BEE552 Biometry Week 1

Maria Feiler

1/25/2022

My Learning Journey

Over the last week, I participated in Biometry in the following ways:

- I asked / answered *4* questions posed in class.
- I asked *2* questions in Slack.
- I answered *3* questions posed by other students on Slack.
- I came to Heather's office hours: *No*
- I came to Jose's office hours: *No*
- I met with Heather or Jose separately from office hours: *No*

Anything not falling into one of the above categories?

- Facilitated zoom working together session over zoom, but no one showed
- Communicated with Heather over slack to discuss potential bugs in my code

On a scale of 1 (no knowledge) to 10 (complete expert), how would I rate my comfort with R programming after this week?

4

Any topics from last week that you are still confused about?

- I would appreciate a look into the stats a little more. How do you get that huge factorial expression?

Problem Set

```
# Vector with M&M colors in order presented
MMcolors <- c("brown", "yellow", "green", "red", "orange", "blue")

# Create vector with sample counts
mybag <- c(5, 1, 2, 2, 1, 5)
names(mybag) <- MMcolors
```

```
My bag's distribution of M&M colors was:
5 brown
1 yellow
2 green
2 red
1 orange
5 blue
```

Question 1

What is the FORMULA (i.e., an equation) for the probability of obtaining any given combination of colors? Please define all variables used.

The probability of getting n_1 brown, n_2 yellow, n_3 green, n_4 red, n_5 orange, and n_6 blue M&Ms from a bag containing n M&Ms of all available colors is the binomial coefficient times the joint probability of getting brown n_1 times, yellow n_2 times, green n_3 times, red n_4 times, orange n_5 times, and blue n_1 times.

$$P(AnyColors) = \frac{n!}{n_1!n_2!n_3!n_4!n_5!n_6!} p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{n_4} p_5^{n_5} p_6^{n_6}$$

Question 2

a. What is the actual PROBABILITY (i.e., a number, not an equation) of obtaining your combination of colors assuming that all colors are equally likely?

$$p_1 + p_2 + p_3 + p_4 + p_5 + p_6 = 1$$

If all colors are equally likely in the population, then…

$$p_{any} = p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6 = 0.167$$

Using the notation from question 1, the probability of getting the distribution found in my bag (sample of n) in any order from an equally distributed population of M&Ms is…

$$\begin{aligned} P(MyColors) &= \frac{n!}{n_1!n_2!n_3!n_4!n_5!n_6!} p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{n_4} p_5^{n_5} p_6^{n_6} \\ &= \frac{n!}{n_1!n_2!n_3!n_4!n_5!n_6!} p_{any}^{n_1+n_2+n_3+n_4+n_5+n_6} \\ &= \frac{16!}{5!1!2!2!1!5!} (0.167^{5+1+2+2+1+5}) \end{aligned}$$

```
# Calculate probability from list of occurrences and known population
# probabilities
find.prob <- function(x = NULL, probs = NULL){
    # Find binomial coefficient
        # Define top value
        top <- factorial(sum(x))

        # Calculate the base factorial values
        base <- c()
        for (i in 1:length(x)) {base[i] <- factorial(x[i])}

        # Get binomial coefficient
        binom <- top/(prod(base))

    # Find probabilities of each option of that number
        prob <- c()
        for (i in 1:length(x)) {prob[i] <- probs[i]^x[i]}

    # Solve for probability
        result <- prod(binom, prod(prob))
}
```

```
# Create vector with probability of all M&M colors given all colors
# are equally likely
equaldist <- c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)
names(equaldist) <- MMcolors

# Find answer
ans2a <- find.prob(mybag, equaldist)
```

```
My bag's probability given an equal distribution of colors is 0.0001287589.
```

b. What is the PROBABILITY of obtaining your combination of colors given the actual distribution of colors (according to the company): 24% blue, 14% brown, 16% green, 19% orange, 13% red, 14% yellow?

