Estimación de parámetros

Fundamentos de Estadística Inferencial

Mg. Leonel Heredia Altamirano

Diplomado en Econometría Aplicada
Módulo
Fundamentos de Estadística

9 de marzo de 2025

Tabla de contenidos

Estimación de parámetros

Al realizar una investigación estadística a menudo se sabe o se supone que la población definida por una variable aleatoria X (discreta o continua) de la cual se selecciona una muestra aleatoria tiene una forma funcional específica $f(x)$ cuyos(s) parámetro(s) se intenta determinar.

Si el parámetro desconocido que queremos determinar es denotado por θ, entonces, la distribución de la población de X será denotado por $f(x, \theta)$.

Los métodos de inferencia estadística, básicamente consisten en seleccionar una muestra aleatoria de la población en estudio, y con la información que se obtenga llegar a:

- Estimar el valor o los posibles valores del parámetro desconocido.
- Tomar la decisión de aceptar o rechazar una afirmación hecha sobre el valor o los posibles valores del parámetro desconocido.

El primero de estos dos procedimientos se denomina ***estimación del parámetro***. El segundo se conoce como ***prueba de hipótesis del parámetro***.

Estimación puntual del parámetro

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño n seleccionada de una población definida X de parámetro θ cuya distribución la denotamos por $f(x, \theta)$. Se denomina **estimador puntual** del parámetro θ a cualquier estadística (función real de la muestra) que la denotaremos por $\hat{\Theta} = H(X)$ cuyo valor $\hat{\theta} = H(x)$ es la estimación puntual del parámetro.

Un estimador puntual del parámetro θ es pues, una variable aleatoria $\hat{\Theta}$ mientras que una estimación puntual es el valor numérico $\hat{\theta}$ del estimador.

Por ejemplo, un **estimador puntual** de la media poblacional μ (esto es $\theta = \mu$), es la estadística media muestral (Variable aleatoria) $\overline{X}$ (esto es $\hat{\Theta} = \overline{X}$) cuyo valor numérico $\hat{x}$ es la estimación puntual del parámetro.

No toda función de la muestra es un **buen estimador** del parámetro. Un buen estimador es aquel que está más cerca del parámetro que se estima. Para que un estimador puntual sea un buen estimador debe cumplir con ciertas propiedades, dos de las cuales son la *insesgabilidad* y la *eficiencia*.

1 Estimaciones Puntuales

- Sólo un valor numérico sirva para estimar el parámetro, es decir, asigna directamente al parámetro el valor obtenido para el estadístico

$$E(X) = \mu$$

- Constituye la inferencia más simple que se puede realizar: asignar al parámetro el valor del estadístico que mejor sirva para estimarlo.

Condiciones para ser un buen estimador

a) **Carencia de Sesgo:** Un estimador será insesgado si su valor esperado coincide con el del parámetro a estimar

b) **Consistencia:** Un estimador será consistente si, conforme aumenta el tamaño muestral, su valor se va aproximando al del parámetro

c) **Eficiencia:** Dados dos posibles estimadores, diremos que el primero es un estimador más eficiente que el segundo si se cumple que el primer estimador tiene una varianza menor que el segundo.

d) **Suficiencia:** Un estimador será suficiente si utiliza toda la información muestral disponible

Método de Máxima Verosimilitud

Sea θ el parámetro a estimar, de una población X distribuida como $f(x, \theta)$ y sea $X_1, X_2, \ldots, X_n$ una muestra al azar de la población X. El procedimiento para obtener el estimador de máxima verosimilitud del parámetro θ es:

1. Determinar la distribución conjunta de la muestra en sus valores observados respectivos $x_1, x_2, \ldots, x_n$. Esta función del parámetro θ conocida también como **función de verosimilitud** está definida por:

$$L(\theta) = f(x_1, x_2, \ldots, x_n, \theta) = f(x_1, \theta) \times \ldots f(x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

2. Obtener el valor de $\hat{\theta}$ del parámetro que maximiza a la función $L(\theta)$. Este valor, que a su vez depende de los valores de la muestra, es la **estimación de máxima verosimilitud** (EVM) del parámetro θ y su función correspondiente es el **estimador de máxima verosimilitud del parámetro**.

