INVESTIGACIÓN & DESARROLLO, No. **10**, Vol. **1**: 131 – 149 (2010)
ISSN 2518-4431

USO DE ÁRBOLES DE DECISIÓN PARA LA ESTIMACIÓN ESTADÍSTICA

Carlos Eduardo Valdivieso Taborga, Roberto Valdivieso Castellón y Oscar Álvaro Valdivieso Taborga
Universidad Privada Boliviana
cvaldivieso@upb.edu
(Recibido el 25 de octubre 2010, aceptado para publicación el 17 de diciembre 2010)

RESUMEN

El artículo muestra el desarrollo de árboles de decisión para facilitar al investigador o profesional la elección del intervalo de confianza adecuado, a fin de estimar parámetros, como la media, proporción o varianza poblacionales o la diferencia de medias, proporciones o cociente de varianzas. Esta elección es compleja, ya que depende de muchos factores, entre los cuales están: el número de muestras comparadas, la normalidad de las poblaciones de las que provienen las muestras, el tamaño de las muestras, el conocimiento de las varianzas poblacionales, la dependencia de las muestras y la igualdad de las varianzas poblacionales.
Se han configurado 4 árboles de decisión tomando en cuenta estos factores y otras consideraciones teóricas y empíricas: Un árbol general que ayuda a establecer el intervalo de confianza adecuado, otro para estimar la media poblacional, otro para estimar la diferencia de medias y el último para la proporción o diferencia de proporciones. Se exponen dos ejemplos para ilustrar su fácil aplicación y se dan conclusiones e implicaciones de su uso.

Palabras Clave: Estadística Inferencial, Árboles de Decisión, Estimación Estadística, Intervalos de Confianza, Estadística Educacional.

1. INTRODUCCIÓN

La elección adecuada de un intervalo de confianza para la estimación de parámetros poblacionales usando la estimación estadística, por parte de investigadores y profesionales que no tienen formación estadística, se hace muy dificultosa, debido a los siguientes aspectos:

- Poca dedicación por parte de los estadísticos a la generación de mecanismos o ayudas adecuadas para facilitar la elección.
- Bibliografía que enfatiza la estadística teórica, la matemática de la estadística y no así su uso en la resolución de problemas.
- Poca experiencia de los investigadores y profesionales en el uso de la estadística para la solución de problemas habituales de empresa e industria, en la toma de decisiones o en la investigación científica.
- Generación de expectativas erradas en cuanto al uso de paquetes estadísticos: en vez de constituirse en un instrumento que ayuda en los cálculos y el procesamiento, se pretende que resuelvan la necesidad de descripción e inferencia del fenómeno, prescindiendo del criterio del investigador.

Esta situación adversa, necesita una respuesta por parte de los estadísticos dedicados a la investigación y a la aplicación, a través del desarrollo de ayudas efectivas, que permitan aplicar la estimación estadística en el análisis de problemas y proposición de soluciones.

El área de la estimación estadística, es tal vez una de las que tiene mayor potencial para la implementación de ayudas, por su complejidad en cuanto a la elección del intervalo de confianza a usar. El investigador con falencias en su formación estadística, al tratar de usar este método inferencial, generalmente tiene bastantes problemas para elegir un intervalo de confianza adecuado al problema planteado, sin ayudas adicionales (resúmenes, cuadros sinópticos, diagramas de flujo, etc.) y de manera rápida, junto con el cálculo de valores críticos, que toman en cuenta distintas distribuciones de probabilidad (t de Student, distribución normal estándar z, chi-cuadrada χ^2, F de Fisher, etc.).

Los libros de Estadística (o libros relacionados como los de Control de Calidad, Análisis y Diseño de Experimentos, Econometría, etc.) que contienen el tema de la estimación estadística, presentan actualmente las siguientes ayudas para la elección:

- Diagramas de flujo: Levin y Rubin [2] y Berenson, Levine y Krehbiel [1], usan flujogramas (aunque no son exhaustivos) para ayudar a la decisión de elegir el intervalo adecuado; sin embargo, este método no tiene la contundencia deseada ya que ocuparía mucho espacio para plasmar todas las alternativas y en el diagrama no se pueden incluir las fórmulas de los intervalos.

DOI: 10.23881/idupbo.010.1-11i

- Tablas o resúmenes: Algunos autores usan tablas o resúmenes (que no son exhaustivas), como Freund y Simon [4]; Miller, Freund y Jonson [5]; Montgomery [12], Gutierrez y de la Vara [13] y Juran y Gryna [15].
- Ninguna ayuda adicional para la elección: La mayoría de los libros, como Mason y Lind [3]; Lobez y Casa [11]; Maisel [9]; García [7], Programa FORD-ITESM [14], Duncan [16], Yamane [19], Novales [20], entre otros, no presentan ninguna ayuda adicional para la elección del intervalo de confianza adecuado cuando se quiere estimar un parámetro poblacional.

Este artículo propone utilizar el árbol de decisión para lograr el objetivo de elección del intervalo de confianza apropiado. Mendenhall [6] es el único que presenta una herramienta similar, pero presenta un árbol muy simple y con pocas alternativas de análisis (Figura 1).

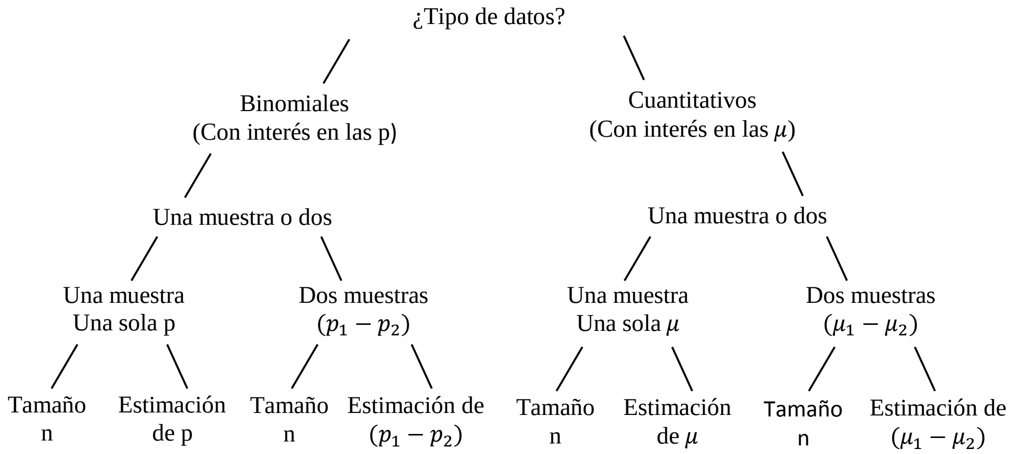


Figura 1 – Árbol de decisiones para la estimación estadística (Fuente: Mendenhall [6])

2. LA ESTADÍSTICA INFERENCIAL

2.1. Conceptos

Es la rama de la Estadística que se ocupa del uso de los conceptos de la probabilidad para afrontar el riesgo y la incertidumbre en la toma de decisiones. Los conceptos básicos de las distribuciones muestrales, sirven como base para la estadística inferencial. Esta comprende dos áreas:

- *La estimación*, que consiste en estimar los valores de los parámetros de la población bajo estudio.
- *Las pruebas de hipótesis*, que constituyen el proceso de aceptar o rechazar declaraciones o supuestos, generalmente relacionadas con características poblacionales.

Las pruebas de hipótesis pueden ser a su vez:

- *Pruebas paramétricas*, si lo que se desea es probar una hipótesis acerca de un parámetro poblacional en estudio. En varias de estas pruebas, se debe realizar la suposición de que las muestras obtenidas por el proceso de muestreo aleatorio deben haber sido extraídas de una población normal.
- *Pruebas no paramétricas*, si lo que se desea es probar una hipótesis en la cual no se requiere el uso de un parámetro poblacional y no se requiere realizar ninguna suposición acerca de la forma de la población.

En la Tabla 1 se muestra los principales métodos estadísticos inferenciales paramétricos y no paramétricos.

