UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

LA ESTADÍSTICA EN LOS GENES: UNA TÉCNICA DE SEGMENTACIÓN PARA EL ESTUDIO DEL DNA

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

A C T U A R I O

P R E S E N T A:

ZINHUE HERNÁNDEZ JUÁREZ

DIRECTOR DE TESIS:
DOCTORA RUTH SELENE FUENTES GARCÍA

2012

T E S I S

La Estadística en los Genes: una técnica de Segmentación para el estudio del DNA

Hernández Juárez Zinhue

Índice

Agradecimientos

Tus promesas siempre se cumplen, y ésta es otra de ellas, gracias por Tu amor y por todo lo que eres en mí, ¡Te amo!

Agradezco a mis padres por su incondicional apoyo, este triunfo es mío como de ellos, éste y todos los éxitos que siguen se los dedico. ¡Gracias por todo!

A mi hermana por su amistad, apoyo y amor incondicional. ¡Gracias!

A mi directora de tesis por su tiempo y apoyo. ¡Muchas gracias!

A mi familia y amigos que siempre han estado ahí para mí. ¡Gracias por su apoyo!

A todos aquellos que me han visto crecer como ser humano y me han brindado una palabra de aliento. ¡Gracias!

Introducción

Objetivo

El objetivo principal del presente trabajo es transmitir al lector la importancia de la estadística en la determinación de enfermedades que se hacen presentes con mayor frecuencia en la actualidad.

En este caso en particular, se presentará una técnica estadística conocida como Segmentación Binaria Circular que permite analizar detenidamente aquellas zonas del DNA[1] que son afectadas por diversas enfermedades. Dado que la gama de afecciones que modifican áreas de los cromosomas es enorme, el enfoque del estudio reside en muestras de personas entre 25 y 50 años de edad que, por diversas causas, sufren cáncer en los pulmones.

En un mundo en el que las enfermedades es un tópico diario, el cáncer es una de las principales causas de mortalidad a nivel mundial; las muertes debidas al cáncer siguen aumentando, se calcula que serán 12 millones para el año 2030 [2].

Durante la división celular, las células se reproducen duplicando su contenido y luego dividiéndose en dos. El ciclo de división es el medio fundamental a través del cual todos los seres vivos se propagan.

El ciclo celular comprende el conjunto de procesos que una célula debe realizar para cumplir la replicación exacta del DNA y la segregación de los cromosomas replicados en dos células distintas.

En este proceso pueden ocurrir alteraciones tales como la presencia de multiplicidad en regiones específicas del DNA, dando como resultado una ganancia en

[1] Deoxyribonucleic Acid, por sus siglas en inglés.

[2] Comentario publicado por la Organización Mundial de la Salud, OMS, en su página de internet.

el número copiado de secuencias (el número de copias de DNA en una región de un conjunto de genes). También es posible hallar una deleción, es decir, pérdida en el número copiado.

A través de diez años de estudio y desarrollo constante en nuevas formas de analizar estos cambios, hay evidencia de que es posible estudiar el cáncer a través del número copiado de secuencias de DNA en las células cancerígenas.

Como se ha descrito anteriormente, si existe una célula sana, el número copiado es igual a dos; cuando la célula es anormal, el número copiado será menor o mayor a dos.

Se han pretendido visualizar, en las diversas investigaciones realizadas, mapas del genoma entero del número de copias involucradas; tecnologías para lograr esto han incluido Hibridación Genómica Comparativa (CGH) y Análisis Diferencial Representativo (RDA), que básicamente detectan variaciones en los cromosomas en el número copiado, ya sea pérdida o ganancia.

El desarrollo de métodos para esta detección han variado a lo largo de su historia, adquiriendo nuevas hipótesis y modificando algunas ya existentes, al principio y a través del avance, se ha detectado que los arreglos de números copiados contienen *ruido*, por lo que algunos detectores no reflejan resultados adecuados. Por lo tanto, es necesario emplear un método para dividir los cromosomas en regiones iguales de copias del DNA, que después de localizar, descarte aquellas regiones que aparentan estar dañadas cuando, en realidad, son normales.

Este método es conocido como **Segmentación Binaria Circular**, éste analiza la información de un arreglo CGH mediante el cual se ubican puntos de cambio en segmentos del cromosoma por medio de pruebas de hipótesis estadísticas[3] y así localizar en qué parte del genoma hay pérdida o ganancia de DNA.

[3]Para mayor información, véase el Apéndice B.

Este método es novedoso ya que proporciona una forma natural de segmentar al cromosoma en regiones iguales.

1. El Genoma

1.1. Definición y características del genoma

Un genoma es la suma de genes que define cómo es, cómo se autofabrica y cómo funciona un ser vivo. Un genoma es el manual de instrucciones, el programa genético del cuerpo de un ser vivo, lo que constituye su único patrimonio hereditario.

El genoma se transmite con variaciones individuales, de generación en generación. El genoma, por ejemplo, determina la especie de ser vivo. Cada especie se distingue por su material genético. No se puede cambiar de especie. Todos los seres vivos, desde los más grandes, como el elefante y la ballena, hasta los más pequeños, como los virus y las bacterias, incluso las plantas y árboles, tienen genoma. El cuerpo del ser humano también está determinado por su genoma. No hay ningún ser vivo que no lo tenga. El genoma es, pues, común a todos los seres vivos.

También en el genoma se encuentran escritas las características hereditarias encargadas de dirigir el desarrollo, crecimiento, maduración y funcionamiento de cada individuo. Las enfermedades hereditarias también están escritas en el genoma.

La mitad del genoma que se hereda proviene del macho y la otra mitad de la hembra (la mitad de la mitad -la cuarta parte- viene del abuelo paterno, otra cuarta parte de la abuela paterna, y las otras cuartas partes restantes del abuelo y abuela materna, etc.).

Así se reconstruye el árbol genealógico de cada ser vivo y la herencia recibida de todos sus antepasados figura en el genoma de cada uno de los cuerpos de los seres vivos. El genoma está escrito en el lenguaje químico del DNA y para el hombre tiene una longitud de unos 2 m.

Al nacer un hombre, se suele decir que tiene ésta característica de su madre, aquélla de su padre y esta otra de su abuela materna, etc. Estas afirmaciones tienen una base científica, pues todos los caracteres están escritos en el genoma que se hereda de los padres, abuelos, bisabuelos, etc., así hasta el primer hombre y mujer.[11]

1.1.1. El genoma, su tamaño y las especies vivas

El tamaño del genoma depende de la especie. El tamaño, es decir, el número de instrucciones o *letras* que contiene, es proporcional en general a la complejidad del ser vivo que lo hereda. Así, los virus y las bacterias tienen genomas más pequeños que los animales y plantas.

Sin embargo, no hay tanta diferencia entre una mosca (2x900 millones de instrucciones) y el ser humano (2x3,000 millones de instrucciones). Ni entre el hombre y el gorila, el ratón, el perro o el caballo. Algunas plantas como el pino (2x68,000 millones), el trigo (2x16,000 millones) o el maíz (2x5,000 millones), tienen genomas muy superiores en tamaño al del ser humano.

Un 0.2 % del genoma separa individualmente a cada uno de los seres humanos y un 2 % de los chimpancés. Aunque todavía no se conoce cuántos genes son únicos de los individuos o de los seres humanos. Ese 0.2 % son *sólo*: ¡6 millones de diferencias! Y el 2 % son: ¡60 millones de diferencias!

1.1.2. Número de instrucciones del genoma humano

En el ser humano el genoma contiene 2x3,000,000,000 instrucciones, es decir, casi tantas instrucciones como habitantes hay actualmente en la tierra.

El genoma humano es un producto químico lineal (DNA) que ocupa 2m de largo. Su espesor es de unas pocas millonésimas de milímetros. En él se supone que están codificados unos 50,000 genes. Cada gen codifica a su vez una proteína. Varias proteínas forman un carácter del cuerpo humano.

1.1.3. Tamaño del genoma

Si se pusiera cada instrucción al tamaño de las letras de un libro, considerando un libro de unas 1,000 páginas de tamaño DIN A4 (297 x 210 mm), el genoma del ser humano ocuparía una biblioteca de más de 1,000 libros. Son el mismo número de letras que el tamaño natural del genoma. Colocadas una detrás de otra a tamaño natural formarían un hilo de 2 metros. Químicamente, este hilo es de DNA.

El genoma completo tendría por lo tanto, 2 librerías, una del padre y otra de la madre (2n). Ello significa que los 50,000 genes (n) que componen a los seres humanos están duplicados (2). Cada gen tiene, por lo tanto, dos copias, paterna y materna, y pueden ser distintas. Sin embargo, sólo una de las copias se expresa como carácter y es lo que caracteriza la parte visible. Unos de los genes que se expresan son los del padre y otros los de la madre, generando así otra fuente de variabilidad individual.

De aquí viene el que, además de llevar caracteres hereditarios en el genoma, se porten otros caracteres no aparentes, pero que pueden pasar en línea de descendencia.

1.1.4. Un gen, una proteína

Un gen es la cantidad de genoma que codifica una proteína. Un gen es cada una de las unidades del genoma con las que se construye y funciona un ser vivo.

Cada gen contiene todas las instrucciones para reunir los componentes básicos de una proteína. Las proteínas son los principales componentes que constituyen todos los seres vivos. Los componentes de las proteínas, los aminoácidos, se van adquiriendo con los alimentos.

Cada gen tiene una cantidad variable de instrucciones que van desde 500 a 5,000 aproximadamente. El orden y el número total de esas instrucciones codifican cada proteína, según el código genético.

Cambiando una sola instrucción de este código se puede cambiar la forma e incluso la longitud de la proteína que codifica el gen.

Se estima que en el ser humano existen unas 50,000 proteínas diferentes, de diversa forma y tamaño. Por lo tanto, al menos, debe haber 50,000 genes distintos que codifiquen esas proteínas que constituyen nuestra huella genética individual.

Las instrucciones del gen están codificadas en el DNA genético. El DNA es un rosario de instrucciones químicas muy pequeñas. Cada una solamente ocupa varias mil millonésimas de milímetro.

Si cada instrucción de DNA fuera una letra, un gen sería una palabra. El código genético es el lenguaje en el que está escrito el genoma. El abecedario del DNA está constituido por 4 letras químicas (A, C. T, G) mientras que el de las proteínas está constituido por 20 letras químicas (los aminoácidos). A partir de la equivalencia entre DNA y proteínas (cada 3 letras de DNA codifican uno de los aminoácidos) se construye el código genético.

El DNA por lo tanto, es lo que se transmite de las características hereditarias de todos los seres vivos y que está constituido por genes, unidades independientes

que contienen la información necesaria para producir cada uno su proteína.

Cada gen contiene información para producir una proteína específica, lo que tiene lugar a través de un proceso de decodificación es decir, se lee el código genético y se traduce el mensaje contenido en el DNA (esto sucede en el citoplasma de la célula) y de ese modo comienza a producirse la proteína en cuestión. Cada gen origina su proteína durante períodos limitados de tiempo, cuando es necesaria o durante el desarrollo embrionario.

El DNA es una molécula formada por dos hebras complementarias y antiparalelas. Una de las primeras dudas que se plantearon fue la de cómo se replicaba el DNA. Al respecto habían dos hipótesis[4]:

- El DNA se replica de manera conservativa. Esto es, cada hebra de DNA forma una copia y una célula hija recibe la molécula original y la otra célula recibe la copia. *Véase la Figura 1.*

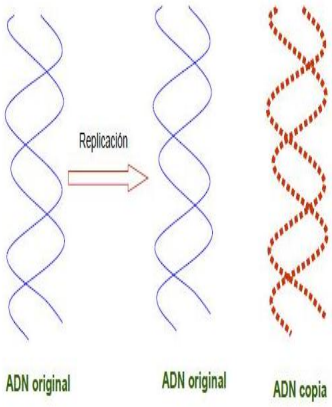


Figura 1: Proceso Conservativo.

[4]Esta controversia fue resuelta por MESELSON y STAHL con una serie de elegantes experiencias, el planteamiento se encuentra en el Apéndice A.

- El DNA se replica de manera semiconservativa. Cada hebra de DNA forma una hebra complementaria y cada célula hija recibe una molécula de DNA que consta de una hebra original y de su complementaria sintetizada de nuevo. *Véase la Figura 2.*

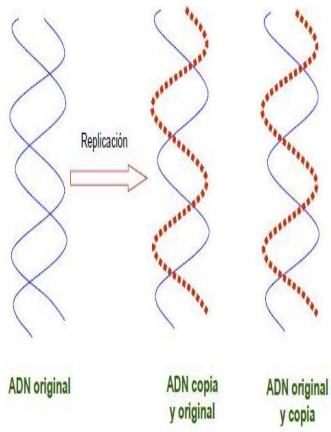


Figura 2: Proceso Semiconservativo.

Cuando una célula se divide, o cuando se originan los gametos, las nuevas células que se forman deben contener la información genética que les permita sintetizar todas las enzimas y el resto de las proteínas necesarias para realizar sus funciones vitales. Ésta es la principal razón por la que el DNA debe replicarse.

La replicación del DNA es el proceso según el cual una molécula de DNA de doble hélice da lugar a otras dos moléculas de DNA con la misma secuencia de bases.

En la célula procariótica[5] la replicación parte de un único punto y progresa

[5]Esta célula es estructuralmente la más simple y pequeña. Como toda célula, está delimitada por una membrana plasmática que contiene pliegues hacia el interior (invaginaciones) algunos de los cuales son denominados laminillas y otro es denominado mesosoma. La célula procariota por fuera de la membrana está rodeada por una pared celular que le brinda protección.

en ambas direcciones hasta completarse. En la célula eucariótica[6] el proceso de replicación del DNA no empieza por los extremos de la molécula sino que parte de varios puntos a la vez y progresa en ambas direcciones formando los llamados ojos de replicación.

Primero se separan las dos hebras y, una vez separadas, van entrando los nucleótidostrifosfato complementarios de cada uno de los de las hebras originales del DNA. Las enzimas DNA polimerasas los unen entre sí formando una hebra de DNA complementaria de cada una de las hebras del DNA original. Se dice que la síntesis de DNA es semiconservativa porque cada una de las moléculas de DNA *hijas* está formada por una hebra de DNA original y otra complementaria sintetizada de nuevo.

1.1.5. Los cromosomas

El genoma está fragmentado y empaquetado en el hombre en 46 partes o cromosomas (23 procedentes del padre y 23 procedentes de la madre).

Los cromosomas X o Y marcan la diferencia de sexo. Aunque pertenecen al mismo *par*, no son iguales entre sí.

1.1.6. El genoma copiado

Las copias del genoma se heredan y se encuentran en el propio cuerpo del ser vivo. En un mismo individuo existen millones y millones de copias de su genoma.

[6]Estas células tienen un modelo de organización mucho máscomplejo que las procariotas. Su tamaño es mucho mayor y en el citoplasma es posible encontrar un conjunto de estructuras celulares que cumplen diversas funciones y en conjunto se denominan organelas celulares. Entre las células eucariotas podemos distinguir dos tipos de células que presentan algunas diferencia: son las células animales y vegetales.

Cada ser vivo está organizado en varios niveles. Su cuerpo está formado por órganos (cerebro, hígado, músculo, etc.), que a su vez están compuestos por tejidos (epitelios, adiposo, etc.) y los tejidos compuestos por células.

De esta forma, en primera aproximación, las dos copias del genoma, los dos metros de información genética están en todas y cada una de las células de un ser vivo. En otras palabras, un genoma se guarda en cada una de las células que componen su ser vivo individual.

1.2. La célula

Cada célula es una esfera de 0.01 mm de diámetro, en la que se guardan, entre otras cosas, las 2 mitades, materna y paterna, del genoma del individuo (2n) en el núcleo. Además, rodeando el núcleo se encuentra el citoplasma donde se traduce el genoma. *Véase la Figura 3.*

Es decir, en el caso del hombre, cada una de las células guarda en unos 2 metros de DNA el único genoma con todas las características del cuerpo material.

Sin embargo esta afirmación es una simplificación, pues hay células en algunos de los tejidos que tienen la mitad del genoma (células reproductoras, n), y otras que tienen parte del genoma (como las de defensa) o ningún genoma (como las células rojas de la sangre).

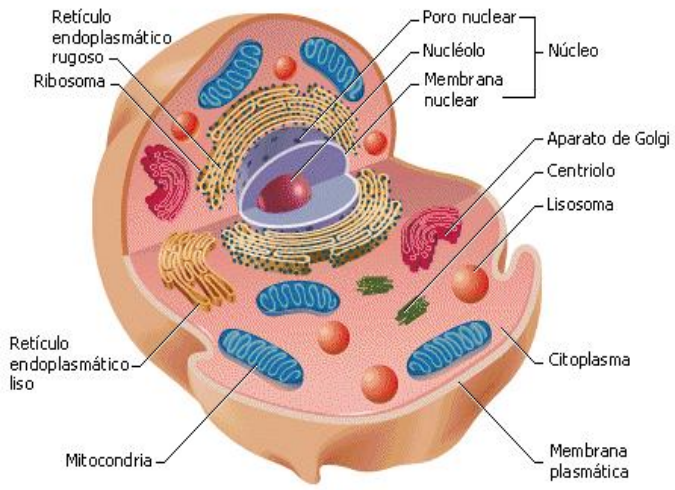


Figura 3: La célula.

La célula es la unidad *anatómica*, *funcional* y *genética* de los seres vivos.

La célula es una estructura constituida por tres elementos básicos:

1. Membrana Plasmática
2. Citoplasma
3. Material Genético (DNA)

y posee la capacidad de realizar tres funciones vitales: *nutrición*, *relación* y *reproducción*.

Se llaman eucariotas a las células que tienen la información genética envuelta dentro de una membrana que forman el núcleo.Un organismo formado por células eucariotas se denomina eucarionte.

Muchos seres unicelulares tienen la información genética dispersa por su citoplasma, no tienen núcleo. A ese tipo de células se les da el nombre de procariotas.[14]

1.2.1. Características generales de las células

Hay células de formas y tamaños muy variados. Algunas de las células bacterianas más pequeñas tienen forma cilíndrica de menos de una micra; en el extremo opuesto se encuentran las células nerviosas, corpúsculos de forma compleja con numerosas prolongaciones delgadas que pueden alcanzar varios metros de longitud.

Casi todas las células vegetales tienen entre 20 y 30 micras de longitud, forma poligonal y pared celular rígida. Las células de los tejidos animales suelen ser compactas, entre 10 y 20 micras de diámetro y con una membrana superficial deformable y casi siempre muy plegada.

Pese a las muchas diferencias de aspecto y función, todas las células están envueltas en una membrana llamada membrana plasmática que encierra una sustancia rica en agua llamada citoplasma.

En el interior de las células tienen lugar numerosas reacciones químicas que le permiten crecer, producir energía y eliminar residuos. El conjunto de estas reacciones se llama metabolismo. Todas las células contienen información hereditaria codificada en moléculas de ácido desoxirribonucleico (DNA); esta información dirige la actividad de la célula y asegura la reproducción y el paso de los caracteres a la descendencia.

Estas y otras numerosas similitudes demuestran que hay una relación evolutiva entre las células actuales y las primeras que aparecieron sobre la tierra.

