

ESTUDIOS EXPERIMENTALES

INTRODUCCIÓN Y CONCEPTO

A diferencia de los estudios que han sido comentados en los capítulos anteriores, todos ellos estudios de observación, los estudios experimentales son estudios prospectivos, en los que se va a valorar el efecto de una o varias intervenciones. El punto crucial de la metodología de los estudios experimentales es cómo va a decidirse quiénes, entre los participantes, van a recibir el elemento nuevo que se introduce. Esta decisión puede ser dejada al azar y de manera aleatoria los sujetos participantes en el estudio se dividen en dos grupos: el experimental, formado por los receptores del nuevo factor introducido por el investigador, y el control, constituido por los restantes participantes. La aleatorización en la distribución de la intervención es lo que define a un estudio experimental como puro o verdadero. Si, por el contrario, la aleatorización no se realiza y es el investigador -influido o no por las características de la experiencia- el que decide quienes recibirán o no la intervención, el estudio es definido como cuasiexperimental.

Cuanto más cuidado se tenga en asegurar que los dos grupos sean tan similares como sea posible, tanto mejor controlado estará un experimento. En este sentido, los estudios experimentales puros ofrecen una gran ventaja sobre los cuasiexperimentales, ya que la aleatorización es una técnica para controlar errores.

Aquellos casos en que la epidemiología experimental introduce un elemento nuevo, o lo suprime, en el devenir natural de un proceso que afecta a una colectividad entera y sigue los resultados que se derivan como consecuencia de tal intervención, reciben el nombre de experiencias en condiciones no controladas (ausencia de grupo control). Un ejemplo de ellas serían las vacunaciones masivas en períodos de epidemia o el control de la eliminación de un vertido.

Las experiencias en condiciones naturales han sido definidas como aquellas que se desarrollan en una población en la que ha aparecido de manera circunstancial y brusca un factor que altera profundamente las características del medio ambiente. Al no haber manipulación por parte del investigador, que tan solo aprovecha otras intervenciones humanas o de la propia naturaleza, son estudios de observación, aunque con un carácter de excepción. Los accidentes de Bhopal o Chernobyl serían ejemplos de esta situación.

Los estudios experimentales son utilizados con mayor frecuencia en la valoración terapéutica de fármacos e intervenciones terapéuticas, mediante el ensayo clínico. Este mismo tipo de diseños puede ser aplicado a la investigación de medidas preventivas de la enfermedad, buscando reducir el riesgo individual. Ejemplos de ello son los estudios que valoran la eficacia de una vacuna o de la educación sanitaria. Una posterior aplicación de los estudios experimentales es la investigación de la etiología de la enfermedad, como confirmación de los resultados encontrados mediante diseños de observación. Por desgracia y debido a los condicionantes éticos, es muy problemático someter a poblaciones humanas a un agente nocivo que presumiblemente es determinante de una enfermedad.

Por ello, los estudios experimentales rara vez son empleados con esta finalidad, por lo que se ha de confiar exclusivamente en los datos proporcionados por los estudios de observación (sobre todo estudios de cohortes y de casos y controles, por este orden). Otra de las utilidades que presentan los estudios experimentales es en la evaluación de la planificación y programación sanitarias.

PRINCIPIOS BÁSICOS

Diseño general

Antes del inicio del estudio debe desarrollarse un protocolo de investigación adecuado de investigación. Por regla general, es recomendable la realización de un estudio piloto, con las finalidades de averiguar la proporción de participantes o la variabilidad en el resultado, para probar el cuestionario (entrevista o protocolo de valoración), y para investigar si es factible su realización. A título ilustrativo se expone a continuación un ejemplo de protocolo en el que se detallan los puntos principales que hay que tener en cuenta antes del inicio de la investigación.

1. Introducción. Exposición del problema con una revisión bibliográfica apropiada.
2. Objetivos del estudio, detallando de manera explícita las variables con las que se medirá el cumplimiento de los objetivos
3. Diseño:
 - a) Población de estudio, definiendo de manera muy clara los criterios de inclusión y de exclusión y la estimación del tamaño de muestra. Las estimaciones del tamaño de muestra de un estudio adquieren su máxima importancia en los diseños experimentales, sobre todo con humanos (por razones éticas). Por el coste y dificultades especiales de este tipo de estudios, la programación del número de sujetos que se van a necesitar es algo que debe estar especificado en el protocolo del estudio.
 - b) Incorporación de los sujetos del estudio: Informe de aceptación (si son humanos) y valoración de que cumplen los criterios de inclusión (mediante un examen basal o inicial). Hay que detallar muy bien como se piensa distribuir la intervención entre los participantes.
 - c) Intervención, muy bien descrita y cómo se medirá el grado de cumplimiento de la misma.
 - d) Seguimiento: Esquema
 - e) Valoración de las variables de respuesta: preparación y recolección de los datos, monitorización de la recolección y control de calidad y análisis de los datos
4. Organización: Investigadores participantes, Administración (Comités y subcomités)
5. Apéndices:
 - a) Definiciones de criterios de incorporación al estudio
 - b) Definiciones de las variables de respuesta

Aleatorización

Es el método electivo ideal para alcanzar la comparabilidad entre los grupos. Fortalece las pruebas estadísticas empleadas, disminuye el sesgo de selección e intenta proteger contra otro tipo de sesgos, fundamentalmente el de confusión (su principal finalidad). Aunque tradicionalmente se

ha afirmado que aleatorización es que cualquier sujeto tenga la misma probabilidad de entrar en cualquiera de las ramas de un ensayo, en realidad la aleatorización se define como el que todos los individuos tengan la misma probabilidad (p) de entrar al grupo de tratamiento y al grupo control ($1 - p$). p no tiene porqué ser igual a $1/2$, en cuyo caso $p = 1 - p$. El que sea $p = 1/2$ viene dado por razones de potencia estadística: cuando los grupos tienen un tamaño de muestra similar la potencia es máxima.

El proceso de aleatorización persigue que el investigador no seleccione los sujetos en los grupos que se comparan. Para ello ha de ser enmascarada (desconocida) para los investigadores que aplicarán la intervención. La responsabilidad de la misma depende del comité de monitorización y ha de ser realizada por una persona ajena a la recogida de información y a la práctica de la intervención. La aleatorización intenta conseguir que todos los factores de confusión se distribuyan por igual en los grupos de estudio. Si el tamaño de muestra es lo suficientemente grande, un experimento bien diseñado puede esperar controlar todos los efectos de distorsión producidos por factores de riesgo extraños al estudio (sean o no de confusión), incluidos aquellos que no sean medidos. El control de factores de confusión por la aleatorización es directamente proporcional al tamaño muestral. A mayor tamaño, menor posibilidad de que un factor extraño persista. Se ha sugerido que en un estudio de 10.000 sujetos (5.000 en cada grupo), por regla general, es muy improbable que exista influencia de variables extrañas, pero ello depende del factor que se estudia. La aleatorización, en definitiva, aumenta la validez interna de un estudio. El control de factores extraños por la aleatorización es difícil de realizar en los estudios comunitarios, donde puede ser prácticamente imposible la manipulación de ciertos factores y características ambientales.

La aleatorización total es el procedimiento normalmente recomendado, pero presenta algunos problemas. El primero es la falta de aceptación por parte de los sujetos el no saber a priori el tratamiento que van a recibir. Esto causa una reducción en el tamaño de muestra que puede ser importante. El segundo de los inconvenientes es que la aleatorización total normalmente produce grupos desiguales en el tamaño muestral, tanto más cuanto más reducido es el número de participantes. Se han propuesto otros esquemas de aleatorización para obviar estos inconvenientes.