2.2. La estimación estadística

2.2.1. Introducción

En muchos experimentos, problemas o investigaciones del mundo real, se quiere saber si la diferencia de medias, proporciones o varianzas entre dos poblaciones es significativa, o a veces si la media, proporción o varianza se conforman a un valor patrón; también se está interesado en conocer un intervalo que se espera contenga el valor de la diferencia de medias, proporciones o varianzas de las dos poblaciones, o de la media, proporción o varianza de una población y, por lo tanto, se deberá usar la estimación estadística.

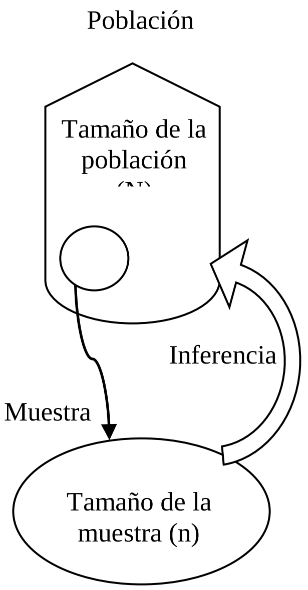
TABLA 1 - MÉTODOS ESTADÍSTICOS INFERENCIALES

Tipo de inferencia	Número de poblaciones		
	Una	Dos	Mayor a 2
Paramétrica	Media Varianza Proporción	Diferencia de medias Cociente de varianzas Diferencia de proporciones	ANOVA simple ANOVA de bloques
No paramétrica	Prueba de signos Prueba de bondad de ajuste	Prueba de signos Prueba U de Mann-Withney Prueba de independencia de atributos	Prueba H de Kruskal Wallis

Fuente: Elaboración propia

2.2.2. El proceso de la estimación estadística

Con los conceptos anteriores, es importante que el investigador entienda el proceso que se debe seguir para realizar una estimación. Se lo presenta en la Figura 2.



1. ¿Qué se quiere estimar?

$$\begin{cases} \text{Media de la población } (\mu) \\ \text{Proporción de la población (P)} \\ \text{Varianza de la población } (\sigma^2) \end{cases}$$

2. ¿Cuál es el procedimiento?

- Se determina el tamaño poblacional (N)
- Se determina el tamaño muestral (n)
- Se extrae una muestra representativa de la población
- Se determinan los estadísticos de la muestra

$$\begin{cases} \text{Media de la muestra } (\bar{x}) \\ \text{Proporción de la muestra } (\hat{p}) \\ \text{Varianza de la muestra } (s^2) \end{cases}$$

- A partir de los estadísticos se calcula el intervalo de confianza para estimar los parámetros poblacionales:

$$\begin{cases} \bar{x} \text{ para estimar } \mu \\ \hat{p} \text{ para estimar } P \\ s^2 \text{ para estimar } \sigma^2 \end{cases}$$

Figura 2 - Proceso de la estimación estadística.

Fuente: Elaboración propia.

El proceso de la estimación estadística, comienza con elegir qué parámetro se quiere estimar: la media, varianza o proporción de la población; o la comparación de medias, varianzas y proporciones poblacionales. Luego se determina el tamaño poblacional. Después, se procede a calcular el tamaño muestral representativo, se extrae una muestra aleatoria de la población en estudio y se determinan los estadísticos muestrales. Como último paso, mediante una estimación puntual o por intervalo, se estima el parámetro poblacional.

2.2.3. Tipos de estimaciones

Un estimador es un estadístico muestral con el cual se estima un parámetro de la población (media, proporción o varianza). Una estimación es un valor específico observado de un estadístico. En la vida real, se desconocen los valores de los parámetros de las poblaciones que se estudian. Se pretende entonces estimarlos mediante información muestral. Para esto puede hacerse una:

- *Estimación puntual*: es un valor único que pretende estimar el valor del parámetro. Dos estimadores puntuales son: la media muestral (para la media poblacional): $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, y la varianza muestral, (para la varianza poblacional): $s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$.

- *Estimación por intervalo*: es un intervalo numérico donde se pretende encontrar el valor del parámetro bajo estudio.

2.2.4. Criterios de un buen estimador [8], [3]

Un buen estimador debe cumplir con los siguientes criterios:

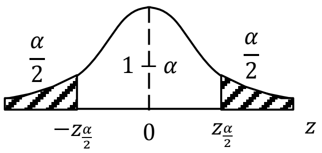
- Imparcialidad, cuando, en promedio, los valores del estimador coinciden con el verdadero valor del parámetro.
- Consistencia, si al aumentar el tamaño de la muestra, se logra una seguridad casi absoluta de que el valor del estadístico se acerca mucho al valor del parámetro.
- Eficiencia, designa el tamaño del error estándar del estadístico. Cuanto menor sea, el estimador es más eficiente.
- Suficiencia, si utiliza la información contenida en la muestra, al punto que ningún otro estimador podría extraer de esta última más información referente al parámetro de la población que va a ser estimado.

Los métodos para determinar los mejores estimadores son los de máxima verosimilitud y el de momentos (Mood/Graybill [8]; Maisel [9]; entre otros). Por estos métodos, se ha determinado que la media muestral es el mejor estimador para la media de la población y la desviación muestral de la desviación poblacional.

2.2.5. Estimación por intervalos

Una desventaja de los estimadores puntuales, es que no se sabe qué tan cerca se encuentran del valor del parámetro que estiman. Con el fin de obtener alguna medida de la precisión en una estimación, se determina un intervalo de valores (intervalo de confianza), que incluirá el valor del parámetro con una probabilidad prefijada $(1-\alpha)$, donde α es la probabilidad de que el intervalo no contenga al verdadero valor del parámetro (nivel de significancia). Los intervalos de confianza pueden ser construidos utilizando el método del pivote (basado en el teorema central del límite y la ley de los grandes números). A continuación se muestra un ejemplo.

De una población normal, de la cual se quiere estimar la media, pero por alguna razón se conoce su desviación estándar, se extrae una muestra aleatoria sin importar su tamaño. La variable tipificada, para una población infinita es: $z=\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$. Sabiendo que tiene una distribución normal:



El intervalo de confianza donde se pretende hallar la media de la población está comprendido entre $-z_{\frac{\alpha}{2}}$ y $z_{\frac{\alpha}{2}}$ y $1-\alpha$ indica la probabilidad de que la media poblacional se encuentre en ese intervalo:

$$P\left(-z_{\frac{\alpha}{2}} \leq z \leq z_{\frac{\alpha}{2}}\right)=1-\alpha$$

Sustituyendo z: $P\left(-z_{\frac{\alpha}{2}} \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right)=1-\alpha$. Multiplicando por el error estándar: $P\left(-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \bar{x}-\mu \leq z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$. Despejando la media de la desigualdad simultánea, se tiene el intervalo deseado:

$$P\left(\bar{x}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

Usando el método del pivote se ha construido un intervalo de confianza para la estimación de la media poblacional, cuando se conoce la varianza poblacional y el número de muestra es grande. Sin embargo, existen muchos otros intervalos para otras condiciones y objetivos.

3. DESARROLLO DE LOS ÁRBOLES DE DECISIÓN

3.1. Preguntas para la elección

Se ha deducido un intervalo de confianza en el acápite 2.2.5. Sin embargo, existen muchas combinaciones, ya que el intervalo depende de varias condiciones:

- ¿Cuántas poblaciones se están investigando? Pudiendo ser una o dos poblaciones.
- Si es una población, ¿qué parámetro poblacional se quiere estimar? Pudiendo existir las siguientes alternativas: media, proporción y varianza.
- Si se quieren comparar dos poblaciones, ¿qué parámetros poblacionales se quieren estimar? Pudiendo existir las siguientes alternativas: diferencia de medias, proporciones o cociente de varianzas.
- Si se quiere estimar la media poblacional, ¿la muestra proviene de una población normal o no?
- Si proviene de una población normal, ¿se conoce la varianza de la población o no? Si no se conoce, ¿el tamaño de la muestra es menor a 30 o mayor que ese valor?
- Si no proviene de una población normal, ¿se conoce la varianza de la población o no? Si se conoce, ¿el tamaño de la muestra es menor a 20 o mayor que ese valor? Si no se conoce, ¿el tamaño de la muestra es menor a 50 o mayor que ese valor?
- Si se quiere estimar la diferencia de medias poblacionales, ¿las varianzas de las poblaciones son conocidas?
- Si no son conocidas, ¿Los tamaños muestrales son menores a 30 o mayores? Si son mayores, ¿las muestras son dependientes o independientes? Si son menores, ¿las muestras son dependientes o independientes? Si son independientes, ¿las varianzas de las poblaciones son iguales o diferentes?
- Si se quiere estimar la diferencia de proporciones, ¿las varianzas son conocidas o no?, ¿son iguales o no?

