UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

Unidad Académica de los Ciclos Profesional y de Posgrado del CCH
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

DESCRIPCION DE DIVERSOS METODOS DE ANALISIS ESTADISTICO PARA LOS ESTUDIOS DE CASOS Y CONTROLES

T E S I S

Que para obtener el grado de:
Maestra en Estadística e Investigación de Operaciones
Presenta la Actuaria
Adriana M. Ducoing Watty

México, D. F. Noviembre 1988

INDICE

INTRODUCCION

El objetivo de este trabajo es presentar una descripción general de los métodos estadísticos utilizados para el análisis de estudios de casos y controles.

Para cumplir tal objetivo, el contenido del trabajo se ha dividido en cuatro capítulos. En el primer capítulo se hace una introducción general al campo de la epidemiología, con la finalidad de situar a los estudios de casos y controles en el contexto de esta área de la investigación médica; con ello, se pretende proporcionar las herramientas conceptuales básicas de dicha área, sin las cuales, no es posible realizar un análisis estadístico.

En el segundo capítulo se describen las características de los tipos de estudios que más frecuentemente se realizan en la investigación epidemiológica, y de los cuales el estudio de casos y controles es una de las múltiples opciones.

En el capítulo tercero se reseñan los aspectos específicos de los estudios de casos y controles, así como los análisis estadísticos que tradicionalmente se han utilizado.

Por último, en el cuarto capítulo, como alternativa a los análisis tradicionales, se presenta la utilización de modelos lineales, los cuales proporcionan, por un lado, una mayor flexibilidad en el análisis y, por otro lado, la potencialidad de utilizar paquetes estadísticos de uso común.

El contenido del trabajo se encuentra estructurado y conceptualizado de tal forma, que en algunas de sus partes sea accesible para personas del área de epidemiología, con conocimientos básicos de estadística y en general, sea ilustrativo e incentivador para personas que deseen adentrarse en las aplicaciones de la estadística en esta área de la investigación médica.

Cabe señalar que la motivación para estudiar las técnicas de análisis para estudios de casos y controles, surge del hecho de que este tipo de estudios son los más viables para investigar la etiología de muchas enfermedades crónicas, particularmente importantes en la actualidad.

CAPITULO I

CONCEPTOS BASICOS DE EPIDEMIOLOGIA

I.1. ¿QUE ES LA EPIDEMIOLOGIA?

La epidemiología estudia la distribución de enfermedades o de condiciones fisiológicas en poblaciones humanas y los factores que determinan tal distribución. En términos más generales, puede decirse que la epidemiología estudia la salud y enfermedad en poblaciones humanas.[1]

Para el conocimiento de la salud y enfermedad se requiere de las contribuciones de varias disciplinas: de las ciencias básicas (bioquímica, fisiología, patología, etc.), de la investigación clínica (ginecología, urología, etc.), de otras disciplinas cuantitativas (como la estadística) y de la medicina de población, preventiva o social.

De acuerdo a lo anterior y si se toma el concepto más amplio de epidemiología (estudio de salud y enfermedad en poblaciones humanas), un epidemiólogo puede ser, desde un individuo que investiga en un laboratorio, cómo detectar y controlar enfermedades, hasta el que realiza medicina de población.

En este trabajo, cuando se hable de epidemiología, se estará haciendo referencia a la investigación de población, es decir, al estudio de los determinantes biológicos, ambientales y de comportamiento, de la distribución de enfermedades[2] en poblaciones humanas, así como de su prevención.

I.1.1 Objetivos

La epidemiología estudia la distribución de enfermedades y sus determinantes para responder a cuatro objetivos generales (Kleinbaum D., Kupper L. y Morgenstern H.(1982), cap. 2):

* Describir el estado de salud de las poblaciones,
* explicar la etiología de las enfermedades,
* predecir la frecuencia de las enfermedades y el estado de salud de las poblaciones y finalmente,
* controlar la distribución de las enfermedades.

De lo anterior se deduce, que la investigación epidemiológica se lleva a cabo algunas veces para comprender un fenómeno exclusivamente y otras veces para

---

[1] El concepto de salud ha sido objeto de múltiples polémicas. La Organización Mundial de la Salud la define como un estado de completo bienestar físico, mental y social y no simplemente como la ausencia de enfermedad. Sin embargo, sigue siendo difícil cuantificar el estado de salud de una población; por eso comúnmente el estado de salud se mide mediante su complemento: la enfermedad y la muerte.

[2] Se habla de enfermedades, porque por un lado, como se mencionó anteriormente, medir el estado de salud es muy difícil y por otro lado, gran parte de los recursos se destinan al estudio de enfermedades, ya que representan problemas inmediatos a solucionar.

intervenir, tomando decisiones en materia de salud pública.

El objetivo de la investigación epidemiológica al nivel de comprensión es hacer generalizaciones acerca de la historia natural de una enfermedad, la cual puede dividirse en cuatro etapas (Mausner J. y Bahn A.(1977) cap. 1.):

Etapa de Susceptibilidad.- En esta etapa la enfermedad no se ha desarrollado pero se encuentran presentes los factores que favorecen su ocurrencia.

Etapa de la Enfermedad Presintomática o Preclínica.- En esta etapa no se ha manifestado la enfermedad pero ya se ha iniciado el proceso patológico.

Etapa de la Enfermedad Clínica.- En esta etapa ya ocurrieron los cambios anatómico-funcionales suficientes como para que haya signos y síntomas reconocibles de enfermedad.

Etapa de Resultado de la Enfermedad.- Esta es la etapa en que se manifiesta el resultado de la enfermedad, el cual puede ser, recobrarse, remitir, cambiar de severidad, incapacidad o muerte.

El objetivo de la investigación epidemiológica al nivel de intervención, es desarrollar prácticas, programas y políticas en materia de salud, con la finalidad de prevenir y controlar las enfermedades y promover la buena salud.

Las estrategias para la prevención de enfermedades se encuentran fuertemente vinculadas a las distintas etapas de la historia natural de la enfermedad. De tal modo que pueden identificarse tres diferentes niveles de prevención:

Prevención Primaria.- Es la prevención de enfermedades mediante la alteración de la susceptibilidad o la reducción de la exposición de individuos susceptibles. Esta prevención es la que se lleva a cabo en la etapa de susceptibilidad del desarrollo de una enfermedad.

Prevención Secundaria.- Consiste en detectar y tratar tempranamente las enfermedades. Con ésto, en ocasiones es posible curar la enfermedad en su etapa más temprana, o hacer lento su progreso. Este tipo de prevención se realiza al comienzo de la enfermedad, es decir, en las etapas preclínica y clínica.

Prevención Terciaria.- En este nivel, la prevención tiene como objetivo hacer menos severo el resultado de la enfermedad.

Cuando lo que se pretende es sólo prevenir enfermedades, no es necesario tener un conocimiento profundo acerca del proceso de la enfermedad para lograrlo, sin embargo, gran parte de la investigación epidemiológica gira en torno a la etiología de las enfermedades para conocer la historia natural de las mismas. La investigación etiológica consiste en la búsqueda de las causas de una enfermedad, las relaciones entre ellas y las magnitudes relativas de sus efectos en ella.

En virtud de la importancia que tiene el concepto de causa en la investigación epidemiológica, se discutirá brevemente a continuación.

I.1.2 Causalidad

El problema central sobre causalidad, es cómo determinar o bajo qué criterio decidir que un factor es causa de otro. Existen varios modelos que se han planteado para resolver este problema (Kleinbaum D., Kupper L. y Morgenstern H. (1982), cap. 2. y Mausner J. y Bahn A., (1977), cap. 2.).

Modelo de causalidad determinística unifactorial

Bajo este modelo se dice que X es causa de Y, si en un sistema donde todos los factores están inicialmente fijos, cualquier manipulación o cambio sólo en X induce un cambio subsecuente sólo en Y. Esta definición requiere que se cumplan dos condiciones:

* La especificidad de la causa.
* La especificidad del efecto.

La especificidad de la causa equivale a decir que "X es causa necesaria y suficiente para Y". La especificidad del efecto equivale a decir que X es causa de Y pero ningún otro efecto es causado por X.

Robert Koch (Kleinbaum D., Kupper L. y Morgenstern H. (1982), pag 27), en un intento por hacer operativo el criterio del determinismo puro, propuso tres condiciones para identificar agentes causales de una enfermedad:

* El agente debe estar presente en cada caso de enfermedad.
* El agente no debe ocurrir en otra enfermedad como un evento fortuito y no patológico.
* El agente debe aislarse del cuerpo en cultivo puro e inducir la enfermedad en un animal susceptible.

El modelo de determinismo puro y en consecuencia los criterios operacionales que de él se derivan, adolecen de muchas limitaciones, como son las que a continuación se describen:

* Las enfermedades por lo general tienen más de una causa, es decir, la etiología es multifactorial y por lo tanto, en cualquier caso, deberá violarse ya sea la condición necesaria o la suficiente para poder identificar una relación causal.
* Se ha encontrado que ciertos factores tienen más de un efecto patológico, es decir, que existe la multiplicidad de efectos y en consecuencia el modelo del determinismo puro es incapaz de abordar tales casos.
* Conceptualiza los supuestos factores causales de una manera limitada, porque por un lado, se afirma que la causalidad depende de un cambio en el factor causal y en tal caso, no queda claro el papel de características fijas como sexo, raza, predisposiciones genéticas, etc., y por otro lado, este modelo tampoco concede un papel claro a factores causales contínuos.
* Cuando se estudia el efecto hipotético de un factor, nunca se está seguro de que los efectos de otros factores no estén presentes, puesto que se tiene un conocimiento incompleto de la enfermedad y una habilidad limitada para

observar y medir el proceso causal.

Modelo de causalidad determinística multifactorial

Este es un modelo de determinismo modificado para permitir múltiples causas de una enfermedad.

En este modelo, se definen conglomerados de factores causales, los cuales son tratados como causas suficientes y cada conglomerado suficiente tiene un efecto en la enfermedad, que es independiente de los efectos de los factores que están en otros conglomerados.

En este modelo, una causa no necesita involucrar un cambio explícito, sino que puede ser un acto, evento o estado de la naturaleza que inicia o permite que el efecto ocurra.

El modelo no considera las limitaciones que existen respecto al conocimiento del proceso de una enfermedad y por lo tanto no alcanza a satisfacer las necesidades empíricas.

Otros Modelos Multicausales

En epidemiología se han elaborado otros modelos menos filosóficos o formales, que tratan de explicar la ocurrencia de una enfermedad en términos de las interaccciones entre factores que afectan el desarrollo de la misma. Estos factores se han dividido en:

Factores del Huésped (Intrínsecos).- Estos factores afectan la susceptibilidad a la enfermedad. Se cuentan entre ellos a factores genéticos, de personalidad, clase social e inmunidad específica.

Factores Ambientales.- Estos factores influyen en la exposición y a veces afectan también indirectamente a la susceptibilidad. Se cuentan entre ellos los factores biológicos tales como, agentes[3] infecciosos, reservorios de infección, plantas, animales, etc.; los factores sociales y los factores del medio ambiente físico tales como calor, luz, aire, agua, radiación, gravedad, presión atmosférica y agentes químicos de todas clases.

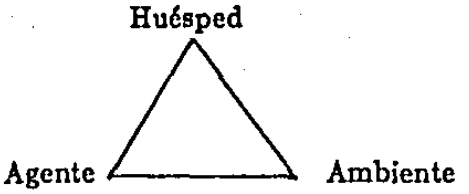
Como se mencionó anteriormente existen modelos que tratan de explicar la forma en que las interacciones entre los factores influyen en la manifestación de enfermedades:

Modelo del Triángulo Epidemiológico.- Consta de tres componentes: el huésped, el ambiente y el agente.

El modelo establece que cada componente debe ser analizado y comprendido para poder captar y predecir las modalidades de una enfermedad. Un cambio en cualquier componente altera el equilibrio existente, aumentando o disminuyendo la frecuencia de la enfermedad. En este modelo se le da mucha importancia al agente, el cual es parte integral del medio ambiente.

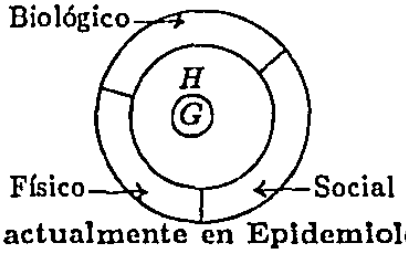
---

[3] Se le llama agente a un factor que siempre que ocurre una enfermedad, él está presente.



Modelo de la Multitud de Causas.- Fue propuesto por MacMahon B. y Pugh T., (1978). En este modelo, los efectos no dependen de causas aisladas sino de cadenas de causas. Desde este punto de vista, se puede evitar el desarrollo de una enfermedad, cortando la cadena causal en diferentes puntos y por lo mismo, no se requiere una comprensión completa de los mecanismos causales para adoptar medidas eficientes para la prevención y el control de las enfermedades.

Modelo de la Rueda.- Este modelo consiste de una rueda cuyo centro es el hombre (huésped), que tiene a su vez como centro su constitución genética. Rodeando al hombre se encuentra el ambiente, dividido en sus tres aspectos: biológico, social y físico. El problema específico que se aborde, determinará el tamaño relativo de las diferentes partes.



Modelo Causal utilizado actualmente en Epidemiología

En epidemiología no se utiliza en la actualidad un modelo causal determinístico, sino uno probabilístico, es decir, se utiliza teoría de probabilidad y técnicas estadísticas para probar y estimar a partir de observaciones, relaciones que se han hipotetizado como causales. Esto no significa que se niegue la existencia de un mundo determinístico, sino que simplemente se utiliza la teoría de probabilidad porque se tiene un conocimiento limitado del proceso causal y una capacidad limitada para observarlo.

En virtud del conocimiento limitado que se tiene del proceso causal, no se utiliza en epidemiología el término factor causal sino factor de riesgo, con el objeto de indicar que se trata de una variable que se "cree que está relacionada" con la probabilidad de que un individuo desarrolle la enfermedad.

En la práctica, se le llama factor de riesgo al agente cuyo efecto en el resultado de interés está siendo estudiado. Esto se hace cuando la exposición a dicho agente es accidental o incontrolable o cuando se aplica para otro propósito diferente al de afectar el factor resultado específico bajo consideración.

En cambio, se le llama factor tratamiento o tratamiento al agente cuyo efecto en el resultado de interés esta siendo estudiado, cuando dicho agente se aplicó específicamente para afectar el factor resultado bajo consideración.

Para elegir a un factor como factor de riesgo se debe satisfacer que éste esté

asociado estadísticamente con el desarrollo de la enfermedad (factor respuesta); que la presencia del factor (o un cambio relevante en él) preceda a la ocurrencia de la enfermedad y finalmente que el efecto del factor (asociación) que se observa no sea completamente explicado por cualquier fuente de error, incluyendo el error por muestreo, la presencia de otros factores de riesgo o en general cualquier problema originado en el diseño o análisis de la investigación.

Determinación de Causalidad en la Práctica

Para determinar si un factor es causa de otro factor, denominado resultado (puede ser por ejemplo la enfermedad), lo primero que hay que lograr es un buen estimador de la asociación que existe entre el supuesto factor causal y el factor respuesta.

Posteriormente se debe determinar si la asociación observada es estadísticamente significativa y de ser así, faltará determinar si existe una relación causal.

En realidad la inferencia causal depende de la sintetización de resultados de múltiples estudios, tanto epidemiológicos como no epidemiológicos. Sin embargo, se han elaborado algunos criterios operacionales que ayudan a determinar si una relación es causal o no lo es (Kleinbaum D.,Kupper L., y Morgenstern H., (1982), cap. 2; Anderson et al., (1980), cap. 14; MacMahon B. y Pugh T. (1978), cap. 2.). Estos criterios operacionales son:

* La fuerza de la asociación.- Mientras más fuerte sea la asociación, mayor es la evidencia de causalidad, es decir, es menos probable que la asociación observada pueda deberse a diversos tipos de errores y no al factor en estudio.

* Claridad en la secuencia de eventos en el tiempo.- Es importante establecer si la causa hipotetizada precedió la ocurrencia del factor resultado (la enfermedad). Esto resulta difícil de determinar en enfermedades con grandes períodos de latencia o factores de estudio que cambian a través del tiempo.[4]

* Existencia de una relación dosis-respuesta.- La existencia de una relación monótona entre el efecto o respuesta y la magnitud del factor causal hipotético, apoya la interpretación causal de la relación.

* Consistencia de los resultados.- Se refuerza la hipótesis causal si todas las investigaciones que estudian la misma relación, producen los mismos resultados y sobre todo si esas investigaciones involucran diferentes poblaciones, métodos y períodos de estudio.

* Factibilidad biológica de la hipótesis.- Si la relación especificada por la hipótesis, es congruente con los conocimientos biológicos que prevalecen, es probable que se acepte una interpretación causal. Sin embargo, debe quedar claro que no debe exigirse que la relación postulada deba ser totalmente explicada biológicamente, ya que el estado de conocimientos prevaleciente podría ser inadecuado para

[4] De la forma en la que se diseñe el estudio, dependerá en mucho la capacidad para determinar la secuencia temporal de los eventos.

explicar lo que se observa.

* Coherencia de los resultados.- Se fortalece una interpretación causal si los resultados obtenidos no entran en conflicto con el conocimiento existente, sobre la historia natural de la enfermedad o con otros hechos ya aceptados.

Es importante aclarar que ninguno de los criterios descritos anteriormente son determinantes y por lo tanto son únicamente auxiliares en la evaluación de factores de riesgo.

En realidad, así como David Hume afirmó, (Kleimbaum D. et al. (1982), cap.2), la inferencia causal depende de la intuición e imaginación humanas como vínculos en el proceso de conectar observaciones y teoría. La inferencia causal depende de algo más que de una prueba de hipótesis basada en un conjunto de observaciones.

I.1.3 Estrategias

La investigación epidemiológica se realiza con base en el método científico, es decir:

* Se examina la teoría y los hechos existentes.
* Se plantean hipótesis conceptuales nuevas o más específicas.
* Se plantean hipótesis operativas.
* Se obtienen observaciones para contrastar las hipótesis.
* Se obtienen conclusiones y se hacen inferencias.
* Se empieza nuevamente el ciclo.

Después de planteadas las hipótesis conceptuales, se tiene que determinar cómo contrastarlas, lo cual involucra al diseño de la investigación. Esta es una etapa clave, pues de ella dependerá esencialmente la validez de los resultados que se obtengan, es decir, si las observaciones obtenidas permiten o no contrastar la hipótesis planteada.

El diseño del estudio puede concebirse como la liga entre una o más hipótesis conceptuales y las hipótesis operacionales. En realidad el diseño del estudio dá origen a las hipótesis operacionales.

Las hipótesis conceptuales siempre diferirán de las operacionales, ya que éstas dependen del diseño del estudio, de aquí que se deba planear un diseño que genere hipótesis operacionales lo más cercanas a las hipótesis que se pretenden contrastar (hipótesis conceptuales). Cabe mencionar que la mayoría de los problemas serios o errores que se cometen en esta etapa, no pueden ser corregidos en etapas subsecuentes de la investigación.

Por lo anterior, el diseño del estudio tiene una importancia central en la investigación epidemiológica, ya qué de él dependerán las hipótesis que se puedan probar, y la calidad y validez de las observaciones obtenidas.

I.2 VALIDEZ DE LAS INVESTIGACIONES EPIDEMIOLOGICAS

Al hablar de la validez de un estudio, se hace referencia a la presencia o ausencia de sesgos originados en el diseño o en el análisis, que hacen que esté distorsionada la estimación del efecto que el supuesto factor causal tiene en el factor resultado.

Es importante hacer notar que el estimador de la medida de un efecto puede estar afectado por dos tipos diferentes de errores:

* Errores aleatorios
* Errores sistemáticos

Los errores aleatorios son errores que se originan precisamente por la naturaleza aleatoria del proceso de muestreo[5] y producen una diferencia entre el estimador de la medida del efecto, $\hat{\theta}$ y la medida del efecto que realmente se está estimando, θ_o (se toma a θ_o como el valor esperado de $\hat{\theta}$). Esta diferencia depende del tamaño de muestra, de las características estadísticas del estimador que se esté utilizando, y de la variabilidad de la característica de interés en la población bajo estudio. Los errores aleatorios se miden mediante la varianza del estimador $\hat{\theta}$, a la cual se hace referencia en múltiples ocasiones como la precisión del estimador.

Los errores sistemáticos, comúnmente conocidos como sesgos, se originan por problemas metodológicos en el diseño o análisis del estudio. Estos errores son diferentes a la variación aleatoria producida por el muestreo y dan origen a la diferencia entre lo que el estimador está realmente estimando, θ_o, y la verdadera medida del efecto de interés que se denotará por θ.

Esquemáticamente, el error total entre el estimador de la medida del efecto y la medida del efecto, se representa de la siguiente manera:

$$(\hat{\theta} - \theta) = \underbrace{(\hat{\theta} - \theta_o)}_{\text{errores aleatorios}} + \underbrace{(\theta_o - \theta)}_{\text{errores sistemáticos}}$$

En los estudios epidemiológicos suele ser más importante controlar la validez que la precisión. Es preferible un estudio libre de sesgos aunque restringido en alcances, a un estudio más general con problemas de validez irresolubles.

Debe observarse que las pruebas de significancia e intervalos de confianza son útiles para obtener el posible efecto de los errores aleatorios en la estimación de la medida del efecto, pero que tales técnicas no sirven para determinar el efecto de un error sistemático. Las pruebas de significancia servirán para inferir sobre la medida del efecto, siempre y cuando, hayan sido eliminados previamente los sesgos en el estimador.

Existen dos tipos distintos de validez en los estudios epidemiológicos: la interna y la externa.

---

[5] En múltiples ocasiones, no se realiza un proceso de muestreo aleatorio estrictamente hablando, sino que se dispone únicamente de un conjunto de elementos de una determinada población. En estos casos, para poder realizar inferencias, se actúa de cualquier forma, como si la muestra hubiera sido una muestra aleatoria de la población que representa.

Para poder definir lo que se entiende por validez interna y externa, es necesario definir primero varias poblaciones sobre las cuales existe interés en extender los resultados obtenidos:

Población real.- Es la población de donde se obtuvo finalmente la muestra.

Población objetivo.- Es la población que se quiere conocer y a la cual se quieren extender los resultados. De esta población se pretende tomar la muestra. En ocasiones, debido a características metodológicas del diseño, la muestra no resulta ser representativa de la población objetivo. Idealmente la población real y la objetivo coinciden.

Población externa.- Es una población en la cual no se realizó el estudio pero se está interesado en extender los resultados a ella, argumentando similaridades con la población objetivo.

Al hablar de validez interna de un estudio epidemiológico, se está cuestionando si el estimador de la medida del efecto de interés está libre de sesgos, de tal forma que los resultados pueden extenderse a la población objetivo. Es decir, si la población real y la objetivo coinciden y el estimador es insesgado, entonces existe validez interna.

Para hablar de validez externa es necesario presuponer que existe validez interna. El cuestionar la validez externa equivale a verificar si los resultados del estudio pueden extenderse a una población que no es la población estudiada. Es decir, existirá validez externa si la población externa es similar a la población objetivo.

I.2.1 Fuentes de Sesgo

Los sesgos se originan en tres fuentes principalmente:

* En la selección de sujetos para el estudio. A los sesgos resultantes se les conoce como sesgos de selección.
* En la medición de las variables de interés. A estos sesgos se les conoce como sesgos de información.
* Por la presencia de factores extraños que mezclan su efecto en el factor respuesta. A estos sesgos se les conoce como sesgos debidos a confusión o más propiamente sesgos introducidos por factores de confusión (o efectos confundidos).

Los sesgos de selección son distorsiones en el estimador que son consecuencia de la forma en la que se seleccionan los sujetos de la población. Muchos de estos sesgos surgen debido a defectos en el diseño del estudio, entre los cuales pueden mencionarse:

* Errores en la selección de grupos que van a ser comparados.
* Errores en la elección del marco de muestreo.
* Errores por no-respuesta y/o datos faltantes en seguimientos. Estos problemas se presentan esencialmente en investigaciones donde se estudia a los sujetos por un determinado período, es decir, se hace un seguimiento de los sujetos de estudio.

* Errores por supervivencia selectiva. Esto significa que las personas que desarrollaron la enfermedad y murieron antes del tiempo en el que se realiza el estudio, no se incluyen en la población de estudio y en consecuencia sólo se tiene información de las personas que logran sobrevivir.

* En estudios de casos y controles se presentan sesgos de selección, cuando el procedimiento para elegir casos y/o controles está relacionado con el estado de exposición.

Se deben hacer todos los esfuerzos posibles por evitar sesgos de selección cuando se está diseñando el estudio; el lograrlo depende del conocimiento que se tenga de las fuentes potenciales de sesgo.

Corregir los sesgos de selección en la etapa de análisis es muy difícil, porque para ello se requiere conocer las probabilidades de selección.

Los sesgos de información son distorsiones en el estimador, que surgen por imprecisiones en la medición de las variables de interés. Cuando las variables están en escala nominal u ordinal, al sesgo de información se le conoce como sesgo por mala clasificación.

Este tipo de sesgos surgen, entre otras cosas por la utilización de instrumentos de medición con defectos, por ejemplo un cuestionario, una entrevista o un índice derivado de ellos, que no mide la característica que se quiere medir. Surgen también por la utilización de una fuente de información con errores o incompleta. Por ejemplo, cuando la información de los sujetos bajo estudio se obtiene a través de expedientes con deficiencias.

También es fuente de sesgos de información el que se utilicen procedimientos para diagnóstico poco precisos, o estudios de seguimiento que presenten sesgos por mala clasificación. Ejemplo de ésto es el que se lleve a cabo una vigilancia con un procedimiento de diagnóstico desigual entre los distintos grupos que se desean comparar.

Al igual que los sesgos de selección, los sesgos de información deben de controlarse preferentemente en la etapa de diseño del estudio, pues corregirlos en la etapa de análisis, es muy difícil ya que se requiere de una estimación de la magnitud del sesgo, excepto en el caso de sesgo por mala clasificación, para el cual se puede estimar su magnitud en términos de dos parámetros muy utilizados que son la sensibilidad y la especificidad.

La sensibilidad para una determinada condición, es la probabilidad de que un sujeto que posee tal condición sea clasificado en el estudio como que la posee.

La especificidad es la probabilidad de que un sujeto que no posee la condición, sea clasificado en el estudio como que no la posee.

I.2.2 Factores de Confusión

Los factores de confusión son factores que anteceden al resultado (llamados factores antecedente), y que usualmente satisfacen dos condiciones:

* El factor antecedente debe presentarse de una manera diferente en los grupos que define el supuesto factor causal en estudio (grupos de riesgo o exposición), es decir, la distribución del factor antecedente debe variar en los diferentes grupos a comparar.

* Existe sustento conceptual para suponer que el factor antecedente afecta al resultado.

Para ilustrar estas dos condiciones se considerará un ejemplo. Supóngase que se quiere investigar si el riesgo (o probabilidad) de tener hijos de bajo peso al nacer, es mayor para mujeres que no recibieron cuidado prenatal durante el embarazo, que para mujeres que si lo recibieron (supóngase que se estudiarán a 100 mujeres que no recibieron cuidado prenatal y a 50 que si lo recibieron).

Por otra parte, un investigador sostiene que en base al conocimiento existente hasta el momento, hay razones suficientes para afirmar que las mujeres que fuman durante el embarazo tienen mayor probabilidad de dar a luz hijos de bajo peso, por lo cual postula a la condición de fumadora o no fumadora como un factor potencial de confusión, ya que en principio, se cumple la segunda condición arriba mencionada. Desde luego, no tendría sentido postular como factor de confusión potencial al hecho de que las mujeres usen o no broches de plástico en el cabello durante el embarazo, ya que no existe ninguna base teórica para suponer que esta situación pueda afectar al peso del producto.

Para que el hecho de fumar o no hacerlo sea en efecto un factor de confusión, se requiere además que se cumpla la primera condición anteriormente mencionada, es decir, que la distribución de fumadoras y no fumadoras sea diferente para mujeres que recibieron cuidado prenatal que para mujeres que no lo recibieron.

Ejemplo 1.- Supóngase que en realidad no existe diferencia en el riesgo de tener hijos de bajo peso entre mujeres que recibieron cuidado prenatal y mujeres que no lo recibieron pero que, como afirma el investigador, el hecho de fumar durante el embarazo sí afecta el peso del niño al nacer. Para ilustrar esto, en la siguiente tabla se suponen unas probabilidades de tener hijos de bajo peso según las diferentes categorías del factor de riesgo en estudio y del factor de confusión potencial.

Probabilidad de tener hijos de bajo peso

	No Fumadora	Fumadora
Sin Cuidado Prenatal	0.1	0.2
Con Cuidado Prenatal	0.1	0.2

Debe observarse que en este caso no existe diferencia en el riesgo de tener hijos de bajo peso por recibir o no cuidado prenatal.

Considérese en primera instancia que la distribución de fumadoras es igual en los dos grupos de riesgo: el 40% de las mujeres no fuma y el 60% si fuma.

Distribución de mujeres según factor de riesgo y factor antecedente

	No Fuman	Fuman	Total
Sin Cuidado Prenatal	40	60	100
Con Cuidado Prenatal	20	30	50

Si no se considera a la condición de fumadora para analizar el riesgo de tener hijos de bajo peso en cada grupo de riesgo y se comparan dichos riesgos, se obtiene que no existe diferencia:

Distribución de mujeres según factor de riesgo y respuesta

	Bajo peso	No bajo peso	Total
Sin cuidado Prenatal	16	84	100
Con cuidado Prenatal	8	42	50

Probabilidad estimada de tener un hijo de bajo peso

$$\text{Sin cuidado Prenatal} \quad \frac{40(0.1)+60(0.2)}{100} = \frac{16}{100} = \frac{4}{25}$$

$$\text{Con cuidado Prenatal} \quad \frac{20(0.1)+30(0.2)}{50} = \frac{8}{50} = \frac{4}{25}$$

Por lo tanto, en este caso, si no se considera al factor antecedente fumar o no en el análisis, no se alteran los resultados y ésto se debe a que al no cumplirse la condición de distribución[6] diferencial en los grupos de riesgo, tal factor antecedente no es de confusión aunque si afecte al factor respuesta en estudio.

Ahora considérese el caso en el que la distribución de fumadoras y no fumadoras difiere en los grupos de riesgo: 90% de las mujeres que no recibieron cuidado prenatal fuma y sólo el 20% de las mujeres que si recibieron cuidado prenatal fuma, lo cual se ilustra en la siguiente tabla:

Distribución de mujeres según factor de riesgo y factor antecedente

	No Fuman	Fuman	Total
Sin cuidado Prenatal	10	90	100
Con cuidado Prenatal	40	10	50

[6] Esta situación, como se analizará más adelante solo se cumple cuando el factor de riesgo no tiene efecto en el factor de respuesta

Entonces, al estimar la probabilidad de tener hijos de bajo peso para cada grupo de riesgo y compararlas se concluye erróneamente que sí hay una diferencia debido al cuidado prenatal aunque en la realidad no la hay:

Distribución de mujeres según factor de riesgo y factor de respuesta

	Bajo peso	No bajo peso	Total
Sin cuidado Prenatal	19	81	100
Con cuidado Prenatal	6	44	50

Sin cuidado prenatal $\frac{10(0.1)+90(0.2)}{100} = \frac{19}{100} = 0.19$

Probabilidad estimada de tener
un hijo de bajo peso

Con cuidado prenatal $\frac{40(0.1)+10(0.2)}{50} = \frac{6}{50} = 0.12$

Si se obtiene la diferencia de probabilidades, resulta que las mujeres sin cuidado prenatal tienen 7% más de probabilidad de tener un hijo de bajo peso, que las mujeres con cuidado prenatal, lo cual no es cierto. Esta diferencia, se debe a que fumar afecta al peso del producto y a que la distribución de fumadoras varía entre grupos de riesgo, lo cual determina que se confunda el efecto de cuidado prenatal con el efecto de fumar.

Es muy importante aclarar que el igualar la distribución del factor de confusión en los grupos de riesgo, no elimina necesariamente el efecto del factor de confusión en la estimación del efecto del factor de riesgo. El eliminarlo dependerá de la medida del efecto que se esté utilizando. Unicamente cuando no hay efecto del factor de riesgo en la respuesta, el igualar la distribución del factor de confusión, sí elimina su influencia en el estimador, independientemente de la medida que se esté utilizando.

En el ejemplo del efecto del cuidado prenatal en el peso del producto, se supuso que no existía en realidad ningún efecto de este factor de riesgo. En este caso, el igualar la distribución de fumadoras y no fumadoras en los grupos de riesgo, sí elimina al efecto de fumar o no fumar en el estimador del efecto del factor de riesgo en el bajo peso. Esto independientemente de la medida que se utilice. Es decir, si se considera por separado para fumadoras y no fumadoras, medidas del efecto del factor de riesgo tales como, diferencia de probabilidades de tener un hijo de bajo peso entre grupos de riesgo, cociente de estas probabilidades o la razón de momios[7], los estimadores serán los mismos que los que se obtendrían, al estimar tales medidas, sin considerar por separado

[7] Un momio se define como la probabilidad de que ocurra un evento, dividida por la probabilidad de su complemento. La razón de momios es el cociente de los momios de dos grupos, por ejemplo el momio para el grupo de mujeres que no recibieron cuidado prenatal entre el momio para el grupo de mujeres que si lo recibieron.

a fumadoras y no fumadoras (siempre que su distribución sea igual en los grupos de riesgo). A continuación se ilustra esto.

Ejemplo 2.

Distribución de mujeres según el factor de confusión y el factor de riesgo

	No Fumadoras	Fumadoras	Total
Sin cuidado Prenatal	40	60	100
Con cuidado Prenatal	20	30	50

Distribución de mujeres según el factor de riesgo
y según el factor de respuesta

	Bajo peso	No bajo peso	Total
Sin cuidado Prenatal	16	84	100
Con cuidado Prenatal	8	42	50

Estimador de la diferencia de probabilidades:

$$\hat{\Delta} = \frac{16}{100} - \frac{8}{50} = 0$$

Estimador del cociente de probabilidades:

$$\widehat{RR} = \frac{16}{100} / \frac{8}{50} = 1$$

Estimador de la razón de los momios

$$\hat{\psi} = \frac{(16)(42)}{(8)(84)} = \frac{672}{672} = 1$$

Distribución de mujeres según el factor de riesgo
y según el factor de respuesta para cada nivel
del factor de confusión

	No Fumadoras			Fumadoras		
	Bajo Peso	No Bajo Peso	Total	Bajo Peso	No Bajo Peso	Total
Sin cuidado Prenatal	4	36	40	12	48	60
Con cuidado Prenatal	2	18	20	6	24	30

$$\hat{\Delta}_{nf} = \frac{4}{40} - \frac{2}{20} = 0 \qquad \hat{\Delta}_{f} = \frac{12}{60} - \frac{6}{30} = 0$$

$$\widehat{RR}_{nf} = \frac{4}{40}/\frac{2}{20} = 1 \qquad \widehat{RR}_{f} = \frac{12}{60}/\frac{6}{30} = 1$$

$$\hat{\psi}_{nf} = \frac{(4)(18)}{(2)(36)} = 1 \qquad \hat{\psi}_{f} = \frac{(12)(24)}{(6)(48)} = 1$$

Se observa entonces, que controlando la distribución del factor de confusión en los grupos definidos por el factor de riesgo, cuando éste no tiene efecto sobre el factor respuesta, se elimina el efecto del factor de confusión en el estimador. Esto, independientemente de la medida del efecto que se utilice.

Por el contrario, cuando el factor de riesgo sí tiene efecto en la respuesta, el eliminar el efecto del factor de confusión en el estimador, al igualar su distribución en los grupos de riesgo, dependerá de la medida que se utilice. A continuación se ilustra esto.

Ejemplo 3.- Supóngase ahora que en el ejemplo anterior, sí hay efecto del cuidado prenatal en el bajo peso y que las probabilidades de tener hijos de bajo peso son:

Probabilidad de tener hijos de bajo peso

	No Fumadora	Fumadora
Sin cuidado Prenatal	0.5	0.9
Con cuidado Prenatal	0.1	0.5

Supóngase además que la distribución de fumadoras y no fumadoras por grupo de riesgo es la misma: 60% de fumadoras y 40% de no fumadoras.

Distribución de mujeres según factor de riesgo
y factor respuesta

	No Fumadora	Fumadora	Total
Sin cuidado Prenatal	40	60	100
Con cuidado Prenatal	20	30	50

La distribución de sujetos según el factor de riesgo y respuesta para cada nivel del factor de confusión es:

Distribución de mujeres según el factor de riesgo
y según el factor de respuesta para cada nivel
del factor de confusión

	No Fumadora			Fumadora		
	Bajo Peso	No Bajo Peso	Total	Bajo Peso	No Bajo Peso	Total
Sin cuidado prenatal	20	20	40	54	6	60
Con cuidado prenatal	2	18	20	15	15	30

$$\hat{\Delta}_{nf} = 0.5 - 0.1 = 0.4 \qquad \hat{\Delta}_{f} = 0.9 - 0.5 = 0.4$$
$$\widehat{RR}_{nf} = 5 \qquad \widehat{RR}_{f} = 1.8$$
$$\hat{\psi}_{nf} = 9 \qquad \hat{\psi}_{f} = 9$$

Sin controlar por la presencia del factor de confusión, se tiene que la distribución de sujetos según el factor de riesgo y respuesta es:

	Bajo Peso	No Bajo Peso	Total
Sin cuidado Prenatal	74	26	100
Con cuidado Prenatal	17	33	50

$$\hat{\Delta} = \frac{74}{100} - \frac{17}{50} = 0.74 - 0.34 = 0.4$$

$$\widehat{RR} = 2.18$$

$$\hat{\psi} = 5.52$$

Se observa entonces, que si se hubiera utilizado la diferencia de probabilidades de tener un hijo de bajo peso entre grupos de riesgo, como medida del efecto del cuidado prenatal en el peso del producto, se hubiera eliminado el efecto del factor de confusión (fumar o no) en el estimador, al igualar la distribución de éste en los grupos de riesgo. Sin embargo, no ocurre tal cosa si se utiliza como medida del efecto, el riesgo relativo RR (cociente de probabilidades) o la razón de momios.

Nótese además, que en el caso de haber utilizado al riesgo relativo como medida, se observa la presencia de interacción.[8]

[8] En la siguiente sección se habla del concepto de interacción.

En general en la literatura (Kleinbaum D. et al. (1982), cap 3; Méndez et al. (1984); Anderson et al.(1980),pag. 7, pags. 23-27; Breslow N. E. y Day N. E.,(1980), pags. 93-104), cuando se habla del concepto de confusión, no se enfatiza el hecho de que la presencia de un efecto confundido en el estimador del efecto que el factor de riesgo tiene en la respuesta, depende en gran parte de la medida que se utilice, ya que como se ha ilustrado en el ejemplo previamente presentado, aunque exista una distribución idéntica del factor de confusión en los grupos de riesgo, puede persistir o no su influencia distorsionante en el estimador dependiendo, de si se utiliza la razón de momios, el cociente de riesgos o la diferencia de riesgos como medida. Por lo tanto aunque un factor de confusión haya sido controlado en la etapa de diseño de la investigación, habrá que controlarlo también en la etapa de análisis.

La determinación de factores de confusión, no es una tarea fácil. La decisión de cuáles factores serán considerados de confusión o tentativamente de confusión, debe tomarse antes de la recolección de los datos, por si se piensa controlarlos en la etapa de diseño de la investigación y/o en la etapa de análisis, puesto que en esta última también se tendrá que captar la información respectiva para cada sujeto en estudio.

En el ejemplo del estudio del efecto del cuidado prenatal en el bajo peso al nacer, se tiene que decidir si se considera al factor fumar o no como un factor de confusión potencial antes de realizar el estudio, para poder recolectar la información respectiva. Esto se requiere porque puede quererse igualar la distribución de fumadoras y no fumadoras, cuando se formen los grupos de riesgo a estudiar o porque se piense realizar un análisis por separado para fumadoras y para no fumadoras, para lo cual, se tendría que obtener para cada mujer en el estudio, información acerca de si fuma o no durante el embarazo.

