TEMA 11 ESTIMACIÓN DE PARÁMETROS

El tema que veremos a continuación nos habla de la segunda gran etapa de la estadística lo que se conoce como estadística inferencial . En este esta segunda etapa de la estadística comenzaremos a conocer lo que realmente importa al investigador que no es otra cosa que la población. Ya que nuestros estudios estadísticos se crean para conocer un aspecto concreto de la población . Así una vez que ya conocemos todas las características de la muestra, el conjunto de sujetos con los que trabajamos, querremos ir más lejos y extrapolar los resultados , para tratar de descubrir leyes generales aplicables de forma universal. Este hecho de extrapolar se le puede denominar también inferencia , por lo tanto esta etapa de la estadística se denomina estadística inferencial.

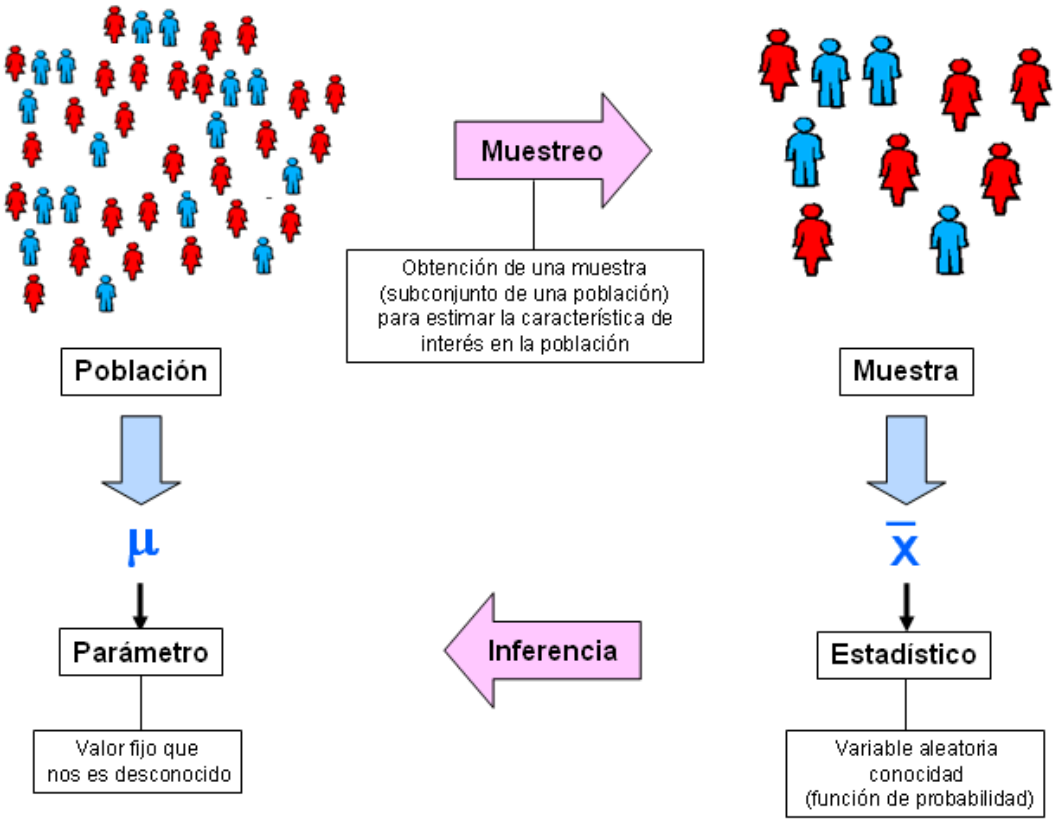
En este tema se abordarán muchos conceptos que ya se han mencionado a lo largo del curso, como son el error típico , el sesgo , el error humano y la homocedasticidad.

