UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA

DEPARTAMENTO DE MATEMÁTICAS

TESIS DOCTORAL

ESTIMACIÓN BOOTSTRAP CON DATOS SECUENCIALES

MARÍA DEL PINO QUINTANA MONTESDEOCA
UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA, 2003

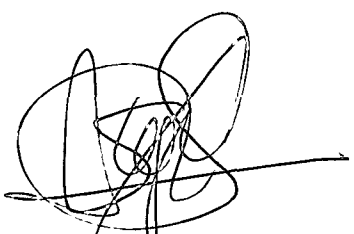
DON PEDRO SAAVEDRA SANTANA, Catedrático de Universidad del Área de Conocimiento de Estadística e Investigación Operativa del Departamento de Matemáticas de la Universidad de Las Palmas de Gran Canaria y

DON ÁNGELO SANTANA DEL PINO, Titular de Universidad del Área de Conocimiento de Estadística e Investigación Operativa del Departamento de Matemáticas de la Universidad de Las Palmas de Gran Canaria,

CERTIFICAN: Que la presente memoria titulada *ESTIMACIÓN BOOTSTRAP CON DATOS SECUENCIALES* ha sido realizada bajo la dirección de ambos por la Licenciada en Matemáticas (Especialidad de Estadística e Investigación Operativa) Doña María del Pino Quintana Montesdeoca, y constituye su Tesis para optar al grado de Doctora en Matemáticas.

Y para que así conste, en cumplimiento de la legislación vigente y a los efectos oportunos a que haya lugar, firmamos la presente en Las Palmas de Gran Canaria, a 4 de abril de dos mil tres.

FDO.: PEDRO SAAVEDRA SANTANA



FDO.: ÁNGELO SANTANA DEL PINO

FDO.: Mª DEL PINO QUINTANA MONTESDEOCA

Es mi deseo expresar al profesor Dr. D. Pedro Saavedra Santana y al profesor Dr. D. Ángelo Santana del Pino mi más sincera gratitud por la entrega, entusiasmo, dedicación, apoyo y tiempo invertido en la dirección de esta memoria. Sus valiosas aportaciones, enseñanzas, consejos y críticas así como sus palabras de ánimo han hecho posible la consecución de este trabajo.

Este agradecimiento lo hago extensivo a mis compañeros del Departamento de Matemáticas, en especial a los profesores Juan José González y Carmen Nieves Hernández, por su apoyo, ayuda y entrega en las tareas de investigación comunes. A Lucía por su ánimos de día a día, a Eduardo, a Rodri, a Nicanor y a muchos más que tengo en mi pensamiento.

Quiero dar las gracias también a mis amigos, a mis padres, a mi Juan por su apoyo y comprensión y, en especial a mi hermana Juana Teresa por haber estado a mi lado en todo momento.

A mi hermana

Índice

Prólogo

El análisis de poblaciones requiere a menudo seleccionar muestras representativas de las mismas a partir de la cuales puedan realizarse inferencias válidas para todo el conjunto poblacional. Cuando en un estudio, el número de elementos de la muestra es fijado a priori y por tanto independiente de los datos observados, se dice que el estudio es de tamaño fijo. Las técnicas aleatorias de muestreo unidas a un tamaño muestral *suficiente* permiten que la muestra represente adecuadamente a la población de referencia. Esta representatividad es necesaria para la validez de los procedimientos inferenciales.

Sin duda, la mayor parte de los procedimientos de análisis estadístico de datos está desarrollada para diseños de tamaño fijo. Es posible que esto se deba originariamente al influjo de R. Fisher, quien realizó una notable contribución a la formulación y desarrollo de los métodos clásicos de estadística inferencial y de análisis de datos. Durante varios años estuvo a cargo del diseño de experimentos y el análisis de datos en la estación agrícola experimental de Rothamsted (Londres). Allí desarrolló los modelos del análisis de la varianza en el contexto de experimentos agrícolas. La naturaleza de estos estudios era tal, que los tratamientos a comparar se asignaban de modo simultáneo a las unidades muestrales (parcelas), siendo también simultánea la recolección de datos y el análisis de los mismos. Tal característica de simultaneidad es la regla en la mayor parte de los estudios experimentales (químicos, biológicos, industriales, etc.). Los estudios observacionales, a pesar de presentar dificultades que en los experimentales pueden ser obviadas, participan generalmente de esta característica de simultaneidad. En los ensayos clínicos sin embargo, la característica de simultaneidad es más la excepción que la regla. Habitualmente, la incorporación de pacientes en los ensayos clínicos se realiza de forma lenta, siendo frecuente que haya datos disponibles para el análisis cuando sólo se ha incorporado al mismo una pequeña proporción de los previstos para la totalidad del ensayo. ¿Qué

hacer si un análisis intermedio de éstos evidencia que uno de los tratamientos es superior al resto?.

El principio ético fundamental en el ejercicio de la medicina consiste en ofertar a cada paciente el mejor tratamiento posible para su enfermedad. La aleatorización en un ensayo clínico (asignar al azar los diversos tratamientos evaluados en el curso del ensayo) podría ser éticamente dudosa salvo que haya incertidumbre acerca de cual es el *mejor tratamiento*. Un nuevo tratamiento puede ser muy prometedor, pero también podría producir efectos adversos graves que habría que determinar en el curso del estudio. Si un análisis intermedio de datos pone claramente de manifiesto la superioridad de alguno de los tratamientos comparados, los investigadores deberían detener el ensayo y ofertar a todos los pacientes el tratamiento que se haya revelado superior. Esto significa que, cuando se prevé lentitud en la incorporación de pacientes al estudio, podría ser aconsejable establecer en el protocolo del ensayo la realización de análisis intermedios (looks) con grupos de datos (generalmente del mismo tamaño) que pudieran conducir a una finalización del ensayo antes de alcanzar el tamaño muestral inicialmente previsto. Cualquier diseño de este tipo recibe el nombre de *diseño secuencial*. Parece claro que una de las ventajas es la posibilidad de concluir el estudio reduciendo el *tamaño muestral*.

En un estudio de diseño secuencial pues, el tamaño muestral es una variable aleatoria la cual recibe el nombre de *regla de parada*. En general, éste depende de los datos que sucesivamente van observándose. Habitualmente la regla de parada está determinada por un contraste de hipótesis. Así por ejemplo, en un ensayo con dos grupos paralelos, cuyo objetivo es decidir si la tasa de respuestas favorables de un tratamiento experimental es superior a un control, podríamos cada vez que se incorporan $2m$ nuevos pacientes (m por brazo de tratamiento), contrastar si alguno de los tratamientos es significativamente superior al resto. En caso de que así fuera, el ensayo se detendría. Aquí por tanto, la regla de parada está marcada por los sucesivos contrastes. Cualquier investigador familiarizado con los contrastes de hipótesis sabe perfectamente que

si se aplica reiteradamente un test con significación nominal α, la significación global del contraste es superior (o muy superior) al α fijado. Por tanto, los contrastes secuenciales requieren una metodología específica que garanticen la significación prevista. En el primer capítulo de esta memoria veremos que, para una significación y potencia especificada, el tamaño muestral esperado para un test secuencial es generalmente menor que el requerido para el contraste análogo de tamaño fijo.

La aproximación secuencial ha estado presente a lo largo de la historia de la experimentación. Los trabajos de Huyghens, Bernoulli, DeMoivre, Laplace y otros sobre sistemas de juegos pueden considerarse precursores de los métodos secuenciales. La teoría moderna de análisis secuencial se debe a A. Wald y a G. Barnard, quienes habían participado en grupos consultivos industriales para la producción y desarrollo de armamento desde 1943 y durante la Segunda Guerra Mundial. Sin duda, el trabajo más relevante fue el de Wald (1947) relacionado con el test secuencial de razón de probabilidades (SPRT). Armitage (1954, 1958, 1975) y Bross (1952, 1958) fueron pioneros en el uso de los métodos secuenciales en el campo de la medicina, especialmente en los ensayos clínicos comparativos. Sus trabajos no tuvieron inicialmente una buena aceptación, debido a que no eran prácticos para la continuación del estudio. Aquí es donde posiblemente radica la mayor dificultad de los diseños secuenciales, a saber: la validez de los datos obtenidos para realizar inferencias sobre los parámetros de los modelos.

En los diseños de tamaño fijo se busca que la muestra seleccionada sea un reflejo razonable de la población de referencia. Esto significa que las distribuciones de las variables a nivel poblacional se transmitan aproximadamente a la muestra. Así por ejemplo, si el 20% de una población es hipertensa, una muestra representativa de ésta debe garantizar que la prevalencia de hipertensión en la muestra sea aproximadamente de esa magnitud. En muchos estudios, el tamaño muestral se determina en orden a garantizar, con una cierta probabilidad, que el error máximo en la estimación de un parámetro de interés sea menor que una cota especificada por el investigador. Una muestra aleatoria, con tamaño

muestral *suficiente* es pues un reflejo aceptable de la población de referencia. De esa forma, la muestra será útil para realizar inferencias de los parámetros poblacionales de interés.

Basta lanzar una breve ojeada a los procedimientos secuenciales que se describen en el primer capítulo de esta memoria para concluir que la muestra obtenida a través de un diseño secuencial podría no ser el reflejo esperado de la población referencial. Podíamos decir que es un reflejo deformado o visto a través de un espejo cóncavo. Supóngase un procedimiento secuencial con una regla de parada consistente en detener el proceso de muestreo cuando la media de una cierta variable supere una cantidad especificada o en su defecto, cuando la muestra alcance las 100 unidades. Parece claro que la media muestral de esa variable podría fácilmente sobrestimar su media poblacional. De esa forma, la muestra no reflejaría adecuadamente a la población. En general, estimadores que son centrados en el contexto de un diseño de tamaño fijo, podrían presentar notables sesgos cuando la muestra se ha obtenido por procedimientos secuenciales. La normalidad asintótica de estadísticos de uso frecuente para tamaño muestral fijo, podría ser discutible con datos obtenidos de un muestreo secuencial. En cualquier caso, la valoración de un procedimiento por sus propiedades asintóticas, no parece muy razonable cuando el procedimiento utilizado es secuencial, toda vez que el interés fundamental de éste es reducir el tamaño muestral. En los problemas de control de calidad el objetivo consiste a menudo en decidir si en un cierto lote de ítems, la proporción de defectuosos supera o no una cierta cantidad. Para tal fin, el contraste secuencial tiene la ventaja de requerir un tamaño muestral esperado inferior al que se requiere para el de tamaño fijo. En un ensayo clínico sin embargo, el objetivo primario puede consistir en decidir si un fármaco experimental tiene acción sobre la enfermedad. Este primer objetivo puede alcanzarse mediante un contraste de grupos secuenciales. Pero aquí se impone al menos cuantificar la superioridad del tratamiento experimental frente al control, pues esto será esencialmente lo que marque la relevancia clínica del posible hallazgo. ¿Cómo realizar inferencias o construir regiones de confianza para los parámetros de interés con datos

procedentes de diseños secuenciales, posiblemente sesgados?. P. O'Brien señalaba hacia finales de los años 70 que una de las razones del escaso uso de los diseños secuenciales radica en las frecuentes dudas de la comunidad científica acerca de la eficiencia de los procedimientos secuenciales.

La introducción del bootstrap por B. Efron en 1979 marcó un importante hito en el desarrollo de los métodos de análisis de datos. Estos procedimientos, llamados también de computación intensiva, permitieron aproximar la distribución de probabilidad de muchos estadísticos cuya obtención por procedimientos clásicos era muy compleja. En muchos casos además, las aproximaciones bootstrap mejoraban las clásicas aproximaciones por la distribución normal.

En el contexto del diseño secuencial, la construcción de intervalos de confianza se basa a menudo en la dualidad test de hipótesis-intervalo de confianza. Emerson y Fleming (1990) consideran el caso en el que los datos observados están normalmente distribuidos con varianza conocida y determinan un procedimiento basado en la referida dualidad para construir intervalos de confianza para la media de la distribución. Este procedimiento no es precisamente de fácil implementación, pero su mayor defecto es su limitación. Mostramos en esta memoria que cuando los datos se apartan notablemente de las referidas hipótesis, los errores de cobertura pueden llegar a ser considerables, sobre todo, cuando los tamaños muestrales son pequeños.

Pensando en la importancia que los análisis intermedios pueden tener en muchos ensayos clínicos y en la realización de inferencias acerca de los parámetros de interés a través de los subsiguientes datos secuenciales, nos planteamos la viabilidad de las aproximaciones bootstrap a la distribución de probabilidad de algunos pivotales clásicos. Tal como esperábamos, las aproximaciones clásicas eran radicalmente inadecuadas. Basándonos en contrastes de la familia de Wang y Tsiatis, exploramos en primer lugar las propiedades de la función de distribución empírica estimada a través de datos secuenciales. Como

cabía esperar, cuando los datos satisfacen la hipótesis nula del contraste, el tamaño muestral tiene una distribución con muy poca variabilidad, por lo cual, las propiedades del estimador son prácticamente similares a las que verifica en diseños de tamaño fijo. Sin embargo, el estimador muestra evidentes sesgos cuando la hipótesis nula falla. En relación con la distribución empírica damos un teorema de consistencia basado en el teorema de Anscombe (1952) y Doeblin (1938). Damos asimismo una cota para el sesgo del estimador. De este resultado se deduce un hecho obvio, a saber: si la varianza de la regla de parada es cero, el estimador es centrado. A lo largo de toda la memoria hemos utilizado la distribución empírica como distribución de remuestreo.

Hemos considerado en esta memoria como parámetros de interés la media de una distribución, la diferencia de proporciones y el riesgo relativo. Para su estimación por intervalos de confianza hemos considerado los pivotales clásicos cuya distribución de probabilidad hemos aproximado a través del bootstrap. Su validez no es desde luego evidente. En el procedimiento bootstrap se requiere definir una regla de parada. Hemos propuesto la misma que para el muestreo de la distribución original, con la diferencia de que el remuestreo se realiza de la distribución empírica. Debido precisamente a los sesgos de ésta, la distribución de la regla de parada bootstrap no imitará en general la real. En los contrastes de la familia de Wang y Tsiatis, cuando los datos se distancian de la hipótesis nula el procedimiento tiende a detenerse. Si la distribución empírica sobrestima la real y como consecuencia de esto conduce a que los datos extraídos de la empírica se apartan más aún de la hipótesis nula, el procedimiento secuencial bootstrap tenderá a parar antes. A pesar de esto sin embargo, la aproximación bootstrap es válida. La razón de esto está en una cadena de compensaciones que analizamos en el capítulo tercero. Diversas simulaciones confirman la validez de las aproximaciones propuestas.

Hemos estructurado la memoria de la siguiente forma. En el primer capítulo se realiza una revisión de los procedimientos secuenciales clásicos, particularmente de los procedimientos de Wang y Tsiatis sobre los que se han

realizado los estudios de simulación. En el capítulo segundo se revisan algunos procedimientos bootstrap para diseños de tamaño fijo. Se consideran también algunas cuestiones acerca de los desarrollos de Edgeworth a través de los cuales se muestra como las aproximaciones bootstrap pueden mejorar las normales. Particularmente interesante es el estudio de la consistencia de los procedimientos bootstrap basada en la métrica de Mallows. Este aspecto se resume en la última sección del segundo capítulo. En el tercer capítulo, tras una revisión del procedimiento de construcción de intervalos de confianza basado en la dualidad test de hipótesis e intervalos de confianza, se analizan las propiedades de la distribución empírica basada en datos secuenciales como estimador de la distribución real. Se introducen seguidamente las aproximaciones bootstrap a las que nos referimos en el párrafo anterior. El análisis de la validez del procedimiento lo hacemos a través de una descomposición del pivotal correspondiente. El cuarto capítulo aplica los procedimientos secuenciales y los métodos de construcción de intervalos de confianza a tres estudios biomédicos. El primero es un ensayo clínico cuyo objetivo es evaluar la efectividad del recubrimiento con antibióticos de los catéteres utilizados por pacientes críticos en orden a disminuir la tasa de colonización por agentes microbianos. Se trata de un ensayo con dos grupos paralelos en el que 139 pacientes fueron aleatoriamente asignados a recibir un catéter impregnado con antibiótico y 145 a un control. De los datos de este ensayo hemos realizado una simulación seleccionado un máximo de 7 looks de tamaño 40 (20 pacientes en cada grupo). La variable principal de valoración fue la indicatriz de colonización. El segundo estudio tenía como objetivo evaluar los genotipos del grupo HLA como predictores de la diabetes (tipo 2). El estudio original se diseñó como de caso-control estratificándose los participantes según fueran o no diabéticos. Los genotipos se clasificaron como de alto y bajo riesgo. En cada grupo se determinó pues la variable binaria indicatriz del grupo HLA. Se realizó una simulación de un estudio secuencial a partir del estudio original en el que 153 personas eran diabéticas y 121 eran controles. En la simulación se consideró un máximo de 6 looks de tamaño 40 cada uno (20 por grupo). En el tercer estudio se evaluó un marcador basado en ultrasonidos (qui-Stifness) como predictor de la osteoporosis. Este último estudio es también de

caso-control con un total de 340 mujeres, de las cuales 149 habían sido diagnosticadas con osteoporosis y 191 se les consideró libres de enfermedad. La variable de valoración fue la referida determinación *qui-stifness*. En la simulación realizada a partir de los datos reales se consideraron también looks de tamaño 40 (20 en cada grupo) siendo el máximo de inspecciones previsto de 7.

1. Contrastes secuenciales

1.1 Introducción.

En el contexto general de la investigación científica la mayor parte de los procedimientos de análisis estadístico de datos está desarrollada para diseños de tamaño fijo, lo que significa que la toma de datos debe llegar hasta el final antes de poder tomar decisiones. Sin embargo, puede ocurrir que los datos disponibles en algún momento, antes de terminar el estudio, señalen que las suposiciones realizadas en el diseño del mismo son falsas y deben replantearse; o bien permiten detectar situaciones que no se tuvieron en cuenta al diseñar éste; o incluso que se llega a disponer de información suficiente para tomar una decisión sin tener que alcanzar el tamaño de muestra previamente establecido. Ello podría motivar en el investigador dilemas éticos, económicos, etcétera, ya que no se justifica seguir con la toma de datos en vista de las evidencias detectadas. Dado que en esta situación no se pueden aplicar las técnicas estadísticas desarrolladas para diseños de tamaño fijo, resulta preciso emplear métodos de análisis de datos o de grupos de datos que tengan en cuenta la naturaleza secuencial del problema.

En este sentido, los procedimientos secuenciales, donde el tamaño de muestra se considera aleatorio, posibilitan que si en un determinado momento de un proceso de recogida de datos se dispone de información suficiente para tomar decisiones respecto a la cuestión o cuestiones que motivaron dicho proceso, éste se detenga y se tome la decisión más conveniente, conservando los niveles de significación y potencia establecidos al inicio de estudio.

En este capítulo describiremos los aspectos más importantes de los procedimientos estadísticos secuenciales: el establecimiento de las *hipótesis* a contrastar, esto es, la cuestión sobre la que se debe decidir; los *instantes de inspección* en los que se lleva a cabo el análisis de los datos disponibles; y la *regla*

de parada que indica, tras cada inspección, si se puede tomar ya una decisión y por tanto se puede terminar el proceso, o si por el contrario, se debe continuar reuniendo información.

Existen numerosas e importantes razones para recurrir al proceso de toma de datos y análisis de resultados de manera secuencial. Un claro ejemplo de ello son las *consideraciones éticas* que se suscitan en el contexto de los ensayos clínicos destinados a la comparación de tratamientos alternativos para una determinada enfermedad. En estos ensayos, frecuentemente, los pacientes se incluyen en el estudio a lo largo de un periodo de tiempo tan dilatado, que se llega a disponer de resultados de los primeros pacientes cuando aún no se ha completado el tamaño muestral previsto para todo el ensayo. Tales datos podrían aportar suficiente información para confirmar la superioridad de uno de los tratamientos evaluados, por lo que no sería éticamente justificable asignar a ninguno de los siguientes pacientes al tratamiento que se ha revelado como el menos eficaz.

Los procedimientos secuenciales resultan también de interés desde el punto de vista *económico*. Así ocurre, por ejemplo, en estudios de control de calidad cuando la inspección supone la destrucción de la unidad muestral, en cuyo caso interesa finalizar el muestreo lo antes posible, sobre todo, si es evidente que el lote va a ser aceptado. En este sentido, es de destacar que el análisis secuencial permite reducir el número medio de datos requeridos en la realización del contraste respecto al diseño de tamaño fijo, para un mismo nivel de significación y una potencia dada para una alternativa especificada.

Otro aspecto de interés en el uso de procedimientos secuenciales es el *administrativo*, dado que el realizar análisis intermedios permite asegurar que el experimento se está ejecutando como se había planificado. Además, la pronta disposición y análisis de los primeros resultados del experimento permiten comprobar las suposiciones realizadas al diseñar el ensayo o experimento, lo que permite advertir y, en su caso, corregir aspectos que no se tuvieron en cuenta al iniciar el mismo.

Los métodos habituales de la inferencia estadística, contrastes de hipótesis y estimación de parámetros, desarrollados para diseños con tamaño de muestra fijo no resultan adecuados en un procedimiento secuencial en el que se realizan exámenes sucesivos de los datos acumulados y existe la posibilidad de la finalización anticipada del mismo, o de algún cambio en su diseño original. A las dificultades que ya *per se* presenta la realización reiterada de contrastes de significación en los sucesivos instantes de inspección, se une la presencia de reglas de parada que dependen de los datos observados y que dan lugar a que los tamaños muestrales de los diseños secuenciales sean aleatorios, lo que modifica las propiedades de los estimadores, en particular introduciendo sesgos en los mismos. El uso en este contexto de métodos pensados para diseños de tamaño fijo conduce a una mala evaluación de los niveles de significación y potencia y por tanto a una incorrecta interpretación de los resultados.

Los métodos estadísticos concebidos específicamente para tener en cuenta las inspecciones sucesivas de los datos a lo largo de un estudio se engloban bajo el término *análisis secuencial*. Algunos autores reservan este término para estudios en que se realiza un análisis de datos (*análisis intermedio*) cada vez que se dispone de una nueva observación, y utilizan el término *análisis secuencial en grupos* cuando los análisis intermedios se realizan tras cada grupo de m observaciones ($m>1$).

1.2. Desarrollo histórico de los métodos secuenciales.

La teoría clásica del diseño experimental trata básicamente con experimentos cuyo tamaño de muestra es fijo, posiblemente porque los pioneros en la materia, particularmente R. A. Fisher, trabajaron en la investigación agrícola, donde los resultados de un ensayo se obtenían bastante tiempo después de que el experimento se hubiera diseñado e iniciado. En este sentido, resulta

interesante preguntarse cómo habría evolucionado la estadística teórica si Fisher hubiera trabajado en investigación médica o industrial.

Gran parte de la metodología estadística utilizada en el diseño y análisis de *ensayos clínicos*, tiene su origen en principios desarrollados en el contexto de ensayos agrícolas durante los años veinte. Una de las diferencias importantes entre éstos y los ensayos clínicos se encuentra en la naturaleza de la acumulación de los datos. En el campo de la agricultura un ensayo se rige por el patrón natural de las estaciones. El ensayo se diseña, los datos muestran la evolución del crecimiento de los cultivos, que son recolectados simultáneamente, con lo cual en el momento de realizar el análisis ya están disponibles todos los datos pertinentes. Si resulta insuficiente la información para aportar una clara conclusión, entonces se planea un nuevo experimento para la próxima estación. Por el contrario, los datos de un ensayo clínico se acumulan gradualmente durante un periodo que puede llegar a durar meses o años, con lo que los resultados de los primeros pacientes incluidos para el estudio pueden estar ya disponibles para la interpretación mientras aún se continúa incluyendo pacientes en el estudio y asignándoles aleatoriamente uno de los tratamientos.

Otro campo en el que el análisis secuencial ha jugado un papel importante es el *control estadístico de la calidad,* cuyo inicio data de 1924 cuando ingenieros norteamericanos de la Bell Telephone se ocuparon de diseñar un procedimiento para detectar unidades defectuosas en el proceso industrial. El objetivo inicial del control estadístico de la calidad fue la detección de estas unidades, bien para su desecho bien para su reciclaje pero, en cualquier caso, para impedir que una unidad mal fabricada pudiera incidir negativamente en la fabricación de otras piezas, si era una unidad componente, o en el grado de satisfacción del cliente, si se trataba de un producto final. En este sentido, una pieza defectuosa representaba para la empresa un coste adicional que necesitaba ser controlado. Para ello, se procedía mediante un plan de inspección con muestreo secuencial, basado en el Test Secuencial de Wald (1947) cuyo tamaño de muestra, en principio una variable aleatoria, quedaba fijado bien cuando el lote se aceptase, bien cuando

fuese rechazado. Con el tiempo, el control de calidad ha dejado de ser exclusivo de los procesos industriales, inundando el mundo de los servicios. En este sentido, instituciones financieras, universidades, hospitales, etc., han integrado este tipo de control como un medio más que permita contribuir a garantizar un nivel de calidad de manera continua.

La teoría moderna del análisis secuencial surge con los trabajos de Abraham Wald (1947) en Estados Unidos y de George Barnand (1946) en Gran Bretaña, quienes participaron como asesores en grupos de investigación industrial para el desarrollo y producción de armas, desde 1943, y durante la Segunda Guerra Mundial. En particular, el trabajo desarrollado por Wald (1947) sobre el Test de Razón de Verosimilitud Secuencial se convirtió en la herramienta de referencia para el posterior desarrollo de los procedimientos secuenciales de análisis de datos.

1.3. Visión general de los procedimientos secuenciales.

Si bien, como hemos visto, son muy variados los campos de investigación en que los métodos estadísticos secuenciales resultan de aplicación inmediata vamos a centrarnos, para fijar ideas, en el campo de los ensayos clínicos. En este contexto, uno de los problemas más usuales es el de comparar un nuevo tratamiento para una enfermedad (tratamiento experimental) con un tratamiento preexistente (control). Normalmente esta comparación se concreta en la construcción de un modelo estadístico de las respuestas de los pacientes, dependiente de un parámetro de interés θ que mide la diferencia en eficacia entre ambos tratamientos. El problema fundamental puede plantearse como un contraste de la hipótesis nula $H_0 : \theta = 0$ de que ambos tratamientos son equivalentes frente a la alternativa de que existen diferencias entre ambos, $H_1 : \theta \neq 0$. Frecuentemente, la hipótesis alternativa que se desea poner a prueba es $H_1 : \theta > 0$, que el tratamiento experimental sea superior al control.

Los métodos de análisis secuencial se basan en el uso de dos estadísticos, que denominaremos Z_k e I_k que miden, respectivamente, la *ventaja* del tratamiento experimental sobre el control en el k-ésimo instante de inspección, y la *cantidad de información* sobre el parámetro θ disponible en ese momento. En cada instante de inspección, la regla para decidir si H_0 debe aceptarse o rechazarse (y por tanto parar el procedimiento), o bien si debe continuarse con la toma de datos se construye siguiendo alguna de las dos siguientes aproximaciones:

- *Métodos basados en el establecimiento de barreras en el plano* (Z,I): los sucesivos valores de los estadísticos Z_k e I_k se van representando en el plano (Z,I); mientras se mantengan dentro de ciertas fronteras predefinidas, la toma de datos continúa. Cuando se alcance alguna de las barreras, el procedimiento se detiene, decidiendo por H_0 o H_1 según cuál sea la frontera rebasada. Estos métodos, entre los que cabe citar el procedimiento secuencial abierto, el restringido y el test triangular de Whitehead se derivan del test de razón de verosimilitudes secuencial original de Wald (1947).