Enmascaramiento

Los estudios experimentales son especialmente sensibles al sesgo de información: a la colaboración por parte de los participantes (especialmente en los ensayos clínicos) o a que los investigadores recojan diferencialmente la información en los grupos que se comparan con la finalidad (consciente o inconsciente) de comprobar sus ideas. El enmascaramiento puede ser del observado, del observador y del analista. El del observado intenta evitar la influencia del conocimiento del grupo en la respuesta del participante; el del observador, que el que recoge la informa no se influya por ese dato (se ha comprobado que hay una tendencia a recoger los datos de manera diferencial, favoreciendo la hipótesis de trabajo); y el del analista, para evitar que la “tortura” de los datos les haga “cantar” en el sentido de la hipótesis. Estas tres opciones se suelen agrupar del modo siguiente:

1. Enmascaramiento simple o ciego simple: El sujeto participante no sabe a qué grupo pertenece, si al grupo experimental o al control. De esta manera se elimina la influencia que podría tener en el resultado el conocimiento que el propio sujeto tenga sobre el factor en estudio. Se aumenta así la comparabilidad entre grupos.
2. Enmascaramiento o ciego doble: En esta situación, además de la circunstancia anterior, el encuestador o persona que obtiene el resultado de las variables de respuesta no conoce el grupo al que pertenecen los sujetos. El doble ciego elimina la posibilidad de que el encuestador obtenga la información deseada de manera diferencial entre ambos grupos, al poner más interés (ya sea inconsciente o no) en un grupo que en otro. Esta es la forma habitual de realización de un estudio.
3. Enmascaramiento triple: En este caso, aparte de un doble ciego, el investigador que analiza los resultados no sabe cual es, entre los datos que se le ofrecen, el que mide el efecto que se analiza.

El enmascaramiento exclusivo del observador o del analista no se llama ciego simple.

TIPOS DE ESTUDIOS EXPERIMENTALES

Estudios experimentales puros o verdaderos

Su característica distintiva es que los sujetos participantes son distribuidos de forma aleatoria entre los grupos de intervención y el que actúa como control. Una de las debilidades de la mayoría de los diseños experimentales puros es el excesivo control de la situación en estudio, debilidad que paradójicamente constituye su fuerza. Otro de los inconvenientes, en cierto modo derivado del anterior, es que las experiencias suelen tener lugar en asentamientos artificiales, pudiendo diferir la población de estudio de la de referencia en algunas características. Ambos hechos suponen que la validez externa (extrapolación de resultados a la población de referencia y a otras poblaciones) se ve mermada. Si el verdadero efecto de una manipulación depende de estas características diferenciales, el efecto observado en la población de estudio puede diferir del efecto que se encontraría en la población diana. Esta diferencia puede tener implicaciones importantes en las actividades de salud pública que se basan en hallazgos experimentales. El esquema general de un diseño experimental podría ser el siguiente:

1. Definición de la población a la que se pretende aplicar los resultados obtenidos en la investigación y establecimiento de los criterios de elegibilidad de la muestra.
2. Selección de la población experimental. De la población diana se elige una muestra representativa, de manera general, mediante aleatorización.
3. A los sujetos de la muestra elegida se les propone la participación tras la explicación de los objetivos del estudio. Tras su aceptación o rechazo se constituye la población de estudio, sobre la que se desarrollará la experiencia.

4. Los sujetos participantes se dividen en dos grupos mediante aleatorización, el experimental (receptor del factor que hay que analizar) y el control o testigo.
5. Puesta en marcha del programa. Administración del factor en estudio en el grupo experimental mientras se administra un placebo, otro factor o nada al grupo control.
6. Lectura del resultado del programa y comparación entre los grupos constituyentes de la experiencia. Observación y valoración de la respuesta de acuerdo con los criterios elegidos. Comprobación de si los resultados esperados del tratamiento han sido alcanzados o no. Existe la posibilidad de que en esta fase los sujetos, por las razones que fueren, cooperen o no en el desarrollo de la investigación. Por ello, cada grupo será subdividido en dos, cooperantes y no cooperantes, ya que la medición del efecto puede estar influida por la actitud de los sujetos participantes y así se introducen errores en la valoración general.
7. Ponderación del posible papel de los sujetos en los que no se pudo realizar la medición del efecto, con independencia de su grado de cooperación. En este estadio los sujetos se clasifican a su vez en otros dos subgrupos, según se haya podido o no realizar la medición del efecto.
8. Análisis de los resultados obtenidos. Es en esta fase cuando únicamente se debe efectuar la evaluación estadística de las posibilidades de generalización de la medida estudiada a la población diana.

Estudios cuasiexperimentales

Estos diseños son más frecuentes en la epidemiología ambiental. La principal diferencia con los estudios experimentales puros reside en que no hay aleatorización en la distribución del factor que se analiza. En este tipo de estudios, el investigador u otras circunstancias no aleatorias determinan qué sujetos recibirán la medida que va a evaluarse y quiénes no. La ventaja más importante que presentan con respecto a los diseños experimentales puros es el menor número de obstáculos prácticos que hay en su realización. Su coste también es menor, sobre todo cuando el tamaño muestral es grande. En algunos casos pueden ser la única alternativa a los estudios experimentales. Tienen, sin embargo, el inconveniente de que la falta de aleatorización supone un menor control sobre los factores extraños (de confusión, etc.), que pueden interferir en el resultado final del análisis. Desde el punto de vista del diseño, presentan diferentes opciones en el modo en que se realiza la comparación.

1. Comparación interna, en la que cada unidad experimental sirve como testigo propio, por la observación de la variable de respuesta antes y después de uno o más tratamientos. Se realizan en estudios históricos en los que se comparan los efectos de un tratamiento anterior con los obtenidos con un tratamiento actual, o la situación previa al tratamiento con el resultado tras el mismo. Es un tipo de comparación poco valorada en el terreno científico, ya que presupone que la población presenta características similares en el pasado y en el presente para sacar alguna conclusión válida. Por desgracia, esto suele ser falso en la mayoría de las circunstancias. Las poblaciones se encuentran sometidas a un cambio y evolución continuos, que pueden deberse o no al tratamiento.

2. Comparación externa, en la que, de manera similar a lo que sucede en los estudios experimentales puros, el grupo índice o experimental es comparado con otro que se comporta como testigo o control.

ANÁLISIS DE LOS RESULTADOS

Al ser estudios prospectivos, todo lo que fue comentado para los estudios prospectivos de cohortes es aplicable a este tipo de diseños, en lo que se refiere al análisis epidemiológico. En los estudios con aleatorización se debe comprobar si ha sido eficaz en la eliminación de factores que puedan influenciar el resultado. En este sentido, es obligado que los autores de un estudio experimental ofrezcan al inicio de sus resultados una tabla en la que se comparen el grupo experimental y el control con respecto a las variables que más podrían influir en el resultado final, aparte del factor en estudio. Esta tabla también debe figurar en los restantes diseños experimentales, para valorar por parte del lector la comparabilidad de los grupos.

En los estudios aleatorizados el control del sesgo de confusión depende de que en el análisis se mantenga el principio de «intención de tratamiento» (*intention-to-treat*), en el que cada individuo permanece en el grupo al que fue asignado inicialmente, con independencia del tratamiento que finalmente haya recibido y de si lo completó o no.

EL ENSAYO CLÍNICO

Sir Austin Bradford Hill, uno de los padres del ensayo clínico tal y como hoy se entiende, lo definió como "un experimento cuidadosa y éticamente diseñado con el fin de poder contestar a preguntas concretas formuladas previamente". Pocock matiza un poco más y lo define como una experiencia científica en la que un tratamiento u otra medida se administra a los sujetos sanos o enfermos con intenciones de diagnóstico, terapéutica o prevención primaria. La ley del Medicamento (BOE 22-XII-1990) lo define de forma similar a Pocock «Evaluación experimental de una sustancia o medicamento, a través de su administración o aplicación a seres humanos, orientada hacia alguno de los siguientes fines: (a) poner de manifiesto sus efectos farmacodinámicos...; (b) establecer su eficacia para una indicación terapéutica, profiláctica o diagnóstica determinada; y (c) conocer el perfil de sus reacciones adversas y establecer su seguridad.» Estas definiciones no establecen que el ensayo clínico tenga que ser controlado, aleatorizado o enmascarado. La definición excluye la valoración de aspectos medioambientales. No obstante, para los interesados en los aspectos de la metodología, conviene recordar que este diseño es uno de los terrenos en los que más se publica en metodología y análisis (sobre todo debido a matemáticos) y que muchos de sus avances son aplicables a otros aspectos, incluido el ambiental, y que en el momento actual se les considera el paradigma de la evidencia científica, el modelo al que hay que aproximarse. En España hay dos publicaciones que lo regulan: la Ley del Medicamento de 1990 y el Real Decreto 561/1993.

