*Paradigmas*
ISSN: 1909-4302
http://publicaciones.unitec.edu.co/ojs/

# Potenciar la utilidad de los métodos estadísticos: algunas sugerencias a partir de un estudio de caso

*Michael Wood, D.Phil.*
University of Portsmouth Business School,
Portsmouth, Reino Unido.

CORRESPONDENCIA AL AUTOR
michael.wood@port.ac.uk
mickofemsworth@gmail.com

INFORMACIÓN DEL ARTÍCULO
**Autorización traducción**: 01.04.2014
**Aceptado**: 04.04.2014

**• Para citar este artículo**
**• To cite this article**
**• Para citar este artigo:**
Wood, M. (2014). Potenciar la utilidad de los métodos estadísticos: algunas sugerencias a partir de un estudio de caso, *Paradigmas*, *6*, 37-73.

Originalmente publicado en *SAGE Open* (2013, enero-marzo), 1-11. doi: 10.1177/2158244013476873.


**Resumen**
En este texto se presenta una crítica a los métodos utilizados por los artículos científicos convencionales. Esto lleva a tres conclusiones generales sobre el uso convencional de los métodos estadísticos: en primer lugar, los resultados suelen presentarse de una manera innecesariamente confusa; en segundo lugar, el paradigma empleado para probar las hipótesis nulas tiene profundas fallas (por lo general, es preferible la estimación de la magnitud de los efectos y citar los intervalos o los niveles de confianza); y, en tercer lugar, hay varios problemas, independientemente de los conceptos estadísticos particulares empleados, los cuales limitan el valor de *cualquier enfoque estadístico*. Los dos primeros problemas se pueden remediar fácilmente, mientras que el tercero significa que, en algunos contextos, emplear ciertos enfoques estadísticos puede que no valga la pena. El estudio de caso que se emplea es un artículo sobre administración, pero problemas similares también surgen en otras ciencias sociales.

**Palabras clave:** Confianza, prueba de hipótesis, prueba de significancia de hipótesis nula, filosofía de la estadística, métodos estadísticos.

# Making Statistical Methods More Useful: Some Suggestions From a Case Study

# Potenciar a utilidade dos métodos estatísticos: algumas sugestões a partir de um estudo de caso

**Abstract**

I present a critique of the methods used in a typical article. This leads to three broad conclusions about the conventional use of statistical methods. First, results are often reported in an unnecessarily obscure manner. Second, the null hypothesis testing paradigm is deeply flawed: Estimating the size of effects and citing confidence intervals or levels is usually better. Third, there are several issues, independent of the particular statistical concepts employed, which limit the value of any statistical approach—for example, difficulties of generalizing to different contexts and the weakness of some research in terms of the size of the effects found. The first two of these are easily remedied—I illustrate some of the possibilities by reanalyzing the data from the case study article—and the third means that in some contexts, a statistical approach may not be worthwhile. My case study is a management article, but similar problems arise in other social sciences.

**Keywords:** confidence, hypothesis testing, null hypothesis significance test, philosophy of statistics, statistical methods

**Resumo**

Neste texto, é apresentada uma crítica aos métodos utilizados pelos artigos científicos convencionais. Isto leva a três conclusões gerais sobre o uso convencional dos métodos estatísticos: em primeiro lugar, os resultados costumam ser apresentados de uma maneira desnecessariamente confusa; em segundo lugar, o paradigma empregado para provar as hipótese nulas tem profundas falhas (geralmente, é preferível a estimação da magnitude dos efeitos e citar os intervalos ou os níveis de confiança); e, em terceiro lugar, há vários problemas, independentemente dos conceitos estatísticos particulares empregados, os quais limitam o valor de *qualquer enfoque estatístico*. Os dois primeiros problemas podem ser remediados facilmente, enquanto que o terceiro significa que, em alguns contextos, empregar certos enfoques estatísticos pode que não valha a pena. O estudo de caso que se emprega é um artigo sobre administração, mas problemas similares também surgem em outras ciências sociais.

**Palavras-chave**: confiança, prova de hipótese, prova de significância de hipótese nula, filosofia da estatística, métodos estatísticos.

# Introducción

Los métodos estadísticos son ampliamente utilizados en investigación. Existe una inmensa cantidad de teorías, consejos prácticos, ejemplos de buenas prácticas, etcétera, que apoyan dichos métodos. Sin embargo, algunos aspectos fundamentales de la manera en la que son usualmente utilizados parecen ser problemáticos en ciertas ocasiones, incluso en estudios publicados en respetadas revistas científicas cuyo proceso de revisión garantiza que se eviten errores obvios. Este artículo considera tres grandes áreas que a menudo generan inconvenientes: la facilidad de compresión por parte del lector de los conceptos empleados, el uso de la prueba de hipótesis y las cuestiones acerca de la utilidad del enfoque estadístico general, las cuales aplican independientemente de los métodos particulares que se empleen. Este artículo propone algunas alternativas posibles para abordar varios de estos aspectos problemáticos; por tanto, será de interés para cualquiera que se preocupe por la utilidad de los resultados estadísticos, ya sea como productores o consumidores del análisis estadístico.

La manera en que abordaré este problema consistirá en emplear un único artículo típico de investigación que haya sido publicado previamente, para luego revisar en él algunos de los problemas con el análisis y la presentación de los resultados, examinando además algunos abordajes alternativos. Así las cosas, el artículo que se convertirá en el estudio de caso proviene de una revista de investigación en el área de la administración,

aunque es probable que los problemas planteados sean relevantes para muchos otros proyectos de investigación en las ciencias sociales. Es claro que no es posible formular conclusiones que se puedan generalizar a partir de una muestra de un único artículo; sin embargo, en un enfoque de estudio de caso como este, mediante el análisis en profundidad de un ejemplo ilustrativo, se pueden sugerir *posibilidades* que pueden tener (y es muy probable que tengan) una aplicación mucho más amplia. En una situación ideal me hubiese gustado analizar una muestra representativa de estudios de investigación, pero el grado de detalle del que depende mi argumento hace que esta estrategia sea impracticable.

Elegí el texto de Glebbeek y Bax (2004) como mi artículo ilustrativo porque fue publicado en una revista respetada *(Academy of Management Journal)*. Dicho artículo trata un tema que puede ser comprendido sin un conocimiento detallado de la literatura, está escrito con claridad y el enfoque estadístico utilizado es bastante típico, al involucrar regresión y prueba de hipótesis. El objetivo no es producir una crítica de este artículo, sino explorar cuestiones de mayor cuidado en el empleo de la estadística en investigación. Estoy muy agradecido con el Dr. Arie Glebbeek por facilitarme los datos, lo cual me ha permitido formular algunas de las sugerencias discutidas en este artículo.

Glebbeek y Bax (2004) "probaron la hipótesis de que la rotación de los empleados y el rendimiento de la compañía tienen una relación en forma de U invertida: la rotación demasiado alta o baja es perjudicial" (p. 277), por lo que el nivel óptimo de rotación se sitúa en algún lugar en el medio. Para ello, analizaron los datos de "110 oficinas de una empresa de trabajo temporal" en los Países Bajos. Uno de sus análisis conduce a la figura 1 de este artículo, en la que cada uno de los puntos dispersos representa una única oficina, y el patrón general muestra cómo el rendimiento ("resultado neto por oficina" en florines holandeses por tiempo completo de empleado por año, en valores de 1995) varía con la rotación de los empleados. La línea continua representa una mejor predicción

para una oficina con una media de nivel de absentismo del 3.9 % y una media de edad del personal de 28.4 años, en una de las tres regiones de estudio. El método utilizado para hacer esta predicción se discute a continuación. Glebbeek y Bax mencionan esta gráfica, pero no la presentan, no obstante gráficas de modelos curvilíneos se presentan en dos artículos posteriores sobre el mismo tema en la misma revista (Shaw, Gupta & Delery, 2005; Siebert & Zubanov, 2009). Dichos autores realizaron algunas variaciones de este análisis; por ejemplo, intentaron relacionar el rendimiento a la rotación actual y a las rotaciones en los dos años anteriores. Sin embargo, para mis propósitos actuales, me centraré solo en los datos que son la base de la figura 1.

