DETERMINACIÓN DEL TAMAÑO MUESTRAL PARA LA SELECCIÓN DE POBLACIONES CON DISTRIBUCIÓN WEIBULL

Alejandro Quiroz Zárate & Enrique Villa Diharce

Comunicación Técnica No I-06-13/28-08-2006
(CC/CIMAT)

Determinación del tamaño muestral para la selección de poblaciones con distribución Weibull.

Alejandro Quiroz Zárate[1]
Enrique Villa Diharce[2]

aquiroz@cimat.mx[1], villadi@cimat.mx[2]

Resumen

La necesidad de reducir costos y elevar la calidad de los productos en la industria, ha llevado a las empresas a la dinámica de ahorrar y mejorar la calidad de sus productos continuamente. Un factor muy importante en esta tarea son los proveedores de los componentes que se utilizan como materia prima. Por esto es que con gran frecuencia se tiene el problema de seleccionar el mejor proveedor de entre varios que ofrecen un componente determinado. De acuerdo a las nuevas iniciativas de calidad, el costo ha dejado de ser el criterio único para esta selección, agregándose ahora la confiabilidad como un factor muy importante. En este artículo se presenta un procedimiento de selección para la distribución Weibull. La elección de esta distribución fue de acuerdo a un criterio de confiabilidad, por ser esta distribución una de las mas comunes en la modelación de tiempos a la falla.

1 Introducción

El problema de selección de poblaciones se encuentra en diferentes contextos de aplicación, en donde contamos con varias poblaciones y debemos elegir una de ellas, que sea la mejor de acuerdo a algún criterio de comparación. En la industria es común disponer de varios proveedores de un determinado componente y tener que elegir al proveedor cuyo producto sea el mejor, de acuerdo a algún criterio de comparación. Cuando la confiabilidad del producto es la característica determinante, entonces la comparación de los proveedores debe llevarse a cabo considerando métricas de confiabilidad, como por ejemplo, la vida media o la confiabilidad a un cierto nivel determinado. Los datos de los componentes que nos dan información de su nivel de confiabilidad, son los tiempos de vida, razón por la requerimos conocer procedimientos de comparación de poblaciones, para poblaciones cuyas distribuciones correspondan a distribuciones de tiempos de vida. Tomando en cuenta que una distribución de tiempos de vida muy frecuente es la distribución Weibull, en este trabajo desarrollamos un procedimiento para determinar el tamaño muestral necesario de las diferentes poblaciones que comparamos, cuando la distribución es Weibull.

En la sección 2, se expone el planteamiento general del problema de selección de poblaciones. En esta sección se dan los elementos necesarios para poder obtener el tamaño muestral necesario para lograr una determinada probabilidad de selección correcta.

En la sección 3, se desarrolla el procedimiento que nos permite determinar el tamaño muestral requerido para seleccionar las mejores s poblaciones de entre un grupo de k ($k > s$) poblaciones con distribución Weibull. La comparación se hace en términos del parámetro de escala de las distribuciones Weibull, considerando que estas tienen el mismo parámetro de forma. También se muestra el funcionamiento general del programa elaborado ($Tamaño\ Muestral$) para este trabajo para el cálculo de los tamaños muestrales.

En la sección 4, se muestra un ejemplo de selección de proveedores de bolsas de aire para automóviles, aquí se comparan seis proveedores y se elijen a los dos mejores.

En la sección 5, se exponen las conclusiones del trabajo y se describe el trabajo a desarrollar posteriormente como continuación del presente estudio.

2 Planteamiento general

Supongamos que tenemos k poblaciones, y que dentro de estas existe una que es la "mejor" población, donde "mejor" es definido por el investigador de acuerdo al contexto del problema. Nuestro objetivo es identificar esa población, a la mejor población. Como cualquier procedimiento estadístico, las técnicas de selección están basadas en los resultados de muestras aleatorias tomadas de las k poblaciones. Para poder identificar a la mejor población necesitamos establecer un procedimiento o regla. Este procedimiento estará sujeto a la distribución de las k poblaciones y al parámetro de interés que tomamos como referencia para definir y seleccionar a la mejor población. Para esto, establecemos una medida de discrepancia entre los parámetros de referencia de las poblaciones bajo comparación.

Debido a que las discrepancias entre las poblaciones se estiman por las discrepancias muestrales correspondientes y estas tienen una variabilidad aleatoria natural, requerimos que las discrepancias muestrales entre dos poblaciones sea mayor que un valor umbral para identificar una diferencia entre estas poblaciones. Este valor umbral lo definirá el investigador de acuerdo al contexto del problema o los instrumentos. Así, si la mejor población difiere de las demás cuando menos en un valor mínimo preestablecido podremos identificarla. Existe la posibilidad de que la población de la cual se obtuvo el mejor parámetro estimado (de acuerdo al procedimiento establecido) no tenga en realidad el mejor parámetro. Por este motivo es que se necesita cuantificar la posibilidad de cometer errores. Primero analizaremos la naturaleza del error que se puede cometer en un procedimiento de selección estadístico, después estableceremos una medida de discrepancia y en base a esta y a la naturaleza del error se basará nuestra cuantificación de los errores.

2.1 La filosofía de la selección de poblaciones

El establecimiento de un procedimiento de selección induce la definición de una variable aleatoria X. Esta variable aleatoria dependerá de la muestra que obtengamos de cada población y del parámetro de interés con el cual efectuaremos la selección. Vamos a suponer que esta variable aleatoria, para simplificar el problema del procedimiento estadístico de selección, tiene alguna distribución $F(x;\theta)$. Las diferencias que se registren en la obtención de las k mediciones en X, reflejan las diferencias entre los k parámetros respectivos, $\theta_1, \theta_2, \ldots, \theta_k$ Las observaciones son tomadas de manera independiente y están distribuidas de la siguiente manera:

Población	1	2	...	k
Distribución	$F(x;\theta_1)$	$F(x;\theta_2)$	...	$F(x;\theta_k)$

Hay que tener en consideración que se está construyendo toda una metodología de selección alrededor de un parámetro totalmente desconocido en cada una de las poblaciones, los θ_i, para poder hacer inferencia sobre su comportamiento.

