Capítulo 62/94

Tratado de psiquiatría clínica
Massachusetts General Hospital
2017 - 2018

ESTADÍSTICA EN INVESTIGACIÓN PSIQUIÁTRICA

Trabajamos para su tranquilidad...

Estadística en investigación psiquiátrica

Lee Baer, PhD

PUNTOS CLAVE

- Se utilizan normalmente tres categorías estadísticaas en la investigación psiquiátrica: la estadística *psicométrica*, que evalúa la fiabilidad y validez de las *entrevistas diagnósticas o escalas de evaluación;* la estadística *descriptiva*, utilizada para describir variables clínicas y demográficas en un *grupo de sujetos*, y la estadística *inferencial*, utilizada para hacer afirmaciones probabilísticas sobre los *efectos de los tratamientos u otras variables en grupos de sujetos*.
- Cuantas más pruebas estadísticas se realizan en un estudio, mayores serán las posibilidades de encontrar que una o más sean significativas, cuando, de hecho, no hay un efecto verdadero en la población de la que se extrajo la muestra (es decir, un resultado falso positivo).
- Los investigadores eligen (dependiendo de la escala de medición de determinadas variables) el método estadístico más adecuado para responder a la pregunta concreta de su investigación. Debe elegirse el método disponible más simple para responder adecuadamente a dicha pregunta.
- El análisis del poder estadístico determinará cuántos sujetos serán necesarios para minimizar los resultados falsos negativos en la estadística inferencial.

Tres categorías estadísticas en investigación psiquiátrica

La palabra «estadística» se deriva de un término utilizado para las «cifras que caracterizan el *estado*», es decir, las estadísticas originales fueron las cifras utilizadas por los gobernantes de los estados para comprender mejor a su población. Así, las primeras estadísticas eran simplemente un recuento de cosas (como la población de las ciudades o la cantidad de grano producida por una ciudad determinada). Hoy en día llamamos a este tipo de cuentas simples o promedios «estadística descriptiva», y esta se usa en casi todos los estudios de investigación para describir las características demográficas y clínicas de los participantes en un determinado estudio.

La investigación psiquiátrica actual también utiliza otras dos categorías estadísticas: la estadística psicométrica y la inferencial. La mayoría de los estudios psiquiátricos incluirán las tres categorías estadísticas.

En la investigación psiquiátrica, las variables demográficas (como el sexo y la altura) pueden medirse objetivamente. Sin embargo, la mayoría de nuestros estudios también requieren la medición de variables que no son tan objetivas (p. ej., diagnósticos clínicos y escalas de evaluación psicopatológica). En este caso, por lo general, no podemos medir directamente las características que realmente nos interesan, por lo que en su lugar nos basamos en la puntuación obtenida de un sujeto, bien a través de escalas de autocumplimentación, o recogidas por el investigador. La psicometría se ocupa de lo reproducible que es la puntuación de un sujeto (es decir, lo fiable que es) y de cuánto mide la característica que realmente nos interesa (es decir, lo válida que es).

Los investigadores en psiquiatría estudian muestras relativamente pequeñas de sujetos, generalmente con la intención de generalizar sus resultados a una población mayor de la cual se extrajo su muestra. Este es el terreno de la estadística inferencial, que se basa en la teoría de la probabilidad. Los investigadores están presentando medidas estadísticas inferenciales cuando se ven indicadores de valores *P* y asteriscos que indican la significación estadística en el texto y en las tablas de los apartados sobre los resultados.

Los tres tipos de estadística (descriptiva, psicométrica e inferencial) están presentes en la mayoría de los trabajos publicados en investigación psiquiátrica, y se presentan en un orden concreto por las siguientes razones. En primer lugar, sin medidas fiables y válidas, ninguno de los otros tipos de estadística tendrá sentido. Por ejemplo, si nos basamos únicamente en los juicios de los médicos para valorar la mejoría de un paciente, cualquier otra estadística carecerá de sentido, ya que los médicos del estudio rara vez coincidirán en si ha mejorado un paciente en particular. Del mismo modo, una medida puede ser obtenida de manera muy fiable, como el número de teléfono móvil de un paciente, pero esta medida no es fiable para cualquiera de los propósitos del estudio. En segundo lugar, se necesita la estadística descriptiva para sintetizar las puntuaciones de muchos sujetos individuales en medidas estadísticas resumidas (como recuentos, proporciones, promedios [o medias] y desviaciones estándar) que luego pueden compararse entre grupos. La estadística inferencial sería imposible de realizar sin tener primero dichas medidas estadísticas. En tercer lugar, sin la estadística inferencial y sus cálculos de los valores de probabilidad, el investigador no puede generalizar resultados positivos fuera del grupo determinado que se está estudiando (y esto es, después de todo, el objetivo normal de un estudio de investigación).

La tabla 62-1 muestra las características de cada categoría, así como el orden en que debe presentarse cada categoría, ya que cada una de ellas de forma sucesiva descansa sobre la base de la precedente.

Tabla 62-1

Tres categorías estadísticas utilizadas en investigación psiquiátrica (por orden de relevancia)

Categoría estadística	Objetivo	Ejemplos
Estadística psicométrica	Medidas de fiabilidad y validez de las escalas de evaluación y otras medidas Una vez demostrado que dichas medidas tienen una fiabilidad y una validez adecuadas, pueden utilizarse como estadísticos descriptivos	Coeficiente de fiabilidad de prueba y repetición Coeficiente de correlación intraclase Coeficiente κ Sensibilidad Especificidad
Estadística descriptiva	Medidas estadísticas utilizadas para resumir las puntuaciones de muchos sujetos a un único recuento o media para describir el grupo en su conjunto. Después de obtener los estadísticos descriptivos de una o varias muestras, estos pueden utilizarse posteriormente para calcular las medidas estadísticas inferenciales en un intento por generalizar estos resultados a una población mayor de donde se obtuvieron las muestras	Media Mediana Desviación estándar Varianza Estimaciones del tamaño del efecto Proporciones Porcentajes Diferencias significativas Cociente de posibilidades
Estadística inferencial	Medidas estadísticas utilizadas para calcular estimaciones de probabilidad utilizadas para generalizar los estadísticos descriptivos a una población mayor de donde se obtuvieron las muestras	Estadístico t Estadístico F Estadístico χ^2 Intervalos de confianza

Ejemplos concretos de las tres categorías estadísticas en un artículo de investigación

Para proporcionar un ejemplo concreto de estos conceptos, a veces abstractos, considere un estudio ficticio basado en el diseño de investigación más simple en investigación psiquiátrica: un estudio doble ciego aleatorizado de un nuevo fármaco frente a un comprimido de placebo para el trastorno obsesivo-compulsivo (TOC).

