

DEPARTAMENTO DE ESTADÍSTICA DE LA UNIVERSIDAD DE BARCELONA

PROGRAMA DE PROBABILITATS I ESTADÍSTICA (BIENNIO 87/89)

Optimización en estudios de Monte Carlo en Estadística: Aplicaciones al Contraste de Hipótesis.

Memoria presentada para optar
al título de Doctor en Biología
por la Universidad de Barcelona,
por

Esteban Vegas Lozano

Visto bueno
El Director,

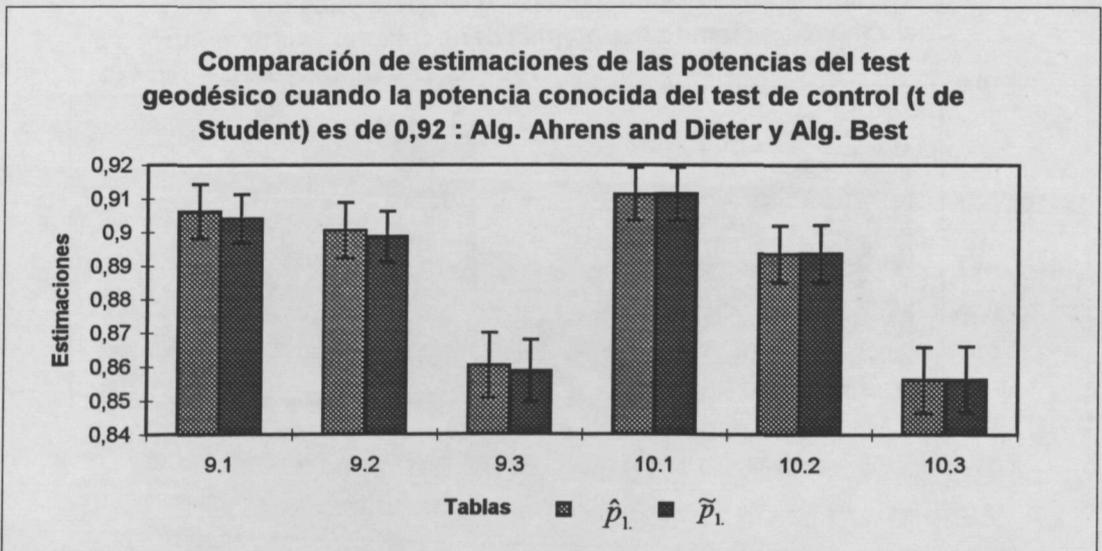
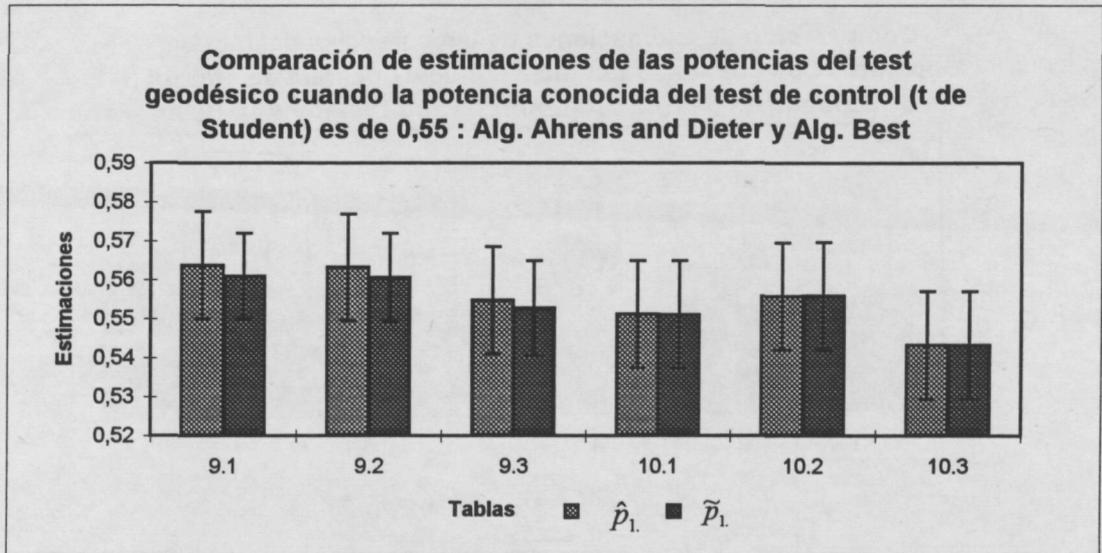
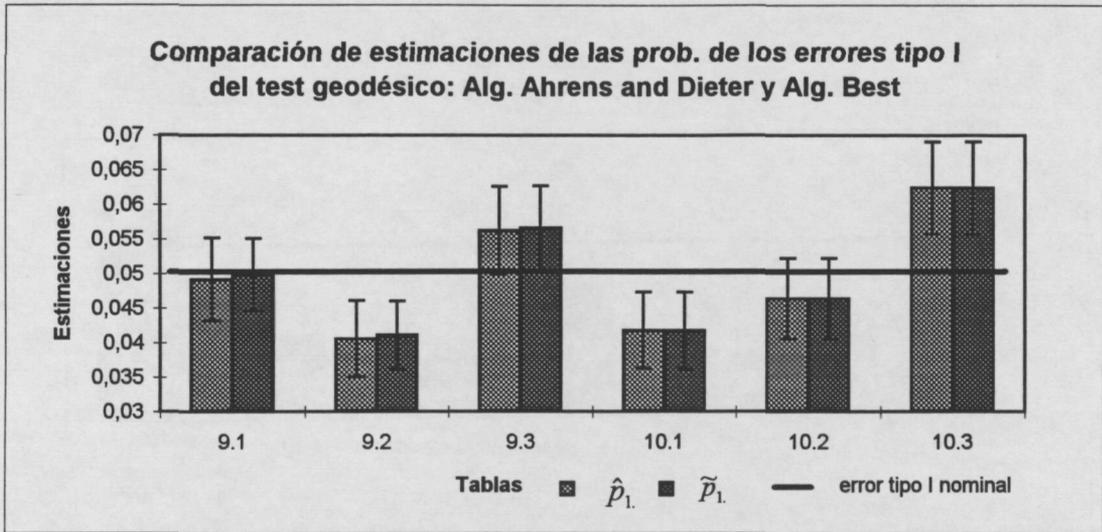
Dr. Jordi Ocaña Rebull,
Catedrático de Universidad.
Departamento de Estadística.
Universidad de Barcelona.

Barcelona, 20 de Junio de 1996

5.2.4 Gráficos

Para mostrar las diferencias entre los resultados obtenidos , en cada tipo de test, con el algoritmo de Best y los obtenidos con el algoritmo de Ahrens and Dieter se ha creado un gráfico para cada potencia (0.05, 0.55, 0.92 del test de control) en el que se presentan las estimaciones de las potencias para ambos algoritmos. Así, en la primera página de gráficos que nos encontramos corresponde a las tres potencias estudiadas para el test geodésico, y en la siguiente página es idéntico pero para el test de Cochran y Cox.

La última página con gráficos muestra, de forma comparativa, los cambios producidos en el tanto por ciento de reducción de la varianza en la estimación de la potencia en los estudios de robustez bajo distribución lognormal y distribución gamma (con los dos tipos de algoritmo) para los dos test estudiados.



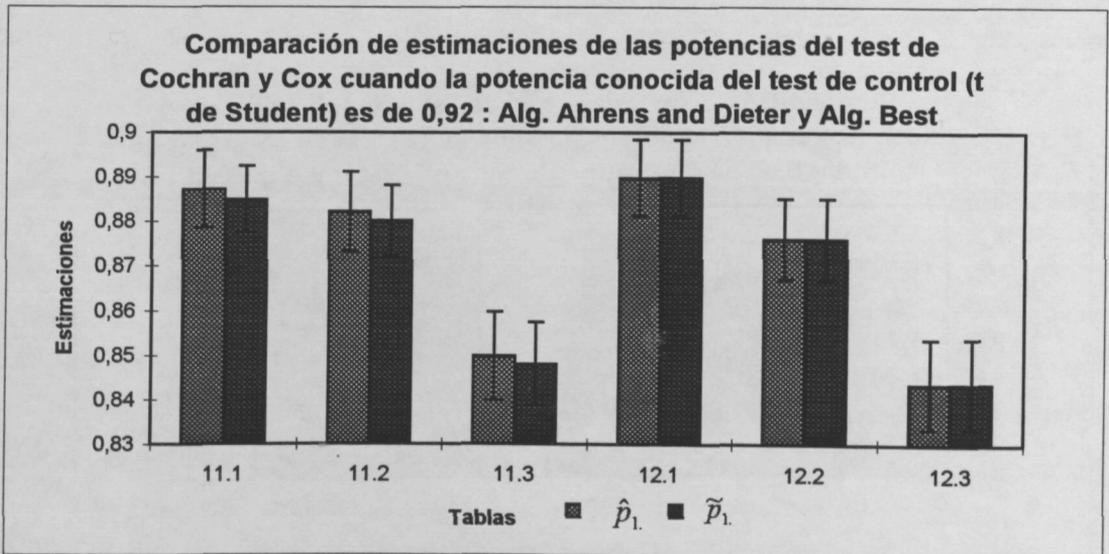
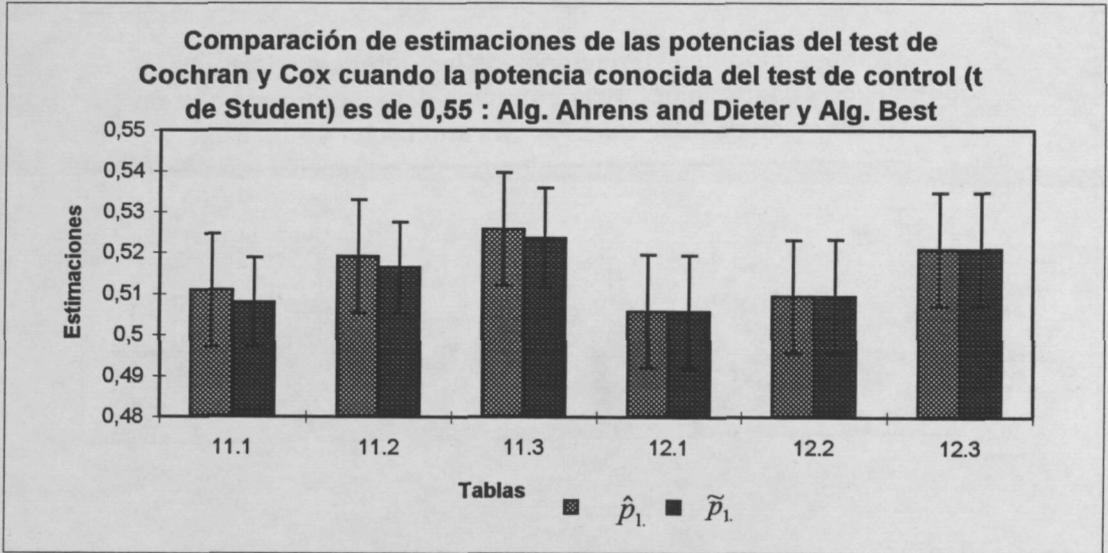
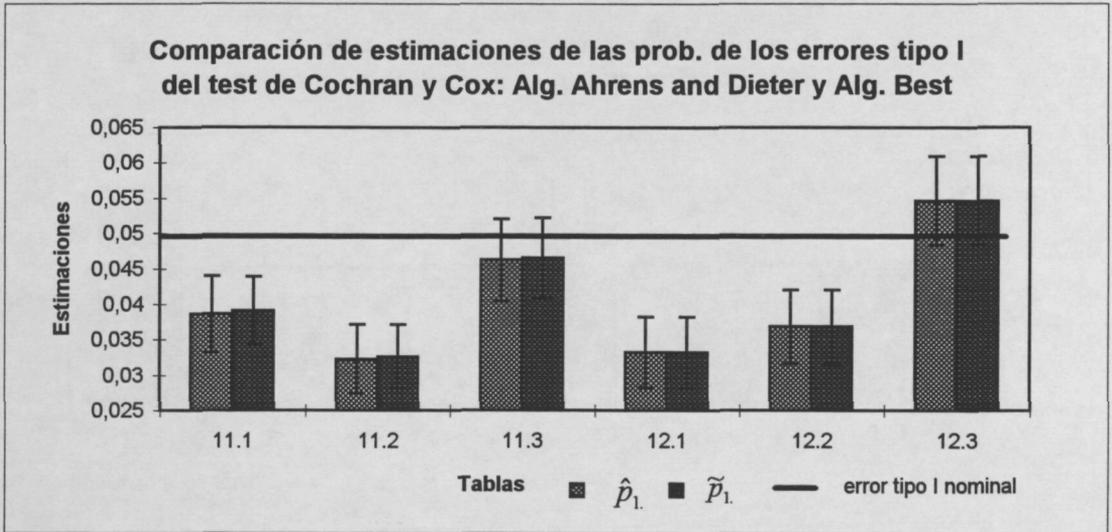


Fig 4: Comparación de estimaciones del % de varianza de las potencias del test geodésico para diferentes potencias y varianzas

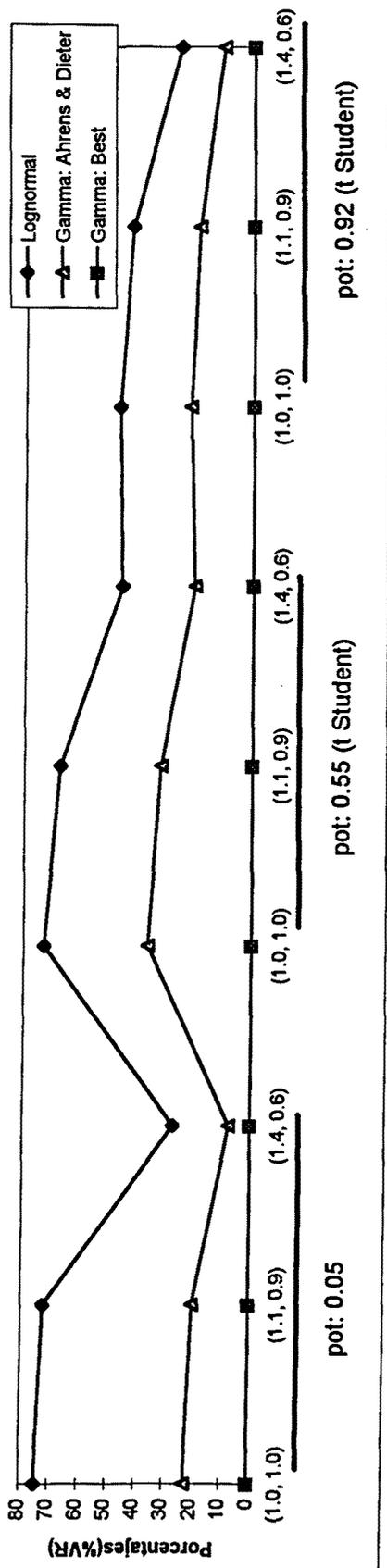
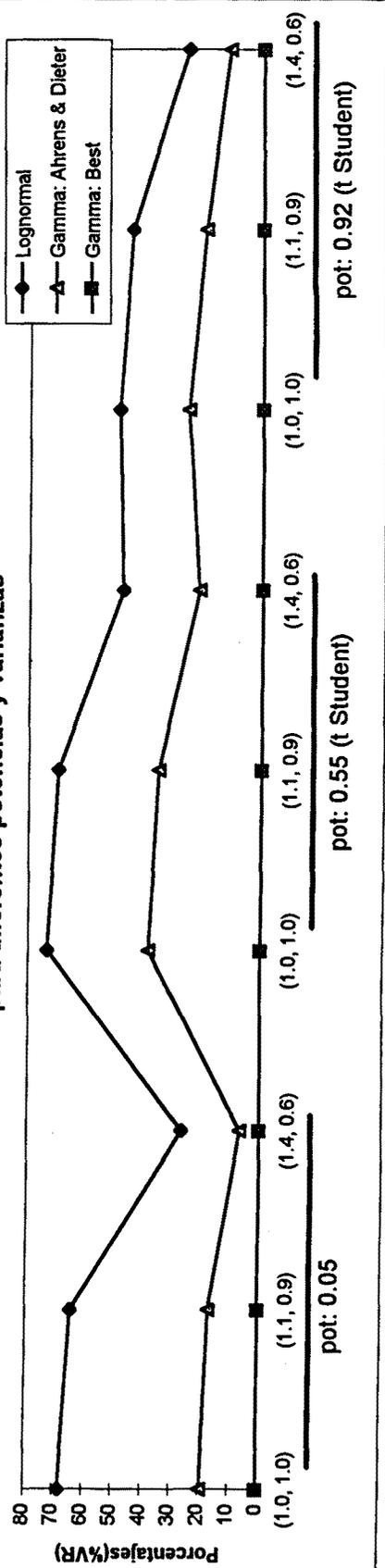


Fig 5: Comparación de estimaciones del % de varianza de las potencias del test de Cochran y Cox para diferentes potencias y varianzas



5.2.5 Conclusiones

Las simulaciones realizadas empleando la rutina G05FFF (Tablas 10 y 12) dan para las diferentes potencias unos valores estimados de reducción de la varianza cercanos a cero y en algunas ocasiones negativos debido a la baja correlación existente entre los valores obtenidos con el test de interés (test geodésico o test de Cochran y Cox) y el test de control (t de Student). Incluso con las mejores condiciones para facilitar la correlación inducida a partir del uso de variables aleatorias comunes, es decir, tamaños muestrales grandes ($n_1 = n_2 = 121$) y varianzas iguales ($\sigma_1^2 = \sigma_2^2 = 1$), no se consigue mejorar los porcentajes de reducción de la varianza.

