QÜESTIIÓ, vol. 21, 1 i 2, p. 293-316, 1997

EL PROBLEMA DE BEHRENS-FISHER EN LA INVESTIGACIÓN BIOMÉDICA. ANÁLISIS CRÍTICO DE UN ESTUDIO CLÍNICO MEDIANTE SIMULACIÓN

ESTEBAN VEGAS LOZANO*

Universitat de Barcelona

En el artículo se hace una revisión del problema de Behrens-Fisher, discutiendo los fundamentos inferenciales asociados a la dificultad de su resolución y exponiendo las soluciones prácticas más comunes, juntamente con una nueva solución basada en conceptos de geometría diferencial. A continuación, se realiza un estudio crítico de una investigación biomédica en donde las verdaderas probabilidades de error son distintas de las supuestas debido a que se ignoran probables diferencias entre las varianzas. En dicha investigación se rechazó la hipótesis nula de igualdad de medias ($p < 0.01$), si bien, la verdadera probabilidad de error de tipo I para valores próximos a los muestrales parece ser otra. Con este fin, se realiza un estudio de Monte Carlo para obtener estas estimaciones según se utilice el test t de Student de comparación de medias o diferentes soluciones más apropiadas para el problema de Behrens-Fisher. En este estudio de simulación se usa una técnica de reducción de la varianza específica para variables de respuesta dicotómica tales como contrastes de hipótesis (aceptar, no aceptar la hipótesis nula). Se presenta brevemente esta técnica y se ilustra como se diseña la simulación de acuerdo con la misma.

The Behrens-Fisher problem in biomedical research. Critical analysis of a clinical study by simulation

Keywords: Simulación de Monte Carlo, Reducción de la varianza, Curvatura estadística, Distancia de Rao, Bootstrap.

Clasificación AMS: 65C05

*Esteban Vegas Lozano. Departament d'Estadística. Facultat de Biologia. Universitat de Barcelona. Diagonal 645, 08028 Barcelona.

– Article rebut el gener de 1997.

– Acceptat el maig de 1997.

1. INTRODUCCIÓN

En la investigación experimental, en particular en la biomédica, es muy común que aparezca el problema de Behrens-Fisher: contrastar la igualdad de medias de dos poblaciones normales, sin hacer ninguna suposición acerca de las varianzas. De una manera más formal, el problema se puede exponer como sigue:

> Sean X_1, X_2 dos variables aleatorias normales independientes con sus respectivas medias $\mu_i, i = 1,2$, y con varianzas desconocidas y arbitrarias $\sigma_i^2, i = 1,2$. Bajo estas premisas se considera el contraste de hipótesis:
>
> $$\begin{array}{lll} H_0: & \mu_1 = \mu_2 = \mu & \text{arbitrarios} \\ H_1: & \mu_1 \neq \mu_2 & \sigma_1 > 0 \ \ \sigma_2 > 0 \end{array}$$
>
> Se supone que el contraste está basado en una muestra de $n_1 + n_2$ valores
>
> $$\mathbf{x} = (x_{11}, \ldots, x_{1n_1}; x_{21}, \ldots, x_{2n_2}),$$
>
> donde cada submuestra viene de n_i realizaciones independientes e idénticamente distribuidas (*iid*) de X_i, para $i = 1,2$.

La diferencia con respecto al conocido test t de Student para la comparación de medias de dos poblaciones normales con varianzas desconocidas pero iguales es no hacer ninguna suposición respecto de las varianzas. Así pues, la diferencia fundamental está en que para aplicar el test t de Student se necesita *homocedasticidad* mientras que en el problema de Behrens-Fisher se incluye, además, el caso de *heterocedasticidad*. Como es bien conocido por los estadísticos, la aparición de heterocedasticidad suele complicar la inferencia estadística (algunos modelos típicos de inferencia clásica no se pueden aplicar como por ejemplo, el modelo lineal normal) y se entra en un campo donde existen numerosos problemas abiertos para los cuales, en el mejor de los casos, se han encontrado algunas soluciones aproximadas pero donde ninguna de ellas es óptima.

Este detalle produce un cambio cualitativo muy importante, que se refleja a un nivel más abstracto en la geometría de ciertas familias paramétricas de densidad. Así, se tiene que tanto el modelo paramétrico correspondiente al caso de varianzas iguales como el del caso de varianzas desiguales se pueden expresar como una familia exponencial de 3 y 4 parámetros, respectivamente. Pero es en el submodelo paramétrico asociado a la hipótesis nula donde se produce la diferencia a remarcar. El espacio de distribuciones de probabilidad asociado a la hipótesis nula en el caso de varianzas iguales es un «subespacio plano» del espacio de la familia exponencial de dimensión

3, mientras que en el caso de varianzas desiguales es un «subespacio curvado» de la familia exponencial de dimensión 4. Como consecuencia, se produce una ruptura de las buenas propiedades para los problemas de inferencia que tiene la familia exponencial, que se acrecienta cuanto mayor es la curvatura. Así, por ejemplo, la varianza del estimador máximo verosímil excede del valor de la cota de Cramer-Rao en proporción aproximada al valor de su curvatura al cuadrado. Tampoco existe un estadístico suficiente con la misma dimensión que la familia exponencial curvada (Efron, 1975).

Las familias exponenciales curvadas suelen aparecer bastante a menudo en la práctica y tiene gran importancia en la discusión de varios conceptos y métodos claves de inferencia estadística (ver Efron, 1975, 1978 y Efron and Hinkley, 1978).

2. SOLUCIONES PRÁCTICAS AL PROBLEMA DE BEHRENS-FISHER

El problema de Behrens-Fisher es una cuestión controvertida que no tiene solución óptima conocida y para el que las diferentes escuelas de pensamiento estadístico dan diferentes soluciones. Como breve resumen, se muestran algunas de las soluciones más habituales en la práctica. Para una revisión más exhaustiva véase Scheffé (1970) y Lee and Gurland (1975).

Para tamaños muestrales mayores que 10, las diferencias entre las diversas soluciones propuestas son generalmente mucho menores que sus diferencias con el test t de Student para la igualdad de medias. Por tanto, el uso de cualquiera de ellos es mejor que el uso del test t de Student, a menos que se garantice la validez de la suposición de igualdad de varianzas.

De todas estas soluciones, las más comúnmente utilizadas son:

- El test de Cochran y Cox.
- El test de Welch o el test de Welch-Aspin.

La primera se puede considerar, con el test de McCullough-Barnerjee, como aproximación a la solución de Behrens-Fisher, mientras que la mayoría de otros test discutidos por Lee and Gurland (1975) se pueden considerar como aproximaciones al test de Welch-Aspin.

En todos los casos, el estadístico de contraste es:

$$T' = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(\hat{s}_1^2/n_1) + (\hat{s}_2^2/n_2)}} \tag{1}$$

donde, para cada submuestra $(x_{i1},\ldots,x_{in_i})$ $i=1$ o 2, $\bar{x_1},\bar{x_2}$ representan las medias muestrales; s_1^2, s_2^2 corresponden a las varianzas muestrales con denominador n_i y $\hat{s}_1^2$, $\hat{s}_2^2$ designa a las varianzas muestrales con denominador n_i-1.

El problema es que el estadístico T' no se distribuye, necesariamente, según el modelo de probabilidad t de Student con n_1+n_2-2 grados de libertad. Por tanto, todo el esfuerzo se centra en conocer de forma aproximada la distribución muestral de T'.

