

Sports Competitions and the Binomial Distribution

Michael P. Brenner

IAP Lecture
January 25 2016

Did the better team win?



Many sporting events have “play-offs”, designed to find out who is the better team.

Football: -- play one game

Baseball: -- World series: best out of 7

Tennis: -- 3 (or 5) sets in a match.

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 259

SEPTMBER 1952

Volume 47

THE WORLD SERIES COMPETITION*

FREDERICK MOSTELLER

Harvard University

Outline:

- (1) The basic mathematical ideas
 - (i) *The binomial distribution*
 - (ii) *Estimating “p” from data.*
- (2) The paper
- (3) Possible Extensions

Mosteller

IF WE compare pairs of teams, products, drugs, or persons on the basis of a fixed number of binomial trials, and identify the member of the pair that wins the majority of trials as the better, we may be in error. For example, if the members of a pair are evenly matched, a decision on the basis of performance is equivalent to coin-flipping. On the other hand, if one of a pair is actually better, it is more likely that the better member will also be the winner. If we carry out such comparisons on many pairs under roughly comparable conditions, it is of interest to estimate the over-all effectiveness of the decision technique in the kinds of situations that have occurred in practice. Data from the World Series are available to illustrate many facets of this type of problem.

About World Series time each year most fans are wondering which team will win. The author is no exception, but he has also wondered about another question: Will the Series be very effective in identifying the better team?

The Simple Model

Playing [baseball] games is like flipping an (unfair) coin!

The better team is the one with the higher probability of winning a single coin flip.

Now, the point of playing a “Series” is that you would like to amplify the chance that the better team (the one with higher winning probability) wins.

Intuitively, it seems clear that the more games you play in a series, the more closely it will amplify the advantage of the better team.

But how good is it? What is the chance that the better team will win a

1 game series

3 game series

5 game series

7 game series

The math problem:

Suppose we give you a coin, and the probability of getting a heads is p . What is the probability of getting m heads out of N trials?

The math problem:

Take $N=3$.

Possible outcomes of 3 coin flips:

HHT TTH
HTH THT
THH HTT
HHH TTT

0 heads	1 head	2 heads	3 heads
1	3	3	1

Probability of 2 heads and 1 tail: $p^2(1-p)$

Hence, probability of getting 2 heads = $3 \times p^2(1-p)$

[There are 3 different ways of getting 2 heads!]

Generalizing: the binomial distribution

$$P(N, m) = \frac{N!}{(N - m)!m!} p^m (1 - p)^{N - m}$$

i.e. $N=3, m=2$

$$\frac{N!}{(N - m)!m!} = \frac{3!}{1!2!} = 3 \times p^2 (1 - p)$$

$$\binom{n}{k} = \frac{n!}{k!(n - k)!}$$

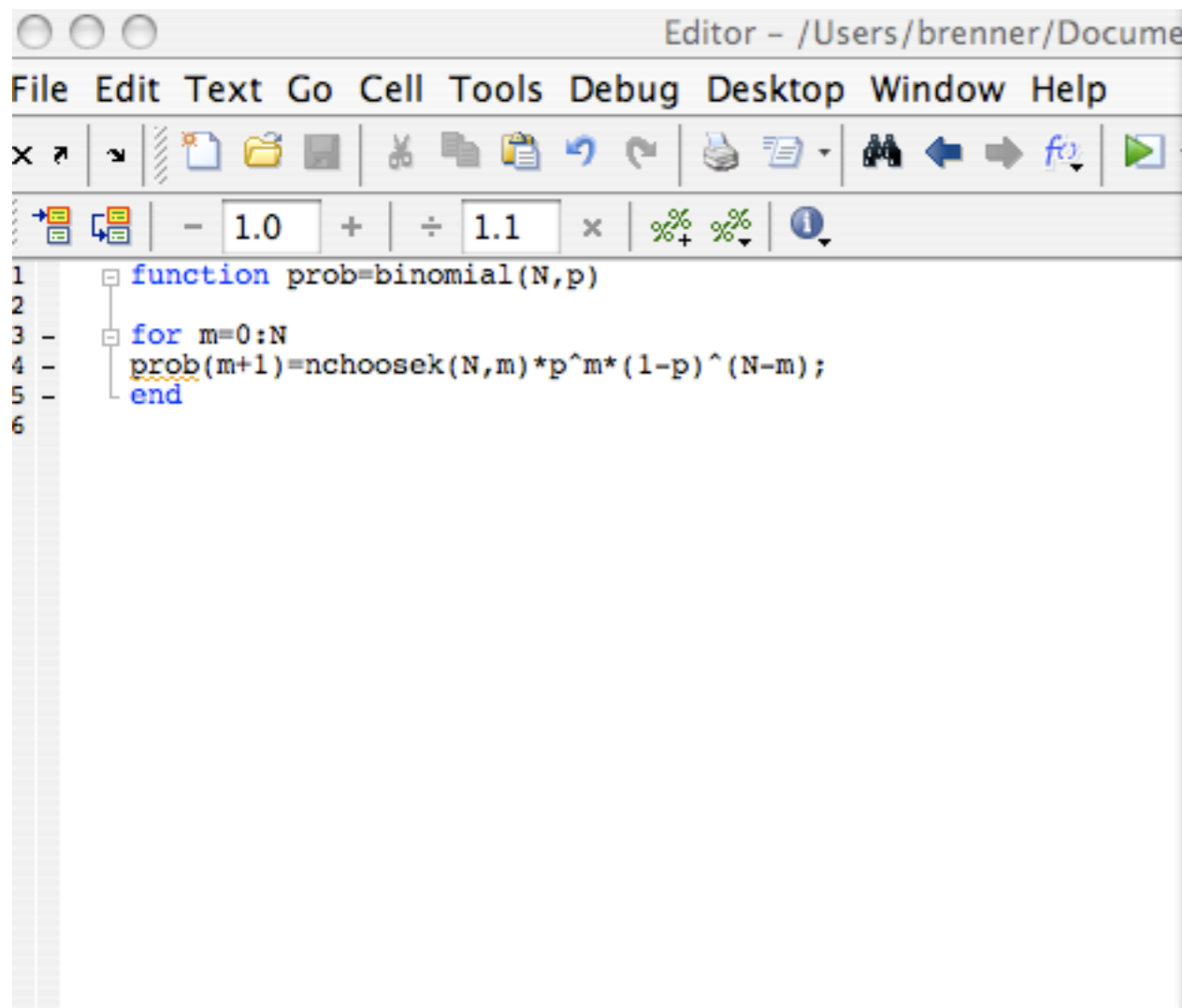
“N Choose k”:

Number of ways of choosing k wins among N games

Generalizing: the binomial distribution

$$P(N, m) = \frac{N!}{(N - m)!m!} p^m (1 - p)^{N - m}$$

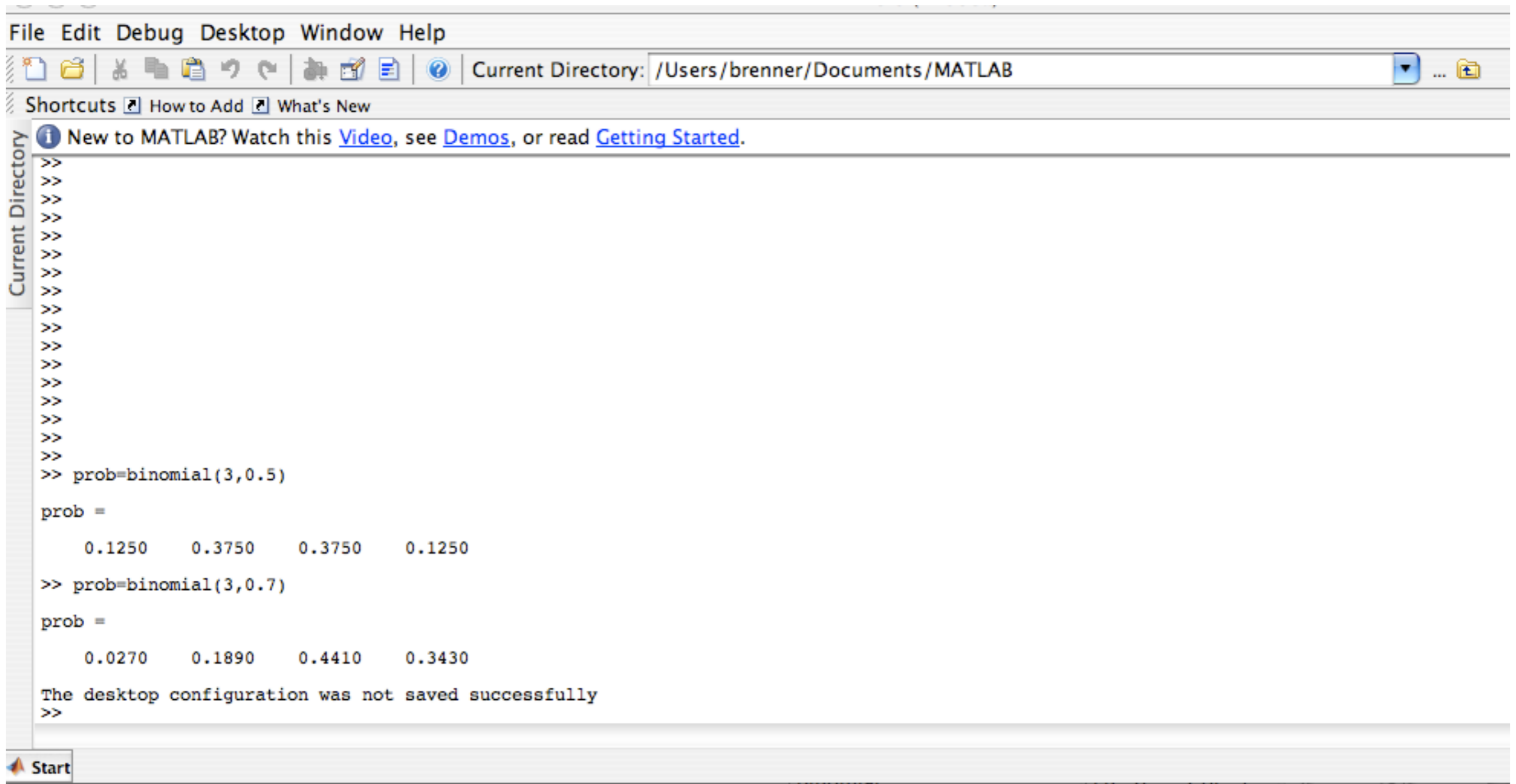
A Matlab Code



The image shows a screenshot of a Matlab editor window. The title bar reads "Editor - /Users/brenner/Docume". The menu bar includes "File", "Edit", "Text", "Go", "Cell", "Tools", "Debug", "Desktop", "Window", and "Help". The toolbar contains various icons for file operations, editing, and execution. Below the toolbar is a numeric keypad with values 1.0, 1.1, and 1. The main editor area contains the following Matlab code:

```
1 function prob=binomial(N,p)
2
3 for m=0:N
4 prob(m+1)=nchoosek(N,m)*p^m*(1-p)^(N-m);
5 end
6
```