- *Métodos basados en la realización repetida de contrastes de significación.* Según que la cantidad de información (número de datos) disponible en cada inspección pueda prefijarse o no de antemano podemos agrupar estos métodos en dos categorías:

  - *Métodos con niveles de información predefinidos*: estos métodos predeterminan el número máximo K de inspecciones que se van a realizar a los datos, así como el número m de datos que se debe observar cada vez. En la k-ésima inspección se realiza un contraste de significación con un nivel de significación nominal α_k. El haber fijado de antemano el valor de K permite elegir los valores de los α_k de forma que se garantice un nivel de significación α global, también predeterminado. El valor de m se elige para garantizar una potencia especificada $1-\beta$ para alguna alternativa de interés $\theta=\delta$. Existen muchas formas de elegir los α_k, lo que da lugar a numerosos

contrastes en esta categoría, entre los que podemos citar el de Pocock, O'Brien & Fleming y Wang & Tsiatis,

- *Métodos con niveles de información arbitrarios*: dada la naturaleza de muchos ensayos clínicos (y de otros procedimientos de muestreo secuenciales) a veces es difícil, o imposible, *a priori* determinar cuál va a ser el número máximo de inspecciones que se va a poder realizar; o cuáles van a ser los números de datos disponibles en las sucesivas inspecciones. Cuando el nivel de información que va a estar disponible en cada inspección es impredecible, los métodos secuenciales más apropiados son los basados en las *funciones de gasto del error tipo I*. Suponiendo que $I_{\max}$ es la cantidad máxima de información que eventualmente se podría llegar a alcanzar durante el desarrollo del procedimiento secuencial, la función de gasto $\alpha(t)$ representa el nivel de significación que se alcanzaría si el método secuencial terminase cuando la cantidad total de información acumulada a lo largo del mismo es una proporción t de $I_{\max.}$. Si α es el nivel de significación que se desea alcanzar en el ensayo, la función $\alpha(t)$ debe ser creciente desde 0 en t=0 hasta α en t=1. De esta forma, $\alpha(t)$ es el nivel de significación que se emplea (se *gasta*) en una inspección en que se ha conseguido reunir una proporción t de la información máxima posible para todo el procedimiento. Una vez que el ensayo se pone en marcha, en la k-ésima inspección se calculan t y el correspondiente nivel de significación $\alpha(t)$. Son posibles muchos perfiles para esta función. Así, por ejemplo, dado que cuanto más pequeño sea el valor de $\alpha(t)$ más difícil será rechazar H_0, funciones de gasto que tomen valores muy pequeños cuando t es pequeño hacen muy difícil rechazar H_0 cuando hay poca información.

Cada uno de estos métodos tiene sus ventajas e inconvenientes. Con las convenientes adaptaciones permiten contrastar hipótesis bilaterales o unilaterales. Cuando la hipótesis nula es falsa, el número de inspecciones requeridas hasta rechazarla es menor que en un diseño con muestra de tamaño fijo. A su vez,

existen también versiones de estos contrastes que permiten aceptar relativamente pronto la hipótesis nula en caso de ser cierta. En general, en los contrastes secuenciales, y esta es su principal ventaja, los tamaños de muestra hasta la terminación del procedimiento (y la consecuente aceptación o rechazo de H_0) son, en media, inferiores al tamaño de muestra que se necesitaría en un diseño de tamaño fijo con un nivel de significación y potencia equivalentes. Como contrapartida, el tamaño de muestra máximo puede ser superior al del diseño de tamaño fijo, por lo que habrá de valorarse adecuadamente la posibilidad de que el test secuencial proceda hasta alcanzar su tamaño máximo. En cada caso, el investigador habrá de evaluar su problema particular (con los condicionantes externos al mismo en cuanto a la facilidad con que se pueden o no alcanzar los sucesivos tamaños muestrales en cada inspección, y los momentos en que éstas pueden llevarse a cabo), y los pros y contras de cada uno de estos métodos para elegir el más adecuado a su caso.

1.4. Test de Razón de Verosimilitud Secuencial. Identidad de Wald.

Fue Abraham Wald quien introdujo la expresión *análisis secuencial* hacia principios de los años cuarenta. En sus trabajos describe una nueva técnica para contrastar una hipótesis nula simple frente a una alternativa simple, consistente en tomar unidades, una tras otra, de una muestra de tamaño aleatorio y (utilizando la razón de verosimilitudes) decidir si la información disponible en cada momento resultaba suficiente para clasificar la densidad de la población en una de dos categorías, o si por el contrario se requería observar una nueva unidad de la muestra. El método propuesto permite controlar el error de tipo I así como la potencia del procedimiento de contraste, y al estar basado en el Test de Razón de Verosimilitudes de Neyman-Pearson recibe el nombre de test de razón de verosimilitudes secuencial (*Sequential Probability Ratio Test: S.P.R.T*).

Para ilustrar la idea de este test, consideremos la familia de funciones de densidad $f_i(x)$; $i=1,2$ y el contraste de hipótesis $H_0: X \sim f_0$ frente a $H_1: X \sim f_1$. Para la secuencia de observaciones $X_1,\cdots,X_n$, no necesariamente independientes[♣], definimos el cociente:

$$\lambda_n = \lambda_n(X) = \frac{f_{1n}(X_1,\cdots,X_n)}{f_{0n}(X_1,\cdots,X_n)}, \tag{1.1}$$

y el SPRT tiene la forma:

Aceptar H_0 *si* $\lambda_n \leq A$, *rechazar* H_0 *si* $\lambda_n \geq B$ *y tomar una nueva observación si* $\lambda_n \in (A,B)$,

siendo A y B constantes especificadas. De modo que, la regla de parada tiene la forma $N = \min\{n; \lambda_n \notin (A,B)\}$. Los valores A y B pueden obtenerse fijando la significación y potencia del test. En efecto, sea:

$$B_n = \{(x_1,\cdots,x_n); \lambda_r \in (A,B), r=1,\cdots,n-1; \lambda_n \geq B\}$$

entonces

$$\alpha = P_0(\lambda_N \geq B) = \sum_{n=1}^{\infty} P_0(N=n; \lambda_n \geq B) = \sum_{n=1}^{\infty} \int_{B_n} f_{0n}(x_1,\cdots,x_n)\,dx_1\cdots dx_n$$

$$\leq \frac{1}{B}\sum_{n=1}^{\infty} \int_{B_n} f_{1n}(x_1,\cdots,x_n)\,dx_1\cdots dx_n = \frac{1}{B}P_1(\lambda_N \geq B) = \frac{1}{B}(1-\beta)$$

Análogamente, sea $A_n = \{(x_1,\cdots,x_n); \lambda_r \in (A,B), r=1,\cdots,n-1; \lambda_n \leq A\}$. Se tiene pues:

♣ *Identidad de Wald.*

$$\beta = P_1(\lambda_N \le A) = \sum_{n=1}^{\infty} P_1(N=n;\lambda_N \le A) = \sum_{n=1}^{\infty} \int_{A_n} f_{1n}(x_1,\cdots,x_n)\,dx_1 \ldots dx_n$$

$$\le A\sum_{n=1}^{\infty} f_{0n}(x_1,\cdots,x_n)\,dx_1\cdots dx_n = A\cdot P_0(\lambda_N \le A) = A\cdot(1-\alpha)$$

Se obtiene, entonces, que A y B deben satisfacer las expresiones:

$$\alpha \le \frac{1}{B}(1-\beta) \quad , \quad \beta \le A\cdot(1-\alpha)\,(1.2) \tag{1.2}$$

y donde se ha supuesto que $\sum_{n=1}^{\infty} P_i(N=n) = \sum_{n=1}^{\infty} \int_{A_n \cup B_n} f_{in}(x_1,\cdots,x_n)\,dx_1\cdots dx_n = 1$, con i=0,1, esto es, la regla de parada es finita con probabilidad uno.

Si las trayectorias del proceso $\{\lambda_n, n \in \mathbb{N}\}$ fuesen continuas y $N < \infty$, obviamente $\lambda_N = A$ ó $\lambda_N = B$. En tal caso, las desigualdades (1.2) serían exactamente igualdades. De este modo:

$$\alpha = \frac{1-A}{B-A} \; ; \qquad 1-\beta = \frac{B(1-A)}{B-A} \tag{1.3}$$

En el caso particular de que $X_1,\cdots,X_n$, sean independientes e idénticamente distribuidas, podemos expresar (1.1) como:

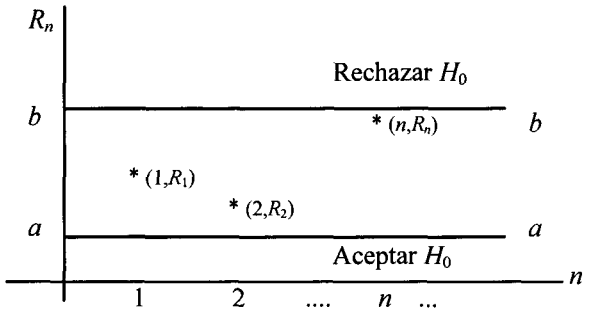
$$\lambda_n = \prod_{i=1}^{n} \frac{f_1(X_i)}{f_0(X_i)} \tag{1.4}$$

y definiendo,

$$Z_i = \log \frac{f_1(X_i)}{f_0(X_i)}$$

entonces $Z_1,\cdots,Z_n$ son también variables aleatorias independientes e idénticamente distribuidas, siendo $\log \lambda_n(X) = \sum_{i=1}^{n} Z_i = R_n$.

En este caso, en términos de R_n el SPRT puede expresarse como sigue: *En cada etapa se calcula* $R_n = \sum_{i=1}^{n} Z_i$ *; entonces si* $\log A = a \leq R_n$ *se acepta* H_0*, si* $\log B = b \geq R_n$ *se rechaza, y si* $a < R_n < b$ *se toma una nueva observación* Z_{n+1}.



La sucesión de variables aleatorias $\{R_n; n \in \mathbb{N}\}$ constituyen un recorrido aleatorio, con $R_n = R_{n-1} + Z_n$, en el cual a y b son barreras absorbentes. Luego, si el SPRT termina con probabilidad uno, entonces $\alpha = P_0(\text{Rechazar } H_0)$ y $1-\beta = P_1(\text{Rechazar } H_0)$ son las probabilidades de que se produzca la absorción cuando la partícula tiene una posición $R_n \geq b$.

Si bien hemos construido aquí el SPRT para contrastar hipótesis simple frente a alternativa simple, resulta sencillo adaptarlo para el contraste de hipótesis unilaterales del tipo $H_0 : \theta \leq \theta_0$ frente a $H_1 : \theta > \theta_0$. Teniendo en cuenta que la probabilidad de decidir a favor de H_0 es máxima para $\theta = \theta_0$, será suficiente con especificar la significación para $\theta = \theta_0$, y la potencia $1-\beta(\theta_1)$ para la alternativa $\theta = \theta_1 > \theta_0$. Sobel & Wald (1949) prueban que el SPRT resulta óptimo también para contrastes unilaterales en el caso de la familia de distribuciones uniparamétricas de tipo exponencial.

Una de las propiedades más interesantes del SPRT es que el muestreo finaliza con probabilidad uno, tal y como se establece en el siguiente resultado:

Lema de Stein: Sea $\{Z_n, n\in\mathbb{N}\}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas y tal que $P(Z_n = 0) < 1$. Sea $-\infty < a < b < \infty$ y $N = \inf\left\{n\ ; \sum_{j=1}^{n} Z_j \notin (a,b)\right\}$, siendo $N = \infty$ si $\sum_{j=1}^{n} Z_j \in (a,b)$, para todo n. Entonces $\exists\ k > 0$ y $0 < \rho < 1$, tales que, para cualquier n, $P(N > n) \le k\rho^n$. En particular, $E[N^j] < \infty$, para todo j =1,2,... y $E[e^{\lambda \cdot N}] < \infty$, para todo $\lambda < \log \rho^{-1}$.

Como consecuencia de este resultado, si $Z_n = \log(f_1(X)/f_0(X))$ es tal que $P(Z_n = 0) < 1$, entonces $P(N < \infty) = 1$. En efecto,

$$P(N < \infty) = \lim_{n\to\infty} P(N \le n) \ge \lim_{n\to\infty} (1 - k\rho^n) = 1, \qquad (1.5)$$

lo que garantiza que con probabilidad uno el SPRT termina.

Wald (1947) demostró que el SPRT era óptimo en el sentido de que entre todos los posibles métodos que permiten tomar una decisión entre H_0 y H_1, para un nivel de significación y una potencia especificada para una alternativa dada, éste requiere el menor *tamaño muestral medio, E*[*N*]. Para la regla de parada $N = \inf\{n\ ;\ \lambda_n \notin (A,B)\}$ puede obtenerse $E[N]$ de manera aproximada cuando $X_1, \cdots, X_n$ son independientes e idénticamente distribuidas. En este caso $\log \lambda_n = \sum_{i=1}^{n} \log(f_1(X_i)/f_0(X_i))$. El siguiente resultado facilita el cálculo de la esperanza $E[\log \lambda_N]$ implicada en la obtención de $E[N]$.

Identidad de Wald.

Sea $\{Z_n,\ n\in\mathbb{N}\}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas y tales que $E[Z_i] = \mu$. Sea N una variable aleatoria con valores enteros tal que el suceso $\{N = n\}$ está determinado por $Z_1, \cdots, Z_n$ y es

independiente de Z_{n+1}, Z_{n+2},.... Supóngase, además, que $E[N]<\infty$. Entonces $E\left[\sum_{j=1}^{N} Z_j\right]=\mu E[N]$.

Ahora, considerando $a = \log A$, $b = \log B$ y $\mu_i = E_i\left[\log\left(f_1(X)/f_0(X)\right)\right]$ ($X \equiv f_i$, $i = 0,1$) y utilizando la identidad de Wald se tiene:

$$E_i\left[\log \lambda_N\right]= \mu_i E_i\left[N\right] \tag{1.6}$$

Si asumimos la aproximación anterior consistente en que $\lambda_N = A$ ó $\lambda_N = B$, entonces:

$$E_i[\log \lambda_N] = a \cdot P_i(\lambda_N \leq A) + b \cdot P_i(\lambda_N \geq B) \tag{1.7}$$

Obsérvese que $P_0(\lambda_N \geq B) = \alpha$ y $P_1(\lambda_N \geq B) = 1-\beta$, de modo que puede obtenerse fácilmente, haciendo además uso de (1.3):

$$E_1[N] = \frac{1}{\mu_1}\left[\frac{a \cdot A \cdot (B-1) + b \cdot B \cdot (1-A)}{B-A}\right] \tag{1.8}$$

$$E_0[N] = \frac{1}{\mu_0}\left[\frac{a \cdot (B-1) + b \cdot (1-A)}{B-A}\right] \tag{1.9}$$

El SPRT presenta el problema de que al ser un procedimiento abierto, es decir con tamaño de muestra no acotado, la distribución del tamaño muestral puede ser bastante sesgada con una varianza elevada, con lo que eventualmente el tamaño muestral final alcanzado puede ser grande. El riesgo de que esto ocurra es aún mayor cuando el verdadero valor del parámetro que se contrasta no es igual a ninguno de los valores considerados en la hipótesis nula o alternativa. Otros procedimientos secuenciales basados en el establecimiento de barreras que resuelven estos problemas son los procedimientos truncado y restringido, y el test triangular de Whitehead. La diferencia fundamental entre ambos métodos es que el procedimiento restringido añade una cota superior al tamaño máximo de la

muestra, evitando así el crecimiento excesivo del tamaño de la misma, mientras que el test triangular establece barreras no paralelas que convergen tras un número predefinido de observaciones, lo que produce forzosamente que dichas barreras sean rebasadas en un tiempo acotado, con la consecuente terminación del proceso secuencial.

1.5. Plan Secuencial Truncado.

En la descripción del test de Wald hemos visto como la decisión a tomar en cada etapa dependía de la posición del punto (n,R_n) respecto a dos barreras horizontales $R=a$ y $R=b$, siendo n el número de datos observados y R_n la log-verosimilitud de la muestra disponible en ese momento. Este procedimiento puede redefinirse en términos de los estadísticos S*core* (S_k) y *Cantidad de Información* (I_k). Para dar una definición adecuada de estos estadísticos supongamos que los datos se obtienen de forma secuencial (individualmente o por grupos) y que la muestra hasta la inspección k-ésima, $\boldsymbol{x} = \{x_1,...,x_{n_k}\}$ tiene una función de verosimilitud $L(\theta,\phi,\boldsymbol{x})$ cuya expresión se conoce salvo un parámetro escalar de interés θ, y un vector de parámetros sin interés $\phi=(\phi_1,...,\phi_r)$ (que en la práctica resultan molestos por cuanto complican la expresión de la verosimilitud). Llamando $\hat{\phi}(\theta)$ al estimador máximo verosímil de ϕ cuando el valor del parámetro de interés es θ, y siendo $l(\theta,\phi)$ la log-verosimilitud anterior, para muestras suficientemente amplias se tiene que $l(\theta,\phi,\mathbf{x}) \cong l\left(\theta,\hat{\phi}(\theta)\right) = l(\theta)$. Asimismo, para θ próximo a cero resulta[1]:

[1] Debe señalarse, no obstante, que este desarrollo puede ser notablemente complejo cuando hay parámetros "molestos" $\phi=(\phi_1,...,\phi_r)$, ya que la presencia del término $\hat{\phi}(\theta)$ obliga a que en la práctica se deba hacer el desarrollo en serie de $l\left(\theta,\hat{\phi}(\theta)\right)$ en un entorno del punto $\left(0,\hat{\phi}(0)\right)$, y

$$l(\theta)=l(0)+l'(0)\theta+\frac{1}{2}l''(0)\theta^2+o(\theta^2) \qquad (1.10)$$

De acuerdo con la idea de *información de Fisher,* la variable *Score* $l'(\theta)$ recoge la variación (sensibilidad) de la verosimilitud ante pequeños cambios de θ indicando, por tanto, si la muestra contiene o no información sobre el parámetro. Asimismo, cuánto más rápidamente varíe $l'(\theta)$ tanto más sensible será dicho score al valor de θ. Una gran sensibilidad indicaría que la muestra contiene mucha información sobre θ, de ahí que la varianza de $l'(\theta)$ reciba el nombre de *cantidad de información* que contiene la muestra sobre θ. Bajo ciertas condiciones de regularidad se tiene además que $\text{var}(l'(\theta))=-E[l''(\theta)]$. Cuando $l''(\theta)$ no depende del valor de θ, si se dan las condiciones citadas, en la expresión anterior el valor de $-l''(\theta)$ se puede identificar con la varianza del *score,* y representa por tanto la cantidad de información que contiene la muestra sobre el parámetro. En este caso, y calculando la verosimilitud con los datos disponibles en la inspección k, denotaremos por $S_k=l'(0)$ y por $I_k=\text{var}(S_k)=-l''(0)$. Por tanto, expresaremos la log-verosimilitud tras la k-ésima inspección como:

$$l(\theta)=l(0)+\theta S_k-\frac{1}{2}\theta^2 I_k+o(\theta^2) \qquad (1.11)$$

A modo de ejemplo, en el caso particular de que la variable bajo observación siga una distribución normal $N(\theta, 1)$, es fácil comprobar que los estadísticos anteriores son:

$$S_k=\sum_{i=1}^{n_k}X_i \qquad I_k=n_k$$

De la expresión del desarrollo en serie de la verosimilitud se sigue que cuando θ es pequeño, $\hat{\theta}^{(k)}=-l'(0)/l''(0)=S_k/I_k$ es el estimador de máxima

por tanto en potencias de θ y $\left(\hat{\phi}(\theta)-\hat{\phi}(0)\right)$, empleando además derivadas parciales respecto a θ y ϕ.

verosimilitud de θ. Por ello $\mathrm{var}\left(\hat{\theta}^{(k)}\right)=\mathrm{var}\left(S_k/I_k\right)=1/I_k$, de donde resulta $I_k=\left(\mathrm{var}\left(\hat{\theta}^{(k)}\right)\right)^{-1}$. Si bien la variable respuesta que se mide durante el test secuencial no tiene por qué ser normal, cuando θ es pequeño y el tamaño de la muestra es grande, la distribución de S_k es aproximadamente $N(\theta I_k, I_k)$. Si las inspecciones secuenciales pudiesen realizarse de modo continuo, el proceso S podría considerarse aproximadamente como un movimiento browniano.

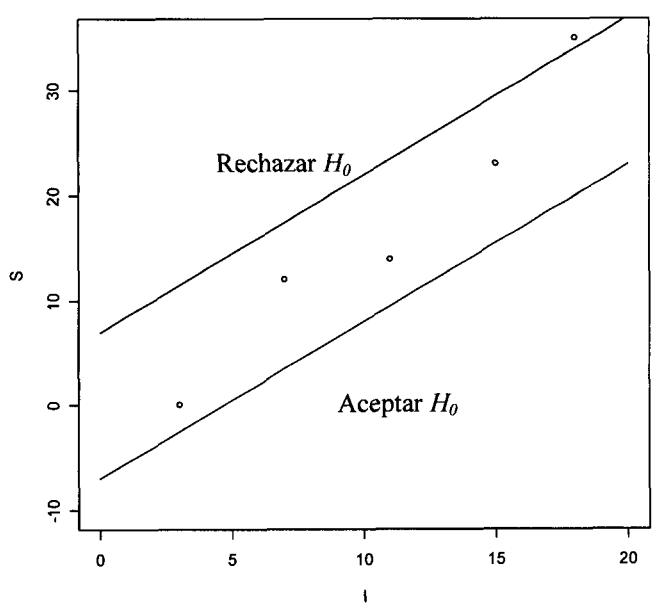
Si en el SPRT para contrastar $H_0:\theta=\theta_0$ frente a $H_1:\theta=\theta_1$ desarrollamos la log-verosimilitud $R_k=\sum_{i=1}^{k}\log(f_1(X_i)/f_0(X_i))$ en función de $\theta=\theta_1-\theta_0$, esto es[2]:

$$R_k=l(0)+\theta S_k-\frac{1}{2}\theta^2 I_k \qquad (1.12)$$

en términos de S_k e I_k la regla de parada del SPRT quedaría entonces expresada como sigue:

> *Si $S_k > a + bI_k$ rechazar la hipótesis nula; si $S_k <- a + bI_k$ aceptar la hipótesis nula. En otro caso tomar una nueva muestra.*

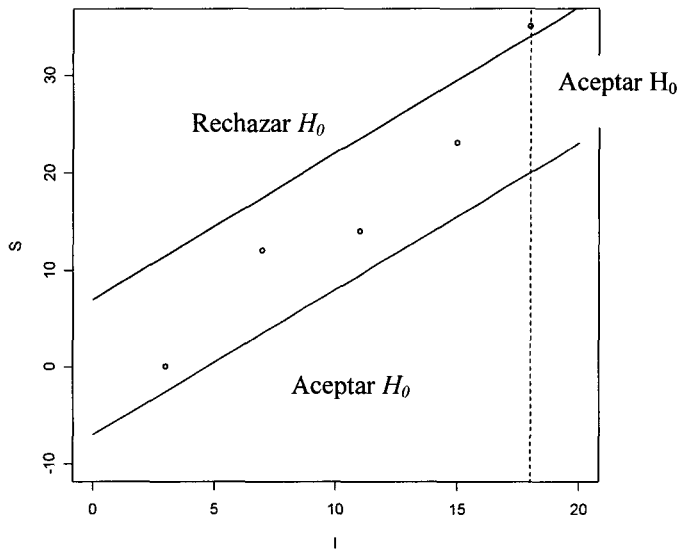
Gráficamente:



El plan secuencial restringido se basa en el SPRT modificándolo mediante la incorporación de una cota superior para el tamaño muestral máximo (lo que resulta equivalente a acotar superiormente el valor máximo de I_k). Si llamamos K a la inspección en que se alcanza el tamaño máximo $N=n$, la regla de parada es:

Mientras $k<K$, *si* $S_k < -a + b \cdot I_k$ *se para y se acepta* H_0; *si* $S_k > a + b \cdot I_k$, *se para y se rechaza* H_0; *si* $-a + b \cdot I_k < S_k < a + b \cdot I_k$ *se toma una nueva muestra. Cuando* $k = K$, *si* $S_k > a + b \cdot I_k$, *se para y se rechaza* H_0; *en caso contrario se para y se acepta* H_0.

Gráficamente:



En el test truncado es preciso recalcular a y b para garantizar que se cumplen las condiciones de nivel de significación y potencia requeridas por el contraste, lo que requerirá en general el uso de métodos numéricos. Armitage (1975) señala como estos valores pueden determinarse mediante la planificación de una serie de contrastes de significación sucesivos a un nivel nominal adecuado que garantice un nivel de significación global preespecificado para todo el contraste. Veremos esta metodología en detalle en el epígrafe 1.9.

[2] A veces el modelo estadístico adecuado para comparar θ_0 con θ_1 puede no utilizar necesariamente la diferencia entre ambos valores; por ejemplo, si son proporciones, podría ser más conveniente usar la *odd-ratio*.

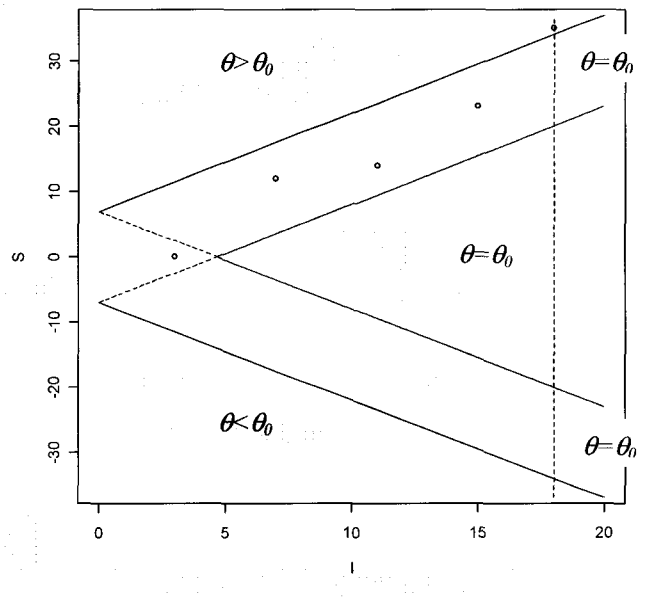
Señalemos, por último, que la introducción del concepto de nivel de información I_k permite que estos contrastes puedan generalizarse para su uso no sólo en diseños secuenciales en que las observaciones se van tomando una a una, sino para diseños en que las observaciones se van tomando en grupos.

1.6. Plan Secuencial Restringido.

El SPRT y el Plan Secuencial Truncado pueden adaptarse fácilmente a contrastes bilaterales de la forma $H_0 : \theta = \theta_0$ frente a $H_1 : \theta \neq \theta_0$, de modo que se alcancen unos objetivos de nivel de significación y potencia predeterminados. Sin embargo, en caso de rechazar H_0, el test no nos dice nada del sentido de la diferencia. Sin embargo hay ocasiones, por ejemplo, cuando se comparan dos tratamientos experimentales en un ensayo clínico, en que es importante decidir no sólo si son o no equivalentes, sino cuál de ellos es mejor. El plan secuencial restringido permite la realización del test bilateral mediante la realización de dos contrastes unilaterales, cada uno de ellos con una potencia prefijada $1-\beta$ para la alternativa correspondiente. Al igual que el test truncado, presenta también una cota superior para el tamaño máximo de la muestra. Si, como en el caso anterior, K es la inspección en que se alcanza el tamaño máximo de muestra, la regla de parada para este test es la siguiente:

Mientras $k<K$, *si* $S_k > a + b \cdot I_k$ *se para y se acepta que* $\theta > \theta_0$; *si* $S_k < -a - b \cdot I_k$, *se para y se acepta que* $\theta < \theta_0$; *si* $a - b \cdot I_k < S_k < -a + b \cdot I_k$ *se para y se acepta* $H_0: \theta = \theta_0$; *en otro caso se continúa hasta la siguiente inspección. Cuando* $k = K$, *si* $S_k > a + b \cdot I_k$, *se para y se acepta que* $\theta > \theta_0$; *si* $S_k < -a - b \cdot I_k$, *se para y se acepta que* $\theta < \theta_0$; *en otro caso se para y se acepta* H_0.

