HAL open science

Ventajas de la estadística bayesiana frente a la frecuentista: ¿por qué nos resistimos a usarla?

Paloma Ruiz-Benito, Enrique Andivia, Juliette Archambeau, Julen Astigarraga, Rafael Barrientos, Verónica Cruz-Alonso, Margarita Florencio, Daniel Gómez, Loreto Martínez-Baroja, Pablo Quiles, et al.

To cite this version:

Paloma Ruiz-Benito, Enrique Andivia, Juliette Archambeau, Julen Astigarraga, Rafael Barrientos, et al.. Ventajas de la estadística bayesiana frente a la frecuentista: ¿por qué nos resistimos a usarla?. Ecosistemas, 2018, 27 (2), pp.136-139. 10.7818/ECOS.1591 . hal-02617924

HAL Id: hal-02617924

https://hal.inrae.fr/hal-02617924

Submitted on 25 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

AEET
ASOCIACIÓN ESPAÑOLA DE ECOLOGÍA TERRESTRE

Ecosistemas 27(2): 136-139 [Mayo-Agosto 2018]
Doi.: 10.7818/ECOS.1591

Artículo publicado en Open Access bajo los términos de Creative Commons attribution Non Comercial License 3.0.

ecosistemas
REVISTA CIENTÍFICA DE ECOLOGÍA Y MEDIO AMBIENTE

ISSN 1697-2473 / Open access disponible en www.revistaecosistemas.net

NOTAS ECOINFORMÁTICAS

Ventajas de la estadística bayesiana frente a la frecuentista: ¿por qué nos resistimos a usarla?

P. Ruiz-Benito[1,*], E. Andivia[1], J. Archambeaou[2], J. Astigarraga[1], R. Barrientos[1], V. Cruz-Alonso[1], M. Florencio[1], D. Gómez[1], L. Martínez-Baroja[1], P. Quiles[1], Z.Rohrer[1], A.M.C. Santos[1], E. Velado[1], S. Villén-Pérez[1], I. Morales-Castilla[1]

(1) Departamento de Ciencias de la Vida, Universidad de Alcalá, Edificio de Ciencias, Campus Universitario, 28805 Alcalá de Henares (Madrid), España.
(2) UMR 1202 BIOGECO, University of Bordeaux, Talence, France.

* Autor de correspondencia: P. Ruiz-Benito [palomaruizbenito@gmail.com]

> Recibido el 04 de junio de 2018 - Aceptado el 05 de junio de 2018

Ruiz-Benito, P., Andivia, E., Archambeaou, J., Astigarraga, J., Barrientos, R., Cruz-Alonso, V., Florencio, M., Gómez, D., Martínez-Baroja, L., Quiles, P., Rohrer, Z., Santos, A.M.C., Velado, E., Villén-Pérez, I., Morales-Castilla, I. 2018. Ventajas de la estadística bayesiana frente a la frecuentista: ¿por qué nos resistimos a usarla?. *Ecosistemas* 27(2): 136-139. Doi.: 10.7818/ECOS.1591

La complejidad de la naturaleza requiere de la aplicación de modelos matemáticos que nos permitan su simplificación, es decir, que nos permitan realizar abstracciones para describir nuestras observaciones (Bolker 2008). La teoría de la probabilidad permite la descripción de la incertidumbre asociada a los sistemas estocásticos naturales y el ajuste de modelos. Para describir la realidad y testar nuestra hipótesis calculamos una serie de parámetros asociados al proceso objeto de estudio, que generalmente incluyen al menos un estadístico de tendencia central que lleva asociado un rango de variación (p. ej. media e intervalos de confianza si usamos una aproximación frecuentista, o la media de la distribución posterior del parámetro y los intervalos de credibilidad en estadística bayesiana). Los tres enfoques principales usados para el cálculo de estos parámetros son la estadística frecuentista, la máxima verosimilitud y la estadística bayesiana (**Tabla 1**). Estos enfoques se diferencian principalmente en los métodos utilizados para extraer conclusiones sobre los datos observados: los p-valores, los estimadores de máxima verosimilitud y la distribución posterior de los parámetros.

