Inferencia estadística del ADN de plantas mediante citometría de flujo

Ramírez Ramírez Lilia Leticia

septiembre del 2000

Contenido

Agradecimientos

Quiero expresar mi agradecimiento al Dr. David A. Sprott por las platicas que tuvo con mi asesora y conmigo, con el fin de analizar los datos de este trabajo, así como a mi asesora, Dra. Eloísa Díaz-Francés, por su guía durante la elaboración de esta tesis, incluyendo su escritura.

Por la beca que recibí durante mis estudios de maestria (septiembre 1997-marzo 1999, número: 119868) agradezco al CONACyT, y por el apoyo económico durante la elaboración de este trabajo, al CIMAT y al Programa general de apoyo y desarrollo teconológico a la cadena productiva Agave-Tequila.

Por sus valiosos comentarios a este trabajo, expreso mi reconocimiento a mis sinodales: Dr. Miguel Nakamura y M. en E. Jose Luis Batún. En especial quiero agradecer al Dr. Nakamura por haberme asesorado durante mis estudios de maestría.

Al Dr. Luis Gorostiza y al Dr. Johan Van Horebeek agradezco sus consejos y apoyo.

Agradezco tanto a la Facultad de Matemáticas de la Universidad de Guanajuato como al Departamento de Estadística del CIMAT el apoyo económico que me brindaron para imprimir esta tesis.

Finalmente, a la Bióloga Irene Quiroz agradezco la revisión que hizo a la Sección 1.1 de esta tesis.

Prefacio

El tema de esta tesis surgió a partir de las asesorías estadísticas que realizó el CIMAT a uno de los proyectos que se encuentran dentro del "Programa general de apoyo y desarrollo tecnológico a la cadena productiva agave-tequila", el cual está encabezado principalmente por el Consejo Nacional de Ciencia y Tecnología (CONACyT), el gobierno del estado de Jalisco y el Consejo Regulador del Tequila (CRT).

El trabajo elaborado en esta tesis se derivó de la colaboración, dentro del programa mencionado, que tuvo el CIMAT con el equipo de biólogos del Instituto de Biología de la UNAM que también participaron en el programa. Uno de los objetivos principales de este proyecto era estimar la cantidad de ADN del agave con el que se produce el licor llamado tequila, usando la metodología denominada citometría de flujo.

El equipo asesor de estadística del CIMAT colaboró en sugerir el diseño del experimento final así como en proponer un modelo estadístico adecuado para analizar los datos experimentales de plantas de agave obtenidos con el citómetro de flujo (CF). El modelo estadístico que aquí se considera es nuevo en tanto a su aplicación a citometría de flujo pero no es nuevo en general pues se basa en la estimación de la razón de dos parámetros de localización, lo cual ha sido ampliamente discutido en la literatura estadística desde hace más de 60 años. La gran ventaja de aplicar este modelo estadístico a los datos del CF es que permite incorporar una característica experimental importantísima que consiste en que el citometrista suele ajustar en tiempo real una cota inferior para las observaciones en cada muestra de tejido analizada con el CF. Los modelos tradicionalmente usados no toman en cuenta esto. En esta tesis se explica detalladamente el modelo estadístico que se propuso para hacer inferencia sobre el contenido de ADN de alguna planta, con base en los datos obtenidos mediante la citometría de flujo.

En el Capítulo 1 se presentan los principales conceptos biológicos y estadísticos necesarios para comprender la aplicación estadística que se describirá más adelante. Para entender los principales resultados que origina el CF, en la Sección 1.1 se explican los principales conceptos del ciclo celular que afectan directamente la cuantificación del ADN nuclear y en la Sección 1.2 se explica cómo funciona el CF y las características distintivas de los datos que produce. En las Secciones 1.3 y 1.4 se presentan los conceptos estadísticos principales que se utilizarán para realizar la inferencia estadística sobre el contenido de ADN.

En el Capítulo 2 se expone la metodología estadística para analizar los datos obtenidos

con el CF y hacer inferencia sobre el contenido de ADN de la planta de interés. La Sección 2.2 describe la primera parte del análisis estadístico, el cual consiste en ajustar un modelo de mezclas de distribuciones a los histogramas de datos de ADN que arroja el CF. La segunda parte del análisis consiste en estimar la cantidad de ADN nuclear de una planta. Esta cantidad, como se mostrará, se relaciona directamente con la razón de medias de dos variables aleatorias normales, β. En la Sección 2.3 se presentan dos modelos para hacer inferencia sobre β, el más difundido (de Fieller) y el que proponemos utilizar, el cual es una generalización y extensión del primero. Este segundo modelo es un caso particular del presentado por Sprott (2000b) y tiene la gran ventaja de que considera el ajuste que efectúa el citometrista sobre el histograma de ADN que produce el CF.

La inferencia estadística sobre el ADN nuclear de la planta de interés está basada en el enfoque científico de la verosimilitud en Sprott(2000a) y se da en términos de intervalos de verosimilitud-confianza. Por esta razón, en la Sección 2.4 se exponen dos posibles aproximaciones a la verosimilitud del ADN nuclear de la planta que servirán para obtener dichos intervalos. En la Sección 2.5 se contrasta el modelo aquí propuesto con la metodología tradicional seguida en algunos trabajos biológicos para analizarlos y exponer sus deficiencias.

En el Capítulo 3 se analizan, con el modelo estadístico aquí propuesto, dos conjuntos de datos de agave obtenidos mediante citometría de flujo, por el Instituto de Biología de la UNAM para el Programa general de apoyo y desarrollo tecnológico a la cadena productiva Agave-Tequila.

Finalmente, en el Capítulo 4 se presentan las conclusiones de este trabajo.

La aportación principal de este trabajo radica en explicar en detalle y aplicar un método no estándar, conocido desde hace mucho tiempo en el área de estadística, de una manera novedosa y muy informativa en el área de citometría para cuantificar el ADN de una planta de interés.

Capítulo 1

Introducción

1.1 Conceptos básicos de biología celular

En esta sección se definirán los principales conceptos sobre la célula y la información genética que contiene, que se requieren para entender mejor la aplicación estadística que se presenta en esta tesis. Para ahondar más sobre estos temas se recomienda consultar el libro de biología molecular de Alberts et al. (1994).

La célula es la unidad básica de la vida, ya que cuenta con la capacidad de digerir nutrientes del medio, asimilarlos, administrarlos y expulsar sus desechos, con el fin de crecer y producir descendencia. Para poder realizar estas funciones, además de contar con la energía que adquiere del medio, debe poseer instrucciones exactas de cómo esta fuente de energía debe ser administrada. Por eso es necesario que esta información sea heredada en el proceso de reproducción celular.

Las células a las que nos referiremos a lo largo de este trabajo son aquellas que tienen núcleo, llamadas eucariontes. Estas forman parte de los organismos pertenecientes a cuatro de los cinco reinos: *Protista, Fungi, Plantae y Animalia*.

El núcleo es el centro de información de las células eucariontes, ya que contiene el ADN (ácido desoxirribonucleico) que es el que posee la información de cómo cada célula debe de utilizar los nutrientes que adquiere, indicándole que proteínas sintetizar[1]. El ADN (figura 1.1) está formado por dos cadenas de polinucleótidos constituidas por un azúcar de cinco carbonos, fosfatos y cuatro bases nitrogenadas: adenina, timina, citosina y guanina, (A,T,C y G respectivamente). Cada una de las cadenas integra una molécula en la que se encuentra la información necesaria para la célula ya que la disposición de las cuatro bases nitrogenadas en una cadena permite conocer la posición exacta de aquellas en la segunda[2]. Las dos cadenas se encuentran enlazadas mediantes puentes de hidrógeno que pueden romperse con facilidad, permitiendo la replicación de la molécula de ADN al

[1]Las mitocondrias y cloroplatos, que son organelos celulares localizados fuera del núcleo, también contienen ADN pero éste representa un código genético diferente del núcleo.

[2]Las cuatro bases se encuentran siempre arregladas formando las parejas: A-T y C-G

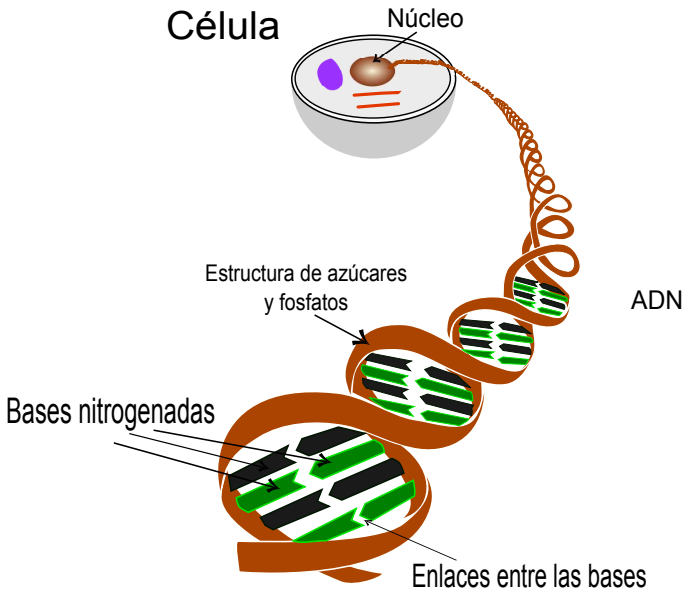


Figura 1.1: Estructura del ADN

sintetizarse nuevas partículas dentro del núcleo que siguen la estructura de información de la ya existente. Es decir, dada una cadena, la segunda se forma de manera complementaria para formar la doble hélice del ADN (figura 1.2). Este proceso se llama *replicación* y es el protagonista en el proceso de transmisión de la información de células madres a hijas.

El ADN de las células eucariontes se encuentra unido estrechamente con las histonas, que son proteínas que permiten, entre otras cosas, el superenrrollamiento de la molécula, que de otra manera, sería incapaz de acomodarse dentro del pequeño núcleo celular. El ADN superenrollado consituye el *cromosoma*, que puede ser simple o estar duplicado (figura 1.3). Los cromosomas duplicados (que son visibles al microscopio) se encuentran formados de dos partes iguales denominadas cromátidas, que representan una cadena de ADN e histonas sumamente condensadas.

Los genes son tramos (no necesariamente conexos) de ADN, es decir, un subconjunto de nucleótidos, que son capaces de ordenar la síntesis de proteínas indispensables para el buen funcionamiento celular. Los genes pueden ser tan pequeños como para contener sólo 1000 pares de bases o tan grandes como para incluir varios cientos de miles de ellos.

Las células tienen un ciclo vital en el que pasan por varias etapas, cada una con característica bioquímicas, fisiológicas e incluso morfológicas diferentes. A este proceso histórico de vida de las células se le llama *ciclo celular*.

Al observar a las células durante su ciclo de vida, los biólogos han observado que los cromosomas se presentan en pares idénticos en forma y en el modo en que están arreglados sus genes. Los cromosomas de un par así se llaman *cromosomas homólogos*.

Todas las especies tienen un número característico de cromosomas en cada célula.

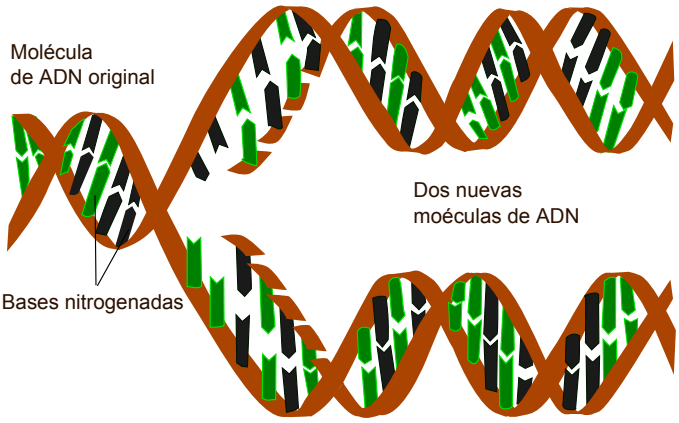


Figura 1.2: Proceso de replicación del ADN celular

En el hombre son 23 pares de cromosomas. El *número haploide* (n) de una especie se define como la cifra que expresa la cantidad de grupos de cromosomas homólogos que posee. Algunas especies, las haploides, poseen n cromosomas; las más numerosas (entre las que se encuentra el ser humano) son diploides, es decir, poseen $2n$ cromosomas[3]. Se diferencian, asimismo, especies triploides ($3n$), tetraploides ($4n$), etc. En general, se denominan como k-ploides a las especies con k mayor que cuatro y a k se le llama el *nivel de ploidía*.

En las especies que se reproducen sexualmente se halla una pareja de cromosomas que, a pesar de no ser iguales, son considerados homólogos. Estos son los cromosomas sexuales, pues son distintos en los seres femeninos que en los masculinos.

Además de encontrar los cromosomas sexuales en los organismos multicelulares con reproducción sexual, existen en ellos dos tipos diferentes de células, las llamadas *somáticas* (que son las que constituyen casi todo el individuo y que están presentes en los individuos con reproducción asexual) y las *sexuales*. Los dos tipos de células tienen ciclo celular diferente y se reproducen por los procesos llamados mitosis y meiosis, respectivamente.

La *mitosis* es el proceso de la división celular en la cual los cromosmas del núcleo de la célula se duplican y posteriormente se separan en dos grupos, para dar origen a dos células idénticas a la original. En la *meiosis* los cromosomas se duplican una vez y dividen dos para originar cuatro células con número de cromosomas a la mitad, es decir, que si una célula diploide se reproduce por meiosis, las cuatro células hijas resultantes serán haploides. Las células sexuales (femeninas y masculinas) deben ser haploides, ya que su función es unirse en la reproducción sexual para formar un nuevo organismo con el número cromosómico de la especie.

Como nos interesa analizar las células somáticas de las plantas a través del Citómetro de Flujo (CF), en este trabajo describiremos el ciclo celular considerando sólo a la mitosis

[3]Para el ser humano se expresa $2n$=46, con número haploide n=23

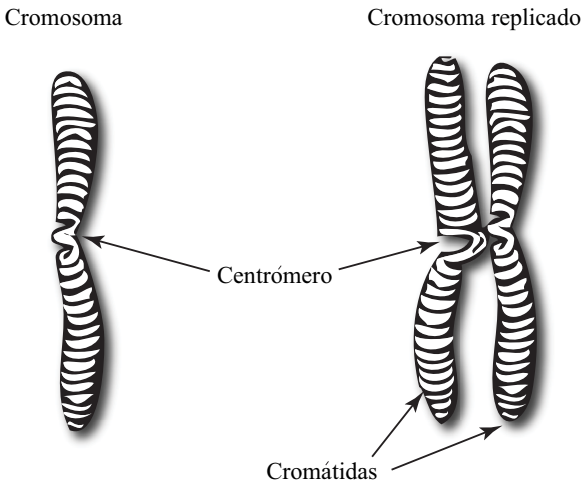


Figura 1.3: Estructura de los cromosomas

como forma de reproducción.

El ciclo celular (figura 1.4) dura aproximadamente 24 horas, aunque este tiempo puede variar con el tipo de célula o especie, y comprende cuatro etapas o fases principales:

1. **Fase G_1-** Periodo de crecimiento célular; los organelos de duplican.

2. **Fase S-** Periodo de duplicación de ADN. La celula sintetiza nuevas cadenas de ADN formándose una réplica de cada cromosoma. En esta etapa el ADN se condensa para formar cromosomas pequeños y enrrollados (figura 1.3), los cuales se empiezan a dividir longitudinalmente para formar dos cromátidas.

3. **Fase G_2-** Crecimiento de la célula. En este periodo la célula se prepara para la mitosis.

4. **Fase M-** Periodo que comprende a la mitosis. Los cromosomas duplicados se separan para disponer un juego de ellos para cada nueva célula.

1.2 El Citómetro de flujo

El Citómetro de Flujo (CF) se desarrolló originalmente, hace más de dos décadas, como un método mucho más rápido que el estándar para analizar las células sanguíneas. Desde entonces sus aplicaciones, no sólo se han incrementado día con día, en investigaciones científicas, tales como botánica, microbiología, ecología, embriología, inmunología, biología molecular y genética, sino que también en aquellas relacionadas con la industria. Su número creciente de aplicaciones se debe a que permite evaluar, fácilmente, las características individuales de miles de células y sus organelos, tales como núcleo y cromosomas,

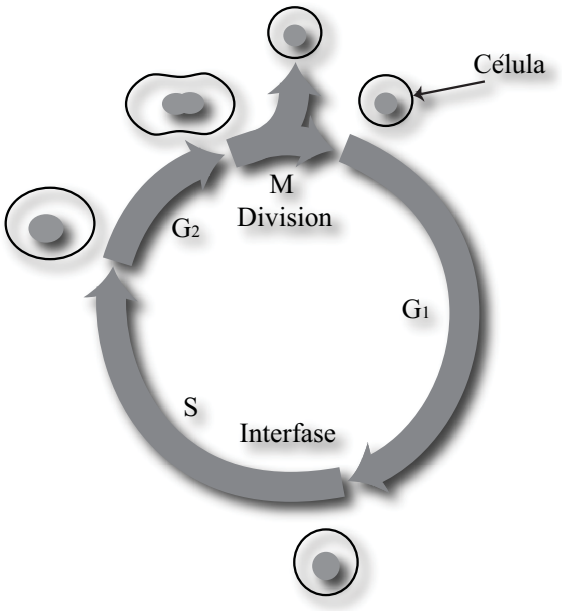


Figura 1.4: Ciclo celular

en cuestión de minutos. Esto último hace a la citometría de flujo, una metodología muy eficiente en comparación con otras que en el pasado requerían meses o años para obtener en parte, la información que el CF es capaz de brindar.

Una de las principales aplicaciones médicas y biológicas del CF se relaciona directamente con el análisis del contenido de ADN, pues gracias a CF se puede determinar el nivel de ploidía de las células (DNA index, Hidderman, et al., 1984), pronosticar casos de pacientes con cancer (Parker, 1988) y, más recientemente, determinar la tasa de producción de células cancerosas (Begg, et al.,1985). El propósito de este trabajo es modelar datos de ADN del CF para determinar la cantidad de ADN de una planta.

El CF (figura 1.5) tiene un mecanismo que consta de un lector óptico (o fotodetector), una fuente de luz intensa y un punto de observación o de enfoque, de ambos. Estos tres componentes conforman la llamada *cámara del CF*, que se considera como la parte principal del aparato.

En la figura 1.5 se observa que el CF está conectado a una computadora, la cual permite que el usuario introduzca datos relacionados a las características de la corrida o ejecución, tales como número de identificación (ID) y nombres de los organismos (plantas o animales) en análisis. También por medio del teclado de la computadora, el usuario puede cambiar o asignar valores (que originalmente determina el programa) de algunas variables, ya sea con fin exploratorio o porque debido a la calidad de los datos, no es posible para el programa definirlos.

Las células o partículas que serán analizadas se preparan con una tinta fluorescente

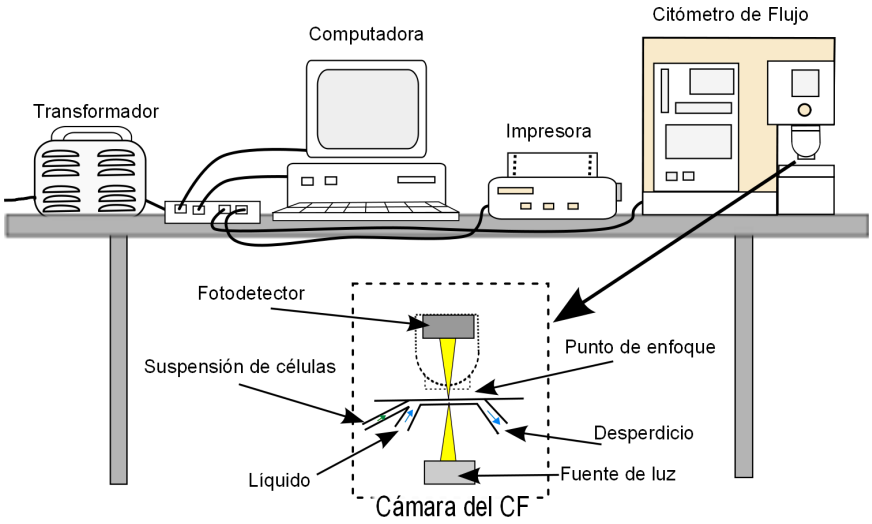
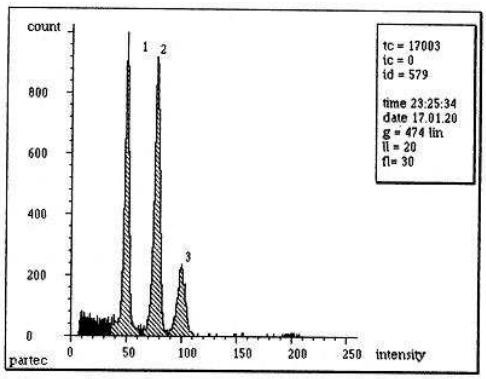


Figura 1.5: Diagramas del equipo periférico y de la cámara del CF

que se fija sobre o dentro de cada subpartícula de interés, para luego ser inyectadas en una suspensión, hacia el centro de la cámara del CF, para que pase justamente por el punto de observación. La suspensión se inyecta en forma de un delgado chorro para exponer rápidamente a cada una de las células o partículas a la luz y al lector óptico. Bajo la excitación producida por la luz del CF, la tinta comienza a emitir señales de luz fluorescente y éstas son registradas por el fotodetector. Posteriormente, las señales provenientes del fotodetector, se transforman en un impulso eléctrico cuya intensidad es proporcional a la de la luz registrada.

Existen varios métodos para preparar las muestras de tejido que serán analizadas por el CF para cuantificar su ADN (algunos han sido propuestos por Otto, 1990 y Dolezel y Göhede, 1995) y en todos ellos hay que teñir con tinta fluorescente y preparar en suspensiones los núcleos celulares. Un ejemplo es el procedimiento que se utilizó para analizar la planta *Agave tequilana* Weber, variedad *azul* (agave azul), que se presentan en el Capítulo 3. Este proceso consiste en liberar los núcleos de la muestra de tejido cortando ésta en pequeños pedazos con un bisturí y después teñirlos con fluorcromos, del tipo de los que se intercalan en la cadena de ADN, para que la intensidad de la fluorescencia originada en el proceso de citometría de flujo sea proporcional a la cantidad de ADN nuclear.

La intensidad de fluorescencia registrada se transforma después en un impulso eléctrico de intensidad proporcional. La señal eléctrica obtenida, a continuación se incrementa a través de una función lineal o logarítmica (y sólo una) para después convertirse a un entero mediante el llamado *convertidor analógico-digital* (ADC por sus siglas en inglés). Entonces el ADC tiene por entrada una variable contínua (el impulso eléctrico) y por salida o resultado una variable discreta llamada *canal*. La resolución del ADC en el CF



ID: 579 17.01.20 23:25:34 TC: 17003
Name: A teq + Zea
Zea peaks 1,3 A teq peak 2
Hoja basal plan embrion somatico I-3
Otto + PI

Peak	Index	Mode	Mean	Area#	Area%	CV%
1	1.00	50	49	6007	35.33	4.08
2	1.54	77	77	6502	38.24	3.25
3	2.00	100	100	2241	13.18	3.50

No cycle evaluated
--

Figura 1.6: Histograma de Frecuencia originado por el CF

determina si existen 256 (del 0 al 255) o 1,024 canales (del 0 al 1,023), y en consecuencia, define la exactitud de los resultados posteriores, Eudey (1996).

El citometrista puede manipular el detector de voltaje que registra el impulso eléctrico para hacerlo más o menos sensible a ésta señal. Este hecho ocasiona que el conjunto de datos, incrementados y discretizados, cambie de posición con respecto a los canales y que la varianza de las señales eléctricas (incrementadas con una función lineal), no sean constantes con respecto a su localización sobre el eje de los canales. Es decir, la señal eléctrica debe incrementarse 10% para registrarse en el canal 11, en lugar del 10, pero necesita sólo un incremento del 1% para saltar del canal 100 al 101. Es por esto que en un CF que incrementa su señal por medio de una función lineal, entre más grande sea el número de canal (o más sensible sea el detector de voltaje), más grande será la resolución de los datos, en término de canales.

Uno de los resultados del análisis del contenido de ADN, que da el CF, es un histograma de frecuencias que describe gráficamente cuántos núcleos o partículas fueron medidos con la intensidad de fluorescencia de cada canal (del 0 al 255 o del 0 al 1,023). Hay que recalcar que la posición del histograma sobre los canales es una medida relativa de la fluorescencia, ya que el citometrista puede modificar su posición simplemente manipulando el detector de voltaje del CF. Después de que el citometrista fija el nivel de detección de voltaje, el CF resume toda la información de la fluorescencia registrada por cada núcleo, en el histograma de frecuencias (ver ejemplo en la figura 1.6) y guarda estos datos en un archivo ascii. Como se puede observar en la figura 1.6, el histograma se compone de varios picos (en el ejemplo son tres y se encuentran enumerados), que se relacionan tanto con la diferente cantidad de ADN presente en las células durante su ciclo de vida, como con las

diferentes plantas (con contenido de ADN diferente), que se analizan simultáneamente.

El ejemplo que describe la figura 1.6 corresponde al histograma e información descriptiva que arroja el software del CF de la compañía Partec CA II GmbH (Münster, Alemania) para una muestra de tejido dada. La gráfica del ejemplo consiste del número de núcleos contabilizados por cada canal de dos plantas diferentes (agave azul y maíz) que se analizan simultáneamente. Los picos 1 y 3 corresponden a las células de la planta de maíz (*Zea mays*) en sus fases G_1 y G_2-M , respectivamente, y el pico 2 representa a la fluorescencia de los núcleos de agave azul en su fase celular G_1.

Como ya se mencionó en la Sección 1.1, la cantidad de ADN en el núcleo de una célula depende de la fase por la que ésta esté transitando. Tienen kn cromosomas en la fase G_1 y $2 \times kn$ cromosomas en las fases G_2 y M (G_2/M), mientras que en la fase S su cantidad de ADN es intermedia a la de las dos anteriores. En la literatura a veces los autores se refieren sólo a tres fases en el ciclo celular (G_1, S y G_2), llamando G_2 a las fases G_2 y M. Pero en la práctica, estos nombres cumplen también con la función de clasificar al ciclo celular según la cantidad de ADN que contiene la célula.

El histograma de ADN suele estar acompañado de información como número de identificación (ID), nombres de las plantas en estudio, número de picos y planta a la que corresponden, medias estimadas y modas de los picos. Los datos como el ID, nombre de las plantas y picos que le corresponden son introducidos por el usuario, mientras que otros, como el número de picos y su localización sobre el rango de canales, son determinados por el CF, con opción a que el usuario los cambie o defina, ya sea con fin exploratorio, porque el CF lo hace erróneamente, o porque el CF no puede darles valores iniciales debido a la mala calidad del histograma obtenido (manual del CF de Partec).

Como el histograma se presenta en la pantalla del CF desde el inicio de la contabilización de la fluorescencia en los núcleos (éste se va actualizando confome el CF adquiere los nuevos datos), se convierte en el instrumento a través del cual interactúa el usuario o citometrista y el software del CF para determinar cuando dar por terminada la corrida del aparato a partir de algún criterio como alcanzar un cierto número de núcleos contabilizados u obtener picos con una marcada simetría. Con el histograma de frecuencias final que aparece en la pantalla del CF el usuario también puede auxiliarse para obtener otros datos tales como el área debajo de la curva, en zonas determinadas por el usuario, y el número de picos a analizar, usando el teclado o el ratón de la computadora.

Como la cantidad de ADN nuclear se mide indirectamente por la intensidad de fluorescencia que produce y ésta es proporcional al número de moléculas fluorescentes adheridas al ADN, es de esperar que la moda del pico del histograma correspondiente a la fase G_2/M se encuentre cerca del canal con el doble de valor de aquel sobre el que se encuentra la moda del pico originado por la fase G_1.

Teóricamente el histograma de frecuencia del contenido de ADN, para la planta de maíz del ejemplo (picos 1 y 3), debería ser como el de la figura 1.7 pero las fuentes de variación de la intensidad de fluorescencia originan un error que se manifiesta en la desviación de los datos alrededor del verdadero valor del contenido de ADN.

