Caltech Department of Mathematics

Ma 3/103
Introduction to Probability and Statistics

KC Border
Winter 2017

Assignment 6: Maximum Likelihood and the World Series

Due Tuesday, February 21 by 4:00 p.m. at 253 Sloan

Instructions:

When asked for a probability or an expectation, give both a formula and an explanation for why you used that formula, and also give a numerical value when available.

When asked to plot something, use informative labels (even if handwritten), so the TA knows what you are plotting, attach a copy of the plot, and, if appropriate, the commands that produced it.

No collaboration is allowed on optional exercises.

Exercise 1 (The World Series Again)

The history of the World Series The first "World Series" was played in 1903, the most recent in 2016. There has been a World Series in every intervening year except two: in 1904 (when the NL champ refused to play the AL champ) and 1994 (the strike year). That makes a total of 112 Series. In 1903, 1919, 1920, and 1921 the Series had a best-of-9 games format. That leaves 108 best-of-7 Series. (The 1919 "Black Sox" scandal was a best-of-9 Series.)

Here are the number of series of each length for those 108 series.

Length of series	Number of series
4 games	21
5 games	25
6 games	24
7 games	38
All	108

(Source: http://en.wikipedia.org/wiki/List_of_World_Series_champions)

The length of a Series These calculations were part of Homework 2.

Let us assume that throughout a World Series a given team has a fixed chance to win each game, that games are independent random experiments. Let us also assume that the this probability is the same for the better team in every World Series. These assumptions may seem unrealistic to you, and they are the source of one of my favorite quotes about statistics. Frederick Mosteller [1] wrote, "It seems worthwhile to examine these assumptions a little more carefully, because any fan can readily think of good reasons why they might be invalid. Of course, strictly speaking, all such mathematical assumptions are invalid when we deal with data from the real world. The question of interest is the degree of invalidity and its consequences."

If the better team always won, then a best-of-7 Series would last only four games. As the probability gets closer to 1/2, one would expect more seven-game Series. The likelihood function depends on p, the probability that the "better" team wins a any particular game, and on N_k where N_k is the number of series where the winning team loses k games, so that the series lasts $4+k$ games, $k = 0, \dots, 3$.

Let $\text{plose}(k, p)$ be the probability that a team loses k games, but still is the first team to win the 4 games needed to win the Series, when its probability of winning each game is p. For this to happen, the team must win 3 and lose k of the first $3+k$ games, and then win the last game:

$$\text{plose}(k,p) = \underbrace{\binom{3+k}{k} p^3 (1-p)^k}_{\text{Prob of winning 3 and losing } k} \times \underbrace{p}_{\text{Prob winning last game}}$$

Let $\text{plen}(k, p)$ denote the probability that the Series lasts $4+k$ games. Since either team

may win the series,

$$\operatorname{plen}(k,p) = \operatorname{plose}(k,p) + \operatorname{plose}(k,1-p) = \binom{3+k}{k}\left[p^4(1-p)^k + p^k(1-p)^4\right] \qquad (k = 0,\dots,3).$$

The Likelihood Function In N Series, let N_k denote the number of Series where the winner loses k games. ($N = N_0 + N_1 + N_2 + N_3$.) The probability that this particular set of lengths occurs is also the likelihood function, and is given by the multinomial probability

$$\begin{aligned} L(p; N_0, N_1, N_2, N_3) &= \frac{N!}{N_0!N_1!N_2!N_3!} \prod_{k=0}^{3} \operatorname{plen}(k,p)^{N_k} \\ &= \underbrace{\frac{N!}{N_0!N_1!N_2!N_3!}\left[\prod_{k=0}^{3}\binom{3+k}{k}^{N_k}\right]}_{\text{independent of } p} \prod_{k=0}^{3}\left[p^4(1-p)^k + p^k(1-p)^4\right]^{N_k} \end{aligned}$$

Since we want to choose p to maximize the likelihood function we may ignore the positive constant term and just concentrate on the part that depends on p:

$$\tilde{L}(p; N_0, N_1, N_2, N_3) = \prod_{k=0}^{3}\left[p^4(1-p)^k + p^k(1-p)^4\right]^{N_k}.$$

Your Assignment The following table summarizes the number of best-of-7 Series where the winning team loses k games (108 in total).

k	0	1	2	3
N_k	21	25	24	38

1. Peruse Mosteller's analysis [1].

2. (5 pts) Graph the likelihood function as a function of p. (If you wish, you may discard the constants and use $\tilde{L}$ instead of L). Graph the log of the likelihood function.

