Applied Math 115/215 — Fall 2025.
Lecture 2

Outline:

(1) Maximum Likelihood Estimation — a simple example

Lecture 1 Notebook

(2) Mostellar's World Series Model

(3) Turning Mostellar's model into code

Lecture 2 Notebook

Maximum Likelihood Estimation

Suppose we have the outcome of some experiment as a series of results $X_1, X_2, \ldots, X_n$ of independent trials of some process. We suspect that a good way to describe our experiments is a probabilistic model, that is our results all come from repeatedly sampling from an underlying probability distribution. Suppose that we are able to identify a good candidate distribution, say a Normal distribution, or a Bernoulli one. We may describe the probability of observing one particular X_i as $f(X_i)$ or, more explicitly, $f(X_i|\theta)$. In this case θ represents the parameters upon which the precise shape of distribution f depends, e.g. θ = (μ,σ) in case of a Normal distribution.

If we only have access to the samplings, we usually want to know also what are the best values of the parameters θ for describing the process we are seeing. One way of doing this is to choose the parameters θ so that the probability of observing exactly what we saw is maximal. The basic idea here is that, no additional information given, the most sensible thing to assume is that we saw (one of) the most probable outcomes of the process. This procedure goes by the name of **Maximum Likelihood Estimation** (or MLE for short).

Given the name of the course, we would very much like to translate all of this nice reasoning in more mathematical terms.

Given our favorite underlying distribution $f(X|\theta)$, the probability of seeing a set of n IID observations $\{X_i\}$ will be:

$$L(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

This function is usually called **Likelihood**, since it describes how likely observed data are given the parameters. Finding the θ that maximize this function amounts to requiring:

$$\hat{\theta} = argmax_\theta \, L(\theta)$$

Mosteller's World Series model

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 259 SEPTEMBER 1952 *Volume 47*

THE WORLD SERIES COMPETITION*

FREDERICK MOSTELLER
Harvard University

Full paper available on Canvas

Mosteller's World Series model

- **Assume that in the World Series, one team wins with probability *p*, like a flipping a unfair coin**
- **The better team is the one with the higher probability of winning a single coin flip**
- **The point of playing a "series" is that you would like to amplify the chance the better team wins**
- **Central question: how often does the better team win?**

Binomial trials

- **Suppose you are given a coin, and the probability of getting a head is *p*. In *N* trials the probability of getting exactly *m* heads is**

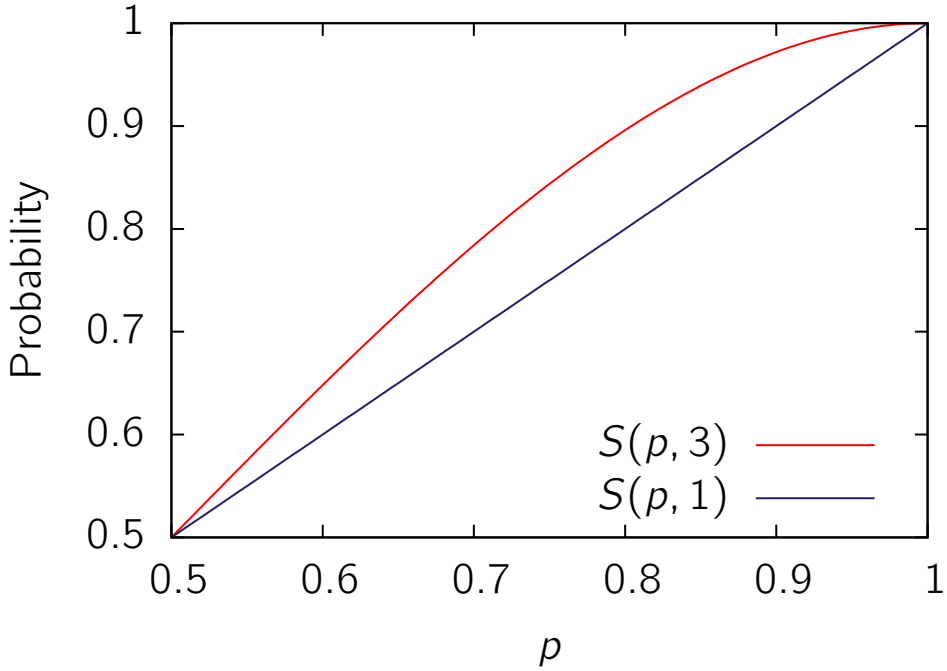
$$P(N,m) = \frac{N!}{(N-m)!m!} p^m (1-p)^{N-m}$$