Para poder establecer el error que se puede cometer en un proceso de selección, supongamos que la mejor población tiene el parámetro $\theta_{(k)}$. Esta notación esta sujeta a la definición de mejor población y al procedimiento de selección. Ahora, intuitivamente es razonable obtener el estimador $\widetilde{\theta}_i$ para cada θ_i de las observaciones correspondientes y establecer que la mejor población es la que tenga $\widetilde{\theta}_{(k)}$. Este procedimiento parece razonable pero la posibilidad de cometer un error siempre existe, puesto que podemos saber de que población proviene $\widetilde{\theta}_{(k)}$, más no sabemos si esa población es la "mejor".

Con el fin de establecer el tipo de error que se puede cometer en un procedimiento estadístico de selección, analicemos los errores en el caso del paradigma de Neyman-Pearson en la teoría de pruebas de hipótesis. No se puede hacer una comparación directa por la diferencia filosófica entre estos procedimientos. En teoría de prueba de hipótesis, bajo el paradigma de Neyman-Pearson, la idea filosófica radica en que el investigador tiene una hipótesis acerca del fenómeno en cuestión y utilizará la maquinaria de este paradigma para corroborar o no la hipótesis. En un procedimiento estadístico de selección, la filosofía es muy distinta, el investigador utiliza lo que ya sabe de las poblaciones para inferir cual es la "mejor" población, más el nunca estableció hipótesis alguna. Es por esto, que lo más que se puede hacer es una analogía, y eso es lo que haremos. Los errores que se pueden cometer en una prueba de hipótesis aplicando el paradigma de Neyman-Pearson son: el de rechazar H_0 (la hipótesis nula) cuando es cierta, llamado *error Tipo I* o el de no rechazar H_0 cuando esta es falsa, llamado *error Tipo II.*

		Situación	Real
		H_0 verdadera	H_0 falsa
Decisión del	No rechazar H_0	No hay error	Error $(TipoII)$
Estadístico	Rechazar H_0	Error $(TipoI)$	No hay error

En una prueba clásica con la hipótesis nula de homogeneidad entre los parámetros, el error de Tipo I solamente se puede cometer en el subconjunto del espacio parametral en donde $\theta_1 = \theta_2 = \ldots = \theta_k$ y el error de Tipo II solamente se puede cometer en el subconjunto del espacio parametral en donde la igualdad no se mantiene. En un procedimiento de selección, la selección es correcta si el valor de θ de la población seleccionada es el "mejor", esto es $\theta_{(k)}$. Hay que tener en cuenta que nuestro objetivo no es estimar $\theta_{(k)}$, ni tomar decisión alguna sobre el valor de $\theta_{(k)}$ sino solamente seleccionar a la población que tiene el valor de θ igual a $\theta_{(k)}$. Por lo que solamente se comete error si la selección es incorrecta, esto es, afirmar que el valor de θ de la población seleccionada es el "mejor" cuando no lo es. Es por esta razón que en los procedimientos de selección existe el análogo del error de Tipo II pero no existe el análogo del error de Tipo I. Supongamos que la población de la cual proviene $\theta_{(k)}$ es la j *ésima*.

		Situación	Real
Decisión del		$\theta_{(k)} = \theta_j$	$\theta_{(k)} \neq \theta_j$
Estadístico	Afirmar que $\widetilde{\theta}_{(k)}$ proviene de la j *ésima* población	✓	×

2.2 Aspectos analíticos del problema de selección de la mejor población

2.2.1 Zona de Indiferencia y Zona de Preferencia

Recordemos el planteamiento general. Supongamos que tenemos k poblaciones y que nuestro objetivo es seleccionar a la "mejor" dentro de éstas. (Estamos suponiendo que $k \geq 2$). Dadas estas suposiciones, el espacio parametral es $\Theta \subset \mathbb{R}^{k \dim(\theta)}$, pues el parámetro θ no necesariamente es unidimensional. Podemos tener varios procedimientos de selección, $h = 1, 2, ..., m$. Estos procedimientos de selección estarán en función del o de los parámetros sobre los que se hará la selección. Para elegir el procedimiento que nos de el menor error de selección, cuantificaremos los procedimientos mediante una probabilidad de selección correcta. Definamos a P_h como la probabilidad de una selección correcta utilizando el procedimiento de selección h. El procedimiento g será el mejor procedimiento de selección si satisface que:

$$P_g \geq P_h, \forall h \in \{1, 2, ..., m\},$$

Sin embargo, como el espacio parametral consta de una infinidad de elementos, habrá zonas de éste en donde g sea el mejor procedimiento y otras zonas en las que k sea el mejor

procedimiento, donde $k \in \{1, 2, ..., m\} - \{g\}$. Esta manera de proceder no es del todo eficiente, pero nos proporciona una manera de proceder. Para poder establecer el mejor procedimiento analizaremos el espacio parametral, dividiéndolo en zonas de valores de $\theta \in \Theta$ en donde podamos establecer la preferencia de hacer una selección correcta y en otras en donde no podemos establecer preferencia entre distintas selecciones. Estas zonas estarán en función de la limitación y la existencia de errores en la medición o en el umbral fijado por el investigador. Las regiones son:

1) una en la cual las distancias entre los parámetros de las poblaciones sean grandes o distinguibles del umbral, llamada *zona de preferencia (ZP)* y

2) otra en donde la distancias entre los parámetros de las poblaciones sean pequeñas o indistinguibles, llamada *zona de indiferencia (ZI).*

Para ejemplificar como se particiona el espacio parametral veamos lo siguiente:

Supongamos que tenemos k poblaciones y que nuestro objetivo es seleccionar a la "mejor" dentro de éstas, donde la mejor será la población que tenga el parámetro θ más grande (Estamos suponiendo que $k \geq 2$). Es decir, en notación de estadísticas de orden: $\theta_{[1]}, \theta_{[2]}, ..., \theta_{[k]}$ tal que $\theta_{[i]} \leq \theta_{[i+1]}$ y queremos seleccionar $\theta_{[k]}$. Bajo esta notación estamos considerando a todos los parámetros $\theta_{[1]}, \theta_{[2]}, ..., \theta_{[k]}$. Ahora, como queremos seleccionar a la población que tenga $\theta_{[k]}$, es natural considerar nada más la diferencia entre $\theta_{[k]}$ y $\theta_{[k-1]}$, pues si la diferencia entre estas nos permite diferenciar a la mejor, también lo hará la diferencia entre las demás. Así, por las propiedades de las estadísticas de orden, hemos reducido la dimensión del espacio parametral. El espacio que ha resultado es

$$\Xi = \{(\theta_{[k-1]}, \theta_{[k]}) | \theta_{[i]} \in \mathbb{R}\} \subset \mathbb{R}^2.$$

Pero, ¿cómo podremos dividir al espacio parametral en las zonas ZP y ZI? Definamos como

$$\delta(\theta_{[k-1]}, \theta_{[k]}) = \theta_{[k]} - \theta_{[k-1]},$$

la medida de discrepancia que utilizaremos. Debido a nuestras limitaciones en los procesos de medición, errores en los experimentos o las condiciones impuestas por el contexto del problema del investigador, se fija un umbral. A partir de este umbral se definen las zonas de preferencia (ZP) y de indiferencia (ZI), de la siguiente manera:

ZP	$\theta_{[k]} - \theta_{[k-1]} \geq \delta^*$
ZI	$\theta_{[k]} - \theta_{[k-1]} < \delta^*$

en donde δ^* es el umbral.

