Stats 604: Project 3

Yizhou Gu, Noah Kochanski, Samuel Rosenberg

10/30/24

1 Introduction

One of the most common household tricks for accelerating the ripening of fruit is to place it in a brown paper bag. The argument goes that there is a cyclical mechanism in which ripening plants reduce ethylene gas which is then trapped by the paper bag, accelerating the ripening process further. If this argument is correct, storing ripening fruit in a variety of containers should accelerate the ripening process, albeit possibly to varying degrees.

Given this household folklore, we were interested in investigating how much veracity there is to this trick. Specifically, we wanted to investigate the following question: *does the storage of bananas in containers such as brown paper bags and plastic kitchen bags, increase the amount of ripening that occurs over the span of four days, all else being equal with standard kitchen characteristics (e.g. lighting, temperature, humidity)?*

2 Experimental Design

We began by purchasing 30 visibly comparable (in terms of ripeness, as determined by color) bananas from a Meijer grocery store. We then weighed and recorded the mass of each banana, matching them into triplets based on their weight such that the heaviest three bananas were in a triplet, the next heaviest three in another, and so on. In this way, we are able to control for unwanted variability in ripening speed that may be due to the mass of the banana rather than the treatment assignment.

With this done, we used the R programming language to assign our three possible treatments (0: control, no container; 1: treatment, paper bag; 2: treatment, plastic Ziploc bag) uniformly at random within each triplet, with a random seed fixed for reproducibility. Both the data and code for treatment assignment, as well as the subsequent measurement process and analysis can be found at this Github repository.

Pre-treatment ripeness measurements were taken using banana color as a proxy for ripeness. To do so, we invited a blinded third-party who was unaware of the treatment assignment to preview randomly shuffled images of the bananas. The photographs were taken all on the same phone (an iPhone 15 Pro Max), using flash and default camera settings with the same camera location and lighting to ensure consistency across images. These images were presented along with a slider with a color gradient background. The color gradient has a darkish green on the left side, a yellow directly in the middle, and a black on the right. The ripeness scores for these given colors would be mapped to a value between 0 and 100. The interface was presented to the rater with the instructions to match color of the banana to the respective color on the slider. The only additional instruction that was given to the rater was to try to be as consistent in their rating as possible. An image of interface used for rating the bananas can be seen in Figure 1.

The bananas were then assigned to their various treatments (i.e. sealed into Ziploc bags, left in brown paper bags, or left out) and placed on the same kitchen counter, ensuring that all units received exposure to typical ripening conditions (such as humidity, temperature, and lighting) which did not vary based on treatment assignment. After four days, photographs were taken again in a similar manner and presented with blinding to the same third-party individual for rating.

With the data collected, we were able to perform statistical inference to answer our research question, as detailed in the following section.

