Exercises 2 – Advanced statistics – Thursday, 8th January 2015

Please return the results by next week, Thursday 15th Jan 2015 1pm, with name and student number on each page. Exercises should be done individually. The whole sheet is worth 30 points.

1. Decision making with simple hypotheses

   (a) Suppose you observe 125 photons in a counting experiment when your average background estimate is 83.4. What is the (one-sided) p-value for this result? What p-value would you obtain be if you approximated the Poisson distribution with a normal distribution? Is the excess significant? (2pt)

   SOLUTION: The one sided p-value is equal to the exceedance probability

$$p = \mathrm{P}(\mathrm{X} \geq 125|\mathrm{H}_0) = \int_{125}^{\infty} \mathcal{P}(x|\mu = 83.4)dx = 8.4 \times 10^{-6}$$

   The mean and variance of the Poisson distribution are $\lambda = 83.4$, therefore the approximation would be distributed as $x \sim \mathcal{N}(x|\lambda, \sqrt{\lambda})$. If we approximate the distribution by a normal distribution, we find $p = 2.6 \times 10^{-6}$. This erroneously raises the significance from 4.3σ to 4.6σ – the distributions' tails are different!

   Under an Anscombe transform, the approximate Gaussian would be $\mathcal{N}(x|18.3, 1)$ and we would find $p = 1.9 \times 10^{-5}$ (4.1σ), again erroneously changing the significance.

```
from scipy.stats import poisson, norm
from math import sqrt

pp = 1. - poisson(83.4).cdf(125)
pn = 1. - norm(83.4, sqrt(83.4)).cdf(125)

print "P-value Poisson distribution: ", pp, " or ", -norm(0,1).ppf(pp), " sigma."
print "P-value Normal distribution: ", pn, " or ", -norm(0,1).ppf(pn), " sigma."

def Anscombe(x):
    return 2*sqrt(x+3.0/8.0)

pa = 1. - norm(Anscombe(83.4)-1/(4*sqrt(83.4)),1).cdf(Anscombe(125))
print "P-value Anscombed Poisson: ", pa, " or ", -norm(0,1).ppf(pa), " sigma."
```

   (b) Consider as null hypothesis a $\chi^2_{k=1}$ distribution, and as alternative a $\chi^2_{k=3}$ distribution. Use the Neyman-Pearson Lemma to define the optimal rejection region for a test with significance $\alpha = 0.05$. (2pt)

   SOLUTION: We want to find the largest likelihood ratio, so we evaluate

$$\frac{P_A}{P_H} = \frac{\chi^2_k(x|k=3)}{\chi^2_k(x|k=1)} = \cdots = x$$

   Where we have used the exact values of $\Gamma(1/2) = \sqrt{\pi}$ and $\Gamma(3/2) = \sqrt{\pi}/2$. The Neyman-Pearson Lemma then prescribes a test region $\left\{x \text{ such that } \frac{\chi^2_3(x)}{\chi^2_1(x)} \geq c\right\}$, where c is implicitly defined by the equation

$$\int_{-\infty}^{+\infty} \chi^2_1(x)\Theta_H\left(\frac{\chi^2_3(x)}{\chi^2_1(x)} - c\right)dx = \int_c^{\infty} \frac{e^{-x/2}}{\sqrt{2\pi x}}dx = \alpha$$

   which can be solved numerically to obtain $c \approx 3.84$:

```
from scipy.stats import chi2
from scipy.optimize import fsolve
fsolve(lambda x: 1 - chi2(1).cdf(x) - 0.05, x0 = 4.)
```

