ERROR ALEATORIO Y ERROR SISTEMÁTICO

Soledad Burgos – Paulina Pino
Escuela de Salud Pública, Universidad de Chile

Los estudios epidemiológicos se orientan a objetivos cuantificables, por lo que el proceso de medición es crucial. Se busca obtener un valor estimado (estimador) de un valor "verdadero" (parámetro). Por ejemplo, la prevalencia de Diabetes Mellitus II en la población total o universo de adultos en Chile es un parámetro que posiblemente no se puede medir. Por ello, se mide - o mejor dicho se estima ese valor - a partir de una muestra de esa población. Son "estimaciones" porque están sujetas a **error**. La naturaleza y magnitud de este error tienen implicancias en la validez y la precisión de la medición y, por lo tanto, en la interpretación de los resultados. Hay dos grandes fuentes de error en las estimaciones: I. **error aleatorio** y II. **error sistemático**.

I. ERROR ALEATORIO

El error aleatorio surge de dos procesos: 1) el error muestral, ya que casi siempre estos estudios se hacen en una MUESTRA y no en todos los individuos (universo) a que se refiere el estudio. 2) el error aleatorio de la medición, proveniente del instrumento o de cambios aleatorios en el sujeto o en el observador.
En epidemiología, así como en otras disciplinas poblacionales, el error muestral es la gran fuente de error aleatorio. Éste ocurre porque la muestra que se obtiene no es idéntica a la población y por lo tanto los valores observados a partir de muchas muestras seleccionadas de idéntica forma, podrían ser *aleatoriamente* mayores, iguales o menores que el "verdadero" parámetro que se hubiera obtenido en toda la población (universo). La magnitud de este error aleatorio puede disminuirse, principalmente, aumentando el tamaño de la muestra. Así, si la muestra incluyera toda la población (universo), no habría error aleatorio. La ausencia de error aleatorio se denomina **precisión**; es decir, cuanto menos error aleatorio, se obtendrá estimaciones más precisas.

La estadística permite valorar la importancia del error aleatorio en las estimaciones, lo que históricamente se ha hecho por *prueba de significancia* o *prueba de hipótesis*. Actualmente, se prefiere un método más informativo, la *estimación del parámetro*.
Supongan que un investigador realiza una intervención (tendrá un Grupo Intervenido (GI) y un Grupo no Intervenido (GnI)), para disminuir las enfermedades gastrointestinales, en lactantes de 0 a 12 meses. Analiza sus datos y encuentra que efectivamente el GnI presenta más enfermedades gastro intestinales que el GI, pero como sabe que puede tener error muestral, decide efectuar una *prueba de hipótesis*, con el siguiente raciocinio:
1) propone una hipótesis nula (H_0), es decir, la hipótesis que niega el efecto de la intervención propuesta en el universo (la verdad): el Programa de Intervención no tiene efecto sobre las enfermedades gastrointestinales, o sea, el Grupo Intervenido (GI), tiene igual frecuencia de enfermedades gastrointestinales que el Grupo no Intervenido (GnI): GI=GnI.
2) determina un valor de probabilidad (valor p) crítico de error Tipo I, o valor α, por ejemplo, 0,05 (5%).
3) se hace la siguiente pregunta: ¿qué probabilidad hay de que se hubiera obtenido este resultado (el resultado observado), *si la hipótesis nula fuera la verdadera*? Es decir, si la VERDAD (en el universo de lactantes de 0-12 meses), fuera que el Grupo Intervenido es igual al Grupo no Intervenido respecto a las enfermedades gastrointestinales, ¿qué tan probable sería este resultado observado o un resultado aún más extremo?