Las figuras 62-1 a 62-3 contienen los apartados «Métodos» y «Resultados» comentados de este estudio ficticio, mostrando cómo se presentan las medidas estadísticas psicométricas en el apartado «Métodos», mientras que las descriptivas se presentan en los apartados «Métodos» y «Resultados», y las inferenciales, en el apartado «Resultados» (para las definiciones de los términos utilizados en estas figuras, consulte el apartado «Términos estadísticos y sus definiciones»).

FIGURA 62-1 Apartado de métodos ficticio comentado para ilustrar las medidas estadísticas psicométricas.

FIGURA 62-2 Apartado de resultados ficticio comentado para ilustrar las medidas estadísticas descriptivas.

FIGURA 62-3 Apartado de resultados ficticio comentado para ilustrar las medidas estadísticas inferenciales.

Tasa de error por experimento

Los investigadores deberían probar solo unas pocas hipótesis cuidadosamente seleccionadas (¡determinadas antes de recoger los datos!) si los valores *P* obtenidos son significativos. Cuantas más pruebas estadísticas realice, mayor será la posibilidad de encontrar por lo menos una significativa únicamente por casualidad (es decir, un resultado falso positivo). La tabla 62-2 muestra este fenómeno.

Tabla 62-2

Error por experimento: ¿encontró el investigador un único resultado significativo solo por casualidad?

Número de pruebas estadísticas realizadas a $P < 0,05$	Probabilidad de obtener al menos un resultado falso positivo
1	0,05
2	0,09
3	0,14

Número de pruebas estadísticas realizadas a $P < 0,05$	**Probabilidad de obtener al menos un resultado falso positivo**
4	0,18
5	0,22
6	0,26
7	0,3
8	0,33
9	0,36
10	0,41
15	0,53
20	0,64
30	0,78
40	0,87

Número de pruebas estadísticas realizadas a $P < 0,05$	Probabilidad de obtener al menos un resultado falso positivo[*]
50	0,92

[*] Tasa de error por experimento.

No debe uno quedarse impresionado por un investigador que, tras llevar a cabo ocho pruebas t, encuentra una de ellas significativa con un valor $P < 0,05$, y procede a interpretar los resultados como confirmación de su teoría. La tabla 62-2 nos muestra que, con ocho pruebas estadísticas con un valor $P < 0,05$, el investigador tenía un 33% de posibilidades de encontrar al menos un resultado significativo por casualidad.

Selección de un método estadístico adecuado

Los dos determinantes clave para la elección de un método estadístico son: 1) el objetivo de la investigación, y 2) la escala de medición de su variable de resultado (o dependiente). La tabla 62-3 muestra las características clave de las distintas escalas de medición y proporciona ejemplos de cada una de ellas.

Tabla 62-3

Escalas de medición de variables

Escalas de medición	Descripción de la escala	Ejemplos
Continuas (también conocidas como intervalo o razón)	Una escala en la que hay intervalos aproximadamente iguales entre las puntuaciones	Escala de depresión de Beck Presión arterial diastólica Edad del sujeto
Ordinales (también conocidas como rangos)	Una escala en la que las puntuaciones están organizadas en orden, pero los intervalos entre las puntuaciones pueden no ser iguales	Clasificación de las clases en la escuela Cualquier medida continua que ha sido convertida en rangos
Nominales (también conocidas como categóricas)	Las puntuaciones son simplemente los nombres de diferentes grupos, pero las puntuaciones no implican su magnitud. A menudo se utilizan para definir grupos basados en tratamientos experimentales o diagnósticos	Categoría diagnóstica Etnia Código postal de lugar de residencia
Dicotómicas (también conocidas como binarias)	Un caso especial de una variable nominal en la que solo hay dos valores posibles	Sexo (hombre o mujer) Supervivencia (sí o no) Respuesta (sí o no)

Una vez que se haya determinado la escala de medición de su variable de resultado, tendrá que decidir si su pregunta de investigación requerirá que compare dos o más grupos diferentes de sujetos o que

compare variables dentro de un mismo grupo de sujetos. Las tablas 62-4 y 62-5 le ayudarán a elegir el método estadístico adecuado una vez que haya tomado estas decisiones. (Obsérvese que estas tablas consideran solo las pruebas estadísticas univariantes; las pruebas multivariantes están fuera del alcance de este capítulo.)

Tabla 62-4

Elección de una prueba estadística adecuada para comparar dos o más grupos, basándose en su objetivo de investigación y en la escala de medición de la variable de resultados

Nuestro objetivo	Escala de medición de la variable de resultados		
	Continua	*Dicotómica*	*Estratificada*
Comparar dos grupos	Prueba t de la diferencia de medias	Tabla de contingencia de proporciones 2 × 2 probada por χ^2	Prueba U de Mann-Whitney de rangos medios
Comparar tres o más grupos	ANOVA	Tabla de contingencia de proporciones probada por χ^2	Prueba de Kruskall-Wallis de rangos medios
Comparar dos o más grupos mientras se controla una o más de las otras variables medidas en ambos grupos	ANCOVA	Prueba de Mantel-Haenszel (no aplicable a más de dos grupos)	N/A
Comparar dos o más grupos estratificados en algunas otras variables	ANOVA factorial	Prueba de Mantel-Haenszel (no aplicable a más de dos grupos)	N/A
Comparar dos o más grupos medidos en repetidas ocasiones	ANOVA o MMRM mixto (o dividido)	MMRM	N/A

ANCOVA, análisis de la covarianza; ANOVA, análisis de la varianza; MMRM, modelo de efecto mixto de medidas repetidas; N/A, no aplicable.

