REVISTA ELECTRÓNICA INTERUNIVERSITARIA DE FORMACIÓN DEL PROFESORADO
CONTINUACIÓN DE LA ANTIGUA REVISTA DE ESCUELAS NORMALES

http://revistas.um.es/reifop

Fecha de recepción: 13 de enero de 2024
Fecha de aceptación: 2 de marzo de 2024

Lizasoain Hernández, L. (2024). El análisis estadístico de datos en la investigación educativa. *Revista Electrónica Interuniversitaria de Formación del Profesorado, 27*(2), 217-232.

DOI: https://doi.org/10.6018/reifop.608261

El análisis estadístico de datos en la investigación educativa

Luis Lizasoain Hernández

Universidad del País Vasco

A Luis Joaristi

In memoriam

Resumen

En este artículo se analizan algunas cuestiones y conceptos básicos del análisis estadístico de datos en el contexto de la investigación educativa. En primer lugar, y como fase o etapa de dicho proceso, se incide en la importancia de una adecuada formulación de objetivos e hipótesis, del diseño de la investigación, del muestreo y de la medida o recogida sistemática de información. Después de una breve descripción de los programas informáticos de análisis estadístico libres y abiertos, se presentan algunos enfoques y modelos. El tercer bloque se centra en la decisión estadística, en los contrastes y pruebas de hipótesis con especial atención a los conceptos de significatividad y tamaño del efecto.

Palabras clave

Análisis estadístico; metodología de la investigación; pruebas de hipótesis; significatividad estadística; tamaño del efecto.

Contacto:

Sabino Muñoz Ledesma, sabinojml@gmail.com. Universidad César Vallejo. Perú.

Statistical analysis of data in educational research

Abstract

In this paper, some basic issues and concepts about statistical data analysis in the context of educational research are analyzed. First, considering the statistical analysis as a phase of the research process, the importance of questions such as the adequate formulation of objectives and hypotheses, the design of the research, sampling and measurement, is emphasized. After a brief description of free and open statistical analysis software, some approaches and models are presented. The third block focuses on the statistical decision, contrasts and hypothesis testing with special attention to the concepts of significance and effect size.

Keywords
Effect size; Hypothesis testing; Research methodology; Statistical analisis; statistical significance.

"Estadísticamente todo se explica, personalmente todo se complica"
Daniel Pennac

Introducción

La finalidad de este trabajo es presentar algunas cuestiones y conceptos básicos concernientes al uso de técnicas estadísticas en la investigación educativa. Dadas las lógicas limitaciones de este tipo de textos, quiero empezar diciendo que no se trata ni de presentar una guía para el correcto uso de este tipo de técnicas, ni -mucho menos- un manual sobre las mismas. Trataré, eso sí, de contextualizar las que se vayan tratando y aportaré referencias para lecturas adicionales.

Dicho esto, conviene desde el principio dejar claros los puntos de partida o las asunciones básicas. Después de algunos quinquenios dedicados a la enseñanza de la estadística aplicada a la investigación en educación, hay algunas -pocas- ideas y principios que conviene señalar y en las que conviene insistir, en la medida en que pueden comprometer la calidad de la investigación.

Lo primero es que tan negativo es un enfoque de mero recetario en el que los programas informáticos al uso (luego hablaremos de ellos) actúan a modo de caja negra, como centrarse exclusivamente en la fundamentación matemática de estas técnicas. Box (1976), al tratar de la relación entre teoría y práctica en la ciencia, y empleando el símil gastronómico, alertaba sobre lo que denominó *cookbookery* y *mathematistry;* y al hablar de la formación de los estadísticos, afirmaba: "*I think there is now a wide readiness to agree that what we want are neither mere theorem provers nor mere users of a cook- book*." (Pág. 798).

El perfil deseable no es ni el de un demostrador de teoremas, ni, mucho menos, el de un usuario de un libro de recetas. Se trata de lograr un equilibrio adecuado entre teoría y práctica. No hay nada más práctico que una buena teoría, por lo que el soporte de la teoría sustantiva es imprescindible. Sin este componente, ninguna investigación contribuirá a la generación de nuevo conocimiento por mucha, pretendidamente sofisticada o muy compleja estadística que emplee.

Si el problema de investigación o si las preguntas de investigación son irrelevantes o no están convenientemente fundamentadas, es difícil que el proyecto llegue a buen puerto. Dicho esto, abordemos el asunto desde el contexto general de las etapas o fases de la investigación educativa de tipo empírico cuantitativo.

El análisis estadístico en el contexto general del proceso de investigación

En primer lugar, son, por tanto, las preguntas de investigación, los objetivos e hipótesis formuladas los elementos que han de guiar la aplicación de las técnicas estadísticas. Esto es claramente una afirmación de Perogrullo, pero estimo conveniente dejar claramente establecido este carácter meramente instrumental del aparataje estadístico.

Segundo, la materia prima son los datos organizados en una matriz. Y, como es sabido, la misma está compuesta de casos y variables. Ambos componentes han de cumplir con los criterios de rigor y calidad suficientes para que el análisis estadístico pueda ser realizado. En lo que a los casos se refiere, esto conlleva la necesidad de un adecuado diseño muestral donde ha de primar la representatividad (y el número suficiente de efectivos) con respecto a la población de referencia.

Una vez seleccionados los casos, llega el siguiente momento crucial: el proceso de medida o, si se quiere, de recogida de información. En primer lugar, el mismo se ha de llevar a cabo con todos los fenómenos objeto de interés, y de nuevo aquí la teoría juega un papel imprescindible. Una vez que se verifica que están contempladas todas las variables necesarias (y ninguna de las prescindibles, no olvidemos el principio de parsimonia), la información, los datos que se recaben han de ser válidos y fiables.

A partir de aquí, y con estos requisitos previos, comienza el análisis estadístico.