(c) Consider the hypotheses from the last exercise, but in the case of a large number of samples (independent measurements). The resulting characteristic function is then a product of the individual χ_k^2 characteristic functions. Use the CLT to derive the mean and variance of the log-likelihood ratio T. To this end, first derive the mean and variance of the log-likelihood ratio for *one* observable

$$T = -2 \ln \frac{P_H(x)}{P_A(x)} \,, \tag{1}$$

by calculating $\langle T \rangle$ and $\langle T^2 \rangle - \langle T \rangle^2$. This is best done by using automatic (analytical or numeric) integration. You can then use the CLT to derive the mean and variance of the normal distribution that T follows in case of a large number of observations, say $n = 100$. What is the appropriate threshold value c for $\alpha = 0.05$? (3pt)

Solution: The null hypothesis for x is a χ_1^2 distribution. The mean and variance of $T(x) = -2\ln(1/x)$ under the null are then

$$\langle T(x) \rangle = \int T(x) \chi_1^2(x) dx = -2.54 \tag{2}$$

$$\langle T^2 \rangle - \langle T \rangle^2 = \int (T(x) - \langle T(x) \rangle)^2 \chi_1^2(x) dx = 19.74 \tag{3}$$

$T^{(n)}$ is the sum of the individual T's: it is asymptotically normally distributed (CLT) so the threshold c for rejecting the null is implicitly defined by

$$\int_c^\infty dT \mathcal{N}(T \mid \langle T \rangle, \sqrt{\langle T^2 \rangle - \langle T \rangle^2}) = \alpha \tag{4}$$

which can be computed numerically $n = 1$ as $c_{n=1} \approx 4.767$. Since for multiple samples both the mean and variance of T scale linearly with N, we find (e.g. with the code below) that $c_{100} \approx -181$.

```
Mathematica
n = 100
T[x_] := -2*Log[1/x]
P[x_] := PDF[ChiSquareDistribution[1],x]
m = NIntegrate[T[x]*P[x],{x,0,Infinity}]
v = NIntegrate[(T[x]-m)^2*P[x],{x,0,Infinity}]
Quantile[NormalDistribution[m*n,Sqrt[v*n]],0.95]
```

(d) Generate a sample of 1000 events from a normal distribution with mean zero and variance one. Perform both a Pearson's chi-squared and a K-S test to see whether the generated events are indeed compatible with the initial normal distribution. Are they also compatible with a distribution with variance 1.1? What are the corresponding p-values for both tests? Which test is more powerful? (4pt)

SOLUTION: An example code solution for this is given below. We find slightly different p-values at every run, but we find that $\mathcal{N}(0,1)$ is a better fit than $\mathcal{N}(0,1.1)$ (for both tests), and we find that the KS test has smaller p-values than the Pearson χ^2. Ks is more likely to reject the null, so it is a stronger test.

```
from numpy import random
from scipy.stats import histogram, norm, kstest
from scipy.stats import chisquare as chi2test

n=1000
samples = random.normal(0,1,n)

h = histogram(samples,numbins=8,defaultlimits=(-4,4))
# This binning is arbitrary but makes calculating the expected frequencies much easier later.
if h[3] != 0:
    print "Warning: with n=1000 we expect a 4 sigma outlier roughly every 15 runs."

freqs1 = [n*(norm(0,1).cdf(i+1)-norm(0,1).cdf(i)) for i in range(-4,4)]
cs1 = chi2test(h[0]/freqs1)
ks1 = kstest(samples,'norm',args =(0,1))

freqs2 = [n*(norm(0,1.1).cdf(i+1)-norm(0,1.1).cdf(i)) for i in range(-4,4)]
cs2 = chi2test(h[0]/freqs2)
ks2 = kstest(samples,'norm',args =(0,1.1))
```

2. Neyman Pearson Lemma

(a) Show that the Neyman Pearson criterium for selecting the rejection region leads to the most powerful statistical test for a given significance level. To this end, consider an interval that satisfies the criterium, and consider the effects of infinitesimal changes of the region boundaries that either satisfy or violate the Neyman Person criterium. (3pt)