Tabla 62-5

Elección de una prueba estadística adecuada para un solo grupo de sujetos, basándose en su objetivo de investigación y en la escala de medición de la variable de resultados

Nuestro objetivo	Escala de evaluación de la variable de resultados		
	Continua	***Dicotómica***	***Estratificada***
Analizar la asociación de una variable continua con:	Coeficiente de correlación de Pearson *(r)*	Coeficiente de correlación biserial puntual	N/A
Analizar la asociación de una variable dicotómica con:	Coeficiente de correlación biserial puntual	Coeficiente de correlación φ	N/A
Analizar la asociación de variable estratificada con:	N/A	N/A	Correlación de rangos de Spearman
Predecir el valor de la medida de resultado a partir de una o más variables predictivas continuas o dicotómicas	Regresión lineal	Regresión logística	N/A
Comparar dos o más grupos medidos en repetidas ocasiones	ANOVA mixta (o dividida)	N/A	N/A
Comparar el cambio en una variable de resultado medida en dos ocasiones	Prueba t dependiente	Prueba de McNemar	Prueba de Wilcoxon
Comparar el cambio en una variable de resultado medida en tres o más ocasiones	ANOVA o MMRM repetidos unidireccionales	MMRM con resultados dicotómicos	Prueba de Friedman

ANCOVA, análisis de la covarianza; ANOVA, análisis de la varianza; MMRM, modelo de efecto mixto de medidas repetidas; N/A, no aplicable.

Por ejemplo, si desea llevar a cabo un estudio comparando un nuevo fármaco con dos situaciones de control, y su medida de resultado es una escala de evaluación continua, la tabla 62-4 indica que normalmente utilizaría un análisis de la varianza (ANOVA, *analysis of variance*) para analizar sus datos. Si desea evaluar la asociación de dos medidas continuas de disociación y ansiedad en una única muestra de pacientes deprimidos, la tabla 62-5 indica que normalmente seleccionaría el coeficiente de correlación de

Pearson. (Obsérvese que los métodos enumerados para las medidas de resultados estratificados son aquellos que se denominan normalmente «pruebas no paramétricas».)

La consideración final en la selección de un método estadístico es si los sujetos se evalúan en más de una ocasión, como en un estudio clínico longitudinal normal. En casos como estos, se utilizan métodos estadísticos especiales para «medidas repetidas». Tradicionalmente, un análisis de varianza de medidas repetidas ha sido el método más utilizado para analizar un estudio que compara dos o más líneas de tratamiento en un diseño longitudinal. Sin embargo, cuando, como suele ocurrir, faltan datos debido al abandono de los sujetos, el método de análisis preferido es un enfoque más complejo denominado modelo de efecto mixto de medidas repetidas (MMRM, *mixed-effect model repeated measure*).

Importancia de evaluar el poder estadístico

Un valor *P* no significativo carece de sentido si el investigador estudia muy pocos sujetos, dando como resultado un bajo poder estadístico. Las tablas 62-6 a 62-8 le ayudarán a estimar el número de sujetos necesarios para tener una posibilidad razonable (generalmente establecida en el 80%, o una potencia = 0,8) de detectar un efecto verdadero (o, dicho de otro modo, una posibilidad del 20% de un resultado falso negativo).

Tabla 62-6

Poder estadístico: ¿incluyó el estudio suficientes sujetos como para poder detectar una diferencia significativa real en la media de los dos grupos?

	Poder estadístico			
Tamaño del efecto (diferencia de medias)	***0,5***	***0,6***	***0,7***	***0,8*** [a]
0,2 DE («pequeño»)	193	246	310	393
0,5 DE («mediano»)	32	40	50	64
0,8 DE («grande»)	13	16	20	26
1,2 DE	7	8	10	12

Tomado de Cohen J. *Statistical power analysis for the behavioral sciences*, ed 2, Hillsdale, NJ, 1988, Lawrence Erlbaum Associates, tabla 2.4.1.

Prueba t para comparar las medias de dos grupos: *N* requerido en cada grupo para obtener varios niveles de poder estadístico (prueba de dos colas a $P < 0,05$).

DE, desviación estándar.

[a] Nivel convencional para un correcto poder estadístico.

Tabla 62-7

Poder estadístico: ¿incluyó el estudio suficientes sujetos como para poder detectar una diferencia significativa real en las proporciones de los dos grupos?

Tamaño del efecto (estadístico «w»)	Poder estadístico				Ejemplos de los valores de «w»
	0,5	*0,6*	*0,7*	*0,8*[-]	
0,1 («pequeño»)	384	490	617	785	45 frente a 55%
0,3 («mediano»)	43	54	69	87	35 frente a 65%
0,5 («grande»)	15	20	25	31	25 frente a 75%
0,7	8	10	13	16	15 frente a 85%

Tomado de Cohen J. *Statistical power analysis for the behavioral sciences*, ed 2, Hillsdale, NJ, 1988, Lawrence Erlbaum Associates, tabla 7.4.6.

Prueba χ^2 2 × 2 para comparar las proporciones de dos grupos: *N* requerido en cada grupo para obtener varios niveles de poder estadístico (prueba de dos colas a $P < 0,05$).

[-] Nivel convencional para un correcto poder estadístico.

Tabla 62-8

Poder estadístico: ¿incluyó el estudio suficientes sujetos como para poder detectar una correlación significativa real en un grupo?

Tamaño del efecto (*r* de Pearson)	Poder estadístico			
	0,5	*0,6*	*0,7*	*0,8*[-]
0,1 («pequeño»)	385	490	616	783
0,3 («mediano»)	42	53	67	85

Tamaño del efecto (*r* de Pearson)	Poder estadístico			
	0,5	***0,6***	***0,7***	***0,8***[a]
0,5 («grande»)	15	18	23	28
0,7	7	9	10	12

Tomado de Cohen J. *Statistical power analysis for the behavioral sciences*, ed 2, Hillsdale, NJ, 1988, Lawrence Erlbaum Associates, tabla 3.4.1.