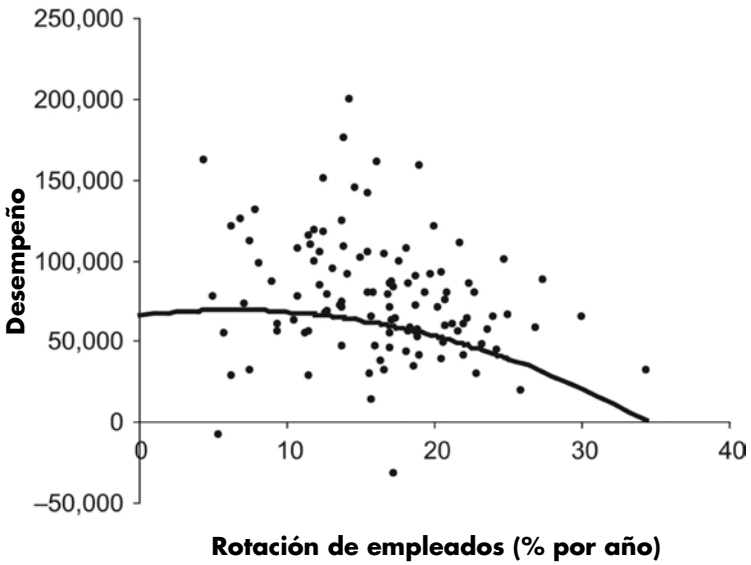


**Figura 1.** *Resultados y predicciones curvilíneas para la región 1 y la media de abstencionismo y edad*

Esta gráfica, y las fórmulas matemáticas que la sustentan, sugieren que el nivel óptimo de la rotación de personal es de aproximadamente el 6 %, esto es, para obtener el mejor nivel posible de rendimiento, el 6 % del personal se retiraría cada año. Cualquier cantidad por encima o por

debajo del 6 % probablemente conduciría a un rendimiento peor, y la figura 1 da una idea de qué tanto caería el rendimiento. La predicción para el rendimiento es de aproximadamente 70 000 unidades si la rotación de personal está en el nivel óptimo, pero solo de alrededor de 3000 unidades si es del 34 %. Esta información tiene un evidente interés práctico para los directivos de los departamentos de recursos humanos.

Para su análisis Glebbeek y Bax (2004) emplearon técnicas de regresión estándares, las cuales se resumen brevemente a continuación. El primer problema que discutiré, por tanto, es que estos métodos son innecesariamente obscuros y confusos; así, creería que a lectores poco familiarizados con estadística matemática les van a parecer muy difíciles algunos de los aspectos que se analizan en el próximo párrafo (muchos artículos en la literatura del área de la administración utilizan métodos estadísticos mucho más complejos, así que la tarea de hacer que el análisis sea claro es más urgente pero, de igual manera, posiblemente más difícil. Mi objetivo es simplemente demostrar las posibilidades que puede ofrecer un ejemplo sencillo).

Los modelos de regresión utilizaron como variable dependiente al "resultado neto por oficina" (Glebbeck & Bax, 2004, p. 281), mientras que a la rotación de personal y al cuadrado de la rotación como variables independientes, así como también tres variables de control (incluir expresiones al cuadrado en la regresión es un método estándar para probar la hipótesis acerca de relaciones en forma de U). Los resultados son presentados de manera convencional por medio de tablas de coeficientes de regresión estandarizadas para los diferentes modelos, complementadas por símbolos para denotar los diferentes rangos de los valores de $p$ (tablas 2 y 3 en Glebbeek & Bax, 2004). En todos los casos, los coeficientes resultaron según lo predicho por la hipótesis de la forma de U invertida: los coeficientes de regresión para las expresiones (lineales) de rotación fueron positivos, mientras que para las expresiones de rotación al cuadrado fueron negativos. Sin embargo, ninguno de los coeficientes de

las expresiones de rotación fueron estadísticamente significativos, aunque tres de los cuatro coeficientes para las expresiones al cuadrado fueron significativos ($p < 5\,\%$ en dos casos y 10 % en el tercero). En la sección "Discusión" del artículo se argumenta que esto ofrece un sustento razonable a la forma de U invertida en el contexto de la agencia de empleo en cuestión, pero que "no se observó con certeza" (Glebbeek & Bax, 2004, p. 277). El modelo de la primera tabla de análisis (A en la tabla 2, que corresponde a la figura 1 de este artículo) proporciona los coeficientes de regresión estandarizados para las expresiones de rotación y del cuadrado de la rotación del modelo con valores de 0.17 y -0.45 respectivamente, pero ninguno de ellos es significativo ($p > 10\,\%$). No hay nada en la tabla de resultados que le indique al lector el nivel óptimo de rotación del 6 % o sobre qué tanta diferencia hacen las desviaciones de esta figura (si bien se menciona información similar en la discusión).

Mi primera tarea es explicar cómo estos resultados pueden ser presentados de una forma más comprensible para el lector y sin perder rigurosidad. La figura 1, que *no* se encuentra en Glebbeek y Bax (2004) es un comienzo; pero es posible ir más allá, por ejemplo, con la tabla 1.

**Tabla 1.** *Parámetros de fácil comprensión para el modelo de la figura 1*

| | Mejor estimación |
|---|---|
| Ubicación del nivel óptimo (la rotación de personal anual para un mejor rendimiento) | 6.3 % |
| Curvatura[a] en forma de U invertida | 86.7 |
| Impacto estimado del aumento del 1 % de en el ausentismo | -3330 florines por ETC |
| Impacto estimado del aumento de un año en la edad promedio | -831 florines por ETC |
| Diferencia estimada entre regiones vecinas (la región 1 con el rendimiento más bajo) | 15 465 florines por ETC |
| Rendimiento máximo estimado para la región 1 y la media del ausentismo (3.8 %) y la edad promedio (28) | 69 575 florines por ETC |
| Estimación global de la exactitud de la precisión ($R^2$ ajustado) | 13 % |

| | Mejor estimación |
|---|---|
| Nivel de confianza para la hipótesis de la forma de U invertida | 65 % |

*Nota:* ETC: equivalente a tiempo completo.
[a] Mide qué tan curva es la línea, donde 0 representa una línea recta y números mayores representan una curvatura en forma de U invertida más pronunciada.

Mi segundo objetivo es revisar el marco de referencia de las pruebas de hipótesis. El artículo de Glebbeek y Bax (2004) prueba la hipótesis de que la relación tiene forma de U invertida. Hay varios problemas aquí: lo más evidente es la gráfica de la figura 1, que apenas si parece tener una forma de U invertida, porque la disminución en el lado izquierdo (rotación de personal baja) es muy leve. Podría de manera plausible interpretarse como un declive levemente curvo de la relación entre las dos variables. La hipótesis es un poco difusa, lo que dificulta una prueba clara.

Como es habitual en la investigación del área de la administración, Glebbeek y Bax (2004) ponen a prueba su hipótesis empleando la prueba de hipótesis nula y valores de *p* resultantes (ambos > 10 % para la figura 1). Sin embargo, hay varios argumentos muy fuertes en contra de esta manera de proceder, los cuales discutiremos a continuación. Una de las alternativas que sugerimos es citar un nivel de confianza para la hipótesis; este equivale solo al 65 % (la fuente de esta cifra se explica más adelante). Esto significa que, basándonos en los datos, podemos tener un 65 % de confianza de que un patrón en forma de U invertida resultaría si analizáramos *todos* los datos de situaciones similares. Esto parece mucho más útil que citar los valores de *p*.

Los datos (y, por tanto, las conclusiones de cualquier análisis) se basan en una organización, en un país y en una época (finales de los noventa) particulares: evidentemente no hay garantía de que se produciría un patrón similar en otros contextos. E incluso teniendo en cuenta esto, la aparente dispersión en la figura 1 sugiere que la rotación de personal

es solo uno de los muchos factores que afectan al rendimiento. Estos son algunos de los problemas más generales que son relevantes, independientemente del enfoque estadístico que se utilice para el análisis de los datos; así que mi tercer objetivo es revisarlos.

