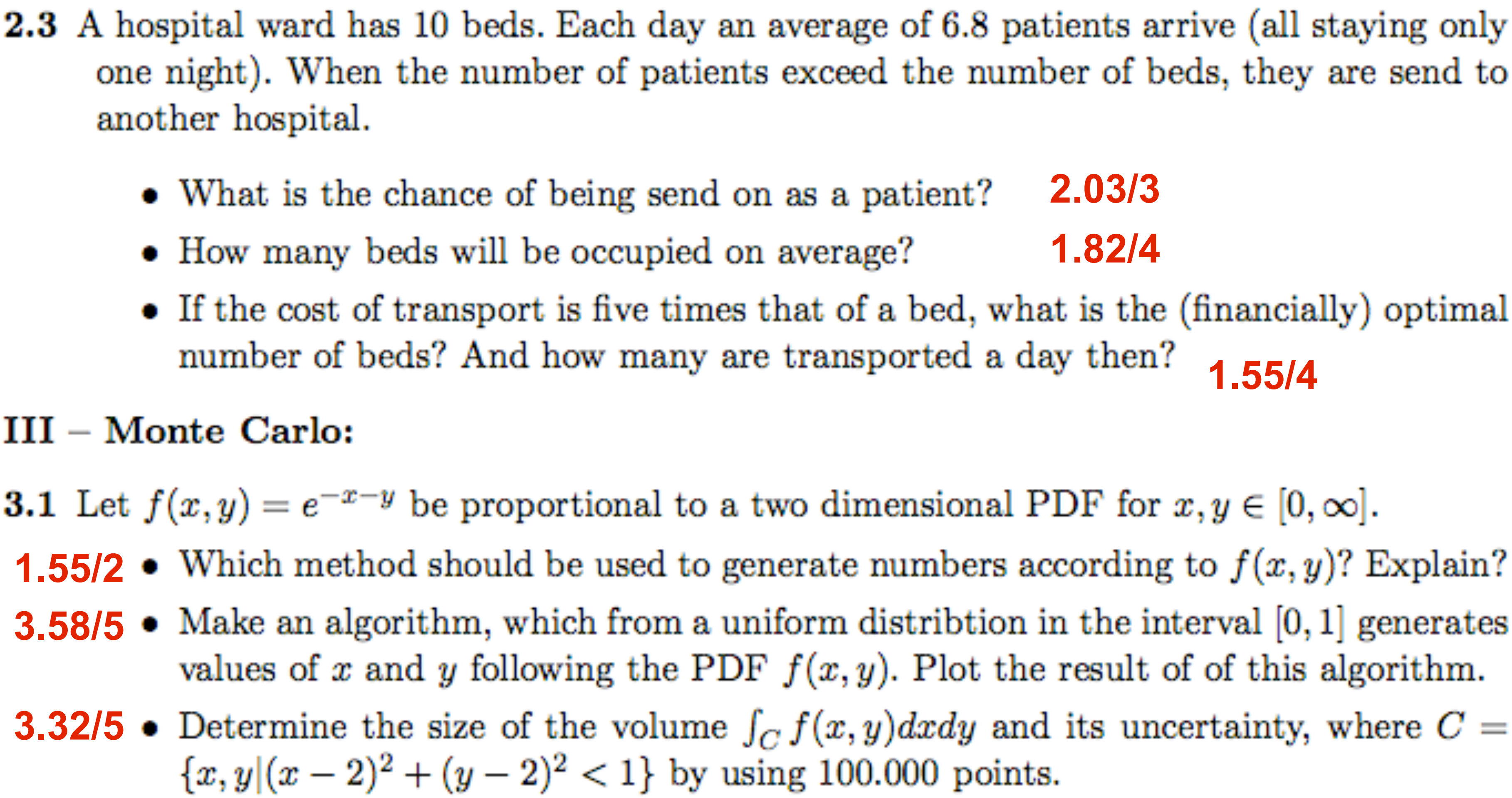
II – Error propagation:

2.1 If the diameter of a circle is known to 1% precision, how well is the area know? 1.61/2

2.2 The effective potential between uncharged atoms can be expressed as: $V(r) = -Ar^{-6} + Br^{-12}$, with $A, B > 0$. An experiment has measured the depth of the potential below zero to be $\epsilon = (0.52 \pm 0.06) \times 10^{-21}$ J and the position of this minimum of the potential to be $r = (2.47 \pm 0.12) \times 10^{-10}$ m. Determine the values and uncertainties of A and B.