Coeficiente de correlación entre dos variables: *N* total requerido para obtener varios niveles de poder estadístico (prueba de dos colas a $P < 0,05$).

[a] Nivel convencional para un correcto poder estadístico.

Por ejemplo, un investigador comunica que ha comparado dos grupos de 12 pacientes deprimidos y ha encontrado que un nuevo fármaco no fue significativamente mejor que placebo a un valor $P < 0,05$, mediante la prueba t. Sin embargo, este resultado negativo no es ilustrativo, ya que, como la tabla 62-6 indica, con solo 12 sujetos por grupo, este investigador tenía un poder estadístico inferior a 0,8 insuficiente para detectar un «gran» efecto; es decir, incluso si el fármaco era realmente eficaz, este estudio tenía menos de un 50:50 de posibilidades de encontrar una diferencia significativa.

Actualmente se requiere el análisis del poder estadístico como parte de prácticamente todas las solicitudes presentadas para subvención y revisión ante comités de ética institucional (CEI).

Términos estadísticos y sus definiciones

Análisis de conglomerados

Se trata de una técnica de reducción de datos utilizada para agrupar sujetos en subgrupos (o «conglomerados») basándose en sus similitudes o diferencias en relación con un conjunto de variables. Esta técnica responde a preguntas como: «¿Los sujetos pueden formar subgrupos?» y «¿qué variables proporcionan un perfil que distingue los subgrupos de sujetos?». Una regla general simple es que «el análisis de conglomerados agrupa a personas, mientras que el análisis factorial agrupa variables». Se considera un método estadístico multivariante porque se analizan simultáneamente muchas variables intercorrelacionadas.

Análisis de la covarianza

El análisis de la covarianza (ANCOVA, *analysis of covariance*) es una forma de ANOVA que prueba la significación de las diferencias entre las medias de los grupos ajustando las diferencias iniciales entre los grupos para una o más covariables. Como ejemplo, un psicólogo interesado en estudiar la eficacia de un programa conductual de pérdida de peso frente a la dieta realizada por el propio sujeto incluye los pesos antes del tratamiento como covariable.

Análisis estadístico univariante

Este análisis se utiliza cuando únicamente una sola variable dependiente debe considerarse en cada análisis. Los ejemplos son la prueba t y el ANOVA.

Análisis estadístico multivariante

Se trata de un procedimiento estadístico en el que se analizan simultáneamente múltiples variables correlacionadas teniendo en cuenta su intercorrelación. Se compara con los análisis univariantes. Algunos ejemplos son el análisis de función discriminante, el análisis factorial y el MANOVA. Puede encontrarse una introducción al uso de la estadística multivariante para datos de neuroimagen en: http://www.jove.com/video/1988/basics-of-multivariate-analysis-in-neuroimaging-data.

Análisis factorial

Esto se utiliza para reducir estadísticamente el número de variables necesarias para explicar o describir un conjunto mayor de variables iniciales basado en la matriz de correlación. Por ejemplo, las 10 puntuaciones de las subescalas de la prueba de personalidad del Minnesota Multiphasic Personality Inventory (MMPI) se obtuvieron de los 567 puntos que componen el cuestionario. Como se ha señalado anteriormente, el «análisis factorial agrupa variables, mientras que el análisis de conglomerados agrupa personas». Se considera un método estadístico multivariante porque se analizan simultáneamente muchas variables intercorrelacionadas. El método de análisis factorial más utilizado actualmente es el «análisis de componentes principales», que es más empírico que el método analítico factorial tradicional basado en la teoría. Después de que el análisis factorial reduce el número de factores (o supervariables) necesarios para resumir adecuadamente un gran grupo de variables, se realiza el método estadístico de rotación factorial para hacer que los factores sean más interpretables.

Análisis de funciones discriminantes

Este es el método adecuado para distinguir estadísticamente dos o más grupos basados en un grupo de variables discriminantes. Es un método importante e infrautilizado. Se considera un método estadístico multivariante porque se analizan simultáneamente muchas variables intercorrelacionadas.

Análisis por intención de tratar

En contraposición con un análisis de datos al final del estudio, en el que solo se analizan los datos de los sujetos que completan un estudio, lo que puede introducir un error ya que los sujetos con más acontecimientos adversos o falta de respuesta es más probable que abandonen el estudio y sus datos sean excluidos del análisis. Los métodos analíticos tradicionales del análisis por intención de tratar han utilizado métodos como la última observación de la variable principal tras tratamiento previo, en el cual la puntuación final registrada de un sujeto que abandona precozmente un estudio se repite en todos los períodos de evaluación posteriores. Sin embargo, estudios de simulación recientes han encontrado que este método tiende a aumentar las tasas de resultados falsos negativos y falsos positivos en muchas situaciones. Por tanto, se prefiere el MMRM.

Análisis multivariante de la varianza

El análisis multivariante de la varianza (MANOVA) es una generalización del ANOVA cuando deben evaluarse simultáneamente múltiples variables dependientes. No se utiliza con frecuencia en la investigación psiquiátrica debido a la dificultad que conlleva su interpretación. En su lugar, a menudo se calcula de forma separada un ANOVA para cada variable dependiente. Cuando se incluye más de una variable predictiva, se considera un método estadístico multivariante porque se analizan simultáneamente muchas variables intercorrelacionadas.

Análisis de la tabla de contingencia por chi cuadrado (χ^2)

Esta es una prueba para determinar si las frecuencias en cada casilla de una tabla de contingencia son diferentes de las proporciones esperadas por casualidad. Se usa más frecuentemente en una tabla de contingencia 2 × 2, representada por cuatro casillas formando un cuadrado.

