



MÉTODOS ALTERNATIVOS ANTE LA VIOLACIÓN DE SUPUESTOS EN DISEÑOS DE EXPERIMENTOS FACTORIALES

Alternative methods in presence of violation of assumptions in factorial experiment designs

Aldo Richard Meza Rodríguez ¹* 

¹ Facultad de Economía y Planificación, Universidad Nacional Agraria La Molina, La Molina 15024, Lima, Perú.

* E-mail: armeza@lamolina.edu.pe

Recibido: 27/10/2020; Aceptado: 02/10/2021; Publicado: 31/12/2021

ABSTRACT

The objective of this research was to compare and test different methods and alternatives to approach factorial designs with fixed effects, when the assumptions of normality or homogeneity of variances are not met. Twenty methods investigated in the literature as an alternative to classical ANOVA (variance analysis) were described and tested, including non-parametric techniques, robust methods, permutations, methods for heterogeneous variances and transformations; which are currently available and implemented in R software. The methods were tested in a 3A2B factorial design, where factor A was varieties of pineapple (Golden, Cayenne Lisa and Hawaiian), factor B was type of crop management (conventional and organic), and the response variable was the average percentage of Brix degrees. Among the proposed methods, 15 rejected the interaction hypothesis, and when comparing the type I error rates through simulations it was found that the permutations methods, the robust methods, the ART, van der Waerder and the BDM yielded error rates by below nominal value. When selecting ART as an alternative to perform the post hoc test, the best combination of treatments was the Lisa Cayena variety in organic management and the Hawaiian variety with conventional management, obtaining with these combinations percentages of Brix degrees above the average.

Keywords. Multiple comparisons | Factorial design | non-parametric methods | robust methods | permutations | R | transformations | violation of assumptions.

RESUMEN

El objetivo de esta investigación fue comparar y probar diferentes métodos y alternativas para abordar diseños factoriales con efectos fijos, cuando no se cumplen los supuestos de normalidad u homogeneidad de varianzas. Se describió y probó 20 métodos investigados en la literatura como alternativa al ANOVA (análisis de varianza) clásico, incluyendo técnicas no paramétricas, métodos robustos, permutaciones, métodos para varianzas heterogéneas y transformaciones; los cuales están disponibles e implementadas actualmente en el software R. Los métodos fueron probados en un diseño factorial 3A2B, donde el factor A fue variedades de piña (Golden, Cayena Lisa y Hawaiana), el factor B tipo de manejo del cultivo (convencional y orgánico), y la variable de respuesta el porcentaje promedio de grados brix. Entre los métodos propuestos, 15 rechazaron la hipótesis de la interacción, y al comparar las tasas de error tipo I mediante simulaciones se encontró que los métodos de permutaciones, los

métodos robustos, el ART, van der Waerder y el BDM arrojaron tasas de error por debajo del valor nominal. Al seleccionar el ART como alternativa para realizar la prueba post hoc, las mejores combinaciones de tratamientos fueron la variedad Lisa Cayena en el manejo orgánico y la variedad Hawaiana en el manejo convencional, obteniendo con estas combinaciones porcentajes de grados brix por encima de la media.

Palabras clave Comparaciones múltiples | diseño factorial | métodos no paramétricos | métodos robustos | permutaciones | R | transformaciones | violación de supuestos

Forma de citar el artículo (Formato APA):

Meza, A. (2021). Métodos alternativos ante la violación de supuestos en diseños de experimentos factoriales. *Anales Científicos*. 81(2), 318-335. <http://dx.doi.org/10.21704/ac.v82i2.1795>

Autor de correspondencia (*): Aldo Richard Meza Rodríguez. Email: armeza@lamolina.edu.pe

© Los autores. Publicado por la Universidad Nacional Agraria La Molina.

This is an open access article under the CC BY

1. INTRODUCCIÓN

En diferentes campos de investigación donde se usan diseños experimentales, con frecuencia se violan los supuestos de normalidad y homogeneidad de varianzas. Luepsen (2018) manifiesta que en general, los investigadores usan las versiones paramétricas (ANOVA) sin verificar los supuestos, asimismo, Konietschke et al. (2020) resalta que las pruebas F correspondientes solo son válidas para datos que se distribuyen normalmente y que tienen varianzas iguales, los cuales son supuestos que frecuentemente no se cumplen. Para Saste et al. (2016), si se utilizan diseños factoriales ignorando los supuestos básicos, los resultados pueden ser mal interpretados e inducir a diferentes errores, también Friedrich et al. (2017) explican que en la práctica los supuestos para diseños factoriales son difíciles de verificar y, a menudo, no se cumplen y su infracción generalmente infla los errores tipo I o II de la prueba F del ANOVA. Esto también concuerda con Thongteeraparp (2019), el cual detalla que, si no se cumplen los supuestos, el error tipo I se desviará del nivel nominal y esto reducirá la potencia de la prueba.

Para diseños experimentales completamente al azar simples o en bloques, o cuando las variables son de tipo ordinales, scores o escalas, existen pruebas alternativas no paramétricas tales como la prueba de Kruskal Wallis o la prueba de Freedman, sin embargo, para diseños factoriales donde se muestran los efectos principales y también los de interacción, la literatura no muestra un método exacto para contrarrestar la asunción de los supuestos. Algunos autores como Gao & Alvo (2005) declaran que cuando no se cumple la

normalidad de la distribución de errores o se desconoce la distribución de errores subyacente, se necesitan métodos alternativos o pruebas no paramétricas para proporcionar inferencias sólidas.

Existen algunas pruebas con bastante precisión para diseños factoriales que controlan el error tipo I cuando no se cumple la homogeneidad de varianzas, sin embargo, estas pruebas no son exactas ante la falta del cumplimiento del supuesto de normalidad. Con el pasar de los años, se han propuesto una variedad de pruebas (generalmente basada en rangos) para probar la interacción en un diseño factorial (De Neve & Thas 2017).

Es importante aclarar que existen situaciones obvias, donde los resultados no siguen una distribución normal por la propia naturaleza de la variable (variable ordinal o un rango), métodos libres de distribución que no se basan en los supuestos restrictivos de las pruebas paramétricas y también cuando hay valores atípicos definidos propios del estudio donde se necesitan pruebas robustas, por lo tanto, no solamente existen pruebas no paramétricas sino también otros procedimientos y métodos en vez el tradicional. Para Feys (2016), los principales métodos para contrarrestar los problemas de asunción de supuestos son: simulaciones, pruebas no paramétricas (pruebas de rangos), procedimientos robustos (para tratar valores atípicos), transformación de datos, y remuestreo, aunque diferentes autores se han acostumbrado a usar el término no paramétricos en forma genérica para todas las situaciones.

Uno de los motivos por lo cual casi no aparece en la literatura latina métodos alternativos a los diseños factoriales es la falta de implementación de las pruebas en softwares comerciales, puesto que, dependiendo de la prueba, los cálculos pueden ser en algunos casos complejos. Recién en el año 2012 aparecen en R las primeras funciones para probar la interacción ante la asunción de supuestos, y ya para el 2020 se ha incrementado la cantidad de pruebas alternativas las cuales se desarrollarán en esta investigación.

2. MATERIALES Y MÉTODOS

Materiales empleados

Los datos fueron obtenidos de un estudio experimental realizado en el departamento de Junín, en la provincia de Chanchamayo, en el distrito de Pichanaki, durante un periodo de 14 meses en el año 2019, trabajando con 3 variedad de piña (Golden, Cayena Lisa y Hawaiana) y 2 tipos de manejo (Orgánico y convencional), evaluando como variable de respuesta el porcentaje promedio de grados brix por cada combinación de tratamientos. Para el procesamiento de datos fue indispensable el uso del software R, en su versión 4.0.2, utilizando diferentes funciones por defecto en el R, las cuales están formalmente en diferentes librerías, así como otras funciones propias y de otros autores que aún no las formalizan pero que están disponibles en plataformas online.

Metodología

El presente estudio se desarrolló mediante la formulación de diferentes procedimientos, los cuales abarcan métodos basados en la mediana, en rangos, transformaciones, métodos robustos y permutaciones.

a) Métodos basados en la mediana

Estos métodos pertenecen al área de la estadística no paramétrica, los cuales construyen el test tomando como medida de referencia la mediana y la aplicación del estadístico chi-cuadrado.

El método de Wilson

Esta metodología fue propuesta por Wilson (1956), quien construyó un test de distribución libre para probar la hipótesis de efectos principales y de interacción, semejante al ANOVA tradicional de dos o más factores para diseños equilibrados o

desequilibrados. Según Jimenes & Pérez (1989) en esta metodología se expuso las bases teóricas, su descripción matemática, su forma y ámbito de aplicación, así como la relación con otras pruebas análogas.

La técnica se basa en la categorización de las variables a partir de la mediana, para luego contrastar la significación de los efectos del diseño a través del estadístico χ^2 . Cada observación de la combinación de tratamientos se divide en valores por encima y por debajo de la mediana general, a partir de ello se construye la tabla de contingencia, formulándose las condiciones de filas y columnas en el diseño factorial. Algunas críticas para esta técnica, las mencionó Sawilowsky (1990), indicando que en ciertas situaciones no es muy poderosa, también Jimenes & Pérez (1989) manifestaron que era algo tediosa, sin embargo, actualmente se puede subsanar esto construyendo la función fácilmente en el R.

