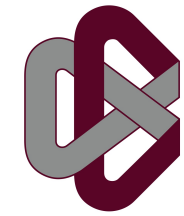

CIMAT

Centro de Investigación
en Matemáticas a.c.

Selección y Ordenamiento con Aplicaciones en Genética

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Maestro en Ciencias con Especialidad en Probabilidad y Estadística

PRESENTA:

Edgar Eduardo Rodríguez Mendoza

DIRECTOR

Miguel Nakamura Savoy

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por haberme proporcionado el apoyo económico sin el cual la realización de mis estudios de postgrado no hubiera sido posible.

Al Dr. Miguel Nakamura Savoy por aceptar ser mi guía en ésta que es mi primera incursión en el camino de la investigación. Así mismo por su paciencia, sus sabios consejos y lecciones y todas las gratas experiencias compartidas durante el desarrollo de este trabajo.

Al Dr. Victor Manuel Pérez-Abreu Carrión por su gran desempeño incondicional como mi tutor académico durante mis estudios de postgrado en CIMAT y cada una de sus valiosas lecciones.

Al Dr. Alexander De Luna y su equipo de trabajo en el Laboratorio de Biología en Sistemas Genéticos de Langebio por aceptar trabajar con nosotros durante la realización del presente trabajo. Así mismo, por su colaboración mediante la base de datos principal con la que se trabajó. Especial agradecimiento también a la Dra. Erika Garay por su asesoría acerca de la misma.

Al Dr. Rogelio Ramos Quiroga y al Dr. Octavio Martínez de la Vega por su valiosa participación como sinodales en el presente trabajo.

A mi madre, Sonia Angélica Mendoza Morfín por su apoyo incondicional en éste y cada uno de mis proyectos; mi ejemplo a seguir y mi pilar de apoyo hasta el final.

A Fiona porque siempre se alegró al verme y me recordó que no hay lugar como el hogar.

Resumen

La motivación general del presente trabajo nació de un problema específico planteado por genetistas del Laboratorio Nacional de Genómica para la Biodiversidad (LANGEBIO). Éste consiste en la identificación de cuántos y cuáles genes del genoma completo de la levadura (Saccharomyces cerevisiae) tienen un mayor impacto de manera positiva o negativa en la longevidad de dicho organismo. Dicho problema forma parte de la frontera de la investigación actual en biología debido a sus potenciales aplicaciones importantes en medicina.

Históricamente, el problema de selección cualitativa de genes se ha identificado como un problema estadístico bona fide y se ha abordado principalmente mediante técnicas relacionadas con pruebas de hipótesis múltiples (PHM). En el presente trabajo, tras la elaboración de una extensiva reseña bibliográfica sobre dichos métodos, se encontró que una PHM podría no ser suficiente para responder a la pregunta de investigación así planteada. No sólo eso, sino que una PHM podría estar dando respuesta a la pregunta equivocada mediante métodos de cuantificación de incertidumbre que no corresponden a los requerimientos del problema original. Bajo estos hallazgos se motivó la búsqueda de una metodología diferente que partiera precisamente de la pregunta original – ¿Cuáles son los mejores individuos (genes)? –y permitiera hacer una selección cualitativa con su correspondiente aseveración de incertidumbre. Es así como se converge al estudio de las Metodologías de Selección y Ordenamiento (RSM) propuestas por Bechhofer [1954].

Aún entonces, tras una nueva revisión bibliográfica se encontró que los métodos clásicos de la RSM resultan ser insuficientes para los casos en los que se tiene un número masivo de datos, como es común en el contexto de genética y en particular el caso de aplicación. Lo anterior motivó la exploración del estado del arte de las nuevas metodologías de Selección y Ordenamiento, principalmente las propuestas en Cui & Wilson [2008]. Se pretende que el presente trabajo funcione como un resumen y una reseña crítica comparativa de dichas metodologías y las técnicas que tradicionalmente se han utilizado en PHM. El resultado final es una disertación crítica con interesantes conclusiones. El presente trabajo muestra además el resultado de la interacción interdisciplinaria con los genetistas, a quienes se propuso como objetivo recomendar herramientas tanto teóricas como computacionales, implementadas y justificadas estadísticamente para resolver su problema. Las recomendaciones se acompañan de una herramienta de software interactivo específicamente diseñado para este problema que hace las funciones de herramienta didáctica y computacional.

Índice general

Capítulo 1

Introducción

Seleccionar una entre varias alternativas constituye un problema recurrente en la vida diaria de cada persona. Algunas selecciones, por su naturaleza, tienen consecuencias de relativamente poca relevancia, por ejemplo, qué color usar en cierta ocasión o qué ordenar para cenar en un restaurante. Algunas otras, en cambio, pueden ser de crucial importancia no sólo para el agente selector sino para todo su entorno, por ejemplo, por quién votar en las elecciones presidenciales o qué medicamentos son seguros para su consumo durante el embarazo.

Bajo cualquier circunstancia, un problema de selección se reduce a identificar la mejor (o la peor) de entre un conjunto determinado de k distintas opciones. Con el fin de seleccionar aquella que mejor corresponda a los intereses del selector se ha de definir algún cierto criterio que le permita discernir cuántas y cuáles, de entre las opciones posibles, constituyen las *mejores* alternativas. Frecuentemente, dicha selección se puede hacer mediante un criterio subjetivo o arbitrario, en el cual intervienen únicamente las preferencias personales del agente selector. Sin embargo, en muchos contextos, especialmente aquellos en los que la selección final tendrá un impacto relativamente significativo, resulta conveniente hacer una selección basada en *méritos* o características cualitativas que induzcan un orden natural en las opciones y permitan seleccionar aquella (o aquellas) que mejor conviene al agente que toma la decisión.

La palabra *méritos* se usa aquí en términos generales, pero puede englobar toda característica que sea de interés para quien hace la selección. Frecuentemente dichas características constituyen resultados de experimentos estadísticos (o eventos aleatorios) lo cual adhiere la noción de incertidumbre y vuelve al problema de selección un problema de optimización estocástica *bona fide*. El propósito principal de este capítulo es dar un primer acercamiento a las nociones principales de dicho problema y delinear las bases del desarrollo teórico y práctico de la tesis que se presentará en los próximos capítulos.

1.1. El Problema General de Selección

1.1.1. Selección como un Problema de Inferencia Diferente

Uno de los principales motivos por los que el problema de selección ha permanecido en constante evolución a través de los años yace en la falta de claridad conceptual que surge de manera natural cuando se plantea. Desde el punto de vista teórico, existe una fina línea entre preguntarse cuál es la mejor opción de entre un conjunto determinado y preguntarse si alguna de ellas en específico lo es.

Para ilustrar esto, supóngase que se desea hacer una predicción sobre los resultados de las elecciones presidenciales en una jornada electoral donde participan únicamente tres partidos políticos A, B y C. La pregunta inmediata en una situación como ésta es, por lo general, «*¿Cuáles serán los resultados de la elección?*». Ante esto, frecuentemente se espera obtener como respuesta una estimación, digamos, de la forma A: 20–25 %, B: 24–29 %, C: 31–36 % en donde a cada intervalo porcentual estimado se le asigna un nivel de confianza α, una medición de la incertidumbre asociada a que los intervalos porcentuales

reportados en realidad contengan el verdadero resultado de la elección. Ahora bien, si se fija para este ejemplo $1 - \alpha = 0{,}95$ y se supone que los valores reportados fueron precisamente los anteriores, es posible afirmar que el candidato C realmente posee el liderazgo de la jornada; sin embargo, no se puede concluir, ni siquiera con un 95 % de confianza, que será el ganador de la elección. La razón de esto yace en que, elegir $1 - \alpha = 0{,}95$, en este caso indica que existe un 95 % de confianza de que el porcentaje que obtendrá el candidato C en la elección sea contenido por el intervalo que va de 31 a 36 unidades porcentuales y de que, de manera análoga, el candidato B obtendrá entre 24 y 29 y el candidato A entre 20 y 25. Sin embargo, esto no dice nada acerca del evento {C>B>A}. Existe, incluso, una probabilidad positiva (que podría ser mucho mayor que 5 %) de que en la verdadera elección el candidato B, o incluso A, obtengan un porcentaje de los votos superior al de C. Por otra parte, si los porcentajes estimados hubieran sido A: 30–35 %, B: 29–34 %, C: 31–36 % este método llevaría (erróneamente) a concluir que no es posible afirmar nada acerca de los resultados de la elección. Esta situación, relativamente común que en el contexto de las elecciones, se conoce como *empate técnico*.

Una alternativa razonable al procedimiento anterior, consistiría en plantear la pregunta hipotética directa «*¿C ganará la elección?*». De esta manera se forma un paradigma binario cerrado en el cual se espera una respuesta que será un «*Sí*» o un «*No*». En el lenguaje de la estadística dicho problema se conoce como prueba de hipótesis y existe una gran gama de procedimientos clásicos y modernos para abordarlo, muchos de los cuales serán expuestos a detalle en el Capítulo 2. De manera heurística, el procedimiento usual consiste en averiguar si se tiene suficiente información para negar la pregunta de investigación (o hipótesis) y si la hay cuantificarla. Naturalmente, si se desea abordar este problema de este modo habría que plantear una hipótesis para cada candidato y es inmediato ver que mientras más candidatos haya más complejo el problema se volverá. De cualquier manera, averiguar si se tiene o no suficiente evidencia para negar que un candidato en particular ganará la elección tampoco responde de manera directa a la pregunta de mayor interés, que radica en cuál de los tres resultará victorioso. En un escenario extremo, en el cual la elección se encuentre suficientemente cercana podría no haber evidencia para negar que cualquiera de los tres candidatos será el ganador, lo cual no aporta ninguna solución útil a la pregunta de interés.

El error no radica en el método, sino en que se está intentando responder la pregunta equivocada. En una situación como la del ejemplo anterior, el interés para el ciudadano promedio radica en el resultado de la elección misma, que se traduce, comúnmente, a cuál de los tres candidatos resultará el ganador. Surge entonces la necesidad de plantear directamente la pregunta de investigación «*¿Cuál de los candidatos ganará la elección?*» y desarrollar un método adecuado para responderla. Intentar hacerlo mediante estimación o mediante una prueba de hipótesis forma una falacia conceptual importante, y bastante frecuente en la práctica, que se discutirá a detalle en el Capítulo 3.

1.1.2. Casos de Aplicación

Las aplicaciones potenciales del problema de selección en contextos reales son muchas y muy variadas. La siguiente colección de ejemplos tiene como propósito ilustrar al lector en algunas de las situaciones en las que un problema de selección puede surgir.

Ejemplo 1.1.1 (Producción de huevo)**:**
Posiblemente el primer ejemplo reportado de aplicación formal de la metodología estadística de selección a un contexto real surgió en la ciencia avícola. En dicho contexto frecuentemente es de interés reportar resultados de análisis comparativos a criadores, clientes y compradores potenciales en relación con la producción de huevo de un número determinado de *stocks* de aves de corral. Becker [1961] consideró el problema de una granja de aves de corral en la cual es de especial interés encontrar aquel *stock* de gallinas que produce la mayor cantidad de unidades de huevo y de mejor calidad (en cuanto a peso, resistencia, tamaño, *etc.*). Para ello considera 10 diferentes *stocks* etiquetados $A, B, ..., J$. Asúmase, sin pérdida de generalidad, que existe un *stock* particular, digamos el *stock* A, que es de mejor calidad

productora que los otros nueve. En la práctica no siempre es posible analizar cada gallina de cada *stock*, por lo que se analizan únicamente muestras aleatorias de cada uno tomadas bajo condiciones similares. Ahora bien, debido a que se sabe que el *stock A* es el de mejor calidad, es de esperarse que la muestra que le corresponde resulte ser siempre de mejor calidad que las demás en los resultados de cualquier prueba de calidad. Tal no es el caso debido a que existen varios factores que podrían afectar las posibilidades de que A en realidad sea identificado como el mejor. Becker [1961] identifica y explica estos tres factores:

1. **El tamaño de muestra.** Se encontró que mientras mayor sea el número de gallinas tomadas de cada *stock*, mayor es la probabilidad de que el *stock A* sea correctamente identificado como el mejor.

2. **La verdadera diferencia entre el *stock A* y el segundo mejor *stock*.** Mientras más distinguible sea el *stock A* en relación con el segundo mejor del conjunto, más fácil será identificarlo correctarmente como el mejor.

3. **La variabilidad.** Si existe gran variabilidad en la producción de huevo dentro de cada *stock* las posibilidades de una selección incorrecta aumentan.

La principal conclusión de Becker [1961] para este ejemplo radica en que, sin importar el número de aves presentes en el estudio, siempre existe una probabilidad positiva de que el mejor stock no produzca la mejor muestra y que, por tanto, de manera incorrecta no sea clasificado como el verdadero mejor. Muchas de las ideas y conclusiones de este artículo tienen una justificación teórica formal que será presentada a detalle en el Capítulo 2.

Ejemplo 1.1.2 (Pacientes con depresión)**:**
Aproximadamente 12 de cada 100 ciudadanos mexicanos entre 18 y 65 años de edad padece de depresión. Como tal, constituye un reto importante para el sector salud del país desarrollar formas efectivas de tratarla a un bajo costo. Para ello se identifican cuatro grupos principales de medicamentos antidepresivos accesibles: Meprobromato, Diazepam, Clordiazepoxido-Hidroclorido (CH) y Thioridazina. Se selecciona a continuación un grupo de pacientes con síntomas de depresión y se separan de forma aleatorizada en 5 grupos de tamaño n de manera que a cada uno de los primeros cuatro grupos se le administre un tipo diferente de medicamento. Debido a que es difícil o imposible determinar si los pacientes mejoran o empeoran debido a los efectos del medicamento mismo o debido a factores psicosomáticos relacionados con su creencia personal de que el medicamento les ayudará, se decidió administrarle al grupo restante un placebo. Es decir, los pacientes del quinto grupo estarán psicológicamente concientes de haber recibido un medicamento antidepresivo sin conocer que éste en realidad no lo es. Para evitar efectos no controlados en el experimentos todos los pacientes son tratados bajo las mismas condiciones y se mantiene oculta la identidad del medicamento recibido de los sujetos experimentales. El problema de interés, planteado mediante este ejemplo en Gibbons et al. [1977], es seleccionar cuáles, de los cuatro medicamentos, presentan un efecto mayor de mejoría en comparación con el placebo.

Otra manera, aparentemente razonable, de abordar este problema podría consistir en comparar cada uno de los cuatro medicamentos con el placebo (de manera individual o conjunta) y plantearse encontrar aquellos medicamentos que resulten ser significativamente distintos del mismo. En el lenguaje de la estadística esta metodología es conocida como comparaciones o pruebas de hipótesis múltiples (PHM) y tiene una cercana relación con el problema de selección. A pesar de que ambas preguntas son similares y podrían incluso reportar conclusiones análogas, son conceptualmente muy diferentes y, dependiendo del contexto, podrían difererir significativamente. Una disertación crítica acerca de ambas metodologías se presentará con detalle en el Capítulo 3.

El siguiente ejemplo tiene como objetivo ilustrar que el problema de selección puede también aplicarse a situaciones en las cuales es de interés encontrar las *peores* poblaciones de un conjunto. Aquí, *peores*,

de manera análoga, se entiende como aquellas que minimizan alguna característica de interés para el agente selector.

Ejemplo 1.1.3 (Cortado de Diamantes)**:**
Los diamantes son minerales preciosos derivados del carbono de gran valor no sólo por su pureza y resistencia sino por su rareza y dificultad de manejo. Cada diamante en bruto requiere pasar por un delicado proceso multi-etapa conocido como cortado en el que se le pule y se le da forma de gema. Debido a la dureza del material este procedimiento es muy delicado y requiere de sofisticadas herramientas, tecnologías y experiencia. Dos de los intereses principales de una compañía cortadora de diamantes son la pérdida de masa absoluta de la gema y la precisión del procedimiento de cortado mismo. Debido a los altos costos de producción a una compañía le interesa seleccionar entre un conjunto de k posibles técnicas distintas de cortado $1, 2, ..., k$, aquella que cumpla de mejor manera con las especificaciones anteriores. Para ello diseña un experimento en el cual toma una muestra aleatoria de tamaño n de gemas cortadas bajo condiciones similares con cada una de las distintas técnicas. Para el primer objetivo de mide la diferencia entre la masa del objeto antes y después del proceso de cortado y para el segundo se mide la variabilidad en las dimensiones del producto final obtenido. La pregunta de investigación para el problema de selección será entonces encontrar aquella técnica que tiene la menor medición en ambos aspectos. El problema de selección de los tratamientos con menor varianza se expone de manera breve en el Capítulo 2 y se trata con mejor detalle en el Capítulo 5 de Gibbons et al. [1977].

1.1.3. Introducción a las Metodologías de Selección y Ordenamiento (RSM)

La estructura general de todo problema de selección parte de la premisa de que se requiere hacer una evaluación cualitativa de un conjunto de k distintas opciones $1, 2, ..., k$ para obtener aseveraciones acerca de cuál (o cuáles) constituye la mejor (o la peor) del conjunto. El término *mejor* (o *peor*) hace referencia a un significado específico dentro del contexto del problema particular. En el lenguaje de la estadística, al conjunto de opciones para seleccionar se le conoce como *poblaciones*.

La metodología estadística clásica de Selección y Ordenamiento (RSM) se originó en Bechhofer [1954]. Como cualquier método de comparación estadística entre k diferentes poblaciones, las metodologías de selección y ordenamiento se basan en un conjunto de observaciones independientes resultantes de un procedimiento de muestreo de cada una de las k poblaciones a comparar. En el contexto de estadística esto es equivalente a tomar una muestra aleatoria $x_1, x_2, ..., x_n$ de cada una de k distintas poblaciones con distribución de probabilidad $F(x, \theta_j)$ para $j = 1, 2, ..., k$. Los detalles técnicos de esta afirmación se reservan para el Capítulo 2.

Con el fin de explicar cada uno de los elementos principales de un problema de selección en términos de RSM se partirá del siguiente ejemplo:

Ejemplo 1.1.4 (Semillas de maíz)**:**
Un ingeniero agrónomo mexicano desea sembrar maíz para un estudio de fertilidad. En la actualidad existen distintas variedades de maíz en el mercado en México, siendo las más cultivadas: amarillo duro (AD), blanco duro (BD), blanco dentado (BE), amarillo dentado (AE), Harinoso y Morocho (HM) y Reventón y Dulce (RD) (CIMMYT [1994]). El ingeniero siembra, en seis parcelas similares, cantidades iguales de cada variedad de maíz bajo las mismas condiciones y posteriormente se asegura de que todas las condiciones de mantenimiento (fertilizante, regado, cantidad de luz, *etc.*) sean lo más homogéneas posibles para cada parcela. El objetivo del estudio del agrónomo será entonces averiguar cuál variedad de maíz produce la mayor cosecha (medida en gramos) al final de la temporada. Por simplicidad se supondrá que el efecto de interacción entre las distintas variedades de maíz cultivadas en una misma parcela es irrelevante y que la variabilidad en la producción se explica completamente en términos del tipo de grano.

Tal como se introduce en Bechhofer [1954], todo problema típico de selección y ordenamiento (RSM)

cuenta básicamente con los siguientes elementos principales:

- k **poblaciones independientes.** En el ejemplo de las semillas de maíz el conjunto de poblaciones es {AD, BD, BE, AE, HM, RD} y por tanto $k = 6$.

- k **muestras independientes de tamaño** n**.** En el ejemplo anterior, cada variedad de maíz fue sembrada en cantidades iguales en cada una de las seis parcelas independientes bajo condiciones similares. Si se denota por $n/6$ la cantidad de semillas de cada variedad de maíz que fueron sembradas en cada parcela obtendremos mediante este procedimiento una muestra aleatoria de tamaño n de cada variedad de maíz que consistirá en las mediciones, en gramos, de la cosecha de cada semilla de cada variedad de grano. Es importante resaltar el supuesto de independencia de las observaciones que en este contexto es asegurado por la homogeneidad de las condiciones del proceso de sembrado y mantenimiento de las plantas de maíz. La mayor parte de la metodología clásica para problemas de selección está cimentada bajo el supuesto de que se ha tomado un tamaño de muestra común de cada tratamiento. Sin embargo, existen variantes en la literatura para casos que no cumplen dicho supuesto. Véase, por ejemplo, Bechhofer [1954].

- **Meta de selección**. La mayoría de las metodologías estadísticas diseñadas para tratar el problema de selección tienen como objetivo general una de las siguientes metas Gibbons et al. [1977]:

  1. Seleccionar la mejor población.
  2. Seleccionar el conjunto (ordenado o no ordenado) de las mejores t poblaciones donde $1 \leq t < k$.
  3. Seleccionar un número aleatorio de poblaciones r que contenga a las mejores t poblaciones.
  4. Seleccionar un número fijo de poblaciones r que contenga a las mejores t poblaciones.
  5. Ordenar todas (o un subconjunto de) las k poblaciones de mejor a peor (o viceversa).
  6. Seleccionar un número aleatorio de poblaciones r que contenga todas las poblaciones que son mejores que un cierto control o estándar.

  Las principales técnicas clásicas y modernas para las metas 1–4 serán expuestas en el capítulo 2. Las metodologías especializadas en las metas 5 y 6 pueden ser consultadas en Gibbons et al. [1977] y Bechhofer et al. [1995].

  En general, la elección de una meta en particular depende del contexto. En el caso del ejemplo del maíz el agrónomo podría estar interesado no precisamente en la mejor variedad de maíz sino en las tres mejores o, posiblemente, si la situación particular lo amerita, un conjunto de tres variedades entre las cuales se encuentra la mejor con cierto nivel de *confianza*. Este último concepto se desarrollará a detalle en el Capítulo 2. En lo que resta de esta sección, por simplicidad, se asumirá que la meta del agrónomo coincide con la Meta 1, es decir, está en busca de la única mejor variedad de grano. La mayoría de los argumentos presentados a continuación son fácilmente expandibles al caso de cualquiera de las otras metas de la lista.

- **Noción de distancia.** Asumiendo que entre el conjunto de k poblaciones existe al menos una mejor, si ésta difiere en al menos un mínimo umbral de las restantes, la prioridad es identificarla. Para ello se ha de definir una forma apropiada de medir *distancia*, una función que permita cuantificar las diferencias entre la población que de desea identificar como la mejor y las restantes.

  Para el ejemplo del agrónomo, supóngase que es sabido que para cada variedad de maíz en particular, la producción X , en gramos, al final de una temporada de cosecha tiene una distribución de probabilidad $G(x;\theta)$ indexada por un parámetro θ. Por ejemplo, $G(x;\theta)$ podría representar la

distribución normal con media θ, la distribución Poisson con intensidad θ, la binomial con probabilidad de éxito θ, *etc.* Las verdaderas diferencias entre las k poblaciones estarán dadas en términos de los parámetros individuales que se denotarán como $\theta_1, \theta_2, ..., \theta_k$. La más sencilla función de distancia entre dos poblaciones i y j estaría dada por la simple diferencia entre sus parámetros, digamos $\delta(i, j) = \theta_i - \theta_j$ pero dependiendo del contexto otras funciones de distancia pueden resultar más convenientes. La información detallada acerca de la interpretación de la función de distancia y distintos ejemplos de la misma se expondrán en el Capítulo 2.

- **Estadísticas Resumen.** Reescrito en el lenguaje estadístico descrito en el punto anterior, el problema de selección se traduce a encontrar la variedad de maíz cuya producción sigue la distribución de probabilidad $G(x; \theta)$ con el mayor parámetro θ. Sin embargo, dado que los parámetros $\theta_1, \theta_2, ..., \theta_k$ son no observables en la práctica, parece razonable utilizar la información disponible en los datos para obtener un estimador $\hat{\theta_j}$ para cada $\theta_j, j = 1, 2, ...k$. En el caso de nuestro ejemplo particular $\hat{\theta_j}$ puede ser la media muestral pero, dependiendo del contexto, puede no ser la única ni la óptima opción. Por ejemplo, si en el caso del agrónomo fuera de interés no sólo la cantidad neta de maíz en gramos obtenida por cada semilla de maíz si no tambien el número (en unidades) de elotes producida por cada semilla una mejor alternativa para $\hat{\theta_j}$ podría ser una media ponderada de los gramos de maíz producidos, donde los pesos $w_1, w_2, ..., w_n$ para cada unidad de maíz estarían dados por el número de elotes obtenidos en cada planta individual de maíz. De cualquier manera, la elección de $\hat{\theta_j}$ debe hacerse de manera que capture la mayor cantidad de información relevante del contexto posible para hacer la selección.

- **Criterio de Selección.** Una vez que se ha seleccionado una estadística resumen $\hat{\theta_j}$ apropiada es necesario definir el criterio de selección que se usará para tomar la decisión de acuerdo con la meta designada de selección. En acuerdo con la metodología clásica RSM esto se hace de manera natural seleccionando aquella población que produjo la mayor $\hat{\theta_j}$ observada como la mejor población.

- **Probabilidad de Selección Correcta (PCS)** A pesar de que el criterio de selección descrito anteriormente es muy simple y razonable, la posibilidad de error siempre estará presente pues la verdadera mejor población, *i.e.* aquella con el parámetro θ mayor, no siempre producirá el mayor valor de entre los estimadores muestrales. La mejor variedad de maíz no siempre será aquella que produzca la mayor cantidad de maíz en una temporada determinada, inclusive aunque el experimento se repita. Tal y como se expuso en el Ejemplo 1.1.1. existe un conjunto de factores que afectan dichas posibilidades; incluso es posible que la probabilidad de que la verdadera mejor variedad de maíz produzca la mayor cosecha sea un valor cercano a cero si ésta está lo suficientemente cerca de la segunda mejor. En términos de las técnicas clásicas de RSM se define una selección como correcta si el valor de θ asociado a la población seleccionada coincide con el de la verdadera mejor. Es decir, en el caso más simple, siempre que las verdaderas mejor y segunda mejor sean estrictamente distintas, sólo una posible selección será correcta y la probabilidad de encontrarla se define como Probabilidad de Selección Correcta (PCS):

$$\text{PCS} = P(\text{máx}_{j=1,2,...,k}\hat{\theta_j} = \text{máx}_{j=1,2,...,k}\theta_j). \tag{1.1}$$

La PCS constituye el objeto central de interés en la teoría de RSM pues en ella se encuentra toda la información relevante acerca de la calidad alcanzable en un determinado proceso de selección y, por consiguiente, con base en ella se han desarrollado numerosos métodos y resultados que forman parte fundamental de esta tesis. Como se discutirá a detalle en el siguiente capítulo, las principales limitaciones de la RSM son básicamente dos: falta de poder computacional y el aumento excesivo del número de poblaciones k. La primera se debe a que, en muchos casos, no existe una expresión numérica exacta cerrada para (1.1), y por lo tanto, es necesaria la implementación de

algoritmos computacionales que no siempre son simples u óptimos. La segunda, y quizás la más importante, pues es la razón por la que el problema de selección y ordenamiento permanece hasta hoy en la frontera de la investigación estadística, es que la teoría clásica de selección iniciada en Bechhofer [1954] resulta ser poco aplicable para casos en los que k es muy grande. Por razones que se detallarán en el siguiente capítulo, en dichos casos PCS tiende a ser cero. Surge entonces la necesidad de definir nuevas metodologías de selección específicas para estos casos, mismas que dan lugar a una revolución conceptual en el tema del problema de selección.

1.1.4. La Hipótesis de Homogeneidad

Un enfoque alternativo al expuesto en la sección anterior consiste en buscar o cuantificar evidencia experimental en contra de la homogeneidad o igualdad teórica de los parámetros θ_j que, como se explicó en la sección anterior, representa un atributo, descripción o respuesta de cada población. En el caso del Ejemplo 1.1.4 el interés yace en si las variedades de maíz difieren en la cantidad de cosecha producida. Si dicha cantidad se modela mediante un parámetro θ_j para la $j-$ésima variedad de maíz, una alternativa razonable podría ser establecer la siguiente expresión, conocida en la literatura como *hipótesis de homogeneidad*:

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_k. \tag{1.2}$$

La llamada *hipótesis alternativa* para este caso es usualmente que los parámetros θ_j no son todos iguales o que existe al menos un par de parámetros l y m tales que $\theta_l \neq \theta_m$.

El procedimiento estadístico usual para el problema así descrito se conoce como *prueba de homogeneidad* y, de manera heurística, consiste en averiguar si existe suficiente información experimental para rechazar H_0. La hipótesis de homogeneidad (1.2) sólo es una instancia de una gran gama de situaciones clásicas en la teoría estadística que se remontan a los tiempos de Fisher. Muchos métodos de solución a problemas de prueba de hipótesis como (1.2) han sido extensivamente estudiados; un breve resumen se encuentra expuesto en la primera parte del Capítulo 2 de este trabajo.

Si la conclusión que ofrece una prueba de homogeneidad resulta resolver satifactoriamente la meta principal y definitiva del estudio, no existe necesidad de buscar métodos alternativos. Sin embargo, existen situaciones en la cuales es necesaria información adicional que una simple prueba de homogeneidad como (1.2) no es capaz de proporcionar. Supóngase, por ejemplo, en el caso del Ejemplo 1.1.4 que el agrónomo realiza una prueba de homogeneidad y que la hipótesis (1.2) es rechazada, es decir, existe evidencia de los datos para rechazar que todas las variedades de maíz son iguales en términos de la cantidad de cosecha producida. A términos prácticos esta conclusión es apenas satisfactoria debido a que:

1. No queda claro cuántas y cuáles variedades de maíz difieren de cuáles y en qué dirección.

2. No se sabe cuáles variedades de maíz pueden considerarse las mejores en términos de producción.

Para el caso de (1) existe un conjunto de metodologías especializadas conocidas como Pruebas o Comparaciones Múltiples (PHM) que pueden ayudarle al agrónomo a hacer inferencia simultánea sobre las variedades de maíz de manera satisfactoria hasta cierto punto. Con el propósito (1) en mente, dichas técnicas se enfocarán en identificar las variedades de maíz que son *diferentes o significativas* y que son distinguibles de las demás (véase Capítulo 2 para los aspectos técnicos de PHM). Como se discutirá en el Capítulo 3 las técnicas PHM pueden también aportar información para el caso (2) pero para este propósito las técnicas de RSM son más apropiadas.

Visto desde el punto de vista anterior, el lector podría estar tentado a pensar que las técnicas de RSM sólo representan una alternativa más de solución para el caso en el que la hipótesis de homogeneidad es rechazada. La realidad es que no es así. En el Ejemplo 1.1.4 el agrónomo está interesado en seleccionar la

mejor (o las dos mejores) variedad de maíz, y reportarla como aquella que produce una mayor cosecha. La prueba de homogeneidad en este caso sólamente podrá decirle si las clases de maíz son equivalentes o no. No está diseñada para dar conclusiones relacionadas con el orden de los individuos experimentales ni determinar cual es el mejor o el peor del conjunto y por tanto no puede resolver directamente el problema del agrónomo. A pesar de que se han propuesto muchas modificaciones a los métodos relacionados con la prueba de homogeneidad y comparaciones múltiples ninguna de ella es capaz de atacar el problema de ordenamiento directamente. Ahora bien, si por el contrario, la hipótesis H_0 no fuese rechazada y hubiera que hacer inferencia acerca del orden cualitativo de las k poblaciones en cuestión, la conclusión disponible para este caso, que no hay evidencia suficiente para determinar que las semillas de maíz no son equivalentes, no sería realista ni útil para el investigador. Las técnicas de RSM has sido diseñadas específicamente para este tipo de problemas. La crítica comparativa completa puede encontrarse en el Capítulo 3.

1.1.5. Actualidad y Relevancia Científica

El problema de selección y ordenamiento *per se* no es un problema nuevo. Desde que Bechhofer [1954] planteó por primera vez la filosofía de selección, el problema mismo de seleccionar aquellas mejores poblaciones de un conjunto de k opciones posibles basado en un análisis estadístico cualitativo ha evolucionado constantemente (véase el Capítulo 2 para una breve reseña histórica de las principales aportaciones publicadas).

Sin embargo, como se discutirá en el siguiente capítulo, muchos de los resultados teóricos más importantes en la RSM se basan en el cálculo y la interpretación de la probabilidad de selección correcta (1.1), y desafortunadamente, la gran mayoría requieren de aproximaciones numéricas que a su vez requieren de alto poder computacional. Si bien una gran parte de los resultados teóricos tenían expresiones analíticas cerradas, la falta del poder de cómputo suficiente había frenado, de alguna manera, las posibilidades de su aplicación. Con el surgimiento de las nuevas tecnologías informáticas ha sido posible la implementación de técnicas de cómputo estadístico intensivo que han facilitado, no sólo el cálculo exacto de la PCS, sino el desarrollo de técnicas alternativas sofisticadas para aproximarla a varios grados de precisión. Lo anterior ha reducido significativamente los tiempos de corrida y ha permitido obtener resultados de manera más eficiente.

La razón principal por la que el problema de selección y ordenamiento ha vuelto a figurar en la frontera de la investigación en estadística yace en que, con el surgimiento mismo del poder computacional y el avance de la ciencia, los avances en la teoría estadística involucran de manera cada vez más frecuente el análisis de conjuntos de datos de tamaño masivo. En finanzas, por ejemplo, existen bases de datos del orden de decenas de millones, o incluso tal vez más. Situaciones similares son cada vez más comunes en áreas de conocimiento como genética, medicina e ingeniería. Con esto en mente, surge la necesidad de la aplicación de un método que permita al analista realizar una *selección* o *filtrado* que auxilie en la identificación de la información verdaderamente relevante entre un conjunto de datos masivo. Cui and Wilson [2008] identifican, por primera vez, que gran parte de la metodología clásica propuesta por Bechhofer [1954] y expandida a través de los años resultaba ineficaz cuando se aplicaba a conjuntos de datos masivos. Esto es debido a que, con el aumento de la cantidad de información disponible, realizar una selección correcta resulta intuitivamente más desafiante. Gran parte de los resultados expuestos en Cui and Wilson [2008] serán resumidos en el Capítulo 2 y posteriormente implementados en el Capítulo 4.

1.2. Objetivos y Estructura de la Tesis

En la sección anterior se describió la necesidad emergente de la implementación de nuevas técnicas de selección a contextos en los cuales el número de poblaciones k es grande. Particularmente, la genética es un área de conocimiento en la cual grandes bases de datos como éstas surgen de manera frecuente debido a su quehacer, y por lo anterior constituye una de las fuentes más interesantes de aplicaciones potenciales para el tema de selección. La motivación de la temática de esta tesis radicó en un problema específico planteado por genetistas, que apuntó de manera natural hacia técnicas de selección y ordenamiento. Lo anterior generó una interacción interdisciplinaria con usuarios potenciales, a quienes hubo que hacer recomendaciones de herramientas, tanto teóricas como computacionales, implementadas y justificadas estadísticamente. El caso particular de aplicación se incluye en el Capítulo 4 junto con la descripción general de las herramientas computacionales desarrolladas específicamente para este problema concreto.

Un segundo objetivo para esta tesis se introdujo en la Sección 1.1.4 mediante una disyuntiva conceptual importante en relación con la elección de la técnica apropiada para resolver un problema de selección. La motivación principal yace en que muchas aplicaciones de este problema en la práctica (véase por ejemplo Golub et al. [1999] y De Luna et al. [2014]) se han implementado técnicas basadas en comparaciones múltiples para resolver un problema que es, en esencia, un problema de selección. El segundo objetivo concreto de la tesis será entonces promover un diálogo concientizador respecto al uso de ambas de técnicas y puntualizar directamente sus diferencias y escenarios objetivo. Para ello se ha desarrollado una reseña extensiva acerca de los principales componentes teóricos de cada metodología y se ha elaborado un ensayo crítico que contrastará de manera objetiva los aspectos descritos anteriormente.

El presente trabajo está conformado por cuatro capítulos. El Capítulo 2 consta de una reseña teórica acerca de los dos pilares principales en la metodología del problema de selección: pruebas de hipótesis múltiples (PHM) (véase Sección 1.1.4) y selección y ordenamiento (RSM) (véase Sección 1.1.3). Es en este capítulo donde se introducirán aspectos conceptuales y notación a detalle que permitan al lector familiarizarse con ambas metodologías desde un punto de vista técnico general. El Capítulo 3 consistirá de una disertación crítica de caracter comparativo entre ambas metodologías, se expondrán las semejanzas y diferencias tanto conceptuales como operativas de ambas metodologías y se describirán los escenarios potenciales en los cuales es más recomendable usarlas y la correspondiente justificación teórica. Finalmente, el Capítulo 4 presenta y describe una herramienta computacional diseñada específicamente para los objetivos de la tesis y un caso de aplicación que se espera, ilustrará varios de los principales aspectos de un problema típico de selección y ordenamiento.

Capítulo 2

Antecedentes Técnicos

Dentro de los principales acercamientos a un método de solución al problema de selección planteado en el capítulo anterior destacan las pruebas de hipótesis (Dudoit et al. [2003]). Éstas constituyen un procedimiento muy popular no sólo por su aparentemente intuitiva interpretación sino también por su relativamente fácil implementación computacional. La idea general de estos procedimientos está íntimamente ligada con la identificación de elementos *significativos* en un entorno multivariado. De este modo, utilizando el concepto de *significativo* como *proxy* se busca identificar a las mejores poblaciones del conjunto de interés.

Existe, por otro lado, la línea de investigación correspondiente a las metodologías de selección y ordenamiento (RSM) propuesta en Bechhofer [1954], que fue concebida, desde sus inicios como un procedimiento distinto del enfoque clásico de pruebas de significancia. En la RSM, el objetivo pasa de buscar elementos *significativos*, a buscar los *mejores* elementos dentro de cierto conjunto en relación a una característica particular de interés. Visto de este modo, la RSM se concibió como una metodología específicamente diseñada para resolver el problema de selección tal y como se introduce en el Capítulo 1. Es de interés, por tanto, comparar ambas metodologías y, de ser distintas, identificar los escenarios en los cuales difieren y las principales razones por las que ésto ocurre.

Antes de pasar a la parte comparativa, que se reserva para el Capítulo 3, es de especial importancia realizar una reseña conceptual de ambas metodologías. En el presente capítulo se presentarán las principales nociones, métodos, escenarios potenciales y posibles conclusiones de ambas metodologías de manera que, en el Capítulo 3, se pueda presentar una crítica comparativa apropiada. El capítulo se divide en tres grandes secciones. La Sección 2.1 estará dedicada al entorno de pruebas de hipótesis múltiples (PHM). La Sección 2.2 a las nociones clásicas de la metodología RSM. Finalmente, la Sección 2.3 reseñará y justificará las metodologías modernas de selección y ordenamiento así como un breve resumen de los problemas abiertos de la investigación actual en el problema de selección.

2.1. Pruebas de Hipótesis Múltiples

Una hipótesis estadística es una conjetura científica que puede ser sujeta a prueba en función a la observación de un proceso modelado a través de variables aleatorias. Dentro del contexto de la inferencia estadística formal, una hipótesis estadística contiene una afirmación específica acerca del espacio parametral Θ que describe el proceso aleatorio que rige el estado o el comportamiendo de una población de interés. El objetivo primordial de una prueba de hipótesis es inferir, con base en la información muestral disponible de la población, acerca de la plausibilidad de la hipótesis de trabajo. El enfoque más común para hacer lo anterior (conocido como Neyman-Pearson) consiste en plantearlo como una decisión binaria: existe evidencia suficiente para rechazar la hipótesis, o no.

Bajo este enfoque, toda hipótesis estadística particiona el espacio parametral Θ en dos regiones,

aquella en la cual la hipótesis es cierta Θ_0 y su contraparte en la cual es falsa Θ_1. De esta manera, en general, una prueba de hipótesis estadística posee la siguiente forma:

$$H_0 : \theta \in \Theta_0; H_1 : \theta \in \Theta_1, \tag{2.1}$$

donde H_0 y H_1 reciben el nombre de *hipótesis nula e hipótesis alternativa* respectivamente. Por ejemplo, si θ representara la media anual de masa en gramos de producción de maíz en un determinado año, podría ser de interés que esté por encima o por debajo de cierto umbral mínimo de tolerancia θ_0. La hipótesis nula para este caso estaría dada por $H_0 : \theta \geq \theta_0$ y la hipótesis alternativa correspondiente sería $H_1 : \theta < \theta_0$.

En términos generales, el procedimiento de prueba de hipótesis consiste de una disyuntiva en la cual el experimentador debe determinar, tras el proceso de muestreo, si rechazar o no H_0 es lo más razonable. Visto de esta forma, un procedimiento de prueba de hipótesis es una regla de decisión que especifica para qué valores observados de la muestra es razonable rechazar H_0. Al subconjunto del espacio muestral para el cual dicha regla de decisión apoya el rechazo de H_0 se le conoce como *región de rechazo* y por ello, su definición es parte fundamental de la teoría de pruebas de hipótesis.

Un debate filosófico frecuente en el tema de pruebas de hipótesis yace en la naturaleza binaria de sus conclusiones. Una cuestión controversial en esta área de conocimiento es aquella que plantea si es equivalente *aceptar* H_0 y *no rechazar* H_0. En el primero de los casos el experimentador estaría dispuesto a admitir una aseveración como verdadera mientras que en el segundo no está seguro acerca de la veracidad de H_0, pero admite no tener suficiente información para asegurar lo contrario. Una situación similar pero incluso más conflictiva conceptualmente ocurre con las aseveraciones *rechazar* H_0 y *aceptar* H_1. Si bien existen referencias bibliográficas en las cuales toda discusión acerca de dicha disyuntiva filosófica se omite en favor de un enfoque más práctico en el cual una acción entre dos posibles alternativas ha de ser tomada (por ejemplo Casella and Berger [2008]), también existen publicaciones que niegan de manera más crítica la suposición de equivalencia entre ambos pares de afirmaciones, por ejemplo Tukey [1991]. Para propósitos del presente trabajo se reservará toda discusión crítica respecto de esta disyuntiva hasta el siguiente capítulo donde se disertarán con más precisión aspectos conceptuales de la filosofía de pruebas de hipotesis.

El resto de esta sección consistirá de una reseña bibliográfica acerca de los principales elementos de una prueba de hipótesis. Para propósitos de la tesis, el objetivo principal es describir la metodología básica de una prueba de hipótesis múltiple (PHM) e identificar los conceptos teóricos principales que intervienen en su desarrollo. La Sección 2.1.1 reseñará el caso más simple en el cual interviene una única hipótesis y servirá como motivación a la Sección 2.1.2 en la cual se tratarán los elementos análogos y nuevos que se introducen con la extensión al caso múltiple. Se finalizará esta sección con la conexión existente entre las metodologías de pruebas hipótesis múltiples y el problema de selección que es el tema principal de la tesis (véase la Sección 1.1.4).

2.1.1. Motivación: Pruebas de Hipótesis Simples

Definiremos una prueba de hipótesis simple como una prueba en la que interviene una única hipótesis nula H_0 y su contraparte, o hipótesis alternativa, H_1. El término *simple* hace referencia a que el problema consiste de una única conjetura a probar; el caso más general en el cual se desean probar múltiples conjeturas de manera simultánea será abordado en la siguiente sección.

Una analogía común en la literatura a una prueba de hipótesis simple es un juicio oral en el cual un ciudadano es acusado de un crimen particular. En dicha situación, el fiscal tratará de probar la culpabilidad del acusado y, sólo cuando haya suficiente evidencia para ello, éste será condenado. El juez, en este caso, se enfrentará a un problema donde intervienen dos hipótesis: H_0 : El acusado es inocente y H_1 : El acusado es culpable. Nótese la importancia conceptual del orden de la elección

de las hipótesis, la hipótesis nula es siempre la hipótesis que se encuentra en prueba directa y cuya veracidad no se está dispuesto a rechazar a menos que haya evidencia suficiente para ello. En el caso del juicio, el acusado permanecerá siendo inocente a menos que haya evidencia suficiente para asumir lo contrario. Visto de esta forma, el juez no quisiera rechazar la hipótesis nula a menos que haya evidencia contundente para ello; rechazarla cuando en realidad es cierta constituiría un error grave pues se enviaría a un individuo inocente a prisión. En el contexto de pruebas de hipótesis este error se conoce como *error tipo I* y es de especial importancia controlar las posibilidades de que ocurra. Análogamente, si el juez decide que no existe evidencia suficiente para condenar al acusado siendo que éste en realidad es culpable estaría cometiendo otro tipo de error, quizás subjetivamente de menor impacto que el error tipo I, que en el contexto de pruebas de hipótesis se denomina *error tipo II*. Típicamente, la elección del orden de H_0 y H_1 se fija de acuerdo con el contexto y se hace de tal manera que reducir el error tipo I sea de mayor prioridad que reducir el error tipo II.

Los posibles resultados del juicio oral se resumen en el Cuadro 2.1.

	H_0 **cierta (Inocente)**	H_0 **falsa (Culpable)**
H_0 **rechazada (Condenado)**	Error Tipo I	Decisión Correcta
H_0 **no rechazada (Libre)**	Decisión Correcta	Error Tipo II

Cuadro 2.1: Dos posibles tipos de error en una prueba de hipótesis simple

Aunque el error tipo I y el error tipo II aparentan representar situaciones independientes, en realidad guardan una relación muy cercana que forma parte clave de la teoría de pruebas de hipótesis. Idealmente, el investigador estaría interesado en reducir las probabilidades de error tipo I y error tipo II de manera simultánea para que ambas fueran lo más pequeñas posibles. Sin embargo, se puede demostrar que guardan una relación inversa que imposibilita este procedimiento.

Para ilustrar lo anterior, supongamos que el juez del ejemplo inicial de la sección desea no cometer ningún error de tipo I en ninguno de los juicios en los que participa, es decir, desea que la probabilidad de que condene a un hombre inocente sea estrictamente cero. Sin embargo, la única forma de lograr esto es ignorar toda la evidencia existente en contra de cualquier individuo y declararlos a todos inocentes. Dicho procedimiento ocasionaría que todos los acusados que son verdaderamente culpables queden libres, y por lo tanto, aumentaría la cantidad de errores de tipo II cometidos. Por otro lado, si el juez adoptara la actitud de no cometer ningún error de tipo II, es decir, no deseara dejar libre a ningún individuo culpable, la única opción posible sería condenar a todos los acusados que lleguen a juicio sin necesidad de evidencia alguna. Ésto implicaría que una cantidad inadmisible de ciudadanos inocentes sean sentenciados a prisión, y por consiguiente, el número de errores de tipo I aumentaría de forma alarmante. En el contexto de las pruebas de hipótesis simples ocurre una situación similar. Es posible demostrar que reducir de manera simultánea ambos tipos de error no será posible pues reducir uno de ellos aumentará el otro, como se especifica a continuación:

$$P(\text{error tipo I}) \to 0 \implies P(\text{error tipo II}) \to 1, \tag{2.2}$$

$$P(\text{error tipo II}) \to 0 \implies P(\text{error tipo I}) \to 1. \tag{2.3}$$

Debido a su naturaleza, controlar la probabilidad de error de tipo I siempre tiene mayor prioridad que controlar la probabilidad de error de tipo II. Como no es posible controlar la probabilidad de error tipo I llevándola a cero, surge la necesidad de especificar una cota superior aceptable para ella. Dicha cota se conoce como *nivel de significancia* y tradicionalmente se denota por α; si la prueba de hipótesis en cuestión cumple con este supuesto decimos que es una prueba de nivel α. Una vez que el investigador se ha asegurado que el procedimiento de prueba que aplicará pertenece a una colección de procedimientos

que controlan la probabilidad de error de tipo I a un cierto nivel α especificado, procederá a seleccionar la prueba de hipótesis con menor probabilidad de error de tipo II posible, entre otras probabilidades deseables. En lenguaje estadístico, las pruebas con mejor desempeño, en términos de reducción de error tipo II, son conocidas como *prueba uniformemente más potente (UMP)*. Existen en la literatura una gran gama de métodos para encontrar una prueba UMP, muchos de los cuales se pueden consultar en Casella and Berger [2008]. Los detalles técnicos no se trabajarán aquí pues van más allá de los objetivos de esta sección.

Una vez que se ha establecido la filosofía de una prueba de hipótesis, el trabajo restante consistirá en caracterizarla a través de los elementos estocásticos que la definen. Para ello se propone el siguiente ejemplo:

Ejemplo 2.1.1:
Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de una población $N(\theta, 1)$. Consideremos que es de interés probar la hipótesis $H_0 : \theta = 0$ contra $H_1 : \theta \neq 0$.

En términos formales, los principales componentes de una prueba de hipótesis simple son los siguientes:

- **Hipótesis de trabajo.** La aseveración acerca del espacio parametral Θ que nos interesa probar, junto con su complemento o hipótesis alternativa. En este caso particular es $H_0 : \theta = 0$ contra $H_1 : \theta \neq 0$ o, en términos del espacio parametral, $H_0 : \theta \in \Theta_0$ contra $H_1 : \theta \in \Theta_1$ donde $\Theta_0 = \{0\}$, $\Theta_1 = \mathbb{R} - \{0\}$ y $\Theta_0 \cup \Theta_1 = \mathbb{R} = \Theta$.

- **Estadístico de Prueba.** Valor calculado en función de la muestra observada, frecuentemente para resumir la información contenida en los elementos observados para propósitos comparativos. La elección del estadístico de prueba conveniente es crucial en toda prueba de hipótesis y por lo general se hace con base en el contexto. El estadístico de prueba escogido debe ser aquel que recoja información de la muestra y sea capaz de dar evidencia, mediante una distribución de probabilidad definida, en contra de la hipótesis nula que pueda ser cuantificada. En el caso del ejemplo 2.1.1 θ representa la media poblacional por lo tanto es razonable tomar $T = \bar{X}$ como el estadístico de prueba. La elección resulta ser acertada pues, si suponemos que la hipótesis nula es cierta, sabemos que $T = \bar{X} \sim N(0, 1/\sqrt{n})$, por tanto, de ser cierta H_0, debería esperarse que el valor observado de T estuviera lo suficientemente cercano a 0, o de lo contrario, habría razones para concluir que en realidad H_0 es falsa.

- **Región de Rechazo.(C)** Región del espacio muestral, *i.e.* el conjunto de valores que puede tomar el estadístico de prueba T , para los cuales se rechaza la hipótesis nula. En otras palabras, H_0 será rechazada si y sólo si, ocurre el evento $\{T \in C\}$. La forma de C puede depender, entre otras cosas, del tamaño muestral, de la distribución de T y del tipo de prueba de hipótesis que se está llevando a cabo.

- **Potencia.** Es la medida de la capacidad de una prueba particular de rechazar correctamente la hipótesis nula. Es decir, la probabilidad de no cometer error de tipo II. En términos del Ejemplo 2.1.1, se expresaría como $\beta = P(T = \bar{X} \in C|\theta \neq 0)$. Frecuentemente la potencia de una prueba se suele describir en términos de la llama *función potencia* que se define como la probabilidad de rechazar la hipótesis nula en función del valor verdadero del parámetro:

$$\beta(\theta^*) = P(T \in C|\theta = \theta^*). \tag{2.4}$$

  Visto de esta manera, esperaríamos que, para que una prueba sea de buena calidad experimental, $\beta(\theta)$ sea lo más pequeña posible para los valores $\theta \in \Theta_0$ y cercana a uno para los valores $\theta \in \Theta_1$.

- **Nivel de Significancia (α).** Es el mayor valor posible para la probabilidad de error de tipo I que el investigador está dispuesto a tolerar. En términos de la función de potencia definida en 2.4 se define mediante la siguiente expresión:

$$\alpha \leq \sup_{\theta \in \Theta_0} \beta(\theta). \tag{2.5}$$

  Naturalmente se espera que α sea lo más pequeño posible, pues constituye una cota superior para el error de tipo I. Sin embargo, valores demasiado cercanos a cero podrían ser inconvenientes debido a que aumentarían la probabilidad de error tipo II a niveles no permisibles. En el Ejemplo 2.5 es fácil demostrar que si fijamos $C = (-\infty, -\frac{1}{\sqrt{n}} Z_{\alpha/2}) \cup (\frac{1}{\sqrt{n}} Z_{\alpha/2}, \infty)$, donde $Z_{\alpha/2}$ es un cuantil adecuado de la distribución normal estándar, obtendremos una prueba de nivel α. Los detalles pueden consultarse en Casella and Berger [2008].

- **p-valor.** Cuando en un determinado problema de prueba de hipótesis la hipótesis nula resulta ser rechazada, es frecuente preguntarse si se hizo mediante un rechazo contundente (fuerte evidencia en contra) o si no lo fue, dicha situación hace referencia a la necesidad de un instrumento que permita medir la intensidad de la evidencia en contra de H_0 presente en la información muestral. El concepto de p-valor surge en respuesta a esta cuestion.

  El p-valor es la probabilidad, asumiendo la hipótesis nula como cierta, de haber observado un valor del estadístico de prueba al menos tan *extremo* como el que se observó. Como tal supone ser la medida fundamental de cuantificación de la evidencia en contra de H_0. Naturalmente el p-valor depende de la distribución del estadístico T bajo el supuesto de que H_0 es cierta, como se exhibe en la siguiente expresión:

$$p = P(T > T_{obs} | \theta \in \Theta_0), \tag{2.6}$$

  donde T_{obs} corresponde al valor específico de T observado en el experimento particular. Mientras más pequeño sea el p-valor, menos razonable resulta haber observado un valor como T_{obs} bajo la hipótesis nula y más fuerte será la evidencia en contra de ella. Es por ésto que frecuentemente en la práctica suele establecerse un umbral, a partir del cual se puede concluir que existe suficiente evidencia en contra de H_0 para rechazarla. Visto de esta manera, es posible redefinir el concepto de región de rechazo en términos del p-valor. Decimos, por consiguiente, que la hipótesis nula de una prueba será rechazada si y solo si su p-valor p es lo suficientemente pequeño, es decir, si $p < \alpha^*$ para algún cierto valor de α^* pre-especificado. Es posible demostrar que α^* tiene una interpretación equivalente al nivel de significancia de la prueba definido anteriormente. Con este paradigma en mente, la definición de un valor conveniente para α^* y su correspondiente interpretación dependen enteramente del contexto. Frecuentemente en la práctica, la pre-especificación de un valor α^* se omite y el resultado de una prueba de hipótesis se reporta simplemente mediante un p-valor, entendiéndose como la fuerza de la evidencia en contra de la hipótesis nula, y con base en ello, tomar una decisión acerca de si la hipótesis ha de ser rechazada o no dependiendo del contexto.

Resumiendo, los pasos para realizar una prueba de hipótesis simple son los siguientes:

1. Plantear la hipótesis nula y la hipótesis alternativa.
2. Seleccionar un nivel de significancia α. El umbral probabilístico bajo el cual la hipótesis será rechazada.
3. Realizar el proceso de muestreo.
4. Elegir la estadística de prueba adecuada T.

5. Encontrar la distribución de T bajo la hipótesis nula.

6. Calcular la región crítica o región de rechazo C. La región del espacio muestral en la cual la hipótesis será rechazada.

7. Encontrar el valor observado del estadístico de prueba T_{obs} de la muestra.

8. Decidir si rechazar o no H_0 en la base a la C especificada en el paso 6.

En términos del p-valor, se puede obtener un proceso alternativo del proceso anterior reemplazando los pasos 6, 7 y 8 por:

6. Encontrar el valor observado del estadístico de prueba T_{obs} de la muestra.

7. Calcular el p-valor asociado como la probabilidad, bajo la hipótesis nula, de observar un estadístico de prueba al menos tan extremo como T_{obs}.

8. Rechazar H_0 si el p-valor obtenido es lo suficientemente pequeño de acuerdo con el nivel de significancia pre-especificado.

