CS114 Final Deep Dive

April 22, 2022

Problem 1

1. First we can find the probability that the certain year is a record high or low. We can conclude that there is an equal probability for any year before and up to the current one to be record high or low. For example, on the interval from 2001 and 2005, there is an equal chance that any of the 5 years are recrod highs or lows. We can make this conclusion because the average annual teperature were samples i.i.d. from a continuous distribution. Thus, in general terms, if n years have passed, the probability of any year being record high or low is $\frac{1}{n}$. To find the expected number of record highs or lows, we can simply sum the probaility of one year to be a record high or low for all 100 years according to the expected value formula:

$E(High) = E(Low) = \sum_{i=1}^{n} \frac{1}{n}$ - expected number of years that are record high. The same formula is applied to find the expected number of record low years.

Now we see that the expression resembles harmonic series, we can use its property to compute this value:

$E(High) = E(Low) = \sum_{i=1}^{n} \frac{1}{n} = ln(n) + \lambda = ln(100) + 0.577 = 5.182$ - expected number of record high year (or record low years)

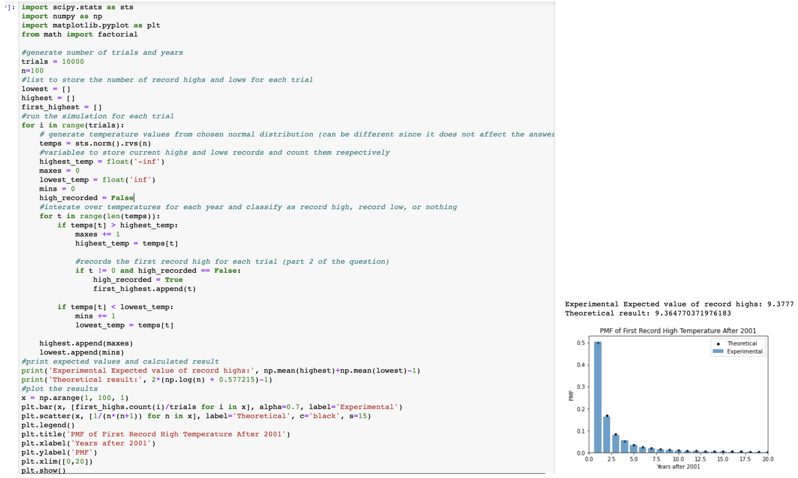
To find the exected value of years being either high or low, we simply multiply the expected value above by 2 (because the expected values for low and high are identical). We also need to subtract 1 because the first year will be counted as both record high and low:

$E(High + Low) = 2 * 5.182 - 1 = 9.36$ years.

2. We need to find the probabillity that the number of years required to get to a new record high after 2001 is greater than a certain value, that is $P(N > n)$. The probability that the number of years to wait is greater than a certain value n is the same as the probabiity that not a single year after 2001 up to and including year 2001+n have been record high. The probability of a single year not being a record high can be found as a complement of a single year being a record high, found above:

$P(H^c) = 1 - P(H) = 1 - \frac{1}{n} = \frac{n-1}{n}$, where $P(H)$ is the probability of a year being record high.

Now, all we need to do is to multiply all these probabilities together to find the probability that all years up



```
import scipy.stats as sts
import numpy as np
import matplotlib.pyplot as plt
from math import factorial

#generate number of trials and years
trials = 10000
n=100
#list to store the number of record highs and lows for each trial
lowest = []
highest = []
first_highest = []
#run the simulation for each trial
for i in range(trials):
    # generate temperature values from chosen normal distribution (can be different since it does not affect the answer
    temps = sts.norm().rvs(n)
    #variables to store current highs and lows records and count them respectively
    highest_temp = float('-inf')
    maxes = 0
    lowest_temp = float('inf')
    mins = 0
    high_recorded = False
    #interate over temperatures for each year and classify as record high, record low, or nothing
    for t in range(len(temps)):
        if temps[t] > highest_temp:
            maxes += 1
            highest_temp = temps[t]

            #records the first record high for each trial (part 2 of the question)
            if t != 0 and high_recorded == False:
                high_recorded = True
                first_highest.append(t)

        if temps[t] < lowest_temp:
            mins += 1
            lowest_temp = temps[t]

    highest.append(maxes)
    lowest.append(mins)
#print expected values and calculated result
print('Experimental Expected value of record highs:', np.mean(highest)+np.mean(lowest)-1)
print('Theoretical result:', 2*(np.log(n) + 0.577215)-1)
#plot the results
x = np.arange(1, 100, 1)
plt.bar(x, [first_highs.count(i)/trials for i in x], alpha=0.7, label='Experimental')
plt.scatter(x, [1/(n*(n+1)) for n in x], label='Theoretical', c='black', s=15)
plt.legend()
plt.title('PMF of First Record High Temperature After 2001')
plt.xlabel('Years after 2001')
plt.ylabel('PMF')
plt.xlim([0,20])
plt.show()
```

```
Experimental Expected value of record highs: 9.3777
Theoretical result: 9.364770371976183
```