ESTUDIOS EXPERIMENTALES EN LA COMUNIDAD

Este tipo de estudios tienen un asentamiento comunitario. Su principal objetivo es producir un cambio a nivel de la población general. Antes de la realización de un estudio de intervención comunitario, al igual que en los ensayos clínicos, deben haberse realizado estudios preliminares (si proceden) en el laboratorio, y opcionalmente en voluntarios, que garanticen de alguna manera la inocuidad de la manipulación que se pretende.

Este tipo de diseños suelen ser cuasiexperimentales, en los que una o varias comunidades recibirán la intervención, mientras que otras servirán como control, en las que se observará el desarrollo natural del proceso que se investiga. En caso de existir aleatorización, son las comunidades enteras y no los sujetos las que se distribuyen en grupo experimental y control. Esto no garantiza en absoluto que la distribución de todos los factores que puedan influir negativamente en el estudio sea similar entre los grupos comunitarios (en realidad, depende de la propia similitud de las colectividades seleccionadas por el investigador). Lo único que elimina es la decisión por parte del investigador de aplicar el tratamiento que ha de estudiarse en una comunidad u otra. Por todo ello se afirma que su diseño es cuasiexperimental y que arrastra un mayor riesgo de sesgo de selección en la formación de los grupos experimental y control. Por su asentamiento comunitario no existen los problemas derivados del consentimiento individual para la participación en la investigación. En este caso, la autorización y el apoyo han de ser dados por los representantes y autoridades de las colectividades involucradas.

La comunidad propuesta para participar como grupo de intervención debe ser representativa de la población en que se aplicarán los resultados obtenidos. Para algunos autores es particularmente importante que el área escogida no tenga recursos excepcionalmente buenos. Si la intervención es satisfactoria en una colectividad con recursos inferiores a la media, se podrá concluir que será factible la introducción del mismo tipo de medidas en otras zonas. Además de considerar los recursos de la comunidad elegida, se recomienda que las experiencias se lleven a cabo en zonas con un riesgo elevado de enfermar por el proceso en estudio y, en su caso, durante el período de incidencia máximo de la enfermedad. La comunidad debe ser informada de los objetivos de la experiencia por razones de ética y porque se conseguirá con ello una mayor cooperación en la investigación.

La comunidad de referencia, que debe ser similar en todos los aspectos a la comunidad de intervención o experimental, ha de estar separada de esta última para evitar contaminaciones por las maniobras realizadas. Si se realiza una campaña de educación sanitaria y las comunidades participantes conviven, es muy fácil que miembros de la comunidad de referencia se "beneficien" de la intervención, sesgando el resultado final. El grupo de referencia debe representar el desarrollo espontáneo del proceso, que incluye los cambios producidos en la sociedad a lo largo de la duración del estudio. Esto no quiere decir que a la comunidad de referencia se la vaya a privar de los adelantos que se produzcan en relación con el proceso bajo estudio a lo largo de la duración de la investigación. Razones éticas lo prohíben.

La duración del seguimiento es importante, ya que un período de tiempo corto puede no ser suficiente para que ocurran cambios permanentes, mientras que una duración demasiado larga del estudio puede conducir a una reducción de las diferencias esperadas entre el grupo experimental y el de referencia. Los cambios en la conducta sanitaria, por ejemplo, se detectan con más rapidez, mientras que los cambios en los niveles de exposición frente a los factores de riesgo suceden más tarde. Los cambios en las cifras de incidencia o mortalidad son los más tardíos en producirse.

En este tipo de estudios, no suele ser rentable la valoración de un determinado aspecto o componente (estudios unifactoriales). Por ello suelen realizarse estudios que abarcan varios factores (estudios multifactoriales).

ÉTICA Y EPIDEMIOLOGIA EXPERIMENTAL

La utilización del personal humano como soporte de la experimentación puede tener resultados beneficiosos y/o nocivos. Con el fin de reducir a un mínimo el riesgo, los estudios experimentales están sometidos a unas reglas generales de ética médica. Los requisitos éticos generales fueron subrayados en la Declaración de Helsinki de 1960, acordada por la Asociación Médica Mundial, y posteriormente revisados en octubre 2000 (disponible en www.wma.net). Dicha declaración hace referencia a los siguientes hechos esenciales.

1. La investigación biomédica debe realizarse con los más estrictos principios científicos, formularse claramente en un protocolo, ser dirigida por personas cualificadas --en las que descansa la responsabilidad del estudio-- y ser sometida a un comité superior para su aprobación.
2. Dicha investigación no deberá emprenderse a menos que la importancia de los objetivos supere los riesgos de la intervención, siendo estos últimos predecibles.
3. Los sujetos participantes deben ser completamente informados acerca de los objetivos, métodos, riesgos y beneficios del programa ... Debe obtenerse su consentimiento (o de la persona responsable si son menores o deficientes), manteniendo la libertad de abandono del estudio en cualquier momento y sin interferir esto último con la relación sujeto-médico.

Sin embargo, la misma declaración recoge en otro apartado que, si el investigador considera esencial no obtener el consentimiento, debe especificar sus razones y someterlas a un comité independiente para que éste decida.

La población humana no constituye un campo abonado para la comprobación de las aptitudes metodológicas, ni un medio para la experimentación no controlada de programas de intervención epidemiológica. La intervención epidemiológica constituye un acto a nivel de toda la población, lo que amplifica cada error cometido en el acto individual. Sin embargo, es menos directamente perceptible, ya que no existe la personalización del "cara a cara" de la relación individual. Por todo ello, es más dependiente de la propia ética del investigador.

En la epidemiología actual existen ciertas situaciones habituales que, aunque no son consideradas en demasía, chocan con un estricto sentido ético. Es el caso de la declaración obligatoria de enfermedades y el de la transferencia de información entre registros, en los que se antepone el bien colectivo al derecho de intimidad del individuo. De todas formas, sea el caso que fuere, la confidencialidad y el secreto profesional deben ser mantenidos y asegurados en el conjunto del estudio y entre sus ejecutores, los investigadores.

ESTUDIOS ECOLÓGICOS

Concepto y organigrama

Es un diseño incompleto que tiene la particularidad de que toma como unidad de análisis al grupo y no al individuo. Esto significa que dentro de cada grupo no se conoce a un nivel individual el estado de exposición y de enfermedad, pero sí se conocen las frecuencias marginales de ambas variables dentro de cada grupo. Se puede ver gráficamente en la tabla siguiente:

| | Enfermos | No enfermos | |
|--------------|----------|-------------|-------|
| Expuestos | i_1 | i_2 | n_1 |
| No expuestos | i_3 | i_4 | n_0 |
| | m_1 | m_0 | n |

Los estudios ecológicos pueden analizar medidas de incidencia, prevalencia o de mortalidad. Lo último es lo más frecuente, ya que son los datos que con más frecuencia se encuentran de forma agregada. De ahí que en las ciencias sociales este tipo de estudios reciban el nombre de estudios ecológicos.