Cabe aclarar, que si no se planea con anterioridad las necesidades de información para realizar el análisis (en este caso información sobre la condición de fumadora), ésta no se captará y por lo tanto no se podrá controlar por la presencia del factor de confusión.

Es importante notar que la determinación de factores de confusión para una investigación epidemiológica, es primordialmente un problema de la materia de la investigación en sí y no un problema estadístico. No deben hacerse pruebas de significancia para determinar si un factor es de confusión, es decir, no debe probarse si la relación entre el factor antecedente y el factor respuesta es estadísticamente significativa, porque el hacerlo pudiera no tener sentido si no existe conceptualmente un sustento que respalde tal asociación. En el ejemplo del efecto del cuidado prenatal en el bajo peso de los hijos, aunque se demostrara que existe una diferencia significativa en la proporción de mujeres que usan broches de plástico en el cabello durante el embarazo, entre los grupos de riesgo, no podría considerarse a este factor como de confusión, ya que no tiene ningún sentido desde el punto de vista de la materia, es decir, no hay sustento teórico de que ese factor afecte al peso del producto.

Sin embargo, existen algunas técnicas estadísticas que pueden auxiliar en la elección de factores de confusión. Por ejemplo Anderson et al. (1980) proponen utilizar

análisis discriminante como vehículo para examinar la distribución conjunta de factores antecedentes.

Entre las estrategias para controlar los factores de confusión a nivel diseño, se tienen las siguientes:

* Restringir el estudio a sujetos que tengan un mismo valor en los factores de confusión.

* Realizar lo que se conoce como aparejamiento (matching).

El objetivo del aparejamiento es formar los grupos de riesgo o tratamiento de tal forma que la distribución de los factores de confusión potenciales sea la misma en cada uno de ellos, evitando con esto, que se cumpla una de las condiciones para que los factores antecedentes sean de confusión. Existen varios métodos de aparejamiento. Algunos de ellos aparejan sujetos con valores similares en los factores de confusión potenciales, es decir, se realiza un aparejamiento individual mediante la búsqueda de sujetos que tengan valores similares en los factores que se quieren controlar y se asigna uno a cada grupo de riesgo. Ejemplos de métodos que trabajan de esta manera son: aparejamiento por tolerancia y aparejamiento por "el más cercano disponible" (Anderson et al. (1980), cap. 6, pags 78-87).

Existen otros métodos que a diferencia de los anteriores no pretenden igualar la distribución de factores de confusión potenciales entre grupos de riesgo, mediante la búsqueda de sujetos con valores específicos en dichos factores, sino, por ejemplo estratificando por los factores antecedente e igualando el número de sujetos en cada estrato para cada grupo de riesgo o igualando las medias de dichos factores entre grupos de riesgo (Anderson et al. (1980), cap. 6, pags. 88-93).

Entre las estrategias para controlar los factores de confusión a nivel análisis se tienen:

Estandarización y estratificación.- Son procedimientos de ajuste que se utilizan cuando los factores de confusión son categóricos o se categorizan en caso de ser numéricos (escala contínua). El objetivo de ambos procedimientos es corregir por las diferencias en las distribuciones de los factores de confusión, entre los grupos de riesgo o tratamiento.

En ambos procedimientos se divide a los sujetos del estudio en grupos (estratos), con base en los valores que tengan los factores de confusión y posteriormente se combina la información de los grupos para proporcionar un estimador de la medida del efecto. Cuando no es posible combinar la información de todos los grupos, se proporciona la medida del efecto para cada grupo. La diferencia entre estratificación y estandarización radica en el criterio que se utiliza para combinar la información de los estratos.

En estratificación se utiliza algún criterio estadístico para la obtención del estimador global. En estandarización se utiliza una población estándar como referencia para combinar la información de los estratos según esa población. A continuación se ilustran estos procedimientos:

En el ejemplo del estudio del efecto del cuidado prenatal en el peso de los niños al nacer, una forma de controlar en el análisis el efecto del factor de confusión fumar, para poder proporcionar un estimador global del efecto del cuidado prenatal, es suponer que la distribución de fumadoras y no fumadoras para cada grupo de riesgo, es igual a la de cierta población estándar. Entonces al aplicar a esta distribución los estimadores de las probabilidades de tener hijos de bajo peso, obtenidos del estudio en curso, se obtiene un estimador corregido de la probabilidad de un hijo de bajo peso, por grupo de exposición. A este procedimiento se le conoce como estandarización y se ilustra a continuación.

Ejemplo 4.- Supóngase que al realizar el estudio en 100 mujeres con cuidado prenatal y 100 mujeres sin cuidado prenatal, se recolecta además de la información sobre el peso de los hijos, información sobre si las madres fumaron o no durante el embarazo, obteniéndose los siguientes resultados:

Distribución de mujeres según el factor de riesgo
y respuesta para cada nivel del factor de confusión

	No Fumadora			Fumadora		
	Bajo Peso	No Bajo Peso	Total	Bajo Peso	No Bajo Peso	Total
Sin cuidado prenatal	2	8	10	27	63	90
Con cuidado prenatal	16	64	80	6	14	20

$$\hat{\Delta}_{nf} = \frac{2}{10} - \frac{16}{80} = 0.2 - 0.2 = 0 \qquad \hat{\Delta}_{f} = \frac{27}{90} - \frac{6}{20} = 0.3 - 0.3 = 0$$

Resulta entonces, que al controlar por el factor de confusión, no existe efecto del cuidado prenatal en el bajo peso. Sin embargo, si se quiere obtener un estimador global del riesgo de tener un hijo de bajo peso para el grupo de mujeres con cuidado prenatal y otro para el grupo de mujeres que no recibieron cuidado prenatal, el hacerlo sin considerar al factor de confusión, conduce a resultados erróneos como ya se había visto anteriormente, es decir,

Estimador de la probabilidad (riesgo) de tener un hijo de bajo peso

$$\text{Sin cuidado Prenatal } \frac{0.2(10)+0.3(90)}{100} = 0.29$$

$$\text{Con cuidado Prenatal } \frac{0.2(80)+0.3(20)}{100} = 0.22$$

$$\hat{\Delta} \text{ Global} = \quad 0.29 - 0.22 = .07$$

Se obtiene entonces, que el grupo de mujeres que no recibieron cuidado prenatal,

tienen un 7% mas de probabilidad de tener un hijo de bajo peso que las mujeres que si lo recibieron. Esto es un error, ya que al controlar por el factor de confusión, se observó que no existe diferencia alguna.

Supóngase ahora, que se toma la distribución de fumadoras y no fumadoras de cierta población (a la que se le llama estándar) en donde el 70% de las mujeres fuma y el 30% no fuma y que dicha distribución se utiliza conjuntamente con los estimadores de los riesgos específicos para cada nivel del factor de confusión y grupo de riesgo, para obtener un estimador global de la probabilidad de tener un hijo de bajo peso para cada grupo de riesgo.

$$\text{Sin cuidado Prenatal } \frac{0.2(30)+0.3(70)}{100} = 0.27$$

Estimador de la probabilidad (riesgo)

de tener un hijo de bajo peso

$$\text{Con cuidado Prenatal } \frac{0.2(30)+0.3(70)}{100} = 0.27$$

$$\hat{\Delta} \quad \text{Global} \quad = 0$$

Resultando entonces, que al estandarizar se obtiene que no hay diferencia en las probabilidades de tener un hijo de bajo peso entre grupos de riesgo, lo cual es correcto.

En cambio, el procedimiento conocido como estratificación, consiste en obtener información sobre la diferencia en los riesgos de tener hijos de bajo peso entre mujeres con cuidado prenatal y sin cuidado prenatal, para el grupo de fumadoras y de no fumadoras por separado y después mediante algún criterio estadístico, obtener un estimador global de la diferencia. Este estimador global se obtiene combinando las diferencias obtenidas en los grupos definidos por los dos niveles del factor de confusión.

Algunos de los problemas enfrentados al estandarizar y estratificar son los siguientes:

* Es difícil determinar cuales de las variables de confusión se incluyen y como categorizarlas.
* Al categorizar se pierde necesariamente información.
* Mientras mayor sea el número de factores de confusión a controlar, menor será la cantidad de observaciones por estrato y en consecuencia, habrá pérdida de precisión.

Se recomienda utilizar estos procedimientos como una aproximación inicial para el control de factores de confusión. Su uso, puede además, proporcionar pistas preliminares acerca de la naturaleza y extensión de cualquier interacción en los datos.

Utilización de modelos matemáticos.- Se utilizan modelos matemáticos para explicar la relación entre el factor respuesta y los factores de riesgo y de confusión, permitiendo al mismo tiempo obtener estimadores de la medida del efecto del factor de riesgo (o tratamiento), ajustados por la presencia de los factores de confusión.

Entre los modelos más utilizados en epidemiología se tienen el modelo logístico lineal y el análisis de covarianza. La utilización de modelos presenta también sus dificultades:

* La elección del modelo.
* Se deben cumplir las suposiciones del mismo, al menos en forma aproximada, y
* La elección de las variables que deban incluirse.

El uso de modelos sin embargo, permite el manejo simultáneo de muchas variables con relativa facilidad, por lo que se recomiendan.

I.2.3 Interacción

Interacción significa que el efecto del factor de riesgo o tratamiento depende de un factor antecedente, es decir, que el factor de riesgo afecta al factor respuesta de una manera distinta en los diferentes niveles del factor antecedente.

Es importante señalar que si existe interacción, no se puede obtener una medida global del efecto de tratamiento y esto porque el efecto de tratamiento no es único, sino que depende del factor antecedente.

Por lo anterior, se recomienda como una estrategia analítica, el que se evalúe la presencia de la interacción antes de cualquier intento para controlar la posible confusión con el factor antecedente. De existir interacción, no tendría sentido, como ya se dijo, obtener un estimador global ajustado por la presencia del factor de confusión, ya que en este caso el efecto del factor de riesgo varía para cada nivel del factor antecedente.

Al igual que la noción de "Factor de Confusión", la noción de "Interacción" puede depender de la medida que haya sido elegida para evaluar el efecto del factor de riesgo en la respuesta. En el ejemplo 3 se observó que el hecho de utilizar el riesgo relativo o la diferencia en riesgos como medida del efecto, hacía que el factor de confusión: fumar o no fumar, interactuase en un caso (cuando se usa riesgo relativo) con el factor de riesgo y no interactuase cuando se escoge como medida la diferencia.

I.3 MEDIDAS USUALES EN EPIDEMIOLOGIA

Entre las medidas que se utilizan en epidemiología se encuentran las medidas de frecuencia y distintas medidas del efecto que el factor tratamiento (o riesgo) tiene en el factor respuesta. En los ejemplos vistos ya se utilizaron tres de estas medidas.

I.3.1 Medidas de frecuencia

Las medidas de frecuencia se utilizan para caracterizar la ocurrencia de una enfermedad o muerte en poblaciones humanas, o más generalmente se podría decir que se utilizan para caracterizar la ocurrencia de un factor respuesta dicotómico.

Estas medidas son importantes tanto en investigaciones descriptivas como etiológicas. Pueden expresar el grado relativo de morbilidad y mortalidad.

Medidas de Morbilidad

Existen dos tipos de medidas de morbilidad:

* Las medidas de incidencia que son medidas de frecuencia basadas en casos nuevos que ocurren en un período de tiempo determinado en una población.

* Las medidas de prevalencia que son medidas de frecuencia basadas en los casos existentes en una población en un período de tiempo determinado.

Las medidas de incidencia reflejan un cambio en el estado de enfermedad y por lo mismo son adecuadas para identificar factores de riesgo. Se prefieren para hacer inferencias acerca de la etiología de una enfermedad (estas medidas se estiman generalmente a partir de estudios de cohortes prospectivos).

Hay dos medidas de incidencia: el riesgo y la tasa.

El riesgo es la probabilidad de que un individuo no enfermo desarrolle una enfermedad específica, en un período de tiempo determinado, pero condicionada a que ese individuo no se esté muriendo por otra enfermedad durante el período. Entre sus características se encuentran:

* Varía entre 0 y 1.

* No tiene dimensión.

* Usualmente se refiere a la primera ocurrencia de la enfermedad para cada persona no enferma, aunque se puede considerar el riesgo de desarrollar la enfermedad dos o más veces, en un período dado. En enfermedades crónicas se trabaja con la primera ocurrencia para cada sujeto en riesgo; en cambio en enfermedades agudas se puede estar interesado en múltiples ocurrencias de la enfermedad.

* Requiere un período de referencia que describa la extensión de tiempo en el cual se detectan los casos nuevos. Este período se puede fijar arbitrariamente o puede variar entre individuos.

Con base en el riesgo, se puede obtener lo que se conoce como momio de riesgos, que es la probabilidad de desarrollar la enfermedad dividida por la probabilidad de no desarrollarla.

Existen varias formas de estimar el riesgo:

i) Método de acumulación simple.- Si el número de sujetos a estudiar está fijo y hay pocas pérdidas de seguimiento durante el período de estudio, entonces el riesgo durante el período de estudio (t_o, t), se puede estimar como:

$$\hat{R}_{(to,t)} = \frac{I}{N_o}$$

donde:

I: número de casos nuevos diagnosticados en el período.

N_o: número de sujetos no enfermos al principio del estudio, es decir, al tiempo t_o.

Al cociente $\frac{I}{N_o}$ se le suele llamar incidencia acumulativa (IA).

ii) Metodo actuarial.- Cuando los períodos de seguimiento son diferentes entre los sujetos que no presentan la enfermedad en el período (t_o, t), entonces se puede estimar el riesgo de la siguiente manera:

$$\hat{R}_{(to,t)} = \frac{I}{N_o - (p/2)}$$

donde:

p: Número de sujetos que se perdieron durante el seguimiento en el período (t_o, t).

$N_o - (p/2)$: Puede interpretarse como el número real de personas en riesgo de desarrollar la enfermedad, suponiendo que en promedio los sujetos que se pierden salen en el punto medio del período de estudio.

iii) Método de densidad.- Este método de estimación se basa en estimadores de tasas promedio de incidencia [9].

Si se tiene una tasa de incidencia constante IP, el número de sujetos sanos al tiempo t, N_t, es una función exponencial con pendiente negativa, es decir,

$$N_t = N_o \quad exp\{-IP(\triangle)\}$$

donde:

$\triangle = t - t_o$: tiempo entre el inicio del estudio, t_o y el tiempo t

$IP(\triangle)$: tasa de incidencia promedio en $\triangle$

Entonces

$$\hat{R}_{(t_o,t)} = 1 - \frac{N_t}{N_o} = 1 - exp\{-\widehat{IP}(\triangle)\}$$

La otra medida para la incidencia es la tasa. La tasa de incidencia es el potencial instantáneo para el cambio en el estado de enfermedad por unidad de tiempo, al tiempo t, relativo al tamaño de la población no enferma al tiempo t. Entre sus características se encuentran:

* Se refiere a una población y por lo tanto no tiene interpretación directa a nivel individual como la tiene el riesgo, ya que el riesgo tiene interpretación probabilística.
* No tiene período de referencia ya que se refiere a un punto en el tiempo y nó a un período de duración del estudio como en el caso del riesgo.
* Sí tiene dimensión ya que se define por unidad de tiempo.

---

[9] Más adelante se describen los estimadores de tasas de incidencia.

* Es mayor que cero y puede ser mayor que la unidad, a diferencia del riesgo que está entre cero y uno.

Es difícil obtener estimadores de las tasas de incidencia instantáneas porque no se puede expresar generalmente el tamaño de la población en estudio como función del tiempo. Lo que se hace, es estimar una tasa promedio para un período dado:

$$\widehat{IP}_{(t_o,t)} = \frac{I}{TP}$$

donde:

I: es el número de casos nuevos en el período (t_o, t)

TP: es la cantidad de tiempo-población libre de enfermedad, acumulado por la población en estudio durante el período (t_o, t).

TP puede calcularse de dos maneras dependiendo de la disponibilidad de información:

* Si se conocen los períodos de seguimiento individuales para toda la población sana en estudio N, entonces

$$TP = \sum_{i=1}^{N} \Delta_{t_i}$$

donde:

Δ_{t_i}: es la duración del período de seguimiento observado para el i-ésimo individuo a partir de su entrada al estudio, hasta la detección de la enfermedad o salida del estudio.

* Si no se conocen los períodos de seguimiento individuales entonces, suponiendo que la duración de los períodos de seguimiento es la misma:

$$TP = N(\Delta_t)$$

donde:

Δ_t: es la duración común de seguimiento

N: tamaño de la población sana y estable[10]

Para determinar cómo medir la incidencia, hay que tomar dos decisiones: la primera de ellas es qué medida teórica se va a usar: riesgo o tasa. La segunda decisión es el cómo se va a estimar esa medida.

---

[10] Estable se refiere a una población que aunque dinámica porque permite la entrada y salida de sujetos, su tamaño y distribución por edad permanece constante durante el período de seguimiento.

El principal criterio para decidir qué medida teórica utilizar, es el objetivo del estudio. Si el objetivo es predecir un cambio individual en el estado de salud con base en ciertas características, se deberá conocer el riesgo de desarrollar la enfermedad.

Si el objetivo, por otro lado, es investigar la etiología de las enfermedades, teniendo en cuenta los posibles efectos de uno o más factores, la elección de la medida de incidencia depende del tipo de enfermedad y del tiempo durante el cual se observa la ocurrencia de casos nuevos.

En enfermedades con largos períodos de riesgo, (en general son enfermedades crónicas con largos períodos de latencia como por ejemplo cáncer pulmonar), deben utilizarse tasas, ya que, el verdadero período de seguimiento para cada individuo es tan sólo una parte del tiempo que la persona esta sujeta al riesgo de desarrollar la enfermedad.

Para enfermedades con períodos de latencia cortos y cuando el factor en estudio es, una exposición de corta duración o un atributo fijo (en general se trata de enfermedades agudas con períodos de latencia cortos como la rabia) entonces se deberá utilizar el riesgo como medida de incidencia, ya que por un lado, el período de riesgo completo para cada sujeto está incluido dentro del verdadero período de seguimiento y por otro lado, las tasas en este tipo de investigaciones cambian mucho durante el curso de la epidemia.

Respecto a la elección de un estimador, debe observarse que para estimar tasas y riesgos puede utilizarse la incidencia promedio, mientras que la incidencia acumulativa sólo puede utilizarse para estimar riesgos. Es importante aclarar que la elección de una medida, depende en gran parte de la posibilidad de estimarla y ésto está determinado por el diseño que se utilice para obtener la información.

Las otras medidas de morbilidad, además de las de incidencia que ya se mencionaron, son las medidas de prevalencia. Como se dijo anteriormente, las medidas de prevalencia son medidas de frecuencia basadas en los casos de enfermedad existentes en una población en un tiempo determinado y nó en los nuevos casos en un período dado, lo que llevaría a medir la incidencia. Las medidas de prevalencia se suelen utilizar en estudios transversales [11] principalmente.

Como los casos de prevalencia representan sobrevivientes de una enfermedad, no son tan adecuados para identificar factores de riesgo como serían los casos de incidencia.

Las medidas de prevalencia son útiles para planear y administrar servicios médicos.

Se utilizan básicamente dos medidas para cuantificar la prevalencia de una enfermedad en una población: la prevalencia puntual y la prevalencia en un período.

La prevalencia puntual es la probabilidad, al tiempo t, de que un individuo en una población, padezca la enfermedad bajo estudio.

---

[11] En el capítulo siguiente se puede consultar lo que se entiende por un estudio transversal.

Como estimador puede utilizarse:

$$\hat{P}_t = \frac{C_t}{N_t'}$$

donde:

N_t': es el tamaño de la población en estudio al tiempo t (incluye a personas que ya estaban enfermas antes del tiempo t)

$C_t = N_t' - N_t$: es el número de casos prevalentes al tiempo t

N_t: es el número de individuos sanos al tiempo t

Con la prevalencia puntual se puede calcular el cociente entre ésta y su complemento a uno; esto es, el cociente entre la probabilidad de estar enfermo al tiempo t y la probabilidad de no estarlo. A esta medida se le conoce como momio de prevalencia.

La prevalencia en un período es la probabilidad de que un individuo en una población esté enfermo, en cualquier tiempo dentro de un período dado (t_o, t) de duración Δ_t, $(\Delta_t = t - t_o)$.

La prevalencia por período, muy frecuentemente se utiliza como aproximación del riesgo de incidencia cuando se desconoce el tiempo exacto de inicio de la enfermedad para cada individuo. La aproximación será muy burda si al principio del período un porcentaje considerable de individuos padecían la enfermedad, y por el contrario la aproximación será excelente si al principio del período sólo un porcentaje muy bajo padecían la enfermedad.

Esta prevalencia se puede estimar de la siguiente manera:

$$\widehat{PP}_{(t_o,t)} = \frac{C_{(t_o,t)}}{N} = \frac{C_o + I}{N}$$

donde:

$C_{(t_o,t)}$: Número de personas que se observó que tenían la enfermedad en cualquier tiempo durante el período de seguimiento (t_o, t). Incluye a los enfermos en t_o y a los casos incidentes (I) detectados durante el período.

Medidas de Mortalidad

La medidas de mortalidad son medidas de frecuencia análogas a las de incidencia, sólo que en este caso el resultado de interés es muerte. Para mortalidad, es evidente que no existe un análogo a las medidas de prevalencia como en el caso de enfermedad.

Como los datos de mortalidad son más fáciles de coleccionar y son generalmente más confiables que los datos de incidencia de enfermedades, muchas veces se utilizan en lugar de ellos para generar y probar hipótesis etiológicas, especialmente con enfermedades altamente fatales.

También se utiliza la información sobre mortalidad, para evaluar programas preventivos y terapéuticos.

Existen tres clases de eventos de mortalidad mutuamente exclusivos en relación al padecimiento de la enfermedad en estudio: la muerte debida a la enfermedad de interés; la muerte debida a otras causas entre los que padecían la enfermedad de interés y la muerte debida a otras causas entre las personas que no tenían la enfermedad de interés.

Se utilizan dos medidas para expresar la frecuencia de muerte en una población: la tasa o fuerza de mortalidad y el riesgo. Estas conjuntamente con los eventos arriba mencionados definen una serie de medidas de mortalidad que son muy similares a las de incidencia de enfermedades. Así mismo, los estimadores de tales medidas son muy similares a los de incidencia de enfermedades, por lo tanto ya no se describirán. (consultar Kleinbaum D. et al, (1982), cap 7).

I.3.2 Medidas del efecto del factor de riesgo

Al hablar de medidas del efecto del factor de riesgo, se está haciendo referencia a todas aquellas medidas que sirvan para evaluar el efecto que el factor de riesgo (o tratamiento) tiene en el factor respuesta, con la finalidad de investigar la etiología, tratamiento y prevención de enfermedades. Son útiles para hacer inferencias causales.

La elección de una medida depende de varios aspectos como son:

* La escala de medición de los factores de riesgo, respuesta y confusión.
* El modelo que se piense utilizar para expresar la forma en que el factor de riesgo (o tratamiento) afecta al factor respuesta.
* Los objetivos del estudio, es decir, qué se va a estudiar, para qué se van a utilizar los resultados, etc.
* La forma en que se diseñe el estudio.

En adición, existen otros criterios que en algunas ocasiones deben considerarse para la elección de una medida adecuada, por ejemplo Cox (1977), pág. 20, menciona los siguientes:

* Que cuando se trabaja con un factor respuesta dicotómico, se requiere que la medida del efecto no se altere o a lo más cambie sólamente de signo si se intercambian la presencia y la ausencia de la condición que se estudia (el éxito o el fracaso).
* Que la medida tenga un significado práctico directo.
* Que sea una medida para la cual la teoría estadística inferencial sea simple.

En función de las escalas con las que se miden los factores de riesgo y respuesta, se analizan a continuación algunos casos.

El primero de ellos es cuando la escala para el factor de riesgo es categórica y la del factor respuesta es dicotómica. Este es un caso que se presenta muy frecuentemente en la práctica y es por ello que aquí se menciona con más detalle.

El segundo caso que se analiza es aquel en el que el riesgo es dicotómico y la respuesta es numérica (contínua).

El tercer caso que se presenta es cuando el riesgo es numérico (contínuo) y la respuesta dicotómica.

Finalmente, se menciona el caso en el que tanto el riesgo como la respuesta son numéricos.

Por simplicidad se excluyó en este breve análisis de casos posibles, la consideración del factor de confusión.

Riesgo Categórico y Respuesta dicotómica

Supóngase que la escala de medición de la variable riesgo es categórica y de la variable respuesta es también categórica pero con dos categorías únicamente.

Las medidas del efecto en este caso, involucran una comparación directa de medidas de frecuencia para diferentes categorías del factor de riesgo. Se compara la medida de frecuencia de cada categoría con la medida de frecuencia de una categoría específica, designada como grupo de referencia.

La elección del grupo de referencia es, hasta cierto punto arbitraria, pero muy importante para la interpretación de los resultados. El grupo de referencia debe ser suficientemente grande como para proporcionar estimadores de frecuencia precisos y además debe ser homogéneo como para que las comparaciones tengan sentido. De ser posible este grupo debe corresponder al grupo de riesgo más bajo.

A las medidas del efecto para este caso, se les puede dividir en medidas de razón y medidas de diferencia.

Una medida de razón es, la medida de frecuencia de un grupo expuesto (E_i) dividida por la misma medida de frecuencia, pero en el grupo de referencia ($i = 0$). Hay tantas medidas de razón como medidas de frecuencia.

A continuación se mencionan dos estimadores de medidas de razón para comparar incidencia: el estimador de la razón de tasas de incidencia y el estimador de la razón de riesgos.

La razón de tasas de incidencia (RTI), se estima con:

$$\widehat{RTI} = \frac{\widehat{IP}_i}{\widehat{IP}_o}$$

donde:

$\widehat{IP}_i$: tasa promedio de incidencia estimada para el grupo i.

$\widehat{IP}_o$: tasa promedio de incidencia estimada para el grupo de referencia.

Nótese que $0 \leq RTI < \infty$, RTI toma el valor de 0 cuando hay una asociación negativa fuerte, toma el valor uno cuando no hay asociación, y toma un valor grande cuando existe una asociación positiva fuerte.

La razón de los riesgos de incidencia o riesgo relativo (RR) se estima con

$$\widehat{RR}_i = \frac{\hat{R}_i}{\hat{R}_o}$$

donde:

$\hat{R}_i$: es el riesgo de incidencia estimado en el grupo i

$\hat{R}_o$: es el riesgo de incidencia estimado en el grupo de referencia.

Es importante mencionar bajo qué condiciones es posible obtener estimadores de estas medidas de razón, en estudios de casos y controles. [12]

En un estudio de casos y controles que utilice casos de incidencia, no es posible (salvo en algunas circunstancias muy especiales) estimar las tasas o riesgos de incidencia específicos para cada grupo de exposición. Esto se debe a que en 'casos y controles' se elige un grupo de enfermos y otro de no enfermos, y en cada uno de estos grupos se determina su exposición al factor de riesgo. Esto en cierto modo es lo contrario de lo clásico en donde a partir de grupos con diversos niveles de exposición al riesgo, se determinan cuantos de cada nivel, resultaron enfermos. Sin embargo, bajo ciertas condiciones que se mencionan a continuación, sí es posible estimar la razón de esos riesgos o tasas.

Si la enfermedad en estudio involucra un período muy largo de riesgo (período entre el inicio de la exposición al factor de riesgo y la manifestación clínica de la enfermedad), comparado con el período de observación, entonces el parámetro que interesa es la razón de tasas de incidencia. Si se supone que la población es estable y que sólo hay dos grupos de exposición, entonces:

$$\widehat{RTI} = \frac{\widehat{IP}_1}{\widehat{IP}_o} = \frac{I_1/N_1(\Delta_t)}{I_o/N_o(\Delta_t)} = \frac{I_1/I_o}{N_1/N_o}$$

donde:

I_i: número de casos nuevos detectados en el i-ésimo grupo de exposición durante el período Δ_t de seguimiento, $i = 0, 1$.

N_i: tamaño de la población sana para el i-ésimo grupo de exposición.

I_1/I_o: es entonces el estimador del momio de estar expuesto entre los casos nuevos y puede obtenerse sin problemas a partir de un estudio de casos y controles.

N_1/N_o: es el estimador del momio de estar expuesto entre los no enfermos y también puede obtenerse de un estudio de casos y controles.

---

[12] En el siguiente capítulo puede consultarse que se entiende por un estudio de casos y controles.

Por lo tanto,[13] si en un estudio de casos y controles los individuos no enfermos son representativos de la población estable de la cual se observaron los enfermos, entonces se puede estimar la razón de tasas de incidencia, mediante el estimador de la razón de momios de exposición ($\widehat{RME}$), es decir, $\widehat{RTI} = \widehat{RME}$.

Si la enfermedad en estudio involucra un período corto de riesgo, comparado con el período de observación del estudio, entonces el parámetro que interesa es la razón de riesgos.

La razón de riesgos se puede expresar en términos de probabilidades condicionales, de la siguiente manera:

$$RR = \frac{P(c|e)}{P(c|\bar{e})} = \frac{P(e|c)}{P(\bar{e}|c)}\left[\frac{P(c)P(\bar{e}|c) + P(\bar{c})P(\bar{e}|\bar{c})}{P(c)P(e|c) + P(\bar{c})P(e|\bar{c})}\right]$$

donde:

$P(c|e)$: probabilidad de que dado que se está expuesto (e) se desarrolle la enfermedad (c = caso).

$P(c|\bar{e})$: probabilidad de que dado que no se está expuesto se desarrolle la enfermedad.

Si el riesgo global de enfermedad es muy bajo, es decir, $P(c)$ es aproximadamente cero, entonces

$$RR \doteq \frac{P(e|c)P(\bar{e}|\bar{c})}{P(\bar{e}|c)P(e|\bar{c})} = \frac{P(e|c)/P(\bar{e}|c)}{P(e|\bar{c})/P(\bar{e}|\bar{c})}$$

Resultando entonces que el riesgo relativo, es aproximadamente igual a la razón de momios de exposición.

En conclusión, si el riesgo de enfermedad es muy bajo y los "no casos" son representativos de la población de la cual se observaron los "casos", entonces se puede estimar el riesgo relativo estimando la razón de momios de exposición, es decir, $\widehat{RR} = \widehat{RME}$.

Existen también medidas de razón para comparar prevalencia y mortalidad, pero ya no se presentan porque se extendería mucho este trabajo. Si se quiere obtener información sobre ellas consultar Kleinbaum D. et al., (1982), cap. 8.

El otro tipo de medidas del efecto para el caso de riesgo categórico y respuesta dicotómica aparte de las medidas de razón, es el de medidas de diferencia.

Una medida de diferencia se calcula restando la medida de frecuencia del grupo de referencia, de la misma medida de frecuencia pero para un grupo expuesto.

Es importante hacer notar que tampoco se pueden obtener estimadores de medidas de diferencia a partir de estudios de casos y controles.

---

[13] Se profundiza más sobre este punto en el capítulo III

Al igual que con las medidas de razón, se describirán únicamente las medidas de diferencia para comparar incidencia de enfermedades. Las medidas de diferencia para comparar prevalencia y mortalidad pueden consultarse en Kleinbaum D. et al. (1982), cap. 8.

Se describen a continuación dos estimadores de medidas de diferencia para comparar incidencia: el estimador de la diferencia de tasas de incidencia y el estimador de la diferencia de riesgos de incidencia.

La diferencia de tasas de incidencia (DTI) se estima como:

$$\widehat{DTI}_i = \widehat{IP}_i - \widehat{IP}_o$$

Obsérvese que $-\infty < DTI < \infty$. DTI es cero cuando no hay asociación; DTI es positiva cuando existe una asociación positiva y negativa cuando la asociación es negativa.

La diferencia de riesgos de incidencia (DR), conocida también como riesgo atribuible, se estima de la siguiente manera:

$$\widehat{DR}_i = \hat{R}_i - \hat{R}_o$$

donde:

$\hat{R}_i$: riesgo de incidencia estimado en el grupo i.

Hasta aquí, sólo se han descrito medidas del efecto cuando la variable riesgo y respuesta son categóricas. En cualquier otro caso las medidas del efecto del tratamiento, son parámetros de los modelos que se elijan para explicar la relación entre el factor de riesgo y el factor respuesta. Los estimadores de las medidas del efecto se obtendrán entonces, obteniéndo los estimadores de los paramétros del modelo elegido.

Se mencionan de manera breve algunos modelos simples con el objeto de ilustrar la utilización del concepto de estimación de un efecto, estimando los parámetros asociados en cada modelo.

Riesgo dicotómico y respuesta continua

Supóngase que la variable riesgo es categórica con dos categorías y que la variable respuesta es numérica. En este caso como medida del efecto puede emplearse la diferencia de medias de la variable respuesta entre los grupos de exposición.

Suponga que para expresar la relación entre la variable riesgo y respuesta, se escoge un modelo de la siguiente forma:

Sea Y : variable respuesta.

$$Y_i = Y_o + \Delta$$

donde:

Y_t: valor de la variable respuesta para un sujeto en el grupo expuesto al riesgo.

Y_o: valor de la variable respuesta para un sujeto en el grupo no expuesto al riesgo.

Bajo este modelo Δ es la medida del efecto, la cual puede estimarse mediante la diferencia de medias.

Si el modelo fuera: $Y_t = \pi Y_o$, es decir, $lnY_t = lnY_o + ln\pi$ entonces $ln\pi$ es la medida del efecto de tratamiento (en la escala logarítmica), la cual puede estimarse mediante la diferencia de las medias de los logaritmos de la variable respuesta entre los dos grupos.

Riesgo numérico y respuesta dicotómica

Supóngase que la variable riesgo es numérica y la variable respuesta es categórica con dos categorías (éxito o fracaso).

En este caso se suele utilizar un modelo conocido como logístico lineal, para explicar la relación entre el factor de riesgo y respuesta, es decir:

$$ln\frac{P(Y_i = 1|X_i)}{1 - P(Y_i = 1|X_i)} = ln\frac{P(X_i)}{1 - P(X_i)} = \alpha + \gamma Xi$$

donde:

$P(X_i)$: Probabilidad de obtener un éxito con el sujeto i que tiene un valor en la variable riesgo igual a X_i.

En este caso el efecto de tratamiento se mide mediante γ que representa el cambio (que se supone constante) que produce un incremento de una unidad en la variable riesgo sobre el logaritmo del momio de éxito. El estimador del efecto, se obtendrá entonces estimando γ.

Riesgo y respuesta numéricos

Cuando las variables riesgo y respuesta son continuas, se utilizan modelos de regresión para explicar la relación entre ellas. Cierto parámetro del modelo representará el efecto de tratamiento, el cual, podrá estimarse mediante el estimador del parámetro correspondiente.

Es importante hacer notar, que en la discusión de medidas del efecto que el factor de riesgo (o tratamiento) tiene en el factor respuesta, se supuso implícitamente que no existían factores de confusión y por supuesto tampoco interacción con algún factor antecedente. Si existieran factores de confusión, habría que controlar por su presencia (antes de obtener estimadores de la medida del efecto) ya sea mediante estratificación, estandarización o incluyéndolos en el modelo que se esté utilizando. Esto con la finalidad de eliminar su influencia en la medida del efecto de tratamiento.

Si existiera interacción, no podría obtenerse un estimador general del efecto de tratamiento, ya que éste variaría en los diferentes niveles de un factor antecedente.

CAPITULO II

TIPOS DE INVESTIGACIONES EPIDEMIOLOGICAS

En este capítulo se pretende presentar un panorama general acerca de las diferentes opciones que existen para diseñar un estudio epidemiológico. Estas opciones resultan al combinar diferentes alternativas para el estudio de las características de interés en los sujetos bajo investigación. La elección de dichas alternativas depende de los objetivos de la investigación, de la disponibilidad de información, de la disponibilidad de recursos en tiempo, dinero y fuerza de trabajo, así como de restricciones de tipo político ó ético.

II.1 CARACTERISTICAS QUE TIPIFICAN A LOS DISEÑOS

A continuación se definen algunos criterios que ayudan a caracterizar los diferentes tipos de diseños que se utilizan en la investigación epidemiológica.

Dependiendo de si se estudia a una sola población o se pretenden comparar varias, un estudio se puede clasificar en descriptivo o comparativo.

En un estudio descriptivo se tiene una sola población, la cual se pretende caracterizar con base en una serie de variables. No existen hipótesis etiológicas centrales. Sin embargo, algunas veces, con la finalidad de explorar relaciones entre variables dentro de la misma población, se plantean un grupo de hipótesis que se refieren a asociaciones entre dichas variables. Las asociaciones encontradas sirven de base para la generación de hipótesis etiológicas específicas para otros estudios posteriores.

Este tipo de estudios se llevan a cabo generalmente cuando se conoce poco acerca de la ocurrencia, la historia natural o los determinantes de una enfermedad. Sus objetivos son entonces, estimar la frecuencia de enfermedades o su tendencia en el tiempo en una población particular y dar origen a hipótesis etiológicas específicas. Se utilizan también cuando lo que se pretende es planificar o llevar a cabo programas de salud.

Los estudios descriptivos son útiles para explorar asociaciones entre factores de riesgo potenciales y una enfermedad, cuando se tienen conocimientos limitados. Se pueden considerar como un paso previo a la realización de estudios donde se tienen muy bien definidas las hipótesis de investigación. En cierto modo los estudios descriptivos tienen una función exploratoria.

En un estudio comparativo se tienen dos o más poblaciones que se pretenden comparar respecto a ciertas características, con la finalidad de contrastar una o varias hipótesis centrales. Se llevan a cabo cuando se tiene un conocimiento relativamente extenso acerca de la enfermedad, antes de la realización de la investigación, de tal modo que se pueden especificar apriori las hipótesis que se desean probar.

Sus objetivos son por lo general la identificación de factores de riesgo para las

enfermedades, la estimación de sus efectos en la misma y el planteamiento de posibles estrategias de intervención.

De acuerdo a la capacidad para manejar el factor de estudio (factor de riesgo o tratamiento), así como por la forma en que se maneje, un estudio puede clasificarse en aleatorizado (o experimental), en quasi-experimental o en observacional.

Un estudio aleatorizado es aquél en el cual las categorías del factor bajo estudio (tratamiento) se asignan aleatoriamente a los sujetos bajo investigación.

En experimentos bien diseñados y con tamaños de muestra grandes es muy probable que queden controlados los efectos de distorsión originados por factores de confusión potenciales, sean estos conocidos por el investigador o no. Este control se logra al igualar la distribución de los factores de confusión potenciales entre grupos de tratamiento mediante la aleatorización.

Al aleatorizar se pueden hacer inferencias estadísticas formales además de que se controlan los sesgos de selección. Si los tamaños de muestra son pequeños, a pesar de la aleatorización, puede ocurrir que la distribución de los factores de confusión potenciales resulte diferente en los diferentes grupos de riesgo. En cualquier caso, si se encuentra que existe alguna distorsión en el estimador del efecto, deberá de corregirse en la etapa de análisis.

Un defecto de este tipo de estudios es que se realizan generalmente en un medio artificial y con una muestra de sujetos muy bien seleccionada, de tal modo que las condiciones en que se realizan pudieran diferir en varias características de las condiciones en las que se realizaría con la población objetivo. Si el verdadero efecto de tratamiento depende de estas características entonces el efecto observado en el estudio aleatorizado no es extrapolable a la población objetivo.

En el trabajo epidemiológico es difícil llevar a cabo un estudio experimental, ya que puede no ser ético o factible el realizarlo.

Un estudio Quasi-Experimental[1] es aquel en el que el investigador asigna las categorías del factor tratamiento a los sujetos bajo investigacción pero no de una manera aleatoria, sino por conveniencia o según la voluntad de los propios sujetos. Al no aleatorizar la asignación de niveles del factor bajo estudio a los sujetos, se tiene poco control sobre la influencia de factores de confusión potenciales y sesgos de selección.

Este tipo de estudios tienen menos obstáculos prácticos que los experimentos y son más baratos.

Un estudio observacional es aquel en el que el investigador no asigna las categorías del factor de estudio (factor de riesgo) a los sujetos bajo investigación sino que simplemente observa la categoría que posee cada sujeto.

