14. Análisis e interpretación de datos

Conceptos y técnicas para manejar, editar, analizar e interpretar los datos de estudios epidemiológicos.

Conceptos claves/expectativas

Este capítulo contiene una gran cantidad de material y va más allá de lo que se espera que tú aprendas en este curso (i.e., para preguntas de examen.) Sin embargo, los temas estadísticos impregnan los estudios epidemiológicos, y puedes encontrar que parte del material que sigue puede serte útil cuando leas la literatura. De manera que si te parece que te estás perdiendo y empiezas a preguntarte que es lo que se supone que *debes* aprender, por favor toma como referencia la siguiente lista de conceptos que esperamos que logres adquirir:

- La necesidad de editar los datos antes de emprender un análisis en serio y captar los errores lo antes posible.
- Opciones para limpiar los datos – verificación de rangos, verificación de consistencia – y lo que estos pueden (y no pueden) lograr.
- Qué significa la codificación de los datos y porqué se realiza.
- Significado básico de varios términos usados para caracterizar los atributos matemáticos de distintos tipos de variables, i.e., nominal, dicotómica, categórica, ordinal, de medición, conteo, discreta, intervalo, razón, continua. Reconocer ejemplos de diferentes tipos de variables y ventajas/desventajas de tratarlas de diferentes maneras.
- Qué significa una variable "derivada" y diferentes tipos de variables derivadas.
- Los objetivos de las pruebas de hipótesis estadísticas ("pruebas de significancia"), el significado de los resultados de dichas pruebas y cómo interpretar un valor p.
- Qué es un intervalo de confianza y cómo debe ser interpretado.
- Los conceptos de error de Tipo I y error de Tipo II, nivel de significancia, nivel de confianza, "potencia" estadística, precisión estadística, y la relación entre estos conceptos y el tamaño muestral.

El cálculo de valores p, intervalos de confianza, potencia o tamaño muestral no será requerido en los exámenes. La prueba exacta de Fisher, pruebas asintóticas, tablas z, pruebas de 1 o 2 colas, correlación dentro del cluster, enfoques Bayesianos versus los frecuentistas, meta-análisis, e interpretación de pruebas de significancia múltiple son todos simplemente para tu información y disfrute, en cuanto a lo que tiene que ver con EPID 168, no para los exámenes. En general, yo promuevo un enfoque no dogmático a la estadística (*advierto* que no soy un estadístico "licenciado"!)

Análisis e interpretación de datos

Los epidemiólogos a menudo hallan el análisis de los datos como la parte más disfrutable de llevar a cabo un estudio epidemiológico, dado que después de todo el duro trabajo y la espera, tienen la oportunidad de encontrar las respuestas. Si los datos no proveen respuestas, es una oportunidad más para la creatividad! De manera que el análisis y la interpretación de los resultados son el "premio" que recompensa el trabajo de recolección de datos.

Los datos, sin embargo, no "hablan por sí mismos". Revelan lo que el analista puede detectar. De manera que cuando el investigador novato, tratando de obtener esta recompensa, se encuentra sólo con el conjunto de datos y ninguna idea de como proceder, la sensación puede ser una de más ansiedad que de entusiasta anticipación. Igual que con otros aspectos de un estudio, el análisis e interpretación del estudio debe relacionarse con los objetivos del mismo y el problema de investigación. Una estrategia, a menudo útil, es comenzar imaginando o hasta trazando el (los) manuscrito(s) que deberían escribirse a partir de los datos.

El enfoque habitual es comenzar con los análisis descriptivos, explorar y lograr "sentir" los datos. El analista luego dirige su atención a las preguntas específicas planteadas en los objetivos o hipótesis de estudio, de los hallazgos y planteos informados en la literatura, y de los patrones sugeridos por los análisis descriptivos. Antes de comenzar el análisis en serio, sin embargo, habitualmente hay que llevar a cabo una cantidad considerable de trabajo preparatorio.

Análisis – objetivos principales

1. Evaluar y realzar la calidad de los datos
2. Describir la población de estudio y su relación con alguna supuesta fuente (justificar todos los pacientes potenciales involucrados; comparación de la población de estudio obtenida con la población blanco)
3. Evaluar la posibilidad de sesgos (p.ej., no-respuesta, negativa a contestar, y desaparición de sujetos, grupos de comparación)
4. Estimar las medidas de frecuencia y extensión (prevalencia, incidencia, media, mediana)
5. Estimar medidas de fuerza de asociación o efecto
6. Evaluar el grado de incertidumbre a partir del azar ("ruido")
7. Controlar y analizar los efectos de otros factores relevantes
8. Buscar una mayor comprensión de las relaciones observadas o no observadas
9. Evaluar el impacto o importancia

Trabajo preparatorio – Edición de datos

En un estudio bien ejecutado, el plan de recolección de datos incluye procedimientos, instrumentos, y formularios, diseñados y ensayados para maximizar su precisión. Todas las actividades de recolección de datos son monitorizadas para asegurar la adherencia al protocolo de recolección de datos y para promover acciones para minimizar y resolver situaciones de datos faltantes o cuestionables. Los procedimientos de monitorización son establecidos al inicio y mantenidos durante todo el estudio, dado que cuanto antes se detecten las irregularidades, mayor la probabilidad de que puedan ser resueltas de manera satisfactoria y más precozmente se puedan establecer medidas preventivas.

Sin embargo, a menudo hay necesidad de "editar" los datos, tanto antes como después de computarizarlos. El primer paso es "manual" o "edición visual". Antes de digitar los formularios (salvo que los datos sean entrados en la computadora en el momento de recolección, p.ej., a través de programas como CATI - entrevistas telefónicas asistidas por computadora (computer-assisted telephone interviewing), los formularios deben ser revisados para identificar irregularidades y problemas que pasaron desapercibidos o no fueron corregidos durante el monitoreo.

Las preguntas abiertas, si están presentes, habitualmente necesitan ser codificadas. También puede ser necesaria la codificación de las preguntas cerradas salvo que las respuestas sean "precodificadas" (i.e., tengan un número o letra que corresponda a cada respuesta elegida.) Aún los formularios que sólo tienen preguntas cerradas con respuestas precodificadas pueden requerir codificación en el caso de respuestas poco claras o ambiguas, múltiples respuestas para un solo ítem, comentarios escritos de parte del participante o del recolector de datos, y otras situaciones que puedan surgir. (La codificación será descrita en mayor detalle más adelante.) Es posible, en esta etapa, detectar problemas con los datos (p.ej., respuestas inconsistentes o fuera del rango), pero habitualmente éstas se manejan en forma sistemática en el momento de, o inmediatamente después, de la introducción de los datos en la computadora. La edición visual también presenta una oportunidad para tener una impresión de qué tan bien fueron completados los formularios y con qué frecuencia se presentaron algunos tipos de problemas.

A continuación los formularios de datos serán digitados, típicamente en una computadora personal o una terminal de computadoras para la cual el programador ha diseñado pantallas de entrada de datos con un formato similar al del cuestionario. Sin embargo, cuando el cuestionario o formulario de recolección de datos es corto, los datos pueden ser introducidos directamente en una planilla de datos o aún en un archivo de texto. Un programa específico de entrada de datos a menudo verifica cada valor en el momento en que es introducido, de manera de evitar que se ingresen valores ilegales en la base de datos. Esta acción sirve para evitar errores de digitación, pero también detectará respuestas ilegales en el formulario que pasaron sin detección en la edición visual. Por supuesto que debe existir un procedimiento para manejar estas situaciones.

Dado que la mayor parte de los estudios epidemiológicos recogen grandes cantidades de datos, la monitorización, edición visual, entrada de datos y consiguiente verificación de datos, típicamente son realizadas por múltiples personas, a menudo con distintos niveles de destreza, experiencia y autoridad, durante un período de tiempo prolongado y en múltiples lugares. Los procedimientos de

procesamiento de datos deben tomar estas diferencias en cuenta, de manera que cuando se detectan problemas o surgen preguntas hay una forma eficiente para resolverlos, y además el personal de análisis y/o los investigadores tengan formas de conocer la información obtenida a través de los múltiples pasos del proceso de edición. Técnicas como las de "batching" (agrupar en lotes), en que los formularios y otros materiales se dividen en conjuntos (p.ej., 50 formularios), se cuentan, posiblemente se suman uno o dos campos numéricos, y se rastrean como grupo, sirven para ayudar a disminuir la pérdida de formularios de datos. El control de calidad y la seguridad son siempre temas críticos. Su cumplimiento se vuelve tanto más complejo cuanto mayor el número de personal participante y cuanto más diversa su experiencia.

Trabajo preparatorio - limpieza de datos

Una vez que los datos son introducidos en la computadora y son verificados (pueden verificarse por introducción por dos personas o por verificación visual) son sometidos a una serie de verificaciones por la computadora para "limpiarlos".

Verificación de rangos

La verificación de rango compara cada dato con un conjunto de valores permitidos y usuales para esa variable. La verificación de rango se usa para:

1. Detectar y corregir valores no válidos
2. Identificar e investigar valores inusuales
3. Señalar valores atípicos o extremos ("outliers") (aún si son correctos, su presencia puede influir sobre los métodos estadísticos a utilizar)
4. Verificar la lógica de las distribuciones y también apreciar sus formas, dado que esto también afectará la selección de procedimientos estadísticos

Verificación de la consistencia

La verificación de la consistencia examina cada par (a veces más) de datos relacionados, en relación con el conjunto de valores habituales y permitidos de las variables como par. Por ejemplo, los hombres no deben haber tenido una histerectomía. Los estudiantes universitarios habitualmente tienen por lo menos 18 años (aunque pueden haber excepciones, por eso se considera que la verificación de la consistencia es un procedimiento "blando", no "duro".) La verificación de la consistencia se usa para:

1. Detectar y corregir las combinaciones no permitidas
2. Señalar e investigar combinaciones inusuales
3. Verificar la consistencia de los denominadores y valores "ausentes" y "no corresponde" (i.e., verificar que los patrones de salteado de llenado han sido cumplidos
4. Verificar la lógica de las distribuciones conjuntas (p.ej., en los gráficos de puntos)

En situaciones en que se encuentran muchas respuestas inconsistentes, el enfoque que se utiliza para manejar la inconsistencia puede tener un impacto notorio sobre las estimaciones y puede alterar comparaciones entre grupos. Los autores deben describir las reglas de decisión utilizadas para manejar la inconsistencia y cómo los procedimientos afectan los resultados (Bauer y Jonson, 2000.)

Trabajo de preparación – codificación de los datos

La codificación de los datos significa la traducción de la información en valores adecuados para ser ingresados en la computadora y para el análisis estadístico. Todo tipo de datos (p.ej., historias clínicas, cuestionarios, pruebas de laboratorio) debe ser codificado, aunque en algunos casos la codificación ha sido realizada previamente. El objetivo es crear variables a partir de la información, con la posibilidad de análisis en mente. Las siguientes interrogantes subyacen las decisiones sobre codificación:

1. ¿Qué información existe?
2. ¿Qué información es relevante?
3. ¿Cómo será probablemente analizada?

Ejemplos de decisiones sobre codificación y edición de datos

- Un criterio típico para la seropositividad para VIH es un ELISA repetidamente positivo (ensayo inmunoenzimático recombinante) para anticuerpos VIH confirmado por una prueba Western Blot con elfin de identificar la presencia de proteínas particulares (p.ej., p24, gp41, gp120/160.) De esta manera los datos del laboratorio pueden incluir todos los siguientes:
  a. Una evaluación global del estado VIH (positivo/negativo/indeterminado)
  b. Pares de resultados de ELISA expresados como:
     i. + + / + – / – – / indeterminado
     ii. densidades ópticas
  c. Resultados de Western Blot (para las personas con resultados de ELISA positivos) expresados como:
     i. (+ / – / indeterminado)
     ii. detección de bandas específicas de proteínas, p.ej., p24, gp41, gp120/160

  ¿Cuánta de esta información debe ser codificada y tecleada?

- ¿Cómo codificar las preguntas abiertas del cuestionario (p.ej., "¿De qué manera ha cambiado su hábito de fumar?", "¿Cuáles son sus razones para dejar de fumar?" , "¿Qué impedimentos al cambio espera encontrar?", "¿Qué hacía en su trabajo?")

- Las preguntas cerradas pueden ser "auto-codificadas" (i.e., el código a ser tecleado está listado al lado de cada opción de respuesta), pero también puede haber:
  a. múltiples respuestas cuando sólo se necesita una – pueden ser:

1. Respuestas inconsistentes (p.ej., "Nunca" y "2 veces o más")
2. Respuestas adyacentes que indican un rango (p.ej., "dos o tres veces" y "cuatro o cinco veces" de parte de un sujeto que no puede elegir entre 2-5 veces).

b. Respuestas salteadas – que deben diferenciar entre

1. Preguntas que no corresponden para este entrevistado (p.ej. edad de la menarca para encuestados de sexo masculino).
2. Encuestados que optan por no contestar (que se puede indicar como "N/C"!)
3. Encuestado que no sabe o no puede recordar
4. Encuestado que saltea una pregunta sin una razón aparente

Es necesario llegar a un balance entre la codificación de lo mínimo y la codificación de "todo".

- La codificación es más sencilla cuando se hace toda de una vez.
- Uno siempre puede ignorar posteriormente las opciones codificadas que se consideran sin importancia.
- La información no codificada no estará a disposición para su análisis (p.ej., la fecha en que se recibió el cuestionario, qué cuestionarios fueron seleccionados al azar para una encuesta de verificación basada en 10% de los cuestionarios totales).
- Mayores detalles significan más re-codificaciones para el análisis lo cual significa más programación y por lo tanto más oportunidades para cometer errores.
- Las decisiones postergadas deben ser concretadas en algún momento, así que porqué no hacerlo de entrada (p.ej., cuando un entrevistado marca respuestas adyacentes como "3. una o dos veces" y "4. Dos a cinco veces" ¿qué debe codificarse – ¿3?,¿4?, ¿3.5? ¿un valor codificado faltante? ¿Un código a ser reemplazado en el futuro cuando se tome una decisión?
- Es importante documentar cómo se realizó la codificación y cómo se resolvieron los problemas, de manera de mantener la consistencia y contestar las inevitables preguntas ("¿Cómo manejamos tal situación?")

Tipos de variables – niveles o escalas de medición

Los constructos o factores en estudio son representados por "variables". Las variables (también llamadas "factores") tienen "valores" o "niveles". Las variables resumen y reducen los datos, tratando de representar la información "esencial".

Técnicas analíticas dependiendo del tipo de variable

Las variables pueden ser clasificadas de diversas maneras. Una ***variable continua*** toma todos los valores dentro de su rango permitido, de manera que entre dos valores cualesquiera dentro del rango hay otros valores legítimos entre ellos. Una variable continua (llamada también a veces "variable de medición") se usa en respuesta a la pregunta "¿cuánto?". Las mediciones como peso, altura, y la

presión arterial pueden, en principio, ser representadas por variables continuas y frecuentemente son tratadas como tales en los análisis estadísticos. En la práctica, por supuesto, los instrumentos utilizados para medir estos y otros fenómenos y la precisión con que se registran los valores permiten sólo un número finito de valores, pero estos pueden ser considerados como puntos en un continuo. Matemáticamente, una ***variable discreta*** puede tomar sólo ciertos valores entre sus valores máximo y mínimo, aún si no hay un límite para el número de dichos valores (p.ej., el conjunto de números racionales es pasible de ser contado aunque es ilimitado en número). Las variables discretas que pueden tomar cualquier valor dentro de un conjunto grande de valores a menudo son tratadas como si fueran continuas. Si los valores de una variable pueden ser ordenados, el hecho de que el analista decida tratar la variable como discreta y/o continua depende de la distribución de la variable, los requerimientos de los procedimientos analíticos disponibles y la opinión del analista sobre la interpretación de los resultados que se pueden obtener.

Tipos de variables discretas

1. ***Identificación*** – una variable que simplemente nombra cada observación (p.ej., un número de identificación en el estudio) y que no se usa en el análisis estadístico;

2. ***Nominal*** – una categorización o clasificación, que no tiene un orden inherente; los valores o la variable son completamente arbitrarios y podrían ser reemplazados por cualquier otro sin afectar los resultados (p.ej., grupos sanguíneos ABO, número de registro en la clínica, etnia). Las variables nominales pueden ser ***dicotómicas*** (dos categorías, p.ej. sexo) o ***politómicas*** (más de dos categorías).