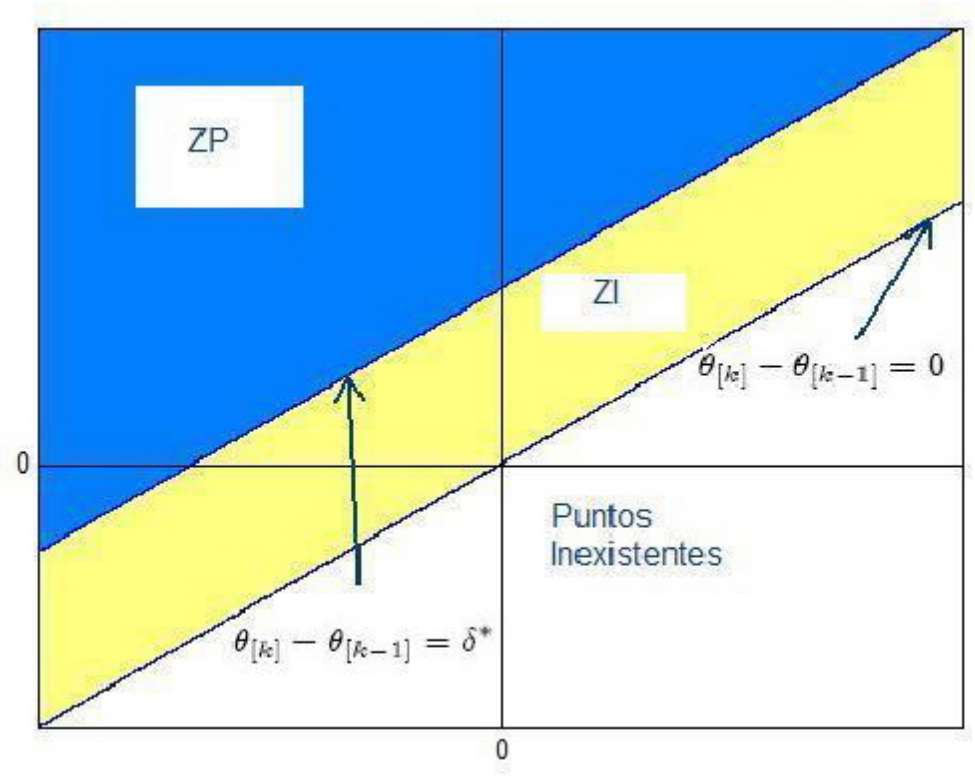

Fig. 1. Ejemplo de la representación gráfica de las zonas de preferencia (ZP) y de indiferencia (ZI) inducidas por la medida de discrepancia

2.2.2 Probabilidad de una selección correcta

De aquí en adelante denotaremos a la selección correcta por SC. Entonces, la probabilidad de una selección correcta bajo cualquier configuración $\theta \in \Theta$ es

$$P(SC|\theta).$$

Esta probabilidad debe de ser grande si $\theta \in ZP$ y pequeña si $\theta \in ZI$. Siendo nuestro objetivo seleccionar a la mejor población, centraremos nuestra atención de SC en ZP. Aunque puede suceder que en la ZP existan infinitas configuraciones lo que complicaría en gran medida el problema de encontrar $P(SC|\theta)$. Sin embargo, puede existir una configuración en ZP en la cual $P(SC|\theta)$ sea mínima. A esta configuración le llamaremos la configuración menos favorable (CMF) y se denotará $\theta_{MF} = (\theta_{1,MF}, \theta_{2,MF}, ...\theta_{k,MF})$. Cuando esta configuración exista centraremos nuestra atención solamente a ella, puesto que nos brinda una cota inferior a todas las probabilidades de una selección correcta. De esta manera habremos simplificado el problema.

$$P(SC|\theta) \geq P(SC|\theta_{MF}), \forall \theta \in ZP.$$

Es importante aclarar que la θ_{MF} puede no ser un vector sino también puede ser un conjunto de vectores, lo que explica el porque de su nombre: *configuración menos favorable,* sin

embargo, es importante notar que $P(SC|\theta) = P^*$ para $\forall\theta \in$ CMF.

Hasta aquí tenemos todos los elementos necesarios para poder implementar un procedimiento de selección estadística. Vamos a ilustrarlo.

Supongamos que tenemos k poblaciones con cierta distribución conocida, y dentro de estas queremos seleccionar a la mejor de acuerdo al valor más grande de un cierto parámetro o ciertos parámetros de un vector de parámetros θ. El primer paso es establecer una estadística apropiada T. Esta estadística deberá ser apropiada de acuerdo al parámetro o parámetros de interés, así como también a la distribución de las poblaciones y al objetivo del procedimiento. Denotemos a T_i como el valor de la estadística de la población i *ésima* con $i \in \{1, 2, ..., k\}$. Denotemos a $T_{[i]}$ a los estadísticos de orden, entonces tendremos

$$T_{[1]} \leq T_{[2]} \leq ... \leq T_{[k]}.$$

Como nuestro objetivo es seleccionar a la población que tenga $\theta_{[k]}$, entonces el procedimiento de selección R es simplemente escoger a la población de la que provenga $T_{[k]}$. Esta manera de proceder es bastante aceptable, aunque surgen de manera natural dos puntos cruciales que están ligados.

1) Estimación del tamaño de las k muestras, elegidas de las poblaciones correspondientes.

2) Estimación de la verdadera probabilidad de una selección correcta.

En este trabajo centraremos nuestra atención en la estimación del tamaño mínimo muestral de las k poblaciones de tal manera que podamos tener una confiabilidad establecida.