4) realiza una prueba estadística adecuada al problema, (en este caso, una prueba de Ji^2 con un grado de libertad), obteniendo un valor observado para esa prueba (Ji^2 observado) y su correspondiente valor p. Por ejemplo, supongan que Ji^2=1,876 y p= 0,14321; es decir, la probabilidad de observar el valor obtenido de Ji^2 *si la hipótesis nula fuera verdadera*, sería 14%. Puesto que es >5% (el valor crítico o alfa estipulado), decide que esa probabilidad es bastante alta y que entonces, no se puede rechazar la (H_0). Otra manera de expresarlo es que la diferencia entre el GI y el GnI *no es estadísticamente significativa.* O también, es que las diferencias observadas *se deben al azar*.
Hasta ahora sólo hemos hablado de error Tipo I, lo que implica que debe haber también, a lo menos, un error Tipo II, que no aparece. Esto es efectivo. Al tomar una decisión dicotómica respecto a la existencia de una asociación a partir del estudio, ocurrirá una de las 4 situaciones ilustradas en la Figura 1, respecto a lo que ocurre en el universo (verdad).

Figura 1. Situaciones alternativas de una prueba de hipótesis a partir de los resultados en un estudio.

Estudio \ Verdad	H_0 Falsa	H_0 Correcta
H_0 Falsa	✔ (a)	Error Tipo I (b)
H_0 Correcta	Error Tipo II (c)	✔ (d)

a. El estudio concluye que se rechaza la nula, dado que la verdad es que hipótesis nula es falsa (decisión correcta): Hay asociación y el estudio correctamente la detecta.
b. El estudio concluye que se rechaza la nula, dado que la verdad es que la hipótesis nula es correcta (error tipo I)*:* no hay asociación pero el estudio incorrectamente la detecta.
c. El estudio concluye que no se rechaza la nula, pero la verdad es que que la hipótesis nula es falsa (error tipo II): Hay asociación, pero el estudio incorrectamente no la detecta.
d. El estudio concluye que no se rechaza la nula, dado que la verdad es que la hipótesis nula es correcta (decisión correcta): no hay asociación y el estudio correctamente no la detecta.

La capacidad de un estudio de rechazar una H_0 dado que la hipótesis alternativa es verdadera ((a) en la Figura 1), se conoce como PODER o potencia del estudio. Es el complemento del error Tipo II, (Poder= 1-error Tipo II). Puesto que es un componente del error aleatorio, el poder del estudio será mayor en las siguientes situaciones: 1) cuanto mayor es la muestra 2) cuanto más grande es el efecto (una asociación o una diferencia), que se quiere detectar y 3) cuanto más precisas son las mediciones. Por lo tanto estos conceptos –poder, magnitud del efecto y variabilidad aleatoria de las mediciones- son muy importantes para el cálculo apropiado del tamaño de muestra de un estudio.
Sin embargo, esta forma de decisión dicotómica en pro o contra de la existencia de una asociación origina muchas críticas, ya que se argumenta que se necesita mucho más que

tomar una decisión de si las diferencias se deben o no al azar; de forma que esta lógica de centrar el análisis en el valor p, muchas veces acaba desvirtuando el sentido de la investigación. Además es raro decidir en pro de una asociación si p=0,0499 y en contra si p=0,0501!

Debido a estas críticas, actualmente se prefiere la estimación estadística, que implica usar los datos del estudio para "estimar" el parámetro de interés. Por ejemplo, en el estudio de intervención para enfermedades gastrointestinales el parámetro de interés es el riesgo relativo (RR), una medida de asociación que compara la incidencia acumulada (riesgo) de los expuestos (Iaexp) con la incidencia acumulada de los no expuestos (IANo exp), a través de un cociente: RR=IAexp/IANo exp. El valor nulo –no hay diferencias entre los grupos- lógicamente será 1. Si el resultado es >1, los expuestos tienen mayor riesgo que los no expuestos y si es <1, significa que la exposición es un factor protector. En el ejemplo de las enfermedades gastrointestinales, si los expuestos son los del GI y si la intervención fuera efectiva esperaríamos valores <1. Por ejemplo, supongamos que RR=0,77.