Los p-valores y los estimadores de máxima verosimilitud, comúnmente utilizados, han sido criticados porque algunas de sus asunciones no se ajustan al funcionamiento de los procesos ecológicos y precisan de normas arbitrarias e inflexibles para aceptar o rechazar hipótesis (Wasserstein y Lazar 2016). Por ejemplo, en muchos estudios se acepta la hipótesis de que dos grupos de datos son diferentes cuando el p-valor asociado al estadístico de contraste es < 0.05; es decir, cuando la probabilidad de obtener una diferencia como la observada, o aún más extrema, si realmente no hubiese diferencias entre ambos grupos es <0.05. Además, el p-valor está muy afectado por el tamaño muestral y su interpretación testando la hipótesis nula puede resultar poco intuitiva (Anderson et al. 2000). Frente al p-valor, los criterios de información como el AIC (*Akaike Information Criterion*), BIC (*Bayesian Information Criterion*) o DIC (*Deviance Information Criterion*) permiten realizar la comparación y selección de modelos testando simultáneamente varias hipótesis específicas (Johnson y Omald 2004). Los Criterios de Información se basan en el principio de parsimonia de forma que se evalúa la verosimilitud del modelo en función de su bondad de ajuste a los datos y de su complejidad (i.e., número de parámetros, Burnham y Anderson 2002). La selección de modelos se hace generalmente en función de diferencias en AIC entre modelos comparables (p.ej. como regla orientativa se usa $\Delta AIC > 2$ para considerar que hay soporte empírico para el modelo más complejo, Burnham y Anderson 2002), y la inferencia de múltiples modelos permite calcular la media de todos ellos cuando no hay un grado de soporte empírico claro para un determinado modelo (Dorman et al. 2018). A pesar de la popularidad creciente de este tipo de enfoques también han sido criticados dado que, si bien serían útiles para descartar los modelos menos informativos, pueden ser similares a otros enfoques basados en p-valores (Murtaugh 2014), no bastarían para identificar el mejor modelo (Kadane y Lazar 2004) y serían necesarios criterios adicionales (Dziak et al. 2017).

La estadística bayesiana es el único enfoque en el que se hace un uso explícito de la probabilidad para cuantificar la incertidumbre de la inferencia (**Tabla 1**). Se trata de un proceso de aprendizaje iterativo en el que se alcanzan conclusiones sobre un fenómeno (probabilidad *a posteriori*) a partir del conocimiento previo sobre el sistema (probabilidad *a priori*) y de nuevas evidencias (información proveniente de los datos). Es decir, los resultados de un nuevo estudio podrían ser usados para actualizar el conocimiento sobre el sistema e incluirlo en estudios posteriores. Por tanto, la estadística bayesiana se basa en la reasignación de credibilidad a través de las probabilidades, donde las probabilidades son los valores de los parámetros en modelos matemáticos biológicamente coherentes (ver manuales de referencia como Bolker 2008; Kéry 2010; Gelman et al. 2013; Kruschke 2015; McElreath 2016).

El uso de un marco de trabajo bayesiano tiene una serie de ventajas sobre la estadística frecuentista (p. ej. Bolker 2008; Clark 2005; Gelman et al. 2013): (i) permite ajustar modelos complejos no abordables por métodos frecuentistas debido a las restricciones de estos modelos; (ii) permite alcanzar estimaciones más exactas de los parámetros cuando el tamaño muestral es pequeño;

© 2018. Los Autores. Editado por la AEET. [Ecosistemas no se hace responsable del uso indebido de material sujeto a derecho de autor]

Tabla 1. *Comparación de las aproximaciones frecuentistas, de máxima verosimilitud y bayesiana para ajustar un determinado modelo (p. ej. Y = a + b X).*
Table 1. *Comparison of frequentist, máximum likelihood and bayesian aproches to fit a certain model (e.g. Y = a + b X).*