SOLUTION: Consider a certain statistical significance such that $\int \mathrm{P_H}(x)\mathrm{d}x = \alpha$. The Neyman-Pearson lemma says that the statistical power is maximised by $\int \mathrm{P_H}(x)\Theta_\mathrm{H}\left(c - \frac{\mathrm{P_A}}{P_\mathrm{H}}\right)\mathrm{d}x = \alpha$, i.e. for all x in this region $\mathrm{P_A} \geq c\mathrm{P_H}$. Note that our choice of α fixes c and thus our integration region. Let us say that $\mathrm{P_A} \geq c\mathrm{P_H}$ holds on $x \in [x_1, x_2]$. Next, let us slightly shift the boundaries, but keeping the statistical significance the same:

$$\int_{x_1+\delta x_1}^{x_2+\delta x_2} \mathrm{P_H}(x)\mathrm{d}x = \alpha$$

We want to see what happens to the statistical power $\int_{x_1}^{x_2} \mathrm{P_A}(x)dx = 1-\beta$ in this new region. We obtain:

$$\int_{x_1+\delta x_1}^{x_2+\delta x_2} \mathrm{P_A}(x)dx = 1-\beta + \underbrace{\int_{x_2}^{x_2+\delta x_2} \mathrm{P_A}(x)dx - \int_{x_1}^{x_1+\delta x_1} \mathrm{P_A}(x)dx}_{=\delta\beta}$$

The NP lemma says that $\delta\beta$ should be smaller than zero. And this is so, since:

$$\mathrm{P_A} \geq c\mathrm{P_H} \forall x \epsilon \, [x_1, x_1 + \delta x_1] \Rightarrow \int_{x_1}^{x_1+\delta x_1} \mathrm{P_A} \mathrm{d}x \geq c \int_{x_1}^{x_1+\delta x_1} \mathrm{P_H} \mathrm{d}x$$
$$\mathrm{P_A} < c\mathrm{P_H} \forall x \epsilon \, [x_2, x_2 + \delta x_1] \Rightarrow \int_{x_2}^{x_2+\delta x_2} \mathrm{P_A} \mathrm{d}x < c \int_{x_2}^{x_2+\delta x_2} \mathrm{P_H} \mathrm{d}x$$

Since we have that $\int_{x_2}^{x_2+\delta x_2} \mathrm{P_H}\mathrm{d}x = \int_{x_1}^{x_1+\delta x_1} \mathrm{P_H}\mathrm{d}x$ (we kept α fixed), $\delta\beta < 0$, so the statistical power decreases. □

3. Estimators

   (a) Show that in case of the gamma distribution with $\alpha = 2$ and $\beta = \zeta^{-2}$,
$$P(x|\zeta) = \frac{x}{\zeta^4} e^{-x/\zeta^2} \theta_H(x) \,, \tag{5}$$
   the MLE is a biased estimator for ζ. Here, $\theta_H(x)$ denotes the Heaviside step function. What is the bias and the variance of this estimator? Define, based on the MLE, a new unbiased estimator. Does the variance of this estimator fulfill the Cramer Rao bound? (3pt)

   SOLUTION: The mean value of x can be calculated to be $\langle x \rangle = 2\zeta^2$ (either by doing the integral by hand, or looking up the mean value of a gamma function). For a given measured value x, the MLE of ζ is given by $\hat{\zeta} = \sqrt{x/2}$ (obtained by differentiating $P(x|\zeta)$ w.r.t. ζ and setting it to zero). The expectation value of $\hat{\zeta}$ is then given by
$$\langle \hat{\zeta} \rangle = \int dx P(x|\zeta)\sqrt{x/2} = \tfrac{3}{4}\sqrt{\frac{\pi}{2}}\zeta \simeq 0.94\zeta \,. \tag{6}$$
   This shows that the MLE is in the present example a *biased* estimator for ζ. An *unbiased* estimator is instead given by $\xi(\zeta) = \zeta^2$ (we perform here a parameter substitution). Using this definition, one can immediately see that $\hat{\xi} = \hat{\zeta}^2 = x/2$ (since the position of the mode does not depend on the parametrization). From this, it follows that $\langle \hat{\xi} \rangle = \langle x/2 \rangle = \zeta^2 = \xi$.