En este tipo de estudios es en donde se tiene aún menos control en el diseño

---

[1] Estrictamente, este tipo de estudios no deberían de realizarse, porque si el investigador tiene la capacidad de manejar el factor tratamiento, debe de utilizar algún mecanismo de aleatorización para asignar los niveles de dicho factor a los sujetos bajo estudio.

por la presencia de factores de confusión potenciales y sesgos de selección, por lo tanto, el investigador deberá eliminar estas influencias distorsionantes en la etapa de análisis.

A pesar de que son los que menos control tienen sobre fuentes de sesgo, en muchas ocasiones son la única manera ética o factible de llevar a cabo la investigación.

Como se llevan a cabo en ambientes más naturales sus resultados son extrapolables más fácilmente a la población objetivo que los de un estudio experimental.

En general no puede decirse que sean menos caros o que se llevan menos tiempo que otro tipo de estudios.

Dependiendo de si la información que se requiere se obtiene en un sólo punto en el tiempo o en varios puntos o si se hace o no un seguimiento a través del tiempo de los sujetos, un estudio puede clasificarse en: transversal o longitudinal.

En los estudios transversales la información que se requiere se obtiene en un sólo punto hipotético en el tiempo y por lo mismo no puede haber un seguimiento de los sujetos bajo investigación.

En los estudios longitudinales se obtiene la información que se requiere, sobre los mismos aspectos, en dos o más puntos hipotéticos en el tiempo, o bien, se hace un seguimiento de los sujetos bajo investigación.

Cuando se realiza un seguimiento, se puede trabajar con lo que se conoce como una cohorte fija o con una población dinámica.

Por cohorte fija se entiende un grupo de sujetos identificado en un punto hipotético en el tiempo y seguido por un período para detección de nuevos casos de enfermedad. No se permite la entrada de sujetos al estudio después de iniciado el seguimiento pero sí la salida ya que pueden ocurrir pérdidas por no participación, migración, muerte, etc.

Por población dinámica se entiende un grupo de sujetos identificado en un punto hipotético en el tiempo y seguido por un período para detección de nuevos casos de enfermedad, pero en este caso sí pueden entrar sujetos al estudio durante el curso del seguimiento. Si el tamaño y la distribución por edad del grupo, permanece constante durante el seguimiento, recibe el nombre de Población Estable.

Dependiendo de si la medición del factor de estudio y del factor respuesta se refieren o no a dos puntos diferentes en el tiempo (es decir que hay una relación temporal entre ellos) y dependiendo también de cual se mide primero, un estudio puede clasificarse en estudio "hacia adelante", "hacia atrás" y "no direccionado".

En un estudio "hacia adelante"[2] se parte de la medición del factor de riesgo o tratamiento en la población (libre de enfermedad) y pasado cierto tiempo se mide el valor del factor respuesta, es decir, se procede de la causa potencial al efecto, respetando

---

[2] En general, en los textos tradicionales de epidemiología a los estudios "hacia adelante" y "hacia atrás" se les llama longitudinales y a los "no direccionados" se les llama transversales, y no se toma en consideración si la información se obtiene en un sólo punto en el tiempo (transversal según la difinición de este trabajo) o en varios puntos en el tiempo (longitudinal según la definición de este trabajo).

la temporalidad de la ocurrencia de los eventos. Nótese que la medición del factor de estudio y respuesta se refieren a dos puntos diferentes en el tiempo.

Este tipo de estudios son los que mejor permiten al investigador determinar si la ocurrencia del factor de estudio verdaderamente antecede a la ocurrencia del factor respuesta. Además usualmente involucran casos de incidencia, lo que conjuntamente con lo anterior da más apoyo a una inferencia causal.

En un estudio "hacia atrás" se parte de la medición del factor respuesta y posteriormente se obtiene la medición sobre la exposición previa al factor de estudio, es decir, se procede, del efecto a la causa potencial pero, respetando la relación temporal entre ellos. Nótese que en este caso también la medición del factor de estudio y respuesta se refiere a dos puntos diferentes en el tiempo.

En un estudio "hacia atrás" que involucre casos de incidencia e información confiable es posible determinar si el factor de riesgo o tratamiento se manifestó antes que el factor respuesta.

En un estudio "no direccionado" [3], la medición sobre el factor de riesgo o tratamiento y sobre el factor respuesta se refieren al mismo punto en el tiempo.

No se puede establecer a partir de un estudio como éste si la causa hipotética se manifestó antes que la ocurrencia de la enfermedad, por lo tanto no son útiles para apoyar inferencias causales.

Dependiendo del tiempo de ocurrencia del factor de riesgo o tratamiento y del factor respuesta relativo al tiempo en el que se realiza la investigación, un estudio puede clasificarse en prospectivo, retrospectivo o "ambipectivo" (ambispective).

Un estudio es prospectivo si al iniciarse éste, todavía no ocurren ni el factor de riesgo o tratamiento ni el factor respuesta.

Este tipo de estudios tienen la ventaja de que permiten que las características de interés se registren de acuerdo a un formato idóneo que responda a los objetivos del estudio. Se pueden evitar muchos errores de medición.

En un estudio retrospectivo el factor de riesgo o tratamiento y el factor respuesta ocurren antes del inicio de la investigación. Presenta la desventaja de que la información se tiene que obtener de registros o a partir de la memoria de las personas, lo que puede generar muchos errores de medición.

En un estudio "ambipectivo" el factor de riesgo o tratamiento ocurre antes del inicio de la investigación y el factor respuesta después.

El hecho de que los eventos de interés hayan ocurrido antes o después del inicio de la investigación, no influencía directamente la habilidad para distinguir antecedente de consecuente si se dispone de información de calidad.

Dependiendo de si se mide un estado de salud o un cambio en el estado de salud durante un período de seguimiento, se pueden tener estudios que trabajan con casos de

---

[3] IBIDEM (2) pag.35.

prevalencia o estudios que trabajan con casos de incidencia.

A la enfermedad se le puede considerar como un estado de salud o un cambio en el estado de salud y en ambos casos la variable enfermedad puede ser contínua o categórica.

La prevalencia es una medida dicotómica de un estado de salud, es decir, es la presencia de una enfermedad en un punto del tiempo hipotético. Los casos de prevalencia son los casos existentes en un momento dado, para los cuales puede desconocerse la duración de la enfermedad.

La incidencia es una medida dicótomica de un cambio en el estado de salud, es decir, es el desarrollo de una enfermedad durante un período dado. Dependiendo del tipo de enfermedad y de los objetivos del estudio, se puede estar interesado en la primera ocurrencia de la enfermedad, en todas las ocurrencias o en la muerte por una o más enfermedades entre la población total o entre casos de enfermedad.

II.2 DISEÑOS MAS UTILIZADOS

De la descripción de las características que ayudan a tipificar un estudio epidemiológico, que se realizó en la sección anterior, puede observarse que al combinar dichas características se obtiene una gama muy amplia de posibilidades para diseñar un estudio epidemiológico.

En esta sección se describirán únicamente las características de los diseños que más se han utilizado tradicionalmente en epidemiología. Si se está interesado en conocer las características de algunas otras opciones, se puede consultar a Kleinbaum, Kupper y Morgenstern (1982) ó Méndez, et al (1984), debiéndose revisar previamente las definiciones de los criterios que tipifican los diferentes diseños, ya que éstas pueden variar de un autor a otro.

II.2.1 Estudio de Cohortes

Por cohorte se entiende un grupo de individuos que tienen o tuvieron una experiencia en común (tratamiento ó exposición a algún agente) o que comparte algún atributo específico (raza, nivel socioeconómico, edad, etc.).

En un estudio de cohortes se consideran dos o más grupos de individuos, definidos según el valor del factor de riesgo o tratamiento en ellos y se siguen a través del tiempo para conocer su evolución, es decir, para conocer qué valores presenta el factor respuesta (enfermedad, muerte, etc.).

En este diseño se conoce necesariamente la información sobre el factor de riesgo o tratamiento al principio del período de seguimiento; es comparativo, observacional, hacia adelante, longitudinal, puede ser retrospectivo, ambipectivo o prospectivo; por lo general se trabaja con casos de incidencia.

Estos estudios de cohortes se realizan con la finalidad de contrastar hipótesis relativas a la etiología de enfermedades. Son útiles para estudiar enfermedades frecuentes y con períodos de duración relativamente cortos entre la primera exposición y el principio clínico y entre el principio clínico y la terminación de la enfermedad (si la

duración fuera grande es posible que no hubiera seguimiento). Se prefieren a otro tipo de diseños, cuando el factor de riesgo es difícil de encontrar en la población objetivo. No debe utilizarse un diseño de cohortes (sobre todo prospectivo o ambipectivo) cuando el resultado estudiado es raro pues se requiere de tamaños de muestra muy grandes y en consecuencia son muy costosos.

Un estudio de cohortes puede involucrar a una poblacción fija o a una dinámica, ya que es un estudio longitudinal.

Un estudio de cohortes retrospectivo es más barato que uno prospectivo. Es también más factible para estudiar una enfermedad rara o con largo período de latencia pero depende de la disponibilidad de información sobre el factor de riesgo o tratamiento (previa a la información sobre el factor respuesta) en una población bien definida, a la que se le hará el seguimiento para la detección de nuevos casos o muertes.

Con un estudio de cohortes prospectivo con casos de incidencia se establece la máxima evidencia de causalidad no experimental.

A continuación se resumen las ventajas y desventajas de un diseño de cohortes.

Entre las ventajas se pueden mencionar:

* Permite explorar la presencia de otros efectos no previstos (principalmente los prospectivos y ambipectivos).
* Permite describir completamente la experiencia subsecuente a la exposición: tasas de progresión, evaluación de la enfermedad, efectos colaterales (principalmente los prospectivos y ambipectivos ya que los retrospectivos en general dependen de información que fue captada para propósitos diferentes a los de investigación en curso).
* Se pueden estimar riesgos y tasas específicas por grupo de exposición y en consecuencia riesgos relativos y atribuibles, razones y diferencias de tasas entre grupos de exposición.
* Proporciona una gran flexibilidad en la selección de variables a estudiar y en su obtención sistemática (principalmente en prospectivos y ambipectivos).
* Se controla más la calidad en la medición de variables (en prospectivo y ambipectivo principalmente).
* Permite el control de factores de confusión potenciales en la etapa de formación de los grupos.
* El nivel del factor de estudio se observa en cada sujeto al principio del período de seguimiento, es decir, antes de que se detecte la enfermedad. En consecuencia, se puede estar razonablemente seguro de que la causa hipotetizada precedió a la ocurrencia de la enfermedad y que el estado de enfermedad no influenció diferencialmente la selección de sujetos por nivel del factor de estudio.

Entre las desventajas de un estudio de cohortes se pueden mencionar:

* Que son de larga duración y alto costo (principalmente las prospectivos y

ambipectivos).

* Que al no aleatorizar los niveles del factor de estudio en los sujetos, no se controla por factores de confusión presentes pero desconocidos.

* Que se pierden sujetos por migración, falta de participación y muerte por causas ajenas a las estudiadas si el período de seguimiento es largo. Este hecho puede producir distorsión en los resultados.

* Que no se pueden generar nuevas hipótesis etiológicas respecto a la enfermedad en estudio ya que la información sobre el factor de riesgo o tratamiento tiene que obtenerse al inicio de la investigación.

* Que el seguimiento en sí, puede influenciar el comportamiento de los sujetos del grupo.

* Que es estadística y prácticamente ineficiente para estudiar enfermedades poco frecuentes debido a que la información sobre el factor de riesgo o tratamiento debe recolectarse en un número muy grande de sujetos de los cuales sólo un número pequeño presentará la enfermedad.

II.2.2 Estudios de Casos y Controles

En este tipo de estudios se forman uno o más grupos de sujetos que presentan un determinado resultado (enfermedad, muerte, sobrevida invalidez, etc.) y otro u otros grupos que no presenta(n) dicho(s) resultado(s). Estos grupos se comparan respecto a la exposición actual o previa a uno o varios factores de riesgo. A los sujetos que presentan el resultado (enfermedad, muerte, etc.) se les conoce como casos y a los que no lo presentan se les conoce como controles o testigos.

Este tipo de estudios se llevan a cabo con la finalidad de contrastar alguna hipótesis etiológica específica. Sin embargo, cuando no se tiene clara una hipótesis etiológica específica, se hacen para explorar los antecedentes de las personas afectadas y no afectadas, en base a varias hipótesis etiológicas posibles.

Un estudio de casos y controles es un estudio comparativo observacional que puede ser sin dirección o hacia atrás; transversal o longitudinal; retrospectivo o ambipectivo y que puede trabajar con casos de incidencia o prevalencia.

Los casos y controles en la práctica se seleccionan de grupos diferentes. Sin embargo para poder hacer una inferencial causal, los controles deben representar a la población de la cual se obtuvieron los casos. En otras palabras los procedimientos de selección deben equivaler a una selección estratificada aleatoria de una población dividida en el estrato de los casos y en el estrato de los controles.

Los casos pueden estar constituidos por:

* Todas (o una muestra de) las personas que presentan el resultado (enfermedad por ejemplo) atendidas en una determinada institución, durante un intervalo de tiempo específico.

* Todas (o una muestra de) las personas que presentan el resultado en una

población más general, como una ciudad, municipio, estado, etc., en cierto período de tiempo. Aunque la muestra de casos no sea estrictamente aleatoria se considera como si lo fuera para efectos de análisis.

Los controles pueden obtenerse a partir de:

* La población de una área geográfica determinada. Esta es una fuente adecuada de controles, cuando los casos fueron extraídos de los casos que ocurrieron en esta población.

* De miembros de las instituciones de donde se obtuvieron los casos, pero que no presenten el resultado que tipifica a los casos. En esta situación se debe tener cuidado con las conclusiones del estudio, ya que puede haber factores que influyan para que los sujetos asistan a tal institución, y por ello no sean representativos de una población más general.

* Se pueden obtener controles entre los parientes de los casos.

* También se pueden obtener controles entre los compañeros de los casos que están expuestos a un ambiente similar.

La forma en la cual se seleccionan controles a partir de las fuentes antes mencionadas puede variar. Se pueden obtener controles mediante un muestreo aleatorio simple sin reemplazo; mediante un muestreo sistemático o estratificado para tener la misma distribución en ciertas características que los casos; se pueden obtener controles para formar una muestra aparejada individualmente con los casos, controlando de esta forma algunos factores de confusión. En cualquier instancia, se considera que se tiene una muestra aleatoria de los sujetos libres de enfermedad, ya sea de una sóla población o de cada subpoblación formada por la estratificación o por factores de aparejamiento.

Es importante hacer notar que en un estudio de casos y controles, éstos podrían ser identificados a través del tiempo como se hace en un estudio de cohortes, en ésta situación se trabajaría con casos de incidencia.

Para probar hipótesis etiológicas es preferible trabajar con casos de incidencia que con casos de prevalencia, porque aunque sea evidente que la exposición ocurrió antes del comienzo del resultado (enfermedad), quizás no sea posible determinar en qué medida una característica particular más frecuente en los casos, está más relacionada con la duración y curso del resultado que con su etiología.

El análisis de un estudio de casos y controles consiste básicamente en la comparación entre los dos grupos respecto a la frecuencia de los factores cuya posible influencia etiológica se está explorando.

Entre las principales ventajas de este tipo de estudios se pueden mencionar las siguientes:

* Son muy útiles para explorar la etiología de enfermedades poco frecuentes y/o con períodos de latencia o expresión grandes.

* En general son más cortos y menos costosos que los estudios de cohortes. Con frecuencia se utilizan como primer paso en la detección de asociaciones entre la

causa potencial y el factor respuesta o para escoger entre varias hipótesis que pudieran explicar las características observadas de la enfermedad.

* Permiten identificar causalidad multifactorial. El investigador puede analizar la influencia de varios factores de riesgo mediante un sólo estudio.

* Son más factibles de llevarse a cabo.

Entre las desventajas de este tipo de estudios se pueden mencionar las siguientes:

* Como en cualquier estudio no aleatorizado no existe la seguridad de que los grupos de casos y controles sean comparables respecto a factores de confusión potenciales y otras fuentes de distorsión. En este caso la inseguridad surge de dos hechos: a) la información sobre el factor de estudio se obtiene después de que la enfermedad (resultado) ha ocurrido; b) los casos y controles se obtienen de grupos que pueden ser de poblaciones diferentes.

* No se puede realizar una eliminación efectiva de algunos factores de confusión potenciales en la relación temporal correcta, puesto que los atributos se igualan al final de la evaluación, esto es, entre casos y controles.

* Pueden existir sesgos de información respecto al factor en estudio (riesgo). Estos surgen cuando la información sobre el factor de estudio no se capta de la misma manera para los casos que para los controles, ya sea porque los registros de donde se obtiene la información de casos, difieren de los de los controles, o porque el recuerdo de la exposición difiere entre unos y otros.

* Pueden existir también sesgos de selección. Estos sesgos ocurren cuando los casos (o controles) tienen diferente probabilidad de ser seleccionados como expuestos y no expuestos de la población objetivo.

* La enfermedad (resultado) debe medirse como una variable categórica. Además debe ser reconocida como el estado de salud de interés antes de que los sujetos sean seleccionados y por lo tanto este tipo de estudios no sirven para explorar los posibles efectos de un determinado factor de riesgo o tratamiento.

* Como no pueden ser hacia adelante, la capacidad del investigador para distinguir si el factor de riesgo precede realmente al factor respuesta, dependerá de la calidad de la información obtenida retrospectivamente sobre el factor de riesgo.

* En general no pueden obtenerse medidas de frecuencia de enfermedad específicas para los grupos de exposición. Además entre las medidas del efecto que el factor de riesgo tiene en el factor respuesta, sólo puede estimarse la razón de momios.

Sin embargo, bajo ciertas condiciones especiales pueden obtenerse algunas medidas de frecuencia específicas o al menos del efecto:

* Cuando los casos representan a todos los casos de la enfermedad en una población determinada (o son una muestra aleatoria de ellos) y los controles son una muestra aleatoria de la misma población, es decir, cuando las

probabilidades de selección de casos y controles son conocidas, entonces sí es posible estimar los riesgos de enfermedad separadamente para los grupos de exposición, y en consecuencia el riesgo relativo y el riesgo atribuible.

* Cuando no se conocen las probabilidades de selección de casos y controles a partir de la población objetivo, aunque se sabe que los controles se obtuvieron de la misma población que originó los casos, no se pueden estimar los riesgos de enfermedad por grupos de exposición. El riesgo relativo puede estimarse en este caso, si se cumplen dos condiciones: Primera, que el resultado sea poco frecuente, lo que implica que el riesgo relativo de la población objetivo es aproximadamente igual a la razón de momios de la población objetivo. Segunda, que no haya sesgos de selección, es decir, que los casos y controles se escojan de tal manera que no se favorezcan a los expuestos ni a los no expuestos (esto significa que la probabilidad de selección de casos expuestos es igual a la de casos no expuestos y que la probabilidad de selección de controles expuestos es igual a la de controles no expuestos). Bajo esta condición el estimador de la razón de momios es un buen estimador de la razón de momios poblacional.

Bajo las suposiciones anteriores se puede estimar el riesgo relativo mediante el estimador de la razón de momios.

Esquemáticamente la situación es la siguiente:

POBLACION

	Expuesto	No Expuesto	
Enfermo	A	B	$A+B$
No Enfermo	C	D	$C+D$
	$A+C$	$B+D$	

MUESTRA

	Expuesto	No Expuesto	
Enfermo	a	b	$a+b$
No Enfermo	c	d	$c+d$

El riesgo relativo es $RR = \dfrac{A/(A+C)}{B/(B+D)}$. Si el resultado es poco frecuente, entonces

$$RR \doteq \frac{A/C}{B/D} = \quad \text{Razón de momios.}$$

Si además, no hay sesgos de selección, entonces:

$$\frac{a}{A} \doteq \frac{b}{B} \quad y \quad \frac{c}{C} \doteq \frac{d}{D}$$

de lo anterior,

$$\frac{a}{b} \doteq \frac{A}{B} \quad y \quad \frac{c}{d} \doteq \frac{C}{D}$$

$$\Longrightarrow \text{Razón de momios} \quad = \frac{AD}{BC} \doteq \frac{ad}{bd} = \quad \text{Razón de momios estimada}$$

En conclusión, la razón de momios estimada[4] se puede utilizar para estimar el riesgo relativo en la población objetivo, si las dos condiciones mencionadas se satisfacen.

II.2.3 Estudios Transversales[5]

En un estudio transversal se obtiene una sola muestra, por lo general aleatoria, de una población objetivo. Después de obtenida la muestra, se obtiene de todos los sujetos seleccionados, información acerca del valor del factor resultado (p. ej. estado de enfermedad) y del nivel actual o pasado del factor de estudio.

En general, el propósito de estudios de este tipo, es describir las características de la población objetivo y encontrar asociaciones entre variables que sirvan de base para la formulación posterior de hipótesis etiológicas.

Un estudio transversal, es un estudio descriptivo, observacional, que puede ser sin dirección o hacia atrás, generalmente retrospectivo. Se mide prevalencia de enfermedad y no incidencia.

Entre sus ventajas pueden mencionarse:

* Son útiles para estudiar características cuantitativas y que pueden variar a través del tiempo, también para estudiar enfermedades relativamente frecuentes de larga duración.
* Son útiles para planear servicios y programas de salud.
* Sirven para generar nuevas hipótesis etiológicas relativas a factores de estudio y/o enfermedades.

Entre las desventajas pueden mencionarse:

---

[4] Obsérvese que dado que se toman muestras de enfermos y no enfermos, entonces $a + b$ y $c + d$ son cantidades fijas, en tanto que $a + c$ y $b + d$ son aleatorias, por lo que no se puede estimar directamente el riesgo relativo como

$$\widehat{RR} = \frac{a/a + c}{b/b + d}$$

[5] En este caso la palabra transversal no tiene el significado que se le dió en la sección II.1, simplemente es un nombre que se ha utilizado comúnmente para estudios con ciertas características particulares.

* Como se utilizan datos de prevalencia éstos no pueden ser utilizadas de la misma manera que los de incidencia, para determinar la dirección de la relación entre el factor de estudio y el factor respuesta, por lo tanto no puede decidirse a partir de los resultados de este estudio si la causa hipotética es antecedente o consecuente del factor respuesta.

* No son útiles para estudiar enfermedades raras o de corta duración.

II.3 CRITERIOS PARA ELEGIR UN DISEÑO

La elección de un diseño depende de varios factores:

* Los objetivos de la investigación.

* Las características de la enfermedad que se estudia.

* La disponibilidad de información.

* La disponibilidad de tiempo y recursos.

Los objetivos de la investigación especifican si lo que se pretende es: probar una hipótesis etiológica específica o describir las características de una enfermedad en una determinada población. También se especifica si se pretende estudiar el componente genético de la etiología de una enfermedad o evaluar el impacto de una intervención o bien planear los programas y servicios de salud, etc. Si por ejemplo, se quiere probar una hipótesis etiológica, se deberá decidir:

* Qué medida de frecuencia de enfermedad y qué medida del efecto se utilizarán con lo cual se eligirá algún diseño (p. ej. cohortes o casos y controles).

* Qué diseño será más adecuado para determinar si el factor de estudio precedió a la enfermedad, en cuyo caso se elegirá un diseño hacia adelante de preferencia y con casos de prevalencia.

* Y, en general se tratará de evitar cualquier fuente de error que pudiera afectar la validez del estudio. Como ejemplo se elegiría un diseño prospectivo, hacia adelante, longitudinal, con casos de incidencia.

La naturaleza de la enfermedad que se estudia es también un factor determinante en la elección de un diseño. Si la enfermedad es poco frecuente o con largos períodos de latencia y/o expresión, no conviene utilizar un estudio de cohortes prospectivo, sería mejor un retrospectivo o un casos y controles. En cambio, si la enfermedad es muy frecuente y/o con cortos períodos de latencia o expresión, sí convendría utilizar un diseño de cohortes prospectivo.

El nivel de conocimiento existente acerca de la enfermedad en estudio determinará a su vez los objetivos de la investigación, es decir, si se tiene en mente una hipótesis etiológica específica o si se piensan explorar asociaciones entre múltiples factores de riesgo o tratamiento y la enfermedad. .

Es importante también considerar la disponibilidad de datos y recursos en general, ya que esto es lo que determinará la viabilidad del diseño.

Puede decirse en conclusión, que no existe el diseño perfecto, el ideal, o el

mejor, puesto que cada circunstancia particular determinará el diseño que responda más adecuadamente a sus necesidades. Sin embargo, siempre deberá tenerse en consideración las bondades y defectos de cada una de las características que tipifiquen al diseño elegido, para evitar errores irreversibles.

CAPITULO III

ESTUDIOS DE CASOS Y CONTROLES

En este Capítulo se describen las características generales de los estudios de casos y controles y se presentan los métodos que tradicionalmente se han utilizado para el análisis de los mismos.

En primer lugar, se presentan los métodos de análisis para estudios en los que no se utiliza en el diseño ningún tipo de aparejamiento para el control de factores de confusión. Estos métodos se aplican también para el análisis de estudios donde el aparejamiento utilizado es por grupo (por ejemplo aparejamiento por frecuencias, ver Capítulo I).

En segundo lugar se presentan los métodos de análisis para estudios en donde el control de factores de confusión se realiza mediante aparejamiento individual (ver Capítulo I).

III.1 CARACTERISTICAS GENERALES

III.1.1 Objetivo

Estos estudios generalmente se llevan a cabo con la finalidad de investigar la etiología de enfermedades poco frecuentes. Las medidas de interés, del efecto que el factor de riesgo tiene en la enfermedad, deben ser la razón de tasas de incidencia si la enfermedad tiene un período de riesgo largo (latencia larga y el factor de riesgo o tratamiento representa una exposición de larga duración) respecto al período de observación o la razón de riesgos si la enfermedad tiene un periodo de riesgo restringido (latencia corta y el factor de riesgo representa un atributo fijo o una exposición de corta duración) respecto al periodo de observación.

Usualmente en estos estudios se pretende contrastar una hipótesis etiológica específica, sin embargo, en las ocasiones en que no se tiene clara una hipótesis específica, se realizan para explorar los antecedentes de las personas afectadas y no afectadas por la enfermedad, en base a varias hipótesis plausibles.

III.1.2 Diseño

Se forman uno o más grupos de sujetos que presentan un determinado resultado (enfermedad, muerte, etc.), y otro u otros grupos que no presentan dicho(s) resultado(s). Estos grupos se comparan respecto a la exposición actual o previa a uno

o varios factores de riesgo o tratamiento.

A los sujetos que presentan el resultado se les conoce como casos y a los que no lo presentan se les conoce como controles o testigos.

Los casos y controles deben de pertenecer a la misma población o, lo que es lo mismo, deben de ser comparables.

Definición y Selección de Casos

Para definir a los casos se requiere:

* Proporcionar una descripción clara y precisa del problema a investigar (muerte, enfermedad, etc.) así como el criterio diagnóstico a utilizar.
* Determinar las fuentes de donde se obtendrán los casos. Los casos pueden estar constituídos por:
  * Todos o una muestra (aleatoria de preferencia) de las personas que presentan el resultado, atendidas en una determinada institución durante un intervalo de tiempo. En esta situación el marco de muestreo pueden ser las hojas de admisión hospitalaria o las listas de diagnóstico de egreso hospitalario.
  * Todos o una muestra de las personas que presentan el resultado en una población más general, como una ciudad, municipio, estado, etc., en cierto período de tiempo.
* Decidir si se trabaja con casos de incidencia o de prevalencia. Es preferible (especialmente si el objetivo es la etiología de una enfermedad) trabajar con casos nuevos, diagnosticados en un período de tiempo determinado. El trabajar con casos de prevalencia desde luego, proporciona un número mayor de casos, pero al tener a estos en diferentes etapas del proceso de enfermedad o con recurrencias, complica la interpretación de hallazgos. Además, en pacientes con un grado avanzado de enfermedad, puede ser difícil diferenciar los factores que son causa de los que son consecuencia de la misma.

Definición y Selección de Controles

El propósito del grupo control es determinar la tasa o riesgo de exposición que tendrían los casos, suponiendo que no hay asociación entre la exposición (factor de riesgo o tratamiento) y la enfermedad. Por lo mismo, los controles deben ser comparables a los casos en todos los aspectos relevantes excepto en que no tienen el resultado (enfermedad, muerte, etc.) que se estudia. Esto permite que las diferencias en las tasas o riesgos de exposición puedan reflejar una asociación verdadera entre el factor de riesgo o tratamiento y la enfermedad.

Al elegir un grupo de controles deben tenerse en cuenta las siguientes consideraciones:

* Seguridad de que la información sobre los factores de riesgo o tratamiento que se estudian, pueda obtenerse para el grupo control, de manera similar a aquella con la que se obtuvo la información respectiva sobre los casos.

* Determinación de si la elección de controles se hará de tal forma que éstos sean similares a los casos respecto a ciertos factores que puedan distorsionar los resultados, es decir, decidir si se realizará un aparejamiento como herramienta para el control de factores de confusión.

* Asegurar que los controles provengan de una población similar a la que originó los casos.

Entre las fuentes para la elección de controles se pueden mencionar:

* La población de un área geográfica determinada. Esta es una fuente adecuada de controles siempre y cuando los casos sean representativos de los casos que ocurren en esta población.

* Los miembros de las instituciones de donde se obtuvieron los casos, pero que no presenten el resultado (muerte, enfermedad, etc.).

* Los parientes de los casos.

De estas fuentes, los controles pueden seleccionarse de diversas formas: con muestreo aleatorio simple, muestreo sistemático, muestreo estratificado aleatorio o una selección que garantice cualquier tipo de aparejamiento para controlar los factores de confusión. En cualquier caso, en la práctica se opera como si se tuviera una muestra aleatoria de controles.

Un requerimiento fundamental en los procedimientos de selección de casos y controles, es que tanto los casos como los controles sean seleccionados independientemente de su condición de exposición al factor de riesgo.

Información sobre la Exposición al Factor de Riesgo

El factor de riesgo o tratamiento debe definirse en términos claros y medibles. Las fuentes para obtener información sobre el mismo, pueden ser: registros hospitalarios, estadísticas vitales, registros de empleo, seguros o servicios sociales o los propios sujetos bajo investigación.

Se debe obtener la misma información y con el mismo método tanto para casos como para controles.

III.1.3 Sesgos y su Control

En los estudios de casos y controles se pueden presentar sesgos de selección

cuando los casos o los controles se seleccionan influenciados por su condición de exposición. Estos sesgos deben controlarse en la etapa del diseño del estudio, ya que para corregir las distorsiones que producen en los estimadores, en la etapa de análisis, se requeriría del conocimiento de las probabilidades de selección de casos y controles como expuestos y no expuestos, y en general no se dispone de ellas.

Los sesgos de información pueden surgir en este tipo de estudios, por errores en la captación de información sobre el factor respuesta que define a los casos y controles (con criterios de diagnóstico no homogéneos por ejemplo) o por errores en la captación de la información del factor de riesgo o tratamiento (por ejemplo, sesgos que surgen por la memoria diferencial sobre exposición entre casos y controles). Deben controlarse en la etapa de diseño, porque de no ser así, se requiere del conocimiento de la magnitud del sesgo para corregir los estimadores en la etapa de análisis.

Finalmente, pueden presentarse también sesgos por la presencia de factores de confusión (ver sección 1.2.2, Capítulo I). El control de factores de confusión puede realizarse en la etapa de diseño de la investigación mediante métodos de aparejamiento (ver sección 1.2.2, del Capítulo I), y desde luego considerar dicho aparejamiento en los métodos de análisis. También puede controlarse la presencia de factores de confusión únicamente en la etapa de análisis mediante estratificación (ver sección 1.2.2, Capítulo I).

Respecto al aparejamiento, hay que tener cuidado de no realizarlo con variables que no son de confusión, porque de ser así, se puede obscurecer la relación entre el factor de riesgo o tratamiento y el factor respuesta. A este tipo de error se le conoce como sobreaparejamiento ("overmatching", ver Capítulo III, Breslow y Day (1982)).

III.1.4 ¿Qué se puede estimar a partir de un estudio de casos y controles?

Para poder comprender que es lo que se puede estimar en un estudio de casos y controles considérese el siguiente desarrollo:

Supóngase que se tiene una población a la cual se estudia por un período determinado $(0,\tau)$. Al principio del período de estudio, se clasifica a los individuos en base a la exposición a un factor de riesgo o tratamiento dicotómico (expuesto o no expuesto) y al final se les clasifica de acuerdo a si desarrollaron o no una determinada enfermedad.

Sea:

p: proporción de individuos expuestos al inicio del período de estudio.

$q = 1 - p$

$P_1 = P_1(\tau)$: la probabilidad o riesgo de que un sujeto expuesto desarrolle la enfermedad durante el período de observación $(0,\tau)$

$Q_1 = 1 - P_1$

$P_0 = P_0(\tau)$: la probabilidad o riesgo de que un sujeto no expuesto de sarrolle la enfermedad durante el mismo período $(0, \tau)$

$Q_0 = 1 - P_0$

$\Delta_1 = \Delta_1(\tau) = \int_0^\tau \lambda_1(t)dt$: tasa de incidencia de la enfermedad, acumulada para el período de duración τ, para el grupo de expuestos.

$\lambda_1(t)$: tasa instantánea de incidencia de la enfermedad al tiempo t, para el grupo de expuestos.

$\Delta_0 = \Delta_0(\tau) = \int_0^\tau \lambda_0(t)dt$: tasa de incidencia acumulada, para el período de duración τ, para el grupo que no está expuesto.

$\lambda_0(t)$: tasa instantánea de incidencia de la enfermedad al tiempo t, para el grupo que no está expuesto.

$\psi = \frac{\frac{P_1}{(1-P_1)}}{\frac{P_0}{(1-P_0)}} = \frac{P_1 Q_0}{P_0 Q_1}$: razón de momios de enfermedad.

Las proporciones esperadas en cada una de las 4 categorías definidas por el factor de riesgo y respuesta son:

Factor de Riesgo			
	Expuesto	No Expuesto	
Enfermo	pP_1	qP_0	$pP_1 + qP_0$
No Enfermo	pQ_1	qQ_0	$pQ_1 + qQ_0$
	p	q	

Si el período de estudio $(0, \tau)$ es pequeño y/o la enfermedad es poco frecuente, resulta que P_1 y P_0 serán pequeñas y entonces se tendrá que:

$$\frac{\Delta_1}{\Delta_0} \doteq \frac{P_1}{P_0}$$

es decir, que la razón de tasas de incidencia acumuladas para el período, será

aproximadamente igual a la razón de riesgos del período [1]

Además, al ser P_1 y P_0 pequeños se tiene que:

$$\frac{\Delta_1}{\Delta_0} \doteq \frac{P_1}{P_0} \doteq \frac{P_1 Q_0}{P_0 Q_1} = \psi$$

Entonces se tiene que con la razón de momios de enfermedad se puede aproximar el riesgo relativo, así como a la razón de tasas de incidencia acumuladas, para el periodo $(0, \tau)$.

Considérese ahora, las probabilidades de exposición dada la condición de enfermedad.

$$P(Expuesto \mid Enfermo) = \frac{pP_1}{pP_1 + qP_0} = P_1'; Q_1' = 1 - P_1'$$

$$P(Expuesto \mid No\ Enfermo) = \frac{pQ_1}{pQ + qQ_0} = P_0'; Q_0' = 1 - P_0'$$

La razón de momios de exposición es entonces:

$$\psi' = \frac{\frac{P_1'}{Q_1'}}{\frac{P_0'}{Q_0'}} = \frac{\frac{pP_1}{qP_0}}{\frac{pQ_1}{qQ_0}} = \frac{P_1 Q_0}{P_0 Q_1} = \psi$$

Resulta entonces que con la razón de momios de exposición (que es igual a la razón de momios de enfermedad) se puede aproximar tanto a la razón de riesgos del período como a la razón de tasas acumuladas para el período, siempre y cuando P_1 y P_0 sean pequeñas.

Las posibilidades para construir estimadores de los diferentes parámetros dependen de como se diseñe la investigación:

* Si se obtiene una muestra aleatoria de la población y se le observa durante $(0, \tau)$, se pueden estimar todos los parámetros: $p, P_1, P_0, \Delta_1, \Delta_0, \psi, \frac{\Delta_1}{\Delta_0}, \frac{P_1}{P_0}, P_1 - P_0, P_1', P_0', \psi'$.

---

[1] Si $P(\tau)$ es el riesgo de enfermedad para un período de longitud τ, éste puede aproximar a la tasa de incidencia acumulada para el período $\Delta(\tau)$: Sea T: tiempo que transcurre entre el inicio del estudio y la manifestación de la enfermedad en un sujeto.

$$\lambda(t) = \lim_{\Delta_t \to 0} \frac{P[t \le T \le \Delta_t \mid T \ge t]}{\Delta_t} = \lim_{\Delta_t \to 0} \frac{P[t \le T \le t + \Delta_t]}{\Delta_t\ P[T \ge t]}$$

$$= \lim_{\Delta_t \to 0} \frac{F_T(t + \Delta_t) - F_T(t)}{\Delta_t\ P[T \ge t]} = \frac{F'(t)}{1 - F_T(t)} = \frac{F'(t)}{(1 - P(t))} = -d(\ln(1 - P(t))$$

Entonces $\Delta(\tau) = \int_0^\tau \lambda(t)dt = -\ln(1 - P(\tau))$ y si $P(\tau) \doteq 0$ entonces $\Delta(\tau) \doteq P(\tau)$

* Si se toman muestras aleatorias de expuestos y no expuestos al principio del estudio, entonces se pueden estimar, $P_1, P_0, \Delta_1, \Delta_0, \frac{P_1}{P_0}, P_1 - P_0, \frac{\Delta_1}{\Delta_0}$.

* Si se realiza un estudio de casos y controles, es decir, se toman muestras aleatorias de casos y controles y se investiga su exposición pasada, se pueden estimar P(Expuesto|caso), P(Expuesto|control) y la razón de momios de exposición. Sin embargo, bajo ciertas circunstancias, que a continuación se mencionan es posible estimar algunos otros parámetros a partir de un diseño como éste:

* Cuando los casos son todos los que existen en una población finita en un determinado período o son una muestra probabilística de los mismos, y los controles son una muestra probabilística de la misma población pero de los sujetos que no padecen la enfermedad en estudio, es decir, cuando las probabilidades de selección de casos y controles son conocidas, entonces si es posible estimar los riesgos de enfermedad separadamente para los grupos de exposición, y en consecuencia el riesgo relativo y la diferencia en riesgos (riesgo atribuible). Considérese el siguiente ejemplo tomado de Mausner y Bahn (1977) (la descripción del procedimiento se ha reanalizado).

En Boston entre 1967-1968 se diagnosticaron 547 casos masculinos de cáncer en la vejiga, de los cuales se seleccionaron aleatoriamente 375 para una investigación sobre el efecto que trabajar en una cierta industria, tiene en dicha enfermedad. En ese mismo período se seleccionaron aleatoriamente de la población ubicada en la misma área y de los grupos de edad y sexo correspondientes a los casos, 368 controles.

* Los resultados del estudio fueron los siguientes:

Trabaja en
la Industria

	Si	No	Total
Caso	118	257	375
Control	69	299	368

Entonces

$\hat{P}$ (Expuesto | caso) = 0.314 $\quad$ $\hat{P}$ (Expuesto | control) = 0.188

En Boston la población masculina de la edad de interés en ese período de tiempo era de 850,000 hombres.

El total estimado de sujetos expuestos en la población no enferma, es: (850,000 - 547) × 0.188 = 159,697.

El total estimado de sujetos no expuestos en la población no enferma, es: (849,453)(0.812) = 689,756

El total estimado de sujetos expuestos entre los casos es de 547 × 0.314 = 172.

El total estimado de sujetos no expuestos entre los casos es 375.

Resumiendo los resultados de los totales poblacionales se tiene:

Trabaja en la Industria

	Si	No	Total
Caso	172	375	547
Controles	159,697	689,756	849,453
Total	159,869	690,131	

de aquí se puede estimar el riesgo de enfermedad para los expuestos como 172/159,869 y el riesgo de enfermedad para los no expuestos como 375/690,131.

