Problem 1: Extreme Temperatures (#Probability, #ModelSelection)

Let $X_1, X_2, \ldots, X_{100}$ be the annual average temperature in Berlin in the years 2001, 2002, …, 2100, respectively. Assume that average annual temperatures are sampled i.i.d. from a continuous distribution.

(Note: For this problem, we assume the temperature distribution doesn't change over time. With global warming, this is not a good assumption.)

A year is a record high if its average temperature is greater than those in all previous years (starting with 2001), and a record low if its average temperature is lower than those in all previous years. By definition, the year 2001 is both a record high and a record low.

1. ***In the 21st century (the years 2001 through 2100, inclusive), find the expected number of years that are either a record high or a record low.***

From the problem, we know that average annual temperatures are sampled independently from the same distribution. In finding the expected value for total record highs and lows over 100 years, we need to find the chance of a record occurring. As these variables are random, each year has a probability of $P = \frac{1}{n}$ of being either a record high or a record low, *n* being the total number of years in the "pool" of comparison. Since we are evaluating records as either "record high" or "record low", we need to identify two indicator variables to equal 1 when the j-th year is a record:

$$H_j = \text{record high (H)}$$

$$L_j = \text{record low (L)}$$

Due to the random pull from the distribution, both record highs and record lows are equally probable. Thus:

$$P(H_j = 1) = P(L_J = 1)$$

Similarly, normal-temperatred years are equally probable:

$$P(H_j = 0) = P(L_j = 0)$$

Knowing that every individual year has a probability of $P = \frac{1}{n}$ of being either a record high or record low, we can adapt this to our indicator variable and j-th year:

$$P(H_j = 1) = \frac{1}{j}$$

$$P(L_j = 1) = \frac{1}{j}$$

Now that we have the probability of any j-th year, we can find the expected amount of record years over a century using the fundamental bridge and linearity of expectation (where TH is our total number of record high and TL is our total number of record low):

$$E[TH] = \sum_{j=1}^{100} P(H_j = 1)$$

$$E[TH] = \sum_{j=1}^{100} \frac{1}{j}$$

Using computational tools:

$$E[TH] \approx 5.19$$

This would look similar to the total number of record low:

$$E[TL] = \sum_{j=1}^{100} P(L_j = 1)$$

$$E[TL] = \sum_{j=1}^{100} \frac{1}{j}$$

$$E[TL] \approx 5.19$$

In total, we would expect a total of:

$$5.19 + 5.19 \approx 10.38$$

$$10.38 = 10$$

Counting in complete years, our expected value is 10 total record years.

```
#PROBLEM 1: EXTREME TEMPERATURES

#1.1: Expected Value
#using: https://localcoder.org/how-to-do-a-sigma-in-python-3

#Using Python to solve the sigma problem for 1.1
from functools import reduce
result = reduce(lambda a, x: a + x**(-1), [0]+list(range(1,100+1)))
print(result)
```

```
5.187377517639621
```

2. ***Let N be an r.v. representing the number of years required to get a new record high after the year 2001. Find*** $P(N > n)$ ***for all positive integers*** n***, and use this to find the PMF of*** N***.***

Approaching this problem, we are first looking for the probability of reaching a record high as time goes on, or as $n + 1$. $N > n$ is required to reach a record high after 2001, where the Nth year is hotter than all previous n years since 2001. Intuitively, we would expect this probability distribution to begin high and quickly jump downwards, as 2001 will always be a record high ($n = 0$), 2002 has a 50/50 chance of being a record high ($n = 1$) (either higher or lower than 2001's temperature), and so on. Finding $P(N > n)$, we can begin with positive integer 1:

$$P(N > 1)$$

Meaning that 2002, the 1st year after 2001, was not a record high. Using indicator variable H_j once more to represent the probability that the j-th term is a high, we can model that the j-th term is not a high using its compliment:

$$P(N > 1) = 1 - P(H_2 = 1)$$

As we stated earlier, we know that this is $\frac{1}{2}$, as 2002 has an equal chance of being above or below 2001's record temperature:

$$P(N > 1) = 1 - P(H_2 = 1) = \frac{1}{2}$$

This can continue with *n* positive integers, multiplying each compliment probability, like the one above:

$$P(N > n) = \prod_{j=1}^{n} 1 - P(H_{j+1} = 1)$$

Simplified, using our known probability of the j-th year to be a record of $\frac{1}{j}$, adapting to accommodate $j + 1$:

$$P(N > n) = \prod_{j=1}^{2} \frac{j}{j+1}$$

Using our *n*th term and simplifying the Pi notation:

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

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

Knowing $P(N > n)$, we can find the CDF of *N* by taking the complement:

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

$$f(n) = \frac{n}{n+1}$$

The PMF, defined for *n* values greater than 1, can be found from the CDF above by subtracting $f(n - 1)$:

$$F(n) = f(n) - f(n - 1)$$

Where $f(n - 1)$ is:

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

Plugged in:

$$F(n) = \frac{n}{n+1} - \frac{n-1}{n}$$

$$F(n) = \frac{n^2}{n(n+1)} - \frac{(n+1)(n-1)}{n(n+1)} = \frac{n^2 - n^2 + 1}{n(n+1)}$$

Solution:

$$F(n) = \frac{1}{n(n+1)}$$

This is applicable to all *n* values of 1 or higher.

3. ***Write a short simulation to check your answers to parts (1) and (2).***

```
#1.3: Simulation
import matplotlib.pyplot as plt
import scipy.stats as sts
import numpy as np
import random

#checking 1.1
trials = 100000
rec = []
Nrec = []

for t in range(trials):
    #generating random values for X_1 to X_100
    annual_temp = [random.random() for _ in range(0, 100)]
    #2001 is a record, this is our starting point
    record_H = 1
    current_H = annual_temp[0]
    for i in range(1, len(annual_temp)):
        if annual_temp[i] > current_H:
            record_H += 1
            if record_H == 2:
                Nrec.append(i-1)
            current_H = annual_temp[i]
    #setting our new high :)
    rec.append(record_H)

print("Simulated Record Highs:", sum(rec)/trials)
print("Theoretical Record Highs: 5.187")
```

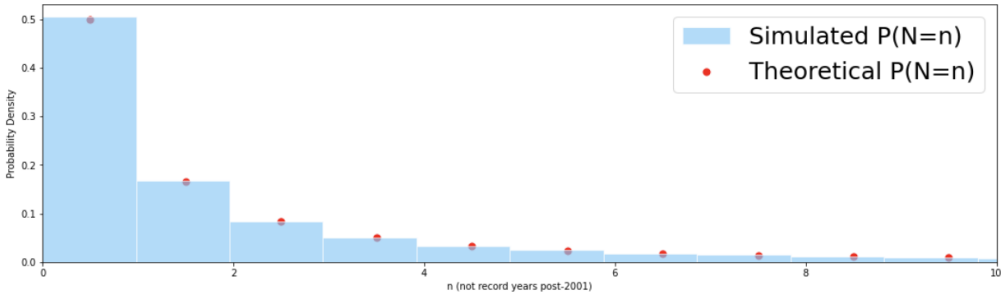
```
#checking 1.2
x_ax = np.arange(1,100,1)
y_ax = []

#PMF for theoretical
for x in x_ax:
    y_ax.append(1/(x*(x+1)))

weights = np.ones_like(Nrec)/len(Nrec)

#plotting figure
plt.figure(figsize = (18,5))
plt.hist(Nrec, bins=100, weights=weights, color='lightskyblue',
         alpha=0.7, edgecolor='white', label='Simulated P(N=n)')
plt.scatter((x_ax-0.5), y_ax, s=50, color='red',
            label='Theoretical P(N=n)')
plt.xlabel('n (not record years post-2001)')
plt.ylabel('Probability Density')
plt.xlim((0, 10))
plt.legend(prop={'size':25})
plt.show()
```

```
5.187377517639621
Simulated Record Highs: 5.19056
Theoretical Record Highs: 5.187
```



4. ***Explain how you could use this model to determine whether or not global warming is really happening.***

Using this model, we can determine a null and alternative hypothesis. The PDF above can be our null because it uses a distribution to randomly decide annual temperature. If global warming were significantly affecting the temperature, then we would expect there to be a deviation from this model. For example, we can compare the number of expected record highs under this distribution and compare it to a historically accurate PMF of temperature highs. Using a significance level, or test statistic, of 0.05 and testing, we can either accept or reject the null hypothesis.