   (b) Show that in case of multiple draws from the flat distribution given by $P(x|\mu) = \theta_H(\frac{1}{2} - |x - \mu|)$, the mean of the measured values is not the best estimator for μ. Do this by showing that the variance of the estimator given by $\hat{\mu} = \frac{1}{2}(\max_i(x_i) + \min_i(x_i))$ is smaller. Use analytical calculations (3pt).

   Hints: First, think about the distribution function of $z^+ \equiv \max_i x_i$ and $z^- \equiv \min_i x_i$ individually (tip: they are a product of CDFs and PDFs). Then calculate the variance of $\hat{\mu}$, assuming that the z^+ and z^- are uncorrelated, which is true for a large number of draws. To simplify your life, you can assume that $\mu = 0$ or $\mu = 1/2$ during the main calculations, and generalize the result at the very end to arbitrary μ.

Solution: Consider the estimator $\hat{\mu} = \frac{1}{2}(z^+ + z^-)$.
The function $\max_i(x_i)$ of an arbitrary number of random variables can be defined recursively in terms of the binary function $\max(x_1, x_2)$ as

$$\max_i(x_i) = \max(x_1, \max(x_2, \max(\cdots, x_n))\cdots))).$$

This suggests that the distribution of $z^\pm$ might be constructed by induction. The fact that these random variables x_i share the same distribution greatly simplifies this construction, since we can guess the answer: We want to show that $z^+_{(n)} \propto \mathrm{PDF}(x|\mu)\mathrm{CDF}^{n-1}(x|\mu)$.

- $n = 2$
  The random variable $z^+_{(2)} \sim \max(x_1, x_2)$ has a distribution

  $$\alpha(\mathrm{CDF}_1 * \mathrm{PDF}_2) + (1-\alpha)(\mathrm{PDF}_1 * \mathrm{CDF}_2)$$

  where α is the probability that $x_1 > x_2$. Since these variables are i.i.d, the above expression reduces to $z^+_{(2)} \propto \mathrm{PDF}(x|\mu)\mathrm{CDF}(x|\mu)$, which has the desired form.
- $n \to n+1$
  Constructing the random variable $z^+_{(n+1)}$ from $\max(x_{n+1}, z^+_{(n)})$ just adds another CDF, because the CDF of $z^+_{(n)}$ is $\mathrm{CDF}^n(x|\mu)$ and the expression we used for $n = 2$ still holds and the α's still cancel because our multiple draws are i.i.d.
- By induction, $z^+_{(n)} \propto \mathrm{PDF}(x|\mu)\mathrm{CDF}^{n-1}(x|\mu)$.

We can find a similar expression for z^-. Therefore

$$\hat{\mu} \propto \frac{\mathrm{PDF}(x|\mu)}{2}\left(\mathrm{CDF}^{n-1}(x|\mu) + (1-\mathrm{CDF})^{n-1}(x|\mu)\right)$$

For convenience, we set $\mu = 1/2$ to calculate the variance of this estimator. Then $\mathrm{PDF}(x|\mu)$ is uniform from zero to one (with height one) and the corresponding CDF is just $\mathrm{CDF}(x|\mu) = \int_0^x 1dx = x$ such that these estimators are distributed as

$$P(z^+_{(n)}) \quad = \quad nx^{(n-1)} \tag{7}$$
$$P(z^-_{(n)}) \quad = \quad n(1-x)^{(n-1)} \tag{8}$$

