López L. 2021. Guía práctica para el análisis de datos biológicos. Revista Sigma, 17 (1). Páginas 32–41.

REVISTA SIGMA

Departamento de Matemáticas y Estadística

Universidad de Nariño

Volumen XVII Nº1(2021), páginas 32–41

Guía práctica para el análisis de datos biológicos

Liliana López Kleine [1]

Abstract: Statistical analysis of biological data is a need in order to have conclusions accepted by the scientific community. Therefore, any study, experiment, biological sampling, etc., requires a statistical analysis of the data. Despite of this need, for biologists and other scientists, undertaking this analysis is a difficult challenge and it is not always done the best possible way without the help of a statistician. Here, some reflections, clarifications and a guide for the proper statistical analysis of biological data are presented with the aim of collecting and complementing knowledge presented in basic textbooks and courses on the subject.

Keywords: biostatistics, biological data, statistical analysis, practical guide.

Resumen: El análisis estadístico de datos biológicos es una necesidad para que las conclusiones sean aceptadas por la comunidad científica. Por esta razón, cualquier estudio, experimento, muestreo biológico, etc., requiere un análisis estadístico de los datos. A pesar de que esto es una necesidad, para los biólogos y otros investigadores, la realización de este análisis es un reto difícil y no se realiza siempre de la mejor manera sin la ayuda de un estadístico. Acá se presentan reflexiones, aclaraciones y una guía general para la realización del análisis estadístico de datos biológicos con métodos estadísticos básico que están incluidos en los cursos básicos de bioestadística de cualquier programa curricular de biología y carreras afines. No se abarcan análisis que no hacen parte de los cursos básicos de bioestadística. Este artículo pretende recopilar y complementar los conocimientos presentados en textos y cursos básicos sobre el tema.

Palabras Clave: bioestadística, datos biológicos, análisis estadístico, guía práctica.

[1]Profesora Titular Departamento de Estadística Universidad Nacional de Colombia - sede Bogotá, email: llopezk@unal.edu.co

1. Introducción

Después de varios años de docencia e investigación en bioestadística, me he dado cuenta de que enseñar a realizar un análisis de datos biológicos correctamente es mucho menos complicado de lo que se piensa y los profesionales de biología y carreras afines poseen herramientas estadísticas sencillas y útiles para resolver la mayoría de sus preguntas gracias a los cursos básicos de dichas carreras. La estadística es muy sencilla, la biología es la que es complicada. El estudio de la vida y de los seres vivos siempre tiene sorpresas ocultas, nada es "blanco y negro", siempre hay una excepción. Por esa dificultad en biología muchas veces los diseños o muestreos son complicados, se quieren tener en cuenta todas las excepciones posibles y responder muchas preguntas a la vez. Adicionalmente, el concepto de modelo como una herramienta para entender de manera sencilla y simple los sistemas complejos (como los biológicos), es algo que no está arraigado ya que se usa mucho menos en biología que en otras ciencias. Por esta razón a un biólogo no le gusta responder preguntas sencillas: 1) se le dificulta planear experimentos sencillos y 2) siempre quiere saber varias cosas al tiempo cuando plantea un experimento.

La estadística es una necesidad para cualquier ciencia natural y lo ha sido por mucho tiempo. La importancia fue mencionada incluso por Platón [13]: "Todas las ciencias son ciencias si son matemáticas". Los estudiantes e investigadores en biología generalmente tienen la motivación de aprender estadística y de usarla correctamente, pero es difícil apropiarse del conocimiento y pasar de los conocimientos y ejemplos de clase a la aplicación de lo aprendido en sus trabajos de investigación. Realizar el puente es un reto.

Este escrito tiene como objetivo exponer algunas reflexiones y servir de guía y complemento para el análisis básico de datos biológicos, sin pretender reemplazar textos o cursos que exponen el tema con rigor y detalle. Solamente se mencionan los análisis básicos que hacen parte de los programas de estadística de pregrado que se ven en las carreras de biología y afines. No pretenden mencionar toda gama de análisis de datos posibles que podrían hacerse a datos biológicos.