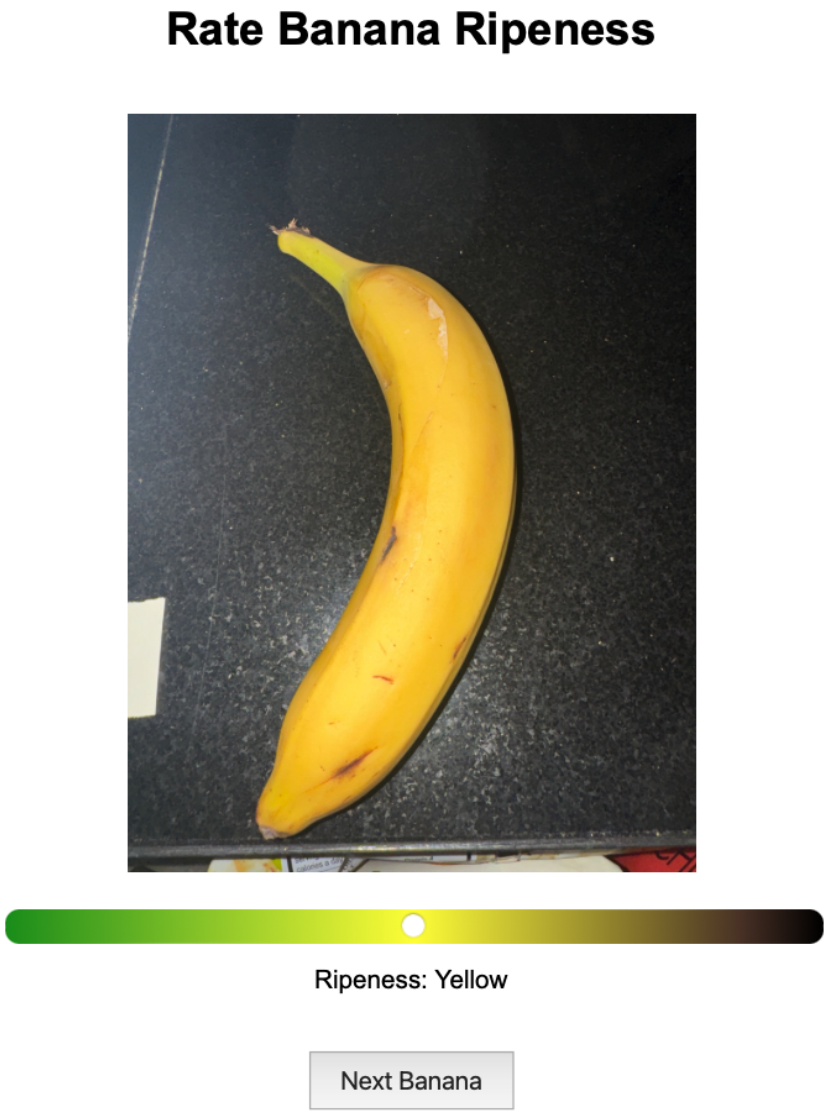

Figure 1: Screenshot of the ripeness tester. An unlabeled banana image was presented with color gradient directly below it.

3 Methods

To answer the research question, we will do three pre-specified hypothesis tests using permutation tests. The three null hypotheses are as follows:

- $H_{0,1}$ = There is no difference in the ripening speed between using paper bag and no bag
- $H_{0,2}$ = There is no difference in the ripening speed between using plastic bag and no bag
- $H_{0,3}$ = There is no difference in the ripening speed between using paper bag and plastic bag

Corresponding to each null hypothesis, we will run a different permutation test. For each banana, we will calculate the difference between post-treatment and pre-treatment ripeness ratings. Since the experiment

was run over a fixed period, this will act as our measure of ripening speed. We will use the difference in means statistic as the test statistic. For each permutation test, we will only take the subset of the data that corresponds to the pair of interest and permute the labels in that subset. We will calculate the observed difference in means between the two groups and then permute the treatment assignments to generate a distribution of differences in means under the null hypothesis. We will then calculate the p-value under a two-sided test, and determine its significance using $\alpha = 0.5$ under a Bonferroni correction.

4 Results

We observe that the difference in means of the ripening speed between paper bag and control is 1.8, resulting in a p-value of 0.616, which is not significant under the specified significance level. The resulting null distribution can be seen in Figure 2.

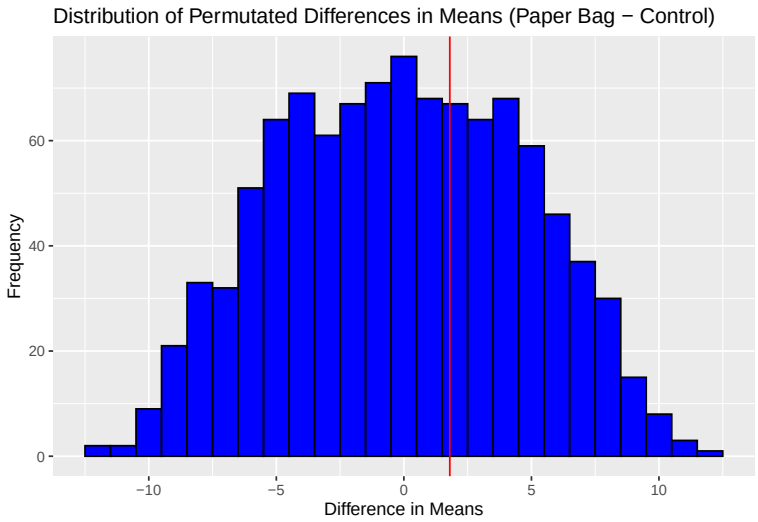

Figure 2: Null distribution generated under $H_{0,1}$ (p-value =0.616).

For the remaining two hypothesis tests, we see statistically significant results under our testing framework. For the second test, as seen in Figure 3, the observed difference in means of the ripening speed between plastic bag and control is -9.7. This results in a p-value of 0.002 under the two-sided test, which is significant. Additionally under the third hypothesis test, the observed difference in means of the ripening speed between plastic bag and paper bag is -9.7. With a p-value of 0.004, this is also significant under the specified significance level. The resulting null distribution under this hypothesis test can be seen in 4.