Gráficamente:



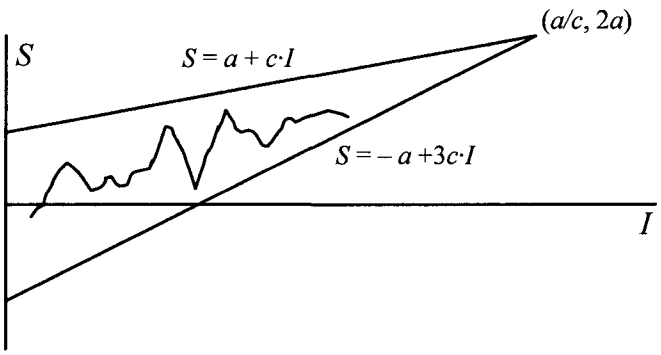
1.7. El test triangular de Whitehead

Este es un contraste unilateral cuyos orígenes se encuentran en el contraste secuencial de $H_0 : \theta = 0$ frente a $H_1 : \theta > 0$ cuando se considera que la trayectoria del proceso S es continua. Las cotas superior e inferior para el contraste así definido quedan determinadas por las rectas:

$$u(I) = \frac{2}{\delta}\log\left(\frac{1}{2\alpha}\right) + \frac{\delta}{4}I, \qquad l(I) = -\frac{2}{\delta}\log\left(\frac{1}{2\alpha}\right) + \frac{3\delta}{4}I, \tag{1.13}$$

siendo $0 < I \leq \left(8/\delta^2\right)\log\left(1/2\alpha\right)$, y el procedimiento secuencial termina en el instante en que, por primera vez, la trayectoria supera la cota superior, $S(I) \geq u(I)$, o queda por debajo de la cota inferior, $S(I) \leq l(I)$. Puesto que el proceso es continuo, la primera condición que se cumpla lo hará con igualdad, y por lo tanto el contraste finaliza rechazando H_0 si $S(I) = u(I)$ y aceptando si $S(I) = l(I)$.

Gráficamente:



donde $a = (2/\delta)\log(1/2\alpha)$ y $c = \delta/4$.

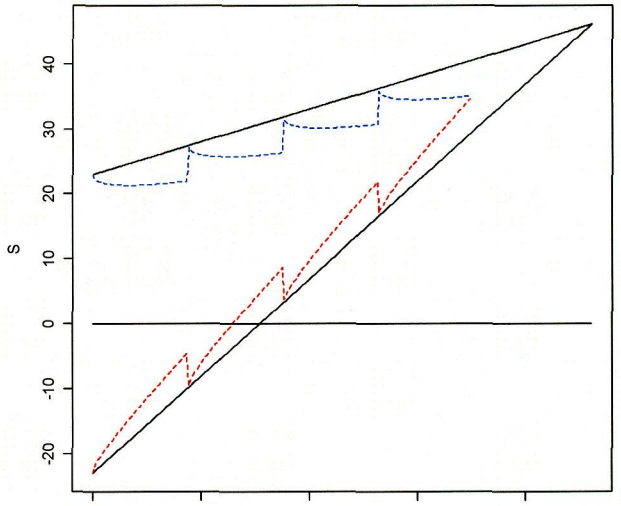
El contraste así descrito fue desarrollado por Lorden (1976), que lo llamó 2-SPRT, y probó que con las cotas que se han definido, y considerando que las trayectorias del proceso $S(I)$ pueden observarse de modo continuo, el nivel de significación es α y la potencia para $\theta = \delta$ es exactamente 1-α. Posteriormente, Whitehead & Stratton (1983), haciendo uso de un resultado debido a Siegmund (1979, 1985) y considerando que el proceso $S(I)$ puede aproximarse mediante un movimiento browniano, adaptaron este contraste a la situación mucho más realista de que las trayectorias del proceso $S(I)$ sólo pueden ser observadas en instantes discretos. Para ello realizan una corrección en las cotas que definen el contraste, consistente en restar el valor esperado del salto que podría producirse en el movimiento browniano durante el tiempo en que éste no está siendo observado. Dicho valor, entre las inspecciones k-1 y k viene dado por $0.583\sqrt{I_k - I_{k-1}}$. De esta forma el efecto neto de esta corrección es un acercamiento mutuo de las cotas superior e inferior, que adoptan una forma que recuerda a un árbol de navidad. Con esta modificación de las cotas el test triangular de Whitehead adopta la forma:

- *Tras la j-ésima inspección,* $j = 1, \cdots, K$-1*:*
  - *Si* $S_j \geq u_j = a + cI_j - 0.583\sqrt{I_j - I_{j-1}}$, *el contraste se para, y se rechaza* H_0.

- *Si* $S_j \le l_j = -a + 3cI_j + 0.583\sqrt{I_j - I_{j-1}}$, *el contraste se para y se acepta* H_0
- *En cualquier otro caso, se continúa con la inspección* $j+1$.

- *Si se alcanza la K-ésima inspección y* $S_K \ge u_K$ *se rechaza* H_0; *en caso contrario se acepta* H_0.

Gráficamente:



Whitehead demuestra que las probabilidades de que el proceso S_j observado en tiempo discreto alcance las cotas corregidas, son aproximadamente las mismas de que el proceso en tiempo continuo alcance las cotas sin corregir, conservándose así los niveles de significación y potencia de aquel caso.

En el supuesto particular de que los niveles de información considerados estén igualmente espaciados, siendo $I_K = I_{\max}$, entonces $I_j = (j/K)I_{\max}$, en el j-ésimo análisis, con $j=1,\cdots,K$. Las cotas superior e inferior para S_j serán ahora, respectivamente, sin más que sustituir en las expresiones anteriores:

$$u_j = \frac{2}{\delta}\log\left(\frac{1}{2\alpha}\right) - 0.583\sqrt{\frac{I_{\max}}{K}} + \frac{\delta}{4}\frac{j}{K}I_{\max}$$

$$l_j = -\frac{2}{\delta}\log\left(\frac{1}{2\alpha}\right) + 0.583\sqrt{\frac{I_{\max}}{K}} + \frac{3\delta}{4}\frac{j}{K}I_{\max}$$

Si tenemos en cuenta que las cotas coinciden en el último análisis, esto es $u_K=l_K$, igualando las dos expresiones anteriores para $j=K$ y despejando $I_{\max}$, se obtiene:

$$I_{\max}=\left[\sqrt{\frac{4\cdot 0.583^2}{K}+8\log\left(\frac{1}{2\alpha}\right)}-\frac{2\cdot 0.583}{\sqrt{K}}\right]^2\frac{1}{\delta^2}$$

Así pues, fijados α y K, se puede obtener el valor $I_{\max}$ y a partir de él los valores de las cotas anteriores u_j y l_j, $j=1,\cdots,K$, de tal forma que para el contraste unilateral de $H_0:\theta=0$ con K análisis, se alcanza un nivel de significación aproximadamente igual a α y una potencia aproximada de 1-α para una alternativa $\theta=\delta$.

Whitehead & Stratton (1983) generalizan el problema para contrastes con significación α, en $\theta=0$ y una potencia 1-β en $\theta=\delta$, cuando $\alpha\neq\beta$. Para ello, en primer lugar observan que un test de tamaño fijo con nivel de significación α y potencia 1-β en $\theta=\delta$ tiene potencia 1-α en $\theta=\xi\delta$, siendo

$$\xi=\frac{2\Phi^{-1}(1-\alpha)}{\Phi^{-1}(1-\alpha)+\Phi^{-1}(1-\beta)}$$

Como las curvas de potencia para los test de tamaño fijo y para los test secuenciales en grupos son muy próximas, si el test triangular se construye como antes, pero utilizando ahora $\tilde{\delta}=\xi\delta$ en lugar de δ en la fórmula de $I_{\max}$, se obtiene un test cuya potencia es aproximadamente 1-β en $\theta=\delta$.

Cuando los incrementos de información entre inspecciones son desiguales (situación bastante común en los ensayos clínicos, en que ocurre que el número de sujetos finalmente observados en cada inspección puede llegar a diferir notablemente de lo planificado al comienzo del ensayo) el test de Whitehead permite mantener aproximadamente los niveles de significación y potencia especificados al comienzo del ensayo.

Una característica fundamental del Test Triangular es que sólo requiere de cálculos elementales para determinar los valores de las cotas y, sólo con ello, dicho contraste consigue satisfacer con bastante precisión, y bajo secuencias de información bastante generales (Jennison & Turnbull, 1999), los requerimientos preestablecidos sobre las probabilidades de ambos tipos de error. Los procedimientos secuenciales basados en el uso de *funciones del gasto o consumo de la probabilidad de error,* que veremos posteriormente, se comportan mejor que el test triangular ante incrementos muy desiguales en el nivel de información entre inspecciones, pero como contrapartida requieren del apoyo de técnicas de cálculo numérico para obtener los sucesivos valores de las cotas. Asimismo, debemos indicar que el test Triangular no conduce a obtener la mejor reducción posible en el *tamaño de muestra esperada*, debido a que su *tamaño de muestra máximo* es bastante alto, ya que el tamaño muestral final requerido por el test depende del patrón de tamaños de muestra que se vaya observando en los sucesivos análisis intermedios. En este sentido, los ya citados contrastes basados en el consumo de la probabilidad de error tipo I necesitan menores tamaños medios de muestra.

1.8. Doble Test Triangular.

Del mismo modo que en el caso del Plan Secuencial Restringido, es posible adaptar el test triangular para realizar un contraste bilateral de la hipótesis $H_0 : \theta = \theta_0$, que permita además decidir unívocamente entre las dos alternativas $\theta > 0$ y $\theta < 0$. El parámetro θ podría medir, por ejemplo, la diferencia entre un tratamiento experimental y un control, e interesaría saber si el experimental es significativamente mejor o peor que el control. Whitehead & Stratton (1983), y posteriormente Whitehead (1997), consiguen este objetivo mediante la construcción del contraste *doble triangular*. Éste se compone de dos tests triangulares diseñados para contrastar la hipótesis nula $H_0 : \theta = 0$, cada uno de ellos con una significación $\alpha/2$; el primero de dichos contrastes tiene como alternativa $H_1 : \theta > 0$, con potencia 1-β en $\theta = +\delta$; para el segundo, la alternativa

es $H_1 : \theta < 0$, siendo su potencia también 1-β en θ=-δ. Las inspecciones continúan hasta que ambos contrastes hayan llegado a una conclusión, rechazando H_0 si ambos contrastes han rechazado $\theta = 0$ y aceptándola en otro caso. Este contraste tiene la ventaja adicional de que si H_0 es cierta, dicha hipótesis se aceptará relativamente pronto.

En términos de los estadísticos Score, S_j con $j = 1,\cdots,K$, y utilizando niveles de información igualmente espaciados, el contraste triangular unilateral de $\theta = 0$ frente a $\theta > 0$ tiene como cotas superior e inferior para S_j, en el j-ésimo análisis:

$$u_j = \frac{2}{\tilde{\delta}} \log\left(\frac{1}{\alpha}\right) - 0.583\sqrt{\frac{I_{\max}}{K}} + \frac{\tilde{\delta}}{4}\frac{j}{K} I_{\max}$$

$$l_j = -\frac{2}{\tilde{\delta}} \log\left(\frac{1}{\alpha}\right) + 0.583\sqrt{\frac{I_{\max}}{K}} + \frac{3\tilde{\delta}}{4}\frac{j}{K} I_{\max}$$

siendo $\tilde{\delta} = 2\Phi^{-1}(1-\alpha/2)\delta / [\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)]$ y donde:

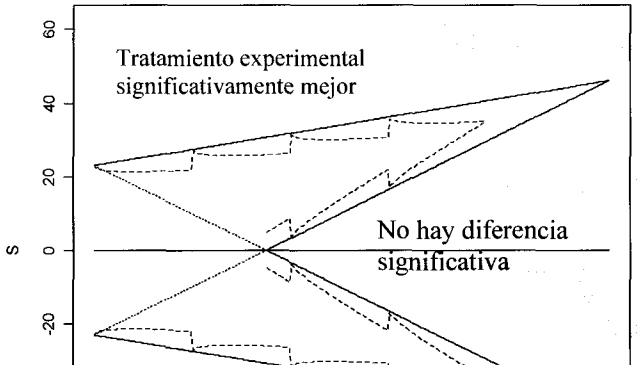
$$I_{\max} = \left[\sqrt{\frac{4 \cdot 0.583^2}{K} + 8\log\left(\frac{1}{2\alpha}\right)} - \frac{2 \cdot 0.583}{\sqrt{K}}\right]^2 \frac{1}{\tilde{\delta}^2}$$

Asimismo, las cotas superior e inferior del contraste $\theta = 0$ frente a $\theta < 0$ se obtienen de modo análogo como $l_j' = -l_j$ y $u_j' = -u_j$, respectivamente. Combinando estos contrastes adecuadamente se obtiene la siguiente regla de parada:

- *Tras el grupo* $j = 1,\cdots, K-1$,
  - *Si* $|S_j| \geq u_j$ *se para y se rechaza* H_0.
  - *Si* $S_i \leq l_i$ *para algún* $i \leq j$ *y* $S_i \geq -l_i$ *para algún* $i \leq j$, *se para y se acepta* H_0,
  - *En otro caso se continúa con el grupo* $j+1$.

- *Tras el grupo K:*
  - *Si* $|S_K| \geq u_K$ *se para y se rechaza* H_0.
  - *En otro caso se para y se acepta* H_0.

Gráficamente:



1.9. Contrastes secuenciales basados en la realización repetida de contrastes de significación: Contrastes RST bilaterales.

Los métodos secuenciales vistos hasta ahora se basan en el establecimiento de barreras en el plano (S,I); las sucesivas observaciones se van representando en dicho plano y la decisión final depende de la barrera alcanzada. Una aproximación diferente al análisis secuencial de datos es la que ofrece la realización repetida de contrastes de significación (RST, *Repeated Significance Tests*). Este procedimiento, introducido por Armitage (1958, 1969, 1971) presupone que en cada etapa de inspección, k, de los datos se lleva a cabo un contraste de significación a nivel α_k. Es preciso tener en cuenta el hecho evidente de que si todos estos contrastes se realizaran con una significación de, digamos, 0.05 a medida que aumenta el número de inspecciones aumenta la probabilidad de que

alguno de los contrastes dé significativo, aún cuando la hipótesis nula puesta a prueba sea verdadera. La única manera de evitar este problema es elegir los valores de α_k adecuadamente para que el nivel de significación global para todo el contraste (esto es, la probabilidad de rechazar en alguna etapa H_0 siendo cierta) se mantenga en un nivel controlado α. Dos son las formas principales de conseguir este objetivo: planificar a priori el número máximo de inspecciones a realizar, así como el tamaño de muestra en cada inspección, y *distribuir* el error α entre todas ellas; o bien, si no es posible anticipar el número de inspecciones que se van a realizar, ó los tamaños de muestra, definir *a priori* una función de *consumo del error tipo I* (*α-spending*) y en cada inspección, una vez que se conoce el nivel de información (tamaño de muestra) efectivamente alcanzado, determinar a partir de dicha función el nivel α del contraste de significación a realizar. Trataremos en primer lugar los contrastes de la primera categoría.

Los primeros procedimientos RST introducidos por Armitage consideraban que los datos iban estando disponibles uno a uno, y que tras cada observación era posible realizar un contraste de significación. Posteriormente Pocock (1977), O'Brien & Fleming (1979), y Wang & Tsiatis (1981) generalizaron este método para la realización de contrastes bilaterales cuando la toma de datos se produce secuencialmente en grupos de igual tamaño. Emerson & Fleming (1989), y Pampallona & Tsiatis (1994) desarrollaron una variación del mismo para la realización de contrastes unilaterales. En los procedimientos descritos por estos autores se fija a priori el número máximo de inspecciones K. En cada análisis intermedio k=1,⋯,K se calcula un estadístico estandarizado, Z_k, a partir de los primeros k grupos de observaciones. En el caso bilateral la regla de parada en los contrastes RST es de la forma:

- *Tras observar el grupo* k=1,⋯,K-1:
  - *Si* $|Z_k| \geq c_k$ *se para y se rechaza* H_0;
  - *en caso contrario se continúa con el grupo* j+1.

- *Tras observar el grupo K*:
  - *Si* $|Z_K| \geq c_K$ *se para y se rechaza* H_0;
  - *en caso contrario se para y se acepta* H_0.

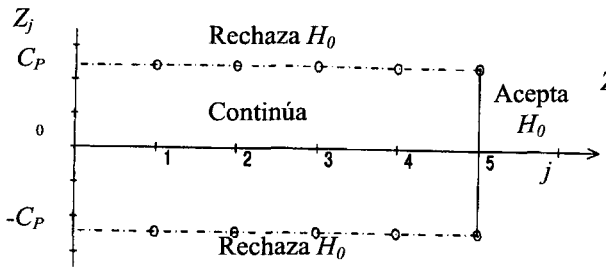
El estadístico Z_k es usualmente el pivote natural que se utiliza en los contrastes equivalentes con muestra de tamaño fijo. A modo de ejemplo, supongamos que tras un adecuado proceso de aleatorización, dos grupos de pacientes son sometidos a dos tratamientos A y B, y que las respuestas observadas para cada tratamiento son normales e independientes, con $X_A \cong N(\mu_A, \sigma)$ y $X_B \cong N(\mu_B, \sigma)$, y que se desea contrastar $H_0 : \mu_A = \mu_B$ frente a $H_1 : \mu_A \neq \mu_B$. Cuando se dispone de los datos de los pacientes de los k primeros grupos, a razón de m pacientes por grupo en cada tratamiento, el estimador de la diferencia $\theta = \mu_A - \mu_B$ es $\bar{X}_A^{(k)} - \bar{X}_B^{(k)}$. Y se define el estadístico estandarizado Z_k como:

$$Z_k = \frac{\sum_{i=1}^{m \cdot k} X_{A_i} - \sum_{i=1}^{m \cdot k} X_{B_i}}{\sqrt{2mk\sigma^2}}, \quad k = 1, 2, \cdots, K$$

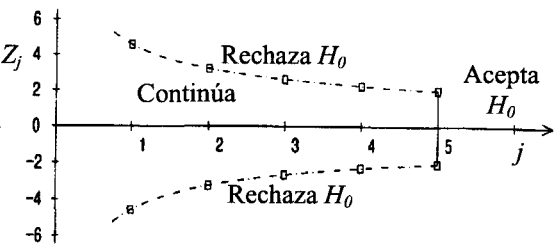
Cuando los valores de c_k de la regla de parada son de la forma:

$$c_k = C_{WT}(K, \alpha, \Delta)(j/K)^{\Delta - 1/2}$$

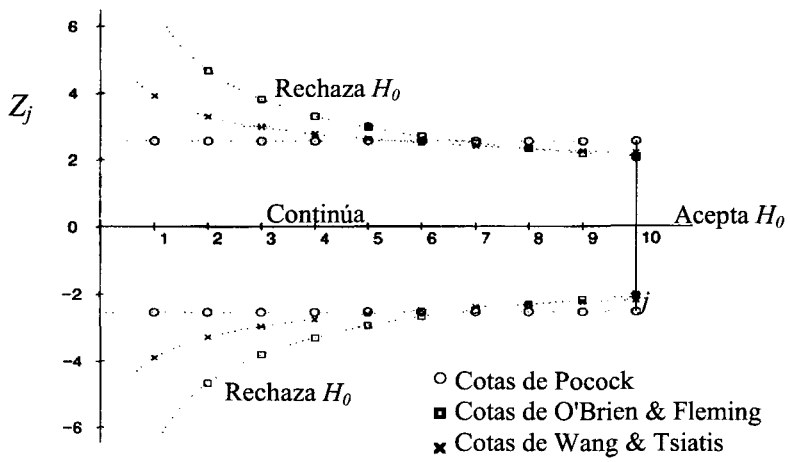
se obtiene el contraste definido originalmente por Wang & Tsiatis. Si se escoge $\Delta = 0.5$ se obtiene el contraste de Pocock y si $\Delta = 0$ resulta el de O'Brien & Fleming. La sucesión de valores críticos $\{c_1, \cdots, c_K\}$ se determina de modo que se alcance globalmente un nivel de significación, especificado al inicio del ensayo. Como puede apreciarse en la expresión anterior, el test de Pocock se caracteriza por unos valores c_k constantes. El test de O'Brien & Fleming produce valores c_k que van decreciendo y que son mayores que los de Pocock en las primeras inspecciones. Otros valores de Δ producen perfiles intermedios:



Contraste de Pocock



Contraste de O'Brien & Fleming



Contrastes de Wang y Tsiatis

Los contrastes secuenciales RST tienen un tamaño de muestra esperado inferior al tamaño requerido para un contraste no secuencial equivalente (con la misma significación y potencia), si bien el tamaño máximo posible es superior. En cualquier caso, cuando la hipótesis nula es falsa, el tamaño de muestra requerido por estos métodos es inferior, por lo que cumplen mejor con el imperativo ético de interrumpir pronto un ensayo clínico si existe una evidencia fuerte a favor de la superioridad de uno de los tratamientos puestos a prueba.

Para valores grandes de $|\mu_A-\mu_B|$ el contraste de Pocock requiere de *tamaños de muestra esperados* más bajos que los del contraste de O'Brien & Fleming. Por el contrario, el contraste de Pocock alcanza *tamaños de muestra máximos* más

grandes que aquél, y sus *tamaños de muestra esperados* son mayores que los de un contraste no secuencial cuando $|\mu_A-\mu_B|$ es pequeña.

En el diseño del test secuencial el incremento del valor del número máximo de grupos a observar, K, en general redunda en una reducción del tamaño de muestra esperado para las alternativas de interés, si bien esta ventaja deberá ponderarse con las dificultades prácticas que conlleve la realización de un mayor número de análisis intermedios. Asimismo, si bien en el caso del test de O'Brien & Fleming el tamaño muestral esperado para valores grandes de $|\mu_A-\mu_B|$ sigue decreciendo continuamente con K, en el caso del test de Pocock dicho valor comienza a crecer de nuevo a partir de cierto K en adelante.

Por su parte, dada la particular estructura de los valores críticos en el test de O'Brien & Fleming, el nivel de significación nominal aplicado al final del análisis es próximo a la probabilidad de error de tipo I global. En consecuencia es muy difícil que se produzca la poco deseable (y confusa, pero posible) situación de aceptar la hipótesis nula cuando un análisis no secuencial del conjunto final de datos la habría rechazado, situación que es más fácil que se produzca en el test de Pocock.

El test de O'Brien & Fleming aplica niveles de significación nominal extremadamente bajos en los primeros análisis, lo que hace difícil rechazar la hipótesis nula muy pronto. Existen poderosas razones prácticas para preferir este tipo de situaciones que impiden que el ensayo termine excesivamente pronto: problemas con la calidad de los datos (que aparecen sobre todo al principio de los ensayos, ya que no se tiene experiencia administrando los tratamientos o midiendo la respuesta de los pacientes), dificultad para verificar la validez de las hipótesis que se han realizado sobre las distribuciones de los datos ya que aún no se han recogido los suficientes para ello, escepticismo de la comunidad científica sobre conclusiones extraídas a partir de pequeños conjuntos de datos. Debe señalarse, no obstante, que estas cuestiones son menos importantes si los tamaños muestrales de los grupos son sustanciales. Por razones éticas, los comités de ensayos clínicos

se muestran reluctantes a continuar con un ensayo si para grupos grandes se observan altos valores del estadístico Z_k en una etapa inicial, aún cuando no se hayan cruzado las cotas de O'Brien & Fleming.

El contraste de Wang & Tsiatis contiene como casos particulares al contraste de Pocock y al de O'Brien & Fleming, y puede considerarse un compromiso entre ambos. Para valores de $0 < \Delta < 0.5$ el contraste de Wang & Tsiatis presenta cotas de forma intermedia entre aquellos dos. En general los tamaños de grupo necesarios en el test de Wang & Tsiatis para unos valores de α y β prefijados para una alternativa concreta, se sitúan entre los tamaños requeridos por los test de Pocock y O'Brien & Fleming. Jennison & Turnbull (2000) comparan estos tres métodos secuenciales y aprecian que el test de Wang & Tsiatis con Δ=0.1 requiere en general un menor tamaño de muestra esperado que el de O'Brien & Fleming, a costa de un pequeño incremento en el tamaño muestral máximo. Con Δ=0.4. se obtienen tamaños muestrales esperados similares o inferiores a los del test de Pocock, y tamaños máximos sensiblemente inferiores. En general, a medida que Δ crece los tamaños muestrales esperados para valores grandes de la diferencia $|\mu_A-\mu_B|$ se reducen, si bien con el coste añadido de que para valores $|\mu_A-\mu_B|$ próximos a cero los tamaños esperados y máximos son también mayores.

Por tanto, la familia de tests de Wang & Tsiatis ofrece unos útiles grados de libertad extra para poder elegir un test secuencial que permita equilibrar lo mejor posible el conflicto entre los objetivos de alcanzar tamaños muestrales máximos y esperados lo más reducidos posible sobre un rango adecuado de valores de $|\mu_A - \mu_B|$.

Como hemos visto en la definición de la regla de parada de estos métodos, la condición de parada depende de la sucesión de valores $\{c_1,\cdots,c_K\}$. El cálculo de estas cantidades depende del nivel de significación α deseado así como del número máximo de inspecciones K a realizar. Para ver como pueden calcularse

estos valores supondremos que se lleva a cabo un estudio secuencial de grupos de casos con un máximo de K análisis intermedios, que produce la sucesión de estadísticos $\{Z_1,\cdots,Z_K\}$. Estos estadísticos siguen la llamada *distribución conjunta canónica* con niveles de información $\{I_1,\cdots,I_K\}$ para el parámetro θ si:

$$\text{i.}\,(Z_1,\cdots,Z_K) \text{ es normal multivariante.}$$

$$\text{ii.}\,E[Z_k]=\theta\sqrt{I_k},\quad k=1,\cdots,K \tag{1.14}$$

$$\text{iii.}\,\operatorname{cov}(Z_i,Z_j)=\sqrt{I_i/I_j}\quad ;\ 1\le i\le j\le K$$

Aunque esta descripción puede parecer restrictiva, de hecho muchos problemas de análisis secuencial de datos (y desde luego los citados más arriba) dan lugar a estadísticos de contraste con esta distribución conjunta. Ello se debe fundamentalmente a dos razones. Por una lado, puede probarse un resultado general (Jennison & Turnbull, 1999) que garantiza que las sucesiones de estadísticos de contraste estandarizados obtenidos a partir de la estimación máximo verosímil de un parámetro en un modelo lineal normal siguen precisamente esta distribución conjunta. Por otra parte, en ausencia de normalidad, estos autores prueban también un resultado asintótico según el cual, en condiciones muy generales y para tamaños de muestra suficientemente grandes se obtiene también la distribución (1.14). En cualquier caso, para la validez de este resultado, ya el primer grupo de observaciones debe ser lo suficientemente grande para que la aproximación asintótica funcione desde la primera inspección.

En el ejemplo que hemos citado más arriba, si definimos el nivel de información sobre θ en este momento como el inverso de la varianza de este estimador resulta:

$$I_k = \left(\frac{2\sigma^2}{km}\right)^{-1}$$

y por tanto el estadístico estandarizado Z_k puede expresarse como:

$$Z_k = \frac{\sum_{i=1}^{m\cdot k} X_{A_i} - \sum_{i=1}^{m\cdot k} X_{B_i}}{\sqrt{2mk\sigma^2}} = \left(\bar{X}_A^{(k)} - \bar{X}_B^{(k)}\right)\sqrt{I_k}, \quad k = 1,\cdots,K$$

Dada la normalidad de las respuestas, el estadístico $(Z_1,\cdots,Z_K)$ es normal multivariante. Además, para cada *k*, es $Z_k \approx N\left(\theta\sqrt{I_k},1\right)$. Asimismo, si $i \leq j$, se tiene que:

$$Z_j = \left(\bar{X}_A^{(j)} - \bar{X}_B^{(j)}\right)\sqrt{I_j} = \frac{1}{jm}\left(\sum_{k=1}^{im}\left(X_{A_k} - X_{B_k}\right) + \sum_{k=im+1}^{jm}\left(X_{A_k} - X_{B_k}\right)\right)\sqrt{I_j} =$$

$$= \frac{1}{jm}\left(im\frac{Z_i}{\sqrt{I_i}} + \sum_{k=im+1}^{jm}\left(X_{A_k} - X_{B_k}\right)\right)\sqrt{I_j}$$

y teniendo en cuenta que Z_i es independiente de las observaciones X_{A_k} y X_{B_k} para $k > im$ y que $\mathrm{var}(Z_i) = 1$, resulta:

$$\mathrm{cov}\left(Z_i, Z_j\right) = \mathrm{cov}\left(Z_i, \frac{i}{j}\frac{\sqrt{I_j}}{\sqrt{I_i}}Z_i\right) = \frac{i}{j}\frac{\sqrt{I_j}}{\sqrt{I_i}} = \frac{\sqrt{I_i}}{\sqrt{I_j}}$$

Por tanto $\{Z_1,\cdots,Z_K\}$ tienen la distribución conjunta canónica (1.14), con niveles de información $\{I_1,\cdots,I_K\}$ para $\theta = \mu_A - \mu_B$. Este resultado se mantiene para este contraste cuando se suponen varianzas distintas en los dos grupos, e incluso cuando los tamaños de grupo son distintos en cada tratamiento. Jennison & Turnbull (1999) muestran como la distribución conjunta canónica aparece asociada a los estadísticos estandarizados a utilizar en muchos otros contrastes secuenciales para datos normales: contraste para la media de una población, contraste de medias en muestras apareadas, ensayos cruzados, etc.