Una vez recopilada la información necesaria en el conjunto de casos de interés, el resultado es, como antes se apuntó, una matriz de datos compuesta de casos y variables. Por variable se considera cualquier fenómeno medible u observable que puede asumir diferentes estados que se denominan valores. Por caso se entiende la unidad de análisis que se haya definido.

Al respecto, conviene adelantar que los datos en educación (y en otros muchos campos) muestran una estructura jerárquica, anidada (Gaviria y Castro, 2005). Por ejemplo, los estudiantes se agrupan en colegios. En muchas investigaciones es conveniente y deseable abordar el diseño de la misma desde una perspectiva multinivel de forma que se puedan analizar *simultáneamente* variables de ambos niveles. En tal caso, se suele hablar de enfoque multinivel y las matrices de datos han de tener en cuenta dicha perspectiva. En ocasiones, en un único archivo de datos habrá variables de ambos niveles; en otras, habrá archivos separados para los casos de cada nivel.

Pues bien, el supuesto básico es que esta matriz contiene información relevante para dar respuesta a nuestras preguntas de investigación. Pero la matriz es información *en bruto*, de alguna manera, ha de ser depurada, analizada, para extraer la información relevante. Y esa es la tarea fundamental del análisis estadístico de datos, tarea que se lleva a cabo mediante la búsqueda de relaciones, diferencias, patrones, regularidades, etc.

La cita de Daniel Pennac con la que se inicia este texto, apunta y da pie a abordar algunas cuestiones importantes. En primer lugar, la inevitable contraposición entre la estadística (que explica) y lo *personal* (que complica), entre el enfoque cuantitativo y el cualitativo.

Se trata de un debate de largo recorrido y dilatada tradición en educación y en las ciencias sociales. Debate que, a decir de algunos, ha aportado más calor que luz. La primera cita que he localizado de tan afortunada expresión, es ¡nada menos! que de 1934 (Bain, 1934). Se trata de la respuesta de dicho autor a House sobre la medición en Sociología publicada en el *American Journal of Sociology*.

Desde mi personal punto de vista, y sobre la base de mi experiencia como investigador en educación, creo que nuestro objeto de estudio e investigación -la educación en todas sus facetas y acepciones- es un fenómeno tan complejo que ninguna mirada sobra. No sé si, como afirma Berliner (2002), la investigación educativa es la ciencia más difícil de todas. Lo que la práctica investigadora me ha demostrado, es que no nos podemos permitir el lujo de prescindir de ningún enfoque o perspectiva porque -con los debidos requisitos de rigor y sistematicidad- todos tienen potencial de aportar luz y de generar conocimiento.

Afortunadamente, parece que la temperatura va disminuyendo, y al respecto, las aportaciones de los métodos o enfoques mixtos (por ejemplo, Cresswell, 2015; y Creswell y Plano, 2018) son de especial relevancia. En un ámbito de investigación tan específico de nuestro quehacer como es la evaluación, la OCDE afirma que "*Good evaluations are almost invariably mixed methods evaluations*" (OECD, 2006).

Dicho esto, y a lo que ahora concierne, ha de quedar claro que no hay que confundir el enfoque cualitativo, con el análisis estadístico de variables cualitativas. La métrica de las variables, su nivel de medida, es una las cuestiones más importantes y que más incidencia tienen a la hora de optar por emplear una u otra técnica estadística, uno u otro procedimiento analítico.

Se opte por un planteamiento dicotómico que distinga entre variables numéricas frente a categoriales (cuantitativas vs. cualitativas), o siguiendo a Stevens (1946) se introduzca, al menos, el nivel de medida ordinal entre las nominales y las cuantitativas (sean de intervalo o de cociente), es imprescindible tener claras las propiedades métricas de las variables de nuestras investigaciones para emplear los procedimientos adecuados.

Es cierto que, en muchos casos, las técnicas de -en su momento- mayor uso y tradición están pensadas para variables cuantitativas (por ejemplo, muchas de las técnicas del modelo lineal general). Pero no lo es menos que, en la actualidad, se dispone de una amplia panoplia de técnicas adecuadas para los diferentes niveles de medida. Por poner dos ejemplos sencillos: un programa estadístico como *jamovi*, en su menú de pruebas T (*T-Tests*), incorpora la t de Student y *también y en el mismo menú*, la U de Mann Whitney o la prueba de rangos de Wilcoxon.

Esta distinción entre las pruebas paramétricas clásicas (como la t de Student) y las no paramétricas, es antigua y no hay que darle muchas más vueltas. Pero veamos un caso de técnicas multivariadas, en concreto, de técnicas de reducción de la dimensionalidad como, por ejemplo, el Análisis de Componentes Principales tan empleado en el análisis factorial exploratorio.

Las versiones habituales requieren que las variables a incorporar en el modelo sean cuantitativas ya que el algoritmo de extracción de factores emplea la matriz de covarianzas o de coeficientes de correlación de Pearson. Pero en programas como JASP es posible seleccionar en su menú "Factor" que, en vez de la misma, se use la matriz de correlaciones policóricas (o tetracóricas) para así, poder emplear, por ejemplo, el ACP con variables ordinales, típicamente ítems de una prueba en escala Likert (Likert, 1932). Si las variables fuesen estrictamente categoriales, puede usarse el análisis de correspondencias simple o múltiple desarrollado por Benzécri (1973) para este tipo de variables. En la colección de

Cuadernos de Estadística de la editorial La Muralla hay un manual introductorio en español sobre este tipo de análisis factorial para variables cualitativas (Joaristi y Lizasoain, 2000).

Les ruego disculpen esta autocita. Aunque viene a colación, su función es, sobre todo, la de recordar y rendir homenaje a Luis Joaristi, compañero de docencia e investigación, amigo entrañable con quien tanto aprendí.

Dicho esto, y para finalizar este apartado, con estos elementales ejemplos quiero resaltar la importancia del criterio del nivel de medida a la hora de optar por una u otra técnica para así emplear los procedimientos adecuados a las propiedades métricas de las variables.

