Harvard John A. Paulson
School of Engineering
and Applied Sciences

Michael P. Brenner

Homework 1

Modelling tournaments

Issued:	September 10, 2025
Due:	September 26, 2025

Note that this is roughly a 2-week exercise. *Do not procrastinate the work.*

Problem 1: Maximum Likelihood Estimation (*10 points*)

Suppose the World Series were best of 3 games. Assume like Mostellar that each game is a Bernoulli trial with probability p. Assume that we see n_0 events where the **losing** team wins 0 games, n_1 events where the **losing** team wins 1 game.

1. Following the derivation in class, derive the Likelihood of our observations.
2. Analytically compute the maximum likelihood for p, again following the derivation in class. **Note: The algebra for this becomes complicated. You can solve this with any program you like, including Mathematica or Wolfram Alpha. Alternatively you will get full credit by simply setting up the problem in the correct way and detailing the main logical steps, without carrying out the algebra. Part of Applied Mathematics is knowing WHEN a calculation is to hard to do expeditiously!**
3. Attached to this homework is data that we have generated for 44 such synthetic competitions. Each competition reports the number of games the losing team wins. From this data, use your derivation for MLE to compute your estimate for p.
4. Additionally, following the code we used in class, carry out a numerical calculation of the maximum likelihood estimate for p and compare to the analytical derivation.

Problem 2: Variants on Mostellar (*10 points*)

1. Mosteller analyzed 44 world series and determined that $p = 0.65$. Take this value of p and carry out a Monte Carlo simulation of 44 different world series (each with at most 7 games), and compare the distribution of the number of games that the losing team wins to Mostellar's results. Now repeat this exercise but with 4400 different world series. Does the fraction of games that the losing team won change significantly by changing the number of Series played? Were Mosteller's measurements for the first fifty years of the twentieth century lucky?
2. In 1976, baseball players were given the right to become free agents, which meant that they could switch teams much more easily. Analyze the probability that the better team wins the world series after 1976, using the colab from class with the data loaded. If it does not hold, please formulate a hypothesis of what a new model might be.
3. In the paper, Mosteller also discussed a model in which p changed from year to year. Modify the code from class to see if this makes much of a difference. A simple way to do this is to add a small random number to p so that it changes from game to game.
4. Identify some reason that you are angry at Mosteller's Model. Think about how you would change the model to assuage your anger.

Problem 3: Adding more features to compute p (*10 points*)

In Problem 1, you estimated the probability p of the better team winning the series. Now, let's explore how additional features in the data could enhance the accuracy of p prediction. In the lecture 2 notebook, we introduced a dataset from Major League Baseball that includes a set of features about teams that go beyond games won and lost, including strikeouts, double plays and so forth. Please look at the notebook and examine how it works. The idea is that **the probability is determined by the properties of the teams, not from performance data alone**. The goal of this problem is to use this data to create a simple model that predicts the probability a team wins the world series.

1. Explain what the code in the notebook does. Include in your explanation the following concepts:
   (a) What does the **dropna** command do in the line transforming the features?
   (b) What is the point of the dataset split into train and test?
   (c) What is the logistic regression fit predicting?
   (d) Explain accuracy, confusion matrix,precision and recall?
2. Create and explain the feature importance plot, identifying the features that influence the probability of the better team winning. Does this make sense?
3. Invent new features and find out if they are more informative for predictions. [Hint: think about some of the stats you hear about when people discuss baseball, e.g. batting average, winning streaks, run differentials....]. Explain why you've chosen these features and transformations and see if the turn out to be important.
4. How could you use this type of analysis to improve Mostellar's model? Note: Just propose ideas, you do not need to carry them out

Problem 4: Simulating a tournament (*10 points*)

Consider a tournament with 32 teams. You are given a draw in the file

`game_results_am115_PS1.csv`

1. From the previous games, estimate p_{ij}, the probability that team i will beat team j in a given game.
2. Given your estimates of p_{ij}, carry out simulations of the draw. For each team, predict the probability that it reaches the semi-finals; the finals and is the tournament champion. Thus you should create a matrix W_{ij}, where the first index labels the team, and the second index W_{i0} is the probability the team makes the semi finals; W_{i1} is the probability that the team makes the finals and W_{i2} is the probability that the team makes the finals.
3. Upload your predictions onto our AM115 Kaggle competition 1 website. We will compare your predictions to the true answer and rank the entries. Note that in this case we know the true answer because we *synthesized* the games for an assumed P_{ij} matrix, and we can thus simulate the tournament probabilities with this matrix.
4. In the first project, one of the directions you can choose is to repeat this exercise but for a REAL sports competition with data – eg the US Open, March Madness, NFL Season, and so forth. Write a brief summary of what you might do when you do this for real, and list some of the challenges.

Problem 5: Apply these ideas to a tennis match (*10 points*)

Use the mathematical formalism that we have introduced to write out a series of steps to create a mathematical model for a tennis match. *You are only required to write out the series of steps, not implement the steps.*

1. Make a model for an individual game, where there is a probability that a player wins a point is a parameter. Note that whoever serves typically has a higher probability of winning a point. Note that there is a serious complication over the world series, which is that a player must win by two points. How can you deal with this?
2. Given the probability that a player wins a game, turn this into the probability that they will win a set. Assume tiebreakers are not allowed, and just assume that (like in the case of the game) that a player must win by two games.
3. A match is 2 out of 3 sets.

Extending these ideas and fitting them to actual data would be a great first miniproject.