Una alternativa más general a las pruebas de hipótesis simples presentadas en esta sección consiste en la prueba simultánea de un número $m > 1$ de hipótesis simples en el mismo marco de referencia. Dicho procedimiento es conocido como prueba de hipótesis múltiple y se reseñará a detalle en la siguiente sección..

2.1.2. Panorama General de una Prueba de Hipótesis Múltiple

En la sección anterior se han desarrollado los fundamentos principales del escenario más simple en el que puede surgir prueba de hipótesis. Éste es aquel en el cual existe una única conjetura, llamada en este contexto hipótesis nula, que el investigador está dispuesto a someter a prueba con base en evidencia procedente de un proceso de muestreo aleatorio. Se cambiará a continuación a un contexto generalizado, en el cual se tiene una colección de $m > 1$ hipótesis $H_{01}, H_{02}, ..., H_{0m}$. El interés en esta ocasión radica en realizar una prueba estadística simultánea para determinar cuántas y cuáles de ellas han de ser rechazadas a un cierto nivel de confianza estadística tal y como se describió de manera individual en la sección anterior. Dicho procedimiento recibe el nombre de *prueba de hipótesis múltiple* o *problema de comparaciones múltiples*. El resto de la presente sección se dedicará a establecer los fundamentos teóricos de dicho procedimiento, estableciendo analogías y contraste con el caso simple. Se finalizará motivando el fuerte vínculo que existe entre esta metodología y el problema de selección.

En términos generales, el algoritmo estándar para resolver un problema de prueba de hipótesis múltiple consiste de una extensión natural del caso de una prueba de hipótesis simple como se discutió en la Sección 2.1.1. Un algoritmo genérico consiste de dos pasos fundamentales (Dudoit et al. [2003]):

Para $j = 1, 2, ..., m$, donde m es el número de hipótesis en prueba:

1. Elegir y calcular un estadístico de prueba T_j para cada hipótesis individual j. (Véase Sección 2.1.1).

2. Aplicar un procedimiento de prueba de hipótesis múltiple para determinar cuáles hipótesis se han de rechazar de manera que se controle de alguna forma específica el error tipo I.

El Problema 1 está dentro del contexto de pruebas de hipótesis simples. Como se discutió a detalle en la Sección 2.1.1, la elección del estadístico de prueba adecuado T_j dependerá enteramente del contexto experimental y del tipo de hipótesis de trabajo que se tenga. Por ejemplo, para pruebas relacionadas con

la media, un estadístico Z o un estadístico t podrían ser adecuados, mientras que, para pruebas relacionadas con observaciones pareadas o de supervivencia se podría requerir estadísticos más sofisticados. No se discutirán aquí detalles relacionados con la elección de dicho estadístico; sólamente se establecerá que T_j es una función determinada de los datos experimentales y que la hipótesis correspondiente H_{0j} será rechazada con base en su valor observado t_j. El problema 2 es la idea central de esta reseña teórica y ha sido, a lo largo de los años, uno de los objetivos primarios en el estudio de las pruebas de hipótesis múltiples. En la mayor parte de las secciones restantes se repasarán las nociones básicas de la transición del caso de pruebas de hipótesis simples al caso múltiple. Se estudiarán a su vez muchas de las principales nociones relacionadas con error tipo I y las formas de controlarlo.

2.1.3. Extensión del caso simple

Existen diversas consideraciones relevantes que deben tomarse en cuenta al extender la metodología para pruebas de hipótesis simples al caso múltiple. La primera es que no en todos los casos es equivalente realizar una prueba simultánea para la colección de hipótesis $H_{01}, H_{02}, ..., H_{0m}$ que realizar m pruebas individuales para cada una de las hipótesis simples. La razón yace en que no existe un supuesto de independencia entre las hipótesis de la colección, de tal manera que, es posible que puedan existir al menos un par de índices i y j tales que el rechazo de H_{0i} podría influir (positiva o negativamente) en las posibilidades del rechazo de H_{0j}. La falta de independencia es, de hecho, un escenario frecuente en la práctica, por ejemplo en problemas relacionados con genética y finanzas donde existen conjuntos masivos de datos altamente correlacionados. Muchos de los procedimientos clásicos dentro de la metodología de comparaciones múltiples requieren el supuesto de independencia entre las hipótesis; sin embargo, existen diversas modificaciones que prueban ser bastante robustas para ciertas instancias de dependencia. En cualquiera de los casos, realizar una prueba múltiple con m hipótesis resulta ser un procedimiento más eficiente y adecuado tanto conceptual como computacionalmente que realizar m pruebas simples, en especial en los casos en los que m es considerablemente grande como ocurre en la mayoría de sus aplicaciones potenciales.

Otro aspecto importante a considerar, y quizás el más importante, es que, al extender el escenario de las pruebas de hipótesis simples al caso múltiple se incorpora el efecto de la multiplicidad al análisis. Es necesario un procedimiento agregado, conocido formalmente como *compensación por multiplicidad*, que busca evitar conclusiones sesgadas basadas en situaciones que ocurren por efectos del azar, como se ilustra en el siguiente ejemplo:

Ejemplo 2.1.2 (Lanzamiento de monedas)**:**
Supóngase que un experimentador desea probar estadísticamente si una moneda determinada es justa. Para ello realiza 10 lanzamientos, de los cuales 9 resultan en cara. Si asumimos como cierta la hipótesis de que la moneda es justa entonces la probabilidad de que se observe un resultado al menos tan extremo como ese sería de $(10+1)(\frac{1}{2})^{10} = 0{,}0107$, con lo que podemos concluir que no es razonable asumir que la moneda es justa con base en la información obtenida, (véase concepto de p-valor en la Sección 2.1).

Si el experimentador deseara repetir la prueba anterior, pero esta ocasión deseara probar a 100 monedas diferentes, se enfrentaría a una prueba de hipótesis múltiple. Dado que la probabilidad de que una moneda justa caiga al menos 9 veces cara cuando se lanza 10 veces es de 0,0107, el experimentador esperaría que observar un resultado como éste al lanzar 100 monedas justas fuera un evento igual de raro; sin embargo, lo cierto es que observar al menos una de las 100 monedas comportarse de esa manera es un evento muy probable, incluso en el caso en que todas sean justas. En efecto, la probabilidad de que en 100 experimentos con monedas justas, al menos una muestre 9 o más caras en 10 lanzamientos es $1-(1-0{,}0107)^{100} = 0{,}6604$, por lo que, aplicar el el criterio anterior para probar la hipótesis de que las 100 monedas son justas constituiría un error importante.

El Ejemplo 2.1.2 exhibe uno de los aspectos más importantes de la teoría de las pruebas de hipótesis

múltiples, y es la necesidad de nuevas formas de cuantificar el error, que en este caso no sólo depende de los errores individuales dentro del contexto de cada hipótesis particular, sino también de la multiplicidad del problema mismo. Como se discutirá a detalle en secciones posteriores, conforme el número de hipótesis incrementa, la noción de error se complica de manera creciente. Por ejemplo, si una prueba simple se hace a un 5 % de confianza, afirmamos que existe un 95 % de probabilidad de que la hipótesis nula sea rechazada incorrectamente. Sin embargo, si se realizan $m = 100$ pruebas de hipótesis simultáneamente, donde todas son ciertas, el número esperado de rechazos incorrectos es 5, mientras que, si las pruebas son independientes, la probabilidad de rechazar al menos una hipótesis incorrectamente es de $1-(0{,}05)^{100} = 0{,}994$. No es difícil ver que, conforme m, el número de hipótesis en prueba, se hace grande, dicha probabilidad se acerca a uno sin importar el nivel de significancia en consideración. En este contexto, el error de rechazar una hipótesis nula que es cierta se conoce comúnmente como *falso positivo* o *error de tipo I* como en el caso de las pruebas de hipótesis simples. Existen en la literatura distintas técnicas para controlar el número de falsos positivos asociados con una prueba de hipótesis múltiple; se pretende ofrecer un panorama general de las técnicas más relevantes en las secciones siguientes, un resumen detallado puede consultarse en Dudoit et al. [2003] y Farcomeni [2008].

2.1.4. Nociones de Error

En la Sección 2.1.1 se introdujo la noción de error en una prueba de hipótesis simple H_0. Se describió el procedimiento general para una prueba de hipótesis que, a grandes rasgos, consiste en la definición y cálculo de un estadística prueba adecuada T y, con base a su valor observado y su distribución de probabilidad bajo la hipótesis nula, cuantificar la evidencia muestral en contra de H_0 y, de ser suficiente, poder rechazarla. Se identificaron entonces dos posibilidades de cometer un error con dicho procedimiento: rechazar H_0 cuando ésta es cierta (error de tipo I) y no rechazar H_0 cuando ésta es falsa (error tipo II).

Cuando extendemos al caso múltiple, reemplazamos la única hipótesis de trabajo H_0, por una colección de hipótesis H_{0j} para $j = 1, 2, ..., m$, por lo que, el concepto de error se vuelve naturalmente más complejo. Bajo este panorama, el interés se generaliza de la probabilidad de rechazar incorrectamente cada hipótesis partícular al número de hipótesis rechazadas incorrectamente que denotaremos por R. Siguiendo la notación de [Benjamini and Hochberg, 1995], el caso en que se prueban m hipótesis puede resumirse en el Cuadro 2.2.

	Hipótesis No Rechazadas	**Hipótesis Rechazas**	**Total**
Hipótesis Verdaderas	U	V	m_0
Hipótesis Falsas	K	S	m_1
	$m - R$	R	M

Cuadro 2.2: Clasificación cruzada decisión x realidad de las hipótesis.

Como se puede ver, el conjunto de m hipótesis de prueba se divide en dos subconjuntos: las hipótesis verdaderas y las hipótesis falsas, de cardinalidades m_0 y $m_1 = 1 - m_0$ respectivamente. Las variables aleatorias U, K, V y S son no observables a diferencia de la variable aleatoria observable R que es de especial importancia. La variable aleatoria V constituye el número de *falsos positivos* o errores de tipo I y la variable aleatoria K el número de *falsos negativos* o errores de tipo II. Idealmente el investigador estará interesado en minimizar V y K pero, al igual como sucede en el caso univariado, realizar esto simultáneamente es imposible. Por tanto, todo procedimiento estándar de prueba de hipótesis múltiple tendrá como prioridad controlar V o una función de V a algún nivel específico de confianza α. La cantidad en función de V que es de interés controlar recibe el nombre de *tasa de error* y existe en la literatura en varias formas que ofrecen distintos grados de control a distintos grados de complejidad.

Esta situación es análoga al caso univariado cuando se decide controlar el error de tipo I restringiéndolo a estar por debajo de un nivel pre-especificado α. En el caso multivariado, sin embargo, la elección de la tasa de error particular que conviene controlar no es única y no existe una forma óptima de control que garantice un procedimiento de prueba óptimo que se pueda generalizar a cualquier situación.

A continuación se describen tres de las formas más comunes en la literatura para la tasa de error Dudoit et al. [2003]:

- **Tasa de Error por Comparación (PCER)**. Consiste de el valor esperado de errores de tipo I dividido entre el número total de hipótesis:

$$\text{PCER} = \text{E}(V)/m. \tag{2.7}$$

  La idea detrás del PCER yace en la intención de *heredar*, al menos en cierto sentido, el nivel de significancia α de las pruebas individuales. Para ver esto, supongamos, por ejemplo, que todas las hipótesis son ciertas y que se prueban individualmente a un nivel de significancia común α. Luego, V es una variable aleatoria cuya distribución es binomial con probabilidad de éxito dada por la probabilidad de rechazar una hipótesis cierta, que es precisamente α. Por tanto, PCER $= \text{E}(V)/m = m\alpha/m = \alpha$. En general, si m hipótesis son probadas a un nivel α de significancia, entonces el PCER será siempre α. En otras palabras, PCER tiende a ignorar el efecto de la multiplicidad y por ello, por lo general no es muy recomendable.

- **Tasa de Error Global (FWER)**. Es la probabilidad de cometer uno o más errores de tipo I:

$$\text{FWER} = P(V \geq 1), \tag{2.8}$$

  o equivalentemente,

$$\text{FWER} = P(V > 0). \tag{2.9}$$

  Hochberg and Tamhane [1987] define el término *familia* como toda colección de inferencias estadísticas para las cuales hace sentido tomar una forma de error combinado o global. La FWER recibe su nombre de una idea similar en la cual es necesario resumir el error global de las pruebas que intervienen en una PHM mediante una cantidad así denominada.

- **Tasa de Falsos Descubrimientos (FDR)**. Propuesto por primera vez en Benjamini and Hochberg [1995], consiste de la proporción esperada de errores entre las hipótesis rechazadas. Formalmente, si definimos la variable aleatoria Q como:

$$Q = \begin{cases} V/R & R > 0 \\ 0 & R = 0 \end{cases}$$

  se tiene que:

$$FDR = \text{E}(Q) = \text{E}(\frac{V}{R})P(R > 0). \tag{2.10}$$

  La propuesta de Benjamini and Hochberg [1995] se colocó como una de las piezas clave en la investigación relacionada con FDR. Lo anterior debido a que, hasta antes de dicha publicación, la mayor parte de la inferencia relacionada con PHM se hacía fundamentalmente con base en métodos relacionados con el control de FWER, o bien técnicas similares derivadas de modificaciones a la misma. FWER posee desventajas importantes que con el surgimiento de conjuntos de datos de mayor tamaño, por ejemplo en el contexto de genética, se hicieron más evidentes (Dudoit et al. [2003]). FDR se propone como una alternativa más potente a casos en los cuales FWER se muestra muy conservativa como se explicará en las próximas secciones.

Por lo general, se busca un procedimiento que sea capaz de dictar un criterio para decidir cuáles hipótesis serán rechazadas mientras se mantiene algún tipo particular de tasa de error debajo de un cierto nivel de significancia α. Si ésto pasa decimos que el procedimiento controla dicha tasa de error a un nivel de significancia α. Por ejemplo, decimos que un procedimiento controla el FDR si FDR$\leq \alpha$ y similarmente para las otras tasas de error.

Naturalmente, la elección de la tasa de error apropiada para controlar no es arbitraria. De manera análoga al caso univariado, lo que se busca es un procedimiento que controle el error de tipo I simultáneamente mantiendo a un nivel aceptablemente bajo la tasa de errores de tipo II. En general, es posible demostrar (véase Benjamini and Hochberg [1995] y Hochberg and Tamhane [1987]) que se cumple la siguiente relación entre las tres tasas de error descritas en esta sección:

$$\text{PCER} \leq \text{FWER} \leq \text{FDR}. \tag{2.11}$$

De esta manera, si mediante un procedimiento determinado controlamos a PCER a un nivel de significancia α, es decir, dicho procedimiento garantiza que PCER $\leq \alpha$, entonces habremos garantizado también el control de FWER y FDR al mismo nivel de significancia. Similarmente, todo procedimiento que controle FWER controlará FDR. Decirmos por tanto que los procedimientos para el control específico del PCER resultan ser más *conservativos*, en el sentido de que rechazan en promedio un menor número de hipótesis, que los procedimientos que controlan a FWER y FDR y así sucesivamente. Esta razón, sumada al hecho de que la mayoría de los procedimientos que controlan el PCER tienden a ignorar el efecto de la multiplicidad (Dudoit et al. [2003]), es por lo que en lo que resta de este capítulo nos enfocaremos principalmente en procedimientos que controlen la FWER y la FDR.

2.1.5. Ajuste de p-valores

En la Sección 2.1.1 se introdujo la noción de p-valor para el caso de las pruebas de hipótesis simples. Se determinó que el p-valor p es la probabilidad de observar una estadística de prueba al menos tan extrema como la que se observó si suponemos la hipótesis nula H_0 como cierta. De esta manera p corresponde a una medida de la fuerza de la evidencia de los datos en contra de H_0. En términos de las nociones introducidas en esta sección, rechazar H_0 si $p < \alpha$ para $\alpha \in [0, 1]$ provee control del error tipo I de la prueba, es decir, la probabilidad del error tipo I asociado a esta prueba estará garantizada a estar debajo del valor de α.

La noción anterior de p-valor puede ser extendida al caso múltiple. Supongamos que se tiene una colección de hipótesis en prueba H_{0j} para $j = 1, 2, ..., m$, y sean t_j y $p_j = P(T_j > t_j | H_{0j}$ es cierta) el valor observado de la estadística de prueba elegida T_j y el p-valor correspondiente obtenido de la $j-$ésima prueba individual. Decimos que H_{0j} será rechazada a un nivel de significancia α si y sólo si $\{t_j \in C_j\}$ o bien si $p_j < \alpha_j$ donde la región de rechazo C_j y el punto de corte α_j se escogen de tal manera que se controle alguna tasa de error (por ejemplo FWER) al nivel pre-especificado α. Nótese que los puntos de corte α_j no necesariamente son iguales y por consiguiente no necesariamente son iguales a α, es decir, visto de esta manera, la regla de decisión o de rechazo para la hipótesis H_{0j} depende no sólo de α si no también de α_j. Un p-valor ajustado, denotado por $\tilde{p}$, es una transformación del p-valor original $\tilde{p}_j(p_j, \alpha_j)$ que incorpora los efectos de multiplicidad de la prueba y el control de la tasa de error elegida de manera que podamos re-escribir la regla de rechazo en términos únicamente del valor nominal de α. Por ejemplo, si se desea realizar una prueba para controlar la FWER a un nivel de significancia α, diremos que H_{0j} será rechazada a nivel FWER α si $\tilde{p} < \alpha$. Nótese cómo esta última manera de escribir el criterio de rechazo es completamente libre de cualquier parámetro excepto α lo cual resulta muy conveniente.

Muchos de los procedimientos estándar de control de FWER y FDR ofrecen expresiones cerradas para los p-valores ajustados correspondientes y en la mayoría de los casos resulta ser computacionalmente más económico expresar el procedimiento completo únicamente en términos de ellos. En la práctica existen

casos en los cuales los p-valores ajustados no pueden calcularse de manera explícita, principalmente aquellos en los cuales resulta difícil o imposible conocer la distribución de las estadística de prueba T_j. Para ello, se puede hacer uso de métodos de remuestreo para estimar $\tilde{p}_j$ de una manera relativamente simple (para detalles técnicos se puede consultar Westfall and Young [1993]).

2.1.6. Control de la FWER

Retomando la notación del Cuadro 2.2 y la ecuación (2.8), definimos en la Sección 2.1.4 a la Tasa Global de Errores (FWER) como la probabilidad de cometer al menos un error de tipo I. Como se explicó en la Sección 2.1.1, la multiplicidad en una PHM se vuelve un problema cuando el número de hipótesis aumenta, ya que ésto causa que también aumente la probabilidad de que un evento raro ocurra y por tanto la probabilidad de cometer un error de tipo I podría incrementarse. Ésto hace surgir la necesidad de procedimientos adecuados para el control de situaciones con multiplicidad. Esta sección estará dedicada a una reseña de aquellos procedimientos específicos para el control de la FWER, reseñas técnicas más extensas y detalladas pueden ser consultadas en Dudoit et al. [2003], Farcomeni [2008] y Hochberg and Tamhane [1987].

El procedimiento más simple y quizás el más conocido de control para la FWER a un nivel α es el procedimiento o correción de Bonferroni. Este algoritmo recibe su nombre en honor al matemático italiano Carlo Emilio Bonferroni en referencia a las desigualdades propuestas en Bonferroni [1936] que son una consecuencia de la desigualdad de Boole utilizada en la demostración. Este procedimiento consiste en rechazar toda hipótesis H_{0j} si

$$p_j < \alpha/m. \tag{2.12}$$

Como consecuencia, los p-valores ajustados $\tilde{p}_j$ para este caso estarán dados por

$$\tilde{p}_j = \min(1, mp_j), \tag{2.13}$$

y de esta manera, equivalentemente, si se rechaza H_{0j} cuando $\tilde{p}_j < \alpha$ se habrá obtenido un procedimiento tal que FWER$< \alpha$ en términos únicamente de α. Para ver que el procedimiento de Bonferroni, descrito anteriormente, en realidad ofrece control sobre la FWER, sea I_0 el conjunto de índices correspondiente a las m_0 hipótesis nulas que en realidad son ciertas (véase el Cuadro 2.2). Entonces se tiene que

$$\text{FWER} = P\{\cup_{I_0}(p_i < \frac{\alpha}{m})\} \leq \sum_{I_0} P\{p_i < \frac{\alpha}{m}\} = m_0\frac{\alpha}{m} \leq \alpha, \tag{2.14}$$

donde el paso clave de la demostración se debe a la desigualdad de Boole $P(\cup E_i) \leq \sum P(E_i)$.

Aunque el procedimiento de Bonferroni es relativamente simple, resulta ser considerablemente conservador. Esto se debe principalmente al hecho de que, conforme aumenta m, aumentan las posibilidades de cometer al menos un rechazo incorrecto, lo cual aumenta rápidamente la FWER. Como consecuencia, exigir control a un nivel tan pequeño como α— especialmente cuando m es considerablemente grande— viene con el costo de cometer un gran número de errores de tipo II, es decir, no rechazar hipótesis que en realidad son falsas. Además, si m es suficientemente grande, la cota α/m se vuelve muy pequeña, de manera que, sólo aquellas pruebas con evidencia en contra excepcionalmente evidente serían detectadas. Otra forma de ver esta misma aseveración es a través del p-valor ajustado, es fácil ver de (2.13) que, si m crece, $\tilde{p}_j$ se acerca a 1 lo cual producirá pocos o ningún rechazo, sin importar el valor de α, incrementando así las posibilidades de cometer errores de tipo II a niveles incómodamente grandes.

Una alternativa conocida al procedimiento de Bonferroni es el llamado procedimiento Šidák, propuesto en Šidák [1967]. Este procedimiento presenta una mejora a la cota de Bonferroni (2.12) proponiendo rechazar H_{0j} si

$$p_j < 1 - (1 - \alpha)^{1/m}. \tag{2.15}$$

No es difícil demostrar (véase Šidák [1967]) que, si los p-valores son variables aleatorias independientes, el procedimiento Šidák así descrito controla la FWER, y que además, el p-valor ajustado correspondiente está dado por

$$\tilde{p}_i = 1 - (1 - p_i)^m. \tag{2.16}$$

Es importante notar los siguientes dos puntos en relación al procedimiento de Šidák.

1. De la ecuación (2.16) se puede ver que, si el número de hipótesis se hace demasiado grande, al igual que en el caso de Bonferroni, el p-valor ajustado para el caso del procedimiento Šidák tiende a ser uno, sin importar el valor de α, por tanto, este procedimiento resulta también ser conservador para los casos en que m es demasiado grande.

2. A diferencia del procedimiento de Bonferroni, el procedimiento de Šidák requiere el supuesto de independencia de los p-valores para garantizar el control de la FWER. Por tanto, a pesar de que ofrece la posibilidad de una prueba más potente que la de Bonferroni, no se puede considerar una mejoría generalizada. Lamentablemente, muchas situaciones reales no obedecen dicho supuesto, por ejemplo los conjuntos de datos en genética y finanzas donde por lo general existe alta correlación.

Una alternativa más al procedimiento de Bonferroni es el procedimiento de Holm, propuesto en Holm [1979]. El método de Holm es una estrategia más elaborada que el procedimiento de Šidák pero, a diferencia de éste, no depende de supuesto alguno de independencia, por lo que representa una mejoría generalizada al procedimiento original de Bonferroni. Sean los p-valores ordenados que se denotan mediante

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(k)}, \tag{2.17}$$

se dice que se rechazará la hipótesis $H_{0(j)}$, la hipótesis correspondiente al p-valor $p_{(j)}$, mediante el método de Holm si

$$p_{(k)} < \frac{\alpha}{m - k + 1} \text{ para } k = 1, 2, ..., j. \tag{2.18}$$

Comparando (2.18) con (2.12) es fácil ver que el procedimiento de Holm representa una mejoría, en términos de potencia, respecto del procedimiento de Bonferroni. Lo anterior debido a que la cota $\alpha/(m - k + 1)$ permite un mayor número de rechazos que α/m mantiendo el mismo nivel de control de la FWER. La idea prinicipal de la demostración de que el procedimiento de Holm controla la FWER puede ser consultada en Efron [2013].

El procedimiento de Holm antes descrito forma parte de un conjunto generalizado de procedimientos de pruebas múltiples conocido como procedimientos multi-etapa (*stepwise procedures*). En un procedimiento de etapa única (*single-step procedure*), cada hipótesis nula es evaluada usando un criterio de rechazo que es independiente del resultado de cualquier otra hipótesis de prueba, tal es el caso en los procedimientos de Bonferroni y de Šidák, donde puede verse de (2.12) y (2.15) que la decisión para el rechazo de cualquier hipótesis específica H_{0j} no depende de ninguna otra. Puede demostrarse que una mejoría, en términos de potencia, preservando el mismo nivel de control de tasa de errores de tipo I puede alcanzarse mediante algoritmos multi-etapa, en los cuales, se permite que el rechazo de una hipótesis en particular dependa de rechazos de otras hipótesis probadas anteriormente. En otra palabras, en un procedimiento multi-etapa, las hipótesis se prueban en forma secuencial y ordenada, por lo general mediante los valores ordenados de los p-valores no ajustados.

Existen dos subclases de procedimientos multi-etapa, los secuenciados hacia arriba (*step-up procedures*) y los secuenciados hacia abajo (*step-down procedures*). En los procedimientos secuenciados hacia abajo las hipótesis se ordenan de manera decreciente considerando primero aquellas cuyas estadísticas de prueba son más significativas, es decir, aquellas cuyos p-valores son más pequeños. El rechazo de cada hipótesis dependerá explícitamente de las hipótesis consideradas anteriormente en la secuencia; en el momento en que una de las hipótesis de la secuencia no es rechazada el procedimiento se detiene y

ninguna otra hipótesis se rechaza. El procedimiento de Holm es un ejemplo particular de este tipo de procedimientos,ya que, como puede verse en (2.18), la regla de decisión para la prueba $H_{0(j)}$ depende de los resultados de todas las hipótesis $H_{0(k)}$ para $k = 1, 2, ..., j$. Existe además una variante multi-etapa del procedimiento de Šidák en la cual se reemplanza la expresión (2.16) con

$$\tilde{p}_j = \max_{k=1,...,j} \{1 - (1 - p_j)^{(m-k+1)}\}. \tag{2.19}$$

Puede demostrarse, de igual manera, que la expresión anterior supone un procedimiento de control más potente que el procedimiento de Šidák tradicional (2.16).

Los procedimientos secuenciados hacia arriba funcionan de manera análoga a su contraparte tomando esta vez las hipótesis ordenadas de manera creciente, respecto de sus p-valores, y evaluando de manera secuencial de la misma manera como se planteó anteriormente. En el momento en el que una hipótesis de la secuencia es rechazada, el proceso se detiene y todas las hipótesis restantes son rechazadas también. Un ejemplo de una propuesta de procedimiento de control de la FWER de este tipo puede encontrarse en Hochberg [1988]. Otros ejemplos importantes de procedimientos multi-etapa para el control de la FWER incluyen los procedimientos secuenciales de min$P$ y max$T$ propuestos en Westfall and Young [1993], entre otros. Los detalles y una revisión bibliográfica comparativa detallada pueden consultarse en Dudoit et al. [2003].

2.1.7. Control de la FDR

La tasa de descubrimientos falsos (FDR) (Benjamini and Hochberg [1995]), como se introdujo en la Sección 2.1.4 es el valor esperado del número de errores de tipo I (V) que se cometen entre las hipótesis rechazadas (R). Es decir, FDR = E(Q) donde $Q = V/R$ si $R > 0$ y $Q = 0$ si $R = 0$ (véase el Cuadro 2.2).

Como consecuencia de (2.11), los procedimientos que garantizan el control de la FDR serán menos conservadores que los procedimientos que controlan la FWER. Esta propiedad resulta ser deseable pues, como se explicó en la Sección 2.1.4 el control de la FWER, en especial para los casos en los que m es suficientemente grande, se vuelve poco práctico pues causa que el número de errores de tipo II se incremente de manera considerable. Benjamini and Hochberg [1995] justifican (2.11), utilizando la idea de que en el caso particular en el cual todas las m hipótesis son ciertas, FWER=FDR. Ésto es fácil de ver pues, en este caso, $Q = V/R$ sólo puede tomar dos posibles valores, $Q = 0$ si $V = 0$ o $Q = 1$ si $V > 0$ ya que todos los rechazos serán equivocados. Es decir, Q representará una variable aleatoria Bernoulli cuya probabilidad de éxito, y valor esperado, será igual a la probabilidad de que ocurra un rechazo equivocado cualquiera $P(V > 0)$, que es precisamente la definición de la FWER. Sin embargo, si algunas de las hipótesis son falsas, entonces $Q = 0$ (si $V = 0$) o bien $0 < Q < 1$ (si $V > 0$), y por tanto FDR= E(Q) $< P(V > 0)$ =FWER. Puede verse además que la desigualdad se vuelve crecientemente más estricta conforme aumenta el número de hipótesis falsas (Dudoit et al. [2003]).

Benjamini and Hochberg [1995] proponen además, el procedimiento estándar más simple de control de la FDR. Dicho procedimiento es de clase secuencial hacia arriba y requiere el supuesto de independencia de las estadísticas de prueba para asegurar el control deseado. Sea $p_{(j)}$ la sucesión de p-valores no ajustados ordenados como en (2.17). El procedimiendo de control de la FDR de Benjamini y Hochberg (B-H) requiere encontrar el índice i más grande para el cual se cumple la condición

$$p_{(i)} \leq \frac{\alpha}{m} i \tag{2.20}$$

y propone rechazar todas las hipótesis ordenadas por p-valor $H_{0(j)}$ para $j = 1, ..., k$. Los detalles de la demostración de que el procedimiento así descrito produce control sobre la FDR se pueden consultar en Benjamini and Hochberg [1995]. El procedimiento B-H resulta ser una mejora considerable respecto de

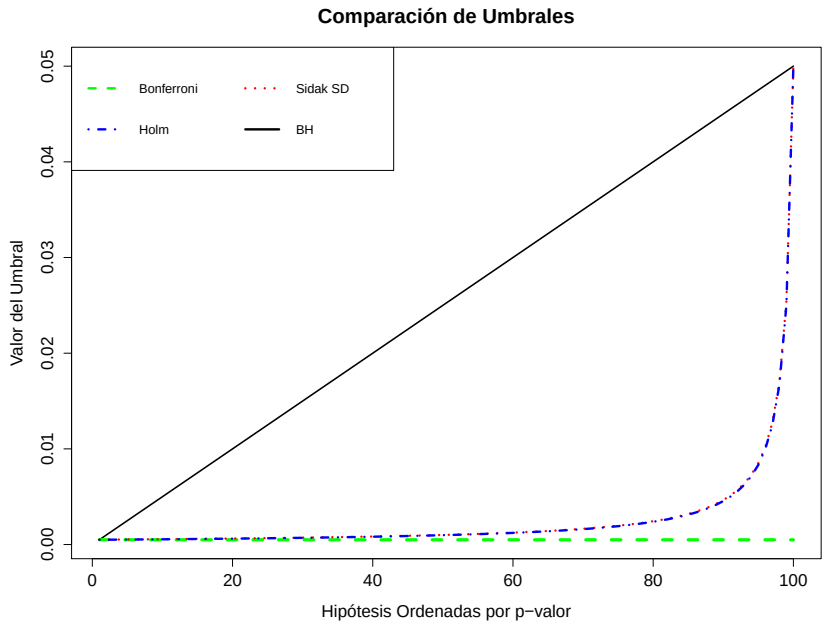

Figura 2.1: Distintas fronteras de rechazo para los procedimientos de control discutidos en la Sección 2.1.4

cualquier procedimiento de control de la FWER en términos de potencia. En la Figura 2.1 se compara la frontera de rechazo, en términos de los p-valores no ajustados, del procedimiento B-H (2.20) con las fronteras análogas de los procedimientos de Bonferroni (2.12), Šidák (SD) (2.16) y Holm (2.18). Aquellos p-valores ordenados que sean graficados por encima de las fronteras marcadas en la figura coresponderán a hipótesis rechazadas de acuerdo con el criterio correspondiente y viceversa. El costo de querer controlar la probabilidad de cometer al menos un error de tipo I se refleja en la notoria diferencia entre los procedimientos que controlan la FWER y la FDR. Es interesante notar también que, a pesar de tener expresiones muy distintas, los procedimientos de Holm y Šidák son muy parecidos. Sin embargo, conceptualmente Šidák viene con un costo más alto debido al supuesto implícito de independencia que requiere. En cualquiera de las instancias se puede concluir que ninguno de los procedimientos que controlan la FWER es capaz de ofrecer la cobertura contra errores de tipo II que el procedimiento de Benjamini and Hochberg [1995] aporta.

Cabe aclarar que una de las principales desventajas potenciales del algoritmo B-H tal y como se propuso en Benjamini and Hochberg [1995] yace en el supuesto de independencia entre las m hipótesis requerido para asegurar el control de la FDR. En respuesta a esa situación Benjamini and Yekutieli [2001] demuestran que el procedimiento B-H también puede ofrecer control de la FDR para los casos en los cuales las hipótesis están positivamente correlacionadas, como es el caso en una gran variedad de aplicaciones como en Génetica y Ecología. Para los casos en los que las hipótesis están negativamente correlacionadas o presentan una estructura de dependencia más compleja Benjamini and Yekutieli [2001] proponen reemplazar m en (2.20) por:

$$m \sum_{i=1}^{m} \frac{1}{i}, \tag{2.21}$$

lo cual creará una modificación naturalmente más conservadora que el procedimiento B-H original y, por tanto, su implementación se recomienda estrictamente para aquellos casos que lo ameriten.

Otras técnicas interesantes para el control de FDR han sido propuestas. La mayoría resultan en modificaciones o extensiones complementarias del procedimiento propuesto en Benjamini and Hochberg

[1995], por ejemplo Benjamini and Hochberg [2000] y Genovese and Wasserman [2002]. Una reseña breve de estas últimas y algunas otras técnicas interesantes puede encontrarse en Koen et al. [2005].

2.1.8. Procedimientos Aumentados de Control

Como se expuso en la sección anterior, los procedimientos de control de la FDR suponen una ventaja inminente sobre aquellos que controlan la FWER, que por lo general resultan ser conservadores (véase la Figura 2.1). Una pregunta natural emergente en estos casos es si es posible establecer una conexión entre ambas tasas de error, en términos de control, de manera que se pueda obtener un nuevo procedimiento de control más generalizado para una tasa de error global, similar a FWER, mantieniendo alta potencia, como en el caso del procedimiento B-H. van der Laan et al. [2004] resume que dicho procedimiento es posible y que es aplicable para el control de versiones similares generalizadas de la FWER y la FDR ofreciendo ciertas propiedades interesantes.

Se define la *tasa de error global generalizada* (gFWER) como la probabilidad de cometer a lo menos un número $k \geq 0$ de errores de tipo I. En términos de las variables aleatorias introducidas en el Cuadro 2.2, la gFWER se escribe como:

$$\text{gFWER}(k) = P(V > k), \tag{2.22}$$

donde es claro que gFWER(0) =FWER, por lo que, la gFWER es esencialmente una versión menos estricta de la FWER.

En contraste, definimos la *probabilidad en la cola de la proporción de falsos positivos* (TPPFP) como la probabilidad de cometer al menos un $q \times 100\,\%$ de errores de tipo I entre las hipótesis rechazadas para $0 < q < 1$. Formalmente esto se escribe como:

$$\text{TPPFP}(q) = P(V/R > q). \tag{2.23}$$

Los procedimientos basados en el control de la TPPFP son de especial interés en aplicaciones en las cuales se tiene un número grande de hipótesis. Esto se debe a que la proporción de falsos positivos V/R se interpreta de la misma manera sin importar el número de hipótesis en prueba, argumento que no es válido para el caso de la gFWER que se basa en el número absoluto de errores de tipo I. Por ejemplo, no es equivalente controlar gFWER(3) para el caso de $m = 5$ que para el caso de $m = 1000$; la primera situación es muy poco estricta mientras que la segunda es muy conservadora. En contraste, controlar, por ejemplo, $TPPFP(0{,}95)$ tiene la misma interpretación sin importar el valor de m.

La $TPPFP$ guarda una cercana relación con la FDR pues ambas tasas de error están basadas en la misma variable aleatoria $Q = V/R$, la proporción de falsos positivos. van der Laan et al. [2004] sostiene que los procedimientos de control de la $TPPFP$ suponen una mejoría respecto al procedimiento clásico de control de la FDR, propuesto en Benjamini and Hochberg [1995]. Esto debido a que, típicamente, el procedimiendo B-H requiere de condiciones especiales como independencia o, en el peor de los casos, una estructura específica de dependencia (Benjamini and Yekutieli [2001]). Los procedimientos para el control de la gFWER y TPPFP resumidos en van der Laan et al. [2004] toman en consideración la distribución conjunta de las estadísticas de prueba y por consiguiente pueden ofrecen control sin importar los supuestos distribucionales o las estructuras de dependencia que éstas puedan tener.

La idea principal y característica distintiva de los procedimientos de control de la gFWER y TPPFP de van der Laan et al. [2004] es que están *montados* sobre un procedimiento de control (general) de la FWER. En otras palabras, se demostró que todo procedimiento de control de la FWER (de etapa única o de multi-etapa) puede ser directamente *aumentado* agregando un número extra, escogido estrategicamente, de hipótesis rechazadas al conjunto de las que ya fueron rechazadas por el primer procedimiento. De esta manera, se presenta una corrección a la desventaja principal de los procedimientos de control de la FWER que resultan ser conservadores. Los detalles técnicos prinicipales, una reseña detallada de los prinicipales métodos y estudios comparativos de simulación pueden ser consultados en van der Laan et al. [2004].

2.1.9. Principales Inconvenientes de una PHM

Una PHM, planteada adecuadamente, constituye una herramienta poderosa en el contexto del análisis estadístico multivariado y la inferencia estadística en general. Sin embargo, en resumen, hay importantes consideraciones y limitaciones que deben hacerse antes de aplicar este procedimiento:

1. **Puede ser (altamente) sensible al número de hipótesis.** Es intuitivo ver que, conforme el número de hipótesis aumenta, las posibilidades de cometer errores al aplicar algún procedimiento de rechazo son cada vez mayores. Es otras palabras, las tasas de error, como la FWER y la FDR, se incrementan significativamente con m, y por consiguiente un gran número de los procedimientos de control discutidos anteriormente se vuelve incómodamente conservador en estos casos. En particular, los procedimientos de control de la FWER como Holm y Bonferroni pierden efectividad cuando m es demasiado grande. Dudoit et al. [2003] plantea también la influencia de la razón del número de hipótesis ciertas entre el número de hipótesis m_0/m y muestra gráficamente que las tasas de error tipo I están positivamente correlacionadas con esta cantidad.

2. **Depende de la elección de estadísticas de prueba.** Debido a que cada procedimiento de prueba de hipótesis múltiple es directamente dependiente de un conjunto de m sub-procedimientos, en los cuales es necesaria la obtención de los m p-valores no ajustados individuales, es clara la dependencia del resultado final de la PHM de la elección particular de las estadísticas T_j para $j = 1, 2, .., m$. En repetidas ocasiones, incluso, es difícil o imposible conocer la distribución exacta de T_j y/o caracterizar la estructura de dependencia subyacente que presentan como conjunto de variables aleatorias. A pesar de que una gran cantidad de procedimientos se han implementado para corregir o rodear esta limitación (vease por ejemplo Westfall and Young [1993]), la alta dependencia del resultado de una PHM del comportamiento probabilístico de la sucesión T_j representa más una limitante que un reto estadístico *per se*.

3. **Tiene interpretación limitada.** Esta limitación es parte de la idea central y la motivación del Capítulo 3. La razón principal proviene del hecho de que una PHM, al igual que una PHS, se encuentra limitada a responder una problemática de tipo *binario* (cuáles hipótesis rechazar y cuáles no), por lo que utilizar una PHM para responder preguntas más generales debe hacerse con especial cuidado y sólo cuando la situación particular lo amerite.

2.1.10. Conexión con el Problema de Selección

A lo largo de las Secciones 2.1.1 – 2.1.8 se ha explorado el escenario correspondiente a las principales nociones de las pruebas de hipótesis múltiples, sus principales elementos, algunos de los métodos más importantes motivados por su problemática y parte de la bibliografía comentada más relevante. La Sección 2.1.9 introduce una crítica a la metodología de las PHM puntualizando algunas de sus más importantes limitaciones, las cuales serán retomadas en el Capítulo 3 y expandidas a un ensayo crítico acerca del papel de PHM en el contexto particular del problema de selección introducido en el Capítulo 1 y tema principal de la tesis.

La conexión entre las técnicas de PHM y el problema de selección es, en realidad, relativamente incidental. Lo es en el sentido de que el problema de selección, en algunos casos particulares, puede permitirse ser reescrito en términos de *significancia*. Decimos que un individuo en un conjunto es estadísticamente *significativo* si existe evidencia estadística suficiente de que es distinguible (o reconocible) entre los individuos del conjunto del que proviene. Es posible reescribir el enunciado anterior en términos de una prueba de hipótesis como H_{0i} : El individuo i es significativo, para cierto criterio de significancia. Para ilustrar esto se sugiere el siguiente ejemplo:

Ejemplo 2.1.3 (Fármacos)**:**
Sean $X_1, X_2, ..., X_m$ los potenciales componentes activos para la patente de un nuevo fármaco para tratar cierta enfermedad. Sea θ_i una medida del efecto del componente i en un paciente enfermo, visto en una escala tal que valores positivos de θ_i indican mejoría y valores negativos de θ_i indican que empeora, mientras que valores cercanos a 0 indican que el componente no induce efectos detectables. Es de interés para la industria farmacéutica encontrar todos los componentes que tienen un efecto relevante (tanto positivo como negativo) en los pacientes, de manera que se pueda concretar una propuesta para la fórmula de la patente del medicamento de manera que se pueda poner en el mercado.

El científico A se plantea la tarea de encontrar, del conjunto de los m componentes potenciales, aquellos que sean *significativos*, es decir, aquellos que tengan un efecto *distinguible* de los demás. Como sabe que los valores θ_i cercanos a cero indican una falta de efecto del componente decide plantearse una PHM como sigue. Para $i = 1, 2, ..., m$:

$$H_{0i} : \theta_i = 0, \tag{2.24}$$

de manera que, si la hipótesis H_{0i} es rechazada, pueda concluir que hay fuerte evidencia estadística en contra de que $\theta_i = 0$. Es decir, es razonable asumir que el $i-$ésimo componente tiene un efecto distinto de cero, o bien positivo o negativo.

El científico B, en cambio, decide plantear el problema anterior como un problema de selección. Para él, la filosofía del problema consiste en *seleccionar* aquellos componentes que tienen un mayor efecto positivo en los pacientes y, viceversa, aquellos componentes que tienen un mayor efecto negativo.

Aunque aparentemente A y B buscan la misma respuesta, en realidad sus procedimientos difieren no sólo computacionalmente sino conceptualmente de manera considerable. La explicación detallada de este hecho va más allá de los propósitos de esta sección pero puede encontrarse en el Capítulo 3. Por ahora, mediante el Ejemplo 2.1.3, se ha establecido que existe una conexión incidental entre el problema de PHM expuesto en esta sección y el problema de selección motivado en el Capítulo 1.

El Ejemplo 2.1.3 constituye una instancia del problema general denominado *filtrado* o *screening* en el cual el experimentador cuenta con un número m de piezas de información (en este caso los componentes farmacéuticos) y desea hacer una evaluación cualitativa para reducir el conjunto de cardinalidad m (que puede ser grande) a un conjunto más pequeño de n elementos, seleccionados mediante algún algoritmo estadístico. Los problemas de filtrado aparecen frecuentemente en contextos de bases de datos masivas, como los microarreglos en genética, en los cuales puede surgir una gran cantidad de atributos de información de manera simultánea y es de interés, para los investigadores, reducir el problema a una situación más simple, reteniendo sólo las piezas de información relevantes del conjunto. Visto desde otro enfoque, aplicando técnicas de PHM podríamos estar interesados en hacer una selección cualitativa de manera que sea posible encontrar todas las piezas de información que son candidatas a consideración en un estudio más simple. Es en esta idea, precisamente, donde se encuentra la conexión con el problema de selección introducido en el Capítulo 1.

Otro tipo de aplicaciones en las cuales los métodos de PHM y el problema de selección podrían tener objetivos comunes en la práctica es en el contexto de *clasificación*. Supóngase que se tiene un conjunto de m atributos o clases $X_1, X_2, ...X_m$ que pueden ser observados en un individuo particular i que puede pertener a una y sólo una de un conjunto de r clases $C_1, C_2, ..., C_r$. El objetivo principal, en estos casos, consiste en desarrollar un método estadístico que incorpore la información de $X_1, X_2, ...X_m$ para poder clasificar a i en una de las clases del conjunto $C_1, C_2, ..., C_r$ de manera acertada. Existen actualmente una gran gama de clasificadores de distintas naturalezas en la literatura desde muy simples como el clasificador lineal de Fisher hasta muy sofisticados como los algoritmos de máquinas de soporte vectorial. Sin embargo, en especial cuando m es muy grande, estos algoritmos pueden resultar costosos computacionalmente. Reducir el número de atributos clasificadores X_m de manera arbitraria podría resultar en pérdidas importantes de información y conducir a resultados sesgados. Es por ello, que surge la necesidad de identificar atributos candidatos a tener un buen poder clasificatorio de manera

que se pueda hacer una buena inferencia sobre la clase de un individuo cualquiera i sin necesidad de la consideración de un número muy grande de atributos clasificadores. Si etiquetamos como Y_i a la variable aleatoria que guarda el valor de la clase correspondiente al individuo i el problema anterior puede establecerse como la identificación de las variables $X_1, X_2, ..., X_m$ que tienen una correlación significativa con Y_i de manera que puedan tomarse como parte de un clasificador simplificado. De esta manera se puede plantear un problema de prueba de hipótesis múltiple donde la hipótesis H_{0i} representará la afirmación de que el $i-$ésimo individuo está correlacionado con Y_i. Frecuentemente en la práctica se escogen estadísticas de la familia t para resolver este problema. Un caso particular aplicado en genética puede consultarse en Golub et al. [1999] y su solución se discute en Dudoit et al. [2003].

Visto desde un punto de vista alternativo, el problema de clasificación descrito anteriormente propone seleccionar un subconjunto $X_1^*, ...X_t^*$ de los *mejores* t atributos clasificadores del conjunto original X_m, donde *mejor* se entenderá como aquellos que estén correlacionados de manera significativa con la variable clase Y_i, y por tanto, este también constituye una instancia del problema de selección original.

Una pregunta natural en estos casos es si estas dos metodologías son equivalentes, es decir, si aplicadas a un mismo conjunto de datos conducen a conclusiones similares que se interpretan de la misma manera. La respuesta es que no es así. La causa principal proviene de la sutileza conceptual entre la identificación de individuos *significativos* (el propósito de un problema de PHM), y la identificación de los *mejores* individuos (el propósito del problema de selección). Como se detallará en el Capítulo 3 existen situaciones particulares en los cuales los individuos significativos no precisamente son los mejores, y la aplicación de una PHM podría resolver de manera imprecisa un problema cuyo interés cae en la selección de los mejores individuos de un conjunto. Antes de pasar a detallar esta discusión, es necesario repasar las principales metodologías estadísticas diseñadas específicamente para problemas en los cuales el interés no está en la identificación de individuos significativos, sino en los mejores. Las Secciones 2.2 y 2.3 presentan, respectivamente, las metodologías estadísticas clásicas y modernas principales de selección y ordenamiento que se trabajarán en la tesis y pretenden, a cierto nivel de profundidad, establecer la base del debate comparativo que surgirá en el Capítulo 3.

2.2. Selección y Ordenamiento Clásico

Esta sección tiene como propósito reseñar formalmente las principales nociones del problema de selección y ordenamiento planteado en el Capítulo 1. Se repasarán los principales conceptos, resultados y aspectos analíticos de la teoría clásica de RSM así como algunos de los escenarios más recurrentes en la literatura y la bibliografía comentada más relevante.

2.2.1. Conceptos Básicos y Notación

En términos formales, todo problema de selección clásico comienza a partir de la situación en que se tiene un conjunto de n observaciones independientes $X_{ij}, j = 1, ..., n$ de cada una de k distintas poblaciones comparables $\pi_i, i = 1, 2, ..., k$ que siguen una función de distribución $G(x-\theta_i), i = 1, 2, ..., k$, indexada por un conjunto de parámetros de localización desconocidos $\theta_i, i = 1, 2, ...k$. Por ejemplo, π_i puede representar distintas variedades de grano, métodos de enseñanza, tipos de medicamentos, *etc.* En cualquiera de los casos el objetivo yace en seleccionar las t poblaciones que corresponden a las t *mejores* poblaciones, en otras palabras, aquellas con los parámetros más grandes $\theta_{(k-t+1)}, ..., \theta_{(k)}$ donde

$$\theta_{(1)} \leq \theta_{(2)} \leq \cdots \leq \theta_{(k)} \tag{2.25}$$

representa el orden de los parámetros. Dado que dicho orden es desconocido, la idea es obtener información suficiente de los datos para hacer inferencia acerca del ordenamiento se la sucesión $\theta_1, \theta_2, ..., \theta_k$ de manera que se pueda hacer una selección correcta. Intuitivamente, parece razonable y apropiado estimar

Población	1	2	...	k
Parámetro verdadero desconocido	θ_1	θ_2	...	θ_k
Observaciones	X_{11}	X_{21}	...	X_{k1}
	X_{12}	X_{22}	...	X_{k2}
	.	.	...	.
	.	.	...	.
	.	.	...	.
	X_{1n}	X_{2n}	...	X_{kn}
Estadística Resumen	Y_1	Y_2	...	Y_k

Cuadro 2.3: Estructura esquemática del modelo de selección

dicha sucesión mediante estadísticas resumen $Y_1, Y_2, ..., Y_k$ respectivamente. De esta manera obtenemos la versión observada de la expresión (2.25):

$$Y_{[1]} \leq Y_{[2]} \leq \cdots \leq Y_{[k]} \tag{2.26}$$

donde $Y_{[i]}$ representa la $i-$ésima estadística resumen ordenada y asumiremos que sigue una función de distribución $F(y - \theta_i), i = 1, 2, ...k$. La estructura esquemática del modelo se resume en el Cuadro 2.3.

Existen contextos en las cuales no es posible realizar el experimento de manera que el tamaño de muestra tomado de cada población sea igual en todos los casos. Es posible tomar en consideración la situación en que se tienen tamaños de muestra distintos $n_1, n_2, ..., n_k$ y existen en la literatura metodologías para tratar específicamente con estos casos. Por el momento se definirá el problema de esta manera y se asumirá, a menos que se especifique lo contrario, que se tiene el mismo número de observaciones de todos los tratamientos.

Una vez obtenidas las estadísticas resumen Y_i se propone una regla de selección que en este caso es muy intuitiva:

$$R : \text{Seleccionar}\, Y_{[k-t+1]}...Y_{[t]} \,\text{como las mejores}\ t \ \text{poblaciones.} \tag{2.27}$$

Si denotamos por $Y_{(i)}$ la estadística resumen correspondiente a la población con parámetro $\theta_{(i)}$, la regla (2.27) producirá una selección correcta si y sólo si ocurre el evento CS_t donde

$$\text{CS}_t = \{\{Y_{[k-t+1]}, ..., Y_{[k]}\} = \{Y_{(k-t+1)}, ..., Y_{(k)}\} \tag{2.28}$$

Es importante notar la diferencia entre $Y_{(i)}$ y $Y_{[i]}$. La primera corresponde a la estadística de prueba que proviene de la población $\pi_{(i)}$ (la verdadera $i-$ésima población), mientras que la segunda corresponde a la $i-$ésima estadística de acuerdo con el orden observado.

El objeto estadístico de mayor interés en la teoría de RSM es la llamada *probabilidad de selección correcta* (PCS_t), que se define como la probabilidad de identificar correctamente las mejores t poblaciones mediante algúna regla de selección determinada R.

$$\text{PCS}_t = P(\text{CS}_t) \tag{2.29}$$

Por lo general usaremos la regla (2.27) pero existen un número infinito de posibilidades de reglas de selección. Supongamos que se tienen dos procedimientos o reglas diferentes de decisión R_1 y R_2 para seleccionar a las mejores t poblaciones y es de interés compararlas. Un criterio inmediato de comparación sería a través de la PCS. Para ello definimos $P_i = P(\text{Selección correcta usando la regla}\, i), i = 1, 2$, donde claramente P_i depende del verdadero vector de parámetros $\theta = \theta_1, ..., \theta_k$. Claramente, si $P_1 \geq P_2$ sin importar el valor de θ escogeremos la regla 1 como la más óptima; similarmente si $P_2 \geq P_1$ sin importar el valor de θ optaremos por la regla 2. Sin embargo, en la mayoría de los casos, existirán algunas

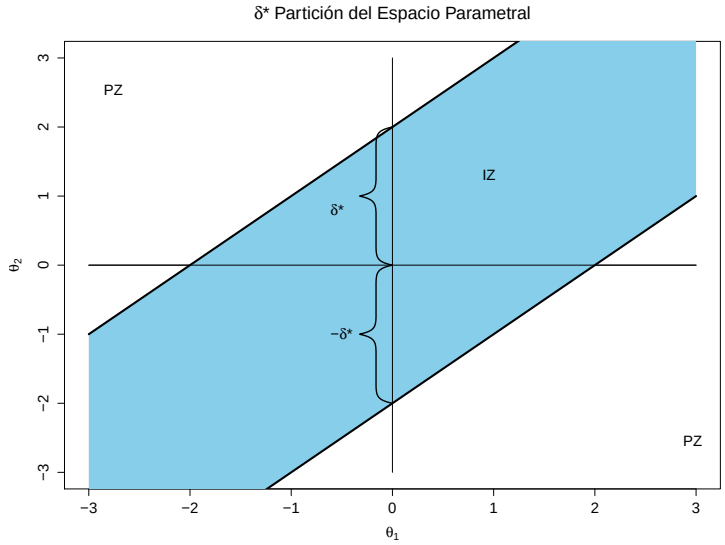

Figura 2.2: Partición del espacio parametral para k=2

configuraciones de θ para las cuales $P_1 \geq P_2$ y algunas otras para las cuales se cumple lo contrario, por lo cual, una comparación a través de la PCS puede hacerse sólo caso a caso para una configuración particular θ_0.

Una manera general de comparar reglas de selección y establecer un criterio para toda buena regla se puede lograr analizando el espacio parametral Θ de manera global, es decir, todas las configuraciones posibles que puede asumir θ, separando aquellas regiones en las que hay una fuerte preferencia por una selección correcta y aquellas en las cuales estamos indiferentes entre dos o más opciones. Estas regiones del espacio parametral reciben el nombre de *zona de preferencia* (PZ) y *zona de indiferencia* (IZ).

La zona de indiferencia fue propuesta por Bechhofer [1954] y se motiva mediante el siguiente razonamiento. Supongamos que estamos interesados en encontrar la mejor población, es decir, fijamos $t = 1$ y buscamos identificar la población correspondiente al parámetro $\theta_{[k]}$. Sin embargo, si la diferencia $\theta_{[k]} - \theta_{[k-1]}$ es muy pequeña podríamos sentirnos *indiferentes* respecto de cuál de las dos poblaciones reportar como la mejor. De esta manera surge la idea de predefinir un valor umbral δ^* que determine cuál es la mínima diferencia que queremos ser capaces de detectar, de modo que, siempre que $\theta_{[k]} - \theta_{[k-1]} \geq \delta^*$ podamos hacer una buena selección. Para el caso particular $t = 1$, aquellos vectores θ cuya configuración satisface $\theta_{[k]} - \theta_{[k-1]} \leq \delta^*$ se dice que pertenecen a la zona de indiferencia, mientras que, en el caso contrario pertenecen a la zona de preferencia.