* Cuando no se conocen las probabilidades de selección de casos y controles, pero se sabe que los controles se obtuvieron de la misma población que originó a los casos, no se puede estimar el riesgo de enfermedad por grupo de exposición. Sin embargo, si no hay sesgos de selección y P_1 y P_0 son pequeños o sea que la enfermedad es poco frecuente (cabe hacer notar que es la situación cuando más se utilizan este tipo de estudios), se pueden estimar la razón de tasas de incidencia acumuladas o el riesgo relativo para el período, aproximados mediante la razón de momios de enfermedad, mediante el estimador de la razón de momios de exposición.

III.1.5 Consideraciones Generales para el Análisis

El objetivo del análisis es determinar el efecto que el factor de riesgo o tratamiento tiene en el factor respuesta (muerte, enfermedad, etc.). Para cumplir tal objetivo es necesario asegurarse de que el efecto observado no es resultado de algún sesgo o de la presencia de factores de confusión.

Además, es muy importante estudiar las posibles interacciones que el factor de riesgo o tratamiento y el factor respuesta tengan con otros factores, para poder de este modo, describir claramente como es que se comporta dicho efecto ante la presencia de esos otros factores. Debe notarse que a este respecto, el objetivo no debe ser eliminar interacciones mediante alguna transformación adecuada, sino comprender su naturaleza.

Análisis Preliminares

Como un primer paso, es conveniente describir el perfil tanto del grupo de casos como el de los controles, en términos de la distribución y de las interrelaciones que en cada uno de dichos grupos, tienen las diferentes variables incluidas en el estudio

así como ciertas características generales que pueden influir en los resultados, como sexo, edad, raza, lugar de nacimiento, método de diagnóstico, etc. El hacer estas descripciones proporciona pautas importantes para la interpretación de los resultados, permite verificar si los procedimientos de aparejamiento se realizaron adecuadamente y explorar fuentes posibles de sesgos. Muchas veces la exposición al factor de riesgo o tratamiento se mide de varias formas (duración, intensidad, etc.). En análisis preliminares conviene explorar la relación de cada una de dichas formas con el factor respuesta para posteriormente obtener una medida que comprenda a todas, pero que no obscurezca los efectos.

En análisis preliminares es importante explorar el efecto que tiene cada uno de los factores de riesgo o tratamiento de interés, en el factor respuesta, pero separadamente. Esto puede realizarse de diferentes maneras dependiendo de la escala de medición del factor de riesgo o tratamiento. Cuando el factor de riesgo es dicotómico, se construye una tabla de 2×2 y a partir de ella se pueden hacer inferencias sobre la razón de momios. Cuando el factor de riesgo es politómico, se elige a un nivel como nivel base de comparación contra el cual se comparan los demás. Se realizan inferencias para la razón de momios que relaciona a cada par de niveles y se comparan éstas. Cuando el factor de riesgo es una variable continua, en primera instancia puede convertirse en una variable categórica ordenada, dividiendo la escala de medición en intervalos (que equivale a un factor de riesgo politómico) con la finalidad de proporcionar una descripción libre de suposiciones, del cambio en la razón de momios al cambiar los niveles del factor de riesgo. En base al comportamiento observado en esta descripción, se puede postular posteriormente un modelo matemático paramétrico, para dicho comportamiento, en el cual la variable es tratada como continua.

Análisis Posteriores

En análisis subsecuentes se investiga en una serie de etapas, la acción combinada de varios factores de riesgo o tratamiento:

* Se puede considerar a cada factor de riesgo o tratamiento separadamente y analizar cómo los otros factores de riesgo modifican su efecto. Esta modificación puede originarse por un efecto de confusión general (la asociación entre los otros factores de riesgo distorsiona la asociación entre el factor de riesgo en estudio y el factor respuesta) o por la presencia de interacciones.

* Se puede examinar el efecto de varios factores de riesgo o tratamiento simultáneamente.

* Cuando se trabaja con muchos factores de riesgo y/o de confusión, es mejor utilizar modelos de regresión para investigar sus efectos en el factor respuesta.

III.1.6 Interpretación de Resultados

Para interpretar los resultados de un estudio de casos y controles se tienen que tener las mismas consideraciones que para cualquier otro estudio observacional:

* Evaluar si las dificultades metodológicas inherentes al diseño del estudio o los problemas que surgieron durante el desarrollo del mismo, no invalidan los resultados (por la presencia de sesgos y confusión).

* Si se encontró que el efecto que el factor de riesgo o tratamiento tiene en el factor respuesta es significativo, hay que decidir si esta relación puede considerarse como causal. No hay criterio estadístico que pueda decidir esto, sin embargo los criterios operacionales que se mencionaron en la sección I.1.2 del Capítulo I, pueden auxiliar en la toma de dicha decisión.

III.2 METODOS CLASICOS PARA EL ANALISIS CUANDO NO SE UTILIZA APAREJAMIENTO INDIVIDUAL

En virtud de que a partir de un estudio de casos y controles el único parámetro que realmente puede estimarse es la razón de momios de exposición [2], todos los métodos que se describen, se refieren a como estimar este parámetro y a cómo realizar pruebas de hipótesis acerca de él, pero bajo diferentes circunstancias.

En primer lugar se describen los métodos exactos y aproximados cuando se tiene un factor de riesgo dicotómico, sin controlar por la presencia de un factor de confusión, posteriormente se describen los métodos respectivos, cuando se controla por la presencia de dicho factor.

Por último se presentan muy brevemente los métodos para el caso de un factor de riesgo politómico, con y sin control por la presencia de un factor de confusión, y algunos comentarios acerca del caso en el que se tienen muchos factores de riesgo y de confusión. La brevedad en la exposición de estos últimos casos obedece a que con métodos tradicionales es prácticamente imposible el realizar los análisis, siendo mucho más fácil realizarlos mediante la utilización de modelos logísticos, los cuales se discutirán en el siguiente Capítulo.

III.2.1 Factor de Riesgo Dicotómico sin Control de Factor de Confusión. Análisis de una Tabla 2 × 2

ANALISIS EXACTO

Modelo

[2] Aunque bajo ciertas circunstancias que se mencionan en III.1 y que muy frecuentemente ocurren cuando se lleva a cabo un estudio de casos y controles, el estimador de la razón de momios estima aproximadamente la razón de riesgos y la razón de tasas de incidencia, en el desarrollo subsecuente únicamente se hará referencia a la razón de momios.

Se considera un estudio con una muestra aleatoria de n_1 casos y una muestra aleatoria de n_0 controles, obtenidas en forma independiente. Se tiene un sólo factor de riesgo o tratamiento dicotómico. La información se resume en una tabla cruzada de 2×2 de la siguiente forma:

Factor de Riesgo			
	Expuestos	No Expuestos	
Casos	a	b	n_1
Controles	c	d	n_0
	m_1	m_0	

Para poder realizar inferencias acerca de la razón de momios que se denotará por ψ, se requiere de un modelo probabilístico para los datos observados. El modelo que surge naturalmente para un estudio de casos y controles como el que se plantea, con factor de riesgo dicotómico, es el de un producto de distribuciones binomiales, puesto que las muestras de casos y controles se obtienen de manera independiente. El modelo resultante es el siguiente:

Sea X: número de expuestos entre los casos.

Y:número de expuestos entre los controles.

$$X \sim B(P_1, n_1) \text{ y } Y \sim B(P_0, n_0)$$

donde P_1: probabilidad de estar expuesto en la población de casos.

P_0: probabilidad de estar expuesto en la población de controles.

Entonces

$$P[X = a, Y = c] = \binom{n_1}{a} \binom{n_0}{c} P_1^a \, (1 - P_1)^{n_1 - a} \, P_0^c (1 - P_0)^{n_0 - c}$$

Esta distribución de probabilidad depende de dos parámetros P_1 y P_0. Como se mencionó anteriormente, en un estudio de casos y controles, el parámetro de interés es la razón de momios $\psi = \frac{\frac{P_1}{1-P_1}}{\frac{P_0}{1-P_0}}$, entonces lo ideal para facilitar las inferencias acerca de ψ, sería encontrar una distribución de probabilidad que dependiera únicamente de este

parámetro. En este caso, se puede encontrar dicha distribución si se aplica el principio de condicionalidad (ver Cox (1979), Cap. 4; Cox y Hinkley (1974), pp. 31-32; Gart (1970), pp. 471-472; Gart (1971), pp. 148-150; Lehmann (1959) pp. 140-143):

La distribución conjunta de X y Y puede reescribirse de la siguiente forma:

$$\begin{aligned} P[X=a, Y=c] &= \binom{n_1}{a}\binom{n_0}{c} Q_1^{n_1} Q_0^{n_0} \left(\frac{P_1}{Q_1}\right)^a \left(\frac{P_0}{Q_0}\right)^c \\ &= \binom{n_1}{a}\binom{n_0}{c} Q_1^{n_1} Q_0^{n_0}\, e^{a \ln \frac{P_1}{Q_1} - a \ln \frac{P_0}{Q_0} + c \ln \frac{P_0}{Q_0} + a \ln \frac{P_0}{Q_0}} \\ &= \binom{n_1}{a}\binom{n_0}{c} Q_1^{n_1} Q_0^{n_0}\, e^{a\left[\ln \frac{P_1}{Q_1} - \ln \frac{P_0}{Q_0}\right] + (a+c) \ln \frac{P_0}{Q_0}} \\ &= \binom{n_1}{a}\binom{n_0}{c} Q_1^{n_1} Q_0^{n_0}\, e^{a \ln \psi + (a+c) \ln \frac{P_0}{Q_0}} \end{aligned}$$

Entonces se observa que la distribución conjunta depende del parámetro que interesa ψ y de P_1 y P_0 que son parámetros de "estorbo".

Se propone como estadística auxiliar para el condicionamiento a $X+Y$. Para simplificar supóngase que $n_1 > m_1$ y $n_0 > m_1$

$$\begin{aligned} P[X+Y=m_1] &= \sum_{j=0}^{m_1} P[X=j,\ Y=m_1-j] \\ &= \sum_{j=0}^{m_1} \binom{n_1}{j}\binom{n_0}{m_1-j} P_1^j\, Q_1^{n_1-j}\, P_0^{m_1-j}\, Q_0^{n_0-m_1+j} \end{aligned}$$

Esta distribución depende de P_1 y P_0, entonces sí se puede utilizar a $X+Y$ como estadística auxiliar para el condicionamiento, ya que al condicionar no depende de P_0 y P_1,

$$\begin{aligned} P[X=a \mid X+Y=m_1] &= \frac{P[X=a,\ Y=m_1-a]}{P[X+Y=m_1]} \\ &= \frac{\binom{n_1}{a}\binom{n_0}{m_1-a} Q_1^{n_1}\, Q_0^{n_0}\, e^{a \ln \psi + m_1 \ln \frac{P_0}{Q_0}}}{\sum_{j=0}^{m_1}\binom{n_1}{j}\binom{n_0}{m_1-j} P_1^j\, Q_1^{n_1-j}\, P_0^{m_1-j}\, Q_0^{n_0-m_1+j}} \\ &= \frac{\binom{n_1}{a}\binom{n_0}{m_1-a} Q_1^{n_1}\, Q_0^{n_0}\, e^{a \ln \psi} \left(\frac{P_0}{Q_0}\right)^{m_1}}{\sum_{j=0}^{m_1}\binom{n_1}{j}\binom{n_0}{m_1-j} \left(\frac{P_1}{Q_1}\right)^j \left(\frac{P_0}{Q_0}\right)^{m_1-j} Q_1^{n_1}\, Q_0^{n_0}} \end{aligned}$$

$$= \frac{\binom{n_1}{a}\binom{n_0}{m_1-a} Q_1^{n_1} Q_0^{n_0} e^{a \ln \psi}}{\sum_{j=0}^{m_1}\binom{n_1}{j}\binom{n_0}{m_1-j}\left(\frac{P_1 Q_0}{P_0 Q_1}\right)^j}$$

$$= \frac{\binom{n_1}{a}\binom{n_0}{m_1-a}\psi^a}{\sum_{j=0}^{m_1}\binom{n_1}{j}\binom{n_0}{m_1-j}\psi^j}$$

La distribución resultante es una hipergeométrica no central, cuyo parámetro de no centralidad es precisamente el parámetro de interés ψ.

Resulta entonces que condicionando a que los marginales n_0, n_1, m_1 y m_0 están fijos, se obtiene un modelo probabilístico para los datos observados que no depende de parámetros de estorbo.

Eliminando la suposición de que $n_1 > m_1$ y $n_0 > m_1$ que se hizo para facilitar la discusión, el modelo queda:

$$P(X = a \mid n_0, n_1, m_0, m_1, \psi) = \frac{\binom{n_1}{a}\binom{n_0}{m_1-a}\psi^a}{\sum_j \binom{n_1}{j}\binom{n_0}{m_1-j}\psi^j}$$

donde $Max\{0, m_1 - n_0\} \leq j \leq Min\{n_1, m_1\}$

Debe observarse que la distribución es la misma si se intercambian los roles de n y m, por lo tanto, se puede aplicar también para el análisis de un estudio de cohortes con factor de riesgo dicotómico. Además, si se expresa la distribución en términos de cualquier otra entrada de la tabla, se obtiene también ψ ó ψ^{-1}.

Cuando $\psi = 1$ el modelo probabilístico que se obtiene es una hipergeométrica central:

$$P(X = a \mid n_0, n_1, m_0, m_1, 1) = \frac{\binom{n_1}{a}\binom{n_0}{m_1-a}}{\binom{n_1 + n_0}{m_1}}$$

Estimación Puntual de ψ

El estimador de máxima verosimilitud de ψ bajo el modelo condicional, después de realizar el álgebra correspondiente, se encuentra resolviendo la ecuación:

$$a = E[X \mid n_1, n_0, m_1, m_0; \psi]$$

Esta ecuación se resuelve con métodos numéricos y por lo tanto cuando las frecuencias en las celdas son muy grandes se convierte en una tarea pesada. En ese caso conviene más utilizar los métodos aproximados de análisis, que se mencionan posteriormente.

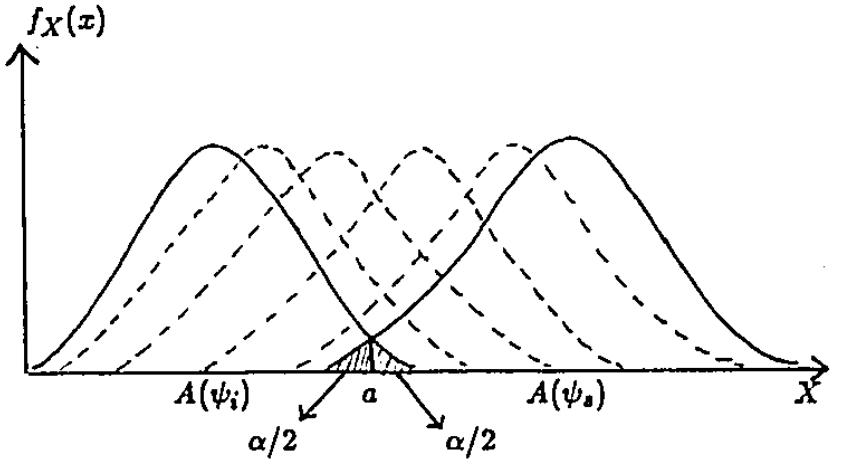
Estimación por Intervalo de ψ

Para obtener un intervalo de confianza al $(1 - \alpha)$ 100% para ψ se tienen que resolver las siguientes ecuaciones:

$$\alpha/2 = \sum_{j \leq a} P[X = j \mid n_1, n_0, m_1, m_0; \psi_s] \quad \text{donde } \psi_s : \text{límite superior del intervalo}$$

$$\alpha/2 = \sum_{j \geq a} P[X = j \mid n_1, n_0, m_1, m_0; \psi_i] \quad \text{donde } \psi_i : \text{límite inferior del intervalo}$$

El porqué se obtienen así los límites puede verse más claramente con un dibujo. Suponiendo una distribución de probabilidad continua $f_X(x)$



Al igual que en el caso de la estimación puntual, si las frecuencias de la tabla son muy grandes, conviene, en términos prácticos utilizar los métodos aproximados en lugar de enfrentar un problema numérico serio, para resolver los sistemas de ecuaciones.

Pruebas de Hipótesis sobre ψ

Para probar hipótesis sobre ψ, lo que se hace es obtener la probabilidad, suponiendo cierta la hipótesis nula Ho, de haber obtenido una tabla como la observada o aún más extrema (es decir hacia la derecha o izquierda de la distribución según sea la

hipótesis alternativa) y si esa probabilidad es menor o igual que el nivel de significancia deseado, se rechaza Ho.

Para probar $Ho : \psi = \psi_0$ vs $Ha : \psi < \psi_0$ rechace si $P_i = \sum_{i \leq a} P[X = i \mid n_1, n_0, m_1, m_0, \psi_0] \leq \alpha$

Para probar $Ho : \psi = \psi_0$ vs $Ha : \psi > \psi_0$ rechace si $P_s = \sum_{i \geq a} P[X = i \mid n_1, n_0, m_1, m_0; \psi_0] \leq \alpha$

Para probar $Ho : \psi = \psi_0$ vs $Ha : \psi \neq \psi_0$ rechace si $min\{P_i, P_s\} \leq \frac{\alpha}{2}$

Cuando la hipótesis nula es $Ho : \psi = 1$, las probabilidades P_i y P_s se calculan a partir de la distribución hipergeométrica central. A la prueba que resulta se le conoce como la prueba exacta de Fisher (ver Everitt (1977), p.p. 15-20 y Conover (1980), p. 167).

En conclusión, los métodos exactos conviene utilizarlos cuando las frecuencias en las celdas de la tabla son pequeñas. Esto suele ocurrir cuando se utiliza en el diseño del estudio, aparejamiento individual o cuando se hace un análisis estratificado muy fino para el control de confusión, ya sea porque en el diseño se realizó un aparejamiento por grupo fino o porque simplemente se quiere controlar la confusión en la etapa de análisis.

ANALISIS APROXIMADO (ASINTOTICO) PARA UNA TABLA DE 2×2

Modelo

En la sección anterior se dedujo el modelo probabilístico para una tabla de 2×2, condicionando a que los totales marginales estén fijos y resultó una distribución hipergeométrica no central. Se mencionó también, que cuando las frecuencias en las celdas eran muy grandes la obtención del estimador y la realización de pruebas de hipótesis se vuelven materialmente impracticables por los problemas numéricos que se tienen que enfrentar. Una manera de resolver este problema es aproximar la distribución hipergeométrica no central mediante una normal, es decir:

Sea X: número de expuestos entre los casos, entonces

$$(X \mid n_1, n_0, m_1, m_0; \psi) \sim \text{Hipergeométrica}(\psi)$$

entonces ahora, se aproxima esta distribución diciendo que

$$(X \mid n_1, n_0, m_1, m_0; \psi) \approx N(A(\psi); V(X \mid n_0, n_1, m_0, m_1, \psi))$$

$A(\psi) \doteq E\{X \mid n_0, n_1, m_0, m_1, \psi\}$ es una media asintótica.[3].

Suponiendo una ψ hipotética, el parámetro $A(\psi) = A$ se obtiene al sustituir el valor observado de X, (a) en la tabla de 2×2 por A y obtener las frecuencias de las celdas restantes B, C, D de tal forma que se obtenga precisamente ψ, es decir, se obtiene A, tal que al obtener B, C, D en la tabla siguiente:

Factor de Riesgo

	Expuesto	No Expuesto	
Casos	A	$B = n_1 - A$	n_1
Controles	$C = m_1 - A$	$D = n_0 - m_1 + A$	n_0
	m_1	m_0	N

se tiene que $\frac{AD}{BC} = \psi$. La ecuación que liga a ψ con A resulta ser una ecuación cuadrática y sólamente una de sus raíces produce un valor posible para A puesto que $ABCD$ deben ser todos positivos.

La varianza de X, cuando se utiliza la aproximación a la normal es:

$$V(X \mid n_0, n_1, m_0, m_1, \psi) \doteq [\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}]^{-1}$$

Cuando $\psi = 1$, se tiene que la media asintótica que es igual a la exacta, es $\frac{m_1 n_1}{N}$, la cual es la media de una hipergeométrica central; sin embargo, la varianza asintótica es $\frac{n_1 n_0 m_1 m_0}{N^3}$, la cual es ligeramente menor que la varianza exacta, o sea la de una hipergeométrica central, que resulta ser $\frac{n_1 n_0 m_1 m_0}{N^2(N-1)}$.

Para que la aproximación a la normal sea buena, se pide generalmente que la $E[X \mid n_0, n_1, m_0, m_1; \psi] \geq 5$, con esto se logra que el valor máximo y el mínimo que toma X, se encuentren al menos a dos desviaciones estándar de su media, con lo cual, al aproximar esta distribución discreta con la normal, se garantiza que los niveles de significancia y de confianza comunmente utilizados en las pruebas de hipótesis e intervalos de confianza, en efecto se alcancen (ver Mantel, N. y Fleiss, J. (1980)).

En lo que sigue, se plantearán métodos de estimación y de pruebas de hipótesis suponiendo una distribución normal para X, con las características anteriormente descritas.

Estimación Puntual de ψ

Bajo alguna hipótesis concreta, como lo es $\psi = \psi_0$, el número $A_0 = A(\psi_0)$ debiera parecerse al número observado $X = a$ en la tabla de 2×2. En esto se basa

[3] Si se quiere profundizar sobre como se lleva a cabo esta aproximación a la normal, se puede consultar a Hannan J. y Harkness W, (1963) y Stevens, W.C., (1951)

la contrastación de hipótesis. En este razonamiento está implícito que para estimar la media asintótica $A(\psi)$ se utilice $\hat{A}(\psi) = a$ lo que determinará $\hat{\psi}$ con $A(\hat{\psi}) = a$. La ψ resultante es precisamente el estimador de máxima verosimilitud. Es decir, bajo la suposición de que $X \sim N(A(\psi); V(X \mid n_0, n_1, m_0, m_1, \psi))$ el estimador de máxima verosimilitud (asintótico) de ψ se obtiene, después de realizar el álgebra correspondiente, resolviendo

$$a = E[X \mid n_0, n_1, m_0, m_1, \psi] = A(\hat{\psi})$$

de donde $\hat{\psi}$ resulta ser $\frac{ad}{bc}$, o sea que la razón de momios empírica es el estimador de máxima verosimilitud asintótico. Esta razón de momios empírica es también el estimador de máxima verosimilitud incondicional de ψ bajo el modelo de producto de binomiales con dos parámetros (ver Breslow, N.E. y Day, N.E. (1980) pág. 131) y bajo el modelo Poisson, producto de multinomiales o multinomial (ver Fienberg S.E. (1978), pág. 18).

Estimación por intervalo de ψ

Límites de Cornfield.-

Cornfield (1956) basándose en la aproximación asintótica a la distribución de $X \mid n_1, n_0, m_1, m_0, \psi$, propone calcular los límites de confianza de la siguiente manera:

Obténgase ψ_i y ψ_s (límite inferior y superior al $(1-\alpha)$ 100%) resolviendo

$$a - A(\psi_i) - \frac{1}{2} = Z_{\alpha/2}\sqrt{V(X \mid n_0, n_1, m_0, m_1, \psi_i)}$$

$$a - A(\psi_s) + \frac{1}{2} = -Z_{\alpha/2}\sqrt{V(X \mid n_0, n_1, m_0, m_1, \psi_s)}$$

El factor de $\frac{1}{2}$ es una corrección por continuidad y $Z_{\alpha/2}$ es el $100(1-\alpha/2)$ porcentil de la distribución normal estándar

Estas ecuaciones son de cuarto grado y deben utilizarse métodos numéricos iterativos para resolverlas.

En virtud de que estos límites se basan en la aproximación a la normal, una manera "burda" de evaluar que tan bien se está aproximando el intervalo de confianza, es averiguar si los intervalos $A(\psi_i) \pm Z_{\alpha/2}\sqrt{V(X \mid n_0, n_1, m_0, m_1, \psi_i)}$ y $A(\psi_s) \pm Z_{\alpha/2}\sqrt{V(X \mid n_0, n_1, m_0, m_1, \psi_s)}$ se encuentran ambos contenidos dentro del rango de valores factibles de X $(Max\{0, m_1 - n_0\} \leq X \leq Min\{n_1, m_1\})$.

Límites Logísticos.-

Woolf, B. (1955) y Gart, J. (1962) sugirieron que en lugar de calcular directamente los límites de confianza para ψ, se obtuvieran los límites de confianza para el $\ln \psi$ que son mucho más fáciles de obtener.

Haldane, J. (1955) y Anscombe, F. (1956) propusieron como estimador del $\ln \psi$

$$\ln \hat{\psi} = \ln\{\frac{(a+\frac{1}{2})(d+\frac{1}{2})}{(b+\frac{1}{2})(c+\frac{1}{2})}\}$$

es decir, incluyeron un factor de corrección por continuidad de $\frac{1}{2}$, y demostraron que $E[\ln \hat{\psi}] \sim \ln \psi$ y Gart, J. y Zweifel, J. (1967) encontraron que un buen estimador de su varianza es:

$$\hat{V}(\ln \hat{\psi}) = \frac{1}{a+\frac{1}{2}} + \frac{1}{b+\frac{1}{2}} + \frac{1}{c+\frac{1}{2}} + \frac{1}{d+\frac{1}{2}}$$

Entonces los límites de confianza para el $\ln \psi$ se aproximan de la siguiente manera [4]:

$$P\{\underbrace{(\ln \hat{\psi} - Z_{\alpha/2}\sqrt{\hat{V}(\ln \hat{\psi})}}_{\ln \psi_i} \leq \ln \psi \leq \underbrace{\ln \hat{\psi} + Z_{\alpha/2}\sqrt{\hat{V}(\ln \hat{\psi})}\}}_{\ln \psi_s} = 1 - \alpha$$

y tomando antilogaritmos se obtienen los límites de confianza para ψ:

$$\psi_i = e^{\ln \hat{\psi} - Z_{\alpha/2}\sqrt{\hat{V}(\ln \hat{\psi})}}$$

$$\psi_s = e^{\ln \hat{\psi} + Z_{\alpha/2}\sqrt{\hat{V}(\ln \hat{\psi})}}$$

Gart, J. y Thomas, D. (1972) evaluaron varios métodos para aproximar los intervalos de confianza para ψ y encontraron que los límites de Cornfield proporcionan una de las mejores aproximaciones a los límites exactos y que los límites logísticos en general son más angostos que otros aunque, como ya se hizo notar, más sencillos de obtenerse.

Límites de Confianza basados en el Valor de una Estadística de Prueba.-

Miettinen (1976) propuso que la $Var(\ln \hat{\psi})$ se estimara con un método "suigéneris" basado en el valor observado de una estadística de prueba.

---

[4] Woolf propuso estos mismos límites pero sin considerar el factor de corrección por continuidad de $\frac{1}{2}$, en el estimador del $\ln \psi$

Bajo el supuesto de que no hay asociación entre el factor de riesgo y la enfermedad (esto es $\psi = 1$) las dos cantidades aleatorias siguientes tienen aproximadamente la misma distribución $\frac{(\ln\hat{\psi}-\ln 1)^2}{Var(\ln\hat{\psi})}$ y $\frac{(|ad-bc|)^2(N-1)}{n_0n_1m_0m_1}$, llámese X^2 a la segunda cantidad.

Entonces al igualarlas, la varianza del $\ln\hat{\psi}$ resulta ser estimada de la siguiente forma:

$$\hat{V}(\ln\hat{\psi}) = \frac{\ln^2(\hat{\psi})}{X^2}$$

Este método desde luego presupone que ambas cantidades están fuertemente relacionadas.

Por lo tanto, los límites de confianza para el $\ln\psi$ quedarían expresados como:

$$\ln\hat{\psi} \pm Z_{\alpha/2}\frac{\ln\hat{\psi}}{X}$$

Tomando antilogaritmos se obtienen los límites de confianza para ψ, es decir,

$$\hat{\psi}^{(1\pm\frac{Z\alpha/2}{X})}$$

Cuando $\hat{\psi} = 1$ entonces $X^2 = 0$ y los límites de confianza así calculados no están definidos.

Debe observarse que el estimador de la $V(\ln\hat{\psi})$ basado en el valor de la estadística de prueba X^2 sólo es estrictamente válido cuando $\psi = 1$. Cuando los tamaños de muestra de casos y controles son iguales, la varianza para otros valores de ψ se encuentra sistemáticamente subestimada (Ver Breslow, N. y Day, N., (1980), pág. 135).

A pesar de esto, los límites de confianza obtenidos con esta estimación de la varianza se utilizan mucho en la práctica, esencialmente porque son muy simples y suelen producir resultados satisfactorios cuando $\hat{\psi}$ no es muy extremo.

Tanto los límites logísticos como los basados en el valor de la estadística X^2, pueden utilizarse como valores iniciales para los procedimientos iterativos mediante los cuales se obtienen límites de confianza más exactos.

Pruebas de Hipótesis sobre ψ

Con base en la aproximación a la normal, de la distribución condicional de X, se obtienen aproximaciones a las probabilidades bajo la hipótesis nula Ho, de tener una tabla como la observada o aún más extrema, según sea la hipótesis alternativa.

Para probar $Ho : \psi = \psi_0$ *vs* $Ha : \psi > \psi_0$, rechace Ho si

$$P_s \approx 1 - \Phi\left(\frac{a - A(\psi_0) - \frac{1}{2}}{\sqrt{V(X \mid n_0, n_1, m_0, m_1, \psi_0)}}\right) \leq \alpha$$

Para probar $Ho : \psi = \psi_0 \quad vs \quad Ha : \psi < \psi_0$, rechace Ho si

$$P_i \approx \Phi\left(\frac{a - A(\psi_0) + \frac{1}{2}}{\sqrt{V(X \mid n_0, n_1, m_0, m_1, \psi_0)}}\right) \leq \alpha$$

Donde Φ es la función acumulativa de la distribución normal estándar.

Para probar $Ho : \psi = \psi_0 \quad vs \quad Ha : \psi \neq \psi_0$, rechace Ho si

$$min\{P_i, P_s\} \leq \frac{\alpha}{2}.$$

Cuando la hipótesis que se prueba es $Ho : \psi = 1 \quad vs \quad Ha : \psi \neq 1$, se puede utilizar para probar la hipótesis nula a:

$$X^2 = \frac{(\mid a - A(1) \mid -\frac{1}{2})^2}{V(X \mid n_0, n_1, m_0, m_1, \psi = 1)} = \frac{(\mid a - \frac{m_1 n_1}{N} \mid -\frac{1}{2})^2}{\frac{n_1 n_0 m_1 m_0}{N^2(N-1)}}$$
$$= \frac{(\mid ad - cb \mid -\frac{1}{2}N)^2(N-1)}{n_0 n_1 m_0 m_1}$$

X^2 es el valor observado de una variable aleatoria con distribución Ji cuadrada con un grado de libertad, bajo $Ho : \psi = 1$.

III.2.2 Factor de Riesgo Dicotómico con Control de un Factor de Confusión: Análisis Estratificado de Tablas 2 × 2

Se considera nuevamente un estudio con una muestra aleatoria de n_1 casos y una muestra aleatoria de n_0 controles, obtenidas en forma independiente. Se tiene un sólo factor de riesgo o tratamiento dicotómico. La diferencia con el caso anterior es que se tiene un factor de confusión politómico y se le quiere controlar en la etapa de análisis, para evitar sesgos en el estimador de la razón de momios ψ.

Uno de los métodos más utilizados para el control de factores de confusión en la etapa de análisis, consiste en agrupar la muestra en conjuntos que son internamente homogéneos respecto al factor de confusión y entonces obtener para cada grupo el cociente de momios, el cual estará libre del sesgo que introduce el factor de confusión puesto que éste es homogéneo dentro de cada grupo. A los grupos así formados se les llama estratos y al análisis que se realiza se le llama análisis estratificado.

Una vez estratificada la muestra y obtenidos los cocientes de momios por estrato, el análisis consiste en:

1o. Determinar si la asociación entre la exposición y la enfermedad (medida a partir de la razón de momios) es razonablemente constante de estrato a estrato. Esto es, que debe hacerse una prueba de hipótesis de no interacción, es decir, de homogeneidad de la razón de momios entre estratos. [5].

2o. Si no se rechaza la hipótesis de homogeneidad de la razón de momios, entonces se procede a obtener un estimador global de ψ (que combine la información de los estratos) y a realizar pruebas de hipótesis sobre ψ, en particular que $\psi = 1$, lo que significa que no hay efecto del factor de riesgo o tratamiento en el factor respuesta.

3o. Si se rechaza la hipótesis de homogeneidad, no tiene sentido el tratar de obtener un estimador global de ψ, ni el realizar pruebas de hipótesis sobre un ψ general puesto que éste no existe. En este caso debe describirse como varía la razón de momios de acuerdo a los cambios en los niveles del factor de confusión (o de cualquier otro factor que interactúe y sea de interés para el estudio).

Modelo

Como se tiene un factor de riesgo dicotómico, un factor respuesta dicotómico y un factor de confusión politómico (supóngase que con I niveles), la información puede resumirse en I tablas 2×2, según se indica a continuación

Estrato i

Factor de Riesgo

	Expuesto	No Expuesto	
Casos	a_i	b_i	n_{1i}
Controles	c_i	d_i	n_{0i}
	m_{1i}	m_{0i}	N_i

donde:

$$i = 1, 2, 3, \ldots I$$

$n_1 = \sum_{i=1}^{I} n_{1i}$: tamaño de muestra de casos.

$n_0 = \sum_{i=1}^{I} n_{0i}$: tamaño de muestra de controles.

Como las muestras entre estratos son independientes, si se condiciona a que los marginales $n_{1i}, n_{0i}, m_{1i}, m_{0i}$ están fijos, la distribución de probabilidad conjunta, es decir, la distribución de probabilidad para los datos observados en todas las I tablas, es el producto de I hipergeométricas no centrales con parámetros de no centralidad ψ_i.

---

[5] Las pruebas que para este efecto se describan, obviamente sirven para probar interacción de cualquier factor (no necesariamente uno que se considere de confusión) con el factor de riesgo y el factor respuesta.

Aunque para llevar a cabo un análisis estratificado lo primero que debe hacerse es una prueba de homogeneidad de las razones de momios, por facilidad en la presentación, se describirán primero los métodos para estimar ψ globalmente y realizar pruebas de hipótesis de no asociación, suponiendo que no hay interacción y después se presentan los métodos para realizar pruebas de hipótesis de no interacción.

Estimación Puntual de la Razón de Momios Común ψ

Estimador de Máxima Verosimilitud.-

Después de realizar los procedimientos para maximizar la función de verosimilitud, resulta que el estimador de la razón de momios común se encuentra resolviendo la siguiente ecuación:

$$\sum_{i=1}^{I} a_i = \sum_{i=1}^{I} E(X_i \mid n_{1i}, n_{0i}, m_{1i}, m_{0i}; \psi)$$

Si se quiere obtener el estimador de máxima verosimilitud condicional (utilizando la distribución de probabilidad condicional exacta para X_i) la $E[X_i \mid n_{1i}, n_{0i}, m_{1i}, m_{0i}; \psi]$, es la de una hipergeométrica no central con parámetro de no centralidad ψ . Esto define ecuaciones polinomiales de alto grado que resultan en un trabajo de cómputo pesado.

Si las frecuencias en las celdas de las tablas de todos los estratos son grandes, se puede utilizar como modelo probabilístico aproximado para la distribución de $(X_i \mid n_{1i}, n_{0i}, m_{1i}, m_{0i}, \psi)$ a la distribución normal, en cuyo caso la $E[X_i \mid n_{1i}, n_{0i}, m_{1i}, m_{0i}; \psi] = A_i(\psi)$. Entonces el procedimiento de estimación requiere de encontrar las frecuencias ajustadas para todas las celdas, de tal forma que el total de casos expuestos observados sea igual al total de casos expuestos ajustados. Se requieren procedimientos iterativos para resolver este problema de estimación.

Cuando hay muchos estratos con pocas observaciones por estrato, éste último estimador de máxima verosimilitud está sesgado (toma valores más alejados de la unidad que la verdadera razón de momios), puesto que la aproximación a la normal es mala.

En general no se utilizan los estimadores de máxima verosimilitud, puesto que el condicional exacto (con la hipergeométrica no central) es muy laborioso de obtener y el condicional con la aproximación normal suele estar sesgado ya que son múltiples las ocasiones en las que los estratos contienen pocas observaciones.

Estimador Logístico.-

Este estimador fue propuesto por Woolf (1955) y consiste en obtener una

combinación lineal ponderada de los estimadores logísticos de las razones de momios por estrato, es decir,

$$\ln \hat{\psi}_l = \sum_{i=1}^{I} \frac{w_i}{\sum_{i=1}^{I} w_i} (\ln \hat{\psi}_i) = \sum_{i=1}^{I} \frac{w_i}{\sum_{i=1}^{I} w_i} \left(\ln \left(\frac{a_i d_i}{b_i c_i} \right) \right)$$

donde $w_i = \frac{1}{\hat{V}(\ln \hat{\psi}_i)} = \{\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}\}^{-1}$

Las ponderaciones son aquellas que minimizan la $V(\ln \hat{\psi}_l)$ y tales que sumen uno.

La varianza estimada del estimador es:

$$\hat{V}(\ln \hat{\psi}_l) = (\sum_{i=1}^{I} w_i)^{-1}$$

El estimador logístico se comporta bien cuando las frecuencias en todas las celdas, de todos los estratos son grandes. Nótese que si en algún estrato, cualquiera de las celdas de la tabla es cero, entonces no están definidos el ln de la razón de momios de ese estrato ni su ponderación correspondiente. Un remedio usual en este caso es sumar un factor de corrección por continuidad de $\frac{1}{2}$ (Ver Gart y Zweifel (1967)) a todas las entradas de la tabla. Sin embargo, si son muchos los estratos que presentan este problema, no se recomienda el uso de este estimador pues presenta sesgos inaceptables.

Estimador de Mantel-Haenszel.-

Mantel y Haenszel (1959) propusieron como estimador de la razón de momios global, una media ponderada de los estimadores de la razón de momios por estrato, es decir:

$$\hat{\psi}_{m-h} = \sum_{i=1}^{I} \frac{\left(\frac{b_i c_i}{N_i}\right)}{\sum_{i=1}^{I} \left(\frac{b_i c_i}{N_i}\right)} \left(\frac{a_i d_i}{b_i c_i}\right) = \frac{\sum_{i=1}^{I} \frac{a_i d_i}{N_i}}{\sum_{i=1}^{I} \frac{b_i c_i}{N_i}}$$

En su artículo, Mantel y Haenszel afirman que dichas ponderaciones aproximan a las varianzas inversas de los estimadores de las razones de momios por estrato (ajustadas para que sumen a la unidad) y que también proporcionan una ponderación razonable por la importancia del estrato.

Este estimador no se ve afectado por la presencia de celdas con cero observaciones y es un estimador consistente, aún cuando se tenga un gran número de estratos con pocas observaciones. Cuando las frecuencias en las celdas son grandes

los resultados que se obtienen con este estimador son parecidos a los que se obtienen por máxima verosimilitud. Su único defecto es que no existe un estimador robusto de la varianza que lo acompañe (Ver Breslow y Day (1982)).

Estimación por Intervalo de la Razón de Momios Común ψ

A continuación se describen únicamente los métodos asintóticos. No se discuten los métodos exactos ya que son extremadamente difíciles de calcular en la mayoría de los casos.

Extensión de los Límites de Cornfield.-

Gart (1970) propuso una extensión de los límites de Cornfield para el caso en el que se tienen una serie de tablas de 2×2, basándose en la distribución asintótica de la $\sum_{j=1}^{I} (X_j \mid n_{0j}, n_{1j}, m_{0j}, m_{1j}; \psi)$.