3. ***Ordinal*** – una clasificación en que los valores pueden ser ordenados o tienen un orden; dado que los valores codificados sólo necesitan reflejar el orden pueden ser reemplazados por cualquier otro conjunto de valores con el mismo ordenamiento relativo (p.ej., 1, 2, 5; 6, 22, 69; 3.5, 4.2, 6.9 podrían todos ser utilizados en vez de 1, 2 ,3). Como ejemplos podemos considerar la severidad de las lesiones y la situación socioeconómica.

4. ***De conteo*** –el número de entidades, eventos, o algún otro fenómeno que puede ser contado, para el cual la pregunta relevante es "cuántos?" (p.ej., paridad, número de hermanos); la sustitución de los valores de la variable por otros números cambiaría su sentido. En el análisis de datos epidemiológicos, las variables de conteo a menudo se tratan como continuas, sobretodo cuando sus posibles valores son muchos.

Tipos de variables continuas

1. ***De intervalo*** – las diferencias (intervalos) entre los valores tienen significado, pero las razones entre los valores no lo tienen. Es decir, que si la variable toma los valores 11-88, con un promedio de 40, tiene sentido afirmar que el puntaje del sujeto A de 60 "se aleja el doble del promedio" que el puntaje de 50 del sujeto B. Pero no tiene sentido decir que el puntaje del sujeto A es "1.5 veces el promedio". La razón es que el punto cero de la escala es arbitrario, de manera que los valores de los puntajes tienen sentido sólo en relación uno con el otro. Sin pérdida de información, la escala puede ser desplazada: 11-88 puede ser transformada en 0-77 restando 11 puntos. Los puntajes de la escala también pueden ser multiplicados por una constante. Después de cualquiera de estas dos transformaciones, el puntaje del sujeto A sigue alejándose el doble de la media que el puntaje del sujeto B, pero el puntaje del sujeto A no es más 1.5 veces el puntaje promedio. Las escalas sicológicas (p.ej.,

para ansiedad, para depresión) a menudo utilizan este tipo de escalas. Un ejemplo que proviene de la física es la temperatura medida en escala de Fahrenheit o Celsius.

2. ***De razón*** – tanto las diferencias como las razones tienen sentido. Tienen un punto cero no arbitrario, de manera que tiene sentido caracterizar un valor como "x" veces el valor del promedio. Cualquier transformación salvo la multiplicación por una constante (p.ej., el cambio de unidades) distorsionará las relaciones de los valores de una variable medida en una escala de razón. Los parámetros fisiológicos como la presión arterial o el colesterol son medidas de razón. La temperatura absoluta o Kelvin es una medida en escala de razón.

Muchas variables importantes en epidemiología son dicotómicas (i.e., nominal con dos niveles) - enfermo vs. sano, expuesto vs. no expuesto. Aunque una variable puede aparentar ser ordinal o continua, el propio fenómeno puede no merecer ser tratado como tal. Sería necesario preguntarse: "¿Es que "más" es realmente más?" y " ¿hay umbrales o discontinuidades involucradas?" De nuevo, la realidad subyacente (o, más bien, el modelo conceptual que tengamos de ella) determina el enfoque de la cuantificación. Los valores de las variables a menudo son agrupados en un pequeño número de categorías para algunos análisis y utilizados en su forma original para otros.

Trabajo preparatorio –reducción de datos

La reducción de datos busca reducir el número de variables para el análisis combinando variables únicas en variables compuestas que cuantifican mejor el constructo. Las variables creadas durante el intento de codificación para reflejar fielmente los datos originales (p.ej., altura, peso.) A menudo se pueden utilizar directamente estas variables para el análisis, pero también es necesario frecuentemente crear variables adicionales para representar constructos de interés. Por ejemplo, el constructo sobrepeso se representa a menudo por una variables que se deriva de los valores para peso y altura. La reducción de datos incluye la simplificación de las variables individuales (p.ej., la reducción de seis posibles valores a un número menor) y la derivación de variables compuestas (p.ej., "nivel socioeconómico" derivado de educación y ocupación.)

En general:

- Lo simple es mejor
- Evitemos detalles superfluos
- Creamos variables adicionales, antes que destruir las originales (nunca hay que sobrescribir los datos crudos!)
- Analicemos los detalles antes de confiar en los resúmenes
- Verificar la precisión de las variables derivadas y las recodificadas estudiando las tablas de cruce de variables entre las variables originales y las derivadas
- Tomemos en cuenta los efectos de umbral, fenómenos de saturación y otras situaciones de no -linealidad

- Creamos categorías basadas en la naturaleza del fenómeno (p.ej., un estudio del Sindrome de Down puede juntar todas las categorías de edad por debajo de 30 años; un estudio de tasas de embarazo va a necesitar una mayor división de las edades por debajo de los 30 años y aún por debajo de los 20 años.)

Tipos de variables derivadas

Escalas – En una escala pura (p.ej., depresión, autoestima) todos los ítems son supuestamente medidas individuales del mismo constructo. El puntaje de la escala es habitualmente la suma de los valores de respuesta de los ítems, aunque los ítems con un sentido inverso (p.ej., "Me siento feliz" en una escala de depresión) deben ser invertidos. El propósito de derivar un puntaje de la escala utilizando múltiples ítems es el de obtener una medida más confiable del constructo que la que es posible a partir de un solo ítem. La confiabilidad de la escala (consistencia interna) se evalúa clásicamente usando el ***coeficiente alfa*** de Cronbach, que se puede considerar como el promedio de todas las correlaciones inter-item. Si los ítems miden efectivamente el mismo constructo de la misma manera y de hecho fueron contestados de idéntica manera, las únicas diferencias en sus valores deberían deberse a errores aleatorios de medición. El alfa de Cronbach da la proporción de la variación total de los puntajes de la escala que no es atribuible al error aleatorio. Valores de 0.80 o más son considerados adecuados para una escala que será utilizada para analizar asociaciones (si la escala es utilizada como instrumento clínico para pacientes individuales, su alfa debe ser de por lo menos 0.90 – ver el texto de Nunally, *Psychometrics*). Cuando la escala consiste de sub-escalas separadas, la consistencia interna puede ser más relevante para las sub-escalas individuales que para la escala como una unidad. Los análisis de las relaciones entre los ítems individuales (correlación inter-ítem o concordancia), entre cada ítem y los restantes ítems (correlación ítem-resto), entre cada ítem y la escala total (correlación escala-ítem), y entre los grupos de ítems (análisis de factores) son métodos habituales para analizar el desempeño de los ítems.

Indices – un índice consiste de un grupo de ítems que están combinados (habitualmente sumados) para dar una medida de un constructo multidimensional. En este caso, cada uno de los ítems mide un aspecto o dimensión diferente, de manera que las medidas de consistencia interna como el alfa de Cronbach o no son relevantes o requieren una interpretación diferente. Ejemplos de índices derivados de varias variables incluyen el estado socioeconómico (p.ej., ocupación, ingresos, educación, barrio), apoyo social (p.ej., estado civil, número de familiares cercanos, número de amigos cercanos), comportamiento de riesgo sexual (número de compañeros, tipo de compañeros, uso de preservativos, sexo anal). Los ítems pueden tener ponderaciones diferentes, dependiendo de su importancia relativa y la escala en que fueron medidos.

Algoritmos – un procedimiento que utiliza un conjunto de criterios según reglas o consideraciones específicas, p.ej., trastorno depresivo mayor, anticoncepción "efectiva" (no he visto, hasta ahora, utilizar este término para denominar este tipo de variable, pero no conozco ningún otro término para este concepto.)

Trabajo preparatorio – Explorando los datos

Trata de "sentir" los datos– analizar la distribución de cada variable. Examina gráficos de correlación bivariados y cruces de variables. ¿Tienen sentido los patrones que aparecen? ¿Son creíbles?

- Observa la forma – simetría vs. asimetría, interrupciones en la forma
- Elige medidas de resumen apropiadas para la distribución y tipo de variable (nominal, ordinal, medida)

  De posición – media, mediana, porcentaje por encima del punto de corte

  Dispersión – desvío estándar, cuantiles
- Busca relaciones entre los datos
- Mira dentro de los subgrupos importantes
- Observa la proporción de valores faltantes

Trabajo preparatorio – Valores faltantes

Los datos que faltan son un estorbo y pueden ser un problema. Por un lado, las respuestas que faltan significan que los denominadores para muchos análisis pueden ser diferentes, lo cual puede confundir y además es tedioso de explicar. Por otro lado los análisis que involucran múltiples variables (p.ej., coeficiente alfa, tabulaciones cruzadas, modelos de regresión) generalmente excluyen la observación entera si le falta el valor para cualquier variable en el análisis (este método se llama ***eliminación por órden de lista*** [N.T. ***listwise deletion*** en inglés]. De esta manera, un análisis que involucra 10 variables, aún si cada una tiene sólo un 5% de valores faltantes, puede resultar en la exclusión de hasta un 50% de la base de datos (si no hay superposición entre las respuestas faltantes)! Es más, salvo que los ***datos falten totalmente al azar*** (en inglés ***missing completely at random MCAR-*** lo cual es equivalente a un patrón de datos faltantes que resultaría de borrar valores en la base de datos sin ninguna sistematización o preferencia) un análisis que no ajusta para los datos faltantes será sesgado, porque ciertos subgrupos estarán sub-representados en los datos disponibles (un tipo de sesgo de selección).

Imputación para los valores faltantes – tema optativo

A medida que, a través de los años, se han desarrollado las teorías, los métodos y el poder de la informática, los métodos analíticos para el manejo de los datos faltantes, para minimizar sus efectos perjudiciales han mejorado también. Estos métodos buscan ***imputar*** los valores para las respuestas faltantes a los ítems de manera de tratar de aumentar la eficiencia estadística (evitando la pérdida de observaciones que tienen uno o unos pocos valores faltantes) y disminuir el sesgo. Métodos antiguos de imputación, abandonados hoy día, incluyen el reemplazo de cada valor faltante por el promedio o la mediana de esa variable. Aunque esas prácticas permiten que todas las observaciones sean utilizadas en los análisis de regresión, estos métodos no disminuyen el sesgo y tienden a introducir una distorsión adicional. Métodos más sofisticados disminuyen el sesgo de los datos faltantes al mismo

tiempo que minimizan la distorsión producida por la imputación. Estos métodos derivan imputaciones que usan los valores de las variables para los cuales los datos están presentes y que están relacionados con las variables imputadas.

Los ***casos con datos completos*** (observaciones que no tienen valores faltantes) sirven típicamente como material crudo para las imputaciones. Los factores que están teóricamente relacionados con las variables imputadas y con las cuales están asociadas en los casos con datos completos, son utilizados para desarrollar modelos "predictivos" para las variables imputadas. Estos modelos luego se aplican a las observaciones restantes, generando valores predichos ("imputados") para las respuestas faltantes. Las imputaciones resultantes se dice que están ***condicionadas a*** las variables en el modelo.

Por ejemplo, supongamos que los datos disponibles muestran una correlación positiva entre la presión arterial y la edad. Al condicionar las imputaciones a la edad, imputamos (en promedio) presiones arteriales mayores a los sujetos de mayor edad a los cuales les falta el dato de presión arterial y presiones arteriales menores a los sujetos de menor edad a los cuales les falta el dato de la presión arterial. Esta técnica mantiene la relación entre edad y presión arterial que existe entre los casos con datos completos. Es más, si los sujetos de mayor edad tienen mayor probabilidad de que les falte la información sobre presión arterial, el condicionamiento disminuye el sesgo que surgiría de analizar sólo los casos completos.

Si el proceso que lleva a la falta de datos es uniformemente aleatorio, salvo porel hecho de estar positivamente correlacionado con factores identificables (p.ej., la edad del sujeto), el proceso de falta de datos se llama ***faltando al azar*** (en inglés missing at random, MAR), más que "faltando totalmente al azar". En esta situación, la presión arterial global promedio para el conjunto completo de datos estará sesgado hacia valores menores (debido a la sub-representación de los sujetos de mayor edad), pero el promedio global basado en las imputaciones condicionadas a la edad no estará sesgado.

Sin embargo, si los valores predichos simplemente se substituyen con los valores faltantes, aunque el sesgo disminuirá, también lo harán los errores estándar. La razón de esto es que los modelos de imputación fueron creados basado en asociaciones (imperfectas) entre las variables condicionantes y las variables que son imputadas. Por el contrario – los valores predichos se calculan directamente a partir del modelo como si, en nuestro ejemplo, la presión arterial fuera completamente determinada por la edad. De hecho, el modelo funciona como "una profecía que se autocumple". Para evitar este problema se introduce una fuente de variabilidad al azar en el proceso de imputación. Por ejemplo, más que sustituir los propios valores predichos con los datos faltantes, los valores imputados pueden ser muestreados de distribuciones cuyas medias son los valores predichos (p.ej., si la media estimada para una respuesta que puede ser si-no fuera 0.30 [donde 1= "si" y 0= "no"], el valor imputado se generaría al azar de una distribución binomial con una proporción de "éxitos" de 0.30).

Además, al usar múltiples imputaciones (generalmente cinco), el analista puede ajustar los errores estándar para reflejar la incertidumbre introducida por el proceso de imputación. El

llevar a cabo múltiples imputaciones significa repetir el proceso de imputación para crear múltiples versiones del conjunto de datos (una para cada imputación), analizar cada conjunto de datos por separado, y combinar los resultados de acuerdo con ciertos procedimientos.

La imputación produce la menor distorsión cuando la proporción de datos faltantes es pequeña, y se consiguen los datos para variables fuertemente asociadas con la variable que es imputada. Perversamente, sin embargo, la imputación es más necesaria cuando la proporción de datos faltantes es importante. Lamentablemente, además, los datos disponibles pueden ser poco orientadores sobre si el proceso por el cual faltan los datos es totalmente aleatorio, aleatorio, o "no despreciable". Puede ser útil prestar atención a las causas de la falta de las respuestas en el proceso de recolección de datos (Heitjan, 1997).

[Me gustaría agradecer a Michael Berbaum y Ralph Folsom por sus pacientes explicaciones de imputaciones y por leer las versiones anteriores de esta sección.]

Análisis descriptivos

En algún momento la exploración de datos se convierte en un análisis descriptivo, para examinar y luego informar las medidas de frecuencia (incidencia, prevalencia) y de extensión (media, tiempo de sobrevida), asociación (diferencias y razones), e impacto (fracción atribuible, fracción de prevención). Estas medidas se calcularán para subgrupos importantes y probablemente para el total de la población de estudio. Pueden ser necesarios procedimientos de estandarización u otros de ajuste para tener en cuenta las diferencias en las distribuciones por edad y otros factores de riesgo, tiempo de seguimiento, etc.

Evaluación de hipótesis

Después del análisis descriptivo viene la evaluación de las hipótesis de estudio, si el estudio ha identificado alguna. En esta etapa se hará una evaluación más formal del potencial fenómeno de confusión, otras formas de sesgo, explicaciones alternativas posibles para lo que ha sido observado. Un aspecto que corresponde tanto al análisis descriptivo como a las pruebas de hipótesis, sobretodo a esta última, es la evaluación de la posible influencia de la variabilidad aleatoria ("azar") sobre los datos. Una gran parte de la disciplina "estadística" se ha desarrollado para tratar este aspecto, al cual nos dedicaremos a continuación.

Evaluando el papel del azar – inferencia

Creamos o no las palabras de Albert Einstein, "el Señor no juega a los dados con el universo", hay muchos eventos en el mundo que atribuimos al "azar". Cuando tiramos un dado, el número que sale habitualmente no es predecible y no sigue un patrón evidente (o por lo menos, no debería hacerlo). De la misma manera cuando sacamos cinco cartas de un mazo recién mezclado, y no marcado, sabemos que algunas cartas tienen más probabilidad de salir (p.ej., un par igual es más probable que tres cartas iguales), pero no podemos predecir que carta vamos a obtener. Las teorías de probabilidad y estadística nacieron en los salones de Monte Carlo y maduraron en los campos de la

campiña británica. La revolución de la computación puso su potencia, para bien o para lo que sea, en manos de todos los que podemos hacer clic con el mouse.

