Fundamentos en Estadística

V. Trujillo

GRC-MERVEX (CO de Vigo. IEO)

Marzo 2021

9.- Inferencia estadística

Introducción

Partimos de la base de que el Método Estadístico se basa en el razonamiento inductivo, o sea que hace proposiciones que van de lo "particular a lo general"[1]. Aplicando el método estadístico, la inferencia estadística nos permite sacar conclusiones sobre la población a partir de una muestra <u>representativa</u>.

El proceso nos permite en general, bajo un determinado marco probabilístico, realizar hipótesis sobre parámetros poblaciones desconocidos y desarrollar formas para comprobar la verosimilitud de tales afirmaciones a través de los estimadores.

Esencia de la Teoría del Muestreo(TM)

En esencia y sencillamente, la TM lo que nos tiene qué decir es:

"Cual es el comportamiento del estadístico **u** en el Muestreo".

Decir cual es el comportamiento, significa saber:

1. Cual es **el valor esperado de u** $\Rightarrow E[u]$
2. Cual es **la varianza de u** $\Rightarrow V[u]\ o\ Var[u]$
3. Cual es **el error de u** $\Rightarrow Error[u]$
4. Cual es **la distribución de u** $\Rightarrow u \cap\ ?$

1. El valor esperado o esperanza $E[u]$ es la media del estadístico **u** en el muestreo. Media de todos los **u** posibles, de todo el conjunto de medias posibles.
2. $V[u] = s_u^2$
3. La desviación típica del estadístico **u** en el muestreo, es el <u>error</u> estadístico[2]. $\sqrt{V[u]} = Error[u] = s_u$
4. $u \cap^3$? ¿Qué tipo de distribución tiene **u**?.

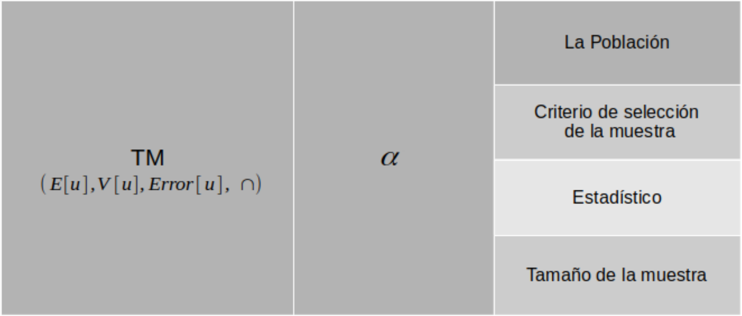
Estas son las cuatro características[4] principales a las que tiene que responder la Teoría del Muestreo (TM).

Pero a veces la teoría no puede responder(demostrar) de forma analítica a estas cuatro preguntas y aquí es cuando entran de lleno las diferentes técnicas de remuestreo por simulación numérica[5]. Lo que hacen simplemente es simular (generar) numéricamente por ordenador un conjunto de muestras posibles en el mundo del muestreo.

Dependencias de la TM

La TM (E, V, Error, $\cap$) depende de:

1. La población, i.e. de los parámetros de la población.
2. Criterio de selección de la muestra.
3. Estadístico elegido.
4. Tamaño de la muestra.


TM
$\left(E[u], V[u], Error[u],\ \cap\right)$
α
La Población
Criterio de selección de la muestra
Estadístico
Tamaño de la muestra

Tipos de inferencias

1. **Estimación** de parámetros de la población
2. **Prueba de hipótesis**. Hipótesis que hacemos sobre los parámetros de la población

Estimación de parámetros

Para estimar un parámetro, tengo que seleccionar un estadístico. Pero, ¿qué estadístico tengo que seleccionar?. El estadístico seleccionado se llama estimador, por ejemplo: si quiero estimar μ (media de la población), selecciono $\bar{x}$ (media aritmética de la muestra), a $\bar{x}$ se le llama estimador de μ, pero existen otros posibles estimadores de μ, como por ejemplo la mediana.

Pero todo esto no nos dice nada. Lo que se quiere saber es si la estimación es buena o mala. Para solucionar este problema, recurrimos, como siempre, a la TM que nos da resultados concretos sobre los estimadores.