Test de la mediana extendida

Esta prueba fue propuesta por Shoemaker (1986), siendo también una extensión de la prueba de la mediana, pero para diseños factoriales de 2 vías. Este procedimiento realiza recuentos en las celdas de una columna que están por encima o por debajo de la mediana de la fila, donde las frecuencias resultantes se presentan en una tabla de contingencia para luego ser contrastadas con el estadístico Chi-cuadrado de Pearson.

Para Sawilowsky (1990), esta prueba tiene ciertas dificultades para preservar el alfa nominal para efectos nulos en muestras pequeñas, sin embargo, la prueba se mejora restando la mediana de la columna de cada observación, demostrando ser robusta y una poderosa alternativa a la prueba Anova.

b) Métodos basados en rangos y datos ordinales

Estas pruebas son consideradas no paramétricas y su principal característica es que utiliza rangos antes de realizar el ANOVA.

Test de Sheirer

Scheirer (1976) formuló en forma teórica y mediante simulaciones un procedimiento de extensión derivada de la prueba de Kruskall-Wallis para el cálculo de efectos de interacción y contrastes lineales. En esta

metodología, las interacciones se obtienen mediante el uso del ANOVA aplicado a los rangos como si fueran independientes entre sí, de tal forma que al multiplicar la estadística obtenida con los cálculos del ANOVA con un factor de corrección, la distribución nula se convierte asintóticamente en una distribución Chi-cuadrado llamada H. Esta prueba usa el numerador de la suma de cuadrados de los rangos (SCRank), resultando el estadístico H para los efectos principales y para la interacción de la siguiente forma:

$$H_A = SCRank_A * (12/N * (N + 1))$$

$$H_B = SCRank_B * (12/N * (N + 1))$$

$$H_{AB} = SCRank_{AB} * (12/N * (N + 1))$$

Una de las críticas a este método las presentaron Toothaker & Chang (1980), quienes afirmaron que el test no puede controlar el alfa en el valor establecido, ni son muy poderosos en presencia de efectos distintos a los que se están probando.

Test de Gao y Alvo

Gao y Alvo (2005) desarrollaron una prueba de rango para la interacción en diseños factoriales, válida según el modelo de cambio de ubicación. Esta prueba compara la suma de los rangos de filas con la suma de los rangos de las columnas. Los autores afirmaron que este enfoque intuitivo funciona bien con y sin la presencia de efectos principales en diferentes distribuciones de errores.

Para De Neve & Thas (2017), este estadístico de prueba es una forma cuadrática generalizada de una combinación lineal de los rankings de los rangos de A y B sobre la interacción. Este método ha recibido cierta crítica por Luepsen (2018), el cual afirma que tiene una buena potencia de prueba, especialmente para muestras pequeñas, pero a costa del control del error tipo 1. La prueba se encuentra en su versión anterior en la función *interaction.test* de la librería StatMethRank, creada por Quinglong (2015).

Pruebas de Puri y Sen (estadística L)

Esta prueba de distribución libre fue propuesta por Puri y Sen (1985), y es conocido como el estadístico L. Es una extensión de la prueba H de Kruskal-Wallis y la prueba de Friedman, basa en la distribución chi-cuadrado. En esta prueba primero se rankea la variable de respuesta, luego la agrupación de los factores A, B

y AB se generalizan mediante las relaciones χ^2 de la siguiente manera:

$$\chi_A^2 = \frac{SC_A}{CM_{Total}} ; \chi_B^2 = \frac{SC_B}{CM_{Total}} ; \chi_{AB}^2 = \frac{SC_{AB}}{CM_{Total}}$$

Sawilowsky (1990) refiere que se han realizado estudios de Monte Carlo de la disposición de ANOVA factorial, encontrándose que la prueba L era robusta y poderosa para detectar pequeños efectos de tratamiento. Una desventaja de esta prueba la presenta Luepsen (2018) el cual afirma que la potencia puede verse afectada para cualquier efecto en el caso de que haya otros efectos no nulos en el modelo, esto a razón de que en el ANOVA estándar, el denominador de los valores F es el cuadrado medio del residual que se reduce por los efectos de otros factores en el modelo, en cambio, el denominador de las pruebas χ^2 del estadístico L es el cuadrado medio del total que aumenta con los efectos de los otros factores; reduciendo así la relación del efecto considerado y, por lo tanto, también la relación χ^2 más pequeña. Las pruebas se encuentran en la función *np.anova*, creada por Luepsen (2020), aún no formalizada en una librería pero con código libre disponible en la web.

Prueba van der Waerden

van der Waerden (1953) propuso una prueba alternativa al ANOVA factorial de Kruskal-Wallis, realizando una transformación normal inversa de la siguiente forma:

$\Phi^{-1} = (R(y)/(n + 1))$, donde $R(y)$ son los rangos de la variable dependiente y n, el número de observaciones. Mansouri y Chang (1995) generalizaron la prueba original de van der Waerden a diseños con varios factores de agrupación. Los procedimientos son similares al cálculo del estadístico L de Puri y Sen, con la diferencia de que los rangos R (y) se transforman en puntajes normales como se describe para el método de transformación normal inversa, antes de calcular la suma de cuadrados y razones χ^2 . Las pruebas se encuentran en la función *np.anova*, creada por Luepsen (2020), aún no formalizada en una librería pero con código libre disponible en la web.

R ANOVA

Hocking (1985) describe un algoritmo utilizado para reducir la dispersión para modelos factoriales con efectos principales y de interacción. Esta prueba es basada en rangos y es una alternativa robusta a los mínimos cuadrados en modelos lineales. Kloke & McKean (2020) desarrollaron esta prueba y la programaron en R a través de la función *raov* de la librería *Rfit*.

ATS y WTS

El ATS (estadístico tipo ANOVA) y el WTS (estadístico tipo Wald) son pruebas basadas en rangos medios para contrastar muestras mediante la comparación de los efectos relativos, los cuales hacen referencia a las distribuciones de las variables evaluadas en el experimento.

Akritas et al. (1997) formularon el ATS (estadístico tipo ANOVA) el cual se distribuye aproximadamente como una F, mientras que Brunner et al. (1997) propuso el WTS (estadístico tipo Wald) el cual se distribuye asintóticamente como una χ^2 . Estas pruebas pueden utilizarse para datos de conteo, datos ordinales y puntuaciones, sin embargo, para Brunner et al. (2018) una de las desventajas de estos métodos radica en el tamaño de celdas (combinación de tratamientos), siendo el WTS ideal para muestras grandes, puesto que en esta prueba se calcula la matriz de covarianzas completa y la inversa generalizada basada en este estimador, lo cual conduce a una mala aproximación cuando los tamaños son pequeños, por el contrario el ATS es ideal para tamaños de muestras pequeñas y moderados. Las pruebas se encuentran en la función *rankFD* de la librería *rankFD* en el R, disponible en su última versión (Konietschke et al., 2020).

RT

El método RT (rango transformado) fue sugerido por primera vez por Conover & Iman (1981), el procedimiento simplemente consiste en rankear la variable dependiente original (en caso de empates dichos rangos se promedian), para posteriormente usar la prueba paramétrica F. Al trabajar con rangos este procedimiento resulta un diseño factorial no paramétrico. Algunos autores como, Higgins & Tashtoush (1994) encontraron que este proceso puede producir resultados inexactos para los efectos de interacción. También Beasley y Zumbo (2009),

hallaron que en esta prueba la tasa de error tipo I de la interacción puede llegar más allá del nivel nominal si hay efectos principales significativos dado que los efectos se confunden.

ART

Ante los problemas proporcionados por el método original RT, Higgins & Tashtoush (1994) proponen la prueba ART (rango alineado transformado), la cual permite que el tratamiento no paramétrico sea más preciso para los efectos principales y la interacción. El ART es un procedimiento de alineamiento, el cual consiste en restar los efectos que no son de interés inicial antes de realizar el ANOVA. Wobbrock *et al.* (2011) describen que este procedimiento calcula primero los residuales a través de un modelo de regresión o también como diferencias de los promedios de celda, luego suma el efecto de interés, transforma la suma en rangos y por último contrasta la prueba F del ANOVA común.

Según Saste et al. (2016), ante los buenos resultados de la prueba, muchos investigadores sugirieron la transformación de rango alineado como un método robusto para experimentos factoriales no paramétricos e incluso se hicieron algunos ajustes. Esta prueba se encuentra en la función *art* o *anova.art* del paquete ARTool disponible en su última versión (Kay & Wobbrock, 2020).

c) Métodos basados en varianzas heterogéneas

Estas pruebas son potentes especialmente cuando no se cumple el supuesto de homogeneidad de varianza, sin importar si se cumpla o no el supuesto de normalidad de los errores.

El test de Brow-Forsythe

Brown & Forsythe (1974) desarrollaron una prueba F producto del análisis de varianza de las desviaciones de la media, mediante la transformación de la variable de respuesta, esta prueba fue formulada específicamente para varianzas heterogéneas para una y dos vías agrupadas. Según Delgado (1992), este método robusto reemplaza el denominador del estadístico F ANOVA por una estimación de la varianza del numerador, la cual corresponde a la "S" de Sheffe, sustituyendo el CME global por el estimador de la varianza concreta de la población, teniendo en cuenta así, el tamaño del grupo y el coeficiente asignado a ese grupo en la

combinación lineal base del contraste. Para este método el cálculo de la interacción es relativamente compleja, haciendo uso de contrastes normalizados y ortogonales. La prueba se puede desarrollar en R a través de la función *bff* de Luepsen (2020).