Como la distribución condicional asintótica de X_j es una normal con media $A_j(\psi)$ y varianza $V(X_j \mid n_{0j}, n_{1j}, m_{0j}, m_{1j}; \psi)$, según se discutió anteriormente, entonces la distribución asintótica de la $\sum_{j=1}^{I} (X_j \mid n_{0j}, n_{1j}, m_{0j}, m_{1j}, \psi)$ es también normal con media $\sum_{j=1}^{I} A_j(\psi)$ y varianza $\sum_{j=1}^{I} V(X_j \mid n_{0j}, n_{1j}, m_{0j}, m_{1j}; \psi)$.

Para encontrar los límites de confianza al $(1-\alpha)$ 100% de la razón de momios común ψ, Gart propone obtener ψ_i y ψ_s (límites inferior y superior al $(1-\alpha)$ 100%) resolviendo:

$$\frac{\sum_{j=1}^{I} a_j - \sum_{j=1}^{I} A_j(\psi_i) - \frac{1}{2}}{\sqrt{\sum_{i=1}^{I} V(X_j \mid n_{0j}, n_{1j}, m_{0j}, m_{1j}, \psi_i)}} = Z_{\frac{\alpha}{2}}$$

$$\frac{\sum_{j=1}^{I} a_i - \sum_{j=1}^{I} A_j(\psi_s) + \frac{1}{2}}{\sqrt{\sum_{i=1}^{I} V(X_j \mid n_{0j}, n_{1j}, m_{0j}, m_{1j}, \psi_s)}} = -Z_{\frac{\alpha}{2}}$$

Para obtener cada límite es necesario resolver I ecuaciones cuadráticas mediante procedimientos iterativos.

Cuando el número de estratos es grande y las observaciones por estrato son

pocas, se pueden utilizar las medias y las varianzas exactas en las ecuaciones anteriores (o sea las de la hipergeométrica no central en cada tabla).

Límites Logísticos.-

En base al estimador puntual logístico, $\ln \hat{\psi}_l$, se construyen los límites de confianza logísticos al $(1-\alpha)$ 100% de la siguiente manera:

$$\ln \psi_s = \ln \hat{\psi}_l + \frac{Z_{\alpha/2}}{\sqrt{\sum_{j=1}^{I} w_j}} \quad : \text{límite superior}$$

$$\ln \psi_i = \ln \hat{\psi}_l - \frac{Z_{\alpha/2}}{\sqrt{\sum_{j=1}^{I} w_j}} \quad : \text{límite inferior}$$

donde

$$\ln \hat{\psi}_l = \sum_{j=1}^{I} \frac{w_j}{\sum_{j=1}^{I} w_j} \ln \frac{a_j d_j}{b_j c_j}$$

$$w_j = \{\frac{1}{a_j} + \frac{1}{b_j} + \frac{1}{c_j} + \frac{1}{d_j}\}^{-1}$$

$$\hat{V}(\ln \hat{\psi}_l) = \left(\sum_{j=1}^{I} w_j\right)^{-1}$$

Tomando antilogaritmo se obtienen los límites de confianza para la razón de momios común ψ:

$$\psi_i = exp(\ln \hat{\psi}_l - \frac{Z_{\alpha/2}}{\sqrt{\sum_{j=1}^{I} w_j}})$$

$$\psi_s = exp(\ln \hat{\psi}_l + \frac{Z_{\alpha/2}}{\sqrt{\sum_{j=1}^{I} w_j}})$$

Como se mencionó anteriormente al hablar del estimador logístico puntual, se presentan problemas al calcular intervalos de confianza mediante este método, cuando se tienen muchos estratos con pocas observaciones en ellos.

Límites de Confianza basados en el Valor de una Estadística de Prueba.-

Al igual que para el caso de una sóla tabla de 2×2, Miettinen (1976) propone estimar la varianza del $\ln \hat{\psi}$ (donde $\hat{\psi}$ puede ser cualquier estimador puntual de la razón de momios global ψ) mediante el valor de una estadística de prueba [6] para contrastar una hipótesis nula de no asociación ($Ho : \psi = 1$), de la siguiente manera:

$$\hat{V}ar(\ln \hat{\psi}) = \frac{\ln^2(\hat{\psi})}{X^2}$$

donde

$$X^2 = \frac{\left(\sum_{j=1}^{I} a_j - \sum_{j=1}^{I} A_j(1)\right)^2}{\sum_{j=1}^{I} Var(X_j \mid n_{0j}, n_{1j}, m_{0j}, m_{1j}; \psi = 1)}$$

Los límites de confianza para el logaritmo natural de la razón de momios común ψ se obtienen de la siguiente manera:

$$\ln \hat{\psi} \pm Z_{\alpha/2} \frac{\ln \hat{\psi}}{X}$$

y tomando antilogaritmos se obtienen los límites de confianza al $100\ (1 - \alpha)\%$ para ψ:

$$\hat{\psi}^{\left(1 \pm \frac{Z_{\alpha/2}}{X}\right)}$$

Breslow y Day (1982) proponen utilizar como estimador puntual de la razón de momios global, al estimador de Mantel y Haenszel cuando se emplee este procedimiento para obtener límites de confianza.

Como se mencionó para el caso de una tabla 2×2, el calcular los límites de confianza como Miettinen propone, tiene problemas cuando la $\hat{\psi}$ está muy alejada de la unidad; además del comentario ya hecho, de que ambas estadísticas de prueba deben estar muy correlacionadas.

---

[6] Esta estadística de prueba se describirá posteriormente, cuando se explique como realizar una prueba de no asociación entre el factor de riesgo y la respuesta

Prueba de Hipótesis de no Asociación entre el Factor de Riesgo y el Factor Respuesta

Se describe un método aproximado para probar únicamente la hipótesis nula $Ho : \psi = 1$, es decir, la hipótesis de que no existe asociación entre el factor de riesgo o tratamiento y el factor respuesta, ya que esta hipótesis es la que con más frecuencia se pretende contrastar en la práctica.

La hipótesis a probar es $Ho : \psi_i = 1 \quad \forall \; i = 1, 2, ..I \quad$ vs $\quad Ha : \psi_i \neq 1$ para al menos una i, donde $i = 1, 2, ..I$.

La distribución asintótica de la $\sum_{i=1}^{I} (X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}, \psi_i)$ es una normal, es decir, $N\left(\sum_{i=1}^{I} A_i(\psi_i), \sum_{i=1}^{I} V(X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}; \psi_i)\right)$, entonces, suponiendo cierta la hipótesis nula de que $\psi_i = 1$ para toda i, se tiene que:

$$\frac{\sum_{i=1}^{I} (X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}, \psi_i = 1) - \sum_{i=1}^{I} A_i(1)}{\sqrt{\sum_{i=1}^{I} V(X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}, \psi_i = 1)}} \approx N(0,1)$$

donde $\quad A_i(1) = \frac{n_{1i} m_{1i}}{N_i}$

$$V(X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}; \psi_i = 1) = \frac{n_{0i} n_{1i} m_{0i} m_{1i}}{N_i^3}$$

Como estadística de prueba se utiliza:

$$X^2 = \frac{\left(\left| \sum_{i=1}^{I} (X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}; \psi_i = 1) - \sum_{i=1}^{I} A_i(1) \right| - \frac{1}{2} \right)^2}{\sum_{i=1}^{I} V(X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}; \psi_i = 1)}$$

la cual, suponiendo cierta $Ho : \psi_i = 1$ para toda i, tiene una distribución aproximada Ji cuadrada con un grado de libertad. Por lo tanto se rechaza Ho si

$$X^2 = \frac{\left(\left| \sum_{i=1}^{I} a_i - \sum_{i=1}^{I} A_i(1) \right| - \frac{1}{2} \right)^2}{\sum_{i=1}^{I} V(X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}; \psi = 1)} > \chi_1^2(\alpha)$$

donde $\quad \chi_1^2(\alpha)$ es el porcentil $(1-\alpha)100$, de una distribución Ji cuadrada con un grado de libertad.

Esta estadística de prueba fue propuesta por Cochran (1954) y posteriormente Mantel y Haenszel (1959) sugirieron que se utilizaran las varianzas exactas en vez de las asintóticas, en el denominador de la misma.

Mantel y Fleiss (1980) sugirieron que para que la aproximación a la normal funcione bien, se requiere que la media asintótica $\sum_{i=1}^{I} A_i(1)$ esté al menos 5 unidades separada del valor mínimo y máximo que puede tomar la $\sum_{i=1}^{I} (X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}, \psi_i)$, es decir, a cinco unidades de $\sum_{i=1}^{I} Max(0, m_{1i} - n_{0i})$ y de $\sum_{i=1}^{I} Min(m_{1i}, n_{1i})$. Con esto se garantiza que los niveles de significancia comunmente utilizados (de 0.05 ó más) para las pruebas de hipótesis, en efecto, se alcancen.

Prueba de Hipótesis de Homogeneidad de la Razón de Momios

Como se mencionó anteriormente el primer paso a realizar cuando se lleva a cabo un análisis estratificado es, investigar si el efecto que tiene el factor de riesgo o tratamiento en el factor respuesta, es razonablemente constante en cada uno de los niveles del factor de confusión. Esto se realiza mediante lo que se conoce como pruebas de homogeneidad o no interacción y permite determinar si existe un sólo efecto general, en cuyo caso se obtiene un estimador de la razón de momios global que lo resuma. Si el efecto que el factor de riesgo tiene en el factor respuesta, se ve modificado por la presencia del factor de confusión, deberá describirse mediante algún modelo, cómo es que se da esta interacción.

A continuación se presenta un método para probar homogeneidad de la razón de momios, que sirve para probar la hipótesis nula de que la razón de momios es igual en todos los estratos, contra la alternativa de que al menos una razón de momios es diferente al resto.

Para probar $Ho : \psi_i = \psi \ \ \forall \ \ i = 1...I$ contra $Ha : \psi_i \neq \psi$ para al menos una i, donde $i = 1...I$, se utiliza la siguiente estadística de prueba:

$$\sum_{i=1}^{I} \frac{((X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}; \hat{\psi}) - A_i(\hat{\psi}))^2}{V(X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}; \hat{\psi})}$$

donde $\hat{\psi}$ es cualquier estimador globalizador de ψ y $A_i(\hat{\psi})$ es el valor esperado de $(X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}; \hat{\psi})$ suponiendo que $\hat{\psi}$ es la razón de momios.

Nótese que esta estadística surge de la suposición de que la distribución de X_i para cada estrato es aproximadamente normal, es decir, que

$$(X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}; \hat{\psi}) \approx N(A_i(\hat{\psi}), V(X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}; \hat{\psi}))$$

de donde

$$\frac{[(X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}; \hat{\psi}) - A_i(\hat{\psi})]^2}{V(X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}; \hat{\psi})} \approx \chi^2_{(1)}$$

y entonces la estadística propuesta se distribuye aproximadamente como una Ji cuadrada con $I - 1$ grados de libertad.

Para que las suposiciones distribucionales se cumplan, se requiere que las observaciones en cada estrato sean razonablemente grandes.

El suponer que $(X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}; \hat{\psi})$ tiene una distribución normal para cada i, es mucho más fuerte que suponer que la $\sum_{i=1}^{I} (X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}; \hat{\psi})$ tenga una distribución normal (como se hizo para probar la hipótesis de no asociación), pero en este caso, como lo que interesa es la homogeneidad de las razones de momios, se requiere evaluar las desviaciones individuales (por estrato) y no en promedio como se hizo para probar la hipótesis de no asociación.

Se rechazará entonces Ho a un nivel de significancia α, si

$$\sum_{i=1}^{I} \frac{(a_i - A_i(\hat{\psi}))^2}{V(X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}; \hat{\psi})} > \chi^2_{I-1}(\alpha)$$

Esta prueba es una prueba muy general, pues sirve para contrastar la hipótesis nula de igualdad de las razones de momios entre estratos, contra una alternativa muy general de que al menos una razón de momios difiere de las demás, es decir contra una alternativa que especifíca que hay diferencias, pero no toma en cuenta si estas diferencias son originadas por algún patrón específico de comportamiento de la razón de momios a través de estratos. Cuando el factor de confusión es una variable continua o al menos medido en una escala ordinal, puede haber interés en contrastar la hipótesis nula de igualdad de razón de momios entre estratos contra una hipótesis alternativa de que la razón de momios se incrementa o decrementa sistemáticamente cuando se pasa de un nivel a otro del factor de confusión. Existen pruebas más potentes para detectar esta alternativa, las cuales se verán en el contexto de modelos logísticos, en el Capítulo IV de este trabajo.

En ocasiones las razones de momios difieren entre estratos pero se sospecha que al dividir los I estratos en H grupos de tamaño $I_1, I_2, I_3 ... I_H$ donde $I_1 + I_2 + I_3 + ... + I_H = I$, se tiene que las razones de momios dentro de grupos son homogéneas no siendo así entre grupos. En este caso para probar homogeneidad global contra la alternativa de interés, conviene utilizar la siguiente estadística de prueba, ya que se tiene mayor potencia para detectar dicha alternativa (Breslow y Day (1980), p. 143):

$$K = \sum_{h=1}^{H} \frac{\{\sum_{i \epsilon I_h} ((X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}; \hat{\psi}) - A_i(\hat{\psi}))\}^2}{\sum_{i \epsilon I_h} V(X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}; \hat{\psi})}$$

Bajo $Ho : \psi_1 = \psi_2 = ... = \psi_H = \psi$ la estadística K tiene un distribución Ji cuadrada con $H - 1$ grados de libertad.

Se rechaza Ho a favor de $Ha : \psi_h \neq \psi$ para al menos una h, donde $h = 1, 2, ...H$, con un nivel de significancia α, si se cumple que:

$$\sum_{h=1}^{H} \frac{\{\sum_{i \epsilon I_h} (a_i - A_i(\hat{\psi}))\}^2}{\sum_{i \epsilon I_h} V(X_i \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}; \hat{\psi})} > \chi^2_{H-1}(\alpha)$$

III.2.3 Factor de Riesgo Politómico, Sin Control de Factor de Confusión: Análisis de Tablas $2 \times K$

Se considera un estudio con una muestra aleatoria de n_1 casos y una muestra aleatoria de n_0 controles obtenidas en forma independiente. Se tiene un sólo factor de riesgo o tratamiento politómico con K niveles, $K > 2$.

La información se resume en una tabla cruzada de $2 \times K$, de la siguiente manera:

Factor de Riesgo

	1	2	3	.....	K	
Casos	a_1	a_2	a_3		a_K	n_1
Controles	c_1	c_2	c_3		c_K	n_0
	m_1	m_2	m_3		m_K	N

Comunmente el análisis que se realiza en esta situación consiste en elegir a uno de los niveles del factor de riesgo como nivel base contra el cual se comparan cada uno de los otros niveles, aplicando los métodos de análisis para tablas de 2×2. Se obtienen estimadores de K razones de momios $\psi_1 = 1, \psi_2, ..., \psi_K$ y se analizan por separado, es decir, se obtienen intervalos de confianza para cada ψ_i y pruebas de hipótesis de que dichas razones de momios son individualmente iguales a la unidad o a otro valor.

Sin embargo, en múltiples ocasiones lo que interesa es determinar si no hay efecto del factor de riesgo o tratamiento en la enfermedad pero globalmente, o lo que es lo mismo probar que las K razones de momios son simultáneamente iguales a la unidad. En este caso, si el factor de riesgo no presenta ningún orden natural entre sus niveles se realiza una prueba similar a la usual para probar homogeneidad de K proporciones. Sin embargo, si los niveles del factor de riesgo presentan algún orden natural (cuando es una variable continua o medida al menos en escala ordinal) conviene utilizar una prueba más sensitiva para detectar si existe alguna tendencia en la razón de momios al incrementar o decrementar los niveles del factor de riesgo.

Prueba de No Asociación entre el Factor de Riesgo o Tratamiento y el Factor Respuesta

Esta prueba sirve para contrastar la hipótesis nula $Ho : \psi_2 = ... = \psi_K = 1$ contra la hipótesis alternativa $Ha : \psi_i \neq 1$ para al menos una i, donde $i = 2...K$.

Suponiendo cierta Ho y condicional a que $n_1, n_0, m_1, m_2....m_K$ están fijos, la distribución de los datos observados, es una hipergeométrica $(K-1)$ dimensional con medias varianzas y covarianzas dadas por (Breslow y Day (1980), p. 147):

$$e_i = E[X_i \mid n_0, n_1, m_0, m_1; \psi = 1] = \frac{m_i n_1}{N}$$

$$Var(X_i \mid n_0, n_1, m_0, m_1; \psi = 1) = \frac{m_i(N - m_i)n_1 n_0}{N^2(N-1)}$$

$$Cov(X_i, X_j \mid n_0, n_1, m_0, m_1, \psi = 1) = -\frac{m_i m_j n_1 n_0}{N^2(N-1)} \quad \text{para} \quad i \neq j$$

La estadística usual para probar igualdad de K proporciones es:

$$\begin{aligned} X^2 &= \sum_{i=1}^{2K} \frac{[(X_i \mid n_0, n_1, m_0, m_1; \psi) - e_i]^2}{e_i} \\ &= \sum_{i=1}^{K} \frac{\{[(X_i \mid n_0, n_1, m_0, m_1; \psi) - e_i]^2}{e_i} \\ &+ \frac{[(m_i - (X_i \mid n_0, n_1, m_0, m_1, \psi) - (m_i - e_i)]^2}{m_i - e_i}\} \\ &= \sum_{i=1}^{K} [(X_i \mid n_0, n_1, m_0, m_1; \psi) - e_i]^2 \{\frac{1}{e_i} + \frac{1}{m_i - e_i}\} \end{aligned}$$

Bajo Ho y condicional a que los marginales están fijos, X^2 se distribuye aproximadamente como una Ji cuadrada con $K-1$ grados de libertad.

Se rechaza entonces Ho, a un nivel de significancia α, si se cumple que

$$\sum_{i=1}^{K} (a_i - e_i)^2 \{\frac{1}{e_i} + \frac{1}{m_i - e_i}\} > \chi^2_{K-1}(\alpha)$$

III.2.4 Factor de Riesgo Politómico con Control de un Factor de Confusión: Análisis Estratificado de Tablas $2 \times K$

Se considera nuevamente un estudio con una muestra aleatoria de n_1 casos y una muestra aleatoria de n_0 controles, obtenidas en forma independiente. Se tiene un sólo factor de riesgo o tratamiento politómico con K niveles, $K > 2$. La diferencia con el caso anterior es que se tiene un factor de confusión politómico con I niveles y se le quiere controlar en la etapa de análisis para evitar sesgos en los estimadores de las razones de momios.

En este caso se lleva a cabo un análisis estratificado. La información se resume en I tablas cruzadas de $2 \times K$ (una tabla de $2 \times K$ para cada estrato, definidos éstos por los niveles del factor de confusión) de la siguiente manera:

Estrato i

Factor de Riesgo

	1	2	3	.....	K	
Casos	a_{1i}	a_{2i}	a_{3i}		a_{Ki}	n_{1i}
Controles	c_{1i}	c_{2i}	c_{3i}		c_{Ki}	n_{0i}
	m_{1i}	m_{2i}	m_{3i}		m_{Ki}	N_i

Como se mencionó anteriormente, el análisis usual cuando se trabaja con un factor de riesgo politómico, consiste en elegir a uno de los niveles del factor de riesgo como nivel base contra el cual se comparan cada uno de los otros niveles, pero aplicando en este caso, los métodos de análisis estratificado para tablas de 2×2, porque se quiere controlar por la presencia de un factor de confusión.

Se tienen entonces K razones de momios, relativas a un nivel base de comparación previamente elegido, $\psi_1 = 1, \psi_2, ..., \psi_K$ cuyo comportamiento se quiere explorar. Lo primero que hay que hacer para cada una de ellas es una prueba de hipótesis de homogeneidad entre estratos. Si se rechaza dicha hipótesis, se deberá proponer algún modelo que describa el comportamiento de dicha razón de momios a través de estratos. Si no se rechaza, se procede a obtener un estimador ajustado de la razón de momios global, se proporcionan sus límites de confianza y se prueba la significancia de su alejamiento de la unidad.

Debe notarse que en este caso los estimadores de las razones de momios presentan una peculiaridad:

Si $\hat{\psi}_4$ es el estimador resumen de la razón de momios que compara el nivel cuatro del factor de riesgo con el nivel uno (el base) y $\hat{\psi}_3$ es el estimador resumen de la razón de momios que compara el nivel tres del factor de riesgo con el nivel uno (el base), entonces su cociente $\frac{\hat{\psi}_4}{\hat{\psi}_3}$ no es idéntico algebráicamente al estimador resumen $\hat{\psi}_{43}$ de la razón de momios que compara el nivel cuatro con el nivel tres (o sea, tomando como nivel base el 3).

Esta peculiaridad no ocurre en el caso en el que se tiene un sóla tabla de $2 \times K$ puesto que:

$$\frac{\hat{\psi}_4}{\hat{\psi}_3} = \frac{\frac{a_4 c_1}{a_1 c_4}}{\frac{a_3 c_1}{a_1 c_3}} = \frac{a_4 c_3}{c_4 a_3} = \hat{\psi}_{43}$$

Tampoco ocurre este problema en el caso en el que, aunque se tenga una serie de tablas $2 \times K$, las razones de momios que comparan cada par de niveles, son las mismas de estrato a estrato.

Esta inconsistencia puede verse entonces como una manifestación de un problema de interacción.

Si se quieren obtener estimadores ajustados de las razones de momios que sí muestren esta consistencia, se debe recurrir a los métodos de regresión logística, con los cuales también pueden realizarse pruebas generales de interacción.

Como ya se había mencionado para el caso de una tabla de $2 \times K$, en múltiples ocasiones lo que interesa es determinar si no hay un efecto global del factor de riesgo o tratamiento en el factor respuesta. Si el factor de riesgo no presenta ningún orden natural entre sus niveles, se realiza una prueba que es una extensión de la que se presentó en la sección anterior y sirve para contrastar la hipótesis nula $Ho : \psi_2 = ... = \psi_K = 1$ contra la hipótesis alternativa $Ha : \psi_j \neq 1$ para al menos una j, donde $j = 2, ..., K$.

La estadística de prueba que se utiliza es (Breslow y Day (1980), p. 149):

$$(\underline{X}_\cdot - \underline{e}_\cdot)^T \quad V_\cdot^{-1} (\underline{X}_\cdot - \underline{e}_\cdot)$$

donde

$$\underline{X}_\cdot = \sum_{i=1}^{I} \underline{X}_i, \quad \underline{X}_i = \begin{pmatrix} (X_{2i} \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}; \psi) \\ \vdots \\ (X_{K,i} \mid n_{0i}, n_{1i}, m_{0i}, m_{1i}; \psi) \end{pmatrix}$$

$$\underline{e}_\cdot = \sum_{i=1}^{I} \underline{e}_i, \quad \underline{e}_i = E[\underline{X}_i]$$

$$V_\cdot = \sum_{i=1}^{I} V_i, \quad V_i \text{ es la matriz de varianzas y covarianzas de } \underline{X}_i \text{ correspondiente a la hipergeométrica multivariada}(\psi_i = 1)$$

Suponiendo cierta la hipótesis nula y condicional a que los marginales están fijos, esta estadística se distribuye aproximadamente como una Ji cuadrada con $K - 1$ grados de libertad.

Se rechaza entonces Ho a un nivel α de significancia, si se cumple que

$$(\underline{a}. - \underline{e}.)^T V_.^{-1} (\underline{a}. - \underline{e}.) > \chi^2_{K-1}(\alpha)$$

donde

$$\underline{a}. = \sum_{i=1}^{I} \underline{a}_i, \quad \underline{a}_i = \begin{pmatrix} a_{2i} \\ \cdot \\ \cdot \\ \cdot \\ a_{K,i} \end{pmatrix}.$$

Si los niveles del factor de riesgo presentan algún orden natural, conviene utilizar una prueba que utilice ese hecho y por ello será más sensitiva para detectar si existe alguna tendencia en la razón de momios al incrementar o decrementar los niveles del factor de riesgo.

III.2.5 Análisis cuando se tienen Varios Factores de Riesgo.-

Con frecuencia interesa estudiar los efectos conjuntos de varios factores de riesgo. Si se define cada nivel de exposición como una combinación particular de los niveles de los factores de riesgo de interés, se pueden utilizar los métodos descritos para estudiar dichos efectos conjuntos en el riesgo de enfermedad.

Si interesa obtener un estimador global de la razón de momios de un determinado factor de riesgo pero ajustado por la presencia de otros factores de riesgo, se pueden incluir a estos últimos como variables para estratificación y realizar el análisis correspondiente.

Aunque el método de estratificación puede utilizarse para analizar los efectos conjuntos de dos o más factores de riesgo, no es el mejor método ya que por un lado requiere de muchos cálculos y por el otro, al dividir las observaciones en muchos estratos se puede terminar con una serie de tablas con muchas celdas vacías.

Es más apropiado en este caso recurrir a los métodos de regresión logística para efectuar los análisis.

III.3 METODOS CLASICOS PARA EL ANALISIS CUANDO SE UTILIZA APAREJAMIENTO INDIVIDUAL

En esta sección se presentan los métodos que tradicionalmente se han utilizado para el análisis de estudios de casos y controles cuando en la etapa de diseño del estudio, se realiza aparejamiento individual como herramienta para el control de factores de confusión.

Del mismo modo que en la sección anterior, los métodos que se describen ahora se refieren a como realizar inferencias acerca de la razón de momios.

En primer lugar se describen los métodos de análisis para un estudio de casos y controles en el que el factor de riesgo es dicotómico y se realiza un aparejamiento individual uno a uno (un control para cada caso) y posteriormente se describen los métodos respectivos, cuando se tiene un factor de riesgo dicotómico pero se aparejan el mismo número de controles por cada caso individual. [8]

No se presentan los análisis clásicos para situaciones más complejas (factor de riesgo politómico con controles múltiples, exposición múltiple con cualquier combinación de controles, etc.) ya que con estos métodos elementales, la obtención de estimadores se vuelve impracticable. Estas situaciones pueden trabajarse en el contexto de modelos logísticos.

III.3.1 Factor de Riesgo Dicotómico y Aparejamiento Individual de un Control por Caso

Modelo

Se considera un estudio de casos y controles en el cual se selecciona una muestra aleatoria de casos y posteriormente con el objeto de controlar los efectos de uno o varios factores de confusión, se elige para cada caso en la muestra un control que tenga las mismas características que él, respecto a los factores de confusión que se pretenden controlar. Se tiene entonces una muestra aleatoria de n pares que consisten de un caso y un control cada uno.

La información se resume en una tabla cruzada de la siguiente forma:

		Control		
		Expuesto	No Expuesto	
Caso	Expuesto	n_{11}	n_{10}	$n_{11} + n_{10}$
	No Expuesto	n_{01}	n_{00}	$n_{01} + n_{00}$

donde

n_{11}: número de pares observados donde tanto el caso como el control están expuestos al factor de riesgo o tratamiento.

n_{10}: número de pares observados donde el caso está expuesto al factor de

[8] Si se considera a cada par o grupo de caso-control como un sólo estrato, se pueden aplicar los métodos exactos de análisis para datos no aparejados individualmente, cuando se controla por la presencia de un factor de confusión, ver sección III.2.

riesgo pero el control no.

n_{01}: número de pares observados donde el caso no está expuesto al factor de riesgo y el control si lo está.

n_{00}: número de pares observados donde ni el caso ni el control están expuestos.

Para poder realizar inferencias acerca de la razón de momios ψ, se requiere de un modelo probabilístico para los datos observados, que dependa únicamente de dicho parámetro. A continuación se describe dicho modelo:

Puede considerarse que cada par (caso, control) constituye un estrato, con tan sólo dos elementos. Se tienen entonces n tablas 2×2.

Los posibles resultados de un par son los siguientes:

	Factor de Riesgo											
	Expuesto +	No Expuesto -		Expuesto +	No Expuesto -		Expuesto +	No Expuesto -		Expuesto +	No Expuesto -	
Caso	1	0	1	1	0	1	0	1	1	0	1	1
Control	1	0	1	0	1	1	1	0	1	0	1	1
	2	0		1	1		1	1		0	2	

Como se mencionó en la sección anterior, el modelo probabilístico más adecuado para realizar inferencias sobre la razón de momios ψ cuando se tiene una tabla 2×2 con pocas observaciones, se obtiene condicionando a que los marginales están fijos y resulta ser una hipergeométrica no central con parámetro de no centralidad precisamente ψ. Entonces las probabilidades condicionales de los posibles resultados de un par, se expresan de la siguiente manera:

$$P(caso+, control+ \mid 1,1,2,0) = \frac{\binom{1}{1}\binom{1}{1}\psi^1}{\binom{1}{1}\binom{1}{1}\psi^1} = 1$$

$$P(caso-, control- \mid 1,1,0,2) = \frac{\binom{1}{0}\binom{1}{0}\psi^0}{\binom{1}{0}\binom{1}{0}\psi^0} = 1$$

$$P(caso+, control- \mid 1,1,1,1) = \frac{\binom{1}{1}\binom{1}{0}\psi^1}{\binom{1}{0}\binom{1}{1}\psi^0 + \binom{1}{1}\binom{1}{0}\psi^1} = \frac{\psi}{1+\psi}$$

$$P(caso-, control+ \mid 1,1,1,1) = \frac{\binom{1}{0}\binom{1}{1}\psi^0}{\binom{1}{0}\binom{1}{1}\psi^0 + \binom{1}{1}\binom{1}{0}\psi^1} = \frac{1}{1+\psi}$$

Debe observarse que cuando una tabla tiene un total marginal igual a 0, la probabilidad condicional del resultado observado es uno y entonces dicha tabla no contribuye con ninguna información acerca de la razón de momios. En consecuencia para realizar el análisis estadístico únicamente se utilizan aquellos pares en los que el nivel del factor de riesgo difiere entre el caso y el control, es decir se trabaja únicamente con los pares discordantes.

Si se considera sólamente a los pares discordantes, sus probabilidades (que serán las mismas que las descritas anteriormente) pueden establecerse alternativamente de la siguiente manera:

Sea

$$P_1 = 1 - Q_1 = P(\text{caso } +)$$
$$P_0 = 1 - Q_0 = P(\text{control } +)$$

Entonces si en lugar de condicionar con los marginales, se condiciona con el hecho de que el par es discordante, se tiene que:

$$P(\text{caso } + \mid \text{par discordante}) = \frac{P_1Q_0}{P_1Q_0 + P_0Q_1} = \frac{\frac{P_1Q_0}{P_0Q_1}}{\frac{P_1Q_0}{P_0Q_1} + \frac{P_0Q_1}{P_0Q_1}} = \frac{\psi}{\psi+1} = \pi$$

$$P(\text{caso - } \mid \text{par discordante}) = \frac{Q_1P_0}{Q_1P_0 + P_1Q_0} = \frac{\frac{P_0Q_1}{P_0Q_1}}{\frac{P_0Q_1}{P_0Q_1} + \frac{P_1Q_0}{P_0Q_1}} = \frac{1}{1+\psi} = 1 - \pi$$

Sea $n_d = n_{10} + n_{01}$, el total de pares discordantes, entonces el modelo probabilístico que describe la distribución condicional del número de pares donde el caso está expuesto (que se denotará por N_{10}) es una binomial con parámetros n_d y π, es decir:

$$P[N_{10} = n_{10} \mid n_d, P_0, P_1] = \binom{n_d}{n_{10}} \pi^{n_{10}}(1-\pi)^{n_{01}}$$

En conclusión, como los pares discordantes son los únicos que proporcionan información acerca de la razón de momios, sólo ellos se consideran en el análisis y el

modelo probabilístico para los datos observados que permite hacer inferencias acerca de la razón de momios ψ es la distribución binomial.

En ocasiones para realizar inferencias acerca de ψ, cuando n_d es grande, se aproxima la distribución condicional de N_{10} con una distribución normal con media $n_d\pi$ y varianza $n_d\pi(1-\pi)$.

Estimación Puntual de ψ

El estimador de máxima verosimilitud del parámetro π de una binomial es

$$\hat{\pi} = \frac{n_{10}}{n_d}$$

Expresando a ψ en términos de π se tiene que:

$$\pi = \frac{\psi}{\psi+1} \Rightarrow \pi = \frac{1}{1+\frac{1}{\psi}} \Rightarrow \psi = \frac{\pi}{1-\pi}$$

Entonces el estimador de máxima verosimilitud de ψ es:

$$\hat{\psi} = \frac{\hat{\pi}}{1-\hat{\pi}} = \frac{n_{10}}{n_{01}}$$

Debe observarse que este estimador resulta ser igual al estimador de Mantel y Haenszel:

$$\hat{\psi}_{m-h} = \frac{\sum_{i=1}^{n} \frac{a_i d_i}{N_i}}{\sum_{i=1}^{n} \frac{b_i c_i}{N_i}} = \frac{\frac{n_{10}(1)}{2} + \frac{n_{01}(0)}{2}}{\frac{n_{10}(0)}{2} + \frac{n_{01}(1)}{2}} = \frac{n_{10}}{n_{01}}$$

Estimación por intervalo de ψ

A continuación se describen métodos exactos y aproximados para obtener los límites de confianza para π. Los límites respectivos para ψ se obtienen realizando una transformación inversa, es decir:

$$\psi = \frac{\pi}{1-\pi}$$

Límites Exactos

Para obtener un intervalo de confianza al $(1-\alpha)$ 100% para π, se tienen que resolver los siguientes sistemas de ecuaciones:

$$\alpha/2 = \sum_{j=0}^{n_{10}} \binom{n_d}{j} \pi_s^j (1-\pi_s)^{n_d-j} \quad \text{donde } \pi_s : \text{límite superior del intervalo}$$

$$\alpha/2 = \sum_{j=n_{10}}^{n_d} \binom{n_d}{j} \pi_i^j (1-\pi_i)^{n_d-j} \quad \text{donde } \pi_i : \text{límite inferior del intervalo}$$

Límites Aproximados

Aproximando la distribución de N_{10} condicional a n_d, mediante una distribución normal con media $n_d\pi$ y varianza $n_d\pi(1-\pi)$ y utilizando la corrección por continuidad, se tiene que:

$$P[n_d\pi - Z_{\alpha/2}\sqrt{n_d\pi(1-\pi)} - \frac{1}{2} < N_{10} < n_d\pi + Z_{\alpha/2}\sqrt{n_d\pi(1-\pi)} + \frac{1}{2}] = 1-\alpha$$

Para encontrar los límites de confianza para π, se tienen que resolver las siguientes desigualdades:

$$(n_{10} - n_d\pi - \frac{1}{2}) < Z_{\alpha/2}\sqrt{n_d\pi(1-\pi)}$$

y

$$-Z_{\alpha/2}\sqrt{n_d\pi(1-\pi)} < n_{10} - n_d\pi + \frac{1}{2},$$

en donde n_{10} es el valor observado de la v.a. N_{10}.

Estas desigualdades son equivalentes a:

$$\frac{n_{10} - n_d\pi - \frac{1}{2}}{Z_{\alpha/2}\sqrt{n_d\pi(1-\pi)}} < 0 \quad \text{y} \quad \frac{n_{10} - n_d\pi + \frac{1}{2}}{-Z_{\alpha/2}\sqrt{n_d\pi(1-\pi)}} > 0$$

Las soluciones se obtienen al igualar a cero y resolver las ecuaciones cuadráticas resultantes:

$$\frac{n_{10} - n_d\pi_i - \frac{1}{2}}{\sqrt{n_d\pi_i(1-\pi_i)}} = Z_{\alpha/2}$$

$$\frac{n_{10} - n_d\pi_s + \frac{1}{2}}{\sqrt{n_d\pi_s(1-\pi_s)}} = -Z_{\alpha/2}$$

La primera ecuación determina el límite de confianza inferior para π, π_i; y la segunda ecuación determina el límite superior para π, π_s..

Pruebas de Hipótesis sobre ψ

La manera más simple de realizar una prueba de hipótesis sobre ψ, es realizarla mediante la prueba de hipótesis correspondiente para π. Como existe una relación uno a uno entre ψ y π, una vez que se fija el valor nulo de ψ, ψ_0 se despeja el valor correspondiente para π, π_0 y se realiza la prueba para π.

Pruebas Exactas

Como la distribución condicional de N_{10} dado n_d es una Binomial (n_d, π), la manera de realizar pruebas de hipótesis sobre π, es la usual para cuando se tiene una población binomial, es decir:

Para probar $Ho : \pi = \pi_0$ vs $Ha : \pi \neq \pi_0$, rechazar Ho si $n_{10} \leq t_1$ ó $n_{10} \geq t_2$ donde t_1 y t_2 se obtienen de tablas de distribución de binomiales de tal forma que cumplan con que: $P[N_{10} \leq t_1 \mid \pi = \pi_0] \doteq \alpha/2 \doteq P[N_{10} \geq t_2 \mid \pi = \pi_0]$

Para probar $Ho : \pi = \pi_0$ vs $Ha : \pi > \pi_0$, rechazar Ho si $n_{10} \geq t$ donde t es tal que $P[N_{10} \geq t \mid Ho : \pi = \pi_0] \doteq \alpha$

Para probar $Ho : \pi = \pi_0$ vs $Ha : \pi < \pi_0$, rechazar Ho si $n_{10} \leq t$ donde t es tal que $P[N_{10} \leq t \mid Ho : \pi = \pi_0] \doteq \alpha$

Si por ejemplo, se quiere probar que no existe asociación entre el factor de riesgo y el factor respuesta, se plantea la hipótesis nula de que $\psi = 1$, la cual corresponde a la hipótesis nula de que $\pi = \frac{1}{2}$ y se realiza la prueba para π correspondiente, que dependerá de la alternativa de interés. La alternativa en ψ debe traducirse a la correspondiente en π.

Pruebas Aproximadas

Cuando n_d es grande, se aproxima la distribución condicional de N_{10} mediante una distribución normal, en consecuencia $\frac{N_{10} - n_d\pi}{\sqrt{n_d\pi(1-\pi)}} \sim N(0,1)$ y ésta es precisamente la estadística de prueba que se utiliza para contrastar las hipótesis sobre π.

Para probar $Ho : \pi = \pi_0$ vs $Ha : \pi > \pi_0$, rechazar Ho si

$$P_s \approx 1 - \Phi\left(\frac{n_{10} - n_d\pi_0 - \frac{1}{2}}{\sqrt{n_d\pi_0(1-\pi_0)}}\right) \leq \alpha.$$

Para probar $Ho : \pi = \pi_0$ vs $Ha : \pi < \pi_0$, rechazar Ho si

$P_i \approx \Phi\left(\frac{n_{10} - n_d\pi_0 + \frac{1}{2}}{\sqrt{n_d\pi_0(1-\pi_0)}}\right) \leq \alpha.$

Para probar $Ho : \pi = \pi_0$ vs $Ha : \pi \neq \pi_0$ rechazar Ho si $min\{P_i, P_s\} \leq \alpha/2$.

Para contrastar este último par de hipótesis se puede utilizar alternativamente como estadística de prueba la siguiente:

$$X^2 = \frac{(\mid N_{10} - n_d\pi_0 \mid -\frac{1}{2})^2}{n_d\pi_0(1-\pi_0)}$$

X^2 tiene una distribución Ji cuadrada con un grado de libertad. Se rechaza entonces Ho a un nivel α de significancia, si

$$\frac{(\mid n_{10} - n_d\pi_0 \mid -\frac{1}{2})^2}{n_d\pi_0(1-\pi_0)} > \chi^2_1(\alpha)$$

Por ejemplo, si se desea probar que no hay asociación entre el factor de riesgo y el factor respuesta, es decir, $Ho : \pi = \frac{1}{2}$ contra $Ha : \pi \neq \frac{1}{2}$, se rechaza H_o si

$$X^2_c = \frac{(\mid n_{10} - \frac{n_d}{2} \mid -\frac{1}{2})^2}{\frac{n_d}{4}} = \frac{(\mid n_{10} - n_{01} \mid -1)^2}{n_d} > \chi^2_1(\alpha)$$

Cuando se utiliza la estadística de prueba X^2 para probar $Ho : \pi = \frac{1}{2}$ contra $Ha : \pi \neq \frac{1}{2}$, a la prueba resultante se le conoce como Prueba de Mc Nemar para probar igualdad de proporciones en muestras apareadas (ver Everitt, B.S. (1977), pág. 21, Conover W.J., pág. 130).