3. Para maximizar $L(\theta)$, a menudo se usa la transformación $L = ln(L(\theta))$. En este caso el valor de $\hat{\theta}$ que maximiza a L y por lo tanto a $L(\theta)$, es la solución de la ecuación:
$$\frac{dL}{d\theta} = 0$$

4. Si la distribución de probabilidades de la población X contiene k parámetros $\theta_1, \theta_2, \ldots, \theta_k$, entonces, la función de verosimilitud se define por:
$$L(\theta_1, \theta_2, \ldots,\ \theta_k) = \prod_{i=1}^{n} f(x_i, \theta_1, \theta_2, \ldots, \theta_k)$$
La estimación de máxima verosimilitud de cada parámetro θ_i es la solución $\hat{\theta}_i$ de la ecuación respectiva:
$$\frac{\partial L}{\partial \theta_1} = 0, \qquad \frac{\partial L}{\partial \theta_2} = 0, \qquad \ldots, \qquad \frac{\partial L}{\partial \theta_k} = 0$$

EJEMPLO 01

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño n, seleccionada de una población X con distribución Bernoulli o binomial (1,p). Obtenga el estimador del parámetro p aplicando el método de máxima verosimilitud.

La distribución de probabilidad de cada variable aleatoria X_i está dada por:

$$f\left(x_i, p\right) = p^{x_i}\left(1-p\right)^{1-x_1}, \qquad \text{donde } \ x_i = 0, 1$$

Entonces, la función de verosimilitud de la muestra aleatoria es:

$$L\left(p\right) = \prod_{i=1}^{n} f\left(x_i, p\right) = \left(p\right)^{\sum X_i}\left(1-p\right)^{n-\sum X_i}$$

Aplicando logaritmos naturales, se tiene:

$$L = ln\left(L\left(p\right)\right) = \sum_{i=1}^{n} x_i \times lnp + \left(n - \sum_{i=1}^{n} x_i\right) \times ln\left(1-p\right)$$

Derivando la función L con respecto a p e igualando a cero se tiene:

$$\frac{dL}{dp} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} = 0, \qquad \text{con } \ p \neq 0 \ \text{ y } \ p \neq 1$$

Resolviendo la ecuación resulta que:

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{p}$$

Por lo tanto, la proporción muestral $\overline{p}$ es la estimación de máxima verosimilitud de la proporción p de éxitos de la población Bernoulli.

EJEMPLO 02

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria escogida de una población X con distribución normal $N\left(\mu, \sigma^2\right)$. Obtener el estimador de la media μ y de la varianza σ^2 por el método de máxima verosimilitud.

La distribución de probabilidad de la población normal de parámetros μ y σ^2 asociada a cada variable aleatoria X_i de la muestra está dada por:

$$f\left(x, \mu, \sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Entonces, la función de verosimilitud de la muestra es:

$$L\left(\mu,\sigma^2\right)=\prod_{i=1}^{n} f\left(x,\mu,\sigma^2\right)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\times\prod_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$L\left(\mu,\sigma^2\right)=\left(2\pi\sigma^2\right)^{-\frac{n}{2}}\times e^{-\frac{1}{2\sigma^2}\times\sum_{i=1}^{n}(x_i-\mu)^2}$$

Aplicando logaritmos naturales se tiene:

$$L=ln\left(L\left(\mu,\sigma^2\right)\right)=-\frac{n}{2}\times ln\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\times\sum_{i=1}^{n}(x_i-\mu)^2$$

Derivando la función L con respecto a μ e igualando a cero se obtiene:

$$\frac{dL}{d\mu}=-0+\frac{1}{\sigma^2}\times\sum_{i=1}^{n}(x_i-\mu)=0$$

De donde resulta:

$$\hat{\mu}=\frac{\sum_{i=1}^{n}x_i}{n}=\overline{x}$$

Por lo tanto, la media muestral $\overline{x}$ es la estimación de máxima verosimilitud de la media μ de la población normal.
Por otro lado, derivando la función L con respecto a σ^2 e igualando a cero se obtiene:

$$\frac{dL}{d\sigma^2} = -\frac{1}{2} \times \frac{n}{\sigma^2} + \frac{1}{2\sigma^4} \times \sum_{i=1}^{n} (x_i - \mu) = 0$$

De donde resulta:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$$

Por lo tanto, la varianza muestral s_n^2 es la estimación de máxima verosimilitud del parámetro σ^2 de la distribución normal.

Estimación de parámetros por intervalo

La estimación de parámetros por intervalo es un método estadístico que consiste en obtener un intervalo de extremos cerrados $[a, b]$, dentro del cual se ubican los posibles valores de un parámetro con una medida de fiabilidad denominado *nivel de confianza*. Los números a y b se obtienen de la distribución muestral de la estadística aplicando el nivel de confianza dado y los resultados de la muestra.