Test WJ (Welch-James)

Esta prueba fue descrita por Algina & Olejnik (1984), sin embargo, requiere de cierta programación algo compleja. Este procedimiento puede usarse para probar hipótesis sobre medias, cuando se dispone de muestras independientes en poblaciones con varianzas heterogéneas.

Según Vallejo et.al (2010) algunos autores como Keselman et al. (1995), usando métodos numéricos dedujeron que el desempeño del test WJ eran adecuados cuando los datos se distribuían normalmente, excepto cuando el tamaño de las celdas era reducido, mientras que cuando los datos estaban sesgados, el método de WJ requería aumentar el tamaño de las celdas para controlar las tasas de error. También más adelante sugirieron corregir los efectos ocasionados por la falta de normalidad, usando el enfoque WJ con estimadores robustos de localización y escala. La prueba se encuentra en el R a través de la función *wj.anova* de Luepsen (2020).

Test BDM

La prueba BDM (Brunner-Dette-Munk) fue propuesta como alternativa cuando no se cumple el supuesto de homogeneidad de varianzas y se basan en permutación de rangos heterocedásticos a través de álgebra matricial. Brunner et al. (1997) inicialmente presentaron una versión heterocedástica de la prueba de Kruskal-Wallis que utiliza la distribución F para una sola vía, y Wilcox (2017) en su última versión describe una extensión bidireccional. Esta prueba es poderosa y asume heteroscedasticidad la cual permite probar los efectos simples e interacciones principales, tal como lo resalta Aho (2013), manifestando que esta prueba al ser robusta a la no normalidad y la heterocedasticidad, permite cálculos de los P-values exactos no tan complejos debido a los valores vinculados. Esta prueba se encuentra en la función *BDM.2way* de la librería *asbio*, en su última versión creada por Aho (2020).

d) Pruebas basadas en métodos robustos

Estos métodos se usan cuando no se cumplen los supuestos de normalidad a causa de valores atípicos definidos, o tal como lo declara Feys (2016), cuando los resultados tienen claros límites de detección. Ante tales situaciones optar por transformaciones u otros métodos resulta inapropiado, siendo la solución probar con procedimientos avanzados conocidos como métodos robustos. Entre los principales autores, Wilcox (2017) presenta algunos métodos robustos sólidos independientemente de la distribución de los errores; conocidos como *Medias recortadas*, *prueba de la mediana* y *M-estimadores*. Estos métodos robustos se utilizan específicamente cuando los valores atípicos afectan al supuesto de normalidad.

Medias recortadas (Trimmed means)

Este método también conocido como medias truncadas descarta los extremos altos y bajos en igual proporción, para luego calcular la media de una determinada distribución de probabilidad. El estimador recortado se obtiene de muestras limitadas al rango $[X_{[n\%]} - X_{[1-n\%]}]$, o como un número fijo de puntos descartados.

Pruebas de mediana (Median test)

Este método utiliza la mediana como medida robusta de tendencia central, evitando la influencia de valores atípicos.

M-Estimadores (M-estimators)

Son estimadores extremos, donde la función objetivo es un promedio de la muestra, estos estimadores pertenecen a una clase general de estadísticas robustas que se obtienen como método de la suma mínima de las funciones de los datos, por ejemplo, los mínimos cuadrados re-ponderados iterados. Adam & Bejda (2018) detallan que la ventaja de estos estimadores es su punto de ruptura, que es el más alto que puede alcanzar un estimador equivalente. Las funciones en R para estos tres métodos robustos se encuentran en *t2way*, *med2way*, *pbad2way* de la librería *WRS2* (Mair y Wilcox, 2020).

e) Pruebas basadas en permutaciones

Mewhort et al. (2010) define a estos métodos como procedimientos que se derivan directamente de experimentos comparativos, los cuales no dependen de una distribución conocida, quienes usan todas las permutaciones posibles de la variable dependiente,

manteniendo fijadas los factores, resultando a veces mucho más sensible que las pruebas paramétricas. Manly (2007) presentó cinco métodos de permutación para interacciones, de los cuales en esta investigación se presentan 3.

Permutación bajo el enfoque de Manly

Este enfoque presentado por Manly (2007) asigna las observaciones aleatoriamente a los tratamientos factoriales sin restricciones, conservando el tamaño de la muestra original para cada tratamiento, siendo las distribuciones de estas permutaciones las que se utilizan para hallar el estadístico F en el factor A, B y la interacción AB. Luego se compara la F obtenida con dicha distribución y se calcula el porcentaje de réplicas bajo la hipótesis nula cuando el remuestreo de las Fs superan a la F obtenida; de la misma forma se procede con los efectos principales.

Permutación de residuos de Still y White

Still y White (1981) utilizan aleatorización restringida en los efectos principales, controlando dichos efectos mediante los residuos (lo que queda después de eliminar los efectos de fila y columna). El enfoque se basa en restar esos efectos y usar los residuos resultantes, para probar la **interacción** sin ser afectada por los efectos principales.

Permutación con el método de Ter Braak

Ter Braak (1992) también utilizó procedimientos de aleatorización en forma similar al enfoque de Still y White, con la diferencia que, al calcular la interacción, los residuos se obtienen sobre el diseño completo en vez de solo el modelo aditivo. Este enfoque elimina la media de la celda de cada observación, creando un conjunto de residuos para luego permutar dichos residuos en todas las celdas del diseño. Para Mewhort et al. (2010) una desventaja de los métodos de permutación completa es el costo computacional (toma más tiempo que cualquier otro método).

Los tres métodos presentados se encuentran detallados por Howell (2013), el cual desarrolló funciones detalladas en R. También en la función *perm.fact.test* de la librería *asbio* se encuentra el enfoque de Manly. Otros métodos de permutación para diseños factoriales se encuentran en la función *lmp* de la librería *lmPerm* (Torchiano, 2016), también en la función *ezPerm* de la librería *ez* (Lawrence, 2015) y la

función *aovperm* de la librería *Permuco* (Frossard y Renaud, 2019).

f) Pruebas basadas en transformaciones

Cuando no se cumplen los supuestos de normalidad u homogeneidad de varianzas una de las opciones que han realizado diferentes investigadores es aplicar transformaciones, donde el objetivo es cambiar la métrica de la variable original por una medida en otra escala (Lawson, 2015). En el transcurso del tiempo se han propuesto diferentes transformaciones, tales como la transformación de la raíz cuadrada, logaritmo natural, transformaciones angulares, transformación Box Cox, entre otras (Melo et. al., 2020).

La transformación Box Cox

Box & Cox (1964) presentaron una técnica de transformación de datos de potencia paramétrica la cual tiene como objetivo reducir diferentes anomalías tales como, no normalidad, heterocedasticidad, y no aditividad. Las transformaciones originales de Box Cox fueron formuladas de la siguiente manera:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{Si } \lambda \neq 0 \\ \log(\lambda) & \text{Si } \lambda = 0 \end{cases}$$

Este método ha recibido algunas críticas por la comunidad científica, por ejemplo, Pelea (2018), resume que, las transformaciones no siempre logran el cumplimiento de los supuestos de los análisis paramétricos, también Ribeiro-Oliveira et al. (2018) muestra que, desde el punto de vista teórico, el problema radica en el procedimiento matemático, el cual puede modificar la distribución original de los datos, mientras que, desde el punto de vista práctico, los investigadores presentan dificultades para discutir e interpretar los resultados en escalas diferentes a la original. En R se encuentra la función *boxcox* de la librería *MASS*, por Ripley (2020) y la librería *trafo* de Kreuzmann (2018) donde se encuentran las funciones *diagnostics* y *trafo_lm* para diferentes opciones de transformaciones.

Modelo Aditivo Lineal

A continuación, se presenta el modelo aditivo lineal con cada uno de sus componentes:

$$Y_{ijk} = \underbrace{\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}}_{\mu_{ij}} + \varepsilon_{ijk} \quad ; \text{ donde:}$$

Y_{ijk} , Porcentaje de grados brix obtenida con la variedad i, en el tipo de manejo j y la repetición k; μ , el efecto del porcentaje de grados brix medio general; α_i , el efecto de la variedad de piña i; β_j , el efecto del tipo de manejo j; $(\alpha\beta)_{ij}$, el efecto de la interacción de la variedad de piña i con el tipo de manejo j; μ_{ij} , el efecto de la media de la combinación (tratamiento) ij y; ε_{ijk} , el efecto del error experimental obtenido con la variedad de piña i, manejo j y repetición k.

$$\text{Tasa de error tipo I} = \frac{\text{n}^\circ \text{ de veces que se rechaza } H_0, \text{ siendo } H_0 \text{ verdadero}}{1000} \times 100$$

3. RESULTADOS Y DISCUSIÓN

La tabla 1 resume el porcentaje promedio de grados brix obtenido al aplicar los tratamientos. Se observa que el diseño es balanceado con 4 repeticiones por celda para la combinación del factor A (Variedades de piña) y el factor B (Tipo de manejo).

