Caltech Department of Mathematics

Ma 3/103
Introduction to Probability and Statistics

KC Border
Winter 2020

Assignment 5

**Due Tuesday, February 18 by 4:00 p.m.
in the dropbox in the lobby of Linde Hall
(the building formerly known as 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. You may use a calculator or software to compute the numerical values.

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.

There are three main exercises, one of which requires the use of scientific computing software. There are coding hints at the end.

Exercise 1 A large flat surface is ruled with parallel lines, spaced 1 unit apart. A needle (model it as a closed line segment) of length $\ell < 1$ units is dropped at random onto the surface. What is the probability that it intersects one of the ruled lines?

1. (10 pts) Specify and justify what "at random" means in this case. Hint: The distribution of the location of the needle's center, and its orientation need to be specified.

2. (5 pts) Compute the probability that the needle intersects more than one of the parallel lines.

3. (20 pts) Compute the probability that the needle intersects at least one of the parallel lines. Draw a picture to illustrate your reasoning.

4. (5 pts) If n needles are dropped independently, what is the expected number that intersect a line. □

Exercise 2 (50 pts) Pitman [2, Review Exercise. 3.30, p. 255.]

Diagonal neighbor random walk. Let (S_n, T_n) denote the position after n steps of a random walk on the lattice of points in the plane with integer coordinates, starting from $(S_0, T_0) = (0, 0)$. Suppose that $S_{n+1} = S_n \pm 1$ and $T_{n+1} = T_n \pm 1$ where the signs are picked by two independent tosses of a fair coin, independently at each step.

1. (20 pts) For $c > 0$, find the limit as $n \to \infty$ of the probability that (S_n, T_n) is inside the square with corners at $(\pm c\sqrt{n}, \pm c\sqrt{n})$.

2. (10 pts) Let $R_n = \sqrt{S_n^2 + T_n^2}$, the distance from the origin after n steps. Find $\boldsymbol{E}(R_n^2)$.

3. (10 pts) Find b, as small as you can, such that $\boldsymbol{E}(R_n) \leqslant \sqrt{bn}$ for every n.

4. (10 pts) Let p_n denote the probability that the random walk is at $(0, 0)$ after n steps. Find p_4 as a decimal.

5. (20 pts) S how that $p_{2m} \sim c/m$ as $m \to \infty$ for a constant c. What is c? □

Exercise 3 (The World Series Again)

The history of the World Series The first "World Series" was played in 1903, the most recent in 2019. (Sadly the Dodgers lost that one to the Red Sox.) 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 year of the players' strike). That makes a total of 115 Series. In 1903, 1919, 1920, and 1921 the Series had a best-of-9 games format. That leaves 111 best-of-7 Series.

Two cheating scandals have marred the results. In the 1919 series, which was a best-of-9 series, players from the Chicago White Sox were found to have taken money from gamblers to lose the series. (The "Black Sox" scandal.) In 2017, "The Dodgers were cheated out of the 2017 World Series championship," by the Houston Astros, as reported in the L.A. Times. Nevertheless, I have left the 2017 results in the sample.

Here are the number of series of each length for those 111 series.

Length of series	Number of series
4 games	21
5 games	26
6 games	24
7 games	40
All	111

(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,

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

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

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

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 (111 in total).

k	0	1	2	3
N_k	21	26	24	40

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?

I have some coding hints for R and Mathematica at the end of the assignment. □

Exercise 4 (10 pts) How much time did you spend on the previous exercises? **Please put the answer to this exercise on the *front page* of your answers and identify it as such.** □

Exercise 5 (Optional Exercise) (60 pts) The set A has n elements, and so has 2^n distinct subsets. For each subset, a ball is labeled with that subset and placed into an urn, and m balls (subsets) $B_1, \ldots, B_m$ are drawn in order at random *with replacement* from the urn. (There are 2^n balls, so each subset has probability $1/2^n$ of being drawn.)

What is the probability that

$$B_1 \subseteq B_2 \subseteq \cdots \subseteq B_m \ ?$$

Explain your answer.

(Hint: The solution is very pretty.) □

Coding Hints

Tips for R Here are some tips for using R. There are additional hints in the incomplete note at http://www.its.caltech.edu/kcborder/Courses/Ma3/Notes/RNotes.pdf

To define a function of variables m, k, and p, for example, to define

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

use

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

(Note: `=` is a synonym for `<-`.) Note that the example 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 save the graphic to a `.png` file, you need to something like

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

The `dev.off()` closes the file. The file will probably be saved in your home directory (Mac or UNIX).

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.png", 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`.

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

[2] J. Pitman. 1993. *Probability*. Springer Texts in Statistics. New York, Berlin, and Heidelberg: Springer.