Control de Factores de Confusión

En múltiples ocasiones en la etapa de análisis de la investigación se desea controlar por la presencia de otros factores de confusión, además de los que ya se controlaron mediante el aparejamiento individual en el diseño del estudio.

Una práctica muy común en el pasado para resolver este problema era restringir el análisis a aquellos pares que presentaban valores similares respecto al factor de confusión que se quería controlar y posteriormente se realizaba un análisis estratificado. La desventaja de este procedimiento es que se desperdicia mucha información al eliminar las parejas, ya que es muy difícil que coincidan en los valores de un factor de confusión que no fue controlado en el diseño.

En la actualidad este problema se resuelve mediante la utilización de modelos de regresión logística multivariados, en los cuales se incluyen tanto a los factores de riesgo como de confusión para analizar sus efectos en el factor respuesta (ver Capítulo IV).

Pruebas de Homogeneidad de las Razones de Momios (No Interacción)

Es importante notar que el hecho de que un factor se utilice para realizar el aparejamiento individual, no impide el que se estudie su interacción con el factor de riesgo y el factor respuesta, es decir, la interacción de un determinado factor no se ve afectada por su uso en el aparejamiento. Se dividen los pares (caso-control) en grupos de acuerdo a los niveles de dicho factor de confusión y se pueden obtener estimadores separados de las razones de momios y compararlos.

La forma más fácil de realizar una prueba de homogeneidad de razones de momios entre estratos cuando se llevó a cabo un aparejamiento individual uno a uno (un control por caso), es a través de sus probabilidades asociadas π:

Supóngase que el factor de confusión cuya interacción con el factor de riesgo y respuesta quiere estudiarse, tiene I niveles. La información disponible se encuentra distribuida en I tablas cruzadas de la siguiente manera:

Estrato i

		Control Expuesto +	Control No Expuesto -
Caso	Expuesto (+)	n_{11i}	n_{10i}
	No Expuesto (-)	n_{01i}	n_{00i}

En cada estrato los pares discordantes son los únicos que proporcionan información acerca de la razón de momios ψ_i, entonces la distribución del número de pares donde sólo el caso está expuesto N_{10i}, condicional a que el total de pares discordantes, $n_{di} = n_{10i} + n_{01i}$ está fijo, es una binomial con parámetros π_i y n_{di} con $\pi_i = \frac{\psi_i}{1+\psi_i}$, entonces si se prueba que $\pi_i = \pi \ \forall_i$ es equivalente a probar que $\psi_i = \psi \ \forall_i$.

Condicional a las n_{di}, para probar la homogeneidad de las I probabilidades, se realiza una prueba de igualdad de proporciones, que equivale a una prueba de independencia. Si el total de pares discordantes se clasifica de la siguiente manera:

Factor de Confusión

	1	2	3	.....	I	
Pares con Caso +	n_{101}	n_{102}	n_{103}		n_{10I}	n_{10}
Pares con Caso -	n_{011}	n_{012}	n_{013}		n_{01I}	n_{01}
	n_{d1}	n_{d2}	n_{d3}		n_{dI}	n_d

se puede entonces realizar una prueba usual de independencia.

El procedimiento que se acaba de describir puede utilizarse para estudiar los efectos de interacción de otros factores diferentes a los que se usaron para realizar el aparejamiento. Sin embargo, no se recomienda su uso puesto que adolece del mismo problema de pérdida de información que se mencionó cuando se pretende controlar en el análisis por la presencia de factores de confusión no utilizados para realizar el aparejamiento.

III.3.2 Factor de Riesgo Dicotómico y Aparejamiento Individual de M Controles por Caso

Modelo

Se considera un estudio de casos y controles en el cual se selecciona una muestra aleatoria de casos y posteriormente con el objeto de controlar los efectos de uno o varios factores de confusión, se elige para cada caso en la muestra un número M fijo de controles que tengan las mismas características que él, respecto a los factores de confusión.

Los resultados del estudio pueden resumirse en una tabla de la siguiente manera:

Número de Controles Expuestos al Factor de Riesgo

		0	1	2	3	4	5	6	.....	M-1	M
Caso	Expuesto (+)	n_{10}	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	n_{16}		n_{1M-1}	n_{1M}
	No Expuesto (-)	n_{00}	n_{01}	n_{02}	n_{03}	n_{04}	n_{05}	n_{06}		n_{0M-1}	n_{0M}

donde

n_{ij}: número de conjuntos aparejados observados en donde i casos están expuestos y j controles están expuestos, $i = 1,0$ $j = 1,...,M$.

A continuación se describe un modelo probabilístico para los datos observados que permite realizar inferencias acerca de la razón de momios ψ.

Los posibles resultados de un conjunto (1 caso - M controles) son los siguientes:

Factor de Riesgo

	+	−	+	−	+	−	+	−	.....	+	−	+*	−
Caso	1	0	1	0	1	0	1	0		1	0	1	0
Control	0	M	1	M−1	2	M−2	3	M−3		M−1	1	M	0
Total de Expuestos	1		2		3		4			M		M+1	

Factor de Riesgo

	+*	−	+	−	+	−	+	−	.....	+	−	+	−
Caso	0	1	0	1	0	1	0	1		0	1	0	1
Control	0	M	1	M−1	2	M−2	3	M−3		M−1	1	M	0
Total de Expuestos	0		1		2		3			M−1		M	

Se observa que para un determinado total de expuestos, existen dos configuraciones alternativas, excepto para las situaciones en las que hay $M+1$ expuestos ó ningún expuesto. Estas excepciones representan las ocasiones en las que el caso y sus controles presentan la misma condición de exposición y por lo tanto no proporcionan ninguna información acerca de la razón de momios. En consecuencia para realizar el análisis, se eliminan todos los conjuntos que presenten estos resultados, los cuales están marcados con un asterisco en las tablas anteriores.

Sean: P_1:probabilidad de que el caso esté expuesto

P_0:probabilidad de que el control esté expuesto

m:total de expuestos en un conjunto aparejado, $m = 1, 2, ...M$

Se tiene entonces, que para un número determinado m de expuestos existen dos resultados posibles:

Factor de Riesgo

	+	−	
Caso	1	0	1
Control	m-1	M-m+1	M
	m	M-m+1	M+1

Factor de Riesgo

	+	−	
Caso	0	1	1
Control	m	M-m	M
	m	M-m+1	M+1

Sus probabilidades están dadas por:

$$P(caso+ \; y \;\; (m-1)controles+) = P(caso+)P((m-1)controles+)$$

$$= P_1[\binom{M}{m-1} P_0^{m-1}(1-P_0)^{M-m+1}]$$

$$P(caso- \;\; y \;\; m \;\; controles+) = P(caso-)P(m \;\; controles+)$$

$$= (1-P_1)[\binom{M}{m} P_0^{m}(1-P_0)^{M-m}]$$

Se tiene entonces que la probabilidad condicional de que el caso esté expuesto es:

$$P(caso+ \mid m \text{ expuestos en total }) = \frac{P(caso+ \; y \;\; (m-1) \text{ controles } +)}{P(m \text{ expuestos en total })}$$

$$= \frac{\binom{M}{m-1} P_1 P_0^{m-1}(1-P_0)^{M-m+1}}{\binom{M}{m-1} P_1 P_0^{m-1}(1-P_0)^{M-m+1} + \binom{M}{m}(1-P_1)P_0^{m}(1-P_0)^{M-m}}$$

$$= \frac{mP_1Q_0}{mP_1Q_0 + (M-m+1)P_0Q_1} = \frac{m\frac{P_1Q_0}{P_0Q_1}}{m\frac{P_1Q_0}{P_0Q_1} + M - m + 1} = \frac{m\psi}{m\psi + M - m + 1}$$

Sean

$N_{1,m-1}$: número de conjuntos donde el caso y $m-1$ controles están expuestos, $m = 1, ..., M$

$T_m = n_{1,m-1} + n_{0,m}$: el total de conjuntos observados donde hay exactamente m expuestos.

entonces

$$P[N_{1,m-1} = n_{1,m-1} \mid T_m; \psi] =$$

$$\binom{T_m}{n_{1,m-1}} \left(\frac{m\psi}{m\psi + M - m + 1}\right)^{n_{1,m-1}} \left(\frac{M-m+1}{m\psi + M - m + 1}\right)^{n_{0,m}}$$

El modelo probabilístico que describe la distribución conjunta (condicional) del número de conjuntos donde el caso está expuesto $N_{1,0}, N_{1,1}, ..., N_{1,M-1}$, es un producto de binomiales:

$$P[N_{10} = n_{10}, N_{11} = n_{11}, ..., N_{1,M-1} = n_{1,M-1} \mid T_1 ... T_M; \psi]$$

$$= \prod_{m=1}^{M} \binom{T_m}{n_{1,m-1}} \left(\frac{m\psi}{m\psi + M - m + 1}\right)^{n_{1,m-1}} \left(\frac{M - m + 1}{m\psi + M - m + 1}\right)^{n_{0,m}}$$

Estimación Puntual de ψ

Estimador de Máxima Verosimilitud

Después de realizar los procedimientos para maximizar la $P[N_{10}, ..., N_{1,M-1} \mid T_1, ..., T_M : \psi]$ resulta que el estimador de máxima verosimilitud condicional se obtiene como solución de la ecuación siguiente:

$$\sum_{m=1}^{M} n_{1,m-1} = \sum_{m=1}^{M} \frac{T_m m\psi}{m\psi + M - m + 1}$$

Para resolverla se requieren procedimientos numéricos iterativos.

Estimador de Mantel-Haenszel

Si se considera a cada conjunto aparejado como un estrato, el estimador de la razón de momios común propuesto por Mantel y Haenszel resulta ser, un estimador de ψ más fácil de calcular y que como se mencionó en la sección III.2, es además un estimador robusto.

$$\hat{\psi}_{m-h} = \frac{\sum_{m=1}^{M} n_{1,m-1}(M - m + 1)}{\sum_{m=1}^{M} n_{0m}(m)}$$

Estimación por Intervalo

Se presentan únicamente procedimientos para obtener límites de confianza aproximados.

La distribución de $(N_{1,m-1} \mid T_m; \psi)$ es $B\left(T_m, \frac{m\psi}{m\psi+M-m+1}\right) \forall\, m = 1, ..., M.$

Se puede entonces aproximar la distribución condicional de $\sum_{m=1}^{M} N_{1,m-1}$ dados T_m y ψ, mediante una normal con media $\sum_{m=1}^{M} E[N_{1,m-1} \mid T_m; \psi]$ y varianza

$\sum_{m=1}^{M} Var(N_{1,m-1} \mid T_m; \psi)$ donde

$$E[N_{1,m-1} \mid T_m; \psi] = T_m[\frac{m\psi}{m\psi + M - m + 1}] \quad \forall \ m = 1...M$$

$$V[N_{1,m-1} \mid T_m; \psi] = T_m[\frac{m\psi(M - m + 1)}{(m\psi + M - m + 1)^2}] \quad \forall \ m = 1...M$$

Para obtener un intervalo de confianza para ψ al $(1 - \alpha) \times 100\%$, se tienen entonces que resolver las siguientes ecuaciones:

$$\frac{\sum_{m=1}^{M}[n_{1,m-1} - E(N_{1,m-1} \mid T_m; \psi_i)] - \frac{1}{2}}{\sqrt{\sum_{m=1}^{M} V(N_{1,m-1} \mid T_m; \psi_i)}} = Z_{\alpha/2}$$

$$\frac{\sum_{m=1}^{M}[n_{1,m-1} - E(N_{1,m-1} \mid T_m; \psi_s)] + \frac{1}{2}}{\sqrt{\sum_{m=1}^{M} V(N_{1,m-1} \mid T_m; \psi_s)}} = -Z_{\alpha/2}$$

donde $n_{1,m-1}$ es el valor observado de la variable aleatoria $N_{1,m-1}$ ψ_i y ψ_s límites inferior y superior del intervalo respectivamente

Límites Logísticos

Es más fácil calcular límites de confianza para el $\ln \psi$. Para muestras grandes el $\ln \hat{\psi}$ se distribuye aproximadamente como una normal (Miettinen (1970)) con media $E[\ln \hat{\psi}] \doteq \ln \psi$ y varianza aproximada $[\sum_{m=1}^{M} \frac{T_m m\psi(M-m+1)}{(m\psi+M-m+1)^2}]^{-1}$ (Breslow y Day (1982), Cap. 5).

Para estimar la varianza se puede utilizar el estimador de máxima verosimilitud de ψ o el de Mantel-Haenszel. Entonces los límites al $(1 - \alpha)100\%$ de confianza se obtienen de la siguiente manera:

$$\ln \hat{\psi} - Z_{\alpha/2}\sqrt{Var(\ln \hat{\psi})} = \ln \psi_i \qquad \text{límite inferior}$$

$$\ln \hat{\psi} - Z_{\alpha/2}\sqrt{Var(\ln \hat{\psi})} = \ln \psi_s \qquad \text{límite superior}$$

Tomando antilogaritmos se obtienen los límites aproximados para ψ.

Pruebas de Hipótesis sobre ψ

Utilizando la aproximación normal de la distribución condicional de la $\sum_{m=1}^{M} N_{1,m-1}$ dados T_m y ψ, se pueden realizar pruebas de hipótesis sobre ψ de la manera usual.

Prueba de Homogeneidad de las Razones de Momios (No Interacción)

Como se mencionó para el caso de aparejamiento individual uno a uno (un control por caso), el procedimiento que a continuación se describe conviene utilizarlo únicamente para probar la interacción (con el factor de riesgo y respuesta) de los factores que hayan sido utilizados para realizar el aparejamiento, ya que de otro modo se pierde mucha de la información disponible. Esto se debe a que es muy difícil que los conjuntos aparejados tengan valores similares en otros factores diferentes a los que se emplearon en el aparejamiento.

Supóngase que se quiere probar la interacción de un factor que tiene H niveles. Los conjuntos aparejados se dividen entonces en H grupos y se pueden obtener estimadores separados de las razones de momios y compararlos.

Sea $N_{1,m,h}$: número de conjuntos aparejados en los que el caso y m controles están expuestos, en el h-ésimo grupo. h = 1,...,M.

$N_{0,m,h}$: número de conjuntos aparejados donde el caso no está expuesto y m controles si están expuestos, en el h-ésimo grupo.

$n_{1,m,h}$ y $n_{0,m,h}$: los valores observados de dichas variables respectivamente.

$T_{m,h} = n_{1,m-1,h} + n_{0,m,h}$: el total de conjuntos aparejados que tienen m expuestos, en el h-ésimo grupo.

ψ_h: la razón de momios en el h-ésimo grupo.

Si se quiere probar $Ho : \psi_1 = \psi_2 = ... = \psi_H = \psi$ contra $Ha : \psi_h \neq \psi$ para al menos una h, se puede utilizar como estadística de prueba [7] la siguiente:

---

[7] Esta prueba es una aplicación de la prueba que se describió al final de la sección III.2.2.

$$K = \sum_{h=1}^{H} \left[\frac{\{ \sum_{m=1}^{M} N_{1,m-1,h} - \sum_{m=1}^{M} E(N_{1,m-1,h} \mid T_{m,h}; \hat{\psi}) \}^2}{\sum_{m=1}^{M} Var(N_{1,m-1,h} \mid T_{m,h}; \hat{\psi})} \right]$$

donde $\hat{\psi}$ es un estimador global de la razón de momios que combina la información de los H grupos y puede ser el de máxima verosimilitud o el de Mantel-Haenzel.

La estadística K tiene una distribución Ji cuadrada con $H-1$ grados de libertad.

CAPITULO IV.

LOS MODELOS LOGISTICOS PARA EL ANALISIS DE CASOS Y CONTROLES

IV.1 DEFINICION GENERAL DEL MODELO LOGISTICO

Introducción

El problema a abordar es el siguiente:

Se quiere estudiar un fenómeno cuyos posibles resultados son solo dos (muerte o vida, éxito o fracaso, enfermedad o no enfermedad, etc.), es decir, la variable a estudiar es una variable binaria. Se dispone además de información de ciertas variables que se supone estan relacionadas con la de interés y que por lo tanto ayudan a explicar su comportamiento.

Supóngase entonces que se tienen $Y_1, \ldots, Y_n$ variables aleatorias independientes que se distribuyen como Bernoulli con parámetro P_i. El objetivo es encontrar métodos que permitan establecer la dependencia de $P_i = P[Y_i = 1] = E[Y_i]$ en ciertas variables explicativas, $X_1, \ldots, X_p$.

Se podría intentar modelar dichas probabilidades P_i considerando un modelo de regresión usual:

$$E[Y_i \mid \underline{X}_i] = P[Y_i = 1 \mid \underline{X}_i] = P_i = \sum_{k=1}^{p} \beta_k X_{ik}$$

donde:

$\underline{X}_i = (X_{i1}, \ldots, X_{ip})$: vector de variables explicativas para la observación i-ésima.

Y entonces aplicar mínimos cuadrados directamente a las observaciones binarias.

Este procedimiento tiene las siguientes limitaciones (Cox, D., 1970):

i) Como la $V(Y_i) = P_i(1 - P_i)$, la condición de igualdad de varianzas que se requiere para aplicar mínimos cuadrados no se cumple, salvo, en el caso poco interesante en el que $P_i = P$ para toda i. Sin embargo, como se sabe que cambios moderados en la $V(Y_i)$ producen una pérdida modesta de eficiencia, si $.2 \leq P_i \leq .8$, se podría aplicar el procedimiento.

   En realidad el problema de heterogeneidad de varianzas no es grave, ya que en caso de que éstas varien mucho, se puede emplear el procedimiento de mínimos cuadrados ponderados, utilizando un esquema iterativo en el que se ajusta en primera instancia el modelo utilizando mínimos cuadrados sin ponderar, obteniendose los valores ajustados $\hat{Y}_i$ y entonces se aplica el método de mínimos cuadrados ponderados con ponderaciones $\{\hat{Y}_i(1 - \hat{Y}_i)\}^{-1}$.

ii) Como la distribución de las Y_i's no es normal, ningún método de estimación

que sea lineal en las Y_i's será completamente eficiente.

iii) La restricción más fuerte al uso de este método surge del requerimiento de que $0 \leq P_i \leq 1$. Si se aplican mínimos cuadrados ponderados o no, directamente a las observaciones binarias, puede ocurrir que se obtenga un vector de valores ajustados con algún componente que no satisfaga esta condición.

Este problema puede resolverse considerando estimadores de mínimos cuadrados modificados, obtenidos al minimizar la suma de cuadrados de residuales sujeto a que se cumpla la restricción de que todos los valores ajustados se encuentren entre 0 y 1.

Este procedimiento presenta un alto grado de dificultad desde el punto de vista computacional.

iv) Otra limitación mucho más seria desde el punto de vista del objetivo de la investigación, es que los parámetros del modelo tienen una interpretación y un rango de validez limitados.

El Modelo Logístico Lineal

Una manera más simple de representar la dependencia de una probabilidad P_i, en variables explicativas, de tal forma que se cumpla la condición $0 \leq P_i \leq 1$, es mediante el siguiente modelo:

Sean $Y_1, \ldots, Y_n$ variables aleatorias independientes que se distribuyen como Bernoulli (P_i), entonces, para $i = 1, \ldots, n$

$$P_i = \frac{e^{\underline{X}_i \underline{\beta}}}{1 + e^{\underline{X}_i \underline{\beta}}}$$

$$1 - P_i = \frac{1}{1 + e^{\underline{X}_i \underline{\beta}}}$$

donde:

$\underline{X}_i = (X_{i1}, \ldots, X_{ip})$

$\underline{\beta}' = (\beta_1 \ldots, \beta_p)$

Estas ecuaciones son equivalentes a:

$$\lambda_i = logit P_i = ln \frac{P_i}{1 - P_i} = \underline{X}_i \underline{\beta} = \sum_{k=1}^{p} \beta_k X_{ik} \qquad \forall \quad i = 1, \ldots, n$$

y en términos matriciales pueden escribirse como

$$\underline{\lambda} = X\underline{\beta} \qquad \text{donde} \qquad X = \begin{pmatrix} \underline{X}_1 \\ \vdots \\ \underline{X}_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

Al modelo $\underline{\lambda} = X\underline{\beta}$ se le conoce como el modelo logístico lineal. Deriva su nombre del hecho de que a la transformación logit o logística de P_i, se le expresa como función lineal de ciertos parámetros asociados con variables explicativas.

Es importante hacer notar que estas variables explicativas pueden ser de cualquier tipo (continuas o discretas).

Los modelos logísticos lineales son muy útiles para el análisis de estudios de casos y controles ya que, permiten explorar el efecto individual y conjunto de muchos factores de riesgo; controlar simultáneamente por la presencia de muchos factores de confusión y analizar interacciones. Además, los parámetros del modelo tienen una interpretación directa en términos de la razón de momios, que es la única medida del efecto que el factor de riesgo tiene en el factor respuesta estrictamente estimable a partir de estudios de este tipo.

IV.2 INTERPRETACION DE PARAMETROS DEL MODELO [1]

A continuación se presentan algunos ejemplos en donde se retoman algunas de las situaciones analizadas con métodos elementales en el Capítulo III y se presentan los modelos logísticos correspondientes así como la interpretación de sus parámetros. Cabe aclarar que la discusión es a nivel parametral y que por el momento no se hablará de los procedimientos de estimación.

Modelo logístico lineal cuando se tiene un factor de riesgo y un factor respuesta dicotómicos: Tabla 2×2

Supóngase que se tiene un factor de riesgo dicotómico: expuesto o no expuesto y un factor respuesta dicotómico: enfermo o no enfermo.

En este caso se tienen dos poblaciones: la de expuestos y la de no expuestos y se observa en cada una de ellas una variable binaria: enfermo o no enfermo. Se supone que para todos los individuos de una misma población la probabilidad de éxito (en este caso enfermo) es la misma.

La situación puede resumirse en una tabla de 2×2 de la siguiente forma:

Factor de Riesgo

	No Expuesto	Expuesto
Enfermo	$P(0)$	$P(1)$
No Enfermo	$Q(0)$	$Q(1)$

[1] En esta sección se supondrá, para facilitar la exposición, que los datos provienen de un estudio de cohortes. Para que los datos de un estudio de casos y controles se analicen del mismo modo que los de un estudio de cohortes se deben cumplir ciertos requisitos los cuales se especifican en la sección IV.5

donde:

$P(X)$: probabilidad de enfermarse dado que se está en el grupo X de exposición, $X = 0, 1$

$Q(X) = 1 - P(X)$

Considerese el siguiente modelo logístico lineal:

$$logit P(X) = ln \frac{P(X)}{1 - P(X)} = \alpha + \beta X$$

entonces

$$logit P(0) = \alpha \quad \text{y} \quad logit P(1) = \alpha + \beta$$

Considérese ahora el logaritmo natural de la razón de momios de enfermedad:

$$ln\psi = ln \frac{\frac{P(1)}{Q(1)}}{\frac{P(0)}{Q(0)}} = logit P(1) - logit P(0)$$

Bajo el modelo propuesto:

$$ln\psi = \alpha + \beta - \alpha = \beta$$

Por lo tanto en el modelo propuesto α es el *logit* de la probabilidad de enfermarse en el nivel base de comparación y β es el logaritmo de la razón de momios. El parámetro de interés en este caso es β puesto que $e^{\beta} = \psi$ es el parámetro objeto de la investigación, ya que mide el efecto relativo al nivel base de comparación de la exposición, en la enfermedad.

El modelo propuesto es un modelo saturado ya que tiene igual número de parámetros que de probabilidades a modelar.

Modelo Logístico cuando se tienen dos factores de riesgo dicotómicos y un factor respuesta dicótomico.

Supóngase ahora que se quiere estudiar el efecto conjunto de dos factores de riesgo A y B dicotómicos. Puede ser que los efectos de A y B sean independientes o que interactúen. Se tienen entonces cuatro categorías de riesgo:

Factor de Riesgo A				
	No Expuesto		Expuesto	
	Factor B		Factor B	
	No Expuesto	Expuesto	No Expuesto	Expuesto
Enfermo	$P(0,0)$	$P(0,1)$	$P(1,0)$	$P(1,1)$
No Enfermo	$Q(0,0)$	$Q(0,1)$	$Q(1,0)$	$Q(1,1)$

donde:

$P(X,Y)$: Probabilidad de enfermarse dado que se esta en el nivel X de A y Y de B,

$$X = 0,1 \quad Y = 0,1$$

$$Q(X,Y) = 1 - P(X,Y)$$

Sin pérdida de generalidad se puede elegir como nivel base de comparación aquel en el que no se esta expuesto ni a A ni a B, es decir el nivel $(0,0)$, con la finalidad de evaluar el efecto que cada una de las otras categorías de riesgo tiene en la enfermedad. Se tienen entonces tres razones de momios por estimar:

$$\psi(0,1) = \frac{\frac{P(0,1)}{Q(0,1)}}{\frac{P(0,0)}{Q(0,0)}} \qquad \text{efecto relativo de } B \text{ en ausencia de } A$$

$$\psi(1,0) = \frac{\frac{P(1,0)}{Q(1,0)}}{\frac{P(0,0)}{Q(0,0)}} \qquad \text{efecto relativo de } A \text{ en ausencia de } B$$

$$\psi(1,1) = \frac{\frac{P(1,1)}{Q(1,1)}}{\frac{P(0,0)}{Q(0,0)}} \qquad \text{efecto relativo de } A \text{ y } B \text{ conjuntamente}$$

El primer punto a investigar es si los factores de riesgo interactuan o no lo hacen, lo cual es equivalente a probar la hipótesis de que $\psi(1,1) = \psi(1,0) \times \psi(0,1)$; esto se ve claramente haciendo uso de los métodos tradicionales para el análisis de interacción. De acuerdo a dichos métodos, para investigar si hay interacción se requiere controlar a cada uno de los factores y ver si el efecto del otro en la respuesta (medido en este caso mediante la razón de momios) no varía al pasar de un nivel a otro del que se controla:

Si se controla al factor B y se analiza el efecto de A en la respuesta:

Factor de Riesgo B				
	No Expuesto		Expuesto	
	Factor A		Factor A	
	No Expuesto	Expuesto	No Expuesto	Expuesto
Enfermo	$P(0,0)$	$P(1,0)$	$P(0,1)$	$P(1,1)$
No Enfermo	$Q(0,0)$	$Q(1,0)$	$Q(0,1)$	$Q(1,1)$

Efecto de A en ausencia de B $\quad \psi(1,0) = \frac{\frac{P(1,0)}{Q(1,0)}}{\frac{P(0,0)}{Q(0,0)}}$

Efecto de A en presencia de B $\quad = \frac{\frac{P(1,1)}{Q(1,1)}}{\frac{P(0,1)}{Q(0,1)}} = \frac{P(1,1)/Q(1,1)}{P(0,1)/Q(0,1)} = \frac{\psi(1,1)}{\psi(0,1)}$

Si no hay interacción se tiene que cumplir que $\psi(1,0) = \frac{\psi(1,1)}{\psi(0,1)}$, o equivalentemente $\psi(1,0)\psi(0,1) = \psi(1,1)$

Si se controla al factor A y se analiza el efecto de B en la respuesta:

Factor de Riesgo A				
	No Expuesto		Expuesto	
	Factor B		Factor B	
	No Expuesto	Expuesto	No Expuesto	Expuesto
Enfermo	$P(0,0)$	$P(0,1)$	$P(1,0)$	$P(1,1)$
No Enfermo	$Q(0,0)$	$Q(0,1)$	$Q(1,0)$	$Q(1,1)$

Efecto de B en ausencia de A $\quad \psi(0,1) = \frac{\frac{P(0,1)}{Q(0,1)}}{\frac{P(0,0)}{Q(0,0)}}$

Efecto de B en presencia de A $\quad = \frac{\frac{P(1,1)}{Q(1,1)}}{\frac{P(1,0)}{Q(1,0)}} = \frac{\psi(1,1)}{\psi(1,0)}$

Para que no haya interacción se requiere que se cumpla que $\psi(0,1) = \frac{\psi(1,1)}{\psi(1,0)}$, que es equivalente a $\psi(1,1) = \psi(0,1)\psi(1,0)$.

Si la hipótesis de no interacción se rechaza, se tienen que obtener estimadores separados de las razones de momios para cada categoría de riesgo. Si por el contrario, dicha hipótesis no se rechaza, se pueden obtener estimadores de las razones de momios para los factores A y B individualmente combinando desde luego toda la información

disponible (por ejemplo con el estimador de Mantel y Haenszel).

A continuación se presentan dos modelos logísticos que podrían proponerse en este caso:

a) Modelo Saturado.- Este modelo considera la interacción de los dos factores:

$$logitP(X_1, X_2) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 X_2$$

donde:

$$X_1 = \begin{cases} 1 & \text{si el sujeto está expuesto a } A \\ 0 & \text{si el sujeto no está expuesto a } A \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{si el sujeto está expuesto a } B \\ 0 & \text{si el sujeto no está expuesto a } B \end{cases}$$

$P(X_1, X_2)$: probabilidad de enfermarse dado que se está en la categoría de exposición (X_1, X_2)

Bajo el modelo propuesto:

$$\begin{aligned} logitP(0,0) &= \alpha \\ ln\psi(1,0) &= logitP(1,0) - logitP(0,0) = \alpha + \beta_1 - \alpha = \beta_1 \\ ln\psi(0,1) &= logitP(0,1) - logitP(0,0) = \alpha + \beta_2 - \alpha = \beta_2 \\ ln\frac{\psi(1,1)}{\psi(1,0)\psi(0,1)} &= logitP(1,1) - logitP(0,0) - logitP(1,0) + logitP(0,0) \\ &\quad - logitP(0,1) + logitP(0,0) \\ &= logitP(1,1) - logitP(1,0) - logitP(0,1) + logitP(0,0) \\ &= \alpha + \beta_1 + \beta_2 + \gamma - \alpha - \beta_1 - \alpha - \beta_2 + \alpha \\ &= \gamma \end{aligned}$$

Por lo tanto, en el modelo propuesto α representa el logit de la probabilidad de enfermarse cuando se esta en el nivel base de comparación, β_1 el logaritmo del efecto relativo de A (en ausencia de B), β_2 el logaritmo del efecto relativo de B (en ausencia de A) y γ el logaritmo del efecto adicional por estar expuesto a ambos factores.

Bajo el modelo, $\psi(X, Y) = e^{\beta_1 X_1 + \beta_2 X_2 + \gamma X_1 X_2}$, nótese entonces que si $\gamma > 0$ el riesgo para la exposición combinada es mayor que el predicho para las exposiciones individuales.

Probar $\gamma = 0$ en el modelo logístico, es equivalente a probar que $\psi_{11} = \psi_{10} \times \psi_{01}$ como ya se mencionó anteriormente.

b) Modelo sin interacción.- Si la hipótesis de interacción se rechaza, esto conduce a plantear el ajuste de un modelo con menos parámetros:

$$logitP(X_1, X_2) = \alpha + \beta_1 X_1 + \beta_2 X_2$$

donde X_1, X_2, y $P(X_1, X_2)$ se definen igual que en el caso anterior.

El que no haya interacción significa como ya se mencionó anteriormente, que el efecto del factor A en el factor respuesta es el mismo esté o no esté presente el factor B, es decir $\psi(1,0) = \frac{\psi(1,1)}{\psi(0,1)}$ y que el efecto del factor B en el factor respuesta es el mismo esté o no esté presente el factor A, es decir $\psi(0,1) = \frac{\psi(1,1)}{\psi(1,0)}$.

Nótese entonces que el modelo propuesto impone estas restricciones, puesto que:

$$ln\psi(1,0) = logitP(1,0) - logitP(0,0) = \beta_1$$
$$ln\frac{\psi(1,1)}{\psi(0,1)} = logitP(1,1) - logitP(0,0) - logitP(0,1) + logitP(0,0) = \beta_1$$

Esto implica que $\beta_1 = ln\psi(1,0) = ln\frac{\psi(1,1)}{\psi(0,1)}$, lo cual significa que β_1 representa el logaritmo natural del efecto relativo de A en la enfermedad, esté o no esté presente el factor B.

Similarmente,

$$ln\psi(0,1) = \beta_2$$
$$ln\frac{\psi(1,1)}{\psi(1,0)} = \beta_2$$

lo que implica que $\beta_2 = ln\psi(0,1) = ln\frac{\psi(1,1)}{\psi(1,0)}$, lo cual significa que β_2 representa el logaritmo natural del efecto relativo de B en la enfermedad, esté o no esté presente el factor A.

El parámetro α, al igual que en el modelo saturado, representa el logit de la probabilidad de enfermarse en el nivel base de comparación.

Es importante hacer notar que la interpretación de parámetros en un modelo varía según los términos que éste incluya. Por ejemplo, en el modelo saturado β_1 representa el efecto de A en ausencia de B mientras que en el modelo sin interacción β_1 representa el efecto de A esté o no esté B presente.

Para estimar β_1 en el modelo saturado se utilizaría unicamente la información generada por una tabla como la del lado izquierdo parte superior de la página 100 y en cambio para estimar β_1 bajo el modelo sin interacción se tendría que combinar la información generada por ambas tablas.

Modelo logístico cuando se tienen dos factores de riesgo uno de ellos politómico y un factor respuesta dicotómico

Supóngase ahora que se quiere estudiar el efecto conjunto de dos factores de riesgo, el factor A con dos niveles: Expuesto ó no expuesto y el factor B con tres niveles: bajo (0), medio (1), alto (2). Se tienen entonces seis categorías de riesgo:

Factor de Riesgo A						
	No Expuesto			Expuesto		
	Factor B			Factor B		
	Nivel 0	Nivel 1	Nivel 2	Nivel 0	Nivel 1	Nivel 2
Enfermo	$P(0,0)$	$P(0,1)$	$P(0,2)$	$P(1,0)$	$P(1,1)$	$P(1,2)$
No Enfermo	$Q(0,0)$	$Q(0,1)$	$Q(0,2)$	$Q(1,0)$	$Q(1,1)$	$Q(1,2)$

donde:

$P(X,Y)$: probabilidad de enfermarse dado que se esta en el nivel X de A y Y de B, $X = 0,1$ y $Y = 0,1,2$.

$$Q(X,Y) = 1 - P(X,Y)$$

Con la finalidad de analizar el efecto que cada una de las categorías de riesgo tiene en el factor respuesta, se elige sin pérdida de generalidad, como nivel base de comparación aquella categoría de riesgo en la que no se está expuesto a A y se está expuesto al nivel 0 de B, es decir, el nivel $X = 0, Y = 0$. Se tienen entonces cinco razones de momios por estimar:

$$\psi(0,1) = \frac{\frac{P(0,1)}{Q(0,1)}}{\frac{P(0,0)}{Q(0,0)}}$$ efecto relativo del nivel 1 de B en ausencia de A

$$\psi(0,2) = \frac{\frac{P(0,2)}{Q(0,2)}}{\frac{P(0,0)}{Q(0,0)}}$$ efecto relativo del nivel 2 de B en ausencia de A

$$\psi(1,0) = \frac{\frac{P(1,0)}{Q(1,0)}}{\frac{P(0,0)}{Q(0,0)}}$$ efecto relativo de A al nivel 0 de B

$$\psi(1,1) = \frac{\frac{P(1,1)}{Q(1,1)}}{\frac{P(0,0)}{Q(0,0)}}$$ efecto relativo de A y el nivel 1 de B conjuntamente

$$\psi(1,2) = \frac{\frac{P(1,2)}{Q(1,2)}}{\frac{P(0,0)}{Q(0,0)}}$$ efecto relativo de A y el nivel 2 de B conjuntamente

El primer punto a investigar es, si los dos factores de riesgo interactuan o no lo hacen, lo cual es equivalente a probar la hipótesis de que $\psi(X,Y) = \psi(X,0)\psi(0,Y), \quad X = 1, \quad Y = 1,2.$

Esto se ve claramente del siguiente desarrollo:

Si se controla al factor A y se analiza el efecto relativo de cada categoría de B en la respuesta:

Factor de Riesgo A						
	No Expuesto			Expuesto		
	Factor B			Factor B		
	Nivel 0	Nivel 1	Nivel 2	Nivel 0	Nivel 1	Nivel 2
Enfermo	$P(0,0)$	$P(0,1)$	$P(0,2)$	$P(1,0)$	$P(1,1)$	$P(1,2)$
No Enfermo	$Q(0,0)$	$Q(0,1)$	$Q(0,2)$	$Q(1,0)$	$Q(1,1)$	$Q(1,2)$

Efecto del nivel 1 de B en ausencia de A $= \psi(0,1) \longleftarrow$ Efecto del nivel 1 de B en presencia de A $\frac{\psi(1,1)}{\psi(1,0)} = \frac{\frac{P(1,1)}{Q(1,1)}}{\frac{P(1,0)}{Q(1,0)}}$

Efecto del nivel 2 de B en ausencia de A $= \psi(0,2) \longleftarrow$ Efecto del nivel 2 de B en presencia de A $\frac{\psi(1,2)}{\psi(1,0)} = \frac{\frac{P(1,2)}{Q(1,2)}}{\frac{P(1,0)}{Q(1,0)}}$

Para que no haya interacción, se requiere que el efecto del nivel 1 de B sea el mismo esté o no A presente, es decir, $\psi(0,1) = \frac{\psi(1,1)}{\psi(1,0)}$ y que el efecto del nivel 2 de B

sea el mismo esté o no A presente, es decir, $\psi(0,2) = \frac{\psi(1,2)}{\psi(1,0)}$.

Estas dos condiciones pueden escribirse como $\psi(1,1) = \psi(1,0)\psi(0,1)$ y $\psi(1,2) = \psi(1,0)\psi(0,2)$

Ahora, si se controla por B y se analiza el efecto relativo de A en la respuesta:

	Factor de Riesgo B					
	Nivel 0		Nivel 1		Nivel 2	
	Factor A		Factor A		Factor A	
	No Expuesto	Expuesto	No Expuesto	Expuesto	No Expuesto	Expuesto
Enfermo	$P(0,0)$	$P(1,0)$	$P(0,1)$	$P(1,1)$	$P(0,2)$	$P(1,2)$
No Enfermo	$Q(0,0)$	$Q(1,0)$	$Q(0,1)$	$Q(1,1)$	$Q(0,2)$	$Q(1,2)$

Efecto de A al nivel 0 de B $\psi(1,0)$

Efecto de A al nivel 1 de B $\frac{\psi(1,1)}{\psi(0,1)}$

Efecto de A al nivel 2 de B $\frac{\psi(1,2)}{\psi(0,2)}$

Para que no haya interacción se requiere también que el efecto de A sea el mismo para cada nivel de B, o lo que es lo mismo que:

$$\psi(1,0) = \frac{\psi(1,1)}{\psi(0,1)} \Longrightarrow \psi(1,1) = \psi(1,0)\psi(0,1)$$

y

$$\psi(1,0) = \frac{\psi(1,2)}{\psi(0,2)} \Longrightarrow \psi(1,2) = \psi(1,0)\psi(0,2)$$

Con esto queda claro que la no interacción se expresa como $\psi(X,Y) = \psi(X,0)\psi(0,Y) \quad X = 1, \quad Y = 1,2.$

Dos modelos logísticos que pueden ajustarse en este caso son los siguientes:

a) Modelo saturado.- En este modelo se contempla la interacción de los dos factores de riesgo:

$$logit P(X,Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \gamma_{12} X_1 X_2 + \gamma_{13} X_1 X_3$$

donde [2]:

$$X_1 = \begin{cases} 1 & \text{si el sujeto está expuesto a } A \\ 0 & \text{si el sujeto no está expuesto a } A; \quad \text{esto es} \quad X_1 = X \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{si el sujeto está expuesto al nivel 1 de } B \\ 0 & \text{si el sujeto no está expuesto al nivel 1 de } B \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{si el sujeto está expuesto al nivel 2 de } B \\ 0 & \text{si el sujeto no está expuesto al nivel 2 de } B \end{cases}$$

$P(X,Y)$: Probabilidad de enfermarse dado que se está en el nivel X del factor A y en el nivel Y del factor B, $X = 0,1$ y $Y = 0,1,2$ (que implica la construcción de X_2 y X_3 como ya se mencionó).