Gibbons et al. [1977] propone visualizar la zona de indiferencia para el caso $k = 2$ (únicamente dos poblaciones a escoger) como en la Figura 2.2. En ella se grafica el espacio parametral (θ_1, θ_2) y se sombrea la zona de inferencia, i.e. el conjunto de puntos del espacio parametral que cumplen la condición $|\theta_{[k]} - \theta_{[k-1]}| \leq \delta^*$ o, equivalentemente, $-\delta^* < \theta_{[k]} - \theta_{[k-1]} < \delta^*$. En este caso la zona de indiferencia tiene forma de franja delimitada por las rectas $\theta_2 - \theta_1 = \delta^*$ y $\theta_2 - \theta_1 = -\delta^*$. Naturalmente, la recta $\theta_1 = \theta_2$ está contenida dentro de la zona de indiferencia como debería ser siempre el caso para $k = 2$. Es importante notar además que, si el valor de δ es muy pequeño, la zona de indiferencia se reduce y el número de configuraciones posibles en la zona de preferencia aumenta. Como se verá posteriormente esto puede ser inconveniente pues detectar diferencias a un nivel tan alto de precisión podría requerir un tamaño de muestra inaccesible.

Nuevamenta para el caso $k = 2$, si reemplazamos el espacio parametral (θ_1, θ_2) por el espacio parametral ordenado $(\theta_{[1]}, \theta_{[2]})$ obtenemos la Figura 2.3. Naturalmente el espacio parametral se reduce a la mitad pues los puntos para los cuales $\theta_{[1]} > \theta_{[2]}$ no existen. En este caso la zona de indiferencia tiene

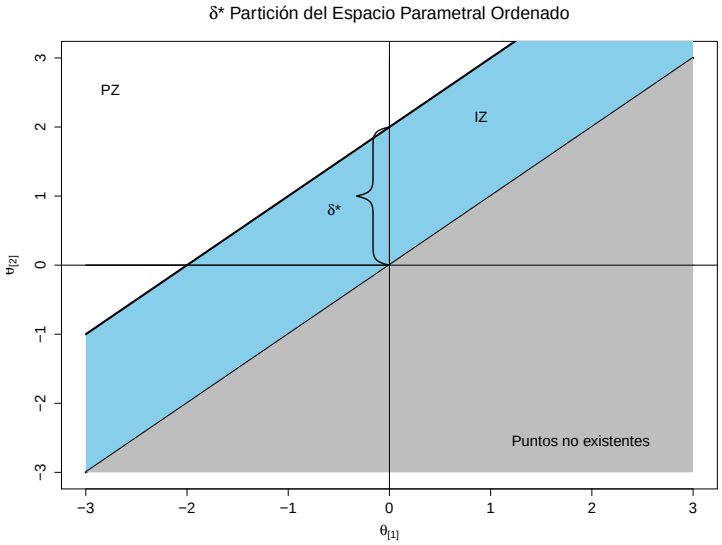

Figura 2.3: Partición del espacio parametral ordenado para k=2

una forma de franja similar a la Figura 2.2 y está delimitada por las rectas $\theta_{[2]} - \theta_{[1]} = \delta^*$ y $\theta_{[1]} = \theta_{[2]}$.

Para el caso de un valor general $k \geq 1$, si $t = 1$, también podemos visualizar la zona de indiferencia en dos dimensiones. Esto es debido a que la zona de indiferencia para este caso se puede definir únicamente en términos de $\theta_{[k]}$ y $\theta_{[k-1]}$ como el conjunto de puntos en el plano $(\theta_{[k]}, \theta_{[k-1]})$ tales que $\theta_{[k]} - \theta_{[k-1]} < \delta^*$. La gráfica para este caso se muestra en la Figura 2.4.

El parámetro δ^* pre-especificado constituye un valor umbral (o máximo permisible) para una determinada medida de distancia δ (véase la Sección 1.1.3) que es una función de los parámetros $\theta_{[k]}$ y $\theta_{[k-1]}$ que establece la noción de distancia entre ellos. Por ejemplo, para el caso de las Figuras 2.3 y 2.4, la función de distancia es la distancia usual, dada por la simple diferencia,

$$\delta = \theta_{[k]} - \theta_{[k-1]}. \tag{2.30}$$

Es importante notar que la forma que toma la zona de indiferencia depende de la función de distancia particular elegida y de los posibles valores que puede tomar θ. En el caso de las Figuras 2.2, 2.3 y 2.4 θ puede tomar cualquier valor en la recta real. Tal es el caso, por ejemplo, cuando representa la media de una distribución normal. Sin embargo, este no es el caso general; el espacio parametral, en algunas situaciones puede estar restringido. Por ejemplo, si θ representa la varianza de una distribución normal, el espacio parametral está restringido a los números reales positivos. Otro caso de particular interés es el de la distribución binomial, donde θ representa la probabilidad de éxito. En este caso, el dominio para θ es el conjunto $[0, 1]$ y, naturalmente, la gráfica de la zona de indiferencia se ve alterada como se ve en la Figura 2.5 para la función de distancia usual $\delta = \theta_{[k]} - \theta_{[k-1]}$.

En el caso binomial es posible considerar una función de distancia alternativa a la usual conocida como la función de *distancia del cociente de momios*,

$$\delta_{OR} = \frac{\theta_{[k]}(1 - \theta_{[k-1]})}{\theta_{[k-1]}(1 - \theta_{[k]})}. \tag{2.31}$$

Esta medida de distancia puede interpretarse como el cociente de los momios de éxito, es decir, cuántas veces es más factible un éxito en la población cuya probabilidad de éxito es $\theta_{[k]}$ comparado con aquella cuya probabilidad de éxito es $\theta_{[k-1]}$. De manera similar a la distancia usual, valores pequeños de δ_{OR} implican indiferencia, y por lo tanto la IZ estará definida por aquellas configuraciones en el espacio

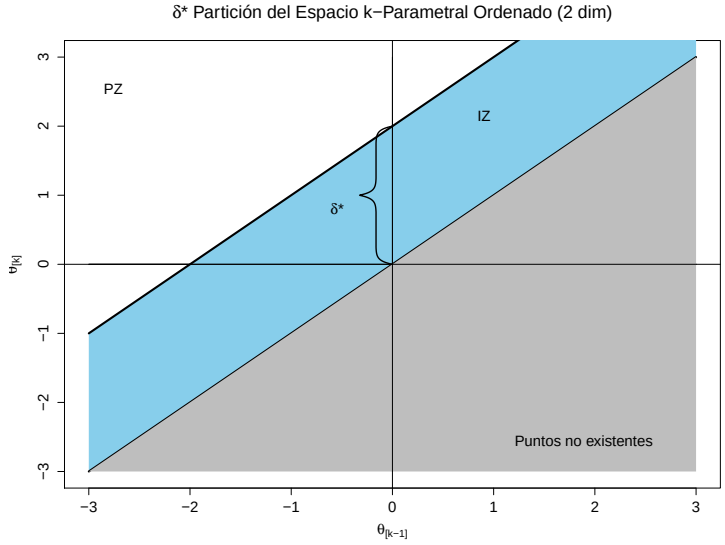

Figura 2.4: Diagrama 2-dimensional de la IZ para el caso de k general

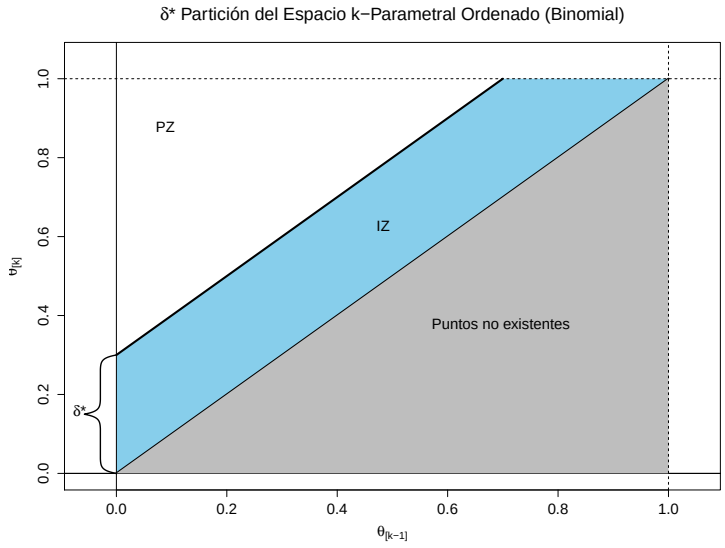

Figura 2.5: Diagrama 2-dimensional de la IZ para el caso binomial y k arbitraria.

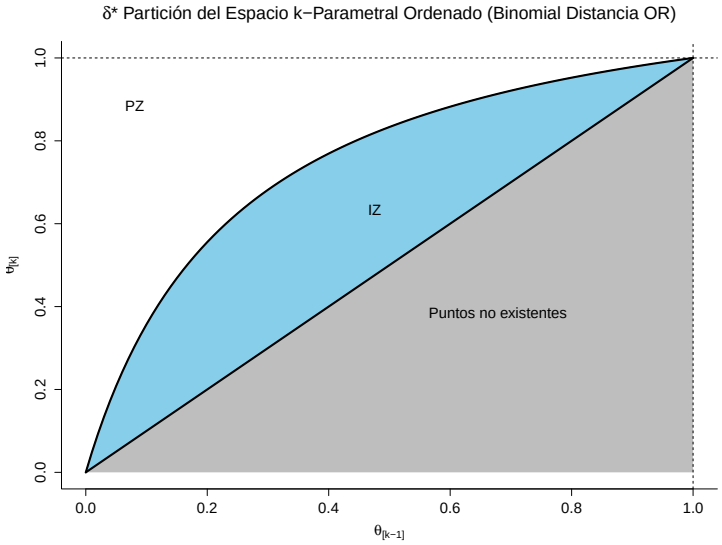

Figura 2.6: Diagrama 2-dimensional de la IZ para el caso binomial y k arbitraria.

parametral tales que $\delta_{OR} < \delta^*$. y la zona de indiferencia por aquellas que cumplen $\delta_{OR} \geq \delta^*$. La gráfica de la zona de indiferencia en 2 dimensiones para este caso se muestra en la Figura 2.6. Es importante notar que en este caso la frontera de la IZ no es una línea recta como en los casos anteriores, sino una curva cerrada que pasa por los puntos $(0, 0)$ y $(1, 1)$.

Las Figuras 2.2 – 2.6 ilustran cómo la función de distancia δ induce una partición del espacio parametral Θ en la cual se identifica explícitamente la región en la cual nos interesa hacer una buena selección o zona de preferencia. La calidad de la selección misma será cuantificada a través de la probabilidad de selección correcta (2.28). Sin embargo, evidentemente dicha probabilidad dependerá de la verdadera configuración θ. Si θ está en la zona de preferencia, nos interesaría que PCS fuera lo más grande posible y, por el contrario, si está en la zona de indiferencia somos indiferentes pues, por ejemplo para el caso $t = 1$, la diferencia entre el mejor tratamiendo y el segundo mejor no es sustancialmente importante para tomar una decisión. Por este motivo, en cuanto al cálculo de PCS concierne, se restringirá la atención únicamente a aquellas configuraciones dentro de la zona de preferencia. Aun así, dado que en la zona de preferencia existen infinitas configuraciones posibles, el problema de encontrar un procedimiento estándar para encontrar la PCS no se simplifica. Sin embargo, es posible encontrar una configuración especial que minimice la PCS, de manera que permita definir una cota inferior generalizada para la PCS bajo cualquier configuración θ. Dicha configuración, de existir, recibe el nombre de *configuración menos favorable* (LFC).

Cuando la LFC existe, podemos únicamente enfocarnos en encontrar PCS para dicha configuración, sin importar el verdadero valor de θ que puede estar en cualquier punto de la zona de preferencia, lo cual representa una gran simplificación del problema original. La probabilidad de selección correcta bajo la LFC se denotará por $P\{\text{CS}|\theta_{\text{LF}}\}$ y cumple que,

$$P\{\text{CS}|\theta_{\text{LF}}\} \leq P\{\text{CS}|\theta\} \tag{2.32}$$

para cualquier vector θ en la zona de preferencia. Bechhofer [1954] encontró la utilidad de la ecuación (2.32) de la siguiente manera. Si nos aseguramos que $P\{\text{CS}|\theta_{\text{LF}}\}$ sea al menos un valor pre-especificado aceptable P^* entonces (2.32) garantizará que $P\{\text{CS}|\theta\} \geq P^*$ independientemente de la verdadera configuración θ. De esta manera, P^* puede especificarse como un parámetro que representa una cota inferior para la PCS que sea de interés calcular. Cabe aclarar que P^* representará una cota conservadora, en el

sentido que se puede escoger de manera que se cumpla que $P\{\text{CS}|\theta\} \geq P^*$ pero bajo la consideración de que el verdadero valor de $P\{\text{CS}|\theta\}$ podría ser muy cercano o muy lejano a P^*.

Por lo general, la LFC puede no ser única y depende enteramente del contexto a través de los parámetros y la función de distancia escogida, por lo cual determinarla no es trivial. Bechhofer [1954] exploró el caso en el que el vector θ representa las medias de k poblaciones normales con la medida de distancia usual $\delta = \theta_{[k]} - \theta_{[k-1]}$. Para este caso θ puede tomar valores en todos los números reales y la zona de indiferencia está dada por todas las configuraciones θ tales que $\theta_{[k]} - \theta_{[k-1]} < \delta^*$ para algún valor $\delta^* > 0$. Bechhofer [1954] encontró que la configuración θ que minimiza $P(\text{CS}|\theta)$ está dada por

$$\theta_{[1]} = \theta_{[2]} = \cdots = \theta_{[k-1]} = \theta^* \text{ y } \theta_{[k]} = \theta^* + \delta^*. \tag{2.33}$$

Notemos que la LFC 2.33 no es única pues θ^* puede tomar cualquier valor en la recta real. Para este caso particular la LFC no depende del tamaño de muestra pero este no siempre es el caso. En la Sección 2.2.3 se mencionarán algunos otros casos particulares interesantes y sus respectivas LFC, entre ellos el caso binomial y el caso de la mínima varianza introducidos en la Sección 1.3.

2.2.2. Aspectos Analíticos

Dos posibles aspectos analíticos independientes pero relacionados que son de interés estudiar en la mayoría de los problemas de selección y ordenamiento se pueden identificar en la mayor parte de la literatura de selección y ordenamiento: la determinación del tamaño de muestra (Bechhofer [1954]) y la estimación de PCS (Olkin et al. [1982]. En esta sección se resumirán las ideas principales de cada uno de ellos. Para ello asumiremos, por simplicidad, que se tiene un problema de selección con $t = 1$ con la medida de distancia usual $\delta = \theta_{[k]} - \theta_{[k-1]}$. Para los casos en que $t > 1$ la explicación es análoga con las respectivas modificaciones.

Determinación del Tamaño de Muestra

Bechhofer [1954] planteó el problema de selección y ordenamiento en primer lugar desde el punto de vista del diseño experimental. Bajo esta idea se visualiza que el objetivo del experimentador es encontrar el tamaño de muestra proveniente de cada población de manera que podamos garantizar que el procedimiento de selección descrito en la Sección 2.2.1 cumpla con requirimientos deseados (alta PCS, ajuste de la IZ, *etc.*).

Formalmente, el problema de Bechhofer [1954] inicialmente asume que se escogerá un tamaño de muestra n común para todos los tratamientos de manera que las estadísticas resumen $Y_j, j = 1, 2, ..., k$ estén basadas en el mismo número de unidades experimentales. La idea será entonces encontrar el mínimo valor de n tal que el procedimiento de selección elegido satisfaga cierto requerimiento de probabilidad, es decir, que la PCS obtenida esté por encima de un cierto valor prefijado P^* para cualquier configuración dentro de la zona de preferencia. Si se define la zona de preferencia de la manera usual $\delta = \theta_{[k]} - \theta_{[k-1]} \geq \delta^* > 0$, lo anterior implica que el experimentador requiere prefijar el par de parámetros (δ^*, P^*) antes de tomar observaciones. Esto se resume mediante el siguiente requerimiento de probabilidad. Encontrar el mínimo entero n tal que la desigualdad

$$P\{\text{CS}|\theta\} \geq P^* \text{ siempre que } \theta_{[k]} - \theta_{[k-1]} \geq \delta^* \tag{2.34}$$

se satisfaga para el par (δ^*, P^*) pre-especificado. En la práctica, el requerimiento (2.34) es verificado mediante la LFC como consecuencia de (2.32). Es decir, por lo general, se busca el mínimo entero n tal que $P\{\text{CS}|\theta_{LF}\} \geq P^*$. Por lo general esta expresión depende del contexto a través θ. Bechhofer [1954] provee tablas calculadas mediante métodos numéricos para el caso normal en algunos de los casos más importantes.

Una pregunta natural es cómo especificar los parámetros (δ^*, P^*). La interpretación de P^* es directa, se escoge de acuerdo con la mínima PCS que se quiera alcanzar con el experimento, bajo la consideración de que si es demasiado grande podría requerir un tamaño de muestra no realista o alcanzable. Es importante considerar que, para el caso $t = 1$, siempre se debe cumplir que $P^* > 1/k$, pues la probabilidad de elegir correctamente la mejor población aleatoriamente sin considerar muestra alguna es $1/k$, por tanto, valores de P^* inferiores a este valor no resultarían interesantes.

La interpretación de δ^* no es tan trivial y, por consiguiente, su especificación es más complicada. Una forma de abordar este subproblema puede ser considerando la formulación alternativa del problema. Supongamos que el tamaño de muestra común n ha sido determinado y que θ_s es el verdadero valor de la población seleccionada de acuerdo con el procedimiento de selección basado en dicho tamaño de muestra. Notemos que en esta situación una selección correcta habrá ocurrido si y sólo si $\theta_s = \theta_{[k]}$. Entonces, con al menos una probabilidad P^*, podemos asegurar que se satisface la desigualdad

$$0 \leq \theta_{[k]} - \theta_s \leq \delta^*. \tag{2.35}$$

O, equivalentemente,

$$P\{\theta_{[k]} - \delta^* \leq \theta_s \leq \theta_{[k]}\} \geq P^*. \tag{2.36}$$

En otras palabras, el experimentador puede decir, con al menos una confianza de P^*, que el tratamiento seleccionado θ_s estará a al menos δ^* unidades del verdadero mejor tratamiento. Por lo cual, una sugerencia para especificar el parámetro δ^* podría ser considerar (2.36) y pensar en δ^* como el máximo error que el experimentador está dispuesto a tolerar con una probabilidad tan alta como P^*. Es importante notar que (2.36) no constituye un intervalo de confianza para $\theta_{[k]}$ pues θ_s no se conoce ni es observable, sólo constituye una aseveración probabilística que en este caso se interpreta como auxiliar en la elección de δ^*.

Finalmente, el procedimiento aquí descrito y detallado en Bechhofer [1954] funciona bajo la premisa de que el experimentador tiene la libertad de escoger el tamaño de muestra n que más se adecue a sus requerimientos probabilísticos, sin embargo este no siempre es el caso. En muchos contextos existen restricciones temporales, monetarias o de otros tipos que limitan o imposibilitan la elección de n. Existen además escenarios en los cuales no es posible asumir un tamaño de muestra común para todos los tratamientos, sino que existen tamaños variables disponibles $n_1, n_2, ..., n_k$ respectivamente. Existen modificaciones particulares para tratar con éstos casos y otras variantes, un resumen de los detalles para diversas situaciones se puede consultar en Bechhofer et al. [1995].

Estimación de PCS

Bajo el procedimiento anterior, sabemos que, mientras $\delta \geq \delta^*$, la PCS siempre estará por arriba de P^* a partir de un cierto umbral para el tamaño de muestra. Este procedimiento puede y debería ser aplicado antes de realizar el procedimiento de muestreo. Sin embargo, si suponemos que el tamaño de muestra n ya ha sido pre-especificado, ya sea por el procedimiento anterior o por cualquier otra especificación (monetaria, temporal, *etc.*), es posible cambiar el enfoque de manera que se aproveche la información muestral de forma más activa. Este enfoque, relativamente más moderno, implica utilizar los datos para estimar la verdadera probabilidad de selección correcta.

Siguiendo la notación introducida en la Sección 2.2.1, Bechhofer [1954] derivó explícitamente la fórmula para la PCS en función del vector ordenado de parámetros verdaderos $(\theta_{(j)}, j = 1, 2, ..., k)$ dada por

$$P(CS_t) = P\left(\max_{1\leq i\leq k-t} Y_{(i)} \leq \min_{k-t+1\leq j\leq k} Y_{(j)}\right) \tag{2.37}$$

$$= \int_{-\infty}^{\infty} \prod_{j=k-t+1}^{k} \bar{F}(y - \theta_{(j)}) d\left\{\prod_{i=1}^{k-t} F(y - \theta_{(i)})\right\} \tag{2.38}$$

donde $\bar{F}(x) = 1 - F(x)$.

La expresión anterior es muy complicada de trabajar analíticamente, incluso para el caso más sencillo $t = 1$ en el cual se puede reducir a:

$$\mathrm{PCS}_1 = \int_{-\infty}^{\infty} \left[\prod_{i=1}^{k-1} F(y + \theta_{(k)} - \theta_{(i)}) \right] dF(y). \tag{2.39}$$

La principal dificultad subyacente de las fórmulas anteriores radica no en su complejidad sino en el hecho de que dependen de los parámetros desconocidos $\theta_{(j)}, j = 1, 2, ...k$. Así surge la necesidad de cambiar el enfoque, de calcular PCS a utilizar los datos para estimarla. Específicamente, si conocemos que la probabilidad de selección correcta depende de $\theta = (\theta_j), j = 1, 2, ...k$ a través de funciones específicas, como en (2.39), una idea intuitiva y razonable es hacer uso de las estadísticas resumen $Y_j, j = 1, 2, ...k$ para obtener un estimador de la verdadera PCS. Se pretende que a este estimador se le puedan también investigar propiedades como insesgadez, robusticidad, consistencia, *etc.*

A lo largo de los años una gran variedad de técnicas de estimación de PCS se han propuesto en la literatura. A continuación se presentará una breve reseña histórica de algunas las principales aportaciones al subproblema de la estimación de PCS. Reseñas más completas, de carácter más técnico, pueden ser consultadas en Wilson [2008] y Gupta and Liang [1998].

El primer estimador para PCS fue propuesto en Olkin et al. [1976]. Este estimador se pensó específicamente para el caso de selección de medias de la distribución normal y consiste, a grandes rasgos, de un estimador tipo *plug-in* basado en la expresión (2.39), dado por

$$\hat{\mathrm{PCS}}_1 = \int_{-\infty}^{\infty} \left[\prod_{i=1}^{k-1} F(y + Y_{[k]} - Y_{[i]}) \right] dF(y). \tag{2.40}$$

Olkin et al. [1982] encontró condiciones de consistencia y asintoticidad para este estimador y derivó cotas inferiores y superiores para $\hat{\mathrm{PCS}}$. Este estimador recibiría el nombre de estimador OST y con base en él se escribieron la mayoría de las publicaciones subsecuentes en relación a la estimación de PCS. La estructura base del los estimadores tipo OST se expresa como

$$\hat{\mathrm{PCS}}_1 = \int_{-\infty}^{\infty} \left[\prod_{i=1}^{k-1} F(y + a(Y_{[k]} - Y_{[i]}) \frac{\sqrt{n}}{\sigma}) \right] dF(y), \tag{2.41}$$

donde σ es la varianza, asumida común y conocida, de los tratamientos. Nótese que el caso $a = 1$ se reduce a (2.40) que es la forma más simple de un estimador tipo OST.

Sin embargo, a pesar de la simplicidad de (2.40), 7 años después, bajo petición de Bechhofer, Faltin and McCulloch [1983] realizaron un estudio de su efectividad en muestras pequeñas y reveló que, para n suficientemente pequeña, el estimador OST era sesgado, incluso en los casos en los que k no es muy grande. Posteriormente, Bofinger [1985] publicó las condiciones bajo las cuales los estimadores de PCS no pueden ser consistentes, lo cual puso aún más en duda la viabilidad del estimador OST y la línea de investigación de estimación puntual de PCS en general.

A pesar de que Bofinger [1985] pareció haber sentado las bases de una recesión en la búsqueda de estimadores consistentes para PCS, algunas propuestas desde otros puntos de vista surgieron. La primera de ellas fue McCulloch and Dechter [1985] que aplicó una visión Bayesiana alternativa al estimador OST y derivó un estimador puntual con la estructura de (2.41) pero con el factor de encogimiento

$$a_M = \left(1 - \frac{(k-3)\sigma^2/n}{\sum_{i=1}^{k} (Y_{[i]} - \bar{X})^2} \right)^+, \tag{2.42}$$

donde $\bar{X}$ es la media global y σ^2 y n son la varianza y el tamaño de muestra respectivamente. Otra propuesta se publicó en Bofinger [1990], donde se aplica nuevamente una técnica Bayesiana para deducir el factor de encogimiento

$$a_B = \eta\sqrt{n}/\sigma, \tag{2.43}$$

donde $\eta^{-2} = V^{-1} + n/\sigma^2$ y V se estima de los datos mediante $V = ((k-1)S^2/Z_B - \sigma^2/n)^+$, donde S^2 es la varianza muestral y Z_B es la mediana de distribución χ^2_{k-1}. En una propuesta más, Edwards [1992] publicó un nuevo estimador basado en el método empírico de Bayes y cuyo factor de encogimiento está dado por

$$a_E = \left(1 - \frac{(k-3)(k-1)\sigma^2/n}{4k(Y_{[i]} - \bar{X})^2}\right)^+, \tag{2.44}$$

donde el factor $k-3$ del numerador se omite si $k = 2, 3$. Finalmente, Sohn and Kahn [1992] propusieron estimadores tipo bootstrap, uno para el caso en que la varianza es común y conocida y otro para el que no se conoce. El algoritmo para estos estimadores se resume en Wilson [2008].

Todas las propuestas mencionadas en esta sección y algunas otras fueron comparadas a profundidad mediante un estudio de simulación para distintos tipos de configuraciones y valores de k en Cui and Wilson [2008]. La conclusión general a la que se llegó es que no existe un estimador universalmente superior a los demás; aún así resaltaron los estimadores aquí mencionados como los más recomendables.

Wilson [2008] también puntualiza la controversia emergida a causa de la interpretación de Bofinger [1985], en el cual se citan las condiciones bajo las cuales ningún estimador para PCS puede ser consistente. Wilson [2008] afirma que, dado que dichas condiciones están basadas en metodologías Bayesianas (donde $\theta = (\theta_i, i = 1, 2, ..., k)$ se asume aleatorio), es posible que bajo ese escenario particular $\hat{\text{PCS}}$ no sea consistente. Por otro lado, en un escenario frecuentista (donde θ se asume fijo), como en el caso de Olkin et al. [1982], no existe razón alguna para asumir que no es posible encontrar un estimador consistente para PCS. A pesar de que dicha distinción es clara en Bofinger [1985], frecuentemente fue malinterpretada en las (pocas) publicaciones posteriores y es considerada una de las razones principales por las que la línea de investigación relacionada con el estimador puntual de PCS dejó de publicarse en 1992. Recientemente, Cui and Wilson [2008] retoma el tema de estimación de PCS para un contexto más general que se desarrollará durante la Sección 2.3. Sin embargo, el problema de la estimación de PCS permanece aún vigente, dado que hasta la fecha no se ha encontrado un estimador puntual que satisfaga simultáneamente las propiedades deseables de insesgadez, asintoticidad y consistencia para cualquier configuración y/o cualquier valor de k o t. Es por esta razón que en 1985, Bechhofer, el padre de la metodología de selección y ordenamiento, lo nombró el #1 en la lista de problemas sin resolver de la RSM.

2.2.3. Principales Escenarios

En esta sección se plantearán varios ejemplos de casos particulares de selección y ordenamiento clásico, entre ellos el caso normal, el caso binomial, la selección multinomial y la selección por subconjunto (propuesta en Gupta [1956]). Se incluirán sólo los detalles principales y las referencias comentadas correspondientes. La intención es meramente informar al lector de las distintas variantes existentes y los principales resultados relacionados con ellas.

Selección de Medias Caso Normal

Posiblemente el ejemplo más simple de un problema de selección y ordenamiento en la literatura es el de seleccionar la población con mayor media de entre un conjunto de k poblaciones independientes con distribución normal. Formalmente y de manera general, utilizando la notación introducida en la Sección 2.2.1, el problema se escribe como sigue

Problema 2.2.1:
$n_1, n_2, ..., n_k$ observaciones independientes son tomadas de las poblaciones $\pi_i, i = 1, 2, ..., k$ respectivamente, denotadas por X_{ij}, donde $X_{ij} \sim N(\theta_i, \sigma_i^2), i = 1, 2, ..., k$. Denotaremos, por $\theta_{[i]}, i = 1, 2, ..., k$ los valores (desconocidos) ordenados de las medias. El objetivo es encontrar las *mejores* t poblaciones $(1 \leq t < k)$ donde *mejor* hace referencia a aquellas con mayor media. Es otras palabras, el objetivo es identificar a aquellas poblaciones con medias $\theta_{[k-t+1]}, ...\theta_{[k]}$.

Bechhofer [1954] fue el primero en plantear y estudiar el problema 2.2.1. Para ello determinó que, dado que el objeto de interés son las medias poblacionales, las correspondientes medias muestrales $Y_i = \bar{X}_i, i = 1, 2, ..., k$ serían consideradas. La regla de selección que propone es muy simple y consiste en seleccionar aquellas poblaciones con medias muestrales $Y_{[k-t+1]}, ...Y_{[k]}$ como las mejores t poblaciones donde la notación $Y_{[i]}, i = 1, 2, ..., k$ hace referencia a las medias ordenadas. En teoría $P(Y_{[i]} = Y_{[j]}) = 0$ para todos los índices $i \neq j$. Sin embargo, en la práctica pueden ocurrir empates debido a truncamientos en las mediciones u otras consideraciones. Para estos casos, Bechhofer [1954] recomienda romperlos mediante un procedimiento aleatorio en el cual se le asigna una probabilidad igual a todas las poblaciones involucradas.

Una vez que el procedimiento de selección ha sido especificado, Bechhofer [1954] propone estudiar dos posibles aseveraciones probabilísticas: la deducción de una fórmula explícita para la PCS y la determinación del tamaño muestral óptimo (véase la Sección 2.2.2). Distinguió además que el caso en el que las condiciones

$$n_1 = n_2 = \cdots = n_k = n \tag{2.45}$$

y

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2 \text{ con } \sigma^2 \text{ conocida} \tag{2.46}$$

constituye la situación más simple de trabajar analíticamente y por tanto dedicó la mayor parte del trabajo de la publicación a dicho caso. Para las situaciones en las que la condición (2.45) no se cumple, Bechhofer [1954] propuso una modificación adecuada al procedimiento anterior bajo ciertas condiciones. Un resumen completo puede ser consultado en Bechhofer et al. [1995]. Para las variantes en las que (2.46) no se cumple, por ejemplo en los casos en que las varianzas son conocidas pero distintas o comunes pero desconocidas, un procedimiento más elaborado es necesario y se puede consultar en Bechhofer et al. [1954] y Dunnett and Sobel [1954]. El caso más general en el cual no se asume igualdad ni conocimiento de las varianzas puede ser consultado en Bechhofer et al. [1995] y Gibbons et al. [1977]. Para propósitos ilustrativos en esta sección únicamente se presentarán las generalidades relacionadas con el caso en el que ambas (2.45) y (2.46) se satisfacen.

Bechhofer [1954] dedujo además que para este caso la LFC está dada por

$$\begin{aligned} \theta_{[k]} - \theta_{[k-t+1]} &= 0 \\ \theta_{[k-t+1]} - \theta_{[k-t]} &= \delta^* \\ \theta_{[k-t]} - \theta_{[1]} &= 0, \end{aligned} \tag{2.47}$$

donde δ^* es un parámetro no negativo pre-especificado. Demostró entonces que, si las condiciones (2.45)–(2.47) se satisfacen, la probabilidad de selección correcta se puede calcular mediante la siguiente expresión analítica cerrada:

$$P(CS_t) = P\left(\max_{1\leq i\leq k-t} Y_{(i)} \leq \min_{k-t+1\leq j\leq k} Y_{(j)}\right) \tag{2.48}$$

$$= tP\left(\max_{1\leq i\leq k-t} Y_{(i)} \leq Y_{(k-t+1)} \leq \min_{k-t+2\leq j\leq k} Y_{(j)}\right) \tag{2.49}$$

$$= t\int_{-\infty}^{\infty} [F(y+d)]^{k-t}[\bar{F}(y)]^{t-1}dF(y), \tag{2.50}$$

donde F es la función de distribución normal estándar y

$$d = \frac{\sqrt{n}\delta^*}{\sigma}. \tag{2.51}$$

Puede demostrarse que (2.50) es un caso particular de la expresión (2.37) que involucra una distribución F general.

Para el subproblema de la determinación del tamaño de muestra n, tal como se introdujo en la Sección 2.2.2, Bechhofer [1954] propone la pre-especificación de una pareja de parámetros (δ^*, P^*) que forma un requerimiento de probabilidad bajo el cual se busca encontrar el mínimo tamaño de muestra (común en este caso) que lo cumple. La idea central intuitiva sería aprovechar la expresión (2.50) (que depende de n), igualarla a P^* y resolver para n de manera que se obtenga el valor de n que garantiza el requerimiento de probabilidad para la LFC. Lo anterior se justifica, pues, como se vio en la Sección 2.2.1, si el requerimiento de probabilidad se satisface para la LFC se satisfacerá para cualquier otra configuración.

La limitación prinicipal del razonamiento anterior radica en que la ecuación (2.50) es difícil de tratar analíticamente por lo que surge la necesidad de determinar medidas alternativas. Afortunadamente, para este caso existe una manera (mediante métodos numéricos) de que, una vez fijada P^* se pueda determinar el valor de d (2.51) que hace que se satisfaga el requerimiento correspondiente, independientemente del valor de δ^*. Una tabulación de los valores de d para distintos valores de P^*, k y t puede encontrarse en Bechhofer [1954]. Una vez determinado el valor de d se obtendrá el tamaño de muestra óptimo n despejándolo de la ecuación (2.51), de modo que, el tamaño de muestra elegido se determina por

$$n = (\frac{d\sigma}{\delta^*})^2. \tag{2.52}$$

Para los casos en los que las poblaciones de interés no son las mejores, sino las peores (aquellas con las menores medias), el problema anterior puede adaptarse mediante una simple transformación. Dado que la distribución normal es simétrica alrededor de la media θ y este constituye un parámetro de localización, se tiene que la distribución de $\bar{X} - \theta$ no depende de θ. Como resultado si reemplazamos cada $\bar{X}$ por $-\bar{X}$ será equivalente a reemplazar $\theta_{[j]}$ por $\theta_{[k-j+1]}$ para $j = 1, 2, ...k$ lo cual provoca el mismo efecto en las estadísticas resumen ordenadas $\bar{X}_{[j]}, j = 1, 2, ..., k$. Visto de esta manera, mediante esta transformación obtenemos un problema equivalente al anterior donde la media más pequeña será ahora la media más grande y así sucesivamente. En general, en términos de PCS, no es equivalente buscar las t mejores poblaciones que buscar las peores $k - t$ poblaciones como se detallará para casos más generales en la Sección 2.3.

Selección Binomial

En esta sección se considerará el problema en el cual se busca seleccionar la mejor población de un conjunto de k posibilidades, donde cada individuo dentro de cada población es clasificado en una y sólo una de dos categorías mutuamente excluyentes, que denominaremos *éxito* y *fracaso* respectivamente. El siguiente ejemplo ilustra una situación particular en la que este modelo surge.

Ejemplo 2.2.1:
Una compañía tiene diez empleados en el área de ventas por catálogo. Estos empleados tienen la tarea de intentar convencer a los clientes de consumir sus productos yendo casa por casa en áreas socio-económicamente equivalentes en distintas ciudades. La empresa está interesada en ofrecerle un bono al empleado que califique como el *mejor vendedor*. Por lo que es de interés seleccionar al empleado con más poder de convenicimiento o posibilidades de éxito al intentar vender los productos.

En cada una de las poblaciones binomiales consideradas anteriormente existe una cierta probabilidad de que un individuo determinado sea clasificado como éxito, denotaremos estas probabilidades como

$p_1, p_2, ..., p_k$. Definiremos de esta manera las *mejores* poblaciones como aquellas que tiene una mayor probabilidad de éxito o, en algunos contextos, las de mayor probabilidad de fracaso. En el Ejemplo 2.2.1 $p_i, i = 1, 2, ..,10$ representa la probabilidad de que un cliente atendido por el empleado i realice una compra, por lo que la empresa estará interesada en encontrar al vendedor que corresponde a $p_{[k]} = p_{[10]}$ donde

$$p_{[1]} \leq p_{[2]} \leq \cdots \leq p_{[k]} \tag{2.53}$$

son las probabilidades (desconocidas) ordenadas. Sin embargo, si por ejemplo, a la empresa le interesara realizar un recorte de personal y despedir a los $t < k$ empleados con peor desempeño entonces podría estar interesada en encontrar a los empleados que corresponden a $p_{[1]}, p_{[2]}, ..., p_{[t]}$. En éstos casos es fácil redefinir el problema tomando $q_i = 1 - p_i, i = 1, 2, ..., k$ y notando que $p_{[1]}, p_{[2]}, ..., p_{[t]} = q_{[t-k+1]}, q_{[t-k+2]}, ..., q_{[k]}$ se constituye un problema equivalente pero escrito en términos de las probabilidades de fracaso.

En términos del modelo, el problema anterior puede escribirse como un conjunto de $n_1, n_2, ...n_k$ individuos de cada población respectivamente donde cada individuo en cada población es clasificado como *éxito* o *fracaso* con probabilidades $p_1, p_2, ..., p_k$ respectivamente. En otras palabras, se tiene observaciones independientes X_{ij} de k poblaciones $\pi_i, i = 1, 2, ..., k$ donde $X_{ij} \sim Bin(n_i, p_i)$ y la meta general es seleccionar aquellas t poblaciones con las mayores probabilidades de éxito $p_{[t-k+1]}, p_{[t-k+2]}, ..., p_{[k]}$.Es necesario enfatizar, como en los casos anteriores, que el interés no es en estimar las más grandes probabilidades de éxito sino en seleccionar a las poblaciones que corresponden a ellas.

Sobel and Huyett [1957] fue la primera publicación en relación al problema de selección binomial. En ella se discernió que este tipo de problema era de naturaleza diferente al caso normal estudiado en Bechhofer [1954] por dos motivos. El primero es que los parámetros de interés $p_i, i = 1, 2, ..., k$ tienen un soporte acotado $[0, 1]$ a diferencia del caso de las medias de poblaciones normales que pueden tomar cualquier valor en la recta real. El segundo motivo es que, a diferencia del caso anterior, no se tiene una noción natural de distancia entre parámetros. Una primera propuesta, análoga al caso normal es la función

$$\delta(p_i, p_j) = p_i - p_j, \tag{2.54}$$

donde naturalmente, es de particular interés la distancia dada por $\delta(p_{[t-k+1]}, p_{[t-k]})$, es decir, la distancia entre las mejores y la población inmediata siguiente de las mejores. Otra propuesta considerada en Sobel and Huyett [1957] es la llamada distancia de la razón de momios (2.31) que, en términos de la notación introducida en esta sección, está dada por

$$\delta_{OR}(p_i, p_j) = \frac{p_i q_j}{p_j q_i}. \tag{2.55}$$

La distancia (2.55) representa la razón entre p_i/q_i y p_j/q_j, es decir, compara cuántas veces es más probable un éxito que un fracaso en la población i contra la población j. Sobel and Huyett [1957] decidieron restringirse únicamente al uso de la distancia (2.54) debido a que determinó que era posible encontrar una solución analítica al problema de selección con ella en el sentido tradicional. En otras palabras, determinó que era posible escoger una pareja de parámetros (δ^*, P^*) con $0 < \delta^* < 1$ y $1/k < P^* < 1$, tales que se puede encontrar un mínimo tamaño de muestra necesario n para garantizar que la PCS será al menos P^* siempre que $\delta \geq \delta^*$. Esta importante propiedad no puede garantizarse para cualquier medida de distancia, en particular no para la distancia (2.55). Lo anterior resulta conveniente pues la interpretación y manejo computacional de (2.54) son más simples que en el caso de la distancia (2.55). Bechhofer et al. [1995] resume algunas de las principales técnicas que pueden aplicarse para los casos en los que se desee emplear la distancia (2.55).

La regla de selección para el caso binomial, propuesta en Sobel and Huyett [1957] en analogía a la propuesta para el caso normal en Bechhofer [1954], consiste en tomar las proporciones de éxitos observados en la muestra de cada población. Supongamos que se tomaron n_i observaciones de la $i-$ésima

población y x_i es el número de éxitos observados en dicha muestra, entonces el conjunto de estadísticas resumen para este caso se propone como

$$Y_i = \frac{x_i}{n_i}. \tag{2.56}$$

De manera análoga al caso normal, las estadísticas resumen se ordenan de menor a mayor $Y_{[i]}, i = 1, 2, ..., k$, y la regla de selección propuesta en Sobel and Huyett [1957] dicta que se seleccionen las t poblaciones correspondientes a las proporciones muestrales mayores como las mejores poblaciones. Nuevamente, en caso de empates estos se romperán de manera completamente aleatorizada.

Debido a que el parámetro p de una distribución binomial no es precisamente un parámetro de localización, existen diversas dificultades al tratar de calcular aspectos analíticos análogos al caso normal. Por ejemplo, para el caso de $t = 1$ y tamaños de muestra iguales $n_1 = n_2 = \cdots = n_k = n$, no es suficiente fijar $p_{[k]} - p[k-1] = \delta^*$ para poder encontrar la LFC. Sobel and Huyett [1957] demostraron analíticamente que la LFC para el caso binomial con la distancia usual, está dada por:

$$p_{[1]} = p_{[2]} = \cdots = p_{[k-1]} = \frac{1}{2} - \frac{\delta^*}{2} \text{ y } p_{[k]} = \frac{1}{2} + \frac{\delta^*}{2}. \tag{2.57}$$

Desarrolló además un cojunto de tablas, gráficos de interpolación y resultados analíticos que complementan el subproblema de selección del tamaño de muestra para el caso de tamaños de muestra iguales. Un resumen de algunos procedimientos más sofisticados para casos más específicos puede consultarse en Bechhofer et al. [1995].

Selección de la Varianza Mínima

Como se explicó en las secciones anteriores el objetivo en un problema de selección no siempre es encontrar las *mejores* poblaciones. Existen muchos contextos en los cuales el interés está en identificar cuáles son las *peores* poblaciones o aquellas que minimizan cierta cantidad de interés. En esta parte se revisita el escenario clásico en el cual se tienen observaciones independientes X_{ij} que provienen de k poblaciones $\pi_i, i = 1, 2, ..., k$ donde $X_{ij} \sim N(\mu_i, \sigma_i^2)$. En esta ocasión, las medias $\mu_i, i = 1, 2, ..., k$ se asumirán como parámetros de estorbo y el principal interés estará en el ordenamiento (desconocido) de las varianzas $\sigma^2_{[1]} \leq \sigma_{[2]} \leq \cdots \leq \sigma^2_{[k]}$. Debido a que la varianza tiene la interpretación de ser la dispersión de los datos alrededor de la media, en muchos contextos es de particular interés seleccionar aquellas poblaciones que tienen la menor varianza, pues una varianza pequeña se traduce a mayor certeza de que los datos están concentrados cerca de la media (en especial si ésta se asume conocida). Es común encontrar contextos en la industria en las cuales la selección de la varianza mínima se aplica; el Ejemplo 1.1.3 ilustra uno de ellos.

La comparación de varianzas de k poblaciones normales es un problema recurrente en la estadística, incluso antes de que surgiera el problema de selección en la literatura por medio de Bechhofer [1954]. Esto se debe a que, en muchas aplicaciones, por ejemplo en diseño experimental, análisis de regresión o incluso en varias instancias de RSM, el supuesto de varianza constante es fundamental. Tradicionalmente este problema se maneja mediante una hipótesis de homogeneidad de varianza $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$ y existen un gran número de pruebas disponibles en la literatura para resolver este problema, por ejemplo las pruebas de Bartlett, Cochran, Hartley o Levene. Bajo este contexto el experimentador quisiera concluir que las varianzas son en realidad iguales. En el contexto de selección y ordenamiento queremos asumir, en cambio, que las varianzas son distintas y estamos interesados en encontrar aquellas que son más pequeñas. En general ambas metodologías difieren significativamente y gran parte de sus diferencias será discutida a detalle en el siguiente capítulo.

Bechhofer and Sobel [1954] estudiaron el problema de selección de la varianza mínima distinguiendo principalmente dos casos: aquél en el que las medias $\mu_i, i = 1, 2, ..., k$ son todas conocidas (no necesariamente iguales) y aquél en el que las medias son todas desconocidas. El caso intermedio en el que

se conocen sólo algunas de las medias resultó ser considerablemente más complicado en teoría, aunque se han propuesto algunos métodos que utilizan los dos casos extremos mencionados anteriormente y utilizarlos como cotas inferior y superior para el caso intermedio respectivamente.

Dado que las varianzas son parámetros estrictamente positivos, una función de distancia más efectiva para este caso que la usual de los casos anteriores es la razón

$$\delta(\sigma_i, \sigma_j) = \frac{\sigma_j}{\sigma_i}, \tag{2.58}$$

donde es de particular interés la distancia $\delta = \delta(\sigma_{[t]}, \sigma_{[t+1]})$. Visto de esta manera, la zona de indiferencia se definirá como todas las configuraciones de las varianzas que cumplan que $\delta < \delta^*$ y la zona de preferencia como $\delta \geq \delta^*$. Dado que (2.58) puede tomar valores en $[1, \infty)$, Gibbons et al. [1977] propone una transformación simple,

$$\Delta(\sigma_i, \sigma_j) = 1/\delta(\sigma_i, \sigma_j) = \frac{\sigma_i}{\sigma_j}, \tag{2.59}$$

de modo que $\Delta = \Delta(\sigma_{[t]}, \sigma_{[t+1]}) = \frac{\sigma[t]}{\sigma_{[t+1]}}$ y $0 < \Delta < 1$. Naturalmente, si se utiliza la distancia (2.59), las desigualdades que definen las zonas de preferencia e indiferencia se invierten.

Bechhofer and Sobel [1954] encontraron que la LFC, para el caso $t = 1$ con la función de distancia (2.58) está dada por

$$\sigma_{[2]} = \sigma_{[3]} = \cdots = \sigma_{[k]} \text{ y } \delta^* = \frac{\sigma_{[2]}}{\sigma_{[1]}}, \tag{2.60}$$

para un cierto valor pre-especificado $\delta^* > 0$. Bechhofer and Sobel [1954] diseñó además, de manera análoga a los casos anteriores, un procedimiento que permita seleccionar la población cuya varianza corresponde a $\sigma^2_{[1]}$ con una probabilidad de al menos $P^* > 1/k$ siempre que $\delta \geq \delta^*$, para una pareja de parámetros pre-especificados (δ^*, P^*).

La regla de selección propuesta por Bechhofer and Sobel [1954] se basa en el hecho de que es posible estimar la varianza de una población normal, σ^2 mediante la varianza muestral. Si X_{ij} representa la $i-$ésima observación de la $j-$ésima población con distribución $N(\mu_j, \sigma^2_j), j = 1, 2, ..., k$ y las medias son todas conocidas entonces podemos estimar a σ^2_j mediante

$$S_j = \frac{\sum_{i=1}^{n_j}(X_{ij} - \mu_j)^2}{n_j}. \tag{2.61}$$

Aquí, de nuevo, $n_1, n_2, ..., n_k$ representan los tamaños de muestra correspondientes a cada población respectivamente. Si, en cambio, todas las medias son desconocidas, el estimador (2.61) se escribe como

$$S^*_j = \frac{\sum_{i=1}^{n_j}(X_{ij} - \bar{x}_j)^2}{n_j - 1}, \tag{2.62}$$

donde $\bar{x}_j$ representa la media muestral correspondiente a la $j-$ésima población que actúa como un estimador del valor desconocido μ_j. Una vez determinadas las estadísticas resumen $S_j, j = 1, 2, ..., k$ (o S^*_j según sea el caso) la regla de decisión consiste en ordenarlas y escoger a las t poblaciones que produjeron las varianzas muestrales más pequeñas como las t poblaciones con mínima varianza. Bechhofer and Sobel [1954] desarrollaron los aspectos analíticos y las tabulaciones correspondientes en relación con la determinación del tamaño de muestra para el caso en que se asumen tamaños iguales $n_1 = n_2 = \cdots = n_k = n$, el procedimiento, en teoría, es análogo al de los casos anteriores.

Selección Multinomial

El modelo multinomial de k categorias consiste en una generalización del modelo binomial (2 categorías: éxito y fracaso). En él se asume que cada observación pertenece a una y sólo una de un conjunto de

k categorías mutuamente excluyentes con probabilidades $p_1, p_2, ..., p_k$ respectivamente donde $\sum_i p_i = 1$. Una observación del modelo multinomial puede entonces considerarse equivalente al lanzamiento de un dado justo de k caras donde cada cara representa una categoría.

En muchos contextos es de especial interés averiguar cuáles de las categorías de un modelo multinomial son las más (o menos) probables, es decir, aquellas con mayor (o menor) probabilidad de ocurrencia p_i. Si ordenamos las probabilidades de acuerdo con la notación del capítulo $p_{[1]} \leq p_{[2]} \leq \cdots \leq p_{[k]}$ la situación anterior puede traducirse a un problema de selección en el cual la meta es seleccionar las $t \leq 1$ categorías que corresponden a las probabilidades $p_{[k-t+1]}, ..., p_{[k]}$.

Es importante notar que, a pesar de que el modelo multinomial es una generalización del modelo binomial, la selección multinomial no es una generalización de la selección binomial cuando $k = 2$. Si hacemos $k = 2$ el problema de la selección multinomial se reduce a la búsqueda de la mejor (o la peor) entre dos categorías de un experimento simple binomial («¿Es más probable un éxito o un fracaso?»), mientras que la selección binomial consiste en encontrar la mejor de k poblaciones binomiales («¿Cuál de las k poblaciones tiene mayor probabilidad de éxito?»). Comparar ambas metodologías en el caso de $k = 2$ no tiene sentido.

Para ilustrar una situación particular en la cual un problema de selección multinomial emerge se propone el siguiente ejemplo.

Ejemplo 2.2.2:
Una compañía fabrica dulces de distintos colores y ha hecho el uso del color parte de su estrategia de mercado más efectiva. Actualmente se producen dulces en colores rojo, azúl, verde, amarillo, verde y rosa. Los directivos de mercadotecnica de la empresa insisten que el color tiene un especial impacto en la percepción del público y que existen colores que atraen más la atención de los clientes potenciales que otros. Para ello se lleva a cabo un estudio de mercado en el cual se requiere seleccionar el/los color (es) que el público objetivo de los dulces prefiere consumir. La compañía prepara entonces un mecanismo de muestreo en el cual se le pide a cada cliente que seleccione cuál es el color que más le apetece de una muestra de dulces preparada equitativamente de manera que todos los colores estén visualmente accesibles. El color que resulte más *popular* en los resultados del estudio será producido en mayores cantidades mientras que el menos preferido será reducido o, si el caso lo amerita, eliminado o reemplazado de las opciones.

Como en el caso de la selección normal y binomial, en la selección multinomial los aspectos analíticos de principal interés en la literatura existente radican en la determinación del tamaño de muestra óptimo n de manera que se cumpla el requerimiento de probabilidad usual en términos de la dupla pre-especificada (δ^*, P^*).

Bechhofer et al. [1959] propusieron para este caso la función de distancia dada por

$$\delta(p_i, p_j) = \frac{p_j}{p_i}. \tag{2.63}$$

Donde nuevamente, es de particular interés la distancia $\delta = \delta(p_{[k-t]}, p_{[k-t+1]})$. De manera análoga al caso de la selección de varianza mínima $\delta \geq 1$ por definición y buscamos encontrar una regla de selección tal que la PCS sea al menos P^* siempre que $\delta \geq \delta^*$ que representa los casos en los cuales nos interesa hacer una buena selección, *i.e.* la zona de preferencia, como se definió anteriormente.

Bechhofer et al. [1959] encontraron además que, para el caso $t = 1$, dadas las condiciones definidas anteriormente, la LFC para el problema de selección multinomial con k categorías está dada por

$$p_{[1]} = p_{[2]} = \cdots = p_{[k-1]} \text{ y } p_{[k]} = \delta^* p_{[k-1]}. \tag{2.64}$$

Además, si se toma en cuenta la condición $\sum_i p_i = 1$ entonces (2.64) se puede rescribir como:

$$p_{[1]} = p_{[2]} = \cdots = p_{[k-1]} = (\delta^* + k - 1)^{-1} \text{ y } p_{[k]} = \delta^*/(\delta^* + k - 1). \tag{2.65}$$

Bechhofer et al. [1959] encontró además la solución al subproblema de la determinación del tamaño de muestra para el caso más sencillo $t = 1$ y publicó las tabulaciones necesarias para distintos valores de δ^* y P^* como se hizo en los casos anteriores.

Más de una década después, Alam and Thompson [1959] demostraron que — en contra de la intuición — el problema de seleccionar las *peores* o menos probables categorías en un modelo multinomial no es equivalente ni reducible del caso descrito anteriormente para seleccionar las mejores categorías. En particular encontraron que la función de distancia (2.63) no podría ser utilizada como función de distancia para $k \geq 3$. La razón de esto radica en que, si estamos interesados en encontrar las t categorías menos probables, *i.e.* aquellas correspondientes a las probabilidades $p_{[1]}, ..., p_{[t]}$, y utilizamos la distancia (2.63), estaremos interesados en el cociente $p_{[t+1]}/p_{[t]}$. Sin embargo, es fácil ver que podemos mantener ese cociente fijo y, al mismo tiempo, hacer las probabilidades $p_{[t]}$ y $p_{[t+1]}$ arbitrariamente pequeñas si hacemos crecer $p_{[k]}$ (y $p_{[k]}$ no necesariamente es igual a $p_{[t+1]}$ dado que $k \geq 3$). De esta manera se encontró un caso particular en el cual $p_{[1]}, ..., p_{[t+1]}$ no pueden distinguirse fácilmente de manera que se cumpla algún requerimiento de probabilidad P^* al elegir las t mejores, sin importar cuánto se incremente el tamaño de muestra. Alam and Thompson [1959] propusieron para estos casos la función de distancia usual (2.54) y encontraron una solución para el caso más simple $t = 1$.

Selección por Subconjunto (SS)

En ocasiones la meta de un problema de selección no consiste explícitamente en encontrar las mejores t poblaciones de un conjunto de k distintas opciones, sino en encontrar un subconjunto (de tamaño aleatorio) que las contenga con alta probabilidad. Este problema se conoce en la literatura como *selección por subconjunto (subset selection)* y aparece por primera vez en Gupta [1956].

Es interesante notar que la selección por subconjunto desde su concepción es un problema distinto al planteado en Bechhofer [1954], en el sentido que, para Bechhofer [1954] el interés principal es *diseñar* un experimento, de tal manera que se determine el tamaño de muestra n que garantizará al experimentador que cierto requerimiento probabilístico se satisfaga. Para Gupta [1956], en cambio, la meta es *filtrar* un conjunto de tratamientos de manera que se seleccione un subconjunto que contenga los mejores con cierta probabilidad, por tanto, es relativamente más recomendable para situaciones en las que se requiere *analizar* un experimento con tamaños de muestra arbitarios.