De hecho, la pregunta o preguntas de investigación en primer lugar, y el nivel de medida de los datos en segundo, son los dos criterios básicos a la hora de emplear un método estadístico determinado. El propio programa jamovi incorpora un módulo denominado *Statkat* que emplea estos dos criterios para ayudar a encontrar la técnica apropiada.

Nota sobre los programas informáticos de análisis de datos

Y ya que volvemos a traer a colación los programas estadísticos, antes de examinar algunas técnicas o enfoques, apuntemos alguna cuestión sobre los programas informáticos de análisis de datos.

No creo equivocarme mucho al afirmar que el paquete SPSS ha sido -¿es aún?- el referente por antonomasia en lo relativo a la informatización de las técnicas estadísticas aplicadas a las ciencias sociales. Se trata de un excelente programa, actualmente propiedad de IBM, que fue creado en 1968 por Nie, Hull y Bent en la Universidad de Chicago. Nos encontramos, por tanto, con un programa de gran recorrido, que en cuatro años cumplirá los 60. Como decía, ha sido la herramienta más empleada en nuestras facultades y por nuestros grupos de investigación.

Pues bien, creo que ha llegado la hora de poder despedirnos de SPSS. Desde el año 2000, en el entorno del *open and free software*, se dispone del programa R que es, con mucho, el más completo. En el momento de escribir estas líneas (marzo 2024), consta de 20.510 paquetes (https://cloud.r-project.org/).

Como es sabido, R, a diferencia de SPSS y otras aplicaciones informáticas, es un programa cuyo funcionamiento se basa en líneas de *comandos* o instrucciones. En el argot informático, es un programa basado en CLI (*Command Language Interface*; interfaz de lenguaje de comandos). A diferencia de los programas basados en interfaces gráficas (GUI, *Graphic User Interface*), su aprendizaje puede resultar lento y algo dificultoso. En mi opinión, este hecho ha impedido que R se haya convertido en una alternativa general al software estadístico *propietario* (hablamos de SPSS, Stata, SAS, etc.).

Pero actualmente se dispone de alternativas libres, abiertas y gratuitas tan completas como estos paquetes y de uso igual de sencillo. Se trata de programas denominados genéricamente R-GUIs. Con este acrónimo se hace referencia a un conjunto de programas cuyo *motor de computación* es R (es decir operan en R), pero incorporan una interfaz gráfica de usuario que hace que su operación sea sencilla y fácil de aprender. En este momento los tres más extendidos son los ya citados jamovi y JASP, y BlueSky. Los dos primeros disponen además de versiones en español (jamovi dispone también de versión en catalán y JASP de versión en gallego).

No es éste el momento ni el lugar para glosar las características y principios del *software* libre y abierto. Aunque solo sea por razones económicas a nivel institucional, y de disponibilidad universal (no hay licencias que caduquen), mi opinión es, como decía, que es el momento de optar por emplear este tipo de plataformas. Como dice el profesor Robert Muenchen, dado que ninguna es claramente superior a las otras, lo mejor es instalar las tres porque a fin de cuentas solo le va a costar espacio en el disco duro. En cualquier caso, mi impresión personal es que, en las facultades españolas de Educación o Psicología, jamovi es el programa hegemónico.

Por último, antes decía que R no consiguió desbancar al software propietario dada la necesidad de aprender a programarlo. Pues bien, ahora, tampoco esto es ya necesario. Y no lo digo solo por los R-GUIs. Prueben a solicitar a ChatGPT que les dé el código en R para ejecutar una tarea determinada y verán… Así que lo mismo en breve plazo tampoco será necesario usar paquetes convencionales sean de pago o abiertos. Bastará un agente de IA y, por ejemplo, R o Python.

Enfoques y modelos

Hechas estas puntualizaciones, volvamos a los que nos ocupa. La afirmación de Pennac de que "estadísticamente todo se explica", nos lleva a abordar uno de los conceptos fundamentales de la estadística cual es el de varianza, y no tanto en su acepción de índice o coeficiente, sino en el de variabilidad, dispersión, diferencia, etc. La estadística es – o puede ser- una herramienta adecuada cuando la variabilidad es la regla y no la excepción.

Desde esta perspectiva, la finalidad de muchos procedimientos y técnicas estadísticas es la *explicación de la varianza*. Dicho en otros términos, ¿qué proporción de la varianza de una variable se explica por la de otra u otras? No se trata, que quede claro, de una explicación causal fuerte en el sentido estricto del término (luego hablaremos de esto al tratar del diseño experimental).

Así, por ejemplo, el principio básico del modelo lineal general se basa en la descomposición aditiva de la varianza. En un modelo sencillo de regresión lineal simple, la varianza total se descompone en dos componentes: la varianza de las puntuaciones estimadas (varianza explicada por la variable predictora o estimadora) y la varianza de los residuos (varianza del error, varianza no explicada). De esta manera, al dividir cada elemento de la igualdad por la varianza total, el primer sumando es la razón entre la varianza de las puntuaciones estimadas y la varianza total, y es, por tanto, la proporción de varianza explicada, índice conocido como el coeficiente de determinación R^2.

Si la variabilidad es el concepto fundamental, las variables son, lógicamente, los elementos clave. En función del número de variables que se analicen simultáneamente, se habla de estadística univariada, bivariada o multivariada.

Si en vez de fijarnos en el número de variables involucradas, el foco se pone en el papel que las mismas juegan en el diseño de la investigación, otra distinción importante que emerge es la que diferencia entre variable dependiente y variable(s) independiente(s). La primera es la variable objetivo, la de resultado, aquella cuya varianza se pretende explicar empleando una o más variables independientes (explicativas, predictoras, covariables). En un modelo de regresión lineal simple como el que anteriormente se señalaba, una es la variable dependiente (en educación, podríamos decir que el rendimiento académico es, por poner un ejemplo, una de las habituales) y otra la independiente o explicativa (el nivel socioeconómico

y cultural familiar, por ejemplo). En este caso, nos encontraríamos con un análisis bivariado bajo un modelo de dependencia.