Cochran y Cox (1950) diseñaron un método de aproximación a los puntos críticos de la distribución muestral T'. El método consiste en obtener estos puntos mediante:

$$t'_p = \frac{t(p,n_1-1)(\hat{s}_1^2/n_1)+t(p,n_2-1)(\hat{s}_2^2/n_2)}{(\hat{s}_1^2/n_1)+(\hat{s}_2^2/n_2)} \tag{2}$$

donde $t(p,\nu)$ es el p-percentil superior de una distribución t de Student con ν grados de libertad. El proceso de decisión es *rechazar* H_0 *si* $|T'| \geq t'_{\alpha/2}$ siendo α el nivel de significación nominal del test.

La principal ventaja de esta solución es su simplicidad. Tal vez es por esta razón que se ha convertido en el test más extensamente utilizado en la práctica, incluso sabiendo que la extensión del test puede diferir substancialmente del nivel de significación nominal, como fue mostrado por Cochran (1964). Afortunadamente, en la mayoría de ocasiones, la extensión del test está por debajo del nivel de significación nominal, es decir, se trata de un test conservador.

Welch (1938) propuso una aproximación alternativa. En esta aproximación T' se concibe como una variable aleatoria distribuida según la t de Student, pero con un número desconocido de grados de libertad. La solución pasa por determinar los grados de libertad (gl') que corresponden a la distribución de T' mediante la expresión:

$$gl' = \frac{\left(\frac{\hat{s}_1^2}{n_1}+\frac{\hat{s}_2^2}{n_2}\right)^2}{\frac{(\hat{s}_1^2/n_1)^2}{n_1-1}+\frac{(\hat{s}_2^2/n_2)^2}{n_2-1}} \tag{3}$$

El resultado obtenido para gl' se redondea al entero más próximo[1]. Se obtienen así unos grados de libertad comprendidos entre un mínimo y un máximo conocidos: el mínimo es el valor más pequeño de n_1-1 y n_2-1; el máximo es n_1+n_2-2. El criterio de decisión es *rechazar* H_0 *si* $|T'| \geq t(\alpha/2, gl')$ donde $t(\alpha/2, gl')$ es el $(\alpha/2)$-percentil superior de una distribución t de Student con gl' grados de libertad.

El test de Welch-Aspin se basa en cálculos asintóticos sobre T'. Welch (1947) obtiene los puntos porcentuales de T' como una serie de potencias en $1/f_i = 1/(n_i - 1)$ para $i = 1,2$ (véase el desarrollo de la fracción P en el apéndice A). Al año siguiente, Aspin (1948) extendió estos resultados hasta el orden 4. Por tanto, para poder utilizar este test se necesita tabular los puntos críticos. Se pueden encontrar en la tabla 11 de *Biometrika Tables* (1976) los valores tabulados para cuatro niveles de probabilidad ($\alpha/2 = 0.05, 0.025, 0.01$ y 0.005) para la cola superior. El criterio de decisión es *rechazar* H_0 *si* $|T'| \geq V(c; f_1, f_2, \alpha/2)$ donde $c = \dfrac{\hat{s}_1^2/n_1}{\hat{s}_1^2/n_1 + \hat{s}_2^2/n_2}$ y el valor $V(c; f_1, f_2, \alpha/2)$ se obtiene en las tablas anteriores.

Las diferencias entre los dos últimos test son mínimas siendo más común utilizar el test de Welch ya que no requiere de unas tablas específicas y en sólo un caso hace falta determinar los grados de libertad gl'. Dado que los valores de la distribución t de Student van disminuyendo a medida que van aumentando los grados de libertad, antes de calcular gl' se puede evaluar T' utilizando el gl' mínimo (es decir, el menor de n_1-1 y n_2-1); si se rechaza $H_0 : \mu_1 = \mu_2$, también se rechazará con el valor proporcionado por (3) para gl'; si no se rechaza H_0, podemos evaluar T' con el gl' máximo (n_1+n_2-2); si se sigue sin rechazar H_0, tampoco se rechazará calculando el valor exacto de gl'. De modo que el único caso en el que se necesita hacer uso de (3) para calcular el valor exacto de gl' será aquel en el que se mantenga H_0 con el gl' mínimo y se rechace con el gl' máximo.

A diferencia del test de Cochran y Cox, tanto el test de Welch como el test de Wech-Aspin, la probabilidad de error es muy cercana al valor nominal en todo el espacio paramétrico, aunque Fisher (1956) los critica por mostrar en un subconjunto relevante la existencia de un sesgo negativo en el sentido de Buehler (1959) (véase apéndice B).

[1] El propio Welch (1947) sugirió posteriormente que hacer:

$$gl' = \left[\frac{\left(\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}\right)^2}{\frac{(\hat{s}_1^2/n_1)^2}{n_1-1} + \frac{(\hat{s}_2^2/n_2)^2}{n_2-1}} \right] - 2$$

puede ofrecer una solución más exacta para gl'. No obstante, la diferencia entre ambas soluciones es, en la mayor parte de los casos, insignificante.

Como una última alternativa, se cita el test denominado geodésico[2] descrito en Vegas y Ocaña (1996) y Vegas (1996) (véase un breve resumen en el apéndice C) que está basado en un enfoque totalmente distinto, concretamente en criterios de geometría diferencial ya que el estadístico propuesto, D^2, expresa el nivel de disconformidad, en términos de distancia geodésica, entre los valores obtenidos en las estimaciones puntuales de la media y la desviación típica de la muestra $(\bar{x_1}, \bar{x_2}, s_1, s_2)$ y el conjunto de todos aquellos puntos paramétricos que cumplen la hipótesis nula de igualdad de medias con desviaciones típicas arbitrarias $(\mu, \mu, \sigma_1, \sigma_2)$.

3. ANÁLISIS CRÍTICO DE UNA INVESTIGACIÓN BIOMÉDICA

En esta sección se realiza un estudio crítico de las conclusiones de una investigación experimental en el entorno de las ciencias médicas y biológicas, en que las verdaderas probabilidades de error son distintas de las supuestas, precisamente a causa de ignorar probables diferencias de las varianzas. Es decir, es un ejemplo donde surge el problema de Behrens-Fisher en el que se obviaron las probables diferencias entre las varianzas.

3.1. Un estudio sobre la trombosis

Los datos que se presentan son una parte de un estudio más amplio que fue expuesto en un artículo de Oost *et al.* (1983). La presente revisión se centra en las medidas obtenidas sobre la excreción de β-tromboglubina urinaria en 12 pacientes normales y 12 pacientes diabéticos.

Normal	Diabético	Normal	Diabético
4.1	11.5	11.5	33.9
6.3	12.1	12.0	40.7
7.8	16.1	13.8	51.3
8.5	17.8	17.6	56.2
8.9	24.0	24.3	61.7
10.4	28.8	37.2	69.2

En el artículo mencionado, se obtuvo como resultado que existía diferencia significativa ($p < 0.01$) en la excreción media de β-tromboglubina urinaria entre los dos tipos de pacientes. A pesar de que no se indica claramente en el artículo, todo parece

[2]Para realizar el test geodésico se implementó un programa en Fortran v.3.2 que da como resultado la aceptación o rechazo de la igualdad de medias poblacionales. El programa esta a disposición de cualquiera que lo desee. Para solicitar una copia dirigirse al autor.

indicar que se realizó un test t de Student de comparación de medias para llegar a tal conclusión.