La estadística que se aplica para obtener el intervalo se denomina **estadística del pivote**. Una estadística es del pivote si está es una función de los valores de la muestra y del parámetro único θ. El método de obtención del intervalo se denomina *método del pivote*.

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño n escogida de una población descrita por una variable X con distribución $f(x, \theta)$, cuyos valores experimentales (o datos) se van a aplicar para estimar el parámetro θ.

Sea además la variable aleatoria $\hat{\Theta} = H(X_1, \ldots, X_n)$ una estadística que es estimador de punto del parámetro θ y cuya distribución de probabilidad sea una distribución conocida.

Dada la probabilidad $1-\alpha$, entonces, en la distribución de $\hat{\Theta}$ se pueden determinar dos valores $A = H_1(X)$ y $B = H_2(X)$ tales que:

$$P\left[A \leq \theta \leq B\right] = 1 - \alpha$$

Por lo tanto, el intervalo $[A, B]$ es el **intervalo estimador** del parámetro θ con probabilidad $1-\alpha$, o que $\theta \in [A, B]$ con probabilidad $1-\alpha$.

Además, si $a = H_1(x_1, x_2, \ldots, x_n)$ y $b = H_2(x_1, x_2, \ldots, x_n)$ son los valores numéricos de A y B respectivamente, entonces se dice que el **intervalo numérico** $[a, b]$ es el **intervalo de confianza** del $(1-\alpha) \times 100\,\%$.

Si se tiene $[a, +\infty[$ se dice que es un intervalo de estimación unilateral del parámetro θ. En caso de tener $]-\infty, b_1]$ es un intervalo de estimación unilateral del parámetro θ.

IC para μ de una población con σ^2 conocida

Si la población X es normal (PRN) o no normal (TLC) se define:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

La estadística Z es la **estadística del pivote** que se aplica para determinar el intervalo de confianza de μ para un nivel de confianza de $1 - \alpha$:

$$P\left[-z_{1-\frac{\alpha}{2}} \leq Z \leq z_{1-\frac{\alpha}{2}}\right] = 1 - \alpha$$

$$P\left[-z_{1-\frac{\alpha}{2}} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}\right] = 1 - \alpha$$

$$P\left[\overline{X} - z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

$$A = \overline{X} - z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \qquad B = \overline{X} + z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

Por definición, el **error de estimación**, al estimar μ por $\overline{x}$ es el valor numérico: $e = |\overline{x} - \mu|$. El valor mínimo del error de estimación es igual a cero, cuando $\overline{x}$ estima exactamente a μ. El valor máximo es igual a $z_0 \times ET$:

$$e = |\mu - \overline{x}| \leq z_0 \times ET$$

Como consecuencia, se tiene la siguiente lectura técnica del intervalo de confianza de la media μ de la población: Si $\overline{x}$ estima a μ, entonces, se tiene una confianza del $(1-\alpha) \times 100\,\%$ de que el error de esta estimación no será superior al valor numérico $e = z_0 \times ET$

Si los elementos de la muestra de tamaño n son escogidos al azar uno por uno sin reposición de una población finita de forma desconocida de tamaño N y si $n \geq 30$ se define la siguiente variable aleatoria:

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}} \underset{TLC}{\sim} N(0,1)$$

Por lo tanto, el intervalo de estimación de μ con nivel de confianza del $(1-\alpha) \times 100\,\%$ tiene los siguientes límites:

$$\overline{x} \pm z_0 \times ET$$

IC para μ de una población con σ^2 desconocida

Sean $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño n escogida de una población definida por la variable X cuya distribución de probabilidades es normal $N(\mu, \sigma^2)$ con ambos parámetros desconocidos. Sean la media y la varianza muestral respectivas:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}$$

La estadística del pivote que se aplica para el IC de μ es:

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

$$P\left[-t_0 \leq T \leq t_0\right] = 1 - \alpha$$

$$P\left[-t_0 \leq \frac{\overline{X} - \mu}{S/\sqrt{n}} \leq t_0\right] = 1 - \alpha$$

$$P\left[\overline{X} - t_0 \times \frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_0 \times \frac{S}{\sqrt{n}}\right] = 1 - \alpha$$