Se podrían considerar cuatro tipos fundamentales de estudios ecológicos:

1. El estudio ecológico exploratorio en el que hay k grupos definidos geográficamente, siendo k grande (> 10). No se obtienen datos sobre la exposición. El objetivo principal es la investigación de patrones especiales que pueden sugerir una etiología ambiental o hipótesis etiológicas más específicas. Suelen ser definidos también como estudios de mapeo.
2. Estudio ecológico de comparación: en este diseño muy común, hay k grupos (normalmente definidos geográficamente), siendo k grande (> 10). El objetivo principal es la valoración de asociaciones entre la frecuencia de la exposición y la frecuencia de enfermedad entre los distintos grupos. Al estar los datos de forma agregada --no conocer a un nivel individual la distribución de la exposición y enfermedad-- es posible la utilización de diferentes fuentes de datos. Por ejemplo, se puede utilizar los informes de declaración obligatoria de enfermedades para cada grupo y los resultados de una encuesta nacional para una determinada exposición. Ejemplo, el estudio entre las tasas de suicidio y la proporción de protestantes existentes en las distintas regiones de Prusia que realizó Durkheim a finales del XIX. Encontró que existía una relación directa entre el número de protestantes y la tasa de suicidios.

3. Estudio ecológico de tendencia: de forma típica suele haber un solo grupo, normalmente una cohorte dinámica geográficamente definida que es seguida a lo largo del tiempo. El objetivo es valorar una asociación posible entre los cambios de la frecuencia de la exposición y los cambios en la frecuencia de la enfermedad. La variable de exposición puede ser un valor agregado de una característica individual o puede ser la presencia o intensidad de un programa de intervención. Recibe también el nombre estudio de series temporales.
4. Estudio ecológico mixto: Este diseño es una mezcla de los dos anteriores. Hay k grupos definidos geográficamente, cada uno de los cuales es seguido a lo largo del tiempo. El objetivo es valorar la posible asociación entre los cambios en la frecuencia de la exposición y los cambios en la frecuencia de la enfermedad entre varios grupos. Se pueden hacer dos tipos de comparaciones entre grupos y dentro de cada grupo a lo largo del tiempo. Un ejemplo sería la relación entre la dureza del agua y la mortalidad por cardiovasculares entre los años 1948 y 1966 en 83 ciudades británicas:

Ventajas e Inconvenientes

a) Ventajas:

- * Relativamente barato y factible, sobre todo con datos secundarios (datos reunidos por otros, procedentes de estadísticas oficiales). Por ello es útil para la generación de hipótesis.
- * Es útil para la evaluación de intervenciones en la población, ya sean programas o políticas sanitarias.

b) Inconvenientes:

- * Tienen una capacidad limitada para realizar inferencias causales, ya que no se observa la asociación a nivel individual (hacia la que se intenta realizar la inferencia). Es la llamada falacia ecológica.
- * Es muy difícil de separar los distintos efectos observados de dos o más exposiciones, por la grandes correlaciones ecológicas entre las exposiciones. Es la multicolinealidad.
- * Ciertos datos, como los historiales clínicos o las actitudes, pueden no estar en forma agregada, o no ser fiables.
- * A menudo no se puede determinar si la exposición precedió a la enfermedad, incluso con datos de incidencia. Es la llamada ambigüedad temporal. En los estudios de tendencia, se tiene que asumir un lapso específico de tiempo entre los cambios en la exposición y el presumible efecto de la enfermedad.
- * En los estudios de tendencia, puede ser difícil el separar la influencia de los cambios en breve plazo de tiempo, tanto de la enfermedad como de la exposición, de los que suceden a

largo plazo. Además, no hay buenas soluciones analíticas para ello, casi todas ellas son muy discutibles.

Análisis

Se partirá de la tabla siguiente, para cada uno de los grupos que componen un estudio ecológico de comparación.

| | Enfermos | Población |
|--------------|----------|-----------|
| Expuestos | $i_j?$ | n_{1j} |
| No expuestos | $i_j?$ | n_{0j} |
| | m_{1j} | n_j |

donde j es cada comunidad, n_j = tamaño de la población a riesgo en la comunidad j

Sea $X = n_{1j}/n_j$ = proporción expuesta en el grupo j

$Y = m_{1j}/n_j$ = tasa de la enfermedad en el grupo j

Se asume un modelo lineal entre ambas variables

$$Y = \beta_0 + \beta_1 X$$

β_1 = pendiente de la recta y β_0 = ordenada en el origen (valor de Y cuando $X = 0$). Estos parámetros se calculan mediante el análisis de regresión de los mínimos cuadrados, que también da el valor del coeficiente de correlación ρ (el paquete estadístico más elemental los proporciona). Es importante señalar que el parámetro que indica asociación en la regresión lineal es la pendiente y no el coeficiente de correlación. Éste indica si la nube de puntos se ajusta a una recta y nada más. Cuanto más distinto sea de 0 el valor de la pendiente mayor es la asociación (si es >0 , indica asociación positiva y si es <0 negativa).

Se pueden estimar el RR y la DRe utilizando los resultados del análisis de regresión. Para ello, se comparará la tasa esperada en el modelo cuando todos los sujetos están expuestos (esto es, $x = 1$) frente a la tasa esperada cuando el nivel de exposición es 0 (esto es, $x = 0$). De esta forma se puede calcular :

$$RR = \frac{R_1}{R_0} = \frac{Y | x = 1}{Y | x = 0} = \frac{\beta_0 + \beta_1 \times 1}{\beta_0 + \beta_1 \times 0} = \frac{\beta_0 + \beta_1}{\beta_0} = 1 + \beta_1/\beta_0$$

$$DRe = (Y | x = 1) - (Y | x = 0) = \beta_0 + \beta_1 - \beta_0 = \beta_1$$

$$FAe = (RR - 1)/RR = \beta_1/(\beta_0 + \beta_1) = 1/(1 + \beta_0/\beta_1)$$

$$FAp = \frac{n_1/n \cdot (RR - 1)}{n_1/n \cdot (RR - 1) + 1} = \frac{n_1/n}{n_1/n + \beta_0/\beta_1}$$

Ejemplo de análisis para variables categóricas

Suponga los datos del ejemplo siguiente:

| Comunidad | Riesgo de enfermedad | Proporción de expuestos |
|-----------|----------------------|-------------------------|
| 1 | 2% | 10% |
| 2 | 3% | 20% |
| 3 | 4% | 30% |

En el análisis se aprecia una correlación perfecta $\rho = 1$, una β_0 de 0,01 y un β_1 de 0,1 (estos coeficientes se obtienen mediante un análisis de regresión lineal).

$RR = 0,1/0,01 + 1 = 11$. Una asociación muy fuerte

$DRe = 0,1$. La exposición supone 10% en valor absoluto del riesgo.

$FAe = (11 - 1)/11 = 0,909$ ó 90,9%. La exposición produce el 90,9% del total de casos que se dan en los expuestos.

Para calcular la FAp habrá que hacerlo para cada comunidad, ya que depende de la frecuencia de exposición y ésta cambia en cada comunidad:

FAp en la comunidad 1 ($n_1/n = 0,1$ ó 10%) = $[0.1(11 - 1)]/[1 + 0.1(11 - 1)] = 0.5$ ó 50%

FAp en la comunidad 2 ($n_1/n = 0,2$ ó 20%) = $[0.2(11 - 1)]/[1 + 0.2(11 - 1)] = 0.667$ ó 66.7%

FAp en la comunidad 3 ($n_1/n = 0,3$ ó 30%) = $[0.3(11 - 1)]/[1 + 0.3(11 - 1)] = 0.75$ ó 75%

Se nota que conforme aumenta la frecuencia de exposición, aumenta la proporción de casos que en la comunidad se deben a la exposición.

Ejemplo de análisis con variables continuas

Basado en el artículo de Schatzkin et al. *Int J Epidemiol* 1989;18:28-31. En este estudio se analiza la relación entre consumo de alcohol (medido en calorías) y mortalidad por cáncer de mama en 46 países. Los datos del análisis de regresión son los siguientes:

$$\beta_0 = 11,02 \text{ por } 100.000$$

$$\beta_1 = 0,0414$$

$$\rho = 0,5, p = 0,002$$

β_0 indica la ordenada en el origen, la tasa de mortalidad del cáncer de mama en ausencia de consumo de alcohol.

