Statistical Analysis of U.S. Avocado Prices

Elinor Holt, Xinran Zhang, Ruixi Xu

2025-11-11

Opening Remarks

This project explores three statistical methods for analyzing avocado prices across different conditions, regions, and time periods using data from the Hass Avocado Board. We apply a nonparametric test, a one-factor ANOVA, and a repeated measures ANOVA to examine how avocado prices vary under different conditions and over time.

The dataset used for this analysis can be found at: https://www.kaggle.com/datasets/vakhariapujan/avocado-prices-and-sales-volume-2015-2023

Thesis 1 (S1: Nonparametric Test)

We hypothesize that there is a significant difference in average avocado prices between organic and conventional avocados sold in Boston during 2020. This will be tested using a Mann–Whitney U test, a nonparametric alternative to the two-sample t-test that does not require normality.

Thesis 2 (S2: One-Factor ANOVA)

We hypothesize that average organic avocado prices differ significantly across four U.S. cities over the 2015–2023 period. This will be examined using a one-factor ANOVA, with corrections for heteroscedasticity if variances differ among groups.

Thesis 3 (S3: Repeated Measures ANOVA)

We hypothesize that average avocado prices across major U.S. cities changed significantly over a six-week period in late 2023. Because the same cities are measured repeatedly, we employ a Repeated Measures ANOVA, with corrections applied to account for violations of sphericity.

S1: Non Parametric Test

The objective of this test is to determine whether there is a significant difference in the average avocado prices between organic and conventional avocados sold in Boston during 2020. Formally, we set up the hypothesis as:

$$H_0 : \mu_{organic} = \mu_{conventional}$$
$$H_a : \mu_{organic} \neq \mu_{conventional}$$

Because the distribution of AveragePrice is not guaranteed to be normal and the sample sizes of the two groups may differ, we apply a Mann–Whitney U test, which is a nonparametric alternative to the two-sample t-test that compares the central tendencies of the two groups using ranked data.

Assumptions

To conduct the test, we need to make sure the data satisfy the assumptions:

(1) Independence within groups
(2) Independence between groups

Independence within groups

The independence within groups assumes that each observation within a group is independent of the others. While the avocado prices are recorded weekly and may exhibit minor temporal correlation, the data are treated as approximately independent because each weekly average represents a separate market observation.Thus, the within-group independence assumption is reasonably met for this nonparametric test.

Independence between groups

The independence between groups assumes that the observations in one group (organic avocados) are independent from those in the other group (conventional avocados).In this analysis, each observation corresponds to a distinct weekly average price for one type of avocado, and the two product types represent separate populations with no overlapping records.Therefore, the between-group independence assumption is satisfied.

Mann–Whitney U test

$$MW_{stat} := \min\{U_1, U_2\}$$

where U_1, U_2 represents the number of pairwise comparisons between two samples in which the observations of one group precede those of the other in ranked order. If two means are significantly different, then U_1, U_2 will be significantly different. The calculation of "partial U" test statistics is:

$$U_j = T_j - \frac{n_j(n_j+1)}{2}$$

where T_j is the total rank. To get T_j, we need to get the ranks for both groups. The calculation of T_j is:

$$T_j = \sum_{i=1}^{n_j} (R_j)_i$$

We can now conduct the Mann-Whitney test to obtain the test statistics and compare it with the critical value to check whether the mean price for both avocados are the same. First we need to calculate the U_1, U_2 below:

```
mydata = read.csv('Avocado.csv')

X = mydata$AveragePrice[substr(mydata$Date, 1, 4) == '2020' &
                          mydata$region == "Boston" &
                          mydata$type=='organic' ]

Y = mydata$AveragePrice[substr(mydata$Date, 1, 4) == '2020' &
                          mydata$region == "Boston" &
                          mydata$type=='conventional' ]
n_x = length(X)
```

```r
n_y = length(Y)
L = list('organic' = X, 'conventional' = Y)
nj = sapply(L, length)
n = sum(nj)
Sall = unlist(L)
idx = rep(names(L), times=nj)
Rall = rank(Sall)
Rj = split(Rall,idx)
Tx = sum(Rj$organic)
Ty = sum(Rj$conventional)
mean_x = mean(Rj$organic)
mean_y = mean(Rj$conventional)
U_x = Tx-(n_x*(n_x+1))/2
U_y = Ty-(n_y*(n_y+1))/2
```

After conducting the U_1, U_2 , we can calculate Mann-Whitney test statistics, below is the result:

```r
MWstat = min(U_x, U_y)
MWstat
```

```
## [1] 0
```

We set the confidence level as 95%. So α=0.05. Then, we'll conduct the critical value as below:

```r
MWcrit = qwilcox(0.05, n_x, n_y, lower.tail = T)
MWcrit
```

```
## [1] 1099
```

Since our test statistics is 0, which is far less than the critical value, this indicates that all organic prices ranked higher than all conventional ones in our sample from 2020 in the Boston region. Therefore, we have sufficient evidence to reject the null hypothesis and suggest the average price for organic and conventional avocados in 2020 in Boston region is different.

Boostrap Sampling

Then we want to investigate the uncertainty of the estimated mean difference in avocado prices between organic and conventional types by using a bootstrap sampling with 10000 replications. We use bootstrap sampling because it does not require the normality assumption and allows us to estimate the uncertainty (sampling distribution) of the mean difference directly from the data.