Puesto que la hipótesis nula se rechaza en la etapa k si $|Z_k|$ (que bajo la hipótesis nula sigue una distribución normal estándar) supera el valor crítico c_k, este método puede interpretarse como una sucesión de contrastes de significación reiterados, con un nivel de significación nominal $\alpha_k = 2(1-\Phi(c_k))$ en la etapa k, siendo $\Phi(x)$ la función de distribución de la normal estándar.

El nivel de significación global del test es entonces:

$$\alpha = P_{\theta=0}\left(\text{Rechazar } H_0\right) = P_{\theta=0}\left(\bigcup_{k=1}^{K}\left\{|Z_k| > c_k\right\}\right) \tag{1.15}$$

Asimismo, la potencia de este test para una alternativa $\theta = \delta$ es:

$$1-\beta = P_{\theta=\delta}\left(\text{Rechazar } H_0\right) = \sum_{k=1}^{K} P_{\theta=\delta}\left(|Z_k| > c_k\right) \tag{1.16}$$

La determinación de las sucesiones $c_1,\cdots,c_K$ en un contraste concreto, para garantizar un nivel de significación α, o la determinación del tamaño de grupo necesario para conseguir una potencia $1-\beta$ para una alternativa $\theta=\delta$, requieren la aplicación de métodos numéricos para evaluar las probabilidades (1.15) y (1.16). A continuación delineamos brevemente la forma en que pueden obtenerse numéricamente tales valores de c_k y m. Para ello supondremos que el número máximo de análisis a realizar es K, y que los estadísticos $\{Z_1,\cdots,Z_K\}$ siguen la distribución conjunta canónica. Si llamamos $\Delta_k = I_k - I_{k-1}$ (k=2,$\cdots$,K), entonces:

$$Z_1 \approx N\left(\theta\sqrt{I_1},1\right) \tag{1.17}$$

$$Z_k\sqrt{I_k} - Z_{k-1}\sqrt{I_{k-1}} \approx N\left(\theta\Delta_k,\Delta_k\right),\ k=2,\cdots,K \tag{1.18}$$

independientemente de $Z_1,\cdots,Z_{K-1}$.

Aunque en nuestros contrastes la región de continuación es de la forma $(-c_k, c_k)$, los resultados que se muestran a continuación son válidos para cualquier región de la forma (a_k, b_k) donde los extremos no tienen por qué ser iguales en valor absoluto. La determinación de los niveles de significación y la potencia requieren el cálculo de la probabilidad de que en alguna etapa k el estadístico Z_k caiga fuera de la región de continuación. En particular, llamemos:

$$U_{k,\theta}(a_1,b_1,...,a_k,b_k) = P_\theta(a_1 < Z_1 < b_1,...,a_{k-1} < Z_{k-1} < b_{k-1}, Z_k > b_k) \quad (1.19)$$

$$L_{k,\theta}(a_1,b_1,...,a_k,b_k) = P_\theta(a_1 < Z_1 < b_1,...,a_{k-1} < Z_{k-1} < b_{k-1}, Z_k < a_k) \quad (1.20)$$

a las probabilidades, bajo θ, de que el test termine en la etapa k con la salida de Z_k, respectivamente, por encima o por debajo de (a_k, b_k). Para calcular estas probabilidades observemos que la densidad de Z_1 es:

$$g_1(z;\theta) = \phi\left(z - \theta\sqrt{I_1}\right) \quad (1.21)$$

siendo $\phi(x) = \exp\left(-x^2/2\right)/\sqrt{2\pi}$ la densidad de la normal estándar. De (1.18) se sigue que para $k = 2,\cdots,K$, la densidad condicional de Z_k, dado $Z_1 = z_1,\cdots,Z_{k-1} = z_{k-1}$ depende sólo de z_{k-1} y es igual a:

$$f_k(z_{k-1},z_k;\theta) = \frac{\sqrt{I_k}}{\sqrt{\Delta_k}}\phi\left(\frac{z_k\sqrt{I_k} - z_{k-1}\sqrt{I_{k-1}} - \theta\Delta_k}{\sqrt{\Delta_k}}\right) \quad (1.22)$$

Por tanto, si construimos:

$$g_k(z;\theta) = \int_{a_{k-1}}^{b_{k-1}} g_{k-1}(u;\theta) f_k(u,z;\theta)\,du, \quad k = 2,\cdots,K \quad (1.23)$$

entonces $g_k(z;\theta)$ representa la densidad de Z_k, supuesto que el procedimiento secuencial ha llegado hasta el grupo k. Asimismo, si denotamos como $p(k, z; \theta)$ la

densidad conjunta de k y Z_k cuando k es la *etapa en que termina* el test, se sigue que:

$$p(k,z;\theta)=\begin{cases} g_k(z;\theta) & z\notin(a_k,b_k) \\ 0 & z\in(a_k,b_k) \end{cases} \tag{1.24}$$

A partir de estas funciones puede calcularse:

$$U_{k,\theta}(a_1,b_1,...,a_k,b_k)=\int_{b_k}^{\infty}p(k,z;\theta)dz=\int_{b_k}^{\infty}g_k(z;\theta)dz \tag{1.25}$$

$$L_{k,\theta}(a_1,b_1,...,a_k,b_k)=\int_{-\infty}^{a_k}p(k,z;\theta)dz=\int_{-\infty}^{a_k}g_k(z;\theta)dz \tag{1.26}$$

Las ecuaciones (1.21), (1.22) y (1.23) permiten evaluar numéricamente, de forma recursiva, las integrales que aparecen en (1.25) y (1.26), mediante un adecuado método de cuadratura, tal como, por ejemplo, el método de Simpson.

Una vez que se dispone de un algoritmo eficiente para el cálculo de las probabilidades $U_{k,\theta}(a_1,b_1,...,\ a_k,b_k)$ y $L_{k,\theta}(a_1,b_1,...,\ a_k,b_k)$ es fácil calcular numéricamente los niveles críticos para los distintos métodos secuenciales en los que es válida la distribución conjunta canónica de los estadísticos de contraste. Por ejemplo, en el caso del test de Pocock, para contrastar la hipótesis nula $\mu=\mu_0$, cuando se observan datos procedentes de una población $N(\mu,\sigma)$, con σ conocida, los niveles de información son de la forma $I_k=n_k/\sigma^2$, y el estadístico estandarizado es:

$$Z_k=\left(\bar{X}^{(k)}-\mu_0\right)\sqrt{I_k},\quad k=1,\cdots,K. \tag{1.27}$$

Considerando $\theta=\mu-\mu_0$ es fácil comprobar que los Z_k siguen la distribución conjunta canónica (1.14). La región de continuación de este contraste es de la forma $(-c,\ c)$. Por tanto la probabilidad α es:

$$\alpha = P_{\theta=0}\left(\text{Rechazar } H_0\right) = P_{\theta=0}\left(\bigcup_{k=1}^{K}\left\{|Z_k| > c\right\}\right) =$$
$$= \sum_{k=1}^{K} P_{\theta=0}\left(|Z_k| > c\right) = \sum_{k=1}^{K}\left\{L_{k,0}(-c,c,....,-c,c) + U_{k,0}(-c,c,....,-c,c)\right\} \tag{1.28}$$

Si se tiene en cuenta la forma de las trayectorias de $Z_1,\cdots,Z_K$ es fácil observar que el error tipo I disminuye a medida que el valor de c aumenta. Así pues, fijados el nivel de significación α y el número máximo de grupos K, el valor $c = C_P(K,\alpha)$ puede obtenerse numéricamente mediante un sencillo algoritmo de búsqueda que vaya variando el valor de c en el cálculo numérico de las expresiones (1.25) y (1.26) hasta que se cumpla (1.28). De igual forma, una vez fijado el valor de c para conseguir el nivel de significación α, puede calcularse también la potencia para una alternativa $\theta = \delta$ simplemente hallando el valor numérico de:

$$1-\beta = P_{\theta=\delta}\left(\text{Rechazar } H_0\right) = \sum_{k=1}^{K} P_{\theta=\delta}\left(|Z_k| > c\right) =$$
$$= \sum_{k=1}^{K}\left\{L_{k,\delta}(-c,c,....,-c,c) + U_{k,\delta}(-c,c,....,-c,c)\right\} \tag{1.29}$$

Si $\theta = \delta$ es mayor que cero y suficientemente grande, la probabilidad de que la trayectoria de las Z_i quede por debajo de $-c$ puede despreciarse, y la expresión de la potencia se reduce a:

$$1-\beta = \sum_{k=1}^{K} U_{k,\delta}(-c,c,....,-c,c) \tag{1.30}$$

Asimismo, si $\theta = -\delta$ es negativo y suficientemente grande en valor absoluto, la potencia se reduce a:

$$1-\beta = \sum_{k=1}^{K} L_{k,-\delta}(-c,c,....,-c,c) \tag{1.31}$$

Si se desea evaluar el tamaño máximo de muestra necesario para realizar el test con un nivel de significación α, una potencia 1-β para una alternativa $\theta = \delta$ prefijada, y un número máximo K de inspecciones, será preciso determinar el valor $I_{\max}$ para el que (1.30) o (1.31) alcanzan el valor 1-β prefijado cuando $I_k = (k/K) I_{\max}$. Esta tarea también puede llevarse a cabo también fácilmente de modo numérico toda vez que dichas expresiones son crecientes con $I_{\max}$. Por último, el tamaño máximo de muestra se determina despejándolo de la expresión de $I_{\max}$ como inverso de la varianza del estimador de θ. Así, en el caso que estamos considerando, como $I_{\max} = n_{\max}/\sigma^2$, se sigue que el número máximo de observaciones a tomar es $n_{\max} = \sigma^2 I_{\max}$, y el número de observaciones por inspección es $n = n_{\max}/K$. Como en general este valor no será entero, se toma como norma redondearlo al entero superior, para garantizar una potencia igual o ligeramente superior a la preespecificada.

Todo lo dicho hasta aquí se aplica a contrastes secuenciales con grupos de idéntico tamaño en cada instante de observación. Cuando los tamaños de grupo son variables, Jennison & Turnbull (1999) realizan, para los distintos contrastes introducidos hasta ahora, un estudio exhaustivo del efecto de disponer de distintos niveles de información en cada instante de observación. Para ello consideran que dichos niveles son de la forma:

$$I_j = \pi (j/K)^r I_{\max}, \quad j = 1, \cdots, K,$$

Aquí π controla el nivel de información final y juega un papel influyente en la potencia del contraste, como era de esperar: si $\pi < 1$ el tamaño final alcanzado es inferior al planificado y el test pierde potencia; por el contrario, si $\pi > 1$ se dispone de tamaños de muestra mayores y la potencia observada puede ser mayor que la planificada. Por su parte, el exponente r afecta el distanciamiento entre los niveles de información: si $r = 1$, los tamaños de los grupos son iguales; si $r < 1$ se producen apelotonamientos en los primeros instantes de observación (grupos de tamaño más grande inicialmente); y si $r > 1$ se comienza inicialmente con grupos

más pequeños que van siendo progresivamente mayores en los sucesivos instantes de observación. Se observa que cuando $r<1$ el nivel de significación global disminuye, y cuando $r>1$ aumenta. En cualquier caso, tanto el incremento como la disminución son muy ligeros.

De esta forma, la conclusión es que aunque los tamaños de los grupos sean desiguales, si la desigualdad no es excesiva, ello no tendrá, en general, demasiado efecto sobre la significación y la potencia del test secuencial. Sin embargo, si es previsible que haya diferencias muy grandes en los tamaños de los grupos, o que éstos sean absolutamente impredecibles, es mejor utilizar otros métodos de contraste secuencial para grupos, entre los que cabe citar el método de la *función del gasto o consumo de la probabilidad de error de tipo I*, que trataremos más adelante.

1.10. Métodos RST basados en la Función del Gasto o Consumo de la Probabilidad de Error de Tipo I: contrastes bilaterales.

Los contrastes descritos hasta ahora están originalmente diseñados para un número predeterminado y fijo, K, de grupos de observaciones de igual tamaño y que dan lugar a niveles de información igualmente espaciados $\{I_1,\cdots,I_K\}$. En la práctica, en el curso de un ensayo clínico no siempre es posible cumplir con este objetivo, ya que es frecuente que en los sucesivos instantes de inspección no se disponga de muestras del mismo tamaño (debido a que la tasa de reclutamiento de pacientes puede ser muy variable en el tiempo, puede haber sujetos que abandonen el ensayo antes de su finalización, etc.); a veces la información acumulada puede invitar a cambiar la frecuencia con que se revisan los datos en algún punto durante el curso del ensayo; podría ocurrir que se produzca un reclutamiento más lento de lo anticipado, lo que forzaría a la extensión del ensayo y, por tanto, a aumentar el número de instantes de inspección.

El procedimiento que vamos a describir a continuación es suficientemente flexible para tratar con secuencias de información impredecibles, y garantiza un nivel de significación exactamente igual a α cualquiera que sea la sucesión de niveles de información observados. Dicho método tiene la ventaja, además, de no requerir un número máximo de análisis fijados a priori. Supondremos, como hemos hecho en secciones anteriores, que el test secuencial se refiere a un parámetro desconocido θ, que en los sucesivos análisis intermedios se dispone de estimadores $\hat{\theta}^{(k)}$ con niveles de información $I_k = \left(\text{var}\left(\hat{\theta}^{(k)}\right)\right)^{-1}$, y que los estadísticos estandarizados $Z_k = \hat{\theta}^{(k)}\sqrt{I_k}$ para contrastar θ=0 siguen la distribución conjunta canónica (1.14) (bien de modo exacto para datos normales, o bien de modo aproximado para otro tipo de datos).

Los métodos secuenciales basados en el *consumo* o *gasto* de la probabilidad de error tipo I requieren la utilización de una función que permita ir cuantificando precisamente cuánta de esa probabilidad se ha ido *consumiendo* o *gastando* en las sucesivas inspecciones a lo largo de un estudio, hasta que se produce su conclusión. Presentamos este método en el contexto del contraste bilateral de la hipótesis $H_0 : \theta = 0$, utilizando el método exacto para garantizar el nivel de significación sugerido por Slud & Wei (1982). Estos autores proponen fijar, antes de comenzar el estudio, el número máximo de análisis, K (se supone que no es posible planificar el número de observaciones en cada inspección), y el error de tipo I, que se reparte en probabilidades $\pi_1, \cdots, \pi_K$ cuya suma total vale α. A medida que se van observando los sucesivos niveles de información $I_1, \cdots, I_K$ se van calculando, condicionados por dichos valores, los valores críticos c_j para los estadísticos estandarizados Z_j, de tal modo que, para cada j=1,⋯,K:

$$P_{\theta=0}\left(|Z_1| < c_1, \cdots, |Z_{j-1}| < c_{j-1}, |Z_j| \geq c_j\right) = \pi_j \qquad (1.32)$$

El contraste procede de acuerdo con la regla de parada habitual:

Rechazar H_0 en el j-ésimo análisis si $|Z_j| \geq c_j$, $j=1,\cdots,K$, o parar y aceptar H_0 si no ha sido rechazada en el análisis K.

Así, π_j representa la probabilidad de parar en la j-ésima inspección rechazando H_0 cuando es cierta; por ello π_j recibe el nombre de *error gastado o consumido en la etapa j*. Nótese que este error no tiene por qué coincidir con el nivel de significación nominal $2\{1-\Phi(c_j)\}$ en el j-ésimo análisis intermedio cuando se emplea la aproximación de los contrastes de significación repetidos con niveles de información predefinidos. Como $\pi_1 + \cdots + \pi_K = \alpha$, la tasa global de error de tipo I es exactamente α.

La condición $P(|Z_1| \geq c_1) = \pi_1$ cuando $\theta = 0$ implica que el primer valor crítico es simplemente:

$$c_1 = \Phi^{-1}(1-\pi_1/2). \tag{1.33}$$

El cálculo de los restantes valores críticos necesita de la resolución de (1.32) para $j=2,\cdots,K$. Para ello, definiendo:

$$G_k(z;\theta) = P_\theta\left(|Z_1| < c_1, \cdots, |Z_{k-1}| < c_{k-1}, Z_k \geq z\right) \tag{1.34}$$

los valores de c_k se pueden obtener recursivamente resolviendo:

$$G_k(c_k;0) = \pi_k/2, \quad k = 2,\cdots,K \tag{1.35}$$

Si llamamos $g_k(z,\theta)$ a la derivada de $G_k(z,\theta)$ respecto a z, entonces $g_k(z,\theta)$ es también la densidad de Z_k condicionada a que el procedimiento secuencial ha llegado hasta la k-ésima inspección. Cuando los $\{Z_1,\cdots,Z_K\}$ siguen la distribución conjunta canónica (1.14), las ecuaciones (1.21), (1.22) y (1.23) permiten evaluar numéricamente la integral de $g_k(z,\theta)$ y por tanto obtener el valor c_k que cumple:

$$\int_{c_k}^{\infty} g_k(z;0)\,dz = \frac{\pi_k}{2} \tag{1.36}$$

La única diferencia con el procedimiento delineado en el epígrafe 1.9 es que allí los valores de los I_k eran conocidos a priori, mientras que ahora debe esperarse a cada inspección para conocer el valor de I_k. Asimismo, la potencia para $\theta = \delta$ puede obtenerse numéricamente como $\pi_1(\delta) + \cdots + \pi_K(\delta)$ donde:

$$\pi_k(\delta) = \int_{-\infty}^{c_k} g_k(z;\delta)dz + \int_{c_k}^{\infty} g_k(z;\delta)dz, \quad k = 1, \cdots, K \qquad (1.37)$$

Aunque el método propuesto por Slud & Wei contiene ya la idea clave del gasto o consumo de error tiene, sin embargo, ciertas limitaciones. Dado que el número máximo de análisis se fija al inicio del estudio, resulta difícil adaptarlo a tasas de recopilación de información inesperadamente altas o bajas y, por tanto, alcanzar el requerimiento de potencia deseado. Asimismo, se podría preferir variar el consumo de error en algún análisis, en respuesta a los niveles de información observados. Por ejemplo, si hay poca información disponible en el primer análisis, debido a la lenta incorporación o al retardo en la respuesta de los pacientes, resulta razonable reducir π_1, reservando el error de tipo I para etapas posteriores en las que se cuente con un mayor nivel de información. El método introducido por Lan & DeMets (1983) y generalizado posteriormente por Kim & DeMets (1987) solventa estos problemas haciendo depender el consumo de error tipo I del nivel de información disponible en cada etapa, y estableciendo una regla de muestreo que hace que el procedimiento continúe hasta alcanzar un nivel máximo de información preespecificado o, si ello no es posible, adaptando el consumo del error a las irregularidades que se van produciendo en el calendario de inspecciones.

En el método propuesto por Lan & DeMets (1983), que asume que el ensayo puede continuar hasta alcanzar un nivel de información máximo predefinido $I_{\max}$, el error de tipo I se reparte de acuerdo con una función de consumo de error, $\alpha^*(t)$, no decreciente y que verifica que $\alpha^*(0) = 0$ y $\alpha^*(t) = \alpha$ para $t \geq 1$. El valor $\alpha^*(t)$ indica la cantidad de error de tipo I acumulado que se deberá consumir cuando se haya obtenido una fracción t de la

información máxima prevista I_{max}. Se asume que mientras no se produzca una condición de parada que lleve a rechazar o aceptar H_0, el procedimiento debe continuar hasta alcanzar el nivel de información I_{max}, momento en que se detiene definitivamente el proceso. Ello obliga a que si, por ejemplo, los pacientes entran en el ensayo a una tasa inferior de la prevista, los organizadores del mismo deberán prolongar en el tiempo la duración del estudio.

La función de consumo de error $\alpha^*(t)$ y el nivel de información máximo $I_{máx}$ han de ser seleccionados antes de iniciar el estudio. Los errores de tipo I asignados a cada análisis son entonces:

$$\pi_1 = \alpha^*(I_1/I_{max})$$

$$\pi_j = \alpha^*(I_j/I_{max}) - \alpha^*(I_{j-1}/I_{max}), \quad j = 2,3,...$$

y los valores críticos c_j se van calculando sucesivamente de manera que satisfagan (1.32), de modo análogo al procedimiento de Slud & Wei (1982). Ahora K es el valor más pequeño de j para el cual $I_j \geq I_{max}$. Dado que, por definición, $\alpha^*(I_K/I_{max}) = \alpha$, tenemos que $\pi_1 + \cdots + \pi_K = \alpha$ y la tasa global de error de tipo I es exactamente α, tal como se deseaba.

Ha habido un gran número de propuestas para la forma de la función de gasto de error de tipo I, $\alpha^*(t)$. Lan & DeMets (1983) señalan que la función:

$$\alpha^*(t) = \begin{cases} 0 & \text{si } t=0 \\ 2-2\Phi(z_{\frac{\alpha}{2}}/\sqrt{t}) & \text{si } 0<t\leq 1 \end{cases}$$

produce valores críticos próximos a los del contraste de O'Brien & Fleming, cuando los tamaños de grupos son iguales. Para obtener valores similares a los de las cotas de Pocock, Lan & DeMets (1983) proponen:

$$\alpha^*(t) = \min\{\alpha\log[1+(e-1)t], \alpha\}$$

Por último, Lan & DeMets (1987) y Jennison & Turnbull (1989, 1990) proponen la familia indexada por el parámetro ρ>0:

$$\alpha^*(t) = \min\{\alpha t^{\rho}, \alpha\}$$

Una selección adecuada del parámetro ρ da lugar a que esta familia de funciones produzca cotas muy similares a las de los contrastes de Pocock (ρ =1), O'Brien & Fleming (ρ =3) y Wang & Tsiatis (por ejemplo ρ =2 da cotas similares a las de Wang & Tsiatis para Δ=0.25). Jennison & Turnbull (1999) muestran que con los contrastes basados en esta función para el gasto de error en el caso de observar niveles de información igualmente espaciados, y eligiendo adecuadamente el valor de ρ se consiguen ventajas sobre los tests citados, concretamente tamaños máximos de muestra menores que aquellos, a la vez que tamaños medios de muestra similares o inferiores.

Para conseguir que este método alcance una potencia prefijada para una alternativa concreta, hay que tener en cuenta que bajo θ = 0 la distribución conjunta de $Z_1,\cdots,Z_K$, condicionada a $I_1,\cdots,I_K$, depende sólo de las proporciones $I_1/I_{\max},\cdots,I_K/I_{\max}$. Por tanto los valores críticos $c_1,\cdots,c_K$ son funciones de estas proporciones. A su vez, cuando $\theta\neq 0$, la distribución de $Z_1,\cdots,Z_K$ depende de los niveles de información absolutos $I_1,\cdots,I_K$. Si bien existe un efecto de la forma en que se produce la sucesión de los niveles de información observados, puede comprobarse empíricamente que este efecto es muy pequeño, y es fundamentalmente el valor de $I_{\max}$ el que finalmente determina la potencia del método. Por ello, una forma conveniente y sencilla de conseguir que el test tenga una potencia muy próxima a la deseada consiste en anticipar un número máximo de inspecciones K y suponer que los niveles de información van a estar equiespaciados (esto es, $I_k = (k/K)I_{\max}$), y calcular el valor de $I_{\max}$ que con esta secuencia de niveles de información producirá la potencia deseada para la alternativa de interés.

Dado que usualmente la parada del método secuencial se produce cuando se ha alcanzado como mínimo el nivel de información $I_{\max}$, la potencia finalmente alcanzada coincide o supera la inicialmente prevista. Obviamente, si razones externas obligan a parar el ensayo antes de haber alcanzado el nivel $I_{\max}$ la potencia alcanzada será normalmente inferior a la utilizada para diseñar el procedimiento. Asimismo, en este caso, para garantizar que el error Tipo I global sea igual al valor α elegido en el diseño del ensayo, el valor de la última cota c_K debe elegirse resolviendo (1.32) para $K=j$, con $\pi_K = \alpha - \alpha^*(I_{K-1}/I_{\max})$.

2. El bootstrap con diseños de tamaño fijo

2.1. Introducción

La publicación en 1979 del primer trabajo de Bradley Efron sobre los métodos Bootstrap constituyó uno de los sucesos de mayor relieve en el campo de la estadística en la década de los años 80. La idea de reemplazar complicadas aproximaciones y a veces poco precisas, a sesgos, varianzas y otras medidas de incertidumbre por simulaciones con el ordenador, fue ampliamente utilizada por investigadores teóricos y usuarios de los métodos estadísticos. Pasado el inicial escepticismo sobre la viabilidad de esta metodología, los investigadores empezaron a poner de manifiesto que estos métodos en muchos casos aproximan mejor que los convencionales.

El bootstrap fue concebido por Efron en el contexto de diseños de tamaño fijo y su planteamiento del bootstrap es muy simple. Dada una muestra aleatoria $\mathbf{X} = (X_1, \cdots, X_n)$ de una distribución de probabilidad F, el objetivo es estimar la distribución de probabilidad de una variable aleatoria especificada $R = R(\mathbf{X}, F)$, posiblemente dependiente de $\mathbf{X}$ y de la distribución de probabilidad desconocida, utilizando los datos observados. Estas cantidades R son a menudo sesgos, varianzas, pivotales o test estadísticos, cuyas distribuciones de probabilidad exacta son difíciles de determinar, o el error de la aproximación es muy considerable. La aproximación de sumas de variables aleatorias independientes y con varianza finita por la distribución normal puede tener errores en la aproximación del orden $n^{-1/2}$. La aproximación bootstrap a la distribución probabilística de R, dada por Efron en el referido trabajo de 1979 es como sigue:

i. Construir a partir de los datos $\mathbf{X}=(X_1,\cdots,X_n)$ la distribución empírica $\hat{F}_n(t)=\frac{1}{n}\sum_{i=1}^{n} I(X_i \le t)$

ii. Extraer una muestra aleatoria de tamaño *n*, $\mathbf{X}^*=(X_1^*,\cdots,X_n^*)$ de la distribución empírica $\hat{F}_n(t)$, la cual recibe el nombre de distribución bootstrap. La muestra bootstrap, a diferencia del jackknife, es aquí una muestra aleatoria con reemplazamiento del conjunto de datos observados $\mathbf{X}=(X_1,\cdots,X_n)$. En el jackknife, se seleccionan muestras aleatorias sin reemplazamiento de tamaño *n*-1.

iii. La distribución de $R(\mathbf{X},F)$ se aproxima finalmente por la distribución bootstrap de $R^*=R(\mathbf{X}^*,\hat{F})$.

Parece claro que la distribución de probabilidad de R^* puede calcularse de la distribución empírica en el mismo modo que la distribución de *R* se calcula de la distribución original *F*. Ahora bien, el recurrir al bootstrap en líneas generales se deberá a la dificultad de obtener la distribución exacta de *R*. La clave del método bootstrap consiste en que, tomando repetidas muestras de la distribución empírica mediante el método de Monte Carlo, se obtiene una muestra de valores de *R* que en si misma constituye la aproximación bootstrap buscada. Naturalmente, esto es posible gracias a la capacidad de los ordenadores modernos. Precisamente los métodos bootstrap reciben también el nombre de métodos de computación intensiva. La idea de generar nuevos datos a partir de los datos observados recordaba al Barón de Munchausen, cuando estando en el fondo de un lago logró salir tirando el mismo de los cordones de sus botas.

2.2. Distribución de remuestreo

La distribución empírica se utiliza en el algoritmo dado por Efron como distribución de remuestreo por ser un estimador consistente de la verdadera

distribución generadora de los datos. Para diseños de tamaño fijo, $\hat{F}_n(t)$ tiene las siguientes propiedades como estimador de $F(t)$.

i. $E\left[\hat{F}_n(t)\right] = F(t)$, para cualquier valor de t.

ii. $\text{var}\left(\hat{F}_n(t)\right) = \dfrac{F(t)(1-F(t))}{n}$.

iii. $\left\|\hat{F}_n - F\right\| \to 0$, a.s., para $n \to \infty$ (Teorema de Glivenko-Cantelli), donde la norma considerada es la del supremo.

Como veremos en el siguiente capítulo, estas propiedades en general no se mantienen cuando los datos se generan a través de un diseño secuencial.