3.2. Comparación de medias entre los dos tipos de pacientes

El primer paso del estudio es comprobar que la comparación de medias entre los dos tipos de pacientes es, razonablemente, un caso del problema de Behrens-Fisher. Se verifica la normalidad utilizando la prueba de Lilliefors. Las tablas utilizadas no son las que obtuvo H. W. Lilliefors (1967) sino, las presentadas en A.L. Mason and C.B. Bell (1986) realizadas con un tamaño muestral de simulación mayor ($n = 20000$). Para ambas muestras, se acepta la hipótesis nula de normalidad para un nivel de significación de 0.05. Por otra parte, en el test F de comparación de varianzas se obtiene que existen diferencias significativas entre ellas ($p < 0.01$).

A continuación, se calcula algunas de las soluciones prácticas habituales para el problema de Behrens-Fisher: test de Cochran y Cox, Welch y Welch-Aspin. Todas ellas se basan en el estadístico T' (1) que da como resultado un valor de 3.3838 que, juntamente con el test geodésico (apéndice C), conducen a rechazar la igualdad de medias.

Incluso si no se tiene en cuenta la existencia de diferencias significativas entre las varianzas y se realiza el test t de Student de comparación de medias también se rechaza la hipótesis nula de igualdad de medias. Es decir, sea cual sea el test utilizado se rechaza la hipótesis nula, así pues, el estudio más interesante consiste en estimar cual es la verdadera probabilidad de error de tipo I.

3.3. Estimación de la verdadera probabilidad de error de tipo I por simulación

Como el resultado del contraste de medias ha sido rechazar la hipótesis nula el único error que podemos cometer es el de tipo I, es decir, rechazar la hipótesis nula cuando realmente la media poblacional de la primera población es igual que la segunda. En teoría esta probabilidad debería ser el nivel de significación nominal del test, que en este caso será de 0.05. Por tanto, lo que se desea verificar es si la extensión del test coincide con el nivel de significación nominal de 0.05.

3.3.1. Procedimiento

Se plantea un estudio de Monte Carlo que analiza el efecto de diferentes cocientes de varianzas (heterocedasticidad) alrededor de los valores muestrales en la estimación de la verdadera probabilidad de error de tipo I. Además, se incorpora una técnica de reducción de la varianza específica para variables de respuesta dicotómica (véase apéndice D para un breve resumen; más información en Ocaña y Vegas, 1995; Vegas 1996) para lograr unas estimaciones más precisas. En cada réplica de la simulación

Monte Carlo se obtiene, además del valor de la variable dicotómica de interés o respuesta Y (en este caso $Y = 1$ si se rechaza la hipótesis nula en el test estudiado, y $Y = 0$ en caso contrario), otra variable dicotómica de control C, correlacionada con Y, cuya esperanza, $E(C)$, es conocida. Esta variable C se basa en el resultado del test t de Student de comparación de medias ($C = 1$ si se rechaza la hipótesis nula y $C = 0$ en caso contrario) ya que este test está correlacionado con el test estudiado, para cualquiera de las soluciones habituales al problema de Behrens-Fisher, y además, se conoce la esperanza de C, es decir, su potencia. Es importante resaltar que interesa que la correlación entre la variable Y y la variable C sea elevada para que la precisión del estimador de la probabilidad de error de tipo I obtenido por esta técnica sea alta.

El algoritmo de simulación se divide en tres módulos:

1. Entrada de parámetros.
2. Obtención de los dos vectores de resultados $(Y_k, C_k)\ \ k = 1, \ldots, n$ donde Y corresponde al resultado del test bajo estudio y C al resultado del test de control (test t de Student).
3. Estimación de la probabilidad de error de tipo I aplicando la técnica de reducción de la varianza indicada anteriormente.

En el apartado de la entrada de parámetros los valores que permanecen fijos se refieren a los tamaños muestrales (n_1, n_2) que son iguales a 12, al nivel de significación del test de respuesta y del test de control que valen 0.05, al número de réplicas de simulación ($n = 5000$) y réplicas del boostrap paramétrico ($B = 1000$) necesario para estimar el punto crítico del test geodésico (véase apéndice C). Por otra parte, se utilizan tres pares de desviaciones típicas

$$(\sqrt{118.28}, \sqrt{84.54}),\ (\sqrt{1427.25}, \sqrt{84.54})\ \text{y}\ (\sqrt{410.87}, \sqrt{84.54})$$

que corresponden a posibles valores de varianzas tales que el cociente de cada pareja son los extremos (1.40, 16.88) y el punto medio (4.86) del intervalo de confianza del 95% para el cociente de varianzas poblacionales, respectivamente. Se escogieron aquellos valores de las varianzas en que al menos una de ellas era igual que la obtenida como varianza muestral a partir de los datos reales obtenidos por Oost *et al.*(1983). Al escoger estos tres pares de desviaciones típicas extremas se intenta observar cómo va variando la estimación de la probabilidad de error de tipo I con un número reducido de simulaciones.

El segundo apartado del algoritmo se divide en (véase fig. 1):

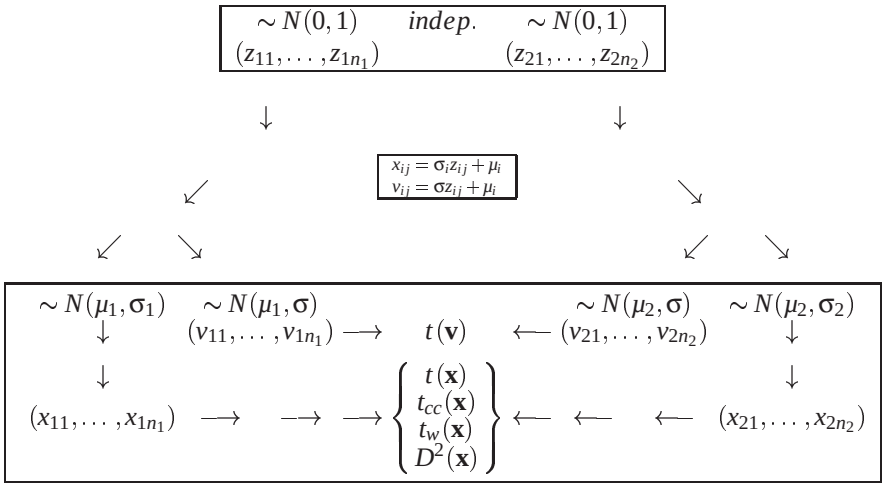
- Generación de las muestras.
- Calcular el test de estudio y el test t de Student de control ($t(\mathbf{v})$).
- Obtención de Y y C.

Como se desea estimar la verdadera probabilidad de error de tipo I para los siguientes tests:

- test t de Student de comparación de medias aunque las muestras provengan de normales con diferentes varianzas ($t(\mathbf{x})$).
- test de Cochran y Cox ($t_{cc}(\mathbf{x})$).
- test de Welch ($t_w(\mathbf{x})$).
- test geodésico ($D^2(\mathbf{x})$).

es necesario realizar tantos tipos de simulaciones como diferentes tipos de tests. Así, se substituye la frase «Calcular el test de estudio», que aparece en el parrafo anterior, por cada uno de ellos.

Generación de las muestras



Parejas diferentes:

$$\begin{array}{cccc} t(\mathbf{v})\to C & t(\mathbf{v})\to C & t(\mathbf{v})\to C & t(\mathbf{v})\to C \\ t(\mathbf{x})\to Y^1 & t_{cc}(\mathbf{x})\to Y^2 & t_w(\mathbf{x})\to Y^3 & D^2(\mathbf{x})\to Y^4 \end{array}$$

Figura 1. Obtención de los dos vectores de resultados $(Y_k, C_k)\ \ k=1,\ldots,n$ en la estimación de la verdadera probabilidad de error de tipo I para varios tests.