Este sería *un estimador puntual*, pero todavía faltaría considerar al error muestral respecto a este único estimador (recuerden que si se obtuvieran varias muestras de igual tamaño y con el mismo método, el RR obtenido en las diferentes muestras sería aleatoriamente diferente). Para considerar el error aleatorio, se calcula un *intervalo de confianza* (IC), que son los valores entre los cuales se encontraría el estimador puntual en el universo, si se obtuvieran infinitas muestras. En este caso no hay una prueba de hipótesis dicotómica sino un *nivel de confianza* que dependerá de la variabilidad y del valor α (de error Tipo I), que el investigador está dispuesto a arriesgar. Ambos están incorporados en la fórmula que conocerán más adelante, la cual permite obtener el valor mínimo (límite inferior) y el valor máximo (límite superior) del intervalo. Un nivel de confianza de 95% (IC95%) implica que se asume un error Tipo I de 5%.

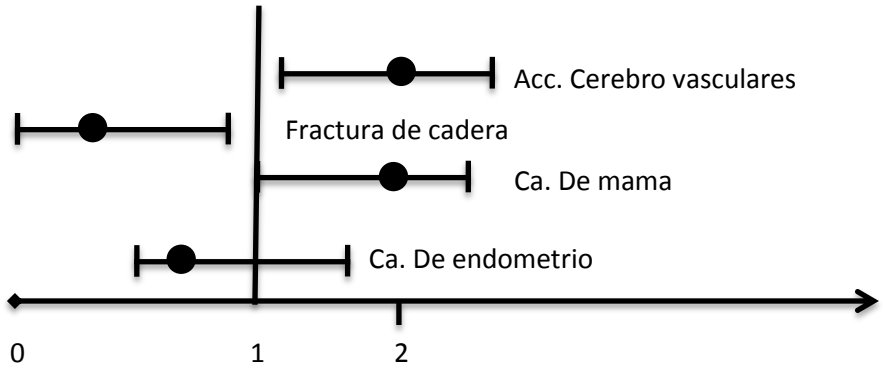
Usando el ejemplo, supongamos que hecho el cálculo, el IC95% para el estimador puntual (RR=0,77), fuera 0,72-0,83. La conclusión será que con un 95% de confianza se puede decir que el verdadero valor del parámetro está entre 0,72 y 0,83. Con eso sabemos más que con la prueba de hipótesis: 1) el sentido del efecto, que es protector; 2) la magnitud del efecto: (el GI (expuesto) tiene 30% (1/0,77), menos riesgo que el GnI de tener enfermedades gastrointestinales entre los 0 y 12 meses); 3) el valor nulo (RR=1), no está incluido entre los valores probables, con 95% de confianza; 4) con 95% de confianza se puede concluir que el parámetro (verdadero valor) se encuentra entre 0,72 y 0,83.

En suma, cuando se estudia una asociación causa-efecto, el intervalo de confianza indicará el rango en el que está la medida de asociación verdadera. Vale decir, que si un mismo estudio se repitiera con diferentes muestras de la población, en 95% de ellos los resultados estarían distribuidos alrededor de ese valor real y tan solo 5% caería fuera de ese rango, en valores aleatoriamente más altos o más bajos.

Un estudio referente a la terapia de reemplazo hormonal (TRH) en mujeres sanas y su relación con la ocurrencia de 4 desenlaces (Figura 2), reporta un **estimador puntual** de mayor riesgo de accidente cerebro-vascular (ACV) y cáncer de mama, y menor riesgo de fractura de cadera y cáncer de endometrio. Sin embargo, el IC del estimador de cáncer de endometrio recorre valores desde protectores a valores de riesgo incluido el valor 1 (no asociación). O sea, según este estudio, con un 95% de confianza no es posible afirmar que la exposición sea ni factor protector ni de riesgo de Ca de endometrio. En cambio este estudio

muestra dos efectos muy claros de la TRH: protector de fractura de cadera y de riesgo de ACV; finalmente, muestra ser un riesgo para cáncer de mama, aún cuando el Límite Inferior (LI) del IC95% coincida con el valor nulo. Nadie descartaría ese dato ya que posiblemente el efecto sería más claramente observable si se hubiera podido reducir el error aleatorio.