- **Alternative notation**

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

"*N* choose *k*" : the number of ways to choose *k* items from *N* things

Success probability



- **Define *S*(*p*,*N*) to the be the probability of winning an *N* game series**
- **Then**

$$\begin{aligned} S(p,3) &= \mathbb{P}(\text{Win three}) + \mathbb{P}(\text{Win two}) \\ &= p^3 + 3p^2(1-p) \end{aligned}$$

- **Winning combinations:**

BBB, BBW, BWB, WBB

Mosteller's notation
B: better team wins
W: worse team

Amplification factor

- Let ϵ be the "advantage," so that $p = \frac{1}{2} + \epsilon$
- Mosteller's calculation:

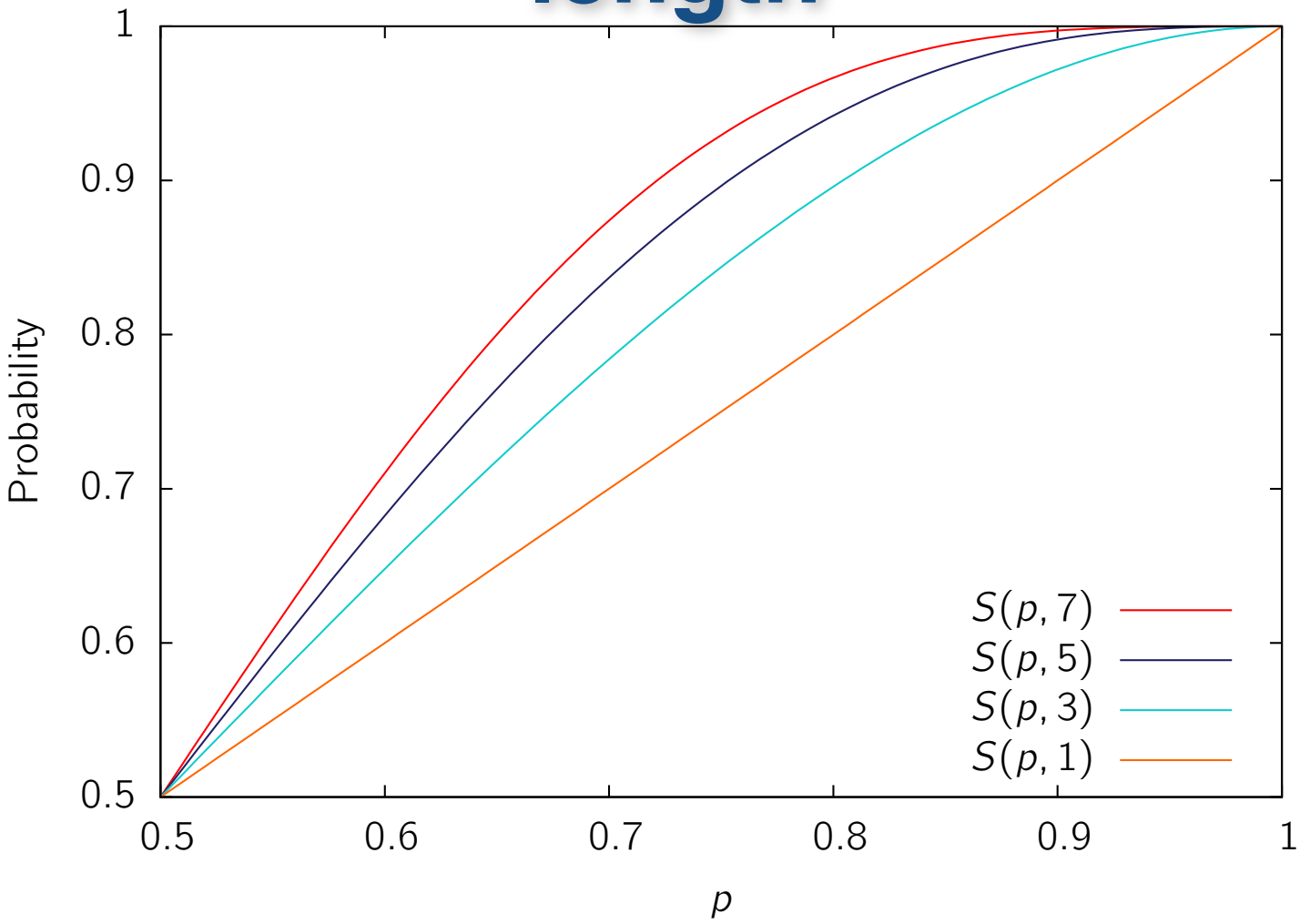
$$\begin{aligned} S(p, 3) &= p^3 + 3p^2(1 - p) = p^2(3 - 2p) \\ &= (\tfrac{1}{2} + \epsilon)^2(2 - 2\epsilon) \\ &= \tfrac{1}{2} + \tfrac{3}{2}\epsilon - 2\epsilon^2, \end{aligned}$$

