Question 2

Naman Agrawal

2025-01-03

Let the observed counts for the six colors be:

1. X_{Red} : Count of red M&Ms
2. X_{Orange} : Count of orange M&Ms
3. X_{Yellow} : Count of yellow M&Ms
4. X_{Green} : Count of green M&Ms
5. X_{Blue} : Count of blue M&Ms
6. X_{Brown} : Count of brown M&Ms

The total count of M&Ms is:

$$n = X_{\text{Red}} + X_{\text{Orange}} + X_{\text{Yellow}} + X_{\text{Green}} + X_{\text{Blue}} + X_{\text{Brown}}$$

For each plant, the observed data follows a multinomial distribution. The likelihood of observing the counts given the plant is:

$$P(X \mid \text{C}) = \frac{n!}{\prod_{i=1}^{6} X_i!} \prod_{i=1}^{6} p_{i,\text{C}}^{X_i}$$

$$P(X \mid \text{H}) = \frac{n!}{\prod_{i=1}^{6} X_i!} \prod_{i=1}^{6} p_{i,\text{H}}^{X_i}$$

where $p_{i,\text{C}}$ and $p_{i,\text{H}}$ are the probabilities of each color according to the respective plant's distribution. Using Bayes' theorem, the posterior probabilities of the packet being from Cleveland or Hackettstown are:

$$P(\text{C} \mid X) = \frac{P(X \mid \text{C})P(\text{C})}{P(X)}$$

$$P(\text{H} \mid X) = \frac{P(X \mid \text{H})P(\text{H})}{P(X)}$$

Assume equal prior probabilities for the plants:

$$P(\text{C}) = P(\text{H}) = 0.5$$

The denominator $P(X)$ is the marginal likelihood, which can be expressed as:

$$P(X) = P(X \mid \text{C})P(\text{C}) + P(X \mid \text{H})P(\text{H})$$

Thus,

$$P(\mathrm{C} \mid X) = \frac{P(X \mid \mathrm{C})}{P(X \mid \mathrm{C}) + P(X \mid \mathrm{H})}$$

$$P(\mathrm{H} \mid X) = \frac{P(X \mid \mathrm{H})}{P(X \mid \mathrm{C}) + P(X \mid \mathrm{H})}$$

So, now we can simply compare $P(\mathrm{C} \mid X)$ and $P(\mathrm{H} \mid X)$ or equivalently $P(X \mid \mathrm{C})$ and $P(X \mid \mathrm{H})$:

If $P(\mathrm{C} \mid X) > P(\mathrm{H} \mid X)$, conclude the packet likely came from Cleveland. Otherwise, conclude Hackettstown.

Tests

We will perform a chi-square goodness-of-fit test to determine whether the observed proportions of M&M colors match the expected proportions for either the Cleveland or Hackettstown plants. The null hypothesis for each test is that the observed data follows the respective plant's color distribution.

Additionally, we will implement the Bayesian method discussed earlier to calculate posterior probabilities for Cleveland and Hackettstown as potential sources of the observed data. This method uses Bayes' theorem and assumes equal prior probabilities for both plants.

```
# observed data
colors <- c("Red", "Orange", "Yellow", "Green", "Blue", "Brown")
observed_counts <- c(108, 133, 103, 139, 133, 96)
total_count <- sum(observed_counts)

# expected proportions for C and H
cleveland_probs <- c(0.131, 0.205, 0.135, 0.198, 0.207, 0.124)
hackettstown_probs <- c(0.125, 0.25, 0.125, 0.125, 0.25, 0.125)
expected_cleveland <- total_count * cleveland_probs
expected_hackettstown <- total_count * hackettstown_probs

# Chi-square goodness-of-fit test
chisq_test_cleveland <- chisq.test(observed_counts, p = cleveland_probs)
chisq_test_hackettstown <- chisq.test(observed_counts, p = hackettstown_probs)

cat("Chi-Square Test Results for Cleveland:\n")
```

```
## Chi-Square Test Results for Cleveland:
```

```
print(chisq_test_cleveland)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed_counts
## X-squared = 6.074, df = 5, p-value = 0.2991
```

```
cat("\nChi-Square Test Results for Hackettstown:\n")
```

```
##
## Chi-Square Test Results for Hackettstown:
```

```
print(chisq_test_hackettstown)
```

```
##
##  Chi-squared test for given probabilities
```

```
##
## data:  observed_counts
## X-squared = 57.652, df = 5, p-value = 3.711e-11
```

```
# calculate likelihoods
log_likelihood_cleveland <- sum(log(cleveland_probs)*observed_counts)
log_likelihood_hackettstown <- sum(log(hackettstown_probs)*observed_counts)

# assuming equal priors
cat("log Posterior Probability (Cleveland) propto :", log_likelihood_cleveland, "\n")
```

```
## log Posterior Probability (Cleveland) propto : -1271.529
```

```
cat("log Posterior Probability (Hackettstown) propto:", log_likelihood_hackettstown, "\n")
```

```
## log Posterior Probability (Hackettstown) propto: -1296.185
```

Issues

The approach described in the article and my Bayesian approach (assuming equal priors for the presence of each plant) lead to same conclusions. However, the article's methodology lacks robustness in scenarios where the plants have a disproportionate presence. In such cases, the Bayesian approach is more suitable because it can incorporate prior knowledge about the relative contributions of the plants, leading to more accurate posterior probabilities.

Additionally, the article does not consider variability in the number of M&Ms per packet, which can influence observed proportions. Larger sample sizes stabilize proportions due to the law of large numbers, but smaller packets exhibit greater variability, potentially skewing conclusions. Ignoring this variability may lead to biased inferences, particularly when drawing conclusions from small or non-representative samples.

Furthermore, the chi-square goodness-of-fit test has critical assumptions that may not always hold. These include: 1. Single Categorical Variable: The test assumes data originates from one categorical variable (e.g., color). However, the packaging process might introduce dependencies between categories (e.g., color counts across packets may not be independent). 2. Independence of Observations: Observations should be independent, but within-packet counts might exhibit correlations due to manufacturing constraints.