Un uso frecuente es responder a la siguiente pregunta: «¿Existe una diferencia entre la aparición de un efecto adverso concreto en el grupo que recibe el fármaco frente al que recibe placebo?». En este caso, en la tabla se dispone el fármaco frente a placebo en las dos filas, y el efecto adverso frente a ningún efecto en las dos columnas. A medida que la diferencia (al cuadrado) entre las frecuencias observadas y esperadas en cada casilla aumenta, el estadístico χ^2 también aumenta y es más significativo el resultado. Si todas las casillas contienen exactamente las frecuencias que se esperaban por casualidad, el estadístico χ^2 es 0. Si las frecuencias difieren mucho de las esperadas por casualidad, el estadístico χ^2 será cada vez mayor. El tamaño del estadístico χ^2 se basa en el número de casillas en la tabla de contingencia (ya que los grados de libertad [gl] = [n.º de filas – 1] × [n.º de columnas – 1], una tabla 2×2 siempre tiene un único grado de libertad).

Análisis de la varianza

Esta es una prueba adecuada para estudiar la significación de la diferencia entre medias de tres o más grupos independientes. Por ejemplo, si un investigador médico quiere comparar los efectos de tres o más fármacos diferentes con una única medida dependiente, calculará una ANOVA unidireccional. La más compleja, ANOVA factorial también evalúa los efectos de la interacción entre múltiples factores. Por ejemplo, si los dos factores que se están evaluando son «fármaco/placebo» y «hombre/mujer», la prueba de interacción ANOVA puede demostrar que el fármaco es más eficaz que placebo solo en las mujeres. La significación del ANOVA se evalúa con el estadístico F.

Análisis de la varianza con medida(s) repetida(s)

Esta es la prueba adecuada para estudiar la significación de comparar variables continuas que se obtienen a través de mediciones repetidas en los mismos sujetos (debido a que las puntuaciones de cada sujeto se encuentran generalmente correlacionadas, el ANOVA normal daría resultados «demasiado significativos»). Un investigador puede seleccionar un diseño de medidas repetidas porque estas son generalmente más sensibles a los efectos del tratamiento (es decir, tienen un alto poder estadístico), puesto que se ignoran las diferencias de puntuación entre los sujetos.

Asignación aleatoria

El método más eficaz para asignar sujetos a grupos o programas terapéuticos es la aleatorización. La aleatorización asegura que no habrá errores sistemáticos en la composición de los grupos, de modo que cada sujeto tenga la misma posibilidad de ser asignado a cualquiera de los grupos. Advertencia: estrictamente hablando, los valores *P* obtenidos al comparar dos grupos son válidos solamente si los sujetos han sido asignados aleatoriamente a los dos grupos.

Coeficiente de correlación de Pearson

Esta es la medida adecuada para estudiar la asociación entre dos medidas continuas. Cuando no hay asociación, $r = 0$; cuando existe una correlación positiva perfecta, $r = 1$; y cuando hay una correlación negativa perfecta, $r = -1$. Puede obtener fácilmente la medida de tamaño del efecto al elevar al cuadrado *r* (que da el porcentaje de varianza compartida por dos variables); por tanto, incluso una correlación de $r = 0{,}05$ puede ser muy significativa con miles de sumas de cuadrados (*S*s), sin embargo, r^2 nos dice que dos variables comparten solo un porcentaje mínimo de su varianza (en este caso, ¡< 3:10 del 1%!). Cuando los investigadores describen la correlación, casi siempre se refieren al coeficiente de correlación de Pearson.

Coeficiente de correlación phi (ϕ)

Este es un caso especial del coeficiente de correlación de Pearson cuando se comparan dos variables dicotómicas. Tenga en cuenta, sin embargo, que debido a que las variables no se distribuyen normalmente, el valor máximo posible de φ es a menudo inferior a 1 o –1.

Comparaciones planificadas

Las comparaciones o contrastes planificados son tipos especiales de pruebas t para un subconjunto específico de hipótesis que incluyen comparaciones de medias que se formulan antes de recopilar datos para un estudio. Los contrastes planificados se realizan normalmente en lugar de un ANOVA completo.

Corrección de Bonferroni

Este es un método conservador para reducir las posibilidades de obtener resultados falsos positivos al evaluar un conjunto de pruebas estadísticas a un valor *P* más conservador. La corrección estándar de Bonferroni divide el valor nominal *P* (digamos $P < 0,05$) entre el número total de pruebas estadísticas que se están realizando. Por ejemplo, con 10 pruebas t realizadas, cada una sería probada a un valor $P < 0,005$ (es decir, 0,05:10) para determinar su significación.

Correlación

Véase el apartado «Coeficiente de correlación de Pearson».

Correlación biserial puntual

Este es un caso especial del coeficiente de correlación de Pearson cuando una variable es continua y la otra es dicotómica. Al igual que con el coeficiente φ, el valor máximo de r_{bp} a menudo no puede alcanzar sus límites de 1 o –1.

Correlación canónica

Esta es una generalización de la regresión múltiple en el caso de múltiples variables independientes y múltiples variables dependientes. Rara vez se usa hoy en día, excepto en estudios de neuroimagen con cientos o miles de medidas correlacionadas. Se considera un método estadístico multivariante porque se analizan simultáneamente muchas variables intercorrelacionadas.

Correlación por rangos

Este es un caso especial del coeficiente de correlación de Pearson cuando se evalúa la asociación entre dos variables con datos estratificados. Con solo unos pocos rangos apareados use el «tau» de Kendall; con muchos rangos apareados use el «rho» de Spearman.

Covariable

Esta es una variable que el investigador cree que puede influir en el resultado o variable dependiente y que se va a ajustar estadísticamente. Por ejemplo, en un estudio de un nuevo antidepresivo, la situación basal de depresión de los sujetos en cada uno de los dos grupos puede usarse como covariable.

Desviación estándar

Véase el apartado «Estadísticos descriptivos».

Efecto de interacción

En estadística se produce un efecto de interacción cuando el efecto simultáneo de dos variables en una tercera no es acumulativo. Por ejemplo, un diseño de investigación para probar un nuevo fármaco podría incluir las dos variables independientes de fármacos (activo frente a placebo) y de la edad del sujeto (personas jóvenes frente a las de edad avanzada). Si se demuestra que el fármaco activo es de media significativamente más eficaz en sujetos más jóvenes, esto expresaría un efecto de interacción significativo del fármaco por la edad del sujeto. Otra forma de expresarlo es que la edad de los sujetos *modera* el efecto del fármaco.

