Sociedad Ecuatoriana de Estadística

"Análisis de Encuestas por Muestreo con R"

Unidad 5: Inferencia estadística.

Andrés Peña M.

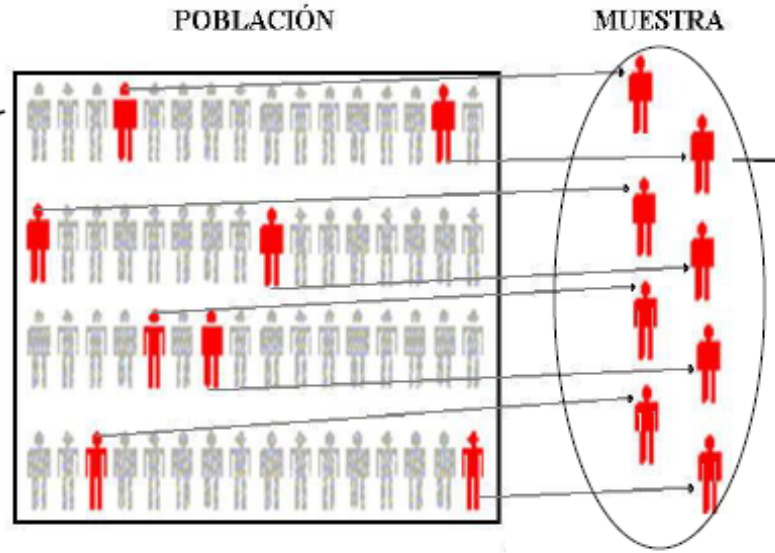
a.pena@rusersgroup.com

Mayo 2021

Tabla de contenidos

1. Estimación de parámetros

Inferencia estadística



PARÁMETROS θ

μ, σ^2, P

ESTIMADORES $\hat{\theta}$

$\bar{x}$, S^2, $\hat{p}$

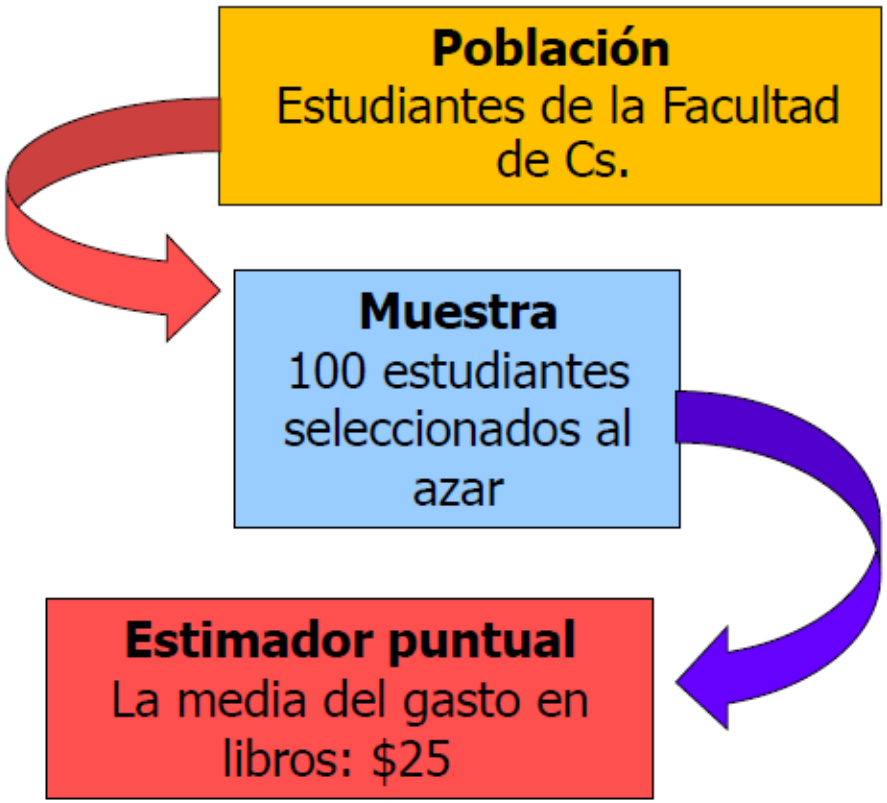
ESTADÍSTICA INFERENCIAL

Estimación de parámetros

Pruebas de hipótesis

Estimación Puntual

El valor de la media muestral de cualquier muestra se puede considerar como una **estimación puntual** ("puntual", porque se trata de un solo número, que corresponde a un solo punto de la recta numérica) de la media poblacional μ.