Para inducir la correlación requerida entre el resultado de cada uno de los anteriores tests ($t(\mathbf{x})$, $t_{cc}(\mathbf{x})$, $t_w(\mathbf{x})$ y $D^2(\mathbf{x})$) con el test t de Student de control ($t(\mathbf{v})$) se utiliza variables aleatorias comunes para evaluar cada uno de ellos.

A partir de n_1 valores normales estándar *iid*, $z_1 = (z_{11}, \ldots, z_{1n_1})$, independientes de otros n_2 valores normales estándar *iid*, $z_2 = (z_{21}, \ldots, z_{2n_2})$, se obtiene, gracias a la transformación $x_{ij} = \sigma_i z_{ij} + \mu_i$, $i = 1,2$ y $j = 1, \ldots, n_i$, los $n_1 + n_2$ valores $x = (x_{11}, \ldots, x_{1n_1}; x_{21}, \ldots, x_{2n_2})$ para calcular los cuatro tests de estudio en diferentes configuraciones de medias y varianzas, incluyendo el test t de Student evaluado bajo condiciones de desigualdad de varianzas. Y aplicando de nuevo la misma transformación, pero siendo la varianza utilizada $\sigma^2 = \dfrac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2}$, un valor común e intermedio entre σ_1^2 y σ_2^2, se obtiene $n_1 + n_2$ valores $\mathbf{v} = (v_{11}, \ldots, v_{1n_1}; v_{21}, \ldots, v_{2n_2})$ que sirven para calcular el test t de Student ($t(\mathbf{v})$) utilizado como variable de control. De esta manera, los $n_1 + n_2$ valores de $\mathbf{v}$ cumplen las condiciones de aplicabilidad del test t de Student ($t(\mathbf{v})$): provenir de dos poblaciones normales con varianza común, y por tanto, se puede utilizar como variable de control al ser conocida su potencia.

Para cada tipo de test tendremos el conjunto de vectores de resultados (Y_k, C_k) $k = 1, \ldots, n$ del cual se estima la probabilidad de error de tipo I aplicando la técnica de reducción de la varianza específica para variables dicotómicas.

3.3.2. Resultados

Los resultados presentados son los obtenidos a partir de $n = 5000$ réplicas de simulación y, en el caso del test geodésico además, se han hecho $B = 1000$ remuestras en cada réplica de simulación para estimar el valor crítico c_α (9, véase apéndice C). Se resumen en la tabla 1 donde se muestran, para cada uno de los test estudiados, las estimaciones de la probabilidad de error de tipo I en los tres niveles de heterocedastidad analizados:

$\tilde{p}_{1.}$**:** Representa la estimación puntual de la verdadera probabilidad de error de tipo I utilizando como estimador (11, véase apéndice D).

Tabla 1

Nivel de significación estimado en el **test geodésico, test de Welch, test de Cochran y Cox y test t de Student**, *respectivamente, en relación al estudio sobre la trombosis con* $n_1 = n_2 = 12$ *bajo distintas combinaciones de parámetros con valores próximos a los muestrales correspondientes a la hipótesis nula.*

	Resultado del test							
	Geodésico		Welch		Cochran y Cox		t de Student	
σ_1^2/σ_2^2	$\tilde{p}_{1.}$	$\pm 1.959964\hat{\sigma}_{GS}$	$\tilde{p}_{1.}$	$\pm 1.959964\hat{\sigma}_{GS}$	$\tilde{p}_{1.}$	$\pm 1.959964\hat{\sigma}_{GS}$	$\tilde{p}_{1.}$	$\pm 1.959964\hat{\sigma}_{GS}$
1.40	.0494	.002568	.0496	.002311	.0405†	.002539	.0504	.002191
4.86	.0535	.004796	.0535	.004775	.0506	.004488	.0507	.004838
16.88	.0546	.005631	.0530	.005540	.0466	.005408	.0620†	.005906

† = significativamente diferente del nivel de significación nominal.

$\pm 1.959964\hat{\sigma}_{GS}$**:** Margen de error asociado a un nivel de confianza del 95% para la verdadera probabilidad de error de tipo I, $p_{1.}$, utilizando como estimador de la varianza de $\tilde{p}_{1.}$ el estadístico $\hat{\sigma}^2_{GS}$ discutido en el apéndice D.

3.3.3. Conclusiones del estudio

Como se puede observar, tanto en las tablas como en el gráfico que se presenta posteriormente, sólo el test geodésico y el test de Welch incluyen el nivel de significación nominal de 0.05 en los respectivos intervalos de confianza aproximados del 95% para el nivel de significación a partir de una buena estimación del error estándar ($\hat{\sigma}_{GS}$, (15)) del estimador ($\tilde{p}_{1.}$) para cualquiera de los tres cocientes de varianzas estudiados.

Figura 2. Comparación de las estimaciones de las probabilidades de los errores de tipo I para los cuatro tipos de test bajo distintas combinaciones de parámetros correspondientes a la hipótesis nula.

El test de Cochran y Cox es un test conservador. Su verdadera extensión tiende a ser menor que el nivel de significación nominal (Lee and Gurland, 1975) y esta tendencia es más fuerte en el caso de $n_1\sigma_1^2 = n_2\sigma_2^2$. Esta afirmación se refleja al estar excluido en el intervalo de confianza aproximado del 95% para el nivel de significación el valor de 0.05 cuando el cociente de varianzas poblacionales es de 1.40. Para valores superiores, sí se incluye el valor de 0.05 de nivel de significación nominal. Gráficamente se observa (fig. 2) como en el primer caso la línea horizontal

de valor 0.05 no corta el segmento que representa el intervalo de confianza aproximado del 95% para el nivel de significación construido a partir de $\tilde{p}_{1.}$ correspondiente al cociente de varianzas poblacional igual a 1.40 para el test de Cochran y Cox mientras que para los valores de 4.86 y 16.88 sí que se corta.

Para el test t de Student ocurre al revés, como era de esperar. Cuando el cociente de varianzas es de 1.40 e incluso de 4.86 el intervalo de confianza aproximado del 95% para el nivel de significación obtenido a partir de $\tilde{p}_{1.}$ incluye el valor de 0.05 de nivel de significación nominal. Sin embargo, cuando es de 16.88 la estimación de la probabilidad de error de tipo I es superior a 0.05.

De todo lo expuesto se puede concluir que tanto el test geodésico como el test de Welch mantienen el nivel de significación nominal de 0.05 para los diferentes cocientes de varianzas y que, además, no existen excesivas diferencias entre las diferentes estimaciones obtenidas entre ellos. En cambio, el test de Cochran y Cox falla con varianzas semejantes y el test t de Student va peor cuanto mayor es el valor del cociente de varianzas poblacionales.

4. CONCLUSIONES

Es bastante frecuente en trabajos experimentales, y en particular médico-biológicos, suponer homocedasticidad entre las poblaciones cuando en realidad un test de comparación de varianzas indicaría todo lo contrario. Así, en algunas ocasiones, cuando surge el problema de Behrens-Fisher se resuelve de manera inadecuada a partir del test t de Student de comparación de medias.

Esto implica un error, que repercute en que los p-valores no son los que corresponden, se rechaza la hipótesis nula de igualdad de medias falsamente más veces, incrementándose esta tendencia cuanto mayor sea el valor del cociente de varianzas (véase la fig. 2). Esto implica que el test t de Student será claramente no recomendable bajo condiciones de extrema diferencia de varianzas o cuando los valores del estadístico de test se sitúen en la frontera entre las dos posibles decisiones (aceptar o rechazar la hipótesis nula).