Para el comportamiento de los factores (Figura 1) hay indicios de tendencia a la interacción, con lo cual se sospecha que, en los modelos planteados la interacción es significativa. También hay evidencia de cierta variabilidad puesto que, las líneas muestran amplitud en algunas combinaciones de tratamientos.

En la Figura 2 la gráfica de Residuales vs Fitted, revela que la distribución de los residuos se concentra alrededor de 0 para valores ajustados grandes, pero se dispersan a medida que disminuyen los valores ajustados, lo cual indica que la varianza es decreciente y no constante en todo el rango. En la gráfica Normal Q-Q, los residuos se desvían de la línea diagonal tanto en la cola inferior como superior, por lo tanto, estas colas tienen valores más grandes de lo que se esperaría en un modelo normal estándar. Al analizar formalmente los supuestos (Tabla 2) las pruebas de Shapiro Wilk y Anderson Darling, utilizadas por su potencia estadística (Wijekularathna et al., 2019), arrojaron p-valores altamente significativos con lo cual se demuestra que no se cumple el supuesto de

Simulación y error tipo I

Varios autores han utilizado simulaciones para estimar las tasas de error tipo I, e incluyo la potencia de la prueba, ver por ejemplo a Brunner et.al (2016), Thongteeraparp (2019), y Afonso & Pereira (2019). En este estudio, para evaluar los modelos se realizó simulaciones con los mismos parámetros del experimento propuesto. Se realizó 1000 réplicas para diferentes tamaños de celdas (n=5, 10, 15, ..., 50) considerando diseños balanceados (igual número de repeticiones en cada celda), considerando como principal intensidad el efecto de la interacción. Se tomó como alfa nominal 5%, evaluando el error tipo I para cada modelo de la siguiente manera:

normalidad de los errores. De la misma forma, la prueba Breusch-Pagan, recomendada por su buena capacidad de detección (Andriani, 2017) y la prueba de Bartlet en interacción (Kassambara, 2019) dan como resultado P-valores muy pequeños con lo cual se concluye que tampoco se cumple el supuesto de homogeneidad o varianza constante.

Los resultados de las pruebas para los efectos principales y la interacción en cada uno de los métodos propuestos (Tabla 3) arrojan buenos resultados para el ANOVA tradicional, donde el p valores para la interacción, así como los efectos principales resultaron significativos, sin embargo, este resultado no sería el adecuado por no cumplirse los supuestos. Casi todas las pruebas alternativas mostraron alta significancia estadística para el factor Variedad, y algunas para el factor manejo, sin embargo, la idea de un diseño factorial es probar la interacción, con lo cual, las pruebas basadas en métodos robustos, en permutaciones, y la transformación Box Cox tuvieron éxito al probar la hipótesis de interacción, sin embargo, en cuanto a las pruebas basadas en rangos, solo la prueba ART, R ANOVA y Gao y Alvo obtuvieron P-valores significativos. Todos los resultados a mayor detalle se encuentran en el siguiente link: <https://rpubs.com/MezaStat/nonparametric>

Tabla 1. Distribución de las unidades experimentales a cada combinación de tratamientos.

Tipo de manejo	Variedades de piña					
	V1 (Golden)		V2 (Cayena Lisa)		V3 (Hawaiana)	
S1 (Orgánico)	16.24	16.14	14.74	15.08	12.56	9.52
	15.56	15.8	14.39	15.18	8.3	10.51
S2 (Convencional)	15.36	15.31	12.02	12.41	9.87	10.8
	15.56	15.65	11.68	11.43	11.58	10.7

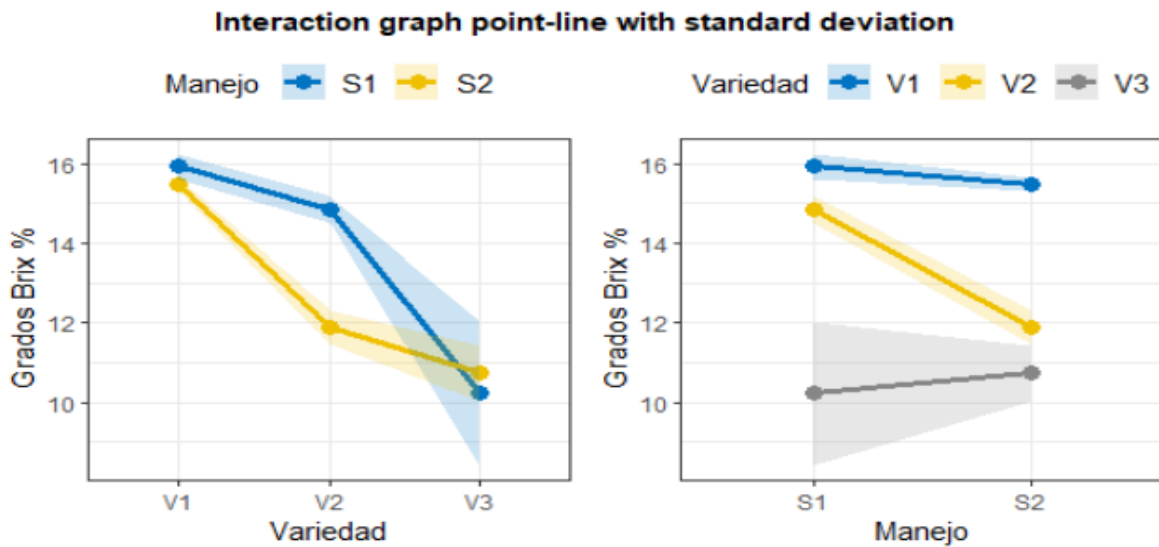


Figura 1. Interacción entre los dos factores.

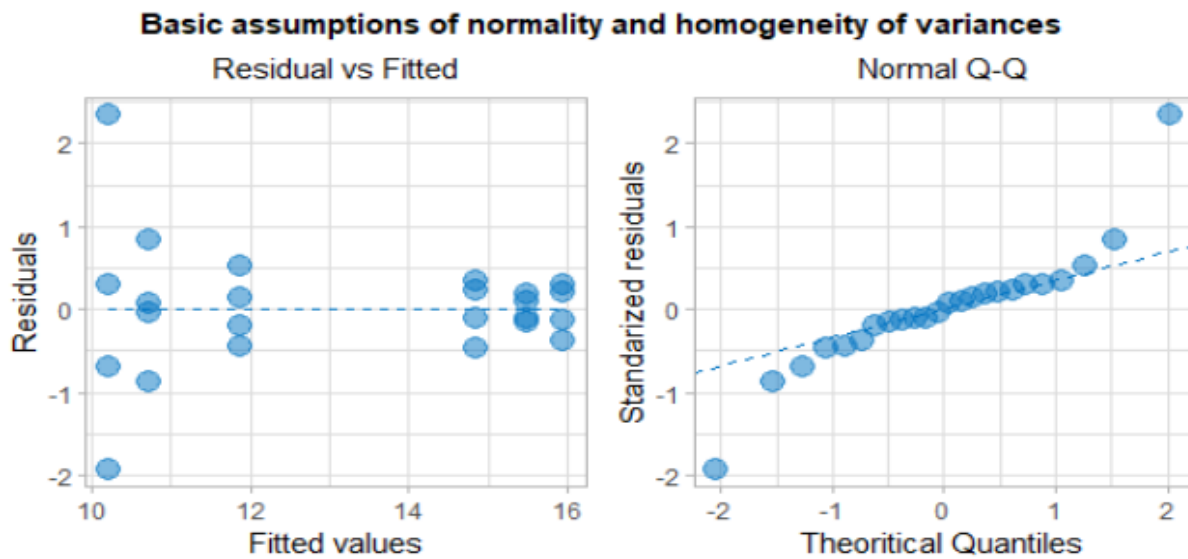


Figura 2. Asunción de supuestos a través de los residuales.

Tabla 2. Pruebas de normalidad y homogeneidad de varianzas en el modelo.

Tipo de prueba	Test	Estadístico	Pvalor	Sig.
Test de normalidad de los errores	Shapiro-Wilk	W=0.86929	0.005095	**
	Anderson-Darling	A=1.1329	0.004627	**
Test de varianza constante (homogeneidad)	Breusch-Pagan test	Chisq.=14.265	0.000159	***
	bartlett.test interaction	Q=18.812	0.002084	**

Los asteriscos indican la significancia: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 '.' 1

Tabla 3. P-valores para los efectos principales e interacción en las diferentes pruebas.