Una estimación puntual no proporciona por sí misma información acerca de la precisión y confiabilidad de la estimación. Por la variabilidad, es poco probable que $\bar{x} = \mu$.
La estimación puntual no indica cuán cerca podría estar la media muestral con respecto a la media poblacional.

Estimación Puntual

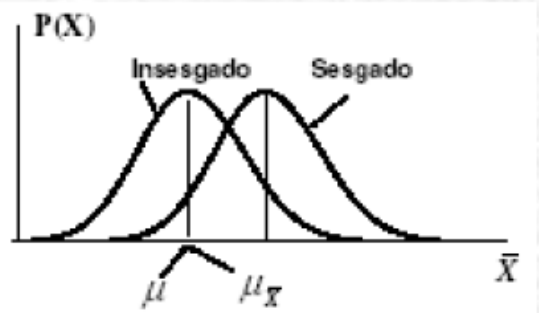
Propiedades de los buenos estimadores:

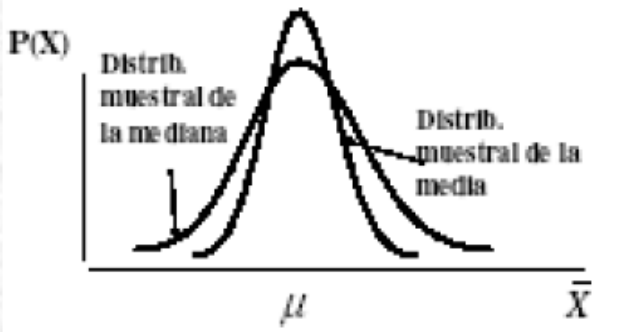
Insesgabilidad $E(\hat{\theta}) = \theta$

Consistencia $\lim_{n\to\infty} \left(\Pr\left|\hat{\theta} - \theta\right| < \varepsilon\right) = 1$

Eficiencia $V(\hat{\theta}_1) < V(\hat{\theta}_2)$

Suficiencia cuando utiliza toda la información que surge de la muestra





Estimación por intervalos

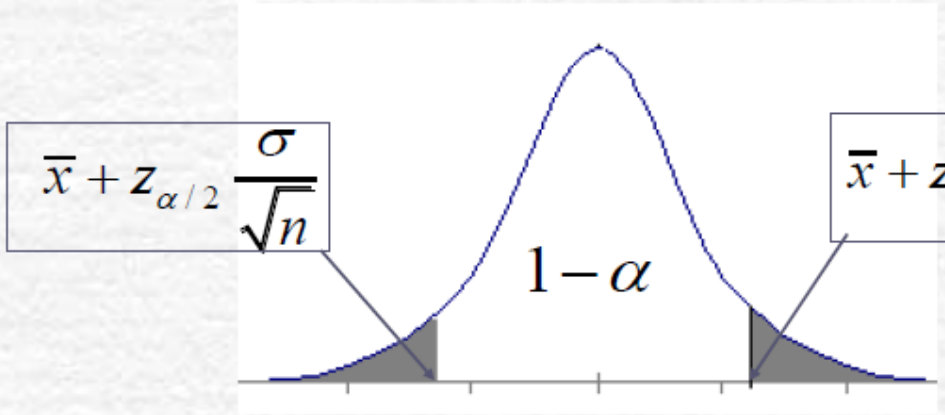
Una alternativa respecto a informar un solo valor sensible del parámetro es calcular un intervalo de valores (intervalo de confianza IC)