Before sampling, we fixed the random seed for reproducibility:

```r
nboot = 10000
set.seed(15000)
boot_diff = numeric(nboot)
for (i in 1:nboot) {
  x_star <- sample(X, length(X), replace = TRUE)
  y_star <- sample(Y, length(Y), replace = TRUE)
  boot_diff[i] <- mean(x_star) - mean(y_star)
}
quantile(boot_diff, c(0.025, 0.975))
```

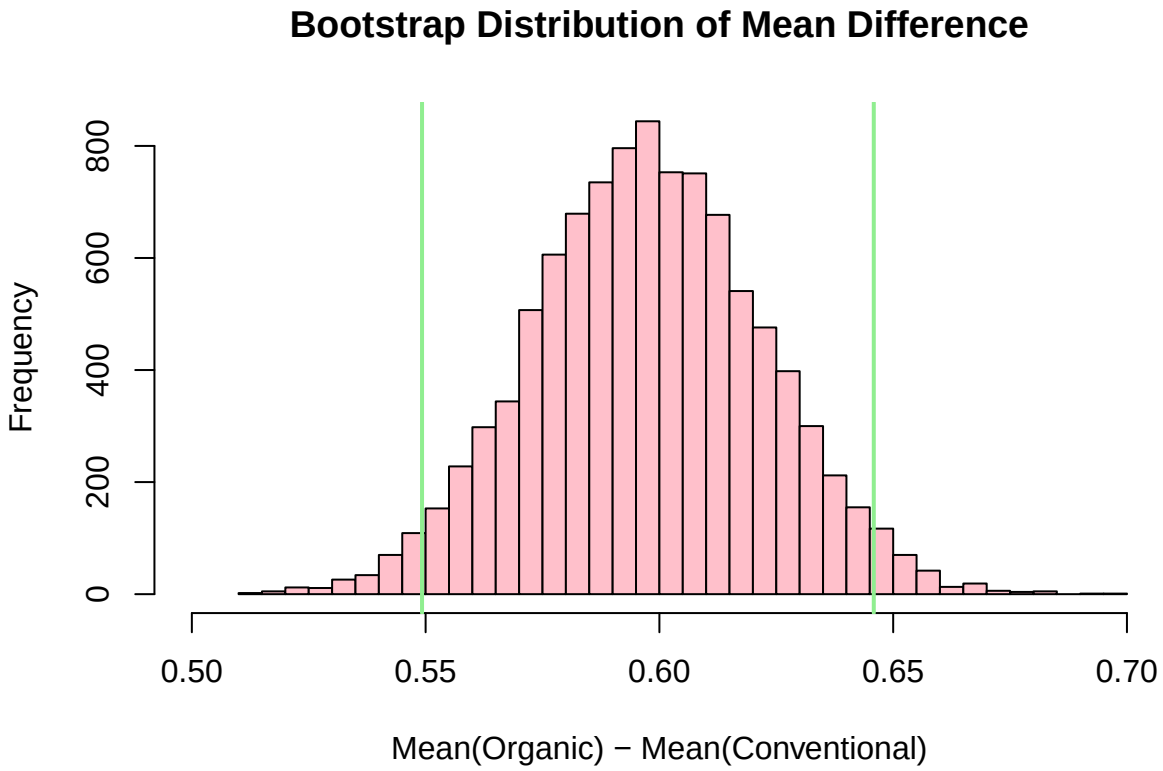
```
##      2.5%     97.5%
## 0.5492310 0.6458507
```

We used set.seed(15000) to ensure that the random resampling in the bootstrap procedure can be reproduced. In other words, this allows the same bootstrap samples (and confidence interval) to be obtained every time the code is run.

The result suggests that we are 95% confident that the mean price difference between organic and conventional avocados in the Boston region in 2020 falls within the confidence interval:

$$CI^{95\%}_{\mu_{organic}-\mu_{conventional}} = (0.5492310,\ 0.6458507)$$

```r
hist(boot_diff, breaks = 30, col = "pink",
     main = "Bootstrap Distribution of Mean Difference",
     xlab = "Mean(Organic) - Mean(Conventional)",
     xlim = c(0.50,0.70))
abline(v = quantile(boot_diff, c(0.025, 0.975)), col = "lightgreen", lwd = 2)
```



The histogram above displays the bootstrap sampling distribution of the mean price difference between organic and conventional avocados in Boston (2020). The distribution is approximately symmetric and centered around 0.6, indicating that organic avocados are, on average, about $0.6 more expensive. The red vertical lines denote the 95% bootstrap confidence interval ([0.55, 0.65]), which does not include zero, suggesting that the difference is statistically significant and that organic prices are consistently higher.

S2: One-Factor ANOVA on Avocado Prices

We want to test whether average avocado prices differ across several regions in a single year.

$$H_0 : \mu_{\text{NewYork}} = \mu_{\text{Jacksonville}} = \mu_{\text{Albany}} = \mu_{\text{Denver}}$$
$$H_a : \text{At least one regional mean price is different.}$$

The response variable is the average price of organic avocados, and the factor is region. We use the Hass Avocado Board dataset, which reports weekly average prices across multiple U.S. cities (New York, Jacksonville, Albany, and Denver) from 2015 to 2023.