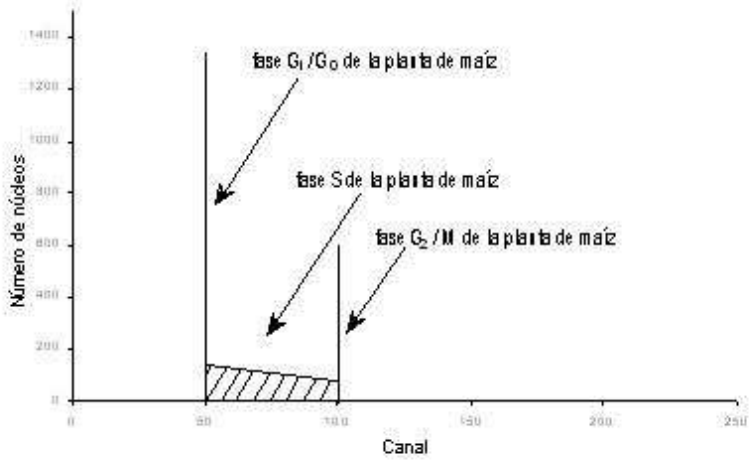


Figura 1.7: Histograma de Frecuencia teórico para el maíz en el ejemplo de la figura 1.6.

Eudey (1996) y Watson (1992) distinguen tres principales fuentes de variación para los datos del CF, que son:

1. Error debido al procesamiento de las células para ser analizadas por el CF, tales como variabilidad en el proceso de teñimiento y en el corte del tejido con el bisturí.

2. Variación perteneciente a cualquier sistema o aparato de medición.

3. Variabilidad biológica entre individuos distintos.

Watson (1992, p. 2) menciona que la tercera fuente de variación es la más importate porque es de la cual, los científicos obtienen información de interés. En Watson (1992, Capítulo 2) se analiza con mayor profundidad las tres fuentes de variación anteriores y otras para los datos del CF; sin embargo, en este trabajo sólamente consideraremos las tres ya mencionadas.

Con el fin de disminuir el error de medición originado por el CF, antes de elaborar las mediciones, los biólogos o citometristas, realizan el procedimiento denominado "de calibración", que consiste en utilizar estándares (como el de células sanguíneas de pollo o pescado) para ajustar el CF con el fin de dar una amplitud de señal máxima y un mínimo de variación a las lecturas del fotodetector, Dolezel (1995).

Uno de los principales indicadores de una mala calibración es la falta de simetría en los picos de las fases G_1 y G_2/M. El manual del CF de Partec señala que la falta de simetría en estos picos es indicio de una mala calibración o de la contaminación de los núcleos a

analizar con los provenientes de otro (u otros) tejidos. En caso del ejemplo presentado en la figura 1.6 se aprecia que los picos del histograma tienen una marcada simetría.

En la modelación estadística es usual suponer que los errores por intensidad de canal no solo son simétricos, sino que tienen distribución normal. Eudey (1996) señala que la mayoría, si no es que todos, los modelos que buscan describir los datos de CF consideran que las intensidades de fluorescencia por núcleo, en fases G_1 y G_2/M se distribuyen, cada uno, normal con medias igual al verdadero valor del contenido de ADN (en términos de canal) en sus fase correspondiente.

Sin embargo, la mayoría de los modelos estadísticos sí difieren en la hipótesis de distribución para la fase S. Algunos consideran que la función de distribución para la fase S es de forma trapezoidal (como se representa en la figura 1.7), otros se basan en la teoría de crecimiento poblacional (Steel, 1968) o consideran el modelo de normales múltiples usados por Fried (1976, 1977) y Fried y Manden (1979). Para el lector interesado en ahondar sobre este tema, puede consultar Watson (1992, Sección 8.3).

A partir de los datos del histograma, el software del CF proporciona algunas estadísticas descriptivas, tales como el total de observaciones en la muestra analizada (*tc*), el valor de las modas (*mode*), un estimador de las medias (*mean*) y el coeficiente de variación (CV) de cada pico del histograma.

El *coeficiente de variación* (CV) es una medida de la dispersión de los datos para cada fase celular (excepto la fase S) en el histograma, que es igual a la desviación estándar (DE) entre la media ($\sigma/\bar{x}$), Watson (1992, p. 94). El CF de Partec calcula los CV's de cada pico del histograma a partir de la definición siguiente:

$$\text{CV} = (0.5)(\text{ancho del pico al } 67\% \text{ de su contenido máximo})/\bar{x}.$$

Para la distribución normal, esta cantidad es igual a $(0.800955)\sigma/\bar{x}$, que es proporcional a la cantidad señalada por Watson.

En el caso de los CF que incrementan su señal elétrica a través de una función lineal (como lo hace el CF de Partec), el uso del CV como indicador de la calidad de las mediciones, en lugar de la desviación estándar (DE), se justifica porque el CV, a diferencia de la DE, no depende de la posición del pico sobre los canales. Por ejemplo, consideremos que una distribución con DE 5 y media de 100. Esto da como resultado un CV igual a 0.05. Ahora si la señal es incrementada linealmente con factor diez, la nueva media es 1000 y una célula que anteriormente aparecía en el canal 101 ahora se registrará en el 1010. Entonces la desviación estándar será de 50 con un CV todavía igual a 0.05, Watson (1992, p. 96). Este resultado, sin embargo, tiene la característica de que la propiedad de permanecer constante del CV estará en función de que el convertidor analógico-digital permita conocer la dispersión de los datos en la distribución. Si la distribución tiene valores cercanos al canal cero, entonces casi todos, si no es que todos, los datos caerán en un solo canal, lo que disminuirá la desviación estándar a cero y también así el CV.

En el caso de que el incremento sea logarítmico, el CV no es constante, pero la desviación estándar de cada pico, más o menos, sí lo es, Watson (1992, p. 95).

Cuando la muestra de tejido analizada es medida con el CF en condiciones adecuadas, se espera que los picos del histograma sean simétricos, que los CV's correspondientes, no sean grandes (usualmente menores que 5) y que exista una baja proporción de partículas o núcleos fragmentados (basura o *debris*), que es lo que se acumula en los primeros canales de la gráfica del histograma de frecuencias. En la figura 1.6 el CF de Partec considera como basura a la información de aproximadamente los primeros 35 canales, que es lo que se encuentra en color negro. Cuando alguna de las condiciones anteriores no se cumple, los biólogos o citometristas repiten la corrida del CF con una nueva preparación de muestra del mismo tejido (o tejidos).

Algunas de las principales ventajas que tiene el CF frente a otras métodologías como la citometría estática, mencionadas en Dolezel (1997), para cuantificar el ADN, se enumeran a continuación.

1. La preparación de las muestras de tejido celular para analizar en el CF es rápida.

2. El CF es muy rápido y permite el análisis de miles de partículas en un solo día de trabajo.

3. En general el análisis mediante el CF no es destructivo, ya que las muestras de tejidos se preparan a partir de unos cuantos miligramos de tejido.

Sin embargo, a pesar de la alta tecnología que utilizan los CF para realizar las lecturas de la fluorescencia, las estadísticas descriptivas que proporcionan los softwares de la mayoría de ellos son muy primitivas y a veces innecesariamente poco informativas. Por ejemplo, el software de Partec calcula la media truncada de cada pico fijando un intervalo arbitrario alrededor de la moda. En este cálculo intervienen miles de observaciones discretas por lo que a pesar de ser truncada debería ser muy informativa; sin embargo algunos softwares del CF redondean la media calculada al entero más cercano, desechando así información valiosa con la que ya se contaba. El redondear el resultado de las medias truncadas pone un obstáculo para estimar la variabilidad real de los datos, en análisis posteriores. Además este tipo de software parece no tomar en cuenta que los datos que arroja el CF, provienen de manera natural de una mezcla de distribuciones.

1.3 Conceptos de inferencia estadística

La ciencia estudia los fenómenos naturales repetibles y su propósito es predecir la naturaleza y cuando sea posible, cambiarla o controlarla, Sprott (2000a, p. 1). A este respecto Fisher (1942) escribió "in order to assert that a natural phenomenon is experimentally desmostrable we need, not an isolated record, but a reliable method of procedure". Es decir, que con el objetivo de afirmar que un fenómeno natural es experimentalmente demostrable, necesitamos no de una sola observación, sino de un método o procedimiento fiable. Esto quiere decir que la demostración de un fenómeno natural no se puede basar

en un solo evento. El procedimiento al que se refiere Fisher se llamará *experimento* y ya sea que tenga o no factores controlados, se requiere que estos puedan realizarse una y otra vez.

En la actualidad, la experimentación científica incorpora los elementos de incertidumbre que se originan al medir u obtener los datos provenientes de los experimentos o los elementos estocásticos que pertenecen directamente al fenómeno de interés. En general se utilizan modelos probabilísticos para describir el proceso que genera a los datos. Para elaborar una descripción probabilística del proceso generador de las observaciones, es necesario utilizar la herramienta estadística, la cual, según Edwards (1922, p. 9), se compone del modelo de probabilidad, un conjunto de hipótesis estadísticas y los datos resultados de los experimentos.

En los problemas de inferencia estadística es usual comenzar con unas observaciones $\boldsymbol{x} = x_1, \ldots, x_n$ y con alguna información de cómo éstas fueron colectadas. Entonces, se intenta formular un modelo de probabilidad $f(\boldsymbol{x}; \theta), \theta \in \Theta$, para el fenómeno que origina los datos. Es importante tratar al conjunto de datos en el contexto del experimento y tener en cuenta lo que se sabe de otras aplicaciones semejantes, Kalbfleisch (1985, p. 2).

En esta Sección se presentan los principales conceptos de inferencia estadística que se utilizarán el Capítulo 2. En la Subsección 1.3.1 se exponen los conceptos básicos de verosimilitud: función relativa de verosimilitud, estimador máximo verosímil, funciones de puntuación y de información, e intervalos de verosimilitud. En la Subsección 1.3.2 se trata la función de verosimilitud perfil para resolver los problemas que surgen cuando en el modelo estadístico existen muchos parámetros desconocidos y la inferencia sólo se quiere realizar para algunos de ellos.

En la Subsección 1.3.3 se define a la verosimilitud originada por un pivotal y a los intervalos de verosimilitud-confianza, los cuales constituyen la principal herramienta de inferencia estadística desde el punto de vista de este trabajo. Finalmente en la Subsección 1.3.4 se presentan algunas aproximaciones a la verosimilitud con el fin de dar origen a los intervalos de verosimilitud-confianza necesarios para hacer inferencia sobre el parámetro de interés.

1.3.1 La función de verosimilitud

La teoría de estimación estadística se relaciona con el problema de cuantificar la incertidumbre o plausibilidad de los valores de las cantidades desconocidas o parámetros, con base en los datos obtenidos por la repetición de experimentos. Sprott (2000a, p. 3) nombra a este tipo de inferencia estadística, *estimación inferencial* para distinguirla de la estimación puntual cuyo objetivo principal es brindar estimadores puntuales óptimos en el sentido de algún concepto, como lo es la función de pérdida o mínima varianza.

Edwards (1992, p. 8) comenta que ha sido reconocido y aceptado que la probabilidad de un resultado en un ensayo o experimento sea una medida racional de creencia, expresada bajo la ignorancia del verdadero valor ocurrido o por ocurrir. En el problema de inferencia

no se puede dar una medida empírica a la hipótesis $H : \theta = \theta_0$ ($\theta_0 \in \Theta$) ya que estas hipótesis no pueden tratarse como si surgieran como resultado de experimentos.

El problema de estimar el verdadero valor del θ llevó a Fisher a introducir el concepto de *verosimilitud*, el cual busca dar un orden de preferencia de los valores de θ en el espacio parametral Θ a la luz de los datos observados. Fisher (1948) comentó:

> "...the mathematical concept of probability is inadequate to express our mental confidence or diffidence in making such inferences, and that the mathematical quantity which appears to be appropriate for measuring our order of preference among different possible populations does not in fact obey the laws of probability. To distinguish it from probability, I have used the term *Likelihood* to designate this quantity",

que en forma más o menos textual dice:

> ...el concepto de probabilidad es inadecuado para expresar la confianza empírica al realizar dichas inferencias [sobre el valor de θ] y la cantidad matemática que parece ser apropiada para medir nuestro orden de preferencia entre las diferentes posibles poblaciones [las poblaciones con diferente valor de θ] no obedece, de hecho, las leyes de la probabilidad. Para distinguirla de la probabilidad, he usado el término de verosimilitud.

Y en 1921 (p. 24) definió a la verosimilitud de la siguiente forma:

> "What we can find from a sample is the likelihood of any particular value or ρ, if we define the likelihood as a quantity proportional to the probability that, from a population having that particular value or ρ, a sample having the observed value r, should be obtained. So defined, probability and likelihood are quantities of an entirely different nature".

La traducción es:

> Lo que encontramos de una muestra es la verosimilitud de cualquier valor particular de θ, θ_0. Si definimos la verosimilitud como una cantidad proporcional a la probabilidad de que en una población con valor de $\theta = \theta_0$ una muestra aleatoria tenga el valor observado x, entonces la probabilidad y la verosimilitud son cantidades de naturaleza totalmente diferente.

Entonces la verosimilitud del parámetro θ dada la muestra observada $\boldsymbol{x} = x_1, \cdots, x_n$ es proporcional a la probabilidad de observar $\boldsymbol{x}$ dado el valor θ. Es decir:

$$L(\theta; \boldsymbol{x}) = c(\boldsymbol{x}) P(\boldsymbol{x}; \theta), \tag{1.1}$$

donde $c(\boldsymbol{x})$ es una función de $\boldsymbol{x}$, positiva y acotada que no depende de θ. Para detallar sobre las diferencias entre la verosimilitud y la probabilidad, ver Edwards (1992, Sección 2.2).

La definición de verosimilitud expresada en (1.1) es exclusiva para variables aleatorias discretas porque se encuentra expresada en términos de probabilidades, sin embargo no pierde generalidad si consideramos que todos los instrumentos de medición generan datos discretos porque tienen precisión finita. Cuando la precisión del instrumento es suficientemente grande como para querer considerar un modelo de probabilidad contínuo, se puede realizar la aproximación que se obtiene en Sprott (2000a, Sección 2.5), la cual resulta en la siguiente expresión

$$L(\theta; \boldsymbol{x}) = c(\boldsymbol{x}) f(\boldsymbol{x}; \theta), \tag{1.2}$$

donde $f(\boldsymbol{x}; \theta)$ es la función de densidad conjunta de $\boldsymbol{x} = x_1, \ldots, x_n$ y $c(\boldsymbol{x})$ es una función de $\boldsymbol{x}$, positiva y acotada, que no depende de θ.

En el caso en que las observaciones registradas provengan de realizaciones independientes de experimentos homogéneos, entonces el conjunto de los n resultados $\{X_i\}_{i=1}^n$, forma una colección de n variables aleatorias independientes e idénticamente distribuidas. Por tanto la definición (1.1) y la aproximación (1.2) pueden expresarse como

$$L(\theta; x) = c(\boldsymbol{x}) \prod_{i=1}^{n} f(x_i; \theta), \tag{1.3}$$

donde $f(x_i; \theta)$ es la probabilidad o función de densidad marginal (dependiendo si el modelo de probabilidad que se utiliza es discreto o contínuo) para la observación x_i, $i = 1, \ldots, n$.

Como la definición de la verosimilitud involucra a una constante $c(\boldsymbol{x})$, no negativa, sus valores absolutos no tienen significado. Sólamente el valor numérico de la razón de verosimilitudes tiene interpretación. Este enfoque sobre la verosimilitud tiene la siguiente traducción frecuentista: si $L(\theta_1; \boldsymbol{x})/L(\theta_2; \boldsymbol{x}) = p$ (para $\theta_1, \theta_2 \in \Theta$), entonces en muestras repetidas de la población definida por $\theta = \theta_1$, la muestra observada $\boldsymbol{x}$ se obtendrá p veces con más frecuencia que las muestras repetidas provenientes de la población con $\theta = \theta_2$. La razón de verosimilitudes es entonces, la medida de plausibilidad de la hipótesis simple $\theta = \theta_1$ relativa a la tambien hipótesis simple $\theta = \theta_2$, basada en la muestra observada.

Debido a que sólo la razón de verosimilitudes tiene significado es conveniente estandarizar la verosimilitud con respecto a su máximo, comparando así todos los valores del parámetro con el valor más plausible, $\hat{\theta}$, el valor de θ que maximiza la función de verosimilitud. Estandarizando a la verosimilitud de esta forma, se obtiene una representación única que no involucra a la constante $c(\boldsymbol{x})$. El resultado, es la *función de verosimilitud relativa*, también llamada la *verosimilitud normalizada*.

$$R(\theta; \boldsymbol{x}) = \frac{L(\theta; \boldsymbol{x})}{\sup_{\theta \in \Theta} L(\theta; \boldsymbol{x})} = \frac{L(\theta; \boldsymbol{x})}{L(\hat{\theta}; \boldsymbol{x})}. \tag{1.4}$$

Considerando esta definición tenemos que $0 \leq R(\theta; \boldsymbol{x}) \leq 1$. El argumento de $L(\theta; \boldsymbol{x})$ en el que se alcanza el máximo, $\hat{\theta}$, se llama el *estimador máximo verosimil de* θ y representa al valor más plausible en el espacio Θ dada la muestra observada $\boldsymbol{x}$. En esta tesis se supondrá que $\hat{\theta}$ existe y es único.

Ejemplo 1.1 (Kalbfleisch, 1985, ejercicio 3, p. 30) Se supone que el tiempo entre la emisión de partículas de una fuente radioactiva tiene distribución exponencial con media θ. Sin embargo, el contador Geiger que se utiliza tarda en activarse una unidad de tiempo después de haber registrado una emisión. Entonces la función de densidad de la v.a. X, el tiempo entre registros consecutivos, es

$$f(x) = \frac{1}{\theta} e^{-(x-1)/\theta}, \qquad \text{para } x \geq 1 \ \text{ y } \ \theta > 0. \tag{1.5}$$

La función de densidad (1.5) se conoce como la distribución exponencial con tiempo de vida garantizado.

Los 10 datos que a continuación se presentan corresponden a 10 observaciones de tiempos entre registros

1.47	1.46	2.20	1.36	2.90
3.71	3.89	1.29	1.86	1.81

y la función de verosimilitud que originan, es la siguiente:

$$L(\theta; \boldsymbol{x}) = \frac{1}{\theta^n} e^{-\sum(x_i-1)/\theta} = \frac{1}{\theta^n} e^{-11.95/\theta},$$

con $c(\boldsymbol{x}) = 1$.

El valor de θ que maximiza a la función de verosimilitud $L(\theta; \boldsymbol{x})$ es igual a

$$\hat{\theta} = \frac{\sum(x_i - 1)}{n} = 1.195,$$

y a partir de éste se obtiene a la función de verosimilitud relativa

$$R(\theta; \boldsymbol{x}) = \left(\frac{1.195}{\theta}\right)^n e^{-11.95/\theta + 10}.$$

La gráfica de la función de verosimilitud relativa del parámetro θ, originada por los 10 registros obtenidos con el contador Geiger se presenta en la figura 1.8.

La gráfica de la función de verosimilitud relativa es una herramienta visual muy útil que permite realizar comparaciones entre las verosimilitudes de todos los valores de θ. Formalmente las cualidades de subconjuntos del espacio paramétrico se resumen en los llamados *intervalos o regiones de verosimilitud*. Un intervalo o región de verosimilitud es el conjunto de valores de θ tales que su verosimilitud relativa es mayor o igual que una costante a. Entonces, si

$$C = \{\theta | R(\theta; \boldsymbol{x}) \geq a, 0 \leq a \leq 1\}.$$

se dice que el nivel de verosimilitud de la región de verosimilitud C es del $100a\%$. Como todo valor de θ en la región anterior tiene verosimilitud relativa $R(\theta; x) \geq a$, y todo

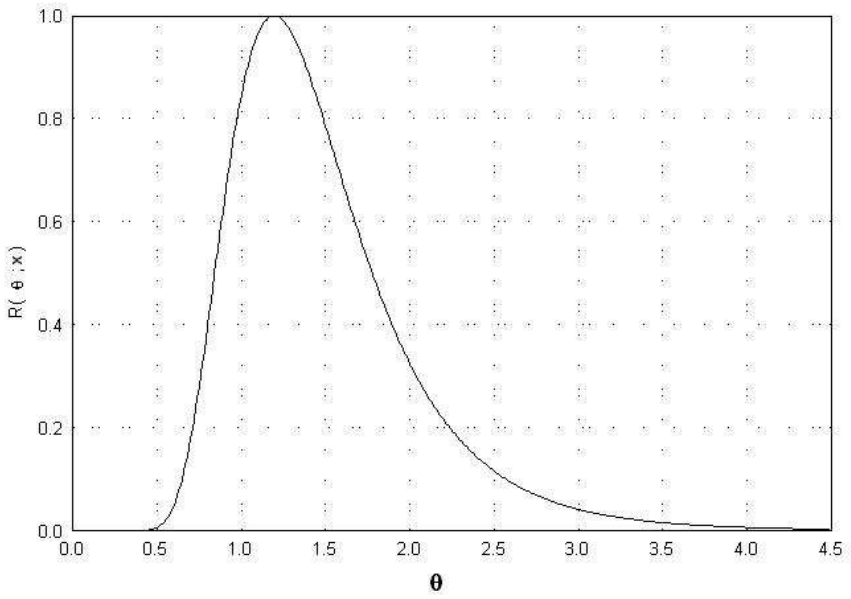


Figura 1.8: Función de verosimilitud relativa de θ correspondiente al ejemplo 1.1

aquel fuera, tiene verosimilitud relativa menor, las regiones entonces separan los valores plausibles de θ de los no plausibles a un nivel a, Sprott (2000a, p. 14). Cuando a es pequeño, por ejemplo igual a 0.03 o 0.01, usualmente se considera que valores del parámetro fuera del intervalo de verosimilitud son insignificantes y se pueden descartar.

Normalmente es deseable tener más de un intervalo de verosimilitud o por lo menos un intervalo y el valor máximo verosimil del parámetro, $\hat{\theta}$ para poder conocer las principales características de la función de verosimilitud (por ejemplo simetría). Para ver algunos ejemplos, referirse a Kalbfleisch (1985, pp. 19-23). La forma en que se puede reproducir la función de verosimilitud relativa a partir de los intervalos de verosimilitud, es dibujando varios intervalos a sus niveles respectivos a de verosimilitud. Los extremos de los intervalos dibujados, delinearán la curva de la función de verosimilitud relativa, como se muestra en la figura 1.9 para el ejemplo 1.1.

Por simplicidad para maximizar la verosimilitud, es común trabajar con el logaritmo natural de la función de verosimilitud

$$l(\theta; \boldsymbol{x}) = \ln L(\theta; \boldsymbol{x}) = \ln c(x) + \ln f(\boldsymbol{x}; \theta), \tag{1.6}$$

por lo que el logaritmo de la función de verosimilitud relativa se puede expresar como

$$r(\theta; \boldsymbol{x}) = \ln R(\theta; \boldsymbol{x}) = l(\theta; \boldsymbol{x}) - l(\hat{\theta}; \boldsymbol{x}). \tag{1.7}$$

Sin embargo, a veces no es posible encontrar el estimador máximo verosímil analíticamente, por lo que se debe recurrir a métodos numéricos. Algunos de éstos utilizan la primera y segunda derivada, con respecto al parámetro, de la función (1.6).

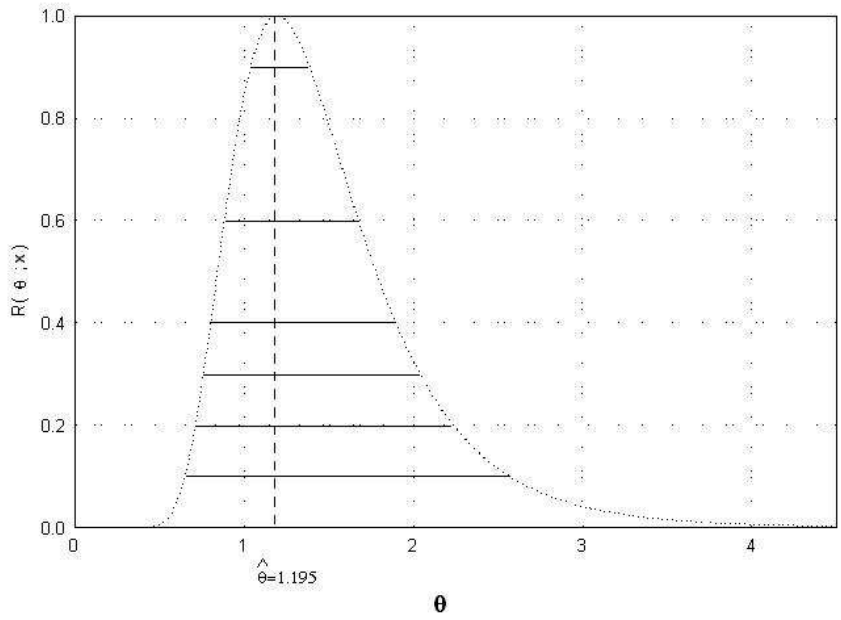


Figura 1.9: Intervalos de verosimilitud anidados correspondientes al ejemplo 1.1

La *función de puntuación* (*score function*, en inglés) se define como la primera derivada del logaritmo de la verosimilitud, con respecto a θ.

$$S(\theta; \boldsymbol{x}) = \frac{\partial l(\theta; \boldsymbol{x})}{\partial \theta}, \tag{1.8}$$

y la *información observada* $I(\theta; x)$ se define como el negativo de la segunda derivada del logaritmo de la verosimilitud

$$I(\theta; \boldsymbol{x}) = -\frac{\partial^2 l(\theta; \boldsymbol{x})}{\partial \theta^2} = -\frac{\partial S(\theta; \boldsymbol{x})}{\partial \theta}. \tag{1.9}$$

Nótese que debido a que $l(\theta; \boldsymbol{x})$ está dada por (1.6), tenemos que ni la función de puntuación ni la función de información dependen de $c(\boldsymbol{x})$.

La *función de información de Fisher*, $\mathcal{I}_\theta$, se define como el valor esperado de la función de información.

$$\mathcal{I}_\theta = E[I(\theta; \boldsymbol{x})],$$

y representa la precisión promedio que puede obtenerse en un gran número de repeticiones del experimento. Fisher (1973, Sección 57.2) señala la importancia de $\mathcal{I}_\theta$ al brindar información antes de realizar algún experimento.

Una vez realizado el experimento, la cantidad que es relevante para hacer inferencia sobre θ es la información observada evaluada en el máximo verosimil $\hat{\theta}$, a la cual generalmente también se le llama información observada, por brevedad. La notación que

usaremos, será para ésta, la siguiente:

$$I_{\hat{\theta}} = I(\hat{\theta}; \boldsymbol{x}) = - \left. \frac{\partial^2 l(\theta; \boldsymbol{x})}{\partial \theta^2} \right|_{\theta=\hat{\theta}} \equiv - \frac{\partial^2 l(\theta; \boldsymbol{x})}{\partial \hat{\theta}^2}.$$

La razón por la que las dos primeras derivadas del logaritmo de la verosimilitud de θ reciben nombres especiales es porque describen la forma de la verosimilitud alrededor del estimador máximo verosimil $\hat{\theta}$. El estimador $\hat{\theta}$ señala la posición de la verosimilitud en el eje correspondiente al parámetro y la información observada describe la curvatura de la función (o precisión) alrededor de $\hat{\theta}$.

Ejemplo 1.2 A continuación se calcula la función de información observada que se origina por el modelo y los datos del ejemplo 1.1.

El logaritmo de la función de verosimilitud del parámetro θ (también conocida como *log verosimilitud* de θ) es

$$l(\theta; \boldsymbol{x}) = -n \ln \theta - \frac{\sum (x_i - 1)}{\theta},$$

con $n = 10$. A partir de esta expresión se obtiene la función de puntuación

$$S(\theta; \boldsymbol{x}) = \frac{\partial l(\theta; \boldsymbol{x})}{\partial \theta} = -\frac{n}{\theta} + \frac{\sum x_i - n}{\theta^2},$$

de la que a su vez se puede calcular la información observada

$$I_{\hat{\theta}} = -\frac{\partial S(\theta; \boldsymbol{x})}{\partial \hat{\theta}} = \frac{2(\sum x_i - n)}{\theta^3} - \left. \frac{n}{\theta^2} \right|_{\theta=\hat{\theta}} = \frac{20(1.195)}{(1.195)^3} - \frac{10}{(1.195)^2} = \frac{10}{(1.195)^2}.$$

La función de verosimilitud tiene tres propiedades muy importantes que se enuncian brevemente a continuación. Ver Sprott (2000a, Sección 2.7).