Desde otro punto de vista, si se comparan los diferentes tests según el grado de heterocedasticidad y utilizando los valores de los parámetros del estudio médico de Oost *et al.*(1983) como marco de referencia, se puede llegar a las siguientes conclusiones:

- Si el cociente de varianzas es elevado (16.88) el test que cometería más error, es decir, aquel que la verdadera probabilidad de error de tipo I se aleja más del

nivel de significación nominal de 0.05 es el test t de Student, ya que el intervalo de confianza del 95% siempre es superior a 0.05, entre 0.056 y 0.068. Por tanto, se rechaza la hipótesis nula cuando realmente es cierta un porcentaje mayor que el 5%. Si se observa los valores estimados en el test de geodésico (0.0546) y el test de Welch (0.0530) tiene cierta tendencia a ser superior a 0.05 aunque sus respectivos intervalos de confianza incluyan el valor de 0.05. Igual que sucede con el test de Cochran y Cox, aunque este caso la estimación puntual, 0.0466, es menor que 0.05.

- En el caso de un cociente de varianzas menor (4.86), todos los intervalos de confianza del 95% de la verdadera probabilidad de error de tipo I en los diferentes tests incluyen el nivel de significación nominal de 0.05. La diferencia entre los test aparece en sus estimaciones puntuales. El test t de Student (0.0507) y el test Cochran y Cox (0.0506) son los que tienen un valor más próximo a 0.05, mientras que, el test geodésico y el test de Welch con un valor de 0.0535 tienen una cierta predisposición a tener valores superiores a 0.05.

- Por último, para el cociente de varianzas (1.40) cercano a la homocedasticidad. El test que se aleja más del nivel de significación nominal es el de Cochran y Cox, el intervalo de confianza no incluye el valor de 0.05 y su estimación puntual vale 0.0405, es decir, que se rechaza la hipótesis de igualdad de medias falsamente un número menor de veces que lo que se esperaría (un 5%). En los otros tests los intervalos de confianza incluyen el nivel significación nominal. Las estimaciones del test geodésico (0.0494) esta más alejada del teórico 0.05 que el test de Welch (0.0496) y el test t de Student (0.0504).

En general, se debe recomendar utilizar tanto el test geodésico (véase apéndice C) como el test de Welch ya que son igualmente adecuados para cualquier valor del cociente de varianzas. El test de Cochran y Cox es muy conservador cuando el cociente de varianzas es próximo a uno, y el peor de todos es el test t de Student. En cualquier caso, las conclusiones finales del artículo que ha servido de base a la presente discusión (no así los «p-valores») son «correctas», en tanto que coinciden con las que se obtienen mediante una técnica estadística más apropiada.

En otros estudios paralelos (Vegas, 1996) en los cuales se hacía perturbar la normalidad no se llegaba a tener los p-valores tan alejados. Por tanto, parece indicar que la heterocedasticidad afecta más que la perturbación de la normalidad en la variación de los p-valores.

Por otro lado, la técnica de reducción de la varianza empleada para el estudio de la probabilidad de error de tipo I mediante simulación de Monte Carlo puede ser exportada hacia otros estudios similares con una eficacia adecuada y un mínimo esfuerzo en la complejidad del diseño de la simulación.

APÉNDICES

A. Desarrollo de la fración P

La fracción P $(0 < P < 1)$ de T' se desarrolla explícitamente hasta orden 2 como

$$\alpha\left[1+\frac{(1+\alpha)^2}{4}\frac{\sum_{i=1}^{2}(\hat{s}_i^4/n_i^2 f_i)}{(\sum_{i=1}^{2}\hat{s}_i^2/n_i)^2}+\frac{(3+5\alpha^2+\alpha^4)}{3}\frac{\sum_{i=1}^{2}(\hat{s}_i^6/n_i^3 f_i^2)}{(\sum_{i=1}^{2}\hat{s}_i^2/n_i)^3}\right.$$

(4)

$$\left.-\frac{(15+32\alpha^2+9\alpha^4)}{32}\frac{\sum_{i=1}^{2}(\hat{s}_i^4/n_i^2 f_i)^2}{(\sum_{i=1}^{2}\hat{s}_i^2/n_i)^4}\right],$$

donde $\alpha = \Phi^{-1}(P)$ y Φ es la función de distribución de la normal estándar.

B. Existencia de sesgo negativo en el sentido de Buehler

Este hecho se muestra en que independientemente de $\hat{s}_1/\hat{s}_2$, $U = (n_1-1)(\hat{s}_1^2/\sigma_1^2) + (n_2-1)(\hat{s}_2^2/\sigma_2^2)$ sigue una distribución ji-cuadrado con n_1+n_2-2 grados de libertad y $(\bar{x}_1-\mu_1)+(\bar{x}_2-\mu_2)$ está distribuido normalmente con media cero y varianza $\sigma_1^2/n_1+\sigma_2^2/n_2$ siendo ambas variables aleatorias estocásticamente independientes. Por consiguiente,

$$T'\sqrt{\frac{\hat{s}_1^2/n_1+\hat{s}_2^2/n_2}{\sigma_1^2/n_1+\sigma_2^2/n_2}\frac{n_1+n_2-2}{U}} \tag{5}$$

tiene una distribución t de Student con n_1+n_2-2 grados de libertad dados $\hat{s}_1/\hat{s}_2, \sigma_1^2, \sigma_2^2$. Cuando $\hat{s}_1/\hat{s}_2 = 1, n_1 = n_2$ y $\sigma_1^2/\sigma_2^2 = w$, la expresión (5) se simplifica a

$$T'\sqrt{\frac{4}{2+1/w+w}}.$$

Por otra parte, puesto que $1/w+w \geq 2$, se tiene

$$\sqrt{\frac{4}{2+1/w+w}} \leq 1$$

y por tanto

$$Pr(|T'| > a|\hat{s}_1/\hat{s}_2 = 1) \geq Pr(|t_{(n_1+n_2-2)}| > a)$$

para todo $a > 0$, donde $t_{(n_1+n_2-2)}$ indica la distribución t de Student con n_1+n_2-2 grados de libertad. En particular, si $n_1 = n_2 = 7$ y $a = 1.74$ (de la tabla 11 de *Biometrika Tables*),

$$Pr(|T'| > 1.74|\hat{s}_1/\hat{s}_2 = 1) \geq Pr(|t_{12}| > 1.74) > 0.1.$$

Entonces el conjunto con $\hat{s}_1/\hat{s}_2 = 1$ es un subconjunto relevante donde el intervalo de confianza basado en el test de Welch-Aspin cubre el verdadero valor de $\mu_1 - \mu_2 = 0$ menos a menudo que el nivel de confianza sugerido.

C. Test geodésico

Se denomina así porque el proceso de decisión del nuevo test se basa en la distancia geodésica[3] para resolver el problema de Behrens-Fisher.

