
Please provide complete and well-written solutions to the following exercises.

Due October 26, 12PM noon PST, to be uploaded as a single PDF document to Gradescope.

Homework 5

Exercise 1. I believe that the number of home runs hit by an MLB baseball player in a single season satisfies a Poisson distribution with some unknown parameter $\lambda > 0$. In this exercise, let's try to find the parameter $\lambda > 0$ that best fits the data, using whatever estimation method you want (e.g. MLE is fine).

Here the data can be found from:

<http://seanlahman.com/download-baseball-database/>

I recommend using the 2020 Version, comma delimited version. The data is in a zip file, and home run data can be found in `Core` then `batting.csv` then the column HR.

After fitting the Poisson distribution to the data, compute the total variation distance of the data from the fitted Poisson distribution. If P, Q are two probability laws on e.g. the positive integers, then the total variation distance between P and Q is

$$\|P - Q\|_{\text{TV}} := \frac{1}{2} \sum_{k=0}^{\infty} |P(k) - Q(k)|.$$

Here P would be the fitted Poisson distribution, and Q would be the probability distribution corresponding to the data. If $\|P - Q\|_{\text{TV}}$ is close to 0, then the Poisson distribution that you found fits well to the data. If $\|P - Q\|_{\text{TV}}$ is far from 0 (perhaps close to 1), then the Poisson distribution that you found does not fit the data well.

Try to answer the same question as above for the number of made 3 point shots among 2020 WNBA players. Data can be found here:

[\(this link\)](#).

The data we are particularly interested in is the column 3P.

Exercise 2. Wikipedia has a list of best selling video games with at least 10 million units sold, sorted by the number of units sold:

[\(this link\)](#)

This is an open-ended question, related to this list.

Plot a histogram of the player counts of this list of games (i.e. the second column of the table). Does this histogram resemble any particular distribution? If so, try to fit that

distribution to the data, as in the previous question, and use the total variation distance as a measure of goodness of fit.

Exercise 3. Suppose you flip a coin 1000 times, resulting in 560 heads and 440 tails. Is it reasonable to conclude that the coin is fair (i.e. it has one half probability of heads and one half probability of tails)? Justify your answer. (Hint: if the coin is fair, then with what probability will you observe at least 560 heads being flipped? That is, is it a rare observation?)

Exercise 4. Suppose the number of typos in my notes in a given year follows a Poisson distribution. In the last few years, the average number of typos was 15, and this year, I had 10 typos in my notes. Is it reasonable to conclude that the rate of typos has dropped this year? Justify your answer. (Hint: if the Poisson random variable X has a mean of 15, then with what probability will you observe that $X \leq 10$? That is, is it a rare observation?)