La base de la incorporación de los frutos de las teorías de probabilidad y estadística en la investigación médica y epidemiológica ha sido relatada por Austin Bradford Hill como sigue:

> "Entre las dos guerras mundiales había motivos importantes para enfatizarle a los clínicos y otros investigadores, la importancia de no dejar pasar desapercibidos los efectos del azar sobre los datos. Talvez las generalidades se basaban demasiado a menudo sobre dos hombres y un perro de laboratorio mientras que el tratamiento de elección se deducía a partir de dos pacientes y podría fácilmente no tener ningún significado. Por lo tanto, era útil que los estadísticos enfatizaran, la aplicación y la enseñanza de la necesidad de las pruebas de significancia estadística solamente para servir de guía, para tener cuidado al sacar una conclusión, antes de extrapolar lo particular a lo general." (pg 299 en El ambiente y la enfermedad: asociación o causa. Procedimientos de la Real Sociedad de Medicina,1965: 295-300. [The environment and disease: association or causation. Proceedings of the Royal Society of Medicine].)

A partir de este comienzo inocente y de sentido común, los procedimientos estadísticos prácticamente invadieron el pensamiento de los investigadores en muchos áreas. Hill continúa:

> "Me pregunto si el péndulo no se ha desplazado demasiado lejos – no sólo con los alumnos atentos sino hasta con los propios estadísticos. Por cierto, debe ser igualmente tonto negarse a llegar a conclusiones sin los errores estándar! Afortunadamente, creo que aún no hemos llegado tan lejos como nuestros amigos en EEUU, donde, me han dicho, algunos editores de revistas devuelven un trabajo porque no se han utilizado pruebas de significancia. De cualquier manera hay numerosas situaciones en que son totalmente innecesarias – porque la diferencia es grotescamente obvia, porque es insignificante, o porque, sea formalmente significativa o no, es demasiado pequeña para ser de importancia práctica. Lo que es peor, los destellos de una tabla t distraen la atención de lo inadecuado del banquete..."

El autor admite que exagera, pero sospecha que la confianza en exceso en las pruebas estadísticas debilita "nuestra capacidad para interpretar datos y tomar decisiones razonables no importa cual sea el valor de P." Hill se refiere a las pruebas de significancia, que son probablemente los procedimientos más comúnmente utilizados para evaluar el rol del azar, o talvez más precisamente, la cantidad de evidencia numérica de que las diferencias observadas no surgirían sólo por azar

Ilustración de una prueba estadística

Tomemos los siguientes datos, del primer trabajo que informó de una asociación entre el adenocarcinoma de la vagina y el uso materno de dietilbestrol (DES). Durante la década de los 60, se observó un grupo de casos de adenocarcinoma de la vagina en mujeres jóvenes, una ocurrencia altamente improbable. La investigación de las historias de las mujeres afectadas mostró que en la mayoría de los casos, la madre de la joven había tomado dietilbestrol (DES) cuando la hija estaba en su útero. En aquel momento el DES había sido indicado por la creencia de que podía prevenir el parto prematuro en mujeres que habían perdido embarazos anteriores. ¿En cuántas pacientes tendría

que ocurrir esta historia para que los investigadores tuvieran confianza en que no era una observación al azar? Esta pregunta habitualmente se contesta por medio de una prueba estadística.

Exposición prenatal al dietilbestrol
entre mujeres jóvenes con adenocarcinoma de la vagina

	Exposición a dietilbestrol?		
	Si	No	Total
Casos	7	1	8
Controles	0	32	32
Total	8	33	40

Fuente: Herbst AL, Ulfelder H, Poskanzer DC. Adenocarcinoma of the vagina. Association of maternal stilbestrol therapy with tumor appearance in young women. *New Engl J Med* 1971; 284:878-881. [From Schlesselman JJ. *Case-Control Studies*. New York, Oxford, l982: 54]

Todos menos uno de los casos tenían el antecedente de exposición intrauterina a dietilbestrol. Por el contrario, ninguno de los controles lo tenía. El riesgo relativo a partir de esta tabla no puede ser calculado directamente por la celda que contiene 0, pero si se agrega 0.5 a las cuatro celdas obtenemos un riesgo relativo (OR) de 325, una asociación más fuerte que la que cualquiera de nosotros puede esperar encontrar en nuestros datos alguna vez en la vida. Sin embargo, este estudio tiene sólo 8 casos. ¿Pueden deberse estos resultados al azar?

Una prueba de significancia estadística es un instrumento para evaluar la cantidad de datos numéricos sobre la cual se basa un patrón observado, para contestar preguntas como, "¿Con qué frecuencia puede surgir una asociación tan fuerte, completamente por azar, en un número infinito de experimentos análogos con el mismo número de sujetos y la misma proporción de casos (o de expuestos)? "Esta pregunta no es idéntica a: "¿qué probabilidad hay de que el azar produjo la asociación en esta tabla?" ni a "¿Cuánto de la asociación se debe al azar?". Pero si una asociación tan fuerte surgiese sólo muy raramente debido exclusivamente al azar, es razonable suponer que por lo menos algún factor potencialmente identificable ha contribuido a la asociación observada. Este factor podría, por cierto, ser un sesgo, más que la exposición, pero por lo menos sería algo distinto al azar. A la inversa, también es posible que asociaciones mucho más fuertes podrían surgir por azar y la que hemos observado puede reflejar un proceso causal. La prueba de significancia simplemente evalúa la fuerza de la evidencia numérica para desechar el azar como una probable explicación suficiente.

Para llevar a cabo una prueba de significancia, necesitamos operacionalizar el concepto de "experimento análogo". Ese es el problema. ¿Qué tipo de experimento es análogo a un estudio epidemiológico, es más, análogo a un estudio observacional? Para la tabla anterior, la prueba de significancia que se usaría sería la Prueba Exacta de Fisher. Aquí, el experimento análogo (***modelo de probabilidad***) es equivalente a lo siguiente:

Supongamos que tú tienes 40 pares de medias – 7 pares de medias rojas y 33 pares de medias azules. Quieres empacar 8 pares de medias en tu valija, de manera que sin mirar tomas 8 pares al azar y las pones en tu bolso. ¿Cuántos pares rojos has empacado para tu viaje?

Cuando este "experimento análogo" se repite un numero suficiente de veces, la proporción de veces en que el bolso tiene 7 pares rojos nos dará la probabilidad de que el azar por si sólo produciría la situación en que hayas empacado 7 pares de medias rojas. Esta probabilidad es el "valor p" de la prueba de significancia de la relación entre el adenocarcinoma de la vagina y el diestilbestrol de la tabla anterior.

Afortunadamente, la distribución del número de pares rojos en la valija ya ha sido desarrollada en forma teórica, de manera que la probabilidad exacta puede ser calculada sin tener que llevar a cabo lo que en este caso sería un número MUY importante de ensayos. La fórmula de la distribución (hipergeométrica) es:

$$\Pr(A=j) = \frac{C_{j}^{n_1} \; C_{(m_1-j)}^{n_0}}{C_{m_1}^{n}} = \frac{n_1! n_0! m_1! m_0!}{n!\, j!\, (n_1-j)!\, (m_1-j)!\, (n_0-m_1-j)!}$$

Donde Pr(A=j) es la probabilidad de obtener j pares rojos en la valija y m_0, m_1, n_0, n_1, y n son los totales de las filas y las columnas de la tabla:

	Color		
	Rojo	Azul	Total
Valija	j	$m_1 - j$	m_1
En cajón	$n_1 - j$	$n_0 - m_1 - j$	m_0
Total	n_1	n_0	n

Así es como se aplica la fórmula:

	Rojo(DES)	Azul	Total
Empacados(casos)	7	1	8
En cajón (controles)	0	32	32
Total	8	33	40

Posibles resultados (Colores de los pares de medias en la valija)		Probabilidad de cada resultado	
Rojo	Azul		
0	8	.181	
1	7	.389	
2	6	.302	
3	5	.108	
4	4	.019	
5	3	.0015	$\frac{7!\ 33!\ 8!\ 32!}{40!\ 5!\ 2!\ 3!\ 30!}$
6	2	.00005	
7	1	4.3 x 10-7	Valor-p
8	0	0	
		1.0000	

Comentarios sobre el modelo de "las medias rojas"

1. Un modelo es un sistema o estructura que tiene como objetivo represenatr las características esenciales de la estructura o sistema que es objeto de estudio. El modelo presentado anteriormente es una representación muy simplificada!

2. El modelo es derivado en base a ciertas constricciones o supuestos (p.ej., en este caso, 8 casos, 7 madres expuestas a DES, y 40 participantes en total – "marginales fijos" – además del hecho de que "todas las permutaciones tienen la misma probabilidad").

3. El modelo subyacente a la prueba de hipótesis supone un experimento repetible y una especificación *a priori* de la "hipótesis" sometida a prueba – una hipótesis "nula" [esto está incorporado en el modelo de permutaciones con "iguales probabilidades"] y una "hipótesis alternativa" [esto trata con los resultados que consideraríamos como inconsistentes con la hipótesis nula].

4. El modelo anterior es tedioso de calcular para tablas grandes, aunque las computadoras han resuelto ese problema.

El Concepto de la prueba de hipótesis (pruebas de significancia)

Lo que realmente queremos saber es: "¿Se debe al azar la asociación observada?", o "¿Qué tan probable es que la asociación observada se deba al azar?". Esta probabilidad es conocida a veces como la "probabilidad posterior [*a posteriori*]", la probabilidad de que la hipótesis es verdadera dados los resultados observados. (La "probabilidad previa [*a priori*]" de que la hipótesis es verdadera es

nuestra creencia de que la hipótesis es verdadera antes de tener los resultados). La escuela frecuentista de estadística, de la cual provienen las pruebas de significancia, no puede contestar esta pregunta directamente. En vez, las pruebas de significancia y los valores p intentan dar una respuesta indirecta, reformulando la pregunta como: " ¿Con qué frecuencia se vería una asociación tan fuerte como la observada sólo por azar?". El rol del azar es llevado a cabo por un modelo adecuado de probabilidad, seleccionado para representar la estructura de probabilidad de los datos y el diseño de estudio. Pero la mayor parte de los estudios epidemiológicos se desvían marcadamente de los modelos probabilísticos sobre los cuales se basan las pruebas estadísticas (p.ej., ver Sander Greenland, Aleatorización, estadística, e inferencia causal [Randomization, statistics, and causal inference]), de manera que aunque la teoría estadística es extremadamente precisa, debe ser aplicada e interpretada con mucho cuidado.

Una versión intermedia de la pregunta que subyace una prueba de significancia es "¿Qué tan consistentes son los datos numéricos con lo que se esperaría "por azar" – según un modelo de probabilidad adecuado". El modelo de probabilidad es frecuentemente uno que supone que no hay diferencia sistemática entre los grupos, en parte porque dichos modelos son más fáciles de derivar y también porque es a menudo conveniente para el marco de la prueba de hipótesis. El resultado de una prueba de significancia es una probabilidad (el ***valor p***) que da una respuesta cuantitativa a esta pregunta intermedia. (Nota: La "hipótesis nula" estadística es pocas veces de interés desde el punto de vista sustancial. Una hipótesis de estudio debe ser planteada en términos de ausencia de asociación sólo cuando es lo que el investigador realmente desea demostrar. De hecho, es bastante difícil demostrar la ausencia de asociación, dado que la evidencia para la ausencia de asociación está relacionada con la probabilidad de error de tipo II (1 – potencia estadística) para el estudio, que es en general considerablemente mayor que el nivel de significancia – ver más adelante).

El valor p por si mismo puede ser considerado como un estadístico descriptivo, un trozo de evidencia que tiene que ver con la cantidad de evidencia numérica para la asociación en estudio. Sin embargo, cuando se necesita tomar una decisión se necesita algún método para asignar una acción al resultado de la prueba de significancia. La toma de decisiones incluye el riesgo de cometer errores. En forma ideal la función de pérdida (los costos de los errores de diverso tipo) se conocen explícitamente. Bajo supuestos ampliamente aplicables, la teoría de la toma de decisiones provee de una técnica para la toma de decisiones basándose en los resultados de la prueba estadística. Esa técnica es la realización de una prueba de hipótesis estadística.

Como se ha señalado, la hipótesis que se prueba es generalmente una "hipótesis nula" (habitualmente indicada como H_0). H_0 es el modelo de probabilidad que hará el rol del azar (por ejemplo, el modelo de las medias rojas). En el contexto actual, ese modelo se basará en la premisa de que no hay asociación. Si hay suficiente evidencia numérica que nos lleve a rechazar la H_0, decidiremos que lo contrario es verdadero, que hay una asociación. La inversa es llamada la "hipótesis alternativa" (H_A). La regla de toma de decision es de rechazar la H_0, a favor de la H_A, si el valor de p es suficientemente pequeño, y sino, aceptar H_0 .

Dado que debemos tomar una decisión entre dos alternativas (H_0 y H_A) podemos cometer dos tipos de errores:

Error Tipo I : Rechazar erroneamente H_0 (i.e., concluir, incorrectamente, que los datos no son consistentes con el modelo)

Error Tipo II No rechazar erróneamente H_0 (i.e., concluir, incorrectamente, que los datos son consistentes con el modelo)

(El creador de estos términos debe haber sido más prosaico que el que creó los términos "significancia", "potencia", "precisión", y "eficiencia"). Tradicionalmente, la probabilidad de error Tipo I ha recibido más atención y se denomina el "***nivel de significancia***" de la prueba.

En un contexto estricto de toma de decisiones, el resultado de la prueba de significancia es "Rechazar la hipótisis nula" o "No rechazar la hipótesis nula". (Señalemos que el "no rechazar la hipótesis nula" no es equivalente a declarar que la hipótesis nula es verdadera.) Sin embargo, muy raramente debe tomarse una decisión basada en un único estudio, de manera que es preferible informar el valor p calculado (probabilidad de que el modelo de probabilidad supuesto produciría datos tan o más extremos que estos). El valor p da más información que la aseveración "los resultados fueron significativos a nivel del 5%", dado que cuantifica el grado al cual los datos son incompatibles con el "azar" (según el modelo probabilístico), permitiendo que el lector ejerza su tolerancia para un error de Tipo 1. Señalemos que el valor p no es un indicador directo de la fuerza de una asociación en el sentido epidemiológico ni de su "significancia" biológica, clínica o epidemiológica. El valor p simplemente evalúa la compatibilidad de los datos observados con el modelo probabilístico supuesto que sirve para representar la H_0.

Hay muchos métodos para obtener un valor p o llevar a cabo una prueba de significancia estadística. La selección depende del nivel de medición de las variables (dicotómica, politómica nominal, ordinal, contínua), el diseño de muestreo del cual se obtuvieron los datos, y otros factores. La prueba estadística ilustrada anteriormente es una prueba "exacta" (Prueba exacta de Fisher), dado que se basa en un modelo que considera todos los posibles resultados y de cuantas maneras puede ocurrir cada una. En una prueba exacta, el modelo probabilístico es claramente aparente.

Ilustración de una prueba asintótica

Las *pruebas asintóticas* son más habitualmente usadas, porque son más sencillas de calcular, (p.ej., Chi cuadrada, prueba t). Las pruebas asintóticas son aproximaciones cuya precisión mejora a medida que aumenta el tamaño muestral y en que los modelos probabilísticos subyacentes tienden a ser más abstractos. En forma típica, las pruebas asintóticas se basan en la distribución "normal" (de Gauss). ¿Por qué la distribución de Gauss? Porque ofrece una serie de ventajas analíticas y, sobre todo, por el Teorema del Límite Central ("uno de los teoremas más sorprendentes de todas las matemáticas", Mood y Graybill, 1963:149). El Teorema del Límite Central mantiene que si tomamos muestras al azar suficientemente grandes de cualquier distribución con una varianza finita, los promedios de esas muestran tendrán una distribución aproximadamente Gaussiana.