1. **No aditividad.** La verosimilitud es una función puntual que va del espacio parametral Θ a los reales. Asigna valores de plausibilidad a cada valor del parámetro y se asocia a una hipótesis simple $H : \theta = \theta_0$. Como la unión de dos hipótesis simples $H_1 : \theta = \theta_1$ y $H_2 : \theta = \theta_2$, no es en general una hipótesis simple, no se puede, siquiera, asociar alguna verosimilitud a su unión. La probabilidad es una función que se define sobre una familia de eventos y a diferencia de la verosimilitud, ésta sí es aditiva. Esta es la diferencia más fuerte entre la verosimilitud y la probabilidad.

2. **Combinación de las observaciones.** La función de verosimilitud es capaz de combinar los datos provenientes de diferentes experimentos, siempre que sean homogéneos, de una forma muy sencilla. Debido a la expresión (1.3), los nuevos datos se incorporan al análisis simplemente multiplicando su verosimilitud a la de los datos anteriores.

3. **Invarianza funcional.** La propiedad de invarianza funcional permite hacer inferencia de cualquier parámetro que sea función 1-1 de θ, $\delta = \delta(\theta)$, a partir de la función de verosimilitud de θ por sustitución algebraica. Esto es muy conveniente ya que en diversos casos, algún otro parámetro función de θ suele ser de mayor interés que θ mismo. El cambio de parámetros también puede simplificar la forma de la función de verosimilitud (simetrizándola, por ejemplo), haciendo los cálculos de inferencia más simples, pero totalmente equivalentes en términos de θ.

Ejemplo 1.3 Para ejemplificar la propiedad de invarianza funcional de la verosimilitud, se retomará el ejemplo 1.1.

Supóngase que se quiere estimar la cantidad $\tau \geq 1$ tal que la probabilidad de que el siguiente registro ocurra en el intevalo $[1, \tau]$ es de $1/2$. Esto es equivalente a decir que τ cumple con la igualdad $\exp(-(\tau - 1)/\theta) = 1/2$. A τ se le conoce como "vida mediana".

La función

$$\tau = -\theta \ln(1/2) + 1$$

es $1-1$ en θ, entonces por la propiedad de invarianza funcional se tiene que la verosimilitud relativa de τ es igual a

$$R_\tau(\tau; \boldsymbol{x}) = R_\theta(\theta(\tau); \boldsymbol{x}),$$

donde la verosimilitud relativa de θ es igual a

$$R_\theta(\theta; \boldsymbol{x}) = \left(\frac{1.195}{\theta}\right)^{10} \exp\left(-\frac{11.95}{\theta} + 10\right).$$

Como $\theta(\tau) = -(\tau - 1)/\ln(1/2)$, entonces la expresión de la verosimilitud relativa de τ es la siguiente:

$$R_\tau(\tau; \boldsymbol{x}) = \left(-\frac{1.195 \ln(1/2)}{\tau - 1}\right)^{10} \exp\left(\frac{11.95 \ln(1/2)}{\tau - 1} + 10\right).$$

En particular

$$\hat{\tau} = -\hat{\theta} \ln(1/2) + 1 = -1.195 \ln(1/2) + 1 \approx 1.8283.$$

La gráfica de la verosimilitud relativa de τ se representa en la figura 1.10.

Ejemplo 1.4 (Kalbfleisch, 1985, ejercicio 7, p. 31) Con el siguiente ejemplo se pretende ilustrar la forma en que se pueden combinar distintos experimentos mediante la función de verosimilitud.

Un laboratorio mide la cantidad de un metal en una solución a través de un método que genera errores independientes con distribución $N(0, \sigma^2)$. Si la concentración verdadera de una solución es μ, entonces las concentraciones registradas, X, provienen de una población con distribución $N(\mu, \sigma^2)$. Con el objetivo de estimar σ, se realizaron varias mediciones de soluciones cuya concentración de metal, μ, es conocida.

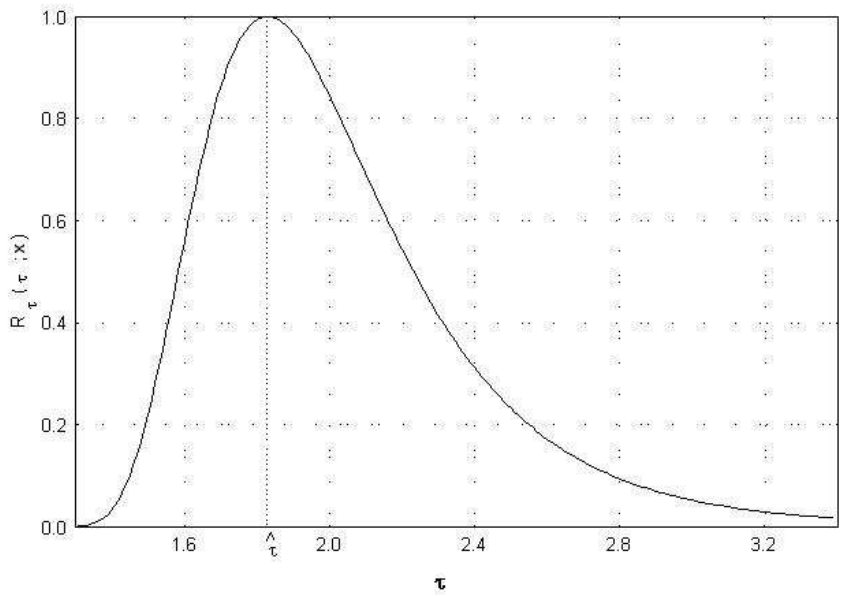


Figura 1.10: Función de verosimilitud relativa de τ

1. Las cinco mediciones, $\boldsymbol{x} = \{9.3, 11.2, 8.7, 10.1, 10.7\}$, se realizaron con una solución que contenía concentración $\mu_1 = 10$.

2. Otras cinco mediciones, $\boldsymbol{y} = \{21.7, 19.9, 20.3, 20.4, 19.7\}$, fueron realizadas con otra solución que contenía concentración $\mu_2 = 20$.

La función de verosimilitud originada por el primer conjunto de cinco observaciones, $\boldsymbol{x}$, es

$$L(\sigma; \boldsymbol{x}) = \frac{1}{\sigma^5} \exp\left(-\sum_{i=1}^{5} \frac{(x_i - \mu_1)^2}{2\sigma^2} \right),$$

con $c(\boldsymbol{x}) = (2\pi)^{5/2}$. De la función de verosimilutud se deduce que el estimador máximo verosímil de σ, en base a los datos $\boldsymbol{x}$, es

$$\hat{\sigma}_{\boldsymbol{x}} = \sqrt{\frac{\sum (x_i - \mu_1)^2}{5}} = \sqrt{\frac{\sum (x_i - 10)^2}{5}} \approx 0.9077.$$

Luego, entonces la función de verosimilitud relativa de σ, cuya gráfica se presenta en la figura 1.11, es igual a

$$R(\sigma; \boldsymbol{x}) = \left(\frac{\sum (x_i - 10)^2}{5\sigma^2} \right)^{5/2} \exp\left(-\frac{\sum (x_i - 10)^2}{2\sigma^2} + \frac{5}{2} \right) = \left(\frac{4.12}{5\sigma^2} \right)^{2.5} \exp\left(-\frac{4.12}{2\sigma^2} + 2.5 \right).$$

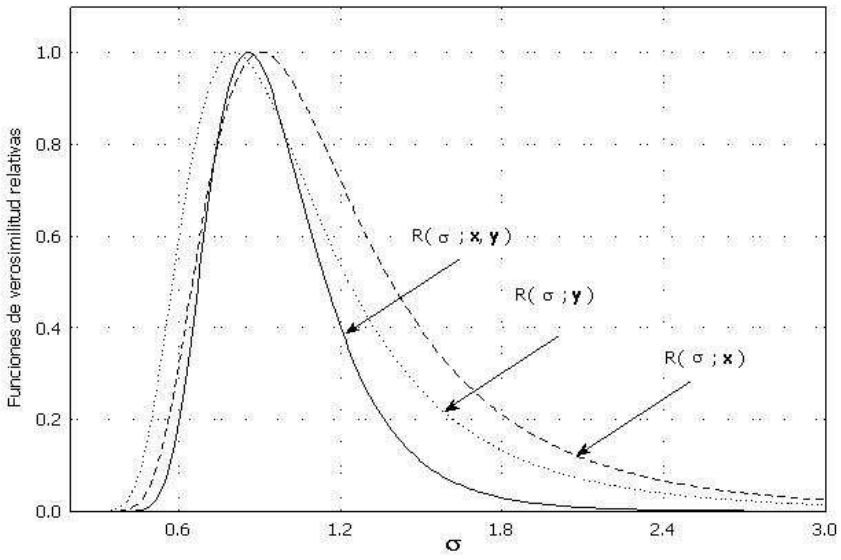


Figura 1.11: Funciónes de verosimilitud relativa de σ originadas por los datos $\boldsymbol{x}$, $\boldsymbol{y}$ y ($\boldsymbol{x}$,$\boldsymbol{y}$)

De forma similar también se puede calcular el estimador $\hat{\sigma}_{\boldsymbol{y}} \approx 0.8050$ y la función de verosimilitud relativa de σ a partir de los datos $\boldsymbol{y}$,

$$R(\sigma;\boldsymbol{y}) = \left(\frac{\sum(y_i-20)^2}{5\sigma^2}\right)^{5/2} \exp\left(-\frac{\sum(y_i-20)^2}{2\sigma^2}+2.5\right) = \left(\frac{3.24}{5\sigma^2}\right)^{2.5} \exp\left(-\frac{3.24}{2\sigma^2}+2.5\right).$$

A partir de la forma de las curvas descritas por las funciones de verosimilitudes relativas de σ generadas por los datos $\boldsymbol{x}$ y $\boldsymbol{y}$ se puede afirmar que los conjuntos de datos provenientes del primer y segundo conjunto de mediciones, son homogéneos y que efectivamente se puede hacer inferencia sobre σ a partir de los dos experimentos juntos. Esta afirmación se basa en algunas características de las curvas como las siguientes: las gráficas de $R(\sigma;\boldsymbol{x})$ y $R(\sigma;\boldsymbol{y})$ son de formas similares (en este caso ligeramente asimétricas y con colas levemente más pesadas hacia la derecha), los estimadores máximo verosimiles $\hat{\sigma}_{\boldsymbol{x}}$ y $\hat{\sigma}_{\boldsymbol{y}}$ no difieren por mucho y las curvas se traslapan fuertemente aun para verosimilitudes pequeñas. Estas características sirven para determinar si los dos conjuntos de datos están dando información sobre la misma cantidad o parámetro y así saber si es válido combinar los dos experimentos mediante las verosimilitudes. Estas características de la gráfica de la verosimilitud se resumirán al final de esta subsección.

Como los dos conjuntos de datos son homogéneos, se puede realizar el análisis del parámetro σ utilizando simultáneamente todos los datos. En este caso, la función de verosimilitud a considerar es igual a

$$L(\sigma;\boldsymbol{x},\boldsymbol{y}) = L(\sigma;\boldsymbol{x})\cdot L(\sigma;\boldsymbol{y}) = \frac{1}{\sigma^{10}}\exp\left(-\frac{\sum(x_i-10)^2+\sum(y_i-20)^2}{2\sigma^2}\right),$$

debido a la propiedad 2 de la verosimilitud descrita en esta subsección.

El estimador máximo verosímil de σ obtenido a partir de $L(\sigma; \boldsymbol{x}, \boldsymbol{y})$ es

$$\hat{\sigma}_{\boldsymbol{x},\boldsymbol{y}} = \sqrt{\frac{\sum(x_i - 10)^2 + \sum(y_i - 20)^2}{10}} \approx 0.8579.$$

La gráfica de la función de verosimilitud relativa obtenida de la combinación de los datos también se presenta en la figura 1.11, y su expresión es la siguiente:

$$\begin{aligned} R(\sigma; \boldsymbol{x}, \boldsymbol{y}) &= \left(\frac{\sum(x_i - 10)^2 + \sum(y_i - 20)^2}{10\sigma^2}\right)^5 \exp\left(-\frac{\sum(x_i - 10)^2 + \sum(y_i - 20)^2}{2\sigma^2} + 5\right) \\ &= \left(\frac{7.3599}{10\sigma^2}\right)^5 \exp\left(-\frac{7.3599}{2\sigma^2} + 5\right). \end{aligned}$$

Como se puede observar en la figura 1.11, el considerar simultámeneamente los datos $\boldsymbol{x}$ y $\boldsymbol{y}$, hace que la función de verosimilitud relativa obtenida sea más informativa que las anteriores, $R(\sigma; \boldsymbol{x})$ y $R(\sigma; \boldsymbol{y})$. Este hecho se refleja directamente en la gráfica de la verosimilitud relativa, al hacer que la curva de $R(\sigma; \boldsymbol{x}, \boldsymbol{y})$ sea más angosta, ya que así origina intervalos de verosimilutud de longitudes menores a los obtenidos a partir de $R(\sigma; \boldsymbol{x})$ o $R(\sigma; \boldsymbol{y})$.

En el ejemplo anterior fue necesario realizar una afirmación sobre la homogeneidad de dos grupos de datos para poder realizar la combinación de la información. Las pruebas de homogeneidad pueden formalizarse a partir de las funciones de verosimilitud, pero en algunos casos basta con observar directamente las gráficas de las funciones de verosimilitud relativa para determinar si experimentos diferentes describen el mismo parámetro de interés. Los criterios que se utilizan para saber si los experimentos son homogéneos son los siguientes:

1. Forma de la verosimilitud. Cuando los experimentos son homogéneos, se espera que la forma de la verosimilitud sea similar (simétrica, asimétrica hacia la derecha o asimétrica hacia la izquierda)

2. Localización de los estimadores máximo verosímiles. En caso de que los experimentos describan al mismo parámetro, se espera que los estimadores originados por cada uno no difieran mucho.

3. Precisión para estimar el parámetro. La apertura de la función de verosimilitud se relaciona con la información que contiene del parámetro y esta generalmente disminuye cuando el número de observaciones crece. En caso de que los experimentos contengan la misma cantidad de observaciones, se espera que la apertura o el grosor de la función de verosimilitud relativa no sea muy diferente, sobre todo cuando el conjunto de datos que se observaron para cada experimento es relativamente grande.

4. Traslape de las funciones de verosimilitud relativa. El traslape de las funciones de verosimilitud se relaciona fuertemente con los tres aspectos anteriores, es decir que si las formas de las curvas, los estimadores máximo verosmiles y la apertura de las funciones son similares, entonces la region de traslape de las curvas es muy grande. Este traslape puede definirse en base de los intervalos de verosmilitud generados por cada experimento. Por ejemplo si los intervalos del 15% se traslapan en más de la mitad de su longitud, se puede decir que efectivamente las curvas se pueden considerar fuertemente traslapadas o coincidentes.

1.3.2 La verosimilitud perfil

Cuando el espacio parametrico Θ es subconjunto de $\mathbf{R}^k$, pueden existir elementos del vector $\boldsymbol{\theta} = \theta_1, \ldots, \theta_k$ que no son de interés ($\boldsymbol{\theta}_R$, $R \subset \{1, \ldots, k\}$) y el considerar a todos nos produce problemas de estimación conjunta. Los parámetros que no son de interés se les llama "parámetros de ruido" o de "estorbo". Para facilitar la inferencia sobre los parámetros que sí son de interés ($\boldsymbol{\theta}_T$, $T = \{1, \ldots, k\} - R$) se han desarrollado diversas metodologías que buscan estimar los parámetros en la presencia de parámetros desconocidos.

El problema de realizar inferencia sobre los parámetros de interés, en ausencia de conocimiento de los parámetros de ruido, está aún vigente y no tiene solución única.

Debido a la propiedad de no aditividad de la verosimilitud, los parámetros de ruido no se pueden separar de la función de verosimilitud integrando sobre los que se quiere eliminar, como puede realizarse con las probabilidades y funciones de densidad, Sprott (2000a, p. 50).

Diversos métodos se han implementado con el fin de poder resolver el problema de estimación por separado que hemos mencionado. Algunos se exponen en Kalbfleisch y Sprott (1970) y en Sprott (2000a, Capítulo 4). A continuación se explicará exclusivamente uno de ellos que se basa en la *verosimilitud perfil* o *maximizada*.

Uno de los procedimientos para hacer inferencia sobre parámetros de interés, que es muy poco restrictivo (Sprott, 2000a, p. 66) es la verosimilitud perfil o maximizada,

$$L_{\max}(\boldsymbol{\theta}_T; \boldsymbol{x}) \propto f(\boldsymbol{x}; \boldsymbol{\theta}_T, \hat{\boldsymbol{\theta}}_R(\boldsymbol{\theta}_T)),$$

donde $\hat{\boldsymbol{\theta}}_R(\boldsymbol{\theta}_T)$ es el estimador máximo verosímil del vector con los parámetros de estorbo, $\boldsymbol{\theta}_R$, restringido a un valor específico de los parámetros de interés, $\boldsymbol{\theta}_T$.

La función de verosimilitud relativa perfil se define como

$$R_{\max}(\boldsymbol{\theta}_T; \boldsymbol{x}) = \frac{f(\boldsymbol{x}; \boldsymbol{\theta}_T, \hat{\boldsymbol{\theta}}_R(\boldsymbol{\theta}_T))}{f(x; \hat{\boldsymbol{\theta}}_T, \hat{\boldsymbol{\theta}}_R)}.$$

Nótese que el estimador máximo verosímil perfil de $\hat{\boldsymbol{\theta}}_T$ coincide con el estimador máximo verosímil global, es decir $\hat{\boldsymbol{\theta}}_R(\hat{\boldsymbol{\theta}}_T) = \hat{\boldsymbol{\theta}}_R$, es el estimador máximo verosimil de $\boldsymbol{\theta}_R$ sin restringir.

Si el parámetro de interés es un escalar ($\theta_T = \theta'$) y el vector de parámetros es igual a ($\boldsymbol{\theta} = (\theta', \boldsymbol{\theta_R})$), entonces Sprott (2000a, p. 183) expone la siguiente igualdad para la información observada perfil de $\theta_T = \theta' = \theta_1$:

$$I_{\hat{\theta}_1} = -\frac{\partial^2}{\partial \hat{\theta}_1^2} \ln L_{\max}(\theta_1; \boldsymbol{x}) = \frac{1}{I^{\theta_1\theta_1}}, \tag{1.10}$$

donde $I^{\theta_1\theta_1} = I^{11}$ es el primer elemento en la matriz inversa de la matriz de información observada $I_{(\theta_1, \boldsymbol{\theta}_R)}$. En Sprott (2000a, Apéndice 9.A.2) se demuestra la igualdad anterior para el caso en el que $\boldsymbol{\theta}_R$ es también un escalar.

Ejemplo 1.5 Supóngase que se tienen n observaciones $x_1, \ldots, x_n$ de una población con distribución $N(\mu, \sigma^2)$, de parámetros μ y σ desconocidos, y se quiere calcular la función de verosimilitud perfil y la información observada de σ^2. En este caso $\boldsymbol{\theta} = (\sigma, \mu)$, el parámetro de interés es $\theta_1 = \sigma$ y el de estorbo, $\theta_2 = \mu$. A partir de la función de verosmilitud $L(\boldsymbol{\theta}; \boldsymbol{x}) = L(\sigma, \mu; \boldsymbol{x})$ se calcula, a continuación, el estimador máximo verosímil de μ, en términos de σ, $\hat{\mu}(\sigma)$. La función de verosimilitud es igual a

$$L(\sigma, \mu; \boldsymbol{x}) = \frac{1}{\sigma^n} \exp\left(-\frac{\sum(x_i - \mu)^2}{2\sigma^2}\right), \quad \text{entonces} \quad \frac{\partial l(\sigma, \mu; \boldsymbol{x})}{\partial \mu} = \frac{\sum(x_i - \mu)}{\sigma^2},$$

y la expresión

$$\frac{\partial l(\sigma, \mu; \boldsymbol{x})}{\partial \mu} = 0 \ \text{ si y solo si } \ \mu = \frac{\sum x_i}{n}.$$

Como $\sum x_i/n$ es el argumento que maximiza a $L(\sigma, \mu; \boldsymbol{x})$ para un valor fijo de μ, entonces $\hat{\mu}(\sigma) = \sum x_i/n \equiv \bar{x}$.

Una vez determinada la expresión de $\hat{\mu}(\sigma)$ se puede obtener la verosimilitud perfil

$$L_{\max}(\sigma; \boldsymbol{x}) = \frac{1}{\sigma^n} \exp\left(-\frac{\sum(x_i - \bar{x})^2}{2\sigma^2}\right)$$

de donde se calcula el estimador máximo verosimil de σ.

$$\frac{\partial \ln L_{\max}(\sigma; \boldsymbol{x})}{\partial \mu} = \frac{\partial l_{\max}(\sigma; \boldsymbol{x})}{\partial \mu} = -\frac{n}{\sigma} + \frac{\sum(x_i - \bar{x})}{n} = 0 \Longleftrightarrow \hat{\sigma} = \sqrt{\frac{\sum(x_i - \bar{x})}{n}}$$

La función de verosimitud relativa perfil tiene entonces la siguiente expresión:

$$R_{\max}(\sigma; \boldsymbol{x}) = \left(\frac{\sum(x_i - \bar{x})^2}{n\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum(x_i - \bar{x})^2}{2\sigma^2} + \frac{n}{2}\right).$$

La información observada perfil se puede calcular en este ejemplo, fácilmente a partir de la expresión, con que ya se cuenta, $\partial l_{\max}(\sigma; \boldsymbol{x})/\partial \mu$.

$$I_{\hat{\sigma}} = -\frac{\partial l_{\max}(\sigma; \boldsymbol{x})^2}{\partial \hat{\sigma}^2} = -\frac{\partial}{\partial \hat{\sigma}}\left(-\frac{n}{\sigma} + \frac{\sum(x_i - \bar{x})}{n}\right) = \frac{2n^2}{\sum(x_i - \bar{x})^2}.$$

1.3.3 La verosimilitud pivotal

Los intervalos de verosimilitud dan una medida de plausibilidad para el parámetro, y por este hecho pueden ser por sí mismos de interés; sin embargo no brindan una medida de incertidumbre sobre la frecuencia con la que en experimentos repetidos contendrán al verdadero valor del parámetro θ. Para poder calcular las probabilidades de que los intervalos contengan el verdadero valor de θ, se utilizarán las llamadas *cantidades pivotales*.

Fisher (1948) definió un pivotal $u(\boldsymbol{x};\theta)$ como una función del parámetro y de los datos cuya distribución no depende de parámetros desconocidos, por lo que se puede conocer numéricamente. La importancia de los pivotales radica en que son la base para asociar un nivel de confianza a un intervalo o región de verosimilitud. Ver Sprott (2000a, pp. 73-74).

Una región $U \subset \Theta$ del $100(1-\alpha)\%$ de confianza ($\alpha \in (0,1)$) para θ, puede originarse a partir de un pivotal u, pero hay que elegir un criterio adicional para seleccionarla, ya que no es única, sino que existe un número infinito de regiones con el mismo nivel de confianza. Dicho criterio se puede basar en la función de verosimilitud, Sprott (2000a, pp. 163-165) y Kalbfleisch (1985, p. 114), pidiendo que el intervalo de $100(1-\alpha)\%$ de confianza también sea de verosimilitud, así la idea es que el conjunto de intervalos de confianza anidados reproduzcan la función de verosimilitud en el sentido de que para cada $a \in [0,1]$ el intervalo de verosimilitud al $100a\%$ y el de confianza obtenido por el pivotal, sean aproximadamente los mismos (en la figura 1.9 se dibujaron los intervalos de verosimilitud). Dichos intervalos de denominan usualmente como de *verosimilitud-confianza*. Ver ejemplos en Sprott y Viveros (1984) y Viveros y Sprott (1987).

El intervalo de verosimilitud-confianza permite conocer así, las propiedades de confianza o frecuencia del intervalo o región, en el sentido de que si el experimento se repitiera muchas veces, el nivel de confianza indicaría aproximadamente la proporción de veces que el intervalo de verosimilitud incluye al verdadero valor del parámetro. Además por ser también intervalo de verosimilitud, da a conocer los atributos de verosimilitud de cualquier valor del parámetro, como se ha señalado en la Subsección 1.3.1.

La definición 1.2 nos dice que la verosimilitud es proporcional a la probabilidad de una cantidad observada $\boldsymbol{x}$. La *verosimilitud pivotal* a diferencia de la verosimilitud, es proporcional a la probabilidad conocida $g[u(\boldsymbol{x};\theta)]$ de un pivotal $u(\boldsymbol{x};\theta)$ que es en general no observable, ya que su expresión involucra al parámetro desconocido.

Un *pivotal lineal* en θ puede expresarse como

$$u_\theta = u(\boldsymbol{x};\theta) = \frac{c(\boldsymbol{x}) - \theta}{d(\boldsymbol{x})}, \tag{1.11}$$

donde $c(\boldsymbol{x})$ y $d(\boldsymbol{x}) > 0$ no involucran a θ. Como el diferencial $du/d\theta$ no depende de θ entonces la transformación no cambia a la verosimilitud y es cierta la relación

$$L_u(\theta;\boldsymbol{x}) \propto g[u(\boldsymbol{x};\theta)], \tag{1.12}$$

por el Teorema de Cambio de Variable.

La verosimilitud pivotal podrá definirse sólamente considerando pivotales lineales como en (1.12), donde $g[u(\boldsymbol{x};\theta)]$ se evalua en $\boldsymbol{x}$ y se considera como función de θ.

La función de verosimilitud relativa inducida por el pivotal lineal $u(\boldsymbol{x};\theta)$ se define como

$$R_u(\theta) = \frac{L_u(\theta)}{\sup_\theta L_u(\theta)}.$$

Ejemplo 1.6 Considere la siguiente cantidad pivotal: $u_\theta = (\hat{\theta} - \theta)\sqrt{I_{\hat{\theta}}}$ que es lineal en θ. Si $u_\theta \sim N(0,1)$ entonces la verosimilitud de u_θ es normal, esto es:

$$f(u_\theta) \propto \exp\left(-\frac{u_\theta^2}{2}\right),$$

y la función de verosimilitud relativa inducida por u es

$$R_{u_\theta}(\theta) \propto \exp\left(-\frac{u_\theta^2}{2}\right) = \exp\left(-\frac{(\hat{\theta}-\theta)^2 I_{\hat{\theta}}}{2}\right).$$

1.3.4 Aproximaciones a la verosimilitud

Si se puede encontrar un pivotal lineal $u_\theta = u(\boldsymbol{x};\theta)$ con densidad $g(u)$ que induzca a una verosimilitud aproximadamente igual a la original, entonces $R_{u_\theta}(\theta)$ y $g(u)$ pueden utilizarse para obtener intervalos de verosimilitud-confianza de θ.

Barnard (1977) define al pivotal $u_\theta = u(\boldsymbol{x};\theta)$ como *eficiente* si la función de verosimilitud originada por éste, aproxima bien a la función de verosimilitud observada

$$R_u(\theta) \approx R(\theta;\boldsymbol{x}). \tag{1.13}$$

Si la aproximación (1.13) se cumple, entonces cada intervalo de verosimilitud de θ hereda la propiedad de cobertura frecuentista correspondiente al intervalo en u_θ, dado por $g(u_\theta)$ y así se obtienen intervalos de verosimilitud-confianza. En este caso, por simplicidad, se puede usar $g(u_\theta)$ para obtener los intervalos de verosimilitud-confianza aproximados en lugar de $r(\theta;\boldsymbol{x})$ la cual puede tener una expresión más complicada.

En la expansión de Taylor de $r(\theta)$ alrededor de $\hat{\theta}$ el pivotal $u_\theta = (\hat{\theta} - \theta)\sqrt{I_{\hat{\theta}}}$ aparece de manera natural. Si esta serie de potencias se reescriben en términos de u_θ, lo que se obtiene es la expansión de Taylor de $r(u_\theta)$ alrededor del cero.

$$\begin{aligned} r(u_\theta) &= -\frac{u_\theta^2}{2} - \frac{u_\theta^3}{3!}F_3 + \frac{u_\theta^4}{4!}F_4 + \cdots \\ &= -\frac{u_\theta^2}{2}\left\{1 + \frac{u_\theta}{3}F_3 - \frac{u_\theta^2}{12}F_4 + \cdots\right\}, \end{aligned} \tag{1.14}$$

donde

$$F_i = \frac{\partial^i r(\theta)}{\partial \hat{\theta}^i} I_{\hat{\theta}}^{-i/2}. \tag{1.15}$$

La verosimilitud se puede resumir entonces mediante las cantidades $\hat{\theta}, I_{\hat{\theta}}$ y F_i, las cuales, a su vez describen la localización, la dispersión y la forma de la verosimilitud, respectivamente. En particular F_3 describe la simetría alrededor de $\hat{\theta}$ de la función de verosimilitud y F_4 brinda información sobre el grosor de las colas de ésta. Estas interpretaciones de F_i son válidas únicamente cuando se realiza la expansión alrededor del valor $\hat{\theta}$. Ver Viveros y Sprott (1987).