Figura 2. Riesgos Relativos e IC95% de cuatro posibles desenlaces en mujeres tratadas con TRH



Así como el valor *p* de la prueba de hipótesis, el intervalo de confianza también depende del tamaño muestral, siendo más estrechos -más precisos- cuanto mayor es el tamaño de muestra, y cuanto más acuciosas sean las mediciones. Por lo tanto, para reducir el error aleatorio y, por ende, aumentar la precisión se puede:

- Incrementar el tamaño de la muestra: es la estrategia más importante; por ello, la estimación del tamaño de muestra con PODER suficiente para detectar un efecto y que minimice el error Tipo I, es crucial en la planificación de un estudio.
- Mejorar las mediciones individuales*:* hay distintas estrategias, como obtener más mediciones de cada individuo, si la característica varía en el sujeto (Ej. Presión arterial), o utilizar instrumentos más precisos (ej. pasar de una balanza que mide en Kg a una que mide en gramos). Así se reduce la variabilidad aleatoria de la medición de la exposición o el efecto.

II. ERROR SISTEMÁTICO

La discusión hasta ahora, se ha referido a un tipo de error -aleatorio- el cual puede ser valorado mediante procedimientos estadísticos. **Sin embargo, eso vale SÓLO SI NO HAY ERROR SISTEMÁTICO (o es mínimo)**, así es que es muy importante entenderlo, evitarlo o, al menos "controlarlo".

El error sistemático *o SESGO,* es definido como *desviaciones del valor real que ocurren en forma sistemática,* vale decir, es un error constante en la medición que -a diferencia del error aleatorio- ocurre en la misma dirección, por ejemplo, SISTEMÁTICAMENTE mayor que el verdadero valor. Según se dijo, los estudios que minimizan el error aleatorio maximizan la PRECISIÓN; a su vez, los estudios que minimizan el error sistemático maximizan la VALIDEZ.

Los SESGOS ocurren por tres razones: 1) por fallas al momento de seleccionar a los participantes de un estudio (sesgo de selección); 2) por fallas en la obtención de las mediciones (sesgo de medición, anteriormente conocido como sesgo de información) y 3) por fallas en identificar con antelación la existencia de otras variables que *confunden* la relación entre la variable de exposición y la variable de efecto (sesgo de confusión). Como

consecuencia de estos sesgos, el estimador obtenido *tenderá (será desviado)* a ser diferente del verdadero valor (parámetro).

Sesgo de selección

Ocurre cuando las personas seleccionadas tienen una probabilidad diferente de ser incluidas en la muestra en base a las características propias de la exposición y/o el efecto a estimar. Por ejemplo, las personas que son expuestas y presentan el efecto (CASOS EXPUESTOS) tienen mayor probabilidad de ser seleccionadas que las otras categorías (CASOS NO EXPUESTOS, NO CASOS EXPUESTOS Y NO CASOS NO EXPUESTOS) (Figura 3).

Exposición	Efecto: Casos	Efecto: No Casos
Expuestos	(a)	(b)
No Expuestos	(c)	(d)

Hay muchos ejemplos en que ocurren estos desbalances. Uno, típico en estudios clínicos, es el sesgo de "vigilancia médica", que, por ejemplo, ocurriría en un estudio de caso-control del efecto de anticonceptivos orales (ACO) sobre una enfermedad con un largo período subclínico como la diabetes mellitus (DM). Por cierto, las mujeres bajo tratamiento ACO tienen posiblemente mayor número de consultas médicas que las mujeres sin ACO, así es que cualquiera enfermedad subclínica será más probablemente diagnosticada en estas mujeres que en las sin ACO. Así al comparar casos (con DM) con No Casos (Sin DM), se podrá encontrar una relación espuria (falsa) con ACO. En otro ejemplo famoso, el sesgo de "sobrevivencia", la categoría sobrerrepresentada es la de No Casos Expuestos y Casos no Expuestos. Por ejemplo, en un estudio de la relación entre exposición a contaminación aérea y asma en ciudades como Puchuncaví (Expuesta) y Con-Con (no Expuesta), las personas asmáticas podrían, por decisión propia, optar por no residir en Puchuncaví *porque no toleran la contaminación*, así es que los casos estarán más concentrados en la ciudad no Expuesta, en tanto que los No casos tenderán a estar en la ciudad expuesta en una mayor proporción. El resultado de un estudio a partir de esta selección producirá distorsiones en la asociación que se estima, sobreestimándola en el primer caso (de las mujeres con ACO) o sub estimándola en el segundo (de Con-Con y Puchuncaví).