La forma general de una prueba así es (ver Rothman, *Modern epidemiology,* p. 139 o Kleinbaum, Kupper, and Morgenstern, *Epidemiologic research*):

$$Z = \frac{a - E(a)}{\sqrt{var(a)}}$$

Donde "a"es el valor observado (p.ej., el número de casos expuestos), E(a) es el valor esperado para "a" bajo la hipótesis nula (también conocido como experimento análogo) y var(a) es la varianza de "a" bajo la hipótesis nula. Por lo tanto, Z es el número de desviaciones estándares por las cuales "a" difiere de lo que se esperaría si no hubiera asociación y tiene una distribución aproximadamente normal. (Z se escribe a veces como χ , llamada "chi", una unidad de distribución normal que es igual a la raíz cuadrada de una distribución chi cuadrada con un grado de libertad).

La probabilidad asociada con el hecho de estar a "Z" desvíos estándares del promedio de una distribución normal puede ser calculada y se obtiene fácilmente en las tablas estadísticas (ver el extracto de tabla más adelante). El valor de una variable aleatoria distribuida normalmente es habitualmente (i.e. una probabilidad de 95%) menor a dos desvíos estándares de su promedio, de manera que si Z es mayor que 1.96 decimos que "$p < .05$", o con mayor precisión, tomamos el valor que hemos calculado para Z, lo buscamos en la tabla de la distribución normal y tomamos el valor correspondiente de p.

El extracto de la tabla más adelante muestra varias probabilidades derivadas de la unidad de la distribución normal. Por ejemplo, la probabilidad asociada con una distancia de 1.645 desvíos estándares por encima del promedio se puede ver en la columna B (0.05) y es idéntica a la probabilidad asociada con una distancia de 1.645 desvíos estándares por debajo del promedio (dado que la distribución normal es simétrica). La probabilidad asociada con la obtención de un valor de z que está por encima o por debajo de 1.645 desvíos estándares del promedio se ve en la columna d (0.10). De manera que si usando la fórmula planteada anteriormente (o una de las presentadas más adelante) obtenemos un valor de Z igual a 1.645, el valor p es 0.05 o 0.10, dependiendo de la hipótesis alternativa.

Extracto de una tabla de la Distribución Normal

z	h	A	B	C	D	E
0.00	0.3989	0.0000	0.5000	0.0000	1.0000	0.5000
0.01	0.3989	0.0040	0.4960	0.0080	0.9920	0.5040
0.02	0.3989	0.0080	0.4920	0.0160	0.9840	0.5080
. . .	. . .	. . .	. . .	. . .	. . .	. . .
0.8416	0.2800	0.30	0.20	0.60	0.40	0.80
. . .	. . .	. . .	. . .	. . .	. . .	. . .
1.282	0.1755	0.40	0.10	0.80	0.20	0.90
. . .	. . .	. . .	. . .	. . .	. . .	. . .
1.645	0.1031	0.45	0.05	0.90	0.10	0.95
. . .	. . .	. . .	. . .	. . .	. . .	. . .
1.960	0.0585	0.475	0.025	0.95	0.05	0.975
. . .	. . .	. . .	. . .	. . .	. . .	. . .
2.576	0.0145	0.495	0.005	0.99	0.01	0.995
. . .	. . .	. . .	. . .	. . .	. . .	. . .
3.090	0.0034	0.499	0.001	0.998	0.002	0.999
. . .	. . .	. . .	. . .	. . .	. . .	. . .

Leyenda:

z =número de desvíos estándares a la derecha del promedio

h = altura de la curva para ese número de desvíos estándares desde el promedio

A =área entre el promedio y z

B = área a la derecha de z (o a la izquierda de –z)

C = área entre -z y +z

D = área más allá de|z| (i.e., a la izquierda de -z y a la derecha de +z)

E = área a la izquierda de z

(Fuente: National Bureau of Standards – Applied Mathematics Series–23, U.S. Government Printing Office, Washington, D.C., 1953, extracto de la Tabla A-4 en Richard D. Remington y M. Anthony Schork, *Statistics with applications to the biological and health sciences*. Englewood Cliffs, NY, 1970.]

Valores p de una cola vs dos colas

Recordemos que el valor p es la probabilidad de obtener una asociación tan fuerte como (o más fuerte que) la asociación observada. Sin embargo, resulta que la expresión "tan fuerte como (o más fuerte que)" es ambigua, porque no especifica si están o no incluidas las asociaciones inversas, i.e., asociaciones en el sentido opuesto a la asociación putativa que motivó el estudio. Por ejemplo, si

observamos un riesgo relativo de 2.5, ¿"tan fuerte como" significa sólo riesgos relativos de 2.5 o más, o también significa riesgos relativos de 0.4 o menos? Si es lo primero (sólo 2.5 y más), el valor p es el que corresponde a una cola. Por el contrario, si H_A es "sea mayor que o igual a 2.5 o [inclusive] menos que o igual a 0.4", está indicado usar un valor p para dos colas. [Sólo los valores p de una cola pueden ser interpretados como la "probabilidad de observar una asociación tan fuerte o más fuerte bajo el modelo del azar" (Rothman and Greenland,185).]

El tema de los valores p de una cola versus valores p de dos colas puede producir emociones muy fuertes. Para un valor calculado de Z, un valor p de una cola es exactamente la mitad del valor p para dos colas. Los que apoyan los valores p de dos colas argumentan que los valores p de una cola dan una medida inflada de la significancia estadística de una asociación (baja probabilidad de obtener los resultados por azar). Las situaciones apropiadas para usar valores p de una cola a veces se caracterizan por ser aquellas en que el investigador no tiene interés en encontrar una asociación en el sentido contrario y la ignoraría aún si ocurriera. Sin embargo, un mensaje en la lista EPIDEMIOL-L solicitando situaciones como las descritas produjo muy pocos ejemplos convincentes.

A continuación hacemos una presentación dramatizada de algunos de los temas que influyen en la selección de valores p de una o dos colas:

> La esposa de un buen amigo ha muerto trágicamente por cáncer de pulmón. Aunque ella nunca fumó en su vida, tu amigo era un gran fumador. Antes de su muerte, ella se había convertido en una activista anti-tabaquismo, y su último deseo fue que tu amigo le hiciera juicio a R. J. Morris Inc, el fabricante de la marca de cigarrillos que tu amigo fumaba. Sabiendo que no puede pagar un asesoramiento por expertos, tu amigo te pide que lo asistas con el juicio.
>
> En la preparación para el juicio, la jueza revisa los estándares de evidencia con todos los participantes. Ella les explica que para que la corte falle a favor del demandante (tu lado) debe concluir que la asociación es apoyada por una "preponderancia de evidencia", que ella caracteriza como "equivalente a 90% de probabilidad de que los cigarrillos de R. J. Morris causaron la enfermedad". El abogado de R.J. Morris presenta objeciones, declarando que, en primer lugar, sólo la probabilidad de que los cigarrillos pueden causar la enfermedad puede ser estimada, y no la probabilidad de que los cigarrillos efectivamente causaron la enfermedad. En el momento en que la jueza está por decir que la interpretación jurídica de probabilidad permite dicha conclusión, el abogado de R.J. Morris plantea su segunda objeción: dado que el demandante está basando su caso en evidencia científica, el caso del demandante debe cumplir con el estándar convencional para la evidencia en ciencias que requiere un nivel de significancia de 5%. [Recuerda que el nivel de significancia es la probabilidad de un error de Tipo I, que en este caso significaría que se encontraría que la compañía es responsable aunque el cáncer de pulmón de la mujer de tu amigo en realidad se debió al azar. Si la corte no encontrara responsable a la compañía, aunque los cigarrillos de la compañía sí causaron el cáncer, eso sería un error de Tipo II.]
>
> Viendo la oportunidad, le pasas una esquela a tu amigo, que se la pasa a su vez a su abogado. Al leerla, el abogado le dice al juez "Su Señoría, mi cliente está de acuerdo con aceptar la insistencia de R. J.Morris sobre el nivel de significancia del 5%, siempre y cuando se base en una hipótesis alternativa de una sola cola". Empezando a lamentar la introducción de la metáfora de

probabilidad, la jueza se dirige al abogado de R. J. Morris, que conversa agitadamente con su bioestadístico. Luego de una rápida consulta el abogado de R. J.Morris acusa indignado al abogado del demandante de intentar, a través del engaño, de obtener menores niveles de evidencia. Acusa que un nivel de significancia de una cola de 5% es en realidad un nivel de significancia de 10%, dado que todo el mundo sabe que las pruebas de dos colas son más apropiadas. El abogado de tu amigo presiente que esta acusación pesará en la opinión de la jueza y busca tu mirada para que le aconsejes como contestar.

Con tu asesoramiento, el abogado de tu amigo responde que una prueba de dos colas está justificada sólo cuando la hipótesis alternativa apropiada (H_A) es de dos colas. La pregunta en este caso es si R.J. Morris es o no responsable, i.e., si sus cigarrillos causaron o no el cáncer. Esta pregunta corresponde a una (H_A) de una cola, i.e., la corte puede (1) rechazar la (H_0) (no hay causa) a favor de la alternativa de que R.J. Morris es responsable o (2) no rechazar la (H_0), si la corte encuentra que la evidencia es insuficiente. "Con su permiso, Señoría" continúa el abogado, " no hay ningún planteo aquí de que el humo de cigarrillo podría haber actuado para prevenir la ocurrencia del cáncer, de manera que el requerimiento de una hipótesis alternativa de dos colas es equivalente a imponer un nivel de significancia de 2.5%, que se acerca más al nivel de un juicio criminal, más que de un juicio civil."

Con el beneficio de consultas adicionales, el abogado de R.J. Morris "objeta enérgicamente". "El demandante puede considerar este caso como de una H_A de una cola, pero no importando el acuerdo sobre tabaquismo propuesto, la Compañía R. J. Morris está preocupada por el hecho de que la relación entre el hábito de fumar y el cáncer aún no ha sido demostrada. Por lo tanto, un hallazgo de que el hábito de fumar puede de hecho prevenir el cáncer es tan relevante como el planteo del demandante de que los cigarrillos fueron responsables."

Naturalmente te sientes indignado por la aseveración del abogado de R.J. Morris de que la relación entre el fumar y el cáncer no está probado, pero tienes que dejar eso de lado cuando el abogado de tu amigo te pregunta si no es correcto que el nivel de significancia es simplemente un mecanismo para decidir cuantos desvíos estándar desde el promedio son necesarios para excluir el azar como explicación. Habitualmente, las personas excluyen el azar cuando la prueba estadística resulta en dos desvíos estándar desde el centro de una distribución normal (en realidad 1.96 desvíos estándar, que corresponde a un nivel de significancia de 5% de dos colas). Si la jueza acepta el nivel de significancia de 5% de una cola, aún con un buen argumento de que porque la H_A apropiada es de una cola de manera que la probabilidad de error Tipo I es realmente sólo de 5%, una decisión que cumple la prueba estando a 1.65 desvíos estándar del promedio, (correspondiendo a un nivel de significancia de una cola de 5%) puede ser vulnerable en una apelación. Dado que la evidencia científica es sólida, ¿sería mejor estar de acuerdo con una prueba de dos colas?

La jueza mira su reloj, y ves gotas de transpiración en la frente del abogado de tu amigo. Mientras tanto tratas de aclarar los temas que surgieron. Acabas de recibir tu título de Epidemiólogo, y no estás muy seguro aún cómo funciona. Es verdad que la corte de apelación puede rechazar la idea de una prueba de una cola, dado que los jueces de apelación suelen ser conservadores, y es seguro que R.J. Morris apelará un juicio negativo para ellos. Pero entonces te viene una idea nefasta a la mente. ¿Y si R.J. Morris ha inventado evidencia que hace parecer de

alguna manera que tu amigo es responsable de la muerte de su esposa por cáncer de pulmón? Sabes que esto es una locura, ¿pero y si pudieran hacerlo? Con una de dos colas, la corte podría rechazarla y hallar a tu amigo responsable, destruyéndolo financiera y emocionalmente. "De una cola!", gritas.... y de golpe te despiertas sobresaltado. El profesor y tus colegas estudiantes te están mirando con perplejidad, preguntándose qué pregunta creías estar contestando. A medida que sales del ensueño esperas no haberte perdido demasiado de la clase y juras acostarte más temprano en el futuro.

Pruebas de significancia en una tabla dos por dos

Para una tabla dos por dos, la fórmula puede ser expresada más fácilmente para el cálculo definiendo a "a" como el contenido de una única celda de la tabla, por convención la celda "a" (la de arriba a la izquierda) de manera que E(a)es el valor esperado para "a" bajo la hipótesis nula (n_1m_1/n), y Var(a) es la varianza de "a" bajo la hipótesis nula $\{(n_1n_0m_1m_0)/[n^2(n-1)]$, basada en la distribución hipergeométrica. Entonces el estadístico de prueba es simplemente:

$$Z = \frac{a - n_1m_1/n}{\sqrt{\{ (n_1\, n_0m_1\, m_0) \,/[\, n^2\, (n-1)]\}}}$$

Una fórmula equivalente pero más fácil de recordar, es:

$$Z = \sqrt{X^2} = \sqrt{\frac{(ad - bc)^2\,(n-1)}{n_1\, n_0m_1\, m_0}}$$

[Nota: puedes ver la fórmula anterior con n, en vez de (n-1) [p.ej., Hennekins y Buring, p. 251 usa T en vez de (n-1)]. La razón de esto es que la fórmula produce un estadístico Chi Cuadrado de Mantel y Haenszel (basado en la distribución hipergeométrica) en vez del estadístico Chi Cuadrado de Pearson (basado en la distribución normal). Para muestras grandes los dos son esencialmente equivalentes. Hay fórmulas paralelas para datos en persona-tiempo.]

	Expuesto a dietilbestrol?		
	Si	No	Total
Casos	a	b	m_1
Controles	c	d	m_0
Total	n_1	n_0	n

No importa cuanta desconfianza le tengamos al modelo estadístico y su aplicación, los resultados con valor p tan pequeño como el obtenido en este estudio, le producirían satisfacción a cualquier investigador que los obtuviese. Pero para apreciar la dinámica del procedimiento, y los problemas de

interpretación que surgen en las circunstancias que se prestan a más dificultad, analicemos lo que subyace un valor p pequeño.

Un valor p pequeño (i.e., una baja probabilidad de que resultados similares a los observados sean producidos por el "azar" [simulado por un modelo estadístico dado]) refleja:

- Una fuerte asociacion observada (o una diferencia observada grande)

o

- Un tamaño muestral grande (hablando en forma general).

Por lo tanto, cuando el valor p no es pequeño, hay dos posibilidades (ignorando las posibilidades del error sistemático, modelo estadístico no adecuado, etc.):

1. La asociación o diferencia observada no es fuerte.
2. La asociación observada es de magnitud respetable pero el tamaño del estudio es demasiado pequeño para considerarlo "significante".

La interpretación de las circunstancias en que no se obtiene un valor p pequeño depende de nuestra interpretación de la magnitud de la asociación observada y de la potencia estadística del estudio para detectar una diferencia verdadera importante.

Si el valor p es pequeño (p.ej., menor al (típico) cinco por ciento, diez por ciento [menos común], o uno por ciento [para los más exigentes o que tienen muchos datos]), los resultados observados son algo inconsistentes con una explicación basada sólo en el azar, de manera que nos inclinamos a considerarlos debidos a algún factor que vale la pena investigar (p.ej., influencias sistemáticas por la manera en que se diseñó o llevó a cabo el estudio, procesos biológicos o sicosociales relacionados a los factores en estudio, etc). Si la diferencia o asociación observada es demasiado pequeña para ser científica o clínicamente significativa (en oposición a estadísticamente significativa), no nos interesará seguir el análisis no importa cual sea el valor de p.

Si el valor p no es pequeño (i.e., los resultados "no son significativos"), ¿se observó una asociación? Si no se observó una asociación, la caracterización apropiada del hallazgo es de "no se observó asociación" (pero, ver más adelante). Si se observa una asociación, podemos decir que "se observó una asociación pero los datos eran insuficientes para descartar el azar como explicación" [no, "no había asociación"!]

Si no se observa asociación, necesitamos preguntarnos además, cuáles eran nuestras posibilidades de detectar una asociación significativa si una existiese. Si la potencia estadística era baja, entonces no podemos decir mucho. Si la potencia estadística era alta, podemos decir que los datos dan evidencia (suponiendo, siempre, que no hay sesgo) en contra de la existencia de una asociación fuerte.