Bajo el modelo propuesto:

$$\begin{aligned}
ln\psi(1,0) &= logitP(1,0) - logitP(0,0) = \alpha + \beta_1 - \alpha = \beta_1 \\
ln\psi(0,1) &= logitP(0,1) - logitP(0,0) = \alpha + \beta_2 - \alpha = \beta_2 \\
ln\psi(0,2) &= logitP(0,2) - logit(0,0) = \alpha + \beta_3 - \alpha = \beta_3 \\
ln\frac{\psi(1,1)}{\psi(1,0)\psi(0,1)} &= logitP(1,1) - logitP(0,0) - logitP(1,0) + logitP(0,0) \\
&\quad - logitP(0,1) + logitP(0,0) \\
&= logitP(1,1) - logitP(1,0) - logitP(0,1) + logitP(0,0) \\
&= \alpha + \beta_1 + \beta_2 + \gamma_{12} - (\alpha + \beta_1) - (\alpha + \beta_2) + \alpha = \gamma_{12} \\
ln\frac{\psi(1,2)}{\psi(1,0)\psi(0,2)} &= logitP(1,2) - logitP(1,0) - logitP(0,2) + logitP(0,0) \\
&= \alpha + \beta_1 + \beta_3 + \gamma_{13} - (\alpha + \beta_1) - (\alpha + \beta_3) + \alpha \\
&= \gamma_{13}
\end{aligned}$$

Por lo tanto en el modelo propuesto, α representa el logit de $P(0,0)$, β_1 el logaritmo del efecto relativo de A (al nivel 0 de B), β_2 el logaritmo del efecto relativo del nivel 1 de B (en ausencia de A), β_3 el logaritmo del efecto relativo del nivel 2 de B (en ausencia de A), γ_{12} el logaritmo del efecto adicional por estar expuesto a A y al nivel 1 de B conjuntamente y γ_{13} el logaritmo del efecto adicional por estar expuesto a A y al nivel 2 de B conjuntamente.

Modelo sin interacción.- Si la hipótesis de interacción se rechaza, esto conduce a plantear el ajuste de un modelo con menos parámetros:

---

[2] Por ser Y una variable que denota a una categoría, esto es $Y = 0$ ó $Y = 1$ ó $Y = 2$, para modelar, se eligen dos variables indicadoras, X_2 y X_3; así, si $Y = 0$, $(X_2, X_3) = (0,0)$; si $Y = 1$, $(X_2, X_3) = (1,0)$ y si $Y = 2$, $(X_2, X_3) = (0,1)$. En general un factor con K niveles puede representarse en un modelo con $K - 1$ variables indicadoras.

$$logitP(X,Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Este modelo impone las restricciones requeridas para que no haya interacción, es decir, que $\psi(1,1) = \psi(1,0)\psi(0,1)$ y que $\psi(1,2) = \psi(1,0)\psi(0,2)$, ya que:

$$ln\psi(0,1) = \beta_2$$
$$ln\frac{\psi(1,1)}{\psi(1,0)} = \beta_1 + \beta_2 - \beta_1 \qquad = \beta_2$$

Esto implica que $\beta_2 = ln\psi(0,1) = ln\frac{\psi(1,1)}{\psi(1,0)}$, lo cual significa que β_2 representa el logaritmo natural del efecto relativo del nivel 1 de B esté o no el factor A presente.

De manera análoga,

$$ln\psi(0,2) = \beta_3$$
$$ln\frac{\psi(1,2)}{\psi(1,0)} = \beta_1 + \beta_3 - \beta_1 = \beta_3$$

Esto implica que $\beta_3 = ln\psi(0,2) = ln\frac{\psi(1,2)}{\psi(1,0)}$, lo cual significa que β_3 representa el logaritmo natural del efecto relativo del nivel 2 de B esté o no el factor A presente y similarmente,

$$ln\psi(1,0) = \beta_1$$
$$ln\frac{\psi(1,1)}{\psi(0,1)} = \beta_1 + \beta_2 - \beta_2 = \beta_1$$
$$ln\frac{\psi(1,2)}{\psi(0,2)} = \beta_1 + \beta_3 - \beta_3 = \beta_1$$

Esto implica que $\beta_1 = ln\psi(1,0) = ln\frac{\psi(1,1)}{\psi(0,1)} = ln\frac{\psi(1,2)}{\psi(0,2)}$, lo cual significa que β_1 representa el logaritmo natural del efecto relativo de A en la respuesta, se esté o no expuesto a cualquier nivel de B.

Nótese que α sigue representando al logit de $P(0,0)$

Modelo logístico cuando se tiene un factor de riesgo dicotómico, un factor de confusión politómico y un factor respuesta dicotómico.- Serie de tablas 2×2

Supóngase que se quiere estudiar el efecto de un factor de riesgo A dicotómico: Expuesto ó no expuesto, en un factor respuesta dicotómico: enfermo ó no enfermo y además se quiere controlar por la presencia de un factor de confusión politómico C, que tiene I niveles.

En primer lugar se considera un modelo que supone que la razón de momios de enfermedad, varia de un nivel a otro del factor de confusión, es decir, que existe interacción:

$$logit P(X,Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_I X_I + \gamma_{12} X_1 X_2 + \ldots + \gamma_{1I} X_1 X_I$$

donde:

$P(X,Y)$: Probabilidad de enfermarse dado que se está en el nivel X del factor A y en el nivel Y del factor C; $X = 0,1$ y $Y = 0,1,2,\ldots,(I-1)$ (que implica la construcción de $I-1$ variables indicadoras).

$$X_1 = \begin{cases} 1 & \text{si el sujeto está expuesto a } A \\ 0 & \text{si el sujeto no está expuesto a } A; \end{cases} \quad \text{esto es,} \quad X_1 = X$$

$$X_{i+1} = \begin{cases} 1 & \text{si el sujeto está expuesto al nivel } i \text{ de } C \\ 0 & \text{si el sujeto no está expuesto al nivel } i \text{ de } C, \\ & i = 1,2,\ldots,(I-1) \end{cases}$$

En este caso β_1 representa el logaritmo del efecto relativo de A en la enfermedad al nivel 0 de C; β_i, $i = 2,\ldots I$, representa el logaritmo del efecto relativo del nivel $i-1$ de C en ausencia de A; γ_{1i}, $i = 2,\ldots,I$, representa el logaritmo natural del efecto adicional por estar expuesto a A y al nivel $i-1$ de C conjuntamente.

Entonces, bajo el modelo,

$$\psi(X,Y) = \frac{\frac{P(X,Y)}{Q(X,Y)}}{\frac{P(0,0)}{Q(0,0)}} = e^{\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_I X_I + \gamma_{12} X_1 X_2 + \ldots + \gamma_{1I} X_1 X_I}$$

Nótese que como el modelo considera interacciones, se tendrán que obtener estimadores de las razones de momios para cada nivel del factor de confusión, como por ejemplo de:

$$\frac{\psi(1,1)}{\psi(0,1)} = \frac{\frac{P(1,1)}{Q(1,1)}}{\frac{P(0,1)}{Q(0,1)}} = e^{\beta_1 + \gamma_{12}}$$

$$\frac{\psi(1,2)}{\psi(0,2)} = \frac{\frac{P(1,2)}{Q(1,2)}}{\frac{P(0,2)}{Q(0,2)}} = e^{\beta_1 + \gamma_{13}}$$

Si la hipótesis de interacción se rechaza, se puede ajustar un modelo con menos parámetros,

$$P(X,Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_I X_I$$

donde $P(X,Y)$ y $X_1 \ldots X_I$ se definen del mismo modo que en el caso anterior.

β_1 representa el logaritmo del efecto relativo de A (independientemente del nivel C al que se este expuesto), o sea que es el logaritmo natural de la medida del efecto que el factor de riesgo tiene en el factor respuesta y es por lo tanto el parámetro de interés en este caso. Similarmente, β_i, $i = 2, \ldots, I$ representa el logaritmo del efecto relativo del nivel $i-1$ de C, en el factor respuesta, esté o no A presente. Como C es un factor de confusión, $\beta_2, \ldots, \beta_I$ no son parámetros que interesen en si mismos, se les incluye en el modelo para controlar su efecto.

Modelo logístico cuando se tiene un factor de riesgo politómico con categorías ordenadas y un factor respuesta dicotómico

Supóngase que se tiene un factor de riesgo categórico con tres categorías ordenadas: 0 bajo, 1 medio, y 2 alto y un factor respuesta dicotómico: enfermo ó no enfermo.

Un modelo logístico que se puede proponer en este caso es:

$$logit P(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2$$

donde:

$P(Y)$: Probabilidad de enfermarse dado que se está en el nivel Y del factor de riesgo, $Y = 0, 1, 2$

$$X_1 = \begin{cases} 1 & \text{si el sujeto está expuesto al nivel 1 del factor de riesgo} \\ 0 & \text{si no lo está} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{si el sujeto está expuesto al nivel 2 del factor de riesgo} \\ 0 & \text{si no lo está} \end{cases}$$

Bajo el modelo propuesto

$$ln\psi(1) = \ln \frac{\frac{P(1)}{Q(1)}}{\frac{P(0)}{Q(0)}} = \beta_1 \quad \text{y} \quad ln\psi(2) = ln \frac{\frac{P(2)}{Q(2)}}{\frac{P(0)}{Q(0)}} = \beta_2$$

Esto significa que β_1 es el ln del efecto relativo al nivel base de comparación (nivel 0), del nivel uno del factor de riesgo, en la enfermedad y β_2 es el ln del efecto relativo al nivel base de comparación, del nivel dos del factor de riesgo, en la enfermedad.

Otro modelo menos general que podría ajustarse en este caso, dado que las categorías estan ordenadas y suponiendo que éstas son "equidistantes" es el siguiente:

$$logitP(Y) = \alpha + \beta Y$$

donde:

$P(Y)$: Probabilidad de enfermarse dado que se está en el nivel Y del factor de riesgo, $Y = 0, 1, 2$

Bajo este modelo

$$ln\psi(1) = \beta$$
$$ln\psi(2) = 2\beta$$

Esto significa que el logaritmo natural del efecto relativo al nivel base de comparación, del nivel dos del factor de riesgo, en la enfermedad, es dos veces el logaritmo natural del efecto relativo al nivel base de comparación, del nivel uno del factor de riesgo, en la enfermedad.

En otras palabras β representa el efecto relativo al nivel base de comparación, por cada unidad de cambio en Y.

Nótese que a diferencia del modelo anterior que es mucho más general, pues supone que cada nivel del factor de riesgo tiene un efecto relativo propio no relacionado con los demás, este modelo supone un efecto relativo constante por unidad de cambio en el factor de riesgo.

Modelo logístico cuando se tiene un factor de riesgo continuo, un factor respuesta dicotómico y un factor de confusión politómico con I niveles

Supóngase que se tiene un factor de riesgo continuo cuyo efecto en un factor respuesta dicotómico: enfermo ó no enfermo, se quiere investigar. Se tiene además un factor de confusión politómico con I niveles y se quiere controlar su efecto.

Si la hipótesis de interacción se rechaza, un modelo que podría proponerse, en el cual se asume que la dependencia del "logit" en el factor de riesgo X es βX (podría haberse propuesto $\beta_1 X + \beta_2 X^2$ o cualquier función lineal en potencias de X), es el siguiente:

$$logitP(X_i, Y) = \alpha + \beta X_i + \gamma_1 Z_1 + \gamma_2 Z_2 + \ldots + \gamma_{I-1} Z_{I-1}$$

donde:

$P(X_i, Y)$: Probabilidad de enfermarse del i-ésimo individuo con valor del factor de riesgo X_i y que presenta el nivel Y del factor de confusión, $Y = 0, 1, 2, \ldots I - 1$,

$i = 1, \ldots, n$

$$Z_j = \begin{cases} 1 & \text{si el sujeto está expuesto al nivel } j \text{ del factor de confusión,} \\ & j = 1, \ldots, I-1 \\ 0 & \text{en caso contrario} \end{cases}$$

α: representa el logit de la probabilidad de enfermarse cuando no se está expuesto al factor de riesgo y se encuentra en el nivel 0 del factor de confusión.

β: representa el ln del efecto relativo al nivel base de comparación, que es constante por cada unidad de cambio en el factor de riesgo y además como no hay interacción con el factor de confusión, es el mismo para todos los niveles del factor de confusión. Este es el parámetro de interés.

γ_j: representa el logaritmo natural del efecto relativo (al nivel base) en el factor respuesta, del nivel j-ésimo del factor de confusión, $j = 1, \ldots, I-1$. En este caso los parámetros γ_j no tienen interés en si mismos. Se les incluye en el modelo para eliminar su efecto.

Si el factor de confusión fuera continuo se puede postular una regresión lineal con él y suponiendo que no hay interacción, resulta un modelo que cumple con las suposiciones estandar de linealidad y paralelismo.

IV.3 METODOS DE ESTIMACION

En esta sección se presentan dos métodos asintóticos de estimación para los parámetros del modelo logístico. Existen desde luego métodos exactos (ver Cox (1970), cap. 4 y 5) pero como se aplican únicamente en el caso en el que se esté interesado en uno de los parámetros del modelo, considerando a los restantes como de estorbo, no se describen en este trabajo.

IV.3.1 Método de la transformación logística empírica

Este método se utiliza cuando las observaciones binarias pueden agruparse en conjuntos, de tal modo que la probabilidad de éxito sea constante para los elementos de un mismo conjunto y además se tenga un número de observaciones razonablemente grande en cada conjunto. Es decir, que este método puede utilizarse cuando se tengan factores de riesgo y confusión categóricos con un número de observaciones suficientemente grande, para cada combinación de niveles definida por ellos. Si se tienen factores continuos se pueden categorizar siempre y cuando se cumpla el requisito de que la probabilidad de éxito sea mas o menos constante dentro de cada categoría.

Cuando se tienen muchos factores de riesgo y/o de confusión es poco probable que pueda aplicarse este método, debido a que se requirirían un gran número de observaciones para que en cada combinación de niveles haya un tamaño de muestra grande.

Descripción del método

Supóngase que la combinación de los niveles de las diferentes variables explicativas (factores de riesgo y factores de confusión) da origen a g grupos y que en cada uno de ellos la probabilidad de éxito, $P_j (j = 1, \ldots, g)$ es constante.

Como se está suponiendo que la probabilidad de éxito es constante dentro de cada grupo, en lugar de trabajar con un modelo probabilistico para las respuestas individuales (Bernoulli), se puede trabajar con un modelo probabilístico para las frecuencias de éxito por grupo (Binomial).

Sea

R_j: Número de éxitos en el grupo j, $j = 1, \ldots, g$

n_j: Número de ensayos en el grupo j, $j = 1 \ldots g$

entonces $R_j \sim$ Binomial (n_j, P_j), $R_j, R_{j'}$ independientes

con $E[\frac{R_j}{n_j}] = P_j \quad V(\frac{R_j}{n_j}) = \frac{P_j(1-P_j)}{n_j}$

Supóngase que la forma en que se relacionan las variables explicativas (factores de riesgo y confusión) con la probabilidad de "éxito" (el enfermarse), se expresa en un modelo logístico lineal de la siguiente forma:

$$\lambda_j = ln\frac{P_j}{1-P_j} = \underline{X}_j\underline{\beta} = \sum_{k=1}^{p} \beta_k X_{jk} \quad \forall \quad j = 1, \ldots, g$$

El método consiste en aplicar una transformación logística a las proporciones de éxitos que se observan, es decir,

Sea $Z_j' = logit(\frac{R_j}{n_j}) = ln\frac{\frac{R_j}{n_j}}{1-\frac{R_j}{n_j}} = ln\frac{R_j}{n_j - R_j} \quad j = 1, \ldots, g$

a Z_j' se le llama la transformación logística empírica de (R_j, n_j).

Si P_j no esta muy cercana a 0 ó 1 y n_j es grande, Z_j' se distribuye aproximadamente normal. Cuando $n_j \to \infty$ su media y varianza asintóticas son:

$$E[Z_j'] = logit P_j = ln\frac{P_j}{1-P_j} \qquad j = 1, \ldots, g$$

$$V(Z_j') = \frac{1}{P_j(1-P_j)n_j} \qquad j = 1, \ldots, g$$

Un estimador consistente de esta varianza es (Cox 1970, pag. 31):

$$\hat{V}(Z_j') = \frac{1}{\frac{R_j}{n_j}(1-\frac{R_j}{n_j})n_j} = \frac{n_j}{R_j(n_j - R_j)} \qquad j = 1, \ldots, g$$

Con estas nuevas variables $Z_1', \ldots, Z_g'$ podemos expresar el modelo logístico como un modelo de regresión lineal usual, es decir,

$$Z_j' = \underline{X}_j\underline{\beta} + e_j$$

con $E[e_j] = 0$, $V(e_j) = \dfrac{1}{P_j(1-P_j)n_j}$ cuando $n_j \to \infty$ y cov $(e_j, e_j') = 0$,

Como las varianzas de los errores no son iguales y además son desconocidas, se aplica mínimos cuadrados ponderados, utilizandose procedimientos iterativos para la obtención de los estimadores de los parámetros del modelo.

Sin embargo, Cox (1970) cap. 3, propone una aproximación al procedimiento anterior, que consiste en utilizar el procedimiento de mínimos cuadrados ponderados pero utilizando como ponderaciones a $\left[\hat{V}(Z_j')\right]^{-1}$. [3] En este caso el análisis ponderado es equivalente a tratar a $\dfrac{Z_j'}{\sqrt{\hat{V}(Z_j')}}$, $j = 1, \ldots, g$, como variables cuya distribución es una normal con media $\sum_{k=1}^{p} \dfrac{\beta_k X_{jk}}{\sqrt{\hat{V}(Z_j')}}$ y varianza unitaria, si se ignoran los errores aleatorios de $\sqrt{\hat{V}(Z_j')}$ (debido a que todo el procedimiento se justifica asintóticamente cuando $n_j \to \infty \quad \forall \quad j$, la variación en los estimadores $\hat{V}(Z_j'), j = 1, \ldots g$ es un efecto de orden secundario). A partir de aquí, se pueden establecer las ecuaciones de mínimos cuadrados, pruebas de hipótesis etc., de la manera usual.

Cox (1970), propone también que si las $\hat{V}(Z_j')$'s no varian mucho, entonces se utilice como aproximación para la estimación de los parámetros del modelo, mínimos cuadrados no ponderados.

Debe observarse que si $R_j = 0$ ó $R_j = n_j$ el método de la transformación logística empírica tal y como se definió previamente no funciona.

A continuación se describen dos modificaciones a la transformación logística empírica para que el método funcione cuando $R_j = n_j$ ó $R_j = 0$.

La primera de ellas, derivada de un trabajo previo de Haldane, J. (1955) y Anscombe, F. (1956), se utiliza en el caso en el que trabaja con combinaciones de las Z_j's con pesos constantes, es decir, cuando se supone que $V(e_j) = V(e_i) \quad \forall \quad i \neq j, \quad j = 1, \ldots, g$ y es la siguiente:

Sea $Z_j = ln\left(\dfrac{R_j+1/2}{n_j-R_j+1/2}\right)$ donde 1/2 es la constante que hace que la $E[Z_j]$ sea lo más cercana a $\lambda_j = ln\dfrac{P_j}{1-P_j}$.

Se han propuesto muchos estimadores de la varianza de Z_j para muestras grandes. Gart, J. y Zweifel, J. (1967) las estudiaron y encontraron que para valores esperados, n_jP_j ó $n_j(1-P_j)$ mayores de 1.5, el siguiente estimador es "casi" insesgado:

---

[3] Esta aproximación equivale a suponer que las varianzas de los errores son conocidas e iguales a las varianzas estimadas de $Z_j' \quad j = 1, \ldots, g$

$$\hat{V}(Z_j) = \frac{(n_j+1)(n_j+2)}{n_j(R_j+1)(n_j-R_j+1)}$$

Cox (1970), menciona que la aproximación $E[Z_j] \doteq \lambda_j$ es mejor que la resultante de la transformación sin modificación, Z_j'.

Otra modificación propuesta por Cox (1970), para el caso en el que se utilizan los recíprocos de las varianzas de las variables Z_j' como pesos, en un análisis de mínimos cuadrados ponderados es la siguiente:

$$\text{para} \quad n_j > 1, \quad \text{sea} \quad Z_j^p = ln\left(\frac{R_j - 1/2}{n_j - R_j - \frac{1}{2}}\right)$$

$$\text{con} \quad \hat{V}(Z_j^p) = \frac{n_j-1}{n_j}\hat{V}(Z_j') = \frac{n_j-1}{n_j}\frac{n_j}{R_j(n_j-R_j)} = \frac{n_j-1}{R_j(n_j-R_j)}$$

Estas modificaciones surgen de la idea de sustituir en las correspondientes ecuaciones normales, a los estimadores, por otros que produzcan insesgamiento en las propias ecuaciones.

El método de la transformación logística empírica ya sea en su versión cruda o modificada se justifica asintóticamente cuando las $n_j \to \infty$. $\forall \quad j = 1,\ldots,g$. Aunque muchas aplicaciones han sugerido que el método proporciona buenos estimadores aún cuando las n_j's sean pequeñas, existe la necesidad de establecer las condiciones bajo las cuales los resultados asintóticos pueden utilizarse con seguridad, especialmente para pruebas de hipótesis y límites de confianza. Por el momento el método ha resultado adecuado siempre que ninguno o muy pocos de los grupos tengan cero éxitos o cero fracasos.

IV.3.2 Método de Máxima Verosimilitud

Mediante este método, que a continuación se describe, se pueden obtener estimadores de los parámetros del modelo logístico.

Sean $Y_1,\ldots,Y_n$ variables aleatorias independientes que se distribuyen como Bernoulli (P_i) y tales que para $i = 1,\ldots,n$, $P_i = \frac{e^{\underline{X}_i\underline{\beta}}}{1+e^{\underline{X}_i\underline{\beta}}} = \frac{e^{\sum_{k=1}^{p} X_{ik}\beta_k}}{1+e^{\sum_{k=1}^{p} X_{ik}\beta_k}}$, entonces, la función de verosimilitud esta dada por:

$$P[Y_1 = y_1, \ldots, Y_n = y_n] = \prod_{i=1}^{n} P_i^{y_i}(1-P_i)^{1-y_i} = \prod_{i=1}^{n} \frac{e^{X_i\underline{\beta}y_i}}{1+e^{X_i\underline{\beta}}} = \frac{e^{\sum_{i=1}^{n} X_i\underline{\beta}y_i}}{\prod_{i=1}^{n}(1+e^{X_i\underline{\beta}})}$$

$$= \frac{e^{\sum_{i=1}^{n}\sum_{k=1}^{p} X_{ik}\beta_k y_i}}{\prod_{i=1}^{n}(1+e^{X_i\underline{\beta}})}$$

Los estimadores de máxima verosimilitud, $\hat{\underline{\beta}}$, son aquellos que maximizan la función de verosimilitud o equivalentemente al logaritmo de dicha función, es decir, son aquellos que maximizan a:

$$L(\underline{\beta}) = \sum_{i=1}^{n} \sum_{k=1}^{p} X_{ik}\beta_k Y_i - \sum_{i=1}^{n} ln(1+e^{X_i\underline{\beta}})$$

o equivalentemente que satisfacen las ecuaciones siguientes:

$$\left[\frac{\delta L(\underline{\beta})}{\delta \beta_k}\right]_{\underline{\beta}=\hat{\underline{\beta}}} = 0$$

La solución a estas ecuaciones se encuentra usualmente mediante procedimientos iterativos como el de Newton-Raphson (Haberman S. J. (1978), cap. 5 describe este procedimiento y Cox (1970), cap. 6 propone dos soluciones aproximadas para iniciar el procedimiento iterativo).

La matriz de varianzas y covarianzas asintótica de los estimadores de máxima verosimilitud, $\hat{\underline{\beta}}$, es la inversa de la matriz de información de Fisher, es decir,

Sea $I(\underline{\beta}) = [I_{jk}(\underline{\beta})]_{k\times k}$ la matriz de información de Fisher para la muestra total,

donde $I_{jk}(\underline{\beta}) = E[-\frac{\delta^2 L(\beta)}{\delta\beta_j \delta\beta_k}] = \sum_{i=1}^{n} \frac{X_{ij}X_{ik}e^{X_i\underline{\beta}}}{(1+e^{X_i\underline{\beta}})^2}$ [4]

entonces la matriz de varianzas y covarianzas de $\hat{\underline{\beta}}$, obtenida del resultado asintótico, esta dada por: $V(\underline{\beta}) \doteq [I(\underline{\beta})]^{-1}$.

Un estimador consistente de esta matriz es $\hat{V}(\underline{\beta}) = V(\hat{\underline{\beta}})$. A partir de quí se construyen intervalos de confianza de la manera usual.

---

[4] En este caso la matriz de información de Fisher y la matriz de información de Fisher observada son iguales

IV.4 BONDAD DE AJUSTE, ELECCION DE UN MODELO Y PRUEBAS DE HIPOTESIS

Respecto a como evaluar la bondad del ajuste de un modelo logístico puede decirse, lo siguiente:

a) Si las variables explicativas del modelo son todas categóricas y la estimación de parámetros se realiza mediante el método de máxima verosimilitud, se puede realizar la prueba de bondad de ajuste de Pearson o la prueba de razón de verosimilitud.

La estadística de bondad de ajuste de Pearson se define (Breslow y Day (1980), pag. 209) de la siguiente manera: [5]

$$X^2 = \sum_{j=1}^{g} \frac{(R_j - E(\widehat{R_j}))^2}{E(\widehat{R_j})} + \sum_{j=1}^{g} \frac{([n_j - R_j] - E[\widehat{n_j - R_j}])^2}{E[\widehat{n_j - R_j}]}$$

donde

g: Número de grupos definidos por la combinación de los niveles de las diferentes variables explicativas.

R_j: Número de éxitos observados en el grupo j

$E(\widehat{R_j})$: Estimador del valor esperado de éxitos en el grupo j bajo el modelo propuesto

$n_j - R_j$: Número de fracasos observados en el grupo j.

$E[\widehat{n_j - R_j}]$: Estimador del valor esperado de fracasos en el grupo j bajo el modelo propuesto.

Si el modelo ajustado es el correcto, X^2 se distribuye asintóticamente como Ji-cuadrada con grados de libertad iguales al número de grupos g menos el número de parámetros en el modelo logístico.

La estadística de razón de verosimilitud se define de la siguiente manera: [6].

$$G^2 = -2 \quad ln \frac{\text{verosimilitud del modelo en prueba}}{\text{verosimilitud del modelo saturado}}$$

Si el modelo en prueba es correcto G^2 se distribuye asintóticamente como Ji-cuadrada con grados de libertad iguales al número de parámetros en el modelo saturado, g, menos el número de parámetros en el modelo en prueba.

Si bien estas dos pruebas proporcionan una evaluación global de que tan bien se ajustan los datos al modelo propuesto, pueden ser poco sensibles a cierto tipo específico

---

[5] Cox define otra estadística X^2 en la que considera únicamente las diferencias entre los valores observados y esperados del número de éxitos (ver Cox 1970 pag.96).

[6] A esta estadística se le conoce como devianza en el paquete GLIM

de alejamiento del modelo.

b) Si las variables explicativas son todas categóricas y la estimación de parámetros se realiza mediante el método de la transformación logística empírica, utilizando mínimos cuadrados ponderados, con pesos estimados empíricamente, se puede utilizar la suma de cuadrados de residuales (ver Cox 1970, cap. 6) para probar la bondad del ajuste, es decir,

Sea

$$Z_j^* = \frac{Z_j'}{\sqrt{\hat{V}(Z_j')}} \quad \text{para} \quad j = 1 \ldots g; \quad Z_j^* \overset{aprox}{\sim} N\left(\sum_{k=1}^{p} \frac{\beta_k X_{jk}}{\sqrt{\hat{V}(Z_j')}}, 1\right)$$

$$e_j^* = \hat{Z}_j^* - Z_j^*$$

Entonces $\sum_{j=1}^{g} e_j^{*2} \overset{aprox}{\sim} \chi^2_{(g-p)}$ si el modelo es correcto.

Se pueden también obtener los residuales estandarizados y graficarse contra variables regresoras que estén en el modelo o contra nuevas variables regresoras de interés potencial o contra los valores ajustados, para explorar patrones extraños de comportamiento lo que indicaría un alejamiento del modelo propuesto.

La distribución global de los residuales estandarizados puede indicar la presencia de observaciones aberrantes.

En general es difícil interpretar la no normalidad en la distribución, a menos que se tenga un conocimiento detallado de la cercanía a la normalidad que se puede esperar cuando el modelo es correcto.

c) Si las variables explicativas son algunas continuas o todas continuas, pero resulta ser razonable el categorizarlas, se puede utilizar alguna de las alternativas mencionadas anteriormente para evaluar la bondad del ajuste.

d) Si las variables explicativas son algunas o todas contínuas y no es posible categorizarlas, se puede hacer un análisis que aproxime al análisis de residuales usual, (Cox 1970, cap. 6) de la siguiente forma:

Sea $Y_i \sim$ Bernoulli (P_i), entonces la variable aleatoria

$$\frac{Y_i - P_i}{\sqrt{P_i(1-P_i)}} = \left(\frac{1-P_i}{P_i}\right)^{\frac{1}{2}} Y_i - \left(\frac{P_i}{1-P_i}\right)^{\frac{1}{2}} (1 - Y_i)$$

tiene media cero y varianza unitaria.

Esta variable no es observable puesto que P_i no se conoce por lo general. Sin embargo, en un modelo logístico lineal P_i es función de parámetros desconocidos $\underline{\beta}$, entonces, si $\hat{\underline{\beta}}$ es el estimador de máxima verosimilitud de $\underline{\beta}$ y $\hat{P}_i = P_i(\hat{\underline{\beta}})$, se puede definir un residual de la siguiente forma:

$$D_i = \left(\frac{1-\hat{P}_i}{\hat{P}_i}\right)^{\frac{1}{2}} \quad Y_i - \left(\frac{\hat{P}_i}{1-\hat{P}_i}\right)^{1/2} \quad (1-Y_i) = \frac{Y_i - \hat{P}_i}{\sqrt{\hat{P}_i(1-\hat{P}_i)}}$$

Las D_i's pueden graficarse contra las variables regresoras en el modelo o contra alguna variable regresora de interés potencial. La estandarización de las D_i's para tener aproximadamente media cero y varianza unitaria, permite tener cierto indicador de la dispersión que se espera, aunque la distribución de frecuencias estará típicamente alejada de la normal.

Respecto a como elegir un modelo puede decirse lo siguiente:

a) Si se tienen dos modelos a comparar, uno de los cuales es un caso particular del otro (que es más complicado) y la estimación de parámetros de los modelos se realiza mediante el método de la transformación logística empírica, se puede evaluar que tan adecuado es el modelo simple, suponiendo que el complicado ajusta bien a los datos, mediante la diferencia de la suma de cuadrados de residuales bajo los dos modelos.

   Esta estadística tiene una distribución asintótica Ji-cuadrada (con grados de libertad igual al número de parámetros en el modelo complejo menos el número de parámetros en el modelo reducido), bajo la hipótesis de que el modelo reducido ajusta bien a los datos, es decir, bajo la hipótesis de que los parámetros adicionales en el modelo complicado son cero.

b) Si se tienen dos modelos a comparar uno de los cuales es un caso particular del otro, pero la estimación de parámetros en los modelos se lleva a cabo mediante el método de máxima verosimilitud, se puede evaluar que tan bien se ajusta el modelo simple bajo la suposición de que el modelo complejo es correcto, mediante la diferencia de las estadísticas de razón de verosimilutud G^2 de ambos modelos.

   Debe observarse que la estadística de razón de verosimilitud G^2 tal y como se definió anteriormente, compara el logaritmo de la verosimilitud de un modelo contra el logaritmo de la verosimilitud del modelo saturado. Por lo tanto el obtener la diferencia de estas estadísticas para dos modelos anidados a comparar, resulta ser equivalente al (menos dos veces) logaritmo de la razón de verosimilitud del modelo simple entre la verosimilitud del modelo complejo.

Esta estadística (la diferencia de G^2 para los dos modelos) tiene una distribución asintótica Ji-cuadrada (con grados de libertad igual al número de parámetros del modelo complejo menos el número de parámetros del modelo reducido) bajo la hipótesis de que los parámetros adicionales en el modelo complicado son cero.

Es importante aclarar que las estadísticas de razón de verosimilitud (G^2) para varios modelos de un mismo caso, no son independientes y por lo tanto no es posible interpretar los niveles de significancia de la manera usual. No se puede simplemente probar la bondad del ajuste de cada modelo separadamente.

Por lo tanto, cuando no se tienen dos modelos específicos a comparar, se requiere de un método que permita seleccionar los términos a ser incluídos en el modelo por ajustar. No existe el mejor método de selección para todos los propósitos.

Uno de los métodos que se han propuesto (ver Fienberg (1978), cap. 4; Breslow y Day (1980), cap. 6) consiste en ajustar una red de modelos jerárquicos cada uno de los cuales contiene al anterior, por ejemplo se puede ajustar.

$$
\begin{aligned}
&(1) \quad \textit{logit} \quad P(\underline{X}) = \alpha \\
&(2) \quad \textit{logit} \quad P(\underline{X}) = \alpha + \beta_1 X_1 \\
&(3) \quad \textit{logit} \quad P(\underline{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2
\end{aligned}
$$

En cada etapa se lleva a cabo una prueba de significancia para los parámetros adicionales, la cual es equivalente a probar si el modelo en curso ajusta mejor que el último. Esta prueba de significancia puede llevarse a cabo obteniendo la diferencia de las estadísticas de razón de verosimilitud para los 2 modelos a comparar.

Sean G_1^2, G_2^2 y G_3^2 las estadísticas de razón de verosimilitud para los tres modelos del ejemplo.

entonces $G_1^2 \leq G_2^2 \leq G_3^2$

Cada modelo es menos restrictivo que el anterior y por lo tanto las probabilidades ajustadas estaran más cercanas a los datos observados.

$G_1^2 - G_2^2$ se utiliza para probar si el modelo (1) ajusta bien los datos, condicional a que el modelo (2) es correcto.

$G_2^2 - G_3^2$ se utiliza para probar si el modelo (2) ajusta bien los datos, condicional a que el modelo (3) es correcto, es decir, con esta estadística se evalúa la contribución de X_2 después de que se ha tomado en cuenta el efecto de X_1.

Obsérvese que la estadística de razón de verosimilitud de bondad de ajuste para el modelo más simple puede escribirse como:

$$G_1^2 = \left(G_1^2 - G_2^2\right) + \left(G_2^2 - G_3^2\right) + G_3^2$$

y cada componente tiene una distribución asintótica χ^2 con los grados de libertad correspondientes.

Entonces, para elegir un modelo, se procede ajustando del modelo más complejo al modelo más simple (en el ejemplo, del modelo (3) al modelo (1)) y se evalua la significancia de los parámetros del modelo anterior. En cada etapa se deben evaluar dos cosas.

a) El valor de la componente apropiada. En el ejemplo, $G_3^2, G_2^2 - G_3^2$ ó $G_1^2 - G_2^2$.

b) El valor acumulado de las componentes examinadas, el cual es equivalente a la estadística G^2 del modelo más simple.

Una posible regla (Fienberg (1978), pag. 50) es parar la búsqueda cuando cualquiera de estos dos valores sea significativo cuando es referido a la distribución Ji-cuadrada apropiada y escoger como el mejor modelo al del paso anterior.

Respecto a las pruebas de hipótesis sobre los parámetros del modelo logístico, éstas pueden realizarse comparando dos modelos, uno que contenga todos los parámetros y otro que no tenga los parámetros cuya significancia quiere evaluarse.

También pueden realizarse pruebas de significancia directas sobre los parámetros de un modelo específico (Breslow y Day(1980) pag. 208) de la manera usual.

IV.5 APLICACION DE LOS MODELOS LOGISTICOS A ESTUDIOS DE CASOS Y CONTROLES

En la descripción que se realizó en las secciones precedentes acerca de los modelos logísticos, las variables explicativas (factores de riesgo y confusión) se consideran fijas y la variable respuesta Y, (enfermo o no enfermo, éxito o fracaso) se considera aleatoria. Este planteamiento de los modelos logísticos se adapta a las características de un diseño de cohortes y no al de un estudio de casos y controles en donde la variable respuesta (enfermo o no enfermo) está fija por diseño y las variables explicativas que corresponden a factores de riesgo, son consideradas como aleatorias.

En esta sección se discuten dos alternativas acerca de cómo adaptar los modelos logísticos para el análisis de estudios de casos y controles. La primera de ellas (análisis prospectivo) consiste en plantear el modelo logístico para el estudio de casos y controles del mismo modo que para un estudio de cohortes, es decir, considerando a las variables explicativas como fijas y a la variable respuesta como aleatoria. La segunda alternativa (análisis retrospectivo) consiste en plantear los modelos logísticos respetando el diseño de casos y controles, en cuyo caso se considera que las variables explicativas que corresponden a factores de riesgo son aleatorias y las variables respuesta y de confusión son fijas.

IV.5.1 Análisis Prospectivo

El modelo logístico puede aplicarse a los estudios de casos y controles del mismo modo que a un estudio de cohortes, siempre y cuando se cumplan ciertas condiciones que se describen a continuación y que fueron propuestas por Mantel N. (1973)[7].

Supóngase que se tiene una población que consiste de N_1 casos y N_0 controles y que de la subpoblación de casos se seleccionan aleatoriamente n_1 de ellos con probabilidad π_1 y de la subpoblación de controles se seleccionan aleatoriamente n_0 de ellos con probabilidad π_0. Aún teniendo muestras aleatorias parece que se está seleccionando individuos con base a su variable respuesta, procedimiento de selección que invalidaría los resultados de un análisis de regresión logística normal. El porqué y bajo condiciones es posible llevar a cabo este análisis se resuelve mediante el siguiente razonamiento:

---

[7] Para una discusión más general sobre dichas condiciones consultar Breslow y Day (1980), capítulo 6.

Sea

$$Z_i = \begin{cases} 1 & \text{si el individuo } i\text{-ésimo está en la muestra} \\ 0 & \text{si no lo está} \end{cases}$$

$$Y_i = \begin{cases} 1 & \text{si el individuo } i\text{-ésimo desarrolla la enfermedad (caso)} \\ 0 & \text{si no la desarrolla} \end{cases}$$

$\underline{X}_i$: vector de variables explicativas para el individuo i

Obsérvese que $\pi_1 = P(Z = 1 \mid Y = 1)$ y $\pi_o = P(Z = 1 \mid Y = 0)$

Los posibles resultados para el individuo i-ésimo de la población son:

a) Desarrollar la enfermedad y estar en la muestra con probabilidad

$$P(Z_i = 1 \mid Y_i = 1, \underline{X}_i) P(Y_i = 1 \mid \underline{X}_i)$$

b) No desarrollar la enfermedad y estar en muestra con probabilidad

$$P(Z_i = 1 \mid Y_i = 0, \underline{X}_i) P(Y_i = 0 \mid \underline{X}_i)$$

c) Desarrollar la enfermedad y no estar en la muestra con probabilidad

$$\Big(1 - P(Z_i = 1 \mid Y_i = 1, \underline{X}_i)\Big) P(Y_i = 1 \mid \underline{X}_i)$$

d) No desarrollar la enfermedad y no estar en la muestra con probabilidad

$$\Big(1 - P(Z_i = 1 \mid Y_i = 0, \underline{X}_i)\Big) P(Y_i = 0 \mid \underline{X}_i)$$

Entonces

$$P[Y_i = 1 | Z_i = 1, \underline{X}_i] = \frac{P(Z_i = 1 | Y_i = 1, \underline{X}_i) P[Y_i = 1 | \underline{X}_i]}{P(Z_i = 1 | Y_i = 1, \underline{X}_i) P[Y_i = 1 | \underline{X}_i] + P[Z_i = 1 | Y_i = 0, \underline{X}_i] P(Y_i = 0 | \underline{X}_i)}$$

$$P[Y_i = 0 | Z_i = 1, \underline{X}_i] = \frac{P(Z_i = 1 | Y_i = 0, \underline{X}_i) P[Y_i = 0 | \underline{X}_i]}{P(Z_i = 1 | Y_i = 0, \underline{X}_i) P(Y_i = 0 | \underline{X}_i) + P[Z_i = 1 | Y_i = 1, \underline{X}_i] P[Y_i = 1 | \underline{X}_i]}$$

Si en efecto se cumple la suposición original de que los n_1 casos y los n_o controles se seleccionan de manera independiente a su condición de exposición $\underline{X}_i$, es decir, si se cumple que:

$$P\Big(Z_i = 1 | Y_i = 1, \underline{X}_i\Big) = P\Big(Z_i = 1 | Y_i = 1\Big) = \pi_1$$

y

$$P\Big(Z_i = 1 | Y_i = 0, \underline{X}_i\Big) = P\Big(Z_i = 1 | Y_i = 0\Big) = \pi_0$$

entonces

$$\frac{P[Y_i=1|Z_i=1,\underline{X}_i]}{P[Y_i=0|Z_i=1,\underline{X}_i]}=\frac{\pi_1 P[Y_i=1|\underline{X}_i]}{\pi_o P[Y_i=0|\underline{X}_i]}=\frac{\pi_1}{\pi_o}\left[\frac{P[Y_i=1|\underline{X}_i]}{P[Y_i=0|\underline{X}_i]}\right]$$

O equivalentemente

$$ln\ \ \frac{P[Y_i=1|Z_i=1,\underline{X}_i]}{P[Y_i=0|Z_i=1,\underline{X}_i]}=ln\ \ \frac{\pi_1}{\pi_o}+ln\ \ \frac{P[Y_i=1|\underline{X}_i]}{P[Y_i=0|\underline{X}_i]}$$

Esto implica que el logaritmo del momio de estar enfermo condicional a que se está en la muestra, tiene la misma dependencia en las variables explicativas que el logaritmo del momio incondicional; solo la ordenada al origen cambia. Cualquiera que sea el modelo paramétrico que se postule, se pueden obtener estimadores de los parámetros involucrados igualmente válidos a partir de un estudio de casos y controles que de un estudio de cohortes.