La literatura de investigación en el área médica ofrece un contraste instructivo al de la administración. Asegurar que los médicos sin formación estadística entiendan los resultados con precisión puede ser una cuestión de vida o muerte, a diferencia de la situación en el campo de la administración, donde la mayoría de los gerentes probablemente ignoran la mayor parte de la investigación que se publica en su campo. La prueba de hipótesis nula se utiliza mucho menos en medicina, y las directrices de las revistas (*British Medical Journal* [BMJ], 2011), así como las autoridades reguladoras (International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use [ICH], 1998), a menudo insisten en que se citen los intervalos de confianza (los cuales se discutirán más adelante). Asimismo, el hecho de que el contexto de la administración sea mucho menos predecible que el cuerpo humano estudiado por la investigación médica también tiene implicaciones en la manera en la que se utiliza la estadística.

En este artículo se analiza solo una investigación y los detalles del análisis son claramente específicos para este estudio en particular. Sin embargo, en la sección "Conclusiones y recomendaciones" infiero algunas recomendaciones más generales derivadas de este ejemplo; dichas generalizaciones deben, por supuesto, ser tentativas.

## Críticas a los métodos estadísticos utilizados en la investigación sobre administración

El análisis estadístico es de utilidad para muchas tareas; por ejemplo, al modelar los precios de viviendas, predice qué clientes potenciales son más propensos a comprar algo y permite, a su vez, el análisis de los resultados de los experimentos (Ayres, 2007). En ejemplos como estos la influencia de las variables de ruido puede ser considerable, pero los métodos estadísticos nos permiten ver con claridad y discernir una tendencia lo suficientemente confiable como para ser útil.

Existe una amplia literatura sobre los pros y los contras de las diferentes maneras de abordar la estadística (sobre todo el enfoque bayesiano y cómo se compara con las alternativas convencionales), así como también sobre la importancia de los métodos particulares y de los problemas con su uso (por ejemplo, Becker, 2005; Cashen & Geiger, 2004; Vandenberg, 2002), sobre la importancia y la dificultad de educar a los usuarios de la estadística y a los lectores de sus conclusiones y, por supuesto, sobre la derivación de nuevos métodos. Sin embargo, sorprendentemente existe muy poca literatura crítica de los métodos estadísticos y de su aplicación en general.

Un artículo que sí hace tal crítica al modelado estadístico —en las ciencias de la administración— es el de Mingers (2006). Él afirma que la estadística, en la práctica, adopta "un punto de vista empobrecido y empirista", con lo que quiere decir que en general no logra ir "por debajo de la superficie para explicar los mecanismos que dan lugar a eventos empíricamente observables". Sin lugar a dudas, esto es cierto en muchos contextos; por ejemplo, la figura 1 indica que la forma en U invertida de Glebbeek y Bax (2004) expresa una tendencia más bien débil que no logra incorporar los factores de ruido, cuya importancia es clara por

la dispersión de la figura 1 (y el valor de $R^2$ ajustado, el cual es del 13 %). Los análisis estadísticos como este proporcionan una explicación parcial o probabilística; si está disponible una explicación determinista satisfactoria, entonces no es necesario un modelo estadístico. En este sentido, la estadística es un método que se emplea como último recurso, pero es un enfoque potencialmente útil cuando no entendemos completamente lo que está sucediendo.

Un problema comúnmente reportado con la estadística es que muchas personas, entre ellas algunos investigadores y algunos de sus lectores, encuentran sus conceptos y técnicas difíciles de entender. Esto es particularmente cierto en relación con la prueba de hipótesis nula, la cual es un concepto complicado que implica tratar de demostrar la "significancia", asumiendo la verdad de una hipótesis nula que es probablemente falsa. El abordaje evidente para tratar con problemas de comprensión es buscar más y mejor formación en estadística, para lo cual existe una amplia literatura y varias revistas sobre este tema.

Un acercamiento alternativo para el problema de la educación es reconocer que el número de técnicas complicadas es demasiado alto como para que los investigadores y los lectores las aprendan todas (Simon, 1996, señala que a las personas generalmente les toma alrededor de diez años para llegar a ser especialistas, y esto impone un límite en la cantidad de experticia que se puede alcanzar), por lo cual se deben hacer esfuerzos para presentar los resultados de manera comprensible con el mínimo nivel de conocimiento técnico posible, sin sacrificar el rigor y la utilidad del análisis (Wood, 2002; Wood, Capon & Kaye, 1998). Esto podría darse a nivel de una redefinición de las medidas de los resultados para que sean más comprensibles o en el uso de métodos cuya base lógica esté más cerca del sentido común que la de los métodos convencionales basados en teoría de la probabilidad; esta es una de las ventajas de los métodos de remuestreo, como el *bootstrapping* (por ejemplo, Diaconis & Efron, 1983; Simon, 1992; Wood, 2005; Wood, Kaye & Capon, 1999). Sin embargo,

en la práctica estas oportunidades rara vez se toman. Por tanto, uno de mis objetivos en este artículo es demostrar algunas de las posibilidades.

**Tabla 2.** *Intervalos de confianza para los modelos lineales (3 en el panel A de la* tabla 2 *en* Glebbeek & Bax, 2004*)*

| | Mejor estimación | Límite inferior de 95% del IC | Límite superior de 95% del IC |
|---|---|---|---|
| Impacto estimado del 1% de aumento en la rotación de personal | −1778 | −3060 | −495 |
| Impacto estimado del 1% de aumento en el ausentismo | −3389 | −6767 | −10 |
| Impacto estimado del aumento de un año de la edad promedio | −731 | −3716 | 2254 |
| Diferencia estimada entre regiones vecinas (región 1 con el rendimiento más bajo) | 15066 | 5607 | 24525 |
| Estimación global de la exactitud de la precisión ($R^2$ ajustado) | 12% | | |

*Nota:* IC = intervalo de confianza.

El asunto de la facilidad de uso y comprensión por parte de los lectores es apenas un aspecto menor en el debate sobre los pros y los contras de los diferentes enfoques estadísticos. Otro asunto que merece ser mencionado aquí es el debate sobre el papel de las pruebas de significación de las hipótesis nulas (y los valores de *p*). Este es el método estándar utilizado en el área de la administración y en la mayoría de las ciencias sociales para responder a cuestionamientos sobre la confiabilidad con la que podemos hacer generalizaciones a partir de una muestra limitada de datos. Sin embargo, hay argumentos muy fuertes —presentados en numerosos libros y artículos en los últimos años— en contra del uso de estas pruebas en bastantes contextos (por ejemplo, Cohen, 1994; Gardner & Altman, 1986; Kirk, 1996; Lindsay, 1995; Morrison & Henkel, 1970; Nickerson, 2000), así como en favor de enfoques alternativos, tales como el uso de intervalos de confianza. Según Cohen (1994), "tras cuatro décadas de

duras críticas, el ritual de las pruebas de significancia de hipótesis nula (la mecánica decisión dicotómica en torno al criterio sagrado del 0.05) aún persiste" (p. 277). Este autor continua refiriéndose a la "mala interpretación casi universal" de los valores de *p*. Más recientemente, Coulson, Healey, Fidler y Cumming (2010) llegaron a la conclusión, basándose en una encuesta a 330 autores de artículos publicados, que la interpretación de los valores de *p* era "generalmente pobre", y no en referencia a los lectores, sino a dichos autores. Aquí no hay espacio para una revisión general de sus argumentos, pero discutiré los temas que se aplican a Glebbeek y Bax (2004) en la siguiente sección.

Por último, es importante tener en cuenta el hecho evidente de que hay alternativas a los métodos estadísticos. El más simple es utilizar estudios de caso para ilustrar y explorar lo que sea posible sin ningún intento de realizar estimaciones estadísticas de la población (Christy & Wood, 1999; Wood & Christy, 1999; Yin, 2003). Este es esencialmente el método que estoy adoptando en este artículo.

Pasemos ahora a la discusión de Glebbeek y Bax (2004), para lo cual voy a comenzar con los problemas de facilidad de uso; luego discutiré el abordaje de la prueba de hipótesis que se adoptó por dichos autores y, finalmente, consideraré las dificultades que enfrenta cualquier enfoque estadístico en este contexto.