3.2. Configuración de los árboles de decisión

Con la información contenida en las preguntas anteriores, se han configurado 4 árboles de decisión que permiten sistematizar la elección del intervalo de confianza adecuado.

- Figura 3: Muestra un árbol de decisión general para elegir el intervalo de confianza adecuado al problema que se quiere resolver. Si la elección es utilizar un intervalo de confianza para la media poblacional, se debe ir a la Figura 4, si es un intervalo de confianza para la diferencia de medias, se deberá ir a la Figura 5, y si es un intervalo de confianza para la proporción o diferencia de proporciones, remitirse a la Figura 6.
- Figura 4: Muestra un árbol de decisión para elegir el intervalo de confianza adecuado para la estimación de la media poblacional, tanto para poblaciones normales como no normales.
- Figura 5: Muestra un árbol de decisión para elegir el intervalo de confianza adecuado para la estimación de la diferencia de medias poblacionales, para poblaciones normales.
- Figura 6: Muestra un árbol de decisión para elegir el intervalo de confianza adecuado para la estimación de la proporción y la diferencia de proporciones poblacionales, para poblaciones normales.

3.3. Bases teóricas y empíricas para el desarrollo de los árboles de decisión

Vale la pena analizar cuáles han sido las bases teóricas y empíricas para la configuración de los tres árboles de decisión. Cabe recalcar primero, la importancia del teorema central del límite: “Si x_1, x_2, …, x_n son variables aleatorias que representan una muestra de una población con media μ y desviación estándar σ, la distribución de la media muestral es aproximadamente normal, con media μ y desviación estándar $\sigma/\sqrt{n}$, si n es suficientemente grande”. Este es un resultado muy importante, ya que asegura que a pesar de que la población bajo estudio no sea normal, la media muestral se distribuye aproximadamente normal, y que la aproximación mejora para valores grandes de n. La mayoría de los autores de textos formales de estadística aconsejan que el tamaño de la muestra no sea inferior a 25 o 30 elementos, para asegurar lo que establece el teorema (Programa FORD-ITESM [14]).

Figura 3: Árbol de decisión general para la estimación de intervalos de confianza

La mayor parte de los libros de consulta utilizados para realizar una estimación estadística, muestran que se puede estimar la media, varianza o proporción poblacionales y, si se quiere realizar comparaciones entre dos poblaciones, se pueden estimar la diferencia de medias, proporciones y el cociente de varianzas. Las consideraciones para la media, se darán cuando se analice el árbol de la Figura 4, ya que existen numerosas opciones.

No todos están de acuerdo con que la varianza pueda ser estimada con diferentes intervalos de confianza dependiendo del tamaño muestral y, aún más, no existe un acuerdo en el tamaño muestral límite para definir una muestra grande de una pequeña. Prácticamente la mayoría de los libros de consulta estiman la varianza poblacional mediante el intervalo de confianza que usa la distribución χ^2, con $n-1$ grados de libertad, sin importar el tamaño muestral, para poblaciones normales:

$$P\left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right) = 1-\alpha$$

Pocos autores (Freund y Simon [4], Duncan [16], el Programa FORD-ITESM [14] y Merril y Fox [18]) consideran que para muestras grandes, es pertinente usar el intervalo de confianza basado en la distribución normal estándar z:

$$P\left(\frac{s}{1+\frac{z_{\frac{\alpha}{2}}}{\sqrt{2n}}} \le \sigma \le \frac{s}{1-\frac{z_{\frac{\alpha}{2}}}{\sqrt{2n}}}\right) = 1-\alpha$$

Por otro lado, para usar este intervalo, el tamaño muestral que recomiendan algunos autores para muestras grandes es de n ≥100 (Merril y Fox [18], Lobez y Casa [11]). Sólo Freund y Simon [4], Duncan [16] y el Programa FORD-ITESM [14], admiten que para muestras $n \ge 30$ ya se puede usar la distribución z, aunque recomiendan preferiblemente valores mucho mayores.

Para el intervalo de diferencia de medias, los análisis teórico-empíricos se tratarán en el árbol de la Figura 5, ya que se presentan varias opciones.

Para el intervalo del cociente de varianzas solo existe una alternativa para cualquier tamaño muestral y está basada en la distribución *F* de Fisher, con n_1-1 y n_2-1 grados de libertad del numerador y el denominador, respectivamente, en poblaciones normales:

$$P\left(\frac{s_1^2}{s_2^2}\frac{1}{F_{\frac{\alpha}{2},n_1-1,n_2-1}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2}\frac{1}{F_{1-\frac{\alpha}{2},n_1-1,n_2-1}}\right) = 1-\alpha$$

Para el intervalo de proporciones o diferencia de proporciones, los análisis teórico-empíricos se tratarán en el árbol de la Figura 6, porque existen muchas opciones.

Figura 4: Árbol de decisión para la estimación de la media poblacional

García [7], Hays y Winkler [10] y muchos otros autores, afirman que cuando el tamaño muestral es mayor a 100 se puede sustituir la curva normal por la distribución *t* como distribución muestral. Sin embargo, también sostienen que empíricamente para $n \ge 30$, la media aritmética muestral presenta ya una distribución que se aproxima suficiente a la normal y que los intervalos de confianza resultantes son casi idénticos a los obtenidos utilizando la varianza verdadera; incluso Novales [20], afirma que para $n \ge 30$ la distribución *t* "coincide" con la normal. Cuando $n < 30$, es mejor utilizar la distribución *t* de Student. Esto se cumple si la muestra proviene de una población normal. Batattacharyya y Johnson [17], afirman que como una regla de trabajo, $n < 30$ es usualmente considerada como muestra pequeña y que se debe tener cuidado en invocar la distribución normal para la estimación. También sostienen que con respecto a las inferencias acerca de la media poblacional usando el estadístico *t*, el efecto de la eficiencia de la estimación no es demasiado serio si el tamaño muestral es moderadamente grande (digamos 15), y se puede considerar que se realiza una inferencia "robusta". Además, Hoel [24], denomina a los métodos de muestras pequeñas como "exactos" ya que también pueden aplicarse a muestras grandes; sin embargo, requieren más conocimientos e información que los métodos de muestras grandes, y por esa razón no pueden desplazarlos completamente.

Si una muestra proviene de una población que no es normal, y la varianza poblacional es conocida, el tamaño muestral para el uso de la distribución z es de 20, ya que es el tamaño establecido para diferenciar muestras grandes y pequeñas en las pruebas no paramétricas (Mason y Lind [3]; Freund y Simon [4]; Miller, Freund y Jonson [5]; Mendenhall [6], entre otros). Si la muestra es menor a 20, se deberá hacer uso del Teorema de Tchebysheff para el cálculo del valor crítico k (Hays y Winkler [10]; Freund y Simon [4]; Larson [21]).