Escala de medición

Véase la tabla 62-3.

Estadísticos descriptivos

Estos son estadísticos utilizados para describir una población individual. Los estadísticos descriptivos utilizados normalmente para resumir la tendencia central de un grupo son la *media* (o media aritmética) para medidas continuas y la *mediana* (o «puntuación media») para medidas ordinales o estratificadas. Los estadísticos descriptivos se usan normalmente para describir la variabilidad dentro de un grupo. Entre estos se encuentran la *varianza* y su raíz cuadrada, la *desviación estándar* para las medidas continuas, y el *rango intercuartílico* para medidas estratificadas. Los investigadores determinan las medidas estadísticas descriptivas primero para obtener un «panorama general» de sus datos (y también para buscar errores en la introducción de datos o valores atípicos obvios).

Fiabilidad

Esta es la credibilidad de una puntuación, o la precisión con la que podemos estar seguros de que una medida puede ser fiable (es decir, ¿cuán reproducible es la puntuación?). Para las escalas de autoevaluación, como los cuestionarios de papel y lápiz, ya que no hay errores de evaluador que se deban tener en cuenta, la principal fuente de falta de fiabilidad que debe evaluarse es la diferencia en la autoevaluación del propio sujeto a lo largo del tiempo. Por ejemplo, si un paciente completa un cuestionario de depresión a las 15.00 h, ¿sería su puntuación con el mismo cuestionario la misma si él o ella realizaran la misma escala a las 16.00 h, asumiendo que no hay cambios en su depresión? Si las puntuaciones fueran idénticas, y este fuese el caso de todos los pacientes, el coeficiente de correlación sería de 1 (en este caso el coeficiente de correlación se denomina «coeficiente de fiabilidad»).

Para las escalas o medidas recogidas por un evaluador, la pregunta principal es: «¿Tendría este paciente la misma puntuación en esta escala de depresión tanto si lo evaluase el doctor A como si lo hiciera el doctor B?». Si la concordancia fuese perfecta para todos los pacientes, el coeficiente de fiabilidad sería 1. Si, por otra parte, hubiese una relación aleatoria entre las puntuaciones de los dos evaluadores, la fiabilidad interevaluador sería 0. La fiabilidad es necesaria, pero no suficiente, para que una escala sea útil. Así, una escala puede ser perfectamente fiable, pero no tener validez para un propósito concreto. Por ejemplo, cada vez que me pregunte mi número de teléfono le daré la misma respuesta (fiabilidad perfecta); sin embargo, si intenta usar mi número de teléfono para predecir mi estado de ansiedad, encontrará una correlación de 0 (sin validez). Si una medida no tiene fiabilidad (es decir, no es reproducible), tiene una fiabilidad de 0. Si tiene una fiabilidad perfecta, o repetitividad, tiene una fiabilidad de 1.

Para una medida continua, esta se evalúa con el coeficiente de correlación, y como regla general es normalmente $r = 0{,}8$ para una fiabilidad adecuada. Para una medida binaria (p. ej., presencia o ausencia de una enfermedad), esta se evalúa a menudo con el coeficiente κ, y por regla general es normalmente $\kappa = 0{,}7$.

Frecuencias

Véase el apartado «Estadísticos descriptivos».

Hipótesis nula

Los investigadores suelen comenzar con la hipótesis de que hay una diferencia de 0 (o «nula») entre dos medias o que existe un coeficiente de correlación de 0 (o «nulo») entre dos medidas. Dado que la estadística no puede probar una hipótesis, solemos expresar la hipótesis nula y presentar los datos que indican que es poco probable que sea verdadera; el valor *P* indica lo poco probable que es. Advertencia: en cualquier muestra muy grande, es probable que la hipótesis nula sea rechazada. Es decir, dos grupos de individuos rara vez tienen exactamente la misma media en cualquiera de dos características; incluso si solo difieren,

digamos, por 0,01 mm, la hipótesis nula no es verdadera, y un tamaño de muestra de miles de casos podría detectar incluso una pequeña diferencia.

Intervalo de confianza

Un intervalo de confianza no solo indica que la diferencia de medias que hemos observado entre dos grupos es de 2,5 puntos y significativa a un valor $P < 0,05$; es mucho más ilustrativo informar que el intervalo de confianza al 95% alrededor de nuestra media observada es de 0,6 a 4,4. Dado que 0 (el valor de la hipótesis nula de la diferencia de media) no está incluido en el intervalo de confianza al 95%, sabemos de un vistazo que la diferencia es significativa para un valor $P < 0,05$, pero también conoceremos los posibles valores de la diferencia de medias real entre los grupos, que van desde tan solo 0,6 puntos hasta un máximo de 4,4 puntos (con un intervalo de confianza al 95%). En el caso de razones de probabilidad, un intervalo de confianza de 0,6 a 4,4 no sería significativo a un valor $P < 0,05$, porque la hipótesis nula indicaría que el cociente de posibilidades es de 1, y esta estaría incluido en el intervalo de confianza calculado al 95%. Muchas revistas científicas requieren expresar los intervalos de confianza, en lugar de usar solo los valores *P*.

MANOVA

Véase el apartado «Análisis multivariante de la varianza».

Matriz de correlación

Esta es una «tabla» o «matriz» de los coeficientes de correlación de todas las variables de interés.

Media

Véase el apartado «Estadísticos descriptivos».

Mediación

En estadística, un modelo de mediación es aquel en el que una tercera variable explicativa subyace en la relación observada entre una variable independiente y una variable dependiente. La mediación suele probarse a través de un conjunto de modelos de regresión. Por ejemplo, un investigador puede encontrar que una nueva psicoterapia cognitiva reduce la depresión en un grupo de pacientes y puede suponer que esta relación significativa está *mediada* por reducciones en la puntuación media de los fallos cognitivos de los pacientes después del tratamiento. Puede encontrarse una buena descripción de la mediación en: http://davidakenny.net/cm/mediate.htm.

Mediana

Véase el apartado «Estadísticos descriptivos».