Es difícil encontrar artículos que sean una guía real para el diseño y análisis de datos biológicos o que hagan reflexiones sobre el tema. Se encuentran buenos libros de bioestadística básica, pero rara vez abarcan el tema de análisis de datos y el de diseño de experimentos en un mismo texto con ejemplos prácticos [4] [6] [11].

A pesar de que no existe un libro o texto de reflexión y guía que sea perfecto, sí podemos encontrar elementos en la literatura científica que expliquen bien los procedimientos, que den lineamientos importantes sobre el diseño y que informen cómo realizar un análisis de datos evitando los errores comunes explicando la importancia de cada paso. Sin embargo, los investigadores y estudiantes de las áreas biológicas no sienten la necesidad de consultar textos al momento de planear su experimento o tampoco de acudir a profesionales en estadística, ya que esperan que sus resultados sean tan claros que incluso sin mucho análisis estadístico se vea la evidencia como lo creía Ernest Rutherford (quien era químico): "If your experiment needs a statistician, you need a better experiment". Sin embargo, hoy en día sabemos que la evidencia estadística es necesaria y es lo aceptado por la comunidad científica. A pesar de que el experimento sea muy claro y sencillo, debe haber una planeación correcta y un análisis estadístico el cual permita controlar el error sobre las conclusiones obtenidas, incluyendo la variabilidad natural de los individuos biológicos. Por eso es un error acudir a textos o estadísticos cuando ya se tienen los datos; es demasiado tarde como bien lo dijo el estadístico Ronald Fisher hace 100 años aproximadamente: "To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem exami-

nation: he may be able to say what the experiment died of." Dada esta situación hay varios aspectos básicos que todos los profesionales de las áreas biológicas ven en sus asignaturas de estadística, relacionados con el diseño de experimentos y el análisis de datos biológicos, sobre los cuales vale la pena reflexionar para proponer una manera sencilla de abordarlos y hacer unas recomendaciones generales para el análisis de datos biológicos. No se mencionan no se incluyen acá análisis de modelamiento o multivariados, que, aunque muy usados en ciencias biológicas no hacen parte de los cursos básicos de estadística en estas profesiones.

2. El diseño y análisis de experimentos paso a paso

2.1. La pregunta biológica

Muchas veces los experimentadores no tienen muy clara cuál es la pregunta que quieren responder. Quieren saber si hay diferencias de crecimiento (por ejemplo) entre condiciones, pero a la vez ver si estas están influenciadas por otras, y también si afectan el metabolismo y tal vez si de pronto las pueden comparar con otro lugar, y en realidad no les interesa solo una cepa sino son varias... Desde el punto de vista estadístico se necesita definir con mucha claridad qué se quiere saber para saber qué se va a medir. Con esto claro, se debe, desde antes de iniciar, saber qué se va a medir o registrar, qué tipo de variable: (cualitativa, numérica, discreta, ...).

Una vez se ha definido con claridad la pregunta y qué información puede servir para responder a la pregunta, hay que tener claro qué método estadístico puede responder a esa pregunta y por lo tanto desde el inicio se debe saber qué análisis de va a utilizar, o por lo menos qué opciones de análisis hay. No es evidente para nadie que esto se deba definir desde el principio, ya que en muchos cursos y textos, se enseña que la estadística se empieza a usar cuando la tabla de datos ya está lista. Sin embargo, hago énfasis en la importancia de que las opciones de análisis deben definirse antes de iniciar la toma de datos.

2.2. ¿Para qué tanta información?

Una vez definida la pregunta, se deben definir la variable o las variables a medir se deben definir. El mensaje en este punto es: solamente medir las variables que realmente se necesitan y responden a la pregunta y también: solamente las réplicas que se necesitan.

Este punto está íntimamente relacionado con el tipo de diseño experimental que se va a hacer y se tratará en el siguiente punto. Sin embargo, primero se abordará el tema de las réplicas técnicas y de los pools que son muy comunes en biología molecular.

2.3. ¿Cuándo hay que hacer réplicas técnicas?