1.2.2. Composición Química

En los organismos vivos no hay nada que contradiga las leyes de la química y la física. La química de los seres vivos, objeto de estudio de la bioquímica, está dominada por compuestos de carbono y se caracteriza por reacciones acae-

cidas en solución acuosa y en un intervalo de temperaturas pequeño. La química de los organismos vivientes es muy compleja, más que la de cualquier otro sistema químico conocido, ésta está dominada y coordinada por polímeros de gran tamaño, moléculas formadas por encadenamiento de subunidades químicas; las propiedades únicas de estos compuestos permiten a células y organismos crecer y reproducirse. Los tipos principales de macromoléculas son las proteínas, formadas por cadenas lineales de aminoácidos; los ácidos nucleicos, DNA y RNA (Ribonucleic Acid), formados por bases nucleotídicas, y los polisacáridos, formados por subunidades de azúcares.

1.2.3. Mecanismos de Transporte

Difusión

La difusión es el fenómeno en donde una sustancia que se encuentra concentrada en sector se difumina hacia otros sectores. Esto mismo pasa en las células. El agua, el oxígeno, el dióxido de carbono y algunos otras moléculas simples difunden con libertad a través de las membranas celulares. La difusión también es uno de los medios principales por los cuales las sustancias se desplazan dentro de la célula. Uno de los de los factores principales que limitan el tamaño celular es su dependencia a la difusión, que es, en esencia un proceso lento, salvo si las distancias son muy cortas. Este proceso adquiere creciente lentitud y menor eficiencia a medida que la distancia cubierta por las moléculas que se difunden aumenta. La rápida diseminación de una sustancia en un volumen grande, no se debe en particular a la difusión. Del mismo modo, en muchas células el transporte de materiales se acelera mediante circulación activa del citoplasma. Para una difusión eficiente no sólo se requiere un volumen relativamente pequeño, sino también un gradiente de concentración acentuado. Las células mantienen estos gradientes con

sus actividades metabólicas, con lo cual se acelera la difusión. Asimismo, dentro de la célula a menudo se producen materiales en un sitio y se les usa en otros.

Endocitosis y Exocitosis

En otros tipos de procesos de transporte participan vacuolos que se forman a partir de la membrana celular o se fusionan con ella. En la endocitosis el material que será captado por la célula se adhiere a las áreas especiales de la membrana celular y hace que ésta se abulte hacia adentro, produciendo un pequeño saco o vacuolo que engloba a la sustancia. Este vacuolo se libera dentro del citoplasma. Este proceso también puede funcionar a la inversa. Por ejemplo muchas sustancias se exportan desde las células en vesículas o vacuolos formados por los cuerpos de Golgi. Los vacuolos se desplazan hasta la superficie de la célula. Al llegar a la superficie celular, la membrana del vacuolo se fusiona con la membrana de la célula y su contenido se expulsa así hacia el exterior. Este proceso es la exocitosis.

La superficie de la membrana que mira hacia al interior de un vacuolo es equivalente a la superficie que mira hacia el exterior de la célula: del mismo modo, la superficie de la membrana del vacuolo que mira hacia el citoplasma es equivalente a la superficie citoplasmática de la membrana celular. El material necesario para la expansión de la membrana celular a medida que crece la célula, sería transportado ya listo, desde los cuerpos de Golgi hasta la membrana, mediante un proceso similar a la exocitosis. También hay evidencia que las porciones de la membrana celular que se utilizan para formar vacuolos endocitóticos vuelven a la membrana en la exócitosis, de modo que los lípidos y proteínas de la membrana se reciclen.

1.2.4. División Celular

La división celular es el proceso por el cual, a partir de una célula madre, se originan dos células hijas con el mismo número de cromosomas y con idéntica información genética que la célula inicial. *Véase la Figura 4.*

La mitosis se divide en cinco fases:[15]

- Interfase. El DNA aparece en forma de cromatina, constituida por largas moléculas filamentosas de DNA. Al final de la interfase, el DNA se duplica, obteniéndose dos moléculas iguales. El centrosoma también se duplica.
- Profase. Comprende tres fases:
  1. Formación de cromosomas o diferenciación de ellos.
  2. Duplicación de cromosomas por división longitudinal, o que las dos cadenas del resultado de la mencionada duplicación se separan.
  3. Formación del huso acromático. Los dos centrosomas migran cada uno a cada polo de la célula, y quedan unidos por fibras.
- Metafase o fase destructora. Comprende dos fases:
  1. Desaparición de la membrana nuclear.
  2. Formación de la estrella madre o placa ecuatorial. Los cromosomas hermanos se colocan en la zona central de la célula y se fijan por el centrómero a las fibras del huso acromático.
- Anafase o fase constructora. Comprende dos fases:
  1. Las fibras del huso acromático se contraen, separando así los cromosomas, y migrando éstos a los polos de la célula, separándose así de los cromosomas hermanos.

2. Los filamentos desaparecen, y los cromosomas permanecen junto a su respectivo centrosoma.

- Telofase o fase final. Comprende dos fases:

1. Aparecen dos núcleos, y cuya membrana envuelve a los cromosomas que desaparecen o se desenrollan, dando lugar a masas de cromatina.
2. División del citoplasma. Hay dos tipos:
   - Por tabicación. Mediante este proceso, propio de las células vegetales, se separa el contenido celular, núcleo y citoplasma, entre las células hijas.
   - Por estrangulamiento. Es un proceso similar al anterior, pero que se da en las células animales. La célula se va estrechando por el centro, hasta tal punto que se divide por la mitad.

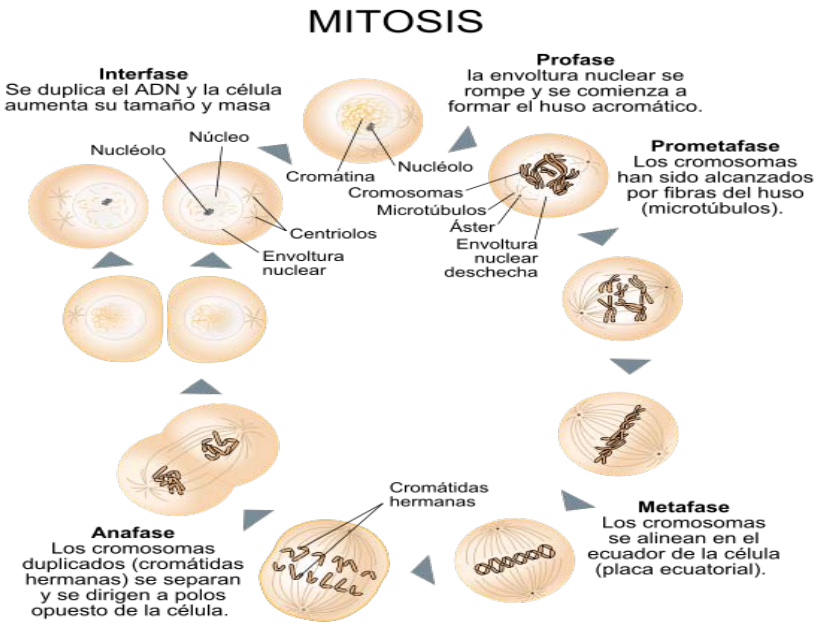


Figura 4: Mitosis.

1.3. El Cáncer

El cáncer no es una enfermedad, sino más bien muchas enfermedades.

Todo cáncer empieza en las células. Las células son las unidades básicas que forman los tejidos del cuerpo.

Para entender mejor qué es el cáncer, es necesario saber cómo las células normales se vuelven cancerosas. *Véase la Figura 5.*

El cuerpo está compuesto de muchos tipos de células, éstas crecen y se dividen para producir nuevas células conforme el cuerpo las necesita. Normalmente las células envejecen, mueren y éstas son reemplazadas por células nuevas pero a veces, este proceso ordenado de división de células se descontrola. Células nuevas se siguen formando cuando el cuerpo no las necesita, llegando a formar una masa de tejido conocida conmúnmente como tumor.[7]

Los tumores se dividen en:

- Benignos. Éstos no son cancerosos, generalmente se pueden extirpar. En la mayoría de los casos, estos tumores no vuelven a crecer. Las células de los tumores benignos no se diseminan o riegan a otros tejidos o partes del cuerpo.
- Malignos. Éstos son cancerosos, las células en estos tumores pueden invadir el tejido a su alrededor y presentar metástasis.

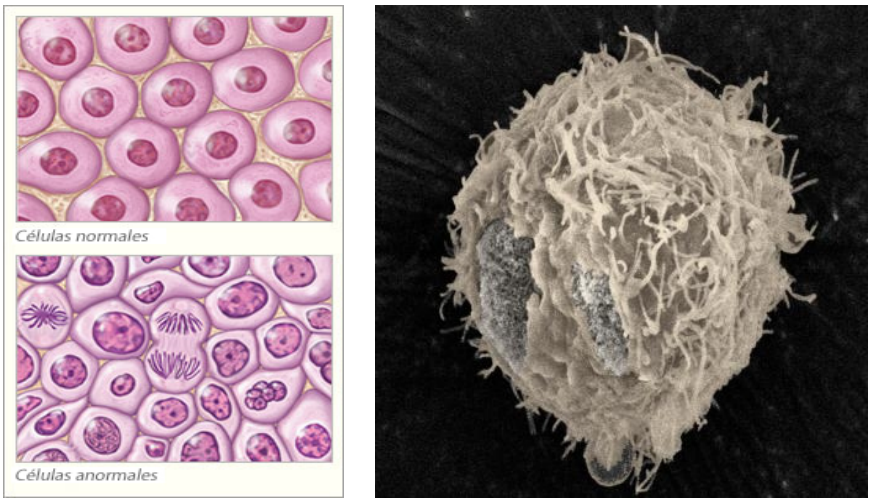


Figura 5: Imágenes de células sanas y cancerosas.

1.3.1. Marcadores de Tumores

Los marcadores de tumores, también conocidos como biomarcadores, son sustancias producidas por las células de tumores o por otras células del cuerpo como respuesta al cáncer o a ciertas afecciones benignas. Estas sustancias se pueden encontrar en la sangre, orina, tejidos de tumor o en otros tejidos. Distintos marcadores tumorales se encuentran en distintos tipos de cáncer, y la concentración de

un marcador tumoral específico varía dependiendo del tipo de cáncer. Además, las concentraciones de los marcadores tumorales no varían en todas las personas con cáncer, especialmente si el cáncer está en una etapa temprana. Pueden variar las concentraciones de algunos marcadores tumorales en pacientes con enfermedades no cancerosas.

Hasta la fecha, los investigadores han identificado más de doce sustancias que parecen expresarse en forma anormal cuando ciertos tipos de cáncer están presentes. También se pueden encontrar algunas de estas sustancias en otras afecciones o enfermedades.

1.3.2. Marcadores de Riesgo

Algunas personas tienen más probabilidad de padecer ciertos tipos de cáncer por haber sufrido un cambio, que se conoce como mutación o alteración, en algunos genes específicos. La presencia de dicho cambio se llama a veces un marcador de riesgo. Las pruebas para detectar los marcadores de riesgo ayudan al médico a estimar la probabilidad de que la persona padezca un cierto tipo de cáncer. Los marcadores de riesgo pueden indicar que es más probable que aparezca el cáncer, mientras que los marcadores tumorales pueden indicar la presencia del cáncer.

1.3.3. Utilización de los marcadores

Los marcadores tumorales se usan para detectar, diagnosticar y manejar ciertos tipos de cáncer. Aunque una concentración anormal de un marcador tumoral pueda sugerir la presencia de cáncer, esto, por sí mismo, no es suficiente para diagnosticar el cáncer. Por lo tanto, las mediciones de los marcadores tumorales se combinan usualmente con otras pruebas, como con una biopsia, para diagnosticar

el cáncer.

Se pueden medir las concentraciones de los marcadores tumorales antes del tratamiento para que los médicos puedan planificar una terapia adecuada. Para algunos tipos de cáncer, las concentraciones de los marcadores tumorales reflejan el estadio (etapa o extensión) de la enfermedad.

La concentración de un marcador tumoral puede usarse también para revisar cómo responde el paciente al tratamiento. Si la concentración disminuye o vuelve a ser normal, puede significar que el cáncer está respondiendo a la terapia. Mientras que un aumento puede indicar que el cáncer no está respondiendo. Después de que termina el tratamiento, la concentración del marcador tumoral puede usarse para vigilar una recidiva (el regreso del cáncer).

1.3.4. Procedimiento de medición

El médico toma una muestra de sangre, orina o tejido y la envía al laboratorio, donde se usan varios métodos para medir la concentración del marcador tumoral.

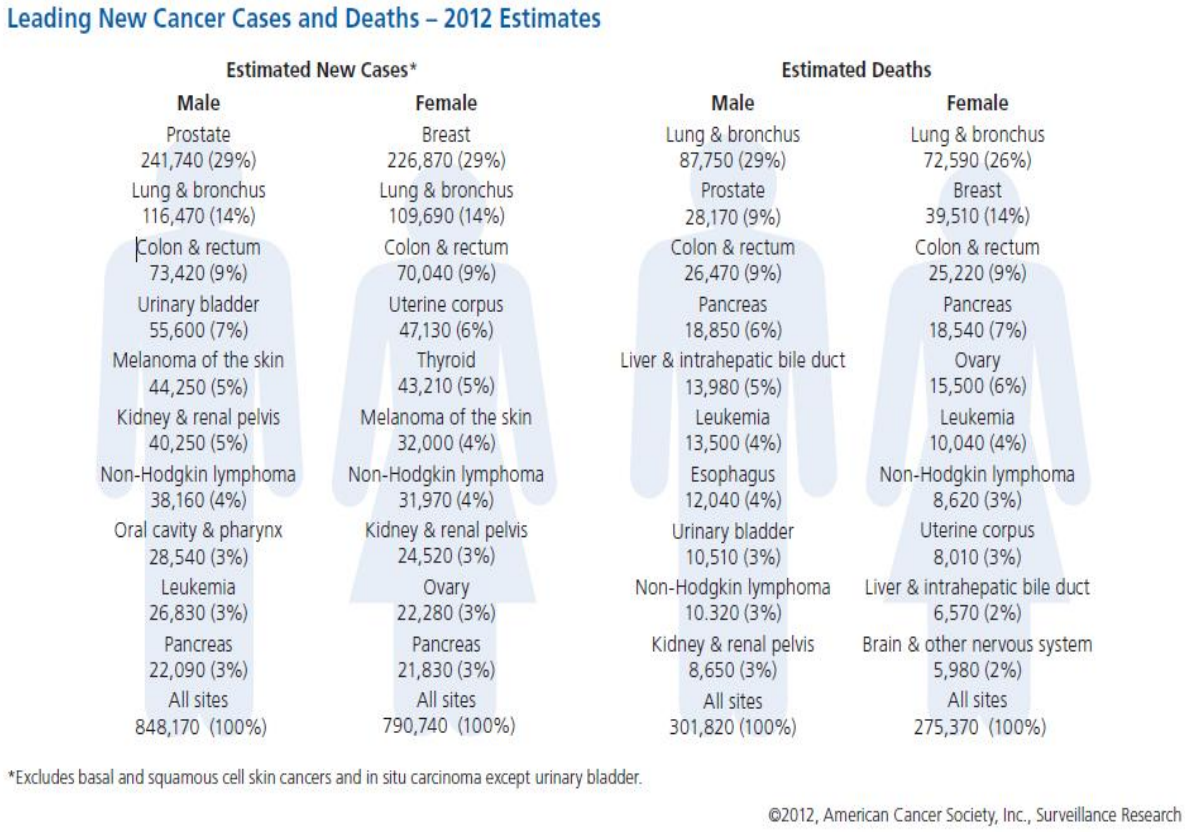
Cuando se usa el marcador tumoral para determinar si el tratamiento está funcionando o si ha regresado el cáncer, se miden generalmente las concentraciones del marcador tumoral durante un periodo de tiempo para ver si están subiendo o bajando. Casi siempre estas mediciones en serie tienen más significado que una sola medición. Las concentraciones de los marcadores tumorales se pueden medir cuando se diagnostica la enfermedad; antes, durante o después de la terapia; y periódicamente para vigilar que no haya recidiva.

1.3.5. Tipos de Cáncer

La lista de cánceres comunes incluye cánceres que se diagnostican con mayor frecuencia entre las mujeres y los hombres hispanos de los Estados Unidos. Las

estadísticas de incidencia de cáncer de la Sociedad Americana del Cáncer y de otras fuentes se usaron para crear esta lista. Para que un cáncer se considere común, el número estimado de casos nuevos para 2012 tiene que ser de 4,000 o más.

La *Figura 6* presenta los números proyectados para 2012 de casos nuevos de cáncer y muertes entre los hispanos en cada uno de los cánceres comunes. Los cánceres se han ordenado de acuerdo al número proyectado de casos nuevos; el cáncer con el número más elevado ocupa el primer renglón.



Leading New Cancer Cases and Deaths – 2012 Estimates

Estimated New Cases*		Estimated Deaths	
Male	**Female**	**Male**	**Female**
Prostate 241,740 (29%)	Breast 226,870 (29%)	Lung & bronchus 87,750 (29%)	Lung & bronchus 72,590 (26%)
Lung & bronchus 116,470 (14%)	Lung & bronchus 109,690 (14%)	Prostate 28,170 (9%)	Breast 39,510 (14%)
Colon & rectum 73,420 (9%)	Colon & rectum 70,040 (9%)	Colon & rectum 26,470 (9%)	Colon & rectum 25,220 (9%)
Urinary bladder 55,600 (7%)	Uterine corpus 47,130 (6%)	Pancreas 18,850 (6%)	Pancreas 18,540 (7%)
Melanoma of the skin 44,250 (5%)	Thyroid 43,210 (5%)	Liver & intrahepatic bile duct 13,980 (5%)	Ovary 15,500 (6%)
Kidney & renal pelvis 40,250 (5%)	Melanoma of the skin 32,000 (4%)	Leukemia 13,500 (4%)	Leukemia 10,040 (4%)
Non-Hodgkin lymphoma 38,160 (4%)	Non-Hodgkin lymphoma 31,970 (4%)	Esophagus 12,040 (4%)	Non-Hodgkin lymphoma 8,620 (3%)
Oral cavity & pharynx 28,540 (3%)	Kidney & renal pelvis 24,520 (3%)	Urinary bladder 10,510 (3%)	Uterine corpus 8,010 (3%)
Leukemia 26,830 (3%)	Ovary 22,280 (3%)	Non-Hodgkin lymphoma 10.320 (3%)	Liver & intrahepatic bile duct 6,570 (2%)
Pancreas 22,090 (3%)	Pancreas 21,830 (3%)	Kidney & renal pelvis 8,650 (3%)	Brain & other nervous system 5,980 (2%)
All sites 848,170 (100%)	All sites 790,740 (100%)	All sites 301,820 (100%)	All sites 275,370 (100%)

*Excludes basal and squamous cell skin cancers and in situ carcinoma except urinary bladder.

Figura 6: Datos de cánceres comúnes. Tabla publicada por la Sociedad Americana de Cáncer.