Pero tratar de abordar (de explicar estadísticamente), un fenómeno tan complejo, tan multifactorial como el rendimiento académico con una sola variable explicativa mediante regresión lineal simple, es, si se me permite la expresión, una aproximación demasiado *simplona*, por muy relevante que -como nos mostró Coleman (1966)- sea el nivel socioeconómico y cultural.

Las cosas se ponen interesantes cuando empleamos modelos y técnicas estadísticas que, en la medida de lo posible, reflejan la complejidad del mundo real. En nuestro caso, la realidad educativa es compleja por multivariada, y, también multinivel.

Si en vez de emplear una única covariable, se emplea un conjunto de ellas, se hablaría de regresión múltiple (Etxeberria, 2007). Si, además, se analizan *simultáneamente* las fuentes de variación a diversos niveles (por ejemplo, estudiantes y centros), sería una regresión múltiple multinivel (Gaviria y Castro, 2005).

Ya que hemos introducido la crucial diferencia entre variables dependientes (VD) e independientes (VI), es el momento de diferenciar entre modelos de dependencia y modelos de interdependencia.

En los primeros, dado un conjunto de k variables involucradas, se realiza una partición de forma que -habitualmente- una se considera como dependiente y las demás como independientes. Es el caso de la regresión múltiple que estamos comentando. Pudiera plantearse también que hubiese más de una variable dependiente, por ejemplo, en la correlación canónica hay un conjunto de variables independientes y otro de dependientes.

Pero si en el diseño de la investigación no se hace esta distinción y todas las variables juegan el mismo papel, entonces se habla de modelos de interdependencia. Tal es el caso, por ejemplo, de -como también antes se señaló- las técnicas factoriales. Su objetivo es analizar (y en su caso reducir) la dimensionalidad subyacente. Aquí, todas las variables (por ejemplo, los ítems de una prueba) juegan el mismo papel sin que quepa hablar de variables dependientes o independientes. Las técnicas de clasificación automática (*cluster*, análisis de conglomerados), se consideran también modelos de interdependencia en la medida en que el objetivo habitual aquí es clasificar los casos en subconjuntos que sean lo más distintos entre sí (heterogeneidad entre), y lo más similares *dentro de cada uno* (homogeneidad intra). La clasificación, la asignación de casos a subconjuntos, se lleva a cabo mediante diversos algoritmos tomando en cuenta la información proporcionada por un conjunto de variables, en un modelo en el que todas juegan el mismo papel (variables clasificatorias), sin que se haga la distinción entre dependientes e independientes.

Otra distinción importante es la que diferencia entre el enfoque descriptivo y el inferencial. En el primer caso, el objetivo de la investigación -y, por ende, de los análisis estadísticos- se centra en el conjunto de casos en el que se han medido las variables, pero sin tratar de generalizar o de ir más allá. En el enfoque inferencial, por el contrario, de lo que se trata es de generalizar, de inferir la información recabada en un subconjunto (muestra) al conjunto al que ésta pertenece (población).

En este último caso, es imprescindible que la muestra sea representativa de la población; y ello suele conllevar técnicas y diseños de muestreo más o menos complejos en las que la selección de los casos que componen la muestra se lleva a cabo mediante procedimientos aleatorios.

Del cruce de todos estos enfoques o criterios de clasificación surge el mapa de las diferentes técnicas estadísticas. Pero, como decía al inicio, no es mi objetivo realizar aquí una cartografía

completa y exhaustiva; a continuación, vamos a detenernos en un primer grupo de técnicas estadísticas que está conformado por lo que se suele denominar pruebas estadísticas de decisión, pruebas de contraste de hipótesis, etc.

Decisión estadística. Contraste y pruebas de hipótesis

Antes ya hemos puesto como ejemplo la t de Student o la U de Mann Whitney. Tal y como acabamos de apuntar, dentro de las mismas se suele distinguir entre pruebas paramétricas y no paramétricas. La distinción se basa, por una parte, en la métrica de las variables implicadas, y, por otra en el cumplimiento o no de una serie de requisitos (por ejemplo, homogeneidad de las varianzas, normalidad de las distribuciones, etc.) En ambos casos, el modelo subyacente es de dependencia ya que se diferencia entre VD y VI.

Si, por ejemplo, hay una VD cuantitativa y una VI cualitativa que solo tiene (o solo se consideran) dos categorías y el estadístico de interés es la media aritmética, la técnica apropiada es la t de Student (para grupos independientes o correlacionados, dependiendo del diseño). Como antes vimos, caso de que el nivel de medida de la VD sea el ordinal (o que no se satisfagan los supuestos), la alternativa no paramétrica es la U de Mann Whitney. Si la VI genera más de dos grupos, se emplea el análisis de varianza de un factor (o su equivalente no paramétrico, la prueba de Kruskal Wallis). Si en vez de un enfoque meramente bivariado, son dos las VI (ambas categoriales), entonces se emplea el análisis de varianza de dos factores, modelo que permite, además de estimar las diferencias entre grupos, analizar posibles efectos de interacción (Fisher, 1935).

Sin entrar en la casuística de si los grupos son independientes o correlacionados, si son dos o más grupos, si la técnica adecuada ha de ser paramétrica o no; todas estas técnicas tienen en común su estrecha vinculación con un tipo de diseño de investigación específico: con los diseños experimentales o cuasi-experimentales.

En el caso de los primeros, veamos un diseño clásico muy sencillo: grupo de control vs. grupo experimental. Por ejemplo, en el contexto de la evaluación del impacto de un programa, al grupo experimental se le aplica un tratamiento, mientras que en el grupo de control no se lleva a cabo ninguna intervención específica. Se puede tomar una medida inicial en cada grupo y, al cabo del tiempo, se vuelve a medir y se evalúa el posible impacto. Por ejemplo, se hipotetiza que la media aritmética en comprensión lectora del grupo de tratamiento es superior a la del de control (H_1: media de la población 1 mayor que la media de la población 2. Contraste unilateral).