We see that as n increases, the maxima are more likely to be large and the minima are more likely to be low. The factor of n is required for these to be normalised.
The variance of the estimator $\hat{\mu}$ is constructed from the variance of $z^\pm$, so we calculate

$$\langle z^+_{(n)}\rangle \quad = \quad \int_0^1 zP(z^+_{(n)})dz = \cdots = \frac{n}{n+1} \tag{9}$$
$$\langle z^-_{(n)}\rangle \quad = \quad \int_0^1 zP(z^-_{(n)})dz = \cdots = \frac{1}{n+1} \tag{10}$$
$$\mathbb{V}(z^+_{(n)}) \quad = \quad \int_0^1 (z - \langle z^+_{(n)}\rangle)^2 P(z^+_{(n)})dz = \cdots \tag{11}$$
$$\mathbb{V}(z^-_{(n)}) \quad = \quad \int_0^1 (z - \langle z^-_{(n)}\rangle)^2 P(z^-_{(n)})dz = \cdots \tag{12}$$

The variance of $\hat{\mu}$ for $\mu = 1/2$ and uncorrelated $z^\pm$ is then simply

$$\mathbb{V}(\hat{\mu}) = \mathbb{V}(\frac{1}{2}(z^+ + z^-)) = \frac{1}{4}\left(\mathbb{V}(z^+_{(n)}) + \mathbb{V}(z^-_{(n)})\right) = \frac{n}{2(n+2)(n+1)^2}$$

By constrast, the estimator for the sample mean $\tilde{\mu} = \frac{1}{n}\sum x_i$ has a variance

$$\mathbb{V}(\tilde{\mu}) = \mathbb{V}(\frac{1}{n}\sum x_i) = \frac{1}{n^2}\sum \mathbb{V}(x_i) = \frac{\mathbb{V}(x_i)}{n}$$

where the variance $\mathbb{V}(x_i)$ is $1/12$ (after solving two more integrals with weight $P(x|\mu) = \Theta(1/2 - |x-\mu|)$). Hence, we need to solve the following inequality over $n \in \mathbb{N}^*$:

$$\frac{n}{2(n+2)(n+1)^2} \leq \frac{1}{12n}$$

which happens to be true for all $n \in \mathbb{N}^*$. Hence, the estimator $\hat{\mu}$ is a more efficient estimator than $\tilde{\mu}$. In fact, this result is true for $\mu \neq 1/2$, since we find that the variances of $\hat{\mu}$ and $\tilde{\mu}$ are both independent of μ.

(c) Confirm the previous result with a simple Monte Carlo (2pt).

```
from __future__ import division
from numpy import *
import pylab as plt

n = 10  # number of mock data events
N = 100  # number MC loops
est1 = 0
est2 = 0
for i in range(N):
    d = random.random(n)-0.5
    est1 += mean(d) ** 2
    est2 += (0.5*(max(d)+min(d))) ** 2

print "Mean estimator:", est1/N
print "Min/max estimator:", est2/N
```

4. Cramér-Rao bound and fisher information

(a) Demonstrate analytically (using integration by parts and the fact that the score has a zero mean) that the Fisher information can be written as

$$\mathcal{I}(\theta) = \left\langle \left(\frac{\partial}{\partial\theta} \ln \mathcal{L}(\theta|c) \right)^2 \right\rangle = -\left\langle \frac{\partial^2}{\partial\theta^2} \ln \mathcal{L}(\theta|c) \right\rangle$$

(2pt)
Hints: As first step, show that taking the derivative of $\int \mathcal{L}(\theta|c)dc = 1$ with respect to θ (on both sides of the equation) yields (remember that $d \ln f(x)/dx = f^{-1} df(x)/dx$)

$$\int \frac{d \ln \mathcal{L}(\theta|c)}{d\theta} \mathcal{L}(\theta|c) dc = \left\langle \frac{d \ln \mathcal{L}(\theta|c)}{d\theta} \right\rangle = 0 \qquad (13)$$

