Practical 3 (11 Feb 2015)

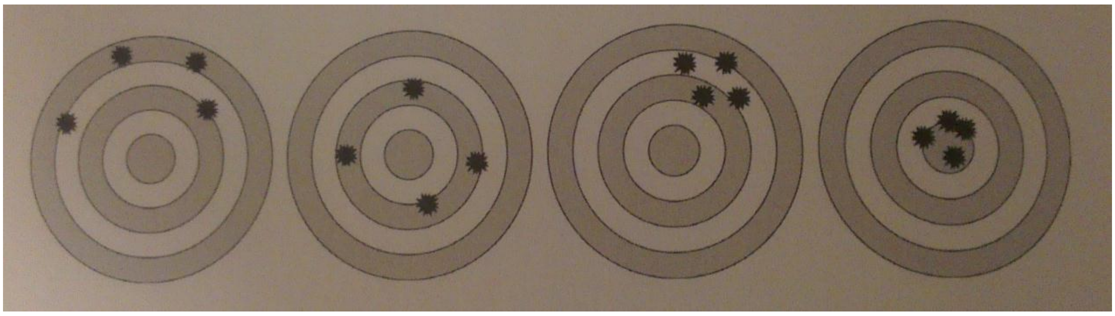
Question 1

What are the four properties of the maximum likelihood estimators? What do they mean in layman's terms?

Question 2

Accuracy: Measure of bias

Precision: Measure of spread



[photo credit: The Signal and the Noise: Why so many predictions fail-but some don't]

Relate the graphs to the statements below

- Accurate but not precise
- Precise but not accurate
- Not accurate and precise
- Both accurate and precise

Question 3

Coin tossing example revisited. If we observe y heads from n independent tosses, then the likelihood function of the unknown parameter p is $L(p) = \binom{n}{y} p^y (1-p)^{n-y}$.

i. If we observed $y = 7$ from $n = 10$ tosses, perform a likelihood ratio test (by hand) to test for $H_0: p = 0.5$ vs $H_1: p \neq 0.5$ at 5% significance level.

ii. Instead if we observed $y = 35$ from $n = 50$ tosses, the MLE is still the same ($\hat{p} = \frac{7}{10} = \frac{35}{50} = 0.7$). Perform a likelihood ratio test to test for $H_0: p = 0.5$ vs $H_1: p \neq 0.5$ at 5% significance level. Are the conclusions the same in two tests? Why?

iii. Find the 95% confidence interval for p for case (ii) above (using R).

Question 4 (to be discussed tomorrow)

A logistic regression question adapted from Mick Crawley's GLM course again. Yesterday we described how we can fit a logistic model under MLE framework. We would like to examine whether the survival of a plant is related to the number of flowers and the size of root. The expectation (hypothesis? Mick Crawley said that? From literature review?) is that plants that produce more flowers are more likely to die in the following winter, and plants with bigger roots are less likely to die.

The dataset `flowering.txt` has three variables: `State`, which indicates the current status for a particular plot of plant (1=alive, 0=dead). `Flowers` and `Root` are two explanatory variables as described above. First let us load the dataset into R and call it `flowering`:

```
flowering<-read.table('flowering.txt', header=T)
names(flowering)
```

Then we would to know how the plots look.

```
par(mfrow=c(1,2))
plot(flowering$Flowers, State)
plot(flowering$Root, State)
```

Because all of the values of the response are just 0's and 1's we get two rows of points: one across the top of the plot at y=1 and another across the bottom at y=0. The plots aren't very informative in this way… It is very hard to see any patterns in plots of binary data. Therefore the first part of this exercise is to fit a logistic regression with the given dataset. I hope you still remember the formulae for logistic regression, but it is not a huge problem if you don't as you can always look up the lecture notes on Day 3 (p.5-8).

$$y_i = Bernoulli(p_i), \text{ where}$$

$$p_i = expit(a + b * Flowers + c * Root)$$

Now let us construct the log-likelihood of the above model in R, namely `logistic.log.likelihood`

```
# TWO ARGUMENTS: parm IS A VECTOR OF PARAMETERS,
# dat IS THE INPUT DATASET
logistic.log.likelihood<-function(parm, dat)
{
# DEFINE PARAMETERS
a<-parm[1]
b<-parm[2]
c<-parm[3]

# DEFINE RESPONSE VARIABLE, WHICH IS THE FIRST COLUMN OF dat
State<-dat[,1]

# SIMILARLY DEFINE OUR EXPLANATORY VARIABLES
Flowers<-dat[,2]
Root<-dat[,3]

# MODEL OUR SUCCESS PROBABILITY
# IF YOU ARE NOT SURE ABOUT THIS, PLEASE REFER TO LECTURE NOTE
p<-exp(a+b*Flowers+c*Root)/(1+exp(a+b*Flowers+c*Root))

# THE LOG-LIKELIHOOD FUNCTION
log.like<-sum(State*log(p)+(1-State)*log(1-p))

return(log.like)
}
```

We may test whether our log-likelihood function is working properly. We may arbitrarily assign a set of parameter value, for instance, `c(0,0,0)`, and see what the function gives you.

```
# TRY
logistic.log.likelihood(c(0,0,0), dat=flowering)
```

You should get a number of around -40.89. This is the log-likelihood value for the parameters `a=0`, `b=0` and `c=0`. If not, please go back to your R code and check for typos. Next we would like to maximise our log-likelihood function with `optim()`. Create an object `M1` to store the output.

```
# MAXIMISE THE LOG-LIKELIHOOD
M1<-optim(????????????????????????????????????)
M1
```

So what are the MLE for the parameters? What is the associated log-likelihood value?

Some suggested that the interaction between Flowers and Root may also be significant in the model. We would like to build a slightly complex model to incorporate the interaction term. The model becomes:

$$p_i = expit(a + b * Flowers + c * Root + d * Flowers * Root)$$

Now, write down the log-likelihood function for this model. Luckily it is largely the same as the one without the interaction term.

```
logistic.log.likelihood.int<-function(parm, dat)
{
# DEFINE PARAMETERS, ONE MORE THIS TIME
????????????????????????????????????????????????

# DEFINE RESPONSE VARIABLE, WHICH IS THE FIRST COLUMN OF dat
State<-dat[,1]
# DEFINE EXPLANATORY VARIABLES
?????????????????????????????????????????????????

# MODEL OUR SUCCESS PROBABILITY
# IF YOU ARE NOT SURE ABOUT THIS, PLEASE REFER TO LECTURE NOTE
p<-????????????????????????????????????????????????

# THE LOG-LIKELIHOOD FUNCTION FOR A SINGLE DATA POINT
# REMAINS THE SAME.
log.like<-sum(State*log(p)+(1-State)*log(1-p))

# THE OVERALL LOG-LIKELIHOOD IS THE SUM OF THE LOG-LIKELIHOODS OF
THE OBSERVATIONS
return(log.like)
}
```

Maximise the log-likelihood function and store the output as M2.

Using the outputs M1 and M2, can you perform a likelihood-ratio test to test for the interaction term at $\alpha = 5\%$?