Ajuste, interpretación y presentación de modelos lineales: el valor *p* no es suficiente

Carlos Lara-Romero1

1. Dpto. de Investigación del Cambio Global. Instituto Mediterráneo de Estudios Avanzados (IMEDEA), Consejo Superior de Investigaciones Científicas, Carrer de Miquel Marquès, 21, 07190 Esporles, Illes Balears, España.

Autor para correspondencia: Carlos Lara-Romero [[carlos.lara.romero@gmail.com](file:///C:\Users\carlos\Downloads\carlos.lara.romero@gmail.com)]

El uso del contraste de hipótesis como herramienta de inferencia estadística está siendo cuestionado en los últimos años (Nuzzo 2014; Reinhart 2015). La principal fuente de críticas proviene del mal uso (y abuso) del valor *p.* Este valor nos indica la probabilidad, bajo la asunción de que no hay ningún efecto o diferencia real entre los tratamientos experimentales, de obtener un resultado igual o más extremo que el observado (Dytham 2011). El valor *p* no puede ser usado, por tanto, para cuantificar la fuerza de la evidencia en favor o en contra en un experimento en particular ni tampoco para estimar el tamaño del efecto de las variables explicativas sobre la variable respuesta (Wasserstein y Lazar 2016). En definitiva, no nos permite sacar ninguna conclusión más allá de si la hipótesis nula puede ser o no rechazada, en base a un valor *p* crítico previamente fijado (a menudo notado como α). El objetivo de esta nota es mostrar un ejemplo sencillo que ilustre la problemática en torno al valor *p* y aportar herramientas que permitan una interpretación y presentación de resultados basados en contrastes de hipótesis lo más objetiva y transparente posible.

Supongamos que queremos evaluar el efecto de un fertilizante sobre el crecimiento de un cultivo. Para ello, cuento con una muestra de 400 semillas que divido en dos grupos de 200. A un grupo le aplico el fertilizante (tratamiento) y a otro no (control). Al cabo de dos años mido el tamaño de todas las plantas (estimado mediante el diámetro máximo de la copa expresado en metros). El siguiente código de R nos permite crear ambos grupos y observar su distribución:

*> set.seed(20)*

*# Fijamos la semilla de los números aleatorios para que rnorm genere siempre la misma distribución*

*> control <- rnorm(200, 3, 1) #Tamaño medio 3, desviación típica 1*

*> tratamiento <- rnorm(200, 3.2, 1) #Tamaño medio 3.2, desviación típica 1*

*> grupo <- gl(2, 200, 400, labels = c("control", "tratamiento"))*

*> tamano <- c(control, tratamiento)*

*> boxplot(tamano ~ grupo)*

***[INSERTAR FIGURA 1]***

Podemos ajustar un modelo lineal para evaluar si existen diferencias “estadísticamente significativas” entre el grupo tratamiento y control fijando el valor crítico para *p* en 0.05 (Dytham, 2011). La hipótesis nula de nuestro experimento es que las plantas del grupo control y del grupo tratado con fertilizante no difieren en tamaño. Sólo necesitamos conocer tres funciones del lenguaje de programación R: *i*) la función *lm* ajusta el modelo lineal mediante el método de los mínimos cuadrados, *ii*) la función *summary* muestra en forma de tabla los parámetros ajustado del modelo (intercepto, pendientes y residuos) y *iii*) la función *anova* realiza un análisis de varianza para probar la hipótesis de que las medias de dos o más grupos son iguales (i.e., obtener el valor *p*).

*> lm1 <- lm(tamano ~ grupo)*

*> summary(lm1)*

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) 3.04733 0.07103 42.904 <2e-16 \*\*\**