Este hecho es debido a la pérdida de sincronía de la serie de valores normales con la serie de valores gamma ya que, como se indicó en la introducción, la técnica utilizada para la generación de la gamma es el método de rechazo. El método de rechazo que se utiliza, Best (1978), necesita generar un número aleatorio de valores (por los menos dos, uno seguiría una t de Student con dos grados de libertad y el otro una distribución uniforme para obtener un valor de gamma). Por contra, la generación de los valores normales es dos a dos, en el sentido de que dos valores uniformes consecutivos producen una pareja de valores normales consecutivos (método polar de Box and Muller (1958)). Por tanto, aunque se parta de los mismos valores en las semillas, rápidamente se pierde sincronía y como consecuencia no se puede inducir la correlación.

En cambio, los resultados obtenidos con la utilización del algoritmo de Ahrens and Dieter (Tablas 9 y 11) son mejores. En cada paso, el valor normal estándar necesario para la posterior evaluación del test t de Student se introduce como entrada en el algoritmo de generación de la gamma. En la mayoría de los casos el valor gamma generado se basa en este valor normal estándar. Consiguiendo, de esta manera, mantener una gran sincronía en la generación de valores normales y gamma. Esta sincronía repercute en el buen funcionamiento de la técnica de reducción de la varianza. Incluso, teniendo en cuenta que los valores normales no están muy correlacionados con los valores gamma porque estos últimos se obtienen, generalmente, como el cuadrado de una función lineal de cada valor normal.

Si ahora se compara estos resultados con los obtenidos en el estudio de robustez bajo lognormalidad presentados en la sección anterior (ver fig. 4 y 5) se advierte

que la variación de los porcentajes de reducción de varianza para diferentes potencias y configuraciones de parejas de varianzas en el estudio de robustez bajo distribución gamma (alg. Ahrens and Dieter) son muy similares a los que se producen en el estudio de robustez bajo lognormalidad. Además, los porcentajes de reducción de la varianza obtenidos con el algoritmo de Ahrens and Dieter son de un orden entre un 25% y 60% de los valores obtenidos bajo lognormalidad que son los mejores posibles ya que la sincronía existente entre la generación de los valores normales y los lognormales es la máxima. Por tanto, los porcentajes de reducción de la varianza obtenidos con el algoritmo de Ahrens and Dieter son más altos que lo que parecen.

Haría falta un estudio más a fondo de la robustez bajo la distribución gamma pero con los resultados presentados se puede lograr tener unos retazos importantes de su comportamiento.

Los resultados obtenidos con ambos algoritmos son similares. En el caso del test de Cochran y Cox se puede observar que sólo para la pareja de varianzas más diferentes (1.4, 0.6) el intervalo de confianza de la probabilidad de error de tipo I incluye el nivel de significación nominal (0.05) en los otros dos casos no (tabla 12). Quedando de manifiesto su carácter más conservador cuanto más iguales sean las varianzas. Para potencias de 0.55 de la *t* de Student muestran un comportamiento similar al del estudio bajo condiciones normales mientras que para potencia de 0.92 es semejante al del estudio de robustez bajo lognormalidad.

En el test geodésico se observa un carácter menos conservador que en el test anterior. Para potencias de 0.55 y 0.92 de la *t* de Student su comportamiento es similar al del estudio de robustez bajo lognormalidad, preferentemente en el caso de la potencia mayor.

En resumen se puede decir que es muy importante, a la hora de aplicar la técnica de reducción de la varianza, que haya la mayor correlación posible entre la variable respuesta y la variable de control. Esto no se consigue simplemente induciendo la correlación a partir del uso de variables comunes, sino que también es importante mantener la sincronía de la serie aleatoria. Si ésta se pierde, no vale la pena aplicar este método. Solamente un buen conocimiento de las técnicas de generación de variables aleatorias, y no una utilización ciega de paquetes comerciales puede permitir diagnosticar y solucionar tales problemas de diseño de las simulaciones.

Capítulo 6

ESTUDIOS DE SIMULACIÓN SOBRE ALGUNOS EJEMPLOS DEL PROBLEMA DE BEHRENS-FISHER EN EL ÁMBITO DE LA EXPERIMENTACIÓN MÉDICO-BIOLÓGICA

Se presentan dos casos con datos reales¹ donde surge el problema de Behrens-Fisher. En concreto se realiza un análisis crítico de las conclusiones de dos estudios experimentales, en el ámbito 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 varianzas. En ambas situaciones, el primer paso será hacer un estudio pormenorizado sobre los resultados diferentes obtenidos mediante varios test para resolver el problema de Behrens-Fisher, incluyendo el nuevo test geodésico y, se contrastarán estos resultados con obtenidos al aplicar el test t de Student. A continuación, según el resultado de los tests, se realizarán estudios de simulación para:

- Si se acepta H_1 : estimar cual es la verdadera probabilidad de error de tipo I.
- Si se acepta H_0 : estimar cual es la verdadera probabilidad de error de tipo II.

Así, en este capítulo se intenta resaltar que tanto el test geodésico como la simulación (mejorada mediante la técnica de reducción de la varianza propuesta en capítulos anteriores) son métodos fácilmente aplicables por cualquier estadístico para problemas reales.

6.1 Un estudio sobre la trombosis

La trombosis, formación de trombos (coágulo sanguíneo intravascular) en el aparato circulatorio, produce una disminución del riego sanguíneo en la zona afectada y, por lo tanto, puede provocar, según la intensidad y el lugar, algunas de las más conocidas causas de muerte dentro de nuestra sociedad: infartos y embolias cerebrales. Por esta razón tiene gran importancia conocer que tipos de personas son más propensas.

¹Proviene del libro D.J. Hand *et al.* (1994) que recopila gran cantidad de datos de diferentes contextos.

Los datos presentados son una parte de un estudio más amplio que fue expuesto en un artículo de van Oost *et al.*(1983). Nosotros nos centraremos en las medidas obtenidas sobre la excreción de β -tromboglobulina 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 -tromboglobulina urinaria entre los dos tipos de pacientes. A pesar de que no se indica claramente en el artículo, todo parece indicar que se realizó un test t de Student de comparación de medias para llegar a tal conclusión.

6.1.1 Comparación de medias entre los dos tipos de pacientes

En la tabla adjunta tenemos una estadística descriptiva mínima de las dos muestras.

	n	\bar{x}	\hat{s}^2
Diabético	12	35.275	410.873
Normal	12	13.533	84.539

En una primera impresión se observa que las varianzas y las medias muestrales son bastante distintas, dando pie a que pueda existir diferencias entre las medias poblacionales. Aunque antes de hacer ningún vaticinio sobre el contraste de medias se debe de verificar la normalidad y la igualdad de varianzas.

Por tanto, haremos la prueba de Lilliefors ya que los parámetros de la teórica distribución normal son desconocidos y como consecuencia hay que estimarlos. Las tablas que utilizaremos 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 existe diferencias significativas entre ellas ($p < 0.01$).

Por consiguiente, estamos en las condiciones requeridas en el problema de Behrens-Fisher. El estadístico T' (4.7) en el que se basan los tests de Cochran y Cox, Welch y Welch-Aspin da como resultado un valor de 3.3838 que, juntamente con el test geodésico², conduce 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, cualquiera de los cuatro, se rechaza la hipótesis nula, así pues, el estudio más interesante consiste en estimar cual es la verdadera probabilidad de error de tipo I.

6.1.2 Estimación de la verdadera probabilidad de error de tipo I

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.

6.1.2.1 Procedimiento

La forma de proceder fue análoga a la del estudio de simulación de la potencia del test geodésico visto en el capítulo denominado "Reducción de la varianza en el estudio de un nuevo test para el problema de Behrens-Fisher".

Como ya vimos, el algoritmo de simulación se divide en tres módulos:

²Para 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. Los resultados obtenidos fueron para un bootstrap paramétrico de 1000 remuestras corriendo en un IBM POWERparallel SP2. El programa es de libre uso, por tanto, está a disposición de cualquiera que lo desee.

1. Entrada de parámetros.
2. Obtención de los dos vectores de resultados (Y_i, C_i) $i = 1, \dots, n$.
3. Estimación de la probabilidad de error de tipo I (con o sin aplicación de la técnica de reducción de la varianza).

En el apartado de la entrada de parámetros los valores fijos serán

$$n_1 = n_2 = 12 \quad \beta_c = 0.05 \quad \alpha = 0.05 \quad n = 5000 \quad B = 1000$$

mientras que utilizaremos 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 (1.39, 16.88) y el punto medio (4.86) del intervalo de confianza del 95% para el cociente de varianzas poblacionales, respectivamente. Se han elegido aquellos valores de las varianzas en que al menos una de ellas sea igual que la alcanzada 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 lograr ver 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 dividirá en (ver 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 ($D^2(\mathbf{x})$).

se necesita realizar tantos tipos de simulaciones como diferentes tipos de tests. Así, se substituirá la frase “Calcular el test de estudio”, que aparece en el párrafo anterior, por cada uno de ellos.

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 ha utilizado variables aleatorias comunes para evaluar cada uno de ellos.

A partir de n_1 valores normales estándar *iid*, $\mathbf{z}_1 = (z_{11}, \dots, z_{1n_1})$, independientes de otros n_2 valores normales estándar *iid*, $\mathbf{z}_2 = (z_{21}, \dots, z_{2n_2})$, se han ido obteniendo, gracias a la transformación $x_{ij} = \sigma_i z_{ij} + \mu_i$, $i = 1, 2$, los $n_1 + n_2$ valores $\mathbf{x} = (x_{11}, \dots, x_{1n_1}; x_{21}, \dots, x_{2n_2})$ 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 un valor común e intermedio (4.25), entre σ_1^2 y σ_2^2 , se obtiene $n_1 + n_2$ valores $\mathbf{v} = (v_{11}, \dots, v_{1n_1}; v_{21}, \dots, v_{2n_2})$ que servirá 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.

Tablas 1 hasta 4. Nivel de significación y grado de reducción de la varianza en el test geodésico, test de Welch, test de Cochran y Cox y 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.

† =significativamente diferente del nivel de significación nominal.

Tabla 1: $n = 5000$, $n_1 = n_2 = 12$ $B = 1000$

σ_1^2/σ_2^2	test geodésico					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
1.40	0.0490	0.005984	0.0494	0.002568	81.58	57.08
4.86	0.0532	0.006221	0.0535	0.004796	40.57	22.91
16.88	0.0544	0.006287	0.0546	0.005631	19.79	10.44

Tabla 2: $n = 5000$, $n_1 = n_2 = 12$

σ_1^2/σ_2^2	test de Welch					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
1.40	0.0492	0.005996	0.0496	0.002311	85.14	61.45
4.86	0.0532	0.006221	0.0535	0.004775	41.10	23.25
16.88	0.0528	0.006199	0.0530	0.005540	20.00	10.63

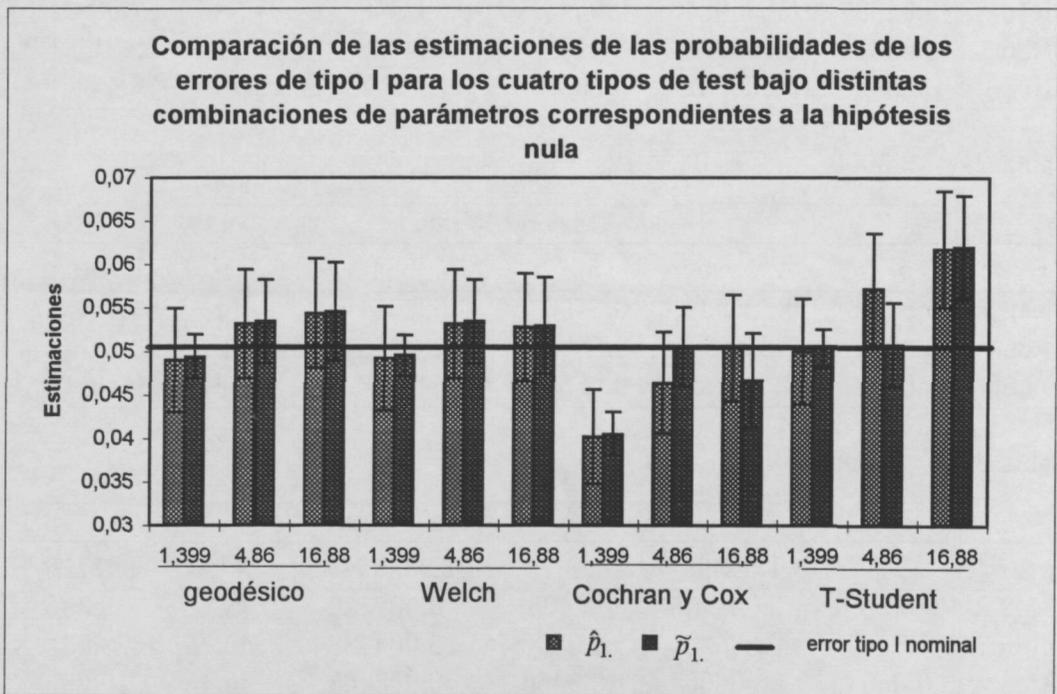
Tabla 3: $n = 5000$, $n_1 = n_2 = 12$

σ_1^2/σ_2^2	test de Cochran y Cox					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
1.40	0.0402†	0.005445	0.0405†	0.002539	78.26	53.38
4.86	0.0464	0.005831	0.0506	0.004488	40.76	23.03
16.88	0.0504	0.006064	0.0466	0.005408	20.48	10.83

Tabla 4: $n = 5000$, $n_1 = n_2 = 12$

σ_1^2/σ_2^2	test t de Student					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
1.40	0.0500	0.006042	0.0504	0.002191	86.84	63.73
4.86	0.0572†	0.006438	0.0507	0.004838	43.51	24.84
16.88	0.0618†	0.006675	0.0620†	0.005906	21.71	11.52

Por otro lado, se expone un gráfico asociado a las anteriores tablas que mantiene la estructura de uno de los gráficos vistos en el estudio de la potencia del test geodésico en el capítulo 4. La única diferencia es que el eje de abscisas indica los tres valores distintos del cociente de varianzas poblacionales (1.399, 4.86 y 16.88) que se han estudiado para cada test.



6.1.2.3 Conclusiones

Como se puede observar, tanto en las tablas como en los gráficos, 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 la estimación del error estándar de la frecuencia relativa (\hat{p}_1) y del error estándar de la aproximación de Grab & Savage del nuevo estimador (\tilde{p}_1) para cualquiera de los tres cocientes de varianzas estudiados. Además, los diferentes valores de la tabla: \hat{p}_1 , $\pm 1.959964\hat{\sigma}$, \tilde{p}_1 , $\pm 1.959964\hat{\sigma}_{GS}$, ... son muy semejantes.

Como ya se indicó en la sección de conclusiones de la potencia del test geodésico el test de Cochran y Cox es conservativo (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 queda reflejada al estar excluido en el intervalo de confianza aproximado del 95% para el nivel de significación el valor de 0.05, tanto para la estimación del error estándar con \tilde{p}_1 como para \hat{p}_1 , cuando el cociente de varianzas poblacionales es de 1.399. Para valores superiores, sí que se incluye el valor de 0.05 de nivel de significación nominal. Gráficamente se observará como en el primer caso la línea horizontal de valor 0.05 no corta los dos segmentos que representan los intervalos de confianza aproximados del 95% para el nivel de significación construidos a partir de \tilde{p}_1 y \hat{p}_1 correspondientes al cociente de varianzas poblacional igual a 1.399 para el test de Cochran y Cox mientras que para los valores de 4.86 y 16.88 sí que lo cortan.