Solamente cuando el experimento es muy variable, cuando la tecnología utilizada tiene limitaciones conocidas y cuando se está calibrando una técnica. Es decir, si se tiene confianza en la técnica, no se necesitan realizar réplicas técnicas. No se deben hacer si no se va a usar la información para responder a la pregunta biológica.

Si el experimento es muy variable en realidad va a ser difícil responder a la pregunta biológica. Si se ha decidido hacer réplicas técnicas, lo primero que hay que verificar es si la variabilidad

entre réplicas técnicas supera la variabilidad biológica (las diferencias entre individuos). Si esto es así, las mediciones no tienen validez. Si esto es así hay tres opciones:

1. Aumentar las réplicas técnicas.
2. Filtrar valores con un criterio estadístico.
3. Cambiar de experimento.

Una vez que se hayan tomado las réplicas técnicas, los valores de un individuo se deben combinar para reportar cuál es la medida definitiva del individuo. Una mala opción es el promedio, generalmente la mediana es una mejor opción, aunque existen métodos intermedios como las medias truncadas, por ejemplo.

Prueba piloto para determinar tamaño de muestra

De manera general, todas las estimaciones de tamaño de muestra requieren información sobre el fenómeno que se está estudiando y se basan en determinar un tamaño de muestra con base en el poder de la prueba. El poder está relacionado con la probabilidad de cometer un error tipo II (β), siendo poder$= 1 - \beta$. Por lo anterior es necesario definir información sobre la hipótesis alternativa. La mejor manera de hacer esto, es realizando una prueba piloto. Una vez se tienen estos datos previos, existen maneras de estimar el tamaño de muestra para diversas pruebas con fórmulas específicas [15] o de manera empírica, usando el tamaño de efecto definido por [3].

2.4. Diseño experimental

Una vez definido el diseño, se deben revisar los supuestos estadísticos que tiene el diseño para identificar el número de réplicas. Dada que la pregunta más común es si hay diferencias en la variable medida entre varios grupos o niveles de tratamiento, los ejemplos usados son la prueba de diferencia de medias (prueba T o prueba de wilcoxon) y el análisis de varianza de un factor. En caso de que se deseen probar varios factores o tratamientos al tiempo, reglas muy similares aplican para el análisis de varianza de dos factores. Sin embargo, para este no existe una alternativa no paramétrica.

¿Si se va a realizar un análisis de varianza (y no interesan o no aportan las interacciones entre factores), se puede hacer un diseño por bloques para bajar número de muestras? Sería bueno evaluar si los bloques deben ser completos o se podría responder a la pregunta con menos datos. ¿Cuántas replicas se deben hacer si se tienen en cuenta pérdidas? Sería bueno hacer una prueba piloto para poder identificar el tamaño de muestra con métodos estadísticos (este no se puede determinar sin prueba piloto).

En varias ocasiones, es más fácil hacer uno o dos experimentos sencillos separados para responder a las preguntas biológicas en vez de un experimento complicado (tipo análisis de varianza de varios factores). Esto muchas veces permite un mejor control de efectos externos que podrían confundirse con los efectos que realmente se quiere identificar (condiciones ambientales incontroladas, efectos de borde, de lote, experimentador, etc.)

Es importante recordar que las tres bases del diseño experimental son: 1) Réplicas, 2) Aleatorización, 3) Bloques [10]. El primer punto permite diferenciar la variabilidad natural de los individuos de las diferencias debidas al tratamiento, el segundo intenta contrarrestar los efectos incontrolados y asegurar la independencia entre individuos. El tercero elimina la variabilidad debida a homogeneidad conocida entre individuos, así como reduce el número

de muestras. Las etapas del diseño que se han mencionado hasta acá están ilustradas en la figura 1.