## Facilidad de uso por parte de los lectores de la estadística en Glebbeek y Bax (2004)

Esto cubre tanto a la claridad de la forma en la que se describen los conceptos estadísticos, como a la claridad de los conceptos mismos. Usted como lector de este artículo podría pensar que los lectores de una

revista técnica de investigación deberían entender los tecnicismos sin ayuda. Sin embargo, dada la extensión de los conocimientos especializados necesarios, se hace razonable presentar los resultados de la manera más clara posible, siempre que ello no dé lugar a un artículo sustancialmente más largo o a sacrificar el rigor y el valor del análisis. Como se explicó en nuestra "Introducción", Glebbeek y Bax (2004) proporcionan los coeficientes de regresión estándares para las expresiones de rotación y cuadrado de la rotación del modelo, siendo dichos valores 0.17 y -0.45, respectivamente, con $p > 10\,\%$ en ambos casos, así como también los coeficientes estandarizados y los valores de $p$ para las tres variables de control. La figura 1 es un paso en la búsqueda de que dicha información sea más comprensible para los lectores, mientras que la tabla 1 presenta un análisis más detallado.

Solo una de las cifras en la tabla 2 de Glebbeek y Bax (2004) aparece en la anterior tabla 1: el valor de $R^2$ ajustado (0.13), el cual he reescrito como 13 % para enfatizar el hecho de que puede ser considerado como una proporción. Esta estadística es una estimación de "la reducción proporcional en el error del modelo nulo (sin variables explicativas) al modelo actual" (King, 1986, p. 676.), a la cual he resumido como "precisión de la estimación global predicha": un 100 % de precisión implicaría que todas las predicciones son precisas en su totalidad, mientras que un 0 % se referiría a una predicción que no hace uso de las variables independientes. Alternativamente, se podría utilizar la expresión "proporción de la variación explicada", pero esta parece menos directa, y la palabra "explicada" puede ser engañosa. El punto general aquí es la conveniencia de utilizar etiquetas que sean lo más informativas posible; es probable que la etiqueta "$R^2$ ajustada" no transmita nada a los inexpertos.

Las otras estadísticas de la tabla 1 son diferentes a las ofrecidas por Glebbeek y Bax (2004), quienes dan coeficientes de regresión *estandarizados*, los cuales son difíciles de interpretar de manera útil (King, 1986); la tabla 1 proporciona los coeficientes equivalentes sin estandarizar para

las variables de control. Por ejemplo, el coeficiente de regresión estandarizado para la primera variable de control —el absentismo— es de -0.19; en la tabla 1 se presenta el coeficiente equivalente no estandarizado (-3330) y su significancia se describe en función del escenario que se está modelando.

En lugar de describir una curva en forma de U invertida, en términos de los coeficientes de regresión estandarizados para las expresiones lineales y al cuadrado (0.17 y -0.45), podríamos utilizar los coeficientes no estandarizados (1097 y -86.7), pero estos aún son un poco difíciles de interpretar. De manera más útil, podríamos citar los coeficientes en la tabla 1: la ubicación del nivel óptimo (6.3 %) y la curvatura en forma de U invertida (86.7). Estas son matemáticamente equivalentes a los resultados de la regresión estándar presentadas por Glebbeek y Bax, en el sentido de que el primero puede calcularse a partir de este último por medio de fórmulas simples y viceversa (Wood, 2012a). No hay pérdida de información, pero están en un formato que facilita el relacionarlos con la realidad.

A modo de ejemplo, para el rendimiento de una oficina con una tasa de rotación de personal del 2 % por encima del nivel óptimo (8.3 %), en la región 1 con la media de absentismo y la media de edad, se haría la siguiente predicción utilizando las ecuaciones en Wood (2012a):

$$69\,575 - 86.7\,(8.3 - 6.3)^2 = 69\,228$$

En esta ecuación la curvatura representa claramente el grado en que el rendimiento disminuye a medida que la tasa de rotación de los empleados se aparta de su valor óptimo. El impacto de las variables de control se puede adicionar fácilmente: si, por ejemplo, el absentismo fuera del 5 % por encima de la media, entonces el rendimiento previsto se reduciría en $5 \times 3330$ a 52 578. Por último, no hay valores *p* en la tabla 1; en su lugar, se presenta un nivel de confianza para la hipótesis. La derivación de esto y la razón para no dar los valores de *p* se discuten en la siguiente sección.

## Problemas con la prueba de hipótesis y las alternativas sugeridas

Glebbeek y Bax "probaron la hipótesis de que la rotación de personal y el rendimiento de las empresas tienen una relación en forma de U invertida: una rotación excesivamente alta o baja son perjudiciales". A primera vista, el formular su objetivo de investigación en función de probar una hipótesis les proporciona un objetivo claro y un buen título para publicar. También es convencional en investigaciones que buscan ser "científicas"; sin embargo, en este caso, que no es en absoluto único, hay tres dificultades evidentes con la idea de poner a prueba esta hipótesis:

1) La hipótesis es bastante *difusa*. La figura 1 apenas si tiene forma de U invertida porque el rendimiento solo decae levemente mientras la rotación cae por debajo del nivel óptimo (y la falta de datos para valores bajos de rotación significa que la evidencia de esta parte de la línea es débil). Los puntos dispersos en la figura 1 bien podrían ser modelados de manera plausible por una línea recta que muestra una tendencia a la baja: en la práctica, estas dos posibilidades se funden la una en la otra. La idea de poner a prueba hipótesis deriva su estatus probablemente de famosas hipótesis como la de Einstein: $E = mc^2$; sin embargo, en este caso la hipótesis de forma de U invertida es mucho menos impresionante.
2) La hipótesis es bastante *obvia*. Si uno imagina una organización donde la tasa de rotación de personal está por encima del 100 %, el sentido común sugiere que probablemente el rendimiento sea relativamente pobre. Sin embargo, si la rotación fuese del 0 %, entonces esto sugeriría que es probable que se deba a una falta de nuevas ideas y energía, o a que la organización está en tan mal estado que nadie puede conseguir otro trabajo. Esto significa que debe haber un nivel óptimo de rotación

en algún lugar entre estos dos extremos, por lo que el patrón *debe* tener forma de U invertida.

3) Simplemente probar la hipótesis *ignora mucha información útil*. La información numérica, como la ubicación del nivel óptimo (6 % para la figura 1) o qué tanta diferencia hacen las desviaciones del nivel óptimo son irrelevantes desde el punto de vista de la prueba de la hipótesis, lo cual es una lástima, porque estas tienden a ser la información más interesante en la práctica. Podría suceder, por ejemplo, que en otros sectores el nivel óptimo de rotación de personal fuera mucho mayor. Este es el tipo de detalles que probablemente sea de interés tanto para los teóricos como para los profesionales.

Estos tres puntos sugieren que, en lugar de poner a prueba una hipótesis bastante difusa y obvia, un objetivo más útil para un proyecto de investigación de este tipo es *medir*, por ejemplo, el nivel óptimo de la rotación de personal, o *evaluar la forma de la relación* entre el rendimiento y la rotación de los empleados como se ilustra en la figura 1.

## Problemas con la prueba de la hipótesis nula

Los tres argumentos anteriores se refieren a la idea de la prueba de hipótesis en general. Como es convencional en la investigación en el área de la administración, el enfoque particular utilizado por Glebbeek y Bax (2004) para poner a prueba la hipótesis es el de crear una hipótesis nula y luego estimar la probabilidad de que los datos, o de manera similar los datos extremos (en tanto que medidos por la prueba estadística), podrían haber resultado de dicha hipótesis nula. Si este valor *p* es bajo, concluimos entonces que los datos no son consistentes con la hipótesis nula, por lo que debe ser falsa, y una hipótesis alterna debe ser verdadera.