Para el test t de Student ocurre al revés, como era de esperar. Cuando el cociente de varianzas es de 1,399 el intervalo de confianza aproximado del 95% para el nivel de significación obtenido a partir de \tilde{p}_1 y \hat{p}_1 incluye el valor de 0.05 de nivel de significación nominal. Cuando es de 4,86 sólo lo incluye para el caso de \tilde{p}_1 . Y cuando es de 16,88 la estimación de la probabilidad de error de tipo I es superior a 0.05 para ambos casos.

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 los distintos estadísticos estudiados 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. Utilizar \tilde{p}_1 en vez del habitual \hat{p}_1 para estimar el nivel de significación ha provocado que los intervalos de confianza

sean más pequeños.

Por otro lado, los porcentajes estimados de reducción de la varianza al utilizar \tilde{p}_1 en vez de \hat{p}_1 son similares en los diferentes tests. Al aumentar el valor del cociente entre varianzas disminuye el porcentaje estimado de reducción de la varianza ya que cada vez hay menos correlación con los valores obtenidos con el test t de Student utilizado como control. Por tanto los valores más altos de reducción de la varianza, para cualquier valor del cociente entre varianzas, son los obtenidos cuando la variable de estudio es el test t de Student, obviamente el más correlacionado con la variable de control, que a su vez es el test t de Student.

6.1.2.4 Implicaciones en cuanto a la interpretación de resultados

Es habitual en trabajos médico-biológicos cometer errores de utilización inadecuada de un test porque no cumple las condiciones requeridas para su uso. Como se vió en la subsección de comparación de medias entre los dos tipos de pacientes se puede aceptar la normalidad de las dos muestras pero no así la igualdad de varianzas. Por consiguiente, se incurre en un error si se utiliza el test t de Student de comparación de medias. Este error 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 (ver gráficas anteriores). De todas formas, si se utiliza el test t de Student cuando las varianzas no son iguales, la probabilidad de error de tipo I solamente es sensiblemente mayor que la nominal cuando la ratio de varianzas es extrema, y solamente algo mayor en caso contrario. 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 situen en la frontera entre las dos posibles decisiones (aceptar o rechazar la hipótesis nula).

En general, se recomendaría al usuario utilizar tanto el test geodésico como el test de Welch, que son igualmente adecuados para cualquier valor del cociente de varianzas. El test de Cochran y Cox es muy conservativo cuando el cociente de varianzas es próximo a uno, y el peor de todos es el test t de Student, posiblemente utilizado en el artículo que ha servido de base a la presente discusión. 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 obtendrían mediante una técnica estadística más apropiada.

6.2 Un estudio sobre la diabetes

La diabetes mellitus constituye un conjunto diverso de enfermedades que desembocan en una situación común: la existencia de valores elevados de glucemia. Esta hiperglucemia puede provocar ceguera, insuficiencia renal, etc. Debido al incremento constante de la obesidad, la vida sedentaria y la longevidad en las sociedades occidentales se propicia la aparición de un número cada vez mayor de nuevos casos cada año siendo actualmente la incidencia en la población general alrededor del 4%. Por tanto, la importancia social que tiene esta enfermedad cada día es más relevante repercutiendo principalmente en los costos sanitarios.

En el artículo de R.E.Dolkar *et al.*(1971) se presenta un estudio para dar una explicación de la incidencia y gravedad de las infecciones en pacientes diabéticos. Así, se estudia la respuesta de los anticuerpos del torrente sanguíneo en ratones normales y con diabetes alloxan con o sin tratamiento de insulina para lo cual se inyecta una proteína (BSA) que hará de marcador. Ya que nuestro interés inicial en estos datos es plantear un ejemplo del problema de Behrens-Fisher en el cual se acepte la hipótesis nula se ha escogido al grupo Normal y al de diabetes Alloxan tratada con Insulina que tienen de tamaño muestral 20 y 19, respectivamente. En la tabla siguiente se presentan los valores obtenidos de la variable cantidad de microgramos de BSA por mililitro de suero en cada grupo.

Normal	Alloxan+Insulina	Normal	Alloxan+Insulina
155.76	82.50	349.14	19.80
282.00	99.66	108.90	100.32
197.34	97.66	143.22	71.94
297.00	150.48	64.02	133.32
115.50	242.88	25.54	464.64
126.72	67.98	85.80	36.96
119.46	227.70	122.10	46.20
29.04	130.68	454.85	34.32
252.78	73.26	655.38	43.56
122.10	17.82	13.86	

Con estos datos, los autores del artículo afirman que no existe diferencia significativa ($p > 0.1$) en la media de microgramos de BSA por mililitro de suero entre los dos tipos de ratones. Aunque no se indica, es casi seguro que se realizó un test t de Student para llegar a la anterior conclusión.

6.2.1 Comparación de medias entre el grupo Normal y el diabetes Alloxan tratado con Insulina

Los valores muestrales obtenidos son presentados en esta tabla:

	n	\bar{x}	\hat{s}^2
Normal	20	186.03	25265.3
Diabetes Alloxan+Insulina	19	112.72	11200.2

Igual que se hizo en el ejemplo de la trombosis se verificará la normalidad y la igualdad de varianzas. Se hace la prueba de Lilliefors para media y varianza poblacional desconocida utilizando las tablas de A.L. Mason and C.B. Bell (1986) con el resultado de rechazo de la normalidad para un nivel de significación de 0.05. Observando la distribución empírica de los dos grupos y recordando que el estudio se centra en la concentración de una sustancia en la sangre, más concretamente en el suero sanguíneo, se comprueba, a partir de la prueba de Kolmogorov-Smirnov, que en cambio se ajusta a una distribución lognormal. Por otro lado, se realizó el test F de comparación de varianzas que dio como resultado la existencia de diferencias significativas entre ellas.

Bajo estas condiciones no estamos en la situación requerida en el problema de Behrens-Fisher pero sí que tenemos un caso para observar la robustez de los diferentes test que se pueden emplear. Tanto en el test geodésico, realizado mediante un programa de ordenador, como los tests de Cochran y Cox, Welch y Welch-Aspin que se sirven del estadístico T' (4.7) el cual obtuvo como resultado el valor de 1.70, coincidieron en la aceptación de la hipótesis nula de igualdad de medias. Se llega a la misma conclusión incluso en el caso de no tener en cuenta la existencia de diferencias significativas entre las varianzas y realizar un test t de Student de comparación de medias. Por tanto, se puede afirmar finalmente que cualquiera de los cuatro test utilizados anteriormente coinciden en la aceptación de la hipótesis nula. Luego un estudio adecuado sería estimar cual es la verdadera probabilidad de error de tipo II a nivel general. Sin embargo, parece más interesante estimar, al tener unos datos concretos, cual es la verdadera probabilidad de error (tipo I o II) alrededor de valores próximos a la media y varianza muestrales.

6.2.2 Estudio de simulación para estimar la verdadera probabilidad de error alrededor de los valores muestrales

La probabilidad de error de tipo II, probabilidad de aceptar la hipótesis nula cuando en realidad las medias poblacionales son diferentes, es la probabilidad complementaria a la potencia del test, así pues, nosotros utilizaremos esta propiedad para presentar estimaciones de potencias en vez de estimaciones de la verdadera probabilidad de error de tipo II.

6.2.2.1 Procedimiento

Se ha debido de pasar por varias etapas hasta conseguir los valores con los cuales se ha realizado el estudio de simulación. Para elegir los valores de media y varianza próximos a los valores muestrales se ha pensado realizar intervalos de confianza del 95% para la media y la varianza de cada grupo (Normal y Alloxan+Insulina). Hay que considerar que, como se vio en la sección precedente, los datos originales no se ajustan a una distribución normal pero sí a una lognormal. Por tanto, se han transformado los datos originales, a partir del logaritmo neperiano de estos, para que siga una distribución normal y así, poder hacer los intervalos de confianza. Una primera aproximación al número de simulaciones que sería necesario realizar se llega escogiendo para cada intervalo de confianza tres valores, los extremos y el central, como representantes de la variabilidad de las medias y varianzas. Haciendo cálculos, necesitaríamos $3 \times 3 \times 3 \times 3$ simulaciones para todas las posibles combinaciones de medias y varianzas ya que utilizamos cuatro intervalos de confianza. Para reducir este número tan elevado de simulaciones se opta por mantener fijo los valores muestrales de un grupo (el Normal), es decir, la media y varianza muestral obtenidas en el grupo Normal se adopta como valores poblacionales, y solo se hace variar los valores de la media y varianza poblacional del grupo Alloxan+Insulina como se ha indicado anteriormente. Por tanto, el número de simulaciones se reduce hasta la cantidad de 3×3 . Una vez se tiene estos 9 casos caracterizados por la pareja de medias y varianzas de cada grupo, $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)^N$, se debe obtener cuales son sus correspondientes valores bajo la distribución lognormal, $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ aplicando la expresión (5.2) vista en el capítulo 5. De esta manera se obtiene los siguientes valores:

Tabla A

Caso	Bajo Lognormalidad				Bajo Normalidad			
	μ_1	σ_1^2	μ_2	σ_2^2	μ_1	σ_1^2	μ_2	σ_2^2
1	205.28	65065.7	115.00	13548.1	4.8575	0.934	4.3923	0.70530
2	205.28	65065.7	172.38	30441.6	4.8575	0.934	4.7971	0.70530
3	205.28	65065.7	76.72	6029.6	4.8575	0.934	3.9875	0.70530
4	205.28	65065.7	99.96	5293.0	4.8575	0.934	4.3923	0.42506
5	205.28	65065.7	149.84	11892.9	4.8575	0.934	4.7971	0.42506
6	205.28	65065.7	66.69	2355.6	4.8575	0.934	3.9875	0.42506
7	205.28	65065.7	182.43	136253.0	4.8575	0.934	4.3923	1.62812
8	205.28	65065.7	273.45	306149.0	4.8575	0.934	4.7971	1.62812
9	205.28	65065.7	121.70	60639.8	4.8575	0.934	3.9875	1.62812

Puesto que los valores de la media y varianza del grupo Normal bajo lognormalidad están por encima de los valores muestrales obtenidos con los datos originales, 186 de media y 25228.5 de varianza, se sustituyen por estos últimos. Además, se observa que el valor de una de las medias del grupo de Alloxan+Insulina bajo lognormalidad para cada tipo de varianza de los valores iniciales bajo normalidad (0.7053, 0.42506, y 1.62812) es cercano al valor de 186. Así, se ha cambiado por este valor para poder estimar la verdadera probabilidad de error de tipo I. Por tanto, con 9 simulaciones estudiaremos la verdadera probabilidad de error de tipo I y II bajo unos valores cercanos a los obtenidos en la muestra, que como parece natural deben de estar próximos a los verdaderos valores poblacionales y por consiguiente tendrá mayor interés para el usuario que hacer un estudio más generalista de la estimación de la verdadera probabilidad de error de tipo II y I.

Los casos que finalmente se estudiarán son:

Tabla B

Caso	Bajo Lognormalidad				Bajo Normalidad			
	μ_1	σ_1^2	μ_2	σ_2^2	μ_1	σ_1^2	μ_2	σ_2^2
1	186	25265.3	115.00	13548.1	4.9516	0.548	4.3923	0.70530
2	186	25265.3	186	30441.6	4.9516	0.548	4.9101	0.63123
3	186	25265.3	76.72	6029.6	4.9516	0.548	3.9875	0.70530
4	186	25265.3	99.96	5293.0	4.9516	0.548	4.3923	0.42506
5	186	25265.3	186	11892.9	4.9516	0.548	5.0780	0.29548
6	186	25265.3	66.69	2355.6	4.9516	0.548	3.9875	0.42506
7	186	25265.3	186	136253.7	4.9516	0.548	4.4272	1.59705
8	186	25265.3	273.45	306149.8	4.9516	0.548	4.7971	1.62812
9	186	25265.3	121.70	60639.8	4.9516	0.548	3.9875	1.62812

Ahora se necesita obtener cual es la potencia teórica del test t de Student ($t(\mathbf{v})$) de control para poder aplicar la técnica de reducción de la varianza. Para poder determinar esta potencia se utiliza el nomograma que aparece en la tabla 10 del *Biometrika Tables* (1976). Sin embargo, se plantea una nueva duda, puesto que tenemos el vector de valores paramétricos bajo la suposición de normalidad, $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)^N$ y sus equivalentes bajo lognormalidad, $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$, ¿cuál de ellos es mejor para utilizarlo como parámetros de la distribución normal que evalúa el test t de Student ($t(\mathbf{v})$) de control? Se recuerda que la varianza común σ^2 para ambas distribuciones normales se calcula como la varianza ponderada respecto a los tamaños muestrales (4.25). Para contestar a la pregunta se realiza un estudio prospectivo que se basa en observar la correlación entre la estimación de la potencia del test utilizado como respuesta en la técnica de reducción de la varianza (Welch, Cochran y Cox, o geodésico) con la estimación de la potencia del test t de Student ($t(\mathbf{v})$) de control en las dos situaciones (valores normales o lognormales). Los resultados indicaban que la correlación era mayor cuando se utilizaba como parámetros los valores de media y varianza lognormales.

Para cada una de las 9 casos anteriores se ha de tener en cuenta que:

$$n_1 = 20 \quad n_2 = 19 \quad \alpha = 0.05 \quad n = 5000 \quad B = 1000$$

Una vez fijado todos los parámetros de la simulación el proceso ha realizar es semejante al ejemplo de trombosis. La diferencia fundamental estriba en que los los valores utilizados para evaluar los tests:

- test t de Student de comparación de medias ($t(\mathbf{x})$).
- test de Cochran y Cox ($t_{cc}(\mathbf{x})$).
- test de Welch ($t_w(\mathbf{x})$).
- test geodésico ($D^2(\mathbf{x})$).

son generados a partir de una distribución lognormal con la esperanza y varianza indicada en cada uno de los 9 casos simulados. La manera de actuar es tal como se explicó en el capítulo 5 en los procedimientos de la sección “Estudio de la robustez para muestras originadas bajo la distribución lognormal”. Además, se ha seguido la misma forma de proceder que se indicó en el caso de la trombosis

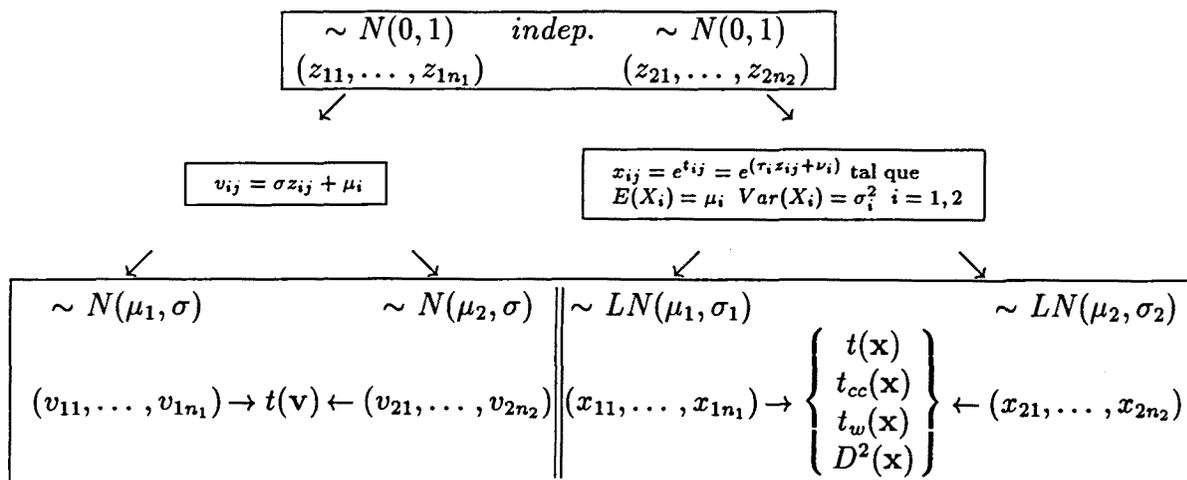
para obtener los valores normales que se emplean en la evaluación del test de control t de Student y para inducir la correlación entre el test de estudio y el test de control.

En la figura 2 se muestra esquemáticamente el proceso de obtención de los dos vectores de resultado (Y_i, C_i) $i = 1, \dots, n$, el cual se debe de realizar para cada situación experimental.

FIG.2: OBTENCIÓN DE LOS DOS VECTORES DE RESULTADOS (Y_i, C_i) $i = 1, \dots, n$ EN LA ESTIMACIÓN DE LA VERDADERA PROBABILIDAD DE ERROR DE TIPO II O I PARA VARIOS TESTS.