Como es sabido, los diseños experimentales son el marco lógico que permite (si se cumplen ciertos supuestos) llegar a conclusiones causales fuertes (no meras explicaciones estadísticas correlacionales). Pero se tienen que cumplir esas condiciones: simplificando mucho, control de la variable independiente, control de variables extrañas y, sobre todo, doble aleatorización (los casos se seleccionan al azar de una población y se asignan aleatoriamente a los grupos). Si se controlan todas las demás variables intervinientes, es decir, si los dos grupos *solo* se diferencian en las categorías de la VI y si los sujetos han sido seleccionados y asignados al azar, si como resultado de la prueba estadística, la media del grupo de tratamiento resulta significativamente superior a la del de control, entonces -y solo entonces- se puede tener una razonable evidencia de que los cambios observados en la VD (la mejora en la comprensión lectora) se deben al tratamiento.

Pero seguro que muchos de ustedes al leer estas líneas habrán pensado que en un entorno escolar real es muy difícil, por no decir imposible, llevar a cabo estudios experimentales en sentido estricto.

Es cierto, que este tipo de estudios experimentales *sensu stricto* son, en muchas ocasiones, de difícil realización en contextos educativos y que, por tanto, son de uso limitado en la investigación educativa. Pero lo que también es incontrovertible es que el diseño experimental permite establecer nexos causales fuertes. A las personas interesadas en este asunto recomiendo la lectura de los artículos publicados en el número monográfico que la revista *Educational Researcher* dedicó a esta cuestión (Vol. 32, Nº 1, 2003). O, más recientemente, el artículo de Taber sobre la investigación experimental y la innovación en la enseñanza (Taber, 2019).

Como sabemos, ni la investigación empírica es el único tipo de indagación científica, ni dentro de la misma, el enfoque cuantitativo es el único posible. A su vez, ni todas las investigaciones cuantitativas son experimentales, ni el diseño experimental es el único diseño de investigación (Campbell y Stanley, 2005; aunque la primera edición de su clásica obra es de 1963).

Por citar ejemplos de otros tipos de diseños, están los cuasi-experimentales (Gopalan, Rosinger y Ahn; 2020), los *ex post facto* (no experimentales), o, los tan frecuentes en nuestro campo, estudios observacionales o correlacionales.

Las pruebas clásicas de contraste de hipótesis se ajustan y corresponden con los diseños experimentales y cuasi-experimentales de ahí que en ocasiones se denominen como técnicas *fisherianas.* Para encarar las dificultades de no poder realizar experimentos *duros*, actualmente se dispone de técnicas estadísticas como el emparejamiento por puntuaciones de propensión o la regresión discontinua (o en discontinuidad) que, aplicadas a los datos de estudios no experimentales, permiten formular también conclusiones causales, que, si bien no son tan *fuertes* como las provenientes de experimentos, sí pueden llegar a tener un grado aceptable de robustez.

Además, y como antes se afirmaba, otra estrategia analítica es la modelización estadística. Se formula un modelo y luego se comprueba hasta qué punto el mismo representa (se ajusta) bien a la realidad. En la investigación educativa, y desde una perspectiva eminentemente correlacional, tienen gran importancia los modelos de regresión. Anteriormente ya apuntamos la obra de Etxeberria (2007) que constituye una excelente primera aproximación a la regresión múltiple. Y también señalamos antes la de Gaviria y Castro (2005) en la que se presentan los modelos jerárquicos lineales que se emplean en la regresión multinivel. Por último, en el manual editado por Touron (2023) se presentan otros modelos más específicos de regresión (logística, de Poisson, etc.)

De cualquier forma, las pruebas de decisión de contraste de hipótesis, son ampliamente usadas en la investigación educativa. Y al respecto, veamos ahora un par de cuestiones importantes relativas al modo en que en ocasiones se usan estas pruebas y sus resultados.

Significativo, ¿relevante?

En el enfoque de recetario elemental, en muchos casos, la regla de decisión parece haber sido la siguiente: "busca en el resultado del programa informático donde ponga "p" o "sign.", y si el número que aparece es menor que .05, o mejor, que .01 (así, con notación

sajona y todo); entonces las diferencias son significativas, se rechaza la hipótesis nula y ya está".

Pues bien, primera cuestión: al margen de la evidente intención simplificadora y del ánimo caricaturesco, esto solo tiene sentido en caso que haya habido muestreo aleatorio, del tipo que sea, pero aleatorio. Es decir, que la selección de los elementos de la muestra se haya realizado por azar, que el azar haya intervenido. Y esto es así porque la tan traída y llevada "p" expresa la probabilidad de que esa diferencia se haya dado por azar en las muestras, en el caso de que a nivel poblacional tales diferencias no existan, es decir, que la hipótesis nula H_0 sea verdadera. Si esa probabilidad es menor que el nivel alfa previamente establecido (0,05 o 0,01), entonces se concluye que esa probabilidad es pequeña y, por tanto, la diferencia (de medias, de varianzas, de frecuencias…) se considera significativa.

En consecuencia, si no ha habido muestreo aleatorio, si el azar no ha intervenido, todo el andamiaje lógico subyacente colapsa. Y no cabe hablar de significatividad. O, en su caso, ha de hacerse extremando las precauciones.

Pero es que, además, la significatividad solo nos informa de lo alta o baja que se considere la probabilidad de que una determinada diferencia se haya dado por azar. Nada menos, pero nada más. No dice nada de la magnitud del impacto que la variable independiente puede haber tenido sobre la dependiente. De hasta qué punto, hasta qué grado, el grupo de tratamiento ha mejorado con respecto al de control. Y esto es, al menos, igual de importante, si no más. Me estoy refiriendo al *tamaño del efecto*. Se trata de una cuestión que, además se exige en la mayoría de las revistas para los trabajos de este tipo. Véase al respecto el trabajo de López Martin y Ardura Martínez (2022) titulado precisamente "El tamaño del efecto en la publicación científica".