Es por esto que el candidato natural para ser el pivotal lineal que ayude a calcular los intervalos de verosimilitud-confianza, es precisamente u_θ. En la práctica, el pimer paso para seleccionar la distribución que tiene u_θ, $g(u_\theta)$, se realiza examinando visualmente la gráfica de la función de verosimilitud relativa. Esta selección se corrobora graficando simultáneamente $r(\theta)$ y la verosimilitud pivotal $r_{u_\theta}(\theta)$ inducida por la $g(u_\theta)$ seleccionada. Si ambas curvas difieren visualmente, entonces esto sugiere que la función $g(u_\theta)$ seleccionada no es adecuada para describir a la verosimilitud observada.

Si consideramos sólamente los primeros cuatro términos de la expansión de Taylor (1.14), entonces en general hay cuatro posibles casos que pueden ocurrir y se describen en detalle en Díaz-Francés (1998, pp. 18-20). Aquí se mencionan brevemente.

1. La verosimilitud observada es asimétrica alrededor de $\hat{\theta}$ pero se puede encontrar una reparametrización $\varphi = g(\theta)$ que la simetrice. Puede ser que si suponemos que u_φ tiene distribución normal, entonces la verosimilitud pivotal normal r_{u_φ} aproxime bien a la verosimilitud observada cuya expansión de Taylor es (1.14). En ese caso la verosimilitud aproximada es igual a

$$r(u_\varphi) = -u_\varphi^2/2. \tag{1.16}$$

2. La verosimilitud observada es simétrica alrededor de $\hat{\theta}$, pero tiene colas más pesadas que la verosimilitud normal (F_3 despreciable y $0 \leq F_4 < 6$), Viveros y Sprott (1987) sugieren considerar la cantidad pivotal

$$u_t = (\hat{\theta} - \theta)\sqrt{I_{\hat{\theta}}\left(\frac{a}{a+1}\right)}, \tag{1.17}$$

cuya distribución es una t de Student con a grados de libertad, donde $a = (6/F_4) - 1$.

Para confirmar que la verosimilitud pivotal R_{u_t} aproxime bien a la verosimilitud observada, hay que verificarlo visualmente.

3. La verosimilitud observada es asimétrica alrededor de $\hat{\theta}$ y $F_3^2 + F_4 \geq 0$. Viveros y Sprott (1987) sugieren considerar la siguiente cantidad pivotal con distribución

$\log F_{(a,b)}$,

$$u_{\log F} = (\hat{\theta} - \theta)\sqrt{I_{\hat{\theta}}\left(\frac{2}{a} + \frac{2}{b}\right)}, \tag{1.18}$$

donde $a = \frac{4}{q(q+F_3)}$, $b = \frac{4}{q(q+F_3)}$ y $q = \sqrt{3F_3^2 + 2F_4}$.

Nuevamente, se verifica visualmente que la verosimilitud inducida por este pivotal $\log F$ aproxime bien a la verosimilitud observada.

4. Ningúno de los casos anteriores se cumple. Hay que considerar a la función de verosimilitud completa o intentar encontrar otro pivotal lineal en el parámetro, con otra distribución, posiblemente en la familia hiperbólica de Barndorff-Nielsen o del Nilo (Chamberlin, 1989). En el caso en que no pueda encontrarse un pivotal lineal que origine a una función de verosimilitud aproximada a la observada, siempre pueden darse intervalos de verosimilitud y encontrar una cota inferior para la frecuencia de cobertura del parámetro, a través de simulaciones.

1.4 Mezclas de distribuciones y el algoritmo EM

En general, se definirá como una *densidad de mezclas* a la función que cumple la igualdad

$$f(x) = \int g(x;\boldsymbol{\theta})dH(\boldsymbol{\theta}), \tag{1.19}$$

donde $g(x;\boldsymbol{\theta})$ es una densidad de probabilidad cuyo parámetro $\boldsymbol{\theta}$, tiene función de acumulación H.

Para los fines de este trabajo, bastará con considerar a las densidades de mezclas denominadas *mezclas finitas de distribuciones*, que consisten en las mezclas de distribuciones en las que Q es una distribución con soporte finito. Entonces la definición dada en (1.19) se puede expresar como

$$f(x) = \sum_{j=1}^{c} g_j(x, \boldsymbol{\theta}_j)Q(\boldsymbol{\theta}_j),$$

donde $\boldsymbol{\theta}_j$ está en el soporte de Q para $j = 1\ldots, c$ y g_j es densidad con parámetro (de dimensión k_j) $\boldsymbol{\theta}_j$. Usualmente $Q_j(\boldsymbol{\theta}_j)$ se expresa como p_j y se le llama como proporción de mezcla. La densidad de mezclas se expresa entonces como

$$f(x;\boldsymbol{\theta}) = \sum_{j=1}^{c} g_j(x, \boldsymbol{\theta}_j)p_j. \tag{1.20}$$

Ver Everitt y Hand (1981, Sección 1.3).

McLachlan y Basford (1988) señalan que el modelo de mezclas de distribuciones, en particular el de mezclas normales, se usa frecuentemente, ya que en la práctica existen muchos datos que surgen de la mezcla (en alguna proporción) de poblaciones. Un ejemplo

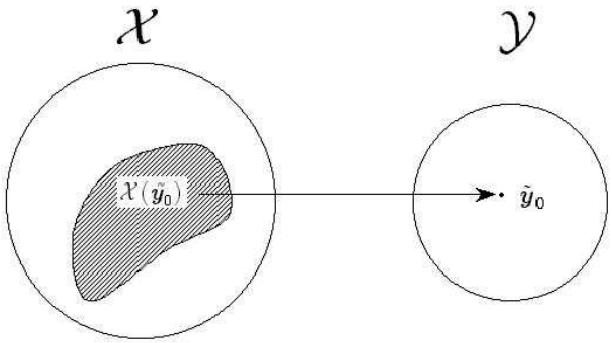


Figura 1.12: Diagrama de la relación entre los elementos que intervienen en el concepto de datos incompletos

simple de una mezcla de dos densidades son las estaturas de personas adultas en un grupo compuesto de hombres y mujeres. Otro ejemplo es el contenido de ADN en células de hígado de ratas, porque el contenido de ADN es diferente, dependiendo de la fase del ciclo de vida por el que cada célula este transitando, Gregor (1969).

Los parámetros de la expresión (1.20) se pueden dividir en dos tipos, uno contiene solamente a c, que es el número de componentes de la mezcla, y el otro comprende a las proporciones de mezcla, p_j, y los vectores $\boldsymbol{\theta}_i$. Esta división obedece a los dos diferentes contextos bajo los cuales se pueden estudiar las mezclas. El primero no considera una estructura de mezcla *a priori*, y desea identificar conglomerados en los datos. Este enfoque se trata en el trabajo de McLachlan y Basford (1988). El segundo identifica el modelo de mezcla, supone que es finita y con c conocida, y lo que intenta es estimar los parámetros de mezcla y los de las distribuciones.

En este trabajo nos inclinaremos por el segundo enfoque. Se supondrá que el número de elementos en la mezcla es conocido y se estimará los p_j's y $\boldsymbol{\theta_j}$'s.

Como el principal problema de una mezcla es identificar cuál de las distribuciones componentes origina un dato en particular, entonces una mezcla de distribuciones puede verse como un problema de datos incompletos.

El término de *datos incompletos* implica, en general la existencia de dos espacios muestrales $\mathcal{X}$ y $\mathcal{Y}$ y de un mapeo (muchos a uno) de $\mathcal{X}$ a $\mathcal{Y}$ ($\tilde{\boldsymbol{x}} \to \tilde{\boldsymbol{y}}(\tilde{\boldsymbol{x}})$). En lugar de observar los "datos completos" $\tilde{\boldsymbol{x}}$ en $\mathcal{X}$, se observan los "datos incompletos" $\tilde{\boldsymbol{y}}(\tilde{\boldsymbol{x}})$ en $\mathcal{Y}$. Ver figura 1.12. Si la función de densidad de $\tilde{\boldsymbol{x}}$ es $h(\tilde{\boldsymbol{x}}, \theta)$, la densidad de $\tilde{\boldsymbol{y}}$ está dada por

$$f(\tilde{\boldsymbol{y}}; \theta) = \int_{\mathcal{X}(\tilde{\boldsymbol{y}})} h(\tilde{\boldsymbol{x}}, \theta) d\tilde{\boldsymbol{x}},$$

donde $\mathcal{X}(\tilde{\boldsymbol{y}}) = \{\tilde{\boldsymbol{x}} \in \mathcal{X} : \tilde{\boldsymbol{y}}(\tilde{\boldsymbol{x}}) = \tilde{\boldsymbol{y}}\}$. Es decir, $\mathcal{X}(\tilde{\boldsymbol{y}})$ es el subconjunto de $\mathcal{X}$ que consta de todos los posibles valores que pudieron tomar los datos faltantes dados los observados $\tilde{\boldsymbol{y}}$.

Si $\boldsymbol{x} = x_1, \ldots, x_n$ son observaciones de una mezcla, entonces la expresión (1.20) puede escribirse como

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} \left(\prod_{j=1}^{c} g_j(x_i; \boldsymbol{\theta}_j)^{\epsilon_{ij}} \right), \tag{1.21}$$

donde $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{ic})$ es un vector indicador cuyos componentes son todos ceros, excepto uno, que es el que indica cuál distribución originó la observación x_i. En la última expresión de (1.21) se puede observar que $\tilde{\boldsymbol{y}} = \boldsymbol{x}$ y que los datos no observados, son $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \ldots \boldsymbol{\epsilon}_n)$.

De los métodos que abordan el problema de estimación de parámetros de un modelo de mezclas, el más eficiente es el de máxima verosimilitud. Bajo este enfoque Dempster, et al. (1977) retomaron ideas ya existentes en la literatura y las formalizaron proponiendo el procedimiento iterativo llamado *algoritmo EM*, que permite calcular los estimadores máximo verosímiles de datos incompletos de manera sencilla. Comprende básicamente dos pasos, a los cuales debe su nombre, el del cálculo de esperanzas (E) y el de maximización (M). La idea central del procedimiento es reemplazar la parte de los datos no observada con el valor esperado (E) bajo el modelo en consideración, para después maximizar (M) la verosimilitud de los datos completados. El paso E se repite ahora considerando el modelo con valores de los parámetros igual a los recien estimados y nuevamente se sigue el paso M. El ciclo finaliza cuando, bajo un criterio, se llega a convergencia.

El trabajo de Dempster, et al. (1977) tiene el mérito, tanto de identificar las áreas de aplicación en estadística, como de demostrar algunas propiedades teóricas del algoritmo EM. Jeff Wu (1983) menciona más propiedades de convergencia del algoritmo.

A grandes rasgos el algoritmo EM se compone de los siguientes pasos:

1. Utilizando un criterio razonable, se asignan valores iniciales a los parámetros.

2. Los valores faltantes se reeemplazan con su valor esperado condicionado a los valores observados y al valor actual de los parámetros.

3. El valor de los parámetros se reestima usando los valores observados y el valor esperado actualizado de los datos faltantes.

4. Se repiten los pasos 2 y 3 hasta que se cumple un criterio de paro.

Para el caso de mezclas de distribuciones, el algoritmo EM se puede resumir en los siguientes procedimientos (para mayores detalles del algoritmo EM aplicado a mezclas finitas de distribuciones, normales y Gumbel, ver Everitt y Hand, 1981 o Díaz Francés, 1985).

1. **Paso E**. Con valor inicial de $t = 1$, calcular las siguientes probabilidades *a posteriori* para cada observación i y componente j

$$P[j|x_i]^{(t)} = \frac{g_j(x_i; \boldsymbol{\theta}_j^{(t)}) p_j^{(t)}}{f(x_i; \boldsymbol{\theta}^{(t)})},\ j = 1, \ldots, c, \quad i = 1, \ldots, n. \tag{1.22}$$

Nótese que cada observación tiene c probabilidades *a posteriori* en este paso.

2. **Paso M**. Resolver las ecuaciones siguientes en términos de las $\boldsymbol{\theta}_j$'s para obtener las que maximizan a la verosimilitud completada, $\hat{\boldsymbol{\theta}}_j^{(t+1)}$.

$$\sum_{i=1}^{n} P[j|x_i]^{(t)} \frac{\partial}{\partial \theta_{jl}} \ln[g_j(x_i; \boldsymbol{\theta}_j)] = 0,\ j = 1, \ldots, c,\ l = 1, \ldots, k_j, \tag{1.23}$$

y usar las $P[j|x_i]^{(t)}$'s obtenidas en el paso 1 para estimar $p_j^{(t+1)}$ como

$$p_j^{(t+1)} = \frac{\sum_{i=1}^{n} P[j|x_i]^{(t)}}{n},\ j = 1, \ldots, c. \tag{1.24}$$

3. Regresar al paso 1 y repetir el ciclo hasta que se cumpla el criterio de paro

Ejemplo 1.7 (Mezcla de una lognormal trasladada y tres normales) A la distribución que llamamos lognormal trasladada también se le conoce con el nombre de distribución de tiempo de vida lognormal garantizado.

Considérese el modelo (1.20) con $c = 4$,

$$g_j(x, \mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right\}, \quad j = 1, 2, 3, \quad i = 1, \ldots, n, \quad x \in \mathbf{R}$$

y

$$g_4(x, \mu_4, \sigma_4^2) = \frac{1}{(x-p)\sqrt{2\pi\sigma_4^2}} \exp\left\{-\frac{(\ln(x-p)-\mu_4)^2}{2\sigma_4^2}\right\}, \quad p > 0, \quad i = 1, \ldots, n, \quad x > p.$$

La función de densidad dada por $g_4(x, \mu_4, \sigma_4^2)$ corresponde a una lognormal trasladada p unidades. La cantidad p es el rezago o el valor de garantía, y es una cantidad conocida.

Para encontrar la solución a las ecuaciones dadas en (1.23), se procede a resolver primero para las μ_j's. Cuando $j = 1, 2, 3$, tenemos

$$\begin{aligned}
&\sum_{i=1}^{n} P[j|x_i]^{(t)} \frac{\partial}{\partial \mu_j} \ln[g_j(x_i; \mu_j, \sigma_j^2)] = \\
&\quad = \sum_{i=1}^{n} P[j|x_i]^{(t)} \frac{\partial}{\partial \mu_j} \left(-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma_j^2) - \frac{(x_i-\mu_j)^2}{2\sigma_j^2}\right) \\
&\quad = \sum_{i=1}^{n} P[j|x_i]^{(t)} \frac{(x_i-\mu_j)}{\sigma_j^2} = 0 \\
\iff &\qquad \mu_j = \frac{\sum_{i=1}^{n} P[j|x_i]^{(t)} x_i}{\sum_{i=1}^{n} P[j|x_i]^{(t)}}.
\end{aligned}$$

Entonces se tiene que

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^n P[j|x_i]^{(t)} x_i}{\sum_{i=1}^n P[j|x_i]^{(t)}}. \tag{1.25}$$

Cuando j es igual a cuatro, se tiene que

$$\begin{aligned}
&\sum_{i=1}^n P[4|x_i]^{(t)} \frac{\partial}{\partial \mu_4} \ln[g_4(x_i; \mu_4, \sigma_4^2)] = \\
&= \sum_{i=1}^n P[4|x_i]^{(t)} \frac{\partial}{\partial \mu_4} \left(-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma_j^2) - \ln(x_i - p) - \frac{(\ln(x_i - p) - \mu_4)^2}{2\sigma_4^2} \right) \\
&= \sum_{i=1}^n P[4|x_i]^{(t)} \frac{(ln(x_i - p) - \mu_4)}{\sigma_4^2} = 0 \\
\iff \quad &\mu_4 = \frac{\sum_{i=1}^n P[4|x_i]^{(t)} \ln(x_i - p)}{\sum_{i=1}^n P[4|x_i]^{(t)}}
\end{aligned}$$

Por lo que

$$\mu_4^{(t+1)} = \frac{\sum_{i=1}^n P[4|x_i]^{(t)} \ln(x_i - p)}{\sum_{i=1}^n P[4|x_i]^{(t)}}. \tag{1.26}$$

Ahora, resolviendo para las σ_j^2's, tenemos

$$\begin{aligned}
&\sum_{i=1}^n P[j|x_i]^{(t)} \frac{\partial}{\partial \sigma_j^2} \left(-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma_j^2) - \frac{(x_i - \mu_j)^2}{2\sigma_j^2} \right) = \\
&= \sum_{i=1}^n P[j|x_i]^{(t)} \left(-\frac{1}{2\sigma_j^2} + \frac{(x_i - \mu_j)^2}{2\sigma_j^4} \right) = 0,
\end{aligned}$$

de donde resulta que

$$\sigma_j^{2(t+1)} = \frac{\sum_{i=1}^n P[j|x_i]^{(t)} (x_i - \mu_j^{(t+1)})^2}{\sum_{i=1}^n P[j|x_i]^{(t)}} \tag{1.27}$$

para $j = 1, 2, 3$.

Análogamente se tiene que

$$\sigma_4^{2(t+1)} = \frac{\sum_{i=1}^n P[4|x_i]^{(t)} (\ln(x_i - p) - \mu_4^{(t+1)})^2}{\sum_{i=1}^n P[4|x_i]^{(t)}}. \tag{1.28}$$

Resumiendo, en este ejemplo el algoritmo EM se reduce a lo siguiente.

1. Se dan valores iniciales a los parámetros $\{(\mu_j, \sigma_j^2)\}_{j=1}^4$ y a las proporciones de mezcla p_j, $j = 1, 2, 3, 4$. Se asigna $t = 1$.

2. **Paso E**. Calcular

$$P[j|x_i]^{(t)} = \frac{g_j(x_i; \mu_j^{(t)}, \sigma_j^{2(t)}) p_j^{(t)}}{f(x_i; \mu_j^{(t)}, \sigma_j^{2(t)})}, \quad j = 1, \ldots, 4, \quad i = 1, \ldots, n.$$

3. **Paso M**. Se actualizan los valores de los parámetros $\{\mu_j, \sigma_j^2\}_1^4$ según (1.25), (1.26), (1.27) y (1.28), y se estima el valor de $p_j^{(t+1)}$ como

$$p_j^{(t+1)} = \frac{\sum_{i=1}^{n} P[j|x_i]^{(t)}}{n}, \ j = 1, \ldots, 4.$$

4. Regresar al paso 2 hasta que el criterio de paro, por ejemplo el siguiente, se cumpla.

$$\max \left\{ \max_{j=1,\ldots,4} \left\{ \left| \frac{\mu_j^{(t+1)} - \mu_j^{(t)}}{\mu_j^{(t)}} \right| \right\}, \max_{j=1,\ldots,4} \left\{ \left| \frac{\sigma_j^{(t+1)} - \sigma_j^{(t)}}{\sigma_j^{(t)}} \right| \right\} \right\} < 0.005.$$

El modelo anterior se ajustó a los datos del CF que se grafican en la figura 1.6, por criterios que se exponen ampliamente en el Capítulo 2, con $p = 6$. Para este ejemplo, se utilizaron los siguientes valores iniciales

j	$\mu_j^{(1)}$	$\sigma_j^{2(1)}$	$p_j^{(1)}$
1	50	9	0.25
2	70	16	0.35
3	100	16	0.11
4	4	1	0.29

y el algoritmo EM programado se completó en nueve iteraciones, reportando los resultados

j	$\hat{\mu}_j = \mu_j^{(9)}$	$\hat{\sigma}_j = \sigma_j^{2(9)}$	$\hat{p}_j = p_j^{(9)}$
1	49.8295	4.6444	0.3143
2	77.0735	7.9428	0.3789
3	99.6055	15.6888	0.1303
4	3.0886	0.8607	0.1765

El buen ajuste realizado de la mezcla evaluada en los parámetros estimados a los datos, se puede ver en la figura 1.13.

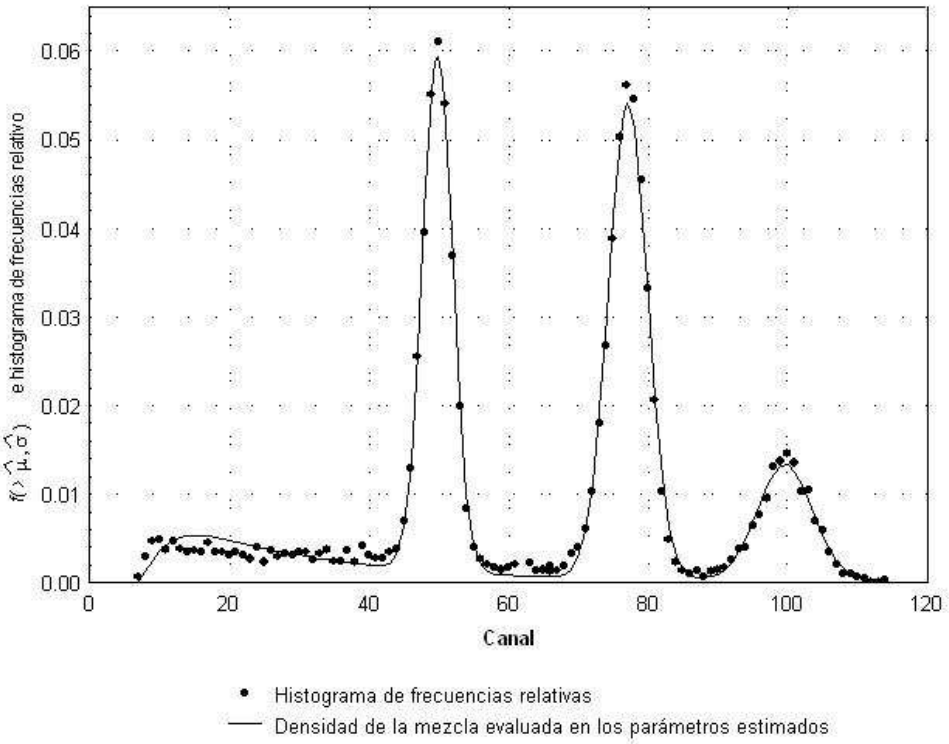


Figura 1.13: Ajuste del modelo de mezcla a los datos del CF correspondientes a la figura 1.6

Capítulo 2

Estimación del ADN nuclear en plantas con citometría de flujo

La creciente cantidad de aplicaciones del CF en áreas científicas e industriales origina un rico campo de investigación estadística. En este trabajo, con el fin de cuantificar el ADN de alguna planta de interés, la modelación estadística se puede dividir en dos partes principales. La primera se encuentra relacionada a la obtención de los estimadores de las medias de contenido de ADN relativo en la fase G_1 de la planta estándar y de interés, y la segunda es referente a la estimación del contenido de ADN de la planta de estudio, usando las medias estimadas.

En la Sección 2.1 se expone, principalmente, la definición de la cantidad de ADN de una planta, en términos de los datos obtenidos por el CF. Esta definición involucra principalmente a las medias de las distribuciones de fluorescencia de las plantas, en su fase G_1. Para obtener los estimadores máximo verosímiles de estas medias, en la Sección 2.2 se trata al modelo de mezcla de distribuciones para ajustarlo a los datos del CF, como se expuso en la Sección 1.4. De esta forma se obtienen los estimadores de las medias que se utilizarán para estimar el contenido de ADN de la planta de interés.

El modelo estadístico que se utilizará para hacer inferencia sobre el contenido de ADN se expone en la Sección 2.3 y dos principales posibles aproximaciones a la verosimilitud del ADN obtenida, por medio de aproximaciones de verosimilitud pivotal, se dan en la Sección 2.4.

En la Sección 2.5 se exponen algunas metodologías seguidas por los biólogos con el fin de puntualizar las principales ventajas que tiene el procedimiento de inferencia descrito en esta tesis en comparación a otras usadas en la literatura biológica.

2.1 Introducción

El citometrista puede manipular un detector de voltaje en el CF, para que éste sea más o menos sensible a la intensidad de señal eléctrica, que proviene del fotodetector. De esta

forma, el usuario del CF puede modificar la posición de los histogramas de frecuencias sobre los canales. Como la posición de los histogramas es, en este sentido, arbitraria, para poder determinar la cantidad de ADN de una planta, se necesita contar con un estándar o referencia, es decir, con células de una planta (o animal) del cual sí se conoce su ADN, (Marie y Brown, 1993).

Existen dos formas de comparar los datos de la planta de referencia con los de la planta de interés; los llamados de estándar externo y estándar interno. El primer método consiste en calibrar el instrumento con los núcleos teñidos de la planta de referencia. El CF se calibra y se fija para después poder comparar la posicion relativa de estos datos con los de las muestras siguientes, las cuales corresponden a la planta de interés. El segundo método consiste en teñir ambos tipos de células (la de planta estándar y la de interés) para ser analizados simultáneamente por el CF. Este segundo procedimiento, según Dolezel (1995), se recomienda para disminuir los errores debidos a la variación en las preparaciones de los núcleos.

Para el CF de Partec, la relación que existe entre la cantidad de ADN y los canales en el intervalo (40,200) es lineal. Entonces la cantidad de ADN de la planta de interés se define como

$$\begin{aligned} &\text{Contenido de ADN de la planta de interés} = \\ &= \frac{\text{Media del pico de la fase } G_1 \text{ de la planta de interés}}{\text{Media del pico de la fase } G_1 \text{ de la planta de referencia}} \times \\ &\times \ \text{Contenido de ADN de la planta de referencia,} \end{aligned} \tag{2.1}$$

Dolezel (1995).

Las unidades del contenido de ADN se dan en picogramos y generalmente los biólogos utilizan el término kC ($k = 1, 2, \ldots$) para enfatizar la ploidía de las células. Por ejemplo, si el contenido de ADN de una planta es igual a 1.14 pg y las células son diploides, se escribirá: $2C$ contenido de ADN = 1.14 pg.

En general se supone que la distribución de las observaciones que conforman los picos de las fases G_1 es simétrica y normal con media igual al verdadero valor del contenido de ADN relativo. En esta tesis también se supondrá esto. Otros trabajos que suponen lo mismo son Gregor (1969), Fried (1976, 1977), Fried y Mandel (1979), Gray, Dean y Mendelsohn (1979), Jett (1978), Vindelov y Christensen (1990) y Watson (1992).

Como la definición para el contenido de ADN (2.1) está dada en términos de la razón de medias multiplicada por una constante que sí se conoce y además se supone que la distribución de los picos G_1 es normal, estadísticamente el problema se reduce a estimar la razón de medias de dos distribuciones normales.

2.2 Análisis de los histogramas del ADN obtenidos por el CF

Como ya se dijo anteriormente, el análisis de los datos del CF se compone, en este trabajo, de dos partes principales. La primera, que es a la que nos referiremos en esta sección, tiene por objetivo estimar por máxima verosimilitud las medias de los picos G_1 de la planta control y la de interés $(\hat{\mu}_X, \hat{\mu}_Y)$, a través de un modelo de mezclas de distribuciones que considere los datos pertenecientes a la basura o *debris*, y los de los picos originados por la intensidad de fluorescencia de núcleos completos.

Como ya se dijo, el software del CF de Partec calcula datos tales como media, moda y CV, a partir del número de picos en el análisis y su posición. Las medias las estima como medias truncadas para cada uno de los picos. Para esto determina un intervalo simétrico alrededor de la moda del pico, y entonces calcula el promedio de los datos dentro de este intervalo; finalmente el promedio obtenido lo redondea al entero más cercano. Los intervalos determinados para el cálculo de las medias truncadas, son ajenos entre sí y usualmente no forman una partición del rango de canales, dejando fuera a muchos datos. Los datos no considerados para calcular las estadísticas descriptivas por el CF están coloreados en negro, como puede apreciarse en la figura 1.6.

Los estimadores de las medias que brinda el sofware del CF, son estimadores muy burdos porque además de ser muy sensibles a la forma en que se seleccionan los intervalos alrededor de la moda, se desecha la información que no cae dentro de algún intervalo. Además, el redondeo al entero más cercano, de la media truncada puede tener como consecuencia que se sobre o subestime la variabilidad real entre y dentro de individuos. Es lamentable que el CF pueda capturar miles de observaciones con gran calidad y precisión y que sin una razón justificada, deseche mucha de esta información al estimar las medias como medias truncadas y redondeadas.