Repite desde $j=1$ hasta n :

Generación de las muestras



Parejas diferentes:

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

6.2.2.2 Resultados

Al examinar la tabla A con los 9 casos tratados se observa que cada tres casos, comenzando desde el primero, tienen como valores de media y varianza bajo normalidad en el grupo Alloxan+Insulina a los correspondientes valores central y extremos del intervalo de confianza de la media y una misma varianza. Posteriormente, tabla B, para cada grupo de tres casos se sustituyó la media y la

varianza bajo normalidad en el grupo Alloxan+Insulina en aquel caso con un valor de media más próximo a 186 (media del grupo Normal) bajo lognormalidad para conseguir este valor. Por ejemplo, los casos 1, 2 y 3 corresponden, tabla A, a una varianza bajo lognormalidad del grupo Alloxan+Insulina de 0.7053 para diferentes valores de medias. Se modificó el caso 2, tabla B, con una media bajo lognormalidad más cercana a 186, para que valiera esta cantidad.

Bajo esta estructura, los resultados se exponen en tres grupos de tablas. Tres casos (2, 5, 7) forman el primer grupo que servirá para la estimación de la verdadera probabilidad de error de tipo I. Los restantes servirán para la estimación de la verdadera probabilidad de error de tipo II. Se dividiran en dos grupos: (1, 4, 8) y (3, 6, 9) que son los valores mayores y menores, respectivamente, de las medias bajo lognormalidad del grupo Alloxan+Insulina en cada conjunto de tres casos sin considerar los casos (2, 5, 7).

En cada uno de los tres grupos de tablas se muestran los resultados obtenidos en los cuatro tests estudiados. Para indicar cada caso y dar una idea de la variabilidad se utilizará el cociente ponderado de varianzas en vez del cociente de varianzas puro, tal como se hizo en el ejemplo de trombosis, ya que los tamaños muestrales son diferentes ($n_1 = 20$, $n_2 = 19$). Como varianzas se emplearán las de la tabla B bajo lognormalidad. Además, en los 6 casos correspondientes a la hipótesis alternativa se incluye la diferencia de medias bajo lognormalidad.

Tablas 5 hasta 8. Nivel de significación y grado de reducción de la varianza en el **test geodésico**, **test de Welch**, **test de Cochran y Cox** y **t de Student**, respectivamente, en relación al estudio sobre la diabetes con $n_1 = 20$ y $n_2 = 19$ bajo distintas combinaciones de parámetros con valores próximos a los muestrales correspondientes a la hipótesis nula (casos 7, 2 y 5, respectivamente).

† =significativamente diferente del nivel de significación nominal.

Tabla 5: $n = 5000$, $n_1 = 20$, $n_2 = 19$, $B = 1000$

$n_1\sigma_1^2/n_2\sigma_2^2$	test geodésico					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
0.20	0.0938†	0.008082	0.09346†	0.007802	6.80	3.46
0.87	0.0436†	0.005661	0.04303†	0.004437	38.55	21.61
2.24	0.0520	0.006155	0.05145	0.005157	29.80	16.21

Tabla 6: $n = 5000$, $n_1 = 20$, $n_2 = 19$

$n_1\sigma_1^2/n_2\sigma_2^2$	test de Welch					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
0.20	0.0938†	0.008082	0.09344†	0.007774	7.47	3.81
0.87	0.0434†	0.005648	0.04281†	0.004318	41.56	23.56
2.24	0.0522	0.006166	0.05165	0.005171	29.67	16.13

Tabla 7: $n = 5000$, $n_1 = 20$, $n_2 = 19$

$n_1\sigma_1^2/n_2\sigma_2^2$	test de Cochran y Cox					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
0.20	0.0880†	0.007853	0.08765†	0.007554	7.47	3.80
0.87	0.0386†	0.005340	0.03804†	0.004100	41.06	23.23
2.24	0.0470	0.005867	0.04647	0.004890	30.53	16.65

Tabla 8: $n = 5000$, $n_1 = 20$, $n_2 = 19$

$n_1\sigma_1^2/n_2\sigma_2^2$	test t de Student					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
0.20	0.0940†	0.008090	0.09363†	0.007761	7.96	4.06
0.87	0.0444	0.005710	0.04379†	0.004297	43.37	24.74
2.24	0.0530	0.006210	0.05245	0.005227	29.15	15.83

Tablas 9 hasta 12. Potencia y grado de reducción de la varianza en el test geodésico, test de Welch, test de Cochran y Cox y t de Student, respectivamente, en relación al estudio sobre la diabetes con $n_1 = 20$ y $n_2 = 19$ bajo distintas combinaciones de parámetros con valores próximos a los muestrales correspondientes a la hipótesis alternativa (casos 8, 1 y 4, respectivamente)

Tabla 9: $n = 5000$, $n_1 = 20$, $n_2 = 19$, $B = 1000$

$\left(\Delta\mu, \frac{n_1\sigma_1^2}{n_2\sigma_2^2}\right)$	test geodésico					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
(87.45, 0.09)	0.0424	0.005586	0.04171	0.005150	14.98	7.80
(-71.00, 1.96)	0.3912	0.013528	0.39334	0.009841	47.08	27.26
(-86.04, 5.02)	0.6198	0.013457	0.62552	0.009788	47.09	27.26

Tabla 10: $n = 5000$, $n_1 = 20$, $n_2 = 19$

$\left(\Delta\mu, \frac{n_1\sigma_1^2}{n_2\sigma_2^2}\right)$	test de Welch					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
(87.45, 0.09)	0.0412	0.005510	0.04053	0.005093	14.54	7.56
(-71.00, 1.96)	0.3948	0.013550	0.39695	0.009824	47.43	27.50
(-86.04, 5.02)	0.6222	0.013440	0.62791	0.009780	47.05	27.23

Tabla 11: $n = 5000$, $n_1 = 20$, $n_2 = 19$

$\left(\Delta\mu, \frac{n_1\sigma_1^2}{n_2\sigma_2^2}\right)$	test de Cochran y Cox					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
(87.45, 0.09)	0.0370	0.005233	0.03637	0.004844	14.30	7.42
(-71.00, 1.96)	0.3742	0.013415	0.37633	0.009723	47.46	27.52
(-86.04, 5.02)	0.6052	0.013550	0.61104	0.009738	48.35	28.13

Tabla 12: $n = 5000$, $n_1 = 20$, $n_2 = 19$

$\left(\Delta\mu, \frac{n_1\sigma_1^2}{n_2\sigma_2^2}\right)$	test t de Student					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
(87.45, 0.09)	0.0532	0.006221	0.05236	0.005651	17.49	9.16
(-71.00, 1.96)	0.3922	0.013534	0.39435	0.009800	47.57	27.59
(-86.04, 5.02)	0.6206	0.013451	0.62634	0.009761	47.34	27.43

Tablas 13 hasta 16. Potencia y grado de reducción de la varianza en el test geodésico, test de Welch, test de Cochran y Cox y t de Student, respectivamente, en relación al estudio sobre la diabetes con $n_1 = 20$ y $n_2 = 19$ bajo distintas combinaciones de parámetros con valores próximos a los muestrales correspondientes a la hipótesis alternativa (casos 9, 3 y 6, respectivamente).

Tabla 13: $n = 5000$, $n_1 = 20$, $n_2 = 19$, $B = 1000$

$\left(\Delta\mu, \frac{n_1\sigma_1^2}{n_2\sigma_2^2}\right)$	test geodésico					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
(-63.30, 0.44)	0.3596	0.013303	0.35889	0.011928	19.60	10.33
(-109.28, 4.41)	0.8392	0.010183	0.84011	0.008529	29.85	16.24
(-119.31, 11.29)	0.9630	0.005233	0.96399	0.004777	16.67	8.71

Tabla 14: $n = 5000$, $n_1 = 20$, $n_2 = 19$

$\left(\Delta\mu, \frac{n_1\sigma_1^2}{n_2\sigma_2^2}\right)$	test de Welch					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
(-63.30, 0.44)	0.3632	0.013332	0.36249	0.011958	19.54	10.30
(-109.28, 4.41)	0.8394	0.010178	0.84031	0.008537	29.64	16.12
(-119.31, 11.29)	0.9618	0.005314	0.96279	0.004865	16.19	8.45

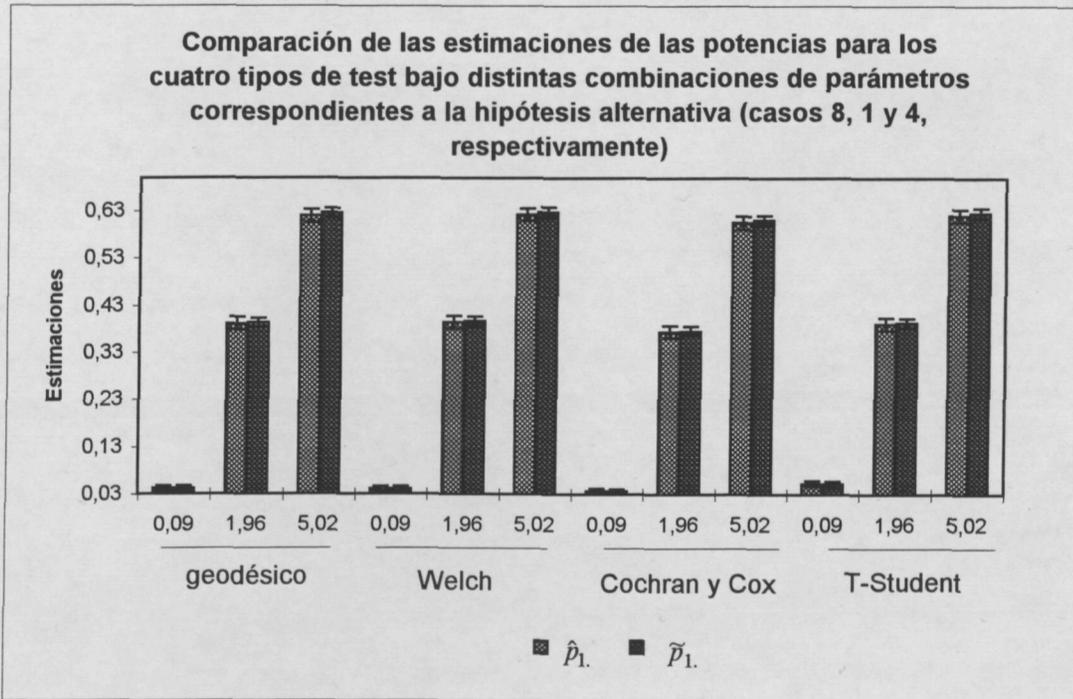
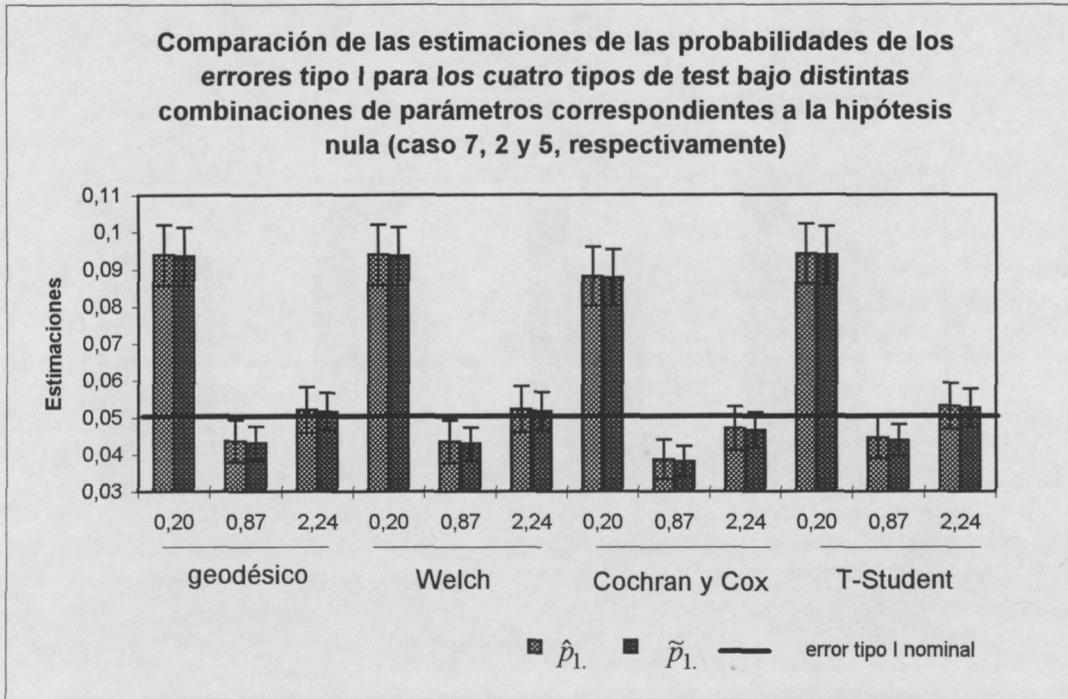
Tabla 15: $n = 5000$, $n_1 = 20$, $n_2 = 19$

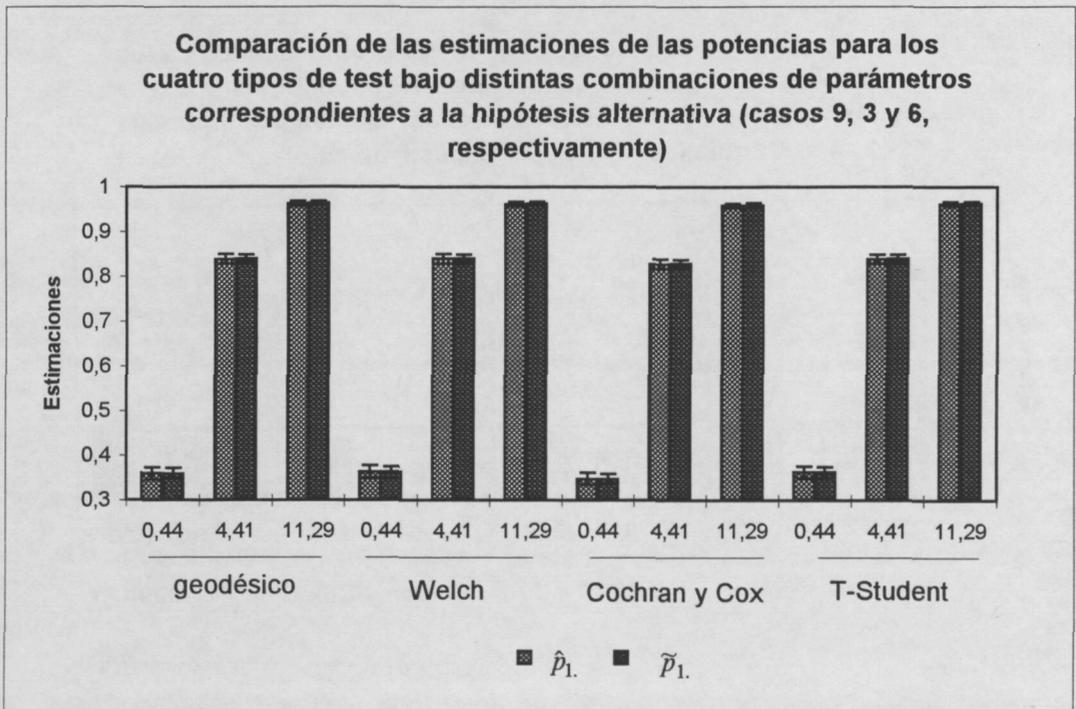
$\left(\Delta\mu, \frac{n_1\sigma_1^2}{n_2\sigma_2^2}\right)$	test de Cochran y Cox					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
(-63.30, 0.44)	0.3476	0.013201	0.34688	0.011767	20.55	10.86
(-109.28, 4.41)	0.8274	0.010476	0.82835	0.008735	30.47	16.62
(-119.31, 11.29)	0.9580	0.005561	0.95910	0.005036	17.99	9.44

Tabla 16: $n = 5000$, $n_1 = 20$, $n_2 = 19$

$\left(\Delta\mu, \frac{n_1\sigma_1^2}{n_2\sigma_2^2}\right)$	test t de Student					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
(-63.30, 0.44)	0.3630	0.013330	0.36228	0.011948	19.66	10.37
(-109.28, 4.41)	0.8386	0.010198	0.83951	0.008529	30.05	16.37
(-119.31, 11.29)	0.9620	0.005300	0.96300	0.004842	16.54	8.65

Además, se muestran tres gráficos asociados a las tablas anteriores con una presentación semejante a la utilizada en el ejemplo de trombosis.