1.4. Hibridación Genómica Comparativa, CGH

A principio de los años 90, y con especial aplicación en tumores sólidos, se describió una nueva técnica de citogenética molecular[7]: la CGH. Esta técnica permite detectar cambios numéricos de secuencias de DNA (pérdidas, deleciones, ganancias y amplificaciones) en un tejido tumoral. Dicha técnica se basa en la hibridación[8] *in situ* del DNA tumoral y de un DNA control[9], marcados con fluorocromos[10] de distinto color sobre metafases normales. *Véase la Figura 7.*

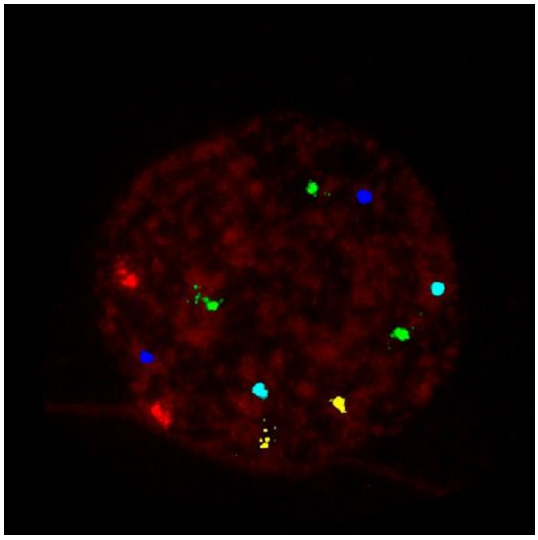


Figura 7: Fluorocromos en una célula.

Después de la hibridación, las variaciones numéricas del DNA tumoral se cuantifican mediante el coeficiente de intensidad de fluorescencia entre el DNA tumoral y el DNA control.

Esta técnica únicamente detecta cambios presentes en una proporción elevada

[7]El estudio de la estructura, función y comportamiento de los cromosomas.

[8]El proceso de unir dos hebras complementarias de DNA.

[9]DNA sano.

[10]Componente de una molécula que hace que ésta sea fluorescente.

de células tumorales (50 %). Por otra parte, no permite detectar translocaciones, inversiones y otras alteraciones de tipo equilibrado que no reflejan ganancias o pérdidas de material genético.

La realización de estas técnicas requiere de un sofisticado y caro soporte informático, por lo que su realización de forma rutinaria supone un gasto importante. En el momento actual, su uso debería regirse por una estrategia razonada, y aplicarse únicamente en casos muy concretos.

Esta técnica presenta el problema de que sólo detecta el número copiado (intensidad del color) en segmentos grandes; esta limitación se resolvió mediante los arreglos de Hibridación Genómica Comparativa (aCGH).[2]

1.5. Arreglos de Hibridación Genómica Comparativa, aCGH

Esta técnica permite realizar un análisis de la secuencia del número copiado en el genoma entero en un sólo experimento. Se analizan las intensidades de cada una de las sondas formadas por cromosomas artificiales bacterianos[11] mejores conocidos como **BACs**, las intensidades de los controles T_z y las intensidades de los casos R_z para cada sonda z en el cromosoma.

Estas características se transforman en variables X_z, para z=1,...,k, con X_z=$log_2(\frac{T_z}{R_z})$ y k el número de sondas.

Dadas las características de la forma en que se duplica la información del DNA, se observa la conveniencia de la transformación que representa $log_2(\bullet)$ al suavizar la forma de los datos y tomando un punto de referencia perfectamente marcado.

La naturaleza de los datos refleja tres escenarios importantes:

[11]Segmentos largos del DNA.

- No se observa alteración alguna en el número copiado en una región del DNA, *i.e.* $X_z = log_2(\frac{2}{2}) = 0$.

- Si se observa pérdida en algún número copiado, se tendría

  $X_z = log_2(\frac{1}{2}) = -1$.

- En el caso de que haya una ganancia, se tendría $X_z = log_2(\frac{3}{2}) = 0.58$.

Ahora, la siguiente parte en la que hay que enfocarse es detectar las regiones con número copiado común y los puntos de cambio, para ello se recurre a métodos estadísticos.

1.6. Puntos de Cambio

Sea X_i, con $i = \overline{1,k}$, una secuencia de variables aleatorias.

Un índice v es llamado **punto de cambio** si $X_1,\ldots,X_v$ tienen una función de distribución común F_0 y $X_{v+1},\ldots,X_k$ tienen una función de distribución diferente común F_1 hasta el siguiente punto de cambio (si éste existe).

En los estudios de arreglos de números copiados la información analizada es naturalmente ordenada por la localización de las sondas a lo largo del cromosoma de interés.

Es necesario observar que pueden haber múltiples puntos de cambio en un cromosoma determinado, cada uno correspondiente a un cambio en el número de copias en la muestra; el objetivo principal es identificar todos los puntos de cambio que luego particionarán al cromosoma en segmentos en un número copiado constante. Una vez que el cromosoma se divide, se puede estimar el número copiado de los segmentos con la ayuda de información adicional, como la ploidía[12]

[12]Número de juegos completos de cromosomas en una célula.

del cromosoma. Esto proporcionará la localización de los problemas en el número copiado.

2. Segmentación Binaria

2.1. Desarrollo de la técnica

Como se ha mencionado anteriormente, la técnica de Segmentación Binaria tiene el objetivo de segmentar recursivamente a un cromosoma en regiones con el mismo número copiado; esto se realizará a través del uso del contraste de pruebas de hipótesis sobre las medias de las variables aleatorias a trabajar. Es decir,

H_0: No existe punto de cambio en la posición v y H_1: Existe un punto de cambio en la posición v.

Sea X_i, con $i = \overline{1,k}$, una secuencia de variables aleatorias independientes.

Si existe un punto de cambio en la región estudiada en la posición v, la media de las variables aleatorias X_1,...,X_v será distinta a la media de las variables aleatorias X_{v+1},...,X_k suponiendo que la varianza es desconocida; mientras que ocurrirá lo contrario si no existe un punto de cambio.[2]

2.1.1. Segmentación Binaria bajo Normalidad con varianza desconocida

Supóngase que $X_i \sim N(\mu_i, \sigma^2)$ para $i = \overline{1,k}$. Téngase en mente que si hay un punto de cambio en la posición v, la media de las variables aleatorias $X_1,\ldots,$ X_v será diferente a la media de las variables aleatorias $X_{v+1},\ldots,X_k$, esto es, $X_y \sim N(\mu, \sigma^2)$ para $y = \overline{1,v}$ y $X_y \sim N(\mu^*, \sigma^2)$ para $y = \overline{v+1,k}$. Entonces, lo que resta es contrastar $H_0 : \mu = \mu^*$, $\sigma^2 > 0$ v_s $H_1 : \mu \neq \mu^*$, $\sigma^2 > 0$.

Primeramente es necesario definir el espacio parametral...

$\Theta = \{(\mu, \mu^*, \sigma^2) \,|\mu, \mu^* \in \Re, \sigma^2 > 0\}$

Entonces...

$\Theta_0 = \{(\mu, \sigma^2) \,|\mu \in \Re, \sigma^2 > 0\}$

$\Theta_1 = \Theta \setminus \Theta_0$

Ahora, al utilizar la prueba de la razón de verosimilitudes generalizada se tiene que:

$$0 \leq \lambda = \frac{\underset{\theta \in \Theta_0}{\text{máx}}\, L(\theta)}{L(\theta_{MV})} \leq 1$$

donde θ es el vector de los parámetros definidos anteriormente, $L(\theta) = \prod_{i=1}^{k} f(x_i, \theta)$ es la función de verosomilitud y θ_{MV} es el estimador máximo verosímil.

Entonces, se sigue que:

$$L(\theta) = \prod_{i=1}^{k} f(x_i, \theta) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{k}{2}} e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{v}(X_i-\mu)^2 + \sum_{i=v+1}^{k}(X_i-\mu^*)^2\right]}$$

Aplicando logaritmo natural:

$$l = lnL(\theta) = \frac{k}{2} ln\left(\frac{1}{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{v}(X_i-\mu)^2 + \sum_{i=v+1}^{k}(X_i-\mu^*)^2\right]$$

$$= -\frac{k}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{v}(X_i-\mu)^2 + \sum_{i=v+1}^{k}(X_i-\mu^*)^2\right]$$

Entonces:

$$\frac{\partial l}{\partial \sigma^2} = -\frac{k}{2} \cdot \frac{\cancel{2\pi}}{\cancel{2\pi}\sigma^2} + \frac{1}{2\sigma^4}\left[\sum_{i=1}^{v}(X_i-\mu)^2 + \sum_{i=v+1}^{k}(X_i-\mu^*)^2\right]$$

Ahora, se busca que la derivada parcial sea cero, entonces...

$$\frac{\partial l}{\partial \sigma^2} = 0 \Longleftrightarrow -\frac{k}{2\sigma^2} + \frac{1}{2\sigma^4}\left[\sum_{i=1}^{v}(X_i-\mu)^2 + \sum_{i=v+1}^{k}(X_i-\mu^*)^2\right] = 0$$

$$\Longleftrightarrow \tfrac{k}{2\sigma^2} = \tfrac{1}{2\sigma^4}\left[\sum_{i=1}^{v}(X_i-\mu)^2 + \sum_{i=v+1}^{k}(X_i-\mu^*)^2\right]$$

$$\iff k\sigma^2 = \sum_{i=1}^{v}(X_i-\mu)^2 + \sum_{i=v+1}^{k}(X_i-\mu^*)^2$$

$$\iff \widehat{\sigma}^2_{MV} = \tfrac{1}{k}\left[\sum_{i=1}^{v}(X_i-\widehat{\mu}_{MV})^2 + \sum_{i=v+1}^{k}(X_i-\widehat{\mu}^*_{MV})^2\right]$$

Luego:

$$\frac{\partial l}{\partial \mu} = -\frac{1}{2\sigma^2}(2)(-1)\sum_{i=1}^{v}(X_i-\mu) = \frac{1}{\sigma^2}\left[\sum_{i=1}^{v}X_i - v\mu\right]$$

$$\Rightarrow \tfrac{\partial l}{\partial \mu} = 0$$

$$\iff \tfrac{1}{\sigma^2}\left[\sum_{i=1}^{v}X_i - v\mu\right] = 0$$

$$\iff \sum_{i=1}^{v}X_i - v\mu = 0 \iff \sum_{i=1}^{v}X_i = v\mu \iff \widehat{\mu}_{MV} = \frac{1}{v}\sum_{i=1}^{v}X_i = \overline{X}_v$$

Y:

$$\frac{\partial l}{\partial \mu^*} = -\frac{1}{2\sigma^2}(2)(-1)\sum_{i=v+1}^{k}(X_i-\mu^*) = \frac{1}{\sigma^2}\left[\sum_{i=v+1}^{k}X_i - (k-v)\mu^*\right]$$

$$\Rightarrow \tfrac{\partial l}{\partial \mu^*} = 0$$

$$\iff \tfrac{1}{\sigma^2}\left[\sum_{i=v+1}^{k}X_i - (k-v)\mu^*\right] = 0$$

$$\iff \sum_{i=v+1}^{k}X_i - (k-v)\mu^* = 0 \iff \sum_{i=v+1}^{k}X_i = (k-v)\mu^*$$

$$\iff \widehat{\mu^*}_{MV} = \tfrac{1}{k-v}\sum_{i=v+1}^{k}X_i = \overline{X}_{k-v}$$

Al sustituir las expresiones anteriores en $L(\theta)$ se tiene:

$$L(\widehat{\theta}) = \left(\frac{1}{2\pi\widehat{\sigma^2}}\right)^{\frac{k}{2}} e^{-\frac{1}{2\widehat{\sigma^2}}\left[\sum_{i=1}^{v}(X_i-\widehat{\mu})^2 + \sum_{i=v+1}^{k}\left(X_i-\widehat{\mu^*}\right)^2\right]}$$

Bajo $H_0\ldots$

$$L(\theta_0) = \left(\frac{1}{2\pi\sigma_0^2}\right)^{\frac{k}{2}} e^{-\frac{1}{2\sigma_0^2}\left[\sum_{i=1}^{v}(X_i-\mu_0)^2 + \sum_{i=v+1}^{k}(X_i-\mu_0)^2\right]}$$

Aplicando la función logarítmo natural:

$$l = lnL(\theta_0) = \frac{k}{2} ln\left(\frac{1}{2\pi\sigma_0^2}\right) - \frac{1}{2\sigma_0^2}\left[\sum_{i=1}^{v}(X_i-\mu_0)^2 + \sum_{i=v+1}^{k}(X_i-\mu_0)^2\right]$$

Así:

$$\frac{\partial l}{\partial \sigma_{0^2}} = -\frac{k}{2}\cdot\frac{\cancel{2\pi}}{\cancel{2\pi}\sigma_0^2} + \frac{1}{2\sigma^4}\left[\sum_{i=1}^{v}(X_i-\mu_0)^2 + \sum_{i=v+1}^{k}(X_i-\mu_0)^2\right]$$

Ahora, se busca que la derivada parcial sea cero, entonces...

$$\frac{\partial l}{\partial \sigma_0^2} = 0 \Longleftrightarrow -\frac{k}{2\sigma_0^2} + \frac{1}{2\sigma_0^4}\left[\sum_{i=1}^{v}(X_i-\mu_0)^2 + \sum_{i=v+1}^{k}(X_i-\mu_0)^2\right] = 0$$

$$\Longleftrightarrow \tfrac{k}{2\sigma_0^2} = \tfrac{1}{2\sigma_0^4}\left[\sum_{i=1}^{v}(X_i-\mu_0)^2 + \sum_{i=v+1}^{k}(X_i-\mu_0)^2\right]$$

$$\Longleftrightarrow k\sigma_0^2 = \sum_{i=1}^{v}(X_i-\mu_0)^2 + \sum_{i=v+1}^{k}(X_i-\mu_0)^2$$

$$\Longleftrightarrow \widehat{\sigma_0}^2_{MV} = \tfrac{1}{k}\left[\sum_{i=1}^{v}(X_i-\widehat{\mu}_{0MV})^2 + \sum_{i=v+1}^{k}(X_i-\widehat{\mu}_{0MV})^2\right]$$

Y:

$$\frac{\partial l}{\partial \mu_0} = -\frac{1}{2\sigma_0^2}(2)(-1)\sum_{i=1}^{v}(X_i-\mu_0) - \frac{1}{2\sigma_0^2}(2)(-1)\sum_{i=v+1}^{k}(X_i-\mu_0)$$

$$= \frac{1}{\sigma_0^2}\left[\sum_{i=1}^{v}(X_i-\mu_0) + \sum_{i=v+1}^{k}(X_i-\mu_0)\right]$$

$$\Rightarrow \tfrac{\partial l}{\partial \mu_0} = 0$$

$$\Longleftrightarrow \tfrac{1}{\sigma_0^2}\left[\sum_{i=1}^{v}(X_i-\mu_0) + \sum_{i=v+1}^{k}(X_i-\mu_0)\right] = 0$$

$$\Longleftrightarrow \sum_{i=1}^{v}X_i - v\mu + \sum_{i=v+1}^{k}X_i - (k-v)\mu = 0 \Longleftrightarrow \sum_{i=1}^{v}X_i + \sum_{i=v+1}^{k}X_i = v\mu + (k-v)\mu$$

$$\therefore \widehat{\mu}_{MV} = \frac{1}{k}\sum_{i=1}^{k}X_i = \overline{X}_k$$

Sustituyendo adecuadamente, se obtiene:

$$L(\hat{\theta}_0) = \left(\frac{1}{2\pi\hat{\sigma_0}^2}\right)^{\frac{k}{2}} e^{-\frac{1}{2\hat{\sigma_0^2}}\left[\sum_{i=1}^{v}(X_i-\hat{\mu}_0)^2+\sum_{i=v+1}^{k}(X_i-\hat{\mu}_0)^2\right]}$$

Ahora,

$$\lambda = \frac{\left(\frac{1}{2\pi\hat{\sigma_0^2}}\right)^{\frac{k}{2}} e^{-\frac{1}{2\hat{\sigma_0^2}}\left[\sum_{i=1}^{v}(X_i-\hat{\mu}_0)^2+\sum_{i=v+1}^{k}(X_i-\hat{\mu}_0)^2\right]}}{\left(\frac{1}{2\pi\hat{\sigma_{MV}}^2}\right)^{\frac{k}{2}} e^{-\frac{1}{2\hat{\sigma_{MV}}^2}\left[\sum_{i=1}^{v}(X_i-\hat{\mu}_{MV})^2+\sum_{i=v+1}^{k}(X_i-\hat{\mu}^*_{MV})^2\right]}}$$

Sean α el numerador y β el denominador de la razón anterior.

Entonces, para α:

$$\left[\frac{1}{\frac{(2\pi)\left(\sum_{i=1}^{v}(X_i-\hat{\mu}_0)^2+\sum_{i=v+1}^{k}(X_i-\hat{\mu}_0)^2\right)}{k}}\right]^{\frac{k}{2}} e^{\left(-\frac{1}{2}\right)\left(\frac{\sum_{i=1}^{v}(X_i-\hat{\mu}_0)^2+\sum_{i=v+1}^{k}(X_i-\hat{\mu}_0)^2}{\sum_{i=1}^{v}(X_i-\hat{\mu}_0)^2+\sum_{i=v+1}^{k}(X_i-\hat{\mu}_0)^2}\right)}{}_{k}$$

$$= \left[\frac{k}{(2\pi)\left(\sum_{i=1}^{v}(X_i-\hat{\mu}_0)^2+\sum_{i=v+1}^{k}(X_i-\hat{\mu}_0)^2\right)}\right]^{\frac{k}{2}} e^{-\frac{1}{2}(k)}$$

$$= \left[\frac{ke^{-1}}{(2\pi)\left(\sum_{i=1}^{v}(X_i-\hat{\mu}_0)^2+\sum_{i=v+1}^{k}(X_i-\hat{\mu}_0)^2\right)}\right]^{\frac{k}{2}}$$

Y para β:

$$\left[\frac{1}{\frac{(2\pi)\left(\sum_{i=1}^{v}(X_i-\hat{\mu})^2+\sum_{i=v+1}^{k}(X_i-\hat{\mu}^*)^2\right)}{k}}\right]^{\frac{k}{2}} e^{(-\frac{1}{2})\left(\frac{\sum_{i=1}^{v}(X_i-\hat{\mu})^2+\sum_{i=v+1}^{k}(X_i-\hat{\mu}^*)^2}{\frac{\sum_{i=1}^{v}(X_i-\hat{\mu})^2+\sum_{i=v+1}^{k}(X_i-\hat{\mu}^*)^2}{k}}\right)}$$

$$= \left[\frac{k}{(2\pi)\left(\sum_{i=1}^{v}(X_i-\hat{\mu})^2+\sum_{i=v+1}^{k}(X_i-\hat{\mu}^*)^2\right)}\right]^{\frac{k}{2}} e^{-\frac{1}{2}(k)}$$

$$= \left[\frac{ke^{-1}}{(2\pi)\left(\sum_{i=1}^{v}(X_i-\hat{\mu})^2+\sum_{i=v+1}^{k}(X_i-\hat{\mu}^*)^2\right)}\right]^{\frac{k}{2}}$$

Entonces:

$$\lambda = \frac{\left[\frac{ke^{-1}}{(2\pi)\left(\sum_{i=1}^{v}(X_i-\hat{\mu}_0)^2+\sum_{i=v+1}^{k}(X_i-\hat{\mu}_0)^2\right)}\right]^{\frac{k}{2}}}{\left[\frac{ke^{-1}}{(2\pi)\left(\sum_{i=1}^{v}(X_i-\hat{\mu})^2+\sum_{i=v+1}^{k}(X_i-\hat{\mu}^*)^2\right)}\right]^{\frac{k}{2}}}$$

$$= \left[\frac{\frac{ke^{-1}}{(2\pi)\left(\sum_{i=1}^{v}(X_i-\hat{\mu}_0)^2+\sum_{i=v+1}^{k}(X_i-\hat{\mu}_0)^2\right)}}{\frac{ke^{-1}}{(2\pi)\left(\sum_{i=1}^{v}(X_i-\hat{\mu})^2+\sum_{i=v+1}^{k}(X_i-\hat{\mu}^*)^2\right)}} \right]^{\frac{k}{2}}$$

$$= \left[\frac{\sum_{i=1}^{v}(X_i-\hat{\mu})^2+\sum_{i=v+1}^{k}(X_i-\hat{\mu}^*)^2}{\sum_{i=1}^{v}(X_i-\hat{\mu}_0)^2+\sum_{i=v+1}^{k}(X_i-\hat{\mu}_0)^2} \right]^{\frac{k}{2}}$$

$$= \left[\frac{\sum_{i=1}^{v}\left(X_i-\bar{X}_v\right)^2+\sum_{i=v+1}^{k}\left(X_i-\bar{X}_{k-v}\right)^2}{\sum_{i=1}^{v}(X_i-\hat{\mu}_0)^2+\sum_{i=v+1}^{k}(X_i-\hat{\mu}_0)^2} \right]^{\frac{k}{2}}$$

Para poder formar una región de rechazo apropiada y obtener las características esenciales para la construcción de la función de distribución, es necesario unificar criterios y procedimientos para simplificar esta tarea. Es importante ver cómo se comporta el numerador con respecto al denominador para proceder con simplificación y agrupación algebráica.