Taking a second time the derivative with respect to θ on the left side of the equation gives two terms that can be rewritten in the form that is requested in this exercise.
Solution: The total probability should be unity, i.e. the integral of the likelihood function is,

$$\int \mathcal{L}(\theta|c)\, \mathrm{d}c = 1$$

Taking the derivative w.r.t. θ on both sides (commutes with the integral) and noting that $x \mathrm{d} \ln x = \mathrm{d}x$:

$$\begin{aligned} 0 &= \int \frac{\mathrm{d}\mathcal{L}(\theta|c)}{\mathrm{d}\theta} \mathrm{d}c \\ &= \int \frac{\mathrm{d} \ln \mathcal{L}(\theta|c)}{\mathrm{d}\theta} \mathcal{L}(\theta|c) \, \mathrm{d}c \\ &= \left\langle \frac{\mathrm{d} \ln \mathcal{L}(\theta|c)}{\mathrm{d}\theta} \right\rangle \end{aligned}$$

Next, taking another derivative and using the product rule we find

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}\theta} \int \frac{\mathrm{d} \ln \mathcal{L}(\theta|c)}{\mathrm{d}\theta} \mathcal{L}(\theta|c) \, \mathrm{d}c \\ &= \int \frac{\mathrm{d}^2 \ln \mathcal{L}(\theta|c)}{\mathrm{d}\theta^2} \mathcal{L}(\theta|c) \, \mathrm{d}c + \int \left(\frac{\mathrm{d} \ln \mathcal{L}(\theta|c)}{\mathrm{d}\theta} \right)^2 \mathcal{L}(\theta|c) \, \mathrm{d}c \end{aligned}$$

Recognising that we again have two expectation values here, we find:

$$\boxed{\left\langle \left(\frac{\mathrm{d} \ln \mathcal{L}(\theta|c)}{\mathrm{d}\theta} \right)^2 \right\rangle = - \left\langle \frac{\mathrm{d}^2 \ln \mathcal{L}(\theta|c)}{\mathrm{d}\theta^2} \right\rangle} \qquad \square$$

(b) Demonstrate in an explicit analytical calculation that any consistent MLE saturates the CRB in the limit of a large number of events. (3pt)

Hints: Start by showing that the Taylor expansion of the MLE condition $\frac{\partial}{\partial a} \ln \mathcal{L}(\hat{a}|x) = 0$ around the true value of a, to first order in $\hat{a}$, yields

$$\frac{\partial}{\partial a} \ln \mathcal{L}(a|x) + (\hat{a} - a) \frac{\partial^2}{\partial a^2} \ln \mathcal{L}(a|x) = 0 \, .$$

In the large-number limit the likelihood function can be approximated by a Gaussian. Notice that then, for different random variables x, the first term (the gradient of the likelihood function at the true value a) as well as the first factor of the second term fluctuate. The second factor of the second term is however approximately independent of x. Rearanging the terms and factors and averaging over c then allows to find an expression for $\langle (\hat{a} - a)^2 \rangle$, which can be conntected to $\mathcal{I}(\theta)$ using the result from the previous exercise.

Solution: By construction we have that for a MLE

$$0 = \frac{\partial}{\partial a} \ln \mathcal{L}(\hat{a}|x) \left(\equiv \frac{\partial}{\partial a} \ln \mathcal{L}(a|x) \Big|_{a=\hat{a}} \right)$$

Taylor expanding the derivative evaluated at $\hat{a}$ about the true value of a, which for clarity I'll call a_0, yields:

$$\begin{aligned}
\frac{\partial}{\partial a} \ln \mathcal{L}(a|x)\bigg|_{a=\hat{a}} &\approx \frac{\partial}{\partial a} \ln \mathcal{L}(a_0|x) + (a - a_0) \frac{\partial^2}{\partial a^2} \ln \mathcal{L}(a_0|x)\bigg|_{a=\hat{a}} \\
&= \frac{\partial}{\partial a} \ln \mathcal{L}(a_0|x) + (\hat{a} - a_0) \frac{\partial^2}{\partial a^2} \ln \mathcal{L}(a_0|x) \\
&= 0
\end{aligned}$$