   You should get graphs that are symmetric about 1/2. In particular, there will be two maxima.

3. (10 pt) Since we are interested in the probability that the better team wins, we should only consider $p \geqslant 0.5$. So find the maximum likelihood estimate of p subject to $p \geqslant 0.5$. Do the same for the logarithm of the likelihood.

4. (15 pt) Using this estimate, what is the probability that the better team wins a best-of-7 series?

5. There are other ways we can estimate p. One is the **method of moments**. Each value of p determines an expected number of losses by the winning team.

   (a) (5 pt) What is the formula for the expected number of losses in a first-to-win-four series as a function of p?

   (b) (5 pt) What is is the actual average number of losses by the winning team in the 108 7-game series?

   (c) (5 pt) What value of p equates the expected number of losses to the actual average number of losses? This is a moment estimator of p.

   Hint: One way to numerically solve an equation like $f(x) = c$ for x is to minimize $\Big(f(x) - c\Big)^2$ with respect to x. If the equation has a solution, then the minimum value is zero, and is attained at the solution.

Tips for R To define a function of variables m, k, and p, e.g.,

$$f(m,k,p) = \log\left(\binom{m}{k} p^k (1-p)^{m-k}\right)$$

use

```
f <- function (m,k,p) log( choose(m,k) * p^k * (1-p)^(m-k) )
```

(Note: `=` is a synonym for `<-`.) Note the example function is not the likelihood function that you want to use.

To graph a function, the `curve` command assumes the argument of the function is named x. To plot it over an interval (a, b), use the option `xlim=c(a,b)`. Axes labels are set with `xlab` or `ylab`. The main title is given by `main`. Here is an example of the syntax. (Note that you may use more than one line to enter a command in R.)

```
curve( f(7,4,x), xlim=c(0,1), xlab="p",
    ylab="Likelihood", main="Likelihood function")
```

To maximize a function f of one variable over the interval (a, b) use

```
optimize( f, interval=c(a,b), maximum=TRUE )
```

The `optimize` command minimizes f if the `maximum=TRUE` option is omitted. Be aware that if f is ill-behaved this may not work, so examine your results carefully. Consider maximizing the logarithm of the likelihood instead of the likelihood. It is typically better behaved numerically.

Tips for Mathematica To define a function of scalar variables m, k, and p, say

$$f(m,k,p) = \log\left(\binom{m}{k} p^k (1-p)^{n-k}\right)$$

use

```
f[m_,k_,p_] := Log[ Binomial[n,k] p^k  (1-p)^(n-k) ]
```

(Note that multiplication symbols, `*`, are optional, just leave space between symbols. Also note that the function's arguments are entered on the left-hand side with trailing underscore characters, and on the right-hand side without them. Finally note that `:=` is used between the left- and right-hand sides, and that functions use square brackets.) Note the example function is not the likelihood function that you want to use.

To graph a function f over the interval (a,b):

```
graphic =
Plot[ f[x], {x,a,b}, PlotLabel->"Likelihood Function",
AxesLabel -> {"p", "Likelihood"}]
Export["File.pdf",graphic]
```

Use the `Export` command to save your plot.

To maximize a function f of one variable over the interval (a,b) try

```
NMaximize[{f[p],  a <= p && p <= b}, {p}]
```

You may need to tweak some of the options to `NMaximize`.

□

Exercise 2 (10 pts) How much time did you spend on the previous exercises? □

Exercise 3 (Optional Exercise) (40 pts) An urn contains 4 balls each of a distinct color. At each step we draw two balls randomly, and change the color of the second one to the color of the first one, then we return the balls to the urn. What is the expected time of arriving to the case where all balls have the same color? (The process of drawing the two balls and replacing them takes place in one time period.) □

References

[1] F. Mosteller. 1952. The world series competition. *Journal of the American Statistical Association* 47(259):355–380. http://www.jstor.org/stable/2281309