Modelo de efecto mixto de medidas repetidas

Este es el método estadístico preferido para analizar estudios clínicos longitudinales mediante el análisis de los gradientes individuales de cambio de cada sujeto, incluyendo aquellos que abandonan un estudio. Este método se prefiere al frecuentemente utilizado análisis por intención de tratar que usa métodos, como la última observación de la variable principal tras tratamiento previo, porque estudios de simulación han encontrado que el MMRM tienen una menor tendencia a proporcionar resultados falsos negativos y falsos positivos.

Moderación

Véase la definición en el apartado «Efecto de interacción». Puede encontrarse una buena descripción de la moderación en: http://davidakenny.net/cm/moderation.htm.

Poder estadístico

Véase el apartado «Potencia».

Potencia

Este refleja la probabilidad de encontrar una verdadera diferencia o relación entre dos o más medidas con un determinado tamaño muestral. Es similar a la sensibilidad de una prueba médica. Una analogía: la potencia de un telescopio (es decir, la magnificación) es análoga a la potencia de un estudio; ambos tienen la capacidad de detectar incluso pequeños objetos o cambios. Si un estudio tiene pocas posibilidades de encontrar una verdadera diferencia entre dos fármacos (digamos, con solo dos sujetos por grupo), la potencia del estudio sería casi de 0. Si el estudio no tiene prácticamente ninguna posibilidad de pasar por alto una verdadera diferencia (digamos, comparando la altura media de 10.000 niños de 2 años frente a 10.000 niños de 18 años), la potencia del estudio sería de casi 1. Una potencia de 0,8 suele ser la mínima requerida por los estadísticos al diseñar un estudio. Advertencia: así como cualquier pequeña diferencia puede resultar significativa solo con tener suficientes sujetos, es posible encontrar enormes diferencias, pero que no son significativas, simplemente por tener demasiados pocos sujetos. Sea especialmente cauto con los falsos negativos en estudios con pocos sujetos (p. ej., < 25 por grupo). Varias revisiones científicas han encontrado que el estudio medio en ciencias de la conducta tiene únicamente una potencia cercana al 40% para detectar un efecto de tamaño medio. Consulte las tablas 62-6 a 62-8 para estimar el poder estadístico de un estudio. Para análisis detallados del poder estadístico, descargue el excelente programa G*POWER en: www.gpower.hhu.de/en.html.

Programas de análisis estadístico

Se trata de programas comerciales o de *software* libre utilizados para la introducción, el manejo y el cálculo de datos estadísticos. Los programas comerciales más utilizados son SPSS, SAS y Stata. El programa gratuito más utilizado es R (descargable en: www.r-project.org).

Prueba de chi cuadrado (χ^2)

Véase el apartado «Análisis de la tabla de contingencia por chi cuadrado (χ^2)».

Prueba t

Esta es la prueba adecuada para determinar si la diferencia entre las medias de dos muestras de sujetos es probable que se deba solo a la casualidad. Si la diferencia entre dos medias es 0, el estadístico t también será 0. Si todas las puntuaciones en ambos grupos son las mismas, ambas tienen desviaciones estándar de 0. En este caso, el estadístico t será infinitamente grande.

Prueba t dependiente

Véase el apartado «Prueba t para medias dependientes».

Prueba t para medias dependientes

Esta es la prueba adecuada para estudiar la significación de la diferencia entre la media de dos grupos o niveles de tratamiento emparejadas o dependientes. Como ejemplo, un grupo de 30 pacientes ansiosos participaron en un estudio con un nuevo fármaco para aliviar la ansiedad. Los niveles de ansiedad se midieron mediante la escala de ansiedad de Hamilton al inicio del estudio y nuevamente después de 3 semanas de tratamiento. En esta prueba se comparan directamente las puntuaciones de cada par de sujetos y se ignora los cambios entre los pares de sujetos. La misma prueba se usa para comparar las medias de dos pares de sujetos emparejados (cuyas puntuaciones se consideran correlacionadas).

Prueba U de Mann-Whitney

Esta es la prueba no paramétrica adecuada para estudiar la significación de diferencias entre las sumas de rangos de dos grupos o niveles de tratamiento independientes.

r

Véase el apartado «Coeficiente de correlación de Pearson».

Regresión

Véase el apartado «Regresión lineal múltiple».

Regresión lineal múltiple

Esta se utiliza para predecir una única variable de resultado continua a partir de un conjunto de variables predictivas continuas o dicotómicas. Es muy flexible porque permite muchas formas de ajuste de la curva no lineal y automáticamente proporciona una medida del tamaño del efecto en forma de R^2 (esto puede ayudar a evaluar la significación práctica además de la significación estadística). Cuando se incluye más de una variable predictiva, se considera un método estadístico multivariante porque se analizan simultáneamente muchas variables intercorrelacionadas.

Regresión logística

Este es un método adecuado para predecir una variable de resultado binario a partir de un conjunto de variables predictivas continuas o binarias. La regresión logística da lugar a un valor de probabilidad y un cociente de posibilidades. A menudo se utiliza en estudios epidemiológicos donde el resultado es la ocurrencia/no ocurrencia de una enfermedad o la supervivencia/muerte. Cuando se incluye más de una variable predictiva, se considera un procedimiento estadístico multivariante porque se analizan simultáneamente muchas variables intercorrelacionadas.

Rho

Véase el apartado «Correlación por rangos».

Tamaño del efecto

Esta es una medida que evalúa la significación práctica del efecto de un tratamiento, en contraposición a la significación estadística. Para comparar dos tratamientos, la medida del tamaño de efecto más frecuentemente utilizada es la «d» de Cohen (la diferencia media entre los grupos en unidades de desviación estándar), con valores de 0,2, 0,5 y 0,8 definidos por Cohen como pequeños, medianos y grandes efectos en las ciencias de la conducta. Los tamaños del efecto son esenciales para calcular el poder estadístico (v. tablas 62-6 a 62-8) y también constituyen la base de los metaanálisis.

Tasa de error por experimento

A medida que se realizan más pruebas estadísticas, es más probable que al menos una de estas pruebas sea significativa por casualidad. Aunque puede realizar cada prueba a un valor $P < 0,05$, la posibilidad de encontrar una de muchas pruebas significativa por casualidad es mucho mayor que este nivel nominal del 5% (v. tabla 62-2).