Estimación por intervalos

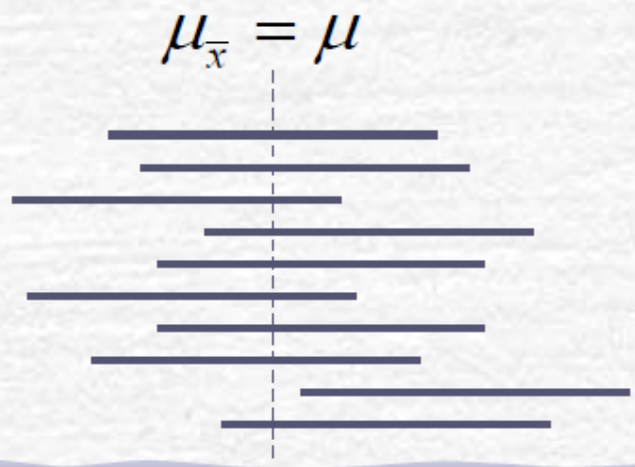
Un intervalo de confianza siempre se calcula al seleccionar primero un **nivel de confianza** que es una medida del grado de confiabilidad del intervalo.

Estimación por intervalos

Distribución muestral de la media



Los intervalos van de $\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ a $\bar{x} + z_{(1-\alpha/2)}\frac{\sigma}{\sqrt{n}}$



100 (1-α) % de los intervalos construidos contienen a μ y 100(α) % no.

Estimación por intervalos

Pasos:

1) Establecer cuál es el parámetro desconocido θ y qué se conoce de la población.
2) Buscar un estimador puntual $\hat{\theta}=g(x_1, x_2, \ldots, x_n)$ función de las observaciones muestrales en una muestra de tamaño *n.*
3) Plantear un estadístico, función del estimador y del parámetro $h(\hat{\theta}, \theta)$ que permita despejar algebraicamente el parámetro (única incógnita de la expresión) y que tenga una distribución de probabilidad conocida.
4) Fijado el nivel de confianza 1-α que indica la probabilidad de que el intervalo construido contenga al parámetro poblacional se determina el intervalo para el estadístico: $P(k_1< h(\hat{\theta}, \theta) <k_2)=1-\alpha$. Los valores k_1 y k_2 se determinan en función de la distribución de probabilidad del estadístico y el nivel de confianza.
5) Se despeja el parámetro, obteniéndose un intervalo aleatorio que tiene una probabilidad $1-\alpha$ de contenerlo.
6) Con los valores de la muestra se realiza la estimación y se obtienen los límites inferior y superior de confianza entre los cuales, con una confianza de $(1-\alpha)\%$ se encuentra θ.

Intervalo de confianza para μ con σ conocida

Comenzaremos con una situación problemática simple. Suponga que el parámetro de interés es una media poblacional μ y que:

La distribución de la población es normal o n $\geq$ 30.

Se conoce el valor de la desviación estándar poblacional σ.

Bajo estas condiciones se usa el estadístico:

Recuerde el muestreo en poblaciones normales

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

La normalidad de la distribución poblacional suele ser una suposición razonable. Sin embargo, si se ignora el valor de μ es poco probable que esté disponible el valor de σ. Posteriormente, desarrollaremos métodos basados en suposiciones menos restrictivas.

Intervalo de confianza para μ con σ conocida

Para una probabilidad igual a 1-α tenemos:

$$P(-z < \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < z) = 1 - \alpha$$

Si despejamos μ queda:

$$P(\bar{x} - z\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

Este intervalo es aleatorio porque los dos puntos extremos del intervalo tienen que ver con una variable aleatoria (la media muestral). Se puede interpretar como "hay una probabilidad de 0.95 de que el intervalo aleatorio incluya o cubra al valor verdadero de μ".

Intervalo de confianza para μ con σ conocida

Si después de tomar una muestra aleatoria se calcula la media muestral y luego se sustituye en la fórmula anterior, el intervalo dejar de ser aleatorio y se llama intervalo de confianza del $(1-\alpha)\%$ para μ. Este intervalo de confianza de puede expresar en forma resumida como:

$$LC = \bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$

Se puede interpretar como "con una confianza de 1- α el intervalo incluirá o cubrirá al valor verdadero de μ".