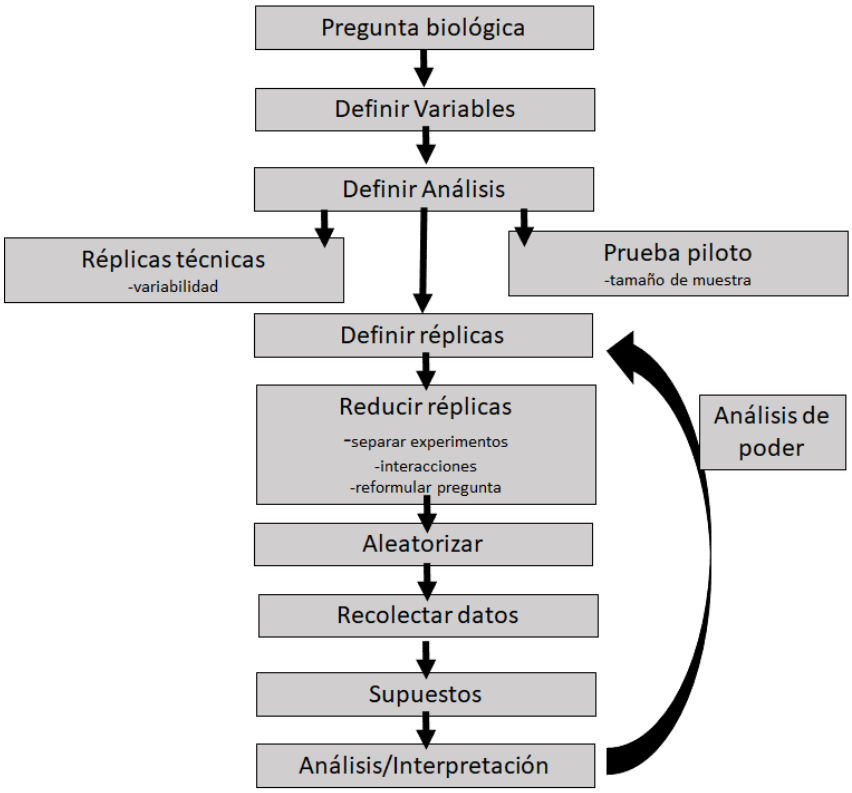

Figura 1: Ilustración de las etapas de planeación e implementación de un experimento biológico desde la formulación de la pregunta hasta el análisis. Para detalles sobre el análisis ver también figura 2

Muchas veces los investigadores no tienen suficiente cuidado con las aleatorizaciones. Estas de deben hacer generando números aleatorios de verdad, ya que en la mayoría de los casos esta etapa es el único proceso para asegurar la independencia entre individuos (supuesto para la mayoría de los análisis estadísticos). Las aleatorizaciones se hacen para asignar los tratamientos en primera instancia, es decir, una vez escogidos los individuos estadísticos (de manera aleatoria), se reparten en grupos como concentraciones, temperaturas, grupo control y tratamiento, etc. Igualmente, se debe aleatorizar el lugar en donde se van a poner, el experimentador que va a medir, el orden en el que se va a medir, etc.

2.5. ¿Cuáles son los supuestos más comunes y por qué son tan importantes?

Existen básicamente tres supuestos que se deben cumplir en la mayoría de los análisis estadísticos clásicos, por ejemplo, para el análisis de varianza: 1) Distribución normal de la variable respuesta, 2) Homocedasticidad o varianzas iguales entre los niveles del factor, 3) Independencia entre los individuos [10].

El supuesto de distribución significa que los datos que se obtuvieron provienen de una

variable con distribución normal. Este es el supuesto más importante, ya que es el que asegura que el estadístico de prueba (en caso del análisis de varianza: estadístico F, en caso de la prueba T: estadístico T) realmente tenga la distribución que se espera. Si este no se cumple, no se puede estar seguro de que el valor P obtenido es real, porque este es calculado en la distribución esperada. Si la variable no tiene una distribución normal, su promedio no tendrá una distribución T y menos la diferencia de promedios [15]. Lo mismo sucede con la distribución F y con todas las otras. La prueba estadística no tiene ninguna validez si no se cumple el supuesto. Es obligatorio que se verifique y que se reporte el resultado de la verificación. Existen alternativas no paramétricas cuando este supuesto no se cumple (por ejemplo la prueba de Wilcoxon [17] o Kruskal & Wallis [8]).