El capítulo comienza hablando de la muestra y la población, o sea que recapitulemos. La población es toda la gente con la que queremos trabajar porque tienen unas características o propiedades concretas. Pero claro está no podemos trabajar con toda la población porque suele ser mucha gente y los recursos y el tiempo no nos permitirán trabajar con todo el mundo. Como no podemos trabajar con todo el mundo tenemos que escoger la muestra , un subconjunto de la población, y que por su tamaño si es factible trabajar con ella, sin que sean demasiado pequeñas y por lo tanto se parezcan a la población de la cual se han extraído. Así realizaremos toda una serie de cálculos sobre esta muestra, los cálculos que se llamarán estadísticos y que se han estudiado en el tema 5.

Ahora tenemos que entender las ideas básicas , yo extraigo una muestra de la población , porque supongo que mi muestra se parecerá a la población , es como una madre y una hija, las hijas vienen de las madres, y se supone que se parecen, pero no se tienen porque parecer en todo y además puede que tengamos mala suerte y justo cojamos una madre y una hija que no se parecen.

Las muestras se deben parecer a las poblaciones para que pueda extrapolar la información que extraigo de ellas y así conocer las poblaciones , que a fin de cuentas es nuestro objetivo

fundamental.



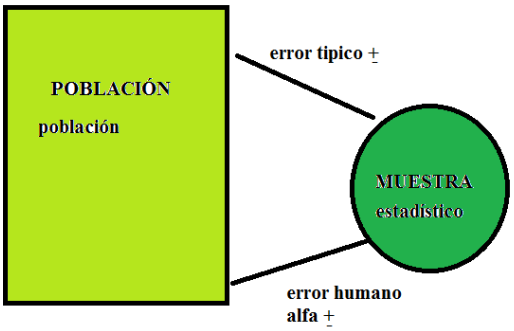
Para que se parezcan tenemos que tener en cuenta que lo ideal es que las muestras sean escogidas al azar y además debe tener un tamaño adecuado porque cuanto más grande sea más se parecerá a la población o mejor dicho, menos diferencias habrá entre ellas.

El problema fundamental es que las muestras no son idénticas a las poblaciones siempre en cualquier cálculo que hagamos ya sea la media , la desviación típica , la proporción… hay una diferencia entre la muestra y la población aunque normalmente la diferencia es pequeña. A esta diferencia se le llama SESGO estadístico y se calcula por medio del error típico concreto del cálculo que estemos trabajando. Dado que hay tantos errores típicos como cálculos se hacen.

A la hora de conocer la población tengo que tener en cuenta el error típico que puede ser por lo alto o por lo bajo , ya que la media de una muestra puede fallar y no ser igual a la población por lo alto y por lo bajo. Pero no solo existirá este problema a la hora de conocer a la población ya que además del error de sesgo existe otro error que cometemos a la hora de trabajar sobre la población que es el error humano. El error humano nos habla de que como cualquier persona , nos podemos equivocar a la hora de hacer un cálculo, recoger un dato, pasarlo a limpio…. Y todos esos fallos los debo tener también en cuenta , a este tipo de errores se llaman error humano o alfa, que tendremos que tenerlo en cuenta a la hora de conocer a la población.

Cuando yo quiero conocer la población lo que hago es juntar estos dos errores posibles que cometemos tanto el error humano como el error típico y así conseguimos todo el error o diferencia que existe entre la muestra y la población. A este error se le denomina error muestral o error máximo. Que se le tendrá que sumar y restar al estadístico (valor de la muestra) para

saber mas o menos cual será el valor de la población (parámetro) , dado que el valor exacto es imposible que lo lleguéis a conocer. Por lo tanto tendremos un intervalo que nos dirá entre que valores se encontrará el parámetro de la población. A este intervalo se le llamara intervalo de confianza, porque lo que hace es tener confianza en que el parámetro esté dento de ese intervalo.



error tipico * error humano = error muestral

estadistico ± error muestral = parametro en intervalo

PROPIEDADES DE LOS ESTIMADORES

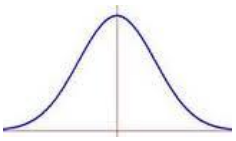
Vamos que la estimación es el proceso para conocer a la población con una determinada probabilidad, porque como todos los cálculos de la estadística no puedo asegurar que un cálculo sea exacto , solo puedo asegurar que hay un tanto por ciento de posibilidades muy elevado de que esto ocurra. Cuando hablamos del estimador es el valor de la muestra que quiero extrapolar , el estadístico que quiero convertir en parámetro, y este estimador debe cumplir toda una serie de características:

- Carencia de sesgo, o que sea insesgado y tenga homocedasticidad, todos estos términos lo que buscan es decir que el estadístico y el parámetro sean similares, no tienen porque ser iguales pero si similares y así cuando calcule el intervalo de confianza , el valor del intervalo este dentro del parámetro en cuestión.
- Eficiencia significa que tiene que haber poca desviación típica de todos los estimadores posibles de todos las muestras posibles.
- Consistencia: cuanto más grande es la muestra más se parecerá el estadístico al parámetro

- Suficiencia: un estimador es suficiente cuando es capaz de obtener de la muestra toda la información que está contenga acerca del parámetro.

DISTRIBUCIÓN MUESTRAL

En el mundo la población se distribuye siguiendo la curva normal , al ser una representación de la tabla de frecuencias, que quiere decir esto, que en un grupo de gente normal la mayoría de gente es normal, valga la redundancia, es media, vamos no es ni muy alta ni muy baja. Si hablamos por ejemplo de cociente intelectual la mayoría de la población tenemos un cociente intelectual normal, habiendo unas pocas personas que se encuentran en la parte derecha o sea son más inteligentes de la media y otro grupo de gente que se encuentra por debajo de la media . Su representación gráfica es lo que conocemos como campana de gauss. Bueno y ya sabemos que una representacion gráfica es una forma atractiva de ordenar la información disponible en la matriz y comprenderla a simple vista. Para realizar un gráfico necesitamos una distribucion de frecuencias, que es como se reparten los valores para cada posible situación. Además podemos escoger un gráfico u otro dependiendo de las variables o sea las características.

Al ver la representación gráfica de una distribucion de frecuencias, se ha descubierto que las normales como dije antes tinen la misma distribucion (forma) que llaman campana de Guass y se dibuja así: 

Es una distribucion teórica simétrica, si doblamos por la mitad la grafica la forma de ambos lados coinciden y es asintotica los valores nunca llegan al máximo posible . En muchos casos las variables psicológicas sobre todo en muestras grandes, tienen esta forma, por eso es tan importante. La curva normal y otras distribuciones teóricas la binomial, la t, F, x^2, son fundamentales en el campo de la inferencia estadística, al permitirnos ver cuando una diferencia es o no estadisticamente significativa , o si el azar no puede explicarla, vamos que se rechaze la hipótesis nula.

La estimación puntual consiste en utilizar el valor del estadístico calculado en la muestra como valor del parámetro que se desea estimar. Mediante este método se utiliza el estadístico obtenido en la muestra y se atribuye tal cual como parámetro de la población

Es poco probable que el valor del estadístico calculado en la muestra concreta coincida exactamente con el verdadero valor del parámetro y por ello es más interesante construir alrededor del estadístico de la muestra un intervalo, definido por su límite inferior y superior. Para ello se tiene en cuenta la precisión del estimador (su error típico), y la proporción de la distribución muestral calculado con una t de student o una z normal del nivel de significación α que significa la confianza que tenemos en nosotros mismos a la hora de hacer el estudio

La distribución muestral se define como la distribución de un estadístico en el muestreo. Está formada por los ∞ valores del estadístico obtenidos en ∞ muestras aleatorias del mismo tamaño extraídas de la misma población.

Si nos remitimos a la distribución muestral de la (media poblacional), está demostrado según el teorema central del límite y para muestras grandes (N > 30), que la distribución se asemeja a una distribución normal. Si el parámetro mediafuese conocido, comprobaríamos como la mayoría de las media de las muestras se encontrarían cerca del valor del parámetro, pero precisamente por ser muestras aleatorias, algunas se alejan un poco y otras, muy pocas, se alejan mucho de ese valor.