Usando como base la distancia de Rao entre dos distribuciones normales univariantes (Atkinson y Mitchell, 1981; Burbea y Rao, 1982), una medida natural de discrepancia entre una muestra compuesta de dos submuestras *independientes*, caracterizadas por los puntos $(\bar{x}_1, s_1)$ y $(\bar{x}_2, s_2)$, y un punto paramétrico compuesto por (μ_1, σ_1) y (μ_2, σ_2) es la *distancia de Rao al cuadrado ponderada* por los tamaños muestrales n_1 y n_2 (es inmediato dada la expresión de d^2 para distribuciones normales independientes, véase Amari *et al.*(1987) página 237)

$$d^2[(\bar{x}_1, \bar{x}_2, s_1, s_2), (\mu_1, \mu_2, \sigma_1, \sigma_2)] = 2\sum_{i=1}^{2} n_i \left\{ \log \frac{1+\Delta_i}{1-\Delta_i} \right\}^2 \tag{6}$$

donde

$$\Delta_i = \left\{ \frac{(\bar{x}_i - \mu_i)^2 + 2(s_i - \sigma_i)^2}{(\bar{x}_i - \mu_i)^2 + 2(s_i + \sigma_i)^2} \right\}^{1/2}. \tag{7}$$

Si se define:

$$D^2 = min_{\mu,\sigma_1,\sigma_2} \left\{ d^2[(\bar{x}_1, \bar{x}_2, s_1, s_2), (\mu, \mu, \sigma_1, \sigma_2)] \right\}, \tag{8}$$

intuitivamente, (8) representa el grado de discrepancia, en términos de distancia geodésica, entre los datos muestrales (resumidos por $\bar{x}_1, \bar{x}_2, s_1, s_2$) y el conjunto de todos los puntos paramétricos $(\mu, \mu, \sigma_1, \sigma_2)$ que satisfacen la hipótesis nula de igualdad de medias. Valores grandes de D^2 proporcionan evidencias contra esta hipótesis. Por

[3]Comúnmente llamada distancia de Rao en honor a su introductor en el ámbito estadístico (Rao, 1945)

tanto, el siguiente procedimiento de rechazo: «*rechazar* H_0 *si* $D^2 \geq c_\alpha$», donde c_α debe satisfacer

$$P\{D^2 \geq c_\alpha | H_0\} = \alpha, \tag{9}$$

producirá un test «geodésico» (en el sentido de Burbea y Oller, (Burbea y Oller, 1989)) con un nivel de significación α para la hipótesis nula de igualdad de medias.

El test estadístico D^2 (para más información véase Vegas y Ocaña, 1996; Vegas 1996) no tiene una forma cerrada, quedando en función de un sistema de ecuaciones

$$\begin{cases} \sigma_1^2 = & s_1^2 + \dfrac{(\bar{x_1} - \mu)^2}{2} \\ \sigma_2^2 = & s_2^2 + \dfrac{(\bar{x_2} - \mu)^2}{2} \\ \dfrac{\sigma_1}{\sigma_2} = & \dfrac{n_1}{n_2} \dfrac{cosh^{-1}(\sigma_1/s_1)}{cosh^{-1}(\sigma_2/s_2)} \end{cases} \tag{10}$$

que debe ser resuelto, numéricamente, para obtener los valores $\hat{\mu}, \hat{\sigma}_1$ y $\hat{\sigma}_2$ que minimizan d^2 en (8), requerido para calcular $D^2 = d^2[(\bar{x_1}, \bar{x_2}, s_1, s_2), (\hat{\mu}, \hat{\mu}, \hat{\sigma}_1, \hat{\sigma}_2)]$. Esto hace que sea muy difícil de resolver el problema distribucional de calcular c_α. Por consiguiente, se opta por aproximar su valor mediante bootstrap paramétrico. Es decir, se genera un número $B(= 1000)$ de remuestras del mismo tamaño, $n_1 + n_2$, del vector original de datos $x = (x_{11}, \ldots, x_{1n_1}; x_{21}, \ldots, x_{2n_2})$. Cada una de las remuestras $\mathbf{x}^*$ se produce por la generación de n_1 valores independientes e idénticamente distribuidos de una distribución normal de media

$$\bar{x} = \frac{n_1 \bar{x_1} + n_2 \bar{x_2}}{n_1 + n_2}$$

y desviación estándar s_1, junto con n_2 valores de una distribución normal con la misma media común, $\bar{x}$, y desviación estándar s_2:

$$\mathbf{x}^* = (x_{11}^*, \ldots, x_{1n_1}^*; x_{21}^*, \ldots, x_{2n_2}^*).$$

Para cada remuestra $\mathbf{x}^*$, el valor correspondiente del test estadístico, D_*^2, se calcula de la misma manera que D^2 y la probabilidad $Prob\left\{D^2 \geq D^2(\mathbf{x}) | H_0\right\}$ se estima por medio de:

$$P^* = \frac{\#[D_*^2 \geq D^2(\mathbf{x})] + 1}{B + 1}.$$

Para un nivel de significación fijo α la hipótesis nula se rechaza ($Y = 1$) si $P^* < \alpha$, y se acepta ($Y = 0$) en caso contrario.

La estimación P^* de $Prob\{D^2 \geq D^2(\mathbf{x}) \mid H_0\}$, en lugar de la frecuencia relativa $(\#[D_*^2 \geq D^2(\mathbf{x})]/B)$, garantiza que el nivel de significación bajo la distribución bootstrap sea realmente α en el sentido de que $Prob\{P^* \leq \alpha | N(\bar{x}, s_1, s_2)\} \leq \alpha$ (Dwass, 1957). No obstante este hecho no garantiza que el verdadero nivel de significación $Prob\{P^* \leq \alpha | N(\mu, \sigma_1, \sigma_2)\}$ sea realmente α.

D. Reducción de la varianza para variables de respuesta dicotómica en simulación

Supongamos que en cada réplica de una simulación de Monte Carlo se obtiene un valor de una variable de respuesta Y con distribución de Bernoulli. Por ejemplo, Y puede ser el resultado final de un test que acepta o rechaza una hipótesis, o un intervalo de confianza el cual incluye o no el verdadero valor del parámetro. Normalmente se realizará la simulación para estimar la esperanza $p_{1.} = E(Y)$, por ejemplo para estimar la verdadera probabilidad de rechazo de la hipótesis nula o la verdadera probabilidad de recubrimiento. El estimador insesgado más obvio es la frecuencia relativa $\hat{p}_{1.}$ del suceso $\{Y = 1\}$. Asumamos que, juntamente con un valor de Y, en cada réplica de simulación se produce un valor de otra variable dicotómica C, correlacionada con Y, cuya esperanza $p_{.1} = E(C) = P(C = 1)$ sea conocida. Por ejemplo C puede ser un test paramétrico cuya curva de potencia sea conocida, relacionado con el nuevo test Y bajo estudio de Monte Carlo o un intervalo de confianza relacionado cuyo verdadero recubrimiento sea conocido.

Al realizar el proceso de simulación se obtienen n pares de datos (Y_k, C_k) que se pueden resumir en forma de una tabla de contingencia 2×2:

Tabla 1

	$C=0$	$C=1$	
$Y=0$	n_{00}	n_{01}	$n_{0.}$
$Y=1$	n_{10}	n_{11}	$n_{1.}$
	$n_{.0}$	$n_{.1}$	n

generada de acuerdo con las siguientes probabilidades p_{ij}

Tabla 2

	$C=0$	$C=1$	
$Y=0$	p_{00}	p_{01}	$p_{0.}$
$Y=1$	p_{10}	p_{11}	$p_{1.}$
	$p_{.0}$	$p_{.1}$	1

En estas condiciones, se puede obtener el estimador máximo verosímil de $p_{1.}$ que es

$$\tilde{p}_{1.} = p_{.0}\frac{n_{10}}{n_{00}+n_{10}} + p_{.1}\frac{n_{11}}{n_{01}+n_{11}} \,. \tag{11}$$

Se ha de notar que este estimador, $\tilde{p}_{1.}$, es válido sólo para valores distintos de cero de $n_{.j} = n_{0j} + n_{1j}$. Por tanto, condicionado por el suceso $\{0 < n_{.0} < n\}$.