El sesgo de selección puede ocurrir al inicio del estudio al escoger los criterios y procedimientos para la selección de los individuos y la conformación de los grupos o, en los estudios de seguimiento (cohortes o estudios experimentales), por pérdidas del seguimiento. Cuando las personas que se pierden durante el seguimiento difieren de las que continúan en el estudio por variables relacionadas con el desenlace (la enfermedad), se producirá un sesgo de selección.

Sesgo de Medición (Información)

Este sesgo incluye cualquier error sistemático en la obtención de la información sobre la exposición, la enfermedad o las variables confusoras. El sesgo de información es, por tanto, una distorsión en la estimación del efecto por *errores de medición* (o clasificación si lo que se mide es una variable dicotómica), *en la exposición o enfermedad*.

Algunos ejemplos de sesgo de información:

1. De los individuos: El sesgo de memoria suele ocurrir cuando se recolecta información de las personas retrospectivamente por lo tanto pueden no recordar detalles de la exposición o el evento. En estudios de casos y controles puede ser más crítico puesto que el hecho de caso o control puede condicionar el recuerdo de los detalles. A menudo los casos recuerdan mejor ciertos eventos pasados que los controles si éstos están vinculados a la enfermedad que padecen.

2. De los instrumentos y su aplicación: El uso de instrumentos con distinta sensibilidad/especificidad puede ocurrir para evaluar tanto la exposición como la enfermedad. En estudios de casos y controles, muchas veces no es posible utilizar los mismos instrumentos puesto que ciertos procedimientos diagnósticos no pueden ser practicados en personas sanas (Ej. Realizar un examen histológico de pulmón para determinar un tipo de cáncer). También puede ocurrir diferencias en quienes aplican los instrumentos, si los examinadores están capacitados diferencialmente para detectar una enfermedad o exposición, también pueden introducir sesgos de información.

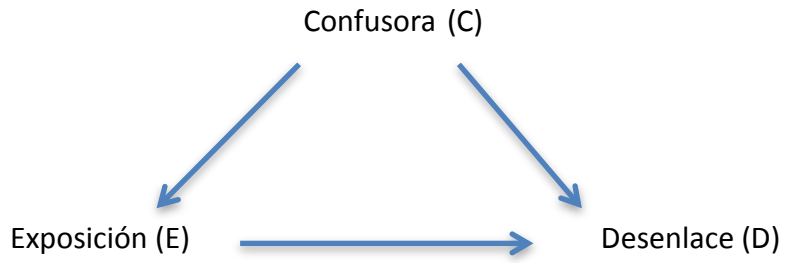
3. Uso de *proxy* o informante indirecto cuando no es posible recolectar la información sobre la exposición o el evento en la persona. Esto puede ocurrir cuando las personas han fallecido o se encuentran inhabilitadas de responder preguntas (Ej. Condición mental o extrema vejez). Se recurre entonces a informantes indirectos quienes pueden disponer información parcial sobre el evento o la historia de la persona.

Sesgo de Confusión

El sesgo de confusión se produce porque al estudiar la relación de interés se ignora la presencia de otra variable que puede incidir en el efecto de asociación que se observa. Por ejemplo, al estudiar la asociación entre obesidad y cáncer de mama, la edad es una VARIABLE DE CONFUSIÓN, ya que ésta es un determinante tanto de la aparición de cáncer de mama, como de la obesidad (a mayor edad, mayor frecuencia de obesidad y mayor frecuencia de cáncer de mama). Si no se considera, la asociación observada entre obesidad y cáncer estará sesgada por el efecto de la edad.