We first read in the data and construct the sample that matches our claim.

```
mydata = read.csv('Avocado.csv')
rows_use = which(mydata$region %in% c("NewYork", "Jacksonville", "Albany", "Denver")
                 & (mydata$type == "organic"))

data_anova = mydata[rows_use, c("AveragePrice", "region")]
data_anova$region = factor(data_anova$region)
price_by_region = split(data_anova$AveragePrice, data_anova$region)
```

Assumptions

In order to conduct a one-factor ANOVA test, our data must satisfy the four required assumptions:

(1) Normality
(2) Homoscedasticity
(3) Independence between groups
(4) Independence within groups

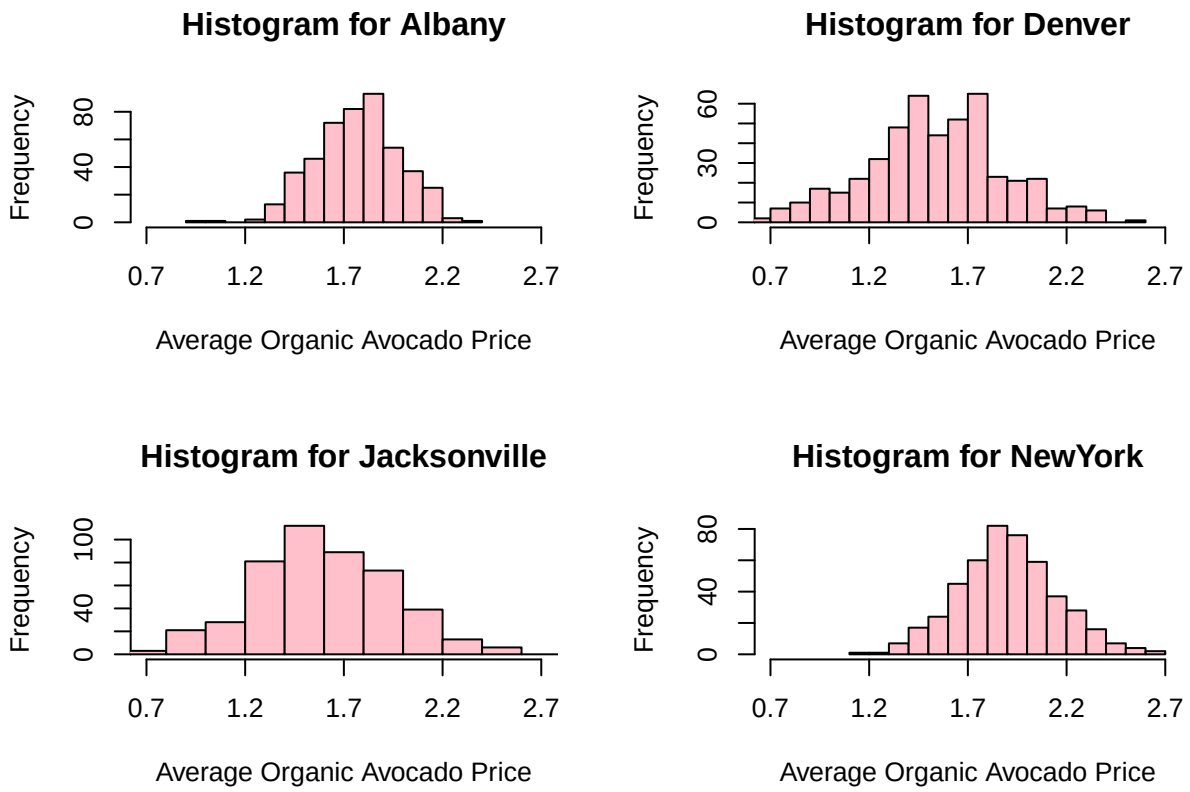
Normality

First we will test the normality of each region's average organic avocado prices. Our hypothesis is :

$$H_0 : P \sim N(\mu, \sigma^2)$$
$$H_a : P \nsim N(\mu, \sigma^2)$$

We can examine histograms to qualitatively check for normality within our samples.

```
par(mfrow = c(2, 2))
invisible(
  lapply(names(price_by_region), function(nm) {
    hist(price_by_region[[nm]],
         breaks = 15,
         col = "pink",
         main = paste("Histogram for", nm),
         xlab = paste("Average Organic Avocado Price"),
         xlim = c(0.7,2.7),
         xaxt = "n")
    axis(1, at = seq(0.7, 2.8, by = 0.5))
  })
)
```



All of the distributions appear to approximately follow a bell curve, suggesting that they are normally distributed. However, it is also important to test normality quantitatively. The Anderson–Darling test was chosen because it provides a powerful and tail-sensitive measure of deviation from the normal distribution. Formally, the Anderson–Darling test statistic is given by:

$$AD_{stat} = \sum_{k=1}^{n} \frac{|F_k - \psi_k|^2}{F_k \cdot (1 - F_k)}$$

where F_k represents the value of the empirical cumulative distribution function (ECDF) at the k-th ordered observation. The formula of F_k is give by:

$$F_k = \frac{k}{n+1}, \quad \text{for } k = 1, 2, \ldots, n$$

ψ_k represents the expected cumulative probability under the theoretical normal distribution, at the same position as the k-th ordered data point, given by

$$\psi_k = \Phi\left(\frac{X_{(k)} - \mu}{\sigma}\right),$$

We can now conduct the AD stat and compare the test statistic with the critical value.

```
AD_stat = function(X) {
  mu_X = mean(X)
  sigma_X = sd(X)
```

```
  X = sort(X)
  n = length(X)
  Fx = seq(1, n, 1) / (n + 1)  # empirical CDF (approx)
  Psix = pnorm(X, mu_X, sigma_X)  # theoretical CDF under N(mu, sigma)
  # Anderson-Darling statistic
  AD_stat = sum(((Fx - Psix)^2) / (Fx * (1 - Fx)))
  return(AD_stat = AD_stat)
}
```

We need to compare the AD test statistics we computed with the critical value under 95% confidence level. We will be using Monte-Carlo simulations to compute the critical value at alpha = 0.05. We use Monte Carlo simulations to estimate the critical value because the Anderson–Darling statistic has no simple theoretical distribution. By simulating many samples under the normality assumption, we can empirically find the threshold beyond which data are considered non-normal. The way to conduct critical value is:

```
Monte_Carlo_Simulation_Norm = function(n_sims, stat_func) {
  results = numeric(n_sims)
  for (i in 1:n_sims) {
    sample_data = rnorm(1000)
    results[i] = stat_func(sample_data)
  }
  return(results)
}

mcn = 10000
alpha = 0.05
MC_dist = Monte_Carlo_Simulation_Norm(mcn, AD_stat)
AD_crit = quantile(MC_dist, 1-alpha)
```