Es interesante destacar la forma que adopta la distribución empírica para datos binarios. Así pues, sea $X_1, \cdots, X_n$ una muestra aleatoria de la distribución $b(1,\pi)$. Obviamente, la distribución empírica es $b(1,\hat{\pi}_n)$, siendo $\hat{\pi}_n = 1/n \sum_{i=1}^{n} X_i$. Nótese que $\text{var}(X_i) = \pi \cdot (1-\pi)$. Si $X_1^*, \cdots, X_m^* \cong b(1,\hat{\pi}_n)$, entonces $\text{var}_*\left(X_i^*\right) = \hat{\pi}_n \cdot (1-\hat{\pi}_n)$, que es justamente el procedimiento plug-in habitual de estimar la varianza de X_i.

Puede alternativamente considerarse también como distribución de remuestreo estimadores de núcleo de la función de distribución, los cuales tienen la forma:

$$\hat{F}_n(x,h) = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x - X_i}{h}\right) \tag{2.1}$$

siendo W una función de distribución y h el parámetro de suavizamiento o bandwidth. La función W recibe el nombre de núcleo integrado si $W'(x)$ existe, es lipschitziana, de soporte compacto, continua y con momento de segundo orden finito.

En general, en todos los estimadores de núcleo, la elección del bandwidth es crucial. El siguiente teorema da las propiedades del estimador alisado.

Teorema 2.1. Sea $W(x)$ un núcleo integrado y $F \in C^2$. Entonces:

i. $$E\left[\hat{F}_n(x;h)\right] - F(x) = \frac{F''(x)\mu_2(K)}{2}h^2 + o(h^2), \quad h \to 0$$

ii. $$\mathrm{var}\left(\hat{F}_n(x;h)\right) = \frac{F(x)(1-F(x))}{n} - \frac{F'(x)\mu_1\left(\left(W^2\right)_\bullet\right)}{n}h + o(h), \quad h \to 0$$

donde $\mu_2(K) = \int x^2 K(x)dx$ y $\mu_1\left(\left(W^2\right)_\bullet\right) = \int y \cdot (\partial/\partial y) W^2(y)dy$. ■

Este resultado es útil para determinar la forma del bandwidth óptimo h. Un criterio de elección de este parámetro se basa en la función error cuadrático medio integrado la cual se define por:

$$\mathrm{mise}(h) = E\left[\int\left\{\hat{F}_n(x,h) - F(x)\right\}^2 dx\right] \tag{2.2}$$

Un criterio habitual de optimalidad para el bandwidth consiste en seleccionar como aquel que minimiza la función mise(h). Teniendo en cuenta que esta función puede también expresarse como:

$$\mathrm{mise}(h) = \int \mathrm{var}\left(\hat{F}_n(x;h)\right)dx + \int\left\{E\left[\hat{F}_n(x;h)\right] - F(x)\right\}^2 dx$$

la forma del bandwidth óptimo es:

$$h_0 = \left\{\frac{\mu_1\left(W_\bullet^2\right)}{\mu_2(K)^2 \cdot \|F''\|_2^2 \cdot n}\right\}^{1/3} \tag{2.3}$$

donde $\|F''\|_2^2 = \int F''(x)^2 dx$.

En la expresión del bandwidth óptimo aparece la función de distribución F, la cual es desconocida. De este modo, el valor exacto del bandwidth no puede determinarse y por tanto habrá de ser estimado. Para tal fin puede utilizarse un método de validación cruzada debido a Bowman *et al.* Este se basa en la función de validación cruzada la cual se define por:

$$\mathrm{cv}(h) = \frac{1}{n}\sum_{i=1}^{n}\int\left\{I_{[0,\infty)}(x - X_i) - \hat{F}_{-i}(x,h)\right\}^2 dx \tag{2.4}$$

donde $\hat{F}_{-i}(x,h)$ representa el estimador de núcleo evaluado en el punto x, pero obtenido omitiendo la observación X_i. El bandwidth óptimo h_0 se estima entonces como $\hat{h}_n$, siendo $\mathrm{cv}\left(\hat{h}_n\right) = \min_h \mathrm{cv}(h)$.

Una propiedad de esta aproximación se obtiene considerando:

$$H(h) = \mathrm{cv}(h) - \frac{1}{n}\sum_{i=1}^{n}\left\{I_{[0,\infty)}(x - X_i) - F(x)\right\}^2 dx \tag{2.5}$$

El término que se sustrae a cv(h) no contiene a h y de esta forma, no afecta al procedimiento de validación cruzada. Es inmediato por otra parte probar que:

$$E[H(h)] = E\left[\int\left\{\hat{F}_{n-1}(x,h) - F(x)\right\}^2\right] \tag{2.6}$$

Esta ecuación sugiere que $H(h)$ puede ser una buena aproximación a mise(h). Bowman prueba que bajo ciertas condiciones, $\hat{h}_n/h_0 \to 1$ con probabilidad 1 para $n \to \infty$, siendo h_0 el bandwidth óptimo y $\hat{h}_n$ el que minimiza a cv(h).

2.3. El bootstrap en la regresión

En este epígrafe haremos una breve revisión del uso del bootstrap en los modelos de regresión con diseño de tamaño fijo en el que nos basaremos para proponer en el siguiente capítulo su uso en diseños secuenciales. Consideremos por tanto un modelo de regresión de la forma:

$$Y_i = g_i(\beta) + e_i \; ; \; i = 1, \cdots, n \tag{2.7}$$

siendo $g_i(\beta)$ una función conocida del parámetro vectorial desconocido β. Supondremos que $e_1, \cdots, e_n$ es una muestra aleatoria de una distribución de probabilidad F, tal que $E[e_i] = \int x \cdot F(dx) = 0$. Para el conjunto de datos observados $Y_1, \cdots, Y_n$, se considera una estimación del vector desconocido β por algún método específico, tal como el de mínimos cuadrados. Sea pues $\hat{\beta}$ el vector que minimiza:

$$\sum_{i=1}^{n} \left[Y_i - g_i(\beta) \right]^2 \tag{2.8}$$

El objetivo ahora es obtener la distribución de probabilidad del estimador $\hat{\beta}$. Esta puede fácilmente aproximarse por el bootstrap de acuerdo con el siguiente algoritmo:

i. Sea $\hat{F}_n$ la función de distribución empírica correspondiente a los residuales centrados. Esto es: sean los residuales $\tilde{e}_i = Y_i - g_i(\hat{\beta})$ y $\hat{e}_i = \tilde{e}_i - 1/n \sum_{i=1}^{n} \tilde{e}_i$. Por tanto, $\hat{F}_n(x) = 1/n \sum_{i=1}^{n} I(\hat{e}_i \leq x)$.

ii. De $\hat{F}_n$ se extrae una muestra aleatoria $e_1^*, \cdots, e_n^*$. A partir de ésta se generan $Y_i^* = g_i(\hat{\beta}) + e_i^*$.

iii. Utilizando nuevamente el método de mínimos cuadrados se obtiene una realización bootstrap β^* minimizando:

$$\sum_{i=1}^{n}\left[Y_i^* - g_i(\beta)\right]^2 \tag{2.9}$$

Repetidas observaciones bootstrap $\left\{\beta^{*i}; i=1,\cdots,B\right\}$ proporcionarán la aproximación bootstrap buscada.

La repercusión práctica de este procedimiento es indudable. En los modelos de regresión no lineal proporcionan aproximaciones a la distribución de probabilidad de pivotales, a través de los cuales es posible obtener intervalos de confianza para β, tales como $\left(\hat{\beta}-E\left[\hat{\beta}\right]\right)/\mathrm{sd}\left(\hat{\beta}\right)$. Si el modelo es lineal, aparentemente no son tan necesarias las aproximaciones bootstrap toda vez que el estimador $\hat{\beta}$ es lineal en las observaciones, por lo que el efecto del teorema central del límite puede proporcionar una aproximación a su distribución de probabilidad. Veremos no obstante en el siguiente epígrafe a través de los desarrollos de Edgeworth como las aproximaciones bootstrap pueden ser mejores que la aproximación normal.

2.4. El Bootstrap y los desarrollos de Edgeworth

En el epígrafe anterior hemos mostrado como a través del bootstrap puede simplificarse notablemente la obtención de la distribución de probabilidad de algunos estadísticos. El efecto del teorema central del límite sobre estadísticos lineales o la normalidad asintótica de los estimadores de máxima verosimilitud, dan aproximaciones a la distribución de tales estadísticos. El bootstrap puede mejorar tales aproximaciones en el sentido de que los errores de aproximación son menores. En orden a mostrar estas ideas, consideremos un pivotal T_n con distribución asintótica normal estándar. En casos regulares, su distribución de probabilidad admite un desarrollo de la forma:

$$G(x) = P(T_n \leq x) = \Phi(x) + n^{-1/2} q(x)\phi(x) + O(n^{-1}) \qquad (2.10)$$

siendo $q(x)$ un polinomio de grado dos. A lo largo de esta memoria, Φ y ϕ representan la función de distribución y función de densidad respectivamente de la distribución normal estándar.

Para una muestra aleatoria $X_1, \cdots, X_n$ de una distribución de probabilidad con media μ y varianza σ^2, el pivotal $T_n = \sqrt{n}(\bar{X} - \mu)/\sigma$ tiene distribución asintótica normal estándar. La validez del desarrollo anterior requiere la siguiente condición (de Cramer):

$$E[X_i^3] < \infty \quad \text{y} \quad \limsup_{|t| \to \infty} |\varphi_{X_i}(t)| < 1$$

siendo φ_{X_i} la función característica de X_i.

Este desarrollo pone de manifiesto que el error de aproximar la distribución probabilística del pivotal T_n por la distribución normal estándar (su distribución asintótica) es del orden $n^{-1/2}$, lo que supone que la opción clásica de la aproximación normal puede ser de escasa validez, sobre todo, para tamaños muestrales relativamente cortos. En este contexto, veamos como las aproximaciones bootstraps pueden mejorar la aproximación normal.

El estimador bootstrap de G obviamente tiene el desarrollo:

$$\hat{G}(x) = P(T_n^* \leq x | X_1, \cdots, X_n) = \Phi(x) + n^{-1/2} \hat{q}(x)\phi(x) + O_p(n^{-1}) \qquad (2.11)$$

siendo T_n^* la versión bootstrap de T_n obtenida de la muestra $X_1, \cdots, X_n$, y $\hat{q}$ es la versión de q que resulta de sustituir los parámetros por sus estimaciones.

Habitualmente, $\hat{q} = q + O_p\left(n^{-1/2}\right)$. De los desarrollos de G y su aproximación bootstrap $\hat{G}$, puede fácilmente obtenerse la siguiente expresión:

$$\hat{G}(x) - G(x) = P\left(T_n^* \leq x \middle| X_1, \cdots, X_n\right) - P\left(T_n \leq x\right) = O_p\left(n^{-1}\right) \tag{2.12}$$

Ello significa que el error de la aproximación bootstrap a la distribución del pivotal T_n es de orden n^{-1} en probabilidad. Podemos por tanto concluir este epígrafe con la afirmación de que el bootstrap no sólo puede permitir aproximar la distribución probabilística de estadísticos de interés cuando la obtención de ésta es compleja, sino que además permite mejorar la aproximación normal de los clásicos estadísticos lineales.

2.5. Consistencia del bootstrap mediante las métricas de Mallows.

La consistencia del método bootstrap requiere la definición de una métrica δ definida sobre un conjunto de distribuciones de probabilidad, de tal forma que la distancia entre la distribución del estadístico R_n y la de su versión bootstrap R_n^* tienda a cero para n tendiendo a infinito. Revisamos en esta sección la consistencia del bootstrap basada en la métrica de Mallows. Definimos en primer lugar la distancia de un conjunto de distribuciones de probabilidad sobre un espacio de Banach. Veremos también un resultado debido a Bickel & Freedman (1981), el cual establece la equivalencia de la convergencia según la métrica de Mallows, con la convergencia débil.

Consideremos un espacio de Banach B cuya norma denotamos por $\|\cdot\|$, y sea $\Gamma_p = \Gamma_p(B)$ el conjunto de las medidas de probabilidad γ definidas sobre la correspondiente σ-álgebra de Borel $\sigma(B)$ que verifican $\int \|x\|^p \gamma(dx) < \infty$. La métrica de Mallows se define en Γ_p del modo siguiente: Dado $\alpha, \beta \in \Gamma_p$

$$d_p(\alpha,\beta) = \inf_{\substack{X\cong\alpha \\ Y\cong\beta}} \left\{ E\left[(X-Y)^p \right] \right\}^{1/p} \tag{2.13}$$

Los siguientes lemas, debidos a Bickel & Freedman, establecen las propiedades básicas de la métrica d_p

Lema 2.2.

i. El ínfimo al que hace referencia la definición de la métrica de Mallows es accesible.

ii. d_p es una métrica sobre Γ_p.

Definición. Sea $\alpha_n, \alpha \in \Gamma_p$. Se dice que α_n converge débilmente a la medida α ($\alpha_n \Rightarrow \alpha$) si $\alpha_n(H) \to \alpha(H)$ para cualquier $H \in \sigma(B)$, tal que $\alpha(\partial H) = 0$ (∂H representa la frontera de *H*).

El siguiente lema da caracterizaciones de la métrica de Mallows.

Lema 2.3. Sean $\alpha_n, \alpha \in \Gamma_p$. Entonces, $d_p(\alpha_n, \alpha) \to 0$, $n \to \infty$, es equivalente a cada una de las siguientes condiciones:

i. $\alpha_n \Rightarrow \alpha$ y $\int \|x\|^p \alpha_n(dx) \to \int \|x\|^p \alpha(dx)$.

ii. $\alpha_n \Rightarrow \alpha$ y $\|x\|^p$ es uniformemente α_n-integrable.

iii. $\int \phi(x)\, d\alpha_n \to \int \phi(x)\, d\alpha$, para cualquier función continua $\phi(x)$, tal que $\phi(x) = O\left(\|x\|^p\right)$.

En lo que sigue emplearemos la siguiente notación: sean *U* y *V* variables aleatorias con valores en el espacio de Banach *B*. Por $d_p(U,V)$ representamos la distancia entre las distribuciones de *U* y *V* respectivamente, supuesto que éstas pertenecen a Γ_p. Una propiedad de escalamiento para la métrica de Mallows es la siguiente:

$$d_p(\alpha U, \alpha V) = |\alpha| \cdot d_p(U,V) \tag{2.14}$$

Consideremos ahora $B = \mathbb{R}$ y p=2. Dada una muestra aleatoria $X_1, \cdots, X_n$ de una distribución de probabilidad F, tal que $\int x^2 F(dx) < \infty$. Es fácil probar que $\hat{F}_n \Rightarrow F$ casi seguramente, siendo $\hat{F}_n$ la distribución empírica correspondiente. Para probar este resultado basta comprobar que $\sum_{n=1}^{\infty} (1/n^2) \cdot \text{var}(I(X_i \le x)) < \infty$. Esto es obvio dado que $\text{var}(I(X_i \le x)) = F(x) \cdot (1 - F(x)) \le 1$, lo cual prueba la convergencia de la serie anterior y por tanto,

$$\hat{F}_n(x) = (1/n) \sum_{i=1}^{n} I(X_i \le x) \to E[I(X_i \le x)] = F(x), \tag{2.15}$$

además, $\int x^2 \hat{F}_n(dx) \to \int x^2 F(dx)$. En efecto,

$$\int x^2 \hat{F}_n(dx) = \frac{1}{n} \sum_{i=1}^{n} X_i^2 \to E[X_i^2] = \int x^2 F(dx), \tag{2.16}$$

casi seguramente.

Aplicando entonces el lema 2.3. se establece que $d_2(\hat{F}_n, F) \to 0$, casi seguramente cuando $n \to \infty$. Ello significa que, para diseños de tamaño fijo, la distribución empírica es un estimador consistente de la verdadera distribución de probabilidad, de acuerdo con la métrica de Mallows.

A través del siguiente lema (también de Bickel & Freedman), puede examinarse la consistencia del método bootstrap aplicado a determinados estadísticos.

Lema 2.4. Sea B un espacio de Hilbert con producto interior $\langle \cdot, \cdot \rangle$, y p=2. Supongamos que U_j son independientes y que también, V_j son independientes perteneciendo ambas sucesiones a Γ_2. Entonces:

$$d_2\left(\sum_{j=1}^{m} U_j, \sum_{j=1}^{m} V_j\right)^2 \leq \sum_{j=1}^{m} d_2\left(U_j, V_j\right)^2 \tag{2.17}$$

Lema 2.5. Sea B un espacio de Hilbert con producto interior $\langle \cdot, \cdot \rangle$, y p=2. Sean U y V variables aleatorias con valores en B, y con norma L_2 finita. Entonces

$$d_2(U,V)^2 = d_2\left(U - E[U], V - E[V]\right)^2 + \left\| E[U] - E[V] \right\|^2 \tag{2.18}$$

Consideremos ahora una muestra aleatoria $X_1, \cdots, X_n$ de una distribución de probabilidad F con media μ y varianza finita σ^2. Consideremos el estadístico $Z_n = \sqrt{n}\left(\bar{X} - \mu\right)$ el cual tiene asintóticamente distribución de probabilidad $N(0, \sigma)$. La versión bootstrap del estadístico es $Z_n^* = \sqrt{n}\left(\bar{X}^* - \bar{X}\right)$. Entonces, de acuerdo con los lemas 2.4. y 2.5 y la propiedad de escalamiento:

$$\begin{aligned} d_2\left(Z_n, Z_n^*\right)^2 &= \frac{1}{n} \cdot d_2\left(\sum_{i=1}^{n} (X_i - \mu), \sum_{i=1}^{n} \left(X_i^* - \bar{X}\right)\right)^2 \\ &\leq \frac{1}{n} \sum_{i=1}^{n} d_2\left(X_i - \mu, X_i^* - \bar{X}\right)^2 = d_2\left(X_i, X_i^*\right)^2 = d_2\left(F, \hat{F}_n\right)^2 \to 0 \end{aligned}$$

casi seguramente, para $n \to \infty$.

Este estudio de consistencia puede fácilmente extenderse a estadísticos pivotales de la forma $\sqrt{n}\left(\bar{X} - \mu\right)/\sigma$, siendo naturalmente la versión bootstrap de este estadístico $\sqrt{n}\left(\bar{X}^* - \bar{X}\right)/\hat{\sigma}_n$, donde $\hat{\sigma}_n^2 = (1/n)\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$. En el epígrafe 2.3. se consideró la aproximación bootstrap a los modelos de regresión. Para el caso de modelos de regresión lineal, Freedman (1981) aproxima la distribución probabilística del estadístico $\sqrt{n}\left(\hat{\beta}_n - \beta\right)$ por el procedimiento bootstrap que se describe en el referido epígrafe, y prueba la consistencia del método utilizando técnicas similares a las que hemos mostrado para el caso del

estadístico Z_n. Franke & Härdle (1992) utilizan también el bootstrap en el contexto de la estimación de la función de densidad espectral para procesos estacionarios lineales. Utilizando la representación asintótica del periodograma para procesos lineales, obtienen una versión bootstrap de un cierto pivotal y prueban la consistencia del método de una forma similar al que utiliza Freedman para los modelos de regresión lineal.

Parece atractivo pensar en la extensión de estos procedimientos a diseños secuenciales. Esto presenta sin embargo enormes dificultades debido a que los tamaños N son aleatorios. No parece además que sea práctico evaluar los procedimientos secuenciales en función de sus propiedades asintóticas debido a que el objetivo de tales procedimientos es generalmente reducir los tamaños muestrales. En el siguiente capítulo propondremos el uso del bootstrap para aproximar la distribución de probabilidad de pivotales ordinarios sobre los que se puedan determinar intervalos de confianza para los parámetros de interés. Allí expondremos de forma empírica la validez de las aproximaciones mediante estudios de simulación.

3. El bootstrap en diseños secuenciales

3.1. El problema de la inferencia con datos secuenciales

Tal como se mostró en el primer capítulo, los diseños secuenciales pueden ser preferibles a los de tamaño fijo en el sentido de que, para unos mismos requerimientos de significación y potencia, conducen a un tamaño muestral esperado inferior al que correspondería a un diseño de tamaño fijo. En algunos estudios tales como los de control de calidad, el objetivo generalmente consiste únicamente en aceptar o no un cierto lote de ítems mediante un contraste de hipótesis. En los ensayos clínicos sin embargo se requiere generalmente, una vez completada la recolección de datos, hacer inferencias acerca de los parámetros de interés. Los diseños secuenciales pueden resultar problemáticos en el sentido de que estimadores que son insesgados cuando se basan en datos obtenidos a través de diseños de tamaño fijo, presentan sesgos cuando éstos se han obtenido de modo secuencial. La normalidad asintótica de muchos pivotales puede también dejar de ser válida. Emerson & Fleming (1990) proponen un método de estimación de la media de una distribución normal con varianza conocida basado en datos obtenidos a través de un diseño de grupos secuenciales. En este capítulo proponemos el uso del bootstrap para aproximar la distribución de probabilidad de los pivotales a través de los cuales puedan obtenerse los intervalos de confianza para parámetros de interés frecuente en ensayos clínicos. Ilustraremos los procedimientos mediante reglas de parada basadas en los procedimientos de Wang & Tsiatis, los cuales fueron descritos en la sección 1.9. La base de cualquier procedimiento bootstrap es la elección de una adecuada distribución de remuestreo. En el contexto de un diseño de tamaño fijo, la distribución empírica puede ser una elección adecuada como vimos en el capítulo anterior. Lamentablemente, las propiedades de la distribución empírica no se mantienen por lo general en el contexto de un diseño secuencial. En la sección 3.3 se analiza la distribución empírica como estimador de la distribución generadora de los

datos. Damos una forma de consistencia en probabilidad, aunque tal resultado no tiene mucho valor práctico, en el sentido de que el objetivo fundamental de los diseños secuenciales es hacer un contraste sobre el parámetro de interés con el mínimo número de datos. Damos también otro resultado a través del cual puede examinarse como la varianza de la regla de parada influye en el sesgo de la distribución empírica, lo cual se muestra también mediante simulación de una regla de parada basada en el procedimiento de Pocock. Utilizaremos el bootstrap para aproximar la distribución de probabilidad de varios pivotales a través de los cuales pueden obtenerse intervalos de confianza para parámetros de frecuente interés en los ensayos clínicos. Hacemos en primer lugar una revisión de un procedimiento de construcción de intervalos de confianza basado en datos obtenidos a través de diseños secuenciales.

3.2. Métodos basados en la dualidad test de hipótesis-intervalo de confianza

Dado que la aproximación normal dada por el teorema de Anscombe y Doeblin es muy poco precisa, revisamos en esta sección los métodos exactos basados en la dualidad de contrastes de hipótesis e intervalos de confianza. Este método requiere la definición de una relación de orden para el estadístico (N,T), donde N representa la etapa de parada y T el test estadístico. El método requiere conocer la forma de la distribución F. Emerson & Fleming (1990) consideran el caso en el que las observaciones $X_{ij} \cong N(\mu,\sigma)$, para $i=1,\cdots,K$, $j=1,\cdots,m_i$. Veremos en primer lugar este procedimiento para diseños de tamaño fijo.

Un intervalo de confianza para un parámetro θ puede definirse como el conjunto de todos los valores del parámetro conjeturados en la hipótesis nula que no serían rechazados para el valor observado de un adecuado test estadístico T. Más concretamente, sea T un test estadístico para el contraste $H_0:\theta=\theta_0$. Para tal contraste determinamos un región de aceptación $A^{\alpha}_{\theta_0}$, tal que

$P\left(T \in A_{\theta_0}^{\alpha} | \theta_0\right) = 1-\alpha$. A partir del contraste y un valor observado del test estadístico T, podemos encontrar una región de confianza $I^{\alpha}(T)$ al nivel 1-α, definida como $I^{\alpha}(T) = \left\{\theta : T \in A_{\theta}^{\alpha}\right\}$. En contrastes de tamaño fijo, la región de aceptación se obtiene usualmente como $A_{\theta}^{\alpha} = \left\{t : P(T \geq t|\theta) > \alpha\right\}$. Para un t observado y un valor del parámetro θ, $P(T > t|\theta)$ es la función p-valor $p(\theta)$ (probabilidad de observaciones más extremas bajo la hipótesis de que el verdadero valor del parámetro es θ). Si esta función es no decreciente, una forma de construir un intervalo de confianza al nivel 1-α es seleccionando valores θ_L y θ_U tales que $p(\theta_L) = \alpha/2$ y $p(\theta_U) = 1-\alpha/2$.

En resumen, en los diseños de tamaño fijo, la información relevante de los datos para un contraste de hipótesis se resume a menudo en un estadístico suficiente unidimensional, y el grado de significación se expresa a través de la función p-valor. La referencia a resultados más extremos requiere una relación de orden en el conjunto de los números reales. Por el contrario, en los diseños secuenciales, la información de los datos se resume en un estadístico suficiente bidimensional; a saber: la etapa de parada N y el test estadístico T. No es ahora inmediato hablar de *resultados más extremos*. Si tal relación está definida, la construcción de la región de confianza se realizaría de la forma natural definiendo previamente la región de aceptación para una hipótesis nula por:

$$A_{\theta}^{\alpha} = \left\{(n,t) : P\left((N,T) \geq (n,t)\right) > \alpha\right\} \tag{3.1}$$

y la correspondiente región de confianza para la observación (N,T) por:

$$I^{\alpha}(N,T) = \left\{\theta : (N,T) \in A_{\theta}^{\alpha}\right\} \tag{3.2}$$

Armitage (1957) propuso una ordenación que posteriormente fue investigada por Jennison & Turnbull (1983) en el contexto de los test de grupos secuenciales. Describimos seguidamente esta relación de orden. Para un test de

grupos secuenciales y para el k-ésimo look representamos por C_k, $S_k^{(0)}$ y $S_k^{(1)}$ las regiones de continuación, aceptación y rechazo, respectivamente. Definimos la relación de orden $(N_1,T_1)<(N_2,T_2)$ si $N_1<N_2$ y $T_1\in S_{N_1}^{(0)}$, o si $N_1=N_2$ y $T_1<T_2$, o si $N_1>N_2$ y $T_2\in S_{N_2}^{(1)}$. Esta relación de orden no está definida para regiones de continuación que no sean intervalos como sucede en el test triangular bilateral de Whitehead & Straton (1983). Emerson & Fleming (1990) propone una familia de relaciones de orden indexada por un parámetro q en los siguientes términos: $(N_1,T_1)<(N_2,T_2)$ si:

$$T_1/\left(\sum_{i=1}^{N_1} n_i\right)^q < T_2/\left(\sum_{i=1}^{N_2} n_i\right)^q \tag{3.3}$$

El significado de esta ordenación parece claro. Para q=1, un resultado es más extremo que otro cuando la media de las observaciones es mayor. La elección de valores para q>1 van en la dirección de penalizar el tamaño muestral acumulado; esto es: un resultado puede tener una media mayor que otro, pero puede ser menos extremo al basarse en un tamaño muestral mucho mayor. No obstante, Emerson & Fleming consideran el valor q=1.

3.3. Distribución empírica en diseños secuenciales

Los algoritmos bootstrap para diseños de tamaño fijo que se presentaron en el capítulo anterior utilizaban la distribución empírica como distribución de remuestreo, aunque cabía la posibilidad de usar estimadores de núcleo de la verdadera función de distribución. Sus propiedades quedaron establecidas en el epígrafe 2.2. En los diseños secuenciales sin embargo, el uso de la función de distribución empírica como distribución de remuestreo puede ser problemático, pues ésta previsiblemente podrá ser sesgada. Veremos como ocurre esto en el contexto de los contrastes de Wang & Tsiatis cuando los datos se apartan de la hipótesis nula. Sin embargo, cuando se aproxima la distribución de probabilidad

de ciertos pivotales mediante el bootstrap utilizando la distribución empírica como distribución de remuestreo, los sesgos de ésta pueden ser compensados, de tal forma que la aproximación bootstrap sea válida. Examinamos pues en primer lugar el comportamiento de la distribución empírica basada en datos secuenciales como estimador de la verdadera distribución de probabilidad.