Si se habla de tamaño del efecto, se hace referencia a la *relevancia* del resultado. Se trata de un conjunto de índices que miden la magnitud del efecto del tratamiento o de las diferencias, en definitiva, del impacto de la variable independiente sobre la dependiente.

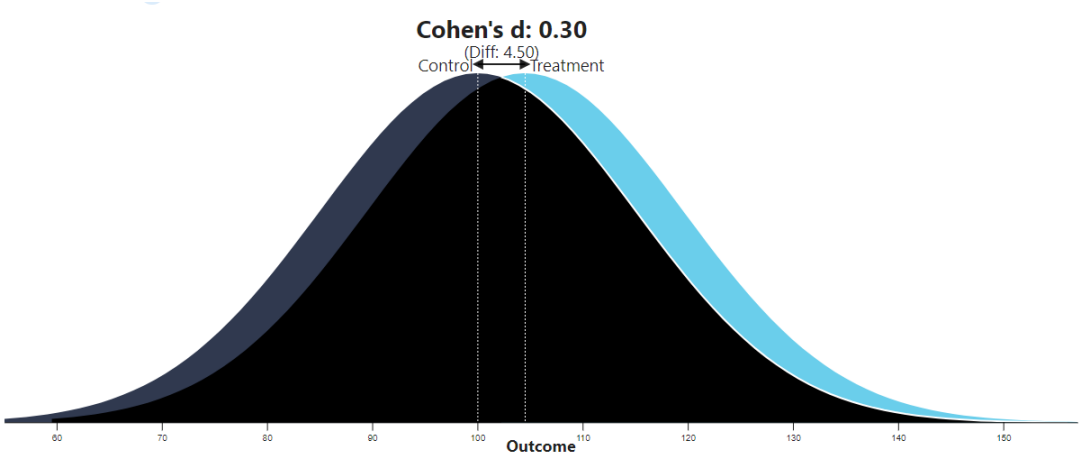
Simplificando la cuestión, podemos afirmar que suele hacerse de dos formas: o bien se tipifica la diferencia entre la media del grupo de control y la del grupo experimental (caso de la d de Cohen, la G de Hedges o la delta de Glass, por ejemplo), o se calcula la correlación entre la VI y las puntuaciones de la VD empleando, por ejemplo, el coeficiente de correlación biserial-puntual para el caso de una VD cuantitativa y una VI dicotómica.

Hasta no hace mucho, este tipo de índices no aparecían en los programas informáticos al uso por lo que había que recurrir a calculadoras *en la nube* y a sitios web especializados. Entre otros, destaca la página de Becker de la Universidad de Colorado (https://www.uccs.edu/lbecker/) a donde remito a quienes estén interesados porque es toda una referencia en el campo. Actualmente, programas como jamovi, JASP o BlueSky proporcionan estos índices en los resultados de las técnicas estadísticas asociadas.

Siguiendo con el ejemplo elemental de la diferencia de dos medias aritméticas, en este caso uno de los índices más usado es la d de Cohen. Sus valores se suelen interpretar como la proporción de casos del grupo de control que quedan ***por debajo de la media*** del grupo experimental. O también, como la proporción de solapamiento (o no solapamiento) entre los casos del grupo experimental y los del control.

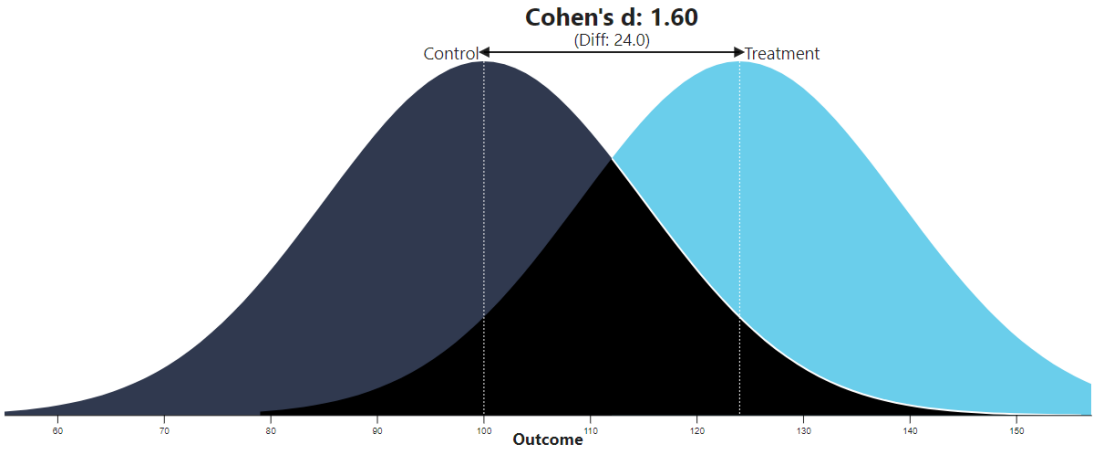
Por ejemplo, un valor de este índice de 0,3 (ver figura 1) denota un tamaño del efecto muy bajo en la medida en que, al estar muy cerca las dos distribuciones (la del grupo de control y la del experimental), la proporción de solapamiento es del 88,1% y solo un 61,8% de los casos del grupo experimental tienen una puntuación superior a la media del grupo de control (U_3 de Cohen).

Figura 1



Por el contrario, un valor de, por ejemplo, 1,6 (ver figura 2) indica un impacto mayor con una proporción de solapamiento de solo el 42,4%, de forma que el 94,5% de los casos del grupo experimental tienen una puntuación superior a la media del grupo de control.