El supuesto de homocedasticidad es importante, pero no afecta la distribución del estadístico de prueba, lo que afecta es la tasa de error (aumenta la probabilidad del error tipo I), aunque no se sabe en cuánto. Hay varios trabajos empíricos al respecto [7] que han medido y verificado esto aumento, pero es altamente dependiente de los datos. Este error baja si el tamaño de muestra en cada nivel del factor es igual (diseño balanceado) [5].

El supuesto de independencia se puede verificar para un análisis de varianza ajustando un modelo lineal e investigando si hay independencia en los residuales, para pruebas como la T se asume si el diseño fue aleatorizado. Es de resaltar que no se debe confundir con la independencia entre variables. Se trata de independencia entre individuos y se refiere es a que los individuos estadísticos no comparten algo que hace que se formen posibles grupos homogéneos (orígenes genéticos, geográficos, etc.).

La homocedasticidad se puede probar con la prueba de Bartlett [1] o la prueba de Levene [9]. El supuesto de normalidad se puede verificar con la prueba de Shapiro-Wilk [14], aunque existen muchas pruebas para hacerlo.

2.6. Análisis de datos

Una vez que se han verificado los supuestos, se puede proceder al análisis de datos. Un resumen de las pruebas para identificar diferencias entre grupos se presenta en la figura 2. En caso de heterocedasticidad la prueba alternativa a la prueba T es la prueba T de Welch[16] que aproxima de manera diferente los grados de libertad de la distribución T. En caso de no cumplir normalidad, existen las pruebas alternativas de Wilcoxon [17] o Kruskal & Wallis [8].

Luego de tener los resultados, la interpretación debe ser apoyada por las gráficas y/o tablas descriptivas que se hicieron originalmente y comparadas con trabajos similares. Finalmente, es muy importante realizar el análisis de poder. Para la mayoría de las pruebas este se basa en el tamaño del efecto introducido por Cohen [3], el cual recoge qué tan fuertes son las diferencias entre grupos, la probabilidad de cometer un error tipo I (α) y el poder, el cual está relacionado con la probabilidad de cometer un error tipo II (β), siendo poder$= 1 - \beta$ (Figura 1).

Todos los análisis mencionados se pueden realizar en el software R [12]. En la última sección se indica un código de ejemplo para cada uno de los pasos del análisis (Figura 2) con datos libres en el software R.

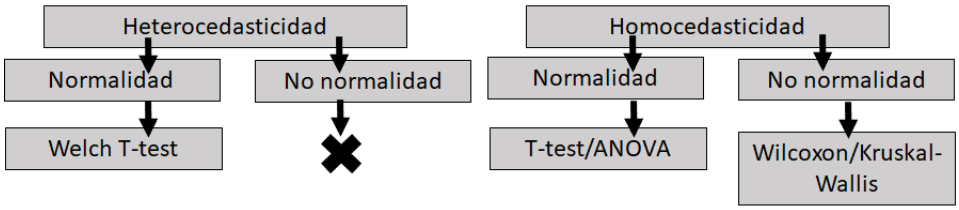

Figura 2: Ilustración de las etapas del análisis de datos desde la verificación de los supuestos para los análisis de diferencia de medias más comunes. ANOVA: análisis de varianza.

2.7. Código de R

Los datos "InsectSprays" contienen el número de picaduras de zancudo en brazos expuestos dependiendo del uso de repelentes diferentes (A,B,C,D,E,F), originalmente fueron publicados por Beall [2]. La figura 3 muestra un boxplot de los datos. Si se asume que solamente se compararon dos de los repelentes, se podría hacer una prueba de comparación de medias como en los ejemplos que se muestran a continuación.