La metodología de SS es de particular utilidad en situaciones en las cuales se busca *eliminar* tratamientos inferiores (por ejemplo aquellos con menor media) y retener únicamente los mejores de manera que puedan pasarse a una segunda etapa en otro estudio, donde pueden ser de interés otros aspectos más específicos. Por ejemplo, se puede buscar seleccionar al mejor de ellos mediante el procedimiento IZ de Bechhofer [1954]. Este es un ejemplo de un escenario donde se puede encontrar la conexión entre ambas metodologías de forma explícita, pero no es el único.

Gupta [1956] estudió a detalle el caso más simple de SS que consiste en considerar n observaciones independientes X_{ij} de cada de una de un conjunto de poblaciones $\pi_i, i = 1, 2, ..., k$ donde $X_{ij} \sim N(\mu_i, \sigma^2)$ y σ^2 se asume conocida. Una característica importante del procedimiento de SS en contraste con IZ es que es posible asumir que σ^2 es desconocida y, mediante una modificación no muy drástica al procedimiento, obtener una solución adecuada. Los detalles para el caso en que se tienen tamaños de muestra distintos pueden consultarse en Gupta and Huang [1976]. El objetivo de la presente sección es meramente ilustrativo por lo que se restringirá la atención únicamente al caso más simple expuesto en Gupta [1956].

Bajo las características mencionadas anteriormente, el objetivo del procedimiento SS, formalmente, es seleccionar un subconjunto (de tamaño aleatorio) que contenga la población con mayor media que denotaremos por $\mu_{[k]}$ siguiendo la notación de las secciones anteriores. Si dicho objetivo se cumple se dice que ocurre un evento de selección correcta (CS) y la probabilidad de que ésto ocurra se denotará como PCS para conservar uniformidad en la notación.

Dado que es de interés que la mejor población quede contenida en el intervalo seleccionado con alta confianza, es necesario el establecimiento de un requerimiento mínimo de probabilidad. De manera análoga a como se hizo en la metodología de IZ, se predefinirá un parámetro P^* tal que $1/k < P^* < 1$ y se requerirá que

$$P(\text{CS}|\mu) \geq P^* \tag{2.66}$$

para todas las configuraciones $\mu = (\mu_1, \mu_2, ..., \mu_k)$.

Claramente es deseable que el subconjunto seleccionado sea lo más pequeño posible ya que de esta manera obtendríamos información más precisa (menos poblaciones erróneas escogidas). Sin embargo, garantizar que la mejor población esté contenida en un subconjunto de tamaño pequeño con alta probabilidad podría ser difícil o imposible para algunos casos, en especial si el tamaño de muestra no es suficientemente grande para permitir que la mejor población se distinga lo suficiente de las restantes.

El algoritmo propuesto en Gupta [1956] para la selección de un subconjunto que contenga a la mejor población normal (en términos de la media) y que cumpla con el requerimiento de probabilidad (2.66) consiste en el cálculo de la media muestral por cada población $\bar{X}_i, i = 1, 2, ..., k$ y ordenarlas $\bar{X}_{[i]}, i = 1, 2, ..., k$. Luego, se incluirá la población π_i en el subconjunto seleccionado si y sólo si se cumple que

$$\bar{X}_i \geq \bar{X}_{[k]} - h\sigma\sqrt{2/n}, \tag{2.67}$$

donde h es una constante positiva que depende de P^* obtenida a partir de una distribución normal multivariada y tabulada en Gupta [1956]. En caso de que σ^2 se desconozca, puede estimarse mediante

$$S_\nu^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2/\nu, \tag{2.68}$$

donde $\nu = t(n-1)$. En este caso la expresión para h es distinta pues depende de una distribución t. Las tablas para ambos casos pueden encontrarse en Gupta [1956] y un resumen más detallado de otras variantes del procedimiento se puede consultar en Bechhofer et al. [1995].

La interpretación del criterio de selección (2.67) es que la población π_i será incluida en el subconjunto seleccionado si y sólo su media muestral $\bar{X}_i$ está lo suficientemente cerca de la mayor media muestral $\bar{X}_{[k]}$ donde la constante $C = h\sigma\sqrt{2/n}$ se utiliza como umbral para determinar cuales poblaciones están suficientemente cerca y cuales no lo están. Notemos que, mientras más grande sea σ o más pequeña sea n, mayor será el umbral y más poblaciones tendrán la posibilidad de ser incluidas, incrementando el tamaño del subconjunto seleccionado. Gupta [1956] demostró además que el procedimiendo (2.67) garantiza que el subconjunto seleccionado cumple con el requerimiento (2.66) y que además, si denotamos por S el tamaño del subconjunto seleccionado por dicho procedimiento, se cumple que

$$\text{E}(S) \to 1 \ \text{ y } \ \text{PCS} \to 1 \ \text{ cuando } \ n \to \infty. \tag{2.69}$$

Lo anterior ilustra de manera explícita el efecto del tamaño de muestra n en la calidad de la selección.

En general SS supone un procedimiento alternativo a IZ para situaciones más específicas, como se explicó al inicio de la sección. Como tal corresponde a un problema distinto con un procedimiento particular y, naturalmente, permite llegar a conclusiones de diferente naturaleza. Entre las ventajas de la utilización de SS para resolver un problema de selección se encuentra el hecho de que permite alcanzar mayores valores de PCS con relativa facilidad. Esto se debe a la flexibilidad que ofrece seleccionar un subconjunto completo de poblaciones que contiene a la mejor, en vez de buscar seleccionar sólo una en específico, razón por la cual permite extraer más información del conjunto de datos en algunos contextos. Por otro lado, el procedimiento SS siempre seleccionará un subconjunto de tamaño aleatorio que depende enteramente del valor de P^* pre-especificado, si P^* es demasiado grande (como en la mayoría de los casos es deseable) el subconjunto resultante puede ser incómodamente grande; esta falta de control sobre

el tamaño del subconjunto podría considerarse una de las desventajas distintivas del procedimiento SS en varios contextos.

A lo largo de los años la teoría de selección por subconjunto ha estado en constante evolución. Recientemente, con el surgimiento de conjuntos de datos cada vez mayores, la visión de Gupta [1956] de realizar un proceso de filtrado que permita retener sólo información selecta ha tomado auge nuevamente. En conjuntos masivos de datos el enfoque se ha visto forzado a cambiar de *seleccionar las mejores* a *seleccionar un conjunto con las mejores*. La adaptación de las técnicas de selección a estos nuevos contextos y las nuevas metodologías que emergieron con ello es la idea central de la Sección 2.3.

2.2.4. Limitantes de la Metodología Clásica

Existen dos tipos de inconvenientes en la metodología clásica de selección y ordenamiento. El primero es la falta de expresiones analíticas cerradas para el cálculo de PCS o, en su defecto, sus estimadores. Una fórmula explícita exacta para PCS se deduce en Bechhofer [1954] y se generaliza en la ecuación (2.37) para cualquier configuración y cualquier distribución F de localización. Sin embargo, es fácil ver que (2.37) no puede ser calculada de manera analítica y por tanto gran parte de los resultados que dependan de ella requerirán la implementación de métodos numéricos (y/o resultados tabulados) que no siempre son económicos computacionalmente. Cui and Wilson [2008] realizó una aproximación a (2.37) utilizando una cuadratura de Gauss-Hermite e implementó el algoritmo correspondiente en R (www.r-project.org), sin embargo, se llegó a la conclusión de que requería gran poder de cómputo para funcionar de manera óptima, especialmente en los casos en los que k no es muy pequeña. Una situación similar ocurre con la expresión para los estimadores tipo OST (2.41). Por esta razón, Cui and Wilson [2008] optaron por la implementación de estimadores tipo Bootstrap para PCS inspirados en las ideas de Sohn and Kahn [1992], estos estimadores resultarían ser más eficientes computacionalmente y brindarían resultados interesantes; los detalles se reservan para la siguiente sección.

El segundo inconveniente, y el más importante, es la imposibilidad conceptual de generalizar los métodos clásicos de RSM para situaciones en las que k es grande o masivamente grande. Es intuitivo ver que no es equivalente el problema de encontrar la mejor población entre un conjunto de $k = 10$, que el problema de encontrarla entre un conjunto de $k = 10000$. Naturalmente, conforme el número de poblaciones a escoger se incrementa, más factible se vuelve el hecho de que sus observaciones se confundan entre sí e, independientemente del caso, si n no es suficientemente grande, la posibilidad de cometer un error se incrementa inevitablemente. Esto ocasiona que PCS se acerque a cero en todos o la mayoría de los métodos clásicos de selección conforme k se incrementa. Con el surgimiento de nuevas bases de datos de tamaño cada vez mayor esta situación se vuelve apremiante y nuevas metodologías son necesarias. Cui and Wilson [2008] respondieron a esta situación con una redefinición conceptual y metodológica de las nociones de PCS, y propusieron un nuevo escenario diseñado específicamente para los casos en los que k es grande. Estas técnicas modernas de selección y ordenamiento se reseñarán a continuación en la siguiente sección y se implementarán computacionalmente en el Capítulo 4.

2.3. Selección y Ordenamiento para k grande

Esta sección informará de manera amplia pero concisa sobre las metodologías modernas de selección y ordenamiento que son motivo propiamente de la tesis. La mayoría de las ideas expuestas aquí serán basadas en lo expuesto en Cui and Wilson [2008].

2.3.1. Motivación

En los últimos 50 años, el surgimiento de nuevas tecnologías de información y procesamiento ha brindado a los estadísticos la posibilidad de almacenamiento y procesamiento de cantidades de información cada vez más grandes de manera simultánea. Al mismo tiempo, con el surgimiento del internet, conjuntos de datos de tamaño cada vez más masivo se han puesto a su disposición y por consiguiente, la demanda en las aplicaciones de la estadística requiere cada vez más de técnicas especializadas en su manejo y análisis.

La existencia de bases de datos de tamaño masivo llama de manera natural a la necesidad de un procedimiento de *filtrado* que permita a los estadísticos retener únicamente la información que es esencial y por tanto *interesante* para el análisis estadístico. En la mayoría de estos casos, el proceso de filtrado no debería basarse en sólo buscar los individuos o piezas de información más distinguibles o *significativos* de los demás, ya que, al tratarse de un conjunto de datos masivo, es sabido que se puede encontrar un conjunto de individuos significativos (que posiblemente sigue siendo masivo) para un determinado tamaño de muestra. En otras palabras, sostener la hipótesis de que en un conjunto masivo de datos no se pueden encontrar diferencias resulta absurdo. En estos casos, el interés del investigador tiene que cambiar de buscar cuáles individuos son significativos a buscar precisamente los *mejores* o más *interesantes* en términos de los objetivos del estudio.

Un proceso de filtrado como se describe anteriormente puede ser visto como una instancia del problema de selección introducido en la Sección 2.2. Bajo este paradigma, se puede pensar en un conjunto de datos masivo como un conjunto de observaciones $X_{ij}, j = 1, 2, ..., n$ de k tratamientos o poblaciones distintas $\pi_i, i = 1, 2, ...k$ donde $k > 1$ puede ser arbitrariamente grande. El interés del investigador recae en seleccionar únicamente las mejores t poblaciones, donde el término *mejor* tiene un significado específico en términos del contexto. Ésto consiste la base de un problema de selección y ordenamiento en la RSM (Bechhofer [1954]).

La gran mayoría del trabajo en la literatura de la RSM se concentra en dos grandes enfoques: la zona de indiferencia (IZ) de Bechhofer [1954] y la selección por subconjunto (SS) de Gupta [1956]. La IZ, desde el punto de vista de diseño experimental, propone prefijar el parámetro δ^* y la mínima PCS P^* deseada para poder determinar el mínimo tamaño de muestra necesario n. En contraste, la selección por subconjunto fija n y P^* y estima un número aleatorio de poblaciones a seleccionar. No es difícil ver que, aplicar dichas metodologías a un caso en el que k es masivamente grande y P^* moderadamente alta, dará resultados insatisfactorios: IZ reportará un tamaño de muestra exageradamente grande y SS seleccionará un subconjunto con demasiados elementos (posiblemente todos los disponibles).

El hecho de que la metodología clásica falle para los casos en los que k es masivamente grande resulta ser intuitivo. En cualquier situación, no es equivalente, en términos de dificultad, buscar el mejor elemento en un conjunto de, digamos, $k = 5$ poblaciones que buscarlo en un conjunto de $k = 10^5$ poblaciones. Visto desde un punto de vista más formal, si recordamos que tanto IZ como SS basan sus criterios de selección en en el conjunto de las estadísticas resumen obtenidas de cada población $Y_i, i = 1, 2, ..., k$, entonces, si k es masivamente grande es intuitivo que las estadísticas Y_i se acercarán cada vez más entre sí dificultando que aquellas que provienen de las verdaderas mejores destaquen como tales lo cual reducirá inevitablemente la PCS alcanzable. Por ejemplo, para el caso más simple de la selección de medias normales (véase la Sección 2.2.3), la Figura 2.7 muestra el histograma de las medias muestrales Y_i para un caso simulado con $k = 5000$ poblaciones normales con medias verdaderas simuladas de una distribución exponencial con media $1/2$ y varianza común $\sigma^2 = 1$. Puede verse que si, por ejemplo, el experimentador estuviera interesado en encontrar las mejores $t < 1000$ con alta probabilidad P^* dicho conjunto de datos podría no permitírselo.

Los problemas de selección con k grande son cada vez más comúnes en la práctica. A manera de ilustración se ofrecen los siguientes ejemplos.

Ejemplo 2.3.1 (Microarreglos)**:**

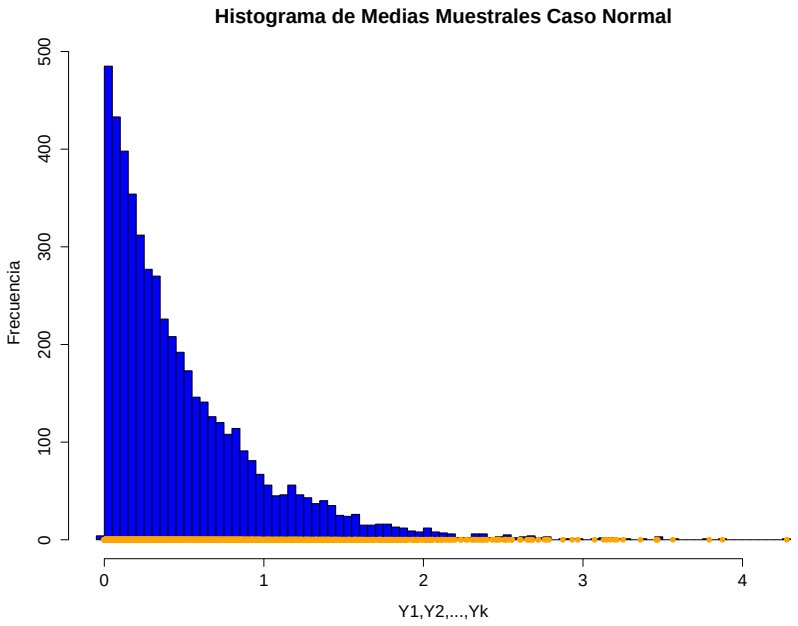

Figura 2.7: Histograma de las medias muestrales de $k = 5000$ poblaciones normales con medias verdaderas simuladas

Un microarreglo es un chip que permite a los genetistas analizar grandes cantidades de material genético de manera simultánea. En particular, un microarreglo permite estudiar los niveles de expresión genética de cada uno de los genes de un organismo en particular. El estudio de los niveles de expresión genética es importante pues permite identificar los genes con mayor impacto en procesos biológicos de interés en diversos organismos como el desarrollo de enfermedades, el crecimiento, la longevidad, *etc.* Dado que el número de genes de un organismo puede estar en el orden de $k \approx 10^3$, las bases de datos de expresión genética frecuentemente son grandes. Un proceso de filtrado en el cual se retengan únicamente aquellos genes que *mejor* contribuyan a un proceso biológico de interés para llevarlos a estudios más especializados frecuentemente es deseable. Es aquí donde se encuentran las bases del problema general de selección.

Ejemplo 2.3.2 (Procesamiento de Imágenes)**:**
En informática, la representación digital de una imágen es a través de un arreglo de pixeles en una cierta escala de colores. Con el surgimiento de nuevas tecnologías de procesamiento de información, el procesamiento de imágenes como dato ha tomado auge. Frecuentemente, en diversos contextos, es de interés identificar características de interés a través de imagenes: defectos de producción, cambios de temperatura, hábitat de especies animales, formación de tumores, aspectos biométricos, etc, por lo que es necesario seleccionar, de un conjunto grande de k pixeles, seleccionar aquellos que mejor representen la característica de interés.

Cui and Wilson [2008] identificaron por primera vez que las nociones clásicas de RSM no podían ser extendidas de manera natural al caso en el que k es grande. Por lo cual, redefine las nociones principales de selección y ordenamiento y propone nuevas metodologías diseñadas específicamente para estos casos con un enfoque principal en genética. Esta sección constituirá, por tanto, una reseña de los principales aportes de Cui and Wilson [2008] y Cui et al. [2010] a la teoría de RSM de manera que el lector pueda construir un puente entre las metodologías clásicas y modernas de esta disciplina.

2.3.2. Notación, Definiciones y Supuestos

Siguiendo la notación introducida en la Sección 2.1 (véase el Cuadro 2.3), partiremos de la situación de que se tienen observaciones independientes $X_{ij}, j = 1, 2, ..., n$ de las poblaciones $\pi_i, i = 1, 2, ..., k$ donde $X_{ij} \sim G(x - \theta_i)$. El interés principal es seleccionar las t mejores poblaciones donde *mejores* hace referencia a aquellas cuya distribución corresponde a los parámetros más grandes $\theta_{(k-t+1)}, ..., \theta_{(k)}$. Para ello se obtiene el conjunto de estadísticas resumen $Y_i \sim F(y - \theta_i), i = 1, 2, ..., k$ que se ordena como en (2.26). La regla de decisión consistirá en elegir a las t poblaciones que produjeron las mayores estadísticas resumen como las t mejores. El instrumento para medir la incertidumbre asociada a esta elección o PCS consiste la base de la teoría de RSM (véase 2.37).

Cui and Wilson [2008] puntualizaron tres tipos de supuestos base para el desarrollo de la extensión de la teoría de RSM al caso de k grande:

1. **Sobre la independencia**. Todos los resultados analíticos expuestos en Cui and Wilson [2008] y Cui et al. [2010] están basados en la independencia tanto entre observaciones como entre poblaciones. Cui and Wilson [2008] cita el estudio de robusticidad ante dependencia de poblaciones como uno de los problemas abiertos de mayor prioridad en su línea de investigación.

2. **Sobre el modelo**. La importancia de la suposición de un modelo de localización para los datos radica en que es deseable la propiedad de que transformaciones del tipo $\theta_i + c, i = 1, 2, ..., k$ no alteren el ordenamiento natural de las poblaciones π_i.

3. **Sobre la varianza**. Sea σ_i^2 la varianza de las observaciones de $\pi_i, i = 1, 2, ..., k$. La situación ideal en la metodología de Cui and Wilson [2008] indica que $\sigma_i^2 = \sigma^2$ para todo $i = 1, 2, ..., k$ con σ^2 conocida.

El supuesto (1) se verifica, por lo general, a través del proceso de muestreo, asumiendo que las observaciones se toman de manera completamente aleatorizada es razonable asumir independencia. Sin embargo, en muchos contextos no se puede garantizar la independencia entre poblaciones. En genética, por ejemplo, donde es conocido que existe un comportamiento agrupado entre genes es altamente posible que exista correlación positiva (y negativa) entre los niveles de expresión de grupos de genes, lo cual no cumpliría completamente con el supuesto (1). En relación al supuesto (2), Cui and Wilson [2009] condujeron un estudio de simulación para estimar el efecto del error de especificación del modelo. Por lo general la mayoría de los resultados resultaron ser robustos bajo condiciones específicas; los detalles pueden consultarse en dicha publicación.

En el caso de que la varianza se asuma común pero sea desconocida (como en la mayoría de las aplicaciones en la práctica) un procedimiento estándar es utilizar una estimación conjunta dada por

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{k}(n_i - 1)s_1^2}{\sum_{i=1}^{k}(n_i - 1)}, \tag{2.70}$$

donde s_i^2 y n_i son la varianza muestral y el números de observaciones disponibles de la población π_i. En el caso en que $n_i = n$ para $i = 1, 2, ..., k$, (2.70) se reduce al promedio de las varianzas muestrales.

Si la varianza no se puede asumir común entre las poblaciones una alternativa razonable es la aplicación de una transformación apropiada tal y cómo suele aplicarse en otros contextos como análisis de regresión y análisis de varianza. La Figura 2.8 (panel izquierdo) ejemplifica un conjunto de datos simulado de $k = 50$ poblaciones normales con distintas medias en las cuales se puede apreciar que existe una clara tendencia ascendente en el gráfico de medias vs varianzas. Dicho escenario es relativamente común en la práctica, en particular en observaciones relacionadas con la escala temporal (tiempo a la falla, tiempo de funcionamiento, tiempo de vida, e*tc.*) , ya que mientras mayores sean las observaciones mayor será la incertidumbre asociada a ellas. El panel derecho muestra el efecto de pre-transformar los

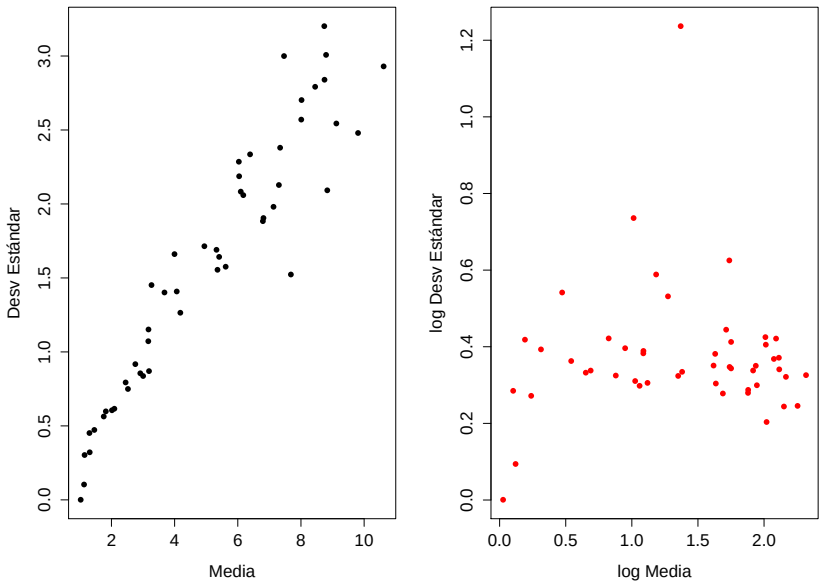

Figura 2.8: Ejemplo del efecto de una transformación a un conjunto de datos simulado.

datos mediante la transformación $h(x) = \log(x)$, puede apreciarse que dicha transformación tiene un efecto *estabilizador* de varianza.

La elección de una transformación adecuada $h(\cdot)$ no es trivial y depende, en general, de la estructura de los datos. Bartlett [1947] estudió la aplicación de transformaciones estabilizadoras de varianza y estableció que si una muestra aleatoria $X_1, X_2, ..., X_k$ cumple que

$$\mathrm{SD}(X_i) = g(\mathrm{E}(X_i)), \forall i = 1, 2, ..., k, \tag{2.71}$$

entonces la transformación

$$h(x) = \int \frac{1}{g(u)} du \tag{2.72}$$

satisface que $\mathrm{V}(h(X_i))$ es aproximádamente constante para $i = 1, 2, .., k$.

Ciertas precauciones deben ser consideradas antes de aplicar una transformación en un problema de selección y ordenamiento. Para ilustrarlo tomemos como ejemplo el caso de la Figura 2.8, que consiste de un conjunto de datos simulado donde es de particular interés hacer inferencia sobre el ordenamiento de las medias $\mu_i, i = 1, 2, ..., k$. Al aplicar la transformación $h(x) = \log(x)$ se obtiene un nuevo conjunto de datos $Z_{ij} = \log(X_{ij}), j = 1, 2, ..., n$ que se puede asumir que tiene varianza constante. Sin embargo, es importante investigar si la inferencia que hagamos sobre el ordenamiento de las medias del conjunto de datos Z_{ij} puede interpretarse en términos del conjunto de datos original X_{ij}, es decir, si el ordenamiento de las medias, en este caso, es *invariante* ante la transformación elegida.

La siguiente proposición indica que, para el caso de un modelo de localización, podemos garantizar invarianza en el ordenamiento ante cualquier transformación monótona creciente.

Proposición 1. *Sean X_1 y X_2 variables aleatorias independientes con función de distribución $F(x-\theta_i)$ y sea $Z_i = h(X_i), i = 1, 2$ una transformación monótona creciente. Luego, si $\theta_i < \theta_2$ entonces $E(Z_1) < E(Z_2)$ y por el contrario, si $\theta_1 > \theta_2$ entonces $E(Z_1) > E(Z_2)$.*

Demostración. Sin pérdida de generalidad supongamos que $\theta_1 > \theta_2$; la demostración es análoga para el caso contrario. Dado que $X_i \sim F(x - \theta_i), i = 1, 2$ entonces es posible escribir

$$X_i = X + \theta_i, \quad \text{para} \quad i = 1, 2, \tag{2.73}$$

donde X es una variable aleatoria tal que $\mathrm{E}(X) = 0$ y $X \sim F(x)$.

Se tiene entonces que,

$$\begin{aligned} E[Z_1] &= E[\log(X_1)] \\ &= E(\log(X + \theta_1)) \\ &= \int_{\mathbb{R}^+} \log(x + \theta_1) dF(x) \\ &> \int_{\mathbb{R}^+} \log(x + \theta_2) dF(x) \\ &= E(\log(X + \theta_2)) \\ &= E[\log(X_1)] \\ &= E[Z_2]. \end{aligned}$$

□

Es importante aclarar que la proposición anterior no garantiza que al aplicar una transformación $h(\cdot)$ a un conjunto de datos bajo un modelo de localización obtendremos un nuevo conjunto de datos que se rigen bajo un modelo de localización. Sin embargo, dado que el ordenamiento de los parámetros es invariante, la proposición anterior garantiza que mediante una transformación monótona creciente se podría obtener un nuevo problema transformado, en el cual, hacer inferencia sobre el ordenamiento de las medias de las variables transformadas Z_{ij} tiene una interpretación en términos del ordenamiento de los parámetros originales θ_i.

Un conjunto conocido de transformaciones monótonas crecientes es la llamada Familia de Transformaciones de Box-Cox (Box and Cox [1964]) que se definen para $x_i > 0$ como

$$h(x_i, \lambda) = \begin{cases} \dfrac{x_i^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0, \\ \ln(x_i) & \text{si } \lambda = 0, \end{cases} \tag{2.74}$$

donde λ es un parámetro real que controla la forma particular de la transformación.

La Figura 2.9 muestra las transformaciones de Box-Cox para distintos valores del parámetro λ. Es importante notar que cualquier transformación de Box-Cox es monótona creciente por lo que cumple con las hipótesis de la Proposición 1. Si $\lambda \geq 0$ se obtendrá una transformación cóncava que puede demostrarse (véase Bartlett [1947]) que estabilizará las varianzas en los casos en los que estas crecen con las medias (como en la Figura 2.8). Si por el contrario las varianzas decrecen con las medias una transformación convexa, obtenida mediante $\lambda < 0$, podría ser una mejor opción. En general el parámetro λ se elige de acuerdo con el contexto; en la práctica se puede hacer de manera informal mediante un gráfico exploratorio de medias *vs.* varianzas o bien con un razonamiento analítico más formal que involucra la verosimilitud perfil de λ como se explica en Box and Cox [1964].

2.3.3. Extensión de las Nociones de PCS

Cui and Wilson [2008] definen tres nuevas metodologías de selección específicamente conceptualizadas para los casos en los que k es grande. Dichos procedimientos se basan en maneras alternativas de calificar si las poblaciones que se escogen como las *mejores* en una determinada selección en realidad son las *mejores*.

Ejemplo 2.3.3:
Para ilustrar las tres metodologías que se definirán a continuación se hará uso de la siguiente tabla que consiste de un conjunto simulado de $k = 6$ poblaciones con sus respectivos parámetros verdaderos y las estadísticas resumen obtenidas de la muestra.

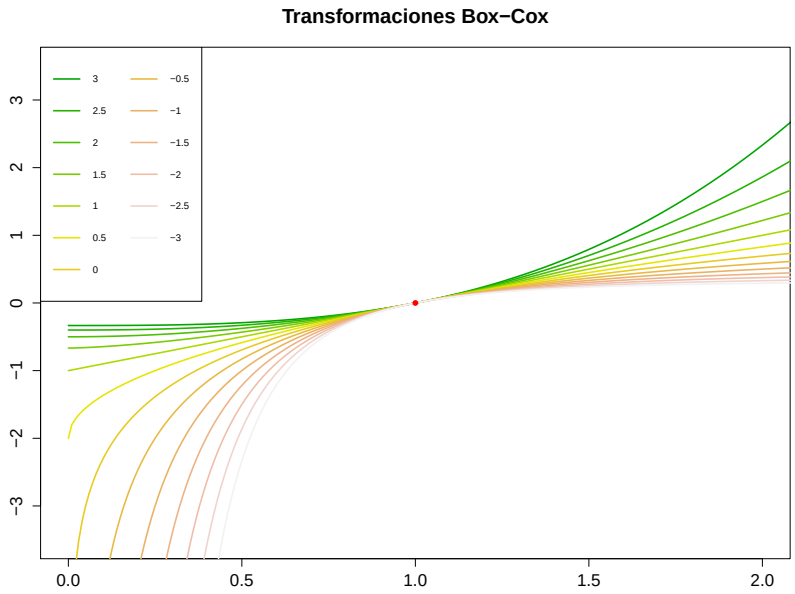

Figura 2.9: Transformación de Box-Cox para distintos valores del parámetro λ.

Población	1	2	3	4	5	6
Parámetro verdadero	$\theta_{(2)} = 1{,}23$	$\theta_{(5)} = 3{,}32$	$\theta_{(3)} = 2{,}85$	$\theta_{(1)} = 1{,}06$	$\theta_{(6)} = 4{,}37$	$\theta_{(4)} = 2{,}93$
Estadística Resumen	$Y_{[2]} = 1{,}37$	$Y_{[6]} = 4{,}05$	$Y_{[4]} = 2{,}21$	$Y_{[1]} = 1{,}23$	$Y_{[5]} = 3{,}54$	$Y_{[3]} = 1{,}90$

$G-$Selección

La $G-$selección está diseñada para aquellos casos en los que al experimentador le interesa encontrar las t mejores poblaciones de un conjunto, pero tiene recursos suficientes para estudiar hasta G poblaciones extra para encontrarlas. En otras palabras, el experimentador escogerá un número fijo $t+G$ de poblaciones, de acuerdo con los valores de las $t+G$ mayores estadísticas resumen, y calificará la selección como correcta si y sólo si ésta contiene las verdaderas t mejores. Formalmente, la definición es como sigue (Cui and Wilson [2008]):

Definición 1. *Sea s el conjunto de los índices de las $t+G$ mayores estadísticas resumen y sea A_t el conjunto de los índices de las t poblaciones que corresponden a los t verdaderos mayores parámetros. El evento de selección correcta se redefine en términos de la $G-$selección como:*

$$CS_{G,t} = \{A_t \subset s\}. \tag{2.75}$$

Todo conjunto s que cumple con (2.75) es llamado $G-$mejor y la probabilidad de encontrar un conjunto de índices con dicha característica se denota por $PCS_{G,t}$ (véase la Figura 2.10).

La interpretación del parámetro G es completamente intuitiva pues representa el número de poblaciones extra o *falsos positivos* que el experimentador está dispuesto a elegir con el fin de encontrar a las t mejores entre ellas. El cociente $r = t/(t+G)$ representará la proporción de los verdaderos mejores que la selección contiene. La característica distintiva de la $G-$selección sobre las otras metodologías es que le ofrece al experimentador la posibilidad de controlar el tamaño de la selección que se hará en función del número de verdaderas mejores poblaciones que desea encontrar (G).

Naturalmente, sería deseable que G fuera lo más pequeña posible, o equivalentemente, que r fuera lo más cercano a uno posible. Sin embargo, dependiendo del contexto, esto no podría ser permisible si se desea obtener una alta $\text{PCS}_{G,t}$. Si se hace $G = 0$ el contexto de $G-$selección se colapsa al escenario original de la metodología clásica de la Sección 2.2 (véase la Figura 2.11). Sin embargo, hacer ésto podría causar un descenso en la $\text{PCS}_{G,t}$, en particular en los casos en los que k es grande. Por ejemplo, seleccionar exactamente a las mejores $t = 100$ poblaciones de un conjunto de $k = 5000$ es altamente improbable (véase la Figura 2.7). Si en cambio, para ese mismo caso, se escoge $t = 90$ y $G = 10$, se estaría admitiendo una selección de *menor* calidad (solo 90 % de las poblaciones elegidas serían verdaderamente las mejores), a cambio de una $\text{PCS}_{G,t}$ más alta.

En el caso del ejemplo 2.3.3, si se toma $t = 1$ y $G = 1$ se seleccionarían $t + G = 2$ poblaciones que formarían el conjunto de índices $s = \{2, 5\}$; esto resultaría en una selección correcta pues s contiene a la verdadera mejor población ($A_t = \{5\}$). Notemos que si se usara la metodología clásica ($G = 0$) se habría obtenido una selección incorrecta, lo cual es evidencia de la ventaja de utilizar $G-$selection para alcanzar una más alta PCS. Cui and Wilson [2008] explora e investiga los detalles técnicos de la demostración de este hecho.

El concepto de $G-$selección es similar al contexto de selección por subconjunto (SS) de Gupta [1956] introducido en la Sección 2.2.3, en el sentido que ambas metodologías seleccionan un subconjunto de poblaciones y admitirán como correcta dicha selección si y sólo si ésta contiene las verdaderas mejores t poblaciones. Sin embargo existen dos diferencias fundamentales que los diferencian:

- SS selecciona un subconjunto de tamaño aleatorio (en ocasiones incontrolablemente grande) mientras que $G-$selección toma un subconjunto de tamaño fijo $t + G$.
- SS requiere el valor prefijado de la mínima PCS deseada P^* (preferiblemente grande), mientras que $G-$selección la calcula o en su defecto la estima con base en los datos y las estadísticas resumen obtenidas.

Para los casos en los que k es grande, prefijar una P^* grande podría ocasionar que el subconjunto seleccionado sea demasiado grande. Mientras que $G-$selección podría adaptarse a dicha situación ajustando el tamaño fijo del subconjunto seleccionado de manera que se pueda alcanzar una PCS considerablemente grande. En el Capítulo 4 se presentará un caso de estudio en el cual se aprecia el efecto del parámetro G y la utilidad de la $G-$selección como herramienta de filtrado.

$d-$selección

La $d-$selección está pensada para aquellos casos en los cuales es de interés seleccionar poblaciones dentro de un cierto *rango* de calidad. A diferencia de la $G-$selección, el número de poblaciones escogidas por $d-$selección es aleatorio y consiste de poblaciones que están a lo más a una distancia d de las verdaderas mejores t poblaciones. Dicha distancia es controlada mediante un parámetro $d > 0$. La definición formal como aparece en Cui and Wilson [2008] es:

Definición 2. *Sea s el conjunto de índices correspondientes a las t mayores estadísticas resumen. Para $d \geq 0$, sean A_{t1} y A_{t2} los conjuntos de índices correspondientes a los parámetros verdaderos en los intervalos $(\theta_{(k-t+1)} + d, \theta_{(k)}]$ y $[\theta_{(k-t+1)} - d, \theta_{(k-t+1)} + d]$ respectivamente (véase la Figura 2.12). El evento de selección correcta en términos de la $d-$selección se define como*

$$ {}_dCS_t = \{A_{t1} \subset s \;\; y \;\; (s - A_{t1}) \subset A_{t2}\}, \tag{2.76}$$

donde $A - B = \{x : x \in A, x \notin B\}$. Todo conjunto que satisface (2.76) es llamado $d-$mejor y la probabilidad de seleccionar un conjunto con esta característica se denota por $P({}_dCS_t)$.

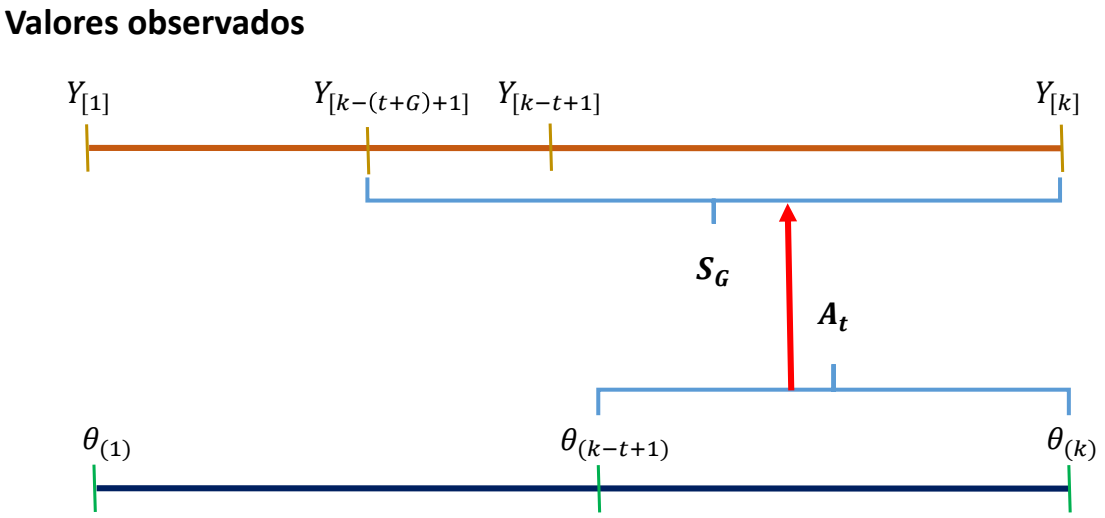

Figura 2.10: Esquema de la $G-$selección.

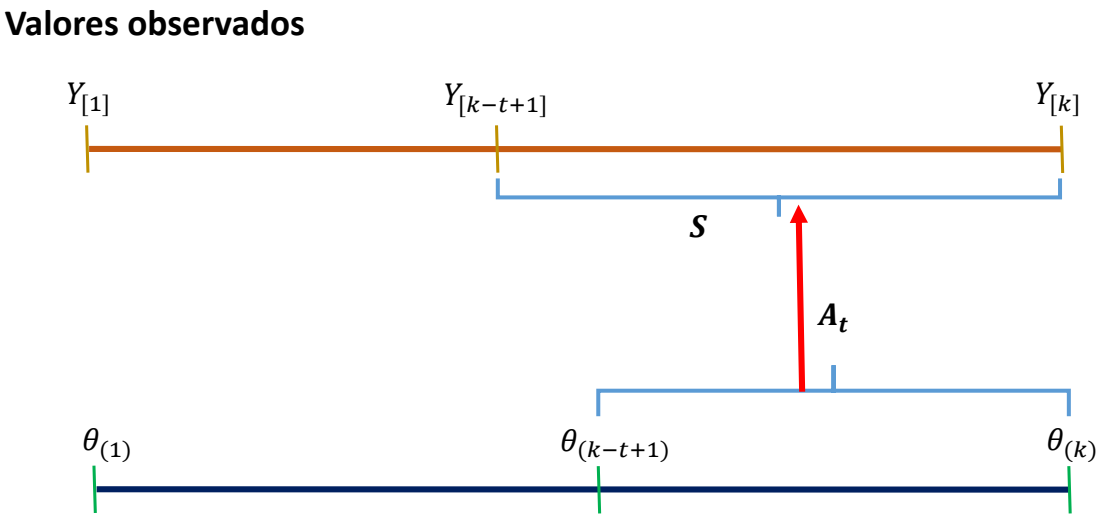

Figura 2.11: Esquema de la selección clásica como se concibió en Bechhofer [1954] y Gupta [1956].

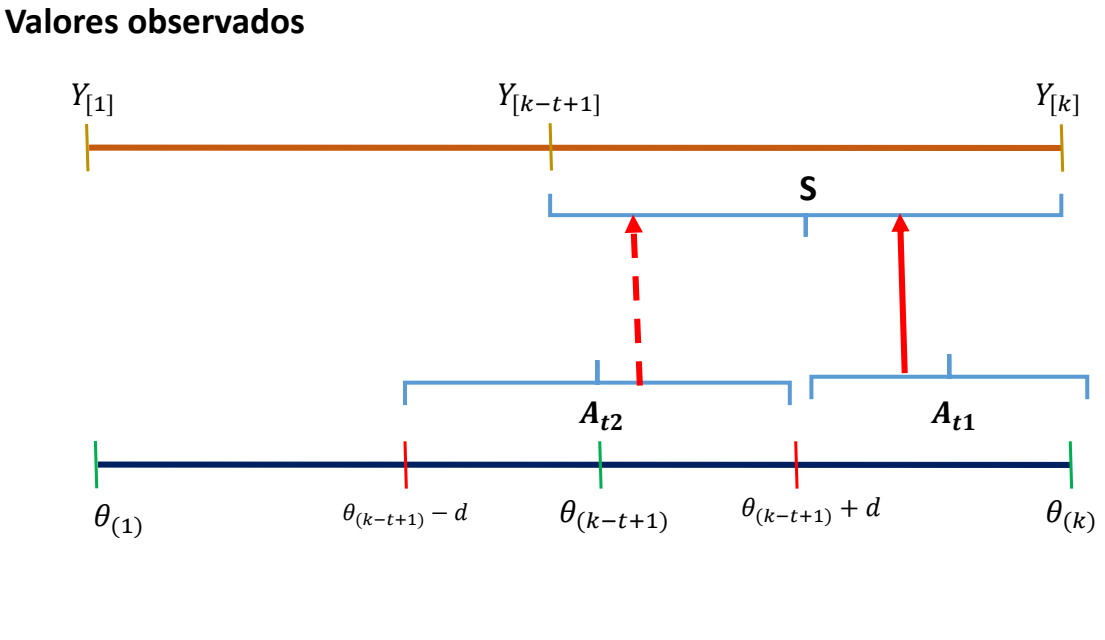

Figura 2.12: Representación gráfica de la $d-$selección. La flecha en línea contínua indica contención completa y la flecha en línea punteada implica contención parcial.

El parámetro d se interpreta como el ancho del intervalo que se obtiene alrededor del parámetro $\theta_{(k-t+1)}$ donde es permisible escoger poblaciones extra o *falsos positivos*. Es posible ver de esta manera que el parámetro d induce un análogo a la zona de indiferencia de Bechhofer [1954], donde el intervalo A_{t2} se puede considerar como una IZ alrededor de $\theta_{(k-t+1)}$ donde consideramos cualquier población igualmente *digna* de ser seleccionada.

La $d-$selección viene con dos limitantes importantes en comparación con la $G-$selección. La primera es el hecho de que el subconjunto de poblaciones elegidas siempre es de tamaño aleatorio, ya que, dependiendo de la elección de d dicho subconjunto puede ser arbitrariamente grande en algunos contextos. La segunda limitante radica en que la elección del parámetro d no es trivial y no siempre es intuitiva. El parámetro d se elige idealmente de acuerdo con el contexto en la escala real del espacio parametral y representa el tamaño del intervalo A_{t2} dentro del cual se permite escoger un número aleatorio de poblaciones como parte de las *mejores*. Lamentablemente, en la práctica, no siempre existe o se conoce tal noción, de manera que se traduzca a predeterminar un parámetro d que desempeñe esas funciones. En el caso de aplicación del Capítulo 4 se presentará una situación de este tipo.

Si se hace $d = 0$ se obtendría el escenario original de la selección clásica de Bechhofer [1954]. Idealmente se desearía que el parámetro d fuera lo más pequeño posible pues ésto reduciría potencialmente el tamaño del subconjunto seleccionado. Sin embargo, en particular para casos en los que k es grande, esto causaría un descenso en la PCS alcanzable (Cui and Wilson [2008]). Esto se puede ilustrar mediante el Ejemplo 2.3.3. Para este caso, si se fija $t = 3$ y $d = 0{,}5$ se obtiene que $s = \{2, 3, 5\}$, y A_{t1} quedaría formado por los índices de las poblaciones cuyos parámetros están en el intervalo $(\theta_{(4)}+0{,}5, \theta_{(6)}] = (3{,}43, 4{,}37]$, entonces $A_{t1} = \{5\}$. Similarmente, A_{t2} consiste de los índices de las poblaciones cuyos parámetros están en el intervalo $[\theta_{(k-t+1)} - d, \theta_{(k-t+1)} + d] = [\theta_{(4)} - 0{,}5, \theta_{(4)} + 0{,}5] = [2{,}43, 3{,}43]$, *i.e.* $A_{t2} = \{2, 6\}$. Así, dado que $A_{t1} = \{6\} \subset \{2, 5, 6\} = s$ y $(s - A_{t1}) = \{2, 6\} \subset \{2, 6\} = A_{t2}$, obtendríamos una selección correcta en este caso. Es importante notar que, para este mismo ejemplo, si se hubiera elegido $d = 0$ se habría obtenido una selección incorrecta a pesar de que las poblaciones 3 y 6 tienen parámetros tan

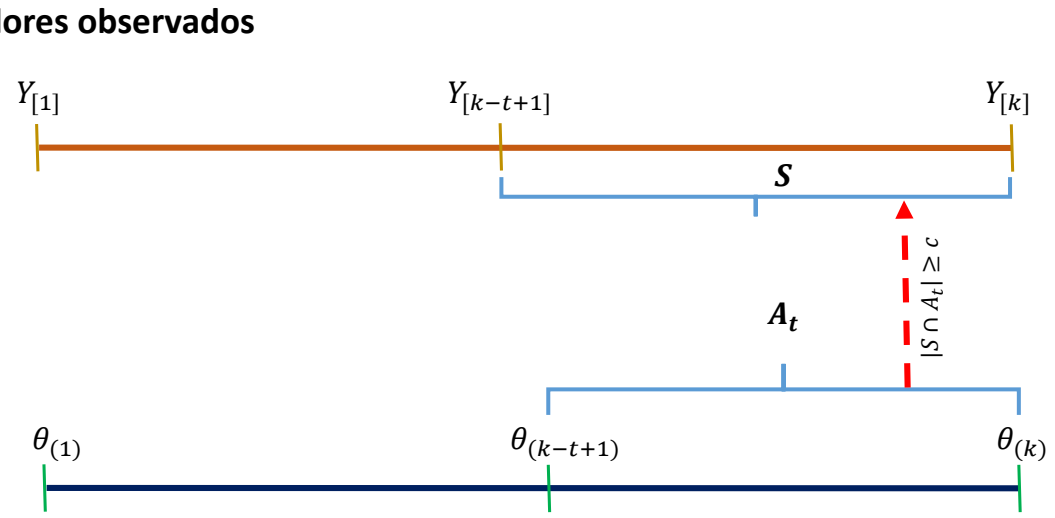

Figura 2.13: Representación gráfica de la $c-$selección. La flecha en línea punteada implica contención parcial bajo la condición indicada.

parecidos que son prácticamente iguales (en términos del parámetro $d = 0{,}5$). Esto da evidencia de que la situación ideal para utilizar la $d-$selección es cuando se desea seleccionar las t mejores poblaciones y una diferencia entre parámetros de d unidades o menos puede ser considerada irrelevante.

$c-$Selección

La $c-$selección (Cui et al. [2010]) surge como una alternativa generalizada de la $G-$selección. En ese panorama, el interés está en seleccionar las t mejores poblaciones de un cierto conjunto y una selección dada se calificará como correcta si contiene al menos una cantidad fija $c \leq t$ de las verdaderas mejores poblaciones.

Definición 3. *Sea s el conjunto de los índices correspondientes a las t estadísticas resumen más grandes y sea A_t el conjunto de índices correspondientes a las poblaciones con los parámetros verdaderos más grandes. Sea además c una constante positiva tal que $1 \leq c \leq t$, el evento de selección correcta se redefine en términos de la $c-$selección (Figura 2.13) como:*

$$CS_{c,t} = \{s : |s \cap A_t| \geq c\}, \tag{2.77}$$

donde $|\cdot|$ indica la cardinalidad de un conjunto. Todo conjunto s que satisface (2.77) es llamado $c-$mejor y la probabilidad de encontrar un conjunto con dicha característica se denota por $P(CS_{c,t})$.

El parámetro c tiene la interpretación directa de ser el número mínimo de elecciones correctas que se desea tener en la selección hecha para ser calificada como buena. Si se define,

$$r^* = c/t, \tag{2.78}$$

se obtendrá una medida estandarizada de la *calidad* de una $c-$selección. Por ejemplo $r^* = 0{,}75$ indica que al menos 75 % de las poblaciones seleccionadas son verdaderas mejores, y esto se interpreta de la misma forma en todos los contextos independientemente del valor de c o de t.

La $c-$selección y la $G-$selección son similares en el sentido que ambas definen un parámetro que permite controlar el número neto de *errores* o *falsos positivos* que se permite entrar en una selección que será calificada como correcta. En el caso de la $G-$selección se definió el cociente $r = t/(t+G)$ como la proporción de poblaciones verdaderamente correctas en una $G-$selección y similarmente se hizo para la $c-$selección en (2.78). Sin embargo, incluso haciendo $r = r^*$, la $c-$selección y la $G-$selección son sutilmente distintas y se puede demostrar que la $c-$selección ofrece una mayor probabilidad de selección correcta. Por ejemplo, para un cierto conjunto de k poblaciones, las expresiones $P(\mathrm{CS}_{G=1,t=4})$ y $P(\mathrm{CS}_{c=4,t=5})$ no son equivalentes, la primera representa la probabilidad de elegir un subconjunto de $t+G = 5 < k$ poblaciones que contenga *exactamente* las 4 mejores poblaciones, mientras que la segunda representa la probabilidad de encontrar un subconjunto de $t = 5 < k$ poblaciones que contenga *al menos* 4 de las mejores 5 poblaciones.

Para ilustrar la situación anterior supongamos que se tiene un conjunto de $k = 10$ poblaciones etiquetadas $A, B, ..., J$ y, sin pérdida de generalidad supongamos que el orden de mejor a peor coincide con el orden alfabético, es decir la mejor población es A, la segunda mejor es B y así sucesivamente. Un conjunto $G-$mejor para $G = 1$ y $t = 4$ tiene entonces la forma

$$s = \{A, B, C, D, x\}, \tag{2.79}$$

donde $x = E, F, ..., J$. Es decir, para este caso existen únicamente 6 conjuntos $G-$mejores. Mientras tanto, un conjunto $c-$mejor para $c = 4$ y $t = 5$ contiene todos los conjuntos que cumplen con (2.79) y algunos más, por ejemplo los conjuntos $\{A, C, D, E, K\}$, $\{A, B, D, E, H\}$, *etc.* Luego es fácil ver que para este caso $P(\mathrm{CS}_{G=1,t=4}) \leq P(\mathrm{CS}_{c=4,t=5})$ a pesar de que $r = t/(t+G) = 4/5 = c/t = r^*$.

Selección sin t específica

En muchos contextos en los cuales un problema de selección está involucrado, la meta general es examinar un conjunto de k poblaciones (genes, tratamientos industriales, variedades de grano, *etc.*) y seleccionar cualitativamente algunas de ellas para investigación adicional específica. Sin embargo, en muchas aplicaciones, el número de poblaciones a seleccionar t es desconocido. Aún así, de acuerdo con conocimientos previos o restricciones económicas o temporales es posible conocer el número mínimo y máximo de poblaciones que se desea elegir, denotados por $T_{\min}$ y $T_{\max}$ respectivamente. Cui et al. [2010] proponen dos metodologías diseñadas para encontrar el valor óptimo t, con un rango predefinido $T_{\min} \leq t \leq T_{\max}$, de tal manera que se obtenga una selección con al menos $r^* = c/t \times 100\,\%$ de las verdaderas mejores poblaciones. El valor r^* será llamado a partir de ahora *razón de selecciones correctas* (CSR).

El primer procedimiento, conocido como r^*-selección, se basa en encontrar el valor de t que optimiza la PCS para una cierta CSR pre-especificada. En otras palabras, se encontrará aquel valor de t que produce una mayor PCS cuando el criterio para que una selección sea correcta se fija en que ésta contenga al menos $r\,\%$ poblaciones correctas.

1. Fijar la CSR r^*, $0 < r^* < 1$.

2. Calcular
$$P_t = P(\mathrm{CS}_{c=\lceil r^* t \rceil, t}) \tag{2.80}$$
para $T_{\min} \leq t \leq T_{\max}$, donde $\lceil \cdot \rceil$ es la función techo.

3. Seleccionar el valor de t que produce la mayor P_t.

Es posible modificar el procedimiento anterior para hacer $G-$selección o $d-$selección simplemente fijando un valor G^* (o d^*) y modificando (2.80) de manera que se obtenga P_t como función de la PCS

para el caso que corresponda. En el caso de estudio que se presentará en el Capítulo 4 se realizará este procedimiento y se darán los detalles.

El segundo procedimiento, llamado P^*-selección, funciona de manera inversa al procedimiento anterior. En este algoritmo se requiere que el experimentador pre-especifique un valor de P^*, la mínima PCS que desea obtener en su selección, y de esta manera encuentre el valor de t que optimice la CSR. Es decir, aquél valor de t que produzca el menor número de errores con una probabilidad de al menos P^*.

1. Fijar P^*, $0 < P^* < 1$.
2. Calcular
$$r_t = (1/t) * \text{máx}\{c : P(\text{CS}_{c,t} \geq P^*) \tag{2.81}$$
para $T_{\min} \leq t \leq T_{\max}$.
3. Seleccionar el valor de t que produce la mayor r_t.

En la práctica, las expresiones para PCS en (2.80) y (2.81) se sustituyen por un estimador apropiado $\hat{\text{PCS}}$. Cui et al. [2010] implementan estos métodos para un caso de simulación y dos casos de estudio y hace uso de técnicas Bootstrap para obtener $\hat{\text{PCS}}$. La siguiente sección hará una reseña sobre la estimación de PCS para las metodologías de selección introducidas anteriormente. La r^*-seleccion y la P^*-selección constituyen herramientas novedosas y útiles pues contienen elementos de optimización, Cui et al. [2010] encontró además que pueden ser útiles como herramientas diagnóstico que permiten visualizar la estructura del conjunto de datos y detectar cuando una selección de buena calidad simplemente no puede realizarse con alta confianza. En el Capítulo 4 se presentarán casos particulares que permitirán ilustrar esta propiedad.

En general las nuevas metodologías de selección introducidas en Cui and Wilson [2008] y Cui et al. [2010] presentan una mejoría considerable respecto de la teoría clásica de Bechhofer [1954] y Gupta [1956] para los casos en los que k es grande. Entre las propiedades deseables de mayor interés destaca el hecho de que PCS no se va a cero conforme k aumenta; esta característica se regula de manera intuitiva mediante los parámetros G, d y c que cuando se aplican de manera conjunta permiten además obtener un panorama general de la estructura de los datos.

2.3.4. Selección de las peores poblaciones

Como se discutió en la Sección 2.2 en muchos contextos es de especial interés no sólo cuáles son las mejores poblaciones de un conjunto sino también las peores, donde *peor* se entiende por aquellas poblaciones que minimizan alguna característica medible de interés, preferentemente que se pueda traducir al valor de un determinado parámetro θ en un modelo de probabilidad.