Figura 2



Estas figuras y ejemplos están tomados de la excelente página web de Magnusson (2023): https://rpsychologist.com/cohend/

El propio Cohen (1988) propuso el siguiente criterio de interpretación de la "d" (tabla 1):

Tabla 1

d de Cohen	Interpretación del efecto
0,20 - 0,30	Débil, pequeño
0,50 - 0,80	Moderado, medio
> 0,80	Grande, fuerte

Para acabar con esta importante cuestión, y de paso ver otro ejemplo, señalemos que además, la significatividad (o no) de un estadístico, en ocasiones depende mucho del tamaño muestral, de forma que, si éste es lo suficientemente grande, una determinada diferencia resultará significativa, al margen de su mayor o menor relevancia. El caso del estadístico de chi cuadrado es paradigmático al respecto.

Veamos el ejemplo: siguiendo en el contexto de las pruebas de hipótesis, cuando ambas variables son categoriales, una prueba habitual es el análisis de independencia usando tablas de contingencia bidimensionales, y el contraste de la igualdad o diferencia entre las frecuencias observadas y las teóricas, se lleva a cabo mediante el estadístico de chi cuadrado.

Supongamos que en una investigación se quiere estudiar las posibles diferencias de género (VI) a la hora de matricularse en los diferentes grados universitarios (VD). Para simplificar el ejemplo, vamos a considerar ambas variables implicadas como dicotómicas diferenciando entre género masculino y femenino por una parte, y carreras científico-técnicas (tipo A) frente al resto (tipo B).

Supongamos que un equipo investigador realiza un muestreo aleatorio de 1.000 estudiantes y que la tabla de contingencia resulta ser la que muestra la tabla 2:

Tabla 2

Tablas de Contingencia

		Género		
Grado		F	M	Total
A	Observado	235	165	400
	Esperado	240	160	400
B	Observado	365	235	600
	Esperado	360	240	600
Total	Observado	600	400	1000
	Esperado	600	400	1000

Pruebas de χ^2

	Valor	gl	p
χ^2	0.434	1	0.510
N	1000		

Como se puede observar, las estudiantes femeninas muestran una cierta tendencia a matricularse en mayor proporción en los grados de tipo A (165 frente a las 160 esperables bajo el supuesto de la independencia de ambas variables); mientras que en los estudiantes masculinos ocurre lo mismo con los grados de tipo B (365 frente a 360). El valor de chi cuadrado es de 0,434 y tiene una probabilidad asociada de 0,51 por lo que en este caso se aceptaría la hipótesis nula.

Queda claro que esta pequeña diferencia, de 5 casos arriba o abajo, no resulta significativa. Pero, ¿qué ocurriría si la muestra, en vez de 1.000 estudiantes, fuese de 10.000?

Tabla 3

Tablas de Contingencia

		Género		
Grado		F	M	Total
A	Observado	2350	1650	4000
	Esperado	2400	1600	4000
B	Observado	3650	2350	6000
	Esperado	3600	2400	6000
Total	Observado	6000	4000	10000
	Esperado	6000	4000	10000

Pruebas de χ^2

	Valor	gl	p
χ^2	4.34	1	0.037
N	10000		

Ahora (tabla 3) las proporciones de las diferencias entre las frecuencias observadas y las esperadas son las mismas que antes. Pero en este caso, el valor de chi cuadrado resulta ser 10 veces mayor, de 4,34 y su probabilidad asociada de 0,037 lo que resultaría significativo al nivel alfa del 0,05 (no así al del 0,01). Vemos pues, como chi cuadrado es un estadístico sensible al tamaño muestral y, desde el punto de vista estrictamente probabilístico, es lógico que sea así, porque sobre 10.000 casos, la distribución observada es menos probable que salga por azar si es que las variables resultan ser independientes, que en el caso de una muestra de 1.000.

Se observa, por tanto, que, dependiendo del tamaño muestral, en una investigación las diferencias resultan no significativas, y en la otra sí. Pero, ¿cómo es el impacto, el efecto del género sobre la elección del grado?

Para este tipo de tablas de 2x2 en el que ambas variables son nominales, se suele usar el coeficiente Phi. Sus valores oscilan entre 0 y 1, y se interpreta como un coeficiente de correlación. Es decir, valores bajos, cercanos a cero, denotan ausencia de relación, independencia entre variables. Pues bien, en ambos casos (ver tabla 4), el valor de Phi es el mismo (0,0208). Esto es lógico, habida cuenta de que la desproporción entre frecuencias observadas y teóricas es la misma en los dos casos. Numéricamente, Phi se calcula dividiendo chi cuadrado entre el número de casos y luego calculando la raíz cuadrada de dicho cociente.

Nos encontramos con un valor muy bajo, casi igual a cero, lo que denota que la relación entre ambas variables es extremadamente débil, prácticamente nula. Es decir, no hay impacto reseñable del género sobre la elección del grado. Y esto es así, *resulte o no significativa la diferencia entre las frecuencias observadas y las teóricas.*

Tabla 4

Nominal

	Valor
Coeficiente Phi	0.0208
V de Cramer	0.0208

En definitiva, la significatividad nos informa sobre la probabilidad de que una determinada diferencia muestral se haya dado por azar, bajo el supuesto de que en la población tal diferencia no se da (hipótesis nula verdadera). En cambio, el tamaño del efecto nos informa de la relevancia de esa diferencia, hasta qué punto la misma se puede considerar grande, mediana o pequeña. Por tanto, y tal y como afirma Blanco (2019) al hablar del consenso sobre nuevas prácticas en la investigación educativa: "… destaca la necesidad de informar siempre del tamaño del efecto…" (Cfr. Pág. 216).

Conclusiones

Es hora de ir acabando. A lo largo de estas líneas he tratado de abordar lo que considero que son algunos puntos importantes que han de tenerse presentes a la hora de emplear técnicas estadísticas en la investigación educativa. En su mayoría, son cuestiones de sobra conocidas y no responden a nuevos planteamientos (habrán observado que, dada su fecha de publicación, algunas de las referencias citadas, se pueden calificar, casi, de venerables). Pero son cuestiones que, en mi opinión, conviene resaltar y deben ser siempre muy tenidas en cuenta a la hora de usar esta metodología en nuestra actividad investigadora.