Por tanto, confiamos en que la muestra elegida sea una de las que el valor de su media esté cercano al valor de media de la población , pero no lo podemos asegurar. Por esta razón, en inferencia siempre hablamos de **nivel de confianza** (*porcentaje de confianza al hacer la estimación*; también puede darse en probabilidad y se denomina 1-alfa) y del **nivel de significación** alfa(*probabilidad de error que estamos dispuestos a asumir en la estimación*). Obviamente, se trata de conceptos complementarios que se refieren a lo mismo.

Para hallar el **intervalo confidencial**, es decir, *los valores entre los cuales es más probable que se encuentre el verdadero valor del parámetro*, necesitaremos calcular la media, a la que sumaremos y restaremos el **error muestral**; es decir, la diferencia más probable entre el estadístico y el parámetro. $IC = \bar{x} \pm EM$

El error muestral nos da una idea de la precisión de nuestra inferencia estadística. Cuanto más grande sea el error muestral, menor será la precisión en la estimación y menor será la utilidad de la estimación.

Una distribución muestral tiene variabilidad, por tanto tendrá su propia **desviación típica** , que recibe el nombre de **error típico**. Es un medida de dispersión con respecto al parámetro, es decir, nos indica la dispersión de las • de las ∞ muestras aleatorias extraídas respecto a la .

Cómo podemos conocer este dato?

Nuevamente lo estimaremos y nos basaremos en los datos de la muestra, en este caso Sx. Para la distribución muestral de medias, la fórmula es

$$\sigma_{\bar{x}} = \frac{S_x}{\sqrt{N-1}}$$

pondremos N-1 en el denominador cuando se haya calculado en la muestra la 2 insesgada, o solo N cuando se haya calculado la 2 sesgada (cuasivarianza o varianza, respectivamente).

El error típico no solo depende de la Sx, sino en gran medida del tamaño de la muestra. Cuanto más grande es N, más pequeño el cociente, más pequeño el error típico y más precisión en la estimación del parámetro.

Cómo se calcula el error muestral?

Previamente hemos tenido que definir el **nivel de significación** □ con el que vamos a realizar la estimación, **calcular la puntuación** z correspondiente a ese nivel (porque la distribución es normal) y aplicamos la fórmula

$$EM = z\alpha_{/2} . \sigma_{\bar{x}}$$

ver fig. 11.2 pág. 230. **Un ejemplo**: sean N = 1.000; $\bar{X}$ = 105 y Sx = 10. Estimar el valor del parámetro μ (intervalo confidencial) con un nivel de confianza del 99%.

IC = $\bar{X} \pm$ EM

nivel de confianza → 1-α = 99% → 0,99

nivel de significación → α = 0,01

Por tanto, α/2 = 0,005. El área bajo la curva normal de valor 0,005, corresponde a una z = - 2,575 (columna C)

$$\sigma_{\bar{x}} = \frac{S_x}{\sqrt{N-1}} = \frac{10}{\sqrt{999}} = 0{,}316$$

$$IC = 105 \pm z_{\alpha/2} . \sigma_{\bar{x}} = 105 \pm (-2{,}575)\ 0{,}316 \begin{cases} 104{,}18 \\ 105{,}81 \end{cases}$$

podemos afirmar con un nivel de confianza del 99%, que el valor de μ se encuentra en el intervalo siguiente: (104,18 ≤ μ ≤ 105,81)

Estimación de μ para pequeñas muestras

Cuando N ≤ 30, la distribución muestral de la media sigue la distribución t de Student. Esta distribución varía en función del número de sujetos, es decir, de sus grados de libertad, aunque se trata también de una distribución simétrica y asintótica. **Cuando N tiende a ∞, la distribución t tiende a la distribución z.**

Solo nos cambia en este caso el estadístico para calcular el error muestral. En lugar de trabajar con $z_{\alpha/2}$, lo haremos con $t_{\alpha/2}$ (ver tablas en pág. 347).

Ejemplo: los mismos datos del ejemplo anterior pero ahora con N = 25.