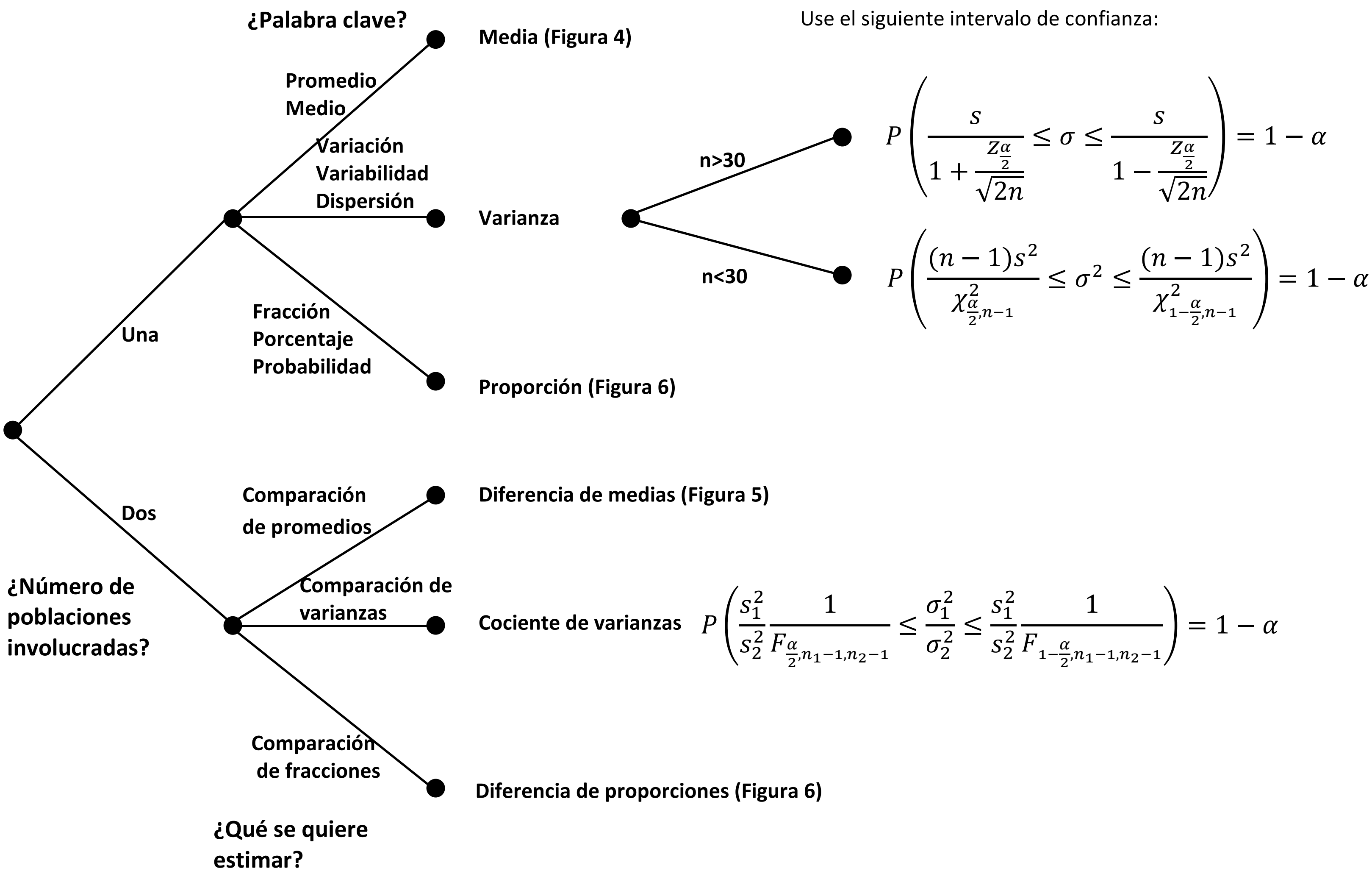
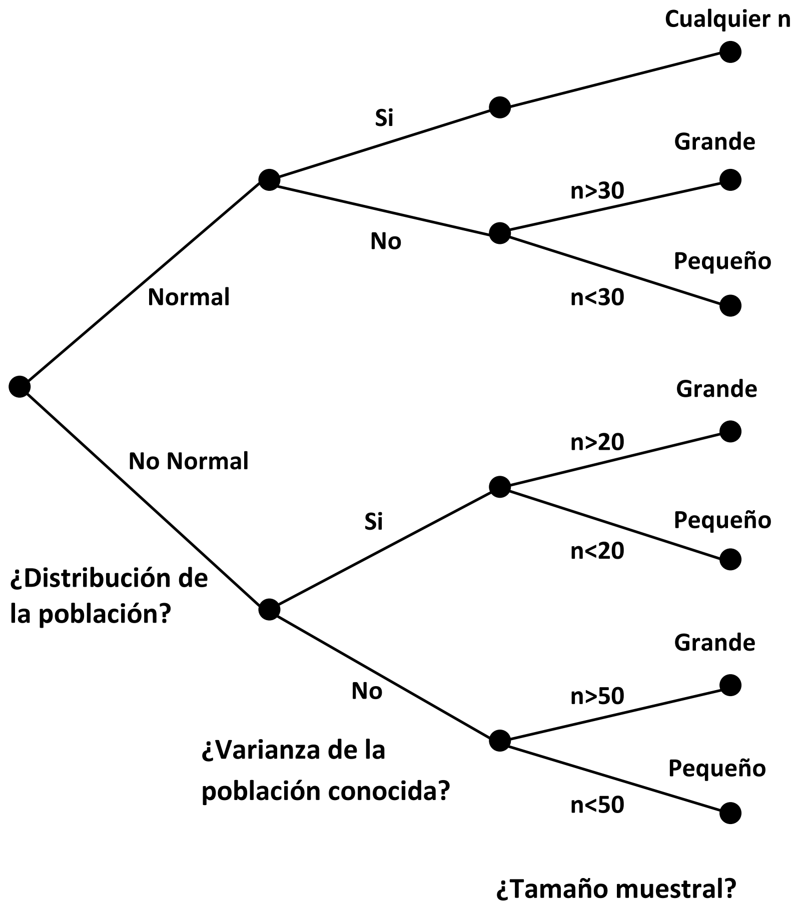


Figura 3 - Árbol de decisión para escoger el intervalo de confianza adecuado.



Use el siguiente intervalo de confianza:

$$P\left(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(\bar{x} - z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(\bar{x} - t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

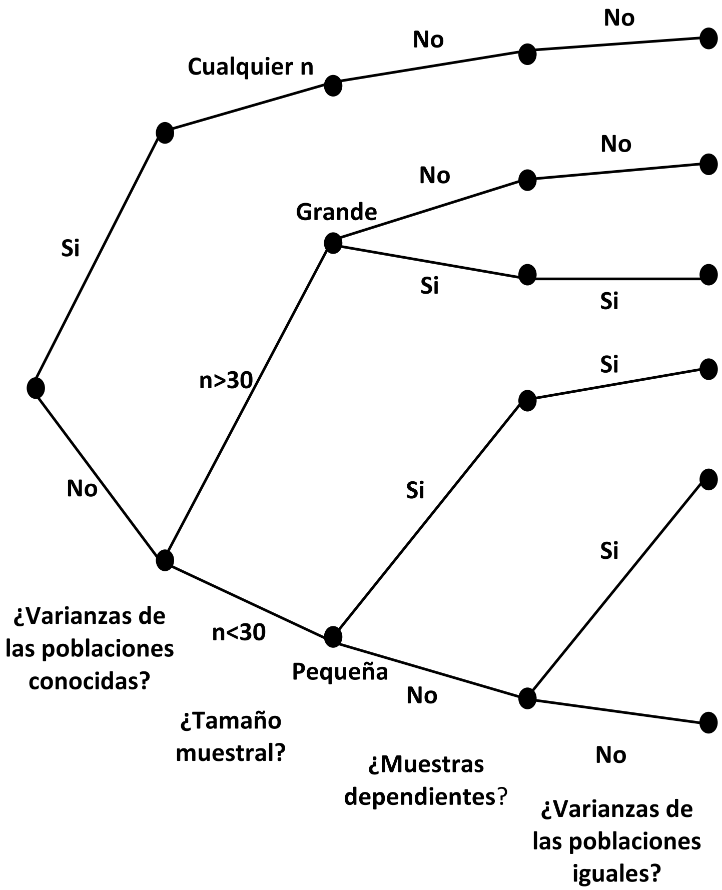
$$P\left(\bar{x} - k\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + k\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(\bar{x} - z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

No existe intervalo de confianza

El valor crítico k, es el valor determinado por el teorema de Tchebysheff.

