Harvard John A. Paulson
School of Engineering
and Applied Sciences

Michael P. Brenner

Homework 3

Extreme Value Distributions, Dimensional analysis and Time Series

Issued:	October 28, 2025
Due:	November 13, 2025

Problem 1: *Extreme Value Distributions*

In class, we discussed extreme value distributions and demonstrated that, for both Gaussian and exponential distributions, the extreme value distributions have the same shape (up to differences in mean and variance).

1. Redo this calculation computationally, following the class/section derivation for the powerlaw distribution with a heavy tail distribution (use scipy.stats.pareto, we suggest $b > 4$ as mentioned in the Lecture notebook) as well as any other distribution you choose to use that is a method in np.random or scipy.stats. Determine whether the shape of the distribution has the same universality property we described in class.

2. Let's now apply the extreme value distribution to actual data. We are providing data about the distribution of running times in the 100 meter dash. Compare the distribution of fastest times to the extreme value distribution. Does it agree? If not, could you hypothesize what the issue might be?

3. Now look at the racing times for Usain Bolt. Can you make any statements about whether his times are consistent with the extreme value distribution?

4. Find another example of a dataset where extreme values occur, and compare it to the extreme value distribution. A prize goes to the most creative solution!

Problem 2: *McMahon's Rowers, revisited*

In class we discussed McMahon's argument for rowers, which is an example of intelligent estimation. We are providing a dataset of rowing speeds in a python notebook taken from the 2022 and 2023 world championship.

1. Please analyze this data and decide whether it obeys the $N^{1/9}$ law from McMahon.
2. List factors in a rowing race that could mess up the basic argument.

Problem 3: *Estimation*

Let's try two estimation questions for yourself. Your goal here is to invent an intelligent estimate to come up with an answer based on facts that you might know.

1. How many total hours do Americans spend commuting to and from work in one year?
2. How many cumulative miles of human hair grow in the United States every 24 hours?

Problem 4: *Covid Epidemiology*

Our discussion of Covid epidemiology included both exploration of the enormous datasets underlying covid as well as exploration of covid models.

1. Using the Lecture 13 notebook, make a plot of the confirmed covid cases over time from the county or region where you grew up.

2. At the same time, plot mobility information for that region. Comment on how people responded to the pandemic, and whether you think case numbers had anything to do with people's behaviors.

3. The list of columns also includes search trends, for example 'search trends common cold'. Figure out if this data exists for your region and comment on its correlation with covid.

4. Lecture 14 notebook implements an SIR model and shows you how to fit it to data. Fit the data from your region for confirmed cases to the SIR model, using this code. How good is the fit?

5. Participate in a Kaggle competition for time series prediction, focusing on predicting COVID-19 cases on a synthetic dataset. The link to the competition will be posted on the Canvas assignment page soon. You do not need to submit your prediction CSV with your pset, as this will be submitted directly to Kaggle. However, you must include a brief description of your strategy, implementation, and approach in your answers for this problem.