El problema de la confusión es crítico en la investigación epidemiológica y gran parte de los avances recientes en esta disciplina se refieren a la forma de identificar las variables confusoras y "controlarlas". Sin embargo, **la base para su identificación es el conocimiento experto del tema que se investiga**, lo que requiere estar al día en el conocimiento que se genera en el área de interés. Es decir, la lectura crítica de artículos científicos, la asistencia a reuniones, congresos y conferencias que permite conocer lo nuevo en la investigación en el área.

Una variable será confusora si y solo si, es "causa común" tanto de la exposición como del desenlace (Figura 4). En el ejemplo anterior, la obesidad es la exposición, el cáncer de mama el desenlace o respuesta y la edad es confusora porque es causa común de E y D. La hipótesis implícita -de que la exposición a obesidad aumenta el riesgo de cáncer de mama- está representada por la flecha entre E y D. Si esa flecha no existiera (o sea, si el cáncer de mama fuera independiente de la obesidad), al analizar los datos se detectaría una falsa asociación entre E y D, debido a la conexión "alterna" que existe entre E y D. Es importante observar que al ser causa común, las flechas se originan en C y se dirigen a E y D (y no de otra forma).



Por supuesto, suele haber más de una variable confusora cuando se estudia una hipótesis epidemiológica. Así es que el proceso de identificación puede ser bastante complejo.

El efecto de estas variables confusoras requerirá ser controlado de alguna forma. Como se discutirá en el capítulo de causalidad, la estrategia epidemiológica más efectiva es la realización de un estudio experimental con asignación aleatoria. Sin embargo, como se discute en ese capítulo, esa es una estrategia limitada ya que la mayor parte del trabajo epidemiológico sólo es posible con estudios observacionales.

En los estudios observacionales, la estrategia de control de confusión consiste en tratar de forzar que "en promedio", los grupos en comparación (Expuestos y No Expuestos en los estudios de seguimiento / Casos y No Casos en los estudios retrospectivos), sean iguales respecto a esa(s) variable(s) de confusión. Hay procedimientos intuitivos en el DISEÑO del estudio y más complejos en el ANÁLISIS.

Por ejemplo, en el diseño, si la variable confusora es la edad, se puede seleccionar a participantes de sólo un grupo de edad (por ejemplo, adultos de 30-45 años); si la variable confusora fuera sexo, el estudio se hará sólo en hombres o sólo en mujeres. Ese procedimiento es conocido como RESTRICCIÓN. Otra estrategia es el PAREAMIENTO de los grupos en comparación: si la variable confusora es sexo y se selecciona una mujer para el grupo de CASOS, se seleccionará también una mujer en el grupo de NO CASOS continuando el proceso con hombres y mujeres hasta completar el número necesario. Son técnicas limitadas porque no permiten controlar muchas variables y por otras desventajas más complejas.

Por lo anterior, la mayor parte de los estudios observacionales controlan el efecto de confusión en la etapa de análisis, mediante procedimientos estadísticos progresivamente más sofisticados: estandarización, estratificación y modelos multivariados. Los últimos permiten considerar el efecto confusor de las variables en forma simultánea, pero requieren un sólido conocimiento para evaluar si los modelos se adecuan o no al problema. El trabajo de equipo que incluya a profesionales estadísticos es la mejor opción, pero aún así, es necesario un conocimiento basal que permita discutir críticamente y elegir las mejores opciones analíticas.

En conclusión, este capítulo resume los grandes desafíos de un estudio observacional respecto a la preocupación básica en epidemiología de identificar y valorar los posibles errores en las estimaciones obtenidas. Estos tópicos han tenido gran desarrollo en las últimas décadas, alcanzando gran complejidad, lo que requiere cursos especializados. Pero esta base permite entender si un artículo científico de interés ha considerado los posibles errores e incluye estrategias para identificarlos, valorarlos o corregirlos en las secciones de metodología y de resultados; y de argumentar, respecto a la suficiencia de tales estrategias, en la sección de discusión del artículo.