Figura 4 - Árbol de decisión para escoger el intervalo de confianza adecuado para la media de la población.



Use el siguiente intervalo de confianza:

$$P\left((\bar{x}_1-\bar{x}_2)-z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\le \mu_1-\mu_2\le(\bar{x}_1-\bar{x}_2)+z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)=1-\alpha$$

$$P\left((\bar{x}_1-\bar{x}_2)-z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\le \mu_1-\mu_2\le(\bar{x}_1-\bar{x}_2)+z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right)=1-\alpha$$

$$P\left(\bar{x}_d-z_{\frac{\alpha}{2}}\frac{s_d}{\sqrt{n}}\le\mu_d\le\bar{x}_d+z_{\frac{\alpha}{2}}\frac{s_d}{\sqrt{n}}\right)=1-\alpha$$

$$P\left(\bar{x}_d-t_{\frac{\alpha}{2},n-1}\frac{s_d}{\sqrt{n}}\le\mu_d\le\bar{x}_d+t_{\frac{\alpha}{2},n-1}\frac{s_d}{\sqrt{n}}\right)=1-\alpha$$

$$P\left((\bar{x}_1-\bar{x}_2)-t_{\frac{\alpha}{2},\nu}\,s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\le \mu_1-\mu_2\le(\bar{x}_1-\bar{x}_2)+t_{\frac{\alpha}{2},\nu}\,s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)=1-\alpha$$

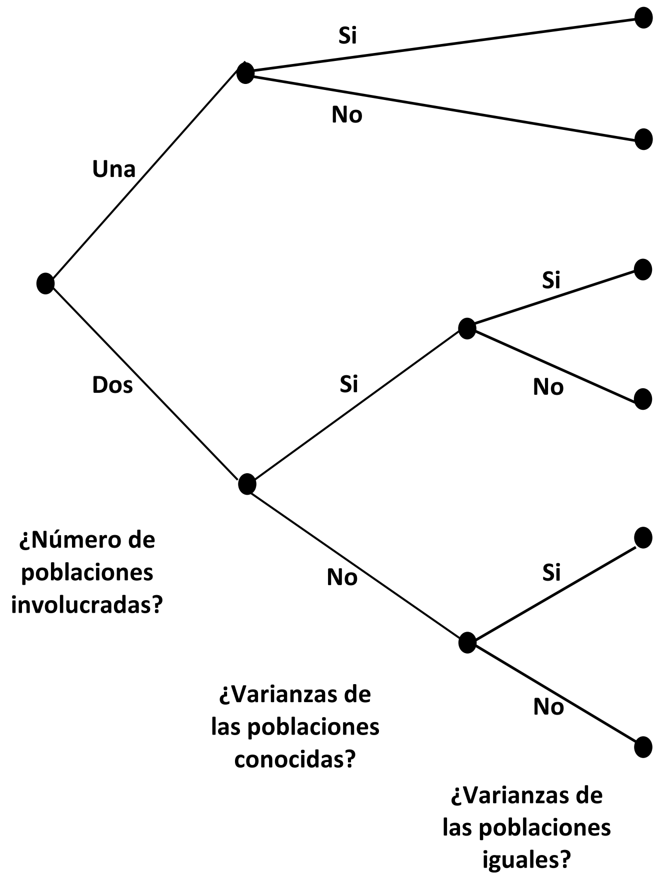
$$s_p=\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\qquad \nu=n_1+n_2-2$$

$$P\left((\bar{x}_1-\bar{x}_2)-t_{\frac{\alpha}{2},\nu}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\le \mu_1-\mu_2\le(\bar{x}_1-\bar{x}_2)+t_{\frac{\alpha}{2},\nu}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right)=1-\alpha$$

$$\nu=\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}-2$$

$\bar{x}_d$ y s_d son la media y la desviación estándar muestrales de la diferencia de datos pareados. μ_d es la diferencia de medias poblacionales de datos pareados.
s_p es la desviación estándar promedio de las muestras. ν representa los grados de libertad de la distribución *t* de Student.

Figura 5 - Árbol de decisión para escoger el intervalo de confianza adecuado para la diferencia de medias de poblaciones normales.



Use el siguiente intervalo de confianza:

$$P\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\ddot{P}(1-\ddot{P})}{n}} \leq P \leq \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\ddot{P}(1-\ddot{P})}{n}}\right) = 1-\alpha$$

$$P\left(\hat{p} - t_{\frac{\alpha}{2},n-1}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq P \leq \hat{p} + t_{\frac{\alpha}{2},n-1}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1-\alpha$$

$$P\left((\hat{p}_1-\hat{p}_2) - z_{\frac{\alpha}{2}}\sqrt{\ddot{P}(1-\ddot{P})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \leq P_1 - P_2 \leq \cdots\right) = 1-\alpha$$

$$P\left((\hat{p}_1-\hat{p}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{\ddot{P}_1(1-\ddot{P}_1)}{n_1}+\frac{\ddot{P}_2(1-\ddot{P}_2)}{n_2}} \leq P_1 - P_2 \leq \cdots\right) = 1-\alpha$$

$$P\left((\hat{p}_1-\hat{p}_2) - t_{\frac{\alpha}{2},\nu}\sqrt{P_p(1-P_p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \leq P_1 - P_2 \leq \cdots\right) = 1-\alpha$$

$$P_p = \frac{x_1+x_2}{n_1+n_2} \qquad \nu = n_1+n_2-2$$

$$P\left((\hat{p}_1-\hat{p}_2) - t_{\frac{\alpha}{2},\nu}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \leq P_1 - P_2 \leq \cdots\right) = 1-\alpha$$

$$\nu = \frac{\left[\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right]^2}{\frac{\left[\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}\right]^2}{n_1-1}+\frac{\left[\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right]^2}{n_2-1}} - 2$$

$\ddot{P}$ es una proporción patrón o estándar de la población.

Figura 6 - Árbol de decisión para escoger el intervalo de confianza adecuado para la proporción y diferencia de proporciones de poblaciones normales.

El Teorema de Tchebysheff manifiesta: "Para cualquier conjunto de datos y cualquier constante k mayor que uno, el porcentaje de los datos que debe caer dentro de k desviaciones estándar de cualquier lado de la media es de por lo menos $1-\frac{1}{k^2}$" (Freund y Simon [4]). Por lo tanto, uno puede estar seguro que, por ejemplo, el $1-\frac{1}{2^2}=75\%$ de los valores de una distribución debe caer dentro de más o menos dos desviaciones estándares. Si se quiere un intervalo de confianza del 95%, el valor de k debe ser de $k=\sqrt{\frac{1}{\alpha}}=\sqrt{\frac{1}{0{,}05}}=4.47$. Se nota que no se realiza ninguna suposición de distribución cuando se usa el teorema de Tchebysheff (Larson [21]).

Si una muestra proviene de una población que no es normal y la varianza poblacional es desconocida, el tamaño muestral para el uso de la distribución z es de 50. Este valor sale de la práctica empírica, ya que se puede demostrar que en una investigación, si se separa un estudio con tamaño muestral n en varios con tamaño 50, las conclusiones son prácticamente las mismas. Por otro lado, Giardina [22] observa que para $n>50$, se logra que el error estándar de la media o de la proporción sea una buena estimación del verdadero valor. Si $n<50$ no hay más salida que desistir de obtener un intervalo de confianza.