Sabiendo el coeficiente β_1 asignado para cada unidad (en el caso del problema es calorías consumidas en alcohol), se puede estimar una tasa de mortalidad para un nivel determinado de consumo. Por ejemplo, ¿qué riesgo supone consumir dos cervezas y dos copas de vino al día? Si una caña tiene 200 ml, a 4°, una copa de vino son 125 ml de alcohol a 12°, la densidad del alcohol es 0,8 g/ml, el contenido en gramos será (2 cañas x 200 ml/caña x 0,04 ml alcohol/ml caña x 0,8 g/ml + 2 vinos x 125 ml/vino x 0,12 ml alcohol/ ml vino x 0,8 g/ml) de 36,8 g (1 g de alcohol = 7,1 calorías), y las calorías serán 261,28. Este valor de X se sustituye en la recta:

$$\text{Mortalidad de cáncer de mama por 100000 mujeres para 261,28 calorías} = 11,02 + 261,28 * 0,0414 = 21,84$$

Para estimar el RR hace falta conocer la mortalidad cuando el consumo es cero, $\beta_0 = 11,02$, que se divide por la anterior:

$$\text{RR} = 21,84 / 11,02 = 1,98 \text{ (casi duplica la mortalidad)}$$

$$\text{DRe} = 21,84 - 11,02 = 10,82 \text{ por cien mil mujeres.}$$

Falacia ecológica

Este error se produce por la propia agregación, que no refleja la asociación a nivel individual de distintas variables.

Puede originarse por un cambio en el nivel de exposición basal en diferentes regiones. Suponga el ejemplo comentado con anterioridad:

| Comunidad | Riesgo de enfermedad | Proporción de expuestos |
|-----------|----------------------|-------------------------|
| 1 | 2% | 10% |
| 2 | 3% | 20% |
| 3 | 4% | 30% |

Cualquier análisis apreciaría una correlación perfecta $\rho = 1$, una β_0 de 0,01 y un β_1 de 0,1. $\text{RR} = 0,1/0,01 + 1 = 11$.

En las tablas siguientes se ofrecen los datos a nivel de persona de cada una de las tres comunidades:

| | | Enf. | Total | Riesgo | RR |
|-------------|---------|------|-------|--------|----|
| | | (a) | (n) | (a/n) | |
| Comunidad 1 | Exp. | 2 | 100 | 2% | 1 |
| | No exp. | 18 | 900 | 2% | |
| | Total | 20 | 1000 | 2% | |
| Comunidad 2 | Exp. | 6 | 200 | 3% | 1 |
| | No exp. | 24 | 800 | 3% | |
| | Total | 30 | 1000 | 3% | |
| Comunidad 3 | Exp. | 12 | 300 | 4% | 1 |
| | No exp. | 28 | 700 | 4% | |
| | Total | 40 | 1000 | 4% | |

Se aprecia que el riesgo en los expuestos y no expuestos es idéntico en cada comunidad, aunque cambia con la comunidad. Esto significa que en cada comunidad la exposición no es un factor de riesgo, aunque el análisis ecológico sí revela una asociación, que es muy fuerte. Esta es la falacia ecológica, imprevisible además.

El uso de variables adicionales ecológicas, para ajustar por ellas, en el análisis multivariable no elimina el error y puede tener como consecuencia desgraciada el que aumente el error. La estandarización de tasas tampoco no elimina la confusión en un análisis ecológico.

VALORACIÓN DE UN INSTRUMENTO DE RECOGIDA DE INFORMACIÓN

Todos los investigadores recogen información y para ello utilizan diferentes instrumentos de medición: pruebas de laboratorio, encuestas, etc. ¿Cuál es la calidad de esas mediciones? Para valorarlas hay que tener en cuenta dos fuentes principales de variación: la del parámetro que se pretende valorar en la población (variación biológica) y la del propio instrumento de medición.

Variación biológica

- Cambios a lo largo del tiempo: una medida en un punto del tiempo intenta ser una estimación representativa de una muestra de medidas tomadas a lo largo de un tiempo. A veces es así, otras no. Para descartar un proceso a veces hay que hacer una valoración durante todo el periodo, para que sea eficaz, porque de otra manera la valoración puntual sería poco eficaz. Piénsese simplemente en la evolución del anhídrido carbónico. Una estimación puntual está demasiado sujeta a variaciones.

- Diferencias entre individuos/ambientes: En muchas ocasiones predomina este tipo de variación.

Sea cual sea la variación de la situación, ésta se ha estudiado a lo largo de la carrera profesional. Sin embargo, la fuente siguiente de variación se ha estudiado menos.

Variación del acto de la medición

Puede ser debida a dos hechos fundamentales: validez y reproducibilidad.

Validez (*validity, accuracy*): grado en el que los resultados de una medición corresponden al fenómeno que se mide, con la verdad. La medición de la prueba problema se compara con un estándar aceptado (*gold criterion, gold standard*). Por ejemplo, la concentración de un contaminante atmosférico (SO₂) medida por un aparato puede ser comparada con la medición obtenida por una muestra con una concentración conocida de SO₂ en un aparato recién calibrado.

Hay situaciones en las que no existen estándares; por ejemplo, cuando se elabora un indicador cuantitativo complejo de repercusión sobre el medio ambiente. La validez se puede valorar de tres formas distintas: a) validez del contenido de las items que lo contienen (es la más fácil en general); b) validez de la escala (*construct validity*): si se observa que varía de acuerdo con la intensidad del fenómeno que se observa; y c) validez del criterio: cuando se compara la medición con el fenómeno que predice y que es observable.

Reproducibilidad o fiabilidad (*reliability*): es el grado en el que las mediciones repetidas de un fenómeno coinciden. Si no hay reproducibilidad difícilmente se podrá valorar la validez de una prueba diagnóstica. La reproducibilidad es el paradigma del consenso. Con demasiada frecuencia no existen criterios ciertos para diagnosticar un proceso y se recurre a la coincidencia entre varios,

en definitiva al consenso. Por ejemplo, los criterios de fiebre reumática o artritis reumatoide. Todo consenso se modifica con el tiempo, cuando hay nueva información que hay que integrar.

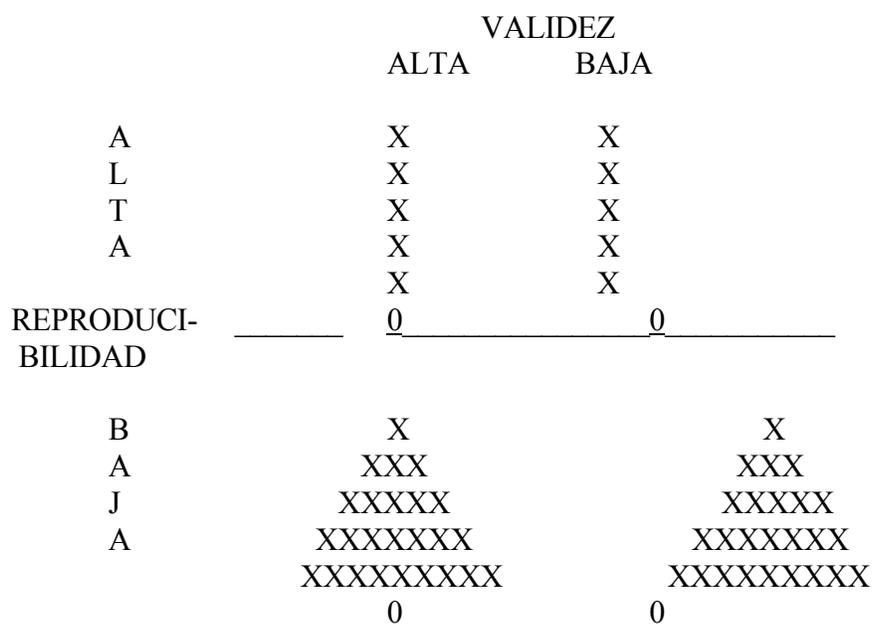


Figura 5. Diferencia entre reproducibilidad y validez.