Next, we check how many of the cities' AD test statistics are greater than the critical value. If they are greater than the critical value, we failed to reject the null hypothesis. By using the code below, we can see the result:

```
avo_ad_stats_1 = lapply(price_by_region, AD_stat)
exceeds_crit = sapply(avo_ad_stats_1, function(stat) stat < AD_crit)
exceeds_crit
```

```
##       Albany.95%       Denver.95% Jacksonville.95%      NewYork.95%
##             TRUE             TRUE             TRUE             TRUE
```

The result shows that none of those cities' AD test statistic is greater than the critical value. Therefore, we have evidence to believe the organic avocado price of Denver, Albany, Jacksonville and New York are normally distributed.

Homoscedasticity

The next assumption to test is homoscedasticity which we will check using the Bartlett test statistic:

```
Bartlett_stat = function(samples, alpha = 0.05) {
  # Group information
  g = length(samples)
```

```
  n_i = sapply(samples, length)
  s_i2 = sapply(samples, var)
  n = sum(n_i)
  grand_mean = mean(unlist(samples))

  # Compute pooled variance
  s2_pooled = sum((n_i - 1) * s_i2) / (n - g)

  # Compute Bartlett's test statistic
  C = (n - g) * log(s2_pooled) - sum((n_i - 1) * log(s_i2))
  denom = 1 + (1 / (3 * (g - 1))) * (sum(1 / (n_i - 1)) - 1 / (n - g))
  B = C / denom

  # Compute p-value & critical value
  p_value = 1 - pchisq(B, df = g - 1)
  B_crit = qchisq(1-alpha, df = g - 1)

  # Return  results
  return(list(
    Bartlett_Statistic = B,
    Critical_Value = B_crit,
    p_Value = p_value))
}

results = Bartlett_stat(price_by_region)
results
```

```
## $Bartlett_Statistic
## [1] 166.4465
##
## $Critical_Value
## [1] 7.814728
##
## $p_Value
## [1] 0
```

Our resulting p-val is very small, approximately zero, so homoscedasticity is violated at alpha = 0.05. We will use a heteroscedastic ANOVA test in order to correct this.

Independence between samples

For this analysis, we assume that the samples from each region are independent. In other words, the average avocado prices collected in New York, Jacksonville, Albany, and Denver are treated as coming from separate and unrelated populations. Since each city represents a distinct market with its own supply, demand, and pricing conditions, there's no direct overlap or dependency between them. While all regions may experience broader national trends, those effects are assumed to impact each region independently for the purpose of this ANOVA test.

Independence of observations within samples

We also assume that observations within each region are independent. Here, that means each recorded price in a given city is assumed not to influence the others, so the within-region errors behave like independent draws from the same distribution.

One Factor Heteroscedastic ANOVA Test

(1) Groups weights: Each group is assigned a weight proportional to its variance, so that groups with smaller variances contribute more to the overall mean

$$w_j = \frac{n_j}{s_j^2}$$

(2) Heteroscedastic Grand Mean: Using these weights, the heteroscedastic grand mean is computed as a weighted average of the group means. This accounts for unequal group sizes and variances.

$$\bar{\bar{x}}_H = \frac{\sum_{j=1}^{g} w_j \bar{x}_j}{\sum_{j=1}^{g} w_j}$$

(3) Between Group Variance (adjusted): The weighted between-group sum of squares measures the variability of each group mean relative to the heteroscedastic grand mean, with each term scaled by its group weight.

$$SSM^* = \sum_{j=1}^{g} w_j \left(\bar{x}_j - \bar{\bar{x}}_H\right)^2$$

(4) Mean Square Between (adjusted): Dividing the weighted sum of squares by its degrees of freedom (g-1) gives the adjusted mean square.

$$MSM^* = \frac{SSM^*}{g-1}$$

(5) Adjusted F-statistic: compares the adjusted between group variance to the within group variance, while incorporating a correction factor that accounts for heteroscedasticity.

$$F^* = \frac{MSM^*}{1 + \frac{2(g-2)}{\nu^*}}$$

(6) Adjusted Degrees of Freedom: The modified degrees of freedom are computed based on group sample sizes and variances. This correction corrects the F-statistic when group variances differ (when homoscedasticity fails).

$$\nu^* = \frac{\left(\frac{n_1}{s_1^2} + \frac{n_2}{s_2^2} + \cdots + \frac{n_g}{s_g^2}\right)^2}{\frac{n_1^2}{s_1^4 \nu_1} + \frac{n_2^2}{s_2^4 \nu_2} + \cdots + \frac{n_g^2}{s_g^4 \nu_g}}$$

We compute the F-statistic below:

```r
Hetero_ANOVA = function(samples, equal_var = FALSE) {
  # Basic group statistics
  g = length(samples)
  n_i = sapply(samples, length)
  x_i = sapply(samples, mean)
  s_i2 = sapply(samples, var)
  n_total = sum(n_i)

  w_i = n_i / s_i2
  grand_mean = sum(w_i * x_i) / sum(w_i)
  SSM_star = sum(w_i * (x_i - grand_mean)^2)
  MSM_star = SSM_star / (g - 1)

  # df (approximate)
  nu = (sum(w_i))^2 / sum((w_i^2) / (n_i - 1))

  # F*
  F_star = MSM_star / (1 + ((2 * (g - 2)) / nu))
  p_value = 1 - pf(F_star, g - 1, nu)

  # results
  return(list(
    F_Statistic = round(F_star, 4),
    p_Value = round(p_value, 6)))
}

output = Hetero_ANOVA(price_by_region)
output
```

```
## $F_Statistic
## [1] 148.5499
##
## $p_Value
## [1] 0
```

From the ANOVA output, the observed F-statistic is approximately 148.5499, and the corresponding p-value is approximately 0, which is far smaller than 0.05. Therefore, the test statistic falls in the rejection region, and we have strong evidence to reject the null hypothesis. We conclude that average avocado prices from 2015 - 2023 differ significantly among the four regions considered.