IC = 105 ± EM

α/2 = 0,005. Se distribuye según t con N-1 grados de libertad. Buscando en tablas, corresponde a un valor de $t_{\alpha/2}$ = 2,797

$$\sigma_{\bar{x}} = \frac{10}{\sqrt{24}} = 2.04$$

$$IC = 105 \pm t_{\alpha/2} . \sigma_{\bar{x}} = 105 \pm (2{,}797)\ 2{,}04 \begin{cases} 99{,}29 \\ 110{,}70 \end{cases}$$

como puede verse, hemos perdido una gran cantidad de precisión al disminuir el tamaño de la muestra, puesto que el intervalo de confianza es ahora mucho mayor.

Estimación del parámetro proporción (π)

Es un caso particular del anterior, en el que la media oscila entre 0 y 1. En las variables dicotómicas ya vimos que **p** corresponde a la proporción de unos y **q** a la proporción de ceros (q = 1-p). En este caso, la σ_p viene dada por la fórmula

$$\sigma_p = \frac{p.q}{\sqrt{N-1}}$$

El intervalo de confianza se establece igual que en el caso de la □, pero ahora partiendo de una proporción. Se emplea el valor de □ • 2 al nivel de confianza correspondiente.

Estimación de la puntuación verdadera en una prueba

Queremos realizar la estimación de la puntuación verdadera de un sujeto en un instrumento, o dicho de otra forma, *entre qué puntuaciones es más probable que se encuentre la verdadera puntuación*, ya que toda medida tiene algún margen de error.

De nuevo se trata de hallar el **intervalo de confianza** en el que es probable que se encuentre la verdadera puntuación del sujeto en la prueba. Sabiendo que la distribución muestral es normal, **necesitamos conocer el error típico de medida:**

$$\sigma_e = S_t.\sqrt{1-r_{xx}}$$

$S_t \rightarrow$ desviación típica total en el instrumento de medida
$r_{xx} \rightarrow$ coeficiente de fiabilidad

$$IC = X_i \pm EM$$

$X_i \rightarrow$ puntuación obtenida en la prueba

Intervalo de confianza de la puntuación estimada en la regresión lineal simple.

Queremos estimar la puntuación de un sujeto en una variable denominada criterio (Y) a partir de las puntuaciones conocidas en otra variable denominada predictora (X). Sabiendo que la distribución muestral de Y´ es la normal, debemos conocer el **error típico de estimaciones**, que para muestras grandes es

$$\sigma_{est} = S_y.\sqrt{1-r_{xy}^2}$$

$$IC = Y' \pm EM$$

Estimación del parámetro correlación de Pearson.

Introducción al concepto de significatividad estadística.

En muchas ocasiones, cuando calculamos una correlación, en realidad estamos más interesados en la relación que existe entre esas dos variables en la población, que en la muestra. **Se trata de calcular el**

intervalo de confianza para la correlación de Pearson, de modo que podamos estimar entre qué valores se encuentra dicho coeficiente entre la población.

Sabemos que la distribución muestral de la correlación de Pearson se asemeja a la distribución normal con el siguiente error típico:

siguiente error típico:

* para muestras grandes (N > 100) → $\sigma_{r_{xx}} = \frac{1}{\sqrt{N-1}}$

* y para muestras pequeñas ----------→ $\sigma_{r_{xx}} = \frac{1-r_{xx}^2}{\sqrt{N-1}}$

y como ya es habitual $IC = r_{xx} \pm EM$; donde $EM = z_{\alpha/2} . \sigma_{r_{xx}}$

Ejemplo: Sea N = 20 y r_{xx} = 0,35. Cómo será la correlación en la población con un nivel de confianza del 95%?.

$$\sigma_{r_{xx}} = \frac{1-(0,35)^2}{\sqrt{20-1}} = \frac{0,8775}{4,358} = 0,201$$

$\alpha = 0,05 \rightarrow \alpha/2 = 0,025$. En tablas, un área bajo la curva de 0,025 (columna C) corresponde a una $z = -1,96$

$IC = 0,35 \pm (-1,96).0,201$ < -0,04 ; 0,74

El intervalo tan amplio es debido al pequeño tamaño de la muestra. Cuanto más pequeña sea la muestra, más imprecisa será la estimación.