2.2.3 Determinación del tamaño de muestra

Para determinar el tamaño de muestra asumimos que el experimento está en la etapa de diseño, donde requerimos determinar el tamaño de muestra de cada población para poder establecer que el procedimiento de selección que hemos descrito nos lleve a una selección correcta con una probabilidad determinada. Sea k el número de poblaciones con las que contamos. Nuestro objetivo es el determinar los tamaños de muestra, n_i, $i \in \{1, 2, ..., k\}$ de tal manera que la probabilidad de una selección correcta sea al menos un P^* especificado. Por lo que, para poder implementar esta metodología el investigador necesita establecer P^* y como se ve en la sección anterior, se necesita también la especificación del umbral δ^*. Hay que notar que se debe tener $P^* > \frac{1}{k}$, pues si no lo es, podemos alcanzar tal confiabilidad sin siquiera tener la necesidad de obtener datos. Utilizando la notación de la sección anterior tenemos que nuestro objetivo es obtener los tamaños de muestra mínimos para poder asegurar lo siguiente:

$$P(SC|\theta) \geq P^*,$$

para $\forall\theta \in ZP$.

2.3 Generalización de los aspectos analíticos del problema de selección de las s mejores poblaciones

La filosofía del problema de selección de las s mejores poblaciones es análoga al contexto de la selección de la "mejor "población. El contexto en este caso es el siguiente: queremos seleccionar al "mejor" grupo de poblaciones, en donde, este grupo esta constituido por s poblaciones de un total de k. Dada esta situación, el espacio parametral cambia, pues la *zona de preferencia (ZP)* y la *zona de indiferencia (ZI)* son modificadas, veamos porque:

tenemos k poblaciones y nuestro objetivo es seleccionar a las s "mejores" dentro de éstas, donde las s mejores serán las que tengan los parámetros θ más grandes (Estamos suponiendo que $k \geq 2$). Es decir, en notación de estadísticas de orden, queremos seleccionar a $\theta_{[k]}, \theta_{[k-1]}, ..., \theta_{[k-s+1]}$.Bajo esta notación estamos considerando a todos los parámetros $\theta_{[1]}, \theta_{[2]}, ..., \theta_{[k]}$. Como queremos seleccionar a las s "mejores" es natural considerar nada más la diferencia entre $\theta_{[k-s+1]}$ y $\theta_{[k-s]}$, pues si la diferencia entre estas nos permite diferenciar a la mejor, también lo hará la diferencia entre las demás. Así, por las propiedades de las estadísticas de orden, hemos reducido la dimensión del espacio parametral. El espacio que ha resultado es

$$\Xi_s = \{(\theta_{[k-s+1]}, \theta_{[k-s]}) | \theta_{[i]} \in \mathbb{R}, i \in \{k-s+1, k-s\}\} \subset \mathbb{R}^2.$$

Al igual que en el contexto de la selección de la mejor población, definiremos como la medida de discrepancia a:

$$\delta(\theta_{[k-s+1]}, \theta_{[k-s]}) = \theta_{[k-s+1]} - \theta_{[k-s]},$$

Por lo ya antes mencionado, por nuestras limitaciones en la metodología de medición y también por las condiciones impuestas por el contexto del problema del investigador, se fija un umbral, a partir del cual se definen las zonas de preferencia (ZP) y de indiferencia (ZI), de la siguiente manera:

ZP	$\theta_{[k-s+1]} - \theta_{[k-s]} \geq \delta^*$
ZI	$\theta_{[k-s+1]} - \theta_{[k-s]} < \delta^*$

en donde δ^* es el umbral definido por el investigador de acuerdo al contexto del problema.

El establecimiento de la probabilidad de selección correcta será análoga al caso de la selección de la "mejor" población, puesto que el espacio parametral Ξ_s es solamente una proyección distinta del espacio paramatral Ξ.

3 Selección de las s mejores poblaciones con distribución Weibull, con los s parámetros de escala más grande con parámetros de forma conocidos

En esta sección desarrollaremos la idea general de la obtención del tamaño muestral que nos permitirá seleccionar las s poblaciones dentro de un conjunto de k poblaciones, con distribución Weibull, que tengan el parámetro de escala más grande. El desarrollo de este procedimiento de selección esta basado en los trabajos para la selección de las s poblaciones, con distribución normal, que tengan la media mas grande dentro de un conjunto de k poblaciones. ([1], [2],[3] y [4]). En el contexto de las poblaciones con distribución normal, se supondrá que las varianzas son conocidas. En este contexto, en que las poblaciones tienen distribución Weibull, suponemos que los parámetros de forma son conocidos e iguales.

La selección de la mejor población dentro de un conjunto de k poblaciones es un caso particular del caso general que aquí se muestra, ya que este caso se tiene cuando $s = 1$.

3.1 Planteamiento del problema

Se tienen k poblaciones independientes con distribución Weibull. De estas se han recolectado las observaciones $T_{i,j} \sim Weibull(\eta_i, \beta)$ donde $1 \leq i \leq k$ y $1 \leq j \leq n_k$.Aquí n_k es el tamaño de la muestra que elegimos de la $k - ésima$ población. En este contexto, estamos asumiendo que el parámetro de escala η_i es desconocido y que el parámetro de forma, β, es conocido. Caracterizaremos a cada población por su parámetro de escala y definiremos como las s mejores poblaciones a las que tengan los s parámetros de escala más grandes. Esta selección se deseará con una probabilidad de al menos P^* y de acuerdo al contexto del problema en general se desea que los s parámetros de escala mas grandes se diferencien de los restantes $k - s$ en al menos δ_u^*.