S3: Repeated Measures ANOVA on avocado prices over six weeks

The goal of this section is to determine whether average avocado prices changed significantly over time across 33 major U.S. cities. Each city acts as one subject in a repeated measures test, and its average price was recorded at six consecutive weekly time points:

2023-10-22, 2023-10-29, 2023-11-05, 2023-11-12, 2023-11-19, 2023-11-26

Because the same cities are observed repeatedly, their average prices are pairwise-dependent. This justifies the use of a Repeated Measures ANOVA (RM-ANOVA). We will test the following hypothesis:

$$H_0 : \mu_{date1} = \mu_{date2} = \mu_{date3} = \mu_{date4} = \mu_{date5} = \mu_{date6}$$
$$H_a : \text{ At least one group mean is not equal}$$

Step 1: Check Assumptions

In order to conduct a RM-ANOVA test, our data must satisfy three assumptions:

(1) Normality
(2) Sphericity
(3) Independence (of)

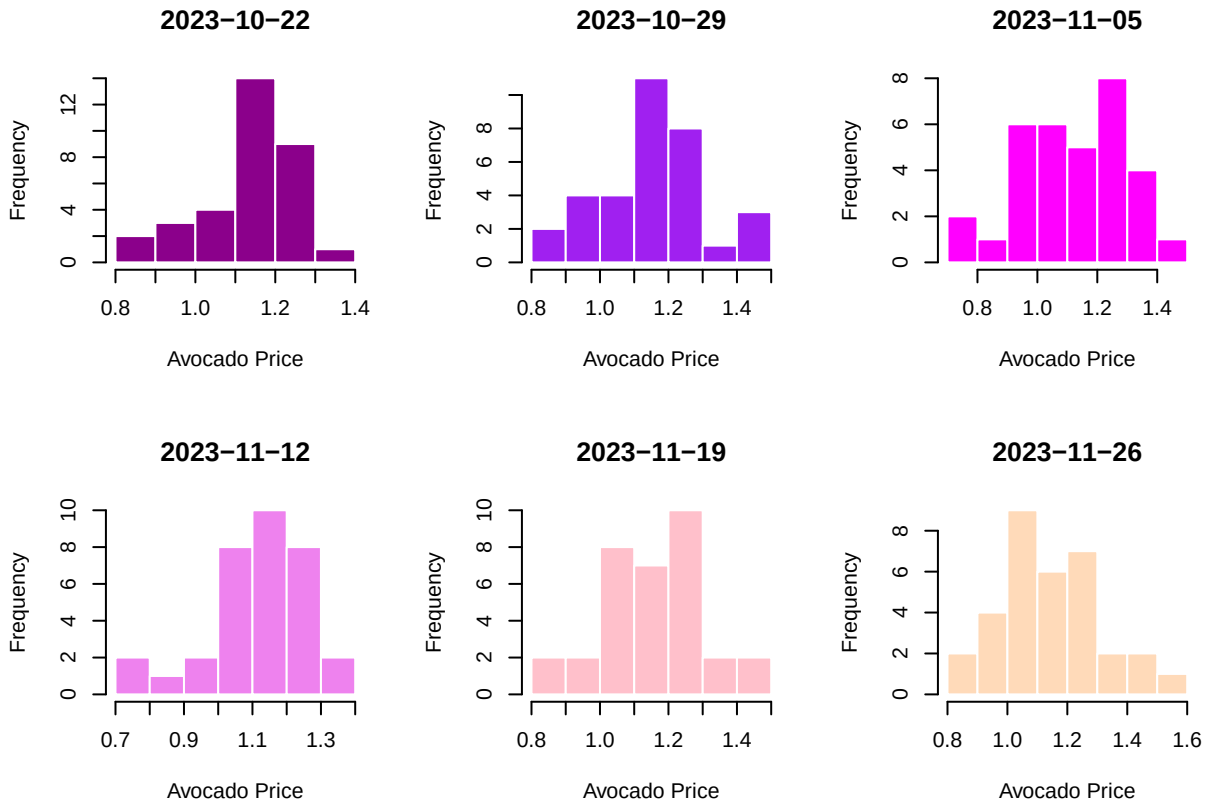
Normality

We will qualitatively and quantitatively analyze our groups to determine if the samples are normally distributed. For each region (subject), we test the following hypotheses:

$$H_0 : P \sim N(\mu, \sigma^2)$$
$$H_a : P \nsim N(\mu, \sigma^2)$$

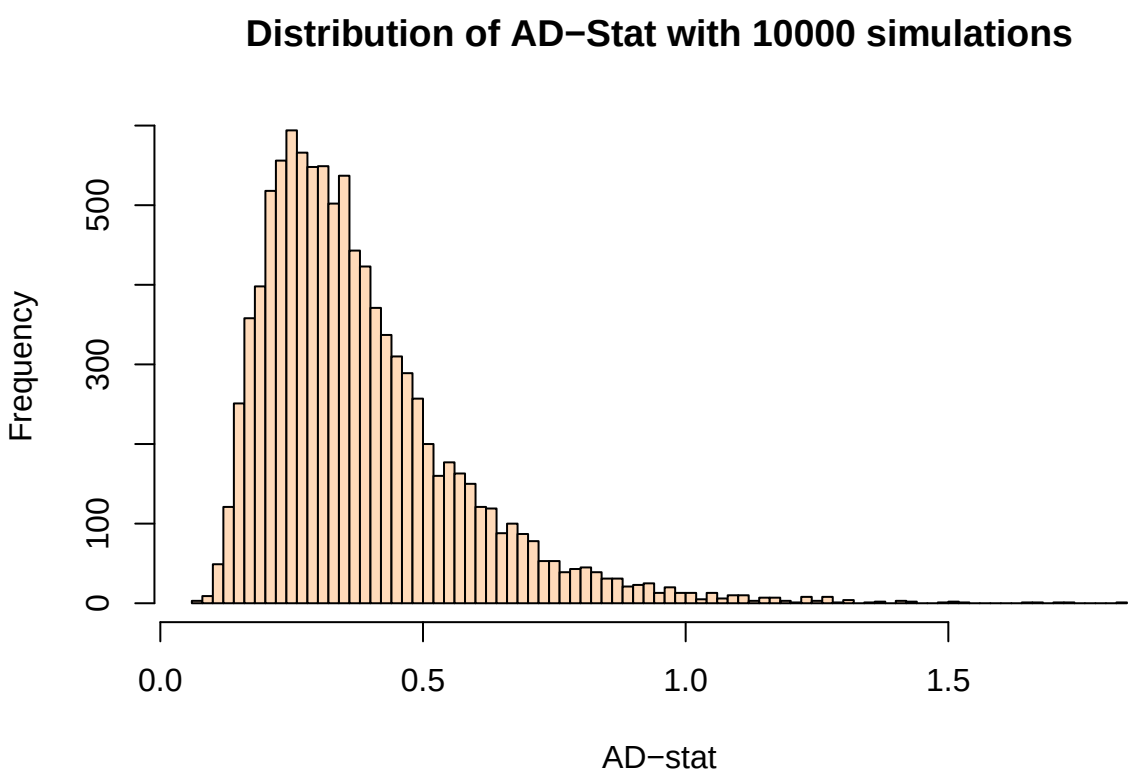
```
# The region variable in this data set also includes broader regions
cities_of_interest = c("Albany", "Atlanta", "Boise", "Boston",
                       "Charlotte", "Chicago","CincinnatiDayton", "Columbus",
                       "DallasFtWorth", "Denver","Detroit", "Houston",
                       "Indianapolis", "Jacksonville", "LasVegas",
                        "LosAngeles", "Louisville", "Nashville", "NewOrleans",
                       "NewYork", "Northeast", "Orlando", "Philadelphia",
                       "Pittsburgh", "Portland","Sacramento", "SanDiego",
                       "SanFrancisco", "Seattle", "SouthCarolina","StLouis",
                       "Syracuse", "Tampa")

major_cities = mydata[mydata$type == "conventional" &
                        mydata$region %in% cities_of_interest, ]

# dates we are grouping by
dates = c("2023-10-22", "2023-10-29", "2023-11-05", "2023-11-12",
          "2023-11-19", "2023-11-26")


# Create a list of AveragePrice vectors, one per date
avo_prices_list = lapply(dates,
        function(d) major_cities$AveragePrice[major_cities$Date == d])
names(avo_prices_list) = dates

# Put our groups into a data frame for future problems
avo_prices = as.data.frame(avo_prices_list)
```

```
# Lets look at the histograms now for normality
par(mfrow = c(2, 3))  # cut it up into sixths
colors = c("darkmagenta", "purple", "magenta", "violet", "pink", "peachpuff")

for(i in seq_along(dates)) {
  hist(avo_prices_list[[i]], col = colors[i], main = dates[i],
       xlab = "Avocado Price",border = "white")
}
```



Based on the histograms, the distributions of average avocado prices at each time point appear approximately normal. We will also test for normality quantitatively using the Anderson-Darling test statistic.

```
mcn = 10000
alpha = 0.05

# Monte_Carlo_Simulation_Norm function definition can be found in One-Factor ANOVA section
MC_dist = Monte_Carlo_Simulation_Norm(mcn, AD_stat)
AD_crit = quantile(MC_dist, 1-alpha)

hist(MC_dist, breaks = 100, main = "Distribution of AD-Stat with 10000 simulations",
     xlab = "AD-stat", col = "peachpuff")
```




```
# AD_stat function definition can be found in One-Factor ANOVA section
avo_ad_stats = sapply(avo_prices_list, AD_stat)

# find which regions fail or pass the normality assumption
passes_normality = avo_ad_stats < AD_crit

# dataframe for nice output
ad_results = data.frame(
  AD_Statistic = round(avo_ad_stats, 4),
  Critical_Value = round(AD_crit, 4),
  Passes_Normality = ifelse(passes_normality, "Yes", "No"))

ad_results
```

```
##            AD_Statistic Critical_Value Passes_Normality
## 2023-10-22       0.3893         0.7444              Yes
## 2023-10-29       0.2957         0.7444              Yes
## 2023-11-05       0.1841         0.7444              Yes
## 2023-11-12       0.3260         0.7444              Yes
## 2023-11-19       0.2366         0.7444              Yes
## 2023-11-26       0.1959         0.7444              Yes
```

The Anderson-Darling test statistic measures how closely a sample follows a normal distribution. Since all our AD-stats are less than our AD-crit computed using a Monte-Carlo simulation with 1000 simulations on a normal distribution, this indicates that the avocado prices for each week are approximately normally distributed (at $\alpha = 0.05$). Therefore, the assumption of normality is satisfied for these data.