Diferencia entre reproducibilidad y validez: Se puede apreciar en la figura 5. En ella, el valor de la prueba diagnóstica se mide en el eje de las abscisas, y se pretende medir el valor verdadero 0. Varias X en el mismo eje indican coincidencia de la medida. De la figura es fácilmente deducible (sobre todo tras la explicación en clase) que la diferencia entre el valor de la referencia y la media de la prueba es un indicador de validez, mientras que la desviación típica sería un indicador de fiabilidad.

Suponga que se quiere acertar en una diana disparando con una carabina (figura 6). Se dispara en varias ocasiones y los disparos se agrupan de la forma que se ilustra en la figura 6a. Todos se encuentran en el cuadrante superior izquierdo, pero bastante agrupados (la agrupación perfecta sería ver un solo orificio por donde han pasado todos los proyectiles). El tirador no es válido (es inexacto o sesgado), ya que en él se observa una tendencia a desviarse hacia arriba y la izquierda; pero es un tirador reproducible, suele dar en el mismo sitio. Probablemente cuando compruebe que sus disparos se han ido altos y a la izquierda corrija y la siguiente tanda sea mejor; el sesgo cuando se conoce es corregible.

Suponga ahora que la agrupación de los disparos es la de la figura 6b. En este caso se aprecia que los disparos se disponen más bien en la periferia de la diana. Está claro que no es un tirador reproducible, unas veces se le van arriba y a la izquierda, otras abajo y a la derecha, etc. Es

más, la falta de reproducibilidad es tan grande que dificulta la apreciación de si hay falta de validez: no se puede observar la existencia de alguna tendencia a desviarse del centro de la diana.

Los disparos pueden presentarse como en la figura 6c. En esta situación no están tan agrupados como en la figura 6a, pero no llegan al grado de dispersión de la figura 6b: este tirador es más reproducible que el de la figura 6b, pero menos que el de la 6a. También se aprecia que tiene una tendencia a desviarse del valor central de la diana, hacia arriba y la derecha, lo que refleja falta de validez.

Por último, la diana puede reflejar el aspecto de la figura 6d, muy agrupados y en el centro mismo de la diana. Es el ejemplo de un tirador que reúne las dos condiciones que se discuten, la reproducibilidad y la validez.

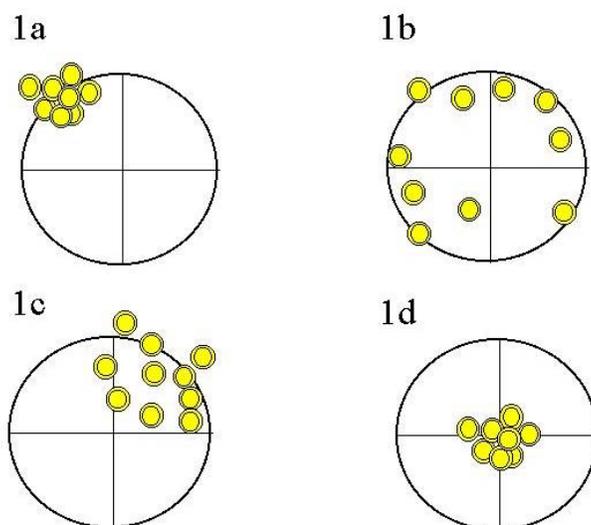


Figura 6. Validez y reproducibilidad (ver texto).

Tipos de reproducibilidad: Existen dos: del mismo observador y entre dos observadores.

Valoración de la reproducibilidad

Para variables continuas no se deben utilizar coeficientes de correlación: la existencia de correlación no indica que dos o más procedimientos coincidan (ver figura 7). Sería mucho mejor que se utilizara una representación gráfica en la que se enfrentara la media de los procedimientos a su diferencia. Se recomienda el uso de los coeficientes de correlación intraclase de acuerdo (y no de consistencia), cuya explicación va más allá de los objetivos del curso. Sería mejor utilizar la

diferencia entre las mediciones (eje de ordenadas) frente a la media de las mismas (eje de abscisas). También se puede calcular una media de las diferencias y su desviación típica, que permiten conocer el rango de concordancia en un 95% de los casos, por ejemplo. Para valorar el sesgo se puede utilizar el error estándar de la media: si el intervalo de confianza con él no incluye el cero, es indicativo de sesgo.

Variables categóricas. Normalmente la realización de una prueba diagnóstica implica la adopción de una decisión: es normal o no. Esto implica que cada variable continua puede ser transformada en una variable categórica y dicotómica. En esta situación en el análisis se parte de los datos de la tabla 9. Hay que recordar que lo que se representan son pares de observaciones. En esta tabla se pueden calcular varios parámetros:

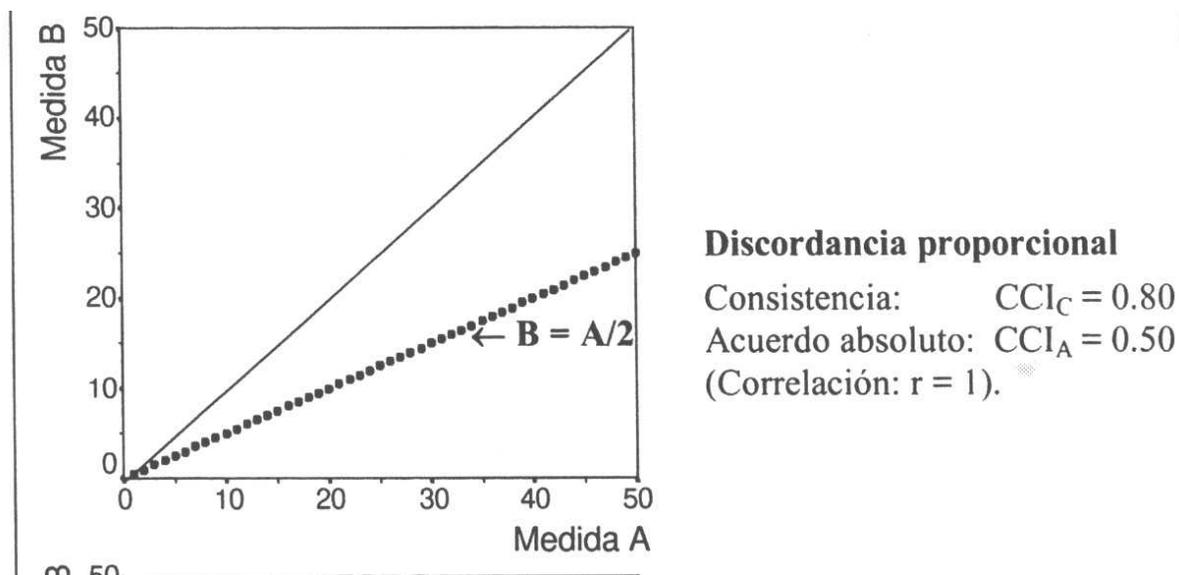


Figura 7. Concordancia de dos medidas: CCI_C : coeficiente de correlación intraclase de consistencia; CCI_A : coeficiente de correlación intraclase de acuerdo.

* Acuerdo específico

- Positivo: $2A/(2A+B+C)$

- Negativo: $2D/(2D+C+D)$

Tabla 9. Reproducibilidad entre dos observaciones para determinar la presencia de una condición

| | | Observación 2ª | | TOTAL |
|----------------|----------|----------------|---------|-------|
| | | Presente | Ausente | |
| Observación 1ª | Presente | A | B | E |
| | Ausente | C | D | F |
| | | G | H | N |

Se cuentan dos A porque son pares de observaciones y sólo se cuentan una B o una C, porque de la pareja sólo la mitad tienen el resultado deseado. Cuando la concordancia es perfecta los dos acuerdos valen 1 (ó 100%).

* Acuerdo total: $(A+D)/(A+B+C+D)$

El acuerdo total es un mal indicador resumen y debe ir siempre acompañado de los acuerdos específicos positivo y negativo. En los datos de la tabla 13-a, el acuerdo global vale un 70%, mientras que el acuerdo positivo es igual a 76,9% y el negativo a 57,1%. Si en vez de tener esa distribución, los datos de las casillas A, B, C, y D fueran 0, 15, 15 y 70, respectivamente, el acuerdo total sería nuevamente igual al 70%, pero el acuerdo positivo sería cero, lo que indicaría un grado de desacuerdo total en una de las opciones de la prueba.