Probar la hipótesis de la forma de U invertida de esta manera es particularmente problemático y será discutido en la próxima sección. Por lo pronto vamos a considerar el valor de *p* para el impacto estimado de la rotación (es decir, el coeficiente de regresión) en el modelo lineal (línea recta) para los datos de la figura 1, el cual es inferior a 0.01 (una cifra más exacta, usando la herramienta de regresión de Excel es 0.007). Esto se lleva a cabo utilizando la hipótesis nula de que la rotación de personal en realidad no tiene un impacto, sea positivo o negativo, sobre el rendimiento. El valor *p* indica que las fluctuaciones al azar darían lugar a un valor de -1778 (el valor observado en realidad) o menos, o +1778 o más, con una probabilidad del 0.7 %. Esta baja probabilidad significa que es poco probable que los datos observados sean consecuencia de la hipótesis nula, por lo cual podemos afirmar, a partir de la evidencia, que hay un verdadero impacto negativo que probablemente se repita si tomamos más muestras. Este es un argumento bastante complicado que la gente a menudo no entiende correctamente. Hay por lo menos tres problemas desde la perspectiva de la facilidad de uso para el lector:

1) La clave para comprender los valores de *p* es la hipótesis nula, no la hipótesis de interés. Glebbeek y Bax ni siquiera mencionan la hipótesis nula, pero esta es la base para la definición de los valores de *p*.
2) Cuanto *más fuerte* es la evidencia de un impacto de la rotación sobre el rendimiento, *menor* es el valor de *p*: como una medida de la fuerza de la evidencia, la escala de valores *p* es *inversa*.
3) Lo que los lectores quieren intuitivamente es una medida de qué tan probable es una hipótesis, y alguna indicación de la naturaleza y la fuerza de la relación entre las dos variables. Si bien los valores de *p* no responden a ninguna de estas preguntas, es casi inevitable que algunos lectores asuman que sí lo hacen. Esto no es, por supuesto, solo un problema desde la perspectiva de la facilidad de comprensión para el lector:

una medición que falla en decirle a las personas lo que estas desean saber no es una buena medición, aunque se entienda correctamente.

Además de los problemas de facilidad de uso hay una serie de problemas con la prueba de hipótesis nula, de los cuales uno es relevante aquí:

4) Puede que existan problemas al elegir una hipótesis nula sensata. Para probar su hipótesis de la forma de U invertida, Glebbeek y Bax (2004) tenían *dos* hipótesis nulas: que los valores de la población de los coeficientes de regresión de las expresiones lineal y al cuadrado eran ambos cero. Esto significa, en efecto, que *no hay* un patrón consistente, recto o curvo, entre las dos variables. Sin embargo, esto no es satisfactorio porque no hay manera obvia de combinar los dos valores de *p* y porque esta hipótesis nula es demasiado fuerte si se la toma literalmente. La figura 1 muestra una clara tendencia decreciente, de modo que la hipótesis nula es claramente falsa, pero esto no significa que la hipótesis curvilínea sea cierta. Las pruebas de hipótesis nulas pueden descartar efectivamente la hipótesis nula, pero no son útiles para proporcionar evidencia en favor de una hipótesis alternativa, si hay más de una de ellas. Como vimos anteriormente, el valor de *p* para el modelo lineal (línea recta) fue del 0.7 %; esto se basa en la hipótesis nula de que el aumento de la rotación de personal *no* tiene impacto en el rendimiento. Sin embargo, una vez más, esto es tan poco probable, que obtener evidencia de que es falsa no presenta realmente interés alguno. En ambos casos, la obvia hipótesis nula utilizada por Glebbeek y Bax no proporciona mucha información interesante.

## La medición de la magnitud del impacto y el uso de intervalos de confianza

Una de las recomendaciones que potencialmente puede abordar *todos* estos problemas es estimar el tamaño y la naturaleza de los efectos de la

rotación de personal en el rendimiento, y posteriormente expresar la incertidumbre sobre esta estimación a través de los intervalos de confianza. Empecemos discutiendo un modelo lineal (modelo 3 en Glebbeek y Bax, 2004), dado que es más sencillo. De acuerdo con el modelo lineal, la mejor estimación del impacto sobre el rendimiento de un 1 % adicional en la rotación del personal es -1778 (el coeficiente de regresión no estandarizado). Esta es una estimación de una cantidad numérica, no implica ninguna hipótesis y evita los problemas de centrarse en una hipótesis nula que sea difusa, obvia y distractora.

Sin embargo, no le hace frente al problema del error del muestreo: es probable que muestras diferentes produzcan resultados diferentes y, además, es poco probable que el resultado de la muestra sea exactamente correcto para toda la población. La prueba de hipótesis nula proporciona una forma poco satisfactoria de abordar este problema; así que los intervalos de confianza a menudo son recomendados como una alternativa (por ejemplo, los escritos de Gardner & Altman, 1986, en la revista *British Medical Journal*; Cashen & Geiger, 2004; y Cortina & Folger, 1998).

En la tabla 2, la mejor estimación para el impacto de la rotación de personal es que cada un 1 % adicional reducirá el rendimiento en 1778. Sin embargo, la cantidad exacta es incierta: el intervalo de confianza sugiere que el impacto real está en algún lugar entre una reducción de 495 unidades y una de 3060 unidades, con una confianza del 95 %. Este intervalo *excluye* el cero, lo que quiere decir que el nivel de significación debe ser inferior a 5 % (100 % - 95 %); de hecho, $p < 1$ % significa que el intervalo de confianza del 99 % también incluiría solo los valores negativos. Sin embargo, el intervalo de confianza del 95 % para el impacto de la edad incluye tanto valores positivos como negativos, es decir, que no es posible rechazar la hipótesis nula de que la edad no tiene algún impacto al nivel de significación del 5 %.

El presentar los impactos estimados y los intervalos de confianza como se hace en la tabla 2 evita cualquier mención de hipótesis nulas y

sus problemas asociados. Así, con ella nos centramos en la relación que realmente nos interesa y no en una hipótesis que casi con seguridad es falsa, y el intervalo de confianza expresa la incertidumbre de una manera mucho más transparente que el valor *p.* Por tanto, como hemos visto, toda la información proporcionada por los valores de *p* se puede derivar de los intervalos de confianza, pero estos también ofrecen una gran cantidad de información adicional.

A pesar de sus ventajas, los intervalos de confianza rara vez son citados en investigaciones en el área de la administración, mientras que la situación es muy diferente en la medicina: los intervalos de confianza son presentados ampliamente y son recomendados por las revistas científicas (por ejemplo, BMJ, 2011) y por las autoridades reguladoras (por ejemplo, ICH, 1998).

## Los niveles de confianza para las hipótesis

Infortunadamente, este abordaje no es tan fácil a la hora de evaluar la confianza de la conclusión de que la curva tiene forma de U invertida, porque esta se mide por dos parámetros, ubicación y curvatura, en la tabla 1 (la ubicación es relevante para la existencia de una forma de U invertida ya que si se produce el nivel óptimo para un valor negativo de la rotación de personal, entonces no habrá una U invertida en la parte positiva y relevante de la gráfica). Podríamos producir intervalos de confianza para la curvatura y para la ubicación de la óptima rotación de personal en la tabla 1, pero el hecho de que aquí existan dos cantidades, haría que esto fuera difícil de manejar y de interpretar. Por tanto, examinemos a continuación cómo podríamos aplicar la idea de confianza a una hipótesis.

El método del *bootstrapping* ofrece una manera fácil de abordar este problema. La idea del *bootstrapping* implica el uso de la muestra de datos

para generar las "remuestras" que imitan otras muestras de la misma fuente. Un grupo de tales remuestras se puede utilizar luego para ver qué tan variables es probable que sean las diferentes muestras y, por tanto, el grado de confianza que podemos tener acerca de la hipótesis. En el presente caso, la figura 2 muestra las curvas de estimación generadas a partir de cuatro remuestras, así como también la estimación a partir de los datos originales (como en la figura 1).

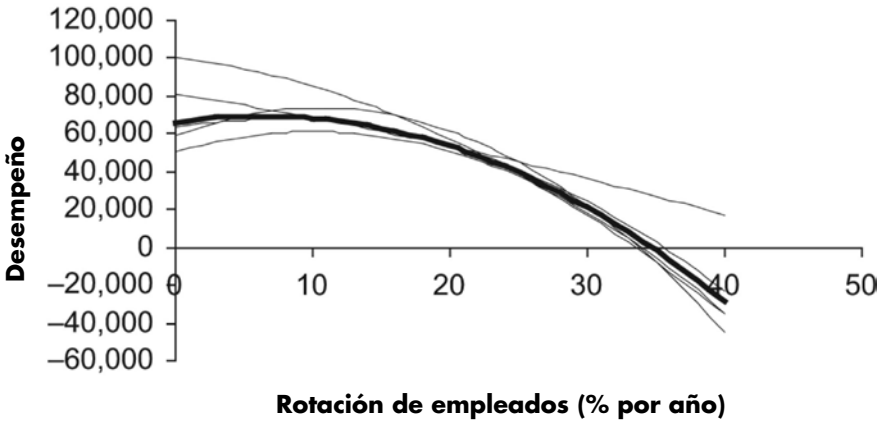


**Figura 2.** *Estimaciones de los datos (negrita) y cuatro remuestras para el modelo de la figura 1*

La figura 2 proporciona una clara demostración del hecho de que la figura 1 puede ser engañosa, simplemente porque dos de las cinco líneas no tienen forma de U invertida. Con 10 000 remuestras, el 65 % de ellas produjo una forma de U invertida (con una curvatura negativa y un valor positivo para la ubicación del nivel óptimo). *Esto sugiere un nivel de confianza para la hipótesis de la forma de U invertida de un 65 %.*

En el apéndice presento una explicación un poco más detallada del procedimiento de *bootstrapping*. Además, es posible encontrar una extensa literatura sobre *bootstrapping*: hay explicaciones simples en Diaconis

y Efron (1983), Simon (1992) y Wood (2005), y una explicación más detallada del procedimiento utilizado aquí en Wood (2012a).