Centrándose en el denominador se tiene que:

$$\sum_{i=1}^{v}(X_i-\hat{\mu}_0)^2 = \sum_{i=1}^{v}\left\{\left(X_i-\bar{X}_v\right)-\left(\hat{\mu}_0-\bar{X}_v\right)\right\}^2$$

$$= \sum_{i=1}^{v}\left\{\left(X_i-\bar{X}_v\right)^2-2\left(X_i-\bar{X}_v\right)\left(\hat{\mu}_0-\bar{X}_v\right)+\left(\hat{\mu}_0-\bar{X}_v\right)^2\right\}$$

$$= \sum_{i=1}^{v} \left(X_i - \bar{X}_v\right)^2 - 2\left(\hat{\mu}_0 - \bar{X}_v\right) \sum_{i=1}^{v} \left(X_i - \bar{X}_v\right) + v\left(\hat{\mu}_0 - \bar{X}_v\right)^2$$

Pero,

$$\sum_{i=1}^{v} \left(X_i - \bar{X}_v\right) = \sum_{i=1}^{v} X_i - v\bar{X}_v = \sum_{i=1}^{v} X_i - \sum_{i=1}^{v} X_i = 0$$

$$\Rightarrow \sum_{i=1}^{v} \left(X_i - \hat{\mu}_0\right)^2 = \sum_{i=1}^{v} \left(X_i - \bar{X}_v\right)^2 + v\left(\hat{\mu}_0 - \bar{X}_v\right)^2$$

Ahora,

$$\hat{\mu}_0 = \frac{1}{k} \sum_{i=1}^{k} X_i = \frac{1}{v + (k - v)} \left(\sum_{i=1}^{v} X_i + \sum_{i=v+1}^{k} X_i \right)$$

$$= \frac{v\bar{X}_v + (k - v)\, \bar{X}_{k-v}}{v + (k - v)} = \frac{v\bar{X}_v}{v + (k - v)} + \frac{(k - v)\, \bar{X}_{k-v}}{v + (k - v)}$$

Entonces,

$$v\left(\hat{\mu}_0 - \bar{X}_v\right)^2 = v\left[\bar{X}_v - \frac{v\bar{X}_v}{v + (k - v)} - \frac{(k - v)\, \bar{X}_{k-v}}{v + (k - v)}\right]^2$$

$$= v\left[\frac{(k - v)\left(\bar{X}_v - \bar{X}_{k-v}\right)}{v + (k - v)}\right]^2 = \frac{v\,(k - v)^2 \left(\bar{X}_v - \bar{X}_{k-v}\right)^2}{\left(v + (k - v)\right)^2}$$

Ahora, análogamente se tiene que:

$$\sum_{i=v+1}^{k} \left(X_i - \hat{\mu}_0\right)^2 = \sum_{i=v+1}^{k} \left(X_i - \bar{X}_{k-v}\right)^2 + (k - v)\left(\bar{X}_{k-v} - \hat{\mu}_0\right)^2$$

Y:

$$(k - v)\left(\bar{X}_{k-v} - \hat{\mu}_0\right)^2 = \frac{v^2\,(k - v)\left(\bar{X}_v - \bar{X}_{k-v}\right)^2}{\left(v + (k - v)\right)^2}$$

De lo anterior se deduce que:

$$v\left(\hat{\mu}_0 - \bar{X}_v\right)^2 + (k - v)\left(\bar{X}_{k-v} - \hat{\mu}_0\right)^2 = \left(\bar{X}_v - \bar{X}_{k-v}\right)^2 \left[\frac{v\,(k - v)}{v + (k - v)}\right]$$

Dado lo anterior, se puede construir la región crítica:

$$\left[\frac{\sum_{i=1}^{v}\left(X_i-\bar{X}_v\right)^2+\sum_{i=v+1}^{k}\left(X_i-\bar{X}_{k-v}\right)^2}{\sum_{i=1}^{v}\left(X_i-\bar{X}_v\right)^2+\sum_{i=v+1}^{k}\left(X_i-\bar{X}_{k-v}\right)^2+\left(\bar{X}_v-\bar{X}_{k-v}\right)^2\left[\frac{v\,(k-v)}{v+(k-v)}\right]}\right]^{\frac{k}{2}} \leq C_\alpha$$

donde C_α es una constante que depende del nivel de significancia. Entonces,

$$\left[\frac{\sum_{i=1}^{v}\left(X_i-\bar{X}_v\right)^2+\sum_{i=v+1}^{k}\left(X_i-\bar{X}_{k-v}\right)^2+\left(\bar{X}_v-\bar{X}_{k-v}\right)^2\left[\frac{v\,(k-v)}{v+(k-v)}\right]}{\sum_{i=1}^{v}\left(X_i-\bar{X}_v\right)^2+\sum_{i=v+1}^{k}\left(X_i-\bar{X}_{k-v}\right)^2}\right] \geq C'_\alpha$$

$$\Leftrightarrow 1+\frac{\left(X_i-\bar{X}_{k-v}\right)^2\left[\frac{v(k-v)}{v+(k-v)}\right]}{\sum_{i=1}^{v}\left(X_i-\bar{X}_v\right)^2+\sum_{i=v+1}^{k}\left(X_i-\bar{X}_{k-v}\right)^2} \geq C'_\alpha$$

$$\Leftrightarrow \frac{\left(X_i-\bar{X}_{k-v}\right)^2\left[\frac{v(k-v)}{v+(k-v)}\right]}{\sum_{i=1}^{v}\left(X_i-\bar{X}_v\right)^2+\sum_{i=v+1}^{k}\left(X_i-\bar{X}_{k-v}\right)^2} \geq C''_\alpha$$

$$\Leftrightarrow \frac{\frac{\left(X_i-\bar{X}_{k-v}\right)^2}{\frac{1}{v}+\frac{1}{k-v}}}{\sum_{i=1}^{v}\left(X_i-\bar{X}_v\right)^2+\sum_{i=v+1}^{k}\left(X_i-\bar{X}_{k-v}\right)^2} \geq C''_\alpha$$

$$\Leftrightarrow \frac{\frac{\left(X_i-\bar{X}_{k-v}\right)}{\sqrt{\frac{1}{v}+\frac{1}{k-v}}}}{\sqrt{\sum_{i=1}^{v}\left(X_i-\bar{X}_v\right)^2+\sum_{i=v+1}^{k}\left(X_i-\bar{X}_{k-v}\right)^2}} \geq C'''_\alpha$$

$$\Leftrightarrow \frac{\frac{\left(X_i-\bar{X}_{k-v}\right)}{\sqrt{\frac{1}{v}+\frac{1}{k-v}}}}{\sqrt{\frac{\sum_{i=1}^{v}\left(X_i-\bar{X}_v\right)^2+\sum_{i=v+1}^{k}\left(X_i-\bar{X}_{k-v}\right)^2}{k-2}}} \geq C_\alpha^{IV}$$

Para poder encontrar el valor de la constante, es necesario verificar la distribución de λ, para esto (bajo H_0):

Como se vió en los supuestos,

$$\bar{X}_v \sim N\left(\mu, \frac{\sigma^2}{v}\right)$$

$$\bar{X}_{k-v} \sim N\left(\mu, \frac{\sigma^2}{k-v}\right)$$

$$\Rightarrow \bar{X}_v - \bar{X}_{k-v} \sim N\left(0, \sigma^2\left(\frac{v+(k-v)}{v\,(k-v)}\right)\right)$$

Y:

$$\frac{\bar{X}_v - \bar{X}_{k-v}}{\sigma\sqrt{\frac{1}{v}+\frac{1}{k-v}}} \sim N\,(0,1)$$

Por otro lado, se sabe que:

$$\frac{\sum_{i=1}^{v}\left(X_i-\bar{X}_v\right)^2}{\sigma^2} \sim \chi^2_{(v-1)}$$

$$\frac{\sum_{i=v+1}^{k}\left(X_i-\bar{X}_{k-v}\right)^2}{\sigma^2} \sim \chi^2_{(k-v-1)}$$

Entonces, por independencia:

$$\frac{\sum_{i=1}^{v}\left(X_i-\bar{X}_v\right)^2}{\sigma^2} + \frac{\sum_{i=v+1}^{k}\left(X_i-\bar{X}_{k-v}\right)^2}{\sigma^2} \sim \chi^2_{(k-2)}$$

Y así,

$$Z_v = \frac{\frac{\left(X_i - \bar{X}_{k-v}\right)}{\sqrt{\frac{1}{v}+\frac{1}{k-v}}}}{\sqrt{\frac{\sum_{i=1}^{v}\left(X_i - \bar{X}_v\right)^2 + \sum_{i=v+1}^{k}\left(X_i - \bar{X}_{k-v}\right)^2}{k-2}}} \sim t_{(k-2)}$$

Como es apreciable, básicamente se trata de la prueba t para diferencia de medias con σ^2 desconocida en dos muestras Normales independientes.

En caso de que se quisiera probar la no existencia de puntos de cambio contra la hipótesis de la existencia de puntos de cambio, entonces se utilizaría la estadística $Z_B = \max_{1 \leq v \leq k} |Z_v|$. Si la hipótesis nula es rechazada, el punto de cambio sería v tal que $Z_B = |Z_v|$.

2.1.2. Segmentación Binaria bajo Normalidad con varianza conocida

Supóngase que los datos tienen una distribución normal con varianza conocida e igual a *uno*. En este caso,

$$L(\theta) = \left(\frac{1}{2\pi}\right)^{\frac{k}{2}} e^{-\frac{1}{2}\left[\sum_{i=1}^{v}(X_i - \mu)^2 + \sum_{i=v+1}^{k}(X_i - \mu^*)^2\right]}$$

y,

$$L(\theta_0) = \left(\frac{1}{2\pi}\right)^{\frac{k}{2}} e^{-\frac{1}{2}\left[\sum_{i=1}^{v}(X_i - \mu_0)^2 + \sum_{i=v+1}^{k}(X_i - \mu_0)^2\right]}$$

De la misma forma que en el caso anterior descrito, se tiene que:

$$\widehat{\mu} = \frac{1}{v}\sum_{i=1}^{v}(X_i) = \overline{X}_v$$

$$\widehat{\mu^*} = \frac{1}{k-v}\sum_{i=v+1}^{k}(X_i) = \overline{X}_{k-v}$$

$$\widehat{\mu_0} = \frac{1}{k}\left(\sum_{i=1}^{v}(X_i) + \sum_{i=v+1}^{k}(X_i)\right) = \frac{v\bar{X}_v}{v+(k-v)} + \frac{(k-v)\bar{X}_{k-v}}{v+(k-v)}$$

Entones,

$$\lambda = \frac{\cancel{\left(\frac{1}{2\pi}\right)^{\frac{k}{2}}} e^{-\frac{1}{2}\left[\sum_{i=1}^{v}(X_i-\widehat{\mu}_0)^2 + \sum_{i=v+1}^{k}(X_i-\widehat{\mu}_0)^2\right]}}{\cancel{\left(\frac{1}{2\pi}\right)^{\frac{k}{2}}} e^{-\frac{1}{2}\left[\sum_{i=1}^{v}(X_i-\widehat{\mu})^2 + \sum_{i=v+1}^{k}(X_i-\widehat{\mu}^*)^2\right]}}$$

$$= \frac{e^{-\frac{1}{2}\left[\sum_{i=1}^{v}(X_i-\widehat{\mu}_0)^2 + \sum_{i=v+1}^{k}(X_i-\widehat{\mu}_0)^2\right]}}{e^{-\frac{1}{2}\left[\sum_{i=1}^{v}(X_i-\widehat{\mu})^2 + \sum_{i=v+1}^{k}(X_i-\widehat{\mu}^*)^2\right]}}$$

$$= \frac{e^{\frac{1}{2}\left[\sum_{i=1}^{v}(X_i-\widehat{\mu})^2 + \sum_{i=v+1}^{k}(X_i-\widehat{\mu}^*)^2\right]}}{e^{\frac{1}{2}\left[\sum_{i=1}^{v}(X_i-\widehat{\mu}_0)^2 + \sum_{i=v+1}^{k}(X_i-\widehat{\mu}_0)^2\right]}}$$

$$= e^{\frac{1}{2}\left[\sum_{i=1}^{v}(X_i-\widehat{\mu})^2 + \sum_{i=v+1}^{k}(X_i-\widehat{\mu}^*)^2 - \sum_{i=1}^{v}(X_i-\widehat{\mu}_0)^2 - \sum_{i=v+1}^{k}(X_i-\widehat{\mu}_0)^2\right]}$$

Entonces, para encontrar la región de rechazo:

$$\lambda = e^{\frac{1}{2}\left[\sum_{i=1}^{v}(X_i-\widehat{\mu})^2 + \sum_{i=v+1}^{k}(X_i-\widehat{\mu}^*)^2 - \sum_{i=1}^{v}(X_i-\widehat{\mu}_0)^2 - \sum_{i=v+1}^{k}(X_i-\widehat{\mu}_0)^2\right]} \leq C_\alpha$$

$$\Leftrightarrow \frac{1}{2}\left[\sum_{i=1}^{v}(X_i-\widehat{\mu})^2 + \sum_{i=v+1}^{k}(X_i-\widehat{\mu}^*)^2 - \sum_{i=1}^{v}(X_i-\widehat{\mu}_0)^2 - \sum_{i=v+1}^{k}(X_i-\widehat{\mu}_0)^2\right] \leq C'_\alpha$$

$$\Leftrightarrow \left[\sum_{i=1}^{v}(X_i-\widehat{\mu})^2 + \sum_{i=v+1}^{k}(X_i-\widehat{\mu}^*)^2 - \sum_{i=1}^{v}(X_i-\widehat{\mu}_0)^2 - \sum_{i=v+1}^{k}(X_i-\widehat{\mu}_0)^2\right] \leq C''_\alpha$$

Sustituyendo las estimaciones de μ y de μ^*:

$$\left[\sum_{i=1}^{v}(X_i-\bar{X}_v)^2+\sum_{i=v+1}^{k}(X_i-\bar{X}_{k-v})^2-\sum_{i=1}^{v}(X_i-\widehat{\mu}_0)^2-\sum_{i=v+1}^{k}(X_i-\widehat{\mu}_0)^2\right]\leq C_\alpha''$$

Recordando que:

$$\sum_{i=1}^{v}(X_i-\widehat{\mu}_0)^2=\sum_{i=1}^{v}(X_i-\bar{X}_v)^2+v(\widehat{\mu}_0-\bar{X}_v)^2$$

y

$$\sum_{i=v+1}^{k}(X_i-\bar{\mu}_0)^2=\sum_{i=v+1}^{k}(X_i-\bar{X}_{k-v})^2+(k-v)(\widehat{\mu}_0-\bar{X}_{k-v})^2$$

Se obtiene que:

$$-v(\widehat{\mu}_0-\bar{X}_v)^2-(k-v)(\widehat{\mu}_0-\bar{X}_{k-v})^2\leq C_\alpha''$$

$$\Leftrightarrow v(\widehat{\mu}_0-\bar{X}_v)^2+(k-v)(\widehat{\mu}_0-\bar{X}_{k-v})^2\geq C_\alpha'''$$

Además se sabe que:

$$v(\widehat{\mu}_0-\bar{X}_v)^2+(k-v)(\widehat{\mu}_0-\bar{X}_{k-v})^2=(\bar{X}_v-\bar{X}_{k-v})^2\left[\frac{v(k-v)}{v+(k-v)}\right]$$

Entonces,

$$(\bar{X}_v-\bar{X}_{k-v})^2\left[\frac{v(k-v)}{v+(k-v)}\right]\geq C_\alpha'''$$

$$\Leftrightarrow \frac{(\bar{X}_v-\bar{X}_{k-v})^2}{\frac{1}{v}+\frac{1}{k-v}}\geq C_\alpha'''$$

$$\Leftrightarrow \frac{\bar{X}_v-\bar{X}_{k-v}}{\sqrt{\frac{1}{v}+\frac{1}{k-v}}}\geq C_\alpha''''$$

Para poder encontrar el valor de la constante, es necesario verificar la distribución de λ, para esto (bajo H_0):

Como se vio en los supuestos,

$$\bar{X}_v \sim N\left(\mu, \frac{1}{v}\right)$$

$$\bar{X}_{k-v} \sim N\left(\mu, \frac{1}{k-v}\right)$$

$$\Rightarrow \bar{X}_v - \bar{X}_{k-v} \sim N\left(0, \frac{v+(k-v)}{v\,(k-v)}\right)$$

Entonces,

$$Z_v = \frac{\left(\overline{X}_v - \overline{X}_{k-v}\right)}{\sqrt{\frac{1}{v} + \frac{1}{k-v}}} \sim N\,(0,1)$$

De la misma forma que el caso anterior, básicamente se trata del estadístico que resulta de la prueba de hipótesis para verificar si dos medias con distribuciones normales independientes con σ^2 concocida es la misma.