Some Numbers for $N=3$

A screenshot of the MATLAB Command Window interface. The window title bar shows 'File Edit Debug Desktop Window Help'. The current directory is '/Users/brenner/Documents/MATLAB'. The command window contains the following text:

```
>>  
>>  
>>  
>>  
>>  
>>  
>>  
>>  
>>  
>>  
>>  
>>  
>>  
>>  
>>  
>>  
>>  
>>  
>>  
>>  
>>  
>>  
>>  
>>  
>>  
>>  
>> prob=binomial(3,0.5)  
prob =  
    0.1250    0.3750    0.3750    0.1250  
>> prob=binomial(3,0.7)  
prob =  
    0.0270    0.1890    0.4410    0.3430  
The desktop configuration was not saved successfully  
>>
```

Mathematical Properties of Binomial Distribution

$$\mu = \sum_m m P_{Binomial}(N, m) = Np$$

$$\sigma^2 = \sum_m (m - \mu)^2 P_{Binomial}(N, m) = Np(1 - p)$$

The Simple Model for Sports Competitions

Assume the probability for the winning team to win is p .

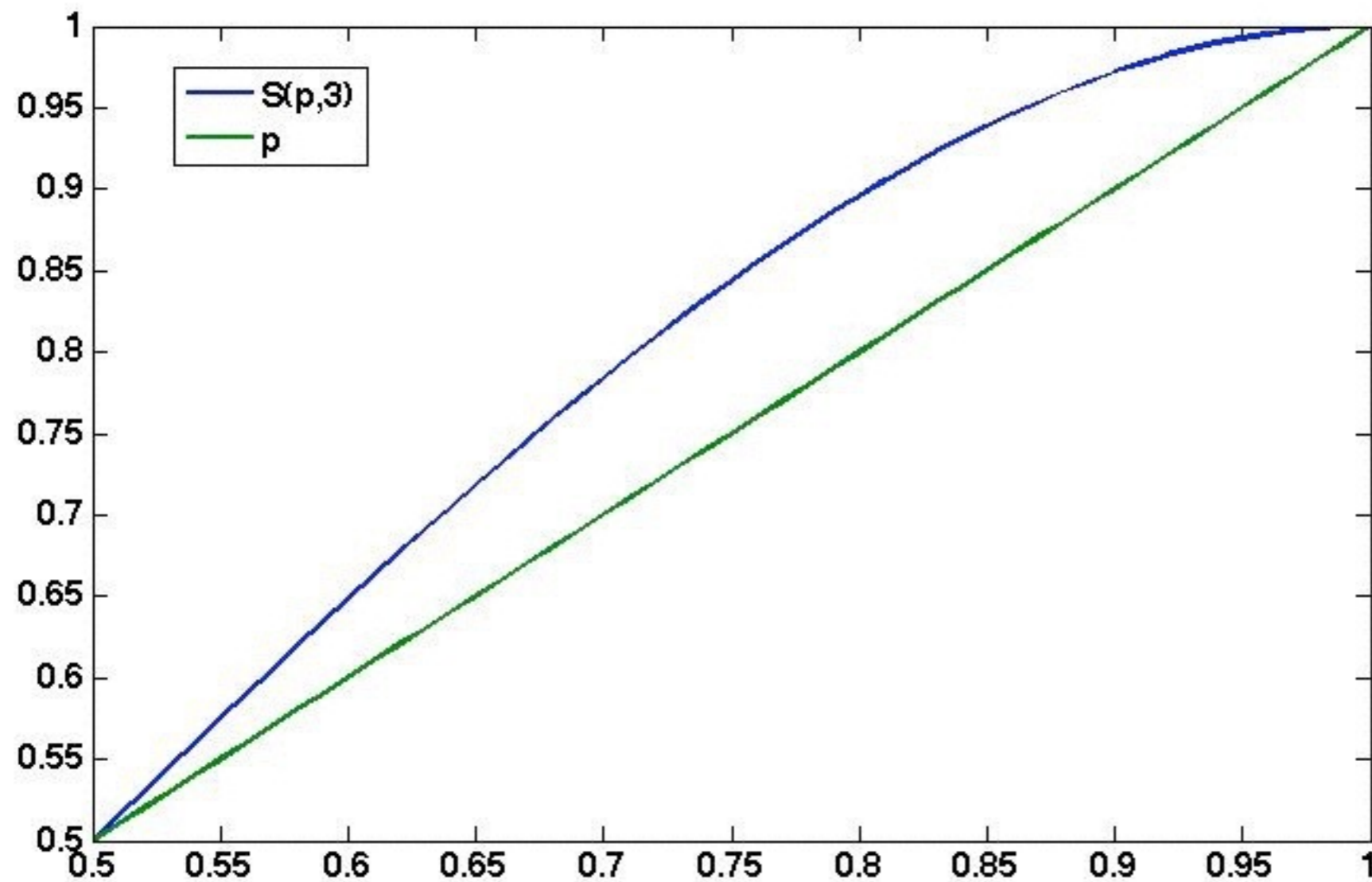
Assume this probability is constant throughout the series.