5 Conclusion

This study examined the impact of different storage methods on banana ripening speed over a four-day period. Results from the permutation test indicate that bananas stored in plastic bags ripen significantly faster than bananas stored in open air or in paper bags. Additionally, the permutation tests indicate that there was no significant difference in speed of ripening from using paper bags. This suggests that the differing speed of ripening may be due to the material used to store bananas. While this study shows these baseline results, there are several limitations to be considered. First, the sample size was relatively small, which may limit the generalizability of our findings. A larger sample would provide more robust estimates of the effects observed here. Additionally, ripeness was assessed using a subjective visual color rating by a single rater, which introduces variability due to individual perception. While color was a primary indicator of ripeness, the photographs used for rating did not capture other important ripeness indicators, such as texture, smell, or softness, which might provide a fuller picture of ripening progression. Future studies could benefit from objective and multi-dimensional ripeness measurements to address these factors more comprehensively.

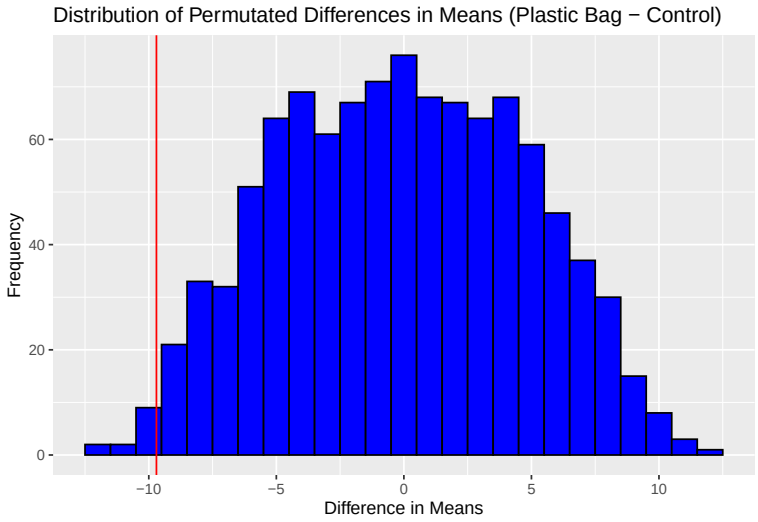

Figure 3: Null distribution generated under $H_{0,2}$ (p-value =0.00195).

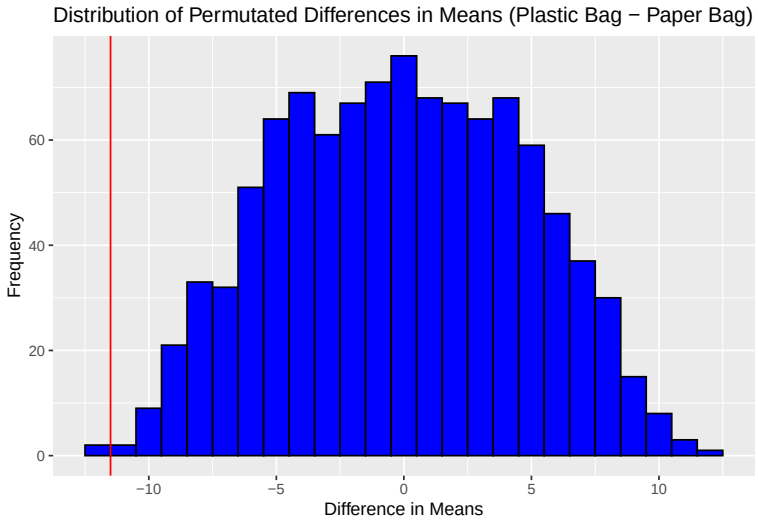

Figure 4: Null distribution generated under $H_{0,3}$ (p-value =0.00391).