Figure 1: Simulation

to n were not record high since we are interested in all these events happening at the same time and each year's temperature is independent:

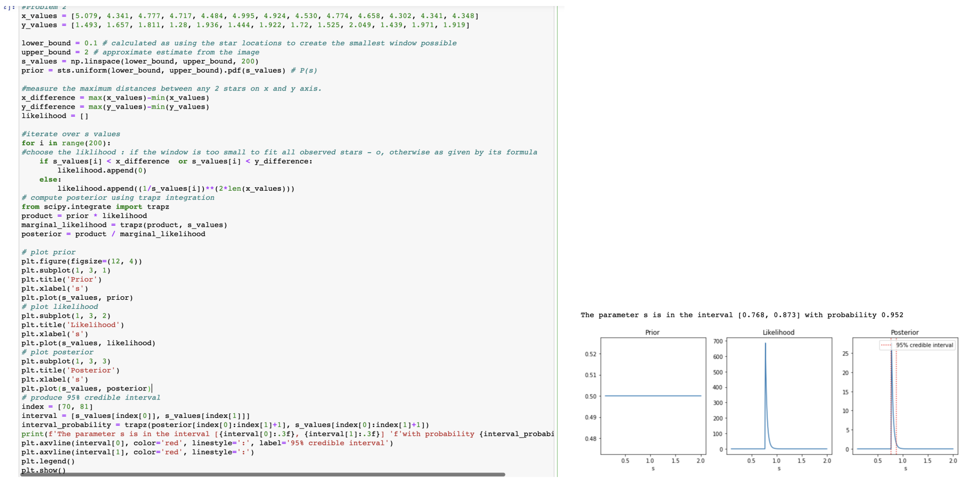
$P(N > n) = \prod_{i=2}^{n} \frac{n-1}{n} = \frac{1}{2} \cdot \frac{2}{3} ... \frac{n-1}{n} = \frac{(n-1)!}{n!} = \frac{1}{n}$

To find the PMF, we can use its definition:

$f(N) = P(n+1 > N > n) = P(N > n) - P(N > n+1) = \frac{1}{n} - \frac{1}{n+1} = \frac{1}{n(n+1)}$

3. If we look at the figure 1 , we can see that the bars (experimental result) perfectly align with the black dots (theoretical result), supporting our results.

4. We generated a bar plot for the probabilities of first record high temperatures to happen after 2001 in a given number of years. We can create the same distributuon for other periods of time before current , such as 1801-1900, 1901-2000. Then, we need to compare all the prvious distribution and the current one to find the pattern. If the distributions for every consequtive 100 year interval become more skewed to the left, it would mean that the probability of the first record high happening faster after a certain year gets higher, indicating that temperatures rise as the time goes (global warming is happening). We should also keep in mind that the model might be flauded in case it does not account for the cyclical nature of temperatures. Another approach would be apply model selection. We can use Bayes equation to calculate posterior $P(M_1|data)$ for the model we already have , which does not account for temperature change over time (increase for globl warming), and create another model that does account for that and find its posterior $P(M_2|data)$. Whichever posterior is greater, that model bests explains the data. If it is the second model, it would provide evidence for global warming.



```
#Problem 2
x_values = [5.079, 4.341, 4.777, 4.717, 4.484, 4.995, 4.924, 4.530, 4.774, 4.658, 4.302, 4.341, 4.348]
y_values = [1.493, 1.657, 1.811, 1.28, 1.936, 1.444, 1.922, 1.72, 1.525, 2.049, 1.439, 1.971, 1.919]

lower_bound = 0.1 # calculated as using the star locations to create the smallest window possible
upper_bound = 2 # approximate estimate from the image
s_values = np.linspace(lower_bound, upper_bound, 200)
prior = sts.uniform(lower_bound, upper_bound).pdf(s_values) # P(s)

#measure the maximum distances between any 2 stars on x and y axis.
x_difference = max(x_values)-min(x_values)
y_difference = max(y_values)-min(y_values)
likelihood = []

#iterate over s values
for i in range(200):
#choose the liklihood : if the window is too small to fit all observed stars - o, otherwise as given by its formula
    if s_values[i] < x_difference  or s_values[i] < y_difference:
        likelihood.append(0)
    else:
        likelihood.append((1/s_values[i])**(2*len(x_values)))
# compute posterior using trapz integration
from scipy.integrate import trapz
product = prior * likelihood
marginal_likelihood = trapz(product, s_values)
posterior = product / marginal_likelihood

# plot prior
plt.figure(figsize=(12, 4))
plt.subplot(1, 3, 1)
plt.title('Prior')
plt.xlabel('s')
plt.plot(s_values, prior)
# plot likelihood
plt.subplot(1, 3, 2)
plt.title('Likelihood')
plt.xlabel('s')
plt.plot(s_values, likelihood)
# plot posterior
plt.subplot(1, 3, 3)
plt.title('Posterior')
plt.xlabel('s')
plt.plot(s_values, posterior)
# produce 95% credible interval
index = [70, 81]
interval = [s_values[index[0]], s_values[index[1]]]
interval_probability = trapz(posterior[index[0]:index[1]+1], s_values[index[0]:index[1]+1])
print(f'The parameter s is in the interval [{interval[0]:.3f}, {interval[1]:.3f}] 'f'with probability {interval_probabi
plt.axvline(interval[0], color='red', linestyle=':', label='95% credible interval')
plt.axvline(interval[1], color='red', linestyle=':')
plt.legend()
plt.show()
```

```
The parameter s is in the interval [0.768, 0.873] with probability 0.952
```