	Frecuentista	Máxima verosimilitud	Bayesiana
Aproximación probabilidad	Contraste de hipótesis basado en el p-valor: probabilidad (*P*) de observar los datos (*D*) dado que la hipótesis nula (*Ho*) es verdadera. $P=P(D\|Ho)$	Probabilidad de observar los datos dado un determinado modelo (con parámetros θ) viene determinado por la verosimilitud (*L: likelihood*). $L(\theta)=P(D\|M(\theta))$	Probabilidad de la hipótesis dados los datos observados. Se basa en el teorema de Bayes: $P(\theta\|D)=\frac{(p(D\|\theta)\ \ p(\theta))}{p(D)}$
Parámetro	P-valor es la probabilidad de observar estos datos (u otros más extremos) bajo la hipótesis nula (¡no es la probabilidad de que la hipótesis nula sea verdadera!). Los parámetros se ajustan minimizando la suma de las diferencias al cuadrado entre cada valor observado y predicho.	Parámetros que hacen que el resultado sea más probable (MLE: estimadores de máxima verosimilitud). No hay definición de probabilidad de los parámetros, pero se definen los intervalos de confianza con dos unidades de probabilidad logarítmica (Edwards 1992).	Media, mediana o moda de la distribución posterior de los parámetros. El parámetro tiene una distribución de probabilidad: esta es la única aproximación con la que podemos estimar la probabilidad de diferentes hipótesis o valores del parámetro.

(iii) la interpretación de los resultados es fácil y directa ya que indican la probabilidad de que un parámetro tome un cierto valor; (iv) se pueden incluir medidas de incertidumbre, datos perdidos y diferentes niveles de variabilidad; (v) permite realizar propagaciones de error; (vi) permite especificar las distribuciones de los parámetros (dependientes a su vez de otros parámetros) cuando *a priori* sabemos cómo se distribuyen (*priors*); (vii) minimiza el uso de límites arbitrarios para tomar decisiones.

A pesar de las múltiples ventajas, la estadística bayesiana tan sólo recientemente comienza a implementarse con cierta frecuencia en el área de Ecología. Hemos realizado una búsqueda en Web of Science incluyendo las palabras "Bayesian" y "Ecology" entre los años 1994-2017 y hemos obtenido 14 495 resultados de los que el 90.7% se han publicado a partir del año 2000 y el 73.6% desde el año 2010 (**Fig. 1a**). Las principales áreas de publicación son: Ciencias Ambientales, Zoología, Genética y Ciencias de la Vida (**Fig. 1b**). El uso de la estadística bayesiana en Ecología puede estar menos generalizado por diversos motivos. En primer lugar, probablemente existe una inercia en el uso de métodos frecuentistas y de máxima verosimilitud, y una falta de conocimiento sobre cómo implementar procedimientos bayesianos. Sin embargo, también es importante resaltar que la inferencia Bayesiana no es el único procedimiento adecuado para los análisis de datos más comunes en Ecología. Por ejemplo, los resultados obtenidos por ajustes bayesianos son similares a los obtenidos mediante estadística frecuentista si no se usa un conocimiento *a priori* de los parámetros o modelos complejos. El conocimiento *a priori* del sistema es discutible y en ocasiones difícil de establecer, lo que ha dado lugar a un fuerte debate en la definición de la distribución *a priori* de los parámetros o la evaluación del ajuste del modelo (Dennis 1996; Gelman y Shalizi 2013). En segundo lugar, los procedimientos bayesianos son computacionalmente complejos y pueden requerir un elevado tiempo de ejecución para ajustar el modelo, especialmente si el tamaño muestral es elevado. Estas desventajas hacen que su uso no sea generalizado, sobre todo cuando los resultados son similares con otros procedimientos en principio más sencillos.