- Increased chance of winning

$$S(p, 3) - S(p, 1) = \frac{\epsilon}{2} - 2\epsilon^3$$

- For p=0.51, so that ϵ = 0.01, this is 0.005

Amplification grows with series length



Exercise: calculate amplification factor for longer series lengths

What's an appropriate *p*?

- To answer the central question, Mosteller aims to estimate a value of *p* based on historical World Series data
- Examine period from 1905 to 1951. During this time, there were 44 seven-game series, plus 4 nine-game series
- Each World Series has one team from each league
  - National League (St. Louis Cardinals, S.F. Giants, *etc.*)
  - American League (Boston Red Sox, New York Yankees, *etc.*)

Mosteller's first estimation

- **Assume that the better has the same *p* each year**
- **Define $q = 1 - p$**
- **Analyze the 44 seven-game series:**

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	
4	3	11	
	Total	44	

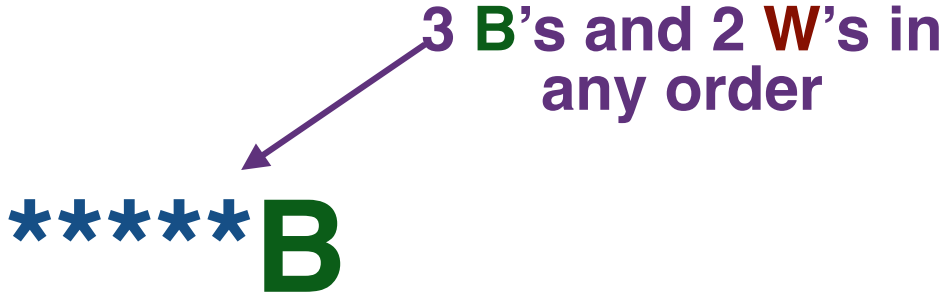
Where does this come from?

