Probability Questions posted by Undergraduate and Master Research Opportunities @ NUS Synthetic Biology

Raven

January 2025

Question 1

The scenario described involves paying \$1 to play a machine that dispenses ice cream that costs \$6. In this case, model the number of independent trials required to get the first success,which is a geometric random variable,$X \sim \text{Geom}(p)$.

The probability mass function (PMF) of the geometric distribution is given by:

$$P(X = k) = (1 - p)^{k-1}p$$

where:

- X is the number of trials (or tries) until the first success,
- p is the probability of success on a single trial (i.e., the probability of winning a tub of ice cream),
- k is the number of trials.

Estimating Parameters

Assume that each trial has an equal chance of success and that the ice cream machine has a fixed number of tubs available for dispensation.

The number of possible trials, T

$$T = \frac{\text{Total Money in Pocket}}{\text{Cost per play}}$$

The probability of success per trial, p

$$p = \frac{1}{\text{Number of possible trials}}$$

For a total number of possible trials T, the probability of success p is:

$$p = \frac{1}{T}$$

And the expected number of first success is:

$$E(X) = \frac{1}{p} = T$$

```python
import numpy as np

cost_per_play = 1.0
money_in_pocket = 20.0 # assume 20 dollars in pocket

# Total number of trials (based on available money)
T = int(money_in_pocket / cost_per_play)

# Probability of success (p) for each trial
p = 1 / T

expected_tries = 1 / p
print(f"Expected number of first success: {expected_tries}")

```

Question 2

For M&Ms, each of the expected proportions are:

Color	Proportion (Cleveland)	Proportion (Hackettstown)
Red	0.131	0.125
Orange	0.205	0.250
Yellow	0.135	0.125
Green	0.198	0.125
Blue	0.207	0.250
Brown	0.124	0.125

Chi-Square Goodness-of-Fit Test

1. **Set Null and Alternative Hypotheses:**
    - Null Hypothesis (H_0): The sample's color distribution follows the proportions from Cleveland.
    - Alternative Hypothesis (H_1): The sample's color distribution do not follow the proportions from Cleveland (but Hackettstown).

2. **Calculate the Observed and Expected Frequencies:**

   To find observed frequency of each color, try shuffle and take a random sample M&Ms, record the total sample size as T and observed frequencies for each color i, as O_i.

   Now given the expected proportions of each color (Red, Orange, Yellow, Green, Blue, Brown):

$$\begin{aligned} p_{\mathrm{r}} &= \text{proportion of M\&M's that are red,} \\ p_{\mathrm{o}} &= \text{proportion of M\&M's that are orange,} \\ p_{\mathrm{y}} &= \text{proportion of M\&M's that are yellow,} \\ p_{\mathrm{g}} &= \text{proportion of M\&M's that are green,} \\ p_{\mathrm{bl}} &= \text{proportion of M\&M's that are blue,} \\ p_{\mathrm{br}} &= \text{proportion of M\&M's that are brown.} \end{aligned}$$

   Assuming the sample comes from Cleveland, the expected frequency of each color i is:

$$E_i = \text{Total Sample Size, T} \times p_i$$

Color	**Observed Frequency**	**Expected Frequency**	**(O-E)^2/E**
Red			
Orange			
Yellow			
Green			
Blue			
Brown			

3. **Conduct a Chi-Square Test:**

   Degrees of Freedom:

$$\mathrm{df} = 6 - 1 = 5$$

   The Chi-Square statistic is calculated as:

$$\sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} \sim \chi_5^2$$

   Where O_i is the observed frequency for color i in the sample and E_i is the expected frequency for that color based on Cleveland.

   Summing up the last column in the previous table, we obtained the Chi-Square statistic, $\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = X$.

4. **Make a Decision:** Compute the p-value using Chi-Square CDF Table for the calculated test statistic X and degrees of freedom 5, which gives:

$$\text{p-value} = P(\chi_5^2 > X)$$

   - If the p-value is less than 0.05 (assuming a 5% significance level), reject the null hypothesis and conclude that the M&M sample likely came from Hackettstown.
   - If the p-value is greater than 0.05, fail to reject the null hypothesis and conclude that the sample likely came from Cleveland.

Remarks

- One major flaw is that no shuffling is mentioned in the article for only one scoop of M&Ms each week.This may result in the distribution being biased towards a certain color that is more concentrated at the top or bottom of the container, thus not representing the actual distribution of colors throughout the entire container.This violates the assumption of the Chi-square test, where each color is assumed independent and randomly selected, which can lead to skewed results.
- One resolution is to take multiple scoops from different corners within the container or by shaking to ensure randomness, record refills and scoops over multiple weeks.