Las limitaciones en el uso de la estadística bayesiana están siendo subsanadas gracias al aumento de las capacidades de computación en paralelo de los equipos informáticos y a la mejora de programas informáticos cada vez más asequibles para usuarios no experimentados. Así, ha habido un reciente desarrollo de programas informáticos de código abierto tales como Stan (www.mc-stan.org), que hacen uso de algoritmos de muestreo más eficientes (p. ej. Hamiltonian Monte Carlo propuesto por Hoffman y Gelman 2014) que los utilizados por anteriores aplicaciones para la estadística bayesiana (p. ej. algoritmo de muestreo de Gibbs propuesto

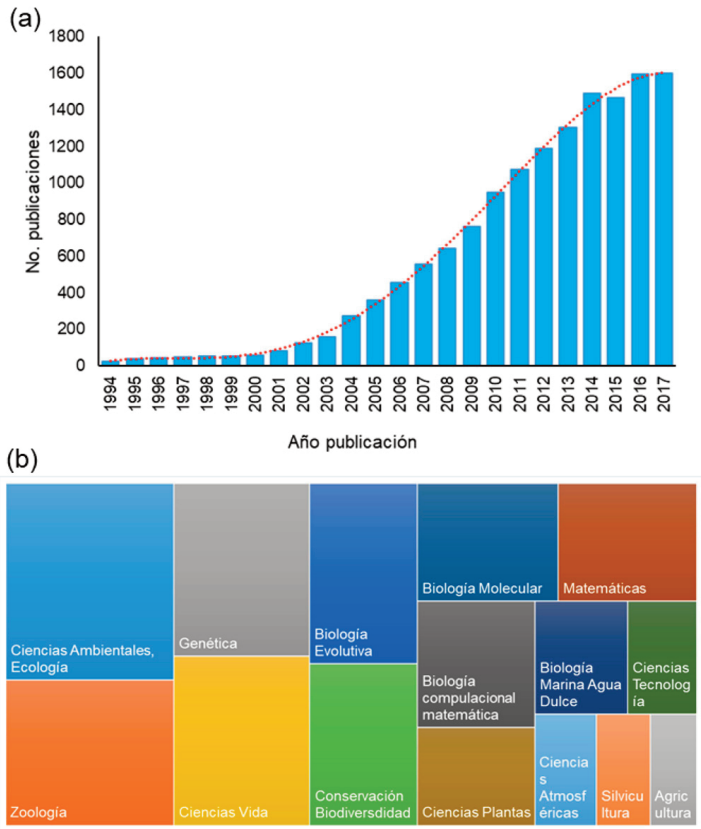

Figura 1. *a) Número de publicaciones en la base de datos Web of Science con las palabras clave "Bayesian" y "Ecology" a lo largo del tiempo. b) Principales áreas temáticas en que se dividen las publicaciones de la búsqueda anterior.*

Figure 1. *a) Number of publications on Web of Science using the key-words "Bayesian" and "Ecology" through time. b) The areas of the publications found in the previous search.*

por Geman y Geman (1984) y utilizado por programas como Wingbugs, o JAGS2, ver **Tabla 2**). Los recursos para implementar distintas aplicaciones bayesianas en R han proliferado de forma notable durante los últimos años, tal y como reflejan grupos de trabajo específicos tales como "*CRAN Task View: Bayesian Inference*" (https://cran.r-project.org/web/views/Bayesian.html), que cataloga exhaustivamente los recursos disponibles en R para realizar inferencia estadística bayesiana. Los recursos disponibles en R para

Tabla 2. *Programas estadísticos comúnmente usados en Ecología para realizar inferencia bayesiana y los paquetes en R disponibles para su ajuste.*
Table 2. *Statistical programmes commonly used in Ecology for Bayesian statistics and the R libraries available for the parameterisation.*

Programa y URL	Descripción	Paquete	Referencias
Stan	Lenguaje de programación probabilístico para la inferencia estadística escrito en C++. Permite el desarrollo de modelos jerárquicos bayesianos usando simulaciones MCMC ("*Markov chain Monte Carlo*") y algoritmos de Monte Carlo Hamiltonianos	brms; rstan	Bürkner (2016), Stan Development Team (2018).
BUGS (Bayesian inference Using Gibbs Sampling)	Programa que incluye programas flexibles como Wingbugs, Openbugs y MultiBugs que usa métodos MCMC y algoritmo de Gibbs	R2WinBUGS; R2OpenBUGS; rbugs; glmmBUGS	Brown y Zhou (2018); Sturtz et al. (2005)
JAGS (Just Another Gibss Sampler)	Programa para el desarrollo de modelos jerárquicos bayesianos usando simulaciones MCMC y algoritmos Gibbs	rjags; R2jags; runjags	Su y Yajima (2015)
BayesX (Bayesian Inference in Structured Additive Regression Models)	Programa para realizar inferencia bayesiana usando modelos de regresión aditivos (p. ej. GAM o "generalized additive models")	BayesX; BayesXsrc; R2BayesX	Belitz et al. (2017); Umlauf et al. (2015)
CUDA (Compute Unified Device Architecture)	Arquitectura de cálculo que aprovecha la potencia de la GPU (unidad de procesamiento gráfico) para proporcionar un incremento del rendimiento del sistema	cudaBayesreg	Ferreira da Silva (2011)