En el caso de la selección clásica de la Sección 2.2 resulta intuitivo que si se desean identificar las *peores* t poblaciones de un conjunto de tamaño k basta con encontrar las mejores $k - t$. Esto se debe a que toda aseveración probabilística que se haga acerca de las $k - t$ mejores poblaciones tiene su correspondiente interpretación en las t peores. Por ejemplo, seleccionar las mejores $t = 5$ de un conjunto de $k = 15$ de ellas y obtener una P^* de al menos 90 % es equivalente a seleccionar las peores $k - t = 10$ poblaciones bajo el mismo requerimiento probabilístico.

Para el caso de las nuevas metodologías de selección introducidas en Cui and Wilson [2008] este no necesariamente es el caso. Para ilustrar esto tomemos como ejemplo la Figura 2.14 que es la representación gráfica de una configuración de $k = 100$ parámetros θ_i ordenadas de mayor a menor, donde las cantidades escritas sobre cada círculo indican el número de parámetros acumulados en dicha zona. Supongamos además que es de interés encontrar los $t = 90$ mejores genes y se aplica $c-$selección con

Figura 2.14: Representación gráfica de una configuración de $k = 100$ parámetros ordenados de mayor a menor.

$r^* = c/t = 0{,}90$, i.e. toda selección con al menos $90 * (0{,}90) = 81$ de las mejores poblaciones será considerada como correcta. Es intuitivo ver que para este caso encontrar un conjunto con dichas características será relativamente *fácil* ya que, sin importar qué tan cercanas estén las poblaciones en el círculo del centro, las 85 poblaciones en el círculo de la izquierda garantizan que casi toda selección de tamaño $t = 90$ incluirá al menos 85 de las mejores poblaciones. Tal no es el caso si se desea seleccionar las peores 10 poblaciones, para que una selección de este tipo sea correcta debe contener al menos $9 * (0{,}9) \approx 9$ de las verdaderas peores poblaciones. Por esta razón, si se desea una selección correcta para este caso será necesario identificar las 4 peores de las 10 poblaciones del círculo del centro lo cual es naturalmente difícil.

Para éstos casos se recomienda en general utilizar la transformación $Z_{ij} = -X_{ij}$ y transformar directamente el problema de seleccionar las peores t poblaciones con base en X_{ij} a seleccionar las mejores t poblaciones con base en Z_{ij}.

2.3.5. Estimación de PCS para k grande

La fórmula explícita para $P(\mathrm{CS}_{G,t})$, $P({}_d\mathrm{CS}_t)$ y $P(\mathrm{CS}_{c,t})$ es la misma excepto que los cálculos se harán de manera distinta dependiendo de la definición de conjunto $G-$mejor, $d-$mejor y $c-$mejor respectivamente. La fórmula, propuesta por Cui and Wilson [2008] como una extensión de la obtenida por Bechhofer [1954], está dada por:

$$\sum_{g=1}^{|S|} \int_{-\infty}^{\infty} \prod_{j=k-t+1}^{k} \bar{F}(y - \theta_{s_{g,j}}) d\{\prod_{j=1}^{k-t} F(y - \theta_{\bar{s}_{g,j}})\}, \tag{2.82}$$

donde S denota el conjunto de todos los conjuntos $G-$mejores, $d-$mejores o $c-$mejores según corresponda. Aquí $\boldsymbol{\theta_g}$ representa la $g-$ésima configuración de parámetros $\theta_{(i)}, i = 1, 2, ..., k$. Este es un conjunto de conjuntos que se particiona en dos $\boldsymbol{\theta_g} = \{\boldsymbol{\theta_{\bar{s}_g}}, \boldsymbol{\theta_{s_g}}\}$, donde $\bar{s}_g \in \bar{S}$ son los conjuntos que NO satisfacen la condición para ser $G-$mejores, $d-$mejores o $c-$mejores según corresponda y $s_g \in S$ los que si la satisfacen. La función $F(y - \theta_{\bar{s}_{g,j}})$ es la distribución de las estadísticas resumen Y. El procedimiento para la obtención de la fórmula (2.82) puede consultarse en Cui and Wilson [2008]. Nótese que la fórmula (2.82) puede ser complicada de trabajar analíticamente en muchos casos.

Cui and Wilson [2008] se proponen además buscar un estimador para PCS, en contraste con la teoría clásica que depende en la mayoría de los casos de encontrar una configuración menos favorable (LFC) para cada caso en específico. En la Sección 2.2.2 se presentó una reseña bibliográfica de las principales aportaciones al contexto de la estimación de PCS; en particular se habló acerca de los estimadores tipo OST (2.40) y de sus posibles variantes determinadas por distintas propuestas de factores de encogimiento. Bajo un enfoque similar, Cui and Wilson [2008] proponen la siguiente extensión de la fórmula propuesta en Olkin et al. [1982] para el estimador de $P(\mathrm{CS}_{G,t})$, $P({}_d\mathrm{CS}_t)$ y $P(\mathrm{CS}_{c,t})$:

$$\sum_{g=1}^{|S|} \int_{-\infty}^{\infty} \prod_{j=k-t+1}^{k} \bar{F}(y - a\hat{\theta}_{s_{g,j}}) d\{\prod_{j=1}^{k-t} F(y - a\hat{\theta}_{\bar{s}_{g,j}})\}, \tag{2.83}$$

donde la notación es la misma que en la fórmula (2.82). Se sustituyen los parámetros desconocidos θ por sus estimadores $\hat{\theta}$ y se introduce el factor de encogimiento a.

Cui and Wilson [2009] conducen un experimento comparativo por simulación de los distintos factores de encogimiento introducidos en la Sección 2.2 para el caso de $d-$selección y $G-$selección. En particular (2.44), (2.43), (2.42) y $a_O = 1$, llamado el factor de encogimiento de Olkin et al. [1982], pues la ecuación original para el estimador de PCS propuesta en dicha publicación no tiene factor de encogimiento (véase 2.40). La conclusión general de Cui and Wilson [2009] es que no existe un factor de encogimiento óptimo para todas las situaciones pero recomienda a_O y a_M por ser los que produjeron menor error. En general, si los supuestos distribucionales se satisfacen, el factor a_O ofrece una precisición razonablemente buena para estimar $P({}_d\mathrm{CS}_t)$ y $P(\mathrm{CS}_{G,t})$. Esto contrasta con las conclusiones en la literatura acerca del estimador OST donde se determinó que para algunos contextos era altamente sesgado al estimar $P(CS_{t=1})$. Sin embargo, para los casos en los que los supuestos distribucionales no se cumplen, Cui and Wilson [2009] determinó que la estimación no era enteramente satisfactoria, por lo que expresa su creencia de que un nuevo factor de encogimiento podría mejorar la precisión de los estimadores para $P({}_d\mathrm{CS}_t)$ y $P(\mathrm{CS}_{G,t})$. Éste y otros problemas abiertos de la línea de investigación de Cui and Wilson [2008] se resumirán al final del presente capítulo.

Si bien los estimadores tipo OST ofrecen una expresión analítica cerrada, y hasta cierto punto precisa, para los casos en los que los supuestos distribucionales correspondientes se satisfacen, su implementación computacional resulta complicada, dado que hacen uso de técnicas numéricas. Cui and Wilson [2008] estudia un nuevo tipo de estimador para PCS basado en las ideas de Sohn and Kahn [1992] que consiste en aplicar técnicas bootstrap o de remuestreo.

Existen dos maneras de calcular un estimador tipo bootstrap para PCS. El bootstrap paramétrico parte del supuesto de que $X_{ij} \sim G(x-\theta_i), j = 1, 2, ..., n$ y θ_i es estimado por la estadística resumen $\hat{\theta}_i = Y_i \sim F(y-\theta_i), i = 1, 2, ...k$. La idea es entonces usar la distribución estimada $F(y-\hat{\theta}_i)$ para obtener una nueva muestra del estadístico Y_i repetidamente, de esta manera se generará un estimador para PCS mediante el siguiente algoritmo:

Fijar un número entero $B > 0$ y repetir para $b = 1, 2, ..., B$:

1. Simular una muestra Y_i^* del estadístico Y_i de $F(y-\hat{\theta}_i)$ para $i = 1, 2, ..., k$.

2. Sea $Y_{[i]}^*$ la $i-$ésima estadística simulada ordenada y $\hat{\theta}_{[i]}$ la $i-$ésima estadística observada ordenada. Verificar el criterio de selección de la $d-$selección, $G-$selección o $c-$selección según corresponda reemplazando $\theta_{(i)}$ por $\hat{\theta}_{[i]}$ (que en este caso es observable) y $Y_{[i]}$ por $Y_{[i]}^*$. Sea m_b una variable aleatoria binaria que toma el valor de 0 si la selección es correcta y 1 si la selección es incorrecta para ese valor particular de b.

El estimador bootstrap paramétrico para PCS estará dado por

$$\hat{\mathrm{PCS}} = (1/B)\sum_{b=1}^{B} m_b. \tag{2.84}$$

Una alternativa al procedimiento anterior es utilizar un procedimiento bootstrap no paramétrico donde se remuestrea directamente de la distribución empírica de los datos de cada población $X_{ij}, j = 1, 2, ..., n$ (asignando probabilidades iguales a cada dato $(1/n)$ y muestreando con reemplazo) en vez de hacerlo de $F(y-\hat{\theta}_i)$. Ésta es una alternativa más libre de restricciones pues no tiene supuestos distribucionales, sin embargo requiere un tamaño de muestra razonablemente grande para dar una buena estimación (Cui and Wilson [2009]).

En general, los estimadores tipo bootstrap resultaron ser más simples de implementar y redujeron considerablemente los tiempos de corrida. Los resultados de un estudio de simulación para el caso no paramétrico pueden consultarse en Cui and Wilson [2009] y los algoritmos están disponibles en la librería PCS en R (www. r-project.org).

2.3.6. Discusión y Problemas Abiertos

Existe un gran número de problemas abiertos en la línea de investigación de selección y ordenamiento. El primero y más importante, nombrado problema de prioridad número uno en la lista de problemas sin resolver de la RSM por Bechhofer, es la obtención de un estimador para PCS que cumpla simultáneamente los requerimientos de insesgadez, consistencia y propiedades asintóticas.

Con el surgimiento de la nueva teoría de selección y ordenamiento para k grande, el tema de estimar PCS volvió a ser retomado por Cui and Wilson [2008]. En dicha publicación se plantean dos nuevos problemas abiertos para la investigación:

- Ideas para la distribución de $\hat{\text{PCS}}$ para los casos de la $d-$selección y $G-$selección que consideraría aspectos como consistencia, sesgo y estimación por intervalo.
- Estimar PCS cuando las poblaciones no se pueden asumir independientes.

Cui and Wilson [2009] realizaron un estudio de simulación para encontrar propiedades interesantes de los estimadores de PCS propuestos en Cui and Wilson [2008]. Al encontrarse con que ningún factor de encogimiento resultó brindar un desempeño óptimo generalizable propone dos nuevos problemas abiertos:

- Encontrar un factor de encogimiento que mejore la precisión de $\hat{P}({}_d\text{CS}_t)$ y $\hat{P}(\text{CS}_{G,t})$ para algunas (o todas) las configuraciones.
- Investigar el efecto de no tener el supuesto de varianza constante entre poblaciones y datos correlacionados.
- Seleccionar t óptimo cuando no se conoce *a priori*.

Finalmente, Cui et al. [2010] responde al tercer problema abierto planteado en Cui and Wilson [2009] y desarrolla dos algoritmos para encontrar t, cuando éste no se conoce, utilizando argumentos de óptimización: la P^*selección y la r^*selección. Posteriormente bajo un caso de simulación y dos de aplicación se plantea como nuevo reto mejorar la estimación del vector $\boldsymbol{\theta}$ e investigar los efectos de distintas técnicas para hacerlo.

A lo largo del presente capítulo se presentaron los principales aspectos técnicos relacionados con las dos técnicas principales utilizadas en la literatura para resolver un problema de selección, las pruebas de hipótesis y las metodologías de selección y ordenamiento (RSM). La pregunta natural de investigación es «¿En qué escenarios es más conveniente la utilización de una técnica sobre la otra y por qué?» La respuesta a dicha cuestión involucra un replanteamiento crítico de las preguntas que cada una de las metodologías es capaz de responder y los posibles modos en que se pueden interpretar las soluciones que son capaces de ofrecer en el contexto del problema de selección. En el siguiente capítulo se presenta un ensayo crítico cuyos objetivos son puntualizar y contrastar las semejanzas y diferencias de ambas técnicas, con la intención de que se pueda esclarecer el debate que se plantea sobre su utilización.

Capítulo 3

Comparación Crítica

En el Capítulo 2 se presentaron las principales nociones de las pruebas de hipótesis. Se hizo especial énfasis en las pruebas o comparaciones múltiples, en las cuales, el objetivo principal es cuantificar la evidencia de un conjunto de datos en contra de un conjunto de hipótesis $H_{0i}, i = 1, 2, ..., k$ de manera simultánea. El procedimiendo usual plantea para cada hipótesis un paradigma binario, en el cual, dependiendo de la fuerza de la evidencia de los datos, habrá que decidir si se rechaza o no. La metodología de las pruebas de hipótesis se construye de tal manera que se controle la tasa de errores de tipo I (hipótesis rechazadas incorrectamente) a un nivel bajo $\alpha > 0$ mientras que la tasa de errores de tipo II se minimiza sujeto a esta restricción. Una colección de procedimientos y tipos de control de tasas de error se presentaron en la Sección 2.1 y se pueden consultar los correspondientes estudios comparativos en Dudoit et al. [2003].

Por otra parte, se presentó en la Sección 2.2 el problema de selección y ordenamiento como la identificación de las *mejores* de un conjunto de k poblaciones comparables. Se pretende que, a través de un análisis cualitativo caracterizado mediante un conjunto de parámetros $\theta_i, i = 1, 2, ..., k$, se realice una *selección* en donde las *mejores* poblaciones correspondan a aquellas poblaciones con los parámetros más grandes. Es la tarea del experimentador inferirlas a través de una muestra aleatoria tomada de cada una de ellas y cuantificar la incertidumbre asociada a la selección correspondiente.

Históricamente, el problema de selección se ha planteado en la práctica a través de las metodologías de pruebas de hipótesis múltiples (Dudoit et al. [2003]). En el Capítulo 2 se explicó esta relación donde a la $i-$ésima población se le plantea una prueba de significancia mediante la hipótesis $H_{0i}, i, i = 1, 2, ..., k$ y con base en el procedimiento estándar de prueba se determina si es estadísticamente *significativa* o no. La pregunta crucial en este caso es «¿Las poblaciones declaradas significativas mediante este procedimiento, en realidad corresponden a las mejores que se están buscando?» De ser así, «¿Cuál es la incertidumbre asociada a esta selección?» Éstas y otras preguntas relacionadas son la base fundamental de la reseña crítica que forma este capítulo.

El ensayo crítico que viene a continuación tiene como propósito motivar, resumir y concientizar acerca de las diferencias entre las metodologías de PHM y RSM tanto a nivel conceptual como a nivel de cuantificación de incertidumbre. Lo anterior se plantea de tal manera que se pueda dar respuesta a la pregunta acerca de cuál de las dos metodologías es más conveniente utilizar y en qué escenarios sucede; ésta será la idea principal de la Sección 3.1. La Sección 3.2 incluirá una colección de ejemplos concretos que motivarán al lector a la identificación de escenarios y a la reflexión acerca de lo establecido en la Sección 3.1. Finalmente, la Sección 3.3 enunciará las principales conclusiones del ensayo.

La mayor parte de la discusión estará basada en Cui and Wilson [2008] y Cui et al. [2010] donde se contrastan por primera vez de manera explícita las ideas de estas dos metodologías y se complementa mediante un estudio formal de simulación con datos reales contrastados con los resultados de Dudoit et al. [2003] acerca de PHM. Se incluyen también ideas de Tukey [1991] quien produce uno de los primeros ensayos críticos acerca de la filosofía de una prueba de hipótesis múltiple.

3.1. Diferencias Conceptuales

En esta primera parte del ensayo se pretende contrastar conceptualmente a las metodologías de PHM y RSM. Se mencionarán el tipo de preguntas que una PHM es capaz de responder y hasta qué alcance y, de existir, se describiran los elementos análogos con RSM. El objetivo primordial de esta parte será conducir al lector a entender dichas diferencias y discernir en cuáles escenarios ambas metodologías pueden llegar a coincidir y en cuáles son totalmente distintas. Los ejemplos concretos se discutirán posteriormente al final de esta sección.

3.1.1. Planteamiento de la Pregunta de Investigación

Supóngase que un genetista desea estudiar un organismo a través de la información genética contenida en sus $k = 3000$ genes. Sin embargo, y como es frecuente en la práctica, los estudios de laboratorio en genética resultan costosos tanto en tiempo como en dinero. Además, el genetista cuenta con un presupuesto limitado para estudiar no más de 100 genes, aunque de ser necesario preferiría enfocarse en la menor cantidad posible, digamos 30 o menos. Estará entonces interesado en hacer una selección cualitativa que le permita aprovechar de la mejor manera posible los recursos limitados que tiene y para ello se plantea la pregunta siguiente:

¿Cuáles genes del organismo son estadísticamente *significativos*?

Para ello, el genetista intuye que es razonable abordar el problema mediante una prueba de hipótesis múltiple (véase la Sección 2.1) planteando para cada gen la prueba de la significancia de la hipótesis:

$$H_{0i} : \text{El } i - \text{ésimo gen no es significativo.} \tag{3.1}$$

Supóngase además que tras la aplicación de un procedimiento estadístico de control de tasa de error de tipo I apropiado el genetista concluye que un número $t < 3000$ de los genes puede considerarse estadísticamente significativo. Ahora bien, si $t > 100$ el genetista se verá obligado a ignorar los resultados de la prueba estadística pues sus recursos no son suficientes para llevar a cabo tal experimento. En cambio si t es muy pequeño, digamos $t = 3$ el genetista podría estar tentado a seleccionar una mayor cantidad de genes independientemente del resultado de la prueba estadística. Si por el contrario, t estuviera en un rango razonable para el genetista, por ejemplo $10 \leq t \leq 100$, podría proceder a realizar el estudio con esa cantidad de genes. Sin embargo, su problema recién comienza, pues no tiene manera de cuantificar la calidad de la selección que acaba de hacer. En otras palabras, no tiene una forma de medir si los genes que acaba de escoger en realidad son los más interesantes para llevarlos a un segundo estudio porque ésta no es la noción de error tipo I que logró controlar al trabajar con la PHM original.

La situación anterior motiva la sutil pero importante diferencia entre *seleccionar* y *buscar significativos*. Cuando se busca aquellas poblaciones que son significativas se busca identificar aquellas que son *distinguibles* entre el conjunto que las contiene. Más áun, la noción de ordenamiento no está presente en ese panorama pues a dicha pregunta simplemente no concierne tal respuesta. El escenario que se visualiza bajo el paradigma de significancia es blanco y negro donde las únicas posibles respuestas son *significativo* o *no significativo*. Por ejemplo, si el genetista declara $t = 20$ genes como significativos, no hay manera de que sepa si en realidad estos corresponden a los mejores 20 de entre el conjunto de k de ellos, es decir, aquellos que en realidad son *interesantes* para el estudio en cuestión. Es incluso muy posible que una (muy pequeña) porción de ellos esté en realidad entre los verdaderos mejores 20.

Las metodologías de selección y ordenamiento (RSM), reseñadas en las Secciones 2.2 y 2.3, permiten, en contraste, estimar la probabilidad de selección correcta (PCS) para cualquier selección de t poblaciones, incluída cualquier selección de t poblaciones declaradas *significativas*. De esta manera, la estimación de PCS complementa y enriquece un análisis de PHM. Por ejemplo, si el genetista *selecciona* $t = 40$

genes con una PCS de 0,95 entonces tendrá alta confianza de que los genes seleccionados en realidad son los verdaderos mejores 40 (de acuerdo con algún significado de mejor, en este caso puede representar a los que aportan la mayor cantidad de información genética del organismo del que provienen).

PHM y RSM no sólo son diferentes en su concepción y metodología sino que también difieren sutilmente en la pregunta de investigación a la cual buscan dar respuesta. Para ver esto, es necesario reescribir la hipótesis (3.1) en términos de homogeneidad. Esto es posible ya que se admite que un individuo es significativo si es estadísticamente *diferente* o *distinguible* del conjunto. Por lo tanto, en general la estructura global de una prueba de significancia como (3.1) es:

$$H: \text{ Las poblaciones son iguales,} \tag{3.2}$$

donde el término *iguales* frecuentemente se representa a través de un modelo paramétrico indexado por los parámetros $\theta_i, i = 1, 2, ..., k$ para cada una de las k poblaciones en prueba. De aquí que la hipótesis (3.2) puede reescribirse en términos de θ_i como:

$$H : \theta_1 = \theta_2 = \cdots = \theta_k. \tag{3.3}$$

Tukey [1991] argumenta que frecuentemente, cuando se plantea una hipótesis múltiple como (3.3), en realidad se plantea la pregunta equivocada «*¿Son iguales las poblaciones?*», y que, por lo general, se está dispuesto a defender una respuesta falsa, «*sí, son iguales*». La hipótesis (3.3) resulta ser poco realista y trivial dado que, a algún número determinado de decimales, se sabe que las poblaciones son diferentes (sus parámetros son diferentes) y un tamaño de muestra suficientemente grande dará evidencia de ello. Más aún la situación anterior se complica cuando el número de hipótesis aumenta pues el número de poblaciones detectadas como *diferentes* es grande (como debería ser dado que todas son diferentes) y el problema original de *seleccionar* cualitativamente un subconjunto de las poblaciones queda sin respuesta.

Para Tukey [1991], la verdadera pregunta de investigación, y la que debería contestarse primero, es: «*¿Es posible determinar las direcciones de las diferencias entre las poblaciones?*» En otras palabras, «*¿Se puede tener confianza acerca del ordenamiento de las poblaciones?*»

Por ejemplo, para el caso $k = 2$ supóngase que se tienen dos poblaciones etiquetadas como A y B con parámetros θ_A y θ_B respectivamente. Las preguntas de interés, según Tukey [1991] deberían ser «*¿Es posible con cierta confianza establecer la dirección de A a B?*» y, de ser así, «*¿Es creciente, decreciente o incierta?*» Si la respuesta a esta pregunta es *incierta* entonces la siguiente pregunta debería ser «*¿Cuáles son los posibles valores de la diferencia $\theta_A - \theta_B$?*» Si la respuesta es, en cambio *ascendente* (o *descendente*) entonces la pregunta inmediata siguiente debería ser «*¿Cual es el mínimo valor de $\theta_A - \theta_B$ (o $\theta_B - \theta_A$)?*» Sin embargo, ninguna de las preguntas anteriores, sin importar las modificaciones necesarias, tiene como objetivo la identificación de la mejor población entre A y B, que es precisamente la pregunta de verdadero interés en un problema de selección. Este mismo caso ocurre (y se complica incluso más) cuando se tiene un número $k > 2$ de poblaciones.

Con acuerdo en el panorama anterior, preguntarse «*¿Cuáles son las mejores poblaciones?*» representa una pregunta totalmente distinta a «*¿Son iguales las poblaciones?*» y a cualquiera de sus variantes. Lo anterior pone en duda el hecho de que resolver un problema de selección mediante una prueba de hipótesis múltiple sea una opción adecuada. En general, debido a su naturaleza binaria, en una prueba de hipótesis no existe la noción de *orden*. Ordenar estadísticas de prueba o p-valores no establece ninguna aseveración probabilística acerca del ordenamiento de los parámetros en cuestión. En consecuencia, un conjunto de poblaciones declaradas *significativas* no puede tener información acerca de si éstas son las *mejores* (o las más interesantes); es posible incluso que difieran considerablemente. En la Sección 3.3 se presentarán algunos ejemplos concretos para ilustrar casos como este.

Las claras diferencias conceptuales y metodológicas entre PHM y RSM llevan inevitablemente al cuestionamiento acerca de si una PHM en realidad responde la pregunta de investigación de un problema

de selección y, si lo hace, cuáles son sus limitantes. Cabe aclarar que lo anterior no constituye una invitación a una crítica calificativa a las metodologías de PHM sino la motivación a un análisis de lo que en realidad son capaces de responder. Esto es precisamente la idea central de este capítulo.

3.1.2. Cuantificación de Incertidumbre

Volviendo al escenario básico de una PHM, supóngase que se tiene un conjunto de k hipótesis $H_{0i}, i = 1, 2, ..., k$ en prueba. El objetivo principal, como se mencionó anteriormente, consistirá en cuantificar la evidencia de los datos en contra de cada hipótesis, y con base en eso decidir si serán rechazadas o no. La tarea global del experimentador en esta instancia de PHM se plantea como:

$$\text{Rechazar un número } t\ (0 \leq t < k) \text{ de hipótesis.} \tag{3.4}$$

Inherente a toda decisión bajo incertidumbre surge de manera natural la noción de *error*. En el caso de la decisión 3.4 puede ocurrir en dos maneras: (I) hipótesis rechazadas incorrectamente (o falsos positivos) e (II) hipótesis no rechazadas incorrectamente. Para el experimentador es importante reducir ambas formas de error al mínimo posible pero, como se ilustró en la Sección 2.1, hacerlo de manera simultánea es imposible, pues reducir una de ellas puede aumentar drásticamente la otra. De esta manera, se fija un umbral $\alpha > 0$, preferentemente pequeño, bajo el cual se acota una tasa permisible de error de tipo I (FWER, FDR, TPPFP, *etc.*), y se espera maximizar la capacidad de la prueba de rechazar correctamente las hipótesis que en realidad son falsas (reducir errores de tipo II). Bajo este paradigma toda la información relacionada con la noción de error en una prueba de hipótesis múltiple queda indexada mediante el parámetro α y su interpretación depende, por lo general, del tipo de tasa de error elegida. Por ejemplo, para el caso de la FDR, FDR$< \alpha$ se interpreta como que la proporción de errores entre las hipótesis rechazadas (falsos positivos) es, en promedio, más pequeña que α. En general, sin importar la tasa de error elegida para controlar, α proporciona una cota superior para alguna función del número de errores de tipo I (o falsos positivos) y la interpretación subsecuente irá siempre en función de dicho concepto.

Por otro lado, en el caso de las técnicas de RSM, el objetivo experimental es seleccionar las mejores t de un conjunto de k poblaciones comparables a través de la información que provee un conjunto de observaciones independientes $X_{ij}, j = 1, 2, ..., n$. La decisión tomada para este caso tiene la forma:

$$\text{Seleccionar las } t\ (1 \leq t < k) \text{ mejores poblaciones del conjunto.} \tag{3.5}$$

La noción de error inherente a la decisión 3.5 consiste en seleccionar incorrectamente un conjunto de t poblaciones que no coincide con las verdaderas mejores. La probabilidad de que esto ocurra se definió como $1-$PCS donde PCS se denominó como probabilidad de selección correcta. En las nuevas metodologías de selección y ordenamiento propuestas en Cui and Wilson [2008] y Cui et al. [2010] la noción de error se redefine de manera específica según el caso particular. Por ejemplo, para el caso de $c-$selección un error consistiría en escoger un conjunto de t poblaciones que contiene menos de c de las verdaderas mejores t poblaciones. En cualquier caso, toda la información relacionada con el error para el contexto de selección queda englobada en el concepto de PCS y como tal se interpreta.

Las técnicas de RSM y PHM se empatan en el caso particular del problema de selección mediante una hipótesis de estructura similar a (3.3). Sin embargo, aunque ambas metodologías se empleen para resolver el mismo problema, sus soluciones no precisamente serán equivalentes. En la sección anterior se explicaron las principales diferencias entre ambas metodologías con relación a la pregunta formulada. Existen, sin embargo, diferencias importantes también en la manera de cuantificar la incerditumbre relacionada a la decisión que implican (el error). En el caso de una PHM se explica a través de α y en el caso de RSM a través de PCS.

Para entender las principales diferencias en el aspecto de cuantificación de incertidumbre para PHM y RSM se considerará el caso análogo entre dos metodologías estadísticas distintas pero relacionadas: estimación y pruebas de hipótesis. Un problema típico de estimación consiste en la inferencia acerca de un parámetro desconocido λ, que es de interés para estudiar el comportamiento de algún fenómeno aleatorio que se modela a través de alguna cierta función de distribución $F(\lambda)$. El modo más simple de resolver un problema de estimación es a través de una muestra aleatoria del fenómeno en cuestión $X_i \sim F(\lambda)$, $i = 1, 2, ..., n$, y con base en ella el cálculo de un estimador puntual $\hat{\lambda}$ que representa una aproximación estadísticamente sustentada del verdadero valor desconocido λ.

Existe, sin embargo, la noción de incertidumbre asociada al valor del estimador $\hat{\lambda}$, pues con cada muestra aleatoria que se obtenga el valor de éste será distinto. Por esta razón se introduce la noción de un *intervalo de confianza* mediante el cual se establece un conjunto de valores (L_1, L_2) que contendrán, con cierto nivel de *confianza* $(1-\alpha)100\,\%$, al verdadero valor del parámetro desconocido λ. Por ejemplo, si $\lambda = \mathrm{E}(X)$ donde $X \sim N(\lambda, \sigma^2)$ con σ conocida, un estimador puntual para λ es $\hat{\lambda} = \bar{X}$ y un intervalo de confianza de $(1-\alpha)100\,\%$ para λ está dado por $L(X) = \bar{X} \pm (Z_{\alpha/2}\sigma)/\sqrt{n}$ pues puede demostrarse que $P(\lambda \in L(X)) = 1 - \alpha$.

Sea $X = (X_1, X_2, ..., X_n)$ una muestra aleatoria observada de $F(\lambda)$ y sea $L(X)$ un intervalo de confianza $(1-\alpha)100\,\%$ para el parámetro desconocido λ. Si se considera a continuación

$$H_0 : \lambda = \lambda_0 \;\; vs. \;\; H_1 : \lambda \neq \lambda_0, \tag{3.6}$$

donde se rechazará H_0 si y sólo si $\lambda_0 \notin L(X)$, se tiene entonces que

$$P(\text{Error Tipo I}) = P(\lambda_0 \notin L(X) | \lambda = \lambda_0) = 1 - P(\lambda \in L(X)) \leq \alpha. \tag{3.7}$$

De esta manera, a partir de un intervalo de confianza se ha construido una prueba de hipótesis con nivel de significancia α.

De manera análoga, supongamos que se tiene un conjunto de pruebas de nivel α para cualquier elección de hipótesis nula de la forma $H_0 : \lambda = \lambda_0$. Sea $X = (X_1, X_2, ..., X_n)$ una muestra aleatoria observada de $F(\lambda)$. Si se considera a continuación:

$$L(X) = \{\lambda_0 : \text{No se rechaza } H_0 \text{ tras observar X}\}, \tag{3.8}$$

se tiene entonces que

$$P(\lambda \in L(X)) = P_\lambda(\text{No se rechace } H_0) \geq 1 - \alpha. \tag{3.9}$$

De aquí que, $L(X)$ es un intervalo de $(1-\alpha)100\,\%$ de confianza para λ construido a partir de una prueba de hipótesis de nivel α.

Las situaciones anteriores ilustran que a partir de todo intervalo de confianza de $(1-\alpha)100\,\%$ para un parámetro desconocido λ se puede construir una prueba de hipótesis con nivel de significancia α y viceversa. Dicho resultado se conoce como el *teorema de inversión* y permite establecer una relación directa entre la noción de *confianza* del enfoque de estimación y la noción de *error de tipo I* de una prueba de hipótesis. En otras palabras, a pesar de que estimación y pruebas de hipótesis son dos problemas distintos, existe una conexión directa entre sus formas de cuantificación de incertidumbre que permite construir soluciones equivalentes.

Tal no es el caso para las pruebas de hipótesis múltiples y las metodologías de selección y ordenamiento. No existe resultado alguno, análogo al teorema de inversión para pruebas de hipótesis y estimación, que permita relacionar directamente sus medidas de cuantificación de incertidumbre que son el nivel de significancia α y PCS. Por lo tanto, no hay manera analíticamente comprobable de obtener soluciones equivalentes mediante ambas metodologías, incluso aunque se apliquen al mismo conjunto de datos.

Tukey [1991] comenta acerca de la discrepancia entre las formas de cuantificar incertidumbre entre una prueba de hipótesis (o intervalo de confianza) y una relación de ordenamiento:

«Si un intervalo de confianza de 95 % es de 70 a 130, un intervalo de 99 % será más ancho; por ejemplo de 60 a 140. Mientras tanto, un intervalo de 50 % será más corto, por ejemplo de 90 a 110. Al menos vagamente, un intervalo de confianza para una tasa de error siempre permite inferior algo acerca de los intervalos de confianza a otras tasas de error. No existe un análogo a esto en el caso de la dirección de confianza. Si se tiene la confianza de que $A > B$ a un 5 % de tasa de error, no es posible saber si la misma desigualdad vale a 1 % o no. Más aún, si no se tiene certeza acerca del signo de $A - B$ a 5 % de tasa de error no se puede saber si se tendrá a 50 %.»

Su observación puntualiza indirectamente que, en general, no es posible obtener una solución enteramente satisfactoria a un problema de ordenamiento a través de algún intervalo de confianza o similarmente, de una prueba de hipótesis. Así mismo, debido a la ausencia de un análogo al teorema de inversión entre PHM y RSM, la aplicación de una PHM para la solución de un problema de ordenamiento llevaría a conclusiones que no son traducibles en términos de un *rankeo* que en muchos casos es el objeto de verdadero interés.

3.1.3. Identificación de Escenarios

Las dos secciones anteriores puntualizaron las principales diferencias sutiles pero críticas entre PHM y RSM. Dado que tanto conceptual como operativamente son distintas, la pregunta natural a responder es «¿Cuándo es apropiado utilizar una PHM para resolver un problema de selección y cuándo no lo es?»

La respuesta yace en la identificación adecuada del escenario y los objetivos del análisis estadístico que se pretente implementar. Supóngase que se tiene un problema de selección sobre un conjunto de k poblaciones independientes comparables:

- Si el objetivo es la identificación de todas las poblaciones *candidatas* a selección, donde es de particular interés detectar todas las poblaciones *distinguibles* del resto del conjunto, una prueba de hipótesis múltiple es apropiada. De esta manera el objetivo de selección y el de identificación de poblaciones significativas se empatan pues ambas metodologías buscan la identificación de exactamente las mismas poblaciones.

- Si el objetivo es la identificación de un grupo selecto de las poblaciones más *interesantes* ordenadas de acuerdo a un criterio *cualitativo* apropiado, una técnica de RSM como las que se proponen en Cui and Wilson [2008] resulta ser más apropiada. Una técnica de PHM no sería útil en este caso pues como se expone en las Secciones 3.1.1 y 3.1.2 no responde a la pregunta de investigación directamente ni admite un modo de cuantificación de incertidumbre que permita obtener una solución equivalente a la que en realidad se busca.

Ejemplos concretos de ambos escenarios se presentarán en la siguiente sección.

3.2. Ejemplos

La siguiente colección de ejemplos tiene como propósito la ilustración de tres escenarios concretos en los cuales un problema de selección se relaciona con una metodología de pruebas de hipótesis. Como podrá verse, en algunos casos esta relación resulta ser fortuitamente apropiada, pero en algunos otros puede llevar a conclusiones súmamente equivocadas como consecuencia de las importantes diferencias conceptuales expuestas en la sección anterior.

Ejemplo 3.2.1:
Supóngase que se tiene un conjunto de $k = 2$ poblaciones etiquetadas como π_1 y π_2 tales que $\mu_i = \mathrm{E}(\pi_i)$, $i = 1, 2$. El objetivo es seleccionar a la mejor, donde mejor en este caso es aquella que tiene mayor

media, *i.e.* aquella cuya media corresponde a $\mu_{[2]}$. Supóngase además que para ello se utiliza una prueba de hipótesis (simple en este caso) planteada de la siguiente manera:

$$H_0 : \mu_1 < \mu_2 \text{ vs. } H_1 : \mu_2 \leq \mu_2. \tag{3.10}$$

Por simplicidad, y sin pérdida de generalidad, supóngase que las poblaciones π_i, $i = 1, 2$ siguen una distribución normal con varianzas desconocidas y que se toman muestras aleatorias X_{ij}, $j = 1, 2, ..., n_i$ de cada una, con tamaños de muestra n_i no precisamente iguales.

El procedimiento estándar de una prueba de hipótesis como (3.10) requiere el cálculo de una estadística de prueba T. Para este caso particular una elección apropiada (Casella and Berger [2008]) tiene la forma:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{3.11}$$

donde $\bar{X}_i$ y s_i^2 son la media y la varianza muestrales de las observaciones tomadas de las observaciones de la población π_i respectivamente. La razón de la elección particular de la estadística (3.11) para este caso es que es posible demostrar que, bajo H_0, ésta se distribuye como una variable aleatoria $t-$Student con grados de libertad dados por una función $g(s_1^2, s_2^2, n_1, n_2)$.

La teoría estadística de pruebas de hipótesis indica que, si se rechaza H_0 siempre que $T \geq t_{\alpha/2}$, donde $t_{\alpha/2}$ es el cuantil correspondiente de la distribución nula de T, una y sólo una de las siguientes conclusiones se puede obtener para cualquier muestra aleatoria obtenida X_{ij}:

- Si se rechaza H_0 con un nivel de significancia α se puede concluir que, si se escoge μ_1 como la mayor media se tendrá al menos un $(1-\alpha)100\,\%$ de confianza de que ésta en realidad lo es. En este escenario, $(1-\alpha)100\,\%$ tiene una interpretación similar a $\text{PCS}_{t=1}$, en el sentido que, proporciona un algoritmo de selección (elegir π_1 como la mejor) y una medida de cuantificación de incertidumbre asociada a dicha elección.

- Si un rechazo no ocurre, la conclusión es que no se tiene evidencia suficiente de los datos en contra de H_0 para rechazarla. En otras palabras, no es posible concluir nada acerca de H_0 en base a la información obtenida. Aceptar la hipótesis nula y *decidir* que μ_2 es la mayor de las medias constituiría un abuso conceptual (Tukey [1991]) y más aún, sería imposible cuantificar la incertidumbre asociada a esta elección.

Nota 1: En el planteamiento anterior se pudo haber supuesto que las varianzas de π_1 y π_2 eran conocidas (o incluso iguales) y que se tomaron tamaños de muestra iguales $n_1 = n_2$, así como un modelo distinto al normal. El discurso anterior, bajo las modificaciones correspondientes a la estadística de prueba y el criterio de rechazo, sería igualmente válido.

Nota 2: La elección del orden de las hipótesis no tiene efecto alguno en el discurso anterior. Se podría reemplazar H_0 en (3.10) por $H_0 : \mu_2 < \mu_1$ y se obtendrían conclusiones análogas.

Nota 3: Si se reemplaza H_0 en (3.10) por una hipótesis de homogeneidad $H_0^* : \mu_2 = \mu_1$, las conclusiones posibles serían incluso más ambiguas. Por un lado, si se rechazara H_0^*, existiría evidencia suficiente para concluir que μ_1 y μ_2 no son iguales. Sin embargo, no sería posible elegir alguna de ellas como la mejor a partir de esta conclusión. Por otro lado, si no se rechazara H_0^* no habría evidencia suficiente para concluir que μ_1 y μ_2 no son iguales, pero tampoco se podría concluir que lo son.

Si alternativamente se hubiera optado por la aplicación de una técnica de RSM la (única) posible interpretación de la conclusión sería más clara y directa. Supóngase que para las muestras obtenidas sucede que $\bar{X}_2 < \bar{X}_1$ y se estima que PCS es 0,9. La conclusión sería para este caso: *La mejor población es* π_1 *con una probabilidad de* 0,9 *de estar en lo correcto.*

PCS aporta información valiosa respecto al problema de selección sin importar su valor. Por ejemplo, si PCS fuera cercana a 0 se podría concluir que μ_1 y μ_2 están tan cercanas que es *difícil* hacer una elección

entre ellas, asumir que son iguales en este caso es completamente intuitivo y razonable. Tal no es el caso en una prueba de hipótesis, donde *no rechazar* H_0 no implica que sea razonable *aceptarla*. En cambio, mientras más se acerque PCS a 1 más certeza se tiene, no sólo de que las medias son distintas sino de que aquella que se está reportando como la mejor en realidad lo es.

PCS constituye un instrumento que responde directamente a la pregunta de investigación en un problema de selección y además cuantifica la incertidumbre asociada a ella. La prueba de hipótesis podría resolver la pregunta de manera indirecta en algunos casos cuando se busca significancia (véase el primer caso en las conclusiones de este ejemplo) pero es completamente incapaz de aportar más información útil en cualquier otro caso.

Ejemplo 3.2.2:
Para este ejemplo se simuló un conjunto de datos X_{ij}, $j = 1, 2, ..., n = 20$ de las poblaciones $\pi_i, i = 1, 2, ..., k = 100$, donde $X_{ij} \sim N(\mu_i, 1)$, de tal manera que $\mu_i = 3, i = 1, 2, ..., 70$ y $\mu_i = 5, i = 71, 72, ..., 100$. El objetivo del ejemplo es la selección de todas las poblaciones con mayor media, que en este caso se sabe de antemano que corresponden a $\pi_i, i = 71, ..., 100$.

Supóngase que se sabe que toda población que pueda considerarse *candidata* a ser parte de las mejores se distingue por tener una media superior a 3. Entonces se plantea una prueba de hipótesis múltiple sobre las medias de la forma:

$$H_{0i} : \mu_i \leq 3. \tag{3.12}$$

El algoritmo empleado a continuación es el siguiente:

- Utilizando una prueba Z se obtuvieron los p-valores sin ajustar para cada una de las hipótesis individuales H_{0i}.
- Utilizando el procedimiento de control de la tasa de descubrimientos falsos (FDR), propuesto en Benjamini and Hochberg [1995], se obtienen los p-valores ajustados $\tilde{p}_i, i = 1, 2, ..., 100$ a partir de los p-valores obtenidos en el paso anterior.
- Se rechazará toda hipótesis cuyo p-valor ajustado esté por debajo del nivel de significancia α pre-especificado.

Con la ayuda de la herramienta computacional interactiva diseñada para propósitos de la tesis y que se describirá en el siguiente capítulo, se graficaron los p-valores ajustados ordenados para todas las hipótesis en la escala $-\log_{10}$ en la Figura 3.1. En ella, la línea punteada horizontal representa $-\log_{10}(0{,}05)$, mientras que los puntos rojos y azúles representan las hipótesis rechazadas y no rechazadas respectivamente. Las hipótesis cuya media verdadera es 3 se etiquetaron mediante la nomenclatura $V1, ..., V70$ y las hipótesis cuya verdadera media es 5 mediante $F1, ..., F30$. En general, la prueba identificó correctamente a todas las mejores poblaciones, excepto la población $V20$ cuya hipótesis nula fue rechazada incorrectamente.

Para este caso es importante notar que, dado que el propósito de investigación es explícitamente la identificación de todas las poblaciones candidatas a ser las mejores, una prueba de hipótesis resuelve de manera indirecta un problema de selección con una tasa de error relativamente baja. Sin embargo, si el escenario hubiera sido diferente y el experimentador estuviera interesado específicamente en la identificación de las $t = 15$ mejores poblaciones, este procedimiendo sería incapaz de darle una respuesta apropiada sin importar las modificaciones posibles que se hagan a la hipótesis (3.12).

Nótese además la importancia de la especificación precisa del umbral $\mu_0 = 3$ en la hipótesis (3.12). Si éste fuera desconocido —como es el caso más frecuente en la práctica— y se escogiera de manera arbitraria, las conclusiones podrían cambiar y ser equivocadas. Por ejemplo, si la hipótesis (3.12) se hubiera planteado con un umbral diferente:

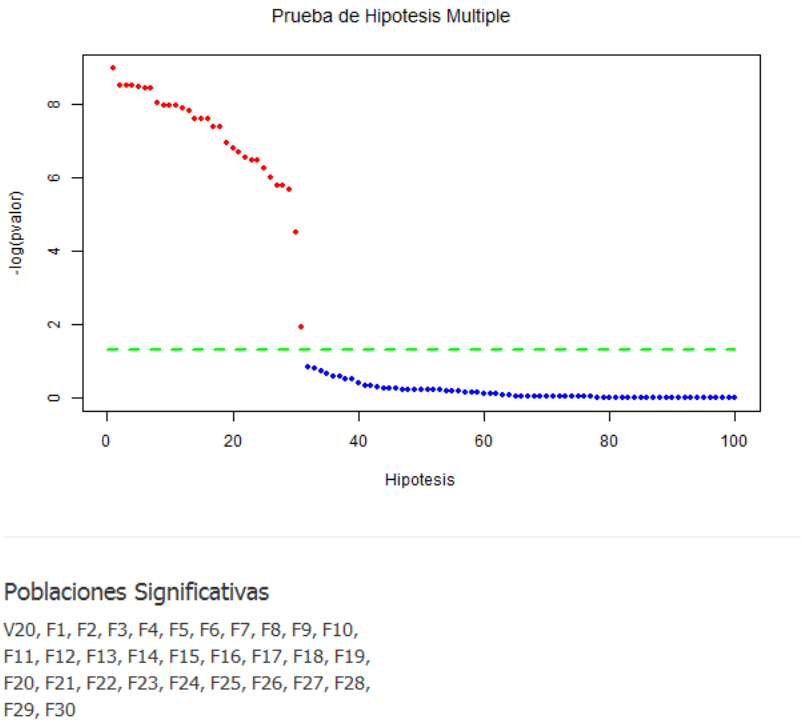

Figura 3.1: Gráfico de p-valores ajustados en la escala logarítimica inversa para la prueba 3.12 y las poblaciones declaradas significativas a un nivel de significancia $\alpha = 0{,}05$

$$H_{0i} : \mu_i \leq 2{,}5, \tag{3.13}$$

se hubiera obtenido el gráfico de la Figura 3.2. En éste puede notarse que hay 30 poblaciones que no fueron rechazadas cuando debieron haberlo sido. Además, la prueba está sugiriendo una selección arbitraria entre las poblaciones con media $\mu_i = 3$ que no tiene razón de ser pues sus medias se simularon iguales desde el principio. En este caso, el experimentador estaría tentado a reportar $t = 70$ poblaciones como las mejores —más aún, a un 95 % de confianza— cuando en realidad ninguna de las dos afirmaciones es cierta. La primera, debido a las diferencias en cuanto a la pregunta formulada entre PHM y RSM (Sección 3.1.1) y la segunda debido a las diferencias en cuanto a cuantificación de incertidumbre (Sección 3.1.2).

Ejemplo 3.2.3:
Para este ejemplo se simuló otro conjunto de datos $X_{ij}, j = 1, 2, ..., n = 3$ de poblaciones $\pi_i = 1, 2, ..., k = 600$, donde $X_{ij} \sim N(\mu_i, 1)$, de tal manera que $\mu_i = 2, i = 1, 2, ..., 200$, $\mu_i = 7, i = 201, ..., 400$ y $\mu_i = 15, i = 401, ..., 600$. En este caso el propósito es la identificación de las $t = 200$ mejores poblaciones.

La Figura 3.3 muestra el gráfico de los p-valores ajustados (cálculados mediante el algoritmo del ejemplo anterior) para un nivel de significancia $\alpha = 0{,}05$ para las hipótesis $H_{0i} : \mu_i \leq 2$ y $H_{0i} : \mu_i \leq 7$ respectivamente. En el caso de $\mu_0 = 2$ ocurrieron 408 rechazos por lo que, de entrada, algunas hipótesis fueron rechazadas de manera incorrecta. Aún así, de entre las hipótesis rechazadas (rojo) es imposible discernir cuántas y cuales son las verdaderas mejores por lo que el experimentador se vería forzado a reportar las $t = 408$ poblaciones significativas como las mejores cuando en realidad existe una cantidad mucho más reducida de ellas ($t = 200$), que en verdad son *interesantes*. Más aún, carecería de una medida de incertidumbre acerca de dicha selección ya que el verdadero significado de $\alpha = 0{,}05$ es que, entre las $t = 408$ hipótesis rechazadas, se espera un 5 % o menos de falsos positivos. Lo anterior en efecto

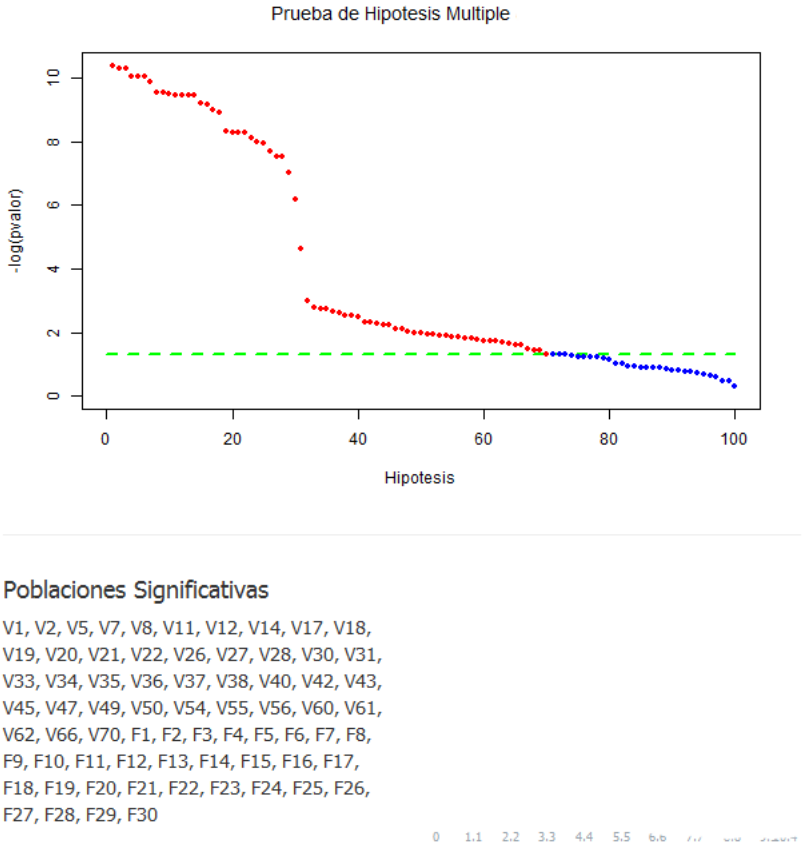

Figura 3.2: Gráfico de p-valores ajustados en la escala logarítimica inversa para la prueba 3.13 y las poblaciones declaradas significativas a un nivel de significancia $\alpha = 0{,}05$

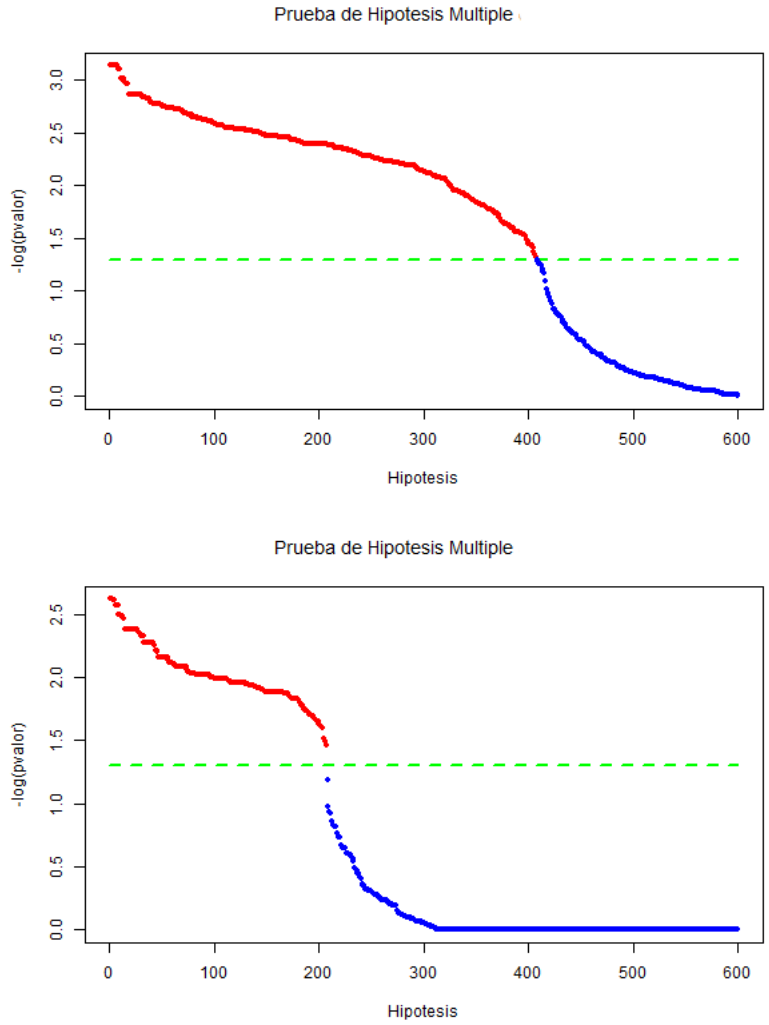

Figura 3.3: Gráfico de p-valores ajustados en la escala logarítimica inversa para las pruebas $H_{0i} : \mu_i \leq 2$ (arriba) y $H_{0i} : \mu_i \leq 7$ (abajo) a un nivel de significancia $\alpha = 0{,}05$

se verifica pues $5/408 \leq 0{,}05$. Sin embargo, α no se relaciona de ninguna manera con la probabilidad de que las $t = 408$ reportadas como las mejores en realidad lo sean.

Para el caso de $\mu_0 = 7$ ocurrieron 207 rechazos, por lo que, nuevamente existe una cantidad de falsos positivos de entrada. Sin embargo, para este caso dicha cantidad no es relativamente importante. Puede verse que, únicamente para el caso en que el experimentador conociera con exactitud que toda población cuya media es específicamente mayor que $\mu_0 = 7$ puede ser considerada como *mejor*, se resolvería el problema de selección de manera satisfactoria mediante una PHM. Cabe aclarar que ambas metodologías coinciden de manera fortuita en este caso porque se está forzando que los conceptos de *mejor* y *significativo* se empaten. Toda población cuya media es mayor que 7 (significativa) es parte de las verdaderas mejores y toda población del conjunto de las verdaderas mejores tiene una media superior a 7 (es significativa). Lamentablemente, en la práctica, dicha situación es poco realista y en muchos casos absurda pues, si *a priori* el investigador conociera las características exactas de toda población digna de considerarse parte de las mejores, posiblemente no sería necesario un análisis estadístico de selección en primer lugar.

Finalmente, la Figura 3.4 muestra el resultado de la aplicación de la $d-$selección, una de las técnicas de RSM propuestas en Cui and Wilson [2008] y resumidas en la Sección 2.3, al mismo conjunto de datos.