Sea pues $N; X_1, \cdots, X_N$ un conjunto de datos secuenciales con regla de parada N y generados por una distribución de probabilidad $F(x)$, y sea:

$$\hat{F}_N(x) = (1/N) \sum_{k=1}^{N} I(X_k \leq x) \tag{3.4}$$

la correspondiente distribución de probabilidad empírica. A través del siguiente teorema (Anscombe (1952) y Doeblin (1938)) probaremos que $\hat{F}_N(x)$ es un estimador consistente para la función de distribución de probabilidad $F(x)$.

Teorema 3.1. Sea $X_1, X_2, \cdots$ variables aleatorias independientes e idénticamente distribuidas con media μ, varianza $\sigma^2 > 0$ y $S_n = \sum_{k=1}^{n} X_k$. Supongamos que $N_c, c \geq 0$ es una sucesión de variables aleatorias entero valoradas y tales que para alguna sucesión de constantes $n_c \to \infty$, $N_c / n_c \xrightarrow{P} 1$. Entonces, para $c \to \infty$

$$P\left\{ N_c^{-1/2} \left(S_{N_c} - N_c \mu \right) \leq t \right\} \to \Phi(t/\sigma) \tag{3.5}$$

siendo Φ la función de distribución normal estándar. ■

Siegmund (1985, pg 23), señala que ésta es una aproximación deficiente. Por otro lado, los procedimientos secuenciales tienen como finalidad generalmente minimizar el tamaño muestral, por lo que puede resultar paradójico el uso de resultados asintóticos.

Corolario 3.2. En las mismas condiciones del teorema anterior, para cualquier x, y $c \to \infty$

$$P\left\{N_c^{1/2}\left(\hat{F}_{N_c}(x) - F(x)\right) \le t\right\} \to \Phi\left(t/\sigma(x)\right) \tag{3.6}$$

siendo $\sigma(x) = \left\{F(x)\left(1-F(x)\right)\right\}^{1/2}$.

De este corolario se sigue inmediatamente este otro.

Corolario 3.3. En las mismas condiciones del teorema 3.1, para cualquier x y para $c \to \infty$, $\hat{F}_{N_c}(x) \to F(x)$ en probabilidad.

El siguiente resultado basado en la identidad de Wald permite analizar el sesgo de $\hat{F}_N(x)$, en el caso de pequeños valores de $\mathrm{var}(N)$.

Proposición 3.4. Para un diseño secuencial con regla de parada N, tal que $P(N>0)=1$ y $\mathrm{var}(N) < \infty$. Entonces, para algún $C > 0$

$$\left|E\left[\hat{F}_N(x)\right] - F(x)\right| \le C \cdot \frac{\mathrm{var}(N)^{1/2}}{E[N]} \tag{3.7}$$

Demostración

Sea $A_N(x) = \sum_{k=1}^{N} I(X_k \le x)$ y $\nu = E[N]$, entonces:

$$\hat{F}_N(x) = \frac{A_N(x)}{\nu \cdot N/\nu} = \frac{A_N(x)}{\nu}\left(1 - \frac{1}{\xi^2}\left(\frac{N}{\nu} - 1\right)\right)$$

siendo ξ una variable aleatoria tal que $|\xi - 1| \le \left|\frac{N}{\nu} - 1\right|$, casi seguramente. De acuerdo con la desigualdad de Wald,

$$E\left[A_N(x)/\nu\right] = F(x).$$

De esta forma:

$$E\left[\hat{F}_N(x)\right] = F(x) - \frac{1}{\nu} E\left[\frac{A_N(x)}{\xi^2} \cdot \left(\frac{N}{\nu} - 1\right)\right]$$

Ahora bien,

$$\left|\frac{1}{\nu} E\left[\frac{A_N(x)}{\xi^2} \cdot \left(\frac{N}{\nu} - 1\right)\right]\right| = \left|\frac{1}{\nu} E\left[\frac{A_N(x)}{\nu\xi^2}(N-\nu)\right]\right|$$

$$\leq \frac{1}{\nu} E\left[\frac{A_N(x)^2}{\nu^2\xi^4}\right]^{1/2} \cdot E\left[(N-\nu)^2\right]^{1/2} = E\left[\frac{A_N(x)^2}{\nu^2\xi^4}\right]^{1/2} \cdot \frac{\mathrm{var}(N)^{1/2}}{\nu}$$

Sea ahora:

$$E\left[\frac{A_N(x)^2}{\nu^2\xi^4}\right] = E\left[\left(\frac{A_N(x)}{N}\right)^2 \left(\frac{N/\nu}{\xi^2}\right)^2\right] \leq E\left[\left(\frac{N/\nu}{\xi^2}\right)^2\right]$$

Dado que $P(N>0)=1$, existe entonces B tal que $P(N/\nu > B)=1$. Se sigue entonces que $P(\xi > B)=1$. Finalmente:

$$E\left[\left(\frac{N/\nu}{\xi^2}\right)^2\right] \leq \frac{1}{B^2}\left(1 + \frac{\mathrm{var}(N)}{\nu^2}\right) = C$$

lo que completa la demostración.

Obviamente, si el diseño es de tamaño fijo, $\mathrm{var}(N)=0$ y de aquí se sigue que la distribución empírica es un estimador centrado para $F(x)$.

En los diseños del tipo de Wang & Tsiatis y bajo la hipótesis nula del contraste, la probabilidad de que se alcance el tamaño muestral máximo es aproximadamente 1-α, siendo α la significación del test dado que en estos contrastes la hipótesis nula no es aceptada en inspecciones intermedias. Es también obvio que la probabilidad de rechazar H_0 en las primeras inspecciones sea muy pequeña, sobre todo en el contraste de O'Brien & Fleming.

Para ilustrar estas ideas, considérese la distribución $N(\mu,\sigma=15)$ y la hipótesis nula $H_0:\mu=85$. Bajo H_0, esto es, siendo el valor real del parámetro $\mu=85$, se ha aplicado el procedimiento de Pocock 30.000 veces, con un número máximo de looks K=8 y el tamaño de cada look m=20. Hemos aproximado $E\left[\hat{F}_N(t)\right]$, promediando las ordenadas de las correspondientes distribuciones empíricas. La figura 3.1.a muestra la función de distribución de probabilidad real $\Phi\left((t-85)/15\right)$ y la aproximación $E\left[\hat{F}_N(t)\right]$. No es posible en el gráfico diferenciar ambas funciones, lo que confirma la práctica insesgadez del estimador. Sin embargo, cuando bajo la misma hipótesis nula $H_0:\mu=85$, el valor real del parámetro es $\mu=90$, la distribución empírica tiende a sobrestimar la distribución real como puede verse en el gráfico 3.1.b. Las figuras 3.1.c y 3.1.d realizan los mismos análisis pero con la distribución exponencial. En la figura 3.1.c se representan simultáneamente la función de distribución exponencial de media $\mu=5$ $(F(t)=1-\exp(-t/5))$ y $E\left[\hat{F}_N(t)\right]$ donde los datos se obtuvieron repitiendo 30.000 veces el mismo procedimiento de Pocock (K=8 y m=20) bajo la hipótesis nula $H_0:\mu=5$. Puede observarse nuevamente que la distribución empírica es prácticamente un estimador centrado la función de distribución de probabilidad. Sin embargo, manteniendo la misma hipótesis nula $H_0:\mu=5$ pero siendo el valor real de la media $\mu=7$, puede observarse en 3.1.d el sesgo de la distribución empírica.

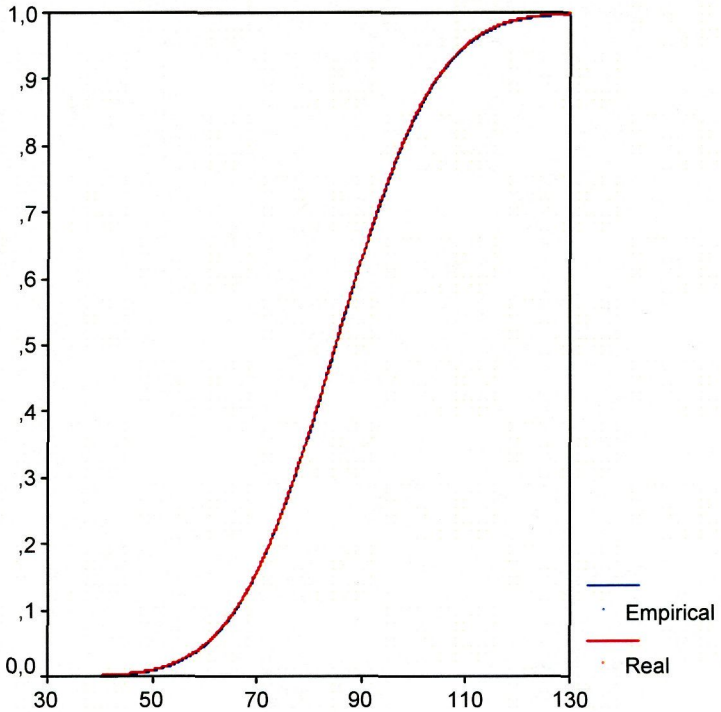


Figura 3.1.a. Distribución $N(85;15)$ y la esperanza de la empírica bajo la hipótesis nula en un procedimiento de Pocock.

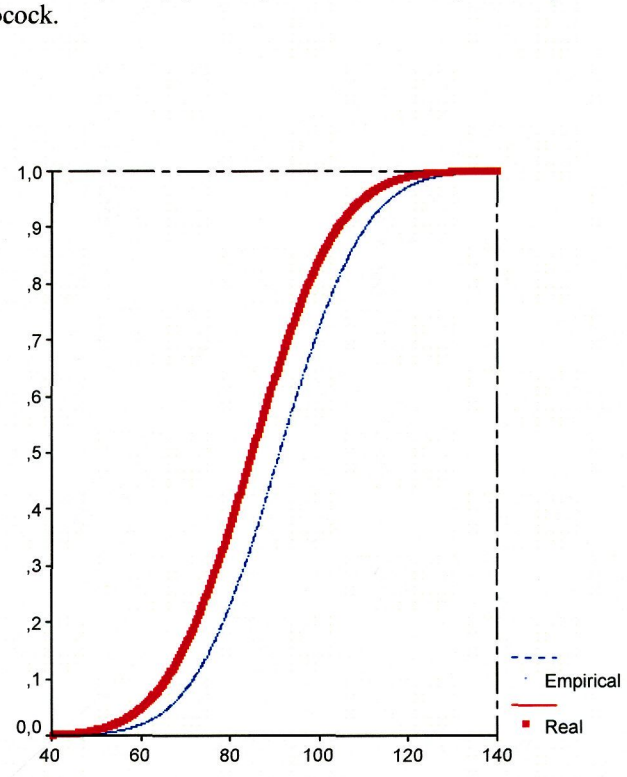


Figura 3.1.b. Distribución $N(90;15)$ y la esperanza de la empírica bajo una alternativa en un procedimiento de Pocock.

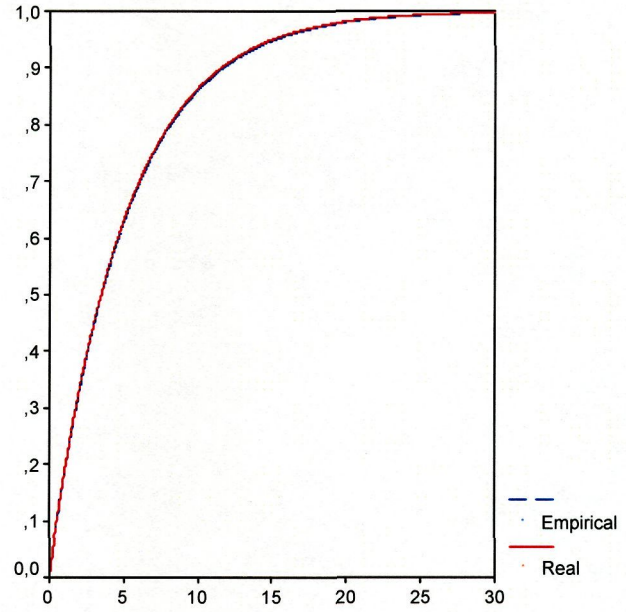


Figura 3.1.c. Distribución exponencial de parámetro 5 y la esperanza de su empírica bajo la hipótesis nula en un procedimiento de Pocock.

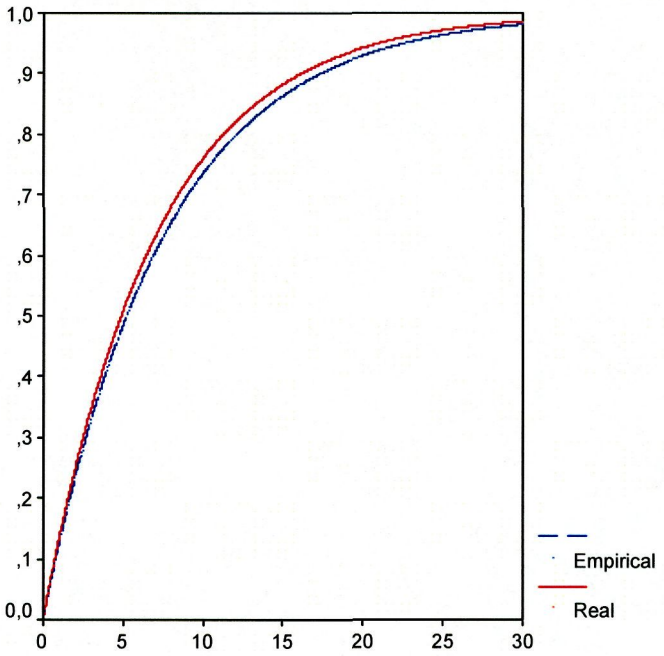


Figura 3.1.d. Distribución exponencial de parámetro 5 y la esperanza de su empírica bajo una alternativa en un procedimiento de Pocock.

Teorema 3.5. Sea $\{X_n : n \in \mathbb{N}\}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas, tales que $E[X_i] = \mu$ y $\text{var}(X_i) = \sigma^2$. Sea N una variable aleatoria con valores en $\mathbb{Z}$ tal que $\{N = n\}$ es un suceso determinado por $X_1, \cdots, X_n$ y es independiente de $X_{n+1}, X_{n+2}, \cdots$ para todo n=1,2,..., y supongamos que $E[N] < \infty$ y $E\left[\left(\sum_{k=1}^{N} |X_k - \mu|\right)^2\right] < \infty$. Entonces,

$$E\left[(S_N - N\mu)^2\right] = \sigma^2 E[N] \tag{3.8}$$

siendo $S_n = \sum_{k=1}^{n} X_k$. ■

Demostración.

$\{N \geq k\}^c = \bigcup_{j=1}^{k-1} \{N = j\}$ es independiente de $X_k, X_{k+1}, \cdots$. Por tanto, las variables X_k y $I(N \geq k)$ son independientes. Sea $Y_k = X_k - \mu$. Entonces:

$$E\left[\left\{\sum_{k=1}^{N} Y_k\right\}^2\right] = E\left[\sum_{k=1}^{N} Y_k^2\right] + E\left[\sum_j \sum_{k \neq j} Y_j Y_k\right]$$

$$= E[N] \cdot E\left[Y_k^2\right] + E\left[\sum_{j \neq k} Y_j Y_k I\left(N \geq \max(j,k)\right)\right]$$

Veamos que el último término es nulo. En efecto

$$E\left[\sum_{j \neq k} Y_j Y_k I\left(N \geq \max(j,k)\right)\right] = 2\sum_{k=2}^{\infty} \sum_{j=1}^{k-1} E\left[Y_j Y_k I(N \geq k)\right]$$

$$= 2\sum_{k=2}^{\infty} E\left[Y_k I(N \geq k) \cdot \sum_{j=1}^{k-1} Y_j\right] = 2\sum_{k=2}^{\infty} E\left[Y_k I(N \geq k) \cdot (S_{k-1} - (k-1) \cdot \mu)\right]$$

$$= 2\sum_{k=2}^{\infty} \left\{E\left[Y_k I(N \geq k) S_{k-1}\right] - (k-1) \cdot \mu \cdot E\left[Y_k I(N \geq k)\right]\right\}$$

Obsérvese que $I(N \geq k)$ está determinada por $X_1, \cdots, X_{k-1}$, por lo que:

$$E\left[Y_k I(N \geq k) S_{k-1}\right] = E\left[I(N \geq k) S_{k-1} E\left[Y_k \middle| X_1, \cdots, X_{k-1}\right]\right] = 0$$

Además, por ser Y_k independiente de $I(k \geq N)$

$$E\left[Y_k I(N \geq k)\right] = 0 \qquad \text{y} \qquad E\left[Y_k \middle| S_{k-1}\right] = E\left[Y_k\right] = 0$$

Por tanto, $E\left[\sum_{j \neq k} Y_j Y_k I\left(N \geq \max(j,k)\right)\right] = 0$

Lo cual completa la demostración.

Corolario 3.6. En las mismas condiciones del teorema 3.5 y siendo además $\text{var}(N) < \infty$, se tiene:

$$\text{var}(S_N) = \sigma^2 E[N] + \mu^2 \text{var}(N) + 2\mu E\left[\left(S_N - N\mu\right)\left(N - E[N]\right)\right] \qquad (3.9)$$

3.4. Aproximación bootstrap con datos secuenciales para la media de una distribución

En esta sección consideraremos el problema de estimar un intervalo de confianza para la media de una distribución de probabilidad basado en datos obtenidos a través de un diseño secuencial. Para la construcción del intervalo utilizamos el método de los pivotales y proponemos un procedimiento bootstrap para aproximar la distribución probabilística del pivotal correspondiente. Examinamos la validez del procedimiento y realizamos un estudio de simulación, en el cual comparamos el método bootstrap propuesto con el de Emerson & Fleming descrito en 3.2.

3.4.1. Aproximación bootstrap

Consideremos el parámetro $\mu = \int x \cdot F(dx)$ correspondiente a una distribución F, con varianza $\sigma^2 = \int (x-\mu)^2 F(dx)$ y sea $N; X_1, \cdots, X_N$ un conjunto de datos obtenidos de F a través de un diseño secuencial siendo N la regla de parada. En orden a obtener un intervalo de confianza para μ definimos el pivotal:

$$T_N = \frac{S_N - N \cdot \mu}{\hat{\sigma}_N \cdot \sqrt{N}} \tag{3.10}$$

donde $S_N = \sum_{i=1}^{N} X_i$, $\overline{X} = S_N / N$ y $\hat{\sigma}_N^2 = (1/N)\sum_{i=1}^{N} (X_i - \overline{X})^2$. Para grandes valores de N y siempre y cuando éste sea fijo, la distribución de T_N puede aproximarse por la t de Student o la normal. De acuerdo con las consideraciones de la sección 2.4, las aproximaciones bootstrap serían mejores que la normal. En cualquier caso, no cabe esperar grandes valores de N si los datos proceden de un ensayo clínico secuencial. Además, la naturaleza aleatoria de N conduce a falta de normalidad en T_N.

La figura 3.2. muestra la función de densidad del pivotal para el caso en el que los datos se hayan extraído de la $N(90;15)$ utilizando el procedimiento de Pocock para contrastar $H_0 : \mu = 85$. Una simple observación de la gráfica, la tabla de percentiles que se muestra a continuación y el contraste de normalidad de Kolmogorov-Smirnov (p=0.001) conduce a descartar similitudes con las distribuciones t o normal.

P5	*P10*	*P25*	*P50*	*P75*	*P90*	*P95*
-1.4283	-1.0192	-.03983	0.3107	0.9771	1.5534	1.9119

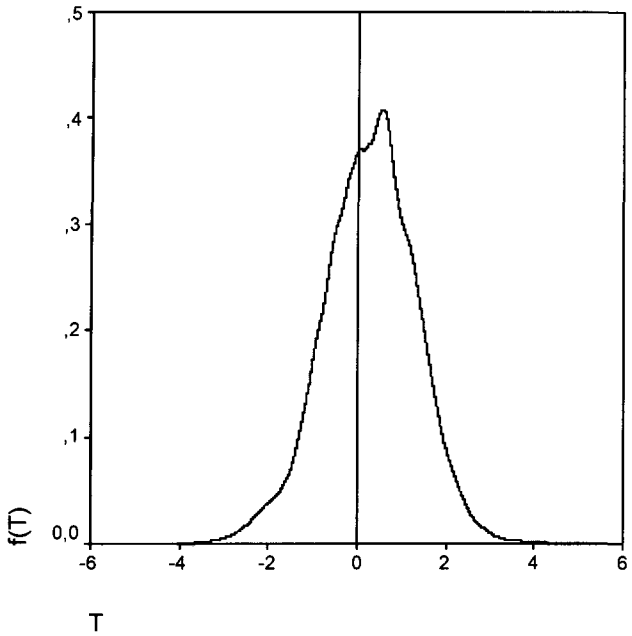


Figura 3.2. Densidad del pivotal T_N obtenida por estimación de núcleo.

Proponemos un método de aproximación bootstrap para la distribución de probabilidad de T_N. Dada la aleatoriedad de N, no son válidos los métodos clásicos de evaluación de la aproximación bootstrap. En el epígrafe 3.4.2 haremos una valoración del método basada en una descomposición del pivotal T_N.

El algoritmo propuesto es el siguiente:

Paso 1. De $X_1, X_2, \cdots, X_N$, definir la distribución de remuestreo. Una posible distribución es la empírica $\hat{F}_N(x)$.

Paso 2. Utilizando la misma regla de parada que dio lugar a la muestra $X_1, X_2, \cdots, X_N$, obtener de la distribución de remuestreo B muestras bootstrap $N^*, X_1^*, X_2^*, \cdots, X_{N^*}^*$.

Paso 3. Obtener B valores del pivotal bootstrap definido por:

$$T_{N^*}^* = \frac{S_{N^*}^* - N^* \cdot S_N / N}{\hat{\sigma}_{N^*}^* \sqrt{N^*}} \tag{3.11}$$

Los B valores bootstrap proporcionarán una aproximación a la distribución de probabilidad de T_N.

De acuerdo con la identidad de Wald, $E[S_N] = E[N] \cdot \mu$. Por tanto, la esperanza del numerador del pivotal T_N es cero. Si se utiliza como distribución de remuestreo la función de distribución empírica, $E_*[X_i^*] = S_N/N$. De esta forma, utilizando nuevamente la identidad de Wald y teniendo en cuenta que $E_*[X_i^*] = S_N/N$, se obtiene:

$$E_*[S_{N^*}^*] = E_*[N^*] \cdot S_N/N \tag{3.12}$$

Como se indicó en la sección 3.3, la regla de parada puede conducir a una estimación sesgada de la media de la distribución, μ, pero simultáneamente produce el mismo tipo de sesgo en la distribución empírica. En realidad, $\hat{\mu}_N = \int x\hat{F}_N(dx) = S_N/N$. Este sesgo influye directamente en el estadístico $S_{N^*}^*$. En la figura 3.1.b se observa que la distribución empírica sobrestima la distribución real. Esto obviamente conduce a que S_N/N sobrestimará a μ y el estadístico $S_{N^*}^*$ tenderá a producir valores mayores que S_N. Sin embargo, ambas sobrestimaciones tienden a compensarse. En efecto:

$$E[S_N - N \cdot \mu] = E_*[S_{N^*}^* - N^* \cdot S_N/N] = 0 \tag{3.13}$$

además, de acuerdo con el teorema 3.5,

$$\text{var}(S_N - N \cdot \mu) = \sigma^2 \cdot E[N] \tag{3.14}$$

y

$$\text{var}_*(S_{N^*}^* - N^* \cdot S_N/N) = \hat{\sigma}_N^2 \cdot E_*[N^*]. \tag{3.15}$$

De esta forma, los denominadores de ambos pivotales estiman en el mismo modo la desviación estándar de los numeradores. En el siguiente epígrafe haremos una valoración más detallada de la aproximación bootstrap.

3.4.2. Evaluación de la aproximación bootstrap

Mostramos ahora una evaluación empírica de cómo la distribución del pivotal T_N^* aproxima la de T_N. Sustituyendo $\hat{\sigma}_N$ por σ y mediante una elemental aplicación del teorema del valor medio tenemos:

$$T_N \approx \frac{S_N - N\mu}{\sigma\sqrt{E[N]}} - \frac{1}{2\sigma E[N]^{3/2}} \cdot (S_N - N\mu)(N - E[N]) \tag{3.16}$$

Nótese que el segundo término es de orden $E[N]^{-3/2}$ mientras que el del primero es $E[N]^{-1/2}$. Obviamente se tiene:

$$E\left[\frac{S_N - N\mu}{\sigma\sqrt{E[N]}}\right] = 0 \tag{3.17}$$

y

$$\text{var}\left(\frac{S_N - N\mu}{\sigma\sqrt{E[N]}}\right) = 1 \tag{3.18}$$

La aproximación análoga del pivotal bootstrap es:

$$T_{N^*}^* \approx \frac{S_{N^*}^* - S_N}{\hat{\sigma}_N\sqrt{E_*[N^*]}} - \frac{1}{2\hat{\sigma}_N E_*[N^*]^{3/2}} \cdot (S_{N^*}^* - S_N)(N^* - E_*[N^*]) \tag{3.19}$$

lo que implica también:

$$E_*\left[\left(S^*_{N^*} - S_N\right)\Big/\hat{\sigma}_N\sqrt{E_*\left[N^*\right]}\right] = 0 \tag{3.20}$$

y

$$\mathrm{var}_*\left(\frac{S^*_{N^*} - S_N}{\hat{\sigma}_N\sqrt{E_*\left[N^*\right]}}\right) = 1 \tag{3.21}$$

Por tanto, la falta de centralidad en el cero de ambos pivotales es atribuible a los segundos términos, más concretamente a la covarianza entre $S_N - N\mu$ y N. El mismo razonamiento cabe para el análogo bootstrap. Parece claro en este punto la importancia de que la regla de parada bootstrap imite la regla de parada original.

Supóngase ahora que la regla de parada está determinada por un contraste unilateral del tipo $H_0 : \mu = \mu_0$ frente a $H_1 : \mu > \mu_0$. Si la hipótesis alternativa es manifiestamente cierta ($\mu > \mu_0$), cualquier test de la familia de Wang & Tsiatis (particularmente el de Pocock) parará en pocas etapas. Si por el contrario, la hipótesis nula se mantiene, previsiblemente el test alcance la inspección máxima. Esto implica una correlación negativa entre N y $S_N - N\mu$ lo cual se mostrará en un estudio de simulación posteriormente. Obsérvese que cuando la hipótesis nula falla, la distribución empírica tiene manifiestos sesgos tal como se ha mostrado en los epígrafes anteriores. En el caso que nos ocupa vimos que sobrestimaba la distribución original. Esto supone que si se usa el mismo criterio de parada, los datos se apartarán aún más de la hipótesis nula lo que sugiere que la distribución de probabilidad del tiempo de parada bootstrap está desplazada a la izquierda en relación con la del tiempo de parada original. Véase por tanto que mantener el mismo criterio de parada en el contexto bootstrap que en el original no conduce a que la distribución del tiempo de parada bootstrap imite a la del tiempo de parada real. Este hecho es no obstante compensado por una mayor distancia entre $S^*_{N^*}$ y $N^*\mu_0$ la cual aparece en la forma de los procedimientos de Wang & Tsiatis. En definitiva, aunque la distribución de probabilidad de N^* no imite a la de N, la

distribución de $\left(S^*_{N^*} - S_N\right)\left(N^* - E_*\left[N^*\right]\right)$ si imita a la de $\left(S_N - N\mu\right)\left(N - E[N]\right)$, lo cual es la **clave** de la validez de la aproximación bootstrap. Esto se pondrá de manifiesto con el estudio de simulación.

3.4.3. Estudio de simulación

Consideremos la distribución $N(\mu, \sigma = 15)$ y la hipótesis nula $H_0 : \mu = 85$. Para contrastar esta hipótesis hemos utilizado el procedimiento de Pocock con número máximo de looks K=8 y tamaño por look m=20. El valor crítico C=2.5 proporciona una significación global para el test de 0.05. El procedimiento se repitió 30.000 veces, tomando μ=85 (hipótesis nula cierta). Esto proporciona una aproximación a la distribución de probabilidad del pivotal T_N. La densidad se ajustó mediante estimación de núcleo. Para obtener su aproximación bootstrap se utilizó como regla de parada la misma que para los datos originales, pero muestreando de la distribución empírica. La figura 3.3.a muestra simultáneamente las funciones de densidad de probabilidad de ambos pivotales, las cuales difícilmente pueden ser distinguidas.