Sphericity

Sphericity means that the variances of all pairwise differences between weeks are equal. Violating this assumption inflates Type I error rates. We will test the following hypothesis:

$$H_0 : \text{The covariance matrix is spherical (sphericity holds)}$$
$$H_a : \text{The covariance matrix is not spherical}$$

(1) The empirical (observed) covariance matrix is given by:

$$\hat{\mathbf{\Sigma}}_0 = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_k^2 \end{bmatrix}$$

(2) Covariance matrix under the assumption of compound symmetry: To construct this, we first compute the trace of the observed covariance matrix and divide it by g to obtain the common variance estimate. We estimate the sample covariance matrix $\hat{\Sigma}_0$ directly from the data, and the idealized compound symmetric matrix $\hat{\Sigma}_x$ assumes equal variances across weeks.

$$\hat{\sigma}^2 = \frac{s_1^2 + s_2^2 + \ ... \ + s_g^2}{g}$$

$$\hat{\mathbf{\Sigma}}_x = \begin{bmatrix} \hat{\sigma}^2 & 0 & \cdots & 0 \\ 0 & \hat{\sigma}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\sigma}^2 \end{bmatrix}$$

$$\text{Note that: } \det(\hat{\mathbf{\Sigma}}_x) = (\hat{\sigma}^2)^g = \left[\frac{\text{tr}(\hat{\mathbf{\Sigma}}_x)}{g}\right]^g$$

(3) Sphericity is tested using Mauchly's Test, which compares the determinant of the observed covariance matrix to that of the compound-symmetric covariance matrix. The test statistic W_{stat} is defined as:

$$W_{stat} = \frac{\det(\hat{\mathbf{\Sigma}}_x)}{\det(\hat{\mathbf{\Sigma}}_0)}$$

(4) A smaller value of W_{stat} indicates a larger deviation from sphericity. There is evidence to believe that sphericity is satisfied if:

$$\left|\det(\hat{\mathbf{\Sigma}}_x) - \det(\hat{\mathbf{\Sigma}}_0)\right| \gg 0$$

(5) Mauchly's Test Statistics (Original and Modified Forms):

Original Mauchly's statistic: $M^0_{stat} = -(n-1) * \ln(W_{stat})$

Modified Mauchly's statistic: $M^m_{stat} = -((n-1) - (((2*g)+1)/6)) * \ln(W_{stat})$

Where $g = 6$ time points, $n = 32$ cities

We perform these calculations below:

```
Sphericity_Test = function(data, alpha = 0.05) {
  # Covariance matrix
  cov_matrix = cov(data)

  # Determinant
  det_cov = det(cov_matrix)
  sqrt_det_cov = sqrt(det_cov)

  # Eignvalues
  eig = eigen(cov_matrix)

  # Trace (ideal)
  sigma2_hat = mean(diag(cov_matrix))
  cov_0 = diag(sigma2_hat, ncol(data))

  # Find W
  W_stat = det(cov_matrix) / det(cov_0)

  # Mauchly's test stat
  n = nrow(data)
  g = ncol(data)
  M_0_stat = -(n - 1) * log(W_stat)
  M_m_stat = -((n - 1) - (((2 * g) + 1) / 6)) * log(W_stat)

  # Chi-square values for original and modified version
  nu_0 = (g * (g - 1)) / 2
  nu_m = ((g * (g - 1)) / 2) - 1

  M_star_stat_0 = qchisq(1 - alpha, df = nu_0)
  M_star_stat_m = qchisq(1 - alpha, df = nu_m)

  # p-val
  p_value_0 = 1 - pchisq(M_0_stat, df = nu_0)
  p_value_m = 1 - pchisq(M_m_stat, df = nu_m)

  # output
  return(list(
    W_stat = W_stat,
    Mauchly_Statistic_Original = M_0_stat,
    Mauchly_Statistic_Modified = M_m_stat,
    Critical_Value_0 = M_star_stat_0,
    Critical_Value_M = M_star_stat_m,
    p_value_0 = p_value_0,
    p_value_M = p_value_m
  ))
}

Sphericity_Test(avo_prices)
```

```
## $W_stat
## [1] 0.001437335
##
## $Mauchly_Statistic_Original
## [1] 209.4389
##
## $Mauchly_Statistic_Modified
## [1] 195.2581
##
## $Critical_Value_0
## [1] 24.99579
##
## $Critical_Value_M
## [1] 23.68479
##
## $p_value_0
## [1] 0
##
## $p_value_M
## [1] 0
```

(6) Reject Sphericity: The resulting p-value is very small, so we reject the null hypothesis. This indicates a severe violation of the sphericity. We will correct using the Greenhouse–Geisser epsilon to adjust degrees of freedom downward, which increases the p-value and reduces Type I error risk for our RM ANOVA test.

Independence

For this analysis, we assume that the samples from each region are independent. In other words, the average avocado prices collected in each region are treated as coming from separate and unrelated populations. Since each city represents a distinct market with its own supply, demand, and pricing conditions, there's no direct overlap or dependency between them. While all regions may experience broader national trends, those effects are assumed to impact each region independently for the purpose of this ANOVA test.