*grupotratamiento 0.10211 0.10045 1.017 0.31*

*> anova(lm1)*

*Response: tamano*

*Df Sum Sq Mean Sq F value Pr(>F)*

*grupo 1 1.04 1.0427 1.0334 0.31*

*Residuals 398 401.56 1.0090*

Si nos fijamos en los coeficientes del modelo (R los denomina e*stimates*) vemos que el intercepto, en el que se encuentra alojado el grupo control tiene un valor de 3.05 ± 0.07. Esto quiere decir que el modelo estima un valor medio de 3.05 m para las plantas procedentes del control con un error típico de 0.07 m. ¿Qué sucede con el grupo tratamiento? Pues bien, el coeficiente del modelo para el tratamiento con fertilizante es 0.1 ± 0.1, lo que implica que el modelo predice que las plantas tratadas con el fertilizante tendrán un tamaño medio de 3.15 ± 0.1 m. A tenor del valor relativamente pequeño del coeficiente para el tratamiento con fertilizante, de su error típico (relativamente grande), y de la similitud de las distribuciones del grupo control y el grupo tratado con fertilizante, parece razonable concluir que el fertilizante no ha tenido un efecto “significativo” sobre el crecimiento de las plantas. El valor *p* obtenido para el estimador así lo sugiere (*F*1,398 = 1.03, *p* = 0.31). Por lo tanto, no parece que exista ningún inconveniente en utilizar únicamente los coeficientes del modelo ajustado y el valor *p* para interpretar los resultados de nuestro experimento.

Pero, ¿Qué sucede si doblo el tamaño muestral?

*> set.seed(20)*

*> control <- rnorm(400, 3, 1)*

*> tratamiento <- rnorm(400, 3.2, 1)*

*> grupo <- gl(2, 400, 800, labels = c("control", "tratamiento"))*

*> tamano <- c(control, tratamiento)*

*> boxplot(tamano ~ grupo)*

***[INSERTAR FIGURA 2]***

El diagrama de cajas muestra que la distribución del tamaño de planta para ambos grupos no parece haber cambiado mucho (como era de esperar ya que hemos mantenido la media y desviación típica para ambos grupos). Pero veamos qué sucede con los coeficientes del modelo y el valor *p*:

*> lm2 <- lm(tamano ~ grupo)*

*> summary(lm2)*

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) 2.99838 0.05126 58.496 < 2e-16 \*\*\**

*grupotratamiento 0.18359 0.07249 2.533 0.0115 \**

*> anova(lm2)*

*Response: tamano*

*Df Sum Sq Mean Sq F value Pr(>F)*

*grupo 1 6.74 6.7408 6.4141 0.0115 \**

*Residuals 798 838.65 1.0509*

El modelo que hemos obtenido ajusta un tamaño de 3 ± 0.05 m para el grupo control y 3.18 ± 0.07 m para el grupo de plantas tratadas con fertilizante. El valor *p* ha cambiado a 0.0115 (F1,798 = 6.41) siendo menor del valor crítico fijado en 0.05. Por lo tanto, a tenor del análisis de la varianza podríamos concluir que rechazamos la hipótesis nula y aceptamos la alternativa. Pero ¿por qué obtenemos ahora un resultado estadísticamente significativo si la magnitud real del efecto del fertilizante sobre el crecimiento es exactamente la misma que en el ejemplo anterior? Principalmente porque el valor *p* nos indica la probabilidad de que la diferencia observada entre grupos sea producto del azar, pero no aporta información sobre la magnitud del cambio en el tamaño que la adición de fertilizante provoca (el tamaño del efecto). En otras palabras, el valor de p depende no sólo de la magnitud real del efecto, sino también del tamaño de muestra.

El coeficiente de determinación (R²) puede ayudarnos a resolver la paradoja frente a la que aparentemente nos encontramos. El principal propósito de R² es determinar la proporción de variación de los resultados que puede explicarse por el modelo. Cuanto mayor sea el efecto del fertilizante sobre el tamaño de la planta, mayor será el porcentaje de varianza del tamaño de las plantas que estará determinada por la adición del fertilizante y, por lo tanto, mayor será el valor de R2. La parte inferior de la salida de la función *summary* contiene el valor de R2 del modelo. Pero también podemos utilizar la función *anova* para calcularlo directamente y así tener mayor control de nuestro análisis. Esta aproximación permite calcular, además, el porcentaje de varianza explicada para cada predictor (en caso de que hubiera más de una), mientras que la salida de la función *summary* muestra el valor de R2 para el modelo completo.