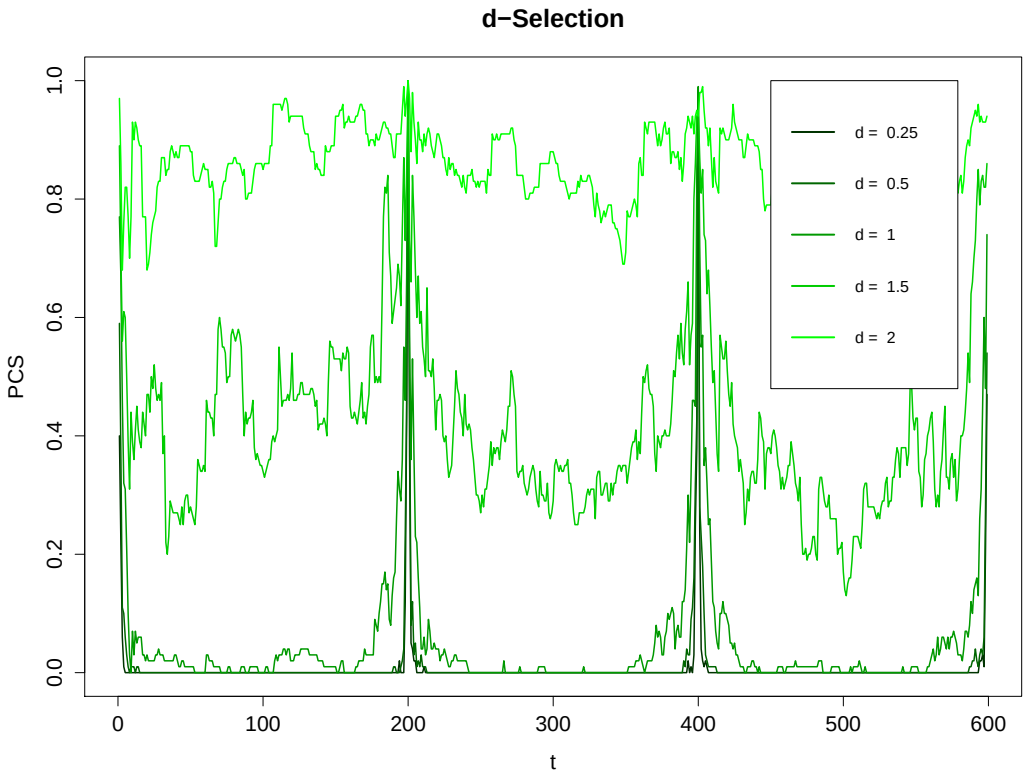

Figura 3.4: Resultado de la aplicación de la $d-$selección de Cui and Wilson [2008] para la estimación de ${}_d\text{PCS}_t$ en el ejemplo simulado, utilizando distintos valores de d y como función t.

Vista como función de t, la ${}_d\text{PCS}_t$ permite visualizar qué valores de t es conveniente utilizar y cuál es la probabilidad de selección correcta aproximada que se puede obtener con ellos. Es interesante notar que, incluso para valores muy pequeños de d, las curvas tienen un máximo local en los puntos $t = 200$ y $t = 400$ que son los valores que, en acuerdo con la simulación representan las elecciones correctas de t. Más aún, ${}_d\text{PCS}_{t=200}$ y ${}_d\text{PCS}_{t=400}$ son cercanas a 1 sin importar el valor de d.

La aplicación de las técnicas de selección restantes de Cui and Wilson [2008] y Cui et al. [2010] a este conjunto de datos simulado será presentada en el siguiente capítulo con motivo de ilustrar las principales funciones del software diseñado para la tesis.

3.3. Discusión y Conclusiones

El problema de selección bajo incertidumbre constituye una línea de investigación específica como problema *bona fide* desde su proposición en Bechhofer [1954] y sus métodos de solución nombrados formalmente metodologías de selección y ordenamiento (RSM). Sin embargo, históricamente el problema de selección ha sido interpretado de diferentes maneras y como consecuencia su solución se ha reformulado y replanteado de distintas maneras. Una de las principales se logra a través de la aplicación de una prueba de hipótesis múltiple (PHM).

Mediante el presente ensayo se puntualizaron las principales diferencias entre una PHM y las técnicas de RSM llegándose a las siguientes conclusiones:

- Una PHM y un problema de selección y ordenamiento son dos problemas distintos desde su concepción.

- La pregunta planteada por una PHM y un problema de selección y ordenamiento es diferente, incluso aunque se apliquen al mismo conjunto de datos. Una PHM busca detectar significancia mientras que las técnicas de RSM buscan una selección cualitativa.
- Los conceptos de *significativo* e *interesante* no son sinónimos y no son intercambiables.
- La forma de cuantificar incertidumbre de una PHM y cualquier método de RSM no es la misma y no existe una relación teórica entre ambas. PCS y nivel de significancia no son equivalentes ni intercambiables y por tanto no se interpretan de la misma manera.
- Existen escenarios selectos en los cuales PHM y RSM pueden conducir fortuitamente a resultados análogos. La razón es porque, en algunos casos, los conceptos de *mejores poblaciones* y *poblaciones significativas* son equivalentes. La recomendación general es saber identificar la pregunta de específico interés y con base en ella seleccionar la técnica más apropiada.

Capítulo 4

Software y Caso de Aplicación

Este capítulo se divide, a grandes rasgos, en dos partes. La primera presenta la herramienta de software desarrollada específicamente para los objetivos de la tesis. El software es una aplicación interactiva elaborada en R utilizando algunos elementos de diseño de interfaces gráficas de Shiny; se pretende demostrar algunas de sus principales funciones a través de un ejemplo simulado en la Sección 4.1. Finalmente, la Sección 4.2 describirá el caso de aplicación principal de la tesis con datos reales. Éste consiste del problema de longevidad en células de levadura propuesto por investigadores de Langebio. Haciendo uso del software de desarrollo propio se plantearán y responderán algunas de las más importantes preguntas en la relación a dicho planteamiento.

4.1. Software

Uno de los objetivos principales de la tesis consiste en la recomendación una metodología estadística adecuada para problemas relacionados con RSM (véase la Sección 1.3). En respuesta a dicho planteamiento se propuso el diseño y desarrollo de una herramienta computacional propia que permita, mediante una interfaz interactiva, establecer un vínculo entre las técnicas estadísticas de RSM y los usuarios potenciales cuyas problemáticas conciernen a dicha metodología.

Cui and Wilson [2008] adaptaron por primera vez las principales nociones de RSM para k grande mediante la librería PCS de R (www.r-project.org). Ésta incluye, principalmente, la implementación computacional de los algoritmos relacionados con la estimación bootsrap (paramétrica y no paramétrica) de PCS (véase la Sección 2.3.5). La herramienta computacional diseñada para los propósitos de este trabajo tiene como base los algoritmos de la librería PCS incorporados en un entorno gráfico diseñado a través de la librería Shiny. El resultado final es una aplicación web interactiva que se considera puede ser muy útil para usuarios potenciales en entornos multidisciplinarios.

La presente sección describirá los elementos principales de la herramienta computacional, sus principales funciones, tipos de entradas y algunos de los conceptos técnicos utilizados para su desarrollo. Se describirán además los tipos de salidas que puede producir y las preguntas a las que se piensa que puede dar respuesta; se explicará mediante un conjunto de datos simulado que se introducirá eventualmente. Cabe aclarar que esta sección no se trata de un manual de usuario *per se*, sino un resumen técnico de la herramienta computacional misma.

4.1.1. Objetivos

Los objetivos del desarrollo del software son tres:

- Proveer una herramienta didáctica que permita entender de manera interactiva las principales técnicas de RSM en el desarrollo de la tesis. Lo anterior se pretende lograr mediante el desarrollo

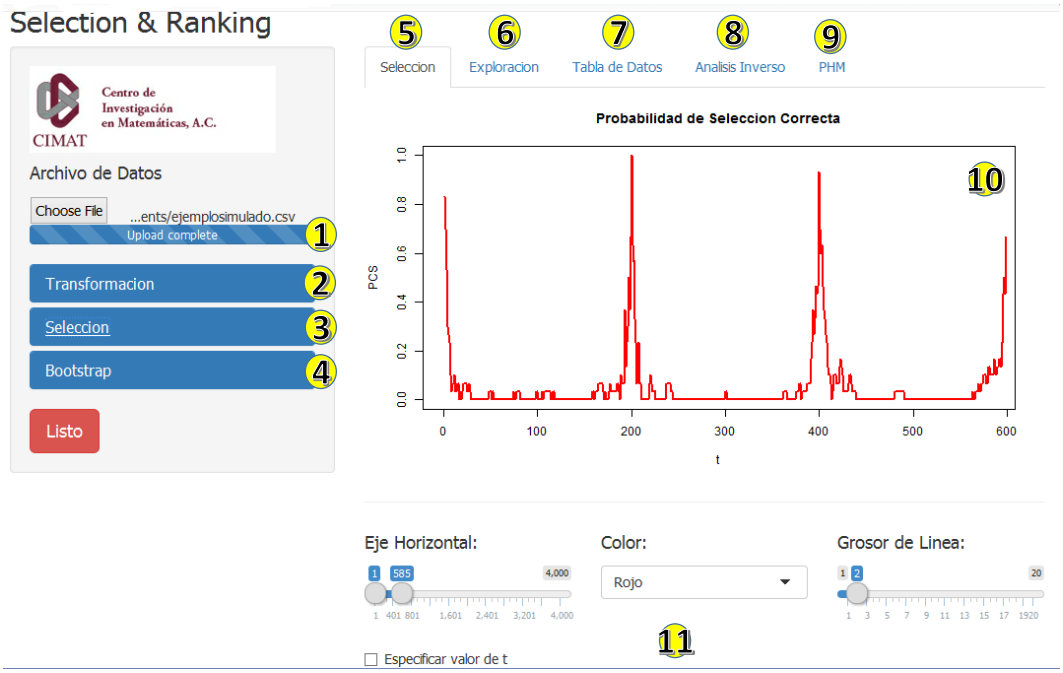

Figura 4.1: Pantalla principal del software con sus principales elementos enumerados.

π_1	π_2	$\cdots$	π_k
X_{11}	X_{21}	$\cdots$	X_{k1}
X_{12}	X_{22}	$\cdots$	X_{k2}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
X_{1n}	X_{2n}	$\cdots$	X_{kn}

Cuadro 4.1: Formato de la base de datos reconocido por el software, donde los encabezados $\pi_i, i = 1, 2, ..., k$ pueden reemplazarse por las etiquetas que el usuario prefiera para nombrar las clases.

de ejemplos de simulación que permitan explorar situaciones varias.

- Auxiliar en problemas de toma de decisiones relacionadas con selección en presencia de incertidumbre.
- Aportar una herramienta de exploración, diagnóstico y verificación de suspuestos de la teoría de RSM que además permita implementar soluciones en casos en los que estos no se cumplan.

4.1.2. Interfaz y Formato de Entrada

Siguiendo la notación introducida en el Capítulo 2, sean $X_{ij}, j = 1, 2, ..., n$ observaciones independientes de la población $\pi_i, i = 1, 2, ..., k$. Para funcionar, el software requiere que las observaciones X_{ij} sean ingresadas en un archivo con extensión .csv en una tabla de dimensiones $(n + 1) \times k$ con el formato ilustrado en el Cuadro 4.1. Sin importar la ubicación del archivo, éste puede buscarse de manera interactiva a través de la interfaz gráfica (Figura 4.1(1)). Es importante que el archivo de datos tenga extensión .csv, ya que de otra manera el software no podrá reconocerlo.

Ejemplo 4.1.1:
Para propósitos ilustrativos, se preparó un conjunto de datos simulados donde $X_{ij} \sim N(\mu_i, 1)$ con $\mu_i = 2$ para $i = 1, 2, ..., 200$, $\mu_i = 7$ para $i = 201, 2, ..., 400$ y $\mu_i = 15$ para $i = 401, 402, ..., 600$. Se simularon $n = 3$ observaciones para cada población y se arreglaron de acuerdo con el formato 4.1.

4.1.3. Herramientas de Diagnóstico y Exploración

Dos buenas prácticas antes de la implementación de toda técnica estadística a un nuevo conjunto de datos son el análisis exploratorio y la verificación de supuestos. El software posee la una gama de herramientas gráficas que pueden utilizarse para ambos objetivos (Figura 4.1(6)). A continuación se describe cada una de ellas y un resumen de su posible interpretación ligada al Ejemplo 4.1.1.

Herramientas de Exploración

- **Histograma de Datos.**
  Calcula un histograma simple de todos los datos ingresados, independientemente de la clase a la que pertenecen. Se pretende que de una idea general al usuario de la estructura que tiene la base de datos que acaba de ingresar. Esta herramienta puede emplearse también para detectar *outliers* o errores de captura en la base de datos. Si el histograma se muestra unimodal con una moda muy marcada o es similar al histograma de una distribución uniforme el procedimiento de selección puede complicarse (más no es imposible). La Figura 4.2 muestra el histograma de datos para el Ejemplo 4.1.1 en el cual se aprecia claramente que existe cierta porción de los datos cuya media es marcadamente mayor y viceversa.

- **Histograma de Tamaños de Muestra.**
  Grafica un histograma de $n_i, i = 1, 2, ..., k$, los números de observaciones que se ingresaron para cada población o tratamiento. Este histograma es particularmente útil cuando se sabe que se tienen datos faltantes o tamaños de muestra desiguales y se pretende cuantificar qué porción de la información está completa. Para el caso de la Figura 4.2 el histograma es trivial pues todas las poblaciones se simularon con tamaños de muestra iguales $n = 3$.

- **Gráfico de Dispersión Datos *vs.* Tratamientos.**
  Muestra un gráfico de dispersión donde sobre el eje horizontal se colocan las poblaciones y sobre ellas cada una de las observaciones correspondientes. Es el análogo al histograma de datos pero haciendo distinción por poblaciones. La Figura 4.2 permite al usuario visualizar que existen tres bloques de tratamientos ordenados cualitativamente.

- **Histograma de Errores.**
  El software incorpora la posibilidad de que cada observación tenga asignado un error estándar relacionado al proceso experimental empleado para su obtención (véase por ejemplo el caso de estudio de la Sección 4.2). Por tanto, esta herramienta le permite al usuario visualizar la estructura de los errores y determinar si existen observaciones con errores asociados relativamente grandes o sistemáticos que pudiera ser deseable revisar o eliminar del estudio.

Herramientas de Diagnóstico

- **Histograma de Estadísticas Resumen.**
  Parte del procedimiento estándar de la RSM involucra el cálculo de estadísticas resumen para cada población $Y_i, i = 1, 2, ..., k$ (Cui and Wilson [2008]). La elección de las estadísticas resumen no es trivial y en general va ligada con la naturaleza de los parámetros θ_i sobre los cuales se desea inferir

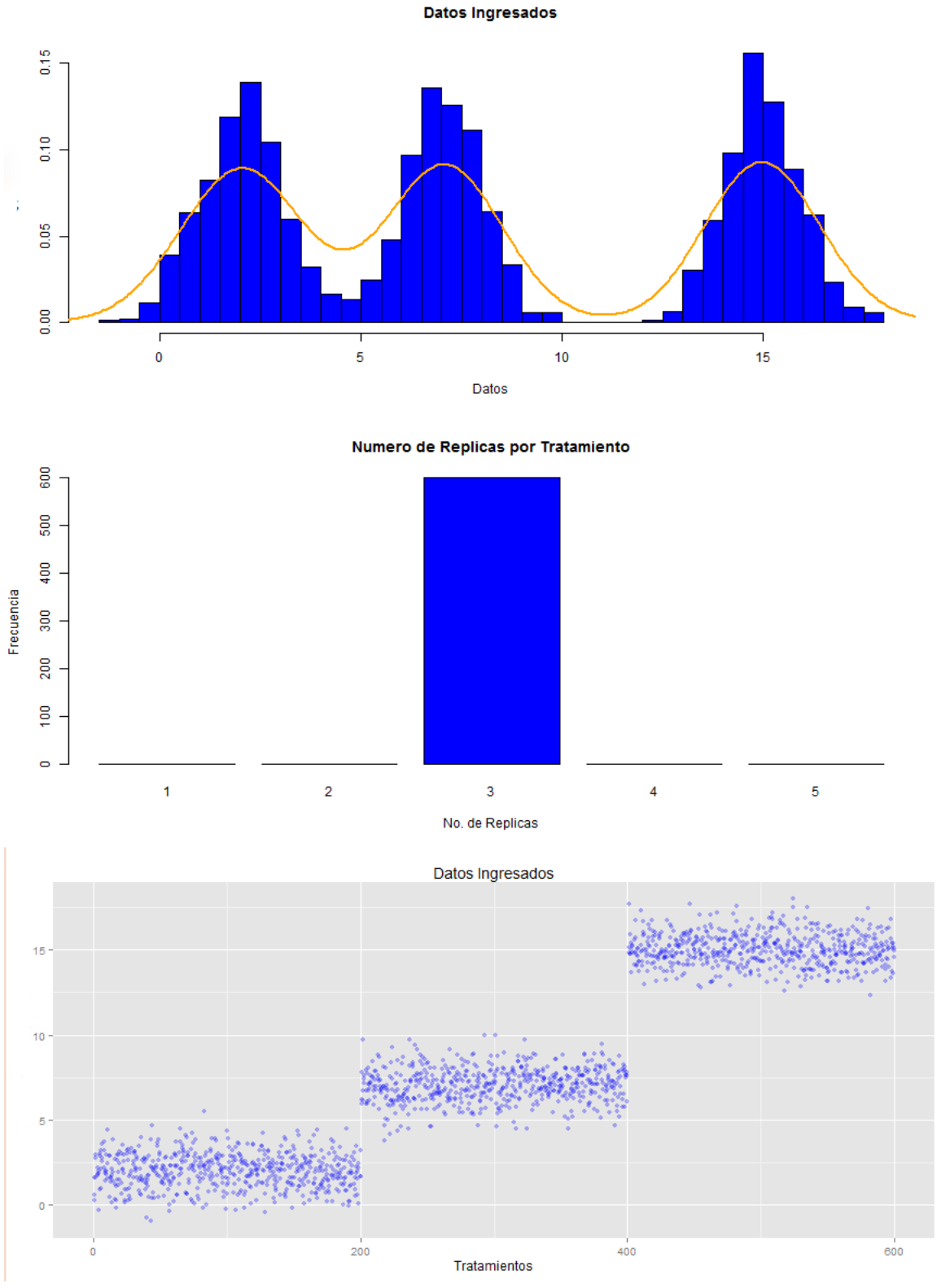

Figura 4.2: Herramientas de exploración incluídas en el software aplicadas a los datos simulados del Ejemplo 4.1.1. **(Arriba)** Histograma de datos. **(Centro)** Histograma de tamaños de muestra. **(Abajo)** Gráfico de dispersión datos *vs.* tratamientos.

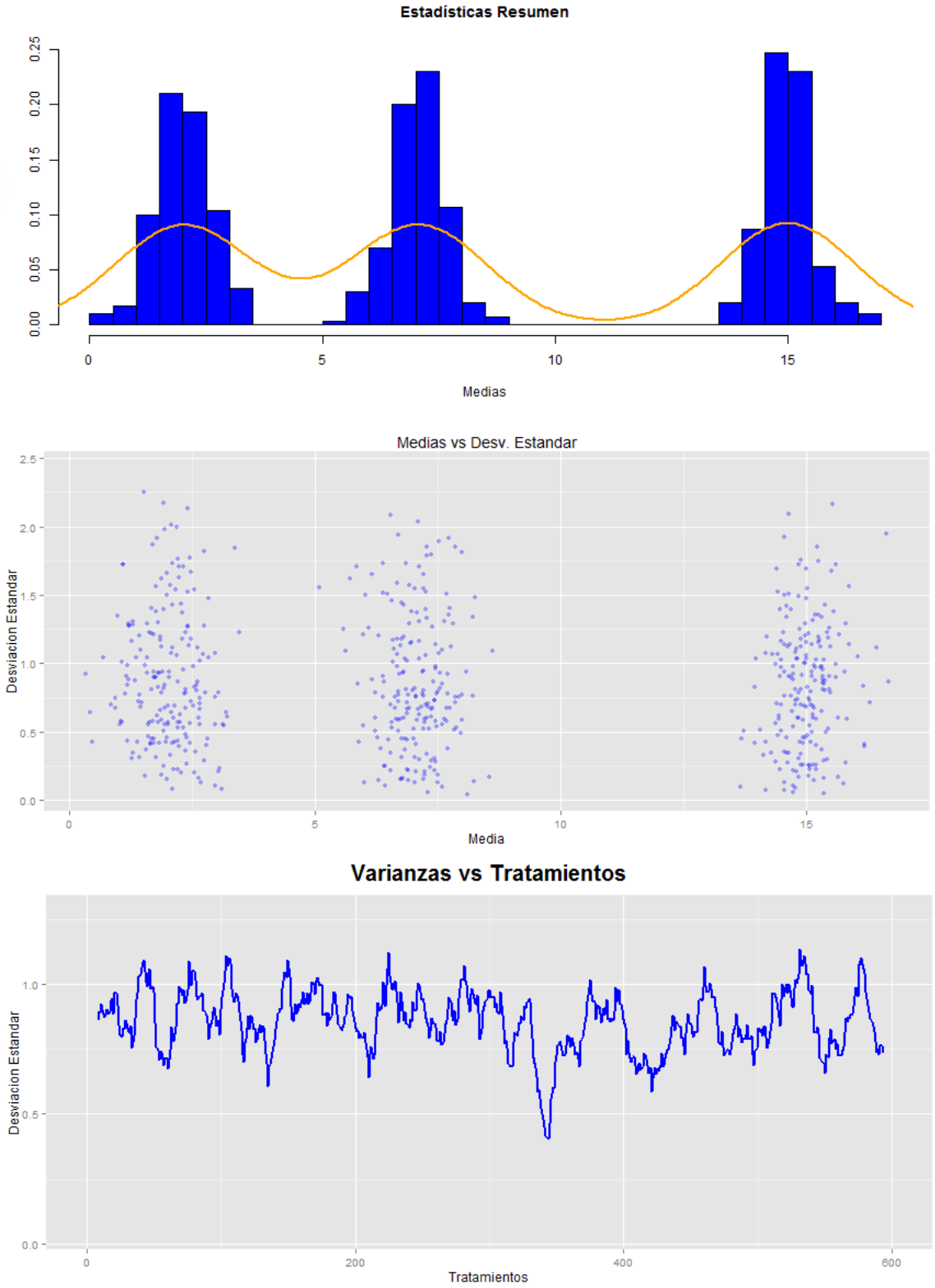

Figura 4.3: Herramientas de diagnóstico incluídas en el software aplicadas a los datos simulados del Ejemplo 4.1.1. **(Arriba)** Histograma de estadísticas resumen (medias muestrales). **(Centro)** Gráfico de dispersión de medias *vs.* desviaciones estándar por población. **(Abajo)** Gráfico de tratamientos *vs.* desviaciones estándar.

un ordenamiento. Es de especial interés que Y_i sirva como un estimador para θ_i de manera que se pueda utilizar para inferir sobre su posición en el ordenamiento.

En el caso particular del Ejemplo 4.1.1 el ordenamiento de interés es sobre las medias $\mu_i, i = 1, 2, ..., 600$, y por tanto, se utilizarán las medias muestrales $Y_i = \bar{X}_i$ como estadísticas resumen. El software tiene implementado, hasta ahora, el cálculo automático de las medias muestrales (con y sin ponderación) como estadísticas resumen. Sin embargo, no es difícil hacer las modificaciones correspondientes si se deseara utilizar estadísticas resumen de otra naturaleza. El histograma de estadísticas resumen para el Ejemplo 4.1.1 se muestra en la Figura 4.3.

- **Gráfico de Dispersión Medias *vs.* Desviaciones Estándar.**
El supuesto de varianza constante entre poblaciones es uno de los más importantes dentro de la teoría de selección y ordenamiento (Cui and Wilson [2008]). Si se calculan las desviaciones estándar muestrales como estimadores de las desviaciones estándar reales (que en la práctica la mayoría de las veces son desconocidas), es posible realizar un gráfico de medias *vs.* varianzas de manera que se aprecie si la varianza en realidad se puede asumir constante entre poblaciones o si, como función de la media, presenta algún comportamiento en particular.

  La Figura 4.3 muestra este tipo de gráfico para el Ejemplo 4.1.1 en el cual puede apreciarse que la media y la desviación estándar no parecen estar correlacionadas para este caso, y por lo tanto, suponer que la varianza es constante es razonable (de hecho lo es ya que se simuló igual para todas las poblaciones). Es importante notar que, para este conjunto de datos no es posible inferir con alto grado de precisión que la verdadera varianza es 1. Esto se debe a que únicamente se simularon $n = 3$ observaciones para cada población, lo cual aumenta la varianza de s^2 como estimador de σ^2 y esto se ve reflejado en el gráfico.

  En el caso en el que el gráfico de medias contra desviaciones estándar de evidencia de correlación (lineal, cóncava o convexa), una alternativa simple es la aplicación de una transformación a la base de datos. El software puede permitirle al usuario explorar distintas transformaciones potencia, indexadas mediante un parámetro λ y observar los efectos que éstas tienen sobre los gráficos exploratorios y de diagnóstico en tiempo real. Estas transformaciones son de la familia Box-Cox (Box and Cox [1964]) y está demostrado que pueden, bajo ciertas condiciones, ayudar a estabilizar la varianza (véase la Sección 2.3).

- **Gráfico Lineal Tratamientos *vs.* Desviaciones Estándar.** En el caso en el que la varianza es por lo general estable para todas las poblaciones excepto para un (generalmente reducido) conjunto de ellas, puede ser de interés para el usuario detectar cuántas y cuáles de ellas están incumpliendo el supuesto, de manera que pueda revisarse o incluso, de ser necesario, se eliminen del estudio. El gráfico de tratamientos contra desviaciones estándar pretende auxiliar al usuario en la identificación de clases atípicas que incumplan flagrantemente el supuesto de varianza constante.

4.1.4. Procedimientos de Selección

En la Sección 2.3 se reseñaron las metodologías modernas de selección y ordenamiento de Cui and Wilson [2008] y Cui et al. [2010]: la $d-$selección, la $G-$selección y la $c-$selección. Parte de la aportación de Cui and Wilson [2008] fue la implementación de dichas metodologías en una librería de R, la cual fue nombrada PCS. La principal función de dicha librería es calcular el estimador para PCS para distintos parámetros, supuestos distribucionales y tamaños de selección. Para esto hace uso, principalmente, de algoritmos bootstrap paramétricos y no paramétricos (Davison and Hinkley [1997]) como el que se resume en la Sección 2.3.5.

El software diseñado y desarrollado para este trabajo, se basa en los algoritmos de la librería PCS de manera que se permita al usuario darle una interpretación gráfica más intuitiva a los resultados. Para

realizar un análisis estadístico de selección y ordenamiento el usuario requiere pre-especificar tres tipos de parámetros:

1. **Parámetros de transformación.** El usuario especifica si desea realizar una transformación a la base de datos antes de realizar el análisis, para efectos de estabilidad de varianza o para buscar las peores poblaciones del conjunto (Figura 4.1(2)).

2. **Parámetros de selección.** El usuario indica el tipo de selección que desea efectuar y el parámetro correspondiente ($d/G/r/P*$) según corresponda (Figura 4.1(3)). Especifica además el supuesto distribucional correspondiente para la distribución F de las estadísticas resumen Y_i (incluye, por el momento, sólo soporte para la Normal y $t-$Student). Finalmente, es necesario especificar el tamaño de la selección, t, que se desea hacer $1 \leq t \leq k$. Si el usuario desconoce el valor de t puede pedirle al software que estime PCS_t para un rango de valores de t y los grafique (Figura 4.1(10)).

3. **Parámetros del bootstrap.** Información relacionada con el remuestreo como el número de muestras bootstrap deseadas B y el tipo de boostrap que se desea implementar (paramétrico o no paramétrico). Es importante notar que mientras mayor sea B mayor será el tiempo de corrida (Figura 4.1(4)).

La salida principal del programa es un gráfico de t *vs.* $\hat{\text{PCS}}_t$ para el rango de valores de t pre-especificado por el usuario. Este gráfico se interpreta buscando el valor de t que maximiza $\hat{\text{PCS}}_t$ para el conjunto de datos particular, de esta manera se encontrará qué tamaño de selección conviene hacer para garantizar la máxima probabilidad de selección correcta posible dentro del rango de valores que el usuario ha pre-especificado.

Las Figuras 4.4–4.6 muestran las salidas para el caso del Ejemplo 4.1.1 para los tres tipos de selección a distintos valores de sus parámetros. Por simplicidad se han agrupado todos los gráficos correspondientes al mismo tipo de selección en una misma ventana pero en la práctica se analizan de manera individual. Un ejemplo de interpretación para cada gráfico en este caso se muestra a continuación:

- ***d*−selección.** La Figura 4.4 ilustra que mientras más pequeño sea el parámetro d más estricto se vuelve el criterio de selección y por tanto disminuye ${}_d\text{PCS}_t$. Para este caso particular donde existe una marcada diferencia entre las medias de las poblaciones puede verse que, incluso para valores pequeños de d, la gráfica tiene puntos máximos locales en $t = 200$ y $t = 400$ que, efectivamente, son los puntos que maximizarían la probabilidad de selección correcta.

- ***G*−selección.** La Figura 4.5 muestra que, similarmente al caso de la $d-$selección, todas las gráficas tienen máximos locales en áreas cercanas a $t = 200$ y $t = 400$. Naturalmente, mientras mayor sea G mayor será el grando de tolerancia permisible y más amplia será la región alrededor de estos puntos con alta PCS. Por ejemplo, para $G = 1$ se obtuvieron máximos locales en $t = 199, 200, 399, 400$ pues al seleccionar las $399 + 1 = 400$ poblaciones se tendrá una alta probabilidad de haber seleccionado las $t = 399$ mejores y así sucesivamente.

- ***c*−selección.** La Figura 4.6 muestra cómo al aumentar el parámetro r, el rango de selección con máxima PCS se reduce hasta que, para el caso más extremo $r = 0{,}99$ se forma una curva con dos máximos locales muy marcados en $t = 200$ y $t = 400$. Es interesante notar cómo se pueden alcanzar valores de PCS relativamente altos en el intervalo $[400 - 600]$ (PCS$>$ 0,7) incluso para $r = 0{,}95$.

El software está programado además para devolverle al usuario un archivo de extensión .csv con los puntos de las gráficas de t *vs.* $\hat{\text{PCS}}_t$. Éste puede ser empleado para un análisis comparativo más detallado

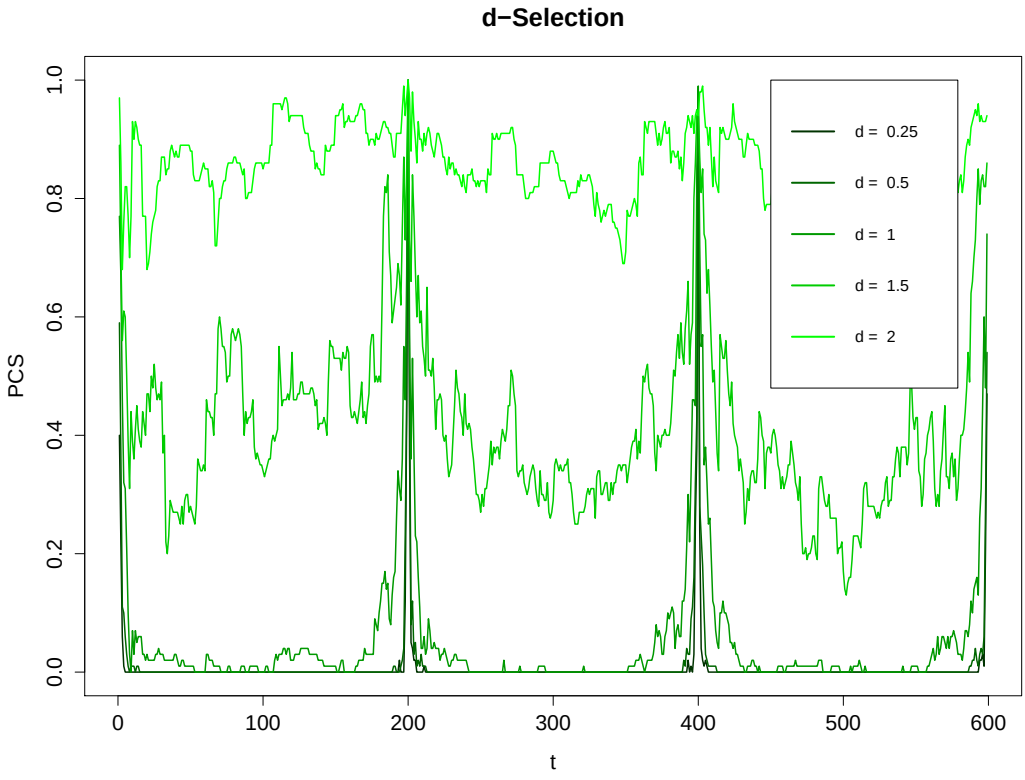

Figura 4.4: Gráfico de t contra ${}_d\hat{\mathrm{PCS}}_t$ producido por el software para el Ejemplo 4.1.1 y para distintos valores del parámetro d.

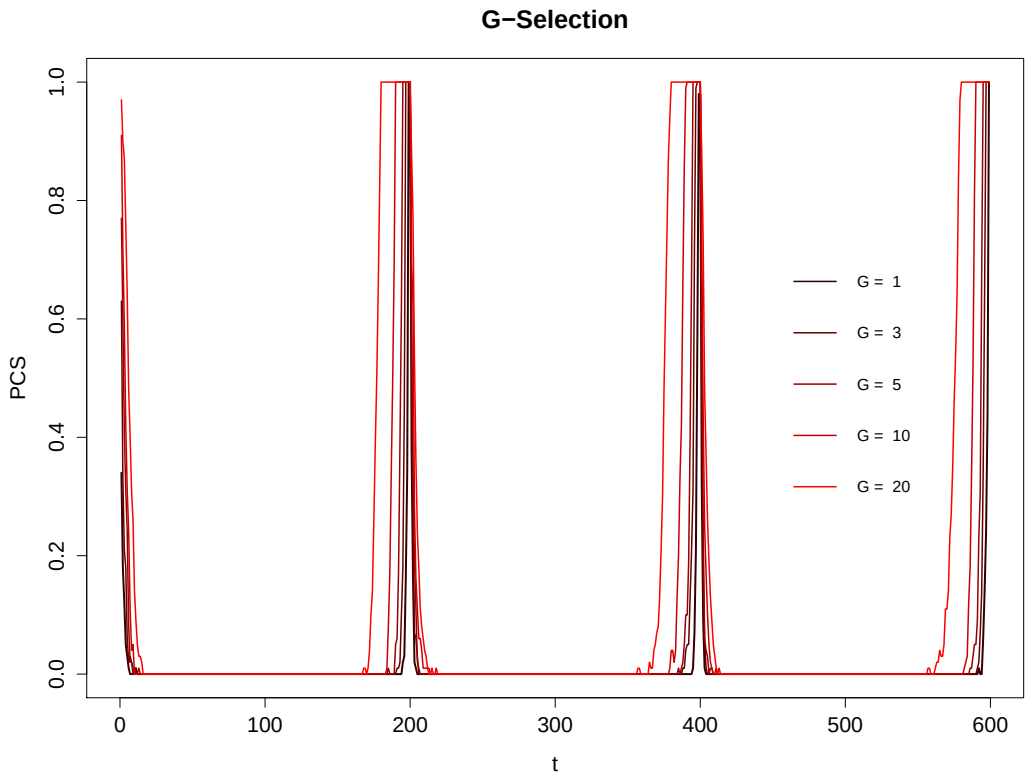

Figura 4.5: Gráfico de t contra $\text{PCS}_{G,t}$ producido por el software para el Ejemplo 4.1.1 y para distintos valores del parámetro G.

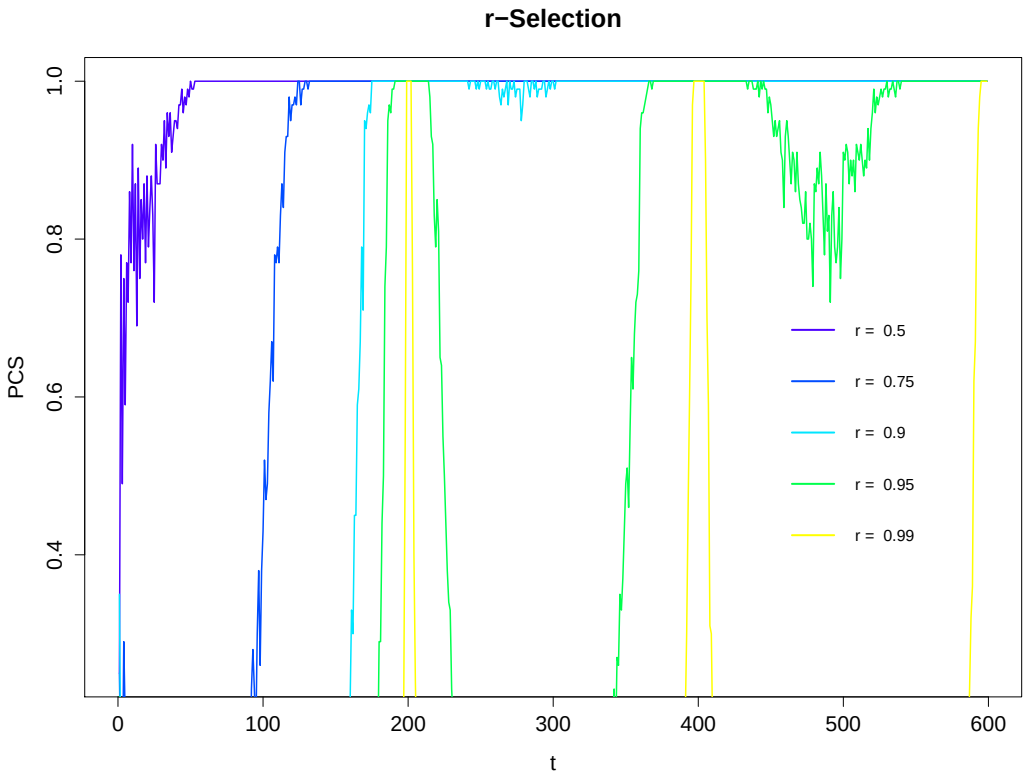

Figura 4.6: Gráfico de t contra $\hat{\text{PCS}}_{\lceil rt \rceil, t}$ producido por el software para el Ejemplo 4.1.1 y para distintos valores del parámetro $r = c/t$.

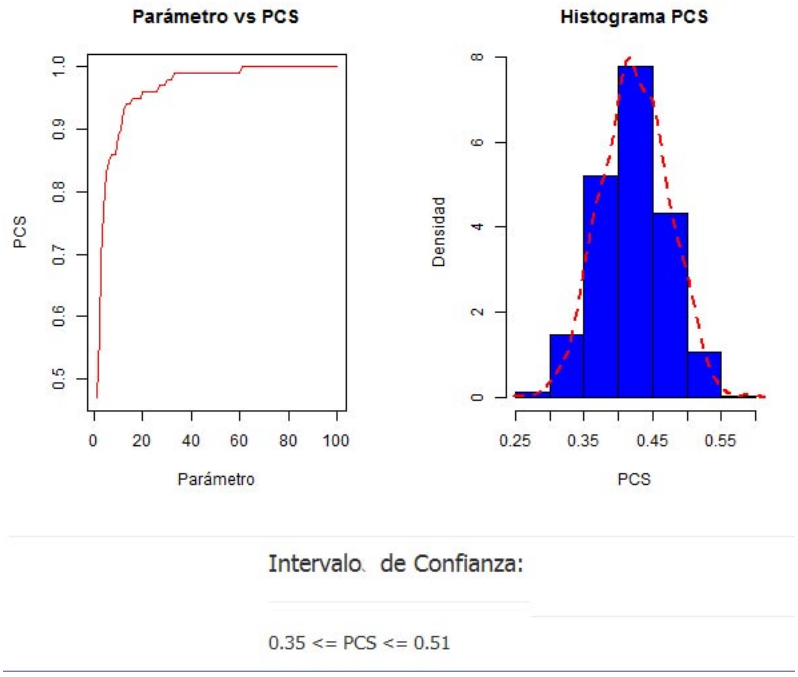

Figura 4.7: (Izquierda) Gráfico de G *vs.* $\hat{\text{PCS}}_{G,t}$ para $t = 120$. (Derecha) Histograma de $\hat{\text{PCS}}_{G=50,t=120}$ obtenido tras calcular 1000 veces el estimador. (Abajo) Intervalo de confianza para $\hat{\text{PCS}}_{G=50,t=120}$ obtenido de la distribución empírica de las 1000 observaciones del estimador.

o para reproducir los gráficos utilizando cualquier otro software más sofisticado (por ejemplo Matlab o incluso Excel).

Para la elaboración de una gráfica de t *vs.* $\hat{\text{PCS}}_t$ se requiere que el usuario fije un valor para el parámetro de selección $d/G/r$, y a continuación, mediante un recurso gráfico como las Figuras 4.4-4.6, tomar una decisión acerca de un valor óptimo para t que maximizará la PCS para el conjunto de datos particular. Es posible, por el contrario, fijar un valor específico para t y realizar un gráfico del parámetro de selección $d/G/r$ *vs.* $\hat{\text{PCS}}_t$, de manera que se tome una decisión acerca de qué parámetro optimiza la PCS para ese tamaño de selección fijo. Nótese que haciendo esto la pregunta cambia de «¿Cuántas poblaciones se deben escoger para obtener una máxima PCS bajo cierto criterio?» a «¿Cuál debe ser el criterio mínimo que se debe adoptar para maximiar PCS si se quiere hacer una selección de las mejores t poblaciones (para t fijo)?». En general la segunda pregunta es más específica y por tanto menos útil en la práctica pero el software ofrece la posibilidad de obtener el gráfico correspondiente (Figura 4.1(8)).

Una última función para el contexto de selección que ofrece el software es el cálculo de histogramas de $\hat{\text{PCS}}_t$ e intervalos de confianza para PCS_t para valores específicos de t y del parámetro de selección $d/G/r$. Un ejemplo de la aplicación de esta herramienta a un caso de estudio real puede consultarse en la siguiente sección. La Figura 4.7 muestra un ejemplo de la salida presentada en la Figura 4.1(8) aplicada al Ejemplo 4.1.1. En ella se muestra el gráfico del parámetro G *vs.* $\hat{\text{PCS}}_{G,t}$ para $t = 120$ donde puede verse el comportamiento creciente de la curva resultante, se puede inferir que para encontrar a las mejores $t = 300$ poblaciones en este caso, el experimentador debe estar dispuesto a escoger hasta $G \approx 70$ poblaciones extra. Se incluye además el gráfico del histograma de $\hat{\text{PCS}}_{G=50,t=120}$ y el intervalo de confianza aproximado obtenido de él.

4.1.5. Pruebas de Hipótesis Múltiples

Para propósitos compárativos, el software cuenta con la posibilidad de realizar una prueba de hipótesis múltiple (PHM) únicamente para pruebas de la forma:

$$H_{0i} : \mu_i = c \text{ para } i = 1, 2, ..., m, \tag{4.1}$$

donde c es una constante fija conocida y m es el número de hipótesis que se desea probar en simultáneo. Cabe aclarar que esta sección no se pensó como una herramienta interactiva para PHM *per se*, pues una prueba de hipótesis múltiple puede ser de formas mucho más generales que (4.1) y tiene procedimientos más sofisticados que los que aquí se incluyen. Esta sección está basada en los algoritmos contenidos en la librería *multtest* de R y su principal función es implementar los métodos de control de la FWER, la FDR, la TPPPFP y la gFWER expuestos en la Sección 2.1 de manera que, al igual que para las técnicas de RSM, se produzca un gráfico que permita al usuario interactuar con estos conceptos y tomar decisiones con base en ellos.

Los detalles de esta sección y los tipos de salidas que es capaz de producir se ilustrarán mediante un ejemplo en la Sección 4.2.4.

4.2. Caso de Estudio

El caso de aplicación de la tesis consiste de una problemática real específica planteada por investigadores del Laboratorio de Biología en Sistemas Genéticos, parte del Laboratorio Nacional de Genómica y Biodiversidad (Langebio). Consiste de un laboratorio interdisciplinario que estudia las funciones de los genes y la evolución a un nivel sistemático. Su principal tarea consiste en analizar grandes conjuntos de mutantes (de levadura principalmente) en un esfuerzo sistemático para entender cómo los genes, su entorno y sus interacciones influyen en procesos complejos como el crecimiento, la supervivencia y la longevidad.

El problema específico consiste en la selección de los genes del genoma de la levadura que tienen una mayor influencia (positiva o negativa) en la longevidad (tiempo de vida) del organismo. Se describirán a continuación el contexto técnico del problema, su relevancia científica y el conjunto de datos específico con el que se trabajó así como una reseña del proceso de muestreo para su obtención. Se detallará además la implementación de la metodología descrita en el Capítulo 2 mediante la aplicación de la herramienta computacional descrita en la Sección 4.1. Se finalizará con la presentación de los resultados y las recomendaciones correspondientes.

4.2.1. Descripción del Contexto

La levadura es un hongo unicelular del cual se conocen en la actualidad más de 1500 especies distintas. La especie más conocida, la levadura de cerveza *Saccharomyces cerevisiae*, es un organismo eucariota, es decir, su célula cuenta con un núcleo y otros organelos con funciones complejas propias dentro de su membrana, de manera análoga a las células de organismos más complejos como los animales y los vegetales. A pesar de esto, la estructura biológica de una célula de levadura es sólo ligeramente más compleja que la de un organismo procariote simple como una bacteria. Esto hace que sea uno de los organismos modelo más adecuados para el estudio de problemas biológicos.

S. cerevisiae fue el primer organismo cuyo genoma completo fue secuenciado. En otras palabras, se conoce la estructura completa y la secuencia de las cadenas de ADN que forman su material genético de manera que es posible enumerar e identificar cada uno de los genes que lo componen. Esto ha permitido la manipulación total de cada uno de sus genes en simultáneo mediante herramientas sofisticadas como microarreglos. Estos permiten estudiar aspectos como la expresión genética (qué tan activo está cada

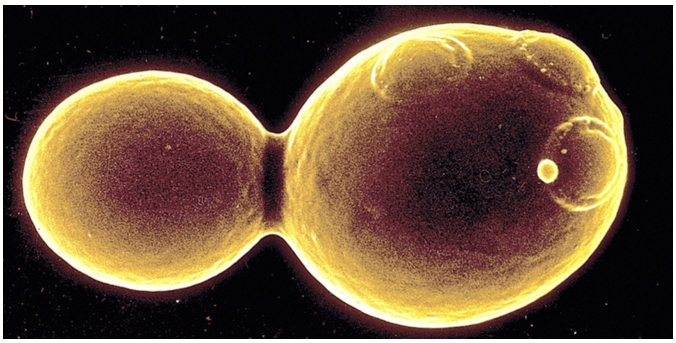

Figura 4.8: Célula de *S. cerevisiae* en la cual se puede apreciar el proceso de gemación y las cicatrices resultantes. (Tomada de http://www.brewshop.co.nz)

gen), la organización funcional (cómo se *agrupan* los genes) y la producción de proteínas (qué función específica desempeña cada gen). Lo anterior coloca a *S. cerevisiae* como uno de los organismos más importantes para el estudio de la biología celular en la actualidad.

Secuenciar el genoma completo de un organismo no es una tarea rápida ni sencilla y se complica mientras se incrementa la complejidad del organismo. Algunos ejemplos de genomas completados o que se espera que sean completados parcialmente en periodos próximos incluyen una gran variedad de gusanos, la vaca, el caballo, el tomate e incluso el hombre. Un hecho importante es que la maquinaria molecular de muchos procesos celulares relevantes es común tanto en vegetales, animales e incluso la levadura. Por ello estudiar el funcionamiento celular en *S. cerevisiae* puede traducirse a importantes descubrimientos en organismos más complejos cuyos genomas no han sido secuenciados completamente y no se espera que sean finalizados en algún momento cercano.

Trabajar con *S. cerevisiae* es, por otra parte, relativamente sencillo por su facilidad de cultivo y velocidad de reproducción. Además, su ausencia de propiedades patológicas implican que puede manipularse con mínimas precauciones. El uso más conocido fuera de la investigación de *S. cerevisiae* es en la industria donde se usa para la fabricación de vinos, cervezas y productos horneados debido a su habilidad para producir dióxido de carbono y etanol al fermentarse.

La característica de interés para este caso particular es la longevidad de la levadura en términos de sus niveles de expresión genética. Es decir, es de interés determinar cuáles genes de su genoma tienen un rol más activo (o mayor influencia) en el tiempo de supervivencia del organismo. Para hacer esto se prepararon una serie de organismos *mutantes*, cada uno de los cuales tiene uno (o más) de sus genes en estado *inactivo* (o *knockouts*), de manera que su tiempo de vida se pueda contrastar con el de organismos cuyo genoma está intacto o silvestres (*wild type (WT)*).

Está reportado en la literatura (véase por ejemplo Longo et al. [2010] y De Luna et al. [2014]) que la longevidad de la levadura no está explicada únicamente en términos de los niveles de expresión genética sino también en términos de otros factores, tanto internos como externos a la célula. La humedad, la temperatura, la cantidad de alimento (glucosa) en el cultivo, la autofagia celular, la presencia o ausencia de estimulación sensorial, entre otros son algunos de los factores que se ha encontrado que tienen mayor influencia en el tiempo de vida de este organismo. Un análisis sistemático de la influencia de estos factores y otras variantes en el tiempo de vida de *S. cerevisiae*, así como la interacción entre ellos se reporta en De Luna et al. [2014].

Existen distintas formas de medir el tiempo de vida de una célula de *S. cerevisiae* dependiendo del interés particular con el que se hace el estudio. Una de las maneras más comúnes en la práctica de hacerlo es a través de su ciclo de vida replicativo (RLS) que consiste en medir el número de veces que una célula se divide antes de su muerte. Una característica particular de las células de *S. cerevisiae* es que se reproducen rápidamente mediante un mecanismo asexual llamado *gemación (budding)*, en el cual

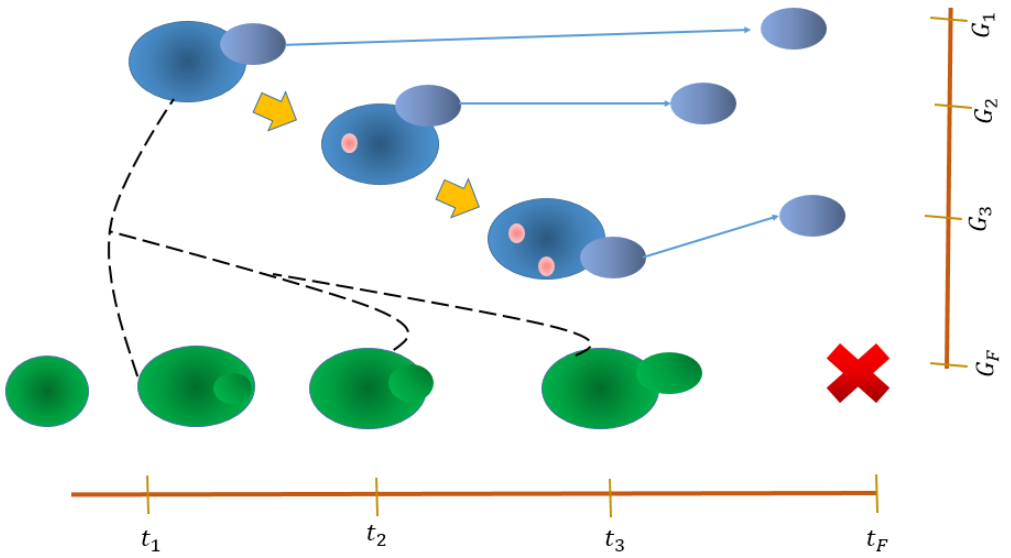

Figura 4.9: Representación gráfica del RLS y el CLS. El RLS (arriba-derecha) se mide en números naturales (número de células hijas) mientras que el CLS (abajo) se mide en la escala temporal.

un nuevo individuo se desarrolla en un sitio particular de la célula de su progenitor, separándose una vez que está maduro dejando una cicatriz en su lugar (Figura 4.8). Esta es la principal razón por la que una célula de *S. cerevisiae* puede dividirse un número limitado de veces (en promedio 24–26 veces) e intuitivamente puede usarse su RLS como forma de medición de longevidad.

En general, el RLS es relativamente fácil de medir. Sin embargo, es altamente sensible a cambios en el método y las condiciones de cultivo. Además, su interpretación en términos de longevidad de células de organismos eucariotes más complejos como animales y vegetales no es directa. Esto se debe a que distintos organismos tienen patrones particulares de división celular (Longo et al. [2010]).

Un método alternativo al RLS comúnmente utilizado para medir la longevidad de una célula de *S. cerevisiae* es su ciclo de vida cronológico (CLS). El CLS se mide en la escala temporal y representa el tiempo de vida de una célula durante su *fase estacionaria*. Es decir, el tiempo que pasa en *reposo* antes de la división celular. En comparación con el RLS, el CLS tiene una interpretación más generalizable a organismos eucariotes más complejos. No obstante, en muchos casos puede ser complicado medirlo con precisión.

La Figura 4.9 muestra una represantación gráfica comparativa del CLS y el RLS para una célula de *S. cerevisiae*. Es importante notar que no se miden en la misma escala y no tienen la misma interpretación. En general el método de medición de longevidad depende del objetivo del estudio particular. Para el caso de estudio de la tesis, De Luna et al. [2014] propone un nuevo método para medir la longevidad de las células mutantes de *S. cerevisiae* basado en el análisis sistemático de co-cultivos de colonias de levadura de tipo silvestre (WT) y mutante (Δx). Los detalles del proceso de medición se darán en la siguiente sección.

El estudio de la longevidad en *S. cerevisiae* forma actualmente parte de la frontera de la investigación en biología no solo por sus potenciales aplicaciones sino por el gran número de problemas abiertos que plantea. Particularmente, la causalidad de la vejez permanece un problema sin resolver hasta la actualidad, en especial en organismos complejos como el hombre. El estudio de la longevidad en *S. cerevisiae* promete ser una de las alternativas más viables para la comprensión y el estudio de dicho fenómeno biológico. Existen a su vez áreas de oportunidad para la estadística, en particular en tareas relacionadas con la cuantificación de incertidumbre de las mediciones, el modelado del proceso de división

celular, técnicas de agrupamiento, selección y análisis discriminante de genes, realización de pronósticos, *etc.*

La posibilidad de extrapolación de los resultados del estudio de la longevidad de *S. cerevisiae* a organismos eucariotes más complejos abre las posibilidades a potenciales aplicaciones interesantes. En particular destaca el desarrollo de fármacos que permitan regular el mecanismo de envejecimiento celular y brindar una mejor calidad de vida a organismos de avanzada edad. Para el caso particular del hombre ofrece la posibilidad del desarrollo de métodos para la prevención y el tratamiento de enfermedades relacionadas con la vejez.

4.2.2. Proceso de Muestreo y Tabla de Datos

La Figura 4.10, extraída de De Luna et al. [2014], muestra el esquema del proceso de medición de longevidad propuesto en dicha publicación. Consiste de una estrategia de medición del CLS de la levadura que permite un análisis cuantitativo de los tiempos de vida de las mutantes (en la escala temporal).

El proceso consta de tres partes:

1. Dos tipos de *cepas* (variantes genéticas) de células de levadura se colocan en co-cultivos en cantidades iguales en 96 placas bajo condiciones específicas. El primer tipo consiste de células de referencia cuyo material genético está intacto, o silvestres (WT), mientras que el segundo tipo corresponde a una mutante o *knockout* específico que denotaremos por Δx. Cada uno de los tipos de cepas fue previamente *etiquetado* mediante una proteína fluorescente que permite distinguir los grupos entre sí. Esto se hace mediante un microscopio de fluorescencia que detecta las señales luminosas emitidas por las células expuestas a la proteína. (Figura 4.10-A)

2. Después de un tiempo, luego de que los cultivos han llegado a fase estacionaria (dejaron de crecer de manera explícita debido al agotamiento de los nutrientes en el medio), se inoculan de manera automática y se llevan a un medio fresco donde se monitorea su crecimiento nuevamente, y este procedimiento se repite varias veces (7–10 o más). Es de esperarse que con el paso del tiempo algunas células dejen de ser viables (ya no se reproduzcan o mueran) y esto se verá reflejado en los cambios en las señales de fluorescencia emitidas. De aquí que, la medida de especial interés es el cambio relativo de la señal de fluorescencia emitida por las células Δx en comparación con la de las células WT (llave magenta en la Figura 4.10-B). Este cambio se aproxima mediante interpolación a un tiempo fijo (línea vertical cyan) bajo un modelo exponencial (los detalles se darán más adelante).