Si la asociación observada es suficientemente fuerte para ser importante si no se debe al azar, la única conclusión a la que podemos llegar es que los datos no proveen suficiente evidencia para

descartar una explicación de debido sólo al azar – lo cual no es equivalente a una conclusión de que "no se observó una asociación" [dado que sí se observó una] o que "la asociación observada se debe al azar" [que nadie sabe si es así]. Otras caracterizaciones a menudo utilizadas también son desafortunadas:

" La asociación observada no es significante" [lo cual tiende a impugnarla]

" La asociación no llegó a la significancia estadística" [que implica que la asociación debería haber sido más fuerte – puede ser tan fuerte como debe ser pero basado en demasiado pocos sujetos.]

Es mejor decir "se observó una asociación de ____, pero los datos eran demasiado escasos para descartar una explicación basada en el azar" o algo similar. [Nota: Cualquier resultado puede volverse "no significativo" si estratificamos lo suficiente.]

Una posibilidad alternativa es que la asociación observada era demasiado débil para ser significativa aún si se hubiera asociado a un valor p pequeño. En este caso nuestra conclusión dependería del tamaño del estudio, i.e., su potencia estadística para detectar una asociación de una magnitud particular. Si la potencia era baja, si la capacidad del estudio para detectar una diferencia que consideraríamos importante es baja, entonces no hay mucho que podamos decir o concluir, salvo que nuestro fracaso en encontrar una asociación podría bien ser debido al azar (i.e., podríamos haber cometido un "error de Tipo II"). Esta incapacidad es una de las razones para desaconsejar a los investigadores a emprender estudios pequeños, salvo que sea como estudio piloto para desarrollar procedimientos e instrumentos. Si la potencia era alta, entonces estamos en mejor posición para interpretar nuestros resultados como evidencia contra la existencia de una verdadera asociación.

Potencia estadística y tamaño muestral

La potencia estadística se refiere a la capacidad de detectar una asociación de interés en el marco de un error de muestreo. Supongamos que hay una verdadera asociación de cierta magnitud y grado, pero por culpa del azar nuestros estudios observarán la asociación como más débil o más fuerte. Para estar razonablemente seguros que nuestro estudio detectará la asociación, el estudio tiene que ser suficientemente grande para que el error de muestreo sea controlado.

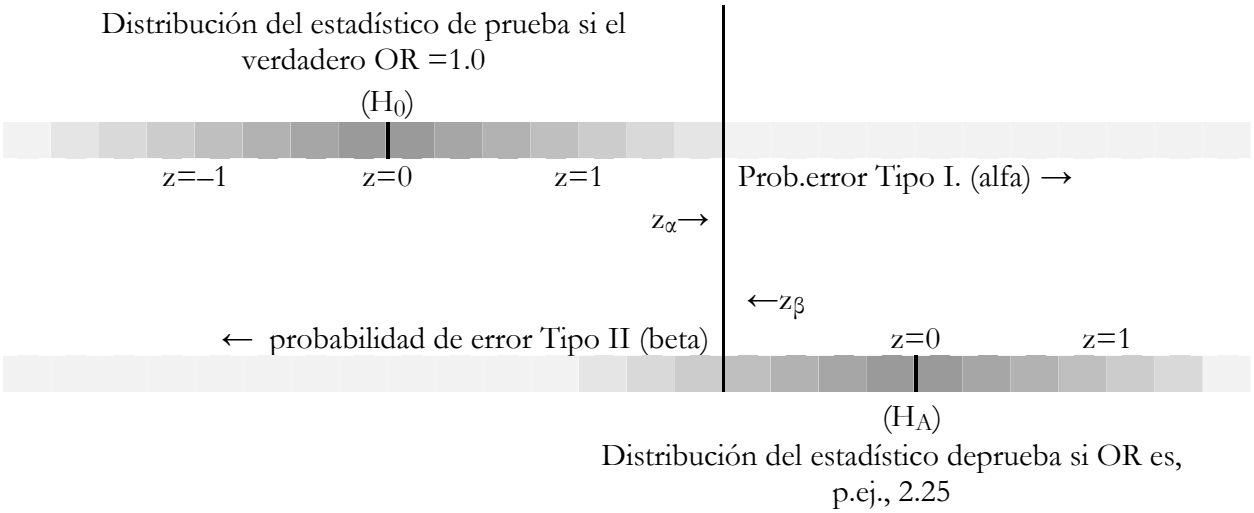
Por ejemplo, supongamos que estamos comparando un grupo de casos de pacientes con enfermedad de Alzheimer con un grupo control para ver si los casos son diferentes con respecto a la presencia de un gen específico. Supongamos también que este gen está en realidad presente en 20% de los casos y en 10% de la población de la cual surgieron los casos (i.e., el OR en un gran estudio caso control no sesgado sería de 2.25). Si estudiamos 20 casos y 10 controles, podríamos encontrar 4 casos con el gen y dos controles con el gen, de manera de estimar correctamente la prevalencia del gen en los casos y en la población y el OR.

Con tan pocos participantes, podríamos sólo obtener 3 casos con el gen y 3 controles con el gen, no detectando la diferencia en la prevalencia (OR = 1.0). De hecho podríamos tener 4 controles con el gen y sólo 2 casos con el gen de manera que pareciera que el gen es protector (OR = 0.44). Por supuesto, no queremos reaccionar a una diferencia o un OR que podría deberse al azar, de manera

que realizaríamos una prueba a cualquier resultado que observemos para asegurarnos de que es mayor del que se esperaría que ocurriera sólo por azar (i.e., "significativo"). Esto significa que descartaríamos cualquier asociación que observemos si es menor de lo que consideramos dentro de lo esperado por azar. (O recordando nuestra fantasía de la corte, una "preponderancia de la evidencia", no solamente una sospecha.)

Por lo tanto, para detectar una asociación, debemos (1) observarla en nuestro estudio y (2) decidir que es poco probable que el azar la hubiera creado. Cada uno de estos requerimientos tiene exigencias sobre el tamaño del estudio. Necesitamos por lo menos un número mínimo de sujetos de manera que (1) tengamos una expectativa razonable de observar una asociación si es que alguna existe (i.e., no cometer un error Tipo II), y (2) creamos poco probable que el azar produzca una asociación de esa magnitud.

Potencia estadística para detectar un OR ≠ 1.0 con una prueba de significancia de una cola



Este diagrama ilustra la superposición entre los sectores centrales de las distribuciones de los estadísticos de prueba (p.ej., Z) esperadas bajo la hipótesis nula (p.ej., verdadero OR es 1.0) y la hipótesis alternativas (p.ej., verdadero OR es 2.25). Cuando obtenemos los resultados del estudio calcularemos el estadístico de prueba (p.ej., Z) y lo compararemos con su distribución bajo la H_0 (la distribución superior de las dos del diagrama). Si el valor calculado de Z es menor que el z_α, i.e., cae a la izquierda del punto de corte que hemos determinado (definido por la probabilidad de error Tipo I, alfa), concluiremos entonces que los datos que observamos vinieron de la distribución superior (la de no asociación, verdadero OR =1.0). Aún si el OR que observamos fuera mayor que 1.0 (que implica que Z es mayor de 0), dado que Z no fue mayor que nuestro punto de corte, consideramos el OR observado como una desviación al azar a partir del 1.0. Si la verdad desconocida es que realmente no hay asociación, nuestra conclusión sería correcta. Si en vez el verdadero OR es realmente 2.25, y los datos que observamos en realidad provienen de la distribución inferior, nuestra conclusión representa un error de Tipo II. El área a la izquierda del punto de corte en la distribución

inferior representa la probabilidad de cometer un error de Tipo II, "beta". La potencia estadística – la probabilidad de detectar una verdadera diferencia- es igual a uno menos beta (i.e.,1 – beta).

A la inversa si observamos un valor de Z a la derecha del punto de corte, concluiremos que los datos que hemos observado no provienen de la distribución superior y que por lo tanto el verdadero OR es mayor que 1.0. Si nos equivocamos – si la asociación que observamos era en realidad un hallazgo casual – nuestra conclusión representa un error de Tipo I. El área a la derecha del punto de corte en la distribución superior representa la probabilidad de cometer un error Tipo I, "alfa".

Si nos horroriza cometer un error Tipo I, podemos correr el punto de corte a la derecha, lo cual reduce alfa – pero aumenta beta. Si preferimos disminuir beta, podemos correr el punto de corte hacia la izquierda – pero eso aumenta alfa. Lo que realmente querríamos hacer es disminuir tanto alfa como beta, haciendo que las distribuciones sean más estrechas (de manera que más del área sombrada se ubica en el centro de cada distribución, simbolizando una mayor precisión de la estimación). El ancho de la distribución es controlado por el tamaño muestral. Con una luz potente podemos distinguir fácilmente por ejemplo, entre una víbora y un palo. Pero con una luz débil, no podemos estar seguros de lo que estamos viendo. Podemos elegir errar en un sentido o el otro, pero la única forma de disminuir nuestra posibilidad de error es obtener una luz más potente.

Los valores habitualmente usados para alfa y beta son, respectivamente, 0.05 y 0.20 (potencia = 0.80), para una probabilidad total de error de 0.25. Si el tamaño del estudio es limitado por la baja incidencia de la enfermedad, la baja prevalencia de la exposición o una limitación en el presupuesto, nuestras estimaciones del estudio serán poco precisas – las distribuciones en el diagrama anterior serán anchas. La probabilidad total de error estará por debajo de 0.25 sólo cuando la distribución se encuentre más a la derecha, i.e., cuando corresponde a una asociación más fuerte.

En esencia, la intolerancia para el error (i.e., alfa y beta pequeños) y el deseo de detectar asociaciones débiles debe pagarse con el tamaño muestral. En nuestro sueño de la corte judicial, cuanto más posibilidad queremos de ganar el caso contra R.J. Morris (nuestra potencia) y/o cuanto más puede R.J. Morris convencer al Juez que aumente el estándar de evidencia (nivel de significancia), mayor el precio que tendremos que pagar para nuestra representación legal (más sujetos de estudio). El Apéndice contiene un sector que traduce estos conceptos en estimaciones de tamaños muestrales.

Sesgo de los estudios pequeños

En términos amplios, los estudios grandes son potentes, los estudios pequeños son débiles. El concepto de "sesgo de los estudios pequeños" ilustra la importancia de comprender la potencia estadística cuando se interpretan investigaciones epidemiológicas.

La idea detrás del sesgo de los estudios pequeños (Richard Peto, Malcolm Pike, y cols., *Br J Cancer* 34:585-612, 1976) es que dado que los estudios pequeños son más fáciles de llevar a cabo que los grandes, muchos más son realizados. Los estudios pequeños que no encuentran resultados "significativos" a menudo no son publicados. Las revistas tienden a no interesarse, dado que como se explicó anteriormente, no hay mucha información en un estudio negativo que tiene poca

potencia. Por el contrario, los estudios grandes son costosos e involucran muchos investigadores. Cualesquiera sean los resultados provenientes de un estudio grande, hay más interés de parte de todos para publicarlo.

En la medida que este escenario describe la realidad, el cuerpo de estudios publicados está formado fundamentalmente por estudios pequeños con resultados "significativos" y estudios grandes con resultados "significativos" y "no significativos". Sin embargo, si hay muchos pequeños (i.e., fáciles, económicos) estudios en marcha, la probabilidad de 5% de cometer un error Tipo I se traduce en un número grande de hallazgos positivos y por lo tanto, de publicaciones. Así, muchos de los pequeños estudios en la literatura están informando errores Tipo I más que verdaderas asociaciones.

El siguiente ejemplo, basado en ensayos aleatorios de tratamientos nuevos, es de un artículo de Peto, Pike, y cols. Supongamos que hay 100 ensayos grandes y 1,000 ensayos pequeños de tratamientos que no difieren en realidad, y 20 ensayos grandes y 200 ensayos pequeños de tratamientos que realmente difieren. Los ensayos grandes tienen una potencia estadística de 95%; los ensayos pequeños tienen una potencia estadística de 25%. El nivel de significancia es de 5%, y sólo los ensayos que tuvieron resultados significativos son publicados. Estos supuestos, algo pesimistas, pero tal vez muy realistas, llevan al siguiente escenario hipotético para el número de ensayos de tratamiento en marcha que son "estadísticamente significativos" ($p < 0.05$):

Tamaño planificado del ensayo	Tasa de mortalidad verdadera en		# de ensayos plantead os	Número que se espera encontrar	
	Controles	Tratamiento		p>0.05	p<0.05
250	50%	50%	100	95 (VN)*	5 (FP)*
250	50%	33%	20	1 (FN)	19 (VP)
25	50%	50%	1,000	950 (VN)	50 (FP)
25	50%	33%	1,000	150 (FN)	50 (VP)

* VN, FP, FN, VP son una analogía para sensibilidad y especificidad (ver más adelante).

En este escenario, 100 ensayos pequeños con resultados "significativos" serán publicados, pero sólo la mitad de ellos reflejarán una diferencia verdadera entre tratamientos. La conclusión a la que llegan Peto, Pike y cols. es de que hay que prestar atención sólo a los ensayos grandes, sobretodo aquellos suficientemente grandes para ser publicados aún si no encuentran diferencias significativos entre tratamientos.

Estos resultados pueden ser considerados en términos de los conceptos de sensibilidad, especificidad, y valores predictivos. En estos conceptos, la sensibilidad corresponde a la potencia estadística para detectar una verdadera diferencia (95% para los ensayos grandes, 25% para los ensayos pequeños), especificidad corresponde a uno menos el nivel de significancia – la probabilidad de identificar correctamente un resultado aleatorio (95% de especificidad para un nivel de

significancia de 5%), y el valor predictivo positivo es la probabilidad de que un resultado "significativo" de hecho refleje una verdadera diferencia en la efectividad del tratamiento.