realizar inferencia estadística abarcan desde paquetes para, por ejemplo, ajustar modelos mixtos en marcos bayesianos (p. ej. paquete MCMCglmm, Hadfield 2010) hasta el desarrollo de paquetes específicos para implementar programas existentes de inferencia bayesiana desde R (**Tabla 2**).

Debido a la dificultad de enfrentarnos en solitario a los retos que supone aplicar marcos bayesianos al estudio de los procesos ecológicos, en la Unidad Docente de Ecología (UDE) del departamento de Ciencias de la Vida de la Universidad de Alcalá hemos constituido un grupo de trabajo EcoBayesUAH sobre estadística bayesiana (EcoBayesUAH@gmail.com). Hemos elegido utilizar Stan y su paquete rstan en R (R Core Team 2018) debido a la facilidad que ofrece para ajustar modelos complejos, su eficiente desarrollo y la buena documentación existente (ver http://mc-stan.org/users/documentation/). Nuestro grupo de trabajo, actualmente compuesto por 15 investigadores pre- y postdoctorales de la UDE, se reúne quincenalmente desde marzo de 2018 con tres objetivos iniciales: (1) fomentar el estudio de aspectos teóricos y aplicados de la estadística bayesiana a problemas en Ecología; (2) ganar experiencia en la implementación de modelos bayesianos mediante la realización y programación de ejemplos y ejercicios propuestos en libros y recursos online; y, (3) solucionar problemas que encontramos al aplicar los modelos bayesianos a nuestros datos, discutiendo los puntos críticos a conocer a la hora de ejecutar un modelo. Estamos encontrando las reuniones del grupo útiles en el proceso de aprendizaje y animamos a que se desarrollen actividades semejantes en otras instituciones. Nuestro grupo fomentará la colaboración con otros grupos de carácter similar o personas interesadas, especialmente para compartir conocimientos y experiencias.

Agradecimientos

Queremos agradecer al grupo de EcoInformática de la AEET por su apoyo para escribir esta nota y en especial a Ignasi Bartomeus, Francisco Rodríguez, Hugo Saiz y Antonio Luque, por la revisión de la misma. A.M.C.S. cuenta con una beca Juan de la Cierva (MINECO, IJCI-2014-19502). I.M-C. y M.F. están financiado por una Ayuda Postdoctoral del Programa Propio de la Universidad de Alcalá. P.R-B. y S.V-P cuentan con ayuda Atracción de Talento de la Comunidad de Madrid (2016-T2/AMB-1665 y 2017-T2/AMB-6035, respectivamente). V.C-A. y E.V-A están financiada por el programa de Becas Predoctorales de Medio Ambiente de la Fundación Tatiana Pérez de Guzmán el Bueno (2015 y 2016, respectivamente). Z.R. está financiada por un convenio entre la Fundación FIRE y LafargeHolcim España.

Referencias

Anderson, D.R., Burnham, K.P., Thompson, W.L. 2000. Null hypothesis testing: problems, prevalence, and an alternative. *The Journal of Wildlife Management* 64: 912-923.

Belitz, C., Brezger, A., Kneib, T., Lang, S., Umlauf, N. 2017. BayesX: Software for Bayesian Inference in Structured Additive Regression Models. Version 1.1. Disponible en: http://www.BayesX.org/.

Bolker, B. 2008. Ecological models and data in R. Princeton University Press. Princeton, New Jersey, Estados Unidos.