3. La razón de cambio de fluorescencia como función del tiempo (la edad en días) se utiliza para obtener un coeficiente de supervivencia s que se usa como un estimador del tiempo relativo de vida de la mutante Δx en comparación con la célula silvestre WT (Figura 4.10-C).

La parte clave del procedimiento descrito anteriormente mediante la Figura 4.10 es la obtención del coeficiente de supervivencia s. Esta cantidad se obtiene mediante la comparación sucesiva del crecimiento de dos poblaciones que se estudian en co-cultivo (WT y Δx). De esta manera, si en el momento en el que el tamaño de ambas poblaciones se estabiliza (fase estacionaria), el tamaño de la población WT es estrictamente mayor que el de Δx habrá evidencia de que el número de células viables (con capacidad de reproducirse) de células WT está tendiendo a ser mayor que el de células mutantes. Lo anterior, sumado al hecho de que se comenzó con el mismo número de células de ambos tipos, implica que las células WT *sobrevivieron* por más tiempo en comparación con las mutantes. El caso es análogo si la situación se invierte. En otras palabras, la metodología de De Luna et al. [2014] implica la utilización del número de células viables a un determinado tiempo t como un *proxy* para hacer inferencia acerca de la supervivencia relativa de cada mutante.

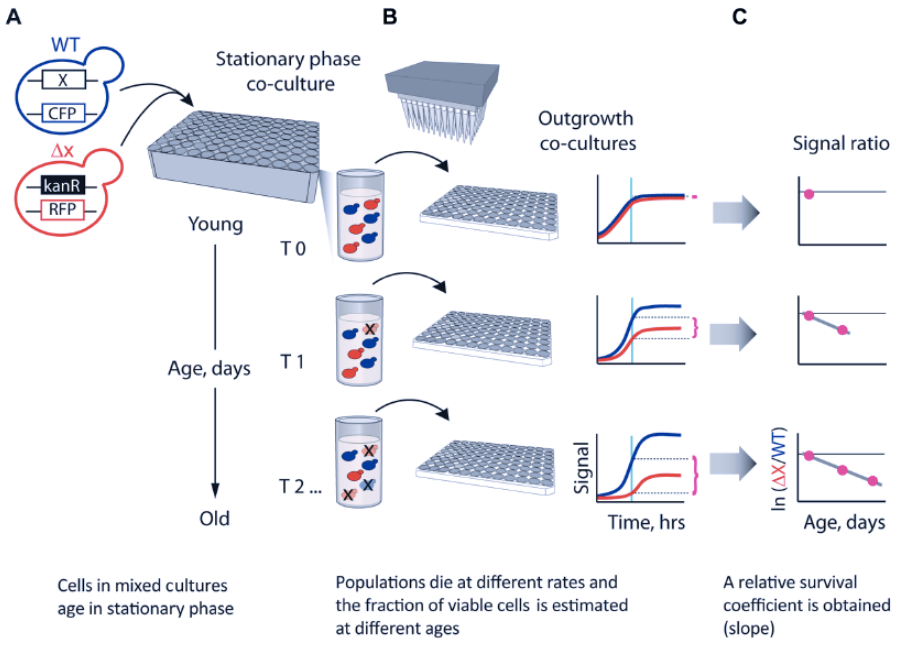

Figura 4.10: Esquema de la estrategia de medición propuesta en De Luna et al. [2014]

Para modelar el crecimiento de las poblaciones de células se supuso en De Luna et al. [2014] un modelo exponencial de la forma:

$$N_x(t) = N_x^0 \exp^{-tg_x(t)} \tag{4.2}$$

donde $N_x(t)$ representa el tamaño (número de células viables) de la población x al tiempo t, N_x^0 es el número inicial de células en el cúltivo y $g_x(t)$ es la tasa de crecimiento de la población x. Nótese que, en el caso más general, la tasa puede depender del tiempo.

Para el caso de De Luna et al. [2014] se tomó el caso más simple que consiste en un modelo exponencial con tasa constante dado por

$$N_x(T) = N_x^0 \exp^{-r_x T}, \tag{4.3}$$

y de esta manera se puede modelar el tamaño de la población x únicamente como función de la edad T.

Luego, si se denota por $N_{wt}(x)$ y $N_x(t)$ el número de células viables de WT y Δx al tiempo t respectivamente, se tiene que el número relativo de células viables de Δx comparado con WT a una edad T es

$$\frac{N_x}{N_{wt}}(T) = \frac{N_x^0}{N_{wt}^0} \exp^{T(r_{wt}-r_x)}. \tag{4.4}$$

Es importante notar que el cociente N_x^0/N_{wt}^0 no depende de la edad T pues se está tomando el modelo exponencial más simple. Aplicando la transformación logaritmo a (4.4) se obtiene una expresión linealizada:

$$\log \frac{N_x}{N_{wt}}(T) = \log \frac{N_x^0}{N_{wt}^0} + T(r_{wt} - r_x). \tag{4.5}$$

Lo que se obtiene es precisamente la medida de interés, el número relativo de células viables como función de la edad T. Si se hace $s_x = r_{wt} - r_x$ y $N_0 = \log \frac{N_x^0}{N_{wt}^0}$ se obtiene el modelo lineal

$$\log \frac{N_x}{N_{wt}}(T) = N_0 + s_x T, \tag{4.6}$$

donde s_x es el coeficiente de supervivencia aparente que engloba la información relacionada con las diferencias en las tasas de crecimiento que se prentenden usar como *proxy* para estimar la longevidad relativa de Δx compado con WT.

La relación (4.6) establece un modelo lineal de regresión respecto de la edad T donde las observaciones de $\log N_x/N_{wt}(T)$ se obtienen de manera experimental mediante extrapolación. Aplicando el algoritmo estándar para la solución de un modelo lineal de regresión se obtiene un estimador $\hat{s_x}$ y su error estándar asociado σ_x y este procedimiento se repitió para cada una de las mutantes.

La interpretación del coeficiente s_x es relativamente simple: valores cercanos a cero indican que la mutante Δx tiene una longevidad muy similar a WT, lo cual se presume da evidencia de que la inactivación del gen que dio lugar a la mutante no causa diferencia en la longevidad del organismo. Análogamente valores grandes de $|s_x|$ indicarían diferencias grandes en los tiempos de vida de Δx en comparación con WT, lo cual daría evidencia de una influencia importante del gen correspondiente en la longevidad del organismo. El signo de s_x determinaría si dicha influencia es negativa o positiva.

La Base de Datos

La base de datos con las que se trabajó en el caso de estudio consta de observaciones correspondientes a 3924 mutantes de levadura e incluye los siguientes atributos para cada mutante:

- **ORF**. Abreviación de *Open Reading Frame*. Representa una secuencia de ADN que índica la *localización* dentro del genoma de la levadura de la secuencia de ADN particular que identifica al gen. Para propósitos de la tesis este atributo no se utiliza, su función es meramente como identificador.
- **Nombre**. El nombre científico que se le asignó al gen en particular.
- **Coeficiente de supervivencia estimado** $\hat{s}$. Es el valor estimado de la pendiente de la recta de regresión (4.6) para cada gen. Por simplicidad se utilizará solo la notación s para denotarlo.
- **Error de ajuste** σ_x. El error estándar de s.

La información anterior se coleccionó en una tabla de tamaño aproximadamente $kn \times 4$. El formato en el cual la base de datos se ingresó al software de la Sección 4.1 se muestra en el Cuadro 4.2, donde M_{ij} representa a la $j-$ésima observación de la $i-$ésima mutante y n_i el número de observaciones disponibles para la $i-$ésima mutante. El software de la Sección 4.1 fue adaptado de tal manera que se pueda transformar automáticamente una tabla con el formato 4.2 en una tabla con el formato 4.1 y se puedan realizar los procedimientos correspondientes sin necesidad de que el usuario modifique manualmente la base de datos para darle el formato estándar.

En este caso particular, las repeticiones n_i no son réplicas experimentales exactas debido a restricciones de tiempo y presupuesto. Sin embargo, representan réplicas realizadas en tres distintos tipos de medios de cultivo: Glutamina, Sulfato y Gaba. El interés del estudio entonces será encontrar la mutante más longeva de manera conjunta en los tres medios.

4.2.3. Análisis Exploratorio

Las siguientes dos secciones presentarán los resultados obtenidos al aplicar la herramienta computacional diseñada para las metodologías de selección y ordenamiento descrita en la Sección 4.1 en el contexto del caso de aplicación. La Figura 4.11 muestra los resultados de la primera parte del análisis exploratorio de los datos. Como se mencionó en la Sección 4.2, la cantidad de interés principal en este caso es el coeficiente de longevidad s. El histograma de s (arriba-izquierda) muestra que la gran mayoría de las observaciones se centran muy cerca de cero y casi la totalidad de los coeficientes observados se

ORF	**Gene**	***s***	**fit.error**
Atributos de la observación M_{11}			
Atributos de la observación M_{12}			
⋮			
Atributos de la observación M_{1n_1}			
Atributos de la observación M_{21}			
⋮			
Atributos de la observación M_{kn_k}			

Cuadro 4.2: Formato de la base de datos tal y como se ingresó en el software descrito en la Sección 4.1 donde cada observación se etiquetó mediante la convención M_{ij} (la $j-$ésima observación de la $i-$ésima mutante).

encuentran en el rango $[-0{,}05, 0{,}05]$, lo cual sugiere que identificar a las mutantes que tienen mayor (o menor) longevidad relativa podría ser relativamente complicado.

El histograma de los errores (Figura 4.11 arriba-derecha) muestra que casi todos los errores son menores a 0,01 y presentan una distribución unimodal con moda en aproximádamente 0,004. En general no se observan estimaciones atípicamente erróneas.

Finalmente, por restricciones experimentales varias, no siempre es posible calcular un estimador para s para cada una de las mutantes en todos los medios (Glutamina, Sulfato y Gaba). Se presenta por tanto un histograma del número de observaciones de s por cada mutante (Figura 4.11 arriba-derecha). Puede verse que la mayoría $\sim 95\,\%$ de las mutantes fueron estudiadas en cada uno de los tres medios al menos una vez. Debido a que uno de los supuestos del procedimiento de selección y ordenamiento que se pretende aplicar es la independencia entre clases, se concluyó que mantener o eliminar del estudio aquellas mutantes que tienen menos de 3 observaciones no debería tener efecto en los resultados. Esto se confirmó mediante un análisis por separado utilizando una reducción de la base de datos.

Sean $s_{ij}, j = 1, 2, ..., n_i$ las observaciones del coeficiente de longevidad de la mutante $x_i, i = 1, 2, ..., k = 3924$ y e_{ij} el error de estimación asociado a s_{ij}. La Figura 4.12 muestra un gráfico de dispersión donde cada punto tiene coordenadas (i, s_{ij}). Mediante este gráfico puede confirmarse que el rango de los datos se encuentra casi totalmente en $[-0{,}05, 0{,}05]$ como se vio en el histograma de s en la Figura 4.11. Además se aprecia, al nivel exploratorio, que no hay evidencia de poblaciones (mutantes) atípicas ni agrupamientos.

Para este caso particular se supuso que $s_{ij}, j = 1, 2, ..., n_i$ son observaciones independientes del coeficiente de longevidad de la mutante x_i que sigue una distribución $G(s - \theta_i), i = 1, 2, ..., 3924$ donde $\theta_i = \mathrm{E}(s_{ij})$. Visto como un problema de selección y ordenamiento (véase las Secciones 2.2 y 2.3) la meta será seleccionar aquellas mutantes que tienen mayor (y menor) coeficiente de longevidad, es decir, aquellas poblaciones x_i cuya distribución corresponde a los mayores (y menores) medias θ_i.

Siguiendo el procedimiento descrito en la Sección 2.3 será necesario el cálculo de un estimador para cada parámetro desconocido θ_i. Dado que θ_i representa la media poblacional de x_i, un estimador razonable es la media muestral $Y_i = \bar{X}_i$. Sin embargo, dado que cada observación s_{ij} tiene asociado un error de estimación e_{ij} se sugiere la utilización de una media ponderada por los recíprocos de los errores, de manera que, mientras más pequeño sea el error de estimación (más precisa sea la observación) mayor sea su contribución a la estimación. De esta manera los estimadores Y_i para θ_i se calcularon de la siguiente manera:

$$Y_i = (\sum_{j=1}^{n_i} w_{ij})^{-1} \sum_{j=1}^{n_i} w_{ij} s_{ij}, \tag{4.7}$$

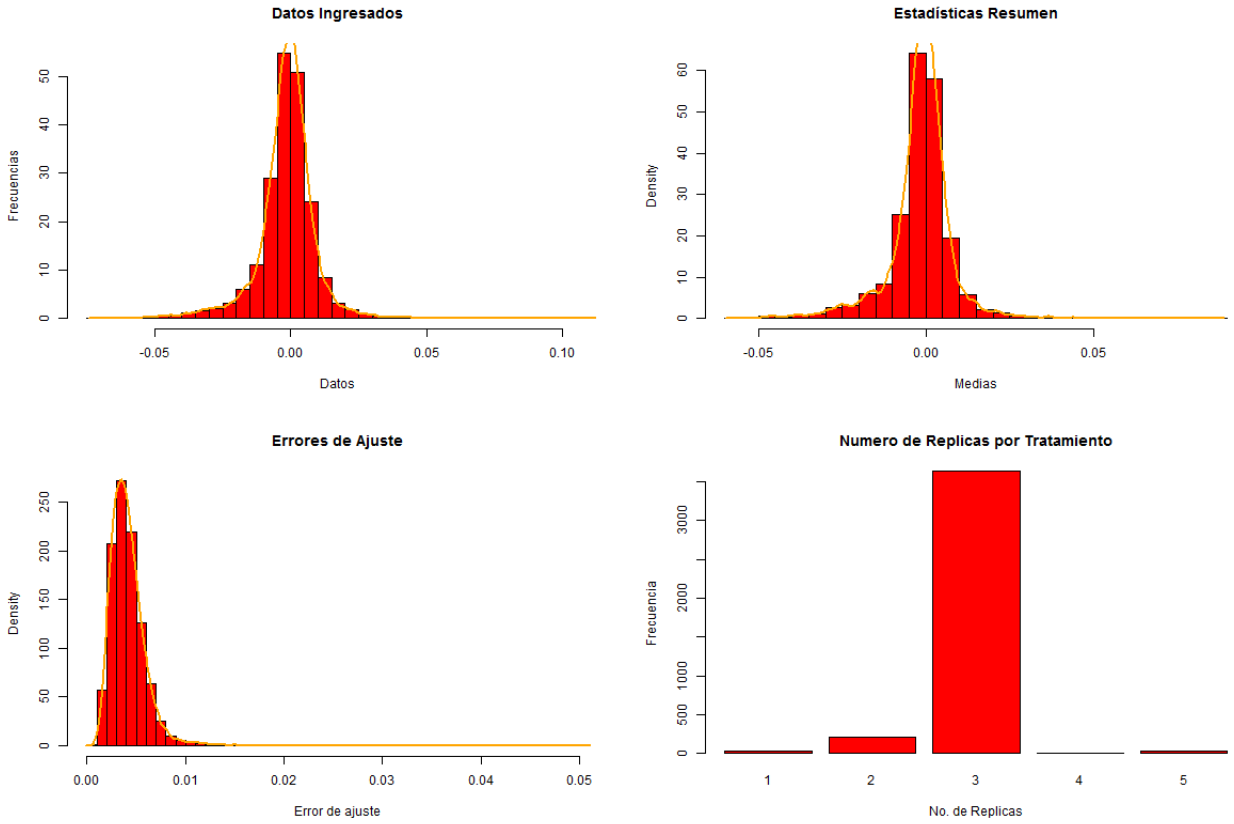

Figura 4.11: (Arriba-Izquierda) Histograma de los coeficientes de longevidad s. (Arriba-Derecha) Histograma de los errores asociados a los coeficientes de longevidad σ_x. (Abajo) Número de observaciones por cada mutante n.

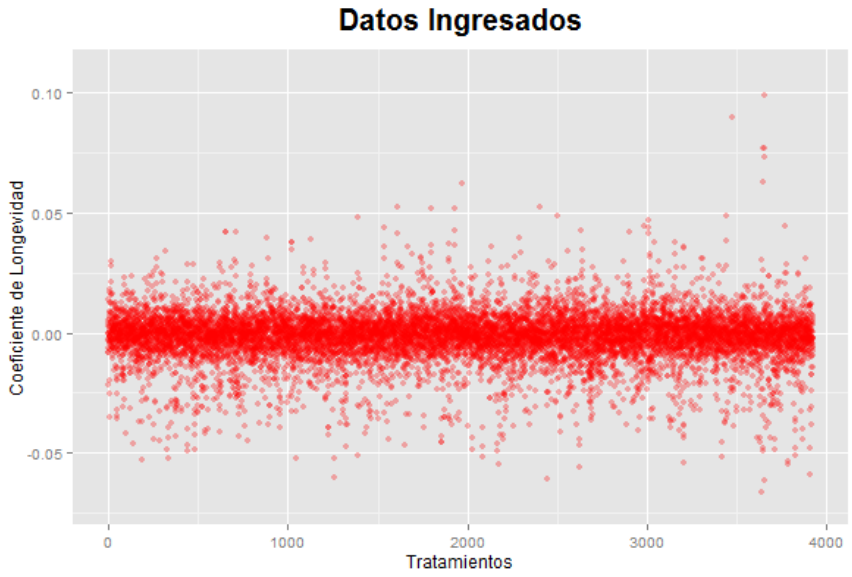

Figura 4.12: Gráfico de dispersión de las observaciones de s *vs.* las mutantes que las produjeron.

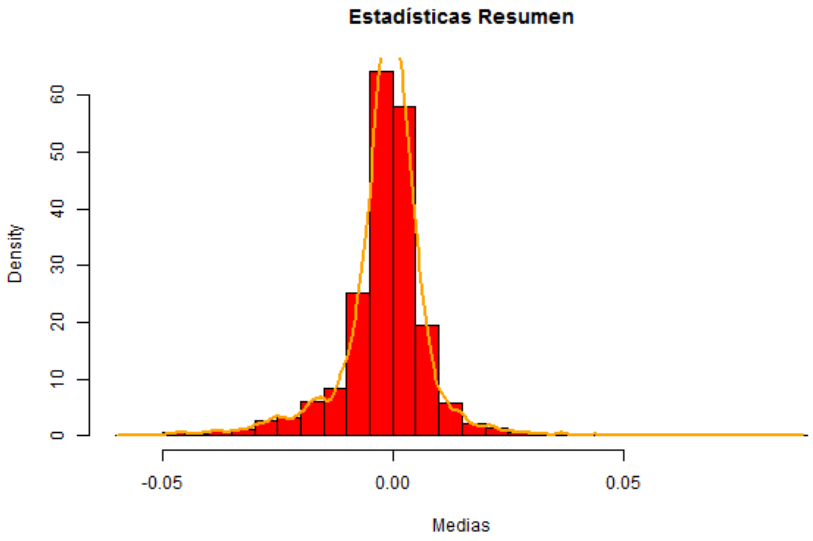

Figura 4.13: Histograma de las medias ponderadas $Y_i, i = 1, 2, ..., 3924$.

donde $w_{ij} = 1/e_{ij}$ y n_i es el número de observaciones correspondientes a la $i-$ésima mutante. La Figura 4.13 muestra un histograma de las medias ponderadas Y_i, donde puede apreciarse que tienen una distribución similar a la de s (Figura 4.11).

Un supuesto importante de la metodología de selección y ordenamiento de Cui and Wilson [2008] es que se tenga varianza común entre poblaciones. Para investigar esto se calcularon las desviaciones estándar muestrales por cada mutante $\hat{\sigma}_i, i = 1, 2, ..., 3924$, y se graficaron en contra de las medias Y_i (Figura 4.14. Puede apreciarse que no existe un patrón particular aparente de los puntos en este gráfico (tendencia creciente o decreciente, linealidad, comportamiento en forma de *cono*, *etc.*), en términos generales la mayoría de los puntos se concentran (en forma dispersa, debido a que únicamente participan 3 observaciones en la estimación de cada desviación estándar) en una zona del gráfico lo cual da evidencia, al nivel exploratorio, de que la varianza no está correlacionada con la media. Lo anterior indica que la varianza puede ser asumida como constante entre poblaciones y descarta la necesidad de pre-transformar la base de datos.

Finalmente, la Figura 4.15 presenta un gráfico donde se comparan las desviaciones estándar muestrales $\hat{\sigma}_i, i = 1, 2, ..., 3924$ por cada población x_i, con la intención de detectar explícitamente poblaciones específicas con varianzas inusualmente grandes. Dicho gráfico confirma que la varianza se encuentra estable entre poblaciones. Para las Figuras 4.14 y 4.15 se omitieron las clases con una sola observación pues su varianza muestral por definición es 0 y no puede usarse como un estimador para la verdadera varianza poblacional σ_i^2.

De acuerdo con las Figuras 4.15 y 4.14, de ahora en adelante se asumirá la varianza σ^2 como constante entre poblaciones y su valor se estimará de manera conjunta mediante una ponderación de todas las clases , tomando

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{3924}(n_i - 1)\hat{\sigma}_i^2}{\sum_{i=1}^{3924}(n_i - 1)} = \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2 + \cdots + (n_{3924} - 1)\hat{\sigma}_{3924}^2}{n_1 + n_2 + \cdots + n_{3924} - 3924}, \tag{4.8}$$

donde $\hat{\sigma}_i$ es la desviación estándar muestral correspondiente a la $i-$ésima población. Nótese que si $n_j = 1$ para algún índice $1 \leq j \leq 3924$ entonces el numerador de (4.8) no incluye el sumando correspondiente a la clase que corresponde a dicho índice. En otras palabras, $\hat{\sigma}^2$ sólo incluye los efectos de aquellas clases que tienen al menos dos observaciones.

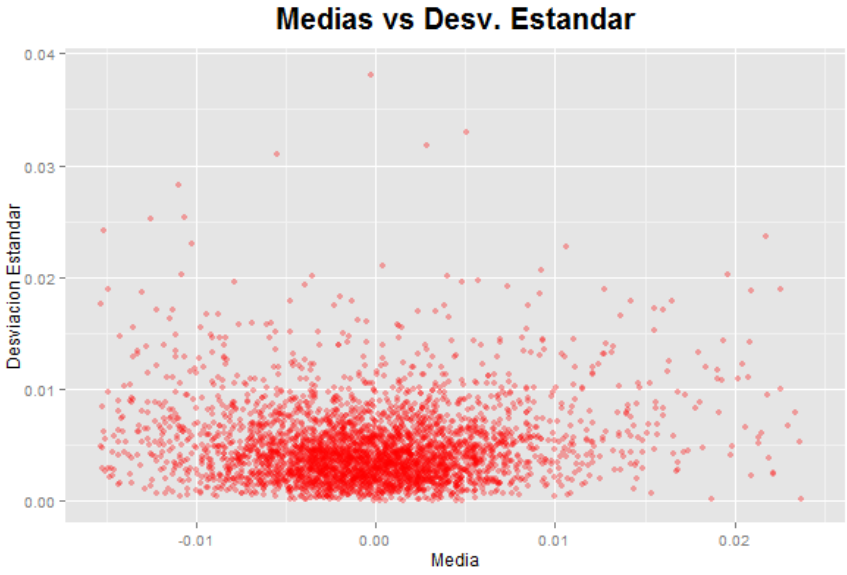

Figura 4.14: Gráfico de dispersión de los puntos $(Y_i, \hat{\sigma}_i)$.

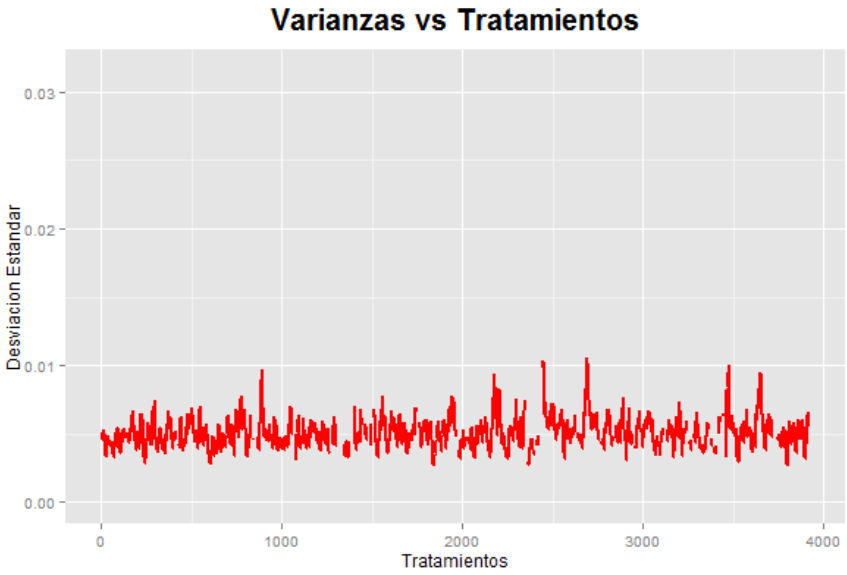

Figura 4.15: Gráfico de $\hat{\sigma}_i$ *vs.* i para $i = 1, 2, ..., 3924$.

4.2.4. Aplicación de la Metodología

En la Sección 2.3 se reseñaron 3 nuevas metodologías de selección propuestas en Cui and Wilson [2008] y Cui et al. [2010] específicamente para los casos en los que el número de poblaciones es grande: la $d-$selección, la $G-$selección y la $c-$selección. Posteriormente, en la Sección 4.1 se presentó una herramienta desarrollada para la implementación computacional de dichas técnicas a conjuntos de datos particulares y se ejemplificaron las principales salidas mediante un caso simulado. A continuación se presentan los resultados de la aplicación de las metodologías al caso de estudio de la longevidad en la levadura.

Cada método de selección se aplicó directamente a la base de datos descrita en la Sección 4.2.2 calculando las estadísticas resumen Y_i y el estimador de la varianza global $\hat{\sigma}^2$ como en (4.7) y (4.8) respectivamente. El procedimiento se hizo en dos fases de acuerdo con los dos objetivos del estudio:

1. Seleccionar las *mejores* mutantes, donde *mejor* se entiende como aquellas que tienen mayores coeficientes de longevidad s.

2. Seleccionar las *peores* mutantes, donde *peor* se entiende como aquellas que tienen menores coeficientes de longevidad.

Debido a lo expuesto en la Sección 2.3.4, no es en general cierto que el segundo objetivo se resuelve al resolver el primero. Por ello, después de resolver el problema correspondiente al primer objetivo, se procedió a pre-multiplicar la base de datos por -1 de manera que las nociones de *peor* y *mejor* se inviertan, y se pueda resolver el objetivo 2 de igual manera como se resolvió el Objetivo 1, con esta base de datos transformada.

Debido a que no se tienen conocimientos previos precisos acerca de un número específico de mutantes *mejores* o *peores* que investigar, no es claro por el contexto qué valor es sensato especificar para el parámetro t. De esta manera, dado que t puede ser cualquier número entero entre $T_{\min} = 1$ y $T_{\max} = 3923$, se procedió a estimar PCS_t para $t = 1, 2, ..., 3924$. Los resultados que se presentan a continuación serán, por tanto, gráficos de t *vs.* PCS_t de manera que se pueda identificar un conjunto de valores óptimos de t que permitan al usuario final hacer una selección a distintos niveles de precisión y calidad. Todas las estimaciones se hicieron con un número de remuestreos bootstrap fijo en $B = 5000$ y con una distribución Normal supuesta para $Y_i, i = 1, 2, ..., 3924$.

$G-$selección

La Figura 4.16 muestra el gráfico de los valores estimados de $\text{PCS}_{G,t}$ como función de t para $G = 1, 3, 5, 10, 50, 100, 500$. Puede apreciarse que, para todos los valores propuestos del parámetro G, $\text{PCS}_{G,t}$ es cero para la mayoría de los valores de t en la parte central. Esto da una idea general acerca de la estructura de la base de datos. Por ejemplo, para $t \in [500, 3000]$, seleccionar las *mejores* t mutantes de manera correcta es un evento muy raro (con probabilidad estimada cercana a 0) incluso aunque se tomen $G = 500$ poblaciones extras. Lo anterior da evidencia de que los parámetros verdaderos θ_i, sobre los cuales queremos inferir un ordenamiento, se encuentran demasiado juntos, para permitir distinguirlos entre sí con alta probabilidad de selección correcta, en el sentido de la $G-$selección.

La Figura 4.17 muestra un acercamiento a la región izquierda de la Figura 4.16 donde puede apreciarse que las curvas correspondientes a valores pequeños de G caen rápidamente incluso para valores pequeños de t; esto parece dar evidencia de que el conjunto de datos sólo permite obtener una buena selección para valores pequeños limitados de t. Las curvas correspondientes a los casos $G = 50, 100, 500$ caen más lentamente que las restantes, pero no representan una mejoría significativa debido a que indican que, para obtener alta PCS con una selección, habría que incluir un gran número de *falsos positivos*. Por ejemplo, para el caso $t = 50, G = 50$ la curva indica alta PCS, pero para ello habría que ceder en tomar una selección de $t + G = 100$ mutantes de las cuales únicamente $t = 50$ serían las mejores.

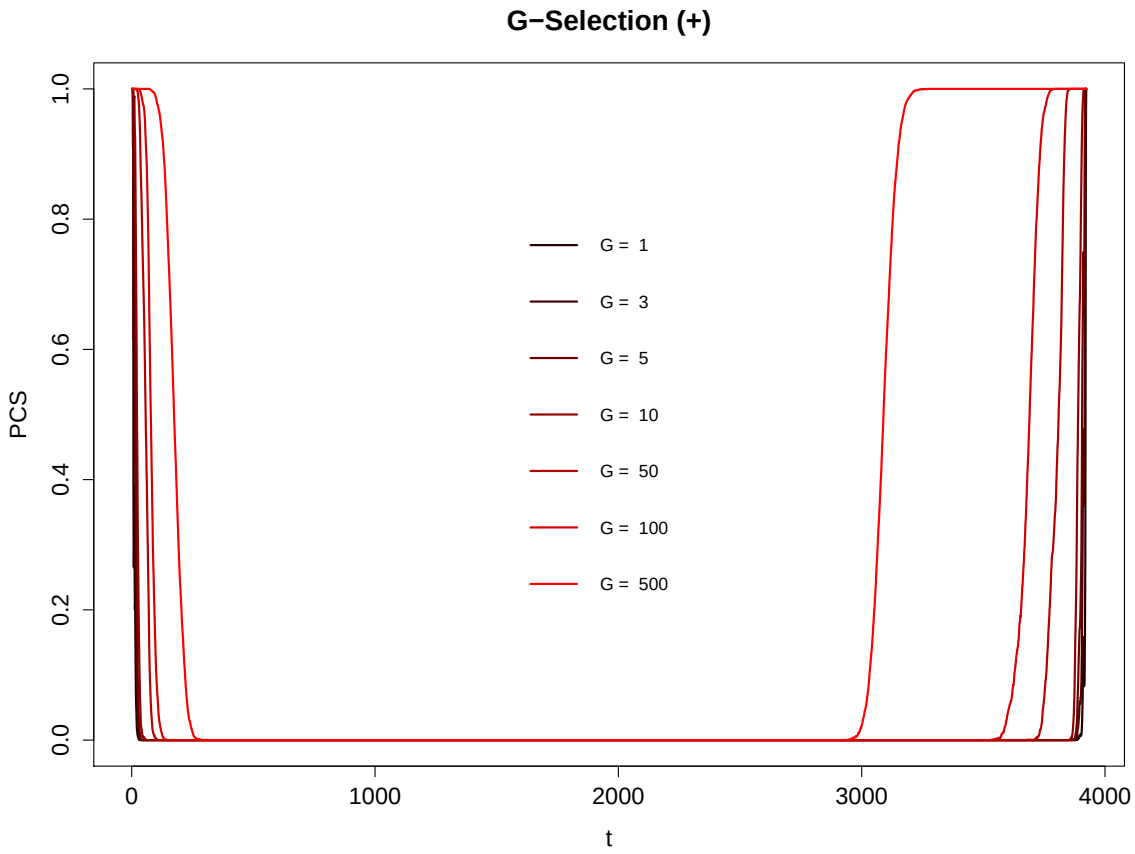

Figura 4.16: Gráfico de t *vs.* $\hat{\text{PCS}}_{G,t}$ para distintos valores de G.

De manera análoga, la Figura 4.18 muestra un acercamiento a la región derecha de la Figura 4.16. En este caso, puede apreciarse que se pueden alcanzar altos valores de PCS para valores muy grandes de t. Sin embargo, estos casos no resultan interesantes ya que una selección demasiado grande en realidad no resuelve el problema.

El experimentador podría estar tentado a pensar que escoger las mejores $t = 3400$ con alta PCS es equivalente a escoger las peores $t = 3924 - 3400 = 524$ con alta PCS. Sin embargo, esto no es necesariamente cierto (véase la Sección 2.3.4). Para comprobarlo, pueden verse las Figuras 4.19–4.21 que son los análogos de las Figuras 4.16–4.18 para la base de datos pre-transformada por un cambio de signo. En general, los resultados para la $G-$selección no difieren demasiado para ambos casos. Un resumen tabulado de los resultados y recomendaciones se presentará en la siguiente sección para distintos valores estratégicos de PCS. La conclusión general de la aplicación de $G-$selección al caso de estudio es que el conjunto de datos es demasiado ruidoso para permitir una $G-$selección de buena calidad (G pequeña) con alta precisión (PCS grande). Cabe aclarar que la aseveración anterior, a pesar de no ser demasiado interesante provee el aprendizaje general acerca del tipo de estructura que tiene la base de datos.

$d-$selección

La $d-$selección implica la pre-especificación de un parámetro $d > 0$ cuya función es determinar una *zona de indiferencia* alrededor de la frontera entre las verdaderas mejores poblaciones y las restantes (véase la Sección 2.3.3 o Cui and Wilson [2008]). De esta manera, se permite que un número (aleatorio)

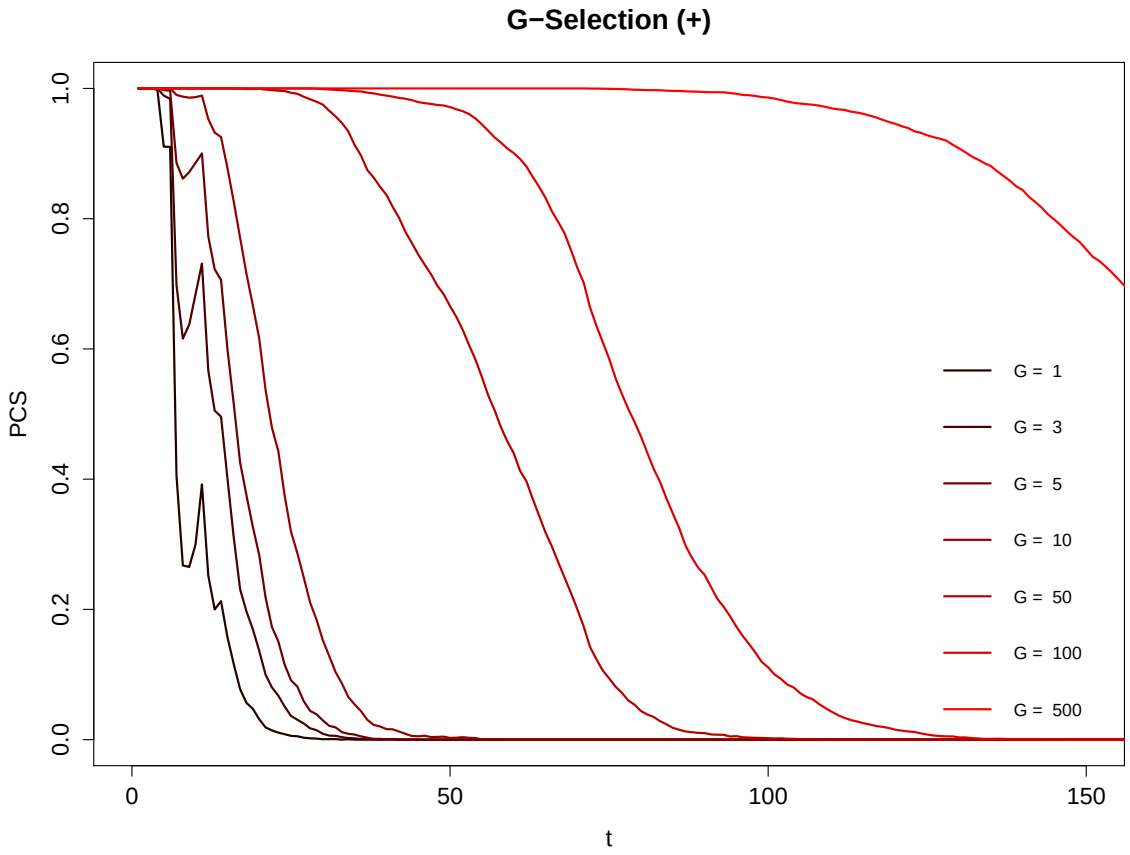

Figura 4.17: Gráfico de t *vs.* $\hat{\text{PCS}}_{G,t}$ para distintos valores de G (Parte Izquierda).

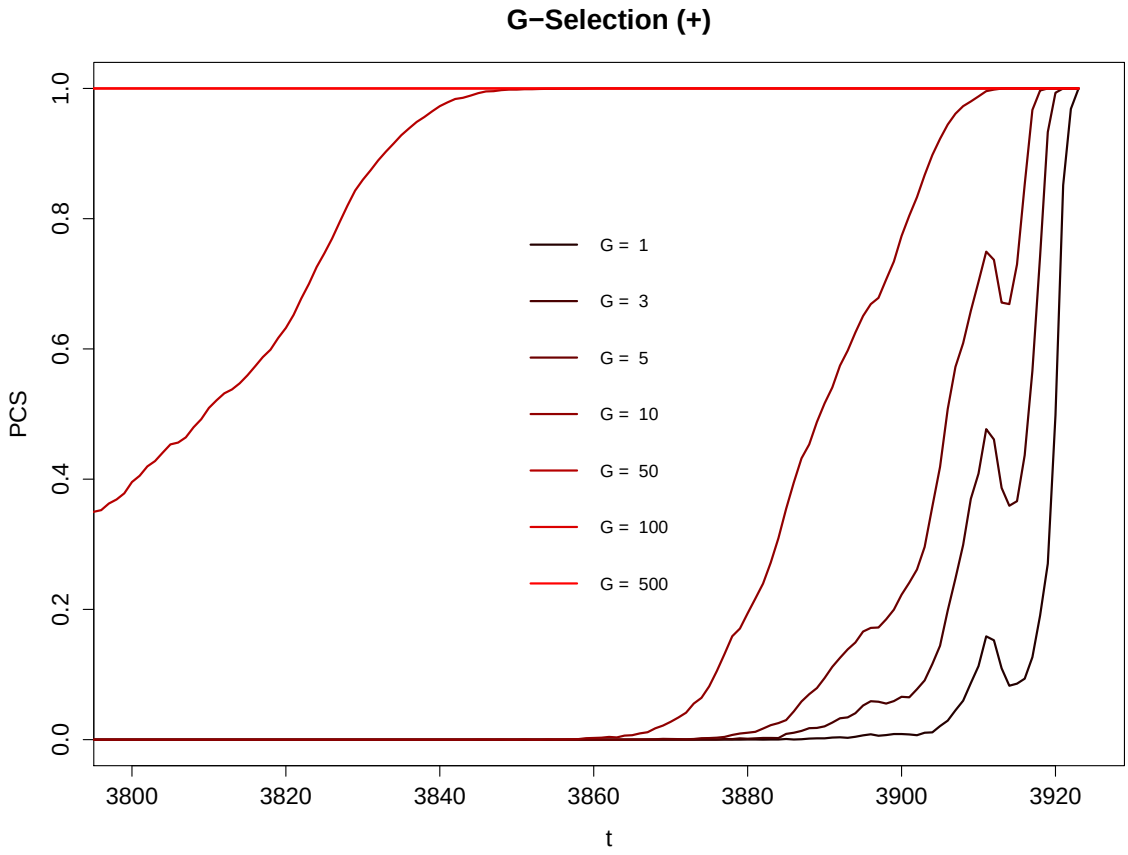

Figura 4.18: Gráfico de t *vs.* $\hat{\text{PCS}}_{G,t}$ para distintos valores de G (Parte Derecha).

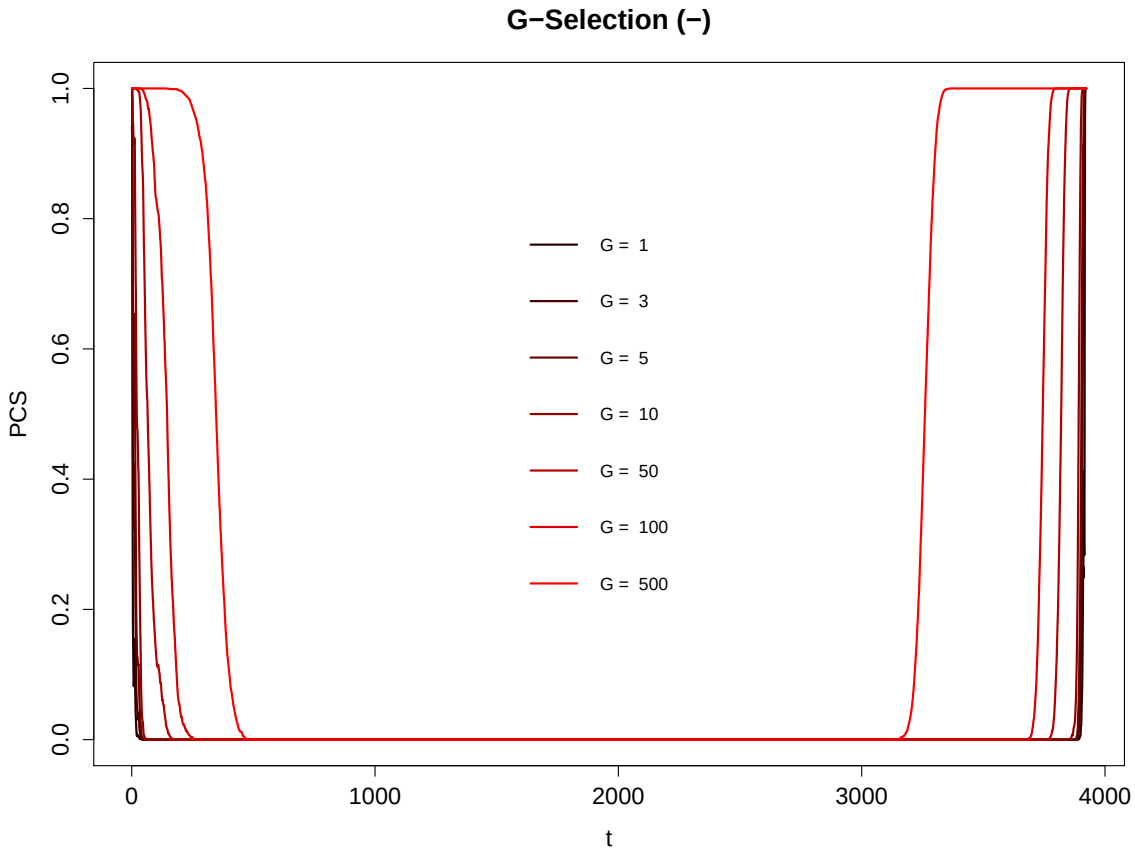

Figura 4.19: Gráfico de t *vs.* $\hat{\text{PCS}}_{G,t}$ para distintos valores de G en el caso inverso.

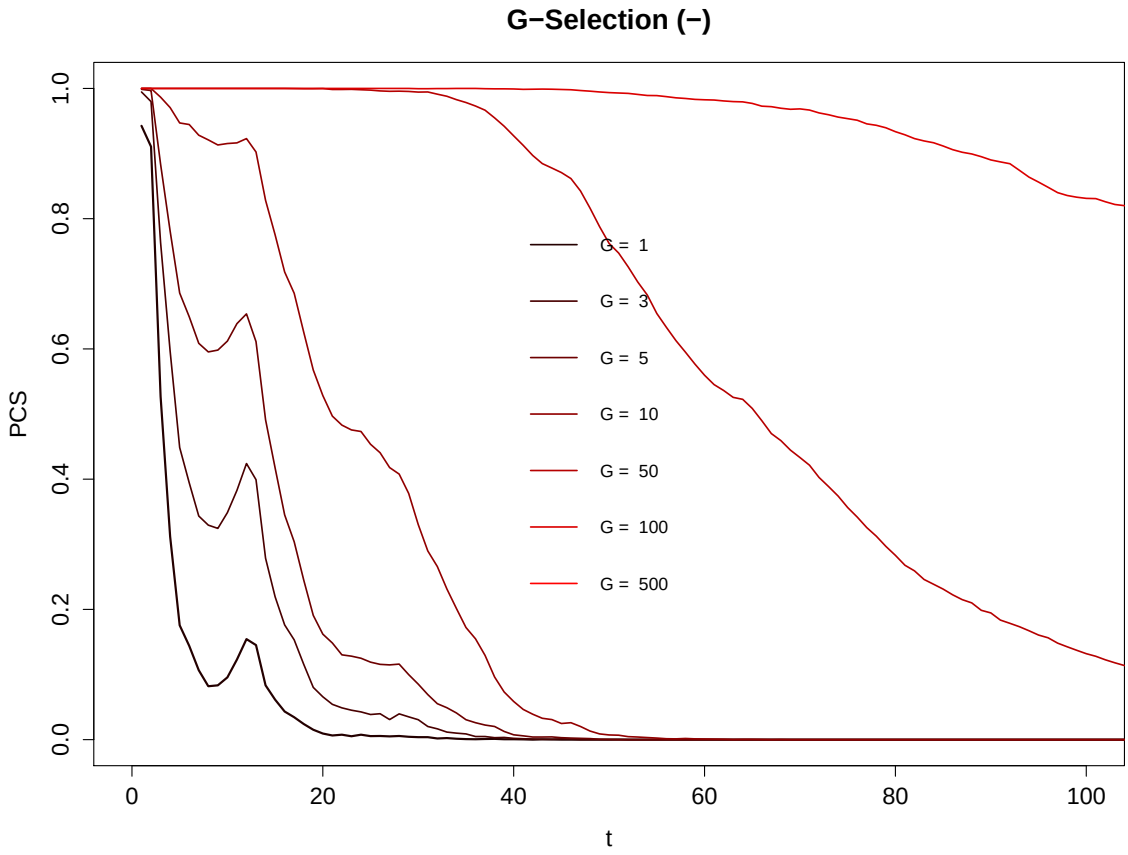

Figura 4.20: Gráfico de t *vs.* $\hat{\text{PCS}}_{G,t}$ para distintos valores de G en el caso inverso (Parte Izquierda).

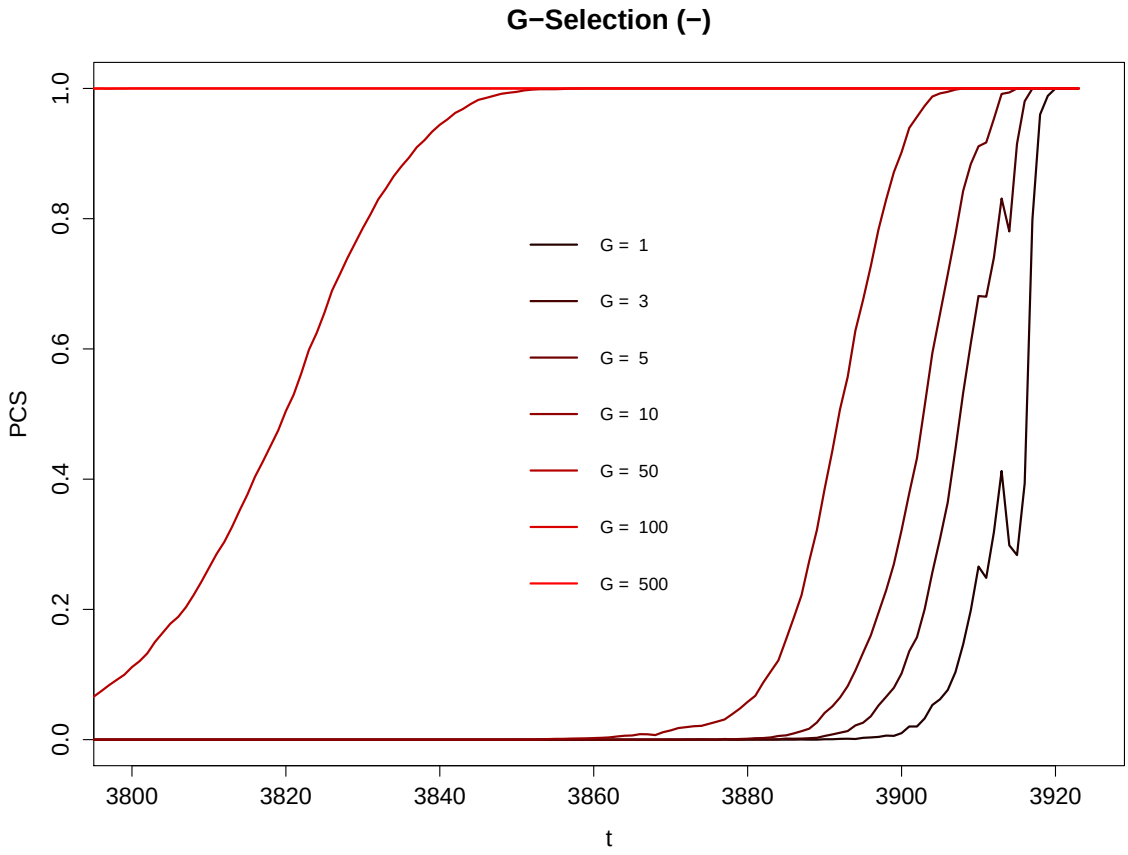

Figura 4.21: Gráfico de t *vs.* $\hat{\text{PCS}}_{G,t}$ para distintos valores de G en el caso inverso (Parte Derecha).

de poblaciones que no son parte de las mejores, pero están lo suficientemente *cerca* de ellas, puedan considerarse parte de una selección correcta. Por tal motivo, la especificación de d, en la escala del espacio parametral, es crucial para realizar un análisis de selección satifactorio. En el caso de estudio, dado que no se tiene la noción de tal valor a partir del cual se pueda considerar una mutante como verdaderamente parte de las *mejores*, la especificación del parámetro d se volvería una cuestión arbitraria y, con altas posibilidades, podría conducir a conclusiones erróneas. Es por ello que se decidió no aplicar $d-$selección para este conjunto de datos.

$c-$selección

La Figura 4.22 muestra el gráfico de $\hat{\text{PCS}}_{\lceil rt \rceil, t}$ como función de r para $r = 0{,}5, 0{,}75, 0{,}85, 0{,}9, 0{,}95, 0{,}99$ (véase el algoritmo de la r^*-selección en la Sección 2.3.3). En este caso r representa la mínima proporción de verdaderos mejores que el experimentador está dispuesto a admitir para calificar una selección determinada como correcta. Un ejemplo de interpretación de estas curvas puede ser para $t = 1000, r = 0{,}85$ (curva violeta), los resultados del análisis indican que si el experimentador busca las $t = 1000$ mutantes, y está dispuesto a calificar como correcta toda selección que contenga al menos $(0{,}85)(1000) = 850$ de ellas, entonces puede alcanzar una PCS de aproximadamente 0,75. Puede apreciarse que la mayoría de las curvas tienen un comportamiento similar a las de $G-$selección (Figura 4.16). Además, puede verse que hay una diferencia muy marcada entre la curva correspondiente a $r = 0{,}75$ y $r = 0{,}85$.

Para estudiar con más detalle estas curvas se puede observar la Figura 4.23 que representa un acercamiento a la región izquierda de la Figura 4.22. En este gráfico puede apreciarse que las curvas para valores $r \geq 0{,}85$ caen rápidamente a partir de valores pequeños de t, $t > 20$ y no vuelven a crecer hasta valores demasiado grandes de t, $t > 2500$ (4.24). Un caso particularmente interesante es la curva correspondiente a $r = 0{,}75$ que se mantiene constantemente alta (PCS $> 0{,}95$) para valores razonablemente grandes de t ($t < 250$). Por ejemplo, si el investigador se propone encontrar las mejores $t = 250$ mutantes a un $r = 0{,}75$ de calidad (permitiendo que toda selección con al menos $(0{,}75)(250) \approx 178$ se califique como correcta) entonces puede alcanzar una PCS superior al 0,95. Debido a la drástica diferencia entre las curvas correspondientes a $r = 0{,}75$ y $r = 0{,}85$ se decidió calcular la curva para $r = 0{,}80$ únicamente restringida al rango de valores de t de la Figura 4.23.

Las Figuras 4.25–4.27 son los análogos de las Figuras 4.22–4.24 para la base de datos transformada con cambio de signo. En este caso, nuevamente las gráficas correspondientes a valores $r \geq 0{,}9$ caen rápidamente para valores pequeños de t ($t < 20$) y no crecen de nuevo hasta valores muy grandes $t > 2700$. La curva de particular interés en este caso es la correspondiente a $r = 0{,}85$. En la Figura 4.23 puede apreciarse que esta curva crece y se mantiene constantemente alta para valores $120 < t < 400$ y después cae relativamente rápido. Esto permite llegar a una conclusión interesante, pues por ejemplo, si se fija $t = 400, r = 0{,}85$ se puede encontrar una selección con al menos $(0{,}85)(400) \approx 340$ de las verdaderas *peores* mutantes a una PCS superior a 0,95.

A diferencia del caso de la $G-$selección, la $c-$selección ha permitido obtener conclusiones más específicas respecto a este caso de estudio. Esto confirma la aseveración hecha en la Sección 2.3.3 acerca de que la $c-$selección no es equivalente a la $G-$selección, sino una generalización menos estricta de su criterio de selección que permite obtener conclusiones interesantes en casos en los que la selección es difícil como el presente caso de estudio. Las recomendaciones principales correspondientes se resumirán en formato tabular en la siguiente sección.