El mismo estudio se repitió tomando como valor real del parámetro $\mu = 90$ y manteniendo como hipótesis nula $H_0 : \mu = 85$. Se obtuvieron asimismo las estimaciones de los pivotales T_N y $T^*_{N^*}$ cuyas funciones de densidad se representan conjuntamente en la figura 3.3.b. Las figuras 3.3.c y 3.3.d muestran las densidades de estos pivotales para el caso $H_0 : \mu = 5$ estando los datos generados por distribuciones exponenciales de parámetro 5 y 7 respectivamente.

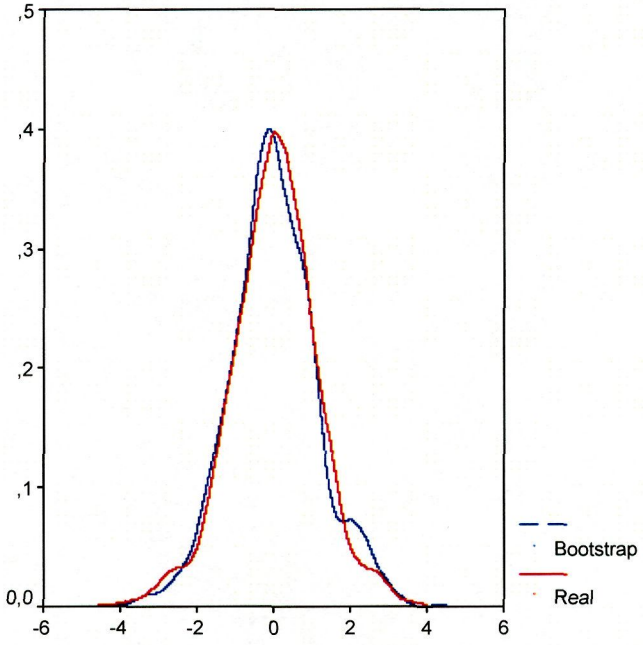


Figura 3.3.a. Distribuciones de probabilidad de T_N y $T_{N^*}^*$. Datos obtenidos secuencialmente de la distribución $N(85;15)$ siendo la regla de parada la determinada por el contraste de Pocock con $H_0 : \mu = 85$.

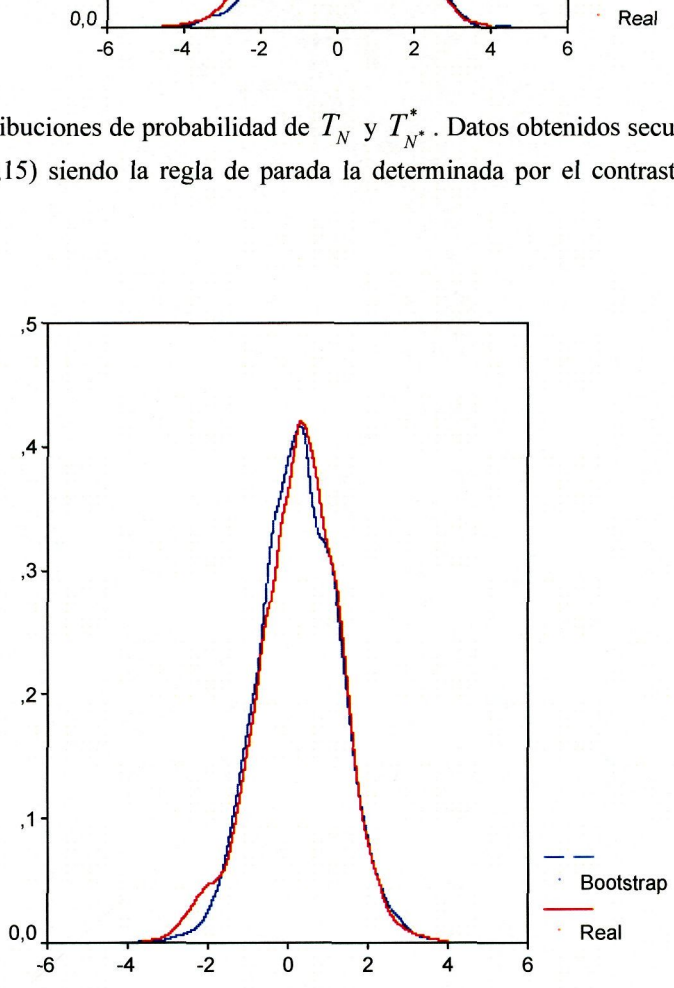


Figura 3.3.b. Distribuciones de probabilidad de T_N y $T_{N^*}^*$. Datos obtenidos secuencialmente de la distribución $N(90;15)$ siendo la regla de parada la determinada por el contraste de Pocock con $H_0 : \mu = 85$.

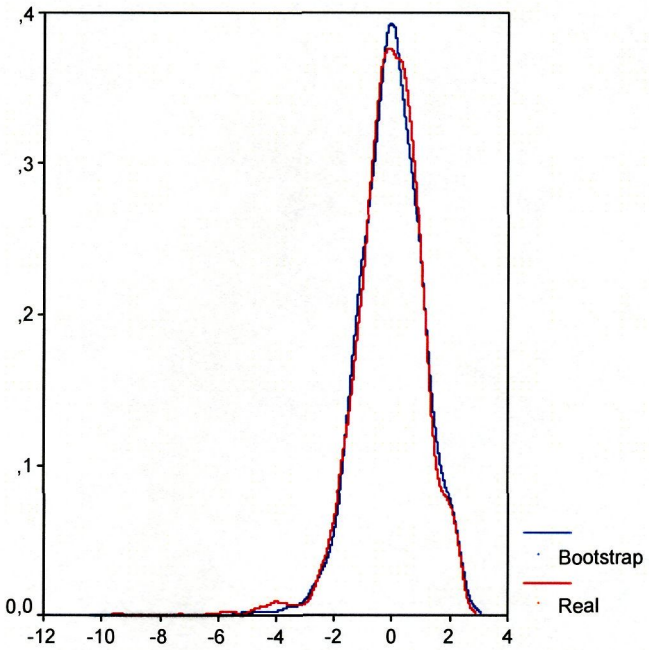


Figura 3.3.c. Distribuciones de probabilidad de T_N y $T_{N^*}^*$. Datos obtenidos secuencialmente de la distribución exp(5) siendo la regla de parada la determinada por el contraste de Pocock con $H_0 : \mu = 5$.

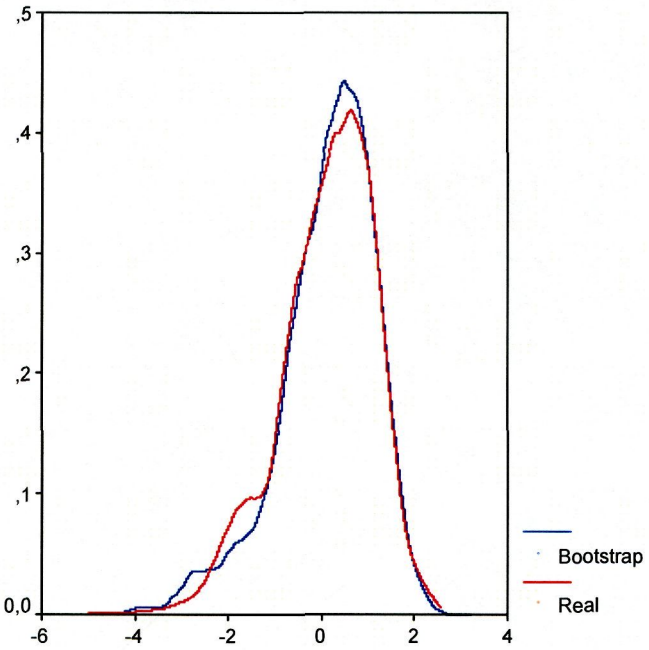


Figura 3.3.d. . Distribuciones de probabilidad de T_N y $T_{N^*}^*$. Datos obtenidos secuencialmente de la distribución exp(7) siendo la regla de parada la determinada por el contraste de Pocock con $H_0 : \mu = 5$.

De las distribuciones de probabilidad se obtuvieron los percentiles 2.5 y 97.5 con el objetivo de obtener intervalos de confianza al nivel nominal del 95%. Mediante 10.000 iteraciones de los datos se estimaron las probabilidades de cobertura para los intervalos obtenidos de T_N y $T_{N^*}^*$ respectivamente. Con objeto de comparar el procedimiento bootstrap con el de Emerson & Fleming dado en 3.2, se estimaron las correspondientes probabilidades de cobertura (mediante 10.000 iteraciones) sustituyendo la varianza de la distribución por la estimada. En la tabla 3.1. se muestran las probabilidades de cobertura estimada para los cuatro casos considerados.

Tabla 3.1.

Hipótesis nula	*Distribución*	*Método*	*% que subestima la media*	*% que contiene la media*	*% que sobreestima la media*
$H_0 : \mu = 85$	$N(85;15)$	Real	2.48%	94.55%	2.97%
		Bootstrap	2.98%	93.77%	3.25%
		E&F	3.44%	92.85%	3.71%
	$N(90;15)$	Real	2.71%	94.74%	2.55%
		Bootstrap	3.01%	94.78%	2.21%
		E&F	3.77%	93.60%	2.63%
$H_0 : \mu = 5$	exp(5)	Real	2.47%	95.28%	2.25%
		Bootstrap	1.31%	94.42%	4.27%
		E&F	1.43%	91.19%	7.38%
	exp(7)	Real	2.42%	94.69%	2.89%
		Bootstrap	2.78%	94.64%	2.58%
		E&F	1.02%	94.26%	4.72%

En relación con el análisis del pivotal dado en 3.4.2, la tabla 3.2. muestra la media, desviación estándar y percentiles notables de los pivotales $(S_N - N\mu)/(\sigma\sqrt{N})$ y el modificado $(S_N - N\mu)/(\sigma\sqrt{E[N]})$. Puede verse

claramente como el modificado está centrado en el origen. También se observa que ocurre lo mismo para la aproximación bootstrap y que los desplazamientos entre los pivotales reales son del mismo orden que entre los bootstrap. La notable coincidencia de los pivotales modificados real y bootstrap, los cuales están centrados en el origen, quedó probada en 3.4.2. Las desviaciones respecto al origen están asociadas con la distribución de probabilidad conjunta de N y $S_N - N\mu$. Los gráficos 3.4.a y 3.4.b muestran la distribución conjunta de N y $S_N - N\mu$ y de sus equivalentes bootstrap N^* y $S^*_{N^*} - S_N$. Puede observarse una coincidencia amplia entre ambas distribuciones.

Tabla 3.2.

Pivotal	*Media*	*Sd*	*P5*	*P25*	*P50*	*P75*	*P95*
$\frac{S_N - N\mu}{\sigma\sqrt{N}}$	0.2850	1.0236	-1.4178	-0.3761	0.3159	0.9866	1.8932
$\frac{S_N - N\mu}{\sigma\sqrt{E[N]}}$	0.0087	1.0298	-2.0719	-0.4245	0.2970	0.6837	1.1883
$\frac{S^*_{N^*} - S_N}{\hat{\sigma}_N\sqrt{N^*}}$	0.3026	1.0369	-1.3495	-0.3837	0.3192	0.9933	1.9490
$\frac{S^*_{N^*} - S_N}{\hat{\sigma}_N\sqrt{E[N^*]}}$	0.0237	1.0256	-1.9669	-0.4326	0.2903	0.6955	1.2359

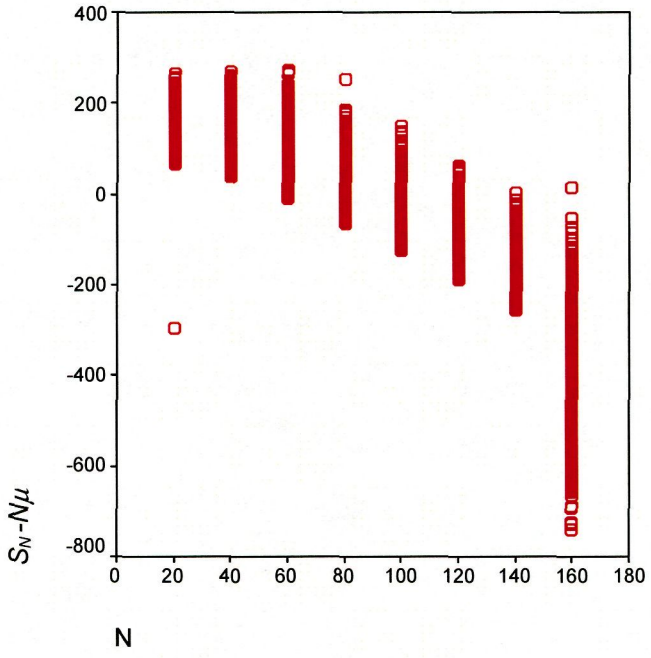


Figura 3.4.a. Distribución de probabilidad conjunta de N y $S_N - N\mu$. De acuerdo con la forma de la regla de parada, resulta obvio que cuando el contraste se detiene con pocos looks, las diferencias entre S_N y $N\mu$ son mayores.

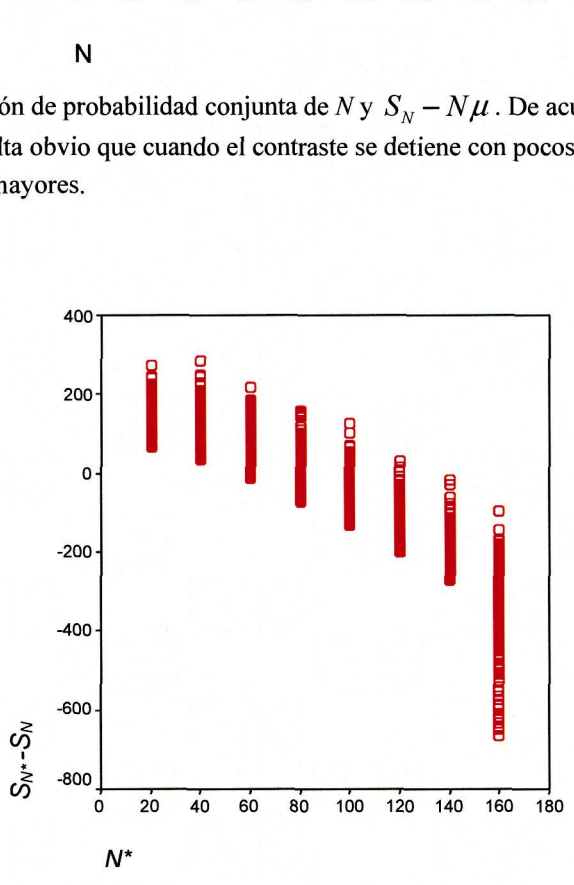


Figura 3.4.b. Distribución de probabilidad conjunta de N^* y $S^*_{N^*} - S_N$. Se produce el mismo efecto que con N y $S_N - N\mu$.

3.5. Aproximación bootstrap con datos secuenciales para proporciones

Consideremos dos proporciones π_1 y π_2 y el conjunto de datos $\{N, Y_{ij}\ ; i=1,2;\ j=1,\cdots,N\}$ obtenidos mediante un diseño secuencial con tiempo de parada N tales que $Y_{ij} \cong b(1,\pi_i)$. Tales datos pueden proceder de un contraste secuencial del tipo de Wang & Tsiatis. El objetivo ahora es obtener un intervalo de confianza para la diferencia de proporciones $\pi_1 - \pi_2$ y otro para el logaritmo del riesgo relativo $\log(\pi_1/\pi_2)$. Para la diferencia de proporciones consideramos el pivotal habitual:

$$T_N = \sqrt{N}\,\frac{\hat{\pi}_{1,N} - \hat{\pi}_{2,N} - (\pi_1 - \pi_2)}{\left\{\hat{\pi}_{1,N}(1-\hat{\pi}_{1,N}) + \hat{\pi}_{2,N}(1-\hat{\pi}_{2,N})\right\}^{1/2}} \tag{3.22}$$

siendo $\hat{\pi}_{i,N} = \sum_{j=1}^{N} Y_{i,j} \big/ N\ ; i=1,2$.

En orden a determinar un pivotal para $\log(\pi_1/\pi_2)$, sea $\rho = \pi_1/\pi_2$. Sea por tanto $\hat{\rho}_N = \hat{\pi}_{1,N}/\hat{\pi}_{2,N}$. De esta forma:

$$\log(\hat{\rho}_N/\rho) = \log\frac{\sum_{j=1}^{N} Y_{1,j}}{N\pi_1} - \log\frac{\sum_{j=1}^{N} Y_{2,j}}{N\pi_2} \tag{3.23}$$

Para un diseño de tamaño fijo (N no aleatorio), se tiene:

$$E\left[\sum_{j=1}^{N} Y_{i,j}\big/(N\pi_i)\right] = 1 \quad \text{y} \quad \operatorname{var}\left(\sum_{j=1}^{N} Y_{i,j}\big/(N\pi_i)\right) = \frac{1-\pi_i}{N\pi_i} \tag{3.24}$$

Dado que la distribución $\sum_{j=1}^{N} Y_{i,j} \big/ (N\pi_i)$ está concentrada alrededor de la unidad, podemos considerar la aproximación:

$$\log\left(\sum_{j=1}^{N} Y_{i,j} \big/ (N\pi_i)\right) \approx \sum_{j=1}^{N} Y_{i,j} \big/ (N\pi_i) - 1 \tag{3.25}$$

Por tanto, para N fijo, se tiene que:

$$\log(\hat{\rho}_N / \rho) \cong N\left(0; \sqrt{(1-\pi_1)/(N\pi_1) + (1-\pi_2)/(N\pi_2)}\right) \tag{3.26}$$

Este resultado sugiere que para el diseño secuencial consideremos el pivotal:

$$R_N = \frac{\log \hat{\rho}_N - \log \rho}{\sqrt{(1-\hat{\pi}_{1,N})/(N\hat{\pi}_{1,N}) + (1-\hat{\pi}_{2,N})/(N\hat{\pi}_{2,N})}} \tag{3.27}$$

La distribución empírica correspondiente a cada muestra observada $Y_{i,1}, \cdots, Y_{i,N}$, para $i = 1,2$ es $b(1, \hat{\pi}_{i,N})$, donde previsiblemente $\hat{\pi}_{i,N}$ es un estimador no centrado para π_i. A pesar de esto, proponemos un procedimiento bootstrap para aproximar la distribución probabilística de los pivotales T_N y R_N.

3.5.1. Aproximación bootstrap para el pivotal T_N.

Paso 1. Para cada $i = 1,2$, considerar como distribución de remuestreo la $b(1, \hat{\pi}_{i,N})$.

Paso 2. Utilizando la misma regla de parada que se utilizó con la distribución original, obtener un conjunto de datos bootstrap $\left\{N^*, Y_{i,j}^* ; i = 1,2; j = 1, \cdots, N\right\}$.

Paso 3. Del conjunto de datos bootstrap, obtener por el método de Monte Carlo la distribución de probabilidad del pivotal:

$$T_N^* = \sqrt{N^*}\,\frac{\hat{\pi}^*_{1,N^*} - \hat{\pi}^*_{2,N^*} - \left(\hat{\pi}_{1,N} - \hat{\pi}_{2,N}\right)}{\left\{\hat{\pi}^*_{1,N^*}\left(1-\hat{\pi}^*_{1,N^*}\right) + \hat{\pi}^*_{2,N^*}\left(1-\hat{\pi}^*_{2,N^*}\right)\right\}^{1/2}} \tag{3.28}$$

Un conjunto de B valores de T_N^* proporcionarán la aproximación buscada.

En orden a justificar la aproximación propuesta, expresaremos el pivotal T_N en la forma alternativa:

$$T_N = \frac{S_N - N\cdot\left(\pi_1 - \pi_2\right)}{\sqrt{N}\left\{\hat{\pi}_{1,N}\left(1-\hat{\pi}_{1,N}\right) + \hat{\pi}_{2,N}\left(1-\hat{\pi}_{2,N}\right)\right\}^{1/2}} \tag{3.29}$$

siendo $S_N = \sum_{j=1}^{N}\left(Y_{1,j} - Y_{2,j}\right)$. De la identidad de Wald se obtiene de forma inmediata:

$$E\left[S_N - N\left(\pi_1 - \pi_2\right)\right] = 0 \tag{3.30}$$

Análogamente, el pivotal bootstrap puede expresarse por:

$$T_N^* = \frac{S^*_{N^*} - N^*\left(\hat{\pi}_{1,N} - \hat{\pi}_{2,N}\right)}{\sqrt{N^*}\left\{\hat{\pi}^*_{1,N^*}\left(1-\hat{\pi}^*_{1,N^*}\right) + \hat{\pi}^*_{2,N^*}\left(1-\hat{\pi}^*_{2,N^*}\right)\right\}^{1/2}} \tag{3.31}$$

donde ahora $S^*_{N^*} = \sum_{j=1}^{N^*}\left(Y^*_{1,j} - Y^*_{2,j}\right)$. Aplicando, a continuación, la identidad de Wald para las distribuciones bootstrap, resulta:

$$E_*\left[S^*_{N^*} - N^*\left(\hat{\pi}_{1,N} - \hat{\pi}_{2,N}\right)\right] = 0$$

Podemos obtener también expresiones para la varianza de los numeradores, aplicando el teorema 3.5. Así pues,

$$\operatorname{var}\left(S_N - N(\pi_1 - \pi_2)\right) = \left\{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)\right\} E[N] \tag{3.32}$$

Obviamente, en el contexto bootstrap se tiene:

$$\operatorname{var}_*\left(S^*_{N^*} - N^*\left(\hat{\pi}_{1,N} - \hat{\pi}_{2,N}\right)\right) = \left\{\hat{\pi}_{1,N}\left(1-\hat{\pi}_{1,N}\right) + \hat{\pi}_{2,N}\left(1-\hat{\pi}_{2,N}\right)\right\} E[N] \tag{3.33}$$

De esta forma y para ambos pivotales, los denominadores estiman la desviación estándar de los numeradores.

Mediante un estudio de simulación, en el subepígrafe 3.5.4 hacemos una evaluación empírica de la aproximación bootstrap propuesta.

3.5.2. Aproximación bootstrap para el pivotal R_N

Paso 1. Para cada $i = 1,2$, considerar como distribución de remuestreo la $b\left(1, \hat{\pi}_{i,N}\right)$.

Paso 2. Utilizando la misma regla de parada que se utilizó con la distribución original, obtener un conjunto de datos bootstrap $\left\{N^*, Y^*_{i,j}\,;\, i = 1,2;\, j = 1,\cdots,N\right\}$.

Paso 3. Del conjunto de datos bootstrap, obtener por el método de Monte Carlo la distribución de probabilidad del pivotal:

$$R^*_{N^*} = \frac{\log \rho^*_{N^*} - \log \hat{\rho}_N}{\sqrt{\left(1-\hat{\pi}_{1,N^*}\right)/\left(N^*\hat{\pi}_{1,N^*}\right) + \left(1-\hat{\pi}_{2,N^*}\right)/\left(N\hat{\pi}_{2,N^*}\right)}} \tag{3.34}$$

Un conjunto de B valores de R_N^* proporcionarán la aproximación buscada.

3.5.3. Evaluación del procedimiento bootstrap para el riesgo relativo

Justificamos en este subepígrafe la aproximación bootstrap propuesta para el pivotal del riesgo relativo. Sea nuevamente:

$$\log(\hat{\rho}_N/\rho) = \log\frac{\sum_{j=1}^{N} Y_{1,j}}{N\pi_1} - \log\frac{\sum_{j=1}^{N} Y_{2,j}}{N\pi_2} \tag{3.35}$$

Según se vio anteriormente, dado que la distribución de $\sum_{j=1}^{N} Y_{i,j}/(N\pi_i)$ está concentrada alrededor de la unidad, consideramos la aproximación:

$$\log\left(\sum_{j=1}^{N} Y_{i,j}/(N\pi_i)\right) \approx \sum_{j=1}^{N} Y_{i,j}/(N\pi_i) - 1 \tag{3.36}$$

Ello significa que:

$$\log(\hat{\rho}_N/\rho) \approx \sum_{j=1}^{N} Y_{1,j}/(N\pi_1) - \sum_{j=1}^{N} Y_{2,j}/(N\pi_2) \tag{3.37}$$

Por tanto:

$$R_N \approx \frac{\log\hat{\rho}_N - \log\rho}{\sqrt{(1-\pi_1)/(N\pi_1)+(1-\pi_2)/(N\pi_2)}} = \frac{\sum_{j=1}^{N} Y_{1,j}/\pi_1 - \sum_{j=1}^{N} Y_{2,j}/\pi_2}{\sigma\sqrt{N}} \tag{3.38}$$

siendo $\sigma = \sqrt{(1-\pi_1)/\pi_1 + (1-\pi_2)/\pi_2}$. De esta forma;:

$$R_N \approx \frac{\sum_{j=1}^{N} Y_{1,j}/\pi_1 - \sum_{j=1}^{N} Y_{2,j}/\pi_2}{\sigma\sqrt{E[N]}\sqrt{N/E[N]}} = \frac{1}{\sqrt{N/E[N]}}\left\{\frac{\sum_{j=1}^{N} Y_{1,j}/\pi_1 - \sum_{j=1}^{N} Y_{2,j}/\pi_2}{\sigma\sqrt{E[N]}}\right\}$$

$$R_N \approx \frac{\sum_{j=1}^{N} Y_{1,j}/\pi_1 - \sum_{j=1}^{N} Y_{2,j}/\pi_2}{\sigma\sqrt{E[N]}} - \frac{1}{2\sigma\left(\sqrt{E[N]}\right)^3}\left(\frac{\sum_{j=1}^{N} Y_{1,j}}{\pi_1} - \frac{\sum_{j=1}^{N} Y_{2,j}}{\pi_2}\right)\left(N - E[N]\right)$$

Mediante la identidad de Wald, es inmediato probar que el primer término de la descomposición anterior tiene media cero. Asimismo, de acuerdo con el teorema 3.5, la varianza correspondiente es uno. El segundo término es de orden inferior en relación con $E[N]^{1/2}$. Para diseños de tamaño fijo el término correspondiente sería nulo. Con diseños secuenciales, este término contribuye a desviar el centro de gravedad de R_N. La versión bootstrap obviamente tiene la forma:

$$R_{N^*}^* \approx \frac{\sum_{j=1}^{N^*} Y_{1,j}^*/\hat{\pi}_{1,N} - \sum_{j=1}^{N^*} Y_{2,j}^*/\hat{\pi}_{2,N}}{\hat{\sigma}_N\sqrt{E_*[N^*]}} - \frac{1}{2\hat{\sigma}_N\left(\sqrt{E_*[N^*]}\right)^3}\left(\frac{\sum_{j=1}^{N^*} Y_{1,j}^*}{\hat{\pi}_{1,N}} - \frac{\sum_{j=1}^{N^*} Y_{2,j}^*}{\hat{\pi}_{2,N}}\right)\left(N^* - E_*[N^*]\right) \quad (3.39)$$

Un análisis comparativo de las descomposiciones obtenidas para R_N y $R_{N^*}^*$ justifica la validez de la aproximación en los mismos términos que el realizado en la sección 3.5.

3.5.4. Estudio de simulación

Para $\pi_1 = \pi_2$=0.7 se ha realizado un estudio de simulación tomando K=8 (número máximo de inspecciones) y m=20 (20 observaciones por grupo e inspección). La regla de parada es la proporcionada por el procedimiento de Pocock para contrastar $H_0 : \pi_1 = \pi_2$. El valor crítico C=2.45 proporciona una significación de 0.05. Para los datos obtenidos hemos determinado la distribución de los pivotales T_N y R_N y sus aproximaciones bootstrap $T_{N^*}^*$ y $R_{N^*}^*$, las cuales se muestran en la figura 3.5.a y 3.5.b. El mismo estudio de simulación se ha repetido

pero con $\pi_1 = 0.7$, $\pi_2 = 0.8$ y la hipótesis nula $H_0 : \pi_1 = \pi_2$. Los resultados para los pivotales T_N y $T_{N^*}^*$ se muestran en la figura 3.5.c, mientras que para los pivotales R_N y $R_{N^*}^*$ aparecen en 3.5.d.