Métodos	Pvalores Variedad	Pvalores Manejo	Pvalores Interacción
Prueba ANOVA tradicional			
ANOVA	0,00000 ***	0,01056 *	0,00172 **
Pruebas basadas en la mediana			
Wilson	0,00034 ***	0,10247	0,06948 ,
Shoemaker	1,00000	1,00000	0,00248 **
Pruebas basadas en rangos			
Sheirer	0,00007 ***	0,37085	0,68843
Pury and Sen	0,00231 **	0,51560	0,68832
van der Waerder	0,00130 **	0,37102	0,5863
Gao y Alvo	0,45103	0,00942 **	0,00743 **
RT	0,00000 ***	0,02569 *	0,09027 ,
ART	0,00000 ***	0,00149 **	0,00033 ***
R ANOVA	0,00000 ***	0,00279 **	0,00021 ***
WTS	0,00000 ***	0,01502 *	0,09161 ,
ATS	0,00003 ***	0,04371 *	0,13381
Pruebas basadas en varianzas heterogéneas			
Brow-Forsythe	0,00000 ***	0,34789	0,02242 *
Wels James	0,00001 ***	0,02151 *	0,00001 ***
BDM	0,00002 ***	0,04371 *	0,13307
Pruebas basadas en métodos robustos			
Trimmead means	0,00100 **	0,02600 *	0,00100 **
Medians	0,00000 ***	0,00240 **	0,00000 ***
M estimator	0,00000 ***	0,01840 *	0,00000 ***
Pruebas basadas en permutaciones			
Manly	0,00000 ***	0,01040 *	0,00280 **
Still_White	0,00000 ***	0,01200 *	0,00160 **
ter_Braak	0,00000 ***	0,01560 *	0,00000 ***
Transformación Box Cox			
BoxCox	0,00000 ***	0,00001 ***	0,00003 ***

Los asteriscos indican la significancia: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 '.' 1

En las Figura 3 y 4 mediante las simulaciones se estimó la tasa de error para cada método propuesto en la interacción, se encontró que, los métodos de Wilson, la mediana extendida de Shoemaker y el test de Sheirer contienen tasas de error tipo I altas en todo el recorrido de la serie de celdas, también otros métodos como Puri y Sen, RT test, R ANOVA test tienen picos de subida y bajada sobre el límite del 5% de tasa de error, por otro lado, el ATS muestra tasas de error bajas para tamaños de celdas pequeñas, sin embargo el error crece, mostrando baja eficiencia conforme el tamaño de celdas va incrementado, confirmando lo dicho por

Brunner et al. (2018); también se observa en los métodos WTS test, WJ (Wels-James), Brow-Forsythe, Median Test, M-estimadores, BDM test, como al inicio tiene una tasa de error no despreciable, pero la tasa disminuye conforme los tamaños de celdas aumentan. Finalmente hay un grupo de métodos como es el caso de las permutaciones (Still White, Ter Braak, Manly), el ART, Trimmed means, van der Waerder y Box Cox los cuales mantienen tasas de error baja para todos los diferentes tamaños de celdas. Estos resultados son similares a los hallazgos de Thongteeraparp (2019) para algunos métodos, pero para tablas de 2x2 y 3x3.

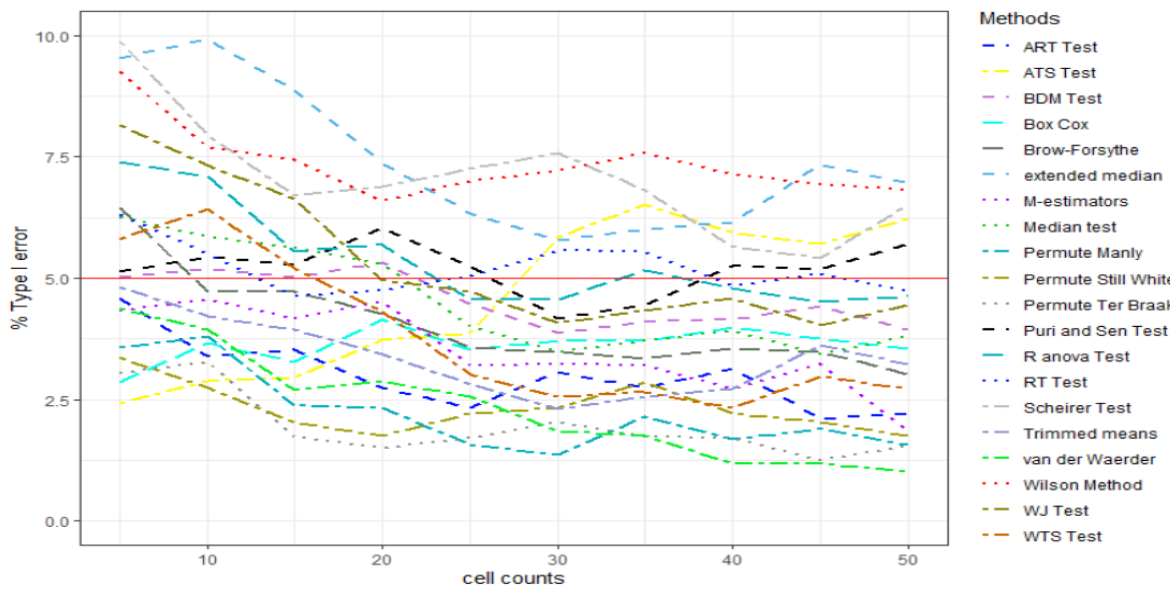


Figura 3. Tasas de error tipo I para la interacción en los métodos evaluados con diferentes tamaños de celdas en diseños balanceados.

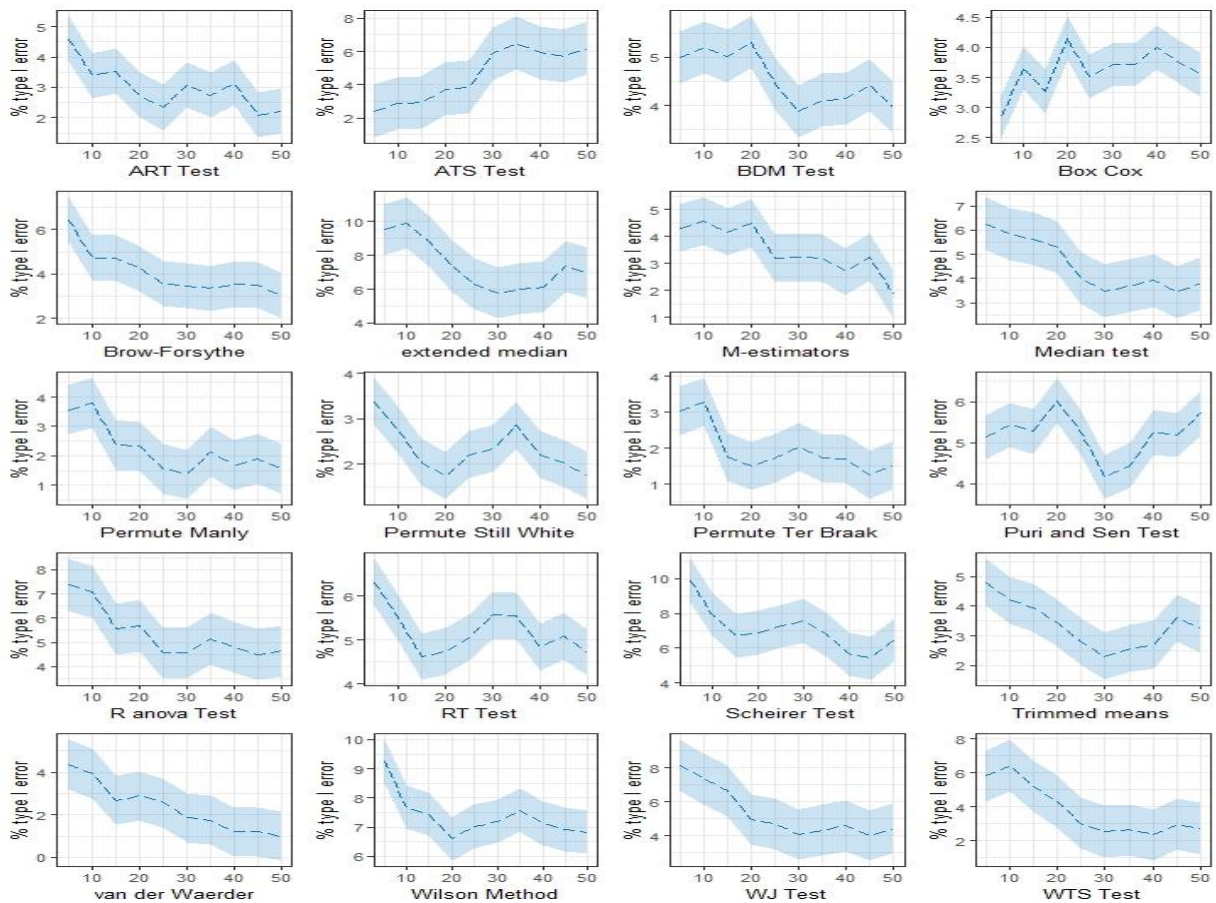


Figura 4. Tasas de error tipo I para la interacción para los métodos evaluado (por separado).

La Tabla 4 resume los valores exactos de los errores tipo I para cada método, los valores en negrita, son tasas que sobrepasan el 5% de error, esto concuerda con las figuras 3 y 4, mostrando una estructura similar da los hallazgos que realizaron Alfonso y Pereira (2019) para datos con tamaños de 3, 5 y 10 celdas en distribuciones no normales.

Puesto que este estudio no se centra en comparar todas las técnicas mediante análisis post-hoc, se eligió un método para ilustrar los resultados y determinar la mejor combinación de tratamientos, el método elegido es el ART (también se puede elegir otros métodos probados como el Trimmed means, alguna permutación, entre otras mencionadas que resultaron ser significativas y poseen una tasa de error baja) por tener la interacción significativa (Tabla 3), poseer una tasa baja de error tipo I baja en diferentes tamaños de celdas (Figura 4 y Tabla 4), y por su gran uso y extensión en los últimos años (Saste et al., 2016).