En las pruebas categóricas es posible el no tener sólo dos resultados. Una prueba que se utiliza mucho en estas circunstancias (también incluso cuando hay sólo dos categorías) es el kappa.

* Prueba de concordancia de kappa.

Tabla 10. Distribución para calcular el kappa: valores absolutos

| | | Medida 2ª | | | | Total | |
|-----------|---|-----------|----------|----------|----------|----------|----------|
| | | 1 | 2 | ... | k | | |
| Medida 1ª | 1 | n_{11} | n_{12} | ... | n_{1k} | $n_{1.}$ | |
| | 2 | n_{21} | n_{22} | ... | n_{2k} | $n_{2.}$ | |
| | | . | . | . | . | . | |
| | | . | . | . | . | . | |
| | | k | n_{k1} | n_{k2} | ... | n_{kk} | $n_{k.}$ |
| Total | | $n_{.1}$ | $n_{.2}$ | ... | $n_{.k}$ | n | |

Tabla 11. Distribución de probabilidades para calcular kappa. Se obtiene dividiendo cada casilla de la tabla 10 por n.

| | | Medida 2 ^a | | | | Total |
|-----------------------|-------|-----------------------|-----------------|-----|-----------------|-----------------|
| | | 1 | 2 | ... | k | |
| Medida 1 ^a | 1 | p ₁₁ | p ₁₂ | ... | p _{1k} | p _{1.} |
| | 2 | p ₂₁ | p ₂₂ | ... | p _{2k} | p _{2.} |
| | . | | | | | |
| | . | | | | | |
| | k | p _{k1} | p _{k2} | ... | p _{kk} | p _{k.} |
| | Total | p _{.1} | p _{.2} | ... | p _{.k} | p |

En la tabla 11 se representan probabilidades. Si se utilizan los datos de la tabla 12-a, y se transforman en probabilidades, los datos serán los de la derecha en la mencionada tabla: cada casilla resulta de dividir el número de sujetos existente en ella por el total de sujetos (de ahí que $p = 100/100 = 1$). En la distribución de la tabla 9 se ha de calcular la concordancia global observada (P_o), que resultará de sumar las probabilidades observadas en la diagonal, probabilidades de concordancia para cada uno de los resultados de la prueba:

$$P_o = \sum p_{ii}$$

En el ejemplo de la tabla 12-a, $P_o = 0,5 + 0,2 = 0,7$. Hay que calcular una concordancia esperada (P_e), ya que hay una probabilidad de que las dos determinaciones coincidan simplemente por azar. Por ejemplo, si una prueba es positiva en el 40% y otra, independiente de la anterior, lo es en el 60%, la probabilidad de que coincidan resulta de multiplicar los dos valores anteriores ($0,6 \times 0,4 = 0,24$). Así, para calcular la concordancia esperada para cada uno de los valores posibles de las pruebas, se multiplica la probabilidad total de cada fila por la de su columna correspondiente:

$$P_e = \sum p_{i.} p_{.i}$$

En el ejemplo de la tabla 12-a, para el resultado positivo será $0,65 \times 0,65 = 0,4225$, ya que la probabilidad de un resultado positivo de la primera prueba (fila 1) es igual a la probabilidad de un resultado positivo para la segunda prueba (columna 1); para el resultado negativo será $0,35 \times 0,35 = 0,1225$. $P_e = 0,4225 + 0,1225 = 0,545$.

La concordancia observada intenta superar a la concordancia esperada. Cuanto más la supere, mejor será la concordancia. Hay un valor máximo de mejora sobre la concordancia esperada. Si la concordancia esperada es 0,545 y la máxima posible es 1, la máxima mejora observable sobre la concordancia esperada será $1 - 0,545$ ($1 - P_e$) = 0,455. Si la mejora apreciada sobre la concordancia esperada, que resulta de restar $P_o - P_e$, se divide por la máxima mejora sobre la concordancia esperada, obtendremos el kappa, que como máximo es 1.

$$K = (P_o - P_e)/(1 - P_e)$$

En el ejemplo de la tabla 12-a, $k = (0,7 - 0,545) / (1 - 0,545) = 0,34$.

Tabla 12. Ejemplos de cálculo de kappa.

| Prueba 1 | Prueba 2 (sujetos) | | | Prueba 2 (probabil.) | | |
|---------------------|--------------------|----|-------|----------------------|------|-------|
| | + | - | Total | + | - | Total |
| Situación a. | | | | | | |
| + | 50 | 15 | 65 | 0,5 | 0,15 | 0,65 |
| - | 15 | 20 | 35 | 0,15 | 0,20 | 0,35 |
| Total | 65 | 35 | 100 | 0,65 | 0,35 | 1 |
| Situación b. | | | | | | |
| + | 5 | 15 | 20 | 0,05 | 0,15 | 0,2 |
| - | 15 | 65 | 80 | 0,15 | 0,65 | 0,8 |
| Total | 20 | 80 | 100 | 0,2 | 0,8 | 1 |
| Situación c. | | | | | | |
| + | 50 | 25 | 75 | 0,5 | 0,25 | 0,75 |
| - | 5 | 20 | 25 | 0,05 | 0,20 | 0,25 |
| Total | 55 | 45 | 100 | 0,55 | 0,45 | 1 |

Un kappa $> 0,75$ es excelente; $< 0,4$ pobre [Fleiss, 1981]. Para Landis-Koch $> 0,8$ muy bueno, 0,61-0,8 bueno, 0,41-6 moderado, 0,21-0,4 bajo, $< 0,21$ pobre.

El kappa se influye por la frecuencia de los resultados positivos y negativos (influidos a su vez por la frecuencia del proceso que se valora). Se puede ver con el ejemplo de la tabla 12-b. En este caso, la frecuencia de positivos para las dos pruebas es del 20%, $kappa = 0,06$, siendo la concordancia global P_o similar a la tabla 12-a, 0,7.

En la situación de la tabla 12-c se puede ver también la presencia de sesgo (existencia de inclinación a un cierto tipo de resultado). Si se observan las casillas discordantes, hay un desequilibrio entre las dos pruebas, en la prueba 1 se observan más positivos. Esto puede indicar que la prueba 1 da más resultados positivos que la 2, o que la 2 da menos resultados positivos, o una mezcla de las dos anteriores. La ausencia de referencia impide decidir entre las tres situaciones anteriores. El kappa en la tabla 12-c es 0,37, el sesgo también influye en el valor de kappa. Obsérvese que la concordancia global P_o es la misma que en la 12-a, 0,7.

La influencia de la prevalencia en el kappa indica que la concordancia será mayor entre dos observadores con las mismas herramientas cuando la enfermedad es más frecuente (es excepcional que pase del 50% la frecuencia).

El kappa puede ser bajo a pesar de existir una buena concordancia, por ello se recomienda no olvidar los índices comentados con anterioridad, y proporcionar los acuerdos específicos positivo y negativo de manera separada.

Validez de una prueba diagnóstica

Cuando se aplica una prueba lo que interesa saber es cual es la probabilidad de acertar cuando una prueba diagnóstica les da un resultado positivo o negativo. Hay que comprobar el grado con esa prueba se acerca a la "verdad". Esto es la validez.