Al aplicar los resultados clásicos de estimación máximo verosímil (Rothery, 1982) y del método delta (Ocaña y Vegas, 1995; Vegas 1996 donde se demuestra también que (11) coincide con el estimador que se obtendría mediante la técnica de reducción de la varianza conocida como «variables de control») se muestra que este estimador es asintóticamente normal y que su varianza puede ser estimada mediante

$$\hat{\sigma}^2_R = \frac{1}{n}\left\{\tilde{p}_{1.}(1-\tilde{p}_{1.}) - \frac{(\tilde{p}_{1.}p_{.1} - \tilde{p}_{11})^2}{(1-p_{.1})p_{.1}}\right\} \,. \tag{12}$$

El estimador (11) tiene como propiedades (Ocaña y Vegas, 1995; Vegas 1996): ser insesgado y con varianza nunca mayor a la varianza de la frecuencia relativa de $\{Y = 1\}$ e igual a

$$var(\tilde{p}_{1.}) = p_{00}p_{10}E\left\{n_{.0}^{-1}\right\} + p_{01}p_{11}E\left\{n_{.1}^{-1}\right\} \tag{13}$$

donde $E\left\{n_{.i}^{-1}\right\}$ es la esperanza de la inversa de la variable binomial $n_{.i}$, truncada a $\{0 < n_{.0} < n\}$. Además, $\tilde{p}_{1.}$ es asintóticamente eficiente.

En Ocaña y Vegas (1995) se demuestra que el estimador (12) tiene sesgo negativo mientras que el estimador

$$\hat{\sigma}^2_U = \tilde{p}_{00}\tilde{p}_{10}\frac{E\left\{n_{.0}^{-1}\right\}}{1-E\left\{n_{.0}^{-1}\right\}} + \tilde{p}_{01}\tilde{p}_{11}\frac{E\left\{n_{.1}^{-1}\right\}}{1-E\left\{n_{.1}^{-1}\right\}} \tag{14}$$

es insesgado. Sin embargo, la esperanza $E\left\{n_{.i}^{-1}\right\}$ no tiene forma cerrada aunque se puede calcular numéricamente. Si se emplea la aproximación $E\left\{n_{.i}^{-1}\right\} \simeq (np_{.i} - (1-p_{.i}))^{-1}$ (Grab and Savage, 1954) se obtiene un nuevo estimador

$$\hat{\sigma}^2_{GS} = \frac{\tilde{p}_{00}\tilde{p}_{10}}{np_{.0} - (1-p_{.0}) - 1} + \frac{\tilde{p}_{01}\tilde{p}_{11}}{np_{.1} - (1-p_{.1}) - 1} \tag{15}$$

que es prácticamente insesgado y de cálculo más sencillo que (14). Por tanto, este último se escoge para ser utilizado como estimador de la varianza de $\tilde{p}_{1.}$ en los estudios de simulación.

Utilizando esta técnica se logran unas reducciones de la varianza importantes, de hasta un 95%, en comparación con las que se obtiene con el estimador habitual, frecuencia relativa ($\hat{p}_{1.}$), con un incremento computacional y complejidad del diseño bajo. Un factor importante para conseguir una alta reducción de la varianza en la variable de estudio Y se basa en escoger adecuadamente la variable de control C de tal manera que exista una alta correlación entre ambas variables.

REFERENCIAS

[1] **Amari, S-I, Barndorff-Nielsen, O.E., Kass, R.E., Lauritzen, S.L.** and **Rao, C.R.** (1987). *Differential geometry in statistical inference*. Volume 10 of *Lecture Notes-Monograph Series*, Institute of Mathematical Statistics, Hayward, California.

[2] **Aspin, A.A.** (1948). «An examination and further development of a formula arising in the problem of comparing two mean values». *Biometrika*, **35**, 88–96.

[3] **Atkinson, C.** and **Mitchell, A.F.S.** (1981). «Rao's distance measure». *Sankhyā*, **43**, A:345–365

[4] **Buehler, R.J.** (1959). «Some validity criteria for statistical inferences». *Ann. Math. Statist.*, **30**, 845–867.

[5] **Burbea, J.** and **Rao, C.R.** (1982). «Entropy differential metric, distance and divergence measures in probability spaces: a unified approach». *J. Multivariate Anal.*, **12**, 575–596.

[6] **Burbea, J.** and **Oller, J.M.** (1989). *On Rao distance asymptotic distribution*. preprint series 67, Universitat de Barcelona, June.

[7] **Cochran, W.G.** (1964). «Approximate significance levels of the Behrens-Fisher test». *Biometrics.*, **20**, 191–195.

[8] **Cochran, W.G.** and **Cox, G.M.** (1950). *Experimental Designs*. John Wiley and Sons, New York.

[9] **Dwass, M.** (1957). «Modified randomization tests for nonparametric hypotheses». *Ann. Math. Stat.*, **28**, 181–187.

[10] **Efron, B.** (1975). «Defining the curvature of a statistical problem (with application to second order efficiency) (with discussion)». *Ann. Statist.*, **3**, 1189–1242.

[11] **Efron, B.** (1978). «The geometry of exponential families». *Ann. Statist.*, **6**, 362–376.

[12] **Efron, B.** and **Hinkley, D.V.** (1978). «Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information». *Biometrika*, **65**, 457–487.

[13] **Fisher, R. A.** (1956). «On a test of significance in Pearson's Biometrika Tables (No. 11)». *J. R. Statist. Soc. Serie B*, **18**, 56–60.

[14] **Grab, E.L.** and **Savage, R.** (1954). «Tables of the expected value of 1/X for positive Bernouilli and Poisson variables». *J. Amer. Statist. Ass.*, **49**, 169–177.

[15] **Lee, A.F.S.** and **Gurland, J.** (1975). «Size and power of tests for equality of means of two normal populations with unequal variances». *J. Amer. Statist. Ass.*, **70**, 933–941.

[16] **Lilliefors, H.W.** (1967). «On the Kolmogorov-Smirnov test for normality with mean and variance unknown». *J. Amer. Statist. Ass.*, **63**, 339–402.

[17] **Mason, A.L.** and **Bell, C.B.** (1986). «New Lilliefors and Srinivasan tables with applications». *Commun. Statist.- Simula.*, **15(2)** 451–477.

[18] **Oost, B.A., Veldhuyzen, B., Timmermans, A.P.M.** and **Sixma, J.J.** (1983). «Increased urinary β-thromboglubin excretion in diabetis assayed with a modified RIA kit-technique». *Thrombosis and Haemostasis*, **49**, 18–20.

[19] **Pearson, E.S.** and **Hartley, H.O. (eds.)** (1976). *Biometrika Tables for Statisticians*. Volume 1, Cambridge University Press, Cambridge.

[20] **Ocaña, J.** and **Vegas, E.** (1995). «Variance reduction for Bernoulli response variables in simulation». *Computational Statistics and Data Analysis*, **19**, 631-640.

[21] **Rao, C.R.**. (1945). «Information and accuracy attainable in the estimation of statistical parameters». *Bull. Calcutta Math. Soc.*, **37**, 81–91.

[22] **Rothery, P.** (1982). «The use of control variates in Monte Carlo estimation of power». *Appl. Statist.*, **31**, 125–129.