IC para la varianza poblacional

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño n escogida de una población X normal con varianza σ^2 (parámetro desconocido). Un estimador puntual de la varianza es la varianza muestral insesgada:

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}$$

La estadística del pivote que se aplica para el IC de σ^2:

$$\chi^2 = \frac{(n-1) \times S^2}{\sigma^2} \sim \chi^2 (n-1)$$

$$P\left[\chi^2_{\frac{\alpha}{2}, n-1} \leq \chi^2 \leq \chi^2_{1-\frac{\alpha}{2}, n-1}\right] = 1 - \alpha$$

$$P\left[\chi^2_{\frac{\alpha}{2}, n-1} \leq \frac{(n-1) \times S^2}{\sigma^2} \leq \chi^2_{1-\frac{\alpha}{2}, n-1}\right] = 1 - \alpha$$

$$P\left[\frac{(n-1) \times S^2}{\chi^2_{\frac{\alpha}{2}, n-1}} \leq \sigma^2 \leq \frac{(n-1) \times S^2}{\chi^2_{1-\frac{\alpha}{2}, n-1}}\right] = 1 - \alpha$$

Intervalo de confianza para la razón de varianzas

Sea S_1^2 y S_2^2 las varianzas de dos muestras aleatorias independientes de tamaños n_1 y n_2 seleccionadas de dos poblaciones normales con varianzas σ_1^2 y σ_2^2. El estadístico de pivote que se construye es:

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(r_1, r_2)$$

Donde $r_1 = n_1 - 1$ y $r_2 = n_2 - 1$. Dado el nivel de confianza $1 - \alpha$, se calcula el intervalo de confianza por:

$$P\left[f_{\frac{\alpha}{2},r_1,r_2} \leq F \leq f_{1-\frac{\alpha}{2},r_1,r_2}\right] = 1 - \alpha$$

$$P\left[f_{\frac{\alpha}{2},r_1,r_2} \leq \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \leq f_{1-\frac{\alpha}{2},r_1,r_2}\right] = 1 - \alpha$$

$$P\left[\frac{S_1^2}{S_2^2} \times f_{\frac{\alpha}{2},r_1,r_2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2}{S_2^2} \times f_{1-\frac{\alpha}{2},r_1,r_2}\right] = 1 - \alpha$$

IC para la diferencia de medias

IC de diferencia de medias para varianzas conocidas

Sean $\overline{X}_1$ y $\overline{X}_2$ las medias de dos muestras aleatorias independientes de tamaños n_1 y n_2 seleccionadas respectivamente de dos poblaciones con medias μ_1 y μ_2 desconocidas. Dado que se conoce σ_1^2 y σ_2^2, se tiene:

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Dado el nivel de confianza $1-\alpha$, la cantidad pivotal es:

$$P\left[-z_{1-\frac{\alpha}{2}} \leq Z \leq z_{1-\frac{\alpha}{2}}\right] = 1-\alpha$$

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

$$P\left[\left(\overline{X}_1 - \overline{X}_2\right) - z_0 \times ET \leq \mu_1 - \mu_2 \leq \left(\overline{X}_1 - \overline{X}_2\right) + z_0 \times ET\right] = 1-\alpha$$

IC de diferencia de medias para varianzas desconocidas $(\sigma_1^2 = \sigma_2^2 = \sigma^2)$

Se define la variable aleatoria independiente:

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_c^2}{n_1} + \frac{S_c^2}{n_2}}} \sim t\left(n_1 + n_2 - 2\right)$$

Donde la varianza común o promedio es:

$$S_c^2 = \frac{(n_1 - 1) \times S_1^2 + (n_2 - 1) \times S_2^2}{n_1 + n_2 - 2}$$

Dado el nivel de confianza $1 - \alpha$, la cantidad pivotal es:

$$P\left[-t_0 \leq T \leq t_0\right] = 1 - \alpha$$

$$P\left[-t_0 \leq \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_c^2}{n_1} + \frac{S_c^2}{n_2}}} \leq t_0\right] = 1 - \alpha$$

$$P\left[\left(\overline{X}_1 - \overline{X}_2\right) - t_0 \times ET \leq \mu_1 - \mu_2 \leq \left(\overline{X}_1 - \overline{X}_2\right) + t_0 \times ET\right] = 1 - \alpha$$

IC de diferencia de medias para varianzas desconocidas $\sigma_1^2 \neq \sigma_2^2$