El estadístico de razón de verosimilitud para probar la hipótesis nula de que no hay ningún cambio contra la alternativa de que hay exactamente un cambio en el v-ésimo lugar desconocido, está dado por $Z_B = \underset{1\leq v<k}{\text{máx}} |Z_v|$. La hipótesis nula de ausencia de cambio se rechaza si el estadístico supera el α-ésimo cuantil superior de la distribución nula de Z_B y la ubicación del punto de cambio se estima sería v tal que $Z_B = |Z_v|$.

El procedimiento de Segmentación Binaria aplica el criterio de forma recursiva hasta que no se detectan más cambios en ninguno de los segmentos obtenidos a partir de los puntos de cambio ya encontrados. Debido a que el método de Segmentación Binaria fue hecho para encontrar un punto de cambio en cada paso, muchas veces se presenta un problema al no detectar puntos de cambio adicionales. Por esta razón y para hacer más rápida la búsqueda, se modificó el método y se creó la **Segmentación Binaria Circular**.

2.2. Segmentación Binaria Circular

A partir de la modificación de la técnica anterior se creó la Segmentación Binaria Circular que permite dividir el cromosoma en tres segmentos al ubicar dos puntos de cambios a la vez.[2]

Sean $X_1,\ldots,X_v,\ldots,X_p,\ldots,X_k$ variables aleatorias independientes tales que $X_i \sim N(\mu_i, \sigma^2)$ para $i = \overline{1,k}$, éstas son ubicadas en forma circular de tal manera que las variables X_1 y X_k estén juntas.

Ahora, se forman dos bloques característicos de la siguiente forma:

- BLOQUE 1: $X_1,\ldots,X_v \wedge X_{p+1},\ldots,X_k$
- BLOQUE 2: $X_{v+1},\ldots,X_p$

Como se hizo anteriormente, y bajo las hipótesis que se han descrito, se busca la existencia de puntos de cambio. Si no existe ninguno, las medias de todas las variables aleatorias serán iguales; si hay dos puntos de cambio, la media de las variables aleatorias del primer y segundo bloques serán distintos.

Estadísticamente esto es, $H_0 : \mu = \mu^*,\ \sigma^2 > 0\ v_s\ H_1 : \mu \neq \mu^*, \sigma^2 > 0$ con $X_i \sim N(\mu, \sigma^2)$ para $i = \overline{1,v}, \overline{p+1,k}$ y $X_i \sim N(\mu^*, \sigma^2)$ para $i = \overline{1+v,p}$.

Como en el caso de Segmentación Binaria, primero es necesario definir el espacio parametral en el que se trabajará: $\Theta = \{(\mu, \mu^*, \sigma^2) \,|\mu, \mu^* \in \Re, \sigma^2 > 0\}$

Entonces: $\Theta_0 = \{(\mu, \sigma^2) \,|\mu \in \Re, \sigma^2 > 0\}$

$\Theta_1 = \Theta \setminus \Theta_0$

Ahora, al utilizar la prueba de la razón de verosimilitudes generalizada se tiene que:

$$0 \leq \lambda = \frac{\max\limits_{\theta \in \Theta_0} L(\theta)}{L(\theta_{MV})} \leq 1$$

donde θ es el vector de los parámetros definidos anteriormente, $L(\theta) = \prod_{i=1}^{k} f(x_i, \theta)$ es la función de verosimilitud y θ_{MV} es el estimador máximo verosímil.

Entonces, se sigue que:

$$L(\theta) = \prod_{i=1}^{k} f(x_i, \theta) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{k}{2}} e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{v}(X_i-\mu)^2 + \sum_{i=v+1}^{p}(X_i-\mu^*)^2 + \sum_{i=p+1}^{k}(X_i-\mu)^2\right]}$$

Aplicando logaritmo natural:

$$l = lnL(\theta) = \frac{k}{2} ln\left(\frac{1}{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{v}(X_i-\mu)^2 + \sum_{i=v+1}^{k}(X_i-\mu^*)^2 + \sum_{i=p+1}^{k}(X_i-\mu)^2\right]$$

$$= -\frac{k}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{v}(X_i-\mu)^2 + \sum_{i=v+1}^{k}(X_i-\mu^*)^2 + \sum_{i=p+1}^{k}(X_i-\mu)^2\right]$$

Entonces:

$$\frac{\partial l}{\partial \sigma^2} = -\frac{k}{2} \cdot \frac{\cancel{2\pi}}{\cancel{2\pi}\sigma^2} + \frac{1}{2\sigma^4}\left[\sum_{i=1}^{v}(X_i-\mu)^2 + \sum_{i=v+1}^{k}(X_i-\mu^*)^2 + \sum_{i=p+1}^{k}(X_i-\mu)^2\right]$$

Ahora, se busca que la derivada parcial sea cero, entonces:

$$\frac{\partial l}{\partial \sigma^2} = 0 \Longleftrightarrow -\frac{k}{2\sigma^2} + \frac{1}{2\sigma^4}\left[\sum_{i=1}^{v}(X_i-\mu)^2 + \sum_{i=v+1}^{k}(X_i-\mu^*)^2 + \sum_{i=p+1}^{k}(X_i-\mu)^2\right] = 0$$

$$\Longleftrightarrow \tfrac{k}{2\sigma^2} = \tfrac{1}{2\sigma^4}\left[\sum_{i=1}^{v}(X_i-\mu)^2 + \sum_{i=v+1}^{k}(X_i-\mu^*)^2 + \sum_{i=p+1}^{k}(X_i-\mu)^2\right]$$

$$\Longleftrightarrow k\sigma^2 = \sum_{i=1}^{v}(X_i-\mu)^2 + \sum_{i=v+1}^{k}(X_i-\mu^*)^2 + \sum_{i=p+1}^{k}(X_i-\mu)^2$$

$$\Longleftrightarrow \widehat{\sigma}^2_{MV} = \tfrac{1}{k}\left[\sum_{i=1}^{v}(X_i-\widehat{\mu}_{MV})^2 + \sum_{i=v+1}^{k}(X_i-\widehat{\mu}^*_{MV})^2 + \sum_{i=p+1}^{k}(X_i-\widehat{\mu}_{MV})^2\right]$$

Luego:

$$\frac{\partial l}{\partial \mu} = -\frac{1}{2\sigma^2}\left[(2)(-1)\sum_{i=1}^{v}(X_i-\mu) + (2)(-1)\sum_{i=p+1}^{k}(X_i-\mu)\right]$$

$$= \frac{1}{\sigma^2}\left[\sum_{i=1}^{v} X_i - v\mu \sum_{i=p+1}^{k} X_i - (k-p)\,\mu\right]$$

$$\Rightarrow \frac{\partial l}{\partial \mu} = 0$$

$$\iff \frac{1}{\sigma^2}\left[\sum_{i=1}^{v} X_i - v\mu + \sum_{i=p+1}^{k} X_i - (k-p)\,\mu\right] = 0$$

$$\iff \sum_{i=1}^{v} X_i - v\mu + \sum_{i=p+1}^{k} X_i - (k-p)\,\mu = 0$$

$$\iff \sum_{i=1}^{v} X_i + \sum_{i=p+1}^{k} X_i = \mu\,(v+k-p)$$

$$\iff \widehat{\mu}_{MV} = \frac{1}{v+k-p}\left[\sum_{i=1}^{v} X_i + \sum_{i=p+1}^{k} X_i\right] = \overline{X}_v + \overline{X}_{k-p}$$

Y:

$$\frac{\partial l}{\partial \mu^*} = -\frac{1}{2\sigma^2}(2)(-1)\sum_{i=p+1}^{k}(X_i - \mu^*) = \frac{1}{\sigma^2}\left[\sum_{i=p+1}^{k} X_i - (k-i)\mu^*\right]$$

$$\Rightarrow \frac{\partial l}{\partial \mu^*} = 0$$

$$\iff \frac{1}{\sigma^2}\left[\sum_{i=p+1}^{k} X_i - (k-p)\mu^*\right] = 0$$

$$\iff \sum_{i=p+1}^{k} X_i - (k-p)\mu^* = 0$$

$$\iff \sum_{i=p+1}^{k} X_i = (k-p)\mu^*$$

$$\iff \widehat{\mu^*}_{MV} = \frac{1}{k-p}\sum_{i=p+1}^{k} X_i = \overline{X}_{k-p}$$

Al sustituir las expresiones anteriores en $L(\theta)$ se tiene:

$$L(\widehat{\theta}) = \left(\frac{1}{2\pi\widehat{\sigma^2}}\right)^{\frac{k}{2}} e^{-\frac{1}{2\widehat{\sigma^2}}\left[\sum_{i=1}^{v}(X_i-\widehat{\mu})^2+\sum_{i=v+1}^{p}(X_i-\widehat{\mu}^*)^2+\sum_{i=p+1}^{k}(X_i-\widehat{\mu})^2\right]}$$

Bajo H_0:

$$L(\theta_0) = \left(\frac{1}{2\pi\sigma_0^2}\right)^{\frac{k}{2}} e^{-\frac{1}{2\sigma_0^2}\left[\sum_{i=1}^{v}(X_i-\mu_0)^2+\sum_{i=v+1}^{p}(X_i-\mu_0)^2+\sum_{i=p+1}^{k}(X_i-\mu_0)^2\right]}$$

Aplicando la función logarítmo natural:

$$l = lnL(\theta) = \frac{k}{2}ln\left(\frac{1}{2\pi\sigma_0^2}\right) - \frac{1}{2\sigma_0^2}\left[\sum_{i=1}^{v}(X_i-\mu_0)^2+\sum_{i=v+1}^{p}(X_i-\mu_0)^2+\sum_{i=p+1}^{k}(X_i-\mu_0)^2\right]$$

$$= -\frac{k}{2}ln(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2}\left[\sum_{i=1}^{v}(X_i-\mu_0)^2+\sum_{i=v+1}^{p}(X_i-\mu_0)^2+\sum_{i=p+1}^{k}(X_i-\mu_0)^2\right]$$

Así:

$$\frac{\partial l}{\partial \sigma_{0^2}} = -\frac{k}{2}\cdot\frac{\cancel{2\pi}}{\cancel{2\pi}\sigma_0^2} + \frac{1}{2\sigma^4}\left[\sum_{i=1}^{v}(X_i-\mu_0)^2+\sum_{i=v+1}^{p}(X_i-\mu_0)^2+\sum_{i=p+1}^{k}(X_i-\mu_0)^2\right]$$

Ahora, se busca que la parcial sea cero, entonces:

$$\frac{\partial l}{\partial \sigma_0^2} = 0$$

$$\iff -\frac{k}{2\sigma_0^2} + \frac{1}{2\sigma_0^4}\left[\sum_{i=1}^{v}(X_i-\mu_0)^2+\sum_{i=v+1}^{p}(X_i-\mu_0)^2+\sum_{i=p+1}^{k}(X_i-\mu_0)^2\right] = 0$$

$$\iff \frac{k}{2\sigma_0^2} = \frac{1}{2\sigma_0^4}\left[\sum_{i=1}^{v}(X_i-\mu_0)^2+\sum_{i=v+1}^{p}(X_i-\mu_0)^2+\sum_{i=p+1}^{k}(X_i-\mu_0)^2\right]$$

$$\iff \widehat{\sigma_0}^2_{MV} = \frac{1}{k}\left[\sum_{i=1}^{v}(X_i-\widehat{\mu}_{0MV})^2+\sum_{i=v+1}^{p}(X_i-\widehat{\mu}_{0MV})^2+\sum_{i=p+1}^{k}(X_i-\widehat{\mu}_{0MV})^2\right]$$

Y:

$$\frac{\partial l}{\partial \mu_0} = -\frac{1}{2\sigma_0^2}(2)(-1)\sum_{i=1}^{v}(X_i-\mu_0) - \frac{1}{2\sigma_0^2}(2)(-1)\sum_{i=v+1}^{p}(X_i-\mu_0) - \frac{1}{2\sigma_0^2}(2)(-1)\sum_{i=p+1}^{k}(X_i-\mu_0)$$

$$= \frac{1}{\sigma_0^2}\left[\sum_{i=1}^{v}(X_i-\mu_0) + \sum_{i=v+1}^{p}(X_i-\mu_0) + \sum_{i=p+1}^{k}(X_i-\mu_0)\right]$$

De donde:

$$\frac{\partial l}{\partial \mu_0} = 0$$

$$\iff \frac{1}{\sigma_0^2}\left[\sum_{i=1}^{v}(X_i-\mu_0) + \sum_{i=v+1}^{p}(X_i-\mu_0) + \sum_{i=p+1}^{k}(X_i-\mu_0)\right] = 0$$

$$\iff \sum_{i=1}^{v}X_i - v\mu_0 + \sum_{i=v+1}^{p}X_i - (p-v)\mu_0 + \sum_{i=p+1}^{k}X_i - (k-p)\mu_0 = 0$$

$$\iff \sum_{i=1}^{v}X_i + \sum_{i=v+1}^{p}X_i + \sum_{i=p+1}^{k}X_i = v\mu_0 + (p-v)\mu_0 + (k-p)\mu_0$$

$$\iff \widehat{\mu_0}_{MV} = \frac{1}{k}\left[\sum_{i=1}^{v}X_i + \sum_{i=v+1}^{p}X_i + \sum_{i=p+1}^{k}X_i\right]$$

Sustituyendo adecuadamente, se obtiene:

$$L(\hat{\theta}_0) = \left(\frac{1}{2\pi\hat{\sigma}_0^2}\right)^{\frac{k}{2}} e^{-\frac{1}{2\hat{\sigma}_0^2}\left[\sum_{i=1}^{v}(X_i-\hat{\mu}_0)^2 + \sum_{i=v+1}^{p}(X_i-\hat{\mu}_0)^2 + \sum_{i=p+1}^{k}(X_i-\hat{\mu}_0)^2\right]}$$

Ahora,

$$\lambda = \frac{\left(\frac{1}{2\pi\hat{\sigma}_0^2}\right)^{\frac{k}{2}} e^{-\frac{1}{2\hat{\sigma}_0^2}\left[\sum_{i=1}^{v}(X_i-\hat{\mu}_0)^2 + \sum_{i=v+1}^{p}(X_i-\hat{\mu}_0)^2 + \sum_{i=p+1}^{k}(X_i-\hat{\mu}_0)^2\right]}}{\left(\frac{1}{2\pi\hat{\sigma}_{MV}^2}\right)^{\frac{k}{2}} e^{-\frac{1}{2\hat{\sigma}_{MV}^2}\left[\sum_{i=1}^{v}(X_i-\hat{\mu}_{MV})^2 + \sum_{i=v+1}^{p}(X_i-\hat{\mu}_{MV}^*)^2 + \sum_{i=p+1}^{k}(X_i-\hat{\mu}_{MV})^2\right]}}$$

Sean α el numerador y β el denominador de la razón anterior.

Entonces, para α, con $\eta = \sum_{i=1}^{v} (X_i - \hat{\mu}_0)^2 + \sum_{i=v+1}^{p} (X_i - \hat{\mu}_0)^2 + \sum_{i=p+1}^{k} (X_i - \hat{\mu}_0)^2$ se tiene:

$$\left[\frac{k}{(2\pi)(\eta)}\right]^{\frac{k}{2}} e^{\left(-\frac{1}{2}\right)\left(\frac{(k)(\eta)}{\eta}\right)} = \left[\frac{k}{2\pi\left(\sum_{i=1}^{v} (X_i - \hat{\mu}_0)^2 + \sum_{i=v+1}^{p} (X_i - \hat{\mu}_0)^2 + \sum_{i=p+1}^{k} (X_i - \hat{\mu}_0)^2\right)}\right]^{\frac{k}{2}} e^{-\frac{k}{2}}$$

es decir,

$$\left[\frac{ke^{-1}}{2\pi\left(\sum_{i=1}^{v} (X_i - \hat{\mu}_0)^2 + \sum_{i=v+1}^{p} (X_i - \hat{\mu}_0)^2 + \sum_{i=p+1}^{k} (X_i - \hat{\mu}_0)^2\right)}\right]^{\frac{k}{2}}$$

Y para β, con $\xi = \sum_{i=1}^{v} (X_i - \hat{\mu}_{MV})^2 + \sum_{i=v+1}^{p} (X_i - \hat{\mu}^*_{MV})^2 + \sum_{i=p+1}^{k} (X_i - \hat{\mu}_{MV})^2$, se tiene:

$$\left[\frac{k}{(2\pi)(\xi)}\right]^{\frac{k}{2}} e^{\left(-\frac{1}{2}\right)\left(\frac{(k)(\xi)}{\xi}\right)}$$

$$= \left[\frac{k}{2\pi\left(\sum_{i=1}^{v} (X_i - \hat{\mu}_{MV})^2 + \sum_{i=v+1}^{p} (X_i - \hat{\mu}^*_{MV})^2 + \sum_{i=p+1}^{k} (X_i - \hat{\mu}_{MV})^2\right)}\right]^{\frac{k}{2}} e^{-\frac{k}{2}}$$

$$= \left[\frac{ke^{-1}}{2\pi\left(\sum_{i=1}^{v} (X_i - \hat{\mu}_{MV})^2 + \sum_{i=v+1}^{p} (X_i - \hat{\mu}^*_{MV})^2 + \sum_{i=p+1}^{k} (X_i - \hat{\mu}_{MV})^2\right)}\right]^{\frac{k}{2}}$$

Entonces, para λ:

$$\lambda = \frac{\left[\frac{ke^{-1}}{2\pi\left(\sum_{i=1}^{v}(X_i-\hat{\mu}_0)^2+\sum_{i=v+1}^{p}(X_i-\hat{\mu}_0)^2+\sum_{i=p+1}^{k}(X_i-\hat{\mu}_0)^2\right)}\right]^{\frac{k}{2}}}{\left[\frac{ke^{-1}}{2\pi\left(\sum_{i=1}^{v}(X_i-\hat{\mu}_{MV})^2+\sum_{i=v+1}^{p}(X_i-\hat{\mu}^*_{MV})^2+\sum_{i=p+1}^{k}(X_i-\hat{\mu}_{MV})^2\right)}\right]^{\frac{k}{2}}}$$

$$= \left[\frac{\frac{\cancel{ke^{-1}}}{\cancel{2\pi}\left(\sum_{i=1}^{v}(X_i-\hat{\mu}_0)^2+\sum_{i=v+1}^{p}(X_i-\hat{\mu}_0)^2+\sum_{i=p+1}^{k}(X_i-\hat{\mu}_0)^2\right)}}{\frac{\cancel{ke^{-1}}}{\cancel{2\pi}\left(\sum_{i=1}^{v}(X_i-\hat{\mu}_{MV})^2+\sum_{i=v+1}^{p}(X_i-\hat{\mu}^*_{MV})^2+\sum_{i=p+1}^{k}(X_i-\hat{\mu}_{MV})^2\right)}}\right]^{\frac{k}{2}}$$

$$= \left[\frac{\sum_{i=1}^{v}(X_i-\hat{\mu}_{MV})^2+\sum_{i=v+1}^{p}(X_i-\hat{\mu}^*_{MV})^2+\sum_{i=p+1}^{k}(X_i-\hat{\mu}_{MV})^2}{\sum_{i=1}^{v}(X_i-\hat{\mu}_0)^2+\sum_{i=v+1}^{p}(X_i-\hat{\mu}_0)^2+\sum_{i=p+1}^{k}(X_i-\hat{\mu}_0)^2}\right]^{\frac{k}{2}}$$

$$= \left[\frac{\sum_{i=1}^{v}\left(X_i-\bar{X}_{v+k-p}\right)^2+\sum_{i=v+1}^{p}\left(X_i-\bar{X}_{p-v}\right)^2+\sum_{i=p+1}^{k}\left(X_i-\bar{X}_{v+k-p}\right)^2}{\sum_{i=1}^{v}(X_i-\hat{\mu}_0)^2+\sum_{i=v+1}^{p}(X_i-\hat{\mu}_0)^2+\sum_{i=p+1}^{k}(X_i-\hat{\mu}_0)^2}\right]^{\frac{k}{2}}$$