As a simple model for the League play-offs, we might suppose that one team has probability p of winning single games, and that the other has probability $1 - p$ and further, that p remains the same for all games in the series under consideration. The reader may easily think of arguments against the assumption of the constancy of p from game to game (pitchers and ball parks are examples of variables that could contribute to variation in p). We will provide some evidence on this issue later. Corresponding to the probability p of winning a single game,

$S(p, n)$ Probability of team with prob p
of winning a single game to win a n game series

$$S(p, 3) = p^3 + 3p^2(1 - p)$$

e.g. for a 3 game series...



Problem:
What about 10 game series

Not that much is gained from a 3 game series for nearly matched teams!

The simple model might be objectionable.

But, does it fit the data?

If it does, you have no logical reason to object!

Model A

The better team has the same p every year.

We do not know who the better team is. We just assume that whoever it is has exactly the same p from year to year.

Data from world series, first 50 years of 20th century

TABLE 5
GAMES WON (SEVEN-GAME SERIES ONLY)

Winner	Loser	Frequency	Theoretical Proportion
4	0	9	$p^4 + q^4$
4	1	13	$4p^4q + 4pq^4$
4	2	11	$10p^4q^2 + 10p^2q^4$
4	3	11	$20p^4q^3 + 20p^3q^4$
		<hr/>	<hr/>
		Total 44	1

TABLE 5
GAMES WON (SEVEN-GAME SERIES ONLY)

Winner	Loser	Frequency	Theoretical Proportion
4	0	9	$p^4 + q^4$
4	1	13	$4p^4q + 4pq^4$
4	2	11	$10p^4q^2 + 10p^2q^4$
4	3	11	$20p^4q^3 + 20p^3q^4$
		Total	44
			1

First term is chance better team won; second term is chance worse team won.

BBBWWB
BBWBWB
BWBBWB
WBBBWB
BBWWBB

BWBWBB
WBBWBB
BWWBBB
WBWBBB
WWBBBB

10 ways for better team to win in 6 games

Want to figure out how to use the data that is given, with the theory to estimate p!

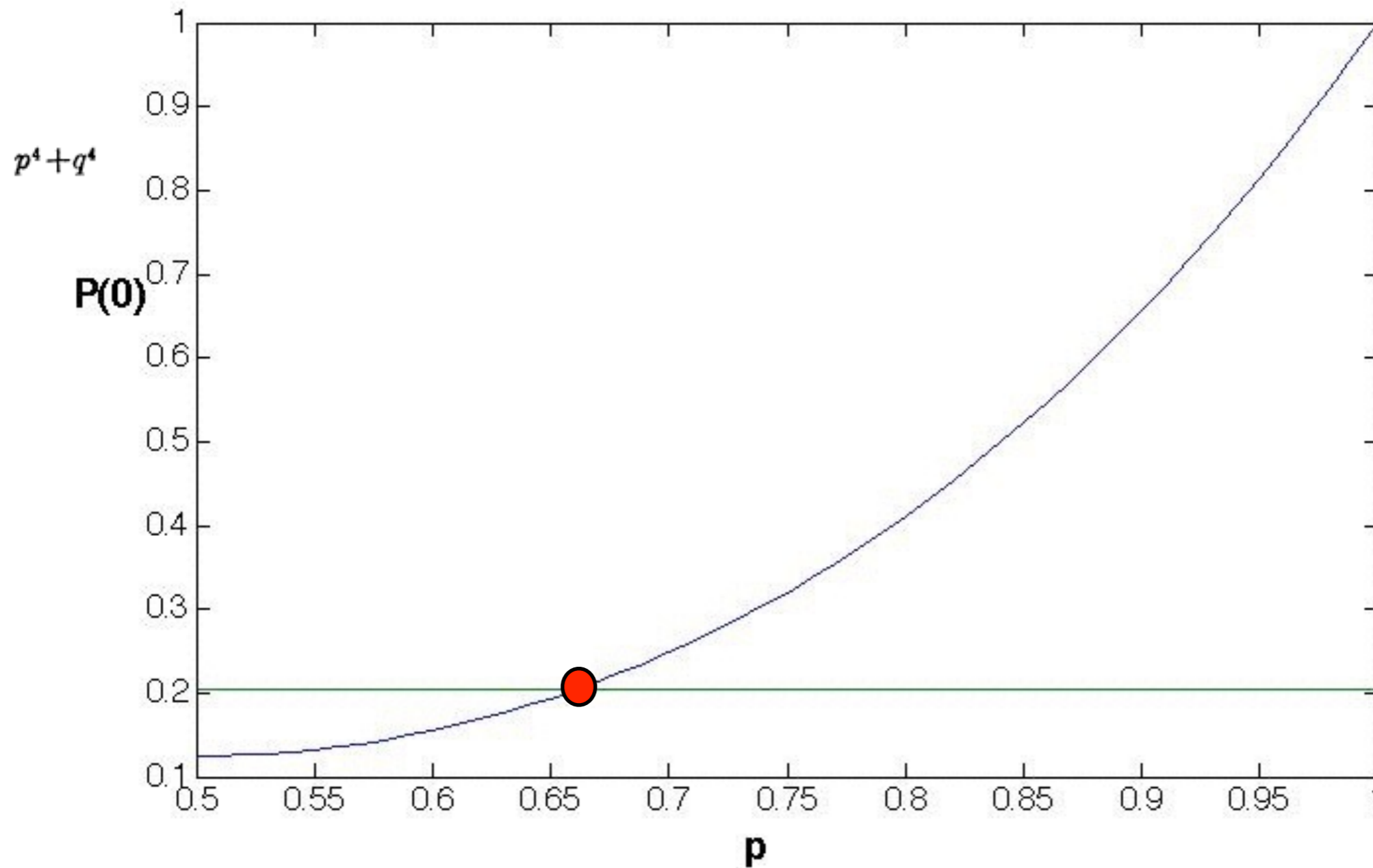
Want to figure out how to use the data that is given, with the theory to estimate p !