Sean $\overline{X}_1$ y $\overline{X}_2$ las medias de dos muestras aleatorias independientes de tamaños n_1 y n_2, con medias μ_1 y μ_2 y varianzas σ_1^2 y σ_2^2 respectivamente:

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t\,(r)$$

Donde los r grados de libertad están dados por:

$$r = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{\left[\frac{s_1^2}{n_1}\right]^2}{n_1 - 1} + \frac{\left[\frac{s_2^2}{n_2}\right]^2}{n_2 - 1}}$$

El intervalo de confianza para una probabilidad $1 - \alpha$ es:

$$P\left[-t_0 \leq T \leq t_0\right] = 1 - \alpha$$

$$P\left[\left(\overline{X}_1 - \overline{X}_2\right) - t_0 \times ET \leq \mu_1 - \mu_2 \leq \left(\overline{X}_1 - \overline{X}_2\right) + t_0 \times ET\right] = 1 - \alpha$$

IC de diferencia de medias para datos pareados

Sea (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) una muestra aleatoria de n datos en parejas escogidas de una población (X, Y). Se supone que $X \sim N(\mu_1, \sigma_1^2)$ e $Y \sim N(\mu_2, \sigma_2^2)$. El estimador insesgado de la diferencia es:

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n} = \frac{\sum_{i=1}^{n} (X_i - Y_i)}{n} = \overline{X} - \overline{Y}$$

La estadística de pivote para la diferencia de medias es:

$$Z = \frac{\overline{D} - (\mu_1 - \mu_2)}{\sigma_D / \sqrt{n}} \sim N(0, 1)$$

Donde el intervalo de confianza para la diferencia de medias es:

$$P\left[\overline{d} - z_0 \times \frac{\sigma_D}{\sqrt{n}} \leq \mu_1 - \mu_2 \leq \overline{d} + z_0 \times \frac{\sigma_D}{\sqrt{n}}\right] = 1 - \alpha$$

Si la varianza σ_D^2 es desconocida, el intervalo de confianza es:

$$P\left[\overline{d} - t_0 \times \frac{\sigma_D}{\sqrt{n}} \leq \mu_1 - \mu_2 \leq \overline{d} + t_0 \times \frac{\sigma_D}{\sqrt{n}}\right] = 1 - \alpha$$

Intervalo de confianza para una proporción

Sea $\overline{P} = \frac{X}{n}$ la variable aleatoria que denota la proporción de éxito de muestras de tamaño n, $X_1, X_2, \ldots, X_n$ extraídas de la población Bernoulli $B\left(1, p\right)$. Entonces, si n es suficientemente grande:

$$Z = \frac{\overline{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \underset{TLC}{\sim} N\left(0, 1\right)$$

Dada la probabilidad $1 - \alpha$ en la distribución Z se tiene el intervalo:

$$P\left[-z_0 \leq Z \leq z_0\right] = 1 - \alpha$$

$$P\left[-z_0 \leq \frac{\overline{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_0\right] = 1 - \alpha$$

$$P\left[\overline{P} - z_0 \times \sqrt{\frac{p\left(1-p\right)}{n}} \leq p \leq \overline{P} + z_0 \times \sqrt{\frac{p\left(1-p\right)}{n}}\right] = 1 - \alpha$$

IC para diferencia de proporciones

Sean $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_n$ dos muestras aleatorias independientes de tamaños n_1 y n_2 seleccionadas respectivamente de dos poblaciones independientes de Bernoulli $B(1, p_1)$ y $B(1, p_2)$ donde p_1 y p_2 son las proporciones de éxito en las poblaciones respectivas. Sean las proporciones muestrales

$$\overline{P}_1 = \frac{\sum_{i=1}^{n_1} X_i}{n_1} = \frac{X}{n_1} \qquad \overline{P}_2 = \frac{\sum_{i=1}^{n_2} Y_i}{n_2} = \frac{Y}{n_2}$$

Para n_1 y n_2 suficientemente grande se obtiene la siguiente aproximación:

$$Z = \frac{\left(\overline{P}_1 - \overline{P}_2\right) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1)$$

Dada la probabilidad $1-\alpha$ en la distribución Z se tiene:

$$P\left[-z_0 \leq Z \leq z_0\right] = 1-\alpha$$

$$P\left[\left(\overline{P}_1 - \overline{P}_2\right) - z_0 \times ET \leq Z \leq \left(\overline{P}_1 - \overline{P}_2\right) + z_0 \times ET\right] = 1-\alpha$$