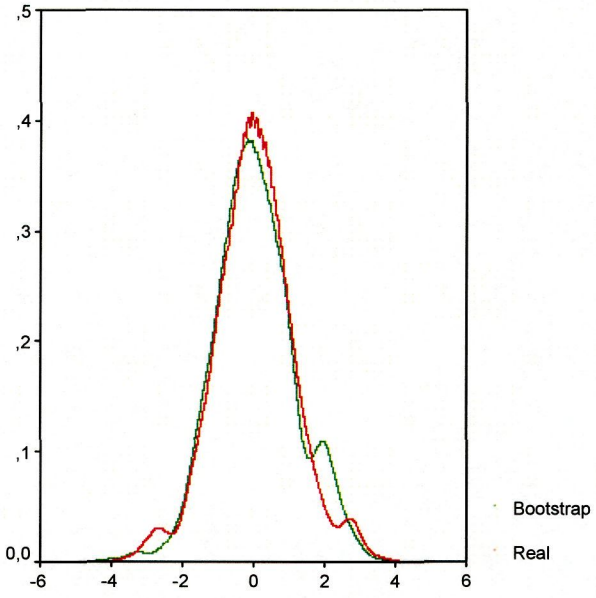


Figura 3.5.a. Densidades de probabilidad para T_N y $T_{N^*}^*$. Los datos están generados por el procedimiento de Pocock bajo H_0.

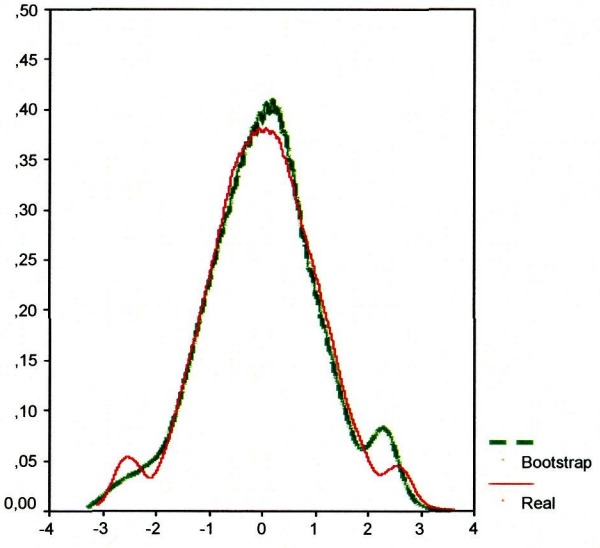


Figura 3.5.b. Densidades de probabilidad para R_N y $R_{N^*}^*$. Los datos están generados por el procedimiento de Pocock bajo H_0.

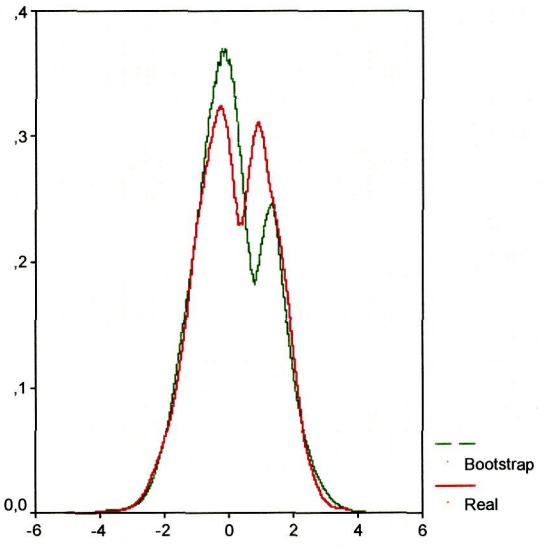


Figura 3.5.c. . Densidades de probabilidad para T_N y $T_{N^*}^*$. Los datos están generados por el procedimiento de Pocock para $\pi_1 = 0.7$ y $\pi_2 = 0.8$.

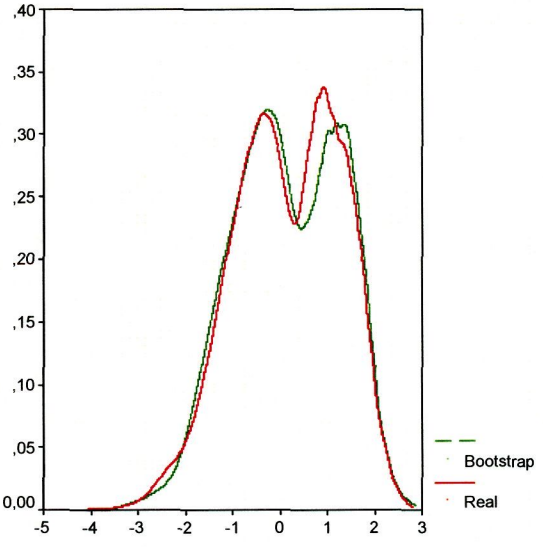


Figura 3.5.d. Densidades de probabilidad para R_N y $R_{N^*}^*$. Los datos están generados por el procedimiento de Pocock para $\pi_1 = 0.7$ y $\pi_2 = 0.8$.

4. Aplicaciones a tres estudios biomédicos

Presentamos ahora tres estudios biomédicos a los cuales aplicaremos los procedimientos secuenciales desarrollados a lo largo de esta memoria. El primero es un ensayo clínico cuyo objetivo fue evaluar la eficacia de un cierto catéter en la reducción de la tasa de colonización y bacteriemia en pacientes ingresados en unidades de cuidados intensivos. El segundo estudio tenía como objetivo establecer el valor pronóstico del grupo de genes HLA para la enfermedad diabética. En el último se analizan determinaciones basadas en ultrasonidos (QUS) como marcadores de la enfermedad osteoporótica. Todos los estudios se desarrollaron originalmente de acuerdo con diseños de tamaño fijo. De sus correspondientes bases de datos se realizaron *submuestreos* secuenciales utilizando los procedimientos de Wang & Tsiatis descritos en el primer capítulo. Considerados en su diseño original y para sus correspondientes variables principales de valoración, en todos se había obtenido significación estadística, por lo que cabía esperar que los tiempos de parada correspondientes a los submuestreos secuenciales se alcanzaran antes de agotar los datos disponibles. Esta conjetura se basó en el hecho de que para las diferencias observadas en los estudios originales, nivel de significación del 5% y potencia del 95%, los tamaños muestrales requeridos para los diseños de tamaño fijo eran menores a los disponibles en las bases de datos. De las consideraciones expuestas en el primer capítulo, los tamaños esperados para los diseños secuenciales serían a su vez inferiores a los requeridos para los diseños de tamaño fijo. En todos los tres estudios se alcanzó finalmente el tiempo de parada antes de agotar los datos disponibles. Para el submuestreo secuencial los datos se mantuvieron en el mismo orden que aparecían en las correspondientes bases de datos. Para los parámetros de interés, se obtuvieron los intervalos de confianza utilizando los procedimientos bootstrap propuestos en el tercer capítulo de esta memoria.

4.1. Estudio 1: Evaluación de la efectividad del recubrimiento de catéteres con antibióticos.

A los pacientes ingresados en las unidades de cuidados intensivos (UCI) se les inserta habitualmente un catéter a través del cual reciben los tratamientos prescritos. Estos catéteres son frecuentemente colonizados por agentes microbianos los cuales pueden pasar a la sangre produciendo una sepsis en el paciente. Con el fin de evitar esta colonización, se ha diseñado un catéter que permite su recubrimiento con antibióticos. Para evaluar su eficacia se han desarrollado diversos ensayos clínicos en los cuales se analiza fundamentalmente si se logra una disminución en las colonizaciones de los catéteres y de ahí, de bacteriemias en los pacientes. Sobre uno de estos ensayos hemos aplicado los procedimientos secuenciales desarrollados en esta memoria.

En el estudio considerado participaron nueve hospitales españoles y se diseñó originalmente como un ensayo clínico de tamaño fijo con dos grupos paralelos, donde un total de 284 pacientes ingresados en las Unidades de Cuidados Intensivos (UCI) fueron aleatorizados a recibir un catéter impregnado en antibiótico o un catéter control. Resultaron evaluables finalmente 139 pacientes en el grupo con catéter impregnado y 145 en el control. Como veremos posteriormente, la tasa de catéteres colonizados en el grupo impregnado se redujo aproximadamente a la mitad en relación a la del grupo control. Sin embargo, no hubo diferencias significativas entre las tasas de bacteriemias de ambos grupos. La siguiente tabla resume el perfil de los pacientes incluidos en el estudio por grupo experimental en el momento de la aleatorización.

Tabla 4.1. Características de los pacientes y catéteres

	Impregnado (*n*=139)	Control (*n*=145)	*p*-valor
Edad media	61 ± 16	59 ± 18	.4
Sexo, Hombre/Mujer	93 / 46	91 / 54	.5
Media APACHE II (admisión)	16 ± 7	15 ± 6	.5
Pacientes			.9
Médicos	87	89	
Quirúrgicos	35	37	
Traumatológicos	17	19	
Localización del catéter			.8
Vena subclavia	63 (45%)	64 (44%)	
Vena yugular interna	76 (55%)	81 (56%)	
Duración de la cateterización (media en días)	10 ± 5	11 ± 5	.5

De esta tabla se concluye que ambos grupos experimentales son perfectamente comparables. Esto permite atribuir las diferencias que surjan entre los parámetros de evaluación a los tratamientos recibidos y no a eventuales factores de confusión. La tabla 4.2 muestra las tasas de colonización de catéteres y las de bacteriemia en pacientes por grupo de tratamiento. Puede observarse una evidente reducción de la colonización en el grupo experimental frente al control. Sin embargo, la reducción en las tasas de bacteriemias no resultó ser significativa.

Tabla 4.2. Tasas de colonización y bacteriemia

	Impregnado (*n*=137)	Control (*n*=145)	*Riesgo Relativo* (IC 95%)
Catéteres colonizados	28 (20,4%)	60(41,4%)	0.49 (0.34 – 0.72)
Bacteriemia	5 (3,6%)	8 (5,5%)	0.66 (0.22 – 1.97)

Si partiendo de los datos de este estudio admitimos que la tasa de catéteres colonizados en el grupo impregnado es del 20% y en el control del 40%, con 130

pacientes por brazo experimental, el test de la ji-cuadrado a un nivel de significación del 5% proporciona una potencia del 95% para la diferencia observada. De acuerdo con las consideraciones expuestas en el primer capítulo de esta memoria, parece razonable pensar que bajo estos supuestos, cualquier test secuencial alcance con una alta probabilidad el tamaño de parada antes de agotar los datos de la base original.

Representamos por π_E la proporción de catéteres en los que se produce colonización en el grupo experimental (catéteres recubiertos con antibiótico) y π_C a la del grupo control. La regla de parada se basa en el procedimiento de Pocock descrito en el primer capítulo de esta memoria aplicado al contraste de hipótesis $H_0 : \pi_E = \pi_C$ frente a $H_1 : \pi_E \neq \pi_C$. El procedimiento secuencial de Pocock exige fijar a priori el número máximo de inspecciones *K*, el tamaño muestral por inspección *m* y el nivel de significación α para realizar el contraste. El valor de *m* depende de la potencia $1-\beta$ con la que se quiera detectar la diferencia mínima que sea considerada relevante. En el epígrafe 1.9 de esta memoria se explicó como podían obtenerse numéricamente estas cantidades. De la base de datos del estudio original se ha realizado el submuestreo secuencial fijando un número máximo de inspecciones *K*=7 con un tamaño de 20 casos por grupo en cada inspección, lo que daría lugar a un tamaño máximo de muestra de 140 casos por grupo, número que coincide aproximadamente con los datos disponibles. Este diseño permitiría detectar, con una potencia 0.9, una diferencia mínima entre las proporciones π_E y π_C de 0.22. Para detectar esta diferencia en un diseño de tamaño fijo serían precisos 112 sujetos en cada grupo.

La aplicación del método de Pocock a estos datos condujo al rechazo de la hipótesis nula en la cuarta inspección, con un tamaño de muestra final de *N*=80 pacientes por brazo de tratamiento. Utilizamos como parámetro principal de valoración el riesgo relativo definido como $\rho = \pi_E / \pi_C$. Las estimaciones de las proporciones de interés son: $\hat{\pi}_E = 0.2125$ y $\hat{\pi}_C = 0.3825$ lo que produce una estimación del riesgo relativo de $\hat{\rho}_N = 0.55$. El intervalo de confianza para el

riesgo relativo puede obtenerse a partir del pivotal correspondiente introducido en el capítulo 3, resultando [0.3344, 0.8616]. Obsérvese que, debido a que el tamaño muestral del proceso secuencial es inferior al diseño original de tamaño fijo, el intervalo de confianza obtenido a partir del procedimiento secuencial es ligeramente más largo.

4.2. Estudio 2: Evaluación de los genotipos HLA como predictores de la diabetes

La diabetes tipo 1 está asociada con los loci HLA, siendo los genes HLA-DRB1, DQB1 y DQA1 los principales implicados. Se ha demostrado que confieren susceptibilidad las moléculas codificadas por ciertos alelos DRB1, así como las combinaciones de alelos DQ que codifican una arginina en la posición 52 de la molécula DQα y cualquier aminoácido distinto del ácido aspártico en la posición 57 de la molécula DQβ (residuos críticos). Diversos estudios (Vicario JL *et al*, Serrano-Ríos M *et al* y Setién Baranda F *et al*) han documentado que los genes del grupo HLA están asociados con la diabetes tipo 1 en la población española. Por otra parte, se ha documentado también una mayor incidencia de diabetes tipo 1 en la isla de Gran Canaria en comparación con el resto de la población española, y una prevalencia mayor de hipertensión y nefropatía en esta población de pacientes.

En orden a establecer en Gran Canaria la asociación entre los genes HLA y la enfermedad diabética se diseñó un estudio de caso-control con diseño de tamaño fijo. El número de participantes fue de 274 personas, de las cuales 114 eran diabéticas y 160 controles. A partir de los genotipos descritos de la HLA, se consideraron dos grupos de riesgo de enfermedad diabética. La siguiente tabla muestra las características de los participantes en el estudio.

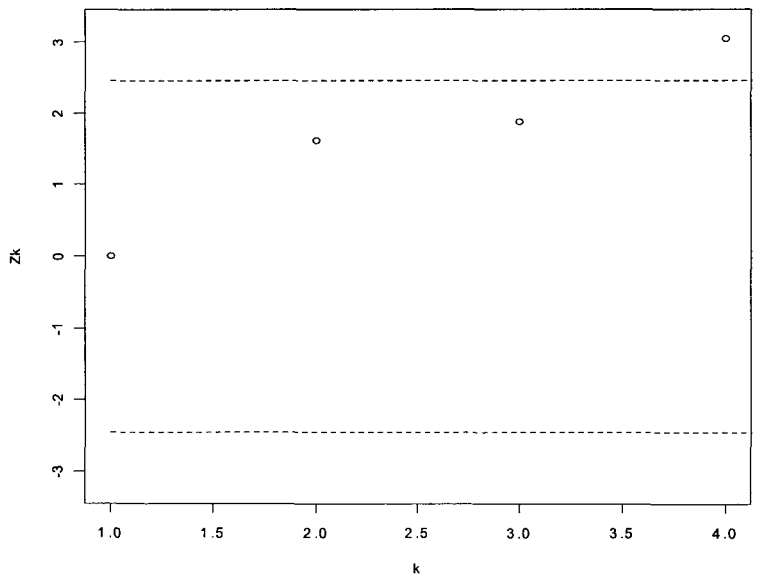
Tabla 4.3. Características de los participantes en el estudio HLA

	Diabéticos (n=114)	Controles (n=160)	p-valor
Edad media	26,1 ± 13,5	25,9 ± 19,8	.9
Sexo, Hombre/Mujer	50 / 64	71 / 89	.9
Grupo riesgo HLA	62 (54,4%)	17 (10,6%)	<.01

Puede observarse que en el grupo de diabéticos, el 54,4% pertenecen al grupo HLA que hemos definido de riesgo, mientras que en el de controles sólo un 10,6% pertenecen a este grupo. Para esta diferencia observada, el test ji-cuadrado con un nivel de significación del 5% alcanza una potencia del 99% seleccionado 33 pacientes en cada uno de los grupos determinados por la presencia o no de enfermedad diabética. Es obvio que hay un número más que suficiente de casos en la base de datos para realizar a partir de esta un muestreo secuencial.

Representamos por π_D y π_C a la proporción de individuos de riesgo HLA en los grupos diabético y control respectivamente. El procedimiento secuencial se basará en el contraste de hipótesis $H_0 : \pi_D = \pi_C$ frente a $H_1 : \pi_D \neq \pi_C$. Para simular la realización del contraste de Pocock en este caso, hemos considerado la realización de 6 inspecciones con un tamaño de 19 casos por grupo en cada inspección, lo que daría lugar a un tamaño máximo de muestra de 114 casos por grupo, número que coincide con el total de diabéticos disponibles en la muestra. Este diseño permitiría detectar, con una potencia 0.9, una diferencia mínima entre las proporciones π_D y π_C de 0.32. Para detectar esta diferencia en un diseño de tamaño fijo serían necesarios 92 sujetos en cada grupo.

La aplicación del método de Pocock a estos datos condujo al rechazo de la hipótesis nula en la cuarta inspección. En el gráfico siguiente podemos observar la evolución del estadístico Z_k en las sucesivas inspecciones:



La diferencia de proporciones finalmente observada fue 0.316. El intervalo de confianza al 95% obtenido por el método de Emerson & Fleming resultó ser [0.02, 0.377], y el intervalo bootstrap [0.158, 0.455].

4.3. Estudio 3. Evaluación de un marcador basado en ultrasonidos como predictor de la osteoporosis

La osteoporosis es una patología consistente en una baja resistencia ósea, cuya principal complicación clínica son las fracturas. Se ha estimado que aproximadamente una tercera parte de mujeres de raza blanca con edad superior a los 50 años sufrirán una fractura lumbar, de cadera o muñeca. La DXA (Dual x-ray absorptiometry) está considerada como el método de referencia para la determinación de la densidad mineral ósea de forma precisa y reproductible, aunque expone a la persona a radiaciones con rayos X. Recientemente se ha propuesto como medida alternativa los ultrasonidos cuantitativos (QUS) (Glüer C,

and F.t.I.Q.U.C. Group). Estas mediciones, además de evitar la exposición a la radiación X, pueden realizarse a través de un aparato de fácil manejo siendo además muy importante la reducción de sus costos. El aparato de medida proporciona dos determinaciones de base llamadas BUA y SOS, a partir de las cuales se define el marcador Qui-Stifness (QUI) como:

$$QU\,I = 0.41*(BUA+ SOS)-571$$

En orden a evaluar el valor discriminante de los QUS para la osteoporosis se diseñó un estudio de caso control en el que participaron 340 mujeres post-menopausicas, de las cuales, 149 habían sido diagnosticadas con osteoporosis de acuerdo con el criterio basado en DXA y 191 eran controles. La siguiente tabla resume algunas características de las mujeres seleccionadas para el estudio.

Tabla 4.4. Características de los participantes en el estudio de osteoporosis

	Osteoporóticas (*n*=149)	Controles (*n*=191)	*p*-valor
Edad (media±sd)	65,5 ± 8,6	60,2 ± 7,9	<.01
Fumadoras	13 (8,7%)	26 (13,6%)	0.16
IMC (media±sd)	26,4 ± 4,0	29,1 ± 4,6	<.01
QUI (media±sd)	64,9 ± 13,3	78,5 ± 15,7	<.01

Denominando μ_O a la media del QUI en la población con osteoporosis (casos), y μ_C a la media de esta variable en la población sana (controles), nos interesa decidir entre las hipótesis $H_0 : \mu_O = \mu_C$ frente a $H_0 : \mu_O \neq \mu_C$. Como hemos hecho en los casos anteriores, resolvemos este contraste aplicando el método de Pocock En este caso el estudio de simulación se ha realizado para considerando K=4 y K=6 (número máximo de inspecciones), en orden a evaluar como dicho número podría haber afectado a los resultados. Asimismo, hemos elegido como nivel de significación el valor habitual de $\alpha = 0.05$. Para fijar el

valor de m hemos considerado la detección de una diferencia mínima relevante $\delta = \mu_C - \mu_O = 7$ ó $\delta = 10$ con una potencia $1-\beta = 0.9$. De este modo en el caso $\delta = 7$, $K = 7$ se podría llegar a utilizar casi toda la muestra disponible (140 casos en cada grupo). Tras cada simulación hemos calculado intervalos de confianza al 95% utilizando tanto por el método de Emerson & Fleming, como el método Bootstrap para diseños secuenciales introducido en el capítulo 3.

El siguiente cuadro resume las características de las cuatro simulaciones que se han realizado. Hemos añadido también en cada caso el tamaño de muestra necesario para realizar un contraste equivalente utilizando un diseño de tamaño fijo:

Tabla 4.5

δ	K	m	*Tamaño muestral máximo (diseño secuencial)*	*Tamaño muestral (diseño fijo)*
7	4	34	136	113
	7	20	140	
10	4	17	68	56
	7	10	70	

Para el procedimiento de Pocock, el valor c_k es constante en k, y depende sólo de K y α, $c_k = C(K, \alpha)$. En nuestras simulaciones resultaron $C(7, 0.05) = 2.361$ y $C(4, 0.05) = 2.485$.

En todos los casos, la hipótesis nula resultó rechazada, detectándose una diferencia significativa en el valor medio de *qui-stifness* entre el grupo de mujeres con osteoporosis y el grupo control. Los resultados obtenidos se muestran en la siguiente tabla:

Tabla 4.6

δ	K	*Etapa de parada*	*Tamaño muestral final alcanzado*	*Intervalo de confianza (Emerson y Fleming)*	*Intervalo de confianza (Bootstrap)*
7	4	2	68	[-16.51, -2.05]	[-17.33, -7.87]
	7	2	40	[-18.18, -2.18]	[-19.94, -6.13]
10	4	3	51	[-17.30, -1.04]	[-20.29, -9.05]
	7	4	40	[-16.41, -0.93]	[-18.88, -5.02]

Como puede observarse, si para este estudio se hubiese elegido un diseño secuencial, en todos los casos el tamaño muestral alcanzado cuando el procedimiento se detiene es inferior al tamaño muestral requerido por el diseño de tamaño fijo. La ventaja relativa es mayor para aquellos diseños que requerían a priori un tamaño de muestra mayor (i.e. en el caso en que se pretende detectar una diferencia media más pequeña). Ello se debe a que se ha rechazado la hipótesis nula ya que, obviamente, en caso de haberse aceptado, el tamaño final alcanzado en el contraste secuencial hubiese superado al necesario en un diseño de tamaño fijo. Asimismo observamos que a menor número de inspecciones, mayores son los tamaños finales de muestra requeridos, tal como se había señalado en el primer capítulo.

Bibliografía

1. Anderson, T.W. (1959). A modification of the sequential probability ratio test to reduce the sample size. *Ann. Math. Statist.*, **31**, 165-197.
2. Armitage, P. (1954). Sequential tests in prophylactic and therapeutic trials. *Quart. J. Med.*, **23**, 255-274.
3. Armitage, P. (1958). Sequential methods in clinical trials. *Am. J. Pub. Health*, **48**, 1395-1402.
4. Armitage, P. (1975). *Sequential Medical Trials.* (2ª Edición). Blackwell Scientific Publications, Oxford.
5. Armitage, P., McPherson, C.K. and Rowe, B.C. (1969). Repeated significance tests on accumulating data. *J.R. Statist. Soc.*, A, **132**, 235-244.
6. Anscombe, F.J. (1952). Large sample theory of sequential estimation. *Proc. Cambrige Philos. Soc.*, **48**, 600-607.
7. Barnard, G. A. (1946). Sequential tests in industrial statistics. *J. Roy. Statist. Soc. Suppl.*, **8**, 1-26.
8. Bickel, P. and Freedman, D. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.*, **9**, 1196-1217.
9. Bowman, A., Hall, P. and Prvan, T. (1998). Banthwith selection for the smoothing of distribution functions, *Biometrika*, **85**, 799-808.
10. Bross, I. (1952). Sequential medical plans. *Biometrics*, **8**, 188-205.
11. Bross, I. (1958). Sequential clinical trials. *J. Chronic Deseases*, **8**, 349-365.
12. DeMets, D.L. and Lan, K.K.G. (1984). An overview of sequential methods and their application in clinicals trials. *Commun. Statist.: Theor. Meth.*, **13**, 2315-2338.
13. DeMets, D.L. and Ware, J.H. (1980). Group sequential methods for clinical trials with one-side hypothesis. *Biometrika*, **67**, 651-660.
14. DeMets, D.L. and Ware, J.H. (1982). Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika*, **69**, 661-663.
15. Doeblin, W. (1938). Sur deux problémes de M. Kolmogoroff concernat les chaines dénombrables, *Bull. Soc. Math. France*, **66**, 218-220

16. Efron, B. (1979). Boostrap methods: another look at the jackknife. *The Annals of Statistics*, **7**, 1-26.
17. Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall. New York.
18. Emerson, S.S. and Fleming, T.R. (1989). Symmetric group sequential designs. *Biometrics*, **45**, 905-923.
19. Emerson, S.S. and Fleming, T.R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika*, **77**, 875-892.
20. Franke, J. and Härdle, W. (1992). On bootstrapping kernel spectral estimates. *The Annals of Statistics*, **20**, 121-145.
21. Freedman, D.A. (1981). Bootstrapping regression models. *The Annals of Statistics*, **9**, 1218-1228.
22. Glüer C, F.t.I.Q.U.C. Group. (1996). Quantitative ultrasound technique for the assessment of osteoporosis: expert agreement on current status. *J Bone Miner Res* ,**11**, 707-730.
23. Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer-Verlag.
24. Jennison, C. and Turnbull, B.W. (1983). Confidence intervals for binomial parameter following a multistage test with application to MIL-STD 105D and medical trials. *Technometrics,* **25**, 49-58.
25. Jennison, C and Turnbull, B.W. (2000). *Group Sequential Methods with Applications to Clinical Trials.* Chapman & Hall/Crc. London.
26. Kim, K. and DeMets, D.L. (1987). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika*, **74**,149-154.
27. Lorden, G. (1976). 2-SPRT's and the modified Keifer-Weiss problem of minimizing an expected sample size. *Ann. Statist.*, **4**, 281-291.
28. McPherson, C.K. and Armitage, P. (1971). Repeated significance tests on accumulating data when the null hypothesis is not true. *J.R. Statist. Soc.*, A, **134**, 15-25.
29. O'Brien, P.C. and Fleming, T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, **35**, 549-556.

30. Pampallona, S. and Tsiatis, A.A. (1994). Group sequential designs for one-sided and two–sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *J. Statist. Planning and Inference*, **42**, 19-35.
31. Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, **64, 2**, 191-199.
32. Rohatgi, V.K. (1976). *An Introduction to Probability Theory and Mathematical Statistics*. Wiley.
33. Serrano-Ríos M, Gutiérrez-López MD and Pérez-Bravo F. (1996). HLA-DR, DQ and anti-GAD antibodies in first degree relatives of type I diabetes mellitus. *Diabetes Res Clin Pract*, **34,** Suppl:S,133-139.
34. Setién Baranda F, Coto E, Menéndez Díaz J, Martínez-Naves E, Álvarez Martínez V and López-Larrea, (1994). C. HLA class II and susceptibility and resistance to insulin-dependent diabetes mellitus in a population from the northwest of Spain. *Eur J Immunogenet*, **21**, 219-229.
35. Siegmund, D. (1979). Corrected diffusion approximations in certain random walk problems. *Advances in Apllied Probability,* **11**, 701-719.
36. Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals.* Springer-Verlag, New York.
37. Sobel, M. and Wald, A. (1949). A sequential decision procedure for choosing one of three hypothesis concerning the unknown mean of a normal distribution. *Ann. Math. Statist.*, **20**, 502-522.
38. Slud, E.V. and Wei, L.J. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *J. Amer. Statist. Assoc.*, **77**, 862-868.
39. Vicario JL, Martínez-Laso J and Corell A. (1992). Comparison between HLA-DRB and DQ DNA sequences and classic serological markers as type I (insulin-dependent) diabetes mellitus predictive risk markers in the Spanish population. *Diabetologia*, **35**, 475-481.
40. Wald, A. (1947). *Sequential Analysis*, New York: Wiley.
41. Whitehead, J. (1997). *The design and Analysis of Sequential Clinical Trials.* (Revised 2nd Edition). John Wiley & Sons Ltd. Chichester.
42. Whitehead, J. and Stratton, I. (1983). Group sequential clinical trials with triangular continuation regions. *Biometrics*, **39**, 227-236.

43. Woodroofe, M. (1992). Estimation after sequential testing: a simple approach for a truncated sequential probability ratio test. *Biometrika*, **79**, **2**, 347-353.