Al igual que en el capítulo anterior, se necesita formar una región de rechazo apropiada para obtener las características deseadas de la función de distribución, entonces, a través de relaciones en las siguientes expresiones ya revisadas a detalle con anterioridad se tiene que:

$$\sum_{i=1}^{v}(X_i-\hat{\mu}_0)^2=\sum_{i=1}^{v}\left(X_i-\bar{X}_{v+k-p}\right)^2+v\left(\bar{X}_{v+k-p}-\hat{\mu}_0\right)^2 \tag{1}$$

$$\sum_{i=v+1}^{p}(X_i-\hat{\mu}_0)^2=\sum_{i=v+1}^{p}\left(X_i-\bar{X}_{p-v}\right)^2+(p-v)\left(\bar{X}_{p-v}-\hat{\mu}_0\right)^2 \tag{2}$$

$$\sum_{i=p+1}^{k}(X_i-\hat{\mu}_0)^2=\sum_{i=p+1}^{k}\left(X_i-\bar{X}_{v+k-p}\right)^2+(k-p)\left(\bar{X}_{v+k-p}-\hat{\mu}_0\right)^2 \tag{3}$$

Además,

$$v\left(\bar{X}_{v+k-p}-\hat{\mu}_0\right)^2+(k-p)\left(\bar{X}_{v+k-p}-\hat{\mu}_0\right)^2=(v+k-p)\left(\bar{X}_{v+k-p}-\hat{\mu}_0\right)^2$$

Y como,

$$\hat{\mu}_0=\frac{(v+k-p)\,\bar{X}_{v+k-p}+(v-p)\,\bar{X}_{p-v}}{v+(k-p)+(p-v)}$$

$$\Rightarrow(v+k-p)\left(\bar{X}_{v+k-p}-\hat{\mu}_0\right)^2=(v+k-p)\left[\bar{X}_{v+k-p}-\frac{\left((v+k-p)\bar{X}_{v+k-p}+(p-v)\bar{X}_{p-v}\right)}{v+(k-p)+(p-v)}\right]^2$$

$$=(v+k-p)\left[\frac{v\bar{X}_{v+k-p}+(k-p)\bar{X}_{v+k-p}+(p-v)\bar{X}_{v+k-p}-(v+k-p)\bar{X}_{v+k-p}-(p-v)\bar{X}_{p-v}}{v+(k-p)+(p-v)}\right]^2$$

$$=(v+k-p)\left[\frac{(p-v)\left(\bar{X}_{v+k-p}-\bar{X}_{p-v}\right)}{v+(k-p)+(p-v)}\right]^2$$

$$=(v+k-p)\,\frac{(p-v)^2\left(\bar{X}_{v+k-p}-\bar{X}_{p-v}\right)^2}{(v+(k-p)+(p-v))^2}$$

De la misma forma,

$$(p-v)\left(\bar{X}_{p-v}-\hat{\mu}_0\right)^2=(p-v)\,\frac{(v+k-p)^2\left(\bar{X}_{p-v}-\bar{X}_{v+k-p}\right)^2}{(v+(k-p)+(p-v))^2}$$

Entonces:

$$(v+k-p)\left(\bar{X}_{v+k-p}-\hat{\mu}_0\right)^2+(p-v)\left(\bar{X}_{p-v}-\hat{\mu}_0\right)^2$$

$$=(v+k-p)\frac{(p-v)^2\left(\bar{X}_{v+k-p}-\bar{X}_{p-v}\right)^2}{(v+(k-p)+(p-v))^2}+(p-v)\frac{(v+k-p)^2\left(\bar{X}_{p-v}-\bar{X}_{v+k-p}\right)^2}{(v+(k-p)+(p-v))^2}$$

$$=\left(\bar{X}_{v+k-p}-\bar{X}_{p-v}\right)^2(v+k-p)(p-v)\left(\frac{\cancel{(p-v)+(v+k-p)}}{(v+(k-p)+(p-v))^{\cancel{2}}}\right)$$

$$=\left(\bar{X}_{v+k-p}-\bar{X}_{p-v}\right)^2\frac{(v+k-p)(p-v)}{v+(k-p)+(p-v)} \tag{4}$$

De esta manera se tiene que, usando en (4) las ecuaciones (1), (2), (3):

$$\lambda=\left[\frac{\varpi}{\varpi+\left(\bar{X}_{v+k-p}-\bar{X}_{p-v}\right)^2\frac{(v+k-p)(p-v)}{v+(k-p)+(p-v)}}\right]^{\frac{k}{2}}\leq C_\alpha$$

Con

$$\varpi=\sum_{i=1}^{v}\left(X_i-\bar{X}_{v+k-p}\right)^2+\sum_{i=v+1}^{p}\left(X_i-\bar{X}_{p-v}\right)^2+\sum_{i=p+1}^{k}\left(X_i-\bar{X}_{v+k-p}\right)^2$$

Y donde C_α es una constante que depende del nivel de significancia. Entonces,

$$\left[\frac{\varpi+\left(\bar{X}_{v+k-p}-\bar{X}_{p-v}\right)^2\frac{(v+k-p)(p-v)}{v+(k-p)+(p-v)}}{\varpi}\right]\geq C'_\alpha$$

$$\Leftrightarrow 1+\frac{\left(\bar{X}_{v+k-p}-\bar{X}_{p-v}\right)^2\frac{(v+k-p)(p-v)}{v+(k-p)+(p-v)}}{\sum_{i=1}^{v}\left(X_i-\bar{X}_{v+k-p}\right)^2+\sum_{i=v+1}^{p}\left(X_i-\bar{X}_{p-v}\right)^2+\sum_{i=p+1}^{k}\left(X_i-\bar{X}_{v+k-p}\right)^2}\geq C'_\alpha$$

Recordando que:

$$\frac{(v+k-p)(p-v)}{v+(k-p)+(p-v)}=\frac{1}{\frac{1}{p-v}+\frac{1}{v+k-p}}$$

Se tiene que:

$$1+\frac{\frac{\left(\bar{X}_{v+k-p}-\bar{X}_{p-v}\right)^2}{\frac{1}{p-v}+\frac{1}{v+k-p}}}{\sum_{i=1}^{v}\left(X_i-\bar{X}_{v+k-p}\right)^2+\sum_{i=v+1}^{p}\left(X_i-\bar{X}_{p-v}\right)^2+\sum_{i=p+1}^{k}\left(X_i-\bar{X}_{v+k-p}\right)^2}\geq C'_\alpha$$

$$\Leftrightarrow \frac{\frac{\left(\bar{X}_{v+k-p}-\bar{X}_{p-v}\right)}{\sqrt{\frac{1}{p-v}+\frac{1}{v+k-p}}}}{\sqrt{\frac{\sum_{i=1}^{v}\left(X_i-\bar{X}_{v+k-p}\right)^2+\sum_{i=v+1}^{p}\left(X_i-\bar{X}_{p-v}\right)^2+\sum_{i=p+1}^{k}\left(X_i-\bar{X}_{v+k-p}\right)^2}{k-2}}}\geq C''_\alpha$$

Para poder encontrar el valor de la constante, es necesario verificar la distribución de λ, para esto (bajo H_0):

Como se vio en los supuestos,

$$\bar{X}_{v+k-p}\sim N\left(\mu,\frac{\sigma^2}{v+k-p}\right)$$

$$\bar{X}_{p-v}\sim N\left(\mu,\frac{\sigma^2}{p-v}\right)$$

$$\Rightarrow \bar{X}_{p-v}-\bar{X}_{v+k-p}\sim N\left(0,\sigma^2\left(\frac{(v+k-p)+(p-v)}{(v+k-p)(p-v)}\right)\right)$$

Entonces:

$$\frac{\bar{X}_{p-v}-\bar{X}_{v+k-p}}{\sigma\sqrt{\frac{1}{p-v}+\frac{1}{v+k-p}}}\sim N(0,1)$$

Además:

$$\frac{\sum_{i=v+1}^{p}\left(X_i-\bar{X}_{p-v}\right)^2}{\sigma^2}\sim\chi^2_{(p-v-1)}$$

Y,

$$\frac{\sum_{i=1}^{v}\left(X_i-\bar{X}_{v+k-p}\right)^2}{\sigma^2}+\frac{\sum_{i=p+1}^{k}\left(X_i-\bar{X}_{v+k-p}\right)^2}{\sigma^2}\sim\chi^2_{(v+k-p-1)}$$

Entonces, por independencia:

$$\frac{\sum_{i=v+1}^{p}\left(X_i-\bar{X}_{p-v}\right)^2}{\sigma^2}+\frac{\sum_{i=1}^{v}\left(X_i-\bar{X}_{v+k-p}\right)^2}{\sigma^2}+\frac{\sum_{i=p+1}^{k}\left(X_i-\bar{X}_{v+k-p}\right)^2}{\sigma^2}\sim\chi^2_{(k-2)}$$

Por lo tanto,

$$Z_{pv}=\frac{\frac{\left(\bar{X}_{v+k-p}-\bar{X}_{p-v}\right)}{\sqrt{\frac{1}{p-v}+\frac{1}{v+k-p}}}}{\sqrt{\frac{\sum_{i=1}^{v}\left(X_i-\bar{X}_{v+k-p}\right)^2+\sum_{i=v+1}^{p}\left(X_i-\bar{X}_{p-v}\right)^2+\sum_{i=p+1}^{k}\left(X_i-\bar{X}_{v+k-p}\right)^2}{k-2}}}\sim t_{(k-2)}$$

Si se deseara probar la hipótesis nula de que no hay puntos de cambio contra la hipótesis alternativa de que hay dos puntos de cambio, independientemente de la posición en la que ocurra, entonces la estadística de prueba está dada por $Z_B=\max_{1\leq v<p\leq k}|Z_{vp}|$. La hipótesis nula de que no hay puntos de cambio es rechazada si la estadística Z_B excede al α-ésimo cuantil de la distribución nula de Z_B y los puntos de cambio detectados serian p y v tales que $Z_B=|Z_{pv}|$.

En el mejor de los casos, se habrán localizado dos puntos de cambio, p y v, y así, este procedimiento se repetirá recursivamente para cada uno de los segmentos resultantes.

2.2.1. Eliminación de Tendencias

Algunas veces se presentan tendencias locales en los datos que llevan a considerar puntos de cambio biológicamente no significativos.

Una forma de solucionar este problema es por medio de la técnica *pruning* en la teoría de Árboles de Clasificación[13].

[13]Específicamente en lo referido a los árboles jerárquicos CART.

Para mayor información, véase el Apéndice E.

Un árbol de clasificación es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde el nodo raíz hasta alguna de sus hojas.

Pruning es el proceso en el cual se eliminan hojas y ramas del árbol para mejorar la clasificación en lo posible. Entonces, esta técnica consiste en eliminar algunos puntos de cambio que se están detectando y que deberían de ser descartados.

Supóngase que se tienen C puntos de cambio. La suma del cuadrado de las desviaciones de los puntos en segmentos alrededor de su segmento promedio puede representarse como *SS(C)*[14].

Entonces, se calcula la suma de los cuadrados correspondientes al mejor conjunto de puntos de cambio de tamaños *1* a *C-1*, i.e., *SS(1)*,. . . ,*SS(C-1)*.

Así $c^* = \min_c \left[\frac{SS(c)}{SS(C)} - 1 < \gamma \right]$, donde γ es una constante predeterminada (0.05 ó 0.10).

Los puntos de cambio son los que llevaron a $SS(^*)$.

2.2.2. Suavización de puntos atípicos

Los puntos atípicos pueden ser causados por errores técnicos en un experimento o por el número de copias mal logradas en una región que abarca únicamente una sonda.

La región de suavizado para cada posición i está dada por $i-R,\ldots,i,\ldots,i+R$, donde $2 \leq R \leq 5$, con $R \in Z$.

Sea m_i la mediana de los datos en la región de suavización y sea $\hat{\sigma}$ la desviación estándar de todos los datos. Si la observación X_i es la máxima o mínima de todas

[14]Esto es equivalente a la suma de cuadrados del error en un ANOVA.

las observaciones en la región de suavización, entonces, se encuentra la observación X_j más cercana a X_i. Si la distancia de X_i a X_j excede $4\hat{\sigma}$ se reemplaza X_i con la expresión $m_i + 2\hat{\sigma} sign(X_i - X_j)$.[15]

2.3. Aproximación de la distribución nula de Z_B

Aún cuando la distribución nula de Z_{vp} sea conocida analíticamente, la distribución de Z_B no es evidente para todo valor de K.

A partir de este problema, se realizaron varias pruebas de una aproximación a la distribución nula de Z_B mediante métodos de simulación.

Para poder estimar la distribución de la estadística de prueba Z_B, se obtiene un número grande de simulaciones de ella.

Se generan N simulaciones de muestras de tamaño k, $X_1, \ldots, X_k$, donde se supone $X_i \sim N(0,1)$, para i=1,...,k.

Para cada una de las N muestras se calcula $Z_B = \max_{1 \leq v < p \leq k} |Z_{vp}|$. Esto es, se obtiene el valor absoluto de las Z_{vp} encontradas y el máximo de esas estadísticas será un valor simulado de la estadística de prueba.

De esta manera se obtienen N valores simulados de la estadística de prueba Z_B, bajo la hipótesis nula.

3. Aplicación de la Técnica

3.1. Introducción

La base de datos contiene una muestra de pacientes que presentaron síntomas que revelaban la presencia de cáncer de pulmón.

[15] Para mayor información sobre la función sign(x), véase el Apéndice D.

Esta base consta de 23 cromosomas de distintos tamaños y formó parte de un programa de lectura de intensidades. La obtención de estos datos se mantendrá en sentido anónimo por petición del facilitador de la muestra.

A la muestra descrita se le aplicará la técnica de Segmentación con toda la preparación a los datos.

3.1.1. Generalidades

Los pulmones son los órganos encargados de oxigenar la sangre y expulsar el dióxido de carbono, un producto de desecho producido por las células del cuerpo. Participando además en otras importantes funciones metabólicas y cardiovasculares. *Véase la Figura 8.* Los bronquios, en tanto transportan el aire inspirado hacia los pulmones. Desde la tráquea se van dividiendo sucesivamente, dando origen a bronquios cada vez más pequeños hasta llegar a los alvéolos. Los alvéolos son pequeños saquitos rodeados de vasos sanguíneos de pequeño calibre (capilares). Entre el aire contenido en los alvéolos y los capilares se produce el intercambio gaseoso.

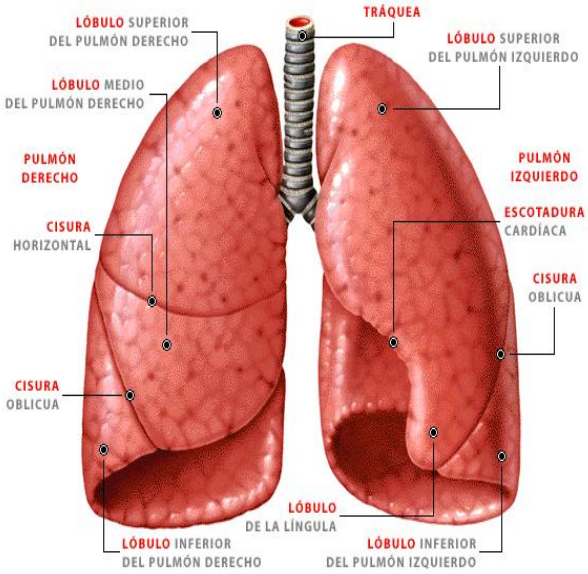


Figura 8: Pulmones ilustrados.

3.1.2. Cáncer de pulmón

El cáncer de pulmón es un tumor maligno que generalmente se origina en las células que recubren los bronquios (epitelio bronquial). Se produce principalmente por la irritación e inflamación crónica del epitelio bronquial por agentes externos (carcinógenos) donde destaca el humo del cigarrillo. Producto de esta irritación crónica y de factores genéticos se producen mutaciones que llevan al crecimiento rápido y descontrolado de algunas células, lo que se denomina transformación neoplásica, generándose así un cáncer.[9] Finalmente al continuar creciendo el tumor, algunas de las células pueden desplazarse hacia otros órganos del cuerpo dando origen a metástasis. *Véase la Figura 9.*

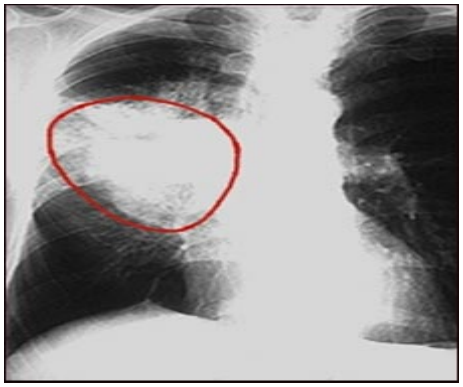


Figura 9: Pulmones con cáncer.

3.1.3. Algunos datos estadísticos

El cáncer pulmonar en la actualidad es el tumor maligno más frecuente en el mundo. En los Estados Unidos de América (EUA), en 1996, fue la principal causa de muerte en ambos sexos. Antes del siglo XX, el cáncer pulmonar era una entidad patológica muy rara. A partir de 1930, su frecuencia ha aumentado. Se estima que para el año 2025 se incrementará el número de muertes en más de 80 %, es decir, a tres y medio millones en países en desarrollo.[16]

La asociación entre tabaco y cáncer de pulmón ha sido bien establecida, el riesgo relativo en fumadores se ha duplicado en hombres y cuadruplicado en mujeres. Existe una predisposición genética, los fumadores con antecedentes familiares de cáncer pulmonar tienen un riesgo relativo de 2 a 2.5 veces mayor en relación con fumadores sin antecedentes familiares.

La dieta es otro factor, el riesgo se incrementa con una dieta alta en colesterol y en consumo de grasas; se ha mencionado un efecto protector de las vitaminas A y C y los betacarotenos.

[16]Area clínica, Instituto Nacional de Enfermedades Respiratorias (INER). Facultad de Medicina, Universidad Nacional Autónoma de México, México.