Por ejemplo:

Parámetro	Estimador	TM	Cualidad
μ	$\bar{x}$	$E[\bar{x}] = \mu$	**no sesgado**

La TM nos dice que el valor esperado de $\bar{x}$ (media del conjunto de todos las $\bar{x}$ posibles) es μ (la media de la población), y nos dice también que es un estimador **no sesgado** o **centrado**, y que además no depende del criterio de selección de la muestra[6].

Que sea no sesgado o no viciado, es una buena característica para un estimador, quiere decir que con un sólo valor de la muestra, podemos utilizarlo como estimador de la población[7]. Si hiciéramos esto miles de veces, por ejemplo: por simulación; tendríamos la misma probabilidad de que el valor se encontrase tanto a la derecha como a la izquierda de la media poblacional.

Con respecto a los estadísticos de dispersión la TM nos dice que:

Parámetro	Estimador	TM	Cualidad
σ^2	s^2	$E[s^2] = \sigma^2$	**no sesgado**

La TM dice que sólo es no sesgado si el criterio es un m.a.s[8] sin reposición

Parámetro	Estimador	TM	Cualidad
σ	s	$E[s] \neq \sigma$	**es sesgado**

La TM dice que el valor esperado de s en el muestreo **no es** la desviación típica poblacional. El estimador es sesgado o no centrado; y el sesgo es grave para un tamaño de muestra menor que 30, si $n > 30$ el sesgo es muy pequeño. Esta cualidad es llamada **consistencia** de un estimador.

La TM dice que: $E[s] \neq \sigma$, pero si: $n \to \infty \Rightarrow E[s] = \sigma$; a medida que el tamaño muestral se incrementa, s es mejor estimador de σ. Se dice entonces que s es un estimador consistente, esta cualidad es más importante que la de sesgado.

Decíamos que si queremos estimar μ podíamos utilizar tanto la media como la mediana($\tilde{x}$). Pero, ¿con cuál de las dos nos quedaríamos?.

Parámetro	Estimador	TM	Cualidad
μ	$\bar{x}$	$E[\bar{x}] = \mu$	no sesgado
μ	$\tilde{x}$	$E[\tilde{x}] = \mu$	no sesgado

Pero si miramos cómo son las varianzas en el muestreo de estos dos estadísticos, vemos que: $V[\bar{x}] < V[\tilde{x}]$.

Esto quiere decir que la media aritmética como estimador de μ es más **eficiente** o **preciso** que la mediana. Que sea más eficiente (formalmente en estadística) significa que la varianza en el muestreo es menor.

Con estos ejemplos se han visto tres[9] de las principales cualidades/propiedades de los estimadores puntuales de parámetros de la población, que son:

1. **Sesgo**[10]
2. **Consistencia**
3. **Eficiencia**[11]

Estimaciones puntuales de los parámetros

Aquí veremos sólo las estimaciones puntuales y por intervalos, pero quedan dos métodos de estimación muy importantes a considerar:

- la estimación de máxima verosimilitud.
- la estimación bayesiana.

Distribución de la media en el muestreo

I.

$$E[\bar{x}] = \mu \tag{1}$$

$\mu =$ media de la población
$\bar{x} =$ media de la muestra
$E =$ media del muestreo

Se ve cómo están relacionados los tres mundos de la estadística:

"La media en el muestreo de la media de la muestra es la media de la población"

II.

$$V[\bar{x}] = \frac{\sigma^2}{n} \tag{2}$$

Esto es cierto, cuando población es infinita o cuándo el muestreo es con reposición.

Si N es finita y el criterio del muestreo es un aleatorio simple(a.s.) sin reposición, entonces[12]:

$$V[\bar{x}] = \left(1 - \frac{n}{N}\right) \frac{S^2}{n} \tag{3}$$

II'. $(N-1)\ S^2 = N\sigma^2$; S^2 es la varianza modificada. Sabiendo una varianza se puede deducir la otra. Normalmente como la corrección para poblaciones finitas $(1-\frac{n}{N})$ es pequeña, salvo en casos excepcionales, se suele usar el conocido:

$$V[\bar{x}] = \frac{s^2}{n} \tag{4}$$

donde s^2 es la varianza de la muestra.