Sin embargo, al contrario del intervalo de confianza en la tabla 2, el indicar simplemente un nivel de confianza del 65 % para la hipótesis de forma de U invertida proporciona poca indicación de cuán aguda es la curva o cuál es el nivel óptimo de rotación de personal. La hipótesis no distingue entre las líneas curvadas ligera o fuertemente.

Podemos hacerle frente a este problema en cierta medida al evaluar un nivel de confianza para una hipótesis más fuerte. Por ejemplo, podríamos insistir en que para una forma de U invertida razonable, el gráfico tendría que decrecer al menos 10 000 unidades del lado izquierdo (el nivel de confianza en este caso llega al 40 %). Sin embargo, el punto de corte elegido es arbitrario porque las hipótesis como esta son inevitablemente nebulosas.

También deberíamos señalar que, en sentido estricto, un nivel de confianza para un intervalo o para una hipótesis *no* es lo mismo que para la probabilidad de la verdad de la hipótesis o de que el verdadero valor del parámetro se encuentre en un intervalo (Nickerson, 2000, pp. 278-280). Al igual que con la prueba de hipótesis nula, los intervalos de confianza están basados en las probabilidades de que los datos de la muestra *proporcionen* la verdad acerca de un parámetro. Para revertir estas probabilidades y encontrar la probabilidad de una hipótesis, *dados* los datos de muestra, tenemos que utilizar el teorema de Bayes y tener en cuenta probabilidades anteriores. Sin embargo, para muchos parámetros, incluyendo la pendiente de una línea de regresión y la diferencia de dos medias, el equivalente bayesiano de un intervalo de confianza —el intervalo de credibilidad— es idéntico al intervalo de confianza convencional (Bayarri & Berger, 2004; Bolstad, 2004, pp. 214-215, 247), siempre que usemos distribuciones de probabilidad *a priori* "planas" (es decir, asumimos una distribución de probabilidad *a priori* uniforme) para el análisis bayesiano. *Esto significa que a menudo es razonable interpretar los intervalos y*

*niveles de confianza en términos de probabilidades:* la única pérdida (desde la perspectiva bayesiana) es que no se incorpora ninguna información de la distribución de probabilidad *a priori*.

## Problemas generales sobre la utilidad del abordaje estadístico

Supongamos ahora que las recomendaciones anteriores se han tenido en cuenta: los resultados se presentan de manera tan comprensible para los lectores como sea posible, se utilizan intervalos de confianza siempre que sea posible y, cuando no lo es, se utilizan los niveles de confianza en vez de valores de *p* para cuantificar la incertidumbre. Los métodos estadísticos se han ajustado de acuerdo con la discusión anterior, pero aún quedan cuestiones importantes acerca de la utilidad de los métodos estadísticos en general (estas son relevantes, independientemente de los métodos y conceptos particulares utilizados).

Las conclusiones de la investigación estadística no son deterministas, pero son cualificadas por probabilidades, promedios o conceptos equivalentes. Es posible que la relación entre la rotación de personal y el rendimiento sea demasiado compleja para una explicación completa y determinista de todas las variables y sus efectos exactos; por tanto, el abordaje estadístico merece tenerse en cuenta a falta de una mejor alternativa. Las siguientes cuestiones son relevantes ante las preguntas sobre el beneficio de un abordaje estadístico.

## Las fortalezas y debilidades de los resultados estadísticos

La figura 1 muestra una forma de U muy débil en el sentido de que el declive del lado izquierdo apenas es ligero. La precisión estimada de la predicción ($R^2$ ajustado) es solo del 13 %, como es evidente a grandes rasgos a partir de la dispersión de la figura 1. Además, el nivel de confianza de que esta forma sea una característica de las poblaciones subyacentes y de que se repetirá en otras muestras es bastante bajo: un 65 %. En consecuencia, los resultados son débiles en todas estas tres dimensiones: *la fuerza del efecto* (qué tan definida es la forma de U invertida), *la consistencia del efecto* (es evidente que hay otros factores además de la rotación de personal que son responsables del buen o del mal rendimiento) y *el nivel de confianza para la hipótesis.* Por otra parte, como se discute en Glebbeek y Bax (2004), la idea predominante es que el rendimiento tiene la tendencia a caer a medida que aumenta la rotación de personal, si bien el sentido común —por las razones expuestas anteriormente— sugiere que de algún tipo de forma de U invertida es casi inevitable. Por todas estas razones la figura 1, y los datos y análisis detrás de esta, parece que aportan poco valor a lo que ya se conoce.

## La naturaleza y la generalidad del contexto objetivo

Si se tuviesen que utilizar la prueba de significancia la de hipótesis nula o los intervalos de confianza para analizar los datos de Glebbeek y Bax (2004), deberíamos asumir que la muestra utilizada es una muestra aleatoria de una población objetivo específica. En la práctica, la utilizada por Glebbeek y Bax fue muestra por conveniencia: los datos concernían a *todas* las sedes de la agencia de empleo en cuestión, la cual fue

elegida simplemente porque estaba disponible. A primera vista, no hay ninguna población objetivo más allá de la muestra, la figura 1 es exacta en relación con esta muestra y no hay incertidumbre debido a un error de muestreo. Entonces, ¿qué sentido puede tener el nivel de confianza del 65 % o los valores de *p*?

Si tomásemos otra organización similar con las mismas fuerzas en juego, o la misma organización en un momento diferente, sería muy poco probable obtener exactamente la figura 1. Una multiplicidad de factores de ruido significaría que la siguiente muestra sería diferente, y quizás similar a una de las cuatro remuestras de la figura 2. Tenemos que saber qué tan variables pueden ser las muestras debido a estos factores aleatorios, de manera que podamos evaluar los niveles de confianza para las conclusiones.

La terminología estándar de las poblaciones aquí es un poco incómoda, así que voy a utilizar la expresión *contexto de destino o contexto objetivo* (*target context*) para referirme al contexto al que la investigación se dirige, del cual la muestra puede considerarse razonablemente una muestra aleatoria y al cual los resultados se pueden generalizar razonablemente. En la ausencia de un proceso de muestreo formal, esta noción es inevitablemente vaga (la población de destino sería una población hipotética de oficinas "similares" a las de la muestra, pero esto parece claramente difícil de visualizar).

La naturaleza de este contexto de destino merece un examen cuidadoso. Los resultados de Glebbeek y Bax (2004) se basan en datos de una sola organización en un solo país (los Países Bajos), en un periodo de tiempo específico (1995-1998), por lo que tal vez cabría preguntar: ¿el contexto de destino deberían ser organizaciones similares en el mismo país, en un momento similar en la historia? Obviamente, un contexto diferente podría dar lugar a un patrón diferente de relación entre la rotación de personal y el rendimiento, por lo que sus conclusiones son cualificadas por palabras tales como "puede". La hipótesis de la forma de U invertida

en ningún sentido está demostrada en términos generales, como lo reconocen Glebbeek y Bax, pero ellos han demostrado que es una posibilidad ya que se aplica a este contexto de destino.

El alcance del contexto de destino es una dificultad clave en esta y en la mayoría de las investigaciones empíricas en el campo de la administración. El problema con hacer que el contexto de destino sea muy amplio es que no es fácil obtener muestras razonables, y los factores contextuales específicos son susceptibles de sumarse a los factores de ruido, haciendo difícil obtener resultados claros. Por otro lado, si el contexto es demasiado estrecho, esto puede llevar a muchos a concluir que la investigación tiene poca relevancia.