[23] **Scheffé, H.** (1970). «Practical solutions to the Behrens-Fisher problem». *J. Amer. Statist. Ass.*, **65**, 1501–1508.

[24] **Vegas, E.** (1996). *Optimización en estudios de Monte Carlo en Estadística: aplicaciones al contraste de hipótesis*. Ph.D. Thesis, Universitat de Barcelona.

[25] **Vegas, E.** and **Ocaña, J.** (1996). «Variance reduction in the study of a new test concerning the Behrens-Fisher problem». *International Journal of Computer Simulation*, in press.

[26] **Welch, B.L.** (1938). «The significance of the difference between two means when the population variances are unequal». *Biometrika*, **29**, 350–362.

[27] **Welch, B.L.** (1947). «The generalization of 'Students' problem when several different population variances are involved». *Biometrika*, **34**, 28–35.

ENGLISH SUMMARY

THE BEHRENS-FISHER PROBLEM IN BIOMEDICAL RESEARCH. CRITICAL ANALYSIS OF A CLINICAL STUDY BY SIMULATION

ESTEBAN VEGAS LOZANO*

Universitat de Barcelona

This paper reviews the Behrens-Fisher problem. Inferential foundations associated with the difficulty of its resolution are discussed and the most common practical solutions are exposed, together with a new solution based on concepts of differential geometry. Next, a critical study of biomedical research is presented in which the true probabilities of error are different from the expected values since probable differences between the variances are ignored. In this research the null hypothesis of equality of mean was rejected ($p < 0.01$), although, the true probability of type I error for values close to sample values may be different from nominal value (0.05). With this purpose, a Monte Carlo study is performed to obtain these estimations according to use of the ***t*** *test for equality of means or other solutions more appropriate for Behrens-Fisher problem. In this simulation study a specific variance-reduction technique for dichotomous response variables such as statistical tests (to accept or reject the null hypothesis) is used. This technique is shown briefly and the design of the simulation is illustrated.*

Keywords: Monte Carlo simulation, Variance reduction, Statistical curvature, Rao's distance, Bootstrap.

AMS Classification: 65C05

---

*Esteban Vegas Lozano. Departament d'Estadística. Facultat de Biologia. Universitat de Barcelona. Diagonal 645, 08028 Barcelona. Espanya.

–Received january 1997.

–Accepted may 1997.

1. INTRODUCTION

The Behrens-Fisher problem, i. e., testing for equality of means of two normal populations without making any assumption about variances, is a complex problem, without a known optimal solution.

The difficulty is reflected, on an abstract level, in the geometry of certain parametrical families of density. Thus, the parametrical submodel associated with the null hypothesis of equality of means is a curved exponential family of order 4. Therefore, this causes a breakdown of very nice properties for estimation, testing, and other inference problems associated with flat exponential families (see Efron (1975), Efron (1978) and Efron and Hinkley (1978)).

Some common practical solutions are: Cochran and Cox test, Welch test and Welch-Aspin test (see Scheffé (1970), Lee and Gurland (1975)). Another alternative is the geodesic test described in Vegas and Ocaña (1996) and Vegas (1996), based on differential geometry. The proposed statistic, D^2, shows the discomformity level, in terms of geodesic distance, between the values obtained in the point estimations of the average and standard deviation of the sample $(\bar{x}_1, \bar{x}_2, s_1, s_2)$, and the set of all the parametric points in accordance with the null hypothesis of equality of means with arbitrary standard deviations $(\mu, \mu, \sigma_1, \sigma_2)$.

2. CRITICAL ANALYSIS OF A BIOMEDICAL STUDY

The data belong to a more extensive study about thrombosis (Oost *et al.*(1983)). This review is based on the measures obtained on urinary β-thromboglubin excretion in 12 normal patients and 12 diabetic patients. The null hypothesis of no difference between means was rejected ($p < 0.01$) using the **t** test. The true probability of type I error for values close to sample values may be different since probable differences between the variances ($p < 0.01$ using the F test of variance comparison) are ignored.

With this purpose, a Monte Carlo study is performed to obtain the estimations of the true type I error probabilities when the **t** test for equality of means or other solutions (more appropriate for Behrens-Fisher problem) are used. Likewise, the effect of different variance quotients (1.40, 16.88 and 4.86, which correspond to the extreme values and middle point of the 95% confidence interval for variance quotients respectively) is analysed.

The precision of the estimations in the Monte Carlo study has been improved by the use of a variance-reduction technique, suitable for dichotomous response variables. This technique is described in Ocaña and Vegas (1995) and Vegas (1996)). In each replication of the simulation, besides the value of the dichotomous response variable

Y (in this case $Y = 1$ if the null hypothesis in the statistic test studied is rejected, and $Y = 0$ otherwise) an additional dichotomous control variable (C) is obtained. C should be a variable correlated with Y with known expectation. In the present study, C corresponds to the final outcome of the **t** test of comparison of means for equal variances.

Common random variates are used to induce the correlation required between the tests under study (**t** incorrectly used over samples from normal populations with different variances ($t(\mathbf{x})$), Cochran and Cox test ($t_{cc}(\mathbf{x})$), Welch test ($t_w(\mathbf{x})$) and geodesic test ($D^2(\mathbf{x})$)) with the «control» **t** test ($t(\mathbf{v})$), evaluated over samples generated according to equal variances (see fig. 1). More specifically, from n_1 standard normal *iid* values, $z_1 = (z_{11}, \ldots, z_{1n_1})$, independent of other n_2 standard normal *iid* values, $z_2 = (z_{21}, \ldots, z_{2n_2})$, are obtained the $n_1 + n_2$ values $x = (x_{11}, \ldots, x_{1n_1}; x_{21}, \ldots, x_{2n_2})$ using the transformation $x_{ij} = \sigma_i z_{ij} + \mu_i$, $i = 1,2$ y $j = 1, \ldots, n_i$. These samples are used to evaluate the four tests under different configurations of means and variances. The same transformation is applied again, but now with common $\sigma^2 = \dfrac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2}$, an intermediate value between σ_1^2 and σ_2^2. These new $n_1 + n_2$ values, $v = (v_{11}, \ldots, v_{1n_1}; v_{21}, \ldots, v_{2n_2})$, are used to evaluate the control **t** test ($t(v)$).

In this way, the $n_1 + n_2$ values of v fulfil the conditions of **t** test applicability ($t(\mathbf{v})$): they come from two normal populations with common variance, and therefore, they may be used as a control variable with known power (that is, expectation).

3. CONCLUSION

The geodesic and Welch tests maintain the nominal significance level of 0.05 under all variance ratios and, additionally, they are very similar. On the other hand, the Cochran and Cox test is very conservative under similar variances and the **t** test becomes worse as the variance quotients diverge from unity (see table 1 and figure 2).

Therefore, in the study of (Oost *et al.*(1983)), the true p-values do not correspond to the nominal ones, the null hypothesis of equality of means is rejected falsely more frequently than expected and this tendency increases with the variance quotient (see fig. 2). This means that the **t** test is not recommendable under conditions of extreme difference between variances or when the statistic values are on the border between the two possible decisions (to accept or reject the null hypothesis). In any case, the final conclusions of the paper (not the «p-values») are «correct», since they are consistent with the results obtained by all statistical techniques that are more appropriate than the **t** test.

On the other hand, the variance-reduction technique used to study the probability of type I error by Monte Carlo simulation may be used in other similar studies with a suitable effectiveness and a minimum effort in the complexity of the simulation design.