Data Analysis with R and Python - Sample Questions

1. Using the `Pulse` variable in the `NHANES` dataset, test the hypothesis that the mean pulse rate is 72 for **black males** between the ages of 20 and 60.
2. Find 90% prediction intervals for the `Pulse` variable in the `NHANES` dataset, for all combinations of `Race1` and `Gender`, for people between the ages of 20 and 60. Display the resulting intervals using a suitable plot.
3. Would you recommend using a power transformation of `Pulse` instead of the untransformed `Pulse` for the previous two questions? Justify your answer. What tools would you use to decide which power transformation to use, if any?
4. Suppose (X, Y) is uniformly distributed over the unit disk S given by

$$S = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$$

   (a) Briefly describe how you would simulate from the distribution of (X, Y).

   (b) Using the approach described in (a), estimate the mean and variance of $Z = \sqrt{X^2 + Y^2}$.

5. Consider a room with one hundred people — forty men and sixty women.

   (a) If ten people are selected from the room, find the probability that exactly six are women. Calculate this probability with and without replacement and compare the results.

   (b) If fifty people are selected from the room, find the probability that exactly thirty are women. Calculate this probability with and without replacement and compare the results.

6. Suppose $X_1, X_2, \ldots, X_n$ are random variables with a common continuous distribution. To test the hypothesis H_0 that the observations are *independent*, a statistician proposes to use the test statistic $T = \sum_{i=2}^{n} M_i$, where

$$M_i = \begin{cases} 1 & \text{if } X_i > \max(X_1, X_2, \ldots, X_{i-1}) \\ 0 & \text{otherwise.} \end{cases}$$

H_0 is rejected if $T \geq c$, where c is chosen such that $P(T \geq c) \approx 0.05$.

(a) Find a resonable value of c when $n = 70$.

(b) The variable `maxTemp` gives the maximum temperature (in Celcius) recorded in an Indian city in each year from 1951 to 2020 (inclusive).

```
maxTemp <-
    c(35.25, 36.61, 36.02, 35.50, 35.94, 37.05, 36.17, 36.25, 36.69,
      35.93, 36.75, 35.78, 35.69, 37.03, 35.95, 36.92, 36.25, 35.63,
      36.86, 36.15, 36.03, 36.41, 37.97, 37.22, 36.66, 37.20, 36.71,
      36.74, 36.69, 37.72, 36.31, 36.92, 37.67, 36.46, 36.79, 36.87,
      37.39, 36.94, 37.00, 37.65, 37.11, 37.41, 38.48, 36.96, 36.63,
      37.74, 36.53, 38.42, 36.99, 38.07, 36.53, 36.97, 37.96, 37.56,
      37.97, 37.72, 36.68, 36.47, 37.65, 37.67, 35.47, 36.53, 37.01,
      36.75, 36.43, 38.93, 37.63, 36.54, 37.53, 36.53)
```

Draw a scatter plot of `maxTemp` against year. Is there any systematic (increasing or decreasing) pattern in the plot? Calculate T for this data. Would you reject H_0 based on T and the cutoff c you obtained in (a)?