Now, in the large number limit we can approximate the likelihood function by a Gaussian, $\mathcal{L}(a|x) \propto e^{\frac{-(x-a)^2}{2\sigma^2}}$, i.e. $\ln \mathcal{L}(a|x) \sim -(x-a)^2/(2\sigma^2)$. Therefore, $\frac{\partial}{\partial a} \ln \mathcal{L}(a|x) = 2(x-a)/(2\sigma^2)$ and $\frac{\partial^2}{\partial a^2} \ln \mathcal{L}(a|x) = -1/(2\sigma^2)$. Since $\hat{a}$ is a function of x, but a_0 is not, only $a \frac{\partial^2}{\partial a^2} \ln \mathcal{L}(a_0|x)$ does not fluctuate when changing x.
Rearranging the above Taylor expansion we get:

$$\hat{a} - a_0 = -\frac{\frac{\partial}{\partial a} \ln \mathcal{L}(a_0|x)}{\frac{\partial^2}{\partial a^2} \ln \mathcal{L}(a_0|x)}$$

Then squaring both sides

$$(\hat{a} - a_0)^2 = \frac{\left(\frac{\partial}{\partial a} \ln \mathcal{L}(a_0|x)\right)^2}{\left(\frac{\partial^2}{\partial a^2} \ln \mathcal{L}(a_0|x)\right)^2}$$

Taking the expectation value of the LHS yields the mean squared error, which for an unbiased estimator is the same as the variance of $\hat{a}$. Now, a little caution has to be taking in taking the variance on the RHS, since the denominator is independent of x we can write:

$$\begin{aligned}
\left\langle \frac{\left(\frac{\partial}{\partial a} \ln \mathcal{L}(a_0|x)\right)^2}{\left(\frac{\partial^2}{\partial a^2} \ln \mathcal{L}(a_0|x)\right)^2} \right\rangle &= \frac{\left\langle \left(\frac{\partial}{\partial a} \ln \mathcal{L}(a_0|x)\right)^2 \right\rangle}{\left(\frac{\partial^2}{\partial a^2} \ln \mathcal{L}(a_0|x)\right)^2} \\
&= \frac{\left\langle \left(\frac{\partial}{\partial a} \ln \mathcal{L}(a_0|x)\right)^2 \right\rangle}{\left\langle \frac{\partial^2}{\partial a^2} \ln \mathcal{L}(a_0|x) \right\rangle^2} \\
&= \frac{1}{\mathcal{I}(a)}
\end{aligned}$$

Again note, that our step in the first line is only allowed because the denominator is independent of x. In the last line we used the result from the previous exercise. Thus we obtain, for an unbiased MLE:

$$\text{var}(\hat{a}) = \frac{1}{\mathcal{I}(a)}$$

5. Confidence belts

(a) Consider a measured flux that is log-normal distributed with a free μ but known $\sigma = 0.2$ (see lecture for definitions). If one measures $x = 10$, what are the 90%CL upper and the 90%CL lower limits on μ? What is the central 90%CL interval (with identical probabilities above and below the confidence belt). (3pt)

SOLUTION: The log-normal distribution is given by

$$P(x|\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} \, . \qquad (14)$$

To obtain the 90% upper limit CL on μ, we have to find the μ for which the probability to measure $x = 10$ or smaller values is 10%; $\int_0^{10} dx P(x|\mu,\sigma) = 0.1$. For the 90% CL lower limit the condition is $\int_{10}^{\infty} dx P(x|\mu,\sigma) = 0.1$. In the case of a central interval, the boundaries are derived by replacing 0.1 by 0.05. We obtain: $\mu \leq 2.56$ (90% CL), $\mu \geq 2.05$ (90% CL) and $\mu \in [1.97, 2.63]$ (90% CL) for the upper, lower and central interval, respectively.