El intervalo tan amplio es debido al pequeño tamaño de la muestra. Cuanto más pequeña sea la muestra, más imprecisa será la estimación.

Significación o significatividad estadística

Hablamos de significación estadística cuando nos referimos al significado de la diferencia entre dos medidas. Decimos también que un coeficiente de correlación es estadísticamente significativo cuando es distinto de cero, es decir, cuando el coeficiente de correlación (*estadístico*) es lo suficientemente grande como para decir que la correlación en la población (*parámetro*) de referencia es distinta de cero. Si queremos conocer cuál es su magnitud en la población, entonces calcularemos el intervalo de confianza.

Cuando estimamos el intervalo de confianza en el cual es probable que se encuentre la verdadera correlación en la población, y este intervalo contiene el valor cero (ausencia absoluta de correlación) diremos que dicha correlación NO es estadísticamente significativa.

Estimación del parámetro diferencia de medias $((\mu_1 - \mu_2)$

Si una universidad encuentra que el CI medio de sus estudiantes es de 110, y otra universidad lo tiene de 105, cómo es la diferencia entre estas dos puntuaciones en la población de referencia?. La diferencia entre estas dos puntuaciones es estadísticamente igual a cero?. Si dicha diferencia es compatible con una diferencia igual a cero, concluimos que ambas muestras tienen el mismo CI, al nivel de confianza fijado, y que la diferencia encontrada es aleatoria.

Si establecemos el intervalo de confianza a partir del estadístico diferencia de medias, obtendremos los límites confidenciales entre los cuales es más probable que se encuentre la diferencia de medias en la población.

Si este intervalo incluye la puntuación cero, entonces dicha diferencia es compatible con una diferencia de medias igual a cero, y en consecuencia, podremos interpretar que dicha diferencia es estadísticamente igual a cero

$$IC = (\bar{X}_1 - \bar{X}_2) \pm EM \qquad EM = z_{\alpha/2} \cdot \sigma_{(\bar{X}_1 - \bar{X}_2)}$$

$$\sigma_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{S_1^2}{N_1 - 1} + \frac{S_2^2}{N_2 - 1}}$$ para muestras grandes e independientes (N > 100)

$$\sigma_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\left(\frac{N_1 S_1^2 + N_2 S_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$ para muestras pequeñas e independientes (N ≤ 100)

Ejemplo:

Universidad 1	Universidad 2
$\bar{X} = 110$	$\bar{X} = 105$
$S_1 = 10$	$S_2 = 12$
$N_1 = 90$	$N_2 = 120$

Calculamos primero el error típico de la diferencia de medias.

$$\sigma_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{90.100 + 120.144}{90 + 120 - 2}\left(\frac{1}{90} + \frac{1}{120}\right)} = \sqrt{\frac{183960}{74880}} = \sqrt{2,4567} = 1,567$$

La $z_{\alpha/2}$ al 95% de nivel de confianza, ya sabemos que corresponde a una $z = -1,96$

EM = (- 1,96) 1,567 = - 3,071 → por tanto:

IC = (110 – 105) ± (- 3,071) < 1,929 ; 8,071

Al ser el intervalo incompatible con el valor cero, concluimos que la diferencia marcada es estadísticamente significativa, es decir, que las diferencias de medias en CI entre ambos grupos no son aleatorias.

Estimación del parámetro diferencia de proporciones $(\pi_1 - \pi_2)$

Lo único que varía en este caso es el error típico. Al igual que en el contraste de medias, lo más interesante suele ser la diferencia entre dos proporciones.

El **error típico de la diferencia de proporciones** es:

$$\sigma_{(p_1-p_2)} = \sqrt{pq.\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}$$

y la estimación del intervalo de confianza y su interpretación será la misma que en el caso anterior, pero con medias en términos de proporción, que frecuentemente se multiplican por 100 para convertirlos en porcentajes.

Estimación de parámetros y contraste de hipótesis

Todo se reduce al contraste de una hipótesis estadística, denominada **hipótesis nula (Ho)**, según la cual NO existen diferencias estadísticamente significativas. Se afirma en ella que no hay efecto presente de la VI sobre la VD.