Ensayos grandes (p.ej., 250 muertes)

	Verdadera tasa de mortalidad en el grupo de tratamiento (suponiendo una tasa de mortalidad de 50% en el grupo control)		
P < 0.05	33%	50%	Total
Si	19	5	24
No	1	95	96
Total	20	100	120

De esta manera, el valor predictivo de una p < 0.05 = 19/24 = 79%

Ensayos pequeños (p.ej., 25 muertes)

	Verdadera tasa de mortalidad en el grupo de tratamiento (suponiendo una tasa de mortalidad de 50% en el grupo control)		
P < 0.05	33%	50%	Total
Si	50	50	100
No	150	950	1,100
Total	200	1,000	1,200

Valor predictivo de P<.05 = 50/100 = 50%

Evaluando el rol del azar - estimación de intervalos

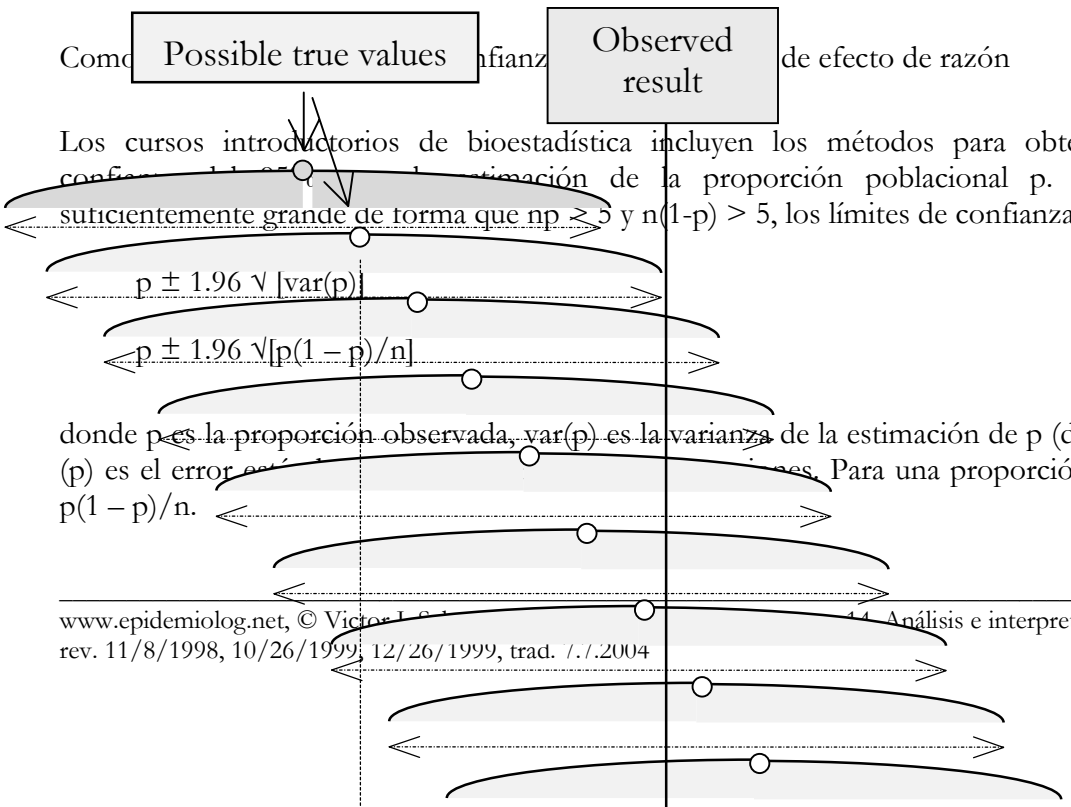
[Los estudiantes de EPID 168 deben conocer de estos conceptos, pero no de los cálculos]

Las pruebas de significancia estadística, con su orientación hacia la toma de decisiones, han caído un poco en desgracia en las investigaciones epidemiológicas. Con la premisa de que un estudio epidemiológico es esencialmente un procedimiento de medición (ver Rothman), se argumenta que el enfoque estadístico más apropiado es uno de estimación (p.ej., de la medida de efecto) más que de prueba de significancia. Por supuesto, igual hay una necesidad de cuantificar el rol del azar, pero en

un marco de estimación el azar es cuantificado por un intervalo de confianza o límites de confianza alrededor de la estimación puntual. Los límites de confianza cuantifican la magnitud de la incertidumbre en una estimación definiendo un intervalo que debería incluir el parámetro poblacional que se intenta estimar (p.ej., medida de efecto) un porcentaje conocido de las veces. Varios autores han argumentado que los intervalos de confianza son superiores a los valores p como mecanismo de cuantificar el grado de error aleatorio subyacente a la asociación.

Los intervalos de confianza contestan la pregunta, "¿qué posibles valores de un parámetro poblacional (p.ej., razón de densidad de incidencia) son consistentes con los resultados observados?" Dicho de otra manera, "¿cuál es rango de verdaderos valores que, cuando son distorsionados por influencias no sistemáticas, podrían producir los resultados observados?" Los intervalos de confianza pueden dar información sobre la precisión de una estimador o estimadores basado en la cantidad de datos disponibles para el estimador. Si no se observó una asociación "significativa", el intervalo de confianza puede dar una idea de que tan fuerte puede ser una asociación existente y sin embargo, por efecto del azar, no ser observada.

La naturaleza de un intervalo de confianza y lo que puede y no puede dar, sin embargo, es un poco complicado (basado en una discusión sobre intervalos de confianza en la lista de internet STAT-L que se prolongó durante semanas y atrajo una cantidad de respuestas y contra-respuestas). La perspectiva frecuentista es que un "intervalo de confianza del 95%" es un intervalo obtenido por un procedimiento que el 95% de las veces produce un intervalo que contiene el verdadero parámetro. En forma ideal, un intervalo del 95% sería aquel que "contiene el parámetro con una probabilidad de 95%". Pero los frecuentistas argumentan que el intervalo es fijado por los datos, y el parámetro poblacional ya existe en la naturaleza. El parámetro puede o no, estar en el intervalo. No hay probabilidades involucradas en eso. Lo único que podemos decir es que el 95% de las veces el procedimiento obtendrá un intervalo que incluye el valor del parámetro (y que el 5% de las veces el procedimiento producirá un intervalo que no lo contiene). Desde esta perspectiva un intervalo de 95% es c que él o ella dará la re una pregunta particular puede ser correcta o incorrecta.



Como nfianz de efecto de razón

Los cursos introductorios de bioestadística incluyen los métodos para obtener intervalos de confianza del 95% de la estimación de la proporción poblacional p. Si la muestra es suficientemente grande de forma que $np \geq 5$ y $n(1-p) > 5$, los límites de confianza son:

$$p \pm 1.96 \sqrt{[var(p)]}$$

$$p \pm 1.96 \sqrt{[p(1-p)/n]}$$

donde p es la proporción observada, var(p) es la varianza de la estimación de p (de manera que √var (p) es el error est nes. Para una proporción, var(p) es igual a $p(1-p)/n$.

Este método puede ser usado para estimar intervalos de confianza para prevalencia, incidencia acumulada, y otras proporciones simples. Muchas medidas epidemiológicas, sin embargo, son razones (p.ej., RIC, RDI, y OR). Dado que las medidas de efecto de razón tienen distribuciones fuertemente asimétricas (la mayor parte de los valores posibles caen a la derecha del valor nulo, 1.0, el enfoque habitual es estimar primero el intervalo de confianza para el logaritmo natural [ln(RIC), ln(RDI),o ln(OR)] y luego tomar el anti-logaritmo (exponente) de los límites de confianza:

$$\text{IC 95\% para } \ln(OR) = \ln(OR) \pm 1.96 \sqrt{\{\text{var}[\ln(OR)]\}}$$

$$\text{IC 95\% para OR} = \exp\{\ln(OR) \pm 1.96 \sqrt{([\text{var}[\ln(OR)])}\}$$

$$= OR \exp\{\pm 1.96 \sqrt{([\text{var}([\ln(OR)])}\}$$

Para obtener la varianza del ln(OR), usamos una fórmula simple (que ha sido derivada por medio de una aproximación de series de Taylor al ln[OR]):

$$\text{var}\{[\ln(OR)] = 1/a + 1/b + 1/c + 1/d\}$$

que funciona bien si a, b, c, y d tienen todos valores de por lo menos 5.

Por lo tanto el intervalo de confianza del 95% para el ln(OR) es:

$$\ln(OR) \pm 1.96 \sqrt{[(1/a + 1/b + 1/c + 1/d)]}$$

y el intervalo de confianza del 95% para el OR es:

$$OR \exp\{\pm 1.96 \sqrt{(1/a + 1/b + 1/c + 1/d)}\}$$

o

$$OR \; e^{+ 1.96 \sqrt{(1/a + 1/b + 1/c + 1/d)}}$$

Las formulas de los intervalos de confianza de la RIC y la RDI se pueden encontrar en Kleinbaum, Kupper y Morgenstern y Rothman y Greenland. Por cierto, si la población de estudio es muy seleccionada (i.e., no representativa de ninguna otra población de interés), ¿qué tan útil es el valor de una estimación?

ADVERTENCIA IMPORTANTE: toda esta sección, obviamente, se ha basado en el supuesto de que el muestreo y la medición son perfectas (no sesgados, independientes). Cualquier cosa que no sea una muestra al azar simple no sesgada y cualquier error de medición invalidará lo anterior por lo menos en alguna medida.

Meta-análisis

El meta-análisis es un enfoque cuantitativo para resumir y sintetizar los hallazgos de distintos estudios sobre una relación particular de interés. El meta-análisis surge del reconocimiento de que el fracaso en encontrar "resultados significativos" puede deberse tanto a una limitación de la potencia

estadística de los estudios individuales como a la ausencia de una relación. La combinación de información a partir de múltiples estudios puede dar una evaluación más precisa y definitiva de la existencia y fuerza de una relación que la que se obtiene de un único estudio o, se ha argumentado, de la revisión no cuantitativa de la literatura.

Hay cuatro pasos en la realización de un meta-análisis: 1) formulación del problema, 2) identificación de los trabajos (publicados y no publicados), 3) codificación y evaluación de los trabajos, y 4) análisis estadístico. Los pasos 2) y 3) son esenciales para la validez del meta-análisis, dado que las conclusiones que surjan del meta-análisis dependerán de lo adecuado que sea la evidencia sobre la relación representada por los trabajos de investigación que son incluidos finalmente en el análisis (la posibilidad de un sesgo de publicación contra los estudios "negativos" implica que se debe realizar un esfuerzo para ubicar los estudios no publicados). La estrategia para el análisis estadístico puede ser similar al del análisis estratificado, tomando cada trabajo como un "estrato" separado. Enfoques más refinados reconocen que los propios trabajos pueden ser considerados una muestra de un universo de trabajos posibles, de manera que el plan de ponderación necesita tomar en cuenta la variabilidad entre estudios además de la variabilidad intra-estudio (como en el modelos de efectos aleatorios del análisis de varianza).

En su forma pura, se predica el meta-análisis basado en el supuesto de que el conjunto de trabajos representa una muestra al azar de observaciones obtenidas en forma equitativa de una asociación, de manera que las diferencias entre los trabajos pueden ser consideradas variabilidad aleatoria (de muestreo). Así una medida de resumen construida por la combinación de estudios nos da una estimación más precisa de la verdadera asociación. En la práctica real, sin embargo, los estudios epidemiológicos raramente son equivalentes, dado que difieren a menudo en cuanto a la población estudiada, las medidas tomadas, y los enfoques analíticos. Aún los estudios que parecen ser equivalentes (p.ej. "estudio caso control basado en población, no apareado, con una medida fisiológica de la exposición y controlado para el mismo conjunto de potenciales factores de confusión") serán diferentes en formas menos obvias: las poblaciones probablemente sean diferentes en maneras desconocidas y no medidas, los sistemas de diagnóstico de la enfermedad pueden ser distintos entre poblaciones, los factores de respuesta en la selección de controles pueden ser diferentes, los procedimientos de recolección y los análisis de laboratorio de la exposición pueden ser diferentes en formas sutiles que, sin embargo, pueden afectar los resultados (p.ej., ver los ejemplos que involucran las pruebas para VIH y los análisis de homocisteína en *J Clin Epidemiol* 2001(5)), y pueden diferir los métodos de recolección de datos y el manejo analítico de los potenciales factores de confusión. Una exploración de la heterogeneidad en los meta-análisis de estudios de Síndrome de Muerte Súbita del Lactante y posiciones al dormir (Dwyer et al, 2001) demuestra algunos de estos temas.

Interpretación de los resultados

Preguntas claves

1. ¿Qué tan buenos son los datos?
2. ¿Podría el azar o algún sesgo explicar los resultados?
3. ¿Cómo se comparan los resultados con los de otros trabajos?

4. ¿Qué teorías o mecanismos podrían explicar los hallazgos?
5. ¿Qué hipótesis nuevas son sugeridas?
6. ¿Cuáles son los próximos pasos de investigación?
7. ¿Cuáles son las implicancias clínicas y de políticas?

Bibliografía

General

Ahlbom, Anders. *Biostatistics for epidemiologists.* Boca Raton, Florida, Lesis Publishers, 1993, 214 pp., $45.00 (reviewed in *Am J Epidemiol*, April 15, 1994).

Bailar, John C., III; Thomas A. Louis, Philip W. Lavori, Marcia Polansky. Studies without internal controls. *N Engl J Med* 1984; 311:156-62.

Bauer UE, Johnson TM. Editing data: what difference do consistency checks make. *Am J Epidemiol* 2000151:921-6.

Bulpitt, C.J. Confidence intervals. *The Lancet* 28 February 1987: 494-497.

Dwyer, Terence; David Couper, Stephen D. Walter. Sources of heterogeneity in the meta-analysis of observational studies: The example of SIDS and sleeping position. *J Chron Dis* 2001;54:440-447.

Feinstein, Alvan R. The fragility of an altered proportion: a simple method for explaining standard errors. *J Chron Dis* 1987; 40:189-192.

Feinstein, Alvan R. X and iprr: An improved summary for scientific communication. *J Chron Dis* 1987; 40:283-288.

Frank, John W. Causation revisited. *J Clin Epidemiol* 1988; 41:425-426.

Gerbarg, Zachary B.; Ralph I. Horwitz. Resolving conflicting clinical trials: guidlines for meta-analysis. *J Clin Epidemiol* 1988; 41:503-509.

Glantz, Stanton A. *Primer of biostatistics.* NY, McGraw-Hill, 1981.

Godfrey, Katherine. Comparing means of several groups. *N Engl J Med* 1985;313:1450-6.

Hertz-Picciotto, Irva. What you should have learned about epidemiologic data analysis. Epidemiology 1999;10:778-783.

Northridge, Mary E.; Bruce Levin, Manning Feinleib, Mervyn W. Susser. Statistics in the journal–significance, confidence, and all that. Editorial. *Am J Public Hlth* 1997;87(7):1092-1095.

Powell-Tuck J, MacRae KD, Healy MJR, Lennard-Jones JE, Parkins RA. A defence of the small clinical trial: evaluation of three gastroenterological studies. *Br Med J* 1986; 292:599-602.

Ragland, David R. Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cutpoint. *Epidemiology* 1992;3:434-440

Rothman - *Modern Epidemiology*, Chapters 9, 10, 14.

Schlesselman - *Case-control studies*, Chapters 7-8. (Especially the first few pages of each of these chapters).

Woolf SH, Battista RN, Anderson GM, Logan AG, et al. Assessing the clinical effectiveness of preventive maneuvers: analytic principles and systematic methods in reviewing evidence and developing clinical practice recommendations. *J Clin Epidemiol* 1990; 43:891-905.

Zeger, Scott L. Statistical reasoning in epidemiology. *Am J Epidemiol* 1991; 134(10):1062-1066.

El papel de las pruebas de hipótesis estadísticas, los intervalos de confianza y otras medidas de resumen de significancia estadística y precisión de las estimaciones

Allan H. Smith and Michael N. Bates. Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies. *Epidemiology* 1992;3:449-452

Browner, Warren S.; Thomas B. Newman. Are all significant P values created equal? *JAMA* 1987; 257:2459-2463.

Fleiss, Joseph L. Significance tests have a role in epidemiologic research: reactions to A.M. Walker (*Am J Public Health* 1986; 76:559-560). See also correspondence (587-588 and 1033).

George A. Diamond and James S. Forrester. Clinical trials and statistical verdicts: probable grounds for appeal. *Annals of Internal Medicine* 1983; 93:385-394

Greenland, Sander. Randomization, statistics, and causal inference. *Epidemiology* 1990;1:421-429.

Maclure, Malcome; Greenland, Sander. Tests for trend and dose response: misinterpretations and alternatives. *Am J Epidemiol* 1992;135:96-104.

Mood, Alexander M. and Franklin A. Graybill. *Introduction to the theory of statistics*. 2ed. NY, McGraw-Hill, 1963.

Oakes, Michael. *Statistical inference*. Chestnut Hill, Mass., Epidemiology Resources, 1986.

Peace, Karl E. The alternative hypothesis: one-sided or two-sided? *J Clin Epidemiol* 1989; 42(5):473-477.

Poole, Charles. Beyond the confidence interval *Am J Public Health* 1987; 77:195-199.

Poole, C. Confidence intervals exclude nothing *Am J Public Health* 1987; 77:492-493. (Additional correspondence (1987; 77:237)).

Savitz DA, Tolo KA, Poole C. Statistical significance testing in the *American Journal of Epidemiology*, 1970-1990. *Am J Epidemiol* 1994;139:1047-.

Thompson, W. Douglas. Statistical criteria in the interpretation of epidemiologic data *Am J Public Health* 1987; 77:191-194.

Thompson, W.D. On the comparison of effects *Am J Public Health* 1987; 77:491-492.

Walker, Alexander M. Reporting the results of epidemiologic studies *Am J Public Health* 1986; 76:556-558.

Woolson, Robert F., and Joel C. Kleinman. Perspectives on statistical significance testing. *Annual Review of Public Health* 1989(10).

Estimación de tamaño muestral

Donner A, Birkett N, and Burk C. Randomization by Cluster: sample size requirements and analysis. *Am J Epidemiol* 1981; 114:706

Snedecor GW, Cochran WG. *Statistical Methods*, 1980 (7th ed) see pages 102-105, 129-130 (Table A is from page 104).

Imputación

Heitjan, Daniel F. Annotation: what can be done about missing data? Approaches to imputation. *Am J Public Hlth* 1987;87(4):548-550.