Como la población suele ser muy grande: $\left(1-\frac{n}{N}\right) * \frac{S^2}{n} \cong \frac{\sigma^2}{n}$
La varianza en el muestreo de $\bar{x}$ es igual a la varianza de la población dividido por el tamaño de la muestra. $V[\bar{x}]$ es más pequeño a medida que el tamaño de la muestra aumenta, es decir: la distribución estará más concentrada entorno a la verdadera media de la población, luego la estimación es más precisa[13].

III.

$$Error[\bar{x}] = \sqrt{V[\bar{x}]} = \frac{\sigma}{\sqrt{n}} \qquad (5)$$

La desviación típica en el muestreo ($Error[\bar{x}]$), cuánto mayor sea el tamaño de la muestra, menor será su error[14].

Si desconocemos σ^2, podremos utilizar la varianza muestral s^2 para calcular el ($\underline{error}[\bar{x}]$). Cuando hacemos esto se denomina $var[\bar{x}]$ en vez de $V[\bar{x}]$.

Si

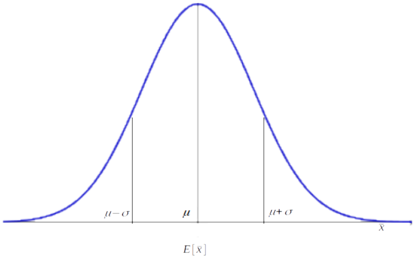
$$var[\bar{x}] = \sqrt{\frac{s^2}{n}} \quad \Rightarrow \quad error[\bar{x}] = \frac{s}{\sqrt{n}} \qquad (6)$$

Cuando usamos esta aproximación ($\frac{s^2}{n}$), la $\bar{x}$ se distribuye como una distribución t de Student. En la práctica es lo que se utiliza, pero ... si $n > 30$ la t-Student se comporta como una distribución Normal[15].

$$\bar{x} \cap N(E,V) \tag{7}$$

la media de la muestra tiene una distribución normal con media, el valor esperado de $(E[\bar{x}])$ y varianza $V[\bar{x}]$. Entonces

$$\bar{x} \;\dot{\cap}\; N(E,V) \Rightarrow N(\mu, \frac{\sigma^2}{n}) \quad ; donde\ \dot{\cap}\ o\ \sim\ \rightarrow \text{"}aproximadamente\text{"}$$



Es decir que la TM nos dice también, que la media muestral en el mundo del muestreo se distribuye como una normal de media la esperanza y desviación típica, el error típico. Es fundamental conocer la distribución de los estadísticos en el muestreo.
Para cualquier población, la aproximación a la Normal es mejor cuando n es grande $(n > 30)$ y cuando la población es simétrica. Deja de ser aproximada para ser exacta, cuando la <u>Población es Normal</u>.

Distribución de la s^2 en el muestreo

I.

$$E[s^2] = \sigma^2 \qquad (8)$$

II. Por ahora vamos a dejar de ver como son: $V[s^2]$ *y el* $Error[s^2]$, siendo de mayor más interés ver cómo es la distribución de s^2:

$$s^2 = \frac{S_{xx}}{(n-1)} \quad ; \quad s^2 \; varianza \; de \; la \; muestra$$

$$\left(\frac{S_{xx}}{\sigma^2}\right) \cap \chi^2_{(n-1)} \Rightarrow E\left[\frac{S_{xx}}{\sigma^2}\right] = n-1 \rightarrow E\left[\frac{S_{xx}}{(n-1)}\right] = \sigma^2 \qquad (9)$$

Por eso decimos que[16]:~$E[s^2] = \sigma^2$

Nota: igual veis un desarrollo análogo, ya que realmente $E[s^2] = \sigma^2$ es una "aproximación"[17]:

$$E[\frac{1}{n}\sum(x_i - \bar{x})^2] = \frac{(n-1)}{n}\sigma^2$$

$$\frac{(n-1)s^2}{\sigma^2} = \frac{1}{\sigma^2}\sum(x_i - \bar{x})^2 \sim \chi^2_{(n-1)} \qquad (10)$$

Si la variable X tiene una distribución normal, la distribución de $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$.

Entre otras propiedades, la varianza muestral no es simétrica, tiene asimetría positiva.