La noción de un contexto de destino se vuelve más sutil aún al tener en cuenta la dimensión del tiempo o cuando extendemos dicha idea para incluir lo *posible*. La mayoría de investigaciones sobre administración —y la de Glebbeek y Bax (2004) no es la excepción— tiene como su propósito final el mejorar algún aspecto de la administración a futuro. La meta de la investigación empírica es probar y comprobar lo que funciona y lo que no funciona. Así las cosas, imaginemos, en aras de la discusión, que teníamos un conjunto de datos similares provenientes de una muestra representativa de un contexto de destino más amplio: todas las grandes organizaciones en Europa durante los últimos diez años. Esta sin duda sería más útil que la muestra que hemos venido manejando; sin embargo, debemos recordar que el contexto podría cambiar en los próximos años, por lo que tendríamos que ser cautelosos al realizar una generalización hacia el futuro. La dificultad con casi cualquier contexto de destino para la investigación estadística en el área de la administración es que depende significativamente de factores contextuales que pueden cambiar en el futuro. Si bien quizás sea una meta loable el ampliar nuestras teorías para incorporar dichos factores contextuales, esto podría hacer que las teorías resultantes sean desordenadas y difíciles de manejar. ¿Será entonces que tal vez deberíamos tratar de centrarnos únicamente en las

verdades centrales e inmutables? La dificultad con ello, por supuesto, es que quizás no haya otra verdad inmutable más allá del hecho de que las cosas varían de una situación a otra —y en este caso el análisis estadístico solo tendría un interés limitado—.

Hacer una comparación con la investigación médica es ilustrativo: allí el contexto de destino podría ser la gente, tal vez de una edad o sexo particulares. Por su parte, en una investigación en administración, el contexto de destino típicamente serían organizaciones o personas en un contexto de negocios en particular. Ahora, el problema que tiene el contexto de negocios, pero no tiene el contexto médico, es que es un contexto artificial que puede diferir radicalmente entre diferentes lugares o diferentes periodos de tiempo, lo que dificulta realizar extrapolaciones de un contexto a otro. Por ejemplo, luego de obtener unas conclusiones sobre cómo la rotación de personal afecta el rendimiento de una agencia de empleo durante un cierto momento de auge económico en los Países Bajos, ¿podríamos asumir que estas aplican a las universidades de Inglaterra del próximo siglo o a los sitios web de redes sociales en California? Es casi seguro que no. En contraposición, en medicina incluso una investigación que tenga una muestra local limitada puede tener un valor mucho más general, simplemente porque las personas son mucho menos variables desde un punto de vista médico, que los entornos empresariales desde un punto de vista administrativo.

## La necesidad de utilizar variables fácilmente medibles

La investigación estadística tiene que centrarse en variables fácilmente medibles; de lo contrario, no es práctico obtener tamaños de muestras útiles. En el presente caso, la rotación de personal, el rendimiento y las variables de control (grado de ausentismo, edad y región) son todas

fáciles de medir y se encuentran disponibles. Evidentemente, es probable que existan variables más "blandas", las cuales podrían no ser tan fácilmente definidas o recolectadas y que, sin embargo, pudieron tener una influencia importante en el rendimiento.

### Alternativas al abordaje estadístico

Por último, debemos recordar otros enfoques, ya sea como alternativa o como adiciones al abordaje estadístico. Los más evidentes son los estudios de caso y la investigación cualitativa, los cuales "pueden proporcionar descripciones robustas y detalladas en contextos de la vida real" (Gephart, 2004, p. 455), y podrían aclarar *cómo* la alta rotación de personal influye en el rendimiento o dar información acerca de escenarios particularmente interesantes (tal vez incluso de cisnes negros [Taleb, 2008]) que no salen a la luz a través de predicciones sobre lo que ocurre en promedio, como en la figura 1. Lo anterior de ninguna manera es un argumento contra el uso de un análisis estadístico, sino más bien podría ser uno en favor de complementar los análisis estadísticos con estudios cualitativos más detallados, pero de muestras más pequeñas. Este principio de métodos mixtos parece ser ampliamente aceptado en teoría, aunque no siempre en la práctica.

## Conclusión y recomendaciones

He revisado tres grupos de problemas relacionados con el análisis estadístico en mi estudio de caso. El primero tiene que ver con su facilidad de uso: esto se puede mejorar mediante el uso de nombres más apropiados para los conceptos (por ejemplo, precisión de la estimación global

predicha en lugar de $R^2$ ajustado) y cambiando los parámetros mismos citados (por ejemplo, la ubicación del nivel óptimo y el grado de curvatura invertida en lugar de los coeficientes de regresión para las expresiones lineales y cuadrados). Algunas posibilidades más se ilustran en las tablas 1 y 2. Los autores incluyen generalmente (pero no siempre) este tipo de información en su discusión, pero mi sugerencia es que los datos recogidos en las tablas de resultados debería presentarse en una forma más fácil de usar para los lectores que la convencional (*véanse* las tablas 1 y 2).

No hay una pérdida de información o de rigor al hacer esto: no es una cuestión de "banalizar", sino más bien mejorar la accesibilidad de la investigación y de aumentar la probabilidad de que los resultados sean interpretados correctamente por el mayor número de lectores que sea posible. He empleado el artículo en el que se basa mi estudio de caso como ejemplo: algunas de las sugerencias se podrían aplicar directamente a otras investigaciones, pero mi objetivo principal es establecer un principio, una tesis.

El segundo problema se refiere a la comprobación de hipótesis. Los artículos de investigación estadística *no* necesitan listas de hipótesis para probar, cuyas verdades o falsedades son a menudo del todo evidentes. Una meta más sensata es evaluar la relación entre variables, como estadísticas numéricas o en forma de gráficas. La investigación cuantitativa convencional basada en la prueba de hipótesis es a menudo —y extrañamente— no cuantitativa, porque a los lectores se les dice muy poco sobre la *magnitud* de los impactos, las diferencias o las relaciones. Por otra parte, para expresar las dudas que resultan de errores de muestreo, en lugar de utilizar los valores de $p$ (los cuales son complejos, poco informativos y ampliamente mal interpretados) a menudo podemos expresarlas como intervalos de confianza. En el ejemplo que revisamos esto resuelve todas las dificultades identificadas con la prueba de hipótesis nula.

A pesar de esto, a veces pueden existir razones para probar una hipótesis. La verdad o la falsedad de la hipótesis de la forma de U invertida

no es fácil de resumir mediante un solo número: no existe una medida obvia para la "forma de U invertida", por lo que no hay más opción que formular los objetivos de investigación en función de una prueba de hipótesis. No obstante, mi sugerencia es que, en lugar de tratar de utilizar los valores de *p* para evaluar la fuerza de la evidencia para esta hipótesis, utilicemos un *nivel de confianza*, lo que para la figura 1 llega al 65 %. También podríamos hacer la hipótesis un poco más fuerte, como se explicó anteriormente (el nivel de confianza para la hipótesis más fuerte es del 40 %). Estos niveles de confianza son mucho más sencillos y fáciles de usar que los valores de *p* convencionales.

En tercer lugar, debemos tener en cuenta el valor de los métodos estadísticos en general. Los aspectos a tener en cuenta son la "fuerza" de los resultados estadísticos, la naturaleza del contexto de destino al que los resultados se pueden generalizar y hasta qué punto la necesidad de utilizar con facilidad variables medibles puede llegar a distorsionar la investigación. La ventaja de los métodos estadísticos es que nos permiten ver —a través de la "niebla" que generan las variables de ruido— patrones como la curva en la figura 1; sin embargo, la figura 1 también nos muestra la falta de claridad de muchas de las hipótesis estadísticas dado que apenas si tienen forma de U invertida. Además, el hecho de que los datos provengan de un muestreo por conveniencia y limitado implica que es difícil generalizar las conclusiones a otras organizaciones, momentos y lugares.

