Exercise/lecture note – Basic Maximum Likelihood Estimation

A)

We want to estimate the prevalence of a parasite in a population. Let us call this unknown parameter for p.

- What is the probability that a random individual in this population is *not* parasitized (expressed as a function of p)?
- What is the probability that three random individuals, sampled independently, are all parasitized?
- If you examine two random individuals, what is the probability that the first one is parasitized and the second one is not?
- If you examine two random individuals, what is the probability that one is parasitized and one is not?
- If you examine 10 random individuals, sampled independently from the population, what is the probability that 2 of them are parasitized? (you don't have to give an exact expression)

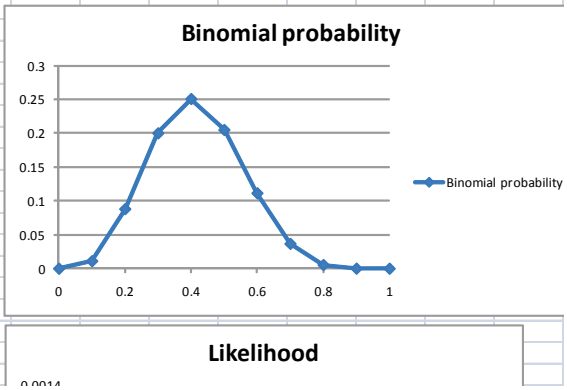
B)

In this exercise we are going to start by using Excel (you can also use R if you prefer that). If you are not familiar with Excel (or if you have problems doing the below in Excel) you can download a ready-made spread-sheet from http://www.finse.uio.no/events/international-workshops/introduction-to-estimation/material/Exercise_TE1.xls. There is also an R script available at the end of this document that can be pasted line-by-line into R, but I recommend trying to make the spread-sheet, or the R-code, yourself.

In a sample of n = 10 individuals k = 4 individuals are parasitized. The probability of getting these data in one particular order (for example, first getting 4 parasitized individuals, then 6 non-parasitized individuals) is $p^4(1-p)^6$, and the probability of getting these data in any order is $\frac{10!}{4!6!}p^4(1-p)^6$ (this is the binomial probability). The first fraction, $\frac{10!}{4!6!}$, is just the number of different sequences you can get 4 parasitized individuals in a sample of 10 and is often shortened to $\binom{10}{4}$ and $4! = 1 \cdot 2 \cdot 3 \cdot 4$.

- Calculate the probability of getting these data given that $p = 0, 0.1, \ldots, 1$ (in Excel you can use the 'BINOMDIST' function, for example "=BINOMDIST(B2,A2,D2,0)" – in R you can use the function 'dbinom(k,n,p)'). Then make a plot of this probability as a function of p. This should look something like the below screen shot.

n	k		p	Binomial probability	Likelihood	Log likelihood
10	4		0	0	0	
			0.1	0.011160261	5.31441E-05	-9.8425
			0.2	0.088080384	0.00041943	-7.77661
			0.3	0.200120949	0.000952957	-6.95594
			0.4	0.250822656	0.001194394	-6.73012
			0.5	0.205078125	0.000976563	-6.93147
			0.6	0.111476736	0.000530842	-7.54105
			0.7	0.036756909	0.000175033	-8.65054
			0.8	0.005505024	2.62144E-05	-10.5492
			0.9	0.000137781	6.561E-07	-14.237
			1	0	0	



- Try different values of sample size (n) and number of parasitized individuals (k) and observe how the graph changes. How can you use this graph to find a "best guess" of the parasite prevalence in the whole population based on these data?
- Try increasing both sample size (n) and number of parasitized individuals (k) while keeping the proportions the same (e.g., k/n = 20/8, 30/12, 40/16,…). How does the shape of the curve change when you increase sample size? How do you interpret this?
- Calculate also the probability of the data given the different values of p when the parasitized individuals appear in one particular order (e.g., first you get 4 parasitized, then 6 non-parasitized) and plot this. In what way are the two curves different? Does this make sense? Do we gain any information about the parameter p by knowing the order in which we sampled parasitized and non-parasitized individuals?

The last plot you made is called the likelihood function of p given the data (k and n) and is written $L(p;k,n) = p^k(1-p)^{n-k}$. Note that this is the same expression as the binomial probability function except that we look at the expression as a function of p instead of a function of k and we have skipped the binomial coefficient $\binom{n}{k}$ because this is just a constant that only affects the elevation of the curve (not the shape or the location of the peak). The value of the parameter(s) that maximize the likelihood function is called the maximum likelihood estimate (often shortened to MLE). This is your "best guess" for the value of the parameter. In this case we can write $\hat{p}_{MLE} = 0.4$ (the 'hat' above the p indicates that this is an estimate of the parameter).

- Usually, the logarithm of the likelihood function, the log-likelihood, is used to find the MLE. The log-likelihood function is in this case $\ell(p;k,n) = k\ln(p) + (n-k)\ln(1-p)$. The log-likelihood function and the likelihood function always have the peak for the same value of the parameter(s). You can confirm this by plotting the log-likelihood function in the spread sheet as well.

C)

In the simple example above you can find the MLE analytically by finding the value of the parameter where the likelihood function is flat (i.e., at the peak), by setting the first derivative to zero, $\frac{\partial \ell}{\partial p} = 0$, and solving this for p. You would then get $\hat{p}_{MLE} = k/n$. However, in more complex models this can often not be done analytically so it has to be done by "trail-and-error" numerically (which is often a lot easier anyway). There are several such functions for such numerical optimization in R (e.g., 'optim' and 'optimize'). It is also possible to do numerical optimization in Excel:

- Write an arbitrary value for p in one cell, and the likelihood function or log-likelihood function based on this p in another cell.
- In Excel 2007, go to "Problem solver" under the 'Data' tab. In older versions of Excel go to 'Tools > Solver' (if you haven't used Solver before, you probably need to go to 'Tools > Add-Ins' and select "Solver Add-in").
- Set up solver to maximize the likelihood by changing the value in the p-cell.
- Try various numbers of n and k

To do this exercise in R, you can use the following code:

```
N = 10
K = 4

# The likelihood function
L = function(p,k,n) p^k*(1-p)^(n-k)

# The log-likelihood function
l = function(p,k,n) k*log(p) + (n-k)*log(1-p)

# Plotting the binomial probability of k given p
p = seq(0,1,0.001)
plot(p, dbinom(k,n,p), type="l")

# Plotting the Likelihood function in a new window
windows()
plot(p, L(p,K,N), type="l")

# Plotting the Likelihood function in a new window
windows()
plot(p, l(p,K,N), type="l")

# The optimization functions in R finds the minimum, not the maximum. We
# therefor must create new functions that return the negavive likelihood
and
# log-likelihood, and then minimize these:

# Minus likelihood:
mL = function(p,k,n) -p^k*(1-p)^(n-k)

# minus log-likelihood:
ml = function(p,k,n) -(k*log(p) + (n-k)*log(1-p))

# Using 'optimize' (NB! if you have more than one parameter, you should
use
# 'optim'
optimize(mL, interval = c(0,1), k=K, n=N)

optimize(ml, interval = c(0,1), k=K, n=N)

K/N
```