En particular si se tiene un modelo paramétrico de la forma

$$ln\ \ \frac{P[Y_i=1|\underline{X}_i]}{P[Y_i=0|\underline{X}_i]}=\alpha+\sum_{k=1}^{p-1}\beta_k X_{ik}$$

entonces $ln\ \ \frac{P[Y_i=1|Z_i=1,\underline{X}_i]}{P[Y_i=0|Z_i=1,\underline{X}_i]}=\alpha^*+\sum_{k=1}^{p-1}\beta_k X_{ik}$

donde $\alpha^*=\alpha+ln\ \ \frac{\pi_1}{\pi_o}$

Mantel hace notar que la selección de casos y controles no tiene que ser estrictamente independiente de la condición de exposición, sino más bien insesgada, es decir, que si la selección de casos y controles depende de la condición de exposición igualmente en ambos, el método sigue funcionando. Sin embargo, si se toma al 100% de los casos, la muestra de controles debe ser independiente de la condición de exposición.

IV.5.2 Análisis Retropectivo

La aplicación de los modelos logísticos respetando el esquema de muestreo que se utiliza en un estudio de casos y controles fué propuesta por Prentice (1976). Bajo este enfoque, la variable dependiente es un indicador dicotómico o politómico de la exposición a un factor de riesgo particular y las variables independientes (o regresoras) representan al estado de enfermedad (caso o control), a otros factores de riesgo y a factores de confusión.

Prentice plantea su proposición en el siguiente contexto:

Supóngase que se tiene un factor de riesgo dicotómico (Expuesto o no Expuesto), un factor respuesta dicotómico (Enfermo o no Enfermo) y un vector de variables explicativas para el individuo i-ésimo, $\underline{X}_i=(X_{1i},\ldots,X_{1p})$

Sea

$$D = \begin{cases} 1 & \text{si el individuo está enfermo (caso)} \\ 0 & \text{si el individuo no está enfermo (control)} \end{cases}$$

$P'(D, \underline{X}_i)$: Probabilidad de que un individuo con condición de enfermedad D y vector de variables explicativas $\underline{X}_i$, esté expuesto al factor de riesgo.

Un modelo logístico lineal para el momio de exposición puede plantearse de la siguiente forma:

$$logit P'(D, \underline{X}_i) = ln \frac{P'(D, \underline{X}_i)}{1 - P'(D, \underline{X}_i)} = \alpha + \beta D + \underline{\gamma}' \underline{X}_i$$

donde:

α: es el logit de la probabilidad de estar expuesto entre los controles, cuando no está presente ningún otro factor de riesgo o confusión.

β: es el logaritmo de la razón de momios de exposición entre casos y controles que poseen un mismo vector de variables explicativas.

$\underline{\gamma}$: representa los efectos en el momio de exposición por estar expuesto al vector de variables explicativas $\underline{X}_i$.

Obsérvese que en este modelo se ha supuesto que no existe interacción entre las variables explicativas y la condición de enfermedad.

Un modelo que incluya la interacción puede plantearse de la siguiente forma:

$$logit \quad P'(D, \underline{X}_i) = \alpha + \beta + \underline{\gamma}' \underline{X}_i + \underline{\delta}' (D \underline{X}_i)$$

Es decir, que los modelos logísticos se construyen en este caso de manera equivalente al enfoque prospectivo.

Breslow N. y Powers W. (1978) comparan el enfoque retrospectivo y prospectivo y sugieren que el propectivo es preferible para estudios que involucran muchos factores de riesgo cuantitativos.

IV.6 CONSIDERACIONES PARA EL ANALISIS DE ESTUDIOS DE CASOS Y CONTROLES CON APAREJAMIENTO INDIVIDUAL O CON ESTRATIFICACION FINA PARA EL CONTROL DE FACTORES DE CONFUSION

En la sección precedente al hablar de la aplicación de los modelos logísticos al análisis de estudios de casos y controles, se supuso que se tenían muestras aleatorias independientes y que bajo la suposición de que no hay sesgos de selección, dichos modelos podrían aplicarse de una manera prospectiva o, en caso contrario de una manera retrospectiva.

Sin embargo, cuando se pretende controlar ciertos factores de confusión ya sea a nivel diseño y/o análisis, puede resultar que el número de estratos definidos por la

combinación de niveles de dichos factores sea muy grande, resultando que se tienen muy pocas observaciones por estrato. En esta situación no es posible aplicar los modelos logísticos tal y como se ha descrito previamente, pues puede ocurrir que haya más parámetros por estimar que observaciones o si no es así, de cualquier forma se obtienen estimadores sesgados de los parámetros.

Los estudios de casos y controles con aparejamiento individual, son un caso extremo de la situación arriba descrita, ya que si por ejemplo, se realiza un aparejamiento de un caso con un control o un caso con M controles, cada par ó conjunto constituye un estrato y con muy pocas observaciones para cada uno. En este caso por lo tanto, tampoco se pueden aplicar los modelos logísticos de la forma descrita en la sección precedente.

Breslow y Day (1980,, pag. 249) presentan un ejemplo simple que a continuación se describe, en donde se evidencía el sesgo en el que se puede incurrir al aplicar la regresión logística, (que llaman incondicional) al análisis de un estudio de casos y controles con aparejamiento de un control por caso:

Supóngase que se seleccionan I casos y para cada caso se selecciona un control con valores similares en las variables consideradas de confusión. Se tienen entonces I estratos con un par por estrato.

Supóngase además que el factor de riesgo es dicotómico y que se denotará por X ($X = 0$ si el sujeto no esta expuesto; $X = 1$ si el sujeto está expuesto).

Sea Y una variable que toma el valor 0 si el sujeto es control y 1 si es caso.

Los posibles resultados de un par son los siguientes:

	Factor de riesgo											
	Expuesto	No Expuesto		Expuesto	No Expuesto		Expuesto	No Expuesto		Expuesto	No Expuesto	
Caso	1	0	1	1	0	1	0	1	1	0	1	1
Control	1	0	1	0	1	1	1	0	1	0	1	1
	2	0		1	1		1	1		0	2	

Bajo la suposición de que la razón de momios es constante en los estratos, el modelo logístico por ajustar es:

$$logit P_i(Y = 1|X) = \alpha_i + \beta X \qquad i = 1, \ldots, I$$

o equivalentemente $$P_i(Y = 1|X) = \frac{e^{\alpha_i + \beta X}}{1 + e^{\alpha_i + \beta X}}$$

La obtención de los estimadores de máxima verosimilitud de los parámetros α y β del modelo propuesto, impone la restricción de que en cada tabla 2×2 los totales marginales de las frecuencias esperadas bajo el modelo, coincidan con los totales marginales observados (ver Fienberg (1978), Everitt (1977)). Resulta entonces que para

los n_{oo} pares donde ni el caso ni el control estan expuestos y para los n_{11} pares donde ambos estan expuestos, el total marginal cero, requiere que los valores esperados y observados sean iguales.

Como ya se había mencionado en la sección 3.3 estos pares no proporcionan información acerca de la razón de momios, ya que cualquiera que sea el valor de β, se puede escoger el parámetro de ruido α_i, de tal modo que las frecuencias esperadas y observadas sean las mismas, es decir:

En una tabla donde el caso y el control están expuestos, si se elige $\alpha_i = -\beta$, las frecuencias observadas y esperadas son las mismas puesto que entonces:

$$P_i(Y = 1|X = 1) = \frac{e^{\alpha_i+\beta}}{1 + e^{\alpha_i+\beta}} = \frac{e^o}{1 + e^o} = \frac{1}{2} = P_i(Y = 0|X = 1)$$

Similarmente, en una tabla donde el caso y el control no están expuestos, si se elige $\alpha = 0$, las frecuencias observadas y esperadas son iguales ya que:

$$P_i(Y = 1|X = 0) = \frac{e^{\alpha_i}}{1 + e^{\alpha_i}} = \frac{e^o}{1 + e^o} = \frac{1}{2} = P_i(Y = 0|X = 0)$$

Las tablas que corresponden a los pares donde sólo el control está expuesto, n_{01} y a los pares donde únicamente el caso está expuesto n_{10}, tienen la misma configuración marginal, entonces los valores esperados bajo el modelo son de la siguiente forma:

Factor de Riesgo			
	Expuesto	No Expuesto	
Caso	μ	$1-\mu$	1
Control	$1-\mu$	μ	1
	1	1	2

donde:

$$\mu = P_i(Y = 1|X = 1) = \frac{e^{\alpha_i+\beta}}{1 + e^{\alpha_i+\beta}}$$

$$1 - \mu = P_i(Y = 1|X = 0) = \frac{e^{\alpha_i}}{1 + e^{\alpha_i}}$$

y entonces la razón de momios es:

$$\psi = \frac{\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}}{\frac{1}{1+e^{\alpha_i+\beta}}} \Big/ \frac{\frac{e^{\alpha_i}}{1+e^{\alpha_i}}}{\frac{1}{1+e^{\alpha_i}}} = \frac{\mu}{1-\mu} \Big/ \frac{1-\mu}{\mu} = \left(\frac{\mu}{1-\mu}\right)^2$$

Otra restricción que satisfacen las frecuencias esperadas es que el total de casos expuestos esperados debe ser igual al total de casos expuestos observados, es decir, que:

$$(n_{10} + n_{01})\mu + n_{11} = n_{10} + n_{11}$$

de donde $\hat{\mu} = \frac{n_{10}}{(n_{10}+n_{01})}$

y entonces, el estimador de la razón de momios es:

$$\hat{\psi} = \left(\frac{\hat{\mu}}{1-\hat{\mu}}\right)^2 = \left(\frac{n_{10}}{n_{01}}\right)^2$$

Resulta entonces, que este estimador es el cuadrado del obtenido bajo un modelo condicional más apropiado para este problema que se presentó en la sección 3.3.

Desde luego que este ejemplo, es uno de los más extremos en cuanto al número de observaciones por estrato, sin embargo, la presencia de sesgos en los estimadores persiste en estratificaciones finas.

Prentice y Breslow (1978) proponen aplicar el modelo de riesgos proporcionales de Cox (1972) para el análisis de estudios de casos y controles. Este enfoque tiene la ventaja de eliminar los parámetros de ruido, permitiendo estimar únicamente los parámetros de interés, por lo cual resulta ser una herramienta muy valiosa para el análisis de estudios de casos y controles con aparejamiento individual y para estudios donde se tienen pocas observaciones por estrato.

El planteamiento de Prentice y Breslow (1978) es el siguiente:

Supóngase que se tiene una población grande de individuos sanos a la cual se sigue por un período de tiempo, para determinar la incidencia de enfermedad.

Sea

T: Tiempo de entrada en la cohorte

D: El estado de enfermedad $D = \begin{cases} 0 & \text{si el individuo está sano} \\ 1 & \text{si el individuo está enfermo} \end{cases}$

$F = [F_1(t), \ldots, F_q(t)]$: q variables que pueden representar factores de riesgo o de confusión ó factores antecedentes que interactúan con los factores de riesgo.

El modelo de riesgos proporcionales supone que la tasa de incidencia de enfermedad, definida como la probabilidad instantánea de ocurrencia de la enfermedad al tiempo t, en un sujeto sano con vector de exposiciones $\underline{F} = \underline{f}$, es de la forma:

$$\lambda(t|f) = \lambda_o(t)e^{\underline{Z}'\underline{\beta}}$$

donde:

$\lambda_o(t)$: Función de incidencia desconocida, a un nivel estandar de exposiciones $\underline{f} = \underline{0}$

$\underline{Z}$: Vector de p componentes que puede incluir transformaciones e interacciones de los componentes de $\underline{f}$, así como interacciones de dichas componentes con t.

$\underline{\beta}$: vector de p parámetros desconocidos.

$e^{\underline{Z}'\underline{\beta}}$: riesgo relativo al nivel base de comparación, asociado al valor del vector de exposiciones $\underline{f}$[8]

Para aplicar este modelo al análisis de un estudio de casos y controles, conceptualícese a la muestra de casos y controles a un tiempo t, como una muestra de la cohorte prospectiva hipotética planteada originalmente.

Sea

$\theta_1 = \theta_1(t)$: La probabilidad de incluir en la muestra de casos, a un individuo que desarrolla la enfermedad al tiempo t.

$\theta_o = \theta_o(t)$: La probabilidad de incluir entre los controles, a una persona sana al tiempo t.

θ_1 y θ_o: se suponen independientes de las variables de exposición $F_1(t), \ldots, F_q(t)$.

$n_1 = n_1(t)$: Número de casos seleccionados al tiempo t, con vectores de exposición $\underline{f}_1, \ldots, \underline{f}_{n_1}$.

$n_o = n_o(t)$: Número de controles seleccionados al tiempo t, con vectores de exposición $\underline{f}_{n_1+1}, \ldots, \underline{f}_{n_1+n_o}$.

Dado que la muestra consiste de individuos con esos $n_1 + n_o$ vectores de exposición, la probabilidad condicional de que los primeros n_1 vectores correspondan a los casos tal y como se observó y el resto a los controles es:

$$\frac{\theta_1^{n_1}\theta_o^{n_o}\prod_{j=1}^{n_1}\lambda(t|\underline{f}_j)}{\sum_{l\in R(n_1,n_0)}\left\{\theta_1^{n_1}\theta_0^{n_o}\prod_{j=1}^{n_1}\lambda(t|\underline{f}_{l_j})\right\}} =$$

$$= \frac{\theta_1^{n_1}\theta_o^{n_o}\lambda_o(t)\prod_{j=1}^{n_1}e^{\underline{Z}_j'\underline{\beta}}}{\sum_{l\in R(n_1,n_o)}\left\{\theta_1^{n_1}\theta_o^{n_o}\prod_{j=1}^{n_1}\lambda_o(t)e^{\underline{Z}_{l_j}'\underline{\beta}}\right\}} = \frac{\prod_{j=1}^{n_1}e^{\underline{Z}_j'\underline{\beta}}}{\sum_{l\in R(n_1,n_o)}e^{\underline{Z}_{l_j}'\underline{\beta}}}$$

donde $R(n_1, n_o)$ es el conjunto de todos los subconjuntos de tamaño n_1 de $\{1, \ldots, n_1 + n_o\}, l = (l_1, \ldots, l_{n_1})$.

---

[8] En la sección 3.1 se discutió que el cociente de tasas de incidencia acumuladas entre dos grupos de exposición, para un período de longitud τ, aproximaba el riesgo relativo para el período, si las probabilidades de ocurrencia de enfermedad en ambos grupos eran pequeñas. Obsérvese que en este modelo, el cociente de tasas instantáneas está definido como el riesgo relativo al tiempo t, lo cual es equivalente al cociente de tasas acumuladas para un período dado. Para ilustrar esto, supóngase que sólo hay dos grupos de exposición y que no hay factores de confusión ni de interacción, en cuyo caso $F = 0(Z = 0)$ si el sujeto no está expuesto, y $F = 1(Z = 1)$ si el sujeto está expuesto, entonces:

$$\frac{\Delta_1}{\Delta_o} = \frac{\Delta(t|F=1)}{\Delta(t|F=0)} = \frac{\int_o^\tau \lambda(t|F=1)dt}{\int_o^\tau \lambda(t|F=0)dt} = \frac{\int_o^\tau \lambda_o(t)e^\beta dt}{\int_o^\tau \lambda_o(t)dt} = e^\beta = \frac{\lambda(t|f=1)}{\lambda(t|f=0)}$$

Obsérvese que la verosimilitud condicional depende únicamente de $\underline{\beta}$.

Si se toman muestras de casos y controles n_{1i} y n_{oi} a tiempos t_i $(i = 1, \ldots, I)$ la verosimilitud condicional para $\underline{\beta}$, es el siguiente producto:

$$\prod_{i=1}^{I} \frac{\prod_{j=1}^{n_{1i}} e^{\underline{Z}'_{ij}\underline{\beta}}}{\sum_{l \in R(n_{1i}, n_{oi})} e^{\underline{Z}'_{lj}\underline{\beta}}}$$

Nótese que la verosimilitud es un producto porque las muestras de casos y controles son independientes entre tiempos.

En el caso de que ciertas variables de confusión $\underline{X}$ se controlen a través de estratificación, en lugar de incorporarlas en $\underline{f}$, el modelo de riesgos proporcionales puede generalizarse a

$$\lambda(t|\underline{f}, \underline{X}) = \lambda_o(t|\underline{X}) e^{\underline{Z}'\underline{\beta}}$$

Los términos de interacción entre $\underline{f}$ y $\underline{X}$ se continúan representado en $\underline{Z}$.

Si las variables de $\underline{X}$ interactúan con el tiempo, resulta que en el estudio prospectivo, los individuos pueden cambiar de un estrato a otro a medida que el tiempo transcurre. Sin embargo, el muestreo retrospectivo en este caso, se lleva a cabo dentro de subpoblaciones que tengan valores similares en $\underline{X}$ y t. (aparejamiento individual ó por frecuencias).

Hasta el momento, la discusión de la aplicación del modelo de riesgos proporcionales al análisis de estudios de casos y controles se ha realizado en términos de una supuesta cohorte prospectiva conceptual, de la cual a diferentes tiempos se toman muestras de casos y controles.

En realidad, cuando se realiza un estudio de casos y controles se toma una muestra de la distribución condicional de $\underline{F}$ dados (T, D), y cuando se realiza un estudio de cohortes prospectivo se toma una muestra de la distribución condicional de (T, D) dada $\underline{F}$.

En la sección 3.1, se vió que para el caso en el que se tiene un factor de riesgo dicotómico (no hay otros factores y no se considera el tiempo, T) la razón de momios obtenida a partir de un estudio de casos y controles es igual a la razón de momios de un estudio de cohortes.

Una relación similar vincula la razón de momios de observar un vector de exposición $\underline{F} = \underline{f}$ cuando se muestrea al tiempo $T = t$, con la razón de momios instantánea de desarrollar la enfermedad al tiempo $T = t$, es decir:

$$\frac{P[\underline{F}=\underline{f}|(T=t,D=1),\underline{X}]/P[F=\underline{0}|(T=t,D=1),\underline{X}]}{P[\underline{F}=\underline{f}|(T=t,D=0),\underline{X}]/P[\underline{F}=\underline{0}|(T=t,D=0),\underline{X}]} = \frac{P[(T=t,D=1)|\underline{F}=\underline{f},\underline{X}]/P[(T=t,D=0)|\underline{F}=\underline{f},\underline{X}]}{P[(T=t,D=1)|\underline{F}=\underline{0},\underline{X}]/P[(T=t,D=0)|\underline{F}=\underline{0},\underline{X}]}$$

donde:

$P[F=\underline{f}|(T=t,D=j),\underline{X}]$: Probabilidad o función de densidad de probabilidad para $\underline{F}=\underline{f}$, para un individuo con estado de enfermedad j ($j=0$ sano, $j=1$ enfermo) y vector de aparejamiento o estratificación $\underline{X}$ al tiempo $T=t$.

$P[(T=t,D=j)|\underline{F}=\underline{f},\underline{X}]$: Probabilidad o función de densidad de probabilidad de que un individuo desarrolle la enfermedad ($D=1$) o no la desarrolle ($D=0$), al tiempo $T=t$ dados $\underline{F}=\underline{f}$ y $\underline{X}$.

En general se tiene que:

$$P[(T=t,D=1|\underline{F}=\underline{f},\underline{X})] = \lambda(t|\underline{F}=\underline{f},\underline{X})P[T=t,D=0|\underline{F}=\underline{f},\underline{X}]$$

entonces, bajo el modelo de riesgos proporcionales:

$$\frac{P[(T=t,D=1)|\underline{F}=\underline{f},\underline{X}]/P[T=t,D=0]|\underline{F}=\underline{f},\underline{X}]}{P[(T=t,D=1)|\underline{F}=\underline{0},\underline{X}]/P[(T=t,D=0|\underline{F}=\underline{0},\underline{X})]} = \frac{\lambda(t|\underline{F}=\underline{f},\underline{X})}{\lambda(t|\underline{F}=\underline{0},\underline{X})} = e^{\underline{Z}'\underline{\beta}}$$

Esta relación implica que:

$$logitP[\underline{F}=\underline{f}|(T=t,D=1),\underline{X}] = \underline{Z}'\underline{\beta} + logitP[\underline{F}=\underline{f}|(T=t,D=0),\underline{X}]$$

Nótese que este modelo no impone restricciones al $logitP[\underline{F}=\underline{f}|(T=t,D=0),\underline{X}]$.

Supóngase que n_1 casos con $\underline{f}_1,\ldots,\underline{f}_{n_1}$ vectores de exposición y n_o controles con $f_{n_1+1},\ldots,f_{n_1+no}$ vectores de exposición, se seleccionan al tiempo t, en un estrato definido por $\underline{X}$, entonces, la probabilidad condicional de que los vectores de exposición $\underline{f}_1,\ldots,\underline{f}_{n_1}$ correspondan a los casos, tal y como se observó, dados los n_1+n_o vectores de exposición, es:

$$\frac{\prod_{i=1}^{n_1} P[\underline{F}_i=\underline{f}_i|(T=t,D=1),\underline{X}] \prod_{i=n_1+1}^{n_1+n_o} P[\underline{F}_i=\underline{f}_i|(T=t,D=0),\underline{X}]}{\sum_{l\in R(n_1,no)} \prod_{i\in l} P[\underline{F}_i=\underline{f}_i|(T=t,D=1),\underline{X}] \prod_{i\notin l} P[\underline{F}_i=\underline{f}_i|(T=t,D=0),\underline{X}]}$$

donde:

$R(n_1, n_o)$: conjunto de todos los subconjuntos de tamaño n_1 de $\{1, \ldots, n_1 + n_o\}$

$l = (l_1, \ldots, l_{n_1})$

Como $logit P[\underline{F}_i = \underline{f}_i | (T = t, D = 1), \underline{X}] = \underline{Z}_i' \underline{\beta} + logit P[\underline{F}_i = \underline{f}_i | (T = t, D = 0), \underline{X}]$, se tiene que:

$$P[\underline{F}_i = \underline{f}_i | (T = t, D = 1), \underline{X}] = e^{\underline{Z}_i' \underline{\beta}} P[\underline{F}_i = \underline{f}_i | (T = t, D = 0), \underline{X}] \left[\frac{P[\underline{F}_i = \underline{0} | (T = t, D = 1), \underline{X}]}{P[\underline{F}_i = \underline{0} | (T = t, D = 0), \underline{X}]} \right]$$

Sea $R[F_i = 0] = \frac{P[\underline{F}_i = \underline{0} | (T = t, D = 1), \underline{X}]}{P[\underline{F}_i = \underline{0} | (T = t, D = 0), \underline{X}]}$

Entonces la verosimilitud condicional puede expresarse de la siguiente manera:

$$\frac{\prod_{i=1}^{n_1} e^{\underline{Z}_i' \underline{\beta}} P[\underline{F}_i = \underline{f}_i | (T = t, D = 0), \underline{X}] R[F_i = 0] \prod_{i=n_1+1}^{n_1+n_o} e^{\underline{Z}_i' \underline{\beta}} P[\underline{F}_i = \underline{f}_i | (T = t, D = 0), \underline{X}] R[F_i = 0]}{\sum_{l \in R(n_1, n_o)} \prod_{i \in l} e^{\underline{Z}_i' \underline{\beta}} P[\underline{F} = \underline{f}_i | (T = t, D = 0), \underline{X}] R[F_i = 0] \prod_{i \notin l} e^{\underline{Z}_i' \underline{\beta}} P[\underline{F}_i = \underline{f}_i | (T = t, D = 0), \underline{X}] R[F_i = 0]}$$

$$= \frac{e^{\sum_{i=1}^{n_1} \underline{Z}_i' \underline{\beta}} \prod_{i=1}^{n_1+n_o} P[\underline{F}_i = \underline{f}_i | (T = t, D = 0), \underline{X}] R[F_i = 0]^{n_1+n_o}}{\sum_{l \in R(n_1, n_o)} e^{\sum_{i=1}^{n_1} \underline{Z}_{li}' \underline{\beta}} \prod_{i=1}^{n_1+n_o} P[\underline{F}_{li} = \underline{f}_{li} | (T = t, D = 0), \underline{X}] R[F_{li} = 0]^{n_1+n_o}}$$

$$= \frac{e^{\sum_{i=1}^{n_1} \underline{Z}_i' \underline{\beta}}}{\sum_{l \in R(n_1, n_o)} e^{\sum_{i=1}^{n_1} \underline{Z}_{li}' \underline{\beta}}}$$

Observése entonces, que respetando el diseño de un estudio de casos y controles, la probabilidad condicional al tiempo t, en un estrato definido por $\underline{X}$, dado que se tienen $f_1, \ldots, f_{n_1+n_o}$ vectores de exposición, de que $f_1, \ldots, f_{n_1}$ correspondan a los casos y $f_{n_1+1}, \ldots, f_{n_1+n_o}$ correspondan a los controles, resultó ser igual a la verosimilitud condicional bajo el modelo hipotético en el que se conceptualiza al estudio de casos y controles como una muestra al tiempo t de una cohorte prospectiva.

Obsérvese también, que esta verosimilitud únicamente depende de $\underline{\beta}$, habiendo quedado eliminados los parámetros de ruido correspondientes a las variables que se utilizan para estratificación y que representan a factores de confusión cuyos efectos en sí mismos no son de interés para el estudio. Lo que interesa es eliminar su efecto en la estimación del efecto que los factores de riesgo tienen en el factor respuesta. Es importante recalcar, que en caso de que dichos factores interactúen con los factores de

riesgo, sus interacciones quedan incorporadas en el modelo a través del vector $\underline{Z}'$, pero los parámetros que representan sus efectos principales y que no son de interés quedan eliminados del modelo.

La verosimilitud condicional global, es decir, considerando que se tomaron muestras de casos y controles a diferentes tiempos t_i y para cada estrato definido por las combinaciones de los niveles de los factores de confusión ($\underline{X}$) (ya sea porque se llevo a cabo un aparejamiento individual o por frecuencias, o porque se estratifica a nivel análisis), es simplemente el producto de la verosimilitud condicional anteriormente descrita, a través de los diferentes tiempos y estratos.

Debe observarse que si el número de casos y controles es grande, el número de sumandos en el denominador de la verosimilitud condicional se incrementa mucho, de tal modo que el procedimiento para la obtención del estimador de máxima verosimilitud condicional de $\underline{\beta}$, resulta ser un procedimiento inoperante desde un punto de vista práctico. En estas situaciones, como los tamaños de muestra son grandes se recomienda aplicar la regresión logística usual, ya que el estimador de $\underline{\beta}$ y su error estándar resultan ser muy parecidos a los que se obtendrían a través de la verosimilitud condicional descrita en esta sección (Breslow y Day (1970), pag. 205).

En un estudio de casos y controles con aparejamiento individual de M_i controles por caso, en donde hay I de estos conjuntos, la verosimilitud condicional toma la siguiente forma:

$$\prod_{j=1}^{I} \frac{e^{\underline{Z}'_{j1}\underline{\beta}}}{\sum_{l=1}^{M_j+1} e^{\underline{Z}'_{jl}\underline{\beta}}} = \prod_{j=1}^{I} \frac{e^{\underline{Z}'_{j1}\underline{\beta}}}{e^{\underline{Z}'_{j1}\underline{\beta}} + \sum_{l=2}^{M_j+1} e^{\underline{Z}'_{jl}\underline{\beta}}} = \prod_{j=1}^{I} \frac{1}{1 + \sum_{l=2}^{M_j+1} e^{(\underline{Z}'_{jl}-\underline{Z}'_{j1})\underline{\beta}}}$$

donde:

Z_{j1}: es el vector de variables de regresión para el caso, en el conjunto (estrato) j.

Z_{jl}: es el vector de variables de regresión para el control l, en el conjunto (estrato) j.

Cuando se realiza un aparejamiento individual de un control por caso, la expresión anterior se simplifica, quedando la siguiente:

$$\prod_{j=1}^{I} \frac{1}{1 + e^{(\underline{Z}'_{j2}-\underline{Z}'_{j1})\underline{\beta}}}$$

Esta verosimilitud puede interpretarse como la verosimilitud (incondicional) para un modelo de regresión logística en donde la unidad de muestreo es el par caso-control y las variables regresoras son las diferencias en los vectores de exposición entre el control y el caso. El término constante α se supone igual a cero y cada par corresponde a un resultado positivo ($Y = 1$).

Esta correspondencia permite utilizar los paquetes que existen para regresión logística usual (incondicional) como GLIM, para ajustar el modelo condicional a estudios de casos y controles con aparejamiento individual de un control por caso.

CONCLUSIONES

Mi objetivo al realizar este trabajo y que espero haber cumplido, fué presentar un panorama general acerca de cómo se puede realizar el análisis de un estudio de casos y controles.

Por panorama general entiendo, no únicamente la presentación de las técnicas estadísticas existentes sino una sensibilización hacia los problemas metodológicos que pueden presentarse en el campo de los estudios observacionales, sin cuya consideración, me parece imposible el que se pueda realizar un análisis estadístico adecuado.

A este último respecto, un problema que considero fundamental en este tipo de estudios y que aún requiere de mucha investigación, es el que se refiere a factores de confusión.

Como hice notar en el capítulo I, la definición misma de factor de confusión presenta problemas y esto obviamente repercute en el hecho de cómo se puede evaluar si hay efectos confundidos. Digo que este problema es fundamental, porque estamos hablando de estudios observacionales, en los cuales no se utiliza ningún mecanismo de aleatorización para la asignación de tratamientos a los sujetos bajo estudio y entonces el control de factores de confusión descansa única y exclusivamente en la habilidad del investigador para definir y controlar a dichos factores. En la práctica epidemiológica es muy difícil elegir qué factores serán considerados como de confusión, pudiendo haber sesgos en el estimador del efecto que un factor de riesgo tiene en un factor respuesta, tanto, porque no se controlan ciertos factores como porque se controlan muchos, de los cuales, algunos no son en realidad de confusión. Por otro lado, es importante señalar que si se observa interacción entre un factor antecedente y el factor de riesgo, no tiene sentido el cuestionarse si dicho factor antecedente es de confusión.

No sólo el problema de factores de confusión requiere atención en este tipo de estudios. Como se mencionó en el capítulo I, existen otras fuentes de sesgos (sesgos de medición y de información) que también requieren control tanto en el diseño como en el análisis.

Otro punto importante que espero haya quedado claro en mi trabajo, es que el diseño que se plantee para la obtención de información, determinará que medida del efecto será estimable y en consecuencia si se responde o no a los objetivos de la investigación.

Por otra parte, en lo que corresponde propiamente al campo de la estadística, considero que aún hay muchas cosas por investigar y aclarar, por ejemplo cómo evaluar la bondad de ajuste de un modelo logístico. Muchos de los procedimientos presentados se justifican asintóticamente y es necesario establecer las condiciones bajo las cuales pueden utilizarse con seguridad o proponer soluciones alternativas. Este es un tema que está discutiendose actualmente en los medios de comunicación internacional, como puede observarse en el artículo de Hirji, K.F. et al (1988), en donde se discute un método para realizar inferencias exactas sobre los parámetros de un modelo de regresión logística condicional.

Por último quiero expresar, que los problemas tanto metodológicos como estadísticos que hay que enfrentar en el campo de la epidemiología no son simples, y que espero que este trabajo sea una invitación a la investigación en esta área.

BIBLIOGRAFIA

ADENA, M.A. Y WILSON, S. R. (1982), *Generalized Linear Models in Epidemiological Research.* The instat Foundation for Statistical Data Analysis.

AGRESTI, A. (1984), *Analysis of Ordinal Categorical Data.* John Wiley and Sons, New York.

ANDERSON, S., AUQUIER, A., HAUCK, W., OAKES, D., VANDAELE, W. Y WEISBERG, H. (1980). *Statistical Methods for Comparative Studies.* John Wiley and Sons, New York.

ANSCOME, F.J. (1956). *"On estimating binomial response relations"*, Biometrika, 43, pp 461-464.

ARMITAGE, P. (1955). *"Tests for linear trends in proportions and frequencies"*. Biometrics, 11, pp 375-386.

BAKER, R.J. Y NELDER, J.A. (1978). *"The GLIM System"*. The numerical Algorithms Group, N.A.G. Central Office, Oxford.

BRESLOW, N.E. Y DAY, N.E. (1980). *Statistical Methods in Cancer Research,* Vol. I, International Agency for Research on Cancer, Scientific Publications No. 32, Lyon.

BRESLOW, N. Y POWERS, W. (1978). *"Are there two logistic Regressions for Retrospective Studies?"*. Biometrics 34, pp 100-105.

COCHRAN, W.G. (1954). *"Some Methods for Strengthening the Common χ^2 tests"*. Biometrics 10, pp 417-451.

COCHRAN, W.G. (1965). *"The planning of Observational Studies of Human Populations"*. Journal of the Royal Statistical Society, Series A, 128, pp 234-266.

CONOVER, W.J. (1980). *Practical Non Parametric Statistics.* John Wiley and Sons, New York.

CORNFIELD, W.J. (1956). *"A Statistical Problem Arising from Retrospective Studies"*. Proc. Third Berkeley Symp. J. Neyman ed., Vol. IV, pp 135-148.

COX, D.R. (1970), *The Analysis of Binary Data.* Chapman and Hall, New York.

COX, D.R. (1972). *"Regressions Models and Life Tables"*. Journal of the Royal Statistical Society, Series B, 34, pp 187-220.

COX, D.R. AND HINKLEY, D.V. (1974). *Theoretical Statistics.* Chapman and Hall, London.

DONNER, A. Y HAUCK, W. (1977). *"Wald's Test as Applied to Hypotheses in Logit Analysis"*. Journal of the American Statistical Association, Vol. 72, No. 360, pp 851-853.

EFRON, B. (1975). *"The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis"*. JASA, Vol. 70, No. 352, pp 892-898.

EVERITT, B.S. (1977). *The analysis of Contingency Tables.* John Wiley and Sons, New York.

FIENBERG, S.E. (1978). *The Analysis of Cross-Classified Categorical Data.* Massachusetts Institute of Technology Press. Cambridge, Mass.

GART, J.J. (1962). *"Aproximate Confidence Limits for the Relaitve Risk".* Journal of the Royal Statistical Society, Series B, 24, pp 454-463.

GART, J.J. Y ZWEIFEL, J.R. (1967). *"On the Bias of Various estimators of the logit and its variance with application to quantal bioassay".* Biometrika, 54, 181-187.

GART, J.J. (1970). *"Point and Interval Estimation of the common odds Ratio in the combination of 2×2 Tables with Fixed Marginals".* Biometrika, 57, 3, pp 471-475.

GART, J.J. (1971). *"The comparison of proportions: A review of significance tests, confidence intervals and adjustments for Stratification".* Review of the International Statistical Institute, Vol. 39, 2, pp 148-169.

GART, J.J. Y THOMAS, D.G. (1972). *"Numerical Results on Approximate Confidence Limits for the Odds Ratio".* J. R. Stat. Soc., B, 34, pp 441-447.

HABERMAN, SHELBY J. (1978). *Analysis of Qualitative Data.* Vol. I. Academic. Press, New York.

HALDANE, J.B.S. (1955). *"The estimation and significance of the logarithm of a ratio of frequencies".* Ann. Hum. Genet., 20, pp 309-311.

HANNAN, J. Y HARKNESS, W. (1963). *"Normal Approximation to the Distribution on two independent Binomials, conditional on Fixed sum".* Ann. Math. Stat., 34, pp 1593-1595.

HIRJI, K.F., MEHTA, C.R., Y PATEL, N.R. (1988). *"Exact Inference for Matched Case-Control Studies".* Biometrics 44, pp 803-814.

KLEINBAUM, D., KUPPER, L. Y MORGENSTERN H., (1982). *Epidemiologic Research, Life Time Learning Publications.* Wadsworth, Belmont, California.

LEHMANN, E.L. (1959). *Testing Statistical Hypotheses.* John Wiley and Sons, New York.

LILIENFELD, M.D. (1976). *Foundations of Epidemiology.* Oxford University Press. New York.

MCMAHON, B. Y PUGH, T. (1978). *Principios y Métodos de Epidemiología.* La Prensa Médica Mexicana. México.

MANTEL, N. Y HAENSZEL, W. (1959). *"Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease".* J. National Cancer Institute, 22, pp 719-748.

MANTEL, N. (1973). *"Synthetic Retrospective Studies and Related Topics".* Biometrics 29, 479-486.

MANTEL, N. Y FLEISS, J. (1980). *"Minimum expected cell size requirements for the Mantel-Haenszel one degree of freedom chi-square test and a related rapid procedure"*. American Journal of Epidemiology, Vol. 112, No.1, pp 129-134.

MAUSNER, J. Y BAHN, A. (1977). *Epidemiología.* Nueva Editorial Interamericana, S. A. de C. V., México.

MC CULLAGH, P. Y NELDER, J. A.. (1983). *Generalized Linear Models.* Chapman and Hall. New York.

MÉNDEZ, I., NAMIHIRA, D., MORENO, L., SOSA, C. (1984).*El Protocolo de Investigación.* Lineamientos para su Elaboración y Análisis. Trillas. México.

MIETTINEN, OLLI, S. (1970). *"Estimation of Relative Risk from Individually Matched Series". Biometrics, 26, pp 75-86.*

MIETTINEN, O. *(1976). " Estimability and Estimation in Case-Referent Studies". Am. J. of Epidemiology, Vol. 103, No. 2, pp 226-235.*

PRENTICE, R. *(1976). "Use of the Logistic Model in Retrospective Studies". Biometrics 32, pp 599-606.*

PRENTICE, R.L. Y BRESLOW, N.E. *(1978). "Retrospective Studies and Failure Time Models". Biometrika, 68, 1, pp 153-158.*

SEIGEL DANIEL, G. Y GREENHOUSE SAMUEL, W. *(1973). "Multiple Relative Risk Functions in Case-Control Studies". American Journal of Epidemiology, Vol. 97, No. 5, pp 324-331.*

STEVENS, W.L. *(1951). "Mean and Variance of an Entry in a Contingency Table". Biometrika 38, pp 468-470.*

WOOLF, B. *(1955). "On Estimating the relation between blood group and disease". Ann. Hum. Genet., 19, pp 251-253.*