Un hecho sumamente importante que hay que recalcar es que el histograma del CF es por naturaleza una mezcla de distribuciones. De aquí que un camino mucho mejor para calcular los estimadores de los parámetros, sea ajustar por máxima verosimilitud un modelo de mezcla de distribuciones para los datos del histograma del CF.

Los datos del impulso eléctrico amplificado de fluorescencia del ADN original, que es una medida continua, no son recuperables, pero sí lo son la discretización de ellos, que están dados en términos de canales y que el software graba en un archivo tipo ascii. A partir de estos datos recuperados se ajustará el modelo de mezclas para calcular los estimadores de las medias que son de interés.

El modelo de mezclas de distribuciones es, por mucho, mejor ya que considera a todos los datos del histograma y además tiene en cuenta y modela el comportamiento de la basura o *debris* que se acumula en los primeros canales. El modelo de mezclas ya ha sido considerado anteriormente para datos provenientes del CF, por Vindelov y Christensen (1990) y Watson (1992).

Para ajustar el modelo de mezclas se tiene que determinar las distribuciones compo-

nentes. Aquí se va a suponer que la distribución de las observaciones que conforman cada uno de los picos, es normal. Ahora, para modelar la basura o *debris*, hay que notar que ésta consiste de los núcleos "maltratados", los cuales han sido partidos o machacados de más al preparar la muestra y por tal razón presentan una fluorescencia diferente a la de cualquier fase. Johnson y Kotz (1970) mencionan que la distribucion lognormal ha sido aplicada para modelar distribuciones de tamaño de partículas en agregados naturales; por tal razón aquí se utilizará esta distribución para modelar el "tamaño" de fluorescencia de núcleos divididos. Pero debido a que en los datos del CF los primeros canales generalmente no tienen observaciones, se sugiere usar una lognormal trasladada.

Una vez que se proponen las distribuciones componentes del modelo de mezclas, se procede a estimar los parámetros de las distribuciones del CF por máxima verosimilitud a través del algoritmo EM, como se explica en la Sección 1.4.

Llamemos $\hat{\mu}_X$ y $\hat{\mu}_Y$ a los estimadores máximo verosímiles de las medias correspondientes a la fase G_1 de la planta estándar y muestral, respectivamente, en el modelo de mezclas. Debido a que los estimadores máximo verosímiles son asintóticamente normales y el número de observaciones que se utilizan para calcular $\hat{\mu}_X$ y $\hat{\mu}_Y$ son casi siempre de orden mayor a diez mil, nos parece adecuado suponer que las variables aleatorias $\hat{\mu}_X$ y $\hat{\mu}_Y$ se distribuyen aproximadamente normales.

En la práctica, los biólogos o citometristas suelen repetir la medición con el CF en un mismo individuo o planta de interés de tres a cinco veces, usando siempre la misma planta de referencia. Las muestras de tejidos se preparan cada ocasión en que se realiza una medición con el CF. Considerando estas repeticiones, aquí se sugiere agrupar, sumando, los datos de los histogramas provenientes de experimentos homogéneos (realizados bajo las mismas circunstancias) para obtener así un histograma global para cada individuo. A este histograma global es al que se le ajusta el modelo de mezcla de distribuciones.

En lo que resta de este trabajo, denominaremos a $\hat{\mu}_{Xi}$ y $\hat{\mu}_{Yi}$ como las medias estimadas por máxima verosimilitud (en el modelo de mezclas de una distribución lognormal trasladada y tres normales) de la fase G_1 de la planta de referencia y la de estudio, respectivamente, para la i-ésima planta de interés.

2.3 Inferencia estadística para la razón de medias de dos poblaciones Normales con datos pareados

La segunda parte del análisis estadístico que se realiza sobre los datos provenientes del CF, en el contexto de esta tesis, se considera a continuación. Esta parte del análisis propone un modelo estadístico de observaciones pareadas para estimar la razón de medias $\beta = E(\hat{\mu}_X)/E(\hat{\mu}_Y)$, que está directamente relacionada con la cantidad de ADN de la planta a estudiar.

El uso del estándar interno, de manera simultánea con la planta de interés, para obtener las muestras estadísticas que servirán para determinar el ADN de la planta a

estudiar, nos lleva a considerar que tenemos muestras pareadas. Esto es, el modelo de mezclas que se ajusta al histograma de datos de ADN produce una pareja de medias de los picos G_1 de las plantas de referencia y de interés $(\hat{\mu}_X, \hat{\mu}_Y)$ que tienen distribución aproximadamente normal. Por otro lado, la definición (2.1) marca como parámetro a estimar a la razón de dos medias de dos distribuciones. En términos de las parejas obtenidas $(\hat{\mu}_X, \hat{\mu}_Y)$ en la primera parte del análisis estadístico, la cantidad sobre la que se quiere hacer inferencia es $\beta = E(\hat{\mu}_Y)/E(\hat{\mu}_X)$, la razón de medias de dos distribuciones normales.

El problema de estimar relaciones funcionales de la forma $E(Y) = \beta E(X) + \delta$, donde X y Y son variables aleatorias normales, ha sido extensamente tratado en la literatura estadística, en trabajos como los de Neyman y Scott (1948), Creasy (1954, 1956), Fisher (1948, 1942), Kendall y Stuart (1961), Lindley y El-Sayyad (1968) y Kalbfleisch y Sprott (1970). En particular, el problema de estimar razones de medias de distribuciones normales (que es el caso cuando $\delta = 0$) es muy antiguo y se han propuesto soluciones desde hace más de 60 años (Fieller, 1940, 1954; Fisher 1948, 1942). Este problema, en el caso de muestras pareadas, se discute ampliamente en Sprott (2000b).

En la Subsección 2.3.1 se expone el llamado pivotal de Fieller, que constituye una de las soluciones más conocidas para estimar, a través de intervalos de confianza, la cantidad $\beta = E(X)/E(Y)$, cuando (X, Y) tiene distribución conjunta normal bivariada. Las condiciones o supuestos que originan el pivotal que Fieller (1940) propuso, no modelan un aspecto muy importante de los datos originados por la estimación de medias $(\hat{\mu}_X, \hat{\mu}_Y)$ del CF y que se relaciona directamente con el ajuste que realiza el citometrista para situar el histograma sobre un rango de canales. Por esta razón en la Subsección 2.3.2 se propone un modelo más adecuado para realizar la inferencia estadística de β que involucra el uso de parámetros adicionales como se mostrará.

2.3.1 El método de Fieller

Para estimar la razón de medias $\beta = E(Y)/E(X)$ de dos distribuciones normales, Fieller (1940) supuso que el vector aleatorio (X, Y) seguía una distribución normal bivariada con vector de medias $(\mu, \beta\mu)$ y matriz de covarianza arbitraria

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix},$$

lo que origina un modelo con cinco parámetros desconocidos. Fieller calculó la verosimilitud bajo este modelo y obtuvo estimadores de máxima verosimilitud para todos los parámetros. En especial $\hat{\mu} = \bar{x}$ y $\hat{\beta} = \bar{y}/\bar{x}$. Utilizando el modelo normal bivariado para (X, Y) encontró que la cantidad $(\bar{y} - \beta\bar{x})$ es una normal con media cero y varianza $(1/n)(\sigma_2^2 - 2\beta\sigma_{12} + \beta^2\sigma_1^2)$. Estandarizó esta cantidad usando los estimadores de máxima verosimilitud para las varianzas de la manera usual, logrando construir así un pivotal $t(\beta)$ con distribución t de Student con $n - 1$ grados de libertad. Este pivotal se conoce como

el *pivotal de Fieller*,

$$t(\beta) = \frac{(\bar{y} - \beta\bar{x})}{\sqrt{\sum(y_i - \bar{y})^2 - 2\beta\sum(y_i - \bar{y})(x_i - \bar{x}) + \beta^2\sum(x_i - \bar{x})^2}}. \tag{2.2}$$

Nótese que del modelo normal bivariado de Fieller se puede obtener una función de verosimilitud, la cual Fieller utiliza sólamente para calcular los estimadores máximo verosímiles. Sin embargo con el conocimiento estadístico que se tiene actualmente, se puede calcular la verosimilitud perfil de β y obtener directamente intervalos de verosimilitud para este parámetro de interés. Hacer esto implica elegir una forma de estimar por separado β en ausencia de conocimiento de los parámetros restantes. En 1940 no se habían desarrollado esto conceptos estadísticos todavía.

Seguramente por la falta de equipo de cómputo y falta de conocimiento de otras herramientas estadísticas adecuadas, Fieller centró la inferencia sobre β en el cálculo de intervalos de confianza obtenidos a partir del pivotal que él propuso.

Es importante notar que el pivotal $t(\beta)$ no es lineal en β y por tanto no puede inducir por sí solo una verosimilitud pivotal de β que sirva para obtener intervalos de verosimilitud de β. Lo que se puede hacer con $t(\beta)$ es encontrar intervalos de confianza para β; sin embargo tales intervalos no son únicos. El problema es que dado un nivel de confianza $(1-\alpha)\%$ no hay una manera única de encontrar dichos intervalos de confianza. Más aún, los intervalos simétricos $(-t_{\alpha/2}, t_{\alpha/2})$ que presentaba Fieller, tales que

$$P[-t_{\alpha/2} \leq t(\beta) \leq t_{\alpha/2}] = 1 - \alpha$$

no son los más cortos en términos de β y mucho menos son de verosimilitud. Además estos intervalos son arbitrarios en el espacio paramétrico, ya que se encuentran determinados sólamente por la forma algebraica de $t(\beta)$.

Un problema adicional que presentan los intervalos generados por el pivotal de Fieller es que siempre existe un nivel de confianza α a partir del cual los intervalos de confianza para β, mayor o igual a α, son toda la línea recta. Obsérvese que el pivotal de Fieller depende explícitamente de los datos a través de $(\bar{x}, \bar{y})$; por tanto una posible justificación de estos intervalos tan poco informativos es que cuando $\bar{x}, \bar{y}$ son muy pequeños no aportan suficiente información sobre la razón β para estimarla con precisión.

Posteriormente en 1948, Fisher notó que se puede llegar al pivotal de Fieller a partir de supuestos más simples y sin tener que considerar el modelo normal bivariado. Por ejemplo, si se supone que las cantidades

$$u_i = \frac{y_i - \beta x_i}{\sigma\sqrt{1+\beta^2}}, \tag{2.3}$$

son independientes y normales estándar (sin tener que suponer nada sobre las variables X y Y), también se puede derivar el pivotal de Fieller $t(\beta)$, cuya expresión en términos de u_i's es

$$t(\beta) = \frac{\bar{u}\sqrt{n}}{s_u}, \tag{2.4}$$

donde $\bar{u} = \sum u_i/n$, $(n-1)s_u^2 = \sum(u_i - \bar{u})^2$.

2.3.2 Un modelo estadístico adecuado para datos del CF

Fisher (1954) enfatizó que pequeñas diferencias en la formulación matemática de un problema pueden tener consecuencias importantes en la inferencias. En nuestro caso, esto se traduce en que la estimación que se realice sobre β puede ser muy diferente conforme se cambien los supuestos sobre X y Y. Siguiendo estas líneas de pensamiento, Sprott (2000b) explora las consecuencias de adoptar distintos supuestos sobre las variables X y Y en la inferencia sobre β. En especial una de las posibilidades que considera (Sección 4.2) reproduce la situación experimental de los datos de CF que aquí se han descrito. Sprott supone que cada pareja de observaciones (x_i, y_i) son independientes, pertenecen a un modelo de localización-escala y tienen parámetros de localización (μ_i, ν_i), $i = 1, ..., n$. Además supone que la razón de parámetros de localización $\beta = \mu_i/\nu_i$, $i = 1, ..., n$ es la misma para todas la parejas.

En la Sección 2.2 se mencionó que es razonable suponer que $\hat{\mu}_{Xi}$ y $\hat{\mu}_{Yi}$ tienen distribución normal para $i = 1, \ldots, n$, donde n es el número de plantas (o individuos) analizadas con el CF. Supondremos que las parejas $(\hat{\mu}_{Xi}, \hat{\mu}_{Yi})$ (que son estocásticamente independientes) se distribuyen normales con medias $(\mu_i, \beta\mu_i)$ y desviaciones $(\sigma, \lambda\sigma)$, respectivamente. Es decir, supondremos que

$$\hat{\mu}_{Xi} \sim N(\mu_i, \sigma^2) \;\text{ y }\; \hat{\mu}_{Yi} \sim N(\beta\mu_i, \lambda^2\sigma^2) \;\text{ son v.a. independientes,} \tag{2.5}$$

donde β, μ_i y σ^2 son parámetros desconocidos. El modelo (2.5) corresponde al modelo descrito en la Sección 4.2 de Sprott(2000b) para el caso particular de que el modelo de localización-escala es la distribución normal. El parámetro λ^2 es la razón de las varianzas que supondremos conocida porque cuando ésta se desconoce, no todos los parámetros del modelo son identificables y se tienen problemas para realizar las estimaciones. Ver Kendall y Stuart (1961). Así entonces, λ no es un parámetro que tenga el mismo nivel lógico de β, sino que describe la diferencia en precisión de las mediciones de las $\hat{\mu}_{Xi}$'s en comparación a las $\hat{\mu}_{Yi}$'s. Este parámetro λ le da flexibililidad al modelo y se pueden evaluar sus efectos sobre la inferencia de β, como se expone hacia el final de esta sección.

La ventaja que tiene este modelo propuesto por Sprott es que es mucho más flexible que el modelo de Fieller pues permite que cada pareja de observaciones (x_i, y_i) tenga medias distintas. Esta característica resulta necesaria para modelar los datos provenientes del CF. El modelo (2.5) supone que β, la razón de las medias para las parejas $(\hat{\mu}_{Xi}, \hat{\mu}_{Yi}), (i = 1, \ldots, n)$, es común (ya que el contenido de ADN de la planta es constante dentro de una misma especie y variedad) y además considera que las medias $\{(\mu_i, \beta\mu_i)\}_{i=1}^n$ para cada pareja de observaciones puede variar, a diferencia del modelo propuesto por Fieller (1940), en el que las medias se consideran iguales para todas la parejas observadas (x_i, y_i). El modelo ahora propuesto es entonces flexible al permitir que las medias de la planta control y la de estudio para cada planta diferente puedan variar, tal y como realmente ocurre en

los datos debido al ajuste que directamente hace el citometrista al seleccionar el canal sobre el cual desea que se distribuyan los datos. Dicha selección busca que la fase G_1 de la planta estándar y de interés estén entre los canales 40 y 200, ya que es entre estos en los que se considera que la relación (2.1) se cumple.

El modelo (2.5) considera el ajuste que realiza el citometrista, pero el precio que se paga ahora, a cambio de esta flexibilidad, es que contamos con n parámetros adicionales $\{\mu_i\}$ de estorbo en el modelo. Sin embargo Sprott propone una manera ingeniosa de deshacerse de ellos para obtener la verosimilitud perfil de β como se muestra a continuación. Este modelo de Sprott no se basa en los datos solamente a través de $\bar{x}, \bar{y}$, sino que pondera a las parejas de acuerdo a qué tanta información contienen sobre el parámetro β.

A continuación se presentan los resultados expuestos en Sprott(2000b, 2000a, Sección 7.8) que son relevantes para esta tesis.

El modelo (2.5), que sugerimos utilizar para estimar la cantidad de ADN de una planta, a partir de los datos de CF, contiene un gran número de parámetros, ya que cada nueva pareja observada, $(\hat{\mu}_{Xi}, \hat{\mu}_{Yi})$, introduce un nuevo parámetro μ_i. Como los parámetros μ_i carecen de significado en el contexto de las observaciones (debido a que la posición absoluta de los picos, en la escala de canales, no aportan información sobre el contenido de ADN, sino que sólo la razón de ellos), se eliminarán las μ_i's para poder realizar inferencia sobre el parámetro de interés β, como sugirió Sprott (2000b).

Como suponemos que $\hat{\mu}_{Xi}$ y $\hat{\mu}_{Yi}$ son v.a. con distribución de localización-escala (en particular con distribución normal), tenemos que $p_i := \hat{\mu}_{Xi} - \mu_i$ y $q_i := \hat{\mu}_{Yi} - \beta\mu_i$ son pivotales de localización con respecto a μ_i, y su distribución depende de los parámetros de escala σ y $\lambda\sigma$, y tal vez de β, pero no de μ_i.

La transformación $1-1$, $(p_i, q_i) \longleftrightarrow (u_i^*, v_i^*)$

$$\begin{aligned} u_i^* &= \frac{q_i - \beta p_i}{\sqrt{\lambda^2+\beta^2}} = \frac{\hat{\mu}_{Yi} - \beta\hat{\mu}_{Xi}}{\sqrt{\lambda^2+\beta^2}}, \\ v_i^* &= \frac{\beta q_i + \lambda^2 p_i}{\sqrt{\lambda^2+\beta^2}} = \frac{\beta\hat{\mu}_{Yi} + \lambda^2\hat{\mu}_{Xi}}{\sqrt{\lambda^2+\beta^2}} - \mu_i\sqrt{\lambda^2+\beta^2}, \end{aligned}$$

consigue aislar a las μ_i's sólamente en el pivotal v_i^*.

Como el Jacobiano de esta transformación es

$$J = \left| \begin{array}{ll} \frac{\partial u_i^*}{\partial p_i} = -\frac{\beta}{\sqrt{\lambda^2+\beta^2}} & \frac{\partial u_i^*}{\partial q_i} = \frac{1}{\sqrt{\lambda^2+\beta^2}} \\ \frac{\partial v_i^*}{\partial p_i} = \frac{\lambda^2}{\sqrt{\lambda^2+\beta^2}} & \frac{\partial v_i^*}{\partial q_i} = \frac{\beta}{\sqrt{\lambda^2+\beta^2}} \end{array} \right| = 1,$$

la función de verosimilitud, que por definición es proporcional a la función de densidad conjunta de las observaciones $\hat{\mu}_{Xi}$ y $\hat{\mu}_{Yi}$, es a su vez proporcional a la densidad conjunta de u_i^* y v_i^*.

$$L(\beta, \sigma, \mu_i; p_i, q_i) \propto h(u_i^*, v_i^*; \beta, \sigma, \mu_i).$$

La densidad $h(u_i^*, v_i^*; \beta, \sigma, \mu_i)$ contiene al parámetro μ_i sólo a través de las v_i^*'s, entonces al obtener la densidad marginal de u_i^* (integrando sobre el soporte de v_i^*), lo podemos eliminar, para obtener la verosimilitud marginal pivotal de β y σ, sólamente.

$$L_{p_i}(\beta, \sigma) \propto \frac{1}{\sigma} h\left(\frac{u_i^*}{\sigma}; \beta, \lambda\right) = \frac{1}{\sigma} h\left(\frac{\hat{\mu}_{Yi} - \beta\hat{\mu}_{Xi}}{\sigma\sqrt{\lambda^2 + \beta^2}}; \beta, \lambda\right). \tag{2.6}$$

La notación $h\left(\frac{u_i^*}{\sigma}; \beta, \lambda\right)$ hace énfasis en que σ solo ocurre en combinación con las u_i^*'s, no así β y λ.

Para una muestra aleatoria de n parejas, tenemos entonces que la función de verosimilitud es el producto de las n verosimilitudes pivotales marginales (2.6),

$$L_p(\beta, \sigma) = \prod_{i=1}^{n} L_{p_i}(\beta, \sigma).$$

Si consideramos que $\hat{\mu}_{Xi} \sim N(\mu_i, \sigma^2)$ y $\hat{\mu}_{Yi} \sim N(\beta\mu_i, \lambda^2\sigma^2)$ y que además son independientes, la función de densidad conjunta de p_i y q_i está dada por

$$f(p_i, q_i; \beta, \sigma) \propto \frac{1}{\sigma^2} \exp\left\{-\frac{p_i^2 + q_i^2/\lambda^2}{2\sigma^2}\right\}.$$

Dado que el jabcobiano de la transformación $(p_i, q_i) \longleftrightarrow (u_i^*, v_i^*)$ es uno, la función de densidad de u_i^* y v_i^*, marginal, es proporcional a

$$h(u_i^*, v_i^*; \beta, \sigma, \mu_i) \propto \frac{1}{\sigma^2} \exp\left\{-\frac{u_i^{*2} + v_i^{*2}/\lambda^2}{2\sigma^2}\right\}.$$

Integrando sobre los reales con respecto a v_i^*, obtenemos la marginal que nos interesa para hacer inferencia sobre β y σ, pues se eliminan así los n parámetros de estorbo μ_i:

$$\begin{aligned} h_m(u_i^*) &\propto \frac{1}{\sigma^2} \exp\left\{-\frac{u_i^{*2}}{2\sigma^2}\right\} \int_{-\infty}^{\infty} \frac{\lambda\sigma}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x^2}{2\sigma^2}\right\} dx \\ &= \frac{\lambda\sigma}{\sigma^2} \exp\left\{-\frac{u_i^{*2}}{2\sigma^2}\right\} \propto \frac{1}{\sigma} \exp\left\{-\frac{u_i^{*2}}{2\sigma^2}\right\}. \end{aligned}$$

Por tanto la verosimilitud basada en las n parejas observadas $\{(\hat{\mu}_{Xi}, \hat{\mu}_{Yi})\}$, cumple la siguiente relación

$$L_p(\beta, \sigma) \propto \frac{1}{\sigma^n} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} \frac{u_i^{*2}}{\sigma^2}\right\}. \tag{2.7}$$

Para obtener la verosimilitud perfil de β sola, sustituimos a σ en (2.7) por $\hat{\sigma}(\beta)$ que es el estimador máximo verosímil de σ para cada valor fijo de β. A continuación calculamos $\hat{\sigma}(\beta)$.

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\ln L_p(\beta,\sigma) &= -\frac{n}{2\sigma^2}+\frac{1}{2}\sum\frac{u_i^{*2}}{(\sigma^2)^2}=0\\ &\Leftrightarrow\ -\frac{n}{2}\sigma^2+\frac{1}{2}\sum u_i^{*2}=0\\ &\Leftrightarrow\ \hat\sigma=\sqrt{\frac{\sum u_i^{*2}}{n}}.\end{aligned}\tag{2.8}$$

Por lo que la verosimilitud perfil-pivotal de β es

$$L_{p,\max}(\beta)\propto\left(\sum u_i^{*2}\right)^{-\frac{n}{2}}=\left[\frac{\sum(\hat\mu_{Yi}-\beta\hat\mu_{Xi})^2}{\lambda^2+\beta^2}\right]^{-\frac{n}{2}}.\tag{2.9}$$

A partir de la función de la verosimilitud (2.7) o (2.9), podemos calcular el estimador máximo verosimil de β, $\hat\beta$, ya que el valor estimado global coincide con el de la perfil. La expresión está dada por

$$\hat\beta=\frac{s_{22}-\lambda^2 s_{11}}{2s_{12}}\pm\sqrt{\left(\frac{s_{22}-\lambda^2 s_{11}}{2s_{12}}\right)^2+\lambda^2},\tag{2.10}$$

donde $s_{11}=\sum\hat\mu_{Xi}^2$, $s_{12}=\sum\hat\mu_{Xi}\hat\mu_{Yi}$ y $s_{22}=\sum\hat\mu_{Yi}^2$. El signo de la raíz es aquel que hace que $\hat\beta$ tenga el mismo signo que s_{12}. Nótese que como los datos del CF son positivos, entonces el signo de la raíz que da el valor de $\hat\beta$ es el positivo.

La función de verosimilitud relativa maximizada de β es entonces,

$$R^\lambda(\beta)\equiv R^\lambda_{p,\max}(\beta)=\left(\frac{\sum(\hat\mu_{Yi}-\hat\beta\hat\mu_{Xi})^2}{\sum(\hat\mu_{Yi}-\beta\hat\mu_{Xi})^2}\right)^{\frac{n}{2}}\cdot\left(\frac{\lambda^2+\beta^2}{\lambda^2+\hat\beta^2}\right)^{\frac{n}{2}}.\tag{2.11}$$

Bajo el modelo (2.5) suponemos que λ es conocida, ya que no se puede hacer inferencia sobre β sin elegir previamente λ, pero aun en el caso de que no lo sea, la expresión (2.9) puede seguir siendo de utilidad. Kalbfleisch y Sprott (1970) comentan que para cualquier valor fijo de β, puede escogerse λ que maximice la función (2.11) con lo cual se obtiene una cota superior para la verosimilitud relativa $R^\lambda(\beta)$. La cota para $R^\lambda(\beta)$ que proponen usar Kalbfleisch y Sprott, formalmente está dada por la expresión:

$$\max_\lambda R^\lambda(\beta)=\begin{cases}R^{\lambda=\infty}(\beta), & \beta\le s_{12}/s_{11}\\ 1, & s_{12}/s_{11}\le\beta\le s_{22}/s_{12}\\ R^{\lambda=0}(\beta), & \beta\ge s_{22}/s_{12}\end{cases}\qquad\textbf{cuando } \boldsymbol{s_{12}>0},\tag{2.12}$$

que es el caso de los datos provenientes del CF, y

$$\max_\lambda R^\lambda(\beta)=\begin{cases}R^{\lambda=0}(\beta), & \beta\le s_{22}/s_{12}\\ 1, & s_{22}/s_{12}\le\beta\le s_{12}/s_{11}\\ R^{\lambda=\infty}(\beta), & \beta\ge s_{12}/s_{11}\end{cases}\qquad\textbf{cuando } \boldsymbol{s_{12}<0}.$$

La interpretación que se puede hacer del modelo cuando λ se considera con valor infinito o toma el valor cero es la siguiente. En el primer caso la variación de $\hat{\mu}_{Yi}$ es mucho más grande que la de $\hat{\mu}_{Xi}$, dominándola, por lo que $\hat{\mu}_{Xi}$ se puede considerar como determinista. Este es el caso de una regresión lineal, donde $\hat{\mu}_{Yi}$ es la variable de respuesta. Se puede demostrar que con $\lambda = \infty$, el estimador de β, $\hat{\beta}$ es igual a s_{12}/s_{11}, que corresponde efectivamente al modelo de regresión. El caso en el que λ es igual a cero, también reduce al modelo a una regresión lineal, en la que, ahora, la variable de respuesta es $\hat{\mu}_{Xi}$ y $\hat{\beta} = s_{22}/s_{12}$.

A continuación se presentará un ejemplo clásico de unos datos apareados de Darwin que se han analizado extensamente en la literatura estadística desde hace al menos 50 años. Estos datos resultan ser muy adecuados para ejemplificar las ideas que se han discutido en esta tesis.

Ejemplo 2.1 (Kalbfleisch y Sprott, 1970) Los datos de maíz de Darwin (tabla 2.1) corresponden a las alturas, en octavos de pulgada, de plantas de maíz. Darwin tenía la hipótesis que las plantas resultantes de fertilizaciones cruzadas (x) son más grandes que las provenientes de autofertilización (y), que en términos de β, equivale a tener la hipótesis $\beta < 1$. Aquí supondremos que las observaciones pareadas (x, y) son independientes y

$$x_i \sim N(\mu_i, \sigma^2) \quad \text{y} \quad y_i \sim N(\beta\mu_i, \lambda^2\sigma^2).$$

La cota superior de la verosimilitud relativa de β, para los 15 datos de maíz de Darwin, $\max_\lambda R^\lambda(\beta)$ se grafica en la figura 2.1 a partir de (2.12) porque $s_{12} > 0$.

No obs	x	y
1	188	139
2	96	163
3	168	160
4	176	160
5	153	147
6	172	149
7	177	149
8	163	122
9	146	132
10	173	144
11	186	130
12	168	144
13	177	102
14	184	124
15	96	144

Tabla 2.1: Datos de maíz de Darwin

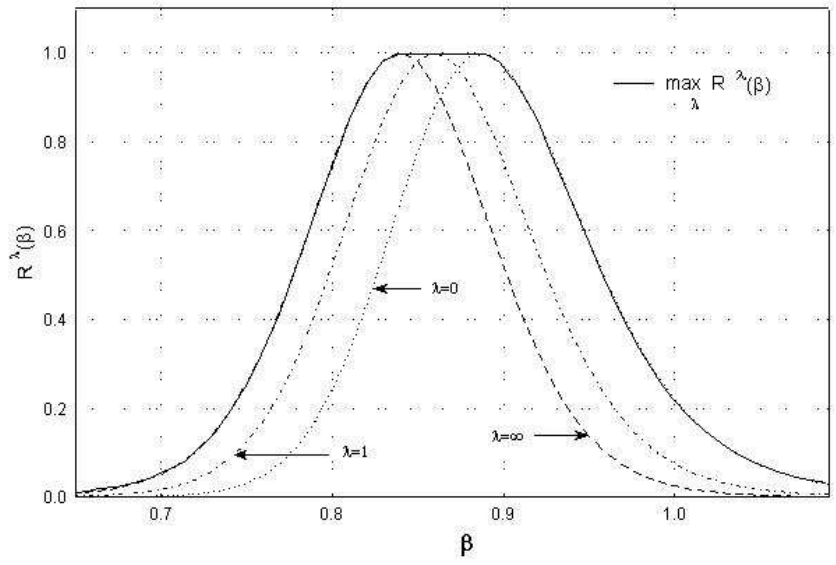


Figura 2.1: Cota superior de las verosimilitudes Relativas de β en los datos de Darwin

En particular cuando se considera que los errores de medición de las plantas tienen variaciones iguales ($\lambda = 1$), se puede apreciar que los valores de beta que son mayores que uno, tienen verosimilitud muy pequeña. Debido a esto, podemos decir que los datos favorecen la hipótesis de Darwin.