There are LOTS of ways of doing this. Before going through his method look at this.

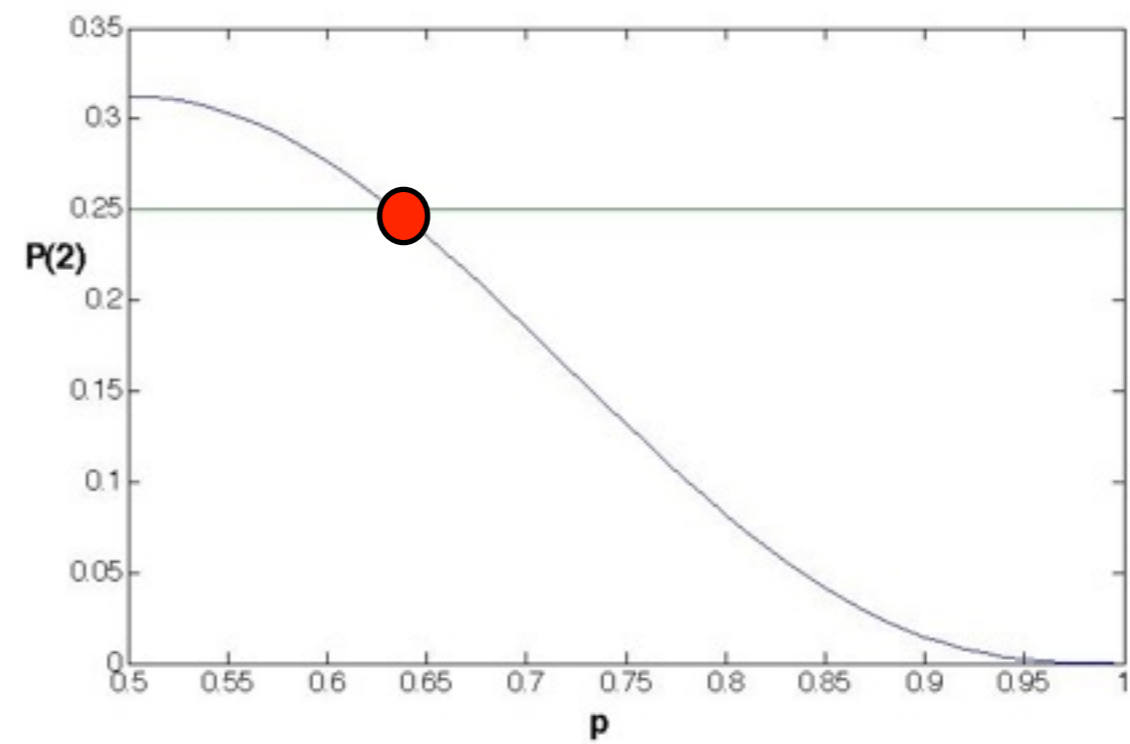