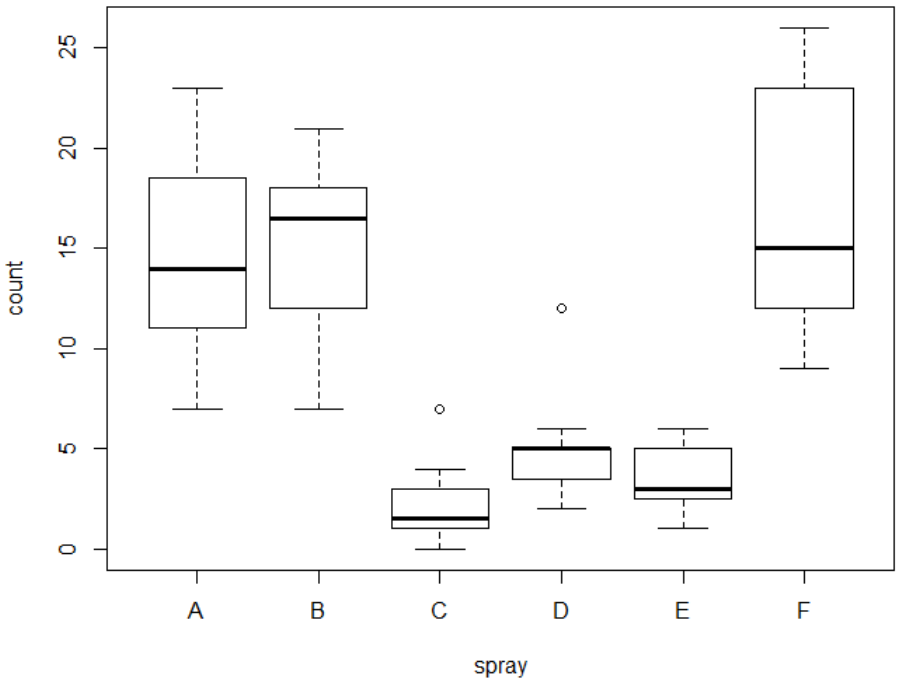

Figura 3: Boxplot de los datos InsectSpray [2] obtenidos en el software R.

Pregunta 1: ¿"Hay diferencias significativas en el número de picaduras entre el repelente A y B"?

Análisis: prueba de comparación de dos medias (T, Welch T o Wilcoxon)

```
data("InsectSprays")
Tinsect<-InsectSprays[1:24,]
attach(Tinsect)
bartlett.test(count ~ spray, data = Tinsect)
Bartlett test of homogeneity of variances
Bartlett's K-squared = 0.10464, df = 1, p-value = 0.7463
by(count,spray,shapiro.test)
spray: A
Shapiro-Wilk normality test
W = 0.95757, p-value = 0.7487
------------------------------------------------
spray: B
Shapiro-Wilk normality test
W = 0.95031, p-value = 0.6415
```

Dado que se cumplen los supuestos de normalidad y homocedasticidad se realiza una prueba T (figura 2). Es de anotar que hay que indicar en R que las varianzas son iguales.

```
t.test(Tinsect[1:12,1],Tinsect[13:24,1],var.equal = TRUE)
Two Sample t-test
t = -0.45352, df = 22, p-value = 0.6546
```

Pregunta 2: ¿"Hay diferencias significativas en el número de picaduras entre el repelente E y F"?

Análisis: prueba de comparación de dos medias (T, Welch T o Wilcoxon)

```
Tinsect<-InsectSprays[49:72,]
attach(Tinsect)
bartlett.test(count ~ spray, data = Tinsect)
Bartlett test of homogeneity of variances
data:  count by spray
Bartlett's K-squared = 13.87, df = 1, p-value = 0.000196
by(count,spray,shapiro.test)
spray: E
Shapiro-Wilk normality test
W = 0.92128, p-value = 0.2967
------------------------------------------------
spray: F
Shapiro-Wilk normality test
W = 0.88475, p-value = 0.1009
```

Dado que se cumplen los supuestos de normalidad, pero no de homocedasticidad se realiza una prueba T de Welch (figura 2).

```
t.test(Tinsect[1:12,1],Tinsect[13:24,1])
Welch Two Sample t-test
t = -7.0711, df = 12.699, p-value = 9.55e-06
```

Pregunta 3: ¿"Hay diferencias significativas en el número de picaduras entre el repelente C y D"?