En México es difícil evaluar la frecuencia del padecimiento. Datos de la Secretaría de Salud (SSA) indican que la mortalidad por cáncer se incrementó de 1.78 %, en 1950, a 9.32 % (tasa bruta por 10,000 habitantes), en 1986. En 14,824 autopsias practicadas, entre 1953 y 1970, en los principales hospitales generales de la ciudad de México, de la entonces Secretaría de Salubridad y Asistencia (SSA), del Instituto Mexicano del Seguro Social (IMSS) y del Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado (ISSSTE), y en el Hospital General de México-SSA, las neoplasias malignas fueron la enfermedad principal en 28.5 % y, de éstas, el cáncer pulmonar ocupó el tercer lugar con 7.4 %. En el IMSS las enfermedades se dividieron por aparatos y sistemas, el aparato respiratorio ocupó el sexto sitio y las neoplasias malignas de éste representaron 9 %; en el ISSSTE, los tumores malignos ocuparon el segundo lugar como enfermedad principal con 17 % y, de éstos, el cáncer broncogénico representó 14 %.

De 1981 a 1985, el carcinoma pulmonar ingresó al grupo de las primeras 20 causas de muerte, y aunque la tasa de mortalidad general disminuyó la de cáncer pulmonar mostró una tendencia al aumento. En 1976, en el IMSS, entre las defunciones por tumores malignos, el carcinoma cervicouterino ocupaba el primer lugar; en 1982, el cáncer pulmonar lo desplazó y alcanzó el primer puesto. La tasa cruda de mortalidad por cáncer pulmonar (por 100,000 habitantes) se incrementó de 5.01, en 1979, a 7.25, en 1993, y esto fue más ostensible en los estados del norte de la República Mexicana.

Se analizaron 1,211 protocolos de autopsias del Hospital General de México, de sujetos mayores de 60 años, estudiados en dos periodos (1960-1965 y 1981-1985), y se encontró disminución de algunas enfermedades como la amibiasis. Las neoplasias malignas persistieron en ambos periodos, el cáncer pulmonar que ocupaba el segundo lugar en los años sesenta, pasó en los ochenta al primer sitio.

Otra investigación del Hospital General de México, informó de 923 casos de carcinoma broncogénico, estudiados entre 1971 y 1990; la relación hombre/mujer fue de 1.95:1. El tipo histológico fue epidermoide en 32 %; adenocarcinoma, en 28 %, y carcinoma de células pequeñas, 13 % en hombres; en mujeres, adenocarcinoma, en 39 % y epidermoide, en 29 %. Asimismo, en el Instituto Nacional de Enfermedades Respiratorias (INER), se encontró una alta frecuencia en pacientes de entre 61 y 70 años, el tipo histológico más frecuente fue el adenocarcinoma.[10]

3.2. Presentación de Datos

Como se había mencionado anteriormente, la base a analizar consta de 23 cromosomas de distintos tamaños; una idea de la forma en que se encuentran estos datos es justamente visualizarlos.

En las gráficas se muestra el número de sondas que poseé cada uno de los 23 cromosomas y su respectiva medida de intensidad.

En la *Cuadro 1* se detalla el número de sondas por cada cromosoma, el mayor número de sondas lo registra el cromosoma 5 con 189 sondas mientras que el menor número lo registra los cromosomas 20 y 21 con 18 sondas.

Cromosoma	**No. de Sondas**
1	163
2	62
3	62
4	137
5	189
6	98
7	137
8	142
9	167
10	180
11	116
12	98
13	61
14	78
15	72
16	69
17	96
18	56
19	35
20	18
21	18
22	62
23	71

Cuadro 1: Relación de Sondas por Cromosoma.

Como se había explicado anteriormente, es necesario hacer un suavizamiento en los datos para evitar datos atípicos. Utilizando del método descrito en la sección 2.2.2 se encontraron dos datos atípicos:

- Cromosoma 21, Sonda 14.
- Cromosoma 23, Sonda 6.

3.3. Empleo de la Técnica

Dado que no se conoce la distribución del máximo del valor absoluto de Z_B, es necesario hacer simulaciones para lograr estimar la distribución que necesitamos.

Para la simulación, el alumno creó una macro en excel que calcula la estadística de prueba; para cada cromosoma se simularon **n** variables aleatorias con distribución normal estándar donde **n** representa el número de sondas. Este proceso se realizó 3,000 veces.

Luego de las 3,000 simulaciones se podrá estimar la función de distribución a través de la utilización del método de Kernel que se explica en el *Apéndice G* con Kernel Gaussiano y segundo recurso.

Ahora es necesario determinar si existen puntos de cambio en los cromosomas y lograr ubicarlos, para esto se requiere estimar el p-valor[17] que se utilizará para la decisión si se acepta o se rechaza la hipótesis nula.

De acuerdo a la literatura de investigación [18] se aconseja utilizar pruebas no paramétricas que resultan muy parecidas a los métodos Bootstrap.

Estas pruebas permiten obtener diferentes muestras a partir de una muestra, en otras palabras, una muestra aleatoria se tomará como la población total, a

[17]Para mayor información, véase el Apéndice C.

[18]Bootstrap methods and their application by Davison-Hinkley (1997)

partir de esta muestra se obtienen varias muestras permutando con reemplazo la muestra.

La implementación de esta técnica ha mostrado gran efectividad en la comparación de dos muestras, es justamente por eso que en la literatura se recomienda esta alternativa, los pasos a seguir son:

1. Se permutan las variables aleatorias $X_1 \dots X_p$ obteniendo una muestra distinta a la original.

2. Se calcula Z_B para esta nueva muestra.

3. Se repiten los pasos anteriores P veces, donde P es mayor que p.

Una vez obtenidas las funciones de distribución empíricas, se repite el proceso para calcular el p-valor y la función de distribución de la estadística de prueba.

Entonces, se utiliza para cada $Z_B \dots$

$$\widehat{F}_p(x) = \frac{Z_B \leq x}{P}$$

para cada uno de los valores obtenidos.

3.4. Resultados

El siguiente cuadro muestra los resultados obtenidos.

Cromosoma	Z_B	**P-valor**
1	6.6639	0.0217
2	3.9163	0.5876
3	3.1562	0.5359
4	5.8908	0.1813
5	8.3391	0.0130
6	6.7638	0.2456
7	6.2137	0.2150
8	7.7926	0.6130
9	10.7387	0.3138
10	39.0143	0.0001
11	5.6203	0.1696
12	5.8462	0.5712
13	4.4733	0.5052
14	14.2230	0.0001
15	4.5290	0.6762
16	4.4845	0.6988
17	5.0045	0.8504
18	4.6609	0.2846
19	3.4341	0.2414
20	1.0432	0.0906
21	2.4719	0.0483
22	10.8131	0.2349
23	6.2292	0.0408

Cuadro 2: Estadísticas de prueba obtenidas con sus respectivos p-valor.

Ahora es necesario decidir si se acepta o se rechaza la hipótesis de que ni v ni p son puntos de cambio. Esta hipótesis se rechaza si Z_B excede el α-ésimo cuantil de la distribución nula o si el p-valor es menor a 0.01.

Dado que los cromosomas 10 y 14 son los que cumplen esta condición, entonces habrá que repetirse el procedimiento anterior segmentando estos cromosomas. Se encontró que para el cromosoma 10 los puntos de cambio están en la sonda 75 y 135 y, para el cromosoma 14 los puntos de cambio están en la sonda 33 y 47.

Replicando este procedimiento, se verifica la existencia de nuevos puntos de cambio. Se obtuvieron los siguientes resultados:

Cromosoma	Z_B	**P-valor**
10	6.7053	0.0978
10	7.7257	0.0107
10	3.8172	0.8811
14	2.5906	0.0188
14	1.8780	0.0547
14	4.2906	0.1850

Cuadro 3: Análisis de nuevas particiones para identificar otros puntos de cambio.

Como todas las particiones tanto del cromosoma 10 como del cromosoma 14 son mayores a 0.01, no se encontraron más puntos de cambio.

Conclusiones

A través del presente trabajo, se establece la importancia que la Estadística está tomando en la vida diaria y, en específico, en el estudio de enfermedades que aquejan a la humanidad desde hace ya mucho tiempo.

Debido a que la aplicación de técnicas estadísticas se usa en innumerables enfermedades, el modelo que se describe a lo largo del trabajo se enfoca en el cáncer.

Para lograr entender la forma en que esta enfermedad actúa sobre las células, se establece un entorno informativo sobre el DNA. Se introduce información sobre los genes y una explicación de su importancia en los seres vivos.

Las réplicas del DNA toman un papel importante para el trabajo y el desarrollo de la técnica que se describe se fundamenta en estas particiones. Luego de esta introducción biológica, se explica de manera detallada el principio y desarrollo del cáncer.

Debido a que el método de Segmentación Binaria fue hecho para encontrar un punto de cambio en cada paso, muchas veces se presenta un problema al no detectar puntos de cambio adicionales. Por esta razón y para hacer más rápida y eficiente la búsqueda, se modificó el método y se creó la Segmentación Binaria Circular.

A partir de este resultado, esta técnica se explica y se desarrolla matemáticamente a detalle, presentándose los resultados a los que se llega.

El trabajo va introduciendo la aplicación de esta técnica en datos reales por lo que, por sugerencia del autor, se necesitan tomar en cuenta arreglos y estudio detenido de los datos a trabajar por lo que se establecen temas importantes como eliminación de tendencias que llevan a la introducción de otros temas estadísticos como los árboles de clasificación; también la necesidad del tratamiento de puntos

atípicos en las muestras.

Con base a toda la teoría desarrollada a través del trabajo, se presenta una muestra de personas entre 25 y 50 años de edad que, por diversas causas, sufren cáncer en los pulmones. Esta base consta de 23 cromosomas de distintos tamaños y formó parte de un programa de lectura de intensidades.

Luego de dar un entorno sobre la afectación en los pulmones por el cáncer, se comienzan a trabajar los datos eliminando tendencias, suavizando datos atípicos, graficando los datos por cromosomas, etc.

El empleo de la técnica se lleva a cabo y se utilizan 3,000 simulaciones para llegar a la distribución de la estadística, se encuentra una primera selección de puntos de cambio y se analizan dichos puntos, luego se aplica la técnica nuevamente para determinar la existencia de nuevos puntos los cuales no fueron encontrados por lo que el análisis terminó.

Con los datos obtenidos se pueden determinar ciertos pros y contras de la técnica, como es de esperarse, la técnica se ha ido mejorando con estudios profundos sobre el análisis de los cromosomas.

Uno de los problemas principales es la interpretación médica significativa de los resultados, este tipo de análisis requiere de profesionistas con formación médica y estadística.

El análisis de datos y toma de decisiones para preparar la información a estudiar para el empleo de la técnica sigue siendo muy subjetivo y reside en experiencia o corazonadas basadas en técnicas no muy comprobadas.

Las células tienen distintos tipos de mutaciones; esta técnica sólo reconocerá pérdidas y ganancias en material genético, por lo que su uso será muy específico.

A pesar de esto, la técnica mostró gran eficiencia para encontrar zonas dañadas

en el DNA, no es la mejor técnica pues se han implementado modificaciones que hacen más eficiente esta búsqueda, pero logra abrir ventanas de acercamiento al estudio más especializado por lo que representa un gran avance en el estudio genético para apoyar a la medicina, y en específico, a la sociedad.

La tesis comprueba el empleo de esta técnica y muestra sus fortalezas y debilidades, definitivamente expone que la Estadística puede ayudar a resolver problemáticas que se presentan día a día.

Apéndices

Apéndice A

Experiencias de Meselson y Stahl

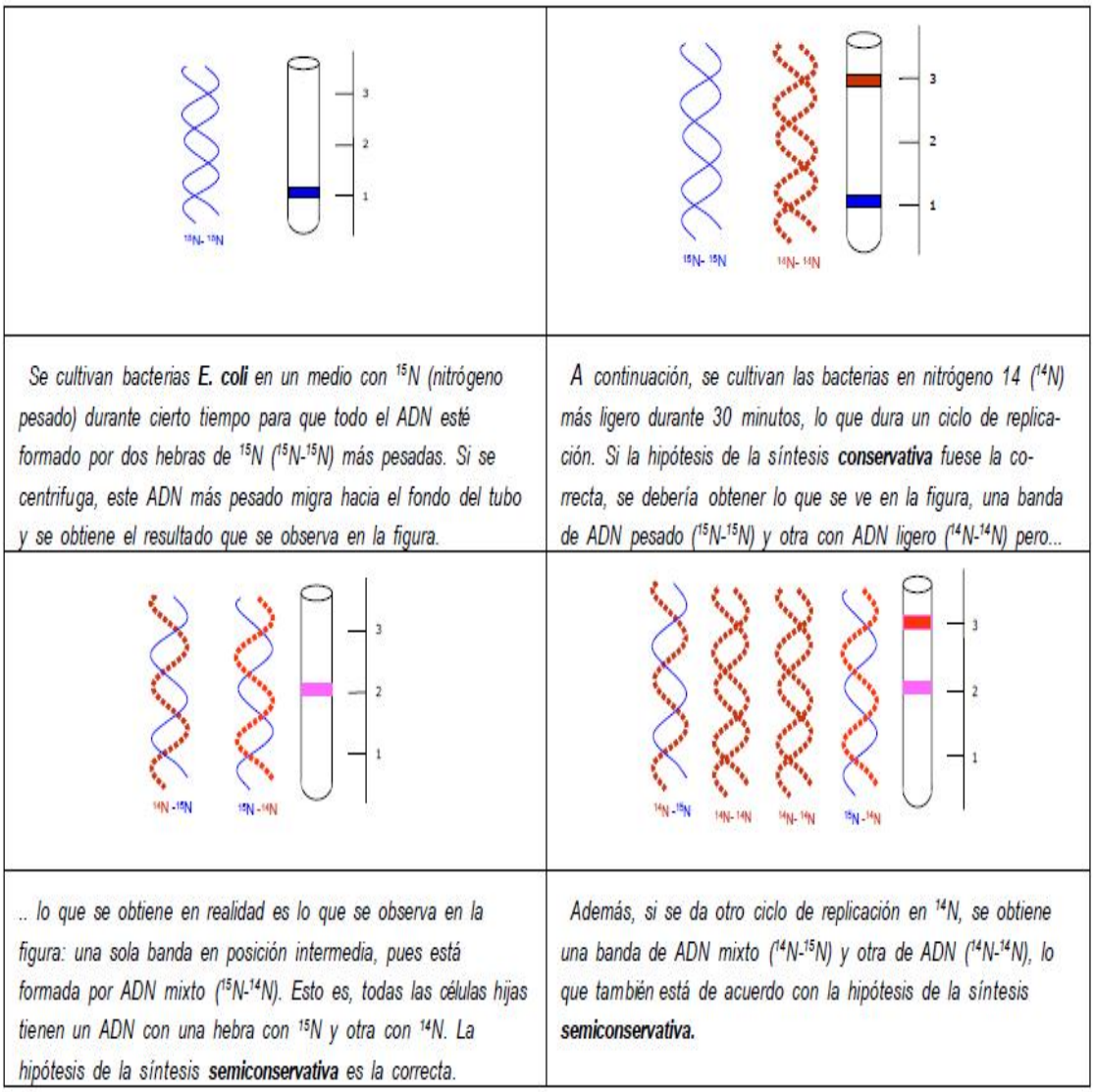


Figura 10: Hipótesis de Meselson y Stahl.

Apéndice B

Pruebas de Hipótesis

Una *prueba de hipótesis* estadística es una conjetura de una o más poblaciones. Nunca se sabe con absoluta certeza la veracidad o falsedad de una hipótesis estadística, a no ser que se examine la población entera. Esto por su puesto sería impráctico en la mayoría de las situaciones. En su lugar, se toma una muestra aleatoria de la población de interés y se utilizan los datos que contiene tal muestra para proporcionar evidencia que confirme o no la hipótesis.

Tipos de error y nivel de significancia

Si se rechaza una hipótesis cuando ésta debiera ser aceptada, se dirá que se ha cometido un error de tipo I.

Si se acepta una hipótesis que debiera ser rechazada, se dirá que se ha cometido un error de tipo II.

En ambos casos se ha producido un juicio erróneo.

	$H_0 escierta$	$H_1 escierta$
Se escogió H_0	No hay error	Error tipo II
Se escogió H_1	Error tipo I	No hay error

Cuadro 4: Tabla que muestra los tipos de errores en las pruebas de hipótesis.

Para que las reglas de decisión sean buenas, deben diseñarse de modo que minimicen los errores de decisión, y no es una cuestión sencilla, por que para cualquier tamaño de la muestra, un intento de disminuir un tipo de error suele ir acompañado de un crecimiento del otro tipo.

En la práctica un tipo de error puede ser más grave que el otro, y debe

alcanzarse un compromiso que disminuya el error más grave, la única forma de disminuir ambos a la vez es aumentar el tamaño de la muestra, que no siempre es posible.

Al contrastar una cierta hipótesis, la máxima probabilidad con la que se está dispuesto a correr el riesgo de cometer un error de tipo I se llama *nivel de significancia*. Esta probabilidad se denota por α, se suele especificar antes de la muestra, de manera que los resultados no influyan en nuestra elección.

En la práctica es frecuente un nivel de significancia de 0.05 ó 0.01, si bien se usan otros valores. Si, por ejemplo, se escoge un nivel de significancia del 5 % ó 0.05 al diseñar una regla de decisión, entonces hay unas cinco oportunidades entre cien de rechazar la hipótesis cuando debiera haberse aceptado; es decir, tenemos un 95 % de confianza de que hemos adoptado la decisión correcta. En tal caso se dice que la hipótesis ha sido rechazada al nivel de significancia 0.05 lo cual quiere decir que la hipótesis tiene una probabilidad del 5 % de ser falsa.

Apéndice C

P-valor

La elección del nivel de significación, tal y como se ha comentado anteriormente, es en cierta forma arbitraria.

Sin embargo, una vez obtenida la muestra, se puede calcular una cantidad que sí permite resumir el resultado del experimento de manera objetiva. Esta cantidad es el *p-valor*, que se denotará como P_v que corresponde al nivel de significación más pequeño posible que puede escogerse, para el cual todavía se aceptaría la hipótesis alternativa con las observaciones actuales. *Véase la Figura 11.*

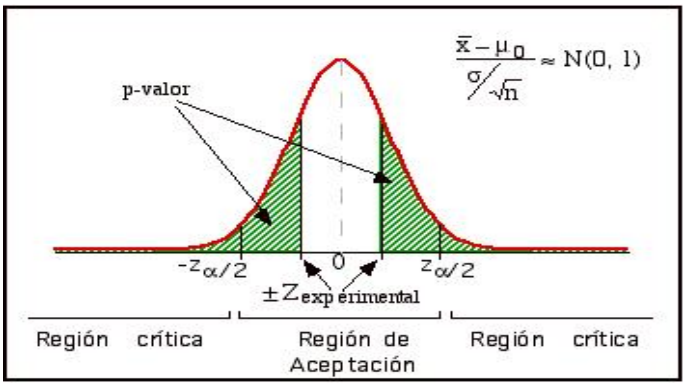


Figura 11: Zonas del p-valor.

El *p-valor* es una medida directa de lo verosímil que resulta obtener una muestra como la actual si es cierta H_0. Los valores pequeños indican que es muy infrecuente obtener una muestra como la actual, en cambio, los valores altos que es frecuente. El *p-valor* se emplea para indicar cuánto (o cuán poco) contradice la muestra actual la hipótesis alternativa.