A pesar de que bajo ciertos criterios podríamos decir que la cota superior de la verosimilitud relativa de β, $\max_\lambda R^\lambda(\beta)$, no difiere mucho de la verosimilitud de β, con $\lambda = 1$ (como el que los estimadores máximo verosímiles son muy cercanos: $\beta^\infty = 0.839577$, $\beta^1 = 0.859572$ y $\beta^0 = 0.887277$), sí existe una marcada diferencia en las conclusiones que se pueden hacer sobre la plausibilidad de $\beta = 1$. Como se puede observar en la figura 2.1, $\beta = 1$ tiene verosimilitud muy pequeña cuando $\lambda = 1$ ($R^1(\beta) = 0.074$) en comparación a la verosimilitud que tiene cuando $\lambda = 0$ y que es $R^0(1) = \max_\lambda R^\lambda(1) = 0.22$.

El uso de la expresión (2.12) es muy útil para hacer inferencia sobre β cuando λ es desconocida, pero para tener mayor precisión en la inferencia sobre β conviene poder restringir a λ a un rango más pequeño, y esto es a veces posible con la información que se tiene de las condiciones con las que se originan los datos. El parámetro λ no se puede estimar, como β o σ, ya que se encuentra a un nivel lógico diferente. Es un parámetro que sirve para ajustar el modelo y que permite cuantificar la diferencia en precisión de las $\hat{\mu}_{Xi}$ con respecto a las $\hat{\mu}_{Yi}$.

A diferencia del modelo (2.5), el pivotal de Fieller no se ve afectado por cambios del parámetro λ. Fieller (1940) no consideró que las precisiones de las mediciones entre x_i

y y_i pudieran variar, pero aun si se generaliza su pivotal (como lo hace Sprott, 2000b) utilizando

$$u_i = \frac{y_i - \beta x_i}{\sigma\sqrt{\lambda^2 + \beta^2}},$$

se puede observar que si se varía λ, los intervalos de confianza de β que se obtienen no se afectan, debido a que los denominadores se cancelan en la expresión del pivotal $t(\beta)$ en (2.4). Sin embargo, los valores que toma λ sí pueden afectar las inferencias sobre β si existen algunas hipótesis sobre las x_i's y y_i's, como es nuestro caso. El valor que toma λ refleja una realidad experimental y siempre es deseable que ésta se tenga en cuenta en la inferencia que se haga para no perder información relevante que pueda afectar los resultados finales.

2.4 Aproximaciónes a la verosimilitud

Como las observaciones del CF son positivas, entonces $\hat{\mu}_{Xi}$ y $\hat{\mu}_{Yi}$ son también positivas para toda i, lo que resulta en que $s_{12} = \sum \hat{\mu}_{Xi}\hat{\mu}_{Yi} > 0$. En este caso la función de verosimilitud relativa está dada por (2.11) con $\hat{\beta} = r + \sqrt{r^2 + \lambda^2}$ y $r = (s_{22} - \lambda^2 s_{11})/(2s_{12})$.

A partir de (2.11) se pueden calcular los intervalos de verosimilitud de β. Para calcular los intervalos de verosimilitud-confianza se considerarán principalmente las aproximaciones a la verosimilitud, normal y t de Student descritas en la Sección 1.3.4.

Cuando la verosimilitud es simétrica y tal que F_i, $i \geq 3$, definidas en (1.15), cercanas a cero, se considera la aproximación normal. Está aproximación está dada por

$$\ln R(\beta) = -\frac{1}{2}(\beta - \hat{\beta})^2 I_{\hat{\beta}}, \tag{2.13}$$

donde

$$\begin{aligned} I_{\hat{\beta}} &= -\left.\frac{\partial^2 \ln R^{\lambda}(\beta)}{\partial \beta^2}\right|_{\beta=\hat{\beta}} \\ &= -\frac{2n\left(\sum(\hat{\mu}_{Yi} - \beta\hat{\mu}_{Xi})\hat{\mu}_{Xi}\right)^2}{\left(\sum(\hat{\mu}_{Yi} - \beta\hat{\mu}_{Xi})^2\right)^2} + \frac{n\sum \hat{\mu}_{Xi}^2}{\sum(\hat{\mu}_{Yi} - \beta\hat{\mu}_{Xi})^2} - \frac{n}{\lambda^2 + \beta^2} + \left.\frac{2n\beta^2}{(\lambda^2 + \beta^2)^2}\right|_{\beta=\hat{\beta}}. \end{aligned}$$

Si la aproximación normal es adecuada, entonces los intervalos de verosimilitud-confianza tendrán forma aproximada

$$\beta = \hat{\beta} \pm \frac{u}{\sqrt{I_{\hat{\beta}}}}, \quad u \sim N(0,1).$$

Si las colas de la verosimilitud son más pesadas, se puede considerar a la aproximación t de Student. Consideramos la forma del pivotal dada en (1.17) en términos de β, donde

$$F_4 = \frac{\partial^4 r(\beta)}{\partial \hat{\beta}^4} I_{\hat{\beta}}^{-2}$$

y $I_{\hat{\beta}}$ es igual que antes. Si la aproximación t de Student describe bien a la verosimilitud, entonces la verosimilitud pivotal aproximada será

$$r(u_t) = -\left(\frac{a+1}{2}\right)\ln\left[1+\frac{u_t^2}{a}\right].$$

2.5 Otros modelos usados por biólogos

El CF es un poderoso instrumento que pone a disposición del usuario gran cantidad de información sobre el contenido del ADN nuclear en células animales o vegetales. Desafortunadamente la interpretación y manejo de dicha información no ha recibido mucho cuidado en trabajos publicados en revistas de biología y en otras relacionadas con biotecnología. La falta de conocimiento estadístico representa una fuerte desventaja para las investigaciones biológicas debido a que muchas conclusiones se vuelven discutibles, no tanto por la forma en que se realiza el experimento, sino por el análisis estadístico y la interpretación que se hace de los datos obtenidos.

Las tres fuentes de variabilidad que intervienen en los resultados del CF (la inherente a cualquier instrumento de medición, la de la preparación de los núcleos y la de la variabilidad entre individuos) deben siempre de considerarse para elegir el modelo estadístico. Se han publicado trabajos (como por ejemplo, Van Duren et al., 1996) en los que descartan las dos primeras fuentes y basan sus conclusiones en los porcentajes de núcleos clasificados en cada una de las fases por el CF. Este enfoque busca estudiar la tercera fuente de variación (entre individuos) considerando, erróneamente, que las preparaciones de los núcleos no tiene variabilidad y que el CF reconoce sin error estas fases.

Otros trabajos (Palomino, G. et al., 1999) calculan para cada planta analizada, la razón de las medias estimadas de los picos G_1, $z_i = \hat{\mu}_{Yi}/\hat{\mu}_{Xi}$ y suponen que estas razones individuales siguen una distribución normal con media β^* (suponiendo falsamente que β^* es el ADN de la planta de interés) y varianza $\sigma^2_{\beta^*}$. Además simultáneamente suponen que $\hat{\mu}_{Xi}$ y $\hat{\mu}_{Yi}$ tienen distribución normal. Este enfoque presenta serios problemas lógicos:

1. Suponer que $\hat{\mu}_Y/\hat{\mu}_X$ o $\hat{\mu}_X/\hat{\mu}_Y$ sean variables aleatorias normales contradice la suposición de que $\hat{\mu}_X$ y $\hat{\mu}_Y$ son variables aleatorias con distribución normal, y suponer que $\hat{\mu}_Y/\hat{\mu}_X$ o $\hat{\mu}_X/\hat{\mu}_Y$ son normales es mutuamente contradictorio puesto que si $\hat{\mu}_Y/\hat{\mu}_X$ es normal, entonces $\hat{\mu}_X/\hat{\mu}_Y$ no puede serlo y viceversa. Esto es un hecho probabilístico.

2. El problema de estimar razones de parámetros de localización β es por su naturaleza invariante en el sentido de que las parejas ordenadas $\{(x_i, y_i)\}$ contienen la misma información sobre β como las parejas ordenadas $\{(y_i, x_i)\}$ sobre $1/\beta$. Esto es un requisito lógico que no depende de la distribución de las observaciones. Los métodos de Sprott y de Fieller presentados en esta tesis cumplen con este requisito de invarianza pero no así el método que los biologos siguen.

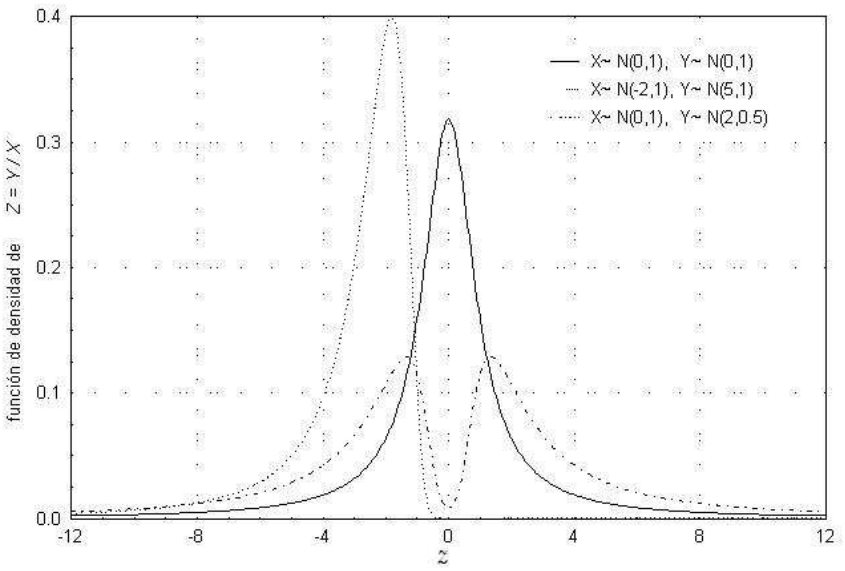


Figura 2.2: Funciones de densidad de Z para diferentes parámetros de las normales

Adicionalmente bajo el supuesto de que las z_i's tienen distribución normal, los biólogos han buscado probar diferencias en el contenido de ADN entre diversos grupos de individuos usando por ejemplo pruebas t (*análisis de varianza* (ANOVA) o pruebas de Tukey) sobre la igualdad de medias, sin verificar el cumplimiento de todos los supuestos del modelo que usan. Por ejemplo, para las pruebas de hipótesis relacionadas al ANOVA o pruebas de Tukey, es crucial que las varianzas en los grupos sean aproximadamente iguales y generalemente este no es el caso en los datos de citometría de flujo.

Como se mencionó, hay dos problemas muy graves e importantes que presenta el enfoque tomado en estos trabajos. El primero es que suponer que $z_i \sim N(\beta^*, \sigma^2_{\beta^*})$ contradice la suposición de que $\hat{\mu}_{Xi}$ y $\hat{\mu}_{Yi}$ son normales y esto último explica bien el mecanismo generador de los datos. Si X y Y son variables aleatorias tal que $X \sim N(\mu_1, \sigma_1^2)$ y $Y \sim N(\mu_2, \sigma_2^2)$, y además son independientes, entonces la distribución de su razón, $Z_{X,Y} = Y/X$, no es normal para todo $\mu_1, \sigma_1, \mu_2, \sigma_2$ en el espacio parametral. La función de densidad exacta de Z está dada por la expresión:

$$f_{Z_{X,Y}}(z) = \left[1 - \sqrt{2\pi \exp[h(z)^2]}\; h(z) \left(\frac{1}{2} - \Phi\left[h(z)\right]\right)\right] \frac{\sigma_1\sigma_2 \exp[-(\mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2)/(2\sigma_1^2\sigma_2^2)]}{\pi(z^2\sigma_1^2 + \sigma_2^2)},$$

donde $h(z) = (z\mu_2\sigma_1^2 + \mu_1\sigma_2^2)/(\sigma_1\sigma_2\sqrt{z^2\sigma_1^2 + \sigma_2^2})$ y $\Phi[\cdot]$ es la función de distribución correspondiente a la distribución normal estándar.

En la figura 2.2 se grafican la función de distribución de Z con diferentes valores de parámetros de medias y varianzas de X y Y. En el caso particular en que $X \sim N(0,1)$ y

$Y \sim N(0,1)$, se tiene que $Z \sim$ Cauchy(1). En todos estos casos queda visualmente muy claro que la distribución de las z_i's está muy lejos de ser normal.

El segundo problema y más importante es que la definición del contenido de ADN de la planta de interés (2.1) marca como parámetro a estimar a β que es la razón de medias de dos distribuciones normales y éste no tiene relación matemática alguna con β^*, la media de las razones de variables aleatorias normales. Más aun, la cantidad β^* no tiene significado alguno en el contexto biológico. Hemos observado mediante simulaciones, que en algunos casos los estimadores máximo verosímiles de β y β^* pueden estar numéricamente, muy cercanos, pero a pesar del parecido que esté presente en esos casos, debemos de tener siempre en mente que ambos resultados surgen de modelos totalmente diferentes y que tal aproximación numérica no se puede garantizar o predecir para todo conjunto arbitrario de datos.

A continuación volvemos a presentar el ejemplo de los datos de Darwin con el objetivo de contrastar el uso de los dos modelos que hemos descrito para describir la verosimilitud de β, β^* y de otro parámetro relacionado β^+ que se definirá también a continuación.

Ejemplo 2.2 Consideremos nuevamente los datos de maíz recolectados por Darwin (tabla 2.1). Darwin decía que las plantas resultantes de autofertilización (y) son más pequeñas que las provenientes de fertilizaciones cruzadas (x), es decir, sostenía que $\beta < 1$.

Como en los datos de CF, las observaciones de la planta control están siempre a la izquierda o la derecha de las observaciones correspondientes a la planta de interés, modificamos los datos de Darwin, de tal forma que esto también ocurra, y hacerlos así, similares a los datos del CF. Se intercambiaron los valores de la altura de la planta proveniente de fertilizaciones cruzadas y la de autofertilización, para las observaciones 2 y 15. Ver tabla 2.2.

A estos datos modificados les aplicamos el modelo (2.5) para obtener la verosimilitud del parámetro $\beta = E(Y)/E(X)$. Veáse la figura 2.3. Como no existe razón alguna para suponer que los errores de mediciones de las dos plantas sean diferentes, se supondrá que el valor de $\lambda = 1$. El estimador máximo verosímil de β es igual a $\hat{\beta} = 0.7905$ y la inferencia sobre $1/\beta$, se puede obtener sustituyendo esta expresión en la función de verosimilitud relativa.

Si suponemos que $Y_i/X_i \sim N(\beta^*, \sigma^2_{\beta^*})$ entonces la verosimilitud de β^* se describe en la figura 2.4, con $\hat{\beta}^* = 0.7878$. En forma simétrica también pudimos haber supuesto que $X_i/Y_i \sim N(1/\beta^+, \sigma^2_{\beta^+})$ y realizar la inferencia sobre $\beta^+ = 1/E(X/Y)$. Es decir, es igualmente válido haber supuesto que las razones inversas son normales pues el problema es simétrico entre X y Y. En la figura 2.4 también se presenta la verosimilitud de β^+ con $\hat{\beta}^+ = 1/1.3029 \approx 0.7675$.

Como se puede observar la inferencia realizada sobre $\beta^* = E(Y/X)$ y $\beta^+ = 1/E(X/Y)$ son diferentes, lo que nos dice que la inferencia considerando a las razones de los datos como normales, no es invariante frente al orden en que se de a las parejas (X_i, Y_i) y pueden llevar a resultados contradictorios.

La razón de presentar como ejemplo a los datos de Darwin se debe a que es un

No obs	x	y	y/x	x/y
1	188	139	0.7394	1.3525
* 2	163	96	0.5889	1.6979
3	168	160	0.9524	1.0500
4	176	160	0.9091	1.1000
5	153	147	0.9608	1.0408
6	172	149	0.8663	1.1544
7	177	149	0.8418	1.1879
8	163	122	0.7485	1.3361
9	146	132	0.9041	1.1061
10	173	144	0.8324	1.2014
11	186	130	0.6989	1.4308
12	168	144	0.8571	1.1667
13	177	102	0.5763	1.7353
14	184	124	0.6739	1.4839
* 15	144	96	0.6667	1.5000

Tabla 2.2: Datos de maíz de Darwin con * modificados

conjunto de datos que ha sido ampliamente analizado en la literatura estadística y siempre utilizando las cantidades $(y_i - \beta x_i)$ en los diversos enfoques. En nuestro conocimiento, nunca se ha propuesto analizar este tipo de datos suponiendo que las razones individuales Y_i/X_i o X_i/Y_i son normales; esto posiblemente debido a las contradicciones lógicas en las que se incurriría como se ha explicado en esta sección.

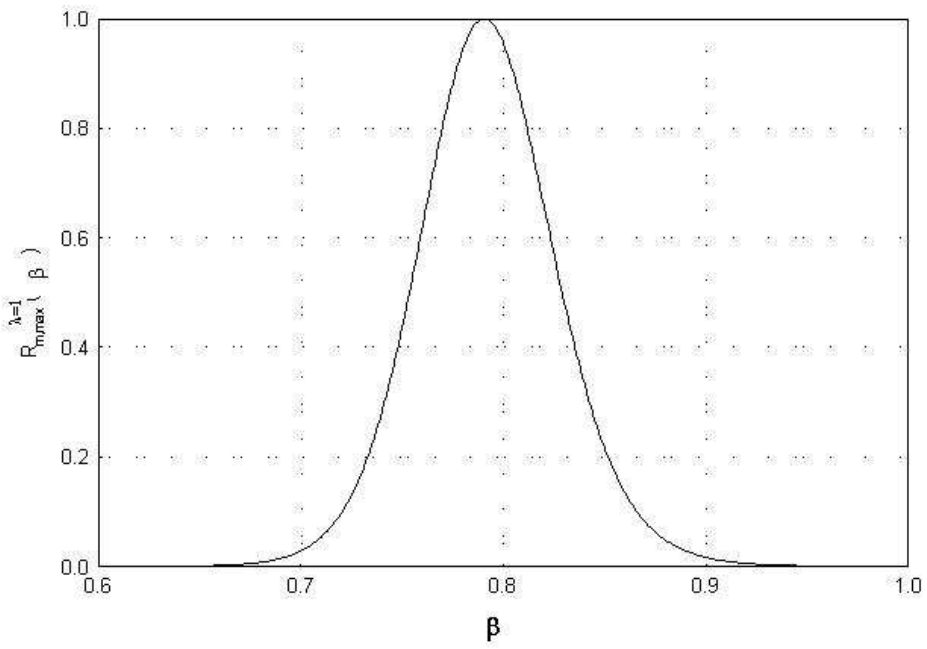


Figura 2.3: Verosimilitud Relativa de β para los datos de maíz de Darwin modificados

En muchos artículos examinados en la literatura biológica, se ha optado erróneamente por el análisis estadístico del estimador β^* (Palomino et al., 1999, etc.) pero es β el parámetro que surge de manera directa a partir de la definición para calcular el ADN.

En cambio el modelo (2.5), propuesto en esta tesis estima adecuadamente el ADN a través del CF, porque explica y modela correctamente la forma en que se generan los datos del CF. Este modelo involucra en forma natural al parámetro de interés β lo cual conduce a llevar a cabo inferencias correctas sobre el ADN de la planta de interés.

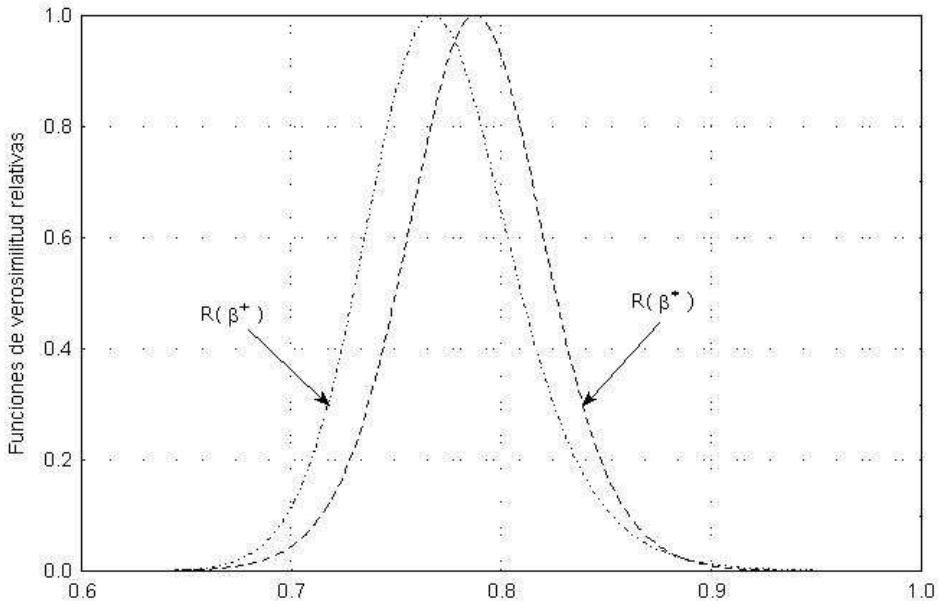


Figura 2.4: Verosimilitudes Relativas de $\beta^* = E(Y/X)$ y $\beta^+ = 1/E(X/Y)$, para los datos de maíz de Darwin modificados

Capítulo 3

Modelación estadística de los datos del *Agave tequilana* Weber, variedad *azul*

La planta *Agave tequilana* Weber, variedad *azul*, comúnmente conocida como agave azul, es de la que actualmente se extrae el licor denominado tequila, el cual representa un fuerte ingreso económico para México. El *Agave tequilana* consta de aproximadamente nueve variedades y a pesar de que algunas de ellas se han utilizado para producir el licor llamado mezcal desde hace más de 200 años, en la actualidad se conocen escasamente. La importancia económica actual del tequila contrasta con el desconocimiento científico que se tiene del *Agave tequilana*; sin embargo el impulso dado a la industria tequilera en el aspecto tecnológico, en recientes fechas, ha originado varios proyectos de investigación de esta planta. La investigación en la industria tequilera que coordina el Consejo Regulador del Tequila junto con otros organismos, ha sido motivada principalmente por la necesidad de obtener mayor información sobre el agave así como encontrar soluciones al problema de susceptibilidad a un par de enfermedades, por el que atraviesa el agave azul.

Para entender el contexto y los objetivos de la modelación estadística de los datos de agave azul obtenidos por citometría de flujo, que se realizará en este capítulo, en la Sección 3.1 se trata brevemente la problemática actual del agave azul y se describe la parte del Programa general de apoyo y desarrollo tecnológico a la cadena productiva Agave-Tequila con la que se relaciona directamente este trabajo. El proyecto del programa, al que nos referimos, se desarrollará por varios años y busca resolver a fondo el problema no solo de la suceptibilidad a los patógenos que actualmente están afectando los cultivos de agave azul, sino los problemas que puedan causar otras enfermedades que en un futuro puedan surgir. Uno de los objetivos contemplados era cuantificar la cantidad total de ADN nuclear del agave azul, que se desconocía. Este se alcanza mediante el análisis de los datos originados por los dos primeros experimentos realizados mediante citometria de flujo. En la Sección 3.2 se elabora el estudio estadístico de los dos conjuntos de datos que fueron obtenidos con objetivos diferentes. Los datos obtenidos en el primer experimento, que llamaremos

preliminares, se originaron con el CF para determinar la región de alguna hoja central de la planta de la cual es conveniente tomar las muestras de tejido de cada individuo a analizar. El estudio estadístico del segundo conjunto de datos está enfocado a caracterizar la cantidad de ADN del agave azul. En ambos casos se realiza el estudio estadístico en sus dos partes: la estimación de la media del contenido de ADN de la planta control y del agave azul en sus fases G_1, y la obtención de la verosimilitud de la cantidad β. Para el segundo juego de datos se realiza además, la aproximación a la verosimilitud descrita en la Sección 2.4 para obtener los intervalos de verosimilitud-confianza de β.

Finalmente en la Sección 3.3, a modo de conclusiones, se comentan los resultados obtenidos.

3.1 Introducción

El tequila es un tipo específico de mezcal, que debe su nombre al pueblo agricultor, del estado de Jalisco, en el que se empezó a producir en el siglo XVII a partir del agave azul. En la actualidad, el tequila está reconocido y protegido como una "denominación de origen". Esto significa que sólamente podrán usar ese nombre, las fábricas, destilerías y envasadoras que estén establecidas en la zona que abarca la protección y que además consuman para la elaboración, exclusivamente plantas de *Agave tequilana* Weber, var. *azul* procedentes también del área de protección a la denominación de origen del tequila. Dicha área se constituye actualmente de algunas regiones específicas de los estados de Nayarit, Tamaulipas, Michoacán y Guanajuato, y todo el estado de Jalisco.

La "denominación de origen" es una figura jurídica reconocida internacionalmente que tiene como finalidad principal evitar que los nombres de algunos productos que han alcanzado un prestigio y reconocimiento importantes, sean convertidos en nombres genéricos que puedan ser utilizados por cualquiera.

El tequila se produce utilizando el corazón o piña de la planta, por eso, a diferencia de la producción de otros licores, la elaboración del tequila es un proceso destructivo pues sólamente se puede utilizar cada planta una única vez. Cuando una planta de agave tiene cuatro o cinco años de vida, produce embriones somáticos, que comúnmente se conocen con el nombre de *hijuelos*, *brotes* o *retoños*, los cuales crecen pegados a la planta madre. El campesino o granjero extrae cada retoño de aproximadamente un año de edad y los replanta, tras haber comprobado que estén en buenas condiciones, en un terreno preparado para albergar a la siguiente generación de agaves. Como el corazón del agave es lo único que se utiliza para elaborar el tequila, los cuidados que se tienen de las nuevas plantitas durante sus maduración, están encaminados a preservar o aumentar el tamaño de su centro o piña.

El agave azul tiene que madurar entre siete y doce años para cosecharse y uno de los principales procedimientos por los que debe pasar para su aprovechamiento son el despunte de sus hojas y el corte de sus flores (en las plantas femeninas). Tras cinco o seis años de crecimiento, las puntas de las hojas (o pencas) se recortan para que el corazón

se haga más grande, ya que los azúcares se concentran así, en el corazón, en lugar de dispersarse en las hojas. Las plantas femeninas entre los cinco y ocho años de edad, florecen, dejando crecer primero un enorme tallo en su centro. Si se dejara crecer el tallo hasta que este alcance su altura máxima, éste secaría el corazón de la planta, porque al igual que las puntas de las pencas, el tallo extrae los azúcares del centro. Para evitar que esto ocurra, se cortan los tallos inmediatamente al brotar.

Los métodos utlizados para cultivar el agave y producir el tequila, han cambiado muy poco desde hace cientos de años, así como la forma de reproducir al agave por brotes. Esta forma de reproducción asexual ha sido la principalmente utilizada para reabastecer los campos de cultivo debido a que el agave femenino muere después de florecer y ya no se puede aprovechar para elaborar tequila.

Los brotes del agave, por originarse directamente de la plata, representan "copias" de la madre, en el sentido de que tienen exactamente las mismas características genéticas, como sucede en el caso de las células hijas originadas por mitosis. Las "copias" o "clones" se replantan y todo el proceso llevado a cabo en la generación de agaves anteriores, se repite masivamente. Estas clonaciones sucesivas, realizadas por largos periodos de tiempo y en zonas extensas, ocasionan la pérdida paulatina de diversidad genética en la especie y tiene como principal consecuencia que los individuos sean todos suceptibles a las agresiones de diversos patógenos. Este reflejo del estrechamiento de la base genética es actualmente el principal responsable de perdidas millonarias en la industria tequilera, ya que fácilmente pueden surgir bacterias, virus u hongos patógenos con el potencial de eliminar proporciones altas de la población de agaves. Actualmente una bacteria y un hongo están afectando a cerca ya del 30% de las plantas ocasionándoles las enfermedades conocidas como "pudrición de tallo" o "pudrición de raiz". Estos patógenos ocasionan primero, una disminución del tamaño del corazón de la planta y finalmente la muerte antes de su maduración.