Dado que las observaciones tienen distribución Weibull se tiene lo siguiente

$$T_j = \sum_{i=1}^{n_j} T_{i,j}^{\beta} \sim \Gamma\left(\eta_j^{\beta}, n_j\right), \tag{1}$$

donde $1 \leq j \leq k$. A partir de lo anterior se tiene que

$$\begin{aligned} E(T_j) &= \mu_j = n_j \eta_j^{\beta} \\ Var(T_j) &= \sigma_j^2 = n_j \eta_j^{2\beta} \end{aligned}$$

Para la selección de las s mejores poblaciones, necesitamos una manera de medir las discrepancias. Una manera de hacerlo es tomando la diferencia de los valores esperados de las estadísticas. Pero si procedemos de esta manera nos enfrentaremos a una complicación, que la distancia estará dependiendo, además del parámetro de escala, del tamaño de la muestra;

y eso es algo que no se desea. Evitamos esta complicación, tomando el cociente de los valores esperados e introduciendo la restricción de que en todas las poblaciones tomaremos el mismo tamaño de muestra. Procediendo de esta manera y teniendo en cuenta el análisis hecho en la sección *1.3,* tendremos lo siguiente:

$$\delta^* \left(\mu_{[k-s+1]}, \mu_{[k-s]}\right) = \frac{\mu_{[k-s]}}{\mu_{[k-s+1]}} = \frac{n\eta^{\beta}_{[k-s]}}{n\eta^{\beta}_{[k-s+1]}} = \frac{\eta^{\beta}_{[k-s]}}{\eta^{\beta}_{[k-s+1]}}$$

Notemos que $\delta^* : \mathbb{R}^2_+ \longrightarrow (0,1]$. Teniendo en cuenta de que el objetivo es seleccionar a las s "mejores" poblaciones que difieran, en cociente, un cierto valor δ^*_u con respecto a los parámetros de escala de la distribución Weibull de las $k-s$ poblaciones restantes, el umbral elegido fue: $(\delta^*_u)^{\beta}$.

Antes de seguir, introduciremos una restricción mas, que consiste en que las varianzas muestrales sean todas iguales: $Var\,(T_j) = c$, para $1 \leq j \leq k$.

3.2 Obtención del tamaño de muestra bajo la probabilidad de una selección correcta

El objetivo es determinar el tamaño muestral de las poblaciones para seleccionar a las s "mejores" poblaciones, que difieran de las demás en un cierto umbral. Esta selección se hará tomando en consideración el error que se pueda cometer, establecido en la sección *1.1.* Este error se cuantificará con la probabilidad con la que se quiera hacer la selección correcta. Por esta razón la probabilidad de selección correcta será función del tamaño muestral (n) y dependerá del umbral que se fije $(\delta^{*\beta}_u)$ y de la probabilidad con la que se quiera hacer la selección correcta (P^*).

La manera de identificar las s "mejores" poblaciones será mediante las estadísticas T_j dadas en (1). Denotaremos por $T_{(j)}$ a las estadísticas de orden, que satisfacen $E\left(T_{(j)}\right) = \mu_{[j]}$, con $1 \leq i \leq k$. Como se quiere seleccionar a las s "mejores" poblaciones que difieran en un umbral dado, queremos calcular la probabilidad de que

$$\max\left(T_{(1)}, T_{(2)}, ..., T_{(k-s)}\right) < T_{(k-s+1)} < \min\left(T_{(k-s+2)}, T_{(k-s+3)}, ..., T_{(k)}\right),$$

puesto que lo único que nos importa es seleccionar a las s "mejores" poblaciones sin importar el orden de éstas siempre y cuando difieran de las $k-s$ restantes en un cierto umbral que se fijará con respecto a la población $k-s+1$. A partir de esto se tiene la expresión siguiente:

$$sP\left[\max\left(T_{(1)}, T_{(2)}, ..., T_{(k-s)}\right) < T_{(k-s+1)} < \min\left(T_{(k-s+2)}, T_{(k-s+3)}, ..., T_{(k)}\right)\right] = P^* \qquad (2)$$

Es fácil ver que (2) es equivalente a

$$sP[\left(T_{(1)}, T_{(2)}, ..., T_{(k-1)}, T_{(k)}\right) \in C] = P^* \qquad (3)$$

en donde $C = \{\Big(T_{(1)}, T_{(2)}, ..., T_{(k-1)}, T_{(k)}\Big) \in \mathbb{R}^k | 0 < T_{(i)} < T_{(k-s+1)} \wedge T_{(k-s+1)} < T_{(j)} < \infty \wedge ...$

$...0 < T_{(k-s+1)} < \infty\}$, con $1 \leq i \leq k-s$ y $k-s+2 \leq j \leq k$. Desarrollando (3) tenemos lo siguiente:

$$sP[\Big(T_{(1)}, T_{(2)}, ..., T_{(k-1)}, T_{(k)}\Big) \in C] =$$

$$s\int_0^\infty \int_{t_{(k-s+1)}}^\infty \cdots \int_{t_{(k-s+1)}}^\infty \int_0^{t_{(k-s+1)}} \cdots \int_0^{t_{(k-s+1)}} \prod_{j=1}^{k} \frac{t_{(j)}^{n_{(j)}-1} e^{-\frac{t_{(j)}}{\eta_{[j]}^\beta}}}{\Gamma(n_{(j)})\Big(\eta_{[j]}^\beta\Big)^{n_{(j)}}} dt_{(j)} dt_{(k-s+1)} =$$

$$s\int_0^\infty \frac{t_{(k-s+1)}^{n-1} e^{-\frac{t_{(k-s+1)}}{\eta_{[k-s+1]}^\beta}}}{\Gamma(n)\Big(\eta_{[k-s+1]}^\beta\Big)^n} \prod_{j=1}^{k-s} \Gamma_{\Big(\eta_{[j]}^\beta, n\Big)}(t_{(k-s+1)}) \prod_{j=k-s+2}^{k} \left[1 - \Gamma_{\Big(\eta_{[j]}^\beta, n\Big)}(t_{(k-s+1)})\right] dt_{(k-s+1)} =$$

$$s\int_0^\infty \gamma_{\Big(\eta_{[k-s+1]}^\beta, n\Big)}(t_{(k-s+1)}) \prod_{j=1}^{k-s} \Gamma_{\Big(\eta_{[j]}^\beta, n\Big)}(t_{(k-s+1)}) \prod_{j=k-s+2}^{k} \left[1 - \Gamma_{\Big(\eta_{[j]}^\beta, n\Big)}(t_{(k-s+1)})\right] dt_{(k-s+1)} = P^*,$$

donde γ es la función de densidad de la distribución Gama y Γ es la función de distribución. Ahora por la restricción de que las varianzas muestrales son todas iguales a una constante c tenemos que:

$$n\eta_j^{2\beta} = c \Leftrightarrow \eta_j^\beta = \sqrt{\frac{c}{n}}.$$

Por el establecimiento de δ_u^* y por el establecimiento de la medida de discrepancia, tenemos que el desarrollo anterior se reduce a:

$$\begin{aligned} P^* &= s\int_0^\infty \gamma_{\Big(\eta_{[k-s+1]}^\beta, n\Big)}(t_{(k-s+1)}) \prod_{j=1}^{k-s} \Gamma_{\Big(\eta_{[j]}^\beta, n\Big)}(t_{(k-s+1)}) \prod_{j=k-s+2}^{k} \left[1 - \Gamma_{\Big(\eta_{[j]}^\beta, n\Big)}(t_{(k-s+1)})\right] dt_{(k-s+1)} \quad (4) \\ &= s\int_0^\infty \gamma_{\left(\sqrt{\frac{c}{n}}, n\right)}(t_{(k-s+1)}) \prod_{j=1}^{k-s} \Gamma_{\left((\delta_u^*)^\beta\sqrt{\frac{c}{n}}, n\right)}(t_{(k-s+1)}) \prod_{j=k-s+2}^{k} \left[1 - \Gamma_{\left(\sqrt{\frac{c}{n}}, n\right)}(t_{(k-s+1)})\right] dt_{(k-s+1)} \\ &= s\int_0^\infty \gamma_{\left(\sqrt{\frac{c}{n}}, n\right)}(t_{(k-s+1)}) \left[\Gamma_{\left((\delta_u^*)^\beta\sqrt{\frac{c}{n}}, n\right)}(t_{(k-s+1)})\right]^{k-s} \left[1 - \Gamma_{\left(\sqrt{\frac{c}{n}}, n\right)}(t_{(k-s+1)})\right]^{s-1} dt_{(k-s+1)}. \end{aligned}$$

El tamaño muestral n que deseamos conocer, es la solución a esta ecuación, la cual no tiene forma cerrada y por lo tanto debemos obtenerla en forma numérica. Para esto, definamos a la siguiente función:

$$G(n) = s\int_0^\infty \gamma_{\left(\sqrt{\frac{c}{n}}, n\right)}(t_{(k-s+1)}) \left[\Gamma_{\left((\delta_u^*)^\beta\sqrt{\frac{c}{n}}, n\right)}(t_{(k-s+1)})\right]^{k-s} \left[1 - \Gamma_{\left(\sqrt{\frac{c}{n}}, n\right)}(t_{(k-s+1)})\right]^{s-1} dt_{(k-s+1)}. \quad (5)$$

Reparametrizamos ésta función de la siguiente manera: sea $t_{(k-s+1)} = \Gamma^{-1}_{\left(\sqrt{\frac{c}{n}}, n\right)}(z_{(k-s+1)})$. Por la regla de la cadena sabemos que $(f^{-1}(x))' = \frac{dx}{f'(f^{-1}(x))}$. Aplicando este cambio de variable

a la función de distribución Gama, tenemos que:

$$\left(\Gamma^{-1}_{\left(\sqrt{\frac{c}{n}},n\right)}(x)\right)' = \frac{dx}{\Gamma'_{\left(\sqrt{\frac{c}{n}},n\right)}\left(\Gamma^{-1}_{\left(\sqrt{\frac{c}{n}},n\right)}(x)\right)} = \frac{dx}{\gamma_{\left(\sqrt{\frac{c}{n}},n\right)}\left(\Gamma^{-1}_{\left(\sqrt{\frac{c}{n}},n\right)}(x)\right)},$$

por lo que (5) se reduce a lo siguiente:

$$G(n) = s\int_0^\infty \gamma_{\left(\sqrt{\frac{c}{n}},n\right)}(t_{(k-s+1)})\left[\Gamma_{\left((\delta_u^*)^\beta\sqrt{\frac{c}{n}},n\right)}(t_{(k-s+1)})\right]^{k-s}\left[1-\Gamma_{\left(\sqrt{\frac{c}{n}},n\right)}(t_{(k-s+1)})\right]^{s-1} dt_{(k-s+1)} =$$

$$s\int_0^\infty \frac{\gamma_{\left(\sqrt{\frac{c}{n}},n\right)}\left(\Gamma^{-1}_{\left(\sqrt{\frac{c}{n}},n\right)}(z_{(k-s+1)})\right)\left[\Gamma_{\left((\delta_u^*)^\beta\sqrt{\frac{c}{n}},n\right)}\left(\Gamma^{-1}_{\left(\sqrt{\frac{c}{n}},n\right)}(z_{(k-s+1)})\right)\right]^{k-s}}{\gamma_{\left(\sqrt{\frac{c}{n}},n\right)}\left(\Gamma^{-1}_{\left(\sqrt{\frac{c}{n}},n\right)}(z_{(k-s+1)})\right)} * \ldots$$

$$\ldots \frac{\left[1-\Gamma_{\left(\sqrt{\frac{c}{n}},n\right)}\left(\Gamma^{-1}_{\left(\sqrt{\frac{c}{n}},n\right)}(z_{(k-s+1)})\right)\right]^{s-1} dz_{(k-s+1)}}{\gamma_{\left(\sqrt{\frac{c}{n}},n\right)}\left(\Gamma^{-1}_{\left(\sqrt{\frac{c}{n}},n\right)}(z_{(k-s+1)})\right)}$$

$$s\int_0^1 \left[\Gamma_{\left((\delta_u^*)^\beta\sqrt{\frac{c}{n}},n\right)}\left(\Gamma^{-1}_{\left(\sqrt{\frac{c}{n}},n\right)}(z_{(k-s+1)})\right)\right]^{k-s}\left[1-z_{(k-s+1)}\right]^{s-1} dz_{(k-s+1)} = P^*.$$

A partir del comportamiento de la función $G(n)$, el cálculo numérico de n se obtiene mediante algoritmos de integración y bisección, estableciendo inicialmente los valores de el umbral de separación de las poblaciones ($\delta_u^{*\beta}$), el número de poblaciones de interés (k y s), la probabilidad con la que se desea hacer la selección correcta (P^*) y el conocimiento de los parámetros de forma de las poblaciones (β). Estos algoritmos fueron implementados el programa *Tamaño Muestral*, codificado en C++.