Estimaciones por intervalos de los parámetros

Intervalos de confianza de la media y la varianza

Al estimar los parámetros obteníamos un valor puntual y como nos dice la teoría de muestreo: la distribución de los mismos. Esta distribución nos permite establecer la precisión de nuestra estimación gracias al cálculo de los intervalos de confianza[18](C) de nuestros estimadores.

La teoría del muestreo nos da los datos necesarios para calcular los intervalos de confianza de los estimadores, diciéndonos que la media se distribuye como una normal de media μ y desviación típica, el error de la media.

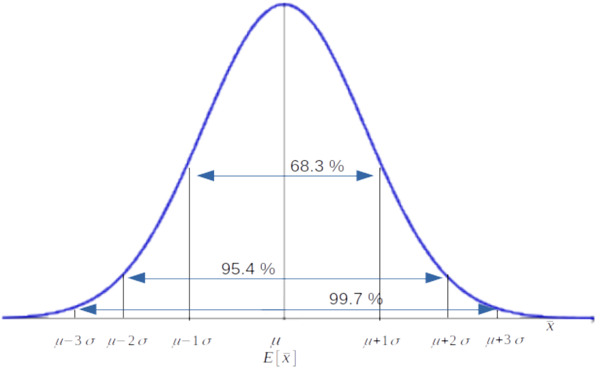
Si tipificamos, se distribuirá como una normal de media cero y desviación típica 1 ($Z \sim N(0,1)$).

Como vimos en la ecuación sobre la función de densidad (en el capítulo 7), definíamos la función de densidad de probabilidad (pdf) para un intervalo dado como:

$$P(a \leq X \leq b) = \int_a^b f(x)dx \tag{11}$$

Que nos permite calcular una densidad entre dos valores dados (a, b) y viceversa, a partir de una determinada área establecer cuales son los valores que acotan tal densidad.

Para una distribución normal tipificada, las densidades correspondientes a 1, 2 y 3 desviaciones(σ), comprenden aproximadamente el 68.3 %, 95.4 % y 99.7 % respectivamente de la densidad. A estas áreas o densidades les podremos denominar **Confianzas** y al resto de la densidad[19] le llamaremos α o **nivel de significación**: $\alpha = 1 - C$, que tiene una enorme importancia en la estadística y que se verá en mayor profundidad en la sección siguiente sobre el Contraste o Prueba de hipótesis.



Intervalo de confianza para la media

En el caso de la desviación normal tipificada (Z), los intervalos de confianza se construirán habitualmente[20] de la siguiente manera:

$$
\begin{aligned}
P(-z \leq Z \leq +z) &= 1-\alpha = C \\
P\left(-z \leq \frac{\bar{x}-\mu}{s/\sqrt{n}} \leq +z\right) &= 1-\alpha = C \\
P\left(-z\frac{s}{\sqrt{n}} \leq \bar{x}-\mu \leq +z\frac{s}{\sqrt{n}}\right) &= 1-\alpha = C \\
P\left(-z\frac{s}{\sqrt{n}} \leq \mu-\bar{x} \leq +z\frac{s}{\sqrt{n}}\right) &= 1-\alpha = C \\
P\left(\bar{x}-z\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x}+z\frac{s}{\sqrt{n}}\right) &= 1-\alpha = C
\end{aligned}
\tag{12}
$$

$$
\begin{gathered}
P(L_1 \leq \mu \leq L_2) = 1-\alpha = C \\
L_1 = \bar{x}-z\frac{s}{\sqrt{n}} \qquad y \qquad L_2 = \bar{x}+z\frac{s}{\sqrt{n}}
\end{gathered}
\tag{13}
$$

L_1 y L_2, son los límites del intervalo de confianza.

Si el tamaño de las muestras es pequeño y no conocemos la desviación típica paramétrica, entonces debemos utilizar la distribución **t**, para usar la desviación típica del muestreo $s_{\bar{x}}$[21]:

$$Var[\bar{x}] = \frac{s^2}{n} \qquad ; \qquad s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Donde s^2 y s son la varianza y desviación típica de la muestra. Siendo los límites del intervalo de confianza:

$$L_1 = \bar{x} - ts_{\bar{x}} \qquad ; \qquad L_2 = \bar{x} + ts_{\bar{x}} \tag{14}$$

Debido a que:

$$\begin{aligned} C &= P(-t \leq t_{n-1} \leq +t) \qquad (15) \\ C &= P(\bar{x} - ts_{\bar{x}} \leq \mu \leq \bar{x} + ts_{\bar{x}}) \end{aligned}$$

Si se supone que los límites son simétricos este intervalo da la mínima amplitud, si se hacen no simétricos la amplitud del intervalo de confianza es mayor para esa confianza. Así que la teoría de muestreo aconseja que sean simétricos.