A modo de resumen, destaquemos cinco aspectos: la importancia de una adecuada fundamentación teórica; el papel crucial del diseño; todo lo relativo al muestreo y a la selección de casos; la necesidad de asegurar la calidad de las medidas y observaciones; y, por último, la conveniencia -casi necesidad- de informar de la relevancia de los resultados, del tamaño de los efectos. En definitiva, la estadística como herramienta instrumental cuya aplicación ha de estar guiada por los objetivos e hipótesis planteadas, teniendo siempre presente que el objetivo es dar respuesta a las preguntas de investigación.

Por razones de espacio, y una vez llegados a este punto, no se puede ni se debe ser más prolijo. Por supuesto que son muchas las técnicas y procedimientos que *se han quedado en el tintero*, a las que hemos citado de pasada o ni siquiera eso: el meta-análisis como síntesis cuantitativa de resultados, los modelos de ecuaciones estructurales, las técnicas de *big data* y minería de datos como, por ejemplo, los árboles de decisión; o, un enfoque tan importante en educación como son los análisis longitudinales y el estudio del cambio. Al respecto, no olvidemos que los procesos de enseñanza-aprendizaje se desarrollan a lo largo de periodos de tiempo más o menos dilatados.

Por último, en el muy improbable caso (alfa muy menor que 0,01) de que alguien se haya quedado con ganas de seguir leyendo sobre estas cuestiones, permítanme para acabar unas breves recomendaciones de manuales introductorios.

El profesor Andy Field tiene dos excelentes manuales sobre estadística, uno centrado en R (Field, 2012) y otro usando SPSS (Field, 2018). Si usan o van a usar el programa jamovi, de la

propia web del programa en el apartado de recursos, es posible descargarse gratuitamente un manual orientado a estudiantes de Psicología y disciplinas afines del que son autores Navarro y Foxcroft (2022). Por último, si lo que usted quiere es estudiar -y aprender- estadística sin programas informáticos, e incluso sin calculadoras, usando solo lápiz y papel, el profesor Marco acaba de publicar un manual titulado "*A pen and paper introduction to statistics*" (Marco, 2024) en el que, desde la perspectiva del modelo lineal y empleando solo lápiz y papel hace un excelente recorrido por muchas de las técnicas y conceptos aquí brevemente apuntados.

Referencias

Bain, R. (1934). Measurement in Sociology (A response to House). *American Journal of Sociology* 40 (1934), 481-488.

Benzécri, J.-P. (1973). *L'Analyse des Données. Volume II. L'Analyse des Correspondances*. París: Dunod.

Berliner, D. C. (2002). Educational Research: The Hardest Science of All. *Educational Researcher*, *31*(8), 18-20.

Blanco Blanco, A. (2018). Estado de las prácticas científicas e investigación educativa. Posibles retos para la próxima década. *Revista de Educación*, 381, 207-232. https://doi.org/10.4438/1988-592X-RE-2017-381-386

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association,* 71 (356): 791–799. doi:10.1080/01621459.1976.10480949

Campbell, D. F. y Stanley, J. C. (2005). *Diseños experimentales y cuasiexperimentales en la investigación social* (1a. edición 9a. reimpresión). Amorrortu Editores.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Earlbaum Ass.

Coleman, J. S. et al. (1966). *Equality of Educational Opportunity*. UD Dept. of Health, Education and Welfare. National Center for Educational Statistics. Us Gov. Printing Office. Washington D.C.

Creswell, J. W. (2015). *A concise introduction to mixed methods research*. Sage.

Creswell, J. W. & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research*. Sage.

Etxeberria, J., Joaristi, L., Lizasoain, L. (1990): *Programación y Análisis estadísticos básicos con SPSS/PC+.* Paraninfo. Madrid.

Etxeberria Murgiondo, J. (2007). *Regresión múltiple* (2a. Ed.). La Muralla.

Field, A. (2012). *Discovering statistics using R*. Sage Publications.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage. CRC Press.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.

Gaviria, J.L. y Castro, M. (2005). *Modelos jerárquicos lineales*. Madrid. La Muralla.

Gopalan, M., Rosinger, K., & Ahn, J. B. (2020). Use of Quasi-Experimental Research Designs in

Education Research: Growth, Promise, and Challenges. *Review of Research in Education*, *44*(1), 218-243. https://doi.org/10.3102/0091732X20903302

Joaristi, L. y Lizasoain, L. (2000). *Análisis de Correspondencias*. La Muralla. Madrid.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22,* 140-155.

López-Martín, E., & Ardura-Martínez, D. (2022). El tamaño del efecto en la publicación científica. *Educación XX1,* 26(1), 9-17. https://doi.org/10.5944/educxx1.36276

Magnusson, K. (2023). A Causal Inference Perspective on Therapist Effects. *PsyArXiv*

Marco, A. (2024). *A pen and paper introduction to statistics.* CRC.

Navarro, D. J. y Foxcroft, D. R. (2022). *Learning statistics with jamovi: a tutorial for psychology students and other beginners. (Version 0.75)*. DOI: 10.24384/hgc3-7p15

OECD (2006) International Workshop on Impact Evaluation for Development. OECD. Disponible en:https://www.oecd.org/dac/evaluation/dcdndep/internationalworkshoponimpactevaluationfordevelopment15november2006-hostedbytheworldbankandthedacevaluationnetwork.htm

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, *103*(2684), 677–680. http://www.jstor.org/stable/1671815

Taber, K. S. (2019). Experimental research into teaching innovations: responding to methodological and ethical challenges, *Studies in Science Education,* 55(1), 69-119, DOI: 10.1080/03057267.2019.1658058

Touron, J. (ed.), Lizasoain, L., Navarro, E. y López-González, E. (2023). *Análisis de Datos y Medida en Educación.* Logroño: UNIR Editorial.