Combinations of games won

- **Four combinations for better team to win in a 4–1 series**

WBBBB, BWBBB, BBWBB, BBBWB

- **How many ways are there for a better team to win a 4–2 series?**



$$\binom{5}{2} = \frac{5!}{2!3!} = 10$$

Mosteller's full table

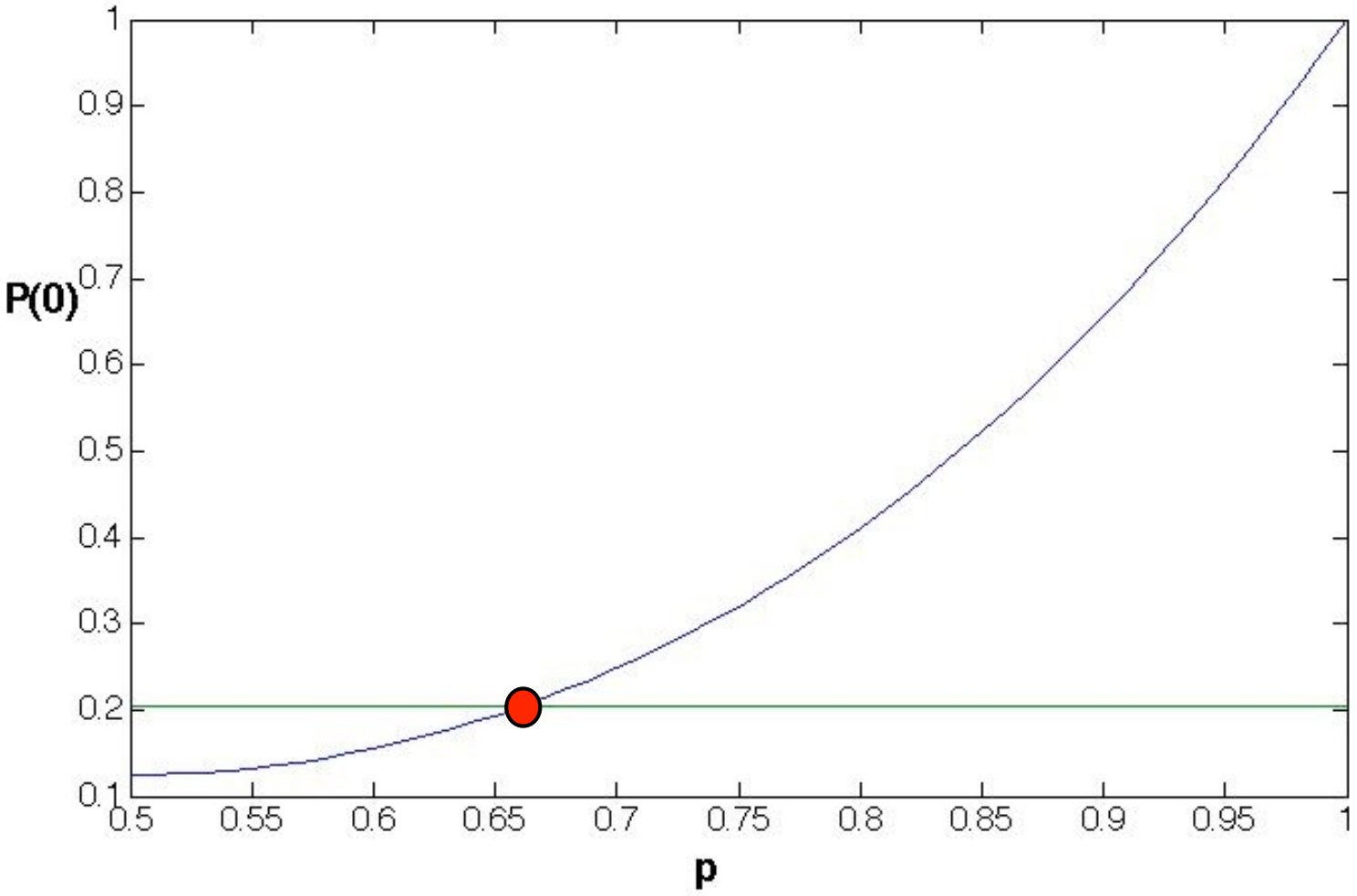
TABLE 5

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

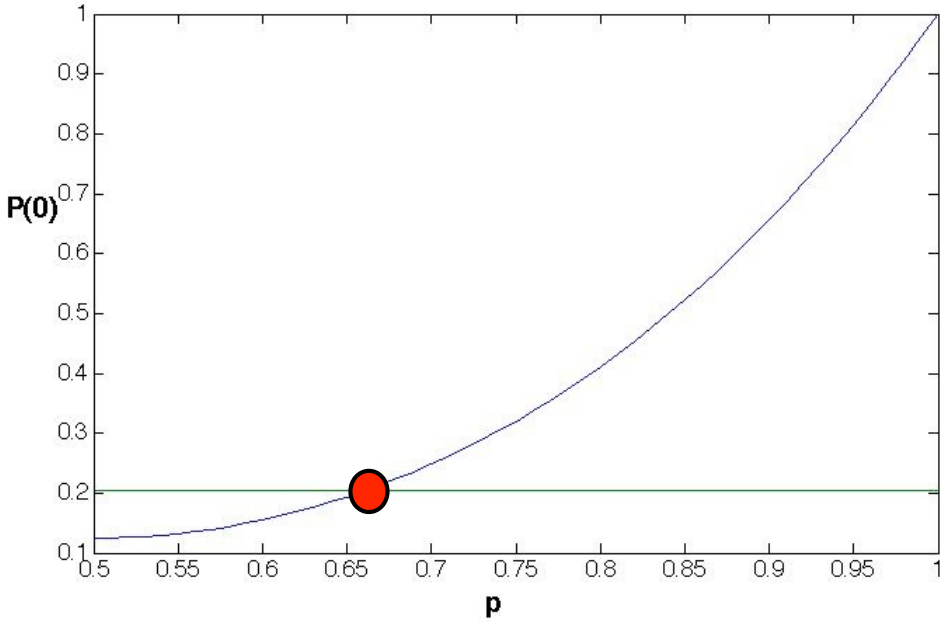
- Can try choosing *p* to fit the observed frequencies
- Let's just examine the 4–0 series, which occur 9/44 = 20.4% of the time