3.3 El programa Tamaño Muestral

El desarrollo de la subsección anterior deja en claro que no existe una expresión cerrada para el cálculo del tamaño muestral para la selección de las s poblaciones con distribuciones Weibull con el parámetro de escala mas grande, dentro de un conjunto de k poblaciones. Es por esta razón que se implementó el desarrollo de la sección anterior en un programa llamado *Tamaño Muestral*, en el cual los cálculos se desarrollan de manera numérica. Se presenta una figura de la ventana del programa para la obtención del tamaño muestral para distribuciones Weibull con el mismo parámetro de forma. Para la determinación del tamaño muestral, al igual que se mostró en la subsección anterior, el programa necesita como entradas lo siguiente:

1) La probablidad P^*, con la cual se requiere realizar la selección correcta de las s poblaciones

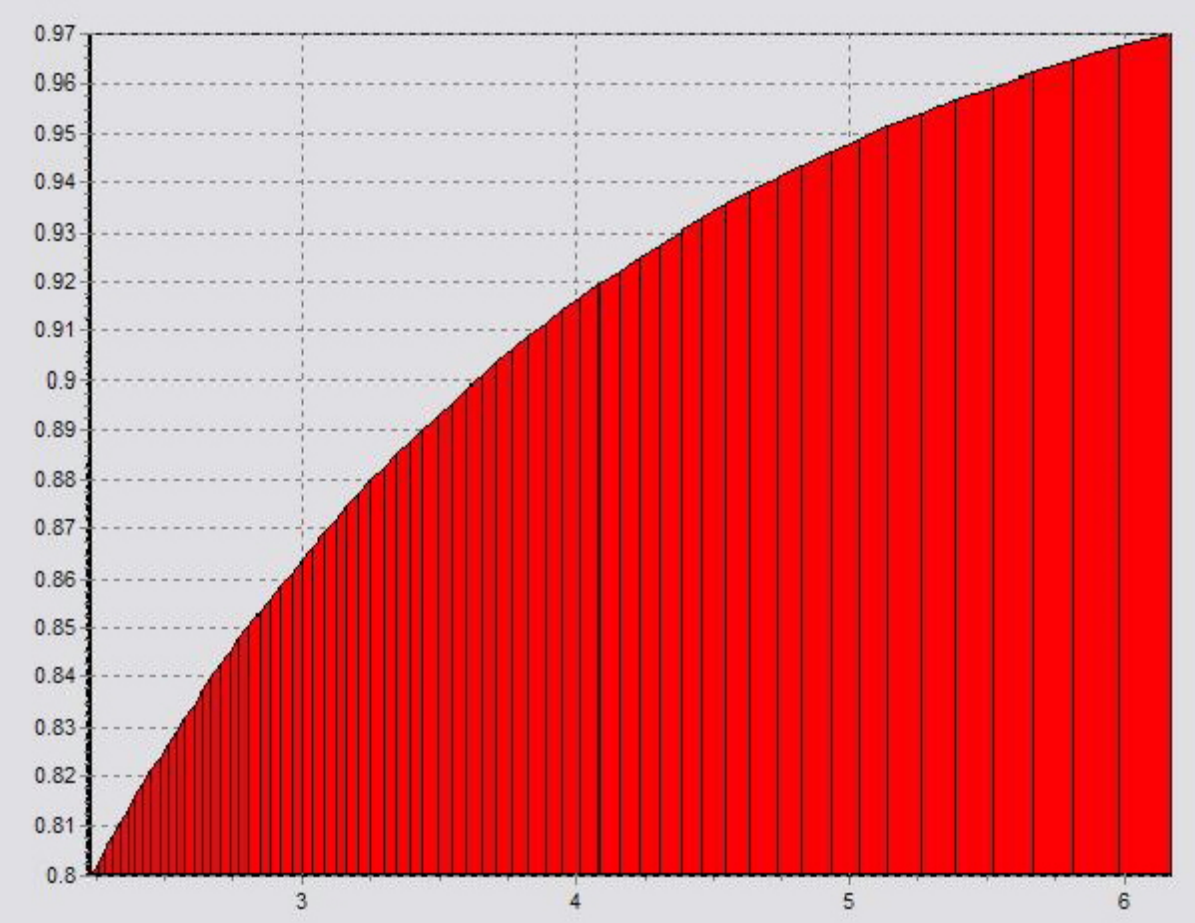

Fig. 2. Ejemplo del comportamiento de la función $G(n)$.

con el parámetro de escala mas grande. En la ventana del programa este parámetro estará representado mediante P^*.

2) Parámetro de forma de las poblaciones Weibull, β, que en este caso puede ser una estimación de este. En la ventana del programa este parámetro se representará como: *Parámetro de forma.*

3) El parámetro de discrepancia, δ_u^*, el cual en la ventana del programa será representado como *Delta.*

4) El número total de poblaciones con el cual se cuenta, K, que tiene la misma representación en la ventana del programa.

5) El número de poblaciones que se quiere seleccionar, S, que tiene la misma representación en la ventana del programa.

Una vez que se ha establecido los parámetros de entrada en la ventana del programa, se procede a calcular el tamaño muestral requerido. El programa, una vez que calcula este tamaño muestral da como resultados una gráfica de la variación de la probabilidad en función