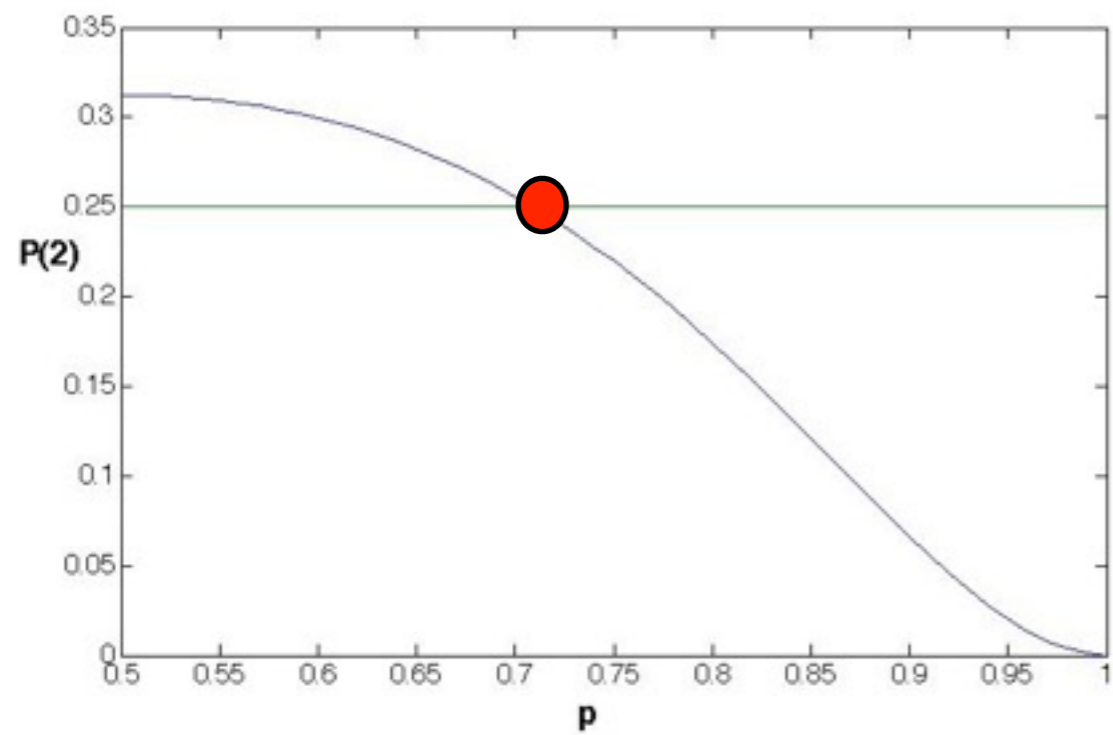
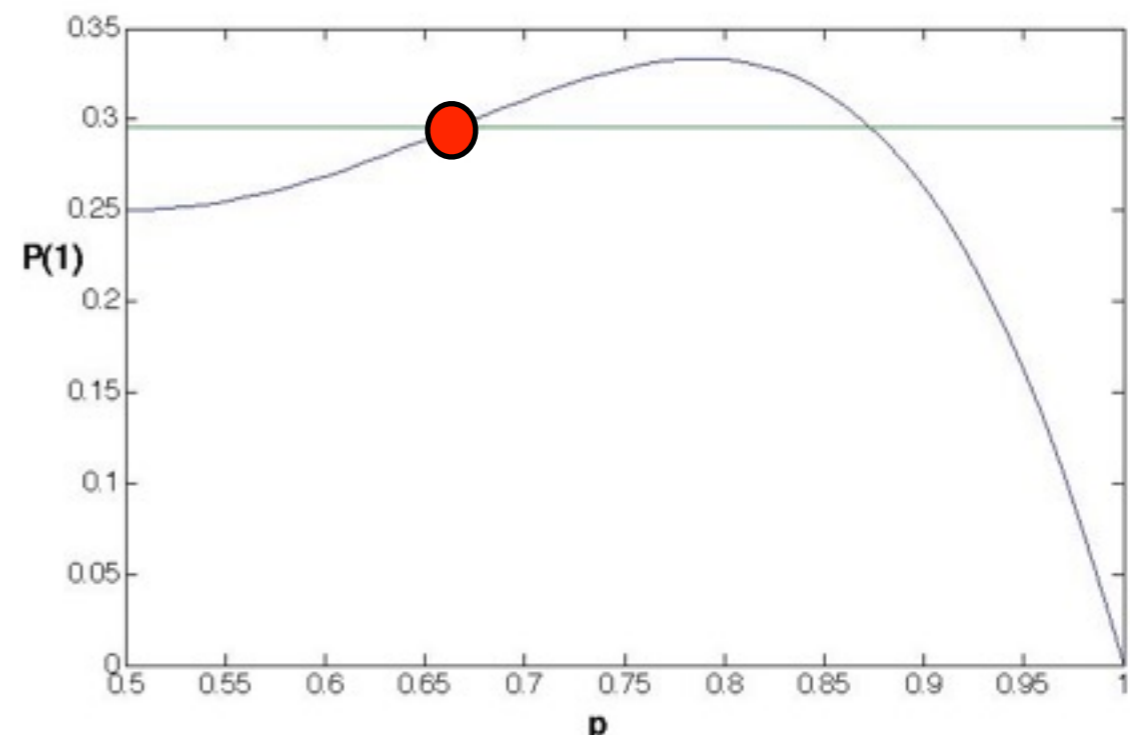
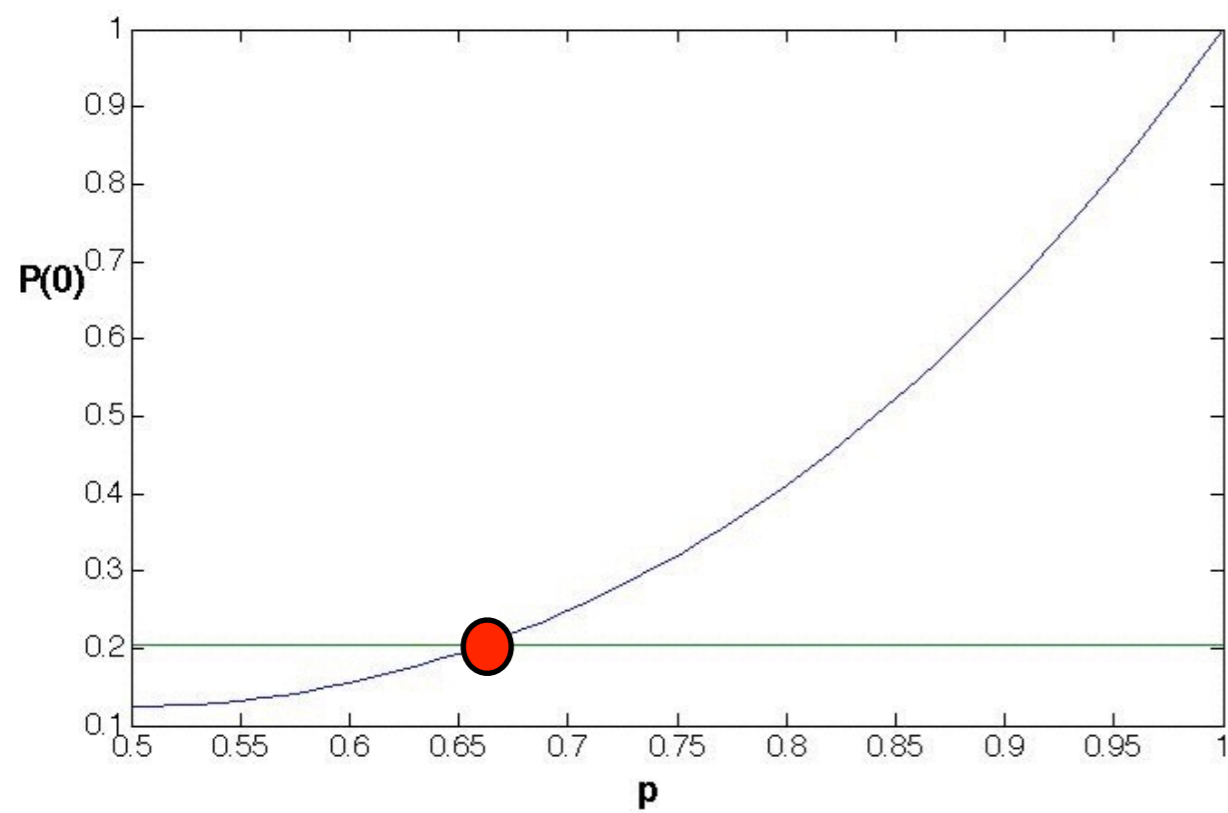
TABLE 5
GAMES WON (SEVEN-GAME SERIES ONLY)