Brown, P.E., Zhou, L. 2018. glmmBUGS: Generalised Linear Mixed Models and Spatial Models with WinBUGS, Jags, and OpenBUGS. R package version 2.4.2. Disponible en: https://CRAN.R-project.org/package=glmmBUGS.

Burnham, K.P., Anderson, D.R. 2002. Model selection and multimodel inference: a practical information-theoretic approach. Springer-Verlag, New York, Estados Unidos.

Bürkner, P.C. 2016. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80: 1-28.

Clark, J.S. 2005. Why environmental scientists are becoming Bayesians. *Ecology Letters* 8: 2-14.

Dennis, B. 1996. Discussion: Should Ecologists Become Bayesians? *Ecological Applications* 6: 1095-1103.

Dziak, J.J., Coffman, D.L., Lanza, S.T., Li, R. 2017. Sensitivity and specificity of information criteria. *PeerJ PrePrints*: 5:e1103v3. DOI:10.7287/peerj.preprints.1103v3

Dormann, C.F., Calabrese, J.M., Guillera-Arroita, G., Matechou, E., Bahn, V., Barto´n, K., Beale, C.M., Ciuti, S., Elith, J., Gerstner, K., Guelat, J., Keil, P., Lahoz-Monfort, J.J., Pollock, L.J. Reineking, B., Roberts, D.R., Schröder B., Tuiller, W., Warton, D.I., Wintle, B.A., Wood, S.N., W¨uest, R.O, Hartig, F. 2018. Model averaging in ecology: a review of Bayesian, information-theoretic and tactical approaches for predictive inference. *Ecological Monographs* DOI: 10.1002/ecm.1309.

Edwards, M.B. 1992. Likelihood. Johns Hopkins University Press. Baltimore, Maryland, Estados Unidos.

Ferreira da Silva, A.R. 2011. cudaBayesreg: Parallel Implementation of a Bayesian Multilevel Model for fMRI Data Analysis. *Journal of Statistical Software* 44: 1-24.

Geman, S., Geman, D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6: 721–741.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian data analysis. Tercera edición. Chapman and Hall / CRC Press, New York, Estados Unidos.

Gelman, A., Shalizi, C.R. 2013. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology* 66(1):8-38.

Hadfield, J.D. 2010. MCMC Methods for Multi-Response Generalized Linear Mixed. *Journal of Statistical Software* 33: 1-22.

Hoffman, M.D., Gelman, A. 2014. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* 15: 1593-1623.

Johnson, J.B. Omland, K.S. 2004. Model selection in ecology and evolution. *Trends in Ecology and Evolution* 19: 101-108.

Kadane, J.B., Lazar, N.A. 2004. Methods and criteria for model selection. *Journal of the American Statistical Association* 99: 279-290.

Kéry, M. 2010. Introduction to WinBUGS for Ecologists. Academic Press. Burlington, MA, Estados Unidos.

Kruschke, J.K. 2015. Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan. Elsevier, USA.

McElreath, R. 2016. Statistical rethinking: A Bayesian Course with examples in R and Stan. CRS Press. Boca Ratón, FL, Estados Unidos.

Murtaugh, P.A. 2014. In defense of P values. *Ecology* 95: 611-617.

R Core Team 2018. R: A language and environment for statistical computing. En: R Foundation for Statistical Computing, Viena, Austria.

Su, Y. S., Yajima, M. 2015. R2jags: Using R to Run 'JAGS' from R. R package version 0.5-7 [usado septiembre 2015]. Disponible en: https://cran.r-project.org/web/packages/R2jags/index.html.

Stan Development Team 2018. RStan: the R interface to Stan. R package version 2.17.3. [usado mayo 2018]. Disponible en: http://mc-stan.org/

Sturtz, S., Ligges, U., Gelman, A. 2005. R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software* 12: 1-16.

Umlauf, N., Adler, D., Kneib, T., Lang, S., Zeileis, A. 2015. Structured Additive Regression Models: An R Interface to BayesX. *Journal of Statistical Software* 63: 1-46.

Wasserstein, R. L., Lazar, N.A. 2016. The ASA's statement on p-values: context, process, and purpose. *The American Statistician*: 70: 129-133