Tau

Véase el apartado «Correlación por rangos».

Validez

Este es el grado de utilidad de una escala de evaluación para un propósito concreto. También es el grado en que la prueba mide lo que se supone que está midiendo. La determinación de la validez normalmente suele requerir criterios independientes, ajenos a lo que la prueba está diseñada para medir. Por ejemplo, si un investigador desarrolla una sola pregunta que pretende ser un buen instrumento de detección de la depresión clínica, las respuestas de los pacientes a esta pregunta deben relacionarse adecuadamente con las mediciones «estándar de referencia» de la depresión clínica, como entrevistas estructuradas y escalas de evaluación bien definidas para la depresión. No hay ninguna regla general para la validez, ya que no hay una medida única. Como mínimo, sin embargo, la escala debe estar al menos correlacionada significativamente con las medidas estándar de referencia para esa característica.

Valor *P*

Refleja la posibilidad de que el resultado de una prueba estadística sea un falso positivo (es decir, la probabilidad de un resultado falso). Si un resultado es casi seguro falso, el cual no sería reproducible en otra muestra, el valor *P* estará cerca del 1. Por otro lado, si el hallazgo, casi con toda seguridad, representa un resultado «verdadero», el valor *P* estará cerca del 0 (los valores *P* pequeños se representan con varios ceros después de la coma decimal, por ejemplo, $P < 0{,}0001$, pero nunca presente un valor *P* como 0, porque esto es imposible, y ¡molesta a los revisores!).

La mayoría de las revistas requieren un valor $P < 0{,}05$ para demostrar la significación. Si se realizan muchas pruebas estadísticas en un estudio, se puede usar un valor *P* más conservador para minimizar el error por experimento (v. tabla 62-2). Advertencia: no se deje impresionar por valores *P* muy bajos. Recuerde que lo que esto quiere indicar es la posibilidad de que la diferencia probablemente no sea 0. Además, recuerde que, con suficientes sujetos, este es fácil de probar. Por tanto, un valor *P* muy bajo no indica necesariamente un gran efecto clínico, sino que representa un efecto muy fiable. Compruebe el tamaño del efecto (el coeficiente de correlación al cuadrado o el tamaño del estadístico t) para tener una idea de la magnitud de la diferencia o la relación.

La mayoría de los valores *P* citados en la investigación publicada son de dos colas o no direccionales. Sin embargo, cuando el investigador plantea *a priori* una fuerte hipótesis direccional (p. ej., «el fármaco será más eficaz que placebo para mejorar la depresión») en lugar de una hipótesis no direccional (p. ej., «el grupo del fármaco y el de placebo serán significativamente diferentes en relación con la variable principal del estudio»), puede ser preferible un valor *P* de una cola, reduciendo las posibilidades de resultados falsos negativos al incrementar el poder estadístico.

Cuando los resultados falsos negativos son más nocivos que los falsos positivos (como al evaluar acontecimientos adversos graves) tiene sentido probar a un valor mayor, digamos, $P < 0{,}1$, para minimizar el riesgo de falsos negativos debido al pequeño tamaño de la muestra.

Variable dependiente

Esta es habitualmente la variable de resultado de interés en un estudio; también se denomina «variable principal». En el estudio de un nuevo fármaco antidepresivo, puede utilizarse una escala de evaluación de la depresión como variable dependiente.

Variable principal

Véase el apartado «Variable dependiente».

Variable de resultado

Véase el apartado «Variable dependiente». También se conoce como «variable de respuesta».

Varianza

Véase el apartado «Estadísticos descriptivos».

Controversias actuales y consideraciones futuras

- Existe un movimiento promovido por la literatura científica de alejarse de presentar simplemente valores *P* y tender a presentar el intervalo de confianza de carácter más ilustrativo.
- Se alienta a los investigadores a mostrar los análisis estadísticos del tamaño de efecto además de los valores *P*.
- Los métodos estadísticos multivariantes (como el análisis discriminante, el análisis de componentes principales, el análisis de conglomerados y la correlación canónica) se utilizan más frecuentemente en investigación médica debido al gran número de variables correlacionadas generadas por la investigación en neuroimagen, genómica y metabolómica.
- La tasa de error por experimento (falso positivo) es aún más preocupante en áreas de investigación (como la neuroimagen, la genómica y la metabolómica) donde son posibles muchas pruebas estadísticas en un conjunto de datos determinado.
- Son preferibles los análisis de MMRM a los abordajes de la última observación de la variable principal en el análisis de estudios clínicos longitudinales con datos perdidos.

Acceda *online* a las preguntas de opción múltiple (en inglés) en https://expertconsult.inkling.com

Lecturas recomendadas

Altman DG. Why we need confidence intervals. *World J Surg*. 2005;29(5):554–556.

Cohen J. *Statistical power analysis for the behavioral sciences*. ed 2. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.

Cohen J, Cohen P, West SG, et al. *Applied multiple regression/correlation analysis for the behavioral sciences*. ed 3. Oxford: Routledge; 2002.

Ellis PD. *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge: Cambridge University Press; 2010.

Gonick L, Smith W. *The cartoon guide to statistics*. New York: HarperCollins; 1993.

Gorsuch R. *Factor analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1983.

Huff D, Geis I. *How to lie with statistics*. New York: WW Norton; 1954.

Keith TZ. *Multiple regression and beyond*. Boston: Pearson Educational; 2006.

Reich D, Price AL, Patterson N. Principal component analysis of genetic data. *Nat Genet*. 2008;40(5):491–492.

Rosenthal R, Rosnow RL. *Essentials of behavioral research: methods and data analysis*. ed 3. New York: McGraw-Hill; 2007.

Siddiqui O, Hung HM, O'Neill R. MMRM vs. LOCF: A comprehensive comparison based on simulation study and 25 NDA datasets. *J Biopharm Stat*. 2009;19(2):227–246.

Tabachnick BG, Fidell LS. *Using multivariate statistics*. ed 6. Pearson; 2012.