Cuanto mayor incertidumbre existe (es decir, la población no es normal, no se conoce la varianza poblacional y el tamaño muestral es pequeño), los valores críticos (determinados por las distribuciones *z*, *t* o el teorema de Tchebysheff, según sea el caso) serán más grandes y, por lo tanto, el intervalo tendrá un rango mayor y poseerá menor precisión. Este fenómeno puede observarse en la Tabla 2, donde se muestran los valores críticos usando las distintas distribuciones (*z*, *t* o el teorema de Tchebysheff), para diferentes probabilidades de confianza y diferentes tamaños muestrales.

TABLA 2 – VALORES CRÍTICOS (z, t, k) PARA DIFERENTES NIVELES DE CERTEZA

$1-\alpha$	Z	k	t (n = 20)	t (n = 30)	t (n = 50)	t (n = 100)
0,90	1,645	3,162	1,729	1,699	1,677	1,660
0,95	1,960	4,472	2,093	2,045	2,010	1,984
0,99	2,576	10,000	2,861	2,756	2,680	2,626

Fuente: Elaboración propia

En la Figura 7 se puede observar que a partir de un tamaño muestral de 15 o 20, los valores de la distribución *t* recién comienzan a diferir de la tendencia constante hasta un tamaño 50 y prácticamente se mantiene invariable hasta 100 o un valor mayor. Este fenómeno también ha sido observado por algunos autores (Batattacharyya y Johnson [17]), que sostienen que los valores de *t* no difieren significativamente de los valores de *z*, para tamaños muestrales iguales o mayores a 15.

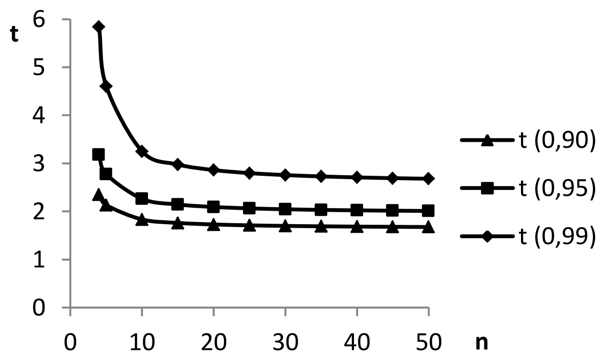


Figura 7 – Valores críticos de la distribución t con distintos niveles de confianza .

Fuente: Elaboración propia

Figura 5: Árbol de decisión para la estimación de diferencia de medias

La elección para la estimación de la diferencia de medias en poblaciones normales depende de varios aspectos:

- El conocimiento de las varianzas poblacionales
- El tamaño de las muestras
- La dependencia de las muestras

- La igualdad de varianzas poblacionales

Si se conocen las varianzas poblacionales el intervalo adecuado se distribuye mediante una normal estándar z, para cualquier tamaño muestral:

$$P\left((\bar{x}_1 - \bar{x}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) = 1 - \alpha$$

Se usa el mismo intervalo si las varianzas no son conocidas pero el tamaño muestral es $n \geq 30$.

Si las varianzas poblacionales son desconocidas, el tamaño es $n \geq 30$ y las muestras son dependientes o pareadas, el intervalo es:

$$P\left(\bar{x}_d - z_{\frac{\alpha}{2}}\frac{s_d}{\sqrt{n}} \leq \mu_d \leq \bar{x}_d + z_{\frac{\alpha}{2}}\frac{s_d}{\sqrt{n}} \right) = 1 - \alpha$$

donde $\bar{x}_d$ y s_d son los estadísticos muestrales de la diferencia de los datos pareados.

Cuando el tamaño es $n < 30$, se usa la distribución *t*, con $n - 1$ grados de libertad:

$$P\left(\bar{x}_d - t_{\frac{\alpha}{2},n-1}\frac{s_d}{\sqrt{n}} \leq \mu_d \leq \bar{x}_d + t_{\frac{\alpha}{2},n-1}\frac{s_d}{\sqrt{n}} \right) = 1 - \alpha$$

Si las varianzas poblacionales son desconocidas, el tamaño es $n < 30$, las muestras son independientes y las varianzas poblacionales son iguales, se usa la distribución t con $\nu = n_1 + n_2 - 2$ grados de libertad:

$$P\left((\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2},\nu}\, s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2},\nu}\, s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) = 1 - \alpha$$

donde s_p es la desviación estándar promedio de las dos muestras:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Si las varianzas de las poblaciones no son iguales, se tiene:

$$P\left((\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2},\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2},\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right) = 1 - \alpha$$

Hay un desacuerdo sobre cómo medir los grados de libertad para la distribución *t* en este intervalo. Autores, como Yamane [19], Programa FORD-ITESM [14], entre otros, realizan el cálculo con la fórmula siguiente:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 + 1}} - 2$$

Otros, como Juran y Gryna [15], entre otros, calculan mediante la siguiente fórmula:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} - 2$$

Ésta es la que usan también los paquetes estadísticos como el Statgraphics y el Análisis de Datos del Excel, entre otros, por lo que se adoptará en este artículo. La razón de la diferencia puede ser debida a la confianza optimista o conservadora para la estimación.

Figura 6: Árbol de decisión para la estimación de proporciones y diferencia de proporciones.

Para el intervalo de la proporción existen dos alternativas y, generalmente, son adecuadas para muestras grandes y proporciones cercanas a 0,5 o por lo menos mayores a 0,1 (Duncan [16]), de modo que se cumpla que $np \geq 5$ (Levin y Rubin [2]). El intervalo de la proporción se basa en una distribución binomial, que puede ser aproximada a una normal sólo bajo esas condiciones. Si la varianza de la población es conocida:

$$P\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\ddot{P}(1-\ddot{P})}{n}} \leq P \leq \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\ddot{P}(1-\ddot{P})}{n}}\right) = 1 - \alpha$$

donde $\ddot{P}$ es un patrón de la población bajo estudio. Si la varianza de la población es desconocida:

$$P\left(\hat{p} - t_{\frac{\alpha}{2},n-1}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq P \leq \hat{p} + t_{\frac{\alpha}{2},n-1}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha$$

que tiene una distribución t con $n - 1$ grados de libertad.

El intervalo de la diferencia de proporciones también supone muestras grandes para que se ajuste a una distribución normal. Muxica [22], es el único autor que plantea varias opciones para la estimación de la diferencia de proporciones, y su análisis es muy similar al de los intervalos para la diferencia de medias, es decir, se toma en cuenta si las varianzas poblaciones son conocidas y si son iguales.

Para varianzas conocidas e iguales:

$$P\left((\hat{p}_1 - \hat{p}_2) - z_{\frac{\alpha}{2}}\sqrt{\ddot{P}(1-\ddot{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \leq P_1 - P_2 \leq \cdots\right) = 1 - \alpha$$

Para varianzas conocidas y no necesariamente iguales:

$$P\left((\hat{p}_1 - \hat{p}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{\ddot{P}_1(1-\ddot{P}_1)}{n_1} + \frac{\ddot{P}_2(1-\ddot{P}_2)}{n_2}} \leq P_1 - P_2 \leq \cdots\right) = 1 - \alpha$$

Para varianzas desconocidas pero iguales:

$$P\left((\hat{p}_1 - \hat{p}_2) - t_{\frac{\alpha}{2},\nu}\sqrt{P_p(1-P_p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \leq P_1 - P_2 \leq \cdots\right) = 1 - \alpha$$

donde $P_p = \frac{x_1 + x_2}{n_1 + n_2}$, y la distribución t tiene $\nu = n_1 + n_2 - 2$ grados de libertad.

Para varianzas desconocidas y no necesariamente iguales:

$$P\left((\hat{p}_1-\hat{p}_2)-t_{\frac{\alpha}{2},\nu}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\leq P_1-P_2\leq\cdots\right)=1-\alpha$$

Los grados de libertad están dados por:

$$\nu=\frac{\left[\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right]^2}{\frac{\left[\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}\right]^2}{n_1-1}+\frac{\left[\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right]^2}{n_2-1}}-2$$

4. EJEMPLOS PARA EL USO DE LOS ÁRBOLES DE DECISIÓN

Se podrá usar el árbol de decisión de la Figura 3 (y si es el caso remitirse a los árboles de las Figuras 4, 5 y 6), cuando tenga que resolver cualquier problema para la estimación de un parámetro poblacional. Su uso es sencillo y se aplicará a dos casos para mayor claridad.