$$\begin{aligned} P(MyColors) &= \frac{n!}{n_1!n_2!n_3!n_4!n_5!n_6!} p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{n_4} p_5^{n_5} p_6^{n_6} \\ &= \frac{16!}{5!1!2!2!1!5!} (0.14^5 0.14^1 0.16^2 0.13^2 0.19^2 0.25^5) \end{aligned}$$

```
# Create vector with probability of all M&M colors given actual
# likelihoods
truedist <- c(0.14, 0.14, 0.16, 0.13, 0.19, 0.24)
names(truedist) <- MMcolors

# Find answer
ans2b <- find.prob(mybag, truedist)
```

```
My bag's probability given the true distribution of colors is 0.0001790201.
```

c. What is the PROBABILITY of obtaining your combination of colors given the actual distribution of colors in your bag of M&Ms?

$$P(MyColors) = \frac{n!}{n_1!n_2!n_3!n_4!n_5!n_6!}p_1^{n_1}p_2^{n_2}p_3^{n_3}p_4^{n_4}p_5^{n_5}p_6^{n_6}$$
$$= \frac{16!}{5!1!2!2!1!5!}(0.3125^5 0.0.625^1 0.1250^2 0.1250^2 0.0625^2 0.3125^5)$$

```
# Create vector with sample proportions
mydist <- c()
for (i in 1:length(mybag)) {mydist[i] <- mybag[i]/sum(mybag)}
names(mydist) <- MMcolors

# Find answer
ans2c <- find.prob(mybag, mydist)
```

```
My bag's probability given my bag's distribution of colors is 0.003076787.
```

d. Comparing 2a-2c, what distribution of colors (equal probability, company-stated probability, sample-based probability) makes your bag of M&Ms more likely to have occurred.

```
# Create vector with answers from 2A, 2B, and 2C
ans2 <- c(ans2a, ans2b, ans2c)
names(ans2) <- c("equal distribution", "true distribution", "my distribution")

# Get the list positions in order from greatest to least
order <- sort.list(ans2, decreasing = TRUE)
```

```
I am most likely to get the distribution from my bag from a population of M&Ms that matches
my distribution of colors.
```

Question 3

Write a short R script to simulate the combinations of colors that would have been possible from your bag of M&Ms assuming the company-stated color distribution (from 2b) and use this script to calculate the probability of obtaining the combination in your bag. There are many ways to solve this problem, but I am looking for you to solve it by sampling hypothetical bags of M&Ms and using that collection of bags to calculate the probability of obtaining your specific bag of M&Ms. [Hint 1: Try the R function 'sample'. Hint 2: Make sure to write your script in a way that makes it possible to also answer Question 4] (Copy and paste your script below.)

Assuming there is an infinite population of M&Ms and sampling does not appreciably change the population's color distribution, I will be sampling with replacement.

```
# Creating pool to sample, setting the colors to be factors and
# removing labels
plaindist <- unname(truedist)
plainMMcolors <- as.factor(unname(MMcolors))

pool <- data.frame(Proportion = plaindist, Color = plainMMcolors)

# Defining a the number of samples to be taken (100000), the number
# of candies to be selected based on my original bag, and the number
# of candy colors
nit <- 100000
nsamp <- sum(mybag)
ncolors <- length(MMcolors)

# Creating dataframe for individual sample iterations
sampit <- list()

# Creating dataframe to store color counts of iterations
allsampit <- data.frame(matrix(ncol = ncolors, nrow = nit))

# Take samples of the colors and store in data.frame
for (i in 1:nit) {
    # Get all sample iterations
        sampit[[i]] <- sample(x = pool[,2],       # Sample colors
                              size = nsamp,       # Number of candies
                              replace = TRUE,     # With replacement
                              prob = pool[,1]     # Define color probs
                              )

    # Collect all iterations into counts data frame
        allsampit[i,] <- table(sampit[[i]])
}

# Run test to get order of color counts and add column names
allsampit[1,] <- table(sampit[[1]])
colnames(allsampit) <- dimnames(table(sampit[[1]]))[[1]]

# Reorder dataframe according to original distribution
allsampit <- allsampit[, MMcolors]

# Determine how many of the samples fit my original bag distribution
samesample <- c()

# Creates list of either 0 (doesn't match) or 1 (matches)
for (i in 1:nit) {
        samesample[i] <- sum(match(paste(allsampit[i,], collapse = ""),
                                   paste(mybag, collapse = ""),
                                   nomatch = 0))
        }
```

```
The probability of obtaining my bag's distribution of colors is 0.00014.
```