La vía más eficiente para el control de patógenos es a través de la localización, identificación y utilización de individuos resistentes, ya que la fuente natural de resistencia se encuentra en la diversidad genética contenida en diferentes individuos. Sin embargo, debido a los métodos de propagación asexual del agave que se han realizado por cientos de años, no sólo se han extinto prácticamente siete variedades diferentes a las conocidas, sino que además la base genética de los individuos que aún se cultivan en los campos, ha disminuido gravemente.

Una de las alternativas que se han contemplado para resolver el problema actual del agave, es la de crear fuentes de variación genética por medio de radiaciones que induzcan mutaciones. La inducción de mutaciones artificiales y el uso de la biotecnología son una opción a la que se ha recurrido con anterioridad en otras especies. La creación de variedad genética a través de mutágenos se recomienda como método de mejoramiento genético sólo en plantas con una variación natural pobre y con el objetivo de resolver un problema específico, Konsak y Mikaelsen (1977).

El objetivo de crear una fuente alterna de variedad genética del agave a través de

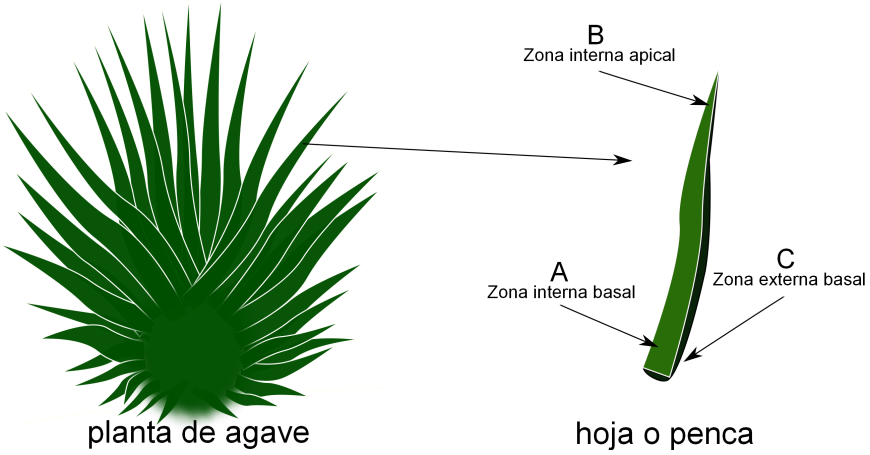


Figura 3.1: Zonas de donde se extrae los tejidos para ser analizados en el CF

inducir mutaciones, para combatir el problema actual de la industria tequilera, es uno de los principales objetivos de un subproyecto que forma parte del Programa general de apoyo y desarrollo tecnológico a la cadena productiva Agave-Tequila. Dentro de este subproyecto se tiene contemplado radiar miles de embriones somáticos en tres dosis diferentes para incrementar la posibilidad de éxito en la localización de individuos resistentes. Los embriones irradiados serán analizados posteriormente con el CF para certificar su identidad genética. Para esto es necesario contar con alguna estimación del contenido de ADN nuclear del agave azul. Como no existe conocimiento previo de esta cantidad, los biologos decidieron aplicar primero la citometría de flujo para caracterizarla.

La introducción de plantas irradiadas a los ambientes naturales, exige un seguimiento de sus características genéticas durante un largo periodo de su vida para verificar que los cambios adquiridos en un principio sean estables a lo largo del tiempo. Esto hace que los análisis de la citometría de flujo también contemplen el obtener datos de plantas de diferentes edades, desde las recién irradiadas hasta las adultas de ya varios años, y comparar los resultados obtenidos.

Para ver si existe alguna variación en las mediciones de ADN que obedecieran a la edad de la planta o a tres lugares diferentes de una hoja de donde obtener las muestras de tejido a analizar con el CF, se realizó un experimento preliminar. El estudio estadístico realizado de estos datos se realiza en la Sección 3.2.1 y considera los datos obtenidos de cuatro agaves azules, de los cuales dos eran jóvenes (1 año) y dos adultos (4 años). Las muestras de tejido se tomaron en tres diferentes lugares de una misma hoja o penca central de cada planta; ver figura 3.1. A partir de este estudio preliminar se busca determinar la mejor posición (A, B o C) para extraer todas las muestras de tejido en análisis subsecuentes.

1. Posición A, Interna Basal. Parte interna de la hoja que se encuentra cerca de su base, situada en la proximidad de la piña

2. Posición B, Interna Apical. Parte interna de la misma hoja que se encuentra cercana a su punta.

3. Posición C, Externa Basal. Parte externa de la misma hoja cercana a su base.

En el análisis de los datos preliminares se quiere evaluar si existen diferencias importantes en las mediciones de ADN entre plantas jóvenes y adultas y evaluar también en qué posición se tiene mayor precisión para estimar el ADN. Así se puede determinar la zona (A, B o C) de donde se realizarán las extracciones de tejido para experimentos subsecuentes.

A cada planta se le realizaron cinco mediciones repetidas por posición excepto para el segundo individuo adulto en la posición A a quien se le tomaron sólo cuatro mediciones.

Una vez definida la posición óptima con los resultados del estudio preliminar, en la Sección 3.2.2 se procede a analizar 12 agaves más para estimar el ADN global del agave azul mediante el CF.

3.2 Análisis estadístico de los datos

Como ya se ha comentado, el análisis estadístico de los datos de contenido de ADN de agave azul obtenidos a través del CF se realizan en este trabajo con dos objetivos principales. El primero busca evaluar si hay diferencias en la medición del ADN con el CF en plantas jóvenes y adultas, y determinar la posición de la hoja de la cual se realizarán todas las mediciones en los experimentos siguientes. El segundo objetivo es el de inferir la cantidad de ADN de la planta *Agave tequilana* Weber, var. *azul*. Las Subsecciones 3.2.1 y 3.2.2 están dirigidas a alcanzar el primer y segundo objetivos respectivamente. En la primera subsección se utilizan los datos originados por el experimento preliminar con los que se concluyó que la mejor posición de la hoja para medir el ADN es la posición A. En la Subsección 3.2.2 se utilizan los datos provenientes de 12 individuos más en la posición A para realizar la estimación del contenido de ADN del agave azul a través de los intervalos de verosimilitud-confianza.

3.2.1 Experimento preliminar

Los datos que llamamos preliminares están destinados a estudiar la variabilidad en las mediciones del CF con respecto a la posición de hoja y edad de la planta y a contrastar la estimación de ADN entre ellas. Estos corresponden a cinco mediciones repetidas por posición en dos individuos jóvenes (un año) y dos adultos (cuatro años), excepto para el segundo individuo adulto en la posición A, del que se realizaron sólamente cuatro repeticiones. Para dar una idea del tipo de salidas que da el software de Partec, en la tabla 3.1 se presentan algunas de las estadísticas descriptivas para cada una de las repeticiones. Nótese que este software del CF redondea las medias de los picos G_1 al entero más cercano como se mencionó anteriormente. Para el análisis que se llevó a cabo

PLANTAS JOVENES (1 AÑO)

		INDIVIDUO I						INDIVIDUO II				
		CV		MEDIA		No.		CV		MEDIA		No.
Posición	ID	MAIZ	AGAVE	MAIZ	AGAVE	NUCLEOS	ID	MAIZ	AGAVE	MAIZ	AGAVE	NUCLEOS
A	780	4.81	4.38	52	80	25,067	781	4.81	4.43	52	79	23,470
	480	5.56	4.71	54	85	9,246	496	4.39	3.49	57	86	17,786
	500	3.00	3.21	50	78	12,028	497	5.26	3.45	57	87	8,148
	501	4.00	3.21	50	78	18,055	498	4.90	3.21	51	78	8,836
	502	4.00	3.85	50	78	17,292	499	3.85	3.05	52	82	8,505
B	503	3.92	3.70	51	81	20,069	527	3.92	3.09	51	81	28,096
	504	4.17	3.25	48	77	9,397	528	3.77	2.38	53	84	29,035
	505	3.00	3.80	50	79	9,502	529	2.94	3.01	51	83	38,382
	506	3.70	2.96	54	84	7,132	708	3.00	3.16	50	79	29,622
	707	2.88	3.21	52	78	22,160	709	3.06	2.56	49	78	23,752
C	508	5.56	4.93	45	71	14,353	513	3.00	3.85	50	78	15,243
	509	5.21	4.73	48	74	15,622	514	4.17	3.95	48	76	12,136
	699	3.85	3.16	52	79	19,079	515	4.08	3.90	49	77	15,424
	700	4.00	4.67	50	75	19,758	516	3.13	3.90	48	77	17,053
	701	4.90	3.90	51	77	19,137	702	3.85	3.13	52	80	24,084

PLANTAS ADULTAS (4 AÑOS)

		INDIVIDUO I						INDIVIDUO II				
		CV		MEDIA		No.		CV		MEDIA		No.
Posición	ID	MAIZ	AGAVE	MAIZ	AGAVE	NUCLEOS	ID	MAIZ	AGAVE	MAIZ	AGAVE	NUCLEOS
A	535	3.00	3.85	50	78	17,669	538	3.06	4.00	49	75	27,217
	536	4.81	3.75	52	80	17,447	539	3.06	3.25	49	77	30,583
	537	2.88	3.75	52	80	14,688	540	3.06	3.90	49	77	20,162
	711	3.70	2.41	54	83	27,408	706	3.77	3.05	53	82	30,618
	712	2.83	3.05	53	82	16,184						
B	530	2.94	2.47	51	81	26,983	541	3.19	3.38	47	74	22,119
	531	2.88	2.47	52	81	26,434	542	3.25	4.00	46	75	27,180
	532	2.88	3.05	52	82	27,220	543	4.17	3.25	48	77	20,170
	703	5.00	4.73	50	74	21,265	705	3.92	3.85	51	78	33,293
	704	2.94	4.61	51	76	16,390	710	2.04	3.38	49	74	18,433
C	517	3.77	3.70	53	81	15,974	760	4.81	3.90	52	77	12,638
	518	2.83	3.70	53	81	15,107	761	4.00	3.25	50	77	16,123
	658	4.00	3.90	50	77	21,991	762	3.92	3.25	51	77	22,039
	520	3.85	3.80	52	79	15,051	763	4.00	3.29	50	76	23,480
	521	3.70	3.01	54	83	13,597	764	3.06	2.67	49	75	16,805

Tabla 3.1: Datos preliminares

en esta tesis no se utilizó nada de la información presentada en la tabla 3.1. Lo que se usaron fueron las frecuencias por canal que forman el histograma del ADN, uno por cada repetición, y que el CF guarda en archivos ascii.

Tras aplicar una prueba de homogeneidad basada en la descrita en Sprott (2000a, Sección 6.5), entre las repeticiones de un mismo individuo, los datos que originan los histogramas de frecuencias correspondientes a cada repetición se sumaron para obtener un solo histograma, como se mencionó en la Sección 2.2. A cada uno de estos histogramas globales se les aplicó el algoritmo EM con el modelo de mezcla descrito en el ejemplo 1.7 de la Sección 1.4, con parámetro de garantía $p = 6$. Los resultados obtenidos por el algoritmo EM escrito en el lenguaje de programación estadístico llamado Gauss para Windows NT/95, versión 3.2.35 (de Aptech System Inc., E. U.), se muestran en la tabla 3.2. El número de iteraciones necesarias para que se cumpliera el criterio de paro fueron entre 8 y 22, con un promedio de 12.58.

Visualmente se verificó el buen ajuste de la mezcla con los parámetros obtenidos,

Plantas Jóvenes (I y II), posiciones A, B y C						
	I A	I B	I C	II A	II B	II C
$\hat{\mu}_1$	50.804	51.250	48.848	53.243	51.858	50.092
$\hat{\mu}_2$	79.304	79.334	76.246	81.075	81.309	78.263
$\hat{\mu}_3$	100.881	101.541	98.683	106.586	101.601	101.360
$\hat{\mu}_4$	2.701	3.183	2.984	2.543	3.246	3.189
$\hat{\sigma}_1^2$	9.194	5.876	13.702	17.836	6.789	7.028
$\hat{\sigma}_2^2$	18.867	13.298	19.403	41.330	17.451	13.785
$\hat{\sigma}_3^2$	37.696	24.590	45.112	39.930	15.736	25.646
$\hat{\sigma}_4^2$	0.990	0.930	0.934	0.899	1.002	0.997
$\hat{p}_1$	0.267	0.347	0.258	0.236	0.159	0.338
$\hat{p}_2$	0.273	0.240	0.364	0.275	0.451	0.232
$\hat{p}_3$	0.120	0.103	0.109	0.137	0.039	0.081

Plantas Adultas (I y II), posiciones A, B y C						
	I A	I B	I C	II A	II B	II C
$\hat{\mu}_1$	53.217	51.470	52.171	48.918	48.929	50.199
$\hat{\mu}_2$	81.070	79.013	79.969	76.784	76.818	76.546
$\hat{\mu}_3$	104.861	102.811	104.158	97.292	96.200	100.232
$\hat{\mu}_4$	2.802	3.172	3.121	2.761	3.219	3.298
$\hat{\sigma}_1^2$	4.768	5.100	7.734	4.588	7.238	4.299
$\hat{\sigma}_2^2$	16.016	25.902	19.612	13.056	17.685	9.426
$\hat{\sigma}_3^2$	22.003	17.492	23.797	15.621	26.644	15.655
$\hat{\sigma}_4^2$	1.017	0.955	0.947	1.000	1.017	1.026
$\hat{p}_1$	0.153	0.306	0.367	0.112	0.147	0.220
$\hat{p}_2$	0.438	0.326	0.341	0.472	0.367	0.243
$\hat{p}_3$	0.074	0.053	0.085	0.091	0.096	0.104

Tabla 3.2: Estimadores máximo verosímiles de los parámetros del modelo de mezcla para los agaves azules obtenidos con el algoritmo EM

graficando la densidad obtenida y el histograma de frecuencias relativo. En la figura 3.2 se muestra el ajuste realizado por el modelo y los parámetros estimados para los datos de la primera planta adulta, en la posición A.

Una vez obtenidos los estimadores de las medias para los datos $\{\hat{\mu}_{Xi}, \hat{\mu}_{Yi}\}$, se puede proceder al análisis estadístico para obtener la verosimilitud de $\beta = E(\hat{\mu}_Y)/E(\hat{\mu}_X)$ considerando cada una de las edades y posiciones de hoja. Para esto, se utilizará la expresión (2.11), obtenida en la Sección 2.3 y que corresponde a la verosimilitud relativa originada a partir de la función de verosimilitud perfil-pivotal de β. Como además consideraremos que $\lambda = 1$ (ya que no existe razón alguna por la cual suponer que las estimaciones de las medias de la fase G_1 de las plantas, tengan variabilidad diferente), la expresión que se utilizará es la siguiente:

$$R^{\lambda=1}(\beta) = \left(\frac{\sum_{i=1}^{2} (\hat{\mu}_{Yi} - \hat{\beta}\hat{\mu}_{Xi})^2}{\sum_{i=1}^{2} (\hat{\mu}_{Yi} - \beta\hat{\mu}_{Xi})^2} \right) \cdot \left(\frac{1+\beta^2}{1+\hat{\beta}^2} \right),$$

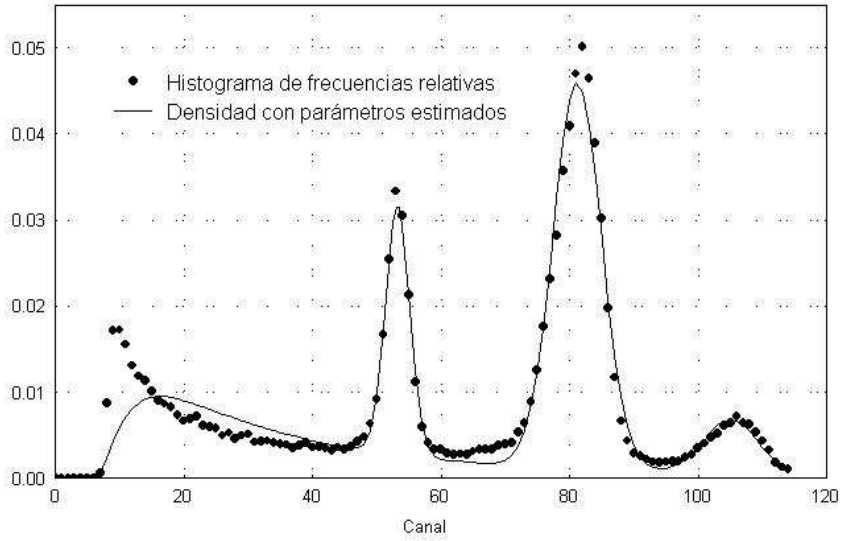


Figura 3.2: Datos preliminares

donde la suma se realiza sobre las medias estimadas de los dos individuos de la misma edad y posición de hoja. Las gráficas de las funciones de verosimilitud relativas obtenidas se exponen por grupos de edad en las figuras 3.3, 3.4 y 3.5, esto con el fin de analizar la precisión y localización de la información aportada sobre la cantidad β por edad de la planta, para cada posición de hoja. Las figuras 3.3, 3.4 y 3.5, que se localizan el las páginas 67 y 68, muestran la verosimilitud relativa de β para plantas jóvenes y adultas en cada posición.

A continuación se analizarán las gráficas para verificar la homogeneidad de experimentos diferentes, según los cuatro criterios descritos en la Subsección 1.3.1.

1. Forma de la verosimilitud. En los seis casos se observa una marcada simetría de la función de verosimilitud relativa.

2. Localización de los estimadores máximo verosímiles. En las posiciones A y B la diferencia de los estimadores, entre las plantas jóvenes y adultas, son menores a 0.01, mientras que para la posición C, esta diferencia es mayor a 0.03.

3. Precisión para estimar β. Se observa que el grupo de edad que tiene funciones de verosimilitud relativas con ancho o apertura más parecidas es la A, siguiendole la B. La función originada por los datos de agaves jóvenes con posición C, tiene un pronunciado pico, lo que señala que esta posición es donde más discrepan las precisiones entre los agaves jóvenes y adultos.

4. Traslape de las funciones de verosimilitud. En las posiciones A y B el traslape de las funciones de verosimilitud relativa es casi total. En contraste, la posición

C presenta un traslape prácticamente nulo, indicando que los valores de β que favorecen las plantas jóvenes son mayores que los que se estiman con las plantas adultas. Es decir, hay indicios de discrepancias fuertes en la estimación del ADN en la posición C parar plantas jóvenes y adultas. Para concluir esta hipótesis, habría que validarla en el futuro con un tamaño de muestra mayor.

En la posición C, la forma de la verosimilitud, la diferencia entre los estimadores máximo verosímiles y la gran precisión para estimar β, principalmente por los datos provenientes de plantas jóvenes, hacen que las curvas de verosimilitud presenten pruebas en contra la homogeneidad de los experimentos que consideran plantas jóvenes y adultas. Mientras que el valor de 1.5616 es el de mayor verosimilitud para los datos que provienen de los individuos jóvenes, la verosimilitud relativa de 1.5616 corresponde a menos del 2% para los datos de los agaves adultos. A diferencia de esta posición, las restantes A y B parecen reproducir la misma información sobre la cantidad β sin importar la edad de las plantas que se analizan en el CF.

A partir del análisis estadístico realizado sobre los datos preliminares y de la interpretación de las gráficas de verosimilitud correspondientes, se determinó que las zonas de la hoja del agave azul de donde se pueden tomar todas las muestras subsecuentes de tejido para experimentos siguientes, son la A o B. Cualquiera de las dos posiciones se pueden utilizar, pero una vez seleccionada una, esta será la que siempre se utilizará para todos los estudios posteriores, esto con el fin de que los resultados sean comparables. Los biólogos del Instituto de Biología de la UNAM decidieron elegir la posición A porque ésta representa (de las tres zonas de hoja) el lugar con tejido más joven en la planta. Así esperan que los resultados de estos experimentos sean comparables con mediciones del CF hechas en estados embriogénicos del agave azul.

3.2.2 Estimación del ADN nuclear del agave azul

Con el objetivo de caracterizar estadísticamente el contenido de ADN del agave azul y con base en los resultados obtenidos en el estudio preliminar, los biólogos de la UNAM utilizaron el CF para obtener los datos de tejido provenientes de las posiciones A de 12 plantas adicionales de agave azul. A los datos nuevos de las 12 plantas (seis jóvenes y seis adultas) se añadieron los obtenidos en el estudio preliminar que también provenían de la posición A, para completar un total de 16 individuos, con los que se realizará la inferencia sobre β.

Nuevamente, para ajustar el modelo de mezclas se sumaron los histogramas de ADN de cada una de las repeticiones de un mismo individuo para tener un histograma global por cada planta. A cada histograma global se ajustó el modelo de mezclas de una distribución lognormal trasladada y tres normales, descrito anteriormente. Del ajuste realizado con el algoritmo EM se obtuvieron los siguientes estimadores máximo verosímiles de las medias de los picos correspondientes a la fase G_1 de la planta de maíz y de agave, respectivamente

Jóvenes			*Adultos*		
i	$\hat{\mu}_{Xi}$	$\hat{\mu}_{Yi}$	i	$\hat{\mu}_{Xi}$	$\hat{\mu}_{Yi}$
1	50.8038	79.3034	**9**	53.2166	81.0697
2	53.2434	81.0752	**10**	48.9179	76.7844
3	52.9687	80.7326	**11**	50.1442	75.3851
4	51.6099	81.7269	**12**	48.8535	78.0533
5	50.5872	76.8985	**13**	50.8301	78.9417
6	51.9030	80.0991	**14**	49.5405	75.0204
7	51.6204	80.3946	**15**	51.2799	78.0825
8	52.4815	81.4230	**16**	49.9109	75.8044

Los datos con índice i igual a 1,2,9 y 10 provienen del estudio preliminar.

La verosimilitud relativa de β para plantas jóvenes y adultas se muestra en la figura 3.6 (página 68). Cada una de estas se originan directamente de la expresión (2.11) con $\lambda = 1$. Los valores de los estimadores máximo verosímiles de β originados por cada juego de datos son 1.5454 y 1.5375, respectivamente. En la figura 3.6 se puede apreciar, de acuerdo a los criterios dados anteriormente, que los grupos por edad para la posición A son homogéneos nuevamente, como se apreció en el análisis de los datos preliminares.

Para realizar el analisis estadístico conjunto de estos datos se debe considerar a la función de verosimilitud originada por la combinación de las observaciones de los dos grupos de edad. Dentro de la modelación matemática suponemos que la variabilidad de los estimadores $\hat{\mu}_X$ y $\hat{\mu}_Y$ son iguales para cualquier grupo de edad (esto es, suponemos que $\lambda = 1$ en todas las edades), pero supondremos que las varianzas de las parejas $\{\hat{\mu}_X^j, \hat{\mu}_Y^j\}$ proveniente de un grupo de edad j, son diferentes a las varianzas de las parejas $\{\hat{\mu}_X^a, \hat{\mu}_Y^a\}$ correspondientes a otro grupo de edad a. La función de verosimilitud perfil-pivotal originada de la combinación de los datos de agaves jovenes y adultos, tiene la siguiente expresión:

$$L_{p,\max}(\beta) = L^j_{p,\max}(\beta) L^a_{p,\max}(\beta) = \left[\frac{\sum_{i=1}^{8}(\hat{\mu}_{Yi} - \beta\hat{\mu}_{Xi})^2}{\lambda^2 + \beta^2}\right]^{-\frac{8}{2}} \left[\frac{\sum_{i=9}^{16}(\hat{\mu}_{Yi} - \beta\hat{\mu}_{Xi})^2}{\lambda^2 + \beta^2}\right]^{-\frac{8}{2}} \quad (3.1)$$

donde $L^j_{p,\max}(\beta)$ y $L^a_{p,\max}(\beta)$ son las funciones de verosimilitud originadas por los datos de los agaves jóvenes y los adultos, respectivamente.

La función (3.1) se origina como el producto de las verosimilitudes perfiles-pivotales de cada uno de los grupos de edad y no debe utilizarse directamente la forma (2.9) para las 16 observaciones $\{\hat{\mu}_{Xi}, \hat{\mu}_{Yi}\}$, porque suponemos que la σ del modelo (2.5) es diferente para los datos de agaves jóvenes que para los datos de agaves adultos. Es decir, suponemos que los datos de agaves jóvenes y adultos cumplen el modelo (2.5), pero el primer grupo con una σ_1 diferente a la σ_2 del segundo.

El estimador $\hat{\beta}$ se obtiene maximizando la función (3.1) para obtener la función de verosimilitud relativa de β. Ver figura 3.7 en la página 69. El estimador máximo verosímil de β que se obtiene a partir de los datos es $\hat{\beta} = 1.5428$.

A partir de la forma de la función de verosimilitud (3.1), se tiene que la información observada $I_{\hat{\beta}}$ de la combinación de los datos de agaves jóvenes y adultos tiene la siguiente expresión

$$I_{\hat{\beta}} = -\frac{\partial^2 \ln L^j_{p,\max}(\beta)}{\partial \hat{\beta}^2} - \frac{\partial^2 \ln L^a_{p,\max}(\beta)}{\partial \hat{\beta}^2}.$$

En nuestro caso, y considerando $\hat{\beta} = 1.542$ y $\lambda = 1$ tenemos que la información observada es igual a $I_{\hat{\beta}} = 17696.440 + 8090.6376 = 25787.078$

La función de verosimilitud relativa de β originada por todos los datos (figura 3.7) es simétrica, de colas muy delgadas y de forma acampanada como la normal. Siguiendo los criterios para proponer una aproximación a la verosimilitud (que se enuncian en la Subsección 1.3.4) se sugirió la aproximación normal con el fin de obtener los intervalos de verosimilitud-confianza de β.

Como también se puede observar en la figura 3.7 la aproximación normal ajusta visualmente muy bien a la verosimilitud original, por lo que los intervalos de verosimilitud-confianza se pueden dar en la forma clásica

$$\beta = \hat{\beta} \pm \frac{u}{\sqrt{I_{\hat{\beta}}}} = 1.5428 \pm \frac{u}{\sqrt{25787.078}} = 1.5428 \pm (0.00622)u,$$

donde u tiene distribución normal estándar.

En la siguiente tabla se dan algunos intervalos de verosimilitud-confianza de β para algunos niveles.

nivel de confianza	nivel de verosimilitud	intervalo
90%	0.2585	[1.5326, 1.5530]
95%	0.1465	[1.5306, 1.5550]
99%	0.0362	[1.5268, 1.5588]

Los intervalos de verosimilitud-confianza del contenido de ADN del agave azul se obtienen fácilmente usando la propiedad de invarianza funcional de la verosimilitud. Simplemente los extremos de los intervalos anteriores se multiplican por la constante contenido de ADN del maíz = 5.433.

nivel de confianza	nivel de verosimilitud	intervalo
90%	0.2585	[8.3264, 8.4376]
95%	0.1465	[8.3158, 8.4482]
99%	0.0362	[8.2950, 8.4690]

Si definimos a δ como la cantidad de ADN nuclear del agave azul, también por la propiedad de invarianza funcional se tiene que el estimador máximo verosímil del contenido de ADN, $\hat{\delta} = 5.433 \times \hat{\beta} = 8.382$. Por tanto, los intervalos de verosimilitud-confianza de δ se puede expresar como:

$$\delta = \hat{\delta} \pm (0.0063)(5.433)u = 8.382 \pm (0.0338)u, \tag{3.2}$$

donde u tiene distribución normal estándar. La función de verosimilitud relativa de δ y su aproximación originada por (3.2) se presentan en la figura 3.8.

3.3 Conclusiones

El modelo de mezclas de distribuciones considerado para describir los histogramas de frecuencias del CF, origina unos estimadores de medias mucho mejores que los que reporta el software del CF porque considera un modelo estadístico adecuado para describir los datos. A esto se suma que los estimadores de máxima verosimilitud que se obtienen a partir de este modelo tienen propiedades óptimas. Por otro lado el algoritmo EM para ajustar la mezcla de distribuciones es fácil de implementar y converge rápidamente.