La "verdad" viene determinada por una prueba (o conjunto de ellas) que indefectiblemente confirman la presencia o ausencia de una determinada alteración. Esto es lo que se ha llamado el *gold standard* o *gold criterion*. Existen no obstante problemas en numerosas ocasiones. Para valorar una prueba se recurre normalmente a una tabla 2x2 similar a la siguiente

Tabla 13. Validez de una prueba diagnóstica.

| | | Criterio de verdad | |
|--------|---|--------------------|----|
| | | + | - |
| Prueba | + | VP | FP |
| | - | FN | VN |

VP: verdaderos positivos; VN: verdaderos negativos; FP: falsos positivos; FN: falsos negativos

Sensibilidad: la sensibilidad se define como la probabilidad que tiene una prueba de dar resultados positivos entre los sujetos o muestras que tienen la condición problema:

$$S = VP/(VP+FN)$$

Especificidad: La especificidad se define como la probabilidad que tiene una prueba de dar resultados negativos entre los sujetos o muestras que no tienen la condición problema:

$$E = VN/(VN+FP)$$

Relaciones entre especificidad y sensibilidad. Curvas *ROC*

Existe por regla general una relación inversa entre la sensibilidad y la especificidad de una prueba diagnóstica. Para entender mejor esto suponga el ejemplo siguiente (figura 8), en el que se observan las distribuciones poblacionales de los sujetos afectados y sanos en lo referente a un parámetro. Es fácil entender que conforme nos movamos más hacia la derecha en el "valor de la

prueba" (es decir, éste sea más alto), menor será el número de FP (la especificidad de la prueba será mayor) y mayor el número de FN (la sensibilidad de la prueba será menor).

Figura 8. Distribución de afectados y sanos según una medición.

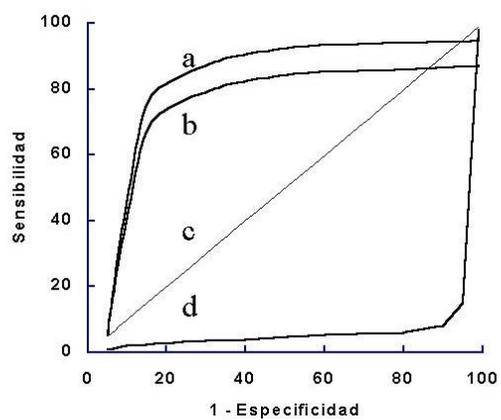
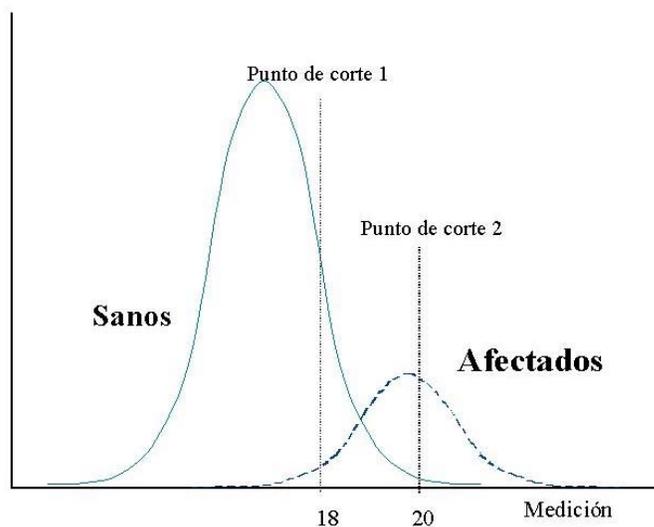


Figura 9. Ejemplos de curvas ROC.

La curva ROC enfrenta la sensibilidad en el eje de las ordenadas frente a la tasa de falsos positivos (uno menos la especificidad) en el eje de las abscisas. Lo que queda por encima de la curva son fallos (FP y FN) y lo que queda por debajo aciertos. Las curvas ROC pueden utilizarse para calcular el punto mejor de corte de una prueba diagnóstica. También permiten comparar diversas curvas diagnósticas entre sí. Dado que el área que queda bajo la curva son los diagnósticos válidos, cuando una curva queda por debajo de otra es peor (la curva *a* es mejor que la *b*). No obstante, de las curvas de la figura 9, la peor es la *c* (bisectriz de la figura), que es como lanzar una moneda al aire. La mejor de todas las curvas es la *d*, y lo que hay que decirle al autor de esa curva es que lo positivo lo considere negativo y viceversa.

Utilidad de la S y de la E

La S y la E son parámetros de validez interna de una prueba. Teóricamente, no deberían variar: una vez establecidos debería ser igual su aplicación en EE.UU. que en China. Sin embargo, ambos parámetros dependen del espectro y del número de situaciones en las que se valoraron. Conforme el número sea mayor, la posibilidad de un error aleatorio disminuye: la estimación es más robusta. Con respecto al espectro de la situación: es fundamental cuando se comunica la S y la E de una prueba el establecer perfectamente el tipo de situaciones en los que se hizo. La relación entre S y E motiva el que para una serie de situaciones se requiere una prueba altamente sensible o muy específica. En general, se quiere una prueba sensible cuando:

- El proceso es grave y no puede permanecer ignorado
- Es tratable
- Los resultados falsamente positivos no entrañan traumatismos psicológicos o económicos en los sujetos/colectividades examinados.

Se desea una prueba muy específica cuando:

- El proceso es serio pero difícilmente tratable.
- El hecho de saber que no se tiene el proceso posee una importancia sanitaria y psicológica.
- Los resultados falsamente positivos pueden ser psicológica y económicamente traumatizantes.

VALORES PREDICTIVOS

Todos los comentarios y afirmaciones dichas y que siguen a continuación parten de que la tabla tetracórica es representativa de una población, esto es, no se han seleccionado por una parte las situaciones problema y por otra las no afectadas.

Hasta el momento presente se ha hablado de lo que representa la S y E de una prueba diagnóstica. Pero ésta es una información de la que habitualmente no dispone el profesional y quizá no esté interesado. Los profesionales en lo que están interesados es en saber la probabilidad que se tiene de equivocarse al afirmar que una situación no existe cuando una prueba da un resultado positivo, y viceversa. Esto es lo que se conoce con el nombre de **VALORES PREDICTIVOS**. Se les considera como probabilidades "posteriores", ya que valoran la probabilidad de tener un proceso una vez que la prueba ha proporcionado un valor, positivo o negativo. La prevalencia de la situación sería la probabilidad anterior.

Valor predictivo del resultado positivo de la prueba (VPP). Se define como la probabilidad que tiene una prueba de detectar la situación problema cuando la prueba da un resultado positivo. En la tabla clásica, su valor sería:

$$VPP = VP/(VP+FP) =$$

Aplicando el teorema de Bayes, se demuestra que el VPP es igual a

$$= (P \cdot S) / [P \cdot S + (1-P)(1-E)]$$

Donde P = prevalencia, S = sensibilidad y E = especificidad. Cuanto más alto sea el valor, más exacta será la prueba en el medio en que se usa. Por ello, el VPP es el valor que gobierna el uso de una prueba. Recuerde que varíe con la prevalencia.

Valor predictivo del resultado negativo de la prueba (VPN). Se define como la probabilidad que tiene una prueba de detectar la ausencia de una situación cuando la prueba da un resultado negativo. En la tabla clásica, su valor sería:

$$VPN = VN/(VN+FN) =$$

Según el teorema de Bayes, el VPN es

$$= [(1-P)E] / [(1-P)E + P(1-S)]$$

Recuerde que varía con la prevalencia.

Valor global (*accuracy*). El valor global es la probabilidad que tiene una prueba de acertar. Algunos creen que es un parámetro de validez y no lo es:

$$VG = (VP + VN) / (VP + VN + FP + FN) = P \cdot S + (1-P)E$$

El VG también depende de la prevalencia. El valor global es un mal criterio porque puede ser alto, con una sensibilidad de cero. No tiene utilidad si no va acompañado de la sensibilidad y de la especificidad.

Relación entre valores predictivos y prevalencia

Conforme la prevalencia aumenta, manteniendo la S y la E constantes (como es de esperar):

- Aumenta el VPP
- Aumenta, en valor absoluto, los FN
- Disminuye, en valor absoluto, los FP
- Disminuye el VPN
- Las modificaciones del VG dependen además de la relación entre sí de la S y de la E.

El aumento del VPP con la prevalencia justifica simplemente la mayor capacidad diagnóstica cuando un proceso es más frecuente. También justifica la utilidad de unas pruebas en un ambiente frente a otro (donde es más infrecuente).