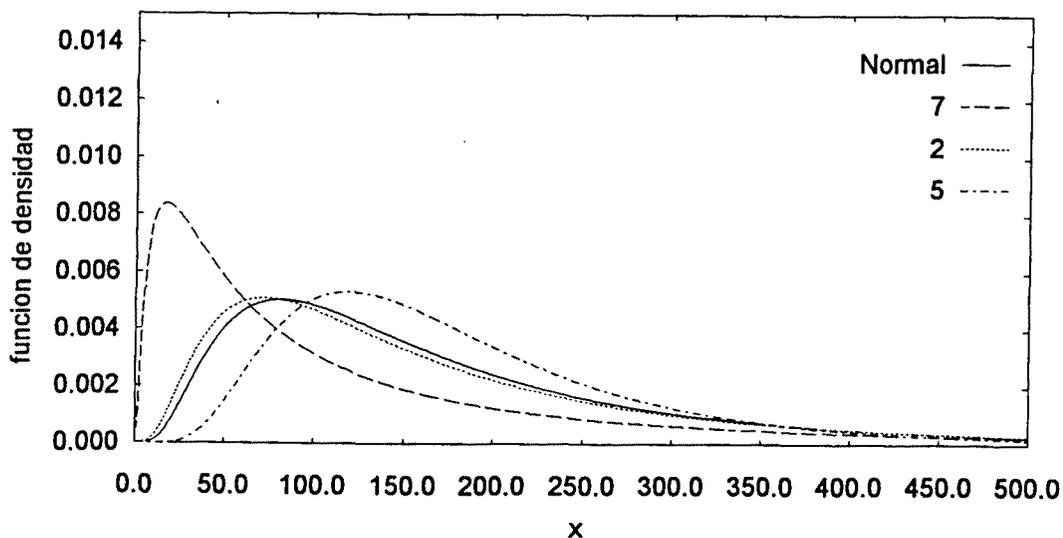




6.2.2.3 Conclusiones

Si primero se analiza los resultados correspondientes a la hipótesis nula, casos 7, 2 y 5 (ordenados por el cociente ponderado de varianzas) se observa que el test geodésico, el de Welch y el test t de Student (calculado sin importar la desigualdad de las varianzas) son muy similares y es el test de Cochran y Cox el que muestra unas estimaciones más bajas, como era de esperar debido a su carácter conservador. Por otro lado, se observa una tendencia a aumentar la estimación de la verdadera probabilidad de error de tipo I al aumentar la diferencia entre varianzas. Así, en el caso más extremo, 0.20, da una estimación alrededor de 0.09. Este comportamiento es explicable por la pareja de distribuciones lognormales que se comparan en cada situación experimental. Para reflejar este hecho, se presenta un gráfico que muestra las distribuciones lognormales correspondientes a la hipótesis nula generadas en cada caso. Tal como se planteó la simulación, en cada pareja de distribuciones lognormales siempre el grupo Normal tiene la distribución fija ($\mu = 186$, $\sigma^2 = 25265.3$) y es el grupo Alloxan+Insulina el que va variando los parámetros de la distribución lognormal según aparece en la tabla B. Así, en cada gráfico tendremos cuatro distribuciones lognormales: la que corresponde al grupo Normal la denominaremos Normal, mientras que las otras tres restantes son para el grupo Alloxan+Insulina en los casos 7, 2 y 5.

Dist. Lognormales correspondientes a la hip. nula



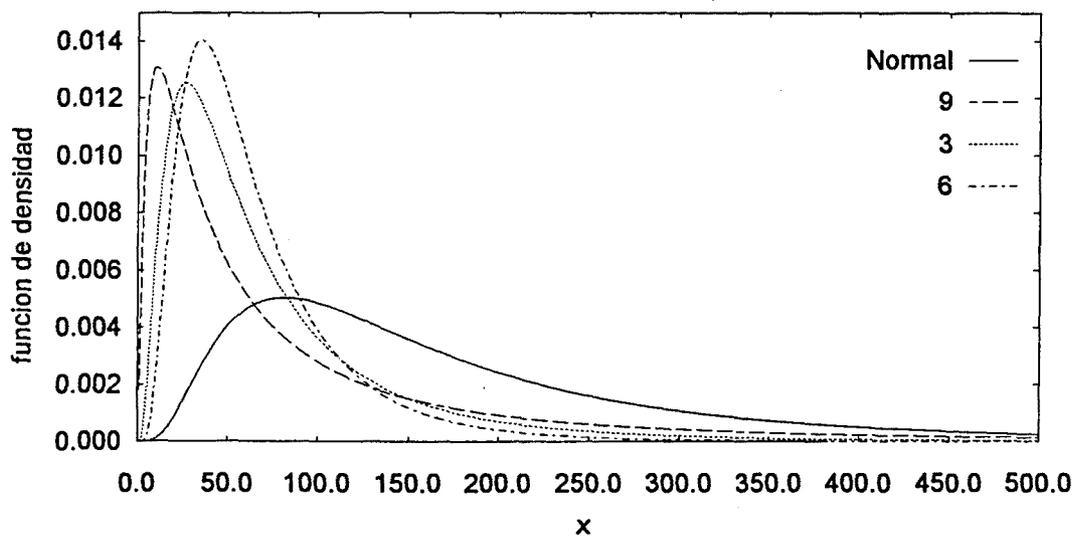
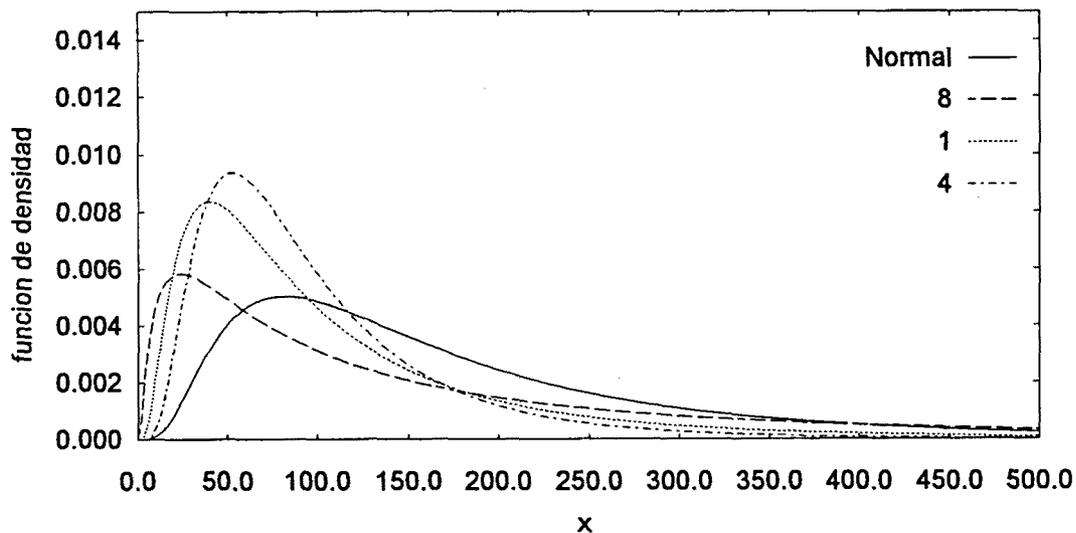
Como puede verse en el gráfico, hay un mayor distanciamiento de la distribución que representa el grupo Normal con la distribución que representa el grupo Alloxan+Insulina cuanto más grande es la diferencia entre varianzas (ver el caso

7). Por tanto, ésto repercute en el valor estimado de sus respectivas potencias y en el porcentaje de reducción de la varianza. Así, cuanto más próximo esté el valor del cociente de varianzas ponderado de uno se tendrá una estimación de la verdadera probabilidad de error de tipo I más baja, y según se observa en los resultados (caso 2) con un valor menor que el nivel de significación nominal de 0.05 y si se aplica la técnica de reducción de la varianza se obtendrá unos valores del 40% de reducción. Sin embargo, cuando nos alejamos de uno, la estimación de la verdadera probabilidad de error de tipo I se va incrementando y la técnica de reducción de la varianza cada vez da peores resultados. Así, en el caso 5 con un cociente ponderado de varianzas de 2.24 incluye el nivel de significación nominal de 0.05 pero en el caso 7 con un cociente de varianzas ponderado de 0.20 (proporción de 1 a 5) se tiene un valor de 0.09.

Si las situaciones experimentales correspondientes a la hipótesis alternativa se ordenan con respecto de las diferencias de medias, $\Delta\mu$, sin considerar el caso 8, se advierte un incremento de la potencia al aumentar la diferencia de medias. Cuando la diferencia está entre -63 y -86 la potencia oscila entre 0.35 y 0.62 y si está entre -109 y -119 la potencia se mueve entre 0.83 y 0.96. El caso 8 nos muestra el efecto distorsionador que aparece al tener unas varianzas muy desiguales. Con una diferencia de medias de un valor de 87.45 se esperaría, al observar los anteriores resultados, una potencia de un orden de 0.5 pero, lo que realmente se observa es un valor de 0.04. Esta gran diferencia se explica al ver en la tabla B que la diferencia entre medias bajo normalidad, -0.1545 , es mucho menor que las que tiene una varianzas más semejantes, por ejemplo el caso 4 ($(\Delta\mu, n_1\sigma_1^2/n_2\sigma_2^2) = (-86.04, 5.02)$), que tiene una diferencia de medias bajo normalidad de 0.5593. A continuación se muestra dos gráficos que representan las distribuciones lognormales generadas para los casos correspondientes a la

hipótesis alternativa.

Dist. Lognormales correspondientes a la hip. alternativa



En las estimaciones de las potencias para los diferentes test sucede lo mismo que en las estimaciones de las probabilidades de error de tipo I; sólo es el test de Cochran y Cox el que tiene unos valores inferiores con respecto a los otros tres test con valores semejantes. El tanto por ciento de reducción de la varianza oscila entre un 14% a un 47% obteniendo unos valores más bajos cuando hay mayor diferencia entre las varianzas.

Por tanto, el resultado del test (aceptar o rechazar la hipótesis nula) puede ser

erróneo debido al efecto distorsionador en las estimaciones de las probabilidades en los dos tipos de error que provoca el cociente ponderado de varianzas, que no es corregido por ninguno de los cuatro test estudiados. Así, una posible solución, sería hacer el logaritmo neperiano de los valores muestrales, con lo cual se conseguiría que si anteriormente provenían de una distribución lognormal ahora se convirtieran en distribuidas normalmente.

6.2.3 Resultados y conclusiones para la estimación de la verdadera probabilidad de error de tipo I y II cuando se ha realizado el logaritmo neperiano de los datos originales

Los resultados muestrales obtenidos son:

	n	\bar{x}	\hat{s}^2
Normal	20	4.857	0.93375
Diabetes Alloxan+Insulina	19	4.392	0.70529

Al hacer el logaritmo neperiano de los datos originales conseguimos reducir la variabilidad y como se puede observar parece factible la no existencia de diferencia significativa entre las varianzas ni entre las medias.

Efectivamente, las hipótesis nulas en el test de Lilliefors para la normalidad y en el test de igualdad de varianzas son aceptadas.

Para poder comparar de la manera más semejante posible con los resultados obtenidos con los datos sin transformar se hará el mismo desarrollo que anteriormente, así pues, al evaluar el estadístico T' (4.7) se obtiene el valor de 1.61, que siempre es menor que los diferentes puntos críticos obtenidos con el test de Cochran y Cox, Welch y Welch-Aspin y por tanto se acepta la hipótesis nula. A la misma conclusión se llega con el test geodésico y con el test t de Student.

Para realizar el estudio de simulación para la estimación de la verdadera probabilidad de error de tipo I y II se introducen como parámetros:

$$n_1 = 20 \quad n_2 = 19 \quad \alpha = 0.05 \quad n = 5000 \quad B = 1000$$

Tabla C

Caso	Bajo Normalidad			
	μ_1	σ_1^2	μ_2	σ_2^2
1	4.8575	0.934	4.3923	0.70530
2	4.8575	0.934	4.8575	0.70530
3	4.8575	0.934	3.9875	0.70530
4	4.8575	0.934	4.3923	0.42506
5	4.8575	0.934	4.8575	0.42506
6	4.8575	0.934	3.9875	0.42506
7	4.8575	0.934	4.3923	1.62812
8	4.8575	0.934	4.8575	1.62812
9	4.8575	0.934	3.9875	1.62812

Se usan como valores de media y varianza (tabla C) los que aparecen en la tabla A bajo normalidad y no los de la tabla B ya que mantienen la idea de fijar μ_1 y σ_1^2 con los valores muestrales e ir modificando sólo μ_2 y σ_2^2 . La única diferencia con la tabla A es la sustitución de los valores de μ_2 más próximos a μ_1 (casos 2, 5 y 8) por su valor. Adviértase que los casos 7 y 8 se utilizarán para estimar las probabilidades de error de tipo II y I, respectivamente, mientras que bajo lognormalidad se hizo al revés. Se recuerda que ésto fue debido a que en el caso 7 de la tabla B la diferencia de medias bajo lognormalidad es menor que en el caso 8.

A diferencia del estudio anterior los valores generados son normales, no lognormales.

A continuación se muestra las tablas de los resultados obtenidos para la estimación de la probabilidad de error de tipo I y II y sus respectivas gráficas.

Tablas 17 hasta 20. Nivel de significación y grado de reducción de la varianza en el test geodésico, test de Welch, test de Cochran y Cox y t de Student, respectivamente, en relación al estudio sobre la diabetes con $n_1 = 20$ y $n_2 = 19$ bajo distintas combinaciones de parámetros con valores próximos a los muestrales correspondientes a la hipótesis nula (casos 8, 2 y 5, respectivamente).

† =significativamente diferente del nivel de significación nominal.

Tabla 17: $n = 5000$, $n_1 = 20$, $n_2 = 19$, $B = 1000$

$n_1\sigma_1^2/n_2\sigma_2^2$	test geodésico					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
0.60	0.0514	0.006121	0.05055	0.003240	71.99	47.08
1.39	0.0508	0.006087	0.04989	0.002447	83.83	59.80
2.31	0.0498	0.006030	0.04900	0.003510	66.12	41.79

Tabla 18: $n = 5000$, $n_1 = 20$, $n_2 = 19$

$n_1\sigma_1^2/n_2\sigma_2^2$	test de Welch					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
0.60	0.0520	0.006155	0.05112	0.003016	75.99	51.00
1.39	0.0506	0.006076	0.04968	0.002297	85.71	62.20
2.31	0.0500	0.006042	0.04919	0.003461	67.19	42.72

Tabla 19: $n = 5000$, $n_1 = 20$, $n_2 = 19$

$n_1\sigma_1^2/n_2\sigma_2^2$	test de Cochran y Cox					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
0.60	0.0464	0.005831	0.04558†	0.002957	74.28	49.29
1.39	0.0440†	0.005685	0.04315†	0.002260	84.19	60.24
2.31	0.0430†	0.005623	0.04226†	0.003309	65.37	41.15

Tabla 20: $n = 5000$, $n_1 = 20$, $n_2 = 19$

$n_1\sigma_1^2/n_2\sigma_2^2$	test t de Student					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
0.60	0.0534	0.006233	0.05251	0.003012	76.65	51.67
1.39	0.0494	0.006007	0.04849	0.002218	86.36	63.07
2.31	0.049	0.005984	0.04819	0.003387	67.97	43.40

Tablas 21 hasta 24. Potencia y grado de reducción de la varianza en el test geodésico, test de Welch, test de Cochran y Cox y t de Student, respectivamente, en relación al estudio sobre la diabetes con $n_1 = 20$ y $n_2 = 19$ bajo distintas combinaciones de parámetros con valores próximos a los muestrales correspondientes a la hipótesis alternativa con $\Delta\mu = -0.4652$ (casos 7, 1 y 4, respectivamente).