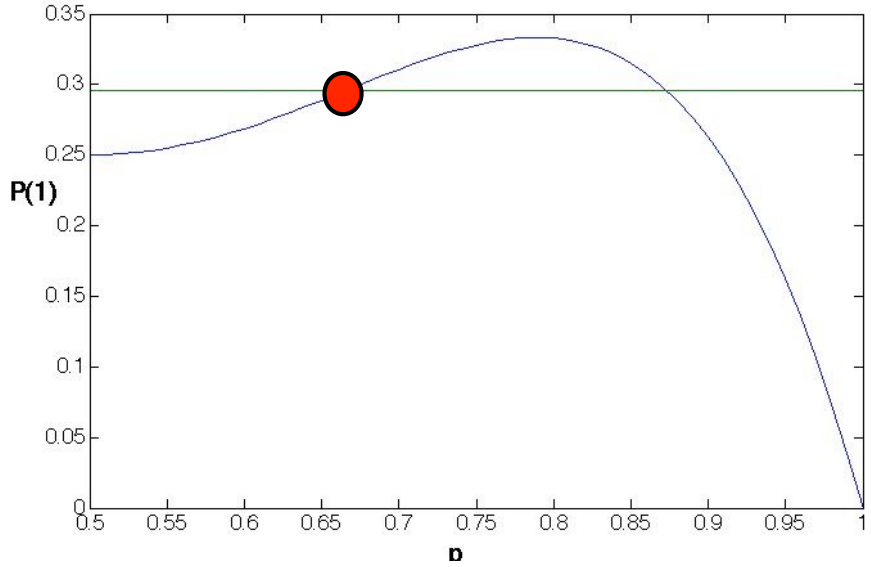
An estimate for *p*

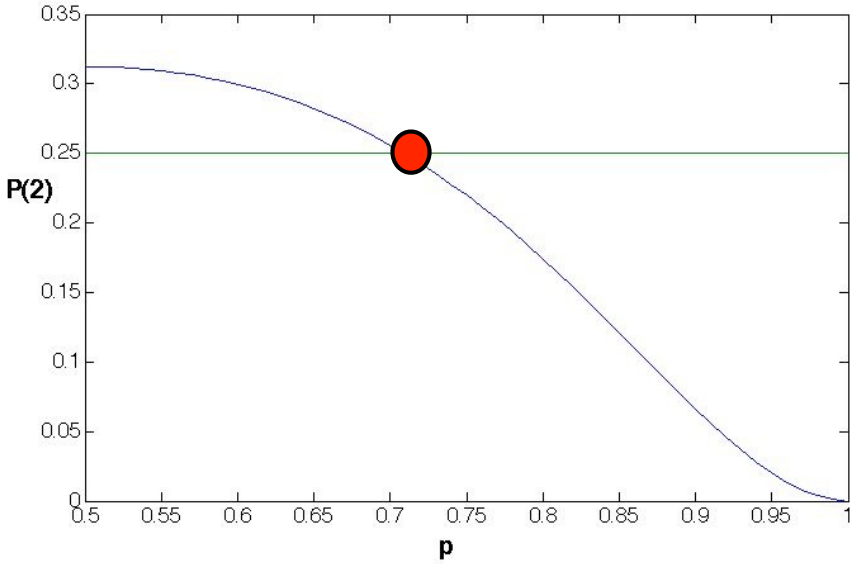


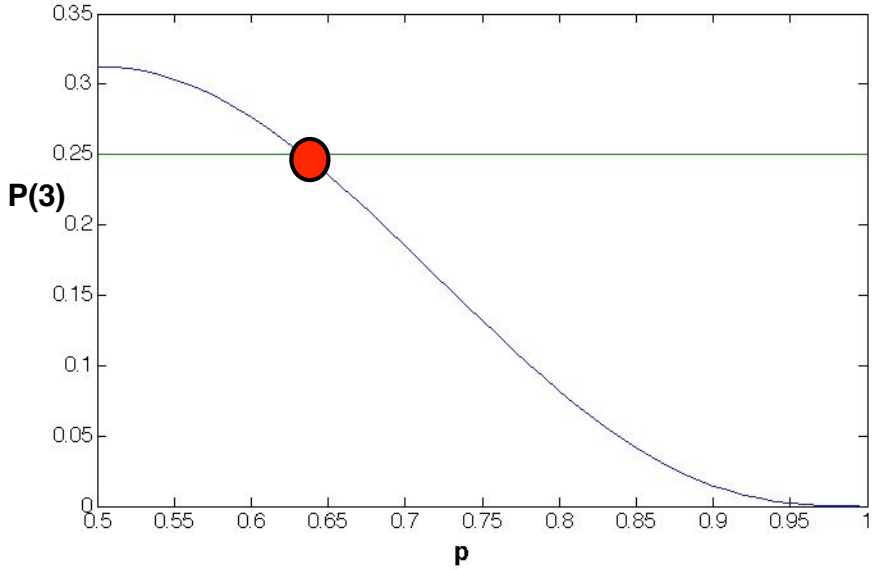
Looks like $p = 0.65$ is pretty good

In fact $p = 0.65$ fits all the data









Mosteller's three estimation methods

- **Mosteller considers three methods of systematically choosing *p*:**
  1. **He makes the average number of games lost by the winning team match**