Little RJA, Rubin DB. *Statistical analysis with missing data.* NY, Wiley, 1987.

Interpretación de múltiples pruebas de significancia estadística

Bulpitt, Christopher J. Subgroup analysis. *Lancet* 1988 (July 2);31-34.

Cupples, L. Adrienne; Timothy Heeren, Arthur Schatzkin, Theodore Coulton. Multiple testing of hypotheses in comparing two groups. *Annals of Internal Medicine* 1984; 100:122-129.

Holford, Theodore R.; Stephen D. Walter, Charles W. Dunnett. Simultaneous interval estimates of the odds ratio in studies with two or more comparisons. *J Clin Epidemiol* 1989; 42(5):427-434.

Jones, David R. and Lesley Rushton. Simultaneous inference in epidemiological studies. *Int J Epidemiol* 1982;11:276-282.

Lee, Kerry L., Frederick McNeer, Frank Starmer, et al. Lessons from a simulated randomized trial in coronary artery disease. *Circulation* 61:508-515, 1980.

Stallones, Reuel A. The use and abuse of subgroup analysis in epidemiological research. *Preventive Medicine* 1987; 16:183-194 (from Workshop on Guidelines to the Epidemiology of Weak Associations)

See also Rothman, *Modern Epidemiology*.

Interpretación de estudios "negativos"

Freiman, Jennie A., Thomas C. Chalmers, Harry Smith, Jr., and Roy R. Kuebler. The importance of beta, the Type II error and sample size in the design and interpretation of the randomized control trial. *N Engl J Med* 1978;299:690-694.

Hulka, Barbara S. When is the evidence for `no association' sufficient? Editorial. *JAMA* 1984; 252:81-82.

Meta-análisis

Light, R.J.; D.B. Pillemer. *Summing up: the science of reviewing research.* Cambridge MA, Harvard University Press, 1984. (very readable)

Longnecker M.P.; J.A. Berlin, M.J. Orza, T.C. Chalmers. A meta-analysis of alcohol consumption in relation to risk of breast cancer. *JAMA* 260(5):652-656. (example)

Wolf, F.M. *Meta-Analysis: quantitative methods for research synthesis.* Beverly Hills, CA, Sage, 1986.

Sesgo

Greenland, Sander. The effect of misclassification in the presence of covariates. *Am J Epidemiol* 1980; 112:564-569.

Walter, Stephen D. Effects of interaction, confounding and observational error on attributable risk estimation. *Am J Epidemiol* 1983;117:598-604.

Apéndice

Estimación del tamaño muestral para comparar dos proporciones o dos promedios

(Adaptado de un resumen preparado por Dana Quade, UNC Departmento de Bioestadísticas, Junio 1984)

Si N es el número de sujetos (unidades de observación) necesarias en **cada** uno de los grupos a ser comparados, entonces

$$N = I \times D \times C$$

Donde:

I = Intolerancia para el error, que depende de:

a. Alfa = nivel de significancia deseado para utilizar en nuestras pruebas estadísticas (p.ej., 5%, dos colas)
b. Beta = error tipo II (p.ej.,.10 – lo mismo que 1- potencia)

Fórmula: $I = (Z_{alfa} + Z_{beta})^2$

Z_{alfa} y Z_{beta} son, respectivamente, los valores críticos correspondientes a alfa y beta de la distribución normal (ver la tabla A en la página siguiente)

D = Diferencia a detectar, que depende de la estrechez de la diferencia entre las verdaderas proporciones o promedios, en relación con el desvío estándar de esa diferencia. D puede ser considerado la inversa de la "razón señal/ruido" – cuanto más débil la señal o más fuerte el ruido, se necesitan más sujetos.

$$D = \frac{\text{"ruido"}}{\text{"señal"}} \quad \text{OR} \quad \frac{p_1(1-p_1) + p_2(1-p_2)}{(p_1 - p_2)^2} \quad \text{OR} \quad \frac{2(\sigma^2)}{(\mu_1 - \mu_2)^2}$$

(para diferencias entre proporciones, en que (p_1 y p_2 son las dos proporciones – ver tabla en la próxima página)

(para diferencias entre medias, en que μ_1 y μ_2 son las dos medias, y σ^2 es la varianza de la diferencia)

C - observaciones en conglomerados, que depende de si las observaciones son seleccionadas en forma independiente o en conglomerados (N.T. en inglés "clusters").

- Si todas las observaciones son muestreadas en forma independiente, C = 1.
- Si las observaciones son muestradas en conglomerados (p.ej., por hogares, escuelas, lugares de trabajo, segmentos censales, etc.), el tamaño muestral debe ser aumentado para compensar el hecho de que las observaciones dentro de un conglomerado son más parecidas entre ellas que a las observaciones en otros conglomerados. Si ro es la correlación intra-conglomerado entre las observaciones dentro del conglomerado, entonces:
- $$C = 1 + (m-1)\, ro$$

  donde m es el tamaño promedio de un conglomerado (i.e., n = km, donde k es el número de conglomerados). C es conocido a menudo como el "efecto de diseño". Si los conglomerados son grandes o si las personas dentro de ellas tienden a ser muy similares, los sujetos individuales contribuyen con poca información y por lo tanto necesitas estudiar un número muy grande de ellos. Si eliges "pensadores independientes", aprenderías más de cada uno.

Tabla A: Intolerancia para el error

Potencia deseada	Prueba de dos colas			Prueba de una cola		
	Nivel de significancia			Nivel de significancia		
	0.01	0.05	0.10	0.01	0.05	0.10
0.80	11.7	7.9	6.2	10.0	6.2	4.5
.90	14.9	10.5	8.6	13.0	8.6	6.6
0.95	17.8	13.0	10.8	15.8	10.8	8.6

Tabla B: Diferencia a detectar

		p_2					
		.10	.20	.30	.40	.50	.60
p_1	.05	55	9.2	4.1	2.4	1.5	1.0
	.10	–	25	7.5	3.7	2.1	1.3
	.15	87	115	15	5.9	3.1	1.8
	.20	25	–	37	10	4.6	2.5
	.25	12.3	139	159	19	7.0	3.5

Complicaciones

1) Tamaños muestrales desiguales

que n sea el tamaño muestral **promedio** = $(n_1+n_2)/2$

que $lambda_1 = n_1/2n$, $lambda_2 = n_2/2n$ $(lambda_1 + lambda_2 = 1)$

$$D = \frac{\frac{p_1(1-p_1)}{2\,lambda_1} + \frac{p_2(1-p_2)}{2\,lambda_2}}{(p_1-p_2)^2} \quad OR \quad \frac{\frac{\sigma^2_1}{2\,lambda_1} + \frac{\sigma^2_2}{2\,lambda_2}}{(\mu_1-\mu_2)^2}$$

2) Covariables

Si las pruebas estadísticas serán llevadas a cabo en forma separada dentro de cada estrato, debe determinarse el n para cada estrato según se describió más arriba.

Si los resultados de los distintos estratos van a ser probados para una asociación global promedio, probablemente sea mejor no tenerlas en cuenta explícitamente en las fórmulas de tamaño muestral, sino aumentar discretamente el n global.

Nota: en la literatura pueden encontrarse fórmulas "más precisas", pero los parámetros necesarios para D y C nunca son realmente conocidos.

Tamaño muestral para estimación de intervalo

Puede usarse la amplitud tolerable para un intervalo de confianza como objetivo para estimar el tamaño muestral necesario para la población de estudio. Supongamos, por ejemplo, que un investigador desea estimar la proporción (p) de uso de preservativo entre los usuarios de una policlínica. Si el investigador puede obtener una muestra al azar simple de esa población, su estimación de la proporción de usuarios de preservativos sería p = u/n, donde u es el número de usuarios en la muestra y n es el tamaño de esa muestra. Como se señaló anteriormente, si np > 5 y n(1-p) > 5, entonces el intervalo de confianza de 95% para p es:

$$p \pm 1.96\ (1/\sqrt{[p(1-p)\ n]})$$

Por ejemplo, si p es 0.50, el intervalo de confianza es:

$$0.5 \pm 1.96\ (1/\sqrt{[\ (0.5)(0.5)/n]}) \quad = \quad 0.5 \pm 1.96\ \frac{(0.5)}{\sqrt{n}}$$

[La raíz cuadrada de (0.5)(0.5) es, por supuesto, 0.5]

Dado que 1.96 × 0.5 es aproximadamente 1, en sentido práctico la expresión es equivalente a:

$0.5 \pm 1/\sqrt{n}$, de manera que los límites de confianza son $(0.5 - 1/\sqrt{n},\ 0.5 + 1/\sqrt{n})$

Por ejemplo, supongamos que n, el tamaño muestral es 100. El intervalo de confianza alrededor de la estimación puntual de 0.5 es:

$$
\begin{aligned}
& (0.5 - 1/\sqrt{100},\ 0.5 + 1/\sqrt{100}) \\
= \quad & (0.5 - 1/10,\ 0.5 + 1/10) \\
= \quad & (0.5 - 0.1, 0.5 + 0.1) \\
= \quad & (0.4, 0.6)
\end{aligned}
$$

La imprecisión a menudo se cuantifica en términos de la mitad de la amplitud del intervalo, i.e., la distancia entre la estimación puntual y el límite superior (o inferior) del intervalo, al cual llamaremos aquí el "margen de error". La mitad de la amplitud del intervalo anterior es 0.1 (i.e., la raíz cuadrada de n) en términos absolutos o 20% (0.1/0.5) en términos relativos. Un margen de error de 0.1 o 20% es adecuado para una estimación grosera de una proporción, pero no mucho más.

Dado que las expresiones anteriores involucran la raíz cuadrada del tamaño muestral, la disminución progresiva de la amplitud del intervalo produce aumentos sustancialmente mayores del tamaño muestral. Por ejemplo, para tener un margen de error absoluto de 0.05 o relativo de 10%, el tamaño muestral debe ser cuadruplicado, a 400:

$$
\begin{aligned}
& (0.5 - 1/\sqrt{400},\ \ 0.5 + 1/\sqrt{400}) \\
= \quad & (0.5 - 1/20,\ \ 0.5 + 1/20) \\
= \quad & (0.5 - 0.05,\ \ 0.5 + 0.05) \\
= \quad & (0.45,\ \ 0.55)
\end{aligned}
$$

De igual manera, un tamaño muestral de 900 da límites de confianza que son un tercio la amplitud de una muestra de 100, una muestra de 2,500 da límites que son un cuarto de ancho que para n = 100, etc.

Estos números se refieren a una estimación puntual de 0.5, que produce el mayor margen de error en términos absolutos. Una estimación puntual menor o mayor tendrá un intervalo más estrecho (en términos absolutos), porque la raíz cuadrada de p(1 – p) no puede exceder 0.5 (¡pruébalo! - o usa cálculo). El margen de error relativo, por otro lado, se relaciona en forma inversa con el tamaño de la estimación puntual. Estudiemos la siguiente tabla:

Estimación puntual	Tamaño muestral	Margen de error (redondeado) Absoluto *	Relativo ** (%)
0.1	100	0.06***	60***
0.2	100	0.08	40
0.3	100	0.09	30
0.4	100	0.096	24
0.5	100	0.10	20
0.6	100	0.096	16
0.7	100	0.09	12
0.8	100	0.08	9.8
0.9	100	0.06	6.5
0.1	400	0.03	30
0.2	400	0.04	20
0.3	400	0.045	15
0.4	400	0.048	12
0.5	400	0.05	10
0.6	400	0.048	8.0
0.7	400	0.045	6.4
0.8	400	0.04	4.9
0.9	400	0.03	3.2

* Aproximadamente la mitad de la amplitud del intervalo de confianza de 95% en términos absolutos

** Aproximadamente la mitad de la amplitud del intervalo de confianza de 95% en términos absolutos, en relación con el tamaño de la estimación puntual

*** Cálculo: $1.96\ (1/\sqrt{[(0.01)(1 - 0.01)\ /\ 100]}) = 1.96\ (0.03) = 0.0588 \approx 0.06$ márgen de error absoluto

Esta tabla ilustra que:

1. cuadruplicando el tamaño muestral disminuye a la mitad el margen de error.
2. el margen de error absoluto disminuye a medida que la estimación puntual de aleja de 0.5.
3. el margen de error relativo está inversamente – y muy fuertemente – relacionado con el tamaño de la estimación puntual

Para estimaciones puntuales muy pequeñas, como se ilustra en la siguiente tabla, son necesarias muestras muy grandes para obtener un margen de error relativo pequeño. Aún un tamaño muestral de 2,500 produce un margen de error relativo de 17% para una proporción de 0.05.

Estimación puntual	Tamaño muestral	Margen de error (redondeado) Absoluto *	Relativo * (%)
0.5	100	0.10	20
0.5	400	0.05	10
0.5	900	0.033	6.6
0.5	1,600	0.025	5.0
0.5	2,500	0.020	4.0
0.05	100	0.043	85
0.05	400	0.021 **	43 **
0.05	900	0.014	28
0.05	1,600	0.011	21
0.05	2,500	0.009	17

* ver tabla anterior

** Cálculo: $1.96 \times (1/\sqrt{[(0.05)(0.95)/400]}) = 1.96 \times 0.0109$

$= 0.0214 \approx 0.021$ margen de error absoluto

Relativo $= 0.0214 / 0.05 = 0.427 = 42.7\%$ (aproximadamente 43%)

Recuerda que esta formula requiere que $nP \geq 5$, que apenas se cumple para P=0.05 and n=100.

¿Qué tan grande es una muestra suficientemente grande? Si el objetivo es fijar un límite superior o inferior en una proporción, un pequeño margen de error absoluto puede ser suficiente. Por ejemplo, si uno está investigando el anticuerpo de la hepatitis C y quiere estar seguro de que la seroprevalencia está por debajo de 5%, un tamaño muestral de 900 producirá un intervalo con un margen de error absoluto no mayor que 0.033 (para una estimación puntual de 0.5 – ver la tabla anterior) y más probablemente de 0.011 (para una estimación puntual de 0.05) o menor. Dado que esperamos que la seroprevalencia sea muy pequeña, el 0.011 es mucho más relevante que el 0.033. Si cuando llevamos a cabo la investigación obtenemos una estimación puntual de exactamente 0.05, el intervalo de confianza del 95% será (0.039, 0.061) que nos dirá que el verdadero valor es por lo menos poco probable que sea mayor de 6%. Si la estimación puntual está por debajo de 0.04, el límite superior de confianza estará por debajo de 5% y estaremos seguros de que la seroprevalencia no es mayor que ese valor.

Señalemos que todo lo anterior se basa en el supuesto de un muestreo y medición perfectos (no sesgados). Cualquiera otra cosa que una muestra al azar simple no sesgada y cualquier error de medición invalidará lo anterior por lo menos hasta cierto punto.

Note that the above is all based on the assumption of perfect (unbiased) simple random sampling and measurement. Anything other than an unbiased simple random sample and any error in measurement will invalidate the above at least to some extent.

Meditaciones sobre pruebas de hipótesis y significancia estadística

La teoría estadística de prueba de hipótesis y evaluación de la "significancia" estadística surge de un análisis de toma de decisiones con respecto a dos hipótesis que compiten: una hipótesis "nula" y una hipótesis alternativa. Dos tipos de errores son posibles.

Tipo I: Rechazar equivocadamente la "hipótesis nula" (H_0), a favor de la hipótesis alternativa (H_A), i.e., rechazar errónemante al azar como explicación suficiente para los resultados observados.

Tipo II: Equivocadamente no logramos rechazar H_0, i.e., erróneamente aceptamos el azar como explicación. [veremos una dicotomía paralela más adelante en el curso cuando discutamos sensibilidad y especificidad.]

Tradicionalmente, la probabilidad de error de Tipo I ha recibido más atención y es conocido como el "nivel de significancia" de la prueba. El error de Tipo I presumiblemente debe su importancia al deseo de la comunidad científica para evitar falsas alarmas, i.e., para evitar reaccionar frente a resultados que bien podrían ser fluctuaciones aleatorias. Por otro lado, la probabilidad de error de Tipo I es más fácil de estimar, dado que la probabilidad de error de Tipo II depende de la identificación del tamaño de la verdadera diferencia que uno busca detectar.