Se establece una zona central alrededor de la media en la curva normal, en la que es más probable que las diferencias encontradas entre las medias de las muestras se deban efectivamente a los efectos del azar. Es la **zona de aceptación** (ó de no rechazo) **de Ho**. Por tanto, al no poder rechazar la Ho, decimos que no es falsa.

Queda de forma inmediata **otra zona** (bilateral o unilateralmente) en la que las diferencias entre las medias resulta muy improbable (al nivel de confianza establecido) que sean aleatorias, por lo que dichas diferencias se atribuyen a los efectos de la VI.

La Ho se expresa así:

Hipótesis nula → Ho : μ1 - μ2 = 0

las diferencias entre las • de los grupos o muestras son estadísticamente iguales a cero;

o bien, que las diferencias empíricas que existen entre las • de las muestras se deben al azar;

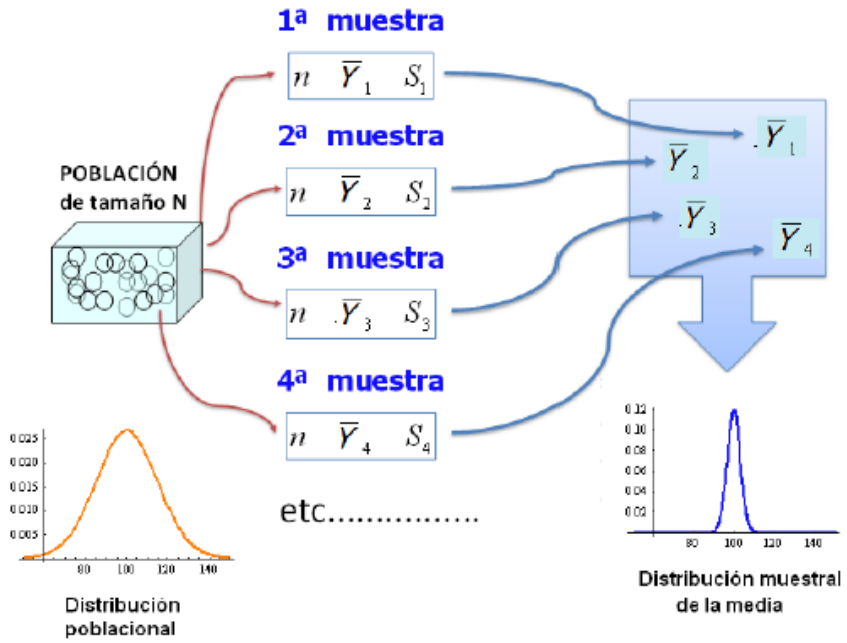
o bien, que los valores paramétricos son iguales;

o bien, que ambas muestras pertenecen a la misma población.

Las fórmulas para hallar **el intervalo de confianza** son las mismas que el epígrafe anterior, para muestras grandes y pequeñas.

Si resulta que nuestra diferencia empírica de medias se encuentra dentro del intervalo de confianza de la distribución muestral conforme a Ho, diremos que nuestra diferencia de medias es compatible con una diferencia de medias igual a cero, y por tanto, que se trata de una diferencia estadísticamente no significativa o igual a cero.

Lo que haremos después será **calcular un estadístico** (**t**, z, etc.) que nos dirá cuántas desviaciones típicas (errores típicos) se aleja nuestra diferencia de medias de una diferencia de medias igual a cero.


1ª muestra
$n \quad \bar{Y}_1 \quad S_1$
2ª muestra
$n \quad \bar{Y}_2 \quad S_2$
3ª muestra
$n \quad \bar{Y}_3 \quad S_3$
4ª muestra
$n \quad \bar{Y}_4 \quad S_4$
etc.....
POBLACIÓN
de tamaño N
$\bar{Y}_1$
$\bar{Y}_2$
$\bar{Y}_3$
$\bar{Y}_4$
Distribución
poblacional
Distribución muestral
de la media