  2. **He uses a maximum likelihood argument**
  3. **He uses a χ^2 statistic**
- **We'll examine the first two**

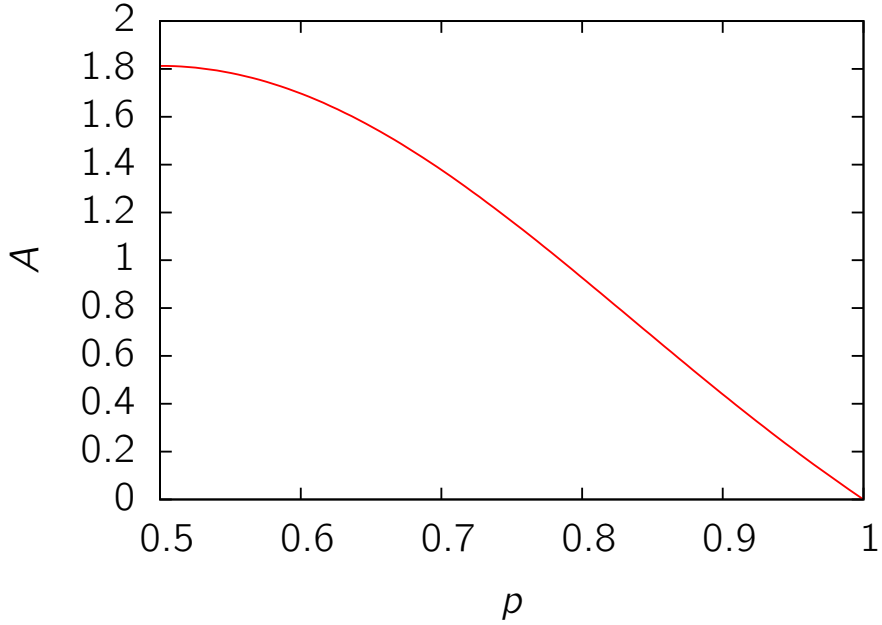
Average number of losing games

$$
\begin{aligned}
A &= \text{Average number of games won per Series by the Series-loser} \\
&= 1(4p^4q + 4pq^4) + 2(10p^4q^2 + 10p^2q^4) + 3(20p^4q^3 + 20p^3q^4) \\
&= 4pq[(p^3 + q^3) + 5pq(p^2 + q^2) + 15p^2q^2].
\end{aligned}
$$