La Figura 5 muestra las comparaciones mediante la prueba de Tukey utilizando el mismo procedimiento descrito por Mangiafico (2016) para la comparación

con el método ART. Si se quisiera analizar los factores por separado, para el efecto del factor “variedad de Piña”, los resultados del experimento (parte lateral izquierda) muestran a la variedad V1 (*Golden*) como el tratamiento que permite obtener mayor porcentaje de grado brix, de igual manera en la parte lateral derecha el tipo de manejo S1 (*Orgánico*) obtiene el mayor porcentaje de grados brix en ese diseño.

Si bien es cierto los efectos principales pueden ser de interés para el investigador, el objetivo de un diseño de experimentos factoriales es analizar en forma exhaustiva la interacción de los factores, esto se logra con el análisis de efectos simples (Figura 6). Para los efectos simples del tipo de manejo sobre la variedad se tiene: en la variedad V1 (*Golden*), a un nivel de significancia del 5%, no existe diferencia significativa entre los dos tipos de manejo; para la variedad V2 (*Cayena Lisa*), existe diferencia altamente significativa entre los tipos de manejo S1 y S2; y para la V3 (*Hawaiiana*) también existe diferencia significativa entre los dos tipos de manejo.

Tabla 4. Tasas de error tipo I para la interacción con diferentes tamaños de celdas).

Métodos	Tamaños de celdas (número de repeticiones para cada combinación de tratamientos)									
	5	10	15	20	25	30	35	40	45	50
ART Test	4,59	3,39	3,53	2,74	2,34	3,07	2,75	3,14	2,09	2,22
ATS Test	2,41	2,89	2,94	3,74	3,87	5,86	6,51	5,93	5,72	6,21
BDM Test	5,01	5,21	5,03	5,31	4,46	3,87	4,11	4,15	4,42	3,95
Box Cox	2,86	3,66	3,27	4,15	3,52	3,71	3,71	3,99	3,75	3,55
Brow-Forsythe	6,44	4,72	4,71	4,25	3,56	3,47	3,34	3,54	3,49	3,01
extended median	9,54	9,92	8,86	7,36	6,33	5,78	6,01	6,14	7,34	6,96
Median test	6,26	5,84	5,63	5,29	4,01	3,49	3,71	3,91	3,44	3,79
M-estimators	4,32	4,57	4,18	4,51	3,21	3,23	3,21	2,71	3,23	1,85
Permute Manly	3,57	3,81	2,38	2,32	1,56	1,37	2,14	1,69	1,89	1,56
Permute Still White	3,37	2,75	2,03	1,75	2,22	2,34	2,86	2,21	2,03	1,75
Permute Ter Braak	3,05	3,27	1,74	1,51	1,71	2,04	1,72	1,71	1,25	1,54
Puri and Sen Test	5,14	5,43	5,28	6,03	5,26	4,17	4,44	5,24	5,19	5,71
R ANOVA Test	7,38	7,08	5,54	5,69	4,56	4,55	5,15	4,78	4,51	4,64
RT Test	6,32	5,51	4,62	4,75	5,05	5,57	5,54	4,84	5,09	4,71
Scheirer Test	9,87	7,94	6,69	6,89	7,26	7,57	6,81	5,65	5,42	6,51
Trimmed means	4,81	4,21	3,95	3,44	2,82	2,31	2,57	2,71	3,61	3,23
van der Waerder	4,37	3,94	2,69	2,89	2,56	1,85	1,75	1,19	1,19	1,01
Wilson Method	9,25	7,68	7,44	6,59	7,01	7,21	7,58	7,13	6,92	6,82
WJ Test	8,15	7,33	6,63	4,96	4,71	4,08	4,33	4,58	4,02	4,44
WTS Test	5,81	6,42	5,21	4,31	3,03	2,55	2,64	2,33	2,97	2,71

Nota: Los valores en negrita indican tasas de error tipo I superiores al 5%

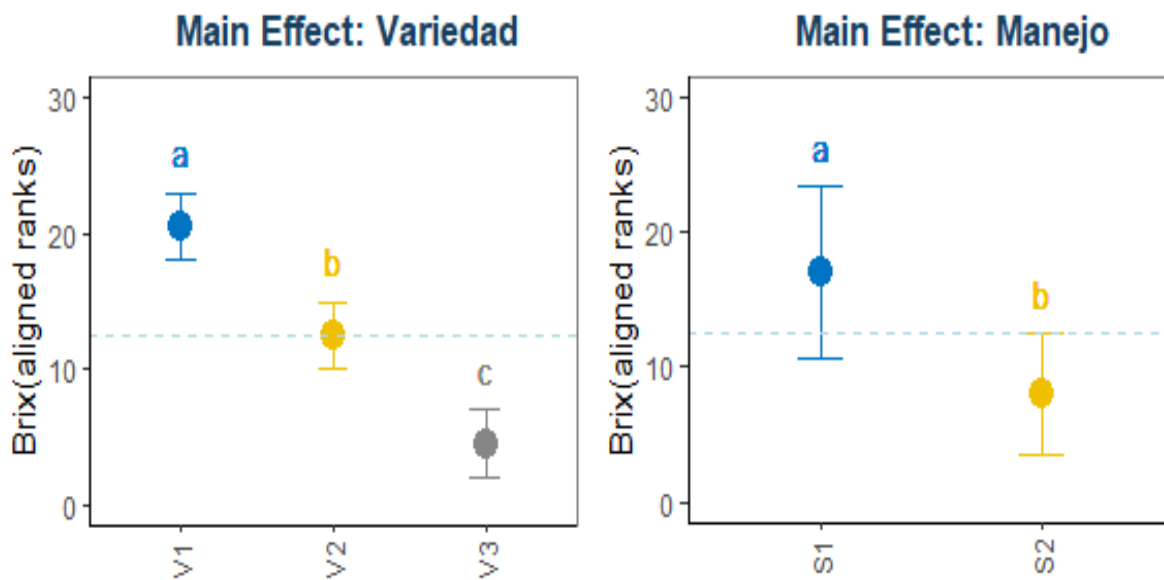


Figura 5. Post hoc mediante la prueba de Tukey para efectos principales.

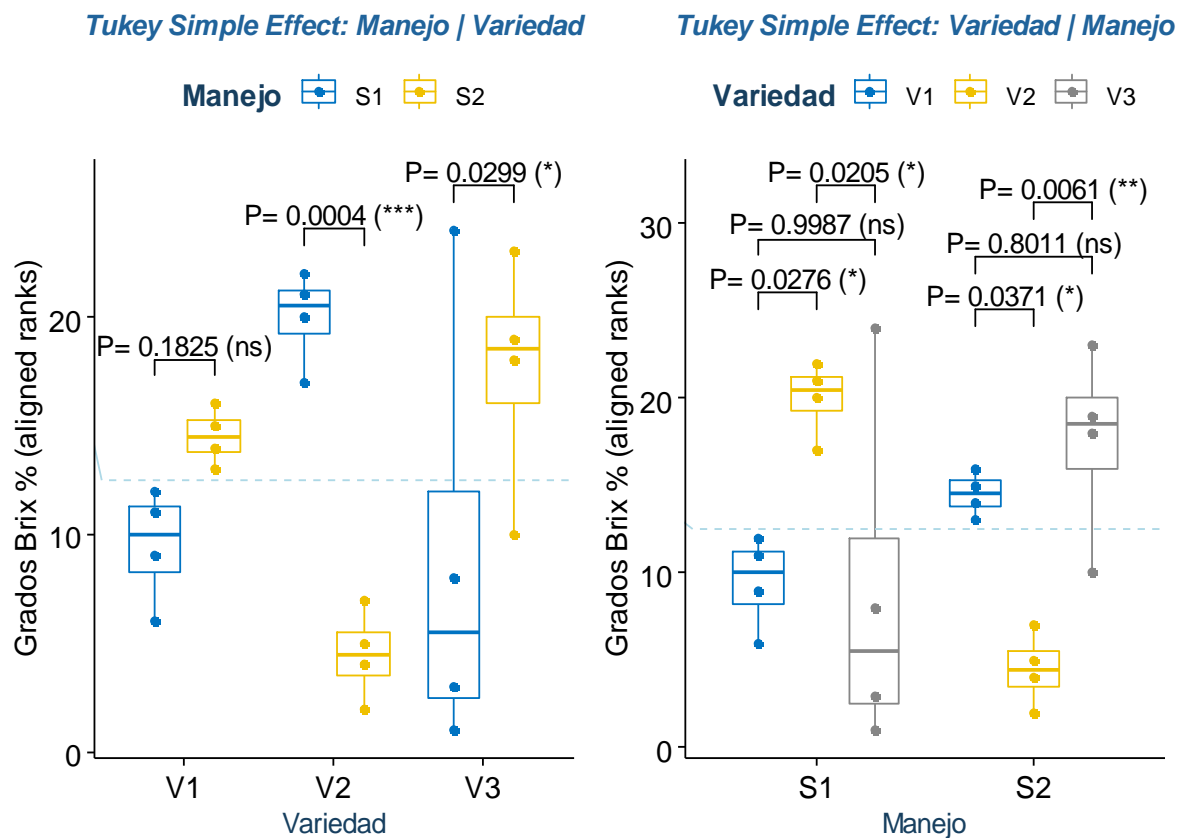


Figura 6. P-valores en efectos simples mediante la prueba de Tukey.