Análisis: prueba de comparación de dos medias (T, Welch T o Wilcoxon)

```
Tinsect<-InsectSprays[25:48,]
attach(Tinsect)
bartlett.test(count ~ spray, data = Tinsect)
Bartlett test of homogeneity of variances
data:  count by spray
Bartlett's K-squared = 0.58466, df = 1, p-value =
0.4445
by(count,spray,shapiro.test)
spray: C
Shapiro-Wilk normality test
W = 0.85907, p-value = 0.04759
------------------------------------------------
spray: D
Shapiro-Wilk normality test
W = 0.75063, p-value = 0.002713
```

Al no haber normalidad, es mejor confirmar la homocedasticidad con la prueba de Levene

```
library(car)
leveneTest(count ~ spray, data=Tinsect)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1    0      1
      22
```

La prueba a realizar en este caso, al cumplir la homocedasticidad, pero no la normalidad es la prueba de Wilcoxon:

```
wilcox.test(Tinsect[1:12,1],Tinsect[13:24,1])
Wilcoxon rank sum test with continuity correction
W = 20, p-value = 0.002651
```

3. Conclusiones

Para realizar un análisis estadístico adecuado que permita responder a las preguntas biológicas que un investigador se plantea, es necesario surtir en orden las diferentes etapas desde la definición del análisis, la realización de una prueba piloto si es posible, pasando por la verificación de los supuestos, el análisis y la interpretación. La importancia del porqué de cada una de estas etapas se presentó a través de este texto, acompañado de algunas reflexiones, aclaraciones y una guía general para la realización del análisis estadístico de datos biológicos.

4. Agradecimientos

Agradezco a todos los estudiantes que he tenido en los diferentes cursos de bioestadística y diseño de experimentos por sus preguntas e interés constante por hacer las cosas bien.

Referencias

[1] Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, *160*(901), 268–282. 37

[2] Beall, G. (1942). The transformation of data from entomological field experiments so that the analysis of variance becomes applicable. *Biometrika, 32(3/4)*, 243-262. 38

[3] Cohen, J. (1988). Statistical power analysis for the behavioral sciences 2nd ed Hillsdale NJ Erlbaum. 35, 37

[4] Fry JC (Editor).(1993) Biological Data Analysis: A Practical Approach (The Practical Approach Series). Oxford Press. 33

[5] Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. *Review of Educational Research, 42*(3), 237–288. 37

[6] Hartvigsen, G. (2014). *A primer in biological data analysis and visualization using R*. Columbia University Press. 33

[7] Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one-and two-factor fixed effects ANOVA cases. *Journal of educational statistics, 17*(4), 315-339. 37

[8] Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association, 47*(260), 583-621. 37

[9] Levene, H. (1961). Robust tests for equality of variances. *Contributions to probability and statistics. Essays in honor of Harold Hotelling*, 279-292. 37

[10] Montgomery, D. (2004). Diseño y análisis de experimentos: Limusa, SA Segunda edición: México DF. 35, 36

[11] Pagano, M., & Gauvreau, K. (2018). *Principles of biostatistics* 2nd Edition. CRC Press. 33

[12] R Core Team. (2018) R: A language and environment for statistical computing. R Foundation for Statistical, Computing, Vienna, Austria. URL https://www.R-project.org/. 37

[13] Sachs, E. (1917) *Die fünf platonischen Körper*, Weidmannsche Buchhandlung. 33

[14] Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3/4), 591-611. 37

[15] Walpole, R. E., Myers, R. H., Myers, S. L., & Cruz, R. (1992). *Probabilidad y estadística* (Vol. 624). México: McGraw-Hill. 35, 37

[16] Welch, B. L. (1947). The generalization of 'STUDENT'S'problem when several different population varlances are involved. *Biometrika, 34*(1-2), 28-35. 37

[17] Wilcoxon F, (1945) *Individual comparisons by ranking methods*. Biometrics Bulletin. 1 (6): 80–83. Doi:10.2307/3001968. 37