Una vez obtenidos los estimadores $(\hat{\mu}_{Xi},\ \hat{\mu}_{Yi})$, los objetivos planteados para los dos primeros experimentos se alcanzan fácilmente utilizando el modelo (2.5). El primer objetivo era encontrar la mejor posición de hoja en donde se puede muestrear tejido para estimar el ADN con mayor precisión y el segundo era estimar el ADN del agave azul. En el primer caso, el poder observar gráficamente la verosimilitud del parámetro por hoja y edad permite determinar las regiones de hoja central que son convenientes para realizar las extracciones de tejido para experimentos posteriores. A partir del análisis de las curvas se puede apreciar que no existen diferencias en las mediciones entre el ADN de jóvenes y adultos en las posiciones A y B, mas sí en la C en donde discrepan fuertemente en localización y en dispersión las curvas de verosimilitud. Para este fin no es necesario calcular las aproximaciones a la verosimilitud ni obtener intervalos de verosimilitud-confianza, sino que basta con analizar visualmente las gráficas de la función de verosimilitud relativas de los individuos jóvenes y adultos de acuerdo a los criterios dados en la Subsección 1.3.1.

Para alcanzar el segundo objetivo, se obtiene a la función de verosimilitud relativa de β que considera los datos de agaves jóvenes y adultos, como se explica en la Subsección 1.3.1. Debido a que se puede aproximar bien esta función con una verosimilitud normal sin necesidad de hacer alguna reparametrización previa, calcular los intervalos de verosimilitud-confianza para β es muy sencillo. Finalmente, el objetivo de este experimento, que es estimar el contenido de ADN del agave azul, se alcanza fácilmente multiplicando los intervalos obtenidos para β por la constante *contenido de ADN del maíz* $= 5.433$ pg. Así se puede presentar la inferencia sobre la cantidad de ADN del agave azul en términos de intervalos de verosimilitud-confianza y del estimador máximo verosímil, 8.382 pg.

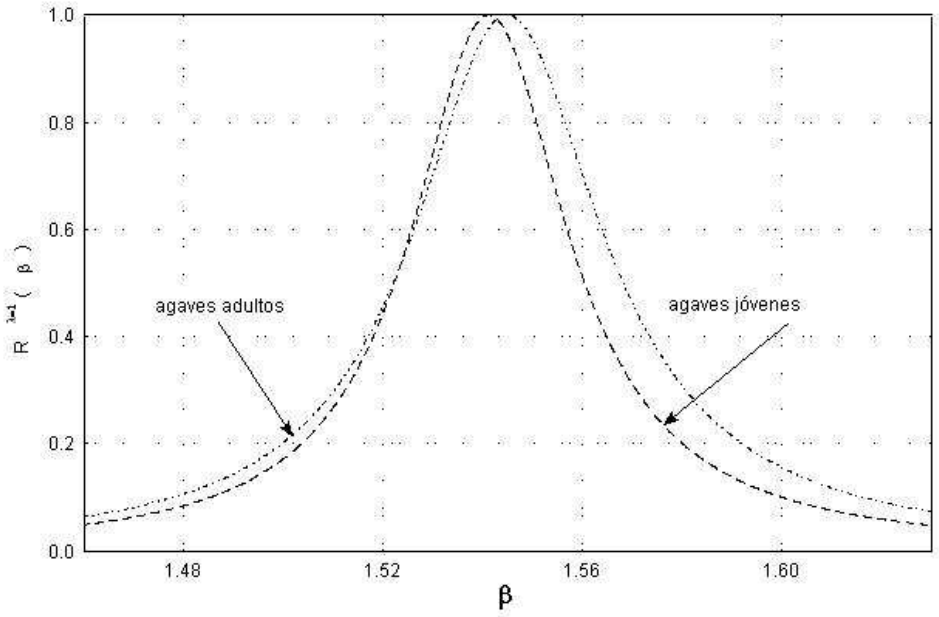


Figura 3.3: Datos preliminares hoja A

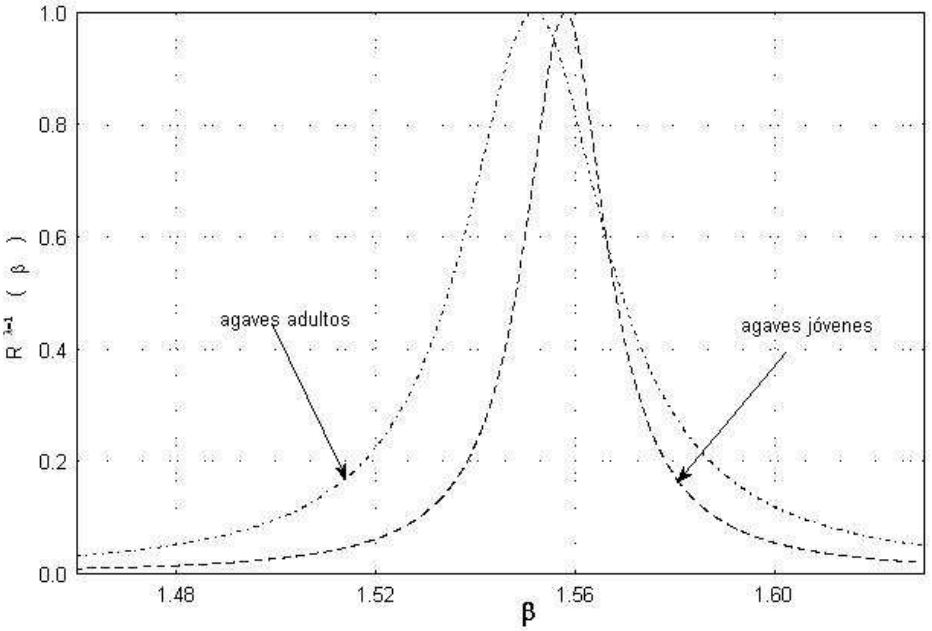


Figura 3.4: Datos preliminares hoja B

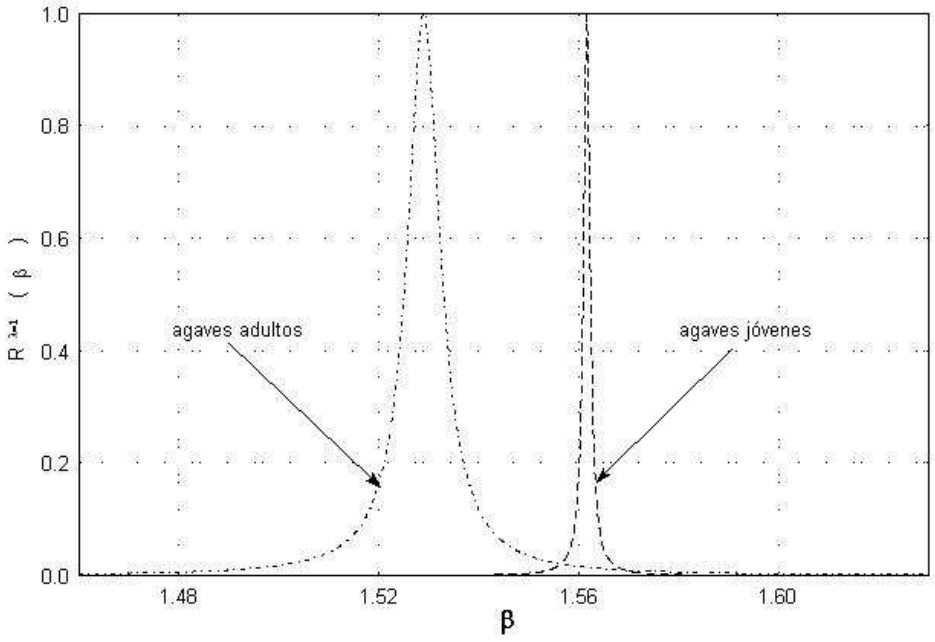


Figura 3.5: Datos preliminares hoja C

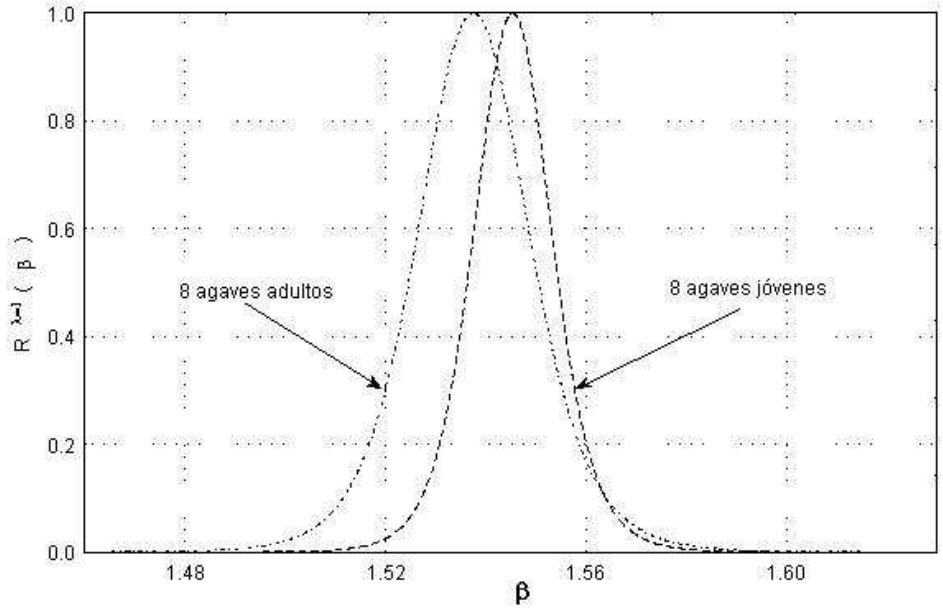


Figura 3.6: Verosimilitud relativa de β para los datos obtenidos de la posición de hoja A

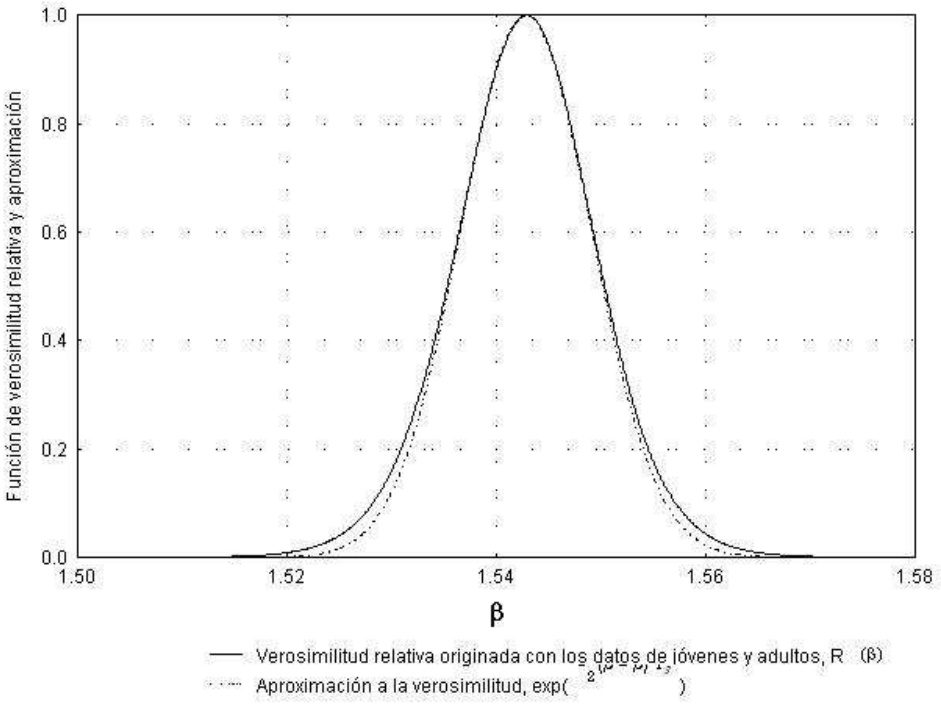


Figura 3.7: Aproximación a la verosimilitud relativa de β para los datos obtenidos de la posición de hoja A

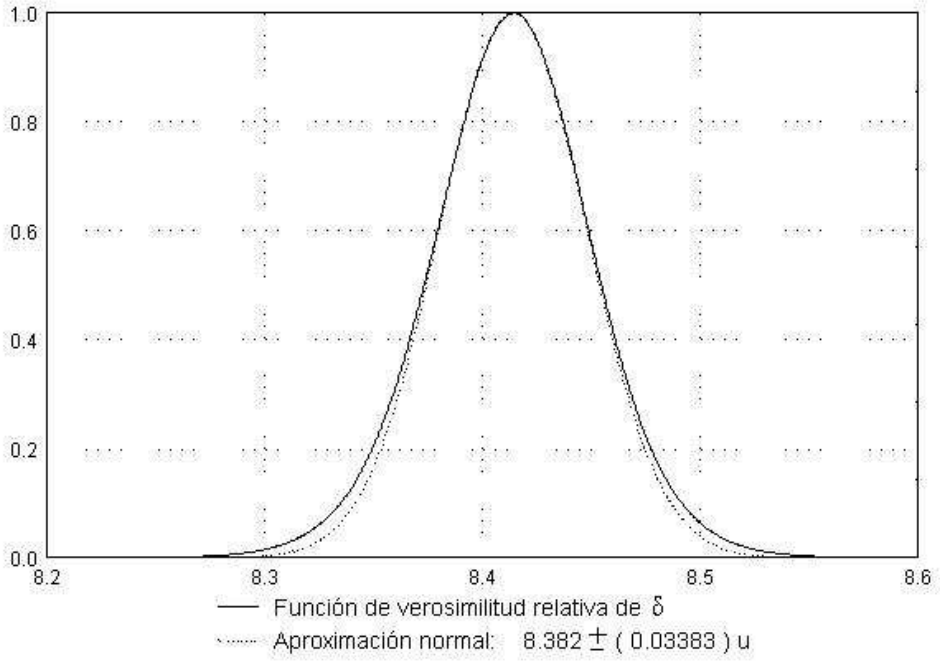


Figura 3.8: Aproximación a la verosimilitud relativa de δ para los datos obtenidos de la posición de hoja A

Capítulo 4

Conclusiones

Un modelo debe ser un reflejo de la realidad y los supuestos del modelo elegido deben tener un sentido y significados científicos. El modelo propuesto en esta tesis se basa en el parámetro biológico de interés y toma en consideración los rasgos distintivos de los datos, así como el mecanismo que los generó. Como se ha mostrado, el procedimiento estadístico que proponemos para cuantificar el ADN con datos provenientes del CF consta de dos partes y cada una modela una característica muy importante de la información que se tiene a partir del CF. La primera parte del procedimiento toma en cuenta que los datos ocurren de manera natural en forma de una mezcla de distribuciones y la segunda considera la manipulación que hace el citometrista para que el histograma se sitúe a partir de un cierto canal.

A continuación se resumiran las principales ventajas del procedimiento estadístico que proponemos en esta tesis para cuantificar la cantidad de ADN de una planta mediante citometría de flujo.

Como se ha comentado con anterioridad, el CF es una aparato muy sofisticado que permite obtener fácilmente y rápidamente información sobre el contenido de ADN de alguna planta, con gran precisión. Sin embargo el software que utiliza para calcular las estadísticas descriptivas desecha innecesariamente mucha de la información obtenida al calcular estimadores muy burdos. La primera parte del análisis estadístico aquí propuesto obtiene estimadores mucho más precisos que los que da el software del CF al considerar un mejor modelo estadístico. Los estimadores máximo verosímiles del modelo de mezclas de distribuciones se obtienen fácilmente con la utilización del algoritmo EM. Este algoritmo tiene además la ventaja de que converge rápidamente. La propiedad asintótica de normalidad de los estimadores máximo verosímiles obtenidos para la medias de los picos G_1, sustenta el supuesto de normalidad que se requiere para la segunda parte del análisis estadístico que se efectúa con las parejas de medias de los picos G_1 de las plantas de referencia y de interés.

El modelo estadístico propuesto para estimar razones de medias de datos normales apareados describe adecuadamente el mecanismo aleatorio de los datos, incluyendo el ajuste que hace el citometrista para fijar el canal a partir del cual se sitúa el histograma.

La introducción de los parámetros adicionales de estorbo μ_i's permite considerar en el modelo esta intervención que realiza el citometrista sobre los datos. Este modelo también considera de forma explícita el parámetro β, que es proporcional al contenido de ADN de la planta de interés. A diferencia de esta metodología, las tradicionales que se utilizan en biotecnología y citometría presentan contradicciones lógicas y probabilísticas. En algunos casos incluso llegan a estimar un parámetro distinto al de interés, el ADN nuclear. Estos métodos frecuentemente suponen que las razónes de las medias estimadas $z_i = \hat{\mu}_{Yi}/\hat{\mu}_{Xi}$ son normales, lo cual contradice el supuesto científicamente más razonable de que las $\hat{\mu}_{Xi}$'s y $\hat{\mu}_{Yi}$'s son normales. Más aún, estos métodos proceden a hacer inferencia sobre $E(z) = \beta^*$ identificando equivocadamente a este parámetro con el ADN de interés, siendo que β^* no se relaciona matemáticamente con el parámetro de interés, el contenido nuclear de ADN de la planta de estudio. No es lo mismo la razón de medias que la media de las razones. El peor defecto al considerar a las $z_i = \hat{\mu}_{Yi}/\hat{\mu}_{Xi}$'s es que no se puede incorporar al modelo el hecho de que el citometrista ajusta el CF para controlar a partir de cual canal se empiezan a acumular las observaciones. En cambio el modelo propuesto en esta tesis sí toma esto en cuenta.

Una característica esencial del problema de estimar razones de parámetros de localización radica en la invarianza lógica y científica que lo constituye. Es decir, la información en las parejas ordenadas $(\hat{\mu}_{Xi}, \hat{\mu}_{Yi})$ sobre β es lógicamente la misma que la contenida en las parejas ordenadas $(\hat{\mu}_{Yi}, \hat{\mu}_{Xi})$ sobre $1/\beta$. El modelo propuesto en esta tesis reconoce esta característica distintiva y es invariante, en este sentido, al orden en como se le alimenten las parejas. Los métodos usados en la literatura biológica examinada y que consideran a las razones de las medias estimadas de los picos G_1, z_i, para cada una de las m plantas, no cumplen con esta propiedad de invarianza. Es decir, dependiendo de que usen $\{\hat{\mu}_Y/\hat{\mu}_X\}_{i=1}^m$ o $\{\hat{\mu}_{Xi}/\hat{\mu}_{Yi}\}_{i=1}^m$ llegan a resultados diferentes para β, lo cual es una contradicción lógica.

Como hemos comentado en la Sección 2.5, los estimadores $\hat{\beta}$ y $\hat{\beta}^*$ pueden ser muy cercanos, dando una falsa impresión de que ambos modelos coinciden en resultados, pero aun en el caso de que esta característica se cumpla para los datos en estudio, las inferencias resultantes de cada modelo pueden ser muy distintas. En un caso, las gráficas de las verosimilitudes de β y β^* pueden tener formas muy diferentes, lo que originaría intervalos de verosimilitud-confianza totalmente distintos y en el otro caso, si sucediera que las gráficas de verosimilitud relativa de β y β^* fueran muy similares, existirá la diferencia con la verosimilitud de β^+, porque el método que considera las z_i's no es invariante, como se menciona en la Sección 2.5.

Con el modelo aquí propuesto, la inferencia sobre el ADN de la planta de interés se puede dar de manera exacta en términos de intervalos de verosimilitud-confianza y del estimador máximo verosímil $\hat{\beta}$. A diferencia de otras metodologías, la aquí propuesta basa la inferencia de β no sólo en un estimador puntual y en intervalos de confianza, sino que además cuantifica la plausibilidad de cada uno de los valores del ADN en estos intervalos, a la luz de la muestra observada. La gráfica de la función de verosimilitud relativa del ADN provee esta información.

Las tablas ANOVA's que se han aplicado en algunos trabajos biológicos y que utilizan a las z_i's incorrectamente para analizar el contenido de ADN, intentan comparar las estimaciones de β^* entre varios grupos de interés, pero no aportan información alguna de la precisión que tienen los datos para estimar β^*. A diferencia de este tipo de análisis, el que se propone en esta tesis no sólo brinda estimadores puntuales e intervalos de verosimilitud-confianza para el parámetro que verdaderamente es de interés para cada grupo, sino que describe la precisión para estimar el ADN a partir de la anchura de las curvas de verosimilitud obtenidas. La interpretación visual de las verosimilitudes permite descubrir características distintivas y estructuras existentes en los datos fácilmente. De esta forma se pueden comparar grupos de edad por posición de hoja del agave de manera objetiva con base en las características de la función de verosimilitud relativa.

Referencias

[1] Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, R. y Watson J. (1994). *Molecular biology of the cell.* 3ra. ed., New York: Garland Publishing, Inc.

[2] Barnard, G. (1977). On ridge regression and the general principles of estimation. *Utilitas Mathematica* **11**, 299-311.

[3] Begg, A. C., McNally, N. J., Schrieve, D. C. y Karcher, H. (1985). A method to measure the duration of DNA synthesis and the potential doubling time from a single sample. *Cytometry* **6**, 620-626.

[4] Chamberlin, S. R. (1989). *Logical foundation and application of inferencial estimation.* Tesis Doctoral. Universidad de Waterloo, Canadá.

[5] Creasy, M. A. (1954). Limits for the ratio of means. *Journal of the Royal Statistical Society* B, **16**, 186-194.

[6] Creasy, M. A. (1956). Confidence limits for the gradient in the lineal functional relationship. *Journal of the Royal Statistical Society* B, **18**, 65-69.

[7] Dempster, A. P., Laird, N. M. y Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* B **39**, 1-38.

[8] Díaz Francés, E. (1995). The EM algorithm. An application to finite mixture distributions. *Comunicación técnica del CIMAT.* D-95-18.

[9] Díaz Francés, E. (1998). *Scientific application of maximum likelihood in multiparametric problems.* Tesis Doctoral. Centro de Investigación en Matemáticas (CIMAT), México.

[10] Dolezel, J. (1995). Flow cytometry: principles and applications in mutation breeding. *Fourteenth IAEA/FAO Interregional training course on advances in plant mutation techniques*, Vienna, Austria, C7-INT-5.135.

[11] Dolezel, J. (1997). Flow cytometry, its applications and potential for plant breeding. *I. T. Lelley cd. Current topics in plant Cytogenetics Related to Plant Improvement* IFA, Tull. Austria 80-90.

[12] Dolezel, J. y Göhede, W. (1995) Sex determination in dioecious plants *Melandrium album* and *M. rubrum* using high-resolution flow cytometry. *Cytometry* **19**, 103-106.

[13] Edwards, A. W. F. (1992). *Likelihood (Expanded Edition)*. Baltimore: The Johns Hopkins University Press.

[14] Eudey, T. L. (1996). Statistical considerations in DNA flow cytometry. *Statistical Science*. **11**, No. 4, 320-334.

[15] Everitt, B. S. y Hand D. J. (1981). *Finite mixture distributions*. Londres: Chapman and Hall.

[16] Fieller, E. C. (1940). The biological standarization of insulin. *Journal of the Royal Statistical Society* B, Suplemento 7, pp. 1-64.

[17] Fieller, E. C. (1954). Some problems in interval estimation (with discussion) *Journal of the Royal Statistical Society* B, **16**, 175-185.

[18] Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* **1**, 3-32.

[19] Fisher, R. A. (1942). *The design of experiments* 3ra ed. Nueva York: Hafner Publishing Company.

[20] Fisher, R. A. (1948). *Statistical methods for research workers* 10ma ed. New York: Hafner Publishing Company.

[21] Fisher, R. A. (1954). Contribution to the symposium on interval estimation. *Journal of the Royal Statistical Society* B **16**, 212-213.

[22] Fisher (1973) *Statistical methods and scientific inference* Hafner Press, New York.

[23] Fried, J. (1976). Method for the quantitative evaluation of data from flow cytofluorometry. *Comp. Biomed. Res.* **9**, 263-276

[24] Fried, J. (1977). Analysis of deoxyribonucleic acid histograms from flow cytometry. *J. Histochem. Cytochem.* **25**, 942-951.

[25] Fried, J. y Mandel, M. (1979). Multi-user system for analysis of data from flow cytometry. *Comput. Programs Biomed.* **10** 218-230.

[26] Gray, J. W., Dean, P. N. y Mendelsohn, M. L. (1979). Quantitative cell-cycle analysis. In M. Melamed, P. Mullaney y M. L. Mendelsohn (eds.). *Flow cytometry and sorting*, p. 383. Nueva York: John Wiley & Sons.

[27] Gregor, J. (1969). An algorithm for the descomposition of a distribution into Gaussian components. *Biometrics*, **25**, 79-93.

[28] Hidderman, W., Schumann, J., Andreeff, M., Barlogie, B., Herman, C. J., Leif, R. C., Mayall, B. H., Murphy, R. F. y Sandberg, A. A.(1984). Convention on nomenclature for DNA cytometry. *Cytometry* **5**, 445-446.

[29] Jeff Wu, C. F. (1983). On the Convergence properties of the EM algorithm. *The Annals of Statistics* **11**, No. 1, 95-103.

[30] Jett, J. H. (1978). Mathematical analysis of DNA histograms for asynchonous and synchronous cell population. In D. Lutz (ed.), *Third International Symposium on Pulse-Cytophotometry*, p. 93. Ghet: European Press.

[31] Johnson, N. L. y Kotz, S.(1970). *Continous univariate distribution* **1**, Nueva York, John Wiley & Sons.

[32] Kalbfleisch, J. G. (1985). *Probability and statistical inference*. 2da ed. **2**, Nueva York: Springer-Verlag.

[33] Kalbfleisch, J. D. y Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society* B **32**, 175-208.

[34] Kendall, M. G. y Stuart, A. (1961) *The advanced theory of statistics*, **2**, pp. 375-418. Londres: Griffin.

[35] Konsak, C. F. y Mikaelsen, K. (1977). *Induced-mutation techniques in breeding seed-propagated species: Selecting parents and handling the MI-M generations for the selection of mutants*. Manual on Mutation Breeding TRS/pub/**119**. 2da ed. International Atomic Energy Agency, Vienna 125-137.

[36] Lindley, D. V. y El-Sayyad, G. M. (1968). The bayesian estimation of a linear functional relationship. *Journal of the Royal Statistical Society* B, **30**, 190-202.

[37] Marie, D. y Brown, S. C. (1993). A cytometric exercise in plant histograms, with $2C$ values for 70 species. *Biol. Cell* **78**, 41-45.

[38] McLachlan, G. J. y Basford, K. E. (1988). *Mixture models. Inference and applications to clustering*. Nueva York: Marcel Dekker, Inc.

[39] Neyman, J. y Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1-32.

[40] Otto (1990). DAPI staining of fixed cells for high-resolution flow cytometry of nuclear DNA, in: H. A. Crissman, Z. Darzynkiewicz (Eds.). *Methods in Cell Biology*, New York, Academic Press **33**, pp. 105-110.

[41] Palomino, G. Dolezel, J., Cid, R., Brunner, I., Méndez, I., Rublo, A. (1999). Nuclear genome stability of *Mammillaria san-angelis* (Cactaceae) regenerants induced by auxins in long-term *in vitro* culture. *Plant Science* **141**, 191-200.

[42] Parker, J. W. (1988). Flow cytometry in the diagnosis of lymphomas. *Cytometry* Suppl. **3**, 38-43.

[43] Sprott, D. A. (2000a). *Statistical inference in science*. Neva York: Springer-Verlag.

[44] Sprott, D. A. (2000b). The estimation of ratios for paired data *Empirical Bayes and likelihood inference, lecture notes in Satistics*, **148**, pp. 141-159, S. E. Ahmendand N. Reid editors. Nueva York: Springer-Verlag.

[45] Sprott, D. A. y Viveros, R. (1984). The interpretation of maximun-likelihood estimation. *The Canadian Journal of Statistics* **12**, 27-38.

[46] Steel, G. G. (1968) Cell loss from experimental tumours. *Cell and Tissue Kinet* **1**, 193-207.

[47] Van Duren, M., Mopurgo, R., Dolezel, J., Afza, R. (1996). Induction and verification of autotetraploids in diploid banana (*Musa acuminata*) by *in vitro* techniques. *Euphytica* **88**, 25-34.

[48] Vindelov, L. L. y Christensen I. J. (1990). A review techniques and results obtained in one laboratory by an integrated system of methods designed for routine clinial flow cytometric DNA analysis. *Citometry* **11**,753-770.

[49] Viveros, R. y Sprott, D. A. (1987). Allowance for skewness in maximun-likelihood estimation with application to the local-scale model. *The Canadian Journal of Statistics*, **15**, No.4, 349-361.

[50] Watson, J. V. (1992). *Flow cytometry data analysis: basic concepts and statistics.* Nueva York: Cambridge University Press.