Tabla 21: $n = 5000$, $n_1 = 20$, $n_2 = 19$, $B = 1000$

$n_1\sigma_1^2/n_2\sigma_2^2$	test geodésico					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
0.60	0.2262	0.011598	0.22584	0.005145	80.31	55.53
1.39	0.3272	0.013006	0.32720	0.004197	89.59	67.73
2.31	0.3892	0.013516	0.38592	0.006845	74.35	49.35

Tabla 22: $n = 5000$, $n_1 = 20$, $n_2 = 19$

$n_1\sigma_1^2/n_2\sigma_2^2$	test de Welch					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
0.60	0.2278	0.011626	0.22745	0.005310	79.14	54.32
1.39	0.3312	0.013047	0.33120	0.003643	92.20	72.08
2.31	0.3886	0.013512	0.38529	0.006657	75.73	50.73

Tabla 23: $n = 5000$, $n_1 = 20$, $n_2 = 19$

$n_1\sigma_1^2/n_2\sigma_2^2$	test de Cochran y Cox					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
0.60	0.2110	0.011311	0.21066	0.005199	78.87	54.03
1.39	0.3090	0.012809	0.30900	0.004341	88.52	66.11
2.31	0.3682	0.013370	0.36489	0.006374	77.27	52.33

Tabla 24: $n = 5000$, $n_1 = 20$, $n_2 = 19$

$n_1\sigma_1^2/n_2\sigma_2^2$	test t de Student					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
0.60	0.2326	0.011712	0.23224	0.005227	80.08	55.37
1.39	0.3308	0.013043	0.33080	0.003723	91.85	71.46
2.31	0.3844	0.013485	0.38109	0.006599	76.05	51.07

Tablas 25 hasta 29. Potencia y grado de reducción de la varianza en el test geodésico, test de Welch, test de Cochran y Cox y t de Student, respectivamente, en relación al estudio sobre la diabetes con $n_1 = 20$ y $n_2 = 19$ bajo distintas combinaciones de parámetros con valores próximos a los muestrales correspondientes a la hipótesis alternativa con $\Delta\mu = -0.87$ (casos 9, 3 y 6, respectivamente).

Tabla 25: $n = 5000$, $n_1 = 20$, $n_2 = 19$, $B = 1000$

$n_1\sigma_1^2/n_2\sigma_2^2$	test geodésico					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
0.60	0.6272	0.013404	0.62809	0.006276	78.07	53.18
1.39	0.8254	0.010524	0.82819	0.003958	85.85	62.39
2.31	0.8890	0.008708	0.88735	0.004672	71.22	46.35

Tabla 26: $n = 5000$, $n_1 = 20$, $n_2 = 19$

$n_1\sigma_1^2/n_2\sigma_2^2$	test de Welch					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
0.60	0.6304	0.013381	0.63130	0.005983	80.01	55.29
1.39	0.8278	0.010466	0.83063	0.003401	89.44	67.50
2.31	0.8900	0.008674	0.88834	0.004535	72.66	47.71

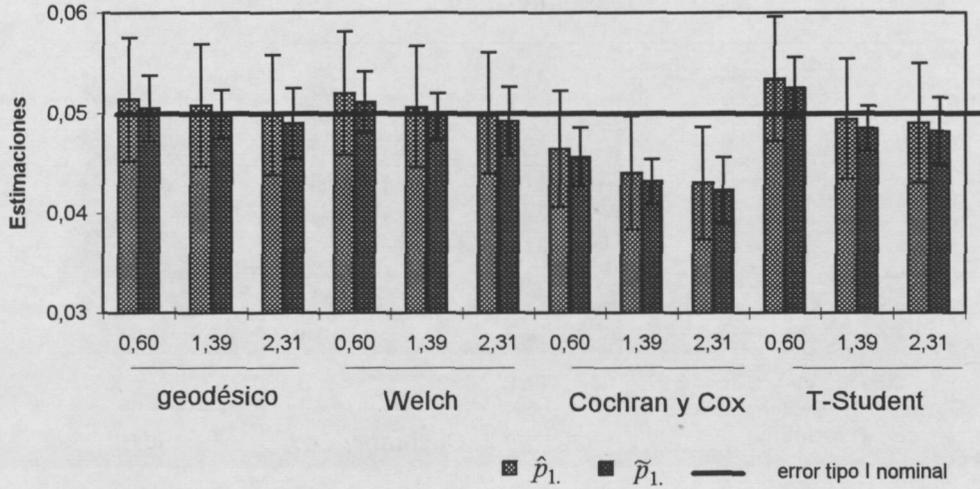
Tabla 27: $n = 5000$, $n_1 = 20$, $n_2 = 19$

$n_1\sigma_1^2/n_2\sigma_2^2$	test de Cochran y Cox					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
0.60	0.6058	0.013547	0.60670	0.006400	77.68	52.76
1.39	0.8100	0.010875	0.81291	0.003864	87.38	64.47
2.31	0.8810	0.008976	0.87928	0.004638	73.30	48.33

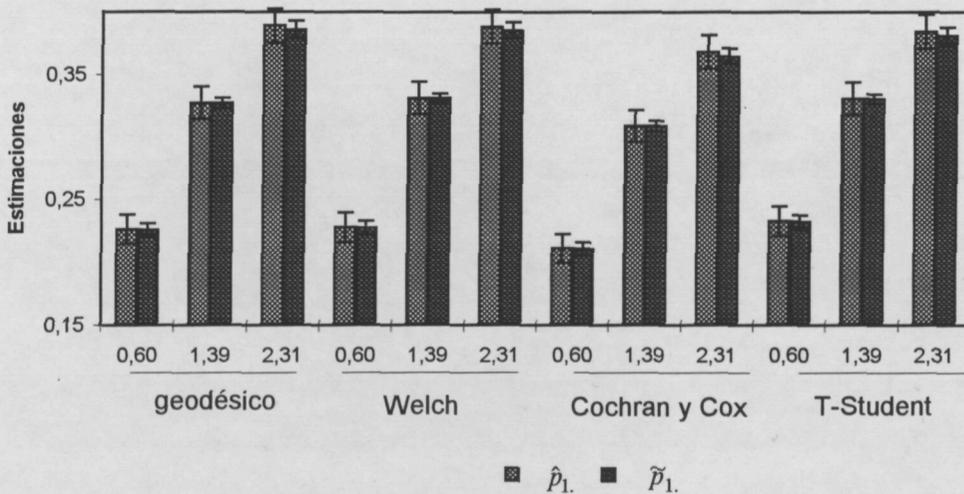
Tabla 28: $n = 5000$, $n_1 = 20$, $n_2 = 19$

$n_1\sigma_1^2/n_2\sigma_2^2$	test t de Student					
	\hat{p}_1	$\pm 1.959964\hat{\sigma}$	\tilde{p}_1	$\pm 1.959964\hat{\sigma}_{GS}$	$\widehat{\%VR}$	$\widehat{\%MER}$
0.60	0.6382	0.013320	0.63910	0.005748	81.38	56.85
1.39	0.8260	0.010509	0.82884	0.003473	89.08	66.95
2.31	0.8892	0.008701	0.88754	0.004573	72.38	47.45

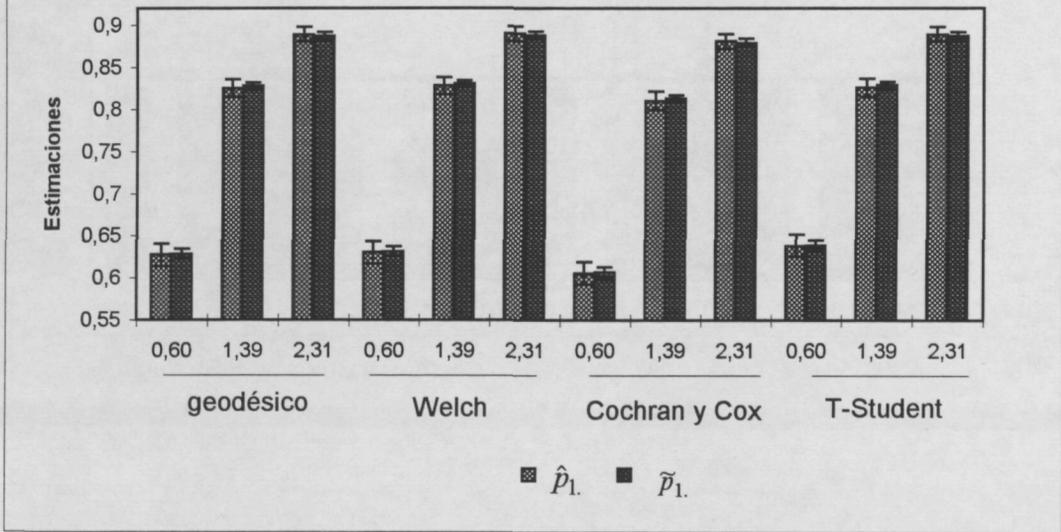
Comparación de las estimaciones de las probabilidades de los errores tipo I para los cuatro tipos de test bajo distintas combinaciones de parámetros correspondientes a la hipótesis nula (casos 8, 2 y 5, respectivamente).



Comparación de las estimaciones de las potencias para los cuatro tipos de test bajo distintas combinaciones de parámetros correspondientes a la hipótesis alternativa (casos 7, 1 y 3, respectivamente)



Comparación de las estimaciones de las potencias para los cuatro tipos de test bajo distintas combinaciones de parámetros correspondientes a la hipótesis alternativa (casos 9, 3 y 6, respectivamente)



En general, las estimaciones de las probabilidades de error de tipo I son muy similares en el test geodésico y en el test de Welch. En el test t de Student para el cociente de varianzas ponderado menor que uno (0.60) se obtienen estimaciones superiores que en los anteriores tests pero, para valores mayores que uno (1.39 y 2.31) se obtienen estimaciones algo inferiores. Para el test de Cochran y Cox se obtienen siempre estimaciones inferiores que las del test geodésico o el test de Welch.

En la estimación de la probabilidad de error de tipo I parece no ser afectada por los diferentes cocientes ponderados de varianzas cuando se utiliza el test geodésico o el test de Welch ya que la estimación puntual no varía mucho y el intervalo de confianza del 95% para la probabilidad de error de tipo I incluye el nivel de significación nominal (0.05) tanto bajo la estimación sin aplicar la técnica de reducción de la varianza, \hat{p}_1 , como cuando se aplica, \tilde{p}_1 . El test t de Student también incluye el nivel de significación nominal mientras que el test de Cochran y Cox no lo hace cuando se utiliza como estimación \tilde{p}_1 . (carácter conservador).

En los casos experimentales correspondientes a la hipótesis alternativa se repara en una tendencia a aumentar el valor de la estimación de la potencia cuanto más grande es el cociente ponderado de las varianzas siendo los resultados más análogos entre ellos los del test geodésico y el test de Welch. Así, para un $\Delta\mu = -0.4652$ en los casos 7, 1 y 4 se pasa de 0.23 a 0.39 y para un $\Delta\mu = -0.87$, de casi el doble que el anterior, en los casos 9, 3 y 6 se pasa de 0.63 a 0.89.

El test de Cochran y Cox tiene unos valores habitualmente por debajo de los anteriores tests. El test t de Student es el que muestra menos variabilidad con respecto a la variabilidad del cociente ponderado de varianzas.

Estimar la probabilidad de error de tipo I o II aplicando la reducción de la varianza, \tilde{p}_1 , produce unos intervalos de confianza mucho más estrechos de un orden entre un 40% a un 60% (ver $\% \widehat{MER}$) consiguiéndose los valores mayores cuando el cociente ponderado de varianzas es más cercano a uno.

Si ahora nos fijamos en el caso que tenga como parámetros los valores más cercanos a los valores muestrales, caso 1, vemos que la potencia es alrededor de 0.33 con un cociente ponderado de varianzas de 1.39 mientras que este mismo número de caso bajo la suposición de lognormalidad da una potencia estimada de 0.39 con un cociente ponderado de varianzas de 1.96. Así que parece que existe un incremento en el error de tipo II al pasar de la distribución lognormal a la

transformación normal. Sin embargo, esta diferencia puede ser originada porque realmente en el caso 1 bajo lognormalidad ($\mu_2 = 115.00; \sigma_1^2 = 13548.1$) no sea exactamente los valores muestrales ($\mu_2 = 112.72; \sigma_1^2 = 11200.2$)

Como conclusión final, se puede decir que con los datos muestrales obtenidos no se producen ningún cambio en la aceptación de la hipótesis nula dada a partir del test t de Student al utilizar otros test más adecuados o hacer la transformación de los datos para ajustarlo a una distribución normal. Además, parece ser que el posible cambio en la estimación de los error de tipo I y II está afectado por el cociente de varianzas y no tanto por la transformación de los datos lognormales a normales. En cualquier caso, se recomendaría al usuario que verificara siempre la igualdad de varianzas antes de realizar el contraste de comparación de medias. En el caso de que se rechazara y, además, los datos no se ajusten a una normal sino a una distribución lognormal es preferible transformar los datos originales para convertirlos en normales que realizar un test como el de Welch para resolver el contraste.

Capítulo 7

CONCLUSIONES

7.1 Conclusiones y resumen de resultados

Se presenta a continuación las conclusiones y un resumen de los principales resultados aportados en la presente memoria.

- El principal resultado es la presentación de una técnica de optimización en estudios de Monte Carlo en Estadística. Se basa en obtener un estimador de la esperanza, \tilde{p}_1 , de una variable dicotómica, Y , que tiene una varianza menor que el estimador habitual, la frecuencia relativa (\hat{p}_1). Este estimador optimizado se basa en el conocimiento de otra variable dicotómica (de control), C , correlacionada con Y y de esperanza conocida, $E(C)$. La aplicación de esta técnica es sencilla de implementar. En simulación de Monte Carlo en estadística es relativamente frecuente disponer de tales variables de control. Así, por ejemplo, en estudios de simulación de la potencia de un nuevo test no paramétrico se puede utilizar en ocasiones un test paramétrico comparable, de potencia conocida.
- Se mostró que el nuevo estimador, \tilde{p}_1 , se puede obtener por diferentes métodos: como estimador máximo verosímil, mediante variables de control ajustadas por regresión, por esperanza condicionada y como el estimador basado en la curva de regresión de la media.
- Se demostró que este estimador es insesgado, se obtuvo la expresión de su varianza y que el modelo probabilístico asociado al vector aleatorio (Y, C) es una familia exponencial curvada y que, por tanto, el estimador propuesto no alcanza la cota de Cramer-Rao aunque es asintóticamente eficiente y asintóticamente normal.
- Se obtuvieron diferentes estimadores de la varianza de \tilde{p}_1 y se estudiaron sus propiedades:
 - $\hat{\sigma}_R^2$: Fue propuesto por Rothery basándose en argumentos estándar de estimación máximo verosímil. Se comprobó que se puede obtener, alternativamente, por la técnica de variables de control o por el método

- delta de primer orden. Se verificó que este estimador de la varianza, $\hat{\sigma}_R^2$, tiene un sesgo negativo.
- $\hat{\sigma}_U^2$: Estimador insesgado sin una forma cerrada ya que está basado en esperanzas de inversas de variable binomiales truncadas. Se puede evaluar, de todas formas, sin dificultad.
 - $\hat{\sigma}_{GS}^2$: Estimador casi insesgado y de cálculo más sencillo que el anterior ya que sí se puede indicar explícitamente, basado en un aproximación propuesta por Grab y Savage.
 - $\hat{\sigma}_B^2$: Estimador basado en la técnica bootstrap.
- Para valorar la reducción de la varianza se propusieron estimadores para la reducción de la varianza y del margen de error relacionados con los anteriores estimadores de la varianza.
 - Para validar los resultados teóricos anteriores y para obtener algunas propiedades se realizaron dos estudios de simulación.
 - Uno de ellos se basa en obtener una estimación de la potencia de un test propuesto por Burbea y Oller (1982) que compara simultáneamente las medias μ_1, μ_2 y las varianzas σ_1^2, σ_2^2 de dos variables aleatorias normales independientes. Se fundamenta en el criterio de distancia geodésica entre distribuciones. El test que se utilizó como control, con potencia conocida, fue el test t de Student. En la preparación de este estudio se comprobó que varios algoritmos de cálculo comunes en estadística y en simulación deben de organizarse de forma distinta de lo habitual para aprovechar las capacidades de los procesadores vectoriales.
 - El otro estudio de simulación pretende validar la técnica presentada sin que corresponda a ningún estudio concreto. Se generan tablas de contingencia 2×2 que representan el resultado final de una “verdadera” simulación de un tamaño n de pares de variables dicotómicas (Y, C) para unos valores de probabilidad de cada celda prefijados.
 - En ambos casos, los mejores resultados se obtienen con los estimadores de la varianza de \tilde{p}_1 : $\hat{\sigma}_B^2, \hat{\sigma}_{GS}^2$ y $\hat{\sigma}_U^2$ mientras que con $\hat{\sigma}_R^2$ se obtiene valores menores que la “verdadera” varianza. Los cuatro diferentes estimadores de la reducción de la varianza dan valores sesgados. Los que utilizan $\hat{\sigma}_B^2, \hat{\sigma}_{GS}^2$ y $\hat{\sigma}_U^2$ dan, de forma sistemática, estimaciones inferiores que la “verdadera”

cantidad de reducción de la varianza, lo cual es preferible a estimaciones superiores a esta cantidad (excesivamente optimista), como pasa en algunas ocasiones con $\hat{\sigma}_R^2$. En cuanto al margen de error está muy próximo al 95% cuando se basan en $\hat{\sigma}_B^2$, $\hat{\sigma}_{GS}^2$ y $\hat{\sigma}_U^2$ mientras que con $\hat{\sigma}_R^2$ también es un valor cercano a este porcentaje pero sistemáticamente está por debajo. Por tanto, por su simplicidad y por la falta de diferencias sustanciales con el estimador insesgado, $\hat{\sigma}_U^2$, en las condiciones de simulación estadística (gran número, n , de réplicas de simulación, y sin tener que estimar valores extremos de p_{ij}) se propone el estimador $\hat{\sigma}_{GS}^2$ como el más apropiado de entre los cuatro estimadores de la varianza de \tilde{p}_1 .