6.08/8

Solving fraction



2.3 A hospital ward has 10 beds. Each day an average of 6.8 patients arrive (all staying only one night). When the number of patients exceed the number of beds, they are send to another hospital.

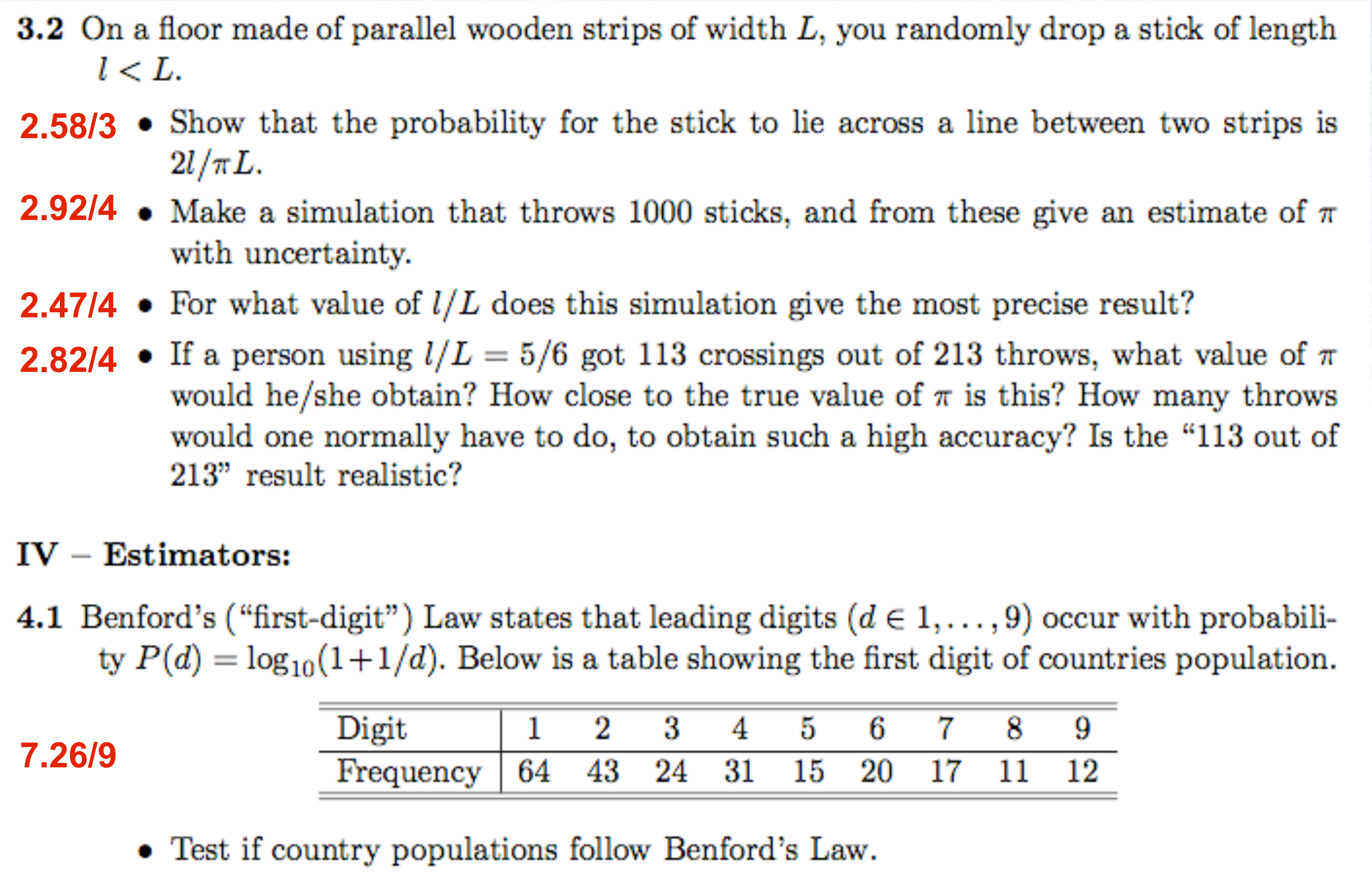
- What is the chance of being send on as a patient? 2.03/3
- How many beds will be occupied on average? 1.82/4
- If the cost of transport is five times that of a bed, what is the (financially) optimal number of beds? And how many are transported a day then? 1.55/4