Caso 1:

Un nuevo dispositivo de filtrado se instala en una planta química. Muestras aleatorias arrojaron la siguiente información del porcentaje de impurezas, antes y después de la instalación:

Muestra/Estadísticos	Media	Varianza	Tamaño
Antes	12,5	101,17	8
Después	10,2	94,73	9

Se trata de establecer:

a) una forma para estimar si el porcentaje de impurezas presenta la misma variabilidad antes y después de la instalación del nuevo dispositivo de filtrado, utilizando $1-\alpha=0{,}95$,
b) si el dispositivo de filtrado ha reducido el porcentaje medio de impurezas de forma significativa,
c) si el porcentaje medio de impurezas permitido en la planta química es de 5%, y si se llegó a cumplir la meta.

Resolución:

a) De la Figura 3, se responde:

- ¿Número de poblaciones? Como existe una muestra antes y otra después, éstas se extrajeron de dos poblaciones.
- ¿Palabra clave? Comparación de varianzas.

Por lo tanto, se utiliza el intervalo para el cociente de varianzas:

$$P\left(\frac{s_1^2}{s_2^2}\frac{1}{F_{\frac{\alpha}{2},n_1-1,n_2-1}}\leq\frac{\sigma_1^2}{\sigma_2^2}\leq\frac{s_1^2}{s_2^2}\frac{1}{F_{1-\frac{\alpha}{2},n_1-1,n_2-1}}\right)=1-\alpha$$

Reemplazando la información muestral y puntos críticos de la distribución de Fisher:

$$P\left(\frac{101{,}17}{94{,}73}\frac{1}{4{,}53}\leq\frac{\sigma_1^2}{\sigma_2^2}\leq\frac{101{,}17}{94{,}73}\frac{1}{0{,}20}\right)=95\%$$

$$P\left(0{,}24\leq\frac{\sigma_1^2}{\sigma_2^2}\leq 5{,}23\right)=95\%$$

La variabilidad del porcentaje de impurezas antes y después de la instalación del nuevo dispositivo de filtrado es la misma, ya que el intervalo contiene el valor de 1.

b) De la Figura 3, se responde:

- ¿Número de poblaciones? Como existe una muestra antes y otra después, éstas se extrajeron de dos poblaciones.
- ¿Palabra clave? Comparación de medias.

Se considera la Figura 5:

- ¿Las varianzas de las poblaciones son conocidas? “No”.
- ¿Tamaños muestrales? “Menores a 30”.
- ¿Muestras dependientes? “No”.
- ¿Varianzas poblacionales iguales? “Si” (Resultado del inciso a)).

Por lo tanto, se usa el siguiente intervalo para la diferencia de medias:

$$P\left((\bar{x}_1-\bar{x}_2)-t_{\frac{\alpha}{2},v}\, s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\leq \mu_1-\mu_2\leq(\bar{x}_1-\bar{x}_2)+t_{\frac{\alpha}{2},v}\, s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)=1-\alpha \text{ con}$$

$$s_p=\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

Reemplazando la información muestral y puntos críticos de la distribución *t*:

$$s_p=\sqrt{\frac{(8-1)101.17+(9-1)94.73}{8+9-2}}=9{,}89$$

$$P\left((12{,}5-10{,}2)-(2{,}13)9{,}89\sqrt{\frac{1}{8}+\frac{1}{9}}\leq\mu_1-\mu_2\leq(12{,}5-10{,}2)+(2{,}13)\,9{,}89\sqrt{\frac{1}{8}+\frac{1}{9}}\right)=95\%$$

$$P(-7{,}94\leq\mu_1-\mu_2\leq 12{,}54)=95\%$$

El dispositivo de filtrado no ha sido efectivo, porque el porcentaje medio de impurezas antes y después es el mismo (el intervalo contiene el valor de 0).

c) De la Figura 3, se responde:

- ¿Número de poblaciones? Como solo se quiere ver si con el nuevo dispositivo se llegó a la meta: “Una”.
- ¿Palabra clave? “Media”.

Se considera la Figura 4:

- ¿Distribución de la población? Vamos a suponer que la muestra proviene de una población normal.
- ¿Varianza de la población conocida? “No”.
- ¿Tamaño muestral? “Pequeño”.

Por lo tanto el intervalo para la media es:

$$P\left(\bar{x}-t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}\leq\mu\leq\bar{x}+t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}\right)=1-\alpha$$

Reemplazando la información muestral y puntos críticos de la distribución *t*:

$$P\left(10{,}2-2{,}31\frac{9{,}73}{\sqrt{9}}\leq\mu\leq 10{,}2+2{,}31\frac{9{,}73}{\sqrt{9}}\right)=95\%$$

$$P(2{,}71\leq\mu\leq 17{,}69)=95\%$$

Se llegó a cumplir la meta del porcentaje de impurezas permitido, ya que el intervalo contiene el valor de 5%.

Nota: Los valores críticos F, t o z para los distintos intervalos de confianza con una probabilidad de certeza, se determinan mediante las tablas estadísticas respectivas que se pueden encontrar en cualquier libro de Estadística.

Caso 2:

Una máquina produce según un estándar de 5% de productos defectuosos. Se quiere ver la eficiencia de un nuevo operario, y para ello se recopiló una muestra aleatoria de 50 días en la que el nuevo operario produjo 10 productos al día. Se analizaron los productos y se comprobó que 30 estaban defectuosos.

a) Se trata de establecer si el nuevo operario se ajusta al patrón de eficiencia de la máquina. Se quiere tener una certeza del 95% en la estimación.

De la Figura 3:

- ¿Número de poblaciones? "Una"
- ¿Qué se quiere estimar? "Una proporción"

De la Figura 6:

- ¿Número de poblaciones? "Una"
- ¿Varianza de la población conocida? "Si"

El intervalo adecuado es:

$$P\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\ddot{P}(1-\ddot{P})}{n}} \leq P \leq \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\ddot{P}(1-\ddot{P})}{n}}\right) = 1-\alpha$$

$$P\left(0{,}06 - 1{,}96\sqrt{\frac{0{,}05(1-0{,}05)}{500}} \leq P \leq 0{,}06 + 1{,}96\sqrt{\frac{0{,}05(1-0{,}05)}{500}}\right) = 95\%$$

$$P(0{,}041 \leq P \leq 0{,}079) = 95\%$$

El operario se ajusta al patrón de eficiencia de la máquina, que es de 0,05.

b) Ahora, se supone que existen dos máquinas, la primera con una eficiencia de 5% de productos defectuosos producidos y la segunda con el 10%. Se quiere comparar si el nuevo trabajador cumple con la diferencia de proporciones de eficiencia de las máquinas, que es del 5%. Los datos muestrales son los siguientes:

	Máquina 1	Máquina 2
Tamaño muestral	500	600
Defectuosos	30	72
Proporción muestral	0,06	0,12

De la Figura 3:

- ¿Número de poblaciones? "Dos"
- ¿Qué se quiere estimar? "Comparación de proporciones"

De la Figura 6:

- ¿Número de poblaciones? "Dos"
- ¿Varianzas de las poblaciones conocidas? "Si"
- ¿Varianzas de las poblaciones iguales? "No"

El intervalo adecuado es:

$$P\left((\hat{p}_1-\hat{p}_2)-z_{\frac{\alpha}{2}}\sqrt{\frac{\ddot{P}_1(1-\ddot{P}_1)}{n_1}+\frac{\ddot{P}_2(1-\ddot{P}_2)}{n_2}}\leq P_1-P_2\leq\cdots\right)=1-\alpha$$

$$P\left((0{,}06-0{,}12)-1{,}96\sqrt{\frac{0{,}05(1-0{,}05)}{500}+\frac{0{,}10(1-0{,}10)}{600}}\leq P_1-P_2\leq\cdots\right)=1-\alpha$$

$$P(-0{,}091\leq P_1-P_2\leq -0{,}029)=1-\alpha$$

$$P(2{,}9\%\leq P_2-P_1\leq 9{,}1\%)=1-\alpha$$

El nuevo operario cumple con el criterio de la diferencia de eficiencias de las máquinas, que es del 5%.