Intervalo de confianza para la varianza

La teoría de muestreo en el caso de la varianza nos informa en el muestreo sobre la esperanza, la varianza y el error, así como su distribución. Por lo tanto, se puede calcular perfectamente el intervalo de confianza de la varianza.

Según la teoría del muestreo:

$$\frac{S_{xx}}{\sigma^2} \;\cap\; \chi^2_{n-1}; \qquad\qquad Error\left(\frac{S_{xx}}{\sigma^2}\right) = \sqrt{2(n-1)}$$

$$C = P\left(L_1 \leq \frac{S_{xx}}{\sigma^2} \leq L_2\right); \qquad\qquad Como\ \frac{S_{xx}}{\sigma^2}\cap\chi^2_{n-1} \;\Rightarrow\; E\left(\frac{S_{xx}}{\sigma^2}\right) = n-1$$

$$\backslash \qquad\qquad\qquad\qquad \backslash$$

$$L_1 = (n-1) - \chi^2_{n-1(\alpha/2)} Error\left(\frac{S_{xx}}{\sigma^2}\right)$$

$$L_2 = (n-1) + \chi^2_{n-1(\alpha/2)} Error\left(\frac{S_{xx}}{\sigma^2}\right)$$

Entonces:

$$\frac{S_{xx}}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} \cap \chi^2_{n-1}$$

$$P(-\chi^2_{n-1} \leq \chi^2 \leq +\chi^2_{n-1})$$

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

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

$$Si \;\; (n{-}1)s^2 = \sum y^2 \quad \Rightarrow \quad \boldsymbol{P\left(\frac{\sum y^2}{\chi^2_{n-1(\alpha/2)}} \leq \sigma^2 \leq \frac{\sum y^2}{\chi^2_{n-1(1-\alpha/2)}}\right) = C}$$

Contraste o Prueba de hipótesis

Prácticamente cualquier suposición puede transformarse en una hipótesis que se puede probar. La hipótesis es lo que queremos probar y en cambio una suposición, al igual que una asunción, no necesita ser probada.

Las hipótesis se hacen siempre sobre el mundo de la Población.

Esto no hay que olvidarlo nunca. La hipótesis se denomina tradicionalmente como H_0, por ej.: podría ser que la media de mi población es igual a $4 \Rightarrow H_0 : \mu = 4$; o, que la varianza de mi población es igual a $5 \Rightarrow H_0 : \sigma^2 = 5$; o que tengo una población (1) que tiene como media μ_1 y otra población(2) que tiene como media μ_2 y por lo tanto podríamos plantear una hipótesis del tipo $\Rightarrow H_0 : \mu_1 = \mu_2$

Siempre sobre la Población, esto es lo que tenemos que probar. Para aceptar (no rechazar)[22] o no(rechazar) una hipótesis tenemos que seguir siempre el mismo proceso, que este consta de 5 partes:

I.- Proceso de realización:

1. Formular la hipótesis $\Rightarrow H_0$ / H_1.
2. Definir el nivel de confianza[23].
   Definir un criterio de selección y el tamaño de la muestra.
   Definir el estadístico que se utilizará.
3. Definir el criterio de decisión sobre rechazar la H_0 o no hacerlo. Esto se hace en base a la TM.
4. Seleccionar la muestra y calcular el valor del estadístico de contraste.
5. Toma de decisión: rechazar o no rechazar H_0.