6 Code Appendix

```
## ----include=FALSE-----------------------------------------------------------
knitr::opts_chunk$set(echo = TRUE)


## ----------------------------------------------------------------------------
library(ggplot2)


## ----------------------------------------------------------------------------
# Unique identifiers of each banana based on weight (and additional letter if weight tie)
banana_ids <- c(
  "130", "131", "140",
  "144", "146A", "146B",
  "147", "148", "155",
  "156", "157", "162",
  "166", "170", "171",
  "173", "175", "176",
  "181", "183A", "183B",
  "184", "185", "188",
  "189", "190", "200A",
  "200B", "209", "211"
)
# Banana weights
banana_weights <- c(
  130, 131, 140,
  144, 146, 146,
  147, 148, 155,
  156, 157, 162,
  166, 170, 171,
  173, 175, 176,
  181, 183, 183,
  184, 185, 188,
  189, 190, 200,
  200, 209, 211
)
# Banana matched triplet assignments based on weight (group based on weight order statistics)
banana_match_ids <- unlist(lapply(1:10, function(n){rep(n,3)}))

banana_df <- data.frame(id = banana_ids, match_id = banana_match_ids, weight = banana_weights)

# Random treatment assignment with seed set for reproducibility
# 0 corresponds to control, 1 to paper bag, 2 to plastic bag
set.seed(123)
banana_df$treatment <- unlist(lapply(1:10, function(x){sample(0:2, 3, replace = FALSE)}))

banana_df


## ----------------------------------------------------------------------------
banana_rating <- read.csv("banana_ratings_final.csv")
# Reorder the banana_rating data frame by the BananaID
banana_rating <- banana_rating[order(banana_rating$BananaID),]
```

```r
# Add a column to the banana_df data frame to store the post treatment rating, the values are the rows
banana_df$post_rating <- banana_rating$Rating[1:30]

# Add a column to the banana_df data frame to store the difference between the pre treatment rating, th
banana_df$pre_rating <- banana_rating$Rating[31:60]

# Add a column to the banana_df data frame to store the difference between the post treatment rating an
banana_df$rating_diff <- banana_df$post_rating - banana_df$pre_rating

banana_df


## ------------------------------------------------------------------------
# Plot the distribution of the rating difference
ggplot(banana_df, aes(x = rating_diff)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black") +
  labs(title = "Distribution of Rating Difference", x = "Rating Difference", y = "Frequency")


## ------------------------------------------------------------------------
ggplot(banana_df, aes(x = rating_diff, y = as.factor(treatment), fill = as.factor(treatment))) +
  geom_boxplot() +
  labs(title = "Distribution of Rating Difference",
       x = "Rating Difference",
       y = "Frequency",
       fill = "Group") +  # Change the legend title to "Group"
  scale_fill_manual(values = c("0" = "white",
                               "1" = "lightblue",
                               "2" = "brown"),   # Set colors for each group
                    labels = c("0" = "Control",
                               "1" = "Plastic Bag",
                               "2" = "Paper Bag")) +  # Map numeric values to labels
  theme_minimal()


## ------------------------------------------------------------------------

# Permutaion Test 1: Control vs. Paper Bag


# Use the difference in means statistic to run a permutation test on control vs. paper bag
# Calculate the observed difference in means
obs_control_paper_diff <- mean(banana_df$rating_diff[banana_df$treatment == 1]) - mean(banana_df$rating

# Create a subset of the data frame with only the control and paper bag treatments
control_paper_df <- banana_df[banana_df$treatment %in% c(0, 1),]

# Create a vector to store the permuted differences in means
perm_diffs <- numeric(2^10)

# Run the permutation test by going through all possible permutations of the treatment assignments
# We are using the first 10 digits of the binary expansion of
# the numbers 0 to 1023 to represent
```

```r
# all possible permutations of the 10 pairs
for(i in 0:(2^10 - 1)){
  # Convert the integer i to a binary representation
  binary_i <- as.integer(intToBits(i))

  # The length of binary_i is 32, but we only need the first 10 digits
  # to represent the permutation of the 10 pairs
  binary_i <- binary_i[1:10]

  # Now, we can use binary_i to permute the data
  permuted_control_diff <- rep(NA, 10)
  permuted_paper_diff <- rep(NA, 10)

  for(j in 1:10){
    if(binary_i[j] == 0){
      permuted_control_diff[j] <- control_paper_df$rating_diff[j*2 - 1]
      permuted_paper_diff[j] <- control_paper_df$rating_diff[j*2]
    } else {
      permuted_control_diff[j] <- control_paper_df$rating_diff[j*2]
      permuted_paper_diff[j] <- control_paper_df$rating_diff[j*2 - 1]
    }
  }