Pruebas de Hipótesis Múltiples

Para propósitos comparativos se decidió analizar el mismo conjunto de datos introducido en la Sección 4.2.2 pero a través de un enfoque de pruebas de hipótesis múltiples. El conjunto de $m = 3924$ hipótesis

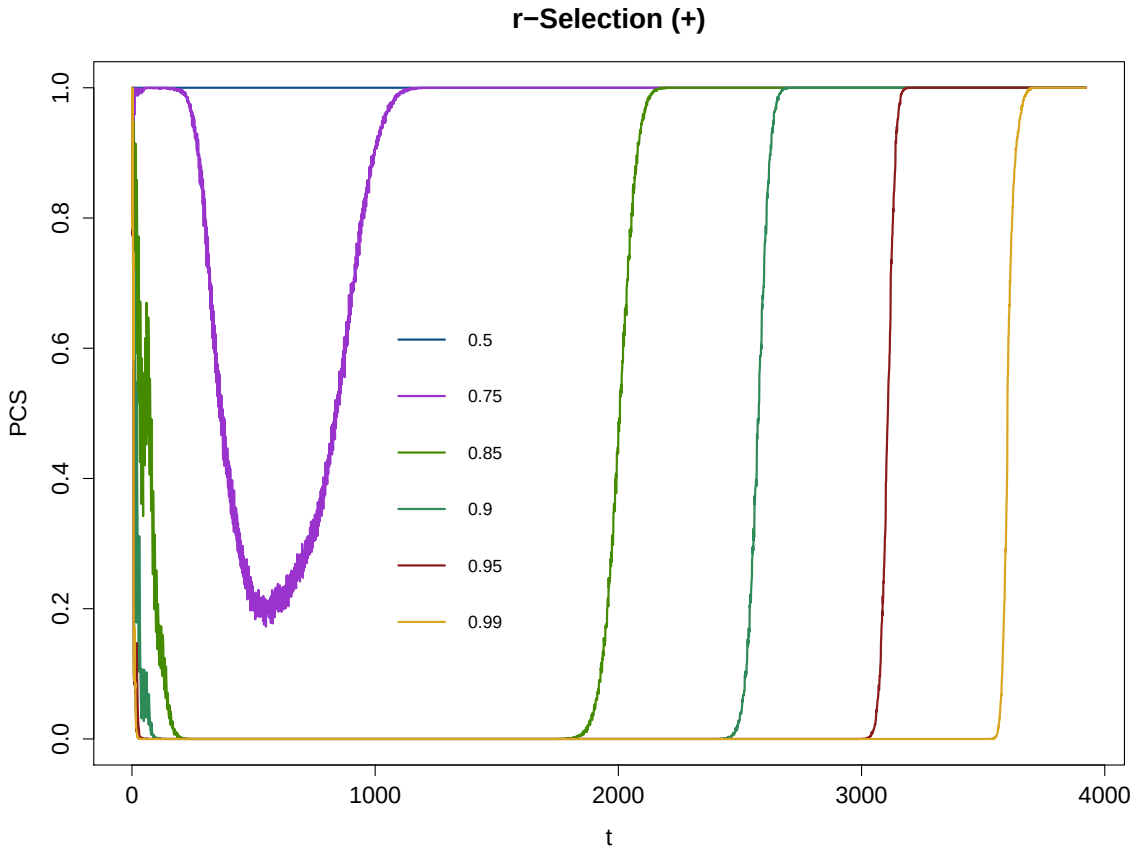

Figura 4.22: Gráfico de t *vs.* $\hat{\mathrm{PCS}}_{\lceil rt \rceil, t}$ para distintos valores de G.

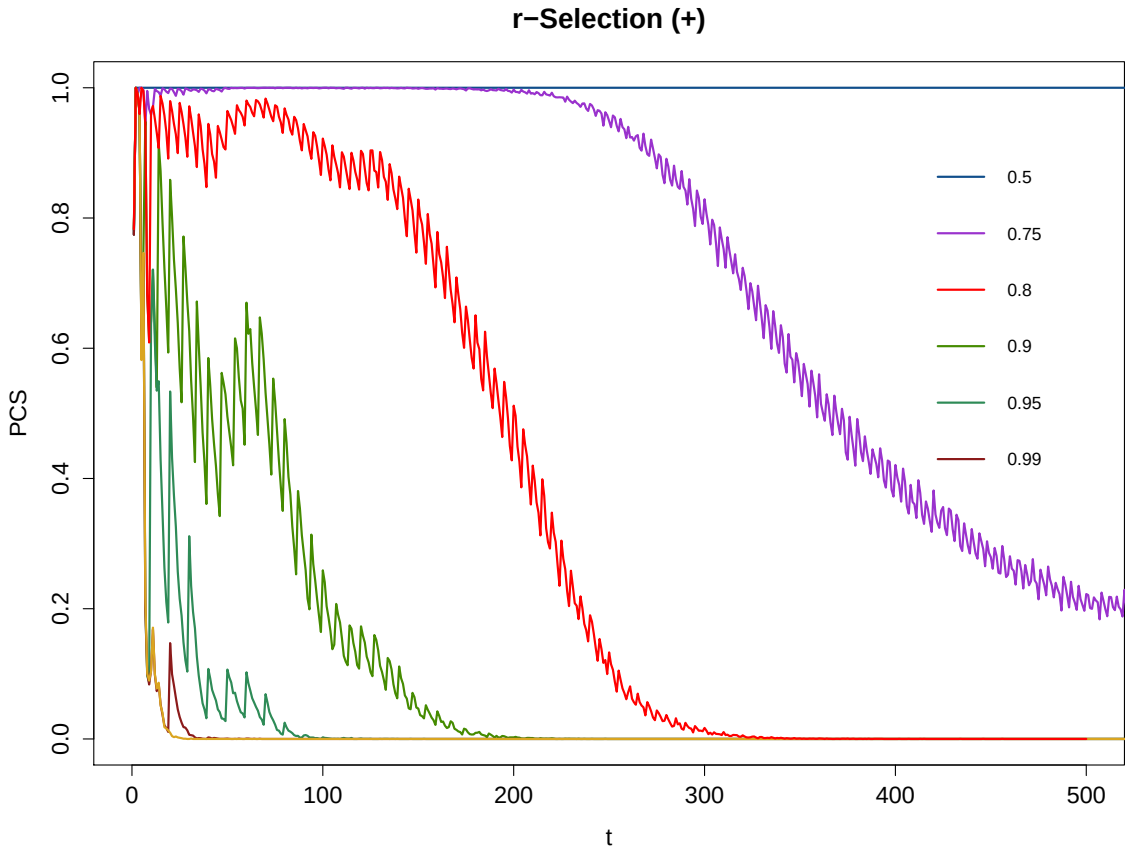

Figura 4.23: Gráfico de t *vs.* $\hat{\text{PCS}}_{\lceil rt \rceil, t}$ para distintos valores de G (Parte Izquierda).

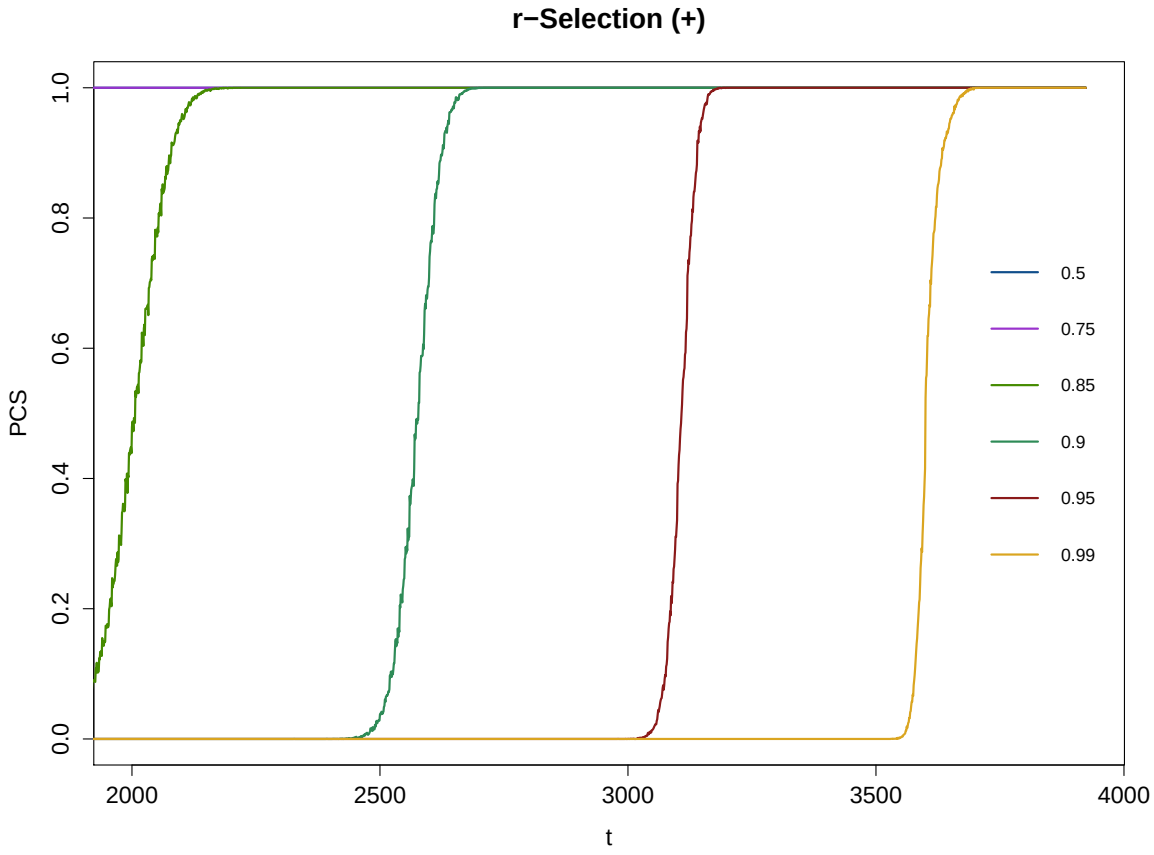

Figura 4.24: Gráfico de t *vs.* $\hat{\text{PCS}}_{\lceil rt \rceil, t}$ para distintos valores de G (Parte Derecha).

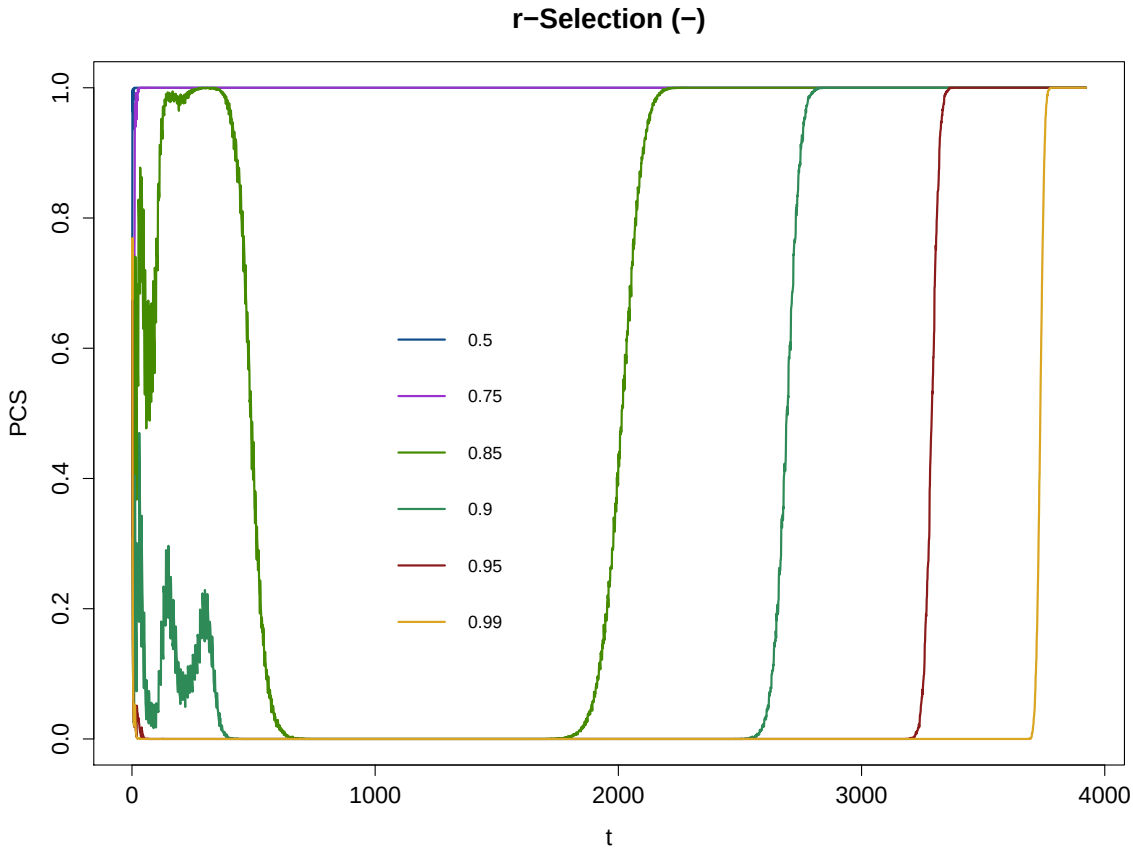

Figura 4.25: Gráfico de t *vs.* $\hat{\text{PCS}}_{\lceil rt \rceil, t}$ para distintos valores de r en el caso inverso.

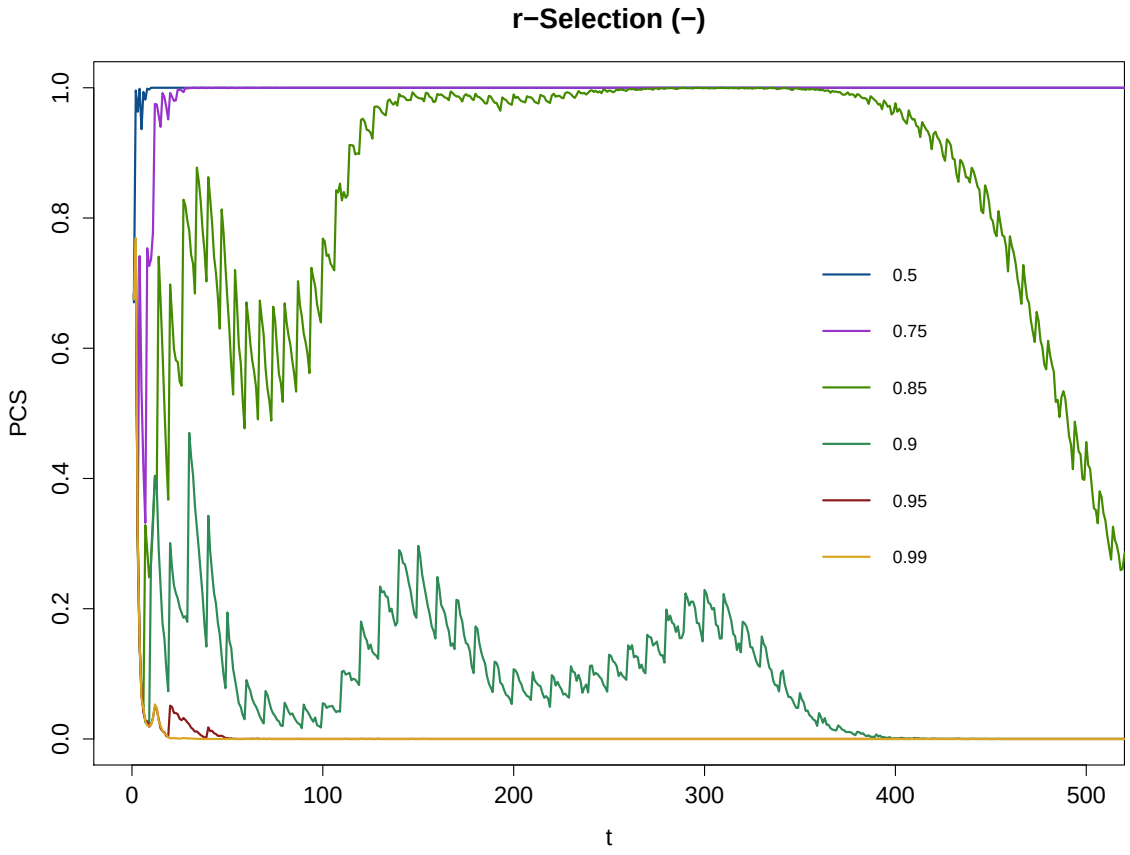

Figura 4.26: Gráfico de t *vs.* $\hat{\text{PCS}}_{\lceil rt \rceil, t}$ para distintos valores de r en el caso inverso (Parte Izquierda).

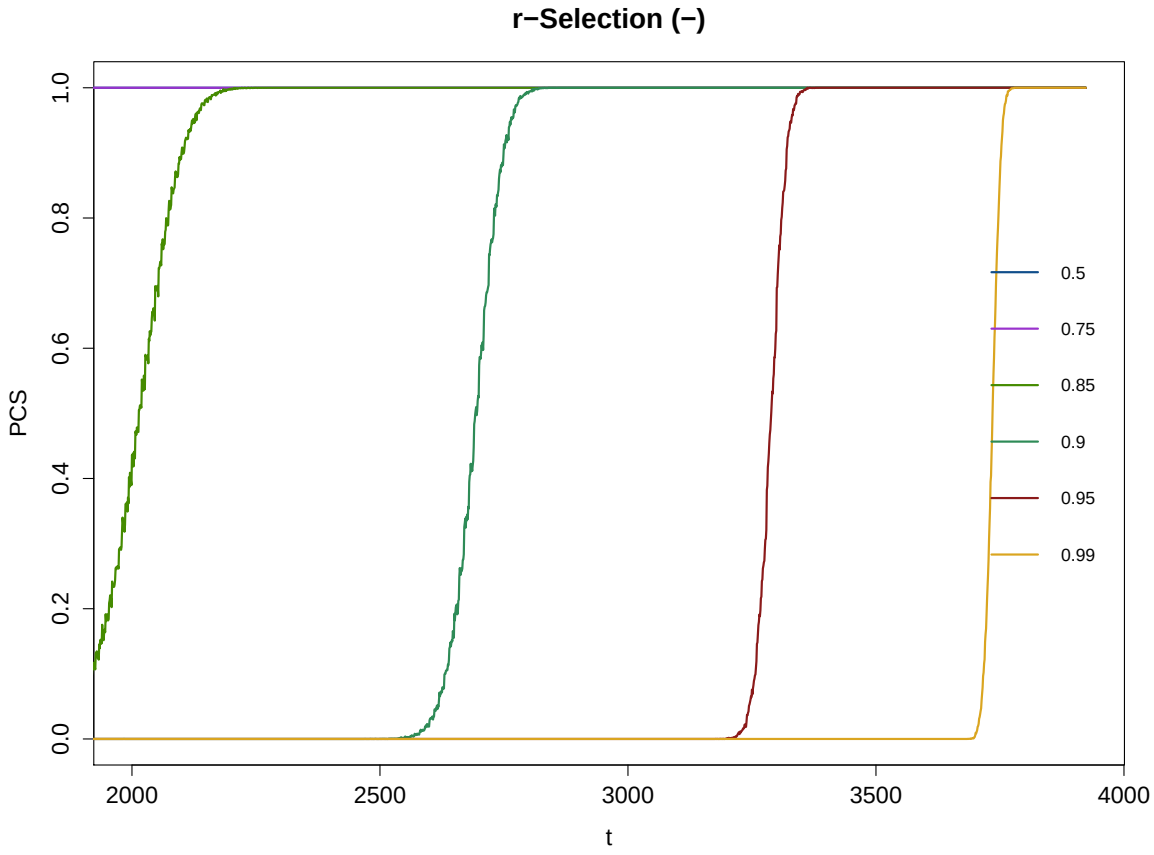

Figura 4.27: Gráfico de t *vs.* $\hat{\mathrm{PCS}}_{\lceil rt \rceil, t}$ para distintos valores de r en el caso inverso (Parte Derecha).

que se plantearon para este problema tienen la forma:

$$H_{0i} : \theta_i = 0 \text{ para } i = 1, 2, ..., 3924, \tag{4.9}$$

donde θ_i, como antes, representa la media del coeficiente de longevidad de la mutante i. La hipótesis (4.9) se planteó de esa manera pues así, todas las mutantes para las cuales H_{0i} sea rechazada tendrán su coeficiente de longevidad significativamente distinto de 0. En otras palabras, habrá evidencia suficiente para determinar que su media es estadísticamente significativa de un WT.

El algoritmo se describe a continuación:

1. Probar la hipótesis (4.9) individualmente para cada una de las m hipótesis y obtener los p-valores no ajustados $p_i, i = 1, 2, ..., m$. Debido a que (4.9) es una prueba de hipótesis sobre la media donde la varianza poblacional es desconocida se utilizó una prueba t.

2. Cada p-valor no ajustado se ajustó para el control de algúna tasa de error particular (FWER, FDR) mediante algún procedimiento particular (Bonferroni, Sidak, Holm, B-H, *etc.*) para obtener un conjunto de p-valores ajustados $\tilde{p}_i, i = 1, 2, ..., 3924$ (véase la Sección 2.1.2).

3. En algunos casos los p-valores ajustados en el paso anterior se re-ajustaron para el control de TPPFP y/o gFWER (van der Laan et al. [2004]).

La Figura 4.28 muestra el gráfico de dispersión del $-\log_{10}(\hat{p}_i)$ contra los p-valores ajustados ordenados $\hat{p}_{(i)}$ para cada uno de los procedimientos de control del FWER descritos en la Sección 2.1.2 (Bonferroni, Sidak y Holm). La línea horizontal representa el umbral $-\log_{10}(\alpha)$ para $\alpha = 0{,}05$ a partir del cual se rechazarían la hipótesis cuyo p-valor ajustado quede por encima. Como puede verse, la conclusión indica que no existe evidencia significativa para rechazar alguna de las hipótesis a un nivel de significancia $\alpha = 0{,}05$. Se decidió experimentar con otros valores de α pero la conclusión fue la misma para todo $\alpha < 0{,}4$ que es un valor relativamente grande.

Para descartar la idea de que no se estén rechazando hipótesis debido a que los métodos de control de FWER son relativamente conservadores (Dudoit et al. [2003]), se decidieron aplicar también el método de control de FDR propuesto en Benjamini and Hochberg [1995] y las técnicas extendidas de control de la TPPFP propuestas en van der Laan et al. [2004]. Los gráficos resultantes se muestran en la Figura 4.29 donde puede apreciarse que, a pesar de que hay una relativa mejoría, no es suficiente para rechazar alguna de las hipótesis a ningún nivel de significancia $\alpha < 0{,}25$.

La conclusión general de la aplicación de las técnicas de PHM al caso de estudio es que no existe evidencia significativa suficiente de los datos para rechazar alguna de las hipótesis $H_{0i}, i = 1, 2, ..., m$ en (4.9). Esto resulta contrastante con las conclusiones obtenidas para RSM expuestas en los Cuadros 4.3 y 4.4 donde sí se lograron identificar las *mejores* y las *peores* mutantes del conjunto.

A pesar de esto, ambas metodologías no se contradicen. Cómo se puntualizó en el Capítulo 3, están respondiendo a preguntas distintas mediante distintas formas de cuantificar incertidumbre. Para el caso de PHM la conclusión dicta que no hay evidencia suficiente para decir que alguna de las mutantes tiene coeficiente de longevidad distinto de 0, o bien, no hay evidencia significativa para afirmar que los coeficientes de longevidad de todas las mutantes no son iguales e iguales al de una célula WT. En contraste, las técnicas de RSM se enfocan en identificar cuáles mutantes son las que tienen mayor (o menor) coeficiente de longevidad.

4.2.5. Presentación de Resultados y Recomendaciones

Tabulación de Resultados

Los Cuadros 4.3 y 4.4 muestran los valores de t que se recomienda al experimentador elegir para distintos valores de los parámetros G y r y algunos valores importantes de PCS. Por ejemplo, si el

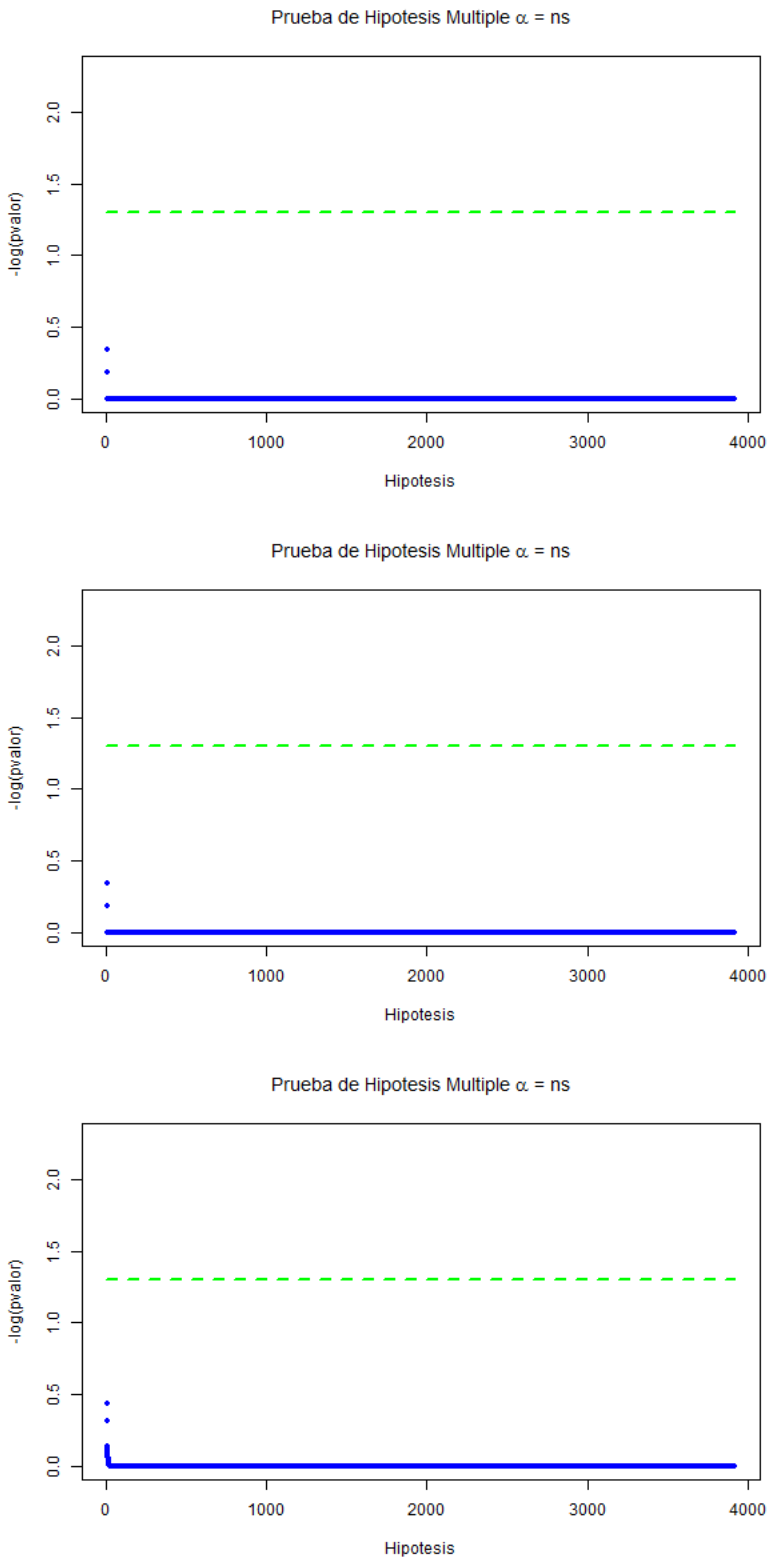

Figura 4.28: Gráfico de dispersión del $-\log_{10}$ de los p-valores ajustados ordenados para los métodos de control de FWER. (De arriba a abajo) Bonferroni, Holm y Sidak

.

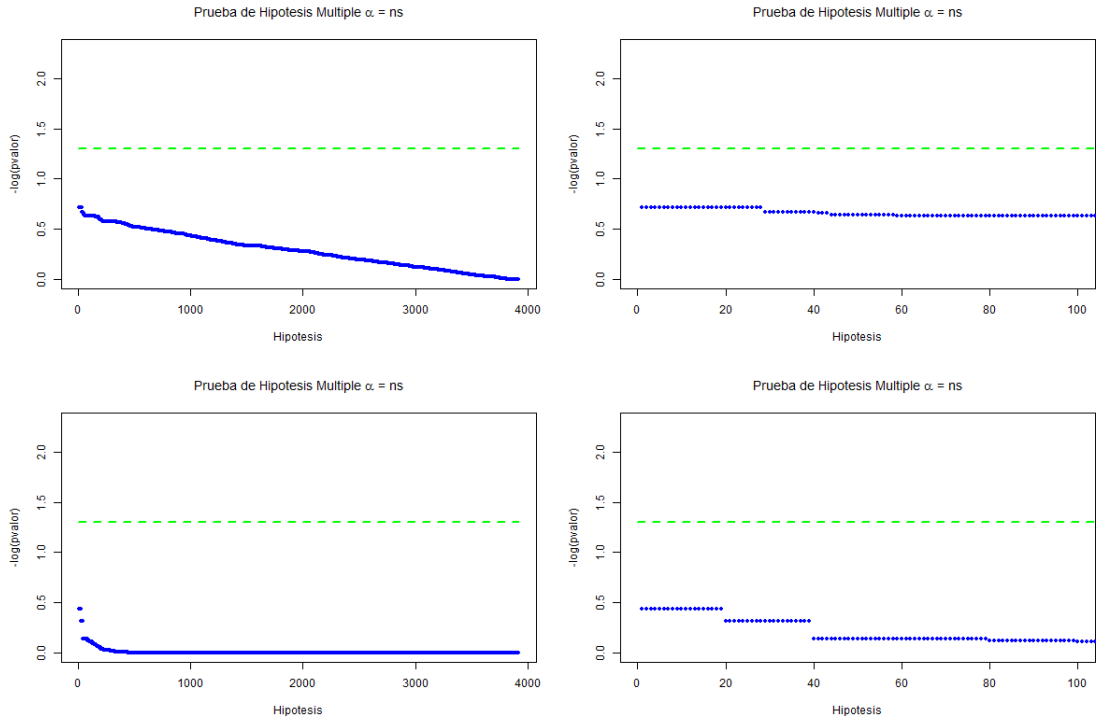

Figura 4.29: (Panel izquierdo) Gráfico de dispersión del $-\log_{10}$ de los p-valores ajustados ordenados para los métodos de control de FDR y TPPFP ($q = 0{,}05$). (Panel derecho) Acercamiento de los gráficos del panel izquierdo.

experimentador desea una PCS de al menos 0,95 y desea que su selección contenga al menos 85 % de las mejores mutantes entonces, el Cuadro 4.4 le indica que debe elegir 7 mutantes para las mejores y cualquier número entero en el intervalo $[126, 406]$ para las peores. En cambio, si deseara una selección mas estricta con 95 % o más de calidad, el Cuadro 4.4 le indica que elija los 4 mejores genes, sin embargo en el caso de los peores el conjunto de datos no permite que haga esa selección bajo esas condiciones. El Cuadro 4.3 se interpreta de manera análoga.

La interpretación de los resultados presentados en el Cuadro 4.4 es intuitiva, en el sentido que, al experimentador le interesará siempre hacer una selección con r y PCS lo más grandes posibles. Tomando en cuenta que al incrementar r PCS decrece, surge un *tradeoff* que sugiere que se debe permitir una proporción (preferentemente pequeña) de falsos positivos $(1-r)100\,\%$ que forme parte de una selección que se considerará como correcta. Por ejemplo, en la tabla 4.4, si el experimentador fija $r > 0{,}9$, tendrá que escoger un número $t \leq 6$ muy pequeño de poblaciones si desea obtener una alta PCS. No es el caso para $r = 0{,}85$ donde se puede ver que se le permite hacer selecciones considerablemente más grandes manteniendo una alta PCS.

La interpretación del Cuadro 4.3 no es tan directa. El parámetro G, que es el número de poblaciones *extra* que el experimentador está dispuesto a tomar para encontrar a las t mejores, no tiene interpretación por si solo. Por ejemplo, $G = 10$ es bastante conveniente si se buscan las $t = 500$ mejores poblaciones pero es sumamente inconveniente si se buscan las $t = 2$ mejores. La elección e interpretación del valor más adecuado para el parámetro G se debe hacer, por tanto, siempre en términos del cociente $r = t/(t+G)$ que representa la proporción de verdaderas mejores poblaciones en la selección realizada con un valor particular del parámetro G. Por ejemplo, $G = 10, t = 500$ produce $r = 500/510 \approx 0{,}98$ mientras que $G = 10, t = 2$ produce $r = 2/12 \approx 0{,}16$.

Para auxiliar en la interpretación del Cuadro 4.3 y la elección del valor de G más conveniente se

$PCS_{G,t}$								
	0.75		**0.90**		**0.95**		**0.99**	
	+	−	+	−	+	−	+	−
$\boldsymbol{G = 1}$	6	2	6	2	4	0	4	0
$\boldsymbol{G = 3}$	6	3	6	2	6	2	5	1
$\boldsymbol{G = 5}$	11	4	6	2	6	2	6	1
$\boldsymbol{G = 10}$	17	15	14	13	12	4	7	2
$\boldsymbol{G = 50}$	44	50	35	41	32	38	26	32
$\boldsymbol{G = 100}$	69	120	60	87	54	76	39	53
$\boldsymbol{G = 500}$	150	320	131	290	118	266	96	221

Cuadro 4.3: Valores de t recomendados para distintos valores de G y PCS. Las columnas con signo + hacen referencia a la elección de los mejores y con el signo − para las peores.

$PCS_{\lceil rt \rceil,t}$								
	0.75		**0.90**		**0.95**		**0.99**	
	+	−	+	−	+	−	+	−
$\mathbf{r = 0.5}$	$\forall t$	$\forall t - \{1\}$	$\forall t$	$\forall t - \{1\}$	$\forall t$	$\forall t - \{1,5\}$	$\forall t$	$\forall t - \{1,3,5,7\}$
$\boldsymbol{r = 0.75}$	313	$t > 10$	273	$t > 12$	256	$t > 15$	216	$t > 24$
$\boldsymbol{r = 0.80}$	165		130		86		5	0
$\boldsymbol{r = 0.85}$	27	[107,461]	14	[120,429]	7	[126,406]	2	[227,372]
$\boldsymbol{r = 0.90}$	6	2	4	0	4	0	2	0
$\boldsymbol{r = 0.95}$	6	2	4	0	4	0	2	0
$\boldsymbol{r = 0.99}$	4	2	4	0	4	0	2	0

Cuadro 4.4: Valores de t recomendados para distintos valores de r y PCS. Las columnas con signo + hacen referencia a la elección de los mejores y con el signo − para las peores. $t = 0$ indica que el conjunto de datos no permite hacer esa selección bajo esas condiciones. Para el caso $r = 0{,}80$ el análisis únicamente se hizo para el caso +, pues, para el caso −, $r = 0{,}85$ proporcionó mejores resultados.

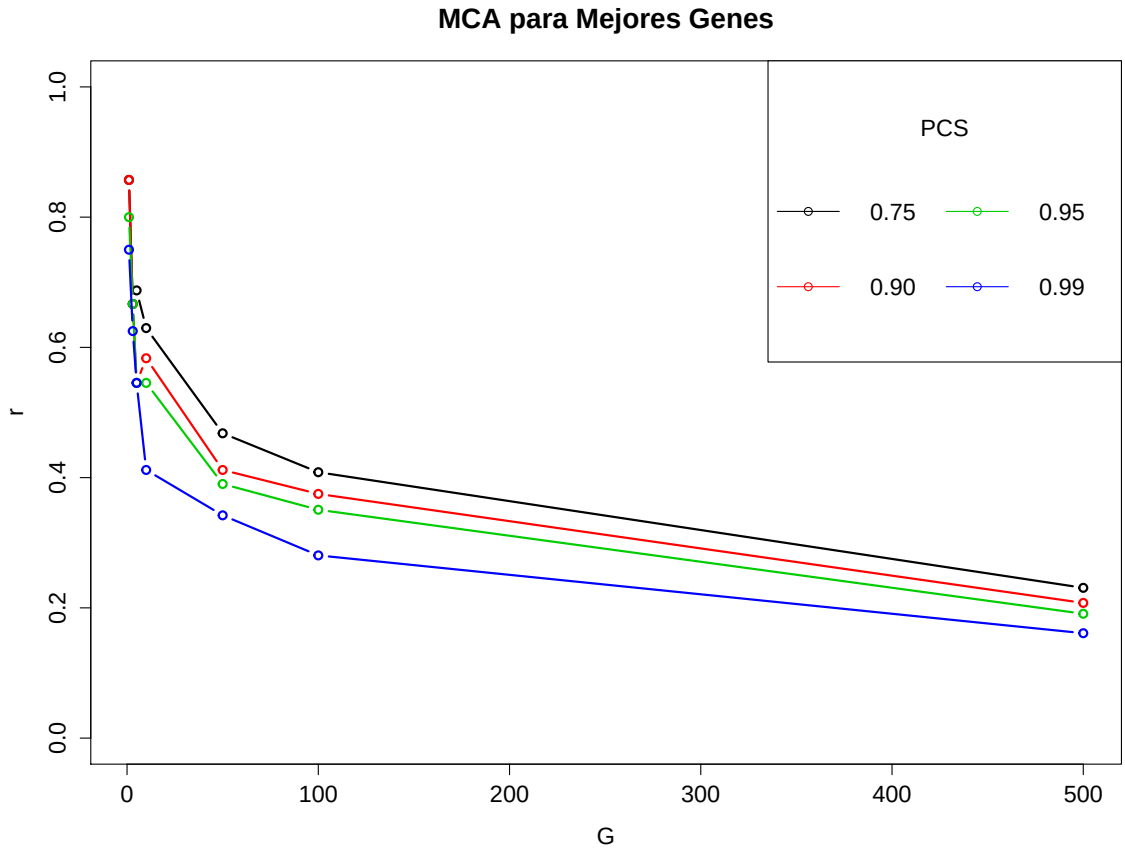

Figura 4.30: Gráfico de G *vs.* $r = t/(t+G)$ para los valores recomendados en el Cuadro 4.3 (+).

propone la utilización de un gráfico de la máxima calidad de selección alcanzable (MCA). Éste grafica los valores de G contra los valores calculados $r = t/(t+G)$ con los valores recomendados de t en el Cuadro 4.3 para cada caso. Las Figuras 4.30 y 4.31 muestran los gráficos de MCA para los valores (+) y (−) del Cuadro 4.3 donde cada línea representa un valor fijo de PCS particular. Es interesante notar que valores grandes de G corresponde a valores pequeños de r sin importar el valor de PCS. Esto es intuitivo dado que mientras más grande es G mayor es el número de *falsos positivos* que se permite entrar en una selección correcta, lo cual provoca que r decrezca. En ambos casos el punto que maximiza r es el que corresponde a $G = 1, \text{PCS} = 0{,}90$ que es el valor que se recomienda tomar para G si se busca minimizar el número de falsos positivos.

Intervalos de Confianza para PCS

Una vez que se ha identificado el tipo de metodología de selección que se desea aplicar ($d-$selección, $G-$selección, $c-$selección), el parámetro de selección $d/G/r$ y el número de mejores (o peores) poblaciones que se desea buscar t, es posible, calcular un estimador puntual $\hat{\text{PCS}}$ para la verdadera probabilidad de selección correcta.

En la Sección 2.3 se expusieron algunas de las principales técnicas de estimación de PCS en la literatura y algunos de los algoritmos principales para su cálculo. Cui and Wilson [2008] desarrollaron la librería PCS en R (www.r-project.org) cuya función principal es calcular el estimador bootstrap para PCS. Sus algoritmos fueron implementados para el desarrollo de una herramienta computacional interactiva (véase la Sección 4.1) que permite al experimentador visualizar el comportamiento de $\hat{\text{PCS}}$ para distintos valores de t de manera simultánea y con base en ello tomar una decisión en relación a la

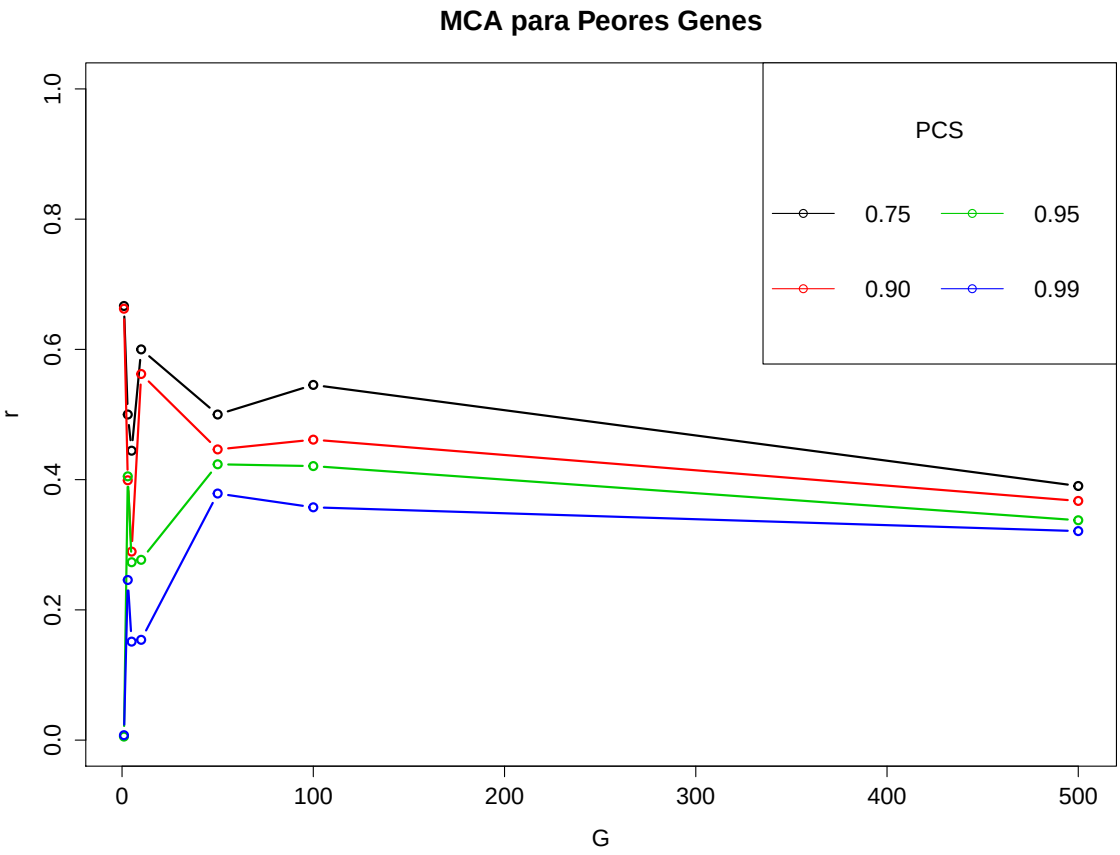

Figura 4.31: Gráfico de G *vs.* $r = t/(t+G)$ para los valores recomendados en el Cuadro 4.3 $(-)$.

selección que hará. Sin embargo, como todo estimador puntual, $\hat{\text{PCS}}$ posee inherentemente la noción de incertidumbre, en este caso acerca de qué tan bien está estimando el verdadero valor de PCS.

Por definición (véase Davison and Hinkley [1997] para los detalles) todo estimador de tipo bootstrap posee dos fuentes principales de incertidumbre:

1. Con relación a qué tan bien aproxima $\hat{\text{PCS}}$ a PCS.

2. Con relación al proceso de remuestreo de un número finito $B > 0$ de muestras que se utilizan como medio para su cálculo.

La noción 2 puede controlarse haciendo el parámetro B tan grande como sea computacionalmente posible. Sin embargo, la noción 1 no es tan fácil de controlar y depende en gran medida de la distribución de $\hat{\text{PCS}}$. Cui and Wilson [2008], Cui et al. [2010] propone la determinación de la distribución de $\hat{\text{PCS}}$ como uno de los problemas abiertos de su línea de investigación, ya que si dicha distribución fuera determinada, sería posible la construcción de intervalos de confianza y con base en ellos hacer inferencia acerca de qué tan bien se está aproximando a PCS mediante $\hat{\text{PCS}}$.

Un primer acercamiento simple a este subproblema planteado durante la tesis, fue fijar $t, d/G/r$ y calcular repetidamente $\hat{\text{PCS}}$, de manera que se construya un conjunto de observaciones $a_i, i = 1, 2, ..., I$ de $\hat{\text{PCS}}$. Luego, mediante las observaciones a_i se construyó un histograma, de tal forma que, al menos al nivel exploratorio, se pueda dar una idea general de cómo es la distribución de $\hat{\text{PCS}}$. Posteriormente, mediante los cuantiles empíricos q_α y $q_{1-\alpha}$ del conjunto de observaciones a_i se construyeron intervalos de confianza aproximados para PCS para distintos casos. La Figura 4.32 muestra ejemplos de histogramas para $I = 1000$ observaciones calculadas para algunos parámetros G/r y algunos valores específicos de t tomados de los Cuadros 4.3 y 4.4, así como barras verticales que indican los límites de un intervalo aproximado de 95 % de confianza para PCS. En todos los casos la distribución aproximada de $\hat{\text{PCS}}$ aparenta ser unimodal y simétrica pero demostrarlo formalmente es un problema abierto que aún sigue vigente.

4.2.6. Discusión y Conclusiones

Las Secciones 4.2.1–4.2.5 presentaron un caso de aplicación de la teoría de RSM moderna reseñada en el Capítulo 2. El problema se planteó y se resolvió con el apoyo de la herramienta computacional descrita en la Sección 4.1 y las conclusiones se reportaron al final de la Sección 4.2.4.

Una de las principales piezas de aprendizaje obtenidas del caso de aplicación anterior es poder reafirmar la sutil pero importante diferencia conceptual entre las pruebas de hipótesis múltiples y la teoría de RSM. Para este caso las técnicas de RSM no sólo proveen una metodología, sustentada estadísticamente, para la toma de decisión que involucren selección bajo incertidumbre (en este caso de cuáles son las mutantes con mayor y menor coeficiente de supervivencia s), sino también brindan un panorama general de la estructura de la base de datos en cuestión.

La aplicación de la $G-$selección y la $r-$selección permitieron visualizar la base de datos desde ángulos que no pueden detectarse mediante herramientas exploratorias simples como gráficos de dispersión e histogramas. Por ejemplo, para el caso de la $r-$selección, el Cuadro 4.4 sugiere que si se fija $r = 0{,}80$ es posible elegir $t = 86$ mutantes entre las cuales se tendrá una certeza de más de 95 % de que estarán al menos $(0{,}80)(86) \approx 69$ de las mejores mutantes. Esto le da al investigador la idea de tomar las 86 mutantes con mayor media muestral Y_i y conducirlas a un segundo estudio, donde posiblemente sea posible un mayor número de observaciones dado que el conjunto de genes se redujo de 3924 a sólamente 86. Posteriormente, si se desea, es posible aplicar nuevamente un análisis de selección para encontrar las verdaderas 69 mejores mutantes de dicho conjunto, entre otros aspectos que puedan ser de interés. Cabe aclarar que, si se hubiera aplicado una PHM, dado que no existe la noción de orden ninguno de

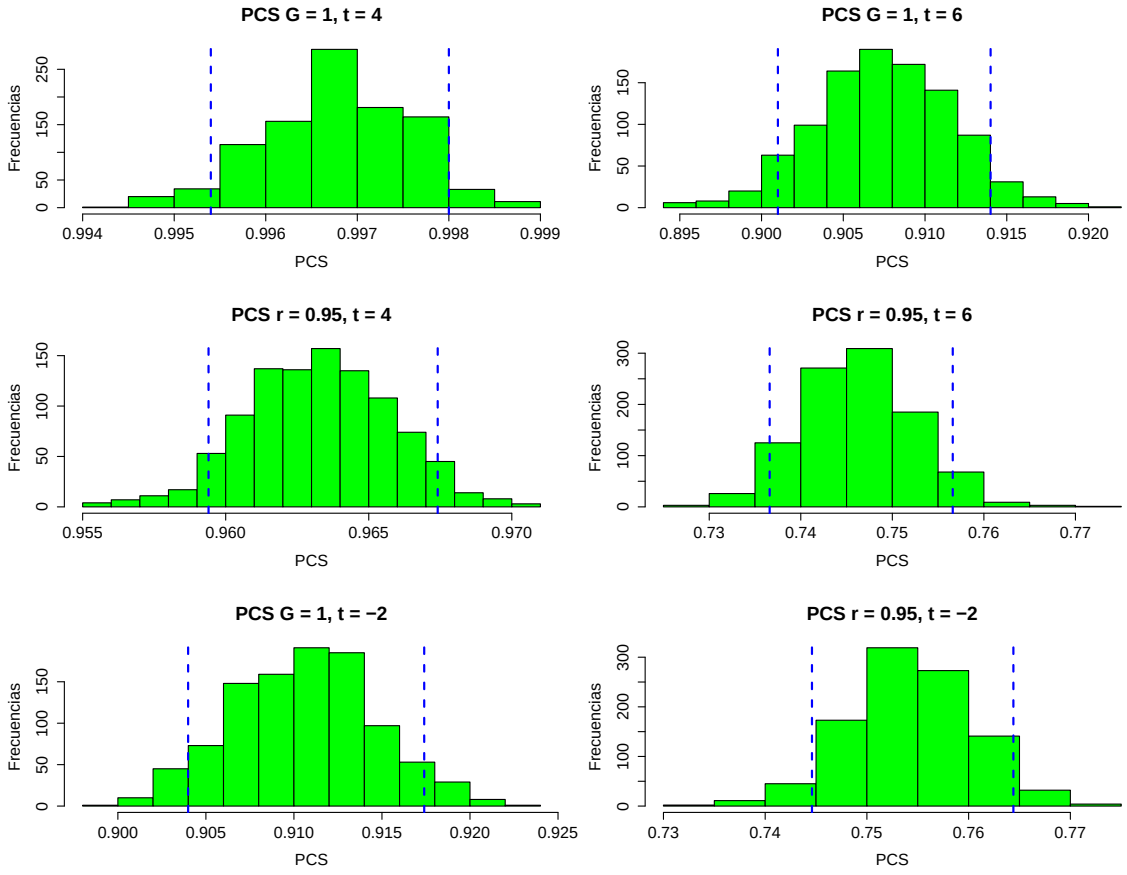

Figura 4.32: Ejemplos de histogramas de las observaciones de $\hat{\text{PCS}}$ para distintos valores de G/r y t. Las barras verticales indican un intervalo de confianza aproximado de 95 % para PCS.

los razonamientos expuestos anteriores hubiera podido ser sostenido ni habría manera de cuantificar la incertidumbre respecto a su planteamiento.

Bibliografía

W.A. Becker. Comparing entries in random sample tests. *Poultry Science*, 40:1507–1514, 1961.

J.D. Gibbons, I. Olkin, and M. Sobel. *Selecting and Ordering Populations*. John Wiley & Sons, 1977.

R.E. Bechhofer. A single-sample multiple decision procedure for ranking means or normal populations with known variances. *The Annals of Mathematical Statistics*, 25:16–39, 1954.

CIMMYT. 1993/94 world maize facts and trends. *CIMMYT*, 1994.

R.L. Bechhofer, T.J. Santner, and D.M. Goldsman. *Design and Analysis of Experiments for Statistical Screening and Multiple Comparisons*. John Wiley & Sons, 1995.

X. Cui and J. Wilson. On the probability of correct selection for large k populations, with applications to microarray data. *Biometrical Journal*, 50(5):870–883, 2008.

T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander. Molecular classification of cancer class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

A. De Luna, E. Garay, S.E. Campos, J. Gonzalez de la Cruz, A.P. Gaspar, and A. Jinich. High-resolution profiling of stationary-phase survival reveals yeast longevity factors and their genetic interactions. *PLoS Genet*, 10(2), 2014.

S. Dudoit, J.P. Shaffer, and J.C. Bodrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103, 2003.

G. Casella and R.L. Berger. *Statistical Inference*. Brooks/Cole, 2008.

J.W. Tukey. The philosophy of multiple comparisons. *Statistical Science*, 6:100–116, 1991.

A. Farcomeni. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, 17:347–388, 2008.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1):289–300, 1995.

Y. Hochberg and A.C. Tamhane. *Multiple Comparison Procedures*. New York: Wiley, 1987.

P.H. Westfall and S.S. Young. *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment*. John Wiley & Sons, 1993.

C.E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 1936.

Z. K. Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

S. Holm. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, 6:65–70, 1979.

B. Efron. *Large-Scale Inference*. Cambridge University Press, 2013.

Y. Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800–802, 1988.

Y. Benjamini and D. Yekutieli. The control of the false-discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29:1165–1188, 2001.

Y. Benjamini and Y. Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Statist.*, 25:60–83, 2000.

C. Genovese and L. Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society B*, 64:499–517, 2002.

J.F. Koen, K.L. Simonsen, and L.M. McIntyre. Implementing false discovery rate control: Increasing your power. *Oikos*, 108:643–647, 2005.

M.J. van der Laan, S. Dudoit, and M.D. Birkner. Multiple testing procedures for controlling tail probability error rates. *U.C. Berkeley Division of Biostatistics Working Paper Series*, 166, 2004.

I. Olkin, M. Sobel, and Y.L. Tong. Bounds for a $k-$fold integral for location and scale parameter models with applications to statistical ranking and selection problems. *Statistical Decision Theory and Related Topics III*, 2:193–212, 1982.

J. Wilson. On the probability of correct selection for large k populations. *PhD Thesis. University of California Riverside*, 2008.

S.S. Gupta and T.C. Liang. Simultaneous lower confidence bounds for probabilities of correct selections. *Journal of Statistical Planning and Inference*, 72(1-2):279–290, 1998.

I. Olkin, M. Sobel, and Y.L. Tong. On finding the probability of correct selection for location and scale families. *Stanford Statistics Department Technical Report 110*, 1976.

F.E. Faltin and C.E. McCulloch. On the small sample properties of the olkin, sobel and tong estimator of the probability of correct selection. *Journal of the American Statistical Asociation*, 78(382):464–467, 1983.

E. Bofinger. On the non-existence of consistent estimators for p(cs). *American Journal Math. Management Sc.*, 5:63–76, 1985.

C.E. McCulloch and A. Dechter. An empirical bayes approach to estimating the probability of correct selection. *Communications in Statistics - Simulation and Computation*, 14:173–186, 1985.

E. Bofinger. Posterior probability of correct selection. *Communications in Statistics: Theory and Methods*, 19(2):599–616, 1990.

H.P. Edwards. Empirical bayes estimators of the probability of correct selection. *The Frontiers of Modern Statistical Inference - American Science Press*, 1992.

J.K. Sohn and S.G. Kahn. On the estimation of the true probability of correct selection. *Communications in Statistics: Simulation and Computation*, 21(2):445–462, 1992.

S.S. Gupta. On a decision rule for a problem in ranking means. *Ph.D. thesis, University of North Carolina*, 150, 1956.

R.E. Bechhofer, C.W. Dunnett, and M. Sobel. A two-sample multiple decision procedure for ranking means of normal populations with a common unknown variance. *Biometrika*, 41:170–176, 1954.

C.W. Dunnett and M. Sobel. A bivariate generalization of student's t distribution, with tables for certain special cases. *Biometrika*, 41:153–169, 1954.

M. Sobel and M. Huyett. Selecting the best one of several binomial populations. *Bell System Technical Journal*, 36:537–576, 1957.

R.E. Bechhofer and M. Sobel. A single-sample multiple decision procedure for ranking variances of normal populations. *The Annals of Mathematical Statistics*, 25:273–289, 1954.

R.E. Bechhofer, S.A. Elmaghraby, and N. Morse. A single-sample multiple-decision procedure for selecting the multinomial event which has the largest probability. *The Annals of Mathematical Statistics*, 30:102–119, 1959.

K. Alam and J.R. Thompson. On selecting the least probable multinomial event. *The Annals of Mathematical Statistics*, 43:1981–1990, 1959.

S.S. Gupta and D.Y. Huang. Subset selection procedures for the means and variances of normal populations: Unequal sample sizes case. *Sankhya*, B38:112–128, 1976.

X. Cui, H. Zhao, and J. Wilson. Optimized ranking and selection methods for feature selection with application in microarray experiments. *Journal of Biopharmaceutical Statistics*, 20(2):223–239, 2010.

X. Cui and J. Wilson. A simulation study on the probability of correct selection for large k populations. *Communications In Statistics - Simulation and Computation*, 38:1244–1255, 2009.

M.S. Bartlett. The use of transformations. *Biometrics*, 3:39–52, 1947.

G. Box and D.R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society*, 26(2): 211–252, 1964.

A.C. Davison and D.V. Hinkley. *Bootstrap Methods and their Application*. Cambridge University Press, 1997.

V.D. Longo, G.S. Shadel, M. Kaeberlein, and B. Kennedy. Replicative and chronological aging in *Saccharomyces cerevisiae*. *Cell Metabolism*, 16:518–31, 2010.