Winner	Loser	Frequency	Theoretical Proportion
4	0	9	$p^4 + q^4$
4	1	13	$4p^4q + 4pq^4$
4	2	11	$10p^4q^2 + 10p^2q^4$
4	3	11	$20p^4q^3 + 20p^3q^4$
		Total	44
			1

9/44~ 20% of the time the series is 4-0



Seems that $p=0.65$ is pretty good

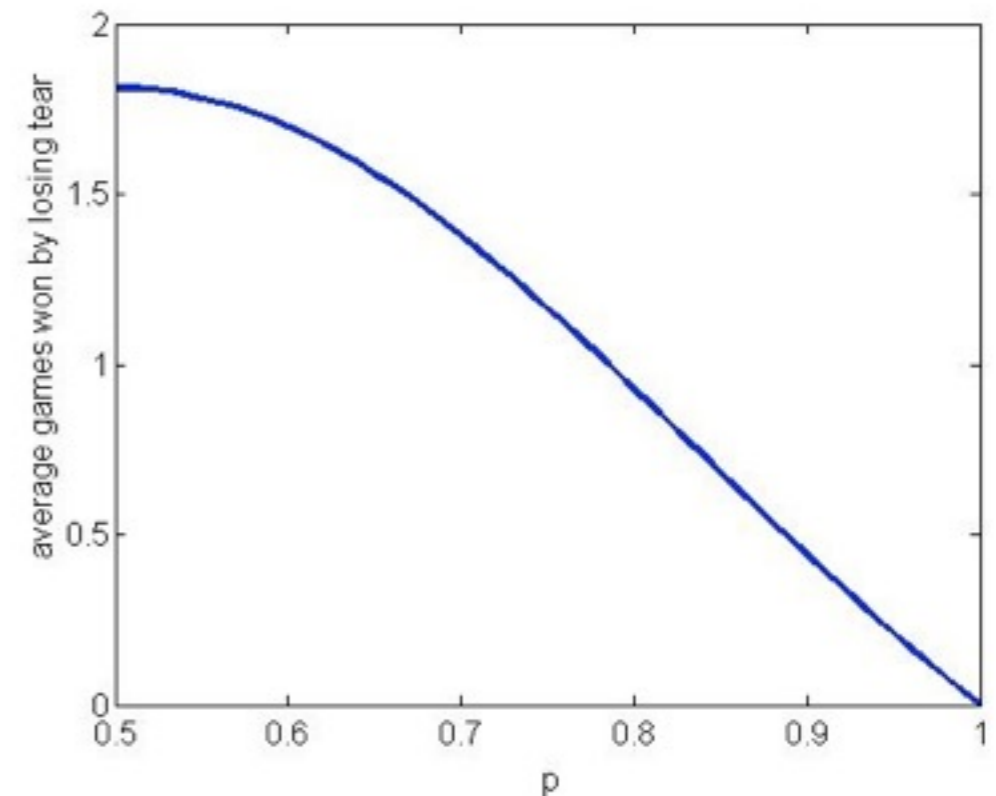


Mosteller uses 3 methods to get the best p

(I) He makes the average number of games lost by the winning team match.