- El siguiente paso fue presentar un nuevo test, geodésico, basado en la distancia de Rao, para resolver el problema de Behrens-Fisher. Primero, se puso de manifiesto la dificultad de resolver el problema de Behrens-Fisher ya que el submodelo paramétrico bajo la hipótesis nula de igualdad de medias es una familia exponencial curvada y, a continuación, se presentaron algunas de las soluciones prácticas actualmente empleadas. El nuevo test no tiene una forma cerrada (4.17), quedando en función de un sistema de ecuaciones que debe ser resuelto numéricamente. Ésto hace que sea muy difícil de resolver el problema distribucional de calcular el punto crítico c_α que debe de satisfacer $P\{D^2 \geq c_\alpha | H_0\} = \alpha$. Como solución se optó por aproximar su valor mediante bootstrap paramétrico.
- A partir de la técnica de optimización presentada anteriormente se estudió la potencia de este nuevo test por simulación teniendo como variable de control, de nuevo, al test t de Student. Para valorar la potencia de este test, se utilizó como referencia la potencia obtenida en el mismo proceso por una de las más habituales soluciones al problema de Behrens-Fisher, el test de Cochran y Cox. Los resultados de simulación señalan que el test geodésico es mejor que el test de Cochran y Cox, especialmente para tamaños muestrales bajos. La extensión del test geodésico es siempre igual o muy próxima al nivel de significación nominal y es más potente que el test de Cochran y Cox. Además, este último test tiene un carácter conservador que sólo se detecta claramente, en tamaños muestrales pequeños, si se utiliza esta nueva técnica de optimización de las simulaciones. Por otro lado, los resultados de reducción de la varianza fueron de un valor mínimo del 75% para varianzas semejantes entre las poblaciones y de un 42% para varianzas muy diferentes de hasta cuatro veces una más grande que la otra. Además,

la utilización de esta técnica supone un mínimo esfuerzo en la complejidad del diseño de la simulación.

- Una cualidad importante en un test, excluyendo su potencia, es su robustez. Se presentaron dos casos estudiados por simulación, uno en el cual las observaciones seguían una distribución lognormal y en otro caso seguían una distribución gamma. Situaciones que pueden ser factibles en casos reales. En el primero se hizo un estudio bastante completo en el que se concluyó que el test geodésico es preferible para tamaños muestrales pequeños, las diferencias se van diluyendo con el test de Cochran y Cox para tamaños muestrales mayores. En cualquier caso, se observa que es mejor utilizar la estimación de la potencia a partir de la técnica de optimización que utilizar la frecuencia relativa. En el segundo caso, más que plantear un estudio de la robustez bajo la distribución gamma lo que se presenta es un ejemplo bajo el cual la técnica de optimización no funciona adecuadamente y cómo puede ser solucionado este problema. Los pocos resultados de simulación obtenidos de la robustez son congruentes con los del caso anterior.

Se concluye que ambos tests son bastante robustos, al menos para los dos casos estudiados y que solamente un buen conocimiento de las técnicas de generación de variables aleatorias, y en general de la estadística computacional, permite diagnosticar y solucionar problemas de diseño de las simulaciones como ocurre con el segundo caso.

- Por último, se exponen dos casos reales, dentro del entorno médico-biológico, donde surge el problema de Behrens-Fisher. En ambos estudios, se ignoraron probables diferencias entre las varianzas, y se aplicó el test t de Student. Por consiguiente, se incurrió en un error que repercute en que los p-valores no son los que corresponden pudiendo llegar a afectar a la conclusión final del test. Además, en el estudio sobre la diabetes la prueba de Lilliefors para contrastar la normalidad fue rechazada. Así, primero se obtuvieron cuales son los resultados para diferentes tests: de Welch, Welch-Aspin, Cochran y Cox, t de Student y test geodésico (se hizo un programa para este fin) y, a continuación, se estimó las verdaderas probabilidades de error de los test anteriores, excepto el de Welch-Aspin, para valores de los parámetros poblacionales alrededor de los resultados muestrales empleando la técnica de optimización.

Se obtuvo que tanto el test geodésico como el test de Welch son igualmente correctos y robustos frente a las desviaciones de la normalidad estudiadas. El test de Cochran y Cox puede enmascarar diferencias entre las medias cuando el cociente de varianzas es próximo a uno, por su carácter conservador. En ambos casos, las conclusiones finales de los artículos que utilizaron estos datos reales son "correctas", en tanto que coinciden con las que se obtuvieron mediante una técnica estadística más apropiada. En general, se puede concluir que el posible cambio en la estimación de los errores de tipo I y II está afectado por el cociente de varianzas y no tanto por la falta de normalidad de los datos.

Es importante resaltar lo sencillo que es utilizar la nueva técnica de reducción de la varianza para el estudio de simulación de cualquier estadístico aplicable a problemas reales.

APÉNDICES

Apéndice A

Obtención de μ_2 a partir del nomograma

Con el conocimiento de la potencia teórica del test de control, t de Student, se puede obtener el valor de μ_2 gracias al uso del nomograma para la determinación de la potencia del t de Student (Tabla 10 de *Biometrika Tables* (1976)).

En el nomograma se muestran las diferentes curvas de potencias que dependen de:

- Potencia del test: β_c
- Nivel de significación del test: α
- Grados de libertad: ν
- ϕ_{β_c}

Gracias a que se sabe la potencia, el nivel de significación y los grados de libertad del test se obtiene, a partir del nomograma, ϕ_{β_c}

$$\phi_{\beta_c} = \frac{(\mu_2 - \mu_1)}{\sigma_y} \frac{1}{\sqrt{2}}$$

donde y es la variable aleatoria $\bar{x}_2 - \bar{x}_1$ siendo $x_1 \sim N(\mu_1, \sigma)$ independiente de $x_2 \sim N(\mu_2, \sigma)$. Así que, $\sigma_y = \sigma \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$

Se despeja μ_2

$$\mu_2 = \mu_1 + \sqrt{2} \sigma \phi_{\beta_c} \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \quad (\text{A.1})$$

y, por tanto, ya se ha conseguido la expresión que se buscaba.

En nuestros estudios de simulación de la potencia de un nuevo test para el problema de Behrens-Fisher utilizamos (A.1) siendo la varianza común, σ^2 , una media ponderada de las varianzas, σ_1^2 y σ_2^2 (4.25):

$$\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2},$$

con lo cual, μ_2 queda como:

$$\mu_2 = \mu_1 + \sqrt{2}\phi_{\beta_c} \sqrt{\frac{\sigma_1^2}{n_2} + \frac{\sigma_2^2}{n_1}} \quad (\text{A.2})$$

En algunas simulaciones se han practicado las siguientes restricciones:

$$\begin{aligned} n_1 &= n_2 = n \\ \mu_1 &= 0 \end{aligned}$$

y la varianza común, σ^2 , calculada como se ha indicado anteriormente, siempre ha sido igual a 1. Así pues, el valor de μ_2 se consigue como

$$\mu_2 = 2\phi_{\beta_c} \frac{1}{\sqrt{n}} \quad (\text{A.3})$$

Apéndice B

Distancia de Rao

En estadística es habitual expresar un conjunto de poblaciones estadísticas a través de un modelo paramétrico

$$M = \{p(x; \theta) \mid \theta \in \Theta \text{ abierto de } \mathbb{R}^q\}$$

donde $p(x; \theta)$ son funciones de densidad respecto de alguna medida de referencia y cada población es representada por $p(x; \theta)$ bajo un $\theta \in \Theta$ fijo.

Bajo esta premisa y utilizando conceptos de geometría diferencial, en 1945 Rao en su artículo "*Information and Accuracy Estimation of Statistical Parameters*" aparte de obtener la desigualdad de Cramér-Rao, introduce en tres apartados finales: "The population space", "The distance between two populations" y "Distance in tests of significance and classification", una distancia Riemanniana en M y su uso en inferencia y análisis de datos. Rao procede como sigue.

Nuestras poblaciones vienen caracterizadas por

$$p(x; \theta), \theta = (\theta_1, \theta_2, \dots, \theta_q) \text{ (parámetros poblacionales)}$$

entonces, variando θ obtenemos un conjunto de poblaciones que puede ser representado en un espacio de q dimensiones (espacio de las poblaciones). Dos poblaciones contiguas difieren (salvo órdenes superiores) en:

$$dp = \sum_{i=1}^q \frac{\partial p}{\partial \theta_i} d\theta_i$$

y si consideramos la discrepancia relativa

$$\frac{dp}{p} = \sum_{i=1}^q \frac{1}{p} \frac{\partial p}{\partial \theta_i} d\theta_i = \sum_{i=1}^q \frac{\partial \log p}{\partial \theta_i} d\theta_i$$

y finalmente su varianza, tendremos la forma cuadrática diferencial, definida positiva:

$$ds^2 = V\left(\frac{dp}{p}\right) = \sum_{i,j=1}^q E\left(\frac{\partial \log p}{\partial \theta_i} \frac{\partial \log p}{\partial \theta_j}\right) d\theta_i d\theta_j = \sum_{i,j=1}^q g_{ij} d\theta_i d\theta_j$$

con

$$g_{ij} = E\left(\frac{\partial \log p}{\partial \theta_i} \frac{\partial \log p}{\partial \theta_j}\right) \quad (\text{B.1})$$

Esta forma cuadrática tiene importantes propiedades:

- Es invariante bajo transformaciones biyectivas (y medibles) de los datos.
- Es invariante bajo cambios (suaves) en la parametrización de las poblaciones.

Así g_{ij} constituye las componentes de un tensor covariante simétrico de segundo orden que proporciona una métrica Riemanniana al espacio de poblaciones.

Dadas, entonces, dos poblaciones A y B (no necesariamente próximas) con parámetros poblacionales θ_A y θ_B respectivamente, podemos encontrar la distancia entre ellas por integración de ds a lo largo de la curva geodésica; las ecuaciones de esta geodésica se pueden expresar como función de un parámetro t ,

$$\theta(t) = (\theta_1(t), \dots, \theta_q(t)),$$

donde $\theta(t)$ es la solución del sistema de ecuaciones diferenciales ordinario de segundo orden conocido como las ecuaciones de Euler-Lagrange

$$\frac{d^2\theta_k}{dt^2} + \sum_{i,j=1}^q \Gamma_{ij}^k \frac{d\theta_i}{dt} \frac{d\theta_j}{dt} = 0 \quad k = 1, \dots, q \quad (\text{B.2})$$

con las condiciones de contorno

$$\theta(t_A) = \theta_A \quad \theta(t_B) = \theta_B$$

y la propiedad que cualquier curva geodésica es tal que su vector tangente tiene norma constante, es decir,

$$\sum_{i,j=1}^q g_{ij} \frac{d\theta_i}{dt} \frac{d\theta_j}{dt} = \text{constante.}$$

Esta constante se puede hacer igual a uno si tomamos como parámetro t la longitud del arco.

En (B.2), Γ_{ij}^k son los símbolos de Christoffel de segundo orden asociado al tensor métrico g_{ij} definidos por

$$\Gamma_{ij}^k = \sum_{l=1}^q g^{kl} [ij, l]$$

con

$$[ij, l] = \frac{1}{2} \left(\frac{\partial g_{il}}{\partial \theta_j} + \frac{\partial g_{jl}}{\partial \theta_i} - \frac{\partial g_{ij}}{\partial \theta_l} \right)$$

que son los símbolos de Christoffel de primer orden y (g^{ij}) es la matriz inversa de (g_{ij})

En resumen, la $\theta(t)$ se obtiene resolviendo las ecuaciones de Euler-Lagrange con las condiciones de contorno y teniendo en consideración que el vector tangente tiene norma constante. Además, los valores de g_{ij} son determinables a partir de (B.1). Por lo tanto, se puede encontrar la distancia por integración de ds desde A hasta B a lo largo de la geodésica:

$$\lambda(A, B) = \int_{t_A}^{t_B} \sqrt{\sum_{i,j=1}^q g_{ij} \frac{d\theta_i}{dt} \frac{d\theta_j}{dt}} dt$$

que denominaremos **distancia de Rao** entre funciones de densidad.

Todos estos conceptos y la metodología asociada se han aplicado para el caso de poblaciones normales de media y varianzas desconocidas

$$p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}$$

(m, σ) , $\mu \in \mathfrak{R}$, $\sigma \in \mathfrak{R}^+$, donde tendremos

$$\begin{aligned} g_{11} &= E \left(\left(\frac{\partial}{\partial \mu} \log p(x; \mu, \sigma) \right)^2 \right) = E \left(\frac{(x - \mu)^2}{\sigma^4} \right) = \frac{1}{\sigma^2} \\ g_{21} = g_{12} &= E \left(\frac{\partial}{\partial \mu} \log p(x; \mu, \sigma) \frac{\partial}{\partial \sigma} \log p(x; \mu, \sigma) \right) = \\ &= E \left(\frac{x - \mu}{\sigma^2} \left(-\frac{1}{\sigma} + \frac{1}{\sigma^3} (x - \mu)^2 \right) \right) = 0 \\ g_{22} &= E \left(\frac{1}{\sigma^2} + \frac{1}{\sigma^6} (x - \mu)^4 - \frac{2}{\sigma^4} (x - \mu)^2 \right) = \\ &= \frac{1}{\sigma^2} + \frac{3\sigma^4}{\sigma^6} - 2\frac{\sigma^2}{\sigma^4} = \frac{2}{\sigma^2} \end{aligned}$$

de manera que

$$ds^2 = \frac{(d\mu)^2}{\sigma^2} + \frac{2(d\sigma)^2}{\sigma^2}.$$

Los símbolos de Christoffel de primer orden son:

$$\begin{aligned} [11, 1] &= \frac{1}{2} \left(\frac{\partial g_{11}}{\partial \mu} \right) = 0 \\ [11, 2] &= -\frac{1}{2} \left(\frac{\partial g_{11}}{\partial \sigma} \right) = \frac{1}{\sigma^3} \\ [12, 1] = [21, 1] &= \frac{1}{2} \left(\frac{\partial g_{11}}{\partial \sigma} \right) = -\frac{1}{\sigma^3} \\ [22, 2] &= \frac{1}{2} \left(\frac{\partial g_{22}}{\partial \sigma} \right) = -\frac{2}{\sigma^3} \\ [22, 1] &= -\frac{1}{2} \left(\frac{\partial g_{22}}{\partial \mu} \right) = 0 \\ [21, 2] = [12, 2] &= \frac{1}{2} \left(\frac{\partial g_{22}}{\partial \mu} \right) = 0 \end{aligned}$$

Como

$$g^{11} = \sigma^2 \quad g^{12} = g^{21} = 0 \quad g^{22} = \frac{\sigma^2}{2}$$

los símbolos de Christoffel de segundo orden son:

$$\begin{aligned} \Gamma_{11}^1 &= g^{11}[11, 1] = 0 \\ \Gamma_{11}^2 &= g^{22}[11, 2] = \frac{1}{2\sigma} \\ \Gamma_{12}^1 = \Gamma_{21}^1 &= g^{11}[12, 1] = -\frac{1}{\sigma} \\ \Gamma_{22}^2 &= g^{22}[22, 2] = -\frac{1}{\sigma} \\ \Gamma_{22}^1 &= g^{11}[22, 1] = 0 \\ \Gamma_{12}^2 = \Gamma_{21}^2 &= g^{22}[21, 2] = 0 \end{aligned}$$

Así, las ecuaciones diferenciales ordinarias de segundo orden de las geodésicas quedan como

$$\begin{cases} \frac{d^2\mu}{dt^2} - \frac{2}{\sigma} \frac{d\mu}{dt} \frac{d\sigma}{dt} = 0 \\ \frac{d^2\sigma}{dt^2} + \frac{1}{2\sigma} \left(\frac{d\mu}{dt} \right)^2 - \frac{1}{\sigma} \left(\frac{d\sigma}{dt} \right)^2 = 0 \end{cases}$$

que habrá que resolver para unas condiciones de contorno adecuadas (que pase por los puntos (μ_A, σ_A) y (μ_B, σ_B)) y además si tomamos la longitud de arco como parámetro t tendremos:

$$\frac{1}{\sigma^2} \left(\frac{d\mu}{dt} \right)^2 + \frac{2}{\sigma^2} \left(\frac{d\sigma}{dt} \right)^2 = 1$$

El último paso es calcular la integral entre los dos puntos A y B en la geodésica

$$\lambda(A, B) = \lambda[(\mu_A, \sigma_A), (\mu_B, \sigma_B)] = \int_{t_A}^{t_B} dt$$

que nos conducirá a obtener la distancia entre ellas (Atkinson y Mitchell, 1981; Burbea y Rao, 1982), la llamada **distancia de Rao entre dos distribuciones normales univariantes**:

$$\begin{aligned} \lambda[(\mu_A, \sigma_A), (\mu_B, \sigma_B)] &= 2\sqrt{2} \tanh^{-1} \Delta(A, B) \\ &= \sqrt{2} \left| \log \frac{1 + \Delta(A, B)}{1 - \Delta(A, B)} \right| \end{aligned}$$

con

$$\Delta(A, B) = \left\{ \frac{(\mu_A - \mu_B)^2 + 2(\sigma_A - \sigma_B)^2}{(\mu_A - \mu_B)^2 + 2(\sigma_A + \sigma_B)^2} \right\}^{1/2}$$

Apéndice C

Bootstrap paramétrico

El bootstrap es un método de estimación de la distribución muestral de un estadístico (o de algunas de sus características, como la desviación típica), propuesto por Efron (véase Efron, 1979 y 1982), que considera a la muestra como si fuese la población y que, en su implementación más habitual, utiliza el método de Monte Carlo para obtener nuevas muestras (remuestras) y a partir de ellas hacer la estimación deseada. Por este motivo, es una técnica que necesita para su aplicación de un ordenador.