  perm_diffs[i] <- mean(permuted_paper_diff) - mean(permuted_control_diff)
}

# Plot the distribution of the permuted differences in means
plot_1 <- ggplot() +
  geom_histogram(aes(x = perm_diffs), binwidth = 1, fill = "blue", color = "black") +
  geom_vline(xintercept = obs_control_paper_diff, color = "red") +
  labs(title = "Distribution of Permutated Differences in Means (Paper Bag - Control)", x = "Difference

# Calculate the p-value
p_value_1 <- sum(abs(perm_diffs) >= abs(obs_control_paper_diff)) / length(perm_diffs)


## ------------------------------------------------------------------------

# Permutation Test 2: Control vs. Plastic Bag

# Use the difference in means statistic to run a permutation test on control vs. plastic bag
# Calculate the observed difference in means

obs_control_plastic_diff <- mean(banana_df$rating_diff[banana_df$treatment == 2]) - mean(banana_df$ratin

# Create a subset of the data frame with only the control and plastic bag treatments
control_plastic_df <- banana_df[banana_df$treatment %in% c(0, 2),]

# Create a vector to store the permuted differences in means
perm_diffs <- numeric(2^10)

# Run the permutation test by going through all possible permutations of the treatment assignments
# We are using the first 10 digits of the binary expansion of
```

```r
# the numbers 0 to 1023 to represent
# all possible permutations of the 10 twins

for(i in 0:(2^10 - 1)){
  # Convert the integer i to a binary representation
  binary_i <- as.integer(intToBits(i))

  # The length of binary_i is 32, but we only need the first 10 digits
  # to represent the permutation of the 10 pairs
  binary_i <- binary_i[1:10]

  # Now, we can use binary_i to permute the data
  permuted_control_diff <- rep(NA, 10)
  permuted_plastic_diff <- rep(NA, 10)

  for(j in 1:10){
    if(binary_i[j] == 0){
      permuted_control_diff[j] <- control_plastic_df$rating_diff[j*2 - 1]
      permuted_plastic_diff[j] <- control_plastic_df$rating_diff[j*2]
    } else {
      permuted_control_diff[j] <- control_plastic_df$rating_diff[j*2]
      permuted_plastic_diff[j] <- control_plastic_df$rating_diff[j*2 - 1]
    }
  }

  perm_diffs[i] <- mean(permuted_plastic_diff) - mean(permuted_control_diff)
}

# Plot the distribution of the permuted differences in means
plot_2<- ggplot() +
  geom_histogram(aes(x = perm_diffs), binwidth = 1, fill = "blue", color = "black") +
  geom_vline(xintercept = obs_control_plastic_diff, color = "red") +
  labs(title = "Distribution of Permutated Differences in Means (Plastic Bag - Control)", x = "Differen

# Calculate the p-value
p_value_2 <- sum(abs(perm_diffs) >= abs(obs_control_plastic_diff)) / length(perm_diffs)


## ------------------------------------------------------------------------

# Permutation Test 3: Paper Bag vs. Plastic Bag


# Use the difference in means statistic to run a permutation test on paper bag vs. plastic bag
# Calculate the observed difference in means
obs_paper_plastic_diff <- mean(banana_df$rating_diff[banana_df$treatment == 2]) - mean(banana_df$rating_

# Create a subset of the data frame with only the paper bag and plastic bag treatments
paper_plastic_df <- banana_df[banana_df$treatment %in% c(1, 2),]

# Create a vector to store the permuted differences in means
perm_diffs <- numeric(2^10)
```

```r
# Run the permutation test by going through all possible permutations of the treatment assignments
# We are using the first 10 digits of the binary expansion of
# the numbers 0 to 1023 to represent
# all possible permutations of the 10 twins

for(i in 0:(2^10 - 1)){
  # Convert the integer i to a binary representation
  binary_i <- as.integer(intToBits(i))

  # The length of binary_i is 32, but we only need the first 10 digits
  # to represent the permutation of the 10 pairs
  binary_i <- binary_i[1:10]

  # Now, we can use binary_i to permute the data
  permuted_paper_diff <- rep(NA, 10)
  permuted_plastic_diff <- rep(NA, 10)

  for(j in 1:10){
    if(binary_i[j] == 0){
      permuted_paper_diff[j] <- paper_plastic_df$rating_diff[j*2 - 1]
      permuted_plastic_diff[j] <- paper_plastic_df$rating_diff[j*2]
    } else {
      permuted_paper_diff[j] <- paper_plastic_df$rating_diff[j*2]
      permuted_plastic_diff[j] <- paper_plastic_df$rating_diff[j*2 - 1]
    }
  }

  perm_diffs[i] <- mean(permuted_plastic_diff) - mean(permuted_paper_diff)
}

# Plot the distribution of the permuted differences in means
plot_3 <- ggplot() +
  geom_histogram(aes(x = perm_diffs), binwidth = 1, fill = "blue", color = "black") +
  geom_vline(xintercept = obs_paper_plastic_diff, color = "red") +
  labs(title = "Distribution of Permutated Differences in Means (Plastic Bag - Paper Bag)", x = "Differ

# Calculate the p-value
p_value_3 <- sum(abs(perm_diffs) >= abs(obs_paper_plastic_diff)) / length(perm_diffs)
```