Informar sobre cuál es el *p-valor* tiene la ventaja de permitir que cualquiera decida qué hipótesis acepta basándose en su propio nivel de riesgo α. Esto no es

posible cuando se informa, como ha sido tradicional, indicando sólo el resultado de la decisión, es decir, si se acepta o se rechaza H_0 con un α fijo.

Al proporcionar el *p-valor* obtenido con la muestra actual, la decisión se hará de acuerdo a la regla siguiente:

$$\text{Si} P_v \leq \alpha, \text{aceptar } H_1$$

$$\text{Si} P_v > \alpha, \text{aceptar } H_0$$

Apéndice D

Función signum

La función signo es una función matemática especial que obtiene el signo de cualquier número real que se tome por entrada. Se representa generalmente mediante sign(x).

La función signo se define de la siguiente manera:

$$sign(x) = \begin{cases} 1 & \text{si} \quad x > 0 \\ 0 & \text{si} \quad x = 0 \\ -1 & \text{si} \quad x < 0 \end{cases}$$

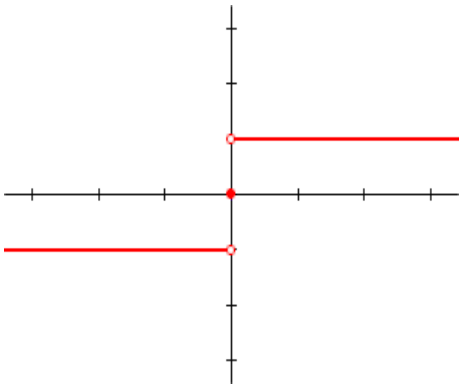


Figura 12: Gráfica de la función signum.

Algunas de sus propiedades se enlistan a continuación:

1. La función signo es una función impar, es decir, $sign(-x)=-sign(x)$.

2. Todo número real x puede expresarse como producto de su valor absoluto y la función signo evaluada en x, es decir, $x = sign(x) \cdot |x|$ con $x \in R$.

Apéndice E

Árboles de Clasificación

Los métodos basados en árboles particionan al espacio en un conjunto de rectángulos y luego les ajusta un modelo simple (como una constante) a cada uno.

Considérese un problema de regresión con variable respuesta continua Y, y entradas X_1, X_2 que toman valores en el intervalo unitario.

En la *Figura 13*, se muestra una partición del espacio por líneas paralelas a los ejes de un plano cartesiano. En cada elemento de la partición puede modelarse Y con una constante diferente pero existe un problema, algunas de las regiones resultantes son difíciles de describir. Para simplificar esto, se restringe el interés a las particiones binarias recursivas justo como en la parte superior derecha de la figura. Primero se dividió al espacio en dos regiones y se modeló la respuesta por la media de Y en cada región.

Se escogió la variable y el punto de división para lograr el mejor ajuste. Después, uno o ambas de estas regiones son divididas en dos regiones nuevas, y este proceso continúa hasta que alguna regla de paro sea aplicada. Por ejemplo, en la *Figura 13* primero se divide en X_1=t_1 luego, la región $X_1 \leq$t$_1$ es dividido en X_2=t_2 y la región X_1>t_1 es dividida en X_1=t_3. Finalmente, la región X_1>t_3 se divide en X_2=t_4; el resultado de este proceso, es una partición en cinco regiones R_i para i=$\{1,\bar{}5\}$ como lo muestra la figura.

El modelo de regresión correspondiente que predice a Y con una constante C_m en la región R_m está dado por:

$$\hat{f}(x) = \sum_{m=1}^{5} C_m I\left\{(X_1, X_2) \in R_m\right\}.$$

Este mismo modelo puede ser representado por el árbol binario en la parte

suroeste de la siguiente figura. Las observaciones que satisfacen la condición en cada ensamblador son asignados a la rama izquierda, y las otras a la rama derecha. Los nodos terminales u hojas del árbol corresponden a las regiones $R_1,\ldots,R_5$. El panel derecho inferior es un *plot* de la perspectiva de la superficie de una regresión de este modelo. (Para ilustrar ésto, se escogieron nodos c_1=-5, c_2=-7, c_3=0, c_4=2, c_5=4 para realizar el *plot*). Una gran ventaja del árbol binario recursivo es su interpretabilidad. La partición del espacio es enteramente descrito por un solo árbol.

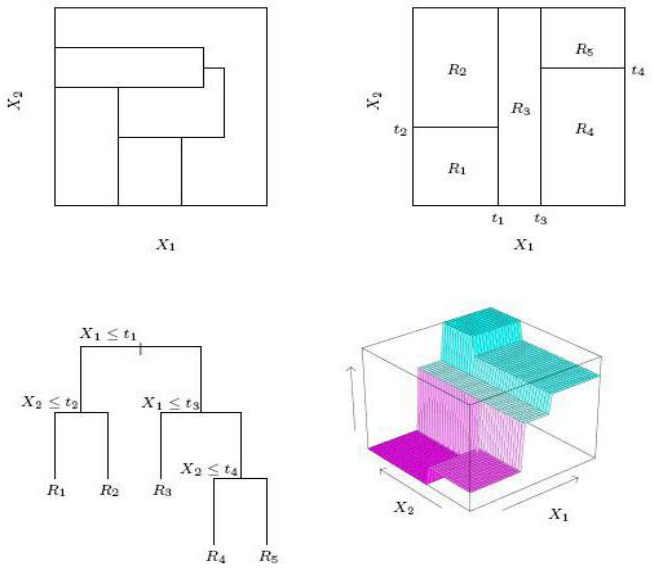


Figura 13: Particiones y CART. El panel superior derecho muestra una partición de un espacio bidimensional por una división recursiva binaria, como se usa en CART. El panel superior izquierdo muestra una partición general que no puede obtenerse de una división recursiva binaria. El panel izquierdo inferior muestra el árbol correspondiente a la partición del panel derecho superior, y el *plot* de una perspectiva de la superficie de una predicción aparece en el panel inferior derecho.

Árboles de Regresión

Los datos que se trabajan con árboles de regresión, consisten en p entradas y una respuesta, por cada una de las N observaciones, i.e. (x_i, y_i) for i = 1, 2, ..., N, con $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$. El algoritmo necesita decidir automaticamente las variables y puntos de división, y también la forma que tendrá el árbol. Supóngase primero que se tienen M regiones $R_1, R_2, \ldots, R_M$, y se modela la respuesta como una constante C_m en cada región:

$$f(x) = \sum_{m=1}^{M} C_m I\left\{x \in R_m\right\}$$

Si se adopta la suma de cuadrados como criterio de minimización es fácil ver que el mejor $\hat{C}_m$ es sólo el promedio de Y_i en la región R_m:

$$\hat{C}_m = promedio(y_i | x_i \in R_m)$$

Empezando con todos los datos, se considera una variable j de división y un punto s de división definiendo el par *half-planes*

$$R_1(j, s) = \{X | X_j \leq s\}$$

y

$$R_2(j, s) = \{X | X_j > s\}$$

Luego, se buscan la variable divisora j y el punto s de división que resuelvan:

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right]$$

Para cada j y s escogida, la minimización interna se resuelve por:

$$\hat{C}_1 = promedio(y_i | x_i \in R_1(j, s))$$

y

$$\hat{C}_2 = promedio(y_i | x_i \in R_2(j, s))$$

Para cada variable de división, la determinación del punto s puede hacerse rápido y mejorarse a través de todas las entradas.

Habiendo encontrado la mejor división, se particiona la información en las dos regiones resultantes y se sigue este proceso.

¿Qué tan grande debe crecer el árbol?

Claramente un árbol muy grande sobreajustará la información, mientras que un árbol pequeño podría no capturar la estructura importante.

El tamaño del árbol es un parámetro de adaptación que gobierna la complexidad del modelo, y el tamaño óptimo del árbol debe ser elegido siendo adaptado por los datos. Un acercamiento sería partir nodos del árbol solamente si la disminución de los suma de cuadrados debido a la división excede un cierto umbral.

La estrategia consiste en crecer un árbol grande T_0, deteniendo el proceso de división solamente cuando algún nodo de tamaño mínimo (digamos 5) se alcanza. Luego este árbol es *podado* usando **podado de costo de complejidad**.

Se define un subárbol $T \subset \mathrm{T}_0$ que puede ser cualquier árbol que pueda obtenerse podando T_0. Los nodos terminales son indexados por m, donde el nodo m representa la región R_m. Sea $|\mathrm{T}|$ el número de nodos terminales en T. Sean:

$$N_m = \#\{x_i \in R_m\},$$

$$\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i$$

$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2$$

Se define el criterio de costo de complejidad:

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|$$

La idea es encontrar, para cada α, el subárbol $T_\alpha \subseteq \mathrm{T}_0$ que minimice $C_\alpha(T)$. El parámetro de adaptación $0 \leq \alpha$ gobierna el intercambio entre el tamaño del árbol y la bondad de ajuste de los datos. Valores más grande de α resultan en árboles más pequeños T_α, e inversamente para valores más pequeños de α. Como sugiere la notación, con $\alpha = 0$ la solución es el árbol entero T_0.

Para cada α se puede mostrar que hay un único subárbol pequeño T_α que minimiza $C_\alpha(T)$. Para encontrar T_α se colapsa sucesivamente el nodo interno que produce el incremento por nodo más pequeño en $\sum_m N_m Q_m(T)$, y continúa hasta que se produzca un *single-node*.

Esto da una secuencia finita de subárboles, y se puede mostrar que esta secuencia debe contener a T_α.

Árboles de Clasificación

Si el objetivo es una salida de clasificación tomando valores 1, 2, ..., K, el único cambio necesitado en el algoritmo del árbol pertenece al criterio de dividir nodos y podar árboles.

Para regresión, se usa la medida de impureza del nodo por el *squared-error* $Q_m(T)$, pero esto no aplica para clasificación. En un nodo m, representando una región R_m con N_m observaciones, sea:

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

la proporción de la clase k-observaciones en el nodo m. Se clasifican las observaciones en el nodo m a la clase $k(m) = argmax_k \hat{p}_{mk}$, la clase mayoritaria en el nodo m. Diferentes medidas $Q_m(T)$ del nodo de impureza incluye lo siguiente:

- Clasificación Errónea

$$\frac{1}{N_m} \sum_{x_i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_m k_m$$

- Índice de Gini

$$\sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

- Devianza

$$-\sum_{k=1}^{K} \hat{p}_{mk} log \hat{p}_{mk}$$

Para dos clases, si p es la proporción en la segunda clase, esas tres medidas son 1-max(p, 1-p), 2p(1-p) y -plogp-(1-p)log(1-p), respectivamente.

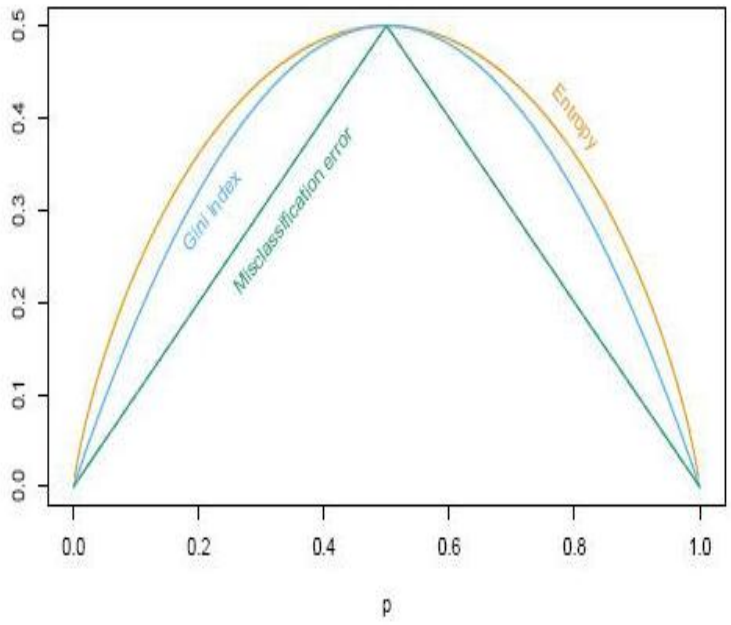


Figura 14: Medidas del nodo de impureza para una clasficación de dos clases, como función de la proporción p en la clase 2. La devianza ha sido escalada para pasar a través de (0.5, 0.5).

Los tres son similares, pero la desvianza y el índice de Gini son diferenciables, y por lo tanto más favorables a la optimización numérica. Además, la devianza y el índice de Gini son más sensibles a los cambios en el nodo de las probabilidades que en la tasa de clasificación errónea.

Apéndice F

Cuantiles

Los cuantiles son aquellos valores de la variable que, ordenados de menor a mayor, dividen a la distribución en partes, de tal manera que cada una de ellas contiene el mismo número de frecuencias.

Los cuantiles más conocidos son:

- Cuartiles (Q_i) Son valores de la variable que dividen a la distribución en cuatro partes, cada una de las cuales engloba el 25 % de las mismas. Se denotan de la siguiente forma: Q_1 es el primer cuartil que deja a su izquierda el 25 % de los datos; Q_2 es el segundo cuartil que deja a su izquierda el 50 % de los datos, y Q_3 es el tercer cuartil que deja a su izquierda el 75 % de los datos.
- Deciles (D_i) Son los valores de la variable que dividen a la distribución en diez partes iguales, cada una de las cuales engloba el 10 % de los datos. En total habrá nueve deciles.
- Centiles o Percentiles (P_i) Son los valores que dividen a la distribución en cien partes iguales, cada una de las cuales engloba el 1 % de las observaciones. En total habrá noventa y nueve percentiles.

Apéndice G

Estimación de Kernel

El Kernel es una función $K(x)$, a partir de la cual se puede establecer el siguiente estimador no paramétrico de cualquier función de densidad $f(x)$

$$\hat{f}(t) = \frac{1}{nh} \sum_{i=1}^{n} \delta \left[\frac{t - X_i}{h} \right]$$

Donde h es el parámetro de alisado y $X_1, \ldots, X_n$ los datos observados.

El parámetro de alisado, también llamado ancho de banda, es un número positivo h que se determina, en general, minimizando algún tipo de error, éste indica cuánto contribuye cada punto muestral al estimado en el punto t.

En general, K y h deben satisfacer ciertas condiciones de regularidad, tales como:

1. $K(z)$ debe ser acotado y absolutamente integrable en $(-\infty, \infty)$

2. Debe cumplirse que:

$$\lim_{n \to \infty} h(n) = 0$$

3. También ...

$$\int_{-\infty}^{\infty} K(x)dx = 1$$

Entre las funciones Kernel más usadas están:

1. El Kernel Rectangular:

$$K(x) = \begin{cases} 0, & \text{si } |x| > 1 \\ \\ \frac{1}{2}, & \text{si } |x| \leq 1 \end{cases}$$

2. El Kernel Gaussiano:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

3. El Kernel Triangular:

$$K(x) = \begin{cases} 1 - |x|, & \text{si } |x| < 1 \\ \\ 0, & e.o.c \end{cases}$$

4. El Kernel Biweight:

$$K(x) = \begin{cases} \frac{15}{16}\left(1 - x^2\right)^2, & \text{si } |x| < 1 \\ \\ 0, & e.o.c \end{cases}$$

5. El Kernel Epanechnikov:

$$K(x) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right), & \text{si } |x| < \sqrt{5} \\ \\ 0, & e.o.c \end{cases}$$

Si h es muy pequeño entonces el estimador de densidad por Kernel degenera en una colección de n picos cada uno de ellos localizado en cada punto muestral. Si h es demasiado grande entonces el estimado se sobresuaviza y se obtiene casi una distribución uniforme. El valor de h también depende del tamaño de la muestra, con muestras pequeñas se debe escoger una h grande y con muestras grandes se puede escoger una h pequeña.

La medida más usada del error de estimación de la función de densidad es el *MISE*[19] definido por,

$$MISE(h) = \int \left[E\left(\hat{f}_h(x) - f(x)\right)\right]^2 dx + \int E\left(\hat{f}_h(x) - E(\hat{f}(x))\right)^2 dx$$

[19]Error Cuadrático Medio Integrado.

La primera parte es el sesgo al cuadrado integrado y la segunda es la varianza integrada. La expresión anterior puede ser escrita como,

$$MISE(h) = \frac{1}{nh}\int K^2(z)dz + \frac{h^4}{4}\int \left(f^{''}(x)\right)^2 dx \left(\int z^2 K(z)dz\right)^2 + O\left(\frac{1}{nh} + h^4\right)$$

Los dos primeros términos del lado derecho forman el *AMISE*, la expansión asintótica del MISE. Minimizando el AMISE en función de h se obtiene la siguiente fórmula para un h óptimo...

$$h = \left[\frac{\int K^2(z)dz}{n\int \left(f^{''}(x)\right)^2 dx \left(\int z^2 K(z)dz\right)^2}\right]^{\frac{1}{5}}$$

A continuación se listan algunas elecciones de h:

- El primer recurso es:

$$h = \frac{rango(X)}{2(1 + log_2 n)}$$

- El segundo recurso es:

$$h = \frac{1.06\,\text{mín}\left(\hat{\sigma}, \frac{R}{1.34}\right)}{n^{\frac{1}{5}}}$$

- El tercer recurso es:

$$h = \frac{1.144\hat{\sigma}}{n^{\frac{1}{5}}}$$

donde $\hat{\sigma}$ es la desviación estándar estimada del conjunto de datos y R representa el rango intercuartílico[20], las constantes provienen de asumir que la densidad desconocida es Normal y un Kernel Gaussiano.

[20]El rango intercuartílico es la diferencia entre el tercer y el primer cuartil: Q_3-Q_1.

Para mayor información, véase el Apéndice F.

Referencias

[1] Venables, W. N. and Ripley, B. D. Ripley. (2002) Modern applied statistics with S.Springer.

[2] Olshen, A., Venkatraman ES, Lucito R and Wigler M. Circular Binary Segmentation for the Analysis of Array- based DNA Copy Number Data.

[3] Ripley, B. D. (1996) Pattern recognition and neural networks. Cambridge University Press.

[4] D. Peña Sánchez De Rivera, Estadística: Modelos y Métodos, 1. Alianza Universidad Textos, Madrid, 1994.

[5] J.H. ZAR, Biostatistical Analysis. Prentice Hall Inc., Englewood Cliffs, 1974.

[6] Davison, A. and Hinkley, D (1997), Bootstrap Methods and Their Applications.

[7] De La Garza, Cáncer guía para médicos, pacientes y familiares. Ed. Trillas, 2006.

[8] http://www.incan.edu.mx/

[9] http://www.cancer.org/Research/CancerFactsFigures/index

[10] http://www.inegi.org.mx/default.aspx

[11] http://uvigen.fcien.edu.uy/utem/genygen/genygen.pdf

[12] Santiago Grisolia, The human genome project available on http://www.cfnavarra.es/salud/anales/textos/vol24/n2/colab.html

[13] http://fp.educarex.es/fp/pruebasacceso/2009/moduloIII/cienciasdelanaturaleza/3nat03.pdf

[14] http://www.omerique.net/pub/euda/naturales/1eso/u11lacelula.pdf

[15] http://personal.us.es/pinero/citohistoma/docencia/citologia/pdf-citologia/1-CICLO20DE20DIVISION20CELULAR.pdf