Step 2: Compute RM-ANOVA

(1) Decompose Total Variability

In repeated measures ANOVA, the total variability in the data is partitioned into three main components:

- Between weeks (conditions) — SSM
- Between subjects (cities) — SSS
- Residual (error) — SSE
- Total — SST

(Note the associated degrees of freedom)

$$SSM = n\sum_{j=1}^{g}(\bar{x}_j - \bar{\bar{x}})^2, \quad df_M = g - 1$$

$$SSS = g \sum_{i=1}^{n} (\bar{x}_i - \bar{\bar{x}})^2, \quad df_S = n - 1$$

$$SSE = \sum_{i=1}^{n} \sum_{j=1}^{g} (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}})^2, \quad df_E = (n-1)(g-1)$$

$$SST = s^2_{\text{total}} \cdot (ng - 1), \quad df_T = ng - 1$$

Where:

- n = number of subjects (cities)
- g = number of repeated measures (weeks)
- x_{ij} = observed price for city i at week j
- $\bar{x}_i$ = mean price for city i across weeks
- $\bar{x}_j$ = mean price for week j across cities
- $\bar{\bar{x}}$ = overall grand mean

(2) Compute the F-statistic

$$F = \frac{MSM}{MSE} = \frac{SSM/(g-1)}{SSE/[(n-1)(g-1)]}$$

We perform these calculations below:

```
RM_ANOVA = function(df) {

  n = nrow(df)         # number of subjects
  g = ncol(df)         # number of repeated measures


  # Total variance
  grand_mean = mean(as.matrix(df))
  SST = sum((as.matrix(df) - grand_mean)^2)
  df_total = n * g - 1

  # Variance within SUBJECTS (SSS)
  subject_means = rowMeans(df)
  SS_subjects = g * sum((subject_means - grand_mean)^2)
  df_subjects = n - 1

  # Variance between groups  (SSM)
  condition_means = colMeans(df)
  SS_conditions = n * sum((condition_means - grand_mean)^2)
  df_conditions = g - 1

  # Variance within groups (SSE)
  SSE = SST - SS_subjects - SS_conditions
  df_error = (n - 1) * (g - 1)

  # Mean squares
  MS_subjects = SS_subjects / df_subjects
  MS_conditions = SS_conditions / df_conditions
  MS_error = SSE / df_error
```

```r
  MS_total = SST / df_total

  # F statistic (for the repeated measure factor)
  F_value = MS_conditions / MS_error

  # p-value
  p_value = 1 - pf(F_value, df1 = df_conditions, df2 = df_error)

  # Return Results
  ANOVA_Table = data.frame(
    Source = c("Subjects_S", "Conditions_M", "Error_E", "Total_T"),
    SS = c(SS_subjects, SS_conditions, SSE, SST),
    df = c(df_subjects, df_conditions, df_error, df_total),
    MS = c(MS_subjects, MS_conditions, MS_error, MS_total),
    F = c(NA, F_value, NA, NA),
    p_value = c(NA, p_value, NA, NA)
  )

  return(list(
    ANOVA_Table = ANOVA_Table
  ))
}

RM_ANOVA(avo_prices)
```

```
## $ANOVA_Table
##         Source         SS  df          MS        F   p_value
## 1   Subjects_S 3.27783410  32 0.102432316       NA        NA
## 2 Conditions_M 0.05485654   5 0.010971309 1.544695 0.1788911
## 3      Error_E 1.13641153 160 0.007102572       NA        NA
## 4      Total_T 4.46910218 197 0.022685798       NA        NA
```

(3) Interpret Results

$$F = 1.544695, \quad p = 0.1788911$$

At $\alpha = 0.05$, we fail to reject H_0, implying that average prices did not significantly change across the six weeks.

Step 3: Greenhouse–Geisser Correction

Since sphericity was violated, we correct the degrees of freedom using the Greenhouse–Geisser ε:

$$C = (I_{g\times g}) - \left(\frac{1}{g}\right) \cdot 1_{g\times g}$$

$$\hat{\Sigma}_c = C \cdot \hat{\Sigma}_x \cdot C$$

$$\hat{\varepsilon}_{GG} = \frac{\left[\text{tr}(\hat{\mathbf{\Sigma}}_c)\right]^2}{(g-1)\cdot \text{tr}(\hat{\mathbf{\Sigma}}_c^2)}$$

This correction reduces the effective degrees of freedom, producing a more conservative (larger) p-value. So it won't actually change our conclusion in this case. We compute the corrected F-statistic below:

```
Nonsphere_RM_ANOVA = function(df, F_value) {
  g = ncol(df)    # number of conditions
  n = nrow(df)    # number of subjects

  # Covariance matrix of the repeated measures
  Sigma_x = cov(df, use = "complete.obs")

  # Centering matrix
  I_g = diag(g)
  one_g = matrix(1, nrow = g, ncol = 1)
  C = I_g - (1 / g) * (one_g %*% t(one_g))

  # Transformed covariance matrix
  Sigma_c_hat = C %*% Sigma_x %*% C

  # Compute traces
  tr_Sigma_c = sum(diag(Sigma_c_hat))
  tr_Sigma_c_sq = sum(diag(Sigma_c_hat %*% Sigma_c_hat))

  # Greenhouse-Geisser epsilon
  epsilon_GG = (tr_Sigma_c^2) / ((g - 1) * tr_Sigma_c_sq)

  # New degrees of freedom
  df1 = g
  df2 = n

  df1_corr = epsilon_GG * df1
  df2_corr = epsilon_GG * df2

  # New p-vaue
  p_corr = 1 - pf(F_value, df1_corr, df2_corr)

  # Return results
  return(Corrected_p_val = p_corr)
}

Nonsphere_RM_ANOVA(avo_prices, 1.5447)
```

```
## [1] 0.2290195
```

After correction, we obtain the result:

$$p = 0.2290195 > \alpha$$

The adjusted p-value reinforces our earlier conclusion that avocado prices did not vary significantly over time. At $\alpha = 0.05$, we still fail to reject H_0. There is no statistically significant difference in average conventional avocado prices across the six weeks for the selected regions.