Estas conclusiones y sugerencias se basan en una única investigación. Evidentemente no se pueden extraer conclusiones definitivas acerca de qué tan típicos son algunos de los problemas descritos, y las sugerencias detalladas pueden ser aplicables solo a Glebbeek y Bax (2004). El método de *bootstrap* para evaluar el nivel de confianza en la hipótesis es útil en este caso pero, por ejemplo, para otras hipótesis relativas a la igualdad o la diferencia de dos medias, puede que sean más factibles otros métodos más simples de evaluación de los niveles de confianza (Wood, 2012b). En el ejemplo anterior, he evaluado la precisión del modelo como

un porcentaje de precisión ($R^2$ ajustado), pero en otros estudios en los que haya un interés en hacer predicciones, puede ser más sensato dar una estimación del error típico en una predicción a partir del modelo (por ejemplo, el error estándar). Por tanto, la utilidad de cada método estadístico depende del contexto y de la finalidad de la investigación.

Sin embargo, me parece que no cabe duda de que muchos de los problemas que hemos destacado son muy comunes, de manera que los comentarios y sugerencias probablemente sean aplicables a muchos otros artículos (por ejemplo, diez de los once artículos de investigación en la edición de septiembre de 2010 de la revista *British Journal of Management* presentaron valores de *p,* y ocho se organizaron en torno a hipótesis formales). Por lo general, se hace muy poco esfuerzo por presentar los resultados en un formato fácil de entender. Los resultados se citan a menudo como la confirmación o no de la hipótesis, las cuales suelen ser difusas o evidentes, y con poca o ninguna discusión sobre la magnitud de los impactos o de los efectos. Asimismo, los resultados estadísticos son a menudo bastante débiles (aunque esto puede ser disfrazado por el uso de valores de *p* y grandes muestras) y pueden estar basados en muestras que hacen que sea difícil extrapolar los resultados de manera creíble a los diferentes ambientes futuros probables. En consecuencia, es probable que en algunos contextos valga la pena cuestionar la idea enraizada de la utilización de métodos estadísticos.

# Apéndice

## El método de *bootstrap* para derivar intervalos y niveles de confianza

Hay 110 registros en los datos sobre la rotación de personal. El método *bootstrap* utiliza remuestreo con reemplazo; esto significa que elegimos uno de estos registros al azar, luego lo reemplazamos, por lo que estamos empezando de nuevo con la muestra original, y luego elegimos otro al azar, y así sucesivamente hasta que tengamos un remuestreo de 110. Todos los registros en el remuestreo provienen de la muestra, pero algunos registros en la muestra pueden aparecer varias veces en el remuestreo, y otros en ningún momento. Sucesivamente repetimos este procedimiento varias veces para generar múltiples remuestras.

Ahora, imaginemos que la población de la que se extrae la muestra real sigue el mismo patrón que la muestra. Esto significa que el 0.91 % (1/110) de la población será como el primer registro de la muestra, el 0.91 % igual que el segundo, y así sucesivamente. Esto a su vez significa que, para tomar una muestra aleatoria de la población, queremos elegir un registro como el primer miembro de la muestra con una probabilidad del 0.91 %, y de manera similar para la segunda, tercera, y así sucesivamente. Pero esto es exactamente lo que el remuestreo con reemplazo logra, por lo que estas remuestras pueden considerarse muestras aleatorias de la población imaginaria. Esta no es la población real, pero parece razonable suponer que es similar, por lo que la variación entre las remuestras da una buena idea del error de muestreo cuando se toma una población real. En la práctica, la experiencia indica que el *bootstrapping* generalmente ofrece resultados muy similares a los métodos convencionales. Más detalles y un vínculo al *software* utilizado para derivar los resultados anteriores se ofrecen en Wood (2012a).

## Agradecimientos

Estoy agradecido con el Dr. Arie Glebbeek por poner sus datos a mi disposición y con los dos árbitros anónimos por sus valiosas sugerencias.

Por su parte, el editor de la revista *Paradigmas* agradece a los Doctores Francia Restrepo y Johan Sebastian Hernández Botero por la revisión final de la traducción de este manuscrito.

## Declaración de conflicto de interés

El autor declara no tener ningún conflicto de interés potencial con respecto a la investigación, la autoría o la publicación de este artículo.

## Financiamiento

El autor no recibió apoyo financiero para la investigación o la autoría de este artículo.

# Referencias

Ayres, I. (2007). *Super crunchers: How anything can be predicted*. Londres: John Murray.

Bayarri, M. J., & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, *19*, 58-80.

Becker, T. E. (2005). Potential problems in the statistical/control of variables in organizational research: A qualitative analysis with recommendations. *Organizational Research Methods*, *8*, 274-289.

*British Medical Journal*. (2011). Research. Recuperado de http://resources.bmj.com/bmj/authors/types-of-article/research

Bolstad, W. M. (2004). *Introduction to Bayesian statistics* (2ª ed.). Hoboken, NJ: Wiley.

Cashen, L. H., & Geiger, S. W. (2004). Statistical power and the testing of null hypotheses: A review of contemporary management research and recommendations for future studies. *Organizational Research Methods*, *7*, 151-167.

Christy, R., & Wood, M. (1999). Researching possibilities in marketing. *Qualitative Market Research*, *2*, 189-196.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003.

Cortina, J. M., & Folger, R. G. (1998). When is it acceptable to accept a null hypothesis: No way Jose? *Organizational Research Methods*, *1*, 334-350.

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but do not guarantee, better inference than statistical significance testing. *Frontiers in Psychology*, *1*, 1-9.

Diaconis, P., & Efron, B. (1983, Mayo). Computer intensive methods in statistics. *Scientific American*, *248*, 96-108.

Gardner, M., & Altman, D. G. (1986). Confidence intervals rather than P values: Estimation rather than hypothesis testing. *British Medical Journal*, *292*, 746-750.

Gephart, R. P. J. (2004). Editor's note: Qualitative research and the academy of management journal. *Academy of Management Journal*, *47*, 454-462.

Glebbeek, A. C., & Bax, E. H. (2004). Is high employee turnover really harmful? An empirical test using company records. *Academy of Management Journal*, *47*, 277-286.

International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. (1998). ICH harmonized tripartite guideline: Statistical principles for clinical trials (E9). Recuperado desde http://www.ich.org

King, G. (1986). How not to lie with statistics: Avoiding common mistakes in quantitative political science. *American Journal of Political Science*, *30*, 666-687.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746-759.

Lindsay, R. M. (1995). Reconsidering the status of tests of significance: An alternative criterion of adequacy. *Accounting, Organizations and Society*, *20*, 35-53.

Mingers, J. (2006). A critique of statistical modelling in management science from a critical realist perspective: Its role within multimethodology. *Journal of the Operational Research Society*, *57*, 202-219.

Morrison, D. E., & Henkel, R. E. (1970). *The significance test controversy*. Londres: Butterworths.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241-301.

Shaw, J. D., Gupta, N., & Delery, J. E. (2005). Alternative conceptualizations of the relationship between voluntary turnover and organizational performance. *Academy of Management Journal*, *48*, 50-68.

Siebert, W. S., & Zubanov, N. (2009). Searching for the optimal level of employee turnover: A study of a large UK retail organization. *Academy of Management Journal*, *52*, 294-313.

Simon, H. A. (1996). *The sciences of the artificial*. Cambridge, MA: MIT Press.

Simon, J. L. (1992). *Resampling: The new statistics*. Arlington, VA: Resampling Stats.

Taleb, N. N. (2008). *The black swan: The impact of the highly improbable*. Londres: Penguin.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods*, *5*, 139-158.

Wood, M. (2002). Maths should not be hard: The case for making academic knowledge more palatable. *Higher Education Review*, *34*, 3-19.

Wood, M. (2005). Bootstrapped confidence intervals as an approach to statistical inference. *Organizational Research Methods*, *8*, 454-470.

Wood, M. (2012a). Bootstrapping confidence levels for hypotheses about regression models. Recuperado de http://arxiv.org/abs/0912.3880v4

Wood, M. (2012b). P values, confidence intervals, or confidence levels for hypotheses? Recuperado de http://arxiv.org/abs/0912.3878v4

Wood, M., Capon, N., & Kaye, M. (1998). User-friendly statistical concepts for process monitoring. *Journal of the Operational Research Society*, *49*, 976-985.

Wood, M., Kaye, M., & Capon, N. (1999). The use of resampling for estimating control chart limits. *Journal of the Operational Research Society*, *50*, 651-659.

Wood, M., & Christy, R. (1999). Sampling for possibilities. *Quality & Quantity*, *33*, 185-202.

Yin, R. K. (2003). *Case study research: Design and methods* (3ª ed.). Thousand Oaks, CA: SAGE.