Para los efectos simples de la Variedad sobre los tipos de manejo se tiene: Para el manejo S1 (*Orgánico*) no existe diferencia entre las V1 y V3, mientras que en las comparaciones V1-V2 y V2-V3 existe diferencia significativa; para el tipo de manejo S2 (*Convencional*) no existe diferencia significativa entre las V1 y V3, mientras que para las comparaciones V1-V2 y V2-V3 existe diferencia significativa.

El resumen general de las comparaciones de la prueba de Tukey para efectos simples e interacción (Figura 7) arrojan que, el mayor porcentaje de grados brix se

obtiene en: variedad V1: (S1 o S2); en la V2: (S1) y en la variedad V3: (S2). Para el tipo de manejo S1:(V3); y para el manejo S2: (V3 o V1) producen mayores porcentajes de grados brix.

En la parte inferior de la misma Figura 7, el efecto de la interacción da como resultado a los tratamientos V2:S1 (*Lisa cayena: Orgánico*) y V3:S2 (*Hawaina: Convencional*), como las combinaciones que producen mayor porcentaje de grados brix (están en la primera posición, comparten la misma letra "a", el mismo color y se ubican por encima de la media general).

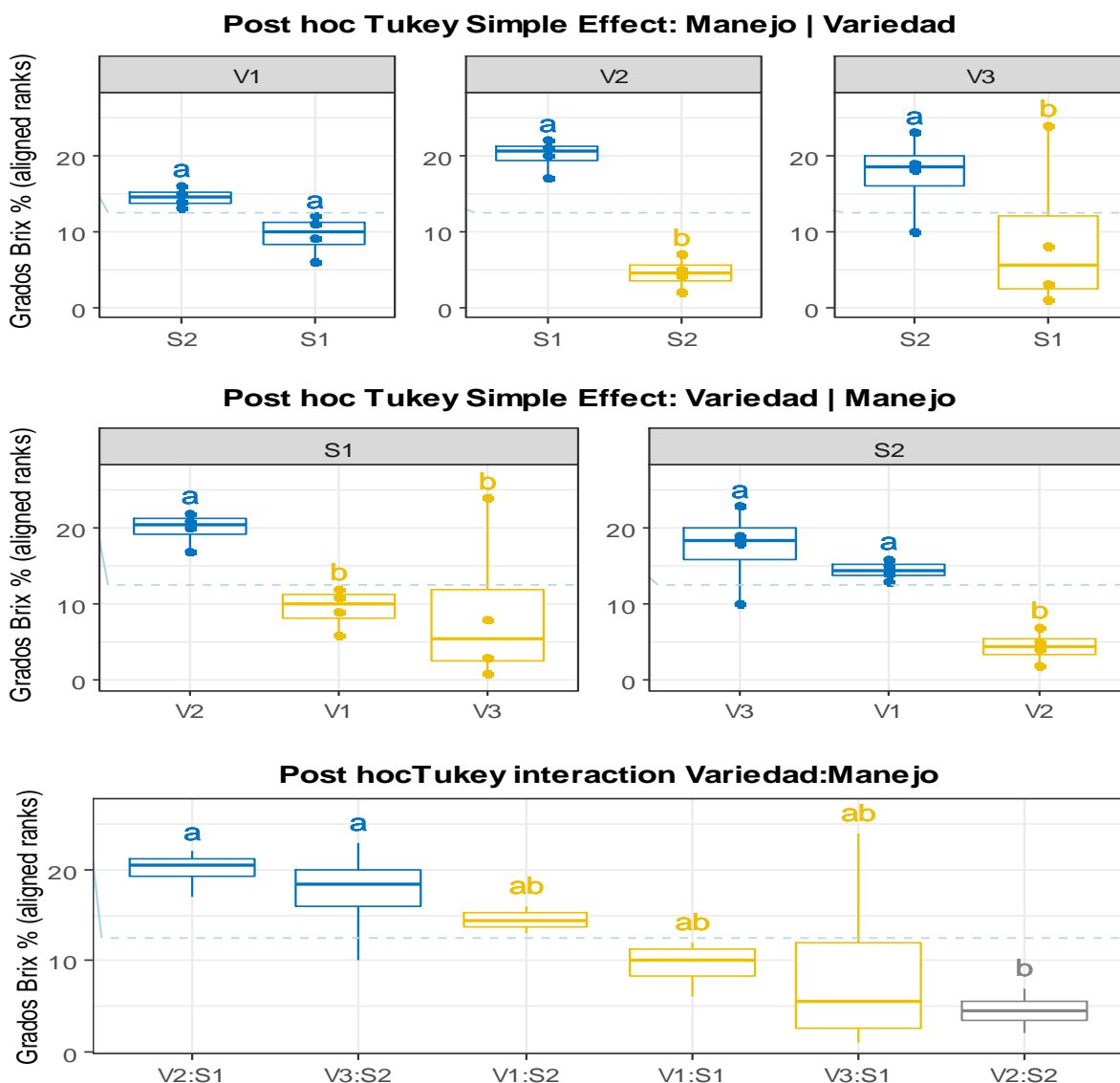


Figura 7. Post hoc mediante la prueba de Tukey para efectos simples e interacción

4. CONCLUSIONES

Se estudió que el efecto de violar los supuestos en un diseño factorial produce una pérdida de potencia lo cual puede conducir a tasas de error altas e imposibilitar encontrar diferencias significativas entre los tratamientos.

Se probó 20 métodos alternativos al ANOVA cuando no se cumplen el supuesto de normalidad u homogeneidad de varianza en los datos experimentales. Si bien es cierto casi todas las pruebas pudieron rechazar las hipótesis de efectos principales, no así con la interacción, tal es el caso de las pruebas basadas en la mediana y la mayoría de las pruebas basada en rangos, las cuales no tuvieron éxito; por lo contrario, las pruebas de permutación, la transformación Box Cox y las pruebas basadas en métodos robustos pudieron rechazar significativamente la interacción.

Al realizar simulaciones para estimar las tasas de error tipo I, los métodos de permutaciones, la transformación Box Cox, el ART, Trimmed means, van der Waerder y el BDM test tuvieron buenos resultados, arrojaron tasas bajas de error para diferentes tamaños de celdas.

Para elegir el método más adecuado me quedo con el análisis reflexivo de Feys (2016), el cual indica que no se debería seleccionar un método por un Pvalue esperanzador pequeño, si no, debería hacerse un análisis detallado y justificado de elección de la prueba. En R actualmente existen varias funciones para probar la interacción en diseños factoriales, se debería elegir la prueba según las características de los datos experimentales, tal como su tamaño (las pruebas no paramétricas suelen ser más efectivas para tamaños pequeños), el balanceo de los datos, su tipo de distribución, su costo computacional (las pruebas de permutaciones tienen una desventaja acá), su capacidad para poder hacer pruebas complementarias de post hoc, entre otras características.

Al realizar el análisis post hoc con un método seleccionado por sus buenas características hacia los datos (ART), se encontró que en el diseño evaluado la combinación de V2:S1 (variedad Cayena Lisa /

Manejo orgánico) y V3:S2 (variedad Hawaiana / Manejo convencional) permiten obtener el mayor porcentaje promedio de grados brix. Se deja en claro que estos resultados son aplicados al diseño investigado y de ninguna manera se puede generalizar a todos los estudios relacionados a la obtención de grados brix en piñas puesto que estos pueden depender de otros factores o de diferentes condiciones.

Conflictos de intereses

El autor del presente trabajo de investigación declara no tener ningún potencial conflicto de interés personal o económico con otras personas u organizaciones que puedan influir indebidamente con el presente manuscrito.

Contribuciones de los autores

Preparación y ejecución: AM; Desarrollo de la metodología: AM; Concepción y diseño: AM; Edición del artículo: AM; Supervisión del estudio: AM.

5. REFERENCIAS

- Adam, L., & Bejda, P. (2018). Robust estimators based on generalization of trimmed mean. *Communications in Statistics - Simulation and Computation*, 47(7), 2139-2151. <https://doi.org/10.1080/03610918.2017.1337136>
- Aho, K. A. (2013). Robust ANOVA. In Chapman & Hall/CRC (Eds.). *Foundational and Applied Statistics for Biologists using R* (1st ed., pp. 493). Taylor & Francis Group. Disponible en: <https://libgen.is/book/index.php?md5=FC7D4BB1EF0423899FFE8E20DED19E09>
- Aho, K. A. (2020). asbio: A collection of statistical tools for biologists. *R package version 1.6-7*. <https://rdr.io/cran/asbio/>
- Akritas, M. G., Arnold, S. F., & Brunner, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *Journal of the American Statistical Association*, 92(437), 258-265. <https://doi.org/10.1080/01621459.1997.10473623>
- Afonso, A., & Pereira, D. G. (2019). Comparação entre métodos não paramétricos para a análise de variância com dois fatores: um estudo de simulação. *Instituto Nacional de Estatística*, 147-158.