Question 4

According to your simulation, what was the most likely combination of colors? What was the probability of getting that most-likely combination?

```
# Get the color counts from samples into one string
shortsamples <- c()

for (i in 1:nit) {
        shortsamples[i] <- paste(allsampit[i,], collapse = "")
}

# Get the sample distribution that was most common
mostcommon3 <- names(sort(table(shortsamples), decreasing = TRUE)[1:3])

# Test to make sure there are not two that are equally likely
length(unique(mostcommon3)) == 3
```

```
## [1] TRUE
```

```
# Get most common distribution by breaking apart the string
mostcommon <- mostcommon3[1]
mostcommondist <- strsplit(mostcommon3[1], split = "")

# Creates list of either 0 (doesn't match) or 1 (matches)
samecommonsample <- c()
for (i in 1:nit) {
        samecommonsample[i] <- sum(match(paste(shortsamples[i], collapse = ""),
                                   paste(mostcommon, collapse = ""),
                                   nomatch = 0))
}
```

```
The most common distribution of colors from the simulation:
2 brown
2 yellow
3 green
2 red
3 orange
4 blue
```

```
The probability of this distribution is 0.002
```

Question 5

If each bag of M&Ms had contained 50 M&Ms, would your combination of colors (whatever they were) be more or less likely to have occurred and why?

```
# Determine what the number of M&Ms of each color would be if the
# bag had 50 candies and matched my distribution of M&Ms from earlier
bag50 <- c()
for (i in 1:length(mydist)) {bag50[i] <- round(mydist[i]*50)}

# Calculate probability of sampling 50 candies with my distribution
# from the true population distribution of M&Ms
ans5 <- find.prob(bag50, truedist)
```

My combination of colors would have been less likely to have occurred if I had sampled 50 from the true population of M&Ms. The probability of getting my distribution in a 16 candy bag is p(MyColors16) = 1.7902e-04. The likelihood of getting my distribution a 50 candy bag is p(MyColors50) = 1.4973e-08.

When you compare the mathematical expressions, the difference between the two probabilities becomes clear:

The probability of my bag was

$$\begin{aligned} P(MyColors16) &= \frac{n!}{n_1!n_2!n_3!n_4!n_5!n_6!} p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{n_4} p_5^{n_5} p_6^{n_6} \\ &= \frac{16!}{5!1!2!2!1!5!} (0.14^5 0.14^1 0.16^2 0.13^2 0.19^2 0.25^5) \end{aligned}$$

while the probability of a bag of 50 candies with my distribution is

$$\begin{aligned} P(MyColors50) &= \frac{n!}{n_1!n_2!n_3!n_4!n_5!n_6!} p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{n_4} p_5^{n_5} p_6^{n_6} \\ &= \frac{50!}{16!3!6!6!3!16!} (0.14^{16} 0.14^3 0.16^6 0.13^6 0.19^3 0.25^{16}) \end{aligned}$$

Question 6

Lets say you have a bag of 20 M&Ms and find the first 19 are all blue. What is the probability that the 20th is also blue?

Assuming that the bag came from the true global population of M&Ms, represented by a distribution presented in question 2b, then the likelihood of selecting another blue would be the same as selecting blue the previous 19 times. This assumes the same as question 4, wherein the selection of one candy does not influence the selection of other candies and the samples are completely independent and not influencing the color distribution of the population significantly. Therefore, the probability of selecting blue is 0.24.