c) También se quiere comparar la eficiencia de producción del operario nuevo con el antiguo en una máquina 3, de la cual no se sabe su eficiencia. Primero, se quiere probar si las varianzas de los dos operarios son iguales o no. Los datos muestrales son:

	Operario 1	Operario 2
Tamaño muestral	200	200
Defectuosos	30	25
Proporción muestral	0,15	0,125

De la Figura 3:

- ¿Número de poblaciones? “Dos”
- ¿Qué se quiere estimar? “Relación de variabilidades”

El intervalo adecuado es:

$$P\left(\frac{s_1^2}{s_2^2}\frac{1}{F_{\frac{\alpha}{2},n_1-1,n_2-1}}\leq\frac{\sigma_1^2}{\sigma_2^2}\leq\frac{s_1^2}{s_2^2}\frac{1}{F_{1-\frac{\alpha}{2},n_1-1,n_2-1}}\right)=1-\alpha$$

Expresado en función de las varianzas para proporciones:

$$P\left(\frac{\hat{p}_1(1-\hat{p}_1)}{\hat{p}_2(1-\hat{p}_2)}\frac{1}{F_{\frac{\alpha}{2},n_1-1,n_2-1}}\leq\frac{\sigma_1^2}{\sigma_2^2}\leq\frac{\hat{p}_1(1-\hat{p}_1)}{\hat{p}_2(1-\hat{p}_2)}\frac{1}{F_{1-\frac{\alpha}{2},n_1-1,n_2-1}}\right)=1-\alpha$$

$$P\left(\frac{0{,}15(1-0{,}15)}{0{,}125(1-0{,}125)}\frac{1}{1{,}32}\leq\frac{\sigma_1^2}{\sigma_2^2}\leq\frac{0{,}15(1-0{,}15)}{0{,}125(1-0{,}125)}\frac{1}{0{,}76}\right)=1-\alpha$$

$$P\left(0{,}88\leq\frac{\sigma_1^2}{\sigma_2^2}\leq 1{,}54\right)=1-\alpha$$

Las varianzas de eficiencia de los dos operarios son iguales.

De la Figura 3:

- ¿Número de poblaciones? “Dos”
- ¿Qué se quiere estimar? “Comparación de proporciones”

De la Figura 6:

- ¿Número de poblaciones? “Dos”
- ¿Varianzas de las poblaciones conocidas? “No”
- ¿Varianzas de las poblaciones iguales? “Si” (Probado anteriormente)

El intervalo adecuado es:

$$P\left((\hat{p}_1-\hat{p}_2)-t_{\frac{\alpha}{2},\nu}\sqrt{P_p(1-P_p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\leq P_1-P_2\leq\cdots\right)=1-\alpha$$

$$P_p=\frac{x_1+x_2}{n_1+n_2}\qquad\qquad P_p=\frac{30+25}{200+200}=0{,}1375$$

$$P\left((0{,}15-0{,}125)-1{,}96\sqrt{0{,}1375(1-0{,}1375)\left(\frac{1}{200}+\frac{1}{200}\right)}\leq P_1-P_2\leq\cdots\right)=1-\alpha$$

$$P(-0{,}042\leq P_1-P_2\leq 0{,}092)=1-\alpha$$

Los operarios tienen eficiencias iguales usando la misma máquina, ya que la diferencia contiene el valor de 0.

5. CONCLUSIONES

Se han configurado árboles de decisión para elegir el intervalo de confianza adecuado en la resolución de problemas de estimación estadística, a partir de bases teóricas y empíricas generalmente aceptadas, que cumplen el objetivo de disminuir la complejidad de la elección. Éstos permiten usar este método inferencial con mayor facilidad y brindar la solución de elección de forma más rápida y exhaustiva que los métodos de flujogramas o tablas resumen, para los fines o los propósitos establecidos.

Por otro lado, esta ayuda se presta para su uso en la informática y para la configuración de un *software* estadístico más funcional y comprensible que los actuales.

La ayuda de los árboles de decisiones para la estimación estadística ha sido adoptada actualmente por muchos investigadores sin formación sólida en Estadística con excelentes resultados.

Por otro lado, los árboles de decisión pueden ser usados de manera similar a la de los intervalos de confianza en otros temas de la estadística inferencial como ser: pruebas de hipótesis, tanto paramétricas como no paramétricas, modelos probabilísticos, modelos de regresión; o en temas de la estadística descriptiva, control estadístico de calidad o el diseño y análisis de experimentos (los autores han desarrollado la ayuda en todas estas áreas).

Se espera que este aporte, logre que cada vez más investigadores y profesionales utilicen el análisis estadístico, como por ejemplo la estimación, para el tratamiento alternativo de la información.

5. REFERENCIAS

[1] Berenson et al. *Estadística para Administración*, Pearson Educación, Segunda Edición, México, 2001.
[2] Levin and Rubin, *Estadística para Administradores*, Prentice Hall Hispanoamericana S.A., Sexta Edición, México, 1996.
[3] Mason and Lind, *Estadística para Administración y Economía*, Alfaomega, Séptima Edición, México, 1995.
[4] Freund and Simon, *Estadística Elemental*, Prentice Hall, Octava Edición, México, 1994.
[5] Miller et al. *Probabilidad y Estadística para Ingenieros*, Prentice Hall Hispanoamericana S.A., Cuarta Edición, México, 1992.
[6] W. Mendenhall. *Estadística para Administradores*, Grupo Editorial Iberoamérica, Segunda Edición, México, 1990.
[7] M. García. *Socioestadística*, Alianza Editorial, Madrid-España, 1985.
[8] Mood/Graybill, *Introducción a la Teoría Estadística*, Aguilar, Cuarta Edición, Madrid-España, 1976.
[9] L. Maisel. *Probabilidad y Estadística*, Fondo Educativo Interamericano, Colombia, 1973.
[10] Hays and Winkler, *Statistics: Probability, Inference and Decision*, Holt, Rinehart and Winston Inc., United States of America, 1971.
[11] Lobez y Casa, *Estadística Intermedia*, Editorial Vicens-Vives, Primera Edición, España, 1967.
[12] D.C. Montgomery. *Control Estadístico de la Calidad*, Grupo Editorial Iberoamerica, S.A., México, 1994.

[13] Gutierrez y De La Vara, *Análisis y Diseño de Experimentos*, McGraw-Hill Interamericana, Primera Edición, México, 2004.
[14] Programa FORD-ITESM, *Inferencia Estadística, Módulo 7*, México, (1989).
[15] Juran y Gryna, *Manual de Control de Calidad*, Volumen II, McGraw-Hill, Cuarta Edición, España, 1993.
[16] Duncan A., *Control de Calidad y Estadística Industrial*, Editorial Alfaomega, México, (1989).
[17] Batattacharyya and Johnson, *Statistical, Concepts and Methods*, John Wiley & Sons, United States of America, 1977.
[18] Merril and Fox, *Introducción a la Estadística Económica*, Amorrortu Editores, Argentina, 1969.
[19] T. Yamane. *Estadística*, Editorial Harla, México, 1974.
[20] A. Novales. *Estadística y Econometría*, MsGraw-Hill Interamericana, España, 1997.
[21] H. Larson. *Introduction to Probability Theory and Statistical Inference*, Second Edition, Wiley International Edition, United States of America, 1974.
[22] B. Giardina. *Manual de Estadística*, Compañía Editorial Continental, México, 1968.
[23] L. Muxica. *Introducción a la Estadística Matemática*, Universidad de Concepción, Publicaciones Docentes, Chile, 1966.
[24] P. Hoel. *Introducción a la Estadística Matemática*, Biblioteca Interamericana de Estadística Teórica y Aplicada, Argentina, 1955.