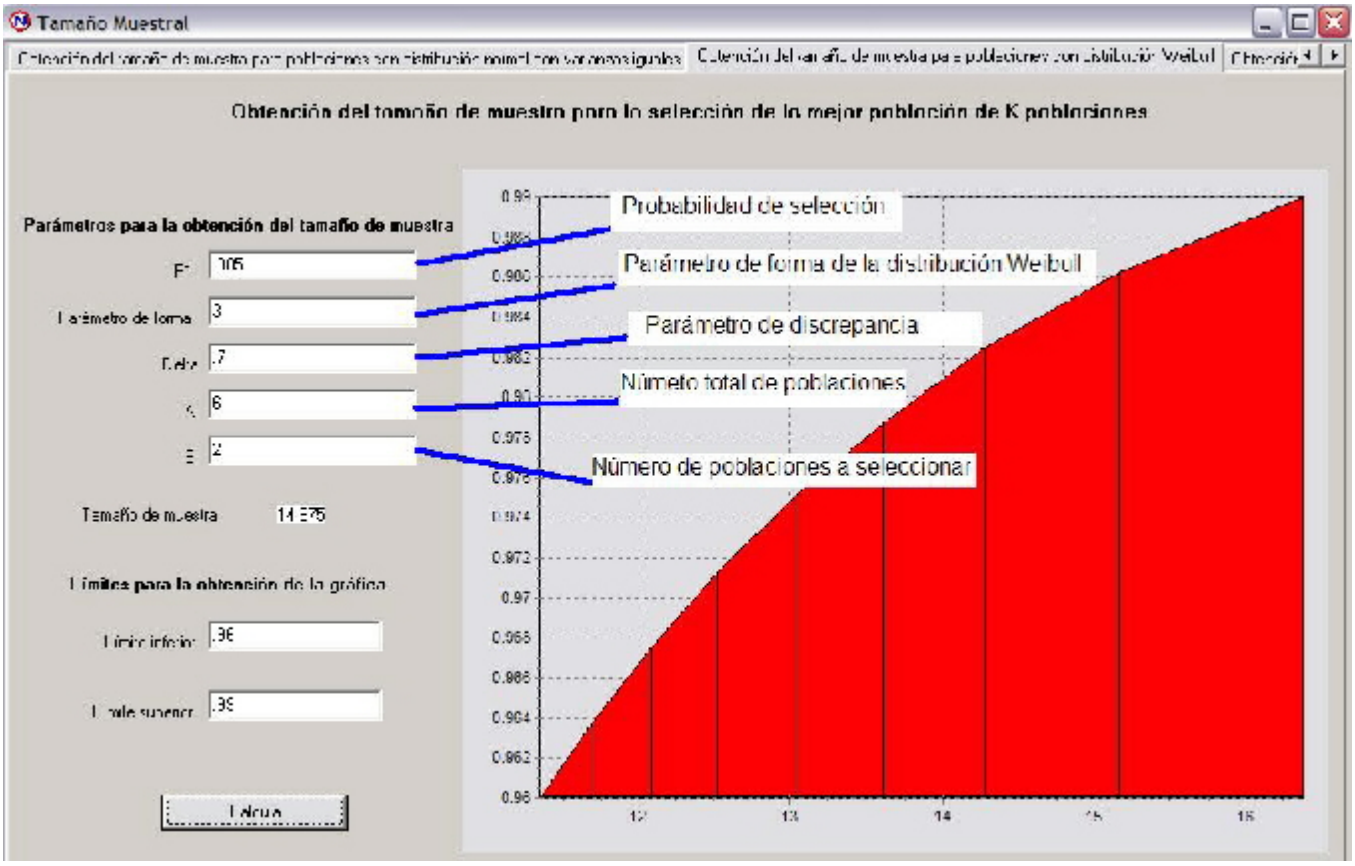

Fig. 3. Ejemplo de la ventana del programa para la determinación del tamaño muestral para la selección de las s poblaciones Weibull con mayor parámetro de escala dentro de k poblaciones. En el ejemplo las entradas del programa son del ejemplo presentado

del tamaño de muestra. El usuario especifica los límites de la gráfica, estableciendo el límite superior de la probabilidad hasta donde uno quiere ver la gráfica al igual que establece el límite inferior de la probabilidad. También el programa da como resultado el tamaño muestral como función de los parámetros de entrada. Se anexa una figura en donde se presenta en funcionamiento el programa *Tamaño Muestral* con los parámetros del ejemplo que se presentará a continuación.

4 Ejemplo

Supongamos que el departamento de asuntos legales de una compañía automovilística en la que trabajamos, nos pide seleccionar a los 2 activadores de las bolsas de aire que duren el mayor tiempo posible en buenas condiciones, para después, haciendo un estudio de mercado se seleccione al activador de bolsas de aire que mas convenga para poder despues, determinar el tiempo de garantia del automóvil al cliente. En el laboratorio, tenemos 6 sistemas de activación de bolsas de aire y los tiempos de vida de los activadores tienen distribución Weibull. Se asume, por la naturaleza del experimento y por los sistemas de activación que

hay en el mercado, que los parámetros de forma de los 6 activadores que se tienen son iguales a 3. Como el departamento legal junto con el departamento financiero han determinado que si se tiene una diferencia en tiempos de vida de un activador o a otro de a lo más en 1 mes y medio, les dará igual seleccionar a uno u a otro, entonces elegimos un valor umbral igual a 1.5 meses. Como no se quiere tener pérdidas de dinero innecesarias y tampoco se quiere un desprestigio a la marca del automóvil, han fijado que se requiere una selección correcta con una probabilidad del 98.5%. ¿Cuantos activadores se necesitan de cada modelo de activadores de bolsas de aire diseñados, para poder hacer una selección correcta?, teniendo en cuenta que para realizar este experimento el costo de todo el inmobiliario del chasis es de \$2000 USD. Los valores de entrada en el programa $Tamaño\ Muestral$ son en este ejemplo los siguientes:

1) La probabilidad de selección correcta: $P^* = 0.985$.

2) $Parámetro\ de\ forma = 3$ ($\beta = 3$).

3) Como la diferencia que queremos observar es de 1.5 meses, tenemos que $\delta_u^* = Delta =$ .6666 (que es el inverso de 1.5).

4) El número total de poblaciones con las que se cuenta es $K = 6$.

5) El número de poblaciones que se quiere seleccionar es $S = 2$.

La solución arrojada por el programa es 11.59375, lo cual fue redondeado a 12 activadores para cada población. Así, deberán someterse a una prueba de vida 12 activadores de cada proveedor y se eligen como los 2 mejores activadores, a los activadores con mayor tiempo de vida característica.

5 Conclusiones

El objetivo de este trabajo es determinar el tamaño muestral necesario para la selección de las mejores poblaciones, cuando estas tienen distribución Weibull. Se consideran procedimientos de selección para la distribución Weibull y se obtiene un procedimiento numérico para determinar el tamaño muestral necesario.

Debido a que en el caso considerado no hay expresiones cerradas para obtener el tamaño muestral, se desarrolló un software ($TamañoMuestral$) para obtener el tamaño muestral necesario para la selección correcta de poblaciones.

Como trabajo a futuro se desprende de aquí el estudio de la determinación del tamaño muestral para la selección correcta de las s de k poblaciones distribuidas como Weibull, con mayor vida característica para parámetros de forma tanto distintos. Una tarea a futuro de interés es investigar la obtención del tamaño muestral para la selección correcta de las mejores poblaciones, en donde la comparación se haga no a través de la vida característica, sino a

través de los cuantiles o a través de la confiabilidad a un tiempo de referencia determinado, que puede ser un tiempo de garantía o un tiempo de vida útil del producto.

6 Bibliografía

[1]. Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Ann. Math. Stat.* **25**, 16-39.

[2]. Bechhofer, R. E., Hayter, A. J. y Tamhane, A. C. (1991). Designing experiments for selecting the largest normal mean when the variances are known and unequal: Optimal sample size allocation. *J. Stat. Plan. Infer.* **28**, 271-289.

[3]. Bechhofer, R. E., Santner, T. J. y Goldsman, D. M. (1995). *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons.* Wiley Series in Probability and Statistics.

[4]. Gibbons, J. D., Olkin, I y Sobel, M. (1977). *Selecting and Ordering Populations: A New Statistical Methodology.* New York: John Wiley & Sons.