Cuadro de decisión:

		HIPÓTESIS	
		VERDADERA	FALSA
DECISIÓN	RECHAZAMOS	Incorrecta	Correcta
	NO RECHAZAMOS	Correcta	Incorrecta

Otra decisión podría ser: ni rechazar, ni no-rechazar[24]. En estadística, nuestras decisiones son tomadas con probabilidad, <u>nunca tenemos la certeza</u> (casi nunca). Lo que tenemos es: la probabilidad de estar en lo correcto o en lo incorrecto (<u>**no**</u>, verdadero o falso). Por eso en estadística, hacemos otro cuadro hablando de probabilidades.

II.- Probabilidades asociadas:

		HIPÓTESIS	
		VERDADERA	FALSA
DECISIÓN	RECHAZAMOS	α Error tipo I	$1-\beta$
	NO RECHAZAMOS	$1-\alpha$	β Error tipo II

Donde $\boldsymbol{\alpha}$ es: **la probabilidad de rechazar una hipótesis verdadera** y $\boldsymbol{\beta}$ es: **la probabilidad de no rechazar una hipótesis falsa**. También son conocidos como error de tipo I (α) y tipo II[25] (β) respectivamente.

$(\mathbf{1}-\boldsymbol{\alpha})$ **es la confianza(C)**[26] que se puede interpretar como: **tener una confianza del 95 % de no rechazar la hipótesis cuando es verdadera**. En prueba de hipótesis, en vez de la confianza, lo habitual es hablar más del error; y a $\boldsymbol{\alpha}$ se le llama <u>nivel de significación</u>. Por lo tanto, tener un $\boldsymbol{\alpha}$ de 5 % implica tener una confianza del 95 %, es decir, significa tener una probabilidad del 5 % de rechazar la hipótesis cuando esta sea verdadera.

$$C = 1 - \alpha = P(no\ rechazar\ /\ H_0\ verdadera)$$
$$\alpha = P(rechazar\ /\ H_0\ verdadera)$$

$$Potencia = 1 - \beta = P(rechazar\ /\ H_0\ falsa)$$
$$\beta = P(no\ rechazar\ /\ H_0\ falsa)$$

Ejemplo[27]:

1. $H_0 : \mu = 4$, con σ^2 conocida $\Rightarrow \sigma^2 = 12{,}1$
2. .- nivel de confianza y significación, 95 % y 5 % respectivamente.
   .- criterio de selección de la muestra: m.a.s sin reposición, de tamaño $n = 10$.
   .- estadístico: $\bar{x}$
3.

$$C = P(\bar{x} - z * s_{\bar{x}} \leq \mu \leq \bar{x} + z * s_{\bar{x}})$$
$$C = P(\mu - z * s_{\bar{x}} \leq \bar{x} \leq \mu + z * s_{\bar{x}})$$

Esta última expresión es la que se usa ya que es la probabilidad de no rechazar H_0 cuando es verdadera y cuando H_0 es verdadera $\mu = 4$.

Ahora se calculan los valores que se necesitan, por ejemplo el error de la media muestral.

$$s_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{12{,}1}{10} = 1{,}21 \quad \Rightarrow \quad s_{\bar{x}} = 1{,}1$$

3. con't

Como sabemos que:

$$\begin{aligned} C &= P(-z \leq Z \leq +z) \\ 0{,}95 &= P(-1{,}96 \leq Z \leq +1{,}96) \quad ; z = 1{,}96 \\ 0{,}95 &= P(4 - 1{,}96 * 1{,}1 \leq \bar{x} \leq 4 + 1{,}96 * 1{,}1) \end{aligned}$$

Quedando finalmente:

$$0{,}95 = P(1{,}84 \leq \bar{x} \leq 6{,}16) \quad / \quad H_0 : \mu = 4 \quad \Rightarrow H_0 : \textit{es verdadera}$$

Si a partir de la la muestra se obtiene una media comprendida entre 1,84 y 6,16, se tiene un 95 % de confianza de no rechazar H_0, y esta es la regla de decisión. Esta decisión tiene una probabilidad del 95 % de no rechazarse y un 5 % de rechazarse.
A la región de rechazo se le denomina también **región crítica** y a la de no rechazo, **región de aceptación**.

4. Selección de la muestra y cálculo de la media.

5. Si la muestra cae en la región de no rechazo entonces H_0 no es rechazada, con un nivel de significación α (H_0 no es rechazada con un nivel de significación del 5 %). Si cae en la región de rechazo, ahí sí que se puede decir que es significativamente diferente.