- **After some manipulations,**

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

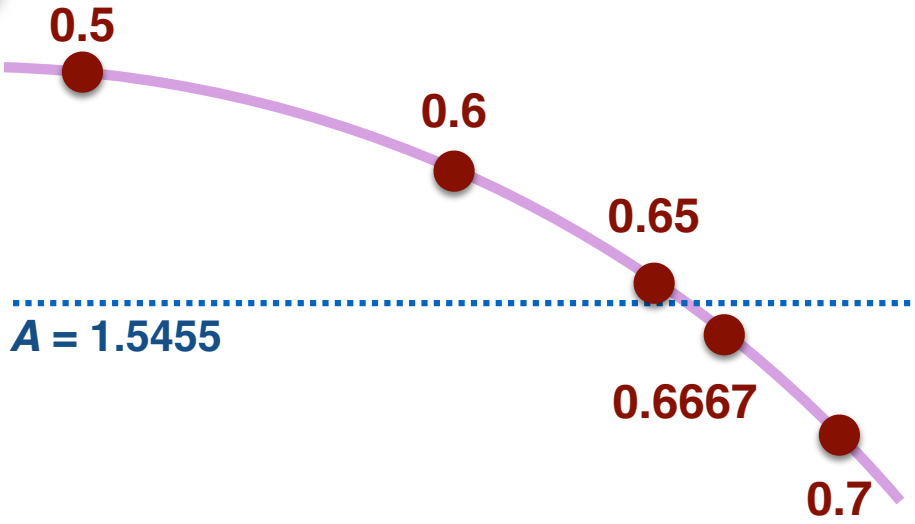
- **Problem: find value of *p* corresponding to the value in the data of *A* = 1.5455**
- **But the right hand side is a sixth order polynomial in *p***



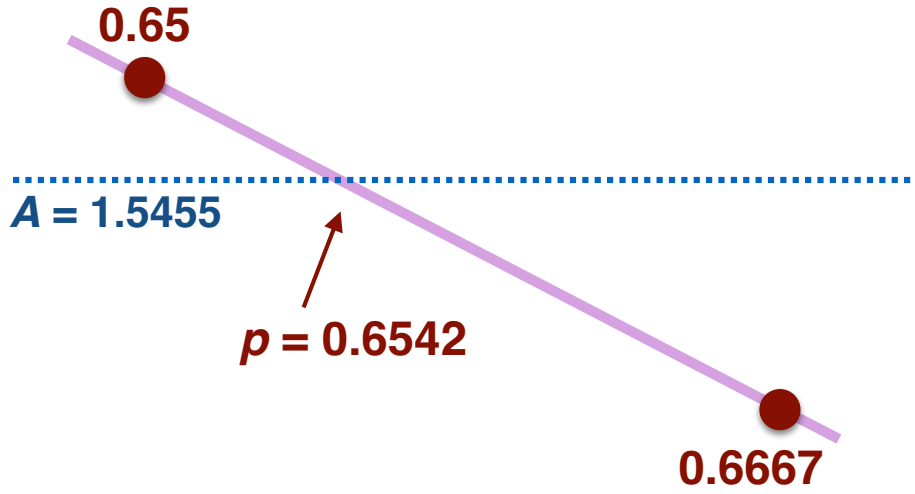
Root finding the old school way

- Mosteller calculates the value of *A* for a number of different values of *p*:

p	*A* = Average Wins Expected by Loser
0.5	1.8125
0.6	1.6973
0.6500	1.5596
0.6667	1.5034
0.7	1.3780



- He then linearly interpolates between the two
- Similar to modern computer root-finding strategies (see "bisection search")



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.

Summary of estimations of *p*

Method	*Estimate*
Average Wins by Series Loser	0.6542
Maximum Likelihood	0.6551
Minimum Chi-square	0.6551

- **All three approaches give almost the same answer, which is very close to our initial guess of $p = 0.65$**
- **Can now answer the original question of how often the better team wins:**

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

- **This means that 20% of the time, the better team doesn't win**

How often does the A.L. have the better team?

- Recall that the A.L. team won 31 out of 48 times
- Let *x* be the proportion of series where the A.L. has the better team. Then

$$0.80x + 0.20(1 - x) = \frac{31}{48} = 0.646$$

- After some algebra,

$$0.60x = 0.446$$
$$x = 0.743.$$

A key assumption in the analysis

- Same p for all world series
- More realistic to say that the better team will have a varying advantage. Sometimes one team may be much better, sometimes both teams may be almost equally matched

Scenario 1
p=0.7 for all games

$S(0.7,7)=0.87$

Scenario 2
p=0.5 for half, p=0.9 for half

$S(0.5,7)=0.5$, $S(0.9,7)=1.00$

Average S = 0.75

- See Mosteller's paper for a more advanced approach with varying p