Problem Set #1: Probability and Statistics

For this problem set, you will use attached csv files containing datasets: `Births.csv` and `Chicken-Weights.csv`. You can find the necessary functions in `numpy` and `scipy.stats`.

1. Baby boom! 44 babies were born in a single day in a single small hospital in Australia in 1997. Various data about each of these births has been recorded. Your goal is to determine which of the four great distributions in Biology fits these data.

    a. Write some code to read the gender (1 = female, 2 = male), weight, and time after midnight for each baby.
    b. Does the number of babies of each gender fit one of the four great distributions? Which one? **Explain your reasoning and assumptions.**
    c. Plot the distribution of weights of the babies. Which of the four great distributions do they fit? **Explain your reasoning and assumptions.**
    d. Plot the number of births per hour (for each one-hour period starting at midnight). Does this fit one of the four great distributions in Biology? **Explain your reasoning and assumptions.**
    e. Calculate and plot the distribution of waiting times between each birth. Does this fit one of the four great distributions in Biology? **Explain your reasoning and assumptions.**

2. As an experiment, six independent groups of chickens were fed six different types of feed. The chickens were weighed individually after eating their new diet, as documented in `Chicken-Weights.csv`. Your goal is to figure out if the weight of these chickens is influenced by their diet.

    a. Write some code to read the weights for each type of feed.
    b. Compute the means and variances of the weights for each type of feed. **Your output should be a list of means and a list of variances.**
    c. The coefficient of variation (CV) is defined as the standard deviation divided by the mean of a dataset. Write your own function in Python to calculate the CV of an array of numbers.
    d. Using the function you just wrote, calculate the coefficient of variation of the weights for each type of feed. **Your output should be a list of CVs for each feed type**.
    e. What is the CV telling you about the variability in the weights of the chickens? **Please explain your answer in a few sentences.**
    f. Discuss two things you would have done differently, if you had designed this experiment yourself (**a few sentences each**).

3. Random number generation and bootstrapping.

    a. Write some code to generate the probability distributions for the baby weights, number of births per hour, and the distribution of waiting times from problem #1. Make plots of each of these. *Hint:* in order to do the simulations, you may want to first calculate the means and standard deviations of the distributions.

b. Using bootstrapping, estimate the mean, standard deviation, and variance of the number of births per hour and the distribution of waiting times between each birth.
c. How do your bootstrapping estimates compare to your expectations based on the identities of these distributions?
d. How many resamplings did you use for bootstrapping, and why?