- Algina, J., & Olejnik, S. F. (1984). Implementing the Welch-James procedure with factorial designs. *Educational and psychological measurement*, 44(1), 39-48. <https://doi.org/10.1177/0013164484441004>
- Andriani, S. (2017). Uji Park Dan Uji Breusch Pagan Godfrey Dalam pendeteksian heteroskedastisitas pada analisis regresi. Al-Jabar. *Jurnal Pendidikan Matematika*, 8(1), 63-72. <https://doi.org/10.24042/ajpm.v8i1.1014>
- Beasley, T. M., & Zumbo, B. D. (2009). Aligned rank tests for interactions in Split- Plot Designs: distributional assumptions and stochastic heterogeneity. *Journal of Modern Applied Statistical Methods*, 8(1), 16-50. <http://dx.doi.org/10.22237/jmasm/1241136180>
- Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2), 211-252. <https://doi.org/10.1111/j.2517-6161.1964.tb00553.x>
- Brown, M. B., & Forsythe, A. B. (1974). The anova and multiple comparisons for data with heterogeneous variances. *Biometrics*, 30(4), 719-724. <https://doi.org/10.2307/2529238>
- Brunner, E., Dette, H., & Munk, A. (1997). Box-Type approximations in nonparametric factorial designs. *Journal of the American Statistical Association*, 92(440), 1494-1502. <https://doi.org/10.1080/01621459.1997.10473671>
- Brunner, E., Konietzschke, F., Pauly, M., & Puri, M. (2016). Rank-Based procedures in factorial designs: Hypotheses about nonparametric treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5), 1463-1485. <https://doi.org/10.1111/rssb.12222>
- Brunner, E.; Bathke, A.; Konietzschke, F. (2018). Two-Factor Crossed Designs. *Rank and Pseudo - Rank Procedures for Independent Observations in Factorial Designs Using R and SAS*. (1st ed, pp. 287-292). Editorial Springer, suiza. url: <http://library.lol/main/96D22A7D31F896490321AEBE84933872>
- Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician* 35(3), 124-129. <https://doi.org/10.2307/2683975>
- Delgado, J. (1992). *Algunos problemas básicos del análisis de varianza*. (1era ed.). Editorial Ediciones Universidad de Salamanca, España.
- De Neve, J., & Thas, O. (2017). A Mann-Whitney type effect measure of interaction for factorial designs. *Communications in Statistics - Theory and Methods*, 46(22), 11243-11260. <https://doi.org/10.1080/03610926.2016.1263739>
- Feys, J. (2016). Nonparametric tests for the interaction in two-way factorial designs using R. *The R Journal* 8(1), 367-378. <https://doi.org/10.32614/RJ-2016-027>
- Friedrich, S., Konietzschke, F., & Pauly, M. (2017). GFD: An R Package for the analysis of general factorial designs. *Journal of Statistical Software*, 79(1), 1-18. <https://doi.org/10.18637/jss.v079.c01>
- Frossard, J., & Renaud, O. (2019). permuco: Permutation tests for regression, (Repeated Measures) ANOVA/ANCOVA and Comparison of Signals. *R package version 1.1.0*. <https://cran.r-project.org/web/packages/permuco/index.html>
- Gao, X., & Alvo, M. (2005). A nonparametric test for interaction in two-way layouts. *Canadian Journal of Statistics*, 33(4), 529-543. doi:10.1002/cjs.5550330405
- Higgins, J., & Tashtoush, S. (1994). An aligned rank transform test for interaction. *Nonlinear World*, 1(2), 201-211.
- Hocking, R. (1985). *The Analysis of Linear Models* (1st ed.). Editorial Brooks/Cole, Monterey, California.
- Howell, D. (2013). *Statistical Methods for Psychology*. Wadsworth (9th ed.). <https://www.uvm.edu/~statdhtx/StatPages/>
- Jimenes, C.; Pérez, J. (1989). Diseños experimentales en ciencias de la conducta: un método de análisis de varianza de libre distribución (no paramétrico). *Anuario de psicología*, 42, 31-48.
- Kassambara, A. (2019). Homogeneity of variance (1st ed. pp. 14). *Practical Statistics in R for Comparing Groups: Numerical Variables*. Editorial Datanovia, Francia.
- Kay, M., & Wobbrock, J. (2020). ARTool: Aligned rank transform. *R package version 0.10.7*. <https://cran.r-project.org/web/packages/ARTool/index.html>.
- Keselman, H. J., Carriere, K. C., & Lix, L. M. (1996). Robust and powerful nonorthogonal analyses. *Psychometrika*, 60(3), 395-418.

- <https://doi.org/10.1007/BF02294383>
- Kloke, J., & McKean, J. (2020). Rfit: Rank-Based estimation for linear models. *R package version 0.24.2*. Disponible en: <https://cran.r-project.org/web/packages/Rfit/index.html>
 - Konietzschke, F.; Friedrich, S.; Brunner, E.; Pauly, M. (2020). rankFD: Rank-Based Tests for General Factorial Designs. *R package version 0.0.5*. Disponible en: <https://cran.r-project.org/web/packages/rankFD/index.html>
 - Kreuzmann, A., Medina, & Rojas, N. (2018). trafo: Estimation, comparison and selection of transformations. *R package version 1.0.1*. Disponible en: <https://cran.r-project.org/web/packages/trafo/index.html>
 - Lawrence, M. (2016). ez: Easy analysis and visualization of factorial experiments. *R package version 4.4.0*. Disponible en: <https://cran.r-project.org/web/packages/ez/index.html>
 - Lawson, J. (2015). Completely randomized designs with one factor (1st ed. Pp. 31-36). *Design and Analysis of Experiments with R*. Editorial CHAPMAN & HALL/CRC, Brigham Young University Provo, Utah, USA. <http://library.lol/main/BE472FED84FD3B4E5D5D53D6C1719C30>
 - Luepsen, H. (2018). Comparison of nonparametric analysis of variance methods: A vote for van der Waerden. *Journal Communications in Statistics - Simulation and Computation*, 47(9), 2547-2576. <https://doi.org/10.1080/03610918.2017.1353613>
 - Luepsen, H. (2020). R Functions for the analysis of variance. Disponible en: <http://www.uni-koeln.de/~a0032/R/>
 - Mair, P., & Wilcox, R. (2020). WRS2: A collection of robust statistical methods. *R package version 1.1-0*. Disponible en: <https://cran.r-project.org/web/packages/WRS2/>
 - Mangiafico, S. (2016). Summary and Analysis of Extension Program Evaluation. *R package version 1.18.1*. Disponible en: rcompanion.org/documents/RHandbookProgramEvaluation.pdf
 - Manly, B. (1997). *Randomization, bootstrap, and Monte Carlo methods in biology* (2nd ed.). Editorial Chapman Hall, London.
 - Mansouri, H., & Chang, G. (1995). A comparative study of some rank tests for interaction. *Computational Statistics & Data Analysis*, 19(1), 85-96. [https://doi.org/10.1016/0167-9473\(93\)E0045-6](https://doi.org/10.1016/0167-9473(93)E0045-6)
 - Melo, O. O., López, L. A., & Melo, S. (2020). Comparaciones múltiples y validación de supuestos (2da ed.). *Diseño de experimentos Métodos y aplicaciones*. Editorial Coordinación de publicaciones - Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, D. C., Colombia.
 - Mewhort, D. J., Brendan, B. T., & Matthew, k. (2010). Applying the permutation test to factorial designs. *Behavior Research Methods*, 42(2), 366-372.
 - Puri, M., & Sen, P. (1985). *Nonparametric Methods in General Linear Models*. (1st ed. pp. 235). Editorial Wiley, New York.
 - Pelea, L. (2018). ¿Cómo proceder ante el incumplimiento de las premisas de los métodos paramétricos? o ¿cómo trabajar con variables biológicas no normales? *Revista Del Jardín Botánico Nacional*, 39, 1-12.
 - Quinglong, L. (2015). StatMethRank: Statistical methods for ranking data. *R package version 1.3*. Disponible en: <https://cran.r-project.org/web/packages/StatMethRank>
 - Ribeiro-Oliveira, J. P., Garcia, D., Pereira, V. & Machado, C. (2018). Data transformation: an underestimated tool by inappropriate use. *Acta Scientiarum-agronomy*, 40, 35-300. <https://doi.org/10.4025/actasciagron.v40i1.35300>
 - Ripley, B., Venables, B., Bates, D., Hornik, K., Gebhardt, A., & Firth, D. (2020). MASS: Support Functions and Datasets for Venables and Ripley's MASS. *R package version 7.3-53*. Disponible en: <https://cran.r-project.org/web/packages/MASS/index.html>
 - Saste, S., Sananse, S., & Sonar, C. (2016). On parametric and nonparametric analysis of two factor factorial experiment. *International Journal of Applied Research*, 2(7), 653-656.
 - Sawilowsky, S. S (1990). Nonparametric Tests of Interaction in Experimental Design. *Review of Educational Research*, 60(1), 91-126. <https://doi.org/10.2307/1170226>
 - Scheirer, C. J., Ray, W. S., & Hare, N. (1976). The analysis of ranked data derived from completely randomized factorial designs. *Biometrics*, 32(2), 429-434. <https://doi.org/10.2307/2529511>

- Shoemaker, L. H. (1986). A *Conference on Human Factors in Computing Systems, Vancouver, British Columbia, New York, 7-12.* <https://doi.org/10.1145/1978942.1978963>
- Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only ANOVA Procedures. In Proceedings of the ACM