*> var <- anova(lm2)*

*> varss <- var$"Sum Sq"*

*> print(cbind(var, PctExp = varss / sum(varss) \* 100))*

*Df Sum Sq Mean Sq F value Pr(>F) PctExp*

*grupo 1 6.7408 6.7408 6.4141 0.0115 0.7974*

*Residuals 798 838.6529 1.0509 NA NA 99.2026*

El porcentaje de varianza explicada es bajo: 0.797% (R2 = 0.00797). ¿Qué implica este resultado? Pues que, aunque según el valor *p* podemos rechazar la hipótesis nula, el efecto del fertilizante en términos de incremento en el tamaño es prácticamente despreciable. ¿Son ambos resultados contradictorios? No hasta cierto punto. Al aumentar el tamaño muestral hemos incrementado nuestro poder estadístico y reducido la probabilidad de que por azar encontremos la diferencia observada entre grupos. Sin embargo, esto no es óbice para que la diferencia entre grupos experimentales sea pequeña, lo que indica que la aplicación del fertilizante no mejora el crecimiento de las plantas de manera relevante (sólo un 0.80% de la varianza en el tamaño de las plantas está determinada por la adición del fertilizante).

Veamos brevemente cómo se comportan los parámetros del modelo y la R2 cuando tenemos un tamaño del efecto grande (grupo control y tratamiento con tamaños muy diferenciados):

*> set.seed(20)*

*> control <- rnorm(400, 3, 1) #Tamaño medio 3, desviación típica 1*

*> tratamiento <- rnorm(400, 5, 1) #Tamaño medio 5, desviación típica 1*

*> grupo <- gl(2, 400, 800, labels = c("control", "tratamiento"))*

*> tamano <- c(control, tratamiento)*

*> boxplot(tamano ~ grupo)*

***[INSERTAR FIGURA 3]***

*> lm3 <- lm(tamano ~ grupo)*

*> summary(lm3)*

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) 2.99838 0.05126 58.50 <2e-16 \*\*\**

*grupotratamiento 1.98359 0.07249 27.36 <2e-16 \*\*\**

*> var<- anova(lm3)*

*> varss <- var$"Sum Sq"*

*> print(cbind(var, PctExp = varss / sum(varss) \* 100))*

*Df Sum Sq Mean Sq F value Pr(>F) PctExp*

*grupo 1 786.9233 786.923327 748.7779 8.41705e-117 48.40888*

*Residuals 798 838.6530 1.050944 NA NA 51.59112*

Los parámetros del modelo muestran que el tratamiento con fertilizante produce plantas más grandes (4.98 ± 0.07 m) que las plantas control (3 ± 0.05 m). El análisis de la varianza muestra resultados significativos (F1,798 = 748.78, p < 0.0001) y el valor de R2 han aumento drásticamente (R2 = 0.484).

Con este ejemplo se muestra que es necesario reportar los coeficientes del modelo y su error, y que no basta simplemente con calcular el valor de la F y el valor *p*. La presentación de los resultados del modelo debería incluir, además, la representación gráfica de las distribuciones de los datos y la estimación de la calidad de los modelos (por ejemplo, a través del cálculo de R2). La incorporación rutinaria de esta información en la presentación de resultados permitiría abandonar prácticas basadas en *piratear* el valor *p* hasta que alcance el deseado 0.05, un juego muy adictivo pero que poco tiene que aportar a la calidad y robustez de nuestras investigaciones. Dytham (2011) es una buena referencia sobre inferencia estadística basada en contraste de hipótesis. El lector también puede investigar aproximaciones estadísticas alternativas. El libro *Statistics Done Wrong,* deAlex Reinhart, puede ser un buen punto de partida.

# Agradecimientos

Han contribuido a su revisión: Ignacio Bartomeus, Nagore García Medina, Sara Varela, Antonio J. Pérez Luque y Francisco Rodríguez Sánchez.

# Referencias

Dytham, C. 2011. Choosing and using statistics. Wiley-blackwell, West sussex, UK.

Nuzzo, R. 2014. Statistical errors: P values, the “gold standard” of statistical validity, are not as reliable as many scientists assume. *Nature* 506:150–152

Reinhart, A. 2015. Statistics done wrong: the woefully complete guide. *No starch press*, San Francisco, USA.

Wasserstein, R. y Lazar, N. 2016. The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician* 70: 129-133.

**Pies de figura**

**Figura 1**. Diagrama de cajas con la distribución de los tamaños de planta para el experimento con bajo tamaño muestral (*n* = 400) y menor efecto del fertilizante sobre el crecimiento.

**Figura 2**. Diagrama de cajas con la distribución de los tamaños de planta para el experimento con mayor tamaño muestral (*n* = 800) y menor efecto del fertilizante sobre el crecimiento.

**Figura 3**. Diagrama de cajas con la distribución de los tamaños de planta para el experimento con mayor efecto del fertilizante sobre el crecimiento (*n* = 800).