III – Monte Carlo:

3.1 Let $f(x,y) = e^{-x-y}$ be proportional to a two dimensional PDF for $x, y \in [0, \infty]$.

- 1.55/2 Which method should be used to generate numbers according to $f(x,y)$? Explain?
- 3.58/5 Make an algorithm, which from a uniform distribtion in the interval $[0,1]$ generates values of x and y following the PDF $f(x,y)$. Plot the result of of this algorithm.
- 3.32/5 Determine the size of the volume $\int_C f(x,y)dxdy$ and its uncertainty, where $C = \{x, y|(x-2)^2 + (y-2)^2 < 1\}$ by using 100.000 points.

Solving fraction



3.2 On a floor made of parallel wooden strips of width L, you randomly drop a stick of length $l < L$.

- 2.58/3 Show that the probability for the stick to lie across a line between two strips is $2l/\pi L$.
- 2.92/4 Make a simulation that throws 1000 sticks, and from these give an estimate of π with uncertainty.
- 2.47/4 For what value of l/L does this simulation give the most precise result?
- 2.82/4 If a person using $l/L = 5/6$ got 113 crossings out of 213 throws, what value of π would he/she obtain? How close to the true value of π is this? How many throws would one normally have to do, to obtain such a high accuracy? Is the "113 out of 213" result realistic?