Figure 2: Parameter Estimation

Problem 2

To find the value of s, we can provide its estimate using 95% credible interval of Bayesian parameter estimation. We need to use the Bayesian equation, which states that posterior is equal to the product of the prior and likelihood, devided by the marginal probability. Since we have no prior knowledge on the size of the window we can assume it follows uniform distribution $Uniform(a, b)$ since there is no reason to prefer one size over the other. The parameters indicate the interval in which the size is distributed : it must be bigger than a, which is 0.01 meters because we can assume that this is the smallest the window can be, and smaller than b, which is 2 meters -the largest the window can be (assumption). We can also look up the likelihood of the uniform distribution, which is

$\prod_{i=1}^{n} \frac{1}{(b-a)^2}$ (we sqaure the denominator to get the area of the window since the problem is in 2 dimensions).

Since a and b are the low and upper bounds of s, their different would provide the estimate for s, making the expression simplify to

$\prod_{i=1}^{n} \frac{1}{(b-a)^2} = \frac{1}{s^{2n}}$

We should note that this expression is only valid for values of s that are bigger than the maximum distance between any 2 stars. Otherwise the likelihood would be 0 and we would not be able to see some of those stars that we see.

From you can see the computation of posterior and 95 percent credible interval. We see that the posterior probability (given the data set) that the true parameter value is in the [0.768, 0.873] interval is 95.2%.

We did not include the center of the square (x,y) into the solution. If we were to do so, we would need to find the posterior $P(s|data)$, but first prior $P(x, y, s)$, likelihood $P(data|x, y, s)$, and marginal $P(data)$. To find the prior, we assume that x, y, and s are independent, meaning it can be written as $P(x)P(y)P(s)$. The variables are given by the following distributions (assuming arbitrary large values for the wall size):

$x \sim Uniform(0, 100)$

$y \sim Uniform(0., 100)$

$s \sim Uniform(0.1, 2)$

Now we can multiply their PDFs with corresponding parameters:

$P(x)P(y)P(s) = \frac{1}{100-0} \cdot \frac{1}{100-0} \cdot \frac{1}{2-0.01} = \frac{1}{19000}$

The formula for likelihood was already defined above, but a new conditions arises that needs to be satisfied: $x - \frac{s}{2} \leq star_x \leq x + \frac{s}{2}$ and $-\frac{s}{2} \leq star_y \leq y + \frac{s}{2}$. This condition implies that the boundaries for the window are given as half the size of the windowm size in both direction from the center point for both axises, and the visible star must be within these boundaries.

To find the marginal probability, we need to intergate the product of likelihood and prior over x, y , and s, using the boundaries as defined by their respective distributions above.

And now, to find the posterior, we simply apply Bayes equation to find $P(x, y, s|data)$, and then integrate with respect to x and y to be left only with s and find $P(s|data)$.

References

[1] Scheffler, K. (n.d.). CS114 Session 14 - [8.1] Synthesis: Parameter estimation. Forum. Retrieved April 20, 2022, from https://forum.minerva.edu/app/courses/2109/sections/8431/classes/63120

[2] Wikimedia Foundation. (2022, March 15). Harmonic series (mathematics). Wikipedia. Retrieved April 20, 2022, from https://en.wikipedia.org/wiki/Harmonic_series_(mathematics)

[3] Zach. (2021, March 2). Maximum likelihood estimation (MLE) for a uniform distribution. Statology. Retrieved April 20, 2022, from https://www.statology.org/mle-uniform-distribution/