El bootstrap típico tiene las siguientes fases:

1. Tomaremos muestras al azar con reemplazamiento de la muestra original de igual tamaño a ésta.
2. Calcularemos en cada muestra el estimador (o más en general, el estadístico) de interés, $\hat{\theta}$.
3. Repitiendo este proceso un número B adecuado de veces, obtendremos una muestra de los valores del estimador $\hat{\theta}$. A partir de esta muestra se podrá calcular el valor medio muestral, la varianza muestral, la distribución empírica ... que, supuestamente, se parecerán a las correspondientes características de la distribución de $\hat{\theta}$.

En nuestro caso, el test geodésico, el problema es obtener una adecuada estimación de la probabilidad de que la distancia mínima de Rao al cuadrado ponderada por los tamaños muestrales n_1 y n_2 (4.15), D^2 , sea mayor o igual que el valor obtenido de esta misma distancia en la muestra original, $D^2(\mathbf{x})$, bajo la suposición

de igualdad de medias (hipótesis nula):

$$P(D^2 \geq D^2(\mathbf{x})|H_0)$$

donde \mathbf{x} representa la muestra original. Como dato a tener en cuenta, sabemos que las variables aleatorias X_1, X_2 son normales. Por tanto, en vez de realizar el típico bootstrap utilizando remuestreo no paramétrico (muestras con reemplazamiento de la muestra original), se tiene en cuenta la información anterior. Así, a partir de la muestra se estiman los valores de los parámetros de la distribución normal que serán utilizados para ir generando remuestras normales de igual tamaño que la muestra original. A esta técnica se la denomina bootstrap paramétrico.

El proceso, más detallado, es el siguiente:

Se generan un número fijo, $B = 1000$, de remuestras paramétricas del mismo tamaño que \mathbf{x} , $n_1 + n_2$. Cada una de las remuestras \mathbf{x}^* se produce por la generación de n_1 valores *iid* 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 :

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

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 $Prob\{D^2 \geq D^2(\mathbf{x})|H_0\}$ se estima por medio de:

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

Para un nivel de significación fijo α (en nuestro caso, $\alpha = 0.05$), la hipótesis nula se rechaza ($Y = 1$) si $P^* < \alpha$, y se acepta ($Y = 0$) en caso contrario.

La estimación P^* de $Prob\{D^2 \geq D^2(\mathbf{x})|H_0\}$, 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, [21]). Lo cual no garantiza que el verdadero nivel de significación $Prob\{P^* \leq \alpha|N(\mu, \sigma_1, \sigma_2)\}$ sea realmente α .

Para escoger el número de remuestras adecuado se han hecho diferentes ensayos de la simulación para $B = 500, 1000, 10000$ donde se observó que con $B = 1000$ se obtenía el mejor valor de estimación bootstrap en relación con el gasto de CPU requerido.

No ha sido probada teóricamente la validez del proceso bootstrap (y difícilmente lo será con las herramientas analíticas actuales, ya que el estadístico tiene forma implícita (4.17)), pero los resultados de simulación sugieren que su comportamiento es bastante bueno.

Bibliografía

- [1] J.H. Ahrens and U. Dieter. Generating gamma variates by a modified rejection technique. *Communications of the ACM*, 25:47–54, 1982.
- [2] C. Arenas, C.M. Cuadras, J. Fortiana. *MULTICUA: paquete no standard de análisis multivariante*. Publicaciones del Departament d'Estadística, n. 4, 1993.
- [3] A. A. Aspin. An examination and further development of a formula arising in the problem of comparing two mean values. *Biometrika* 35:88–96, 1948.
- [4] C. Atkinson and A.F.S. Mitchell. Rao's distance measure. *Sankhyà*, 43 A:345–365, 1981.
- [5] V. D. Barnett. *Random Negative Exponential Deviates*. Curso para ordenadores XXVII. C.U.P., 1965.
- [6] D.J. Best. Letter to the Editor. *Appl. Statist.*, 29:181, 1978.
- [7] G.M. Birtwistle, O.J. Dahl, B. Myhrhaug and K. Nygaard. *SIMULA BEGIN*. Auerbach, Philadelphia, Pa., 1973.
- [8] Y.M.M. Bishop, S.E. Fienberg and P.W. Holland. *Discrete multivariate analysis: theory and practice*. The MIT Press, 1975.
- [9] G.E.P. Box and M.E. Muller. A note on the generation of random normal deviates. *Ann. Math. Stat.*, 29:610–611, 1958.
- [10] R.J. Buehler. Some validity criteria for statistical inferences. *Ann. Math. Statist.*, 30:845–867, 1959.
- [11] J. Burbea and C.R. Rao. Entropy differential metric, distance and divergence measures in probability spaces: a unified approach. *J. Multivariate Anal.*, 12:575–596, 1982.

- [12] J. Burbea and J.M. Oller. *On Rao distance asymptotic distribution*. preprint series 67, Universitat de Barcelona, June 1989.
- [13] J.C.P. Bus and T.J. Dekker. Two efficient algorithms with guaranteed convergence for finding a zero of a function. *ACM Transactions on Mathematical Software.*, 1:330–345, 1975.
- [14] W.G. Cochran. Approximate significance levels of the Behrens-Fisher test. *Biometrics.*, 20:191–195, 1964.
- [15] W.G. Cochran and G.M. Cox. *Experimental Designs*. John Wiley and Sons, New York 1950.
- [16] H. Cramer. *Métodos matemáticos de estadística*. Ediciones Aguilar, 1953.
- [17] P.H. Crowley. Resampling methods for computation-intensive data analysis in ecology and evolution. *Annu. Rev. Ecol. Syst.*, 23:405–447, 1992.
- [18] O.J. Dahl and K. Nygaard. SIMULA - An ALGOL-Based Simulation Language. *Communications of the ACM*, 9:671–678, 1966.
- [19] G. Dahlquist and A. Bjorck. *Numerical Methods*. Prentice-Hall, 1974.
- [20] R.E. Dolkart, B. Halperin and J. Perlman. Comparison of antibody responses in normal and alloxan diabetic mice. *Diabetes*, 20:162–167, 1971.
- [21] M. Dwass. Modified randomization tests for nonparametric hypotheses. *Ann. Math. Stat.*, 28:181–187, 1957.
- [22] B. Efron. Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Stat.*, 3:1189–1242, 1975.
- [23] B. Efron. The geometry of exponential families. *Ann. Stat.*, 6:362–376, 1978.
- [24] B. Efron. Bootstrap methods: another look at the jackknife. *Ann. Stat.*, 7:1–26, 1979.
- [25] B. Efron. *The Jackknife, the Bootstrap, and Other Resampling Plans*. SIAM CMNS-Natl. Sci. Found. Monogr 38. Philadelphia:SIAM, 1982.
- [26] B. Efron. and D.V. Hinkley. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, 65:457–487, 1978.

- [27] A. L. Finkner. Further investigation on the theory and application of sampling for acarcity items. *Institute of Statistics*, University of North Carolina, Mimeo. Series 30, 1950.
- [28] R. A. Fisher. On a test of significance in Pearson's Biometrika Tables (No. 11). *J. R. Statist. Soc. Serie B*, 18:56-60, 1956.
- [29] M. J. Flynn. Some computer organizations and their effectiveness. *IEEE Trans. Comput.*, 21:948-960, 1972.
- [30] D.P. Gaver and G.L. Thompson. *Programming and Probability Models in Operations Research*. Wadsworth:Belmont, CA, 1973.
- [31] G. Gordon. *The Application of GPSS V to Discrete System Simulation*. Prentice-Hall, 1975.
- [32] G. Gordon. *System Simulation*. Prentice-Hall, 1978.
- [33] E.L. Grab and R. Savage. Tables of the expected value of $1/X$ for positive Bernouilli and Poisson variables. *J. Amer. Stat. Ass.*, 49:169-177, 1954.
- [34] R. Guigo, J. Ocaña. An Interactive Suport Software for Design and Use of Simulation Packages. *J. Math. Modeling*, 1205-1207, 1988.
- [35] R. Guigo. *Models de competència en medis compartimentats. Una aportació a la metodologia de la simulació* Tesis Doctoral, 1988.
- [36] J.M. Hammersley and D.C. Handscomb. *Monte Carlo Methods*. Methuen, London, 1964.
- [37] D.J. Hand. Expert systems in statistics. *Knowledge Engrg. Rev.*, 1:1-10, 1986.
- [38] D.J. Hand, F. Daly, A.D. Lunn, K.J. McConway and E. Ostrowski. *A Handbook of Small Data Sets*. Chapman & Hall, 1994.
- [39] E.J. Hannan. The variance of the mean of a stationary process. *J. Roy. Statist. Soc. B*, 19:282-285, 1957.
- [40] P.J. Huber. Robust Estimation of a Location Parameter. *Ann. Math. Stat.*, 35:73-101, 1964.
- [41] P.J. Huber. *Robust Statistics*. Wiley, 1981.

- [42] H.W. Karr, H. Kleine and H. Markowitz. *SIMSCRIPT I.5*. Consolidated Analysis Centers, Inc. Santa Monica, Ca., 1965.
- [43] P.J. Kiviat, J. Villanueva and H. Markowitz. *The SIMSCRIPT II Programming Language*. Prencince-Hall, 1969.
- [44] P.J. Kiviat, J. Villanueva and H. Markowitz. *The SIMSCRIPT II.5 Programming Language*. Consolidated Analysis Centers, Los Angeles, Ca., 1973.
- [45] J.P.C. Kleijnen. *Statistical Techniques in Simulation*. Part I. Dekker, New York, 1974.
- [46] A. Koch and U. Haag. The Statistical Software Guide. *Computational Statistics and Data Analysis*, 15:241–262, 1993.
- [47] D.E. Knuth. *The art of computer programming. Vol. 2. Seminumerical algorithms*. Addison-Wesley, 1981.
- [48] A.F.S. Lee and J. Gurland. Size and power of tests for equality of means of two normal populations with unequal variances. *J. Amer. Stat. Ass.*, 70:933–941, 1975.
- [49] E.L. Lehmann. *Testing Statistical Hypotheses*. John Wiley and Sons, New York 1986.
- [50] E.L. Lehmann. *Theory of Point Estimation*. John Wiley and Sons, New York 1983.
- [51] P.A.W. Lewis and E.J. Orav. *Simulation Methodology for Statisticians, Operations Analysts, and Engineers*. Wadsworth & Brooks/Cole Advanced Books & Software, California 1989, Vol I.
- [52] H.W. Lilliefors. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *J. Amer. Statist. Ass.*, 63:339–402, 1967.
- [53] G. Marsaglia and A. Zaman. A new class of random number generators. *Ann. Appl. Prob.*, 1:462–480, 1991.
- [54] R. L. Martino, C.A. Johnson, E.B. Suh, B.L. Trus and T.K. Yap. Parallel computing in Biomedical Research. *Science*, 265:902–908, 1994.
- [55] A.L. Mason and C.B. Bell. New Lilliefors and Srinivasan tables with applications. *Commun. Statist.- Simula.*, 15(2):451–477, 1986.

- [56] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.
- [57] Z. Michalewicz ed. *Statistical and Scientific Database Management*. In: *Proc. 5th Internat. Conf. SSDBM*. Lecture Notes in Comput. Sci. Vol. 420, Springer, 1990.
- [58] P.A.P. Moran. The estimation of standard errors in Monte Carlo simulation experiments. *Biometrika*, 62:1–4, 1975.
- [59] B.L. Nelson and B.W. Schmeiser. Decomposition of some well-known variance reduction techniques. *J. Stat. Comp. and Simul.*, 23:183–209, 1986.
- [60] J. Ocaña, G. Alonso and C. Ruíz de Villa. *El lenguaje ECOGEN de simulación genética. Manual preliminar*. Publicaciones de Bioestadística y Biomatemática. Vol 15. Universidad de Barcelona, 1986.
- [61] J. Ocaña and E. Vegas. Variance reduction for Bernoulli response variables in simulation. *Computational Statistics and Data Analysis*, 19:631–640, 1995.
- [62] B. A. van Oost, B. Veldhuyzen, A.P.M. Timmermans and J.J. Sixma. Increased urinary β -thromboglobulin excretion in diabetes assayed with a modified RIA kit-technique. *Thrombosis and Haemostasis*, 49:18–20, 1983.
- [63] E. S. Pearson and H.O. Hartley.(eds.). *Biometrika Tables for Statisticians*. Volume 1, Cambridge University Press, Cambridge, 1976.
- [64] C.R. Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37:81–91, 1945.
- [65] B.D. Ripley. *Stochastic Simulation*. Wiley, 1987.
- [66] P. Rothery. The use of control variates in Monte Carlo estimation of power. *Appl. Statist.*, 31:125–129, 1982.
- [67] A. Sánchez, J. Ocaña, and M.C. Ruíz de Villa. An environment for Monte Carlo simulation studies (EMSS), in: Y. Dodge and J. Whittaker (Eds), *Computational Statistics* Vol. 2 (Physica Verlag, 1992) 195–199.
- [68] J.D. Schaffer (Ed.). *Proceedings of the Third Conference on Genetic Algorithms*. Morgan Kaufman Publishers, 1989.

- [69] H. Scheffé. Practical solutions to the Behrens-Fisher problem. *J. Amer. Statist. Ass.* 65:1501-1508, 1970.
- [70] G. Simon. Computer simulation swindles, with applications to estimates of location and dispersion. *Appl. Stats.* 25(3):266-274, 1976.
- [71] W. Stahel and S. Weisberg. *Directions in Robust Statistics and Diagnostics. Part I and II.* Springer-Verlag, 1991.
- [72] J.W. Tukey. *Exploratory Data Analysis.* Addison-Wesley, 1977.
- [73] E. Vegas and J. Ocaña. Variance reduction for Bernoulli response variables in simulation. In: Y. Dodge and J. Whittaker (Eds), *Computational Statistics* Vol. 2 (Physica Verlag, 1992) 103-106.
- [74] W.N. Venables and B.D. Ripley. *Statistics with S.* Springer-Verlag, 1994.
- [75] B. L. Welch. The significance of the difference between two means when the population variances are unequal. *Biometrika* 29:350-362, 1938.
- [76] B. L. Welch. The generalization of Student's problem when several difference population variances are involved. *Biometrika* 34:28-35, 1947.
- [77] J. R. Wilson. Variance reduction techniques for digital simulation. *American Journal of Mathematical and Management Sciences.* 4:277-312, 1984.
- [78] D. Wishart. *CLUSTAN User Manual.* Program Library Unit, University of Edinburgh, 1978.
- [79] D.A. Wooff. Bounds on reciprocal moments with applications and developments in Stein estimation and post-stratification. *J.R. Statist. Soc. Serie B* 47:362-371, 1985.