Method 1. We wish to obtain the average number of games won by the Series-losing team in terms of p . We multiply the theoretical proportions from Table 5 by the number of games won by the loser and add. This operation gives

$$A = 4pq[1 + 2pq + 5p^2q^2].$$



(2) “Maximum likelihood”.

Method 2. If $P(0)$, $P(1)$, $P(2)$, $P(3)$ are the probabilities that the Series-losing team wins 0, 1, 2, or 3 games respectively in a Series, then the maximum likelihood approach involves finding the value of p that maximizes

$$[P(0)]^9[P(1)]^{13}[P(2)]^{11}[P(3)]^{11}.$$

The numbers 9, 13, 11 and 11 are the frequencies tabulated in Table 5 and the $P(x)$ are given in algebraic form in the Theoretical Proportions column in Table 5. Although tedious, this maximization was done, and the estimate obtained was 0.6551, encouragingly close to that obtained from Method 1.

(3) Chi squared.

All give basically same answer

<i>Method</i>	<i>Estimate</i>
Average Wins by Series Loser	0.6542
Maximum Likelihood	0.6551
Minimum Chi-square	0.6551

This is because of what our plots showed.

Probability that best team wins

$$S(0.65, 7) = 0.80$$

Is this a good model?

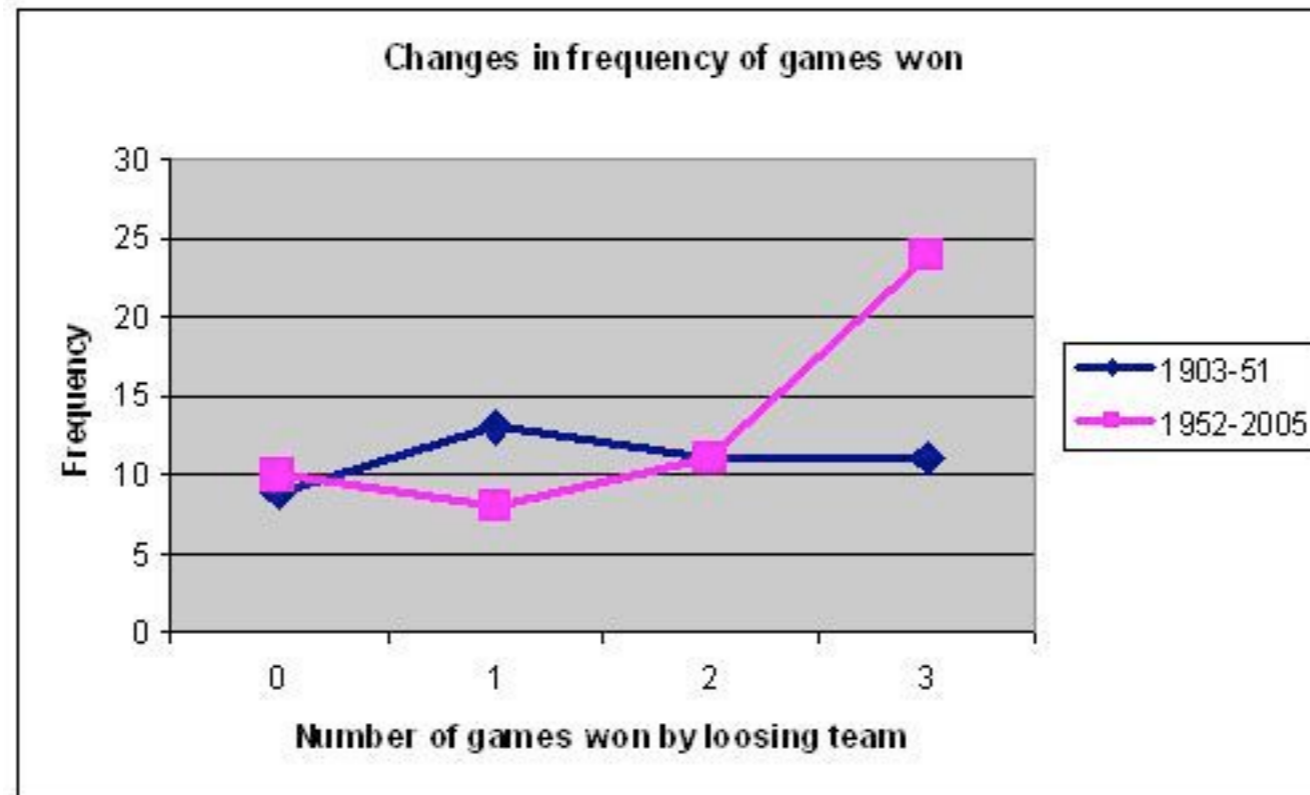
Issues

(1) Teams do better at home than away

(2) Winning one game affects whether you win the next game

(3) ...

The second half of the 20th century!



Huge amplification of 4-3 games!
Clear breakdown of binomial model...why?

Homework:

Playing “baseball” with MATLAB.

The game. Take $p=0.65$.

Draw a random number r uniformly distributed in $[0, 1]$

If $r < p$, you win!

Play a 7 game series. Record number of times that you win.

Do this 50 times. Compare to Mosteller’s data and theory.
Now do it 500 times. Compare again.

A challenge:

Use your simulations to find a simple explanation for the large number of 4-3 games in the 2nd half of the 20th century