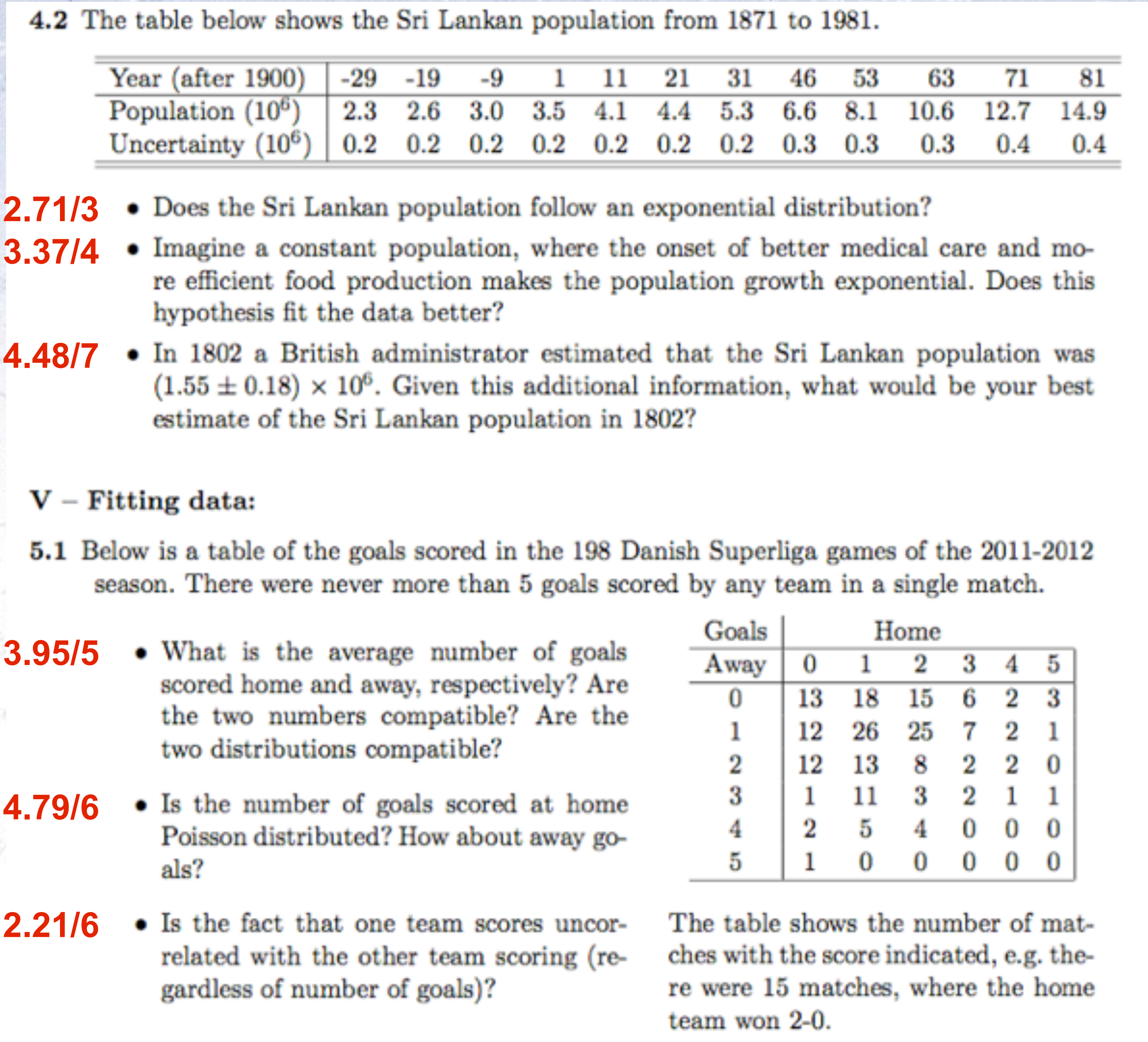
IV – Estimators:

4.1 Benford's ("first-digit") Law states that leading digits ($d \in 1, \ldots, 9$) occur with probability $P(d) = \log_{10}(1+1/d)$. Below is a table showing the first digit of countries population.

7.26/9

Digit	1	2	3	4	5	6	7	8	9
Frequency	64	43	24	31	15	20	17	11	12

- Test if country populations follow Benford's Law.



4.2 The table below shows the Sri Lankan population from 1871 to 1981.

Year (after 1900)	-29	-19	-9	1	11	21	31	46	53	63	71	81
Population (10^6)	2.3	2.6	3.0	3.5	4.1	4.4	5.3	6.6	8.1	10.6	12.7	14.9
Uncertainty (10^6)	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.4	0.4

2.71/3
- Does the Sri Lankan population follow an exponential distribution?

3.37/4
- Imagine a constant population, where the onset of better medical care and more efficient food production makes the population growth exponential. Does this hypothesis fit the data better?

4.48/7
- In 1802 a British administrator estimated that the Sri Lankan population was $(1.55 \pm 0.18) \times 10^6$. Given this additional information, what would be your best estimate of the Sri Lankan population in 1802?

V – Fitting data:

5.1 Below is a table of the goals scored in the 198 Danish Superliga games of the 2011-2012 season. There were never more than 5 goals scored by any team in a single match.

3.95/5
- What is the average number of goals scored home and away, respectively? Are the two numbers compatible? Are the two distributions compatible?

4.79/6
- Is the number of goals scored at home Poisson distributed? How about away goals?

2.21/6
- Is the fact that one team scores uncorrelated with the other team scoring (regardless of number of goals)?

Goals	Home					
Away	0	1	2	3	4	5
0	13	18	15	6	2	3
1	12	26	25	7	2	1
2	12	13	8	2	2	0
3	1	11	3	2	1	1
4	2	5	4	0	0	0
5	1	0	0	0	0	0

The table shows the number of matches with the score indicated, e.g. there were 15 matches, where the home team won 2-0.