En las últimas décadas, el cálculo y la presentación de valores p (que dan información sobre la probabilidad del error de Tipo I) se han vuelto de rigor en la literatura científica empírica. De hecho, hay un número significativo (!) de personas que se niegan a prestar atención a resultados que tienen valores de p mayores de .05 (probabilidad de un error de Tipo I).

Esta actitud es un buen artefacto para generar trabajo, pero tal vez sea un poco cruel. Después de todo, un resultado con un valor p de .10 resultaría de un proceso aleatorio en sólo uno de diez ensayos. ¿Debería descartarse semejante hallazgo? Es más, dado que el valor p refleja el número de sujetos además del tamaño de la diferencia observada, un pequeño estudio tendría valores p muy pequeños sólo en el caso de diferencias observadas muy grandes (¿y tal vez poco realistas?) Si el tamaño de la diferencia observada es extraordinariamente grande, podemos sentir cierta sospecha sobre el hallazgo, a pesar de un valor p pequeño. Si la diferencia observada es plausible, pero el valor p es "no significativo" porque el estudio es pequeño, podríamos tal vez prestarle algo de atención.

Otra razón para un enfoque reflexivo de los valores p (y la estadística inferencial en general) es que las propias estimaciones de probabilidad son precisas sólo con respecto a los modelos que los subyacen. No sólo ocurre que los modelos matemáticos pueden no captar adecuadamente la verdadera situación, sino que además el contexto en que son utilizados puede confundir aún más la situación. Un supuesto crítico es el del muestreo al azar o aleatorización (como en un ensayo

aleatorio controlado).Aunque este supuesto es la base de toda la teoría estadística de prueba de hipótesis e intervalos de confianza, raramente se logra en los estudios observacionales y las limitaciones que impone sobre la interpretación de las pruebas estadísticas a menudo son poco apreciadas (Greenland S. Randomization, statistics, and causal inference *Epidemiology* 1990;1:421-249).

Los problemas de interpretación existen aún en los ensayos clínicos aleatorizados. Por ejemplo, el valor p de un único resultado en un único estudio puede ser 5 por ciento. Pero eso significa que 20 estudios independientes de dos fenómenos idénticos observarían, en promedio, una diferencia que resulta "significativa" al nivel de cinco por ciento. Un investigador prolífico que lleva a cabo 200 estudios en su vida profesional puede esperar que diez sean "significativos" sólo por azar. Es más, un estudio a menudo investiga múltiples resultados, incluyendo maneras diferentes de definir las variables involucradas.

Estas "comparaciones múltiples" aumentan la probabilidad de que diferencias al azar sean consideradas "significativas". Pero los procedimientos estadísticos para manejar esta "inflación de significancia" tienden, igual que las medidas para controlar la inflación de precios o inflación de notas académicas, a producir recesión o aún depresión [de los hallazgos del estudio.] ¿Debería obligarse a un investigador a jurar que (1) especificó una hipótesis a priori, incluyendo los procedimientos para definir y manipular todas las variables, decisiones sobre todas las relaciones a examinar, qué factores controlar, etc; (2) procedió directamente a la prueba estadística pre-especificada sin mirar los demás datos; y (3) no llevará a cabo más pruebas estadísticas con los mismos datos? (Ver Modern Epidemiology para más discusión sobre estos temas.)

¿Y qué ocurre con las llamadas "expediciones de pesca" (N.T. "fishing expeditions" en inglés) en que un investigador (o su computadora) estudian una base de datos para encontrar relaciones "significativas"? ¿Debe caracterizarse este procedimiento como "busca y encontrarás" o más bien "búsqueda y destrucción"? Algunos analistas recomiendan ajustar el nivel de significancia para tomar en cuenta dichas "comparaciones múltiples", pero un investigador con energías puede llevar a cabo suficientes pruebas de manera que el nivel de significancia ajustado es imposible de lograr. Otros autores (p.ej., Rothman, Poole) aseguran que no es necesario ajustar – que una vez que los datos están incluidos, el número de pruebas no es relevante. Otros (p.ej. Greenland) han propuesto enfoques más sofisticados al ajuste. Tal vez el mejor camino en este momento es doble:

(1) Si estás llevando a cabo una investigación, por ejemplo, en ensayo aleatorizado, en que tienes grandes posibilidades de satisfacer los supuestos de una prueba de hipótesis estadística y esperas probar una hipótesis específica, sobretodo una que pueda ayudar a tomar una decisión, probablemente sea mejor adherir lo mejor posible al formato de prueba de hipótesis de Neyman-Pearson. Este enfoque asegura el máximo impacto de tus resultados;

(2) Si estás llevando a cabo una investigación con algunas de las características anteriores, o ya has completado la prueba de hipótesis establecida a priori, analiza todo lo que quieras pero sé cándido describiendo como has procedido. Así los lectores podrán interpretar los resultados como consideren más apropiado.

La potencia aparente (calculada) raramente se logra porque amenudo supone que no hay errores en la clasificación de los participantes. Un estudio con una potencia anunciada de 90% podría haber tenido una probabilidad mucho menor de detectar una verdadera diferencia dada por la dilución producida por un sesgo de información. De igual manera, podemos en principio mejorar la potencia efectiva de un estudio si podemos aumentar la precisión con que las variables importantes son medidas.

Louis Guttman ha escrito que la estimación y aproximación, nunca olvidando la replicación, pueden ser más productivos que la prueba de significancia para el desarrollo de la ciencia. . [Louis Guttman. What is not what in statistics. *The Statistician* 25(2):81-107.]

La replicación independiente es el pilar del conocimiento científico.

Enfoque Bayesiano de la interpretación de un valor p

La utilización de los conceptos de sensibilidad, especificidad, y valor predictivo para interpretar pruebas de hipótesis estadísticas sugiere una analogía entre las pruebas estadísticas y las pruebas diagnósticas (ver Browner y Newman, 1987; Diamond y Forrester, 1983; y Feinstein, *Clinical Biostatistics*). Así como la interpretación de una prueba diagnóstica depende de la prevalencia de la enfermedad (la "probabilidad *a priori* de que el paciente tiene la enfermedad") la interpretacion de las pruebas estadísticas puede ser considerado como dependiente de "la prevalencia de la verdad", i.e., la razonabilidad de la hipótesis.

Como señalamos anteriormente, nos gustaría que la inferencia estadística nos diera una estimación de la probabilidad de que la hipótesis de interés (H) es verdadera dados los resultados observados. En vez, el valor p nos da la probabilidad de observar un resultado extremo bajo una hipótesis nula (clásicamente la inversa de la hipótesis de interés). El enfoque bayesiano de la interpretación de los valores p trata de dar una respuesta que se acerca más al objetivo original. En el enfoque bayesiano, comenzamos con una probabilidad previa para la verdad de la hipótesis y luego ajustamos esa probabilidad basándonos en los resultados de una investigación, para obtener una probabilidad posterior. El efecto que pueden tener los resultados de estudio sobre nuestra evaluación de la credibilidad de la hipótesis depende de nuestra evaluación originad de su credibilidad.

T significa que una prueba estadística es "significativa". Según el Teorema de Bayes, si Pr(H) es la probabilidad "*a priori*" de H, i.e., la probabilidad de que H sea verdadera se basa sólo en información previa, entonces la probabilidad de H *a posteriori* (la probabilidad de que H sea verdadera basado en información previa y el resultado de la actual prueba) es:

$$\Pr(H|T) = \frac{\Pr(H)\,\Pr(T|H)}{\Pr(H)\,\Pr(T|H) + \Pr(h)\,\Pr(T|h)}$$

[donde Pr(T|h) significa la probabilidad de una prueba positiva dada una hipótesis que no es verdadera] lo cual puede escribirse como:

$$Pr(H|T) = \frac{Pr(H)\ Pr(T|H)}{Pr(H)\ Pr(T|H) + [1 - Pr(H)]\ Pr(T|h)}$$

Dado que Pr(T|H) es la potencia estadística (la probabilidad de una prueba positiva dada una hipótesis verdadera) y Pr(T|h) es el valor p (la probabilidad de una prueba positiva dada una hipótesis que no es verdadera), la probabilidad posterior puede expresarse como:

$$Pr(H|T) = \frac{Pr(H)\ (\text{potencia})}{Pr(H)\ (\text{potenciar}) + [1 - Pr(H)]\ (\text{valor p})}$$

Pr(H|T) es por lo tanto una función de la probabilidad "*a priori*" de la hipótesis, la potencia estadística y el valor p. Por lo tanto el valor p tiene más impacto sobre Pr(H|T) cuando Pr(H) es pequeña (i.e., cuando una hipótesis no es respaldada por investigación previa o datos de laboratorio) (ver Diamond y Forrester).

Para tener una idea de cómo funcionan estas fórmulas con los valores típicos para los múltiples elementos, veamos las siguientes tablas:

Evaluación de la probabilidad posterior basada en la probabilidad previa, la potencia estadística y el valor p

	Probabilidad previa (Antes del estudio) Pr(H)	Potencia estadística del estudio Pr(T|H)	Valor P (Hallazgos del estudio) Pr(T|h)	Probabilidad Posterior (Después del estudio) Pr(H|T)	
Hipótesis creíble	0.60	0.8	0.100	0.92	
	0.60	0.8	0.050	0.96	Alta potencia
	0.60	0.8	0.001	1.00	
	0.60	0.5	0.100	0.88	
	0.60	0.5	0.050	0.94	Baja potencia
	0.60	0.5	0.001	1.00	
_Hipótesis poco creíble	0.05	0.8	0.100	0.30	
	0.05	0.8	0.050	0.46	Alta potencia
	0.05	0.8	0.001	0.98	
	0.05	0.5	0.100	0.21	
	0.05	0.5	0.050	0.34	Baja potencia
	0.05	0.5	0.001	0.96	

En esta tabla, por ejemplo, un valor p muy fuerte (p.ej., 0.001) da una alta credibilidad (probabilidad posterior) aún para una hipótesis poco creible estudiada en una investigación de poca potencia estadística. Un valor p que es "apenas significativo", sin embargo, no hace que la hipótesis sea altamente creíble salvo que se considere más probable que no, antes del estudio. Aún un valor p "no significativo" (p.ej., 0.10) nos aumenta en algo la credibilidad de la hipótesis, de manera que en el pensamiento bayesiano un valor p de 0.10 no se consideraría un resultado "negativo" que hiciera dudar la existencia de una asociación. El meta-análisis, donde los resultados de múltiples estudios son combinados para obtener una evaluación cuantitativa de la asociación del total del cuerpo de evidencia, también toma en cuenta la evidencia a favor de la asociación de estudios que observaron una asociación pero que tuvieron un valor p mayor de 0.05. El uso formal de los métodos Bayesianos en el trabajo diario, sin embargo, está algo restringido por la ausencia de un método obvio para obtener una probabilidad previa.

Más meditaciones sobre la interpretación depruebas de significancia estadísticas

Algunos conceptos de la interpretación de las pruebas estadísticas de significancia pueden talvez ser ilustrados a través de un ejemplo basado en el glorioso origen de la teoría de probabilidad – los juegos de azar. Supongamos que un amigo te dice que tiene una intuición sobre la rueda de la ruleta. Mirando al que hace girar la rueda, tu amigo puede, según asegura, predecir donde caerá la bola dentro de un margen muy pequeño. Si para simplificar el ejemplo, la ruleta tiene los números 1-100, tu amigo dice que puede predecir los números en que caerá la bola. Quiere que le entregues dinero para mandarlo a Monte Carlo para hacer una fortuna para todos.

Naturalmente te entusiasma la idea de la riqueza instantánea pero también estás un poco escéptico. Para verificar la afirmación de tu amigo, llevas a cabo una prueba estadística. Le das a tu amigo $5 para que demuestre su habilidad en el casino local, y esperas los resultados para ver que pasa.

La hipótesis nula para tu prueba estadística es que tu amigo no tiene una habilidad especial de manera que sus posibilidades de predecir el lugar en que caerá la bola en cualquier vuelta son simplemente 1 en 100 (.01). La hipótesis alternativa de una cola es que tu amigo sí tiene esa habilidad y puede predecir el número correcto en forma más frecuente que 1 en 100 veces. [la hipótesis alternativa de dos colas es que tu amigo va predecir el lugar en que cae la bola más veces que las esperadas por el azar, o menos veces que lo esperado.]

Tu amigo vuelve con $400. Sabiendo que la probabilidad de que estuviera en lo cierto en cualquier vuelta sólo por azar es sólo 1%, tú estás impresionado. ¡Su desempeño fue "significativo al nivel .01"! ¿Le financias el viaje a Monte Carlo? ¿Cómo interpretas sus predicciones correctas?

¿Es correcto decir que hay sólo una probabilidad de 1% que la precisión de su predicción se debió a la "suerte"? No exactamente. Según la interpretación frecuentista, la predicción fue hecha y la rueda de la ruleta ya ha girado. La precisión se debió al "azar" ("suerte") o a la habilidad de tu amigo, pero sólo uno de los dos fue realmente resposable en ese momento. De manera que la probabilidad de que la predicción correcta se debió al azar es cero (i.e., tu amigo puede predecir) o uno (tu amigo no puede predecir.) ¡El único problema es que no sabes cuál es el caso aquí!

Puedes decir (antes de que gire la rueda y suponiendo que es una rueda balanceada) que si tu amigo no tenía una habilidad especial había sólo una probabilidad de uno por ciento de que hiciera una predicción correcta y que por lo tanto el hecho de que haya ganado es evidencia en contra de la hipótesis nula (de no habilidad) y a favor de la hipótesis alternativa (habilidad de predecir). Si tienes que decidir ese mismo día, puedes calcular que valdría la pena financiarle el viaje a Monte Carlo, pero estarías al tanto de que su predicción correcta podría deberse al azar porque había una probabilidad de uno por ciento de que en ausencia de cualquier clarividencia su predicción hubiera sido correcta (no es exactamente lo mismo que una probabilidad de uno por ciento de que su predicción correcta se debió al azar.) De manera que le das a tu amigo $2,000. Te lo agradece efusivamente, y al partir, te comenta que en realidad le llevó 30 intentos para realizar la predicción correcta - pidió prestado el dinero para los otros 29 intentos.

Esta información te hace pensar. Seguramente no te hubiera impresionado tanto si te hubiera dicho que podía hacer una predicción correcta en 30 intentos. Si la probabilidad de una predicción correcta (i.e., adivina correctamente) en ausencia de cualquier habilidad especial es 0.01, la probabilidad de una o más adivinanzas correctas en 30 intentos es 0.26 (1.0 menos la cantidad 0.99 elevado a la potencia 30). Veintiséis por ciento sigue siendo menor que 50%, i.e., la probabilidad de ganar usando la tirada de una moneda, pero no en forma impresionante. La evidencia en contra de la hipótesis nula no es ahora tan fuerte. Este cambio en tu interpretación demuestra los puntos que surgen en relación con pruebas de significancia múltiples y sesgos de los estudios pequeños.

Es posible, usando la teoría estadística, ajustar los niveles de significancia y los valores p para tomar en cuenta el hecho de que se han realizado múltiples pruebas de significancia independientes. Pero hay varios problemas prácticos para la aplicación de dichos procedimientos, uno de los cuales es la falta de independencia entre las múltiples pruebas en un conjunto particular de datos. Por ejemplo, si tu amigo explicara que hace una predicción incorrecta tan raramente que cuando le ocurre se molesta tanto que le lleva una hora entera (y 29 predicciones más) recuperar su capacidad de predecir, aunque sigas siendo escéptico te sería muy difícil calcular un valor p ajustado para tu prueba si creyeras que te está diciendo la verdad. De igual manera, en un conjunto dado de datos, ¿el hecho de que el investigador pruebe la misma diferencia de distintas maneras (p.ej., obesidad indexada por peso/altura2 [índice de Quetelet], peso/altura 3 [índice ponderal], por ciento encima del peso ideal, grosor del pliegue, y densidad corporal) debilita los hallazgos de cada prueba? Si también mirara las diferencias de presión arterial, ¿eso debilitaría la credibilidad de la significancia estadística de las diferencias en obesidad?

"Pagas tu dinero, y eliges lo que quieras".