Analice las diferencias en la interpretación con respecto a un intervalo aleatorio.

El nivel de confianza no es una declaración acerca de algún intervalo particular, sino que indica lo que sucedería si se construyera un número muy grande de intervalos en iguales condiciones.

Intervalo de confianza para μ con σ desconocida

Cuando σ es desconocida el estadístico planteado en el punto anterior no puede ser aplicado, ya que existirían dos parámetros desconocidos (μ y σ) . El resultado en el que se basan las inferencias introduce una nueva familia de distribuciones de probabilidad llamada familia de distribuciones ***t.***

Bajo estas condiciones se usa el estadístico:

$$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

que tiene distribución t con n-1 grados de libertad (gl).

El intervalo de confianza para μ puede expresar en forma resumida como:

$$LC = \bar{x} \pm t.\frac{S}{\sqrt{n}}$$

Para poder utilizar el estadístico t es necesario que la distribución de la población sea Normal o aproximadamente Normal.

Intervalo de confianza para p

Sea p la proporción de "éxitos" en una población. Recuerde que si n es grande X tiene aproximadamente una distribución normal.

$$\lim_{n\to\infty} P\left(\frac{\hat{p}-p}{\sqrt{\frac{p.q}{n}}} < z\right) = F(z)$$

Siempre que $np \geq 5$ y $nq \geq 5$

Para una probabilidad igual a 1-α tenemos: $P\left(-z < \frac{\hat{p}-p}{\sqrt{\frac{p.q}{n}}} < z\right) = 1-\alpha$

Si despejamos p queda: $P(\hat{p} - z\sqrt{\frac{p.q}{n}} < p < \hat{p} + z\sqrt{\frac{p.q}{n}}) = 1-\alpha$

El problema es que los límites de confianza contienen al parámetro desconocido. Una buena aproximación puede lograrse:

$$LC = \hat{p} \pm z\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Intervalo de confianza para σ^2

Sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria de una distribución normal con parámetros μ y σ. Entonces la v.a.

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

tiene distribución Chi cuadrado con n-1 grados de libertad.

El intervalo para este caso será:

$$P(\chi_1^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_2^2) = 1-\alpha$$

Lo que en forma sintética resulta:

$$\frac{(n-1)S^2}{\chi_2^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_1^2}$$

Determinación de la muestra con reemplazo

- Estimación de la media

$$P(-z < \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < z) = 1 - \alpha$$

siendo $|\bar{x} - \mu|$ el **error de estimación (e).** Al despejar resulta:

$$P(|e| < z.\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

Existe una probabilidad igual a $1 - \alpha$ que el error de estimación sea como máximo $z.\frac{\sigma}{\sqrt{n}}$.

Analicemos:

- ¿qué elemento refleja el riesgo de equivocarse?
- ¿cómo incide la dispersión?
- Si aumentamos *n* ¿qué pasa con el error máximo?
- ¿Cómo influye **e** en el tamaño muestral?

$$|e| = z.\frac{\sigma}{\sqrt{n}}$$

$$n = \frac{z^2\sigma^2}{e^2}$$

Determinación de la muestra con reemplazo

- Estimación de la proporción

$$\left(-z < \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} < z\right) = 1 - \alpha$$

siendo $|\hat{p} - p|$ el **error de estimación (e).** Al despejar resulta:

$$P(|e| < z.\sqrt{\frac{\hat{p}\hat{q}}{n}}) = 1 - \alpha$$

$$|e| = z.\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$n = \frac{z^2\hat{p}\hat{q}}{e^2}$$

SOCIEDAD ECUATORIANA ESTADÍSTICA

2. Pruebas de hipótesis

Contrastes de hipótesis

Lo que afirmamos **versus** lo que vemos. La Ho la tomaremos como cierta hasta que no se demuestre lo contrario. Ej: juicio.

Contrastes de hipótesis

Lo que afirmamos **versus** lo que vemos. La Ho la tomaremos como cierta hasta que no se demuestre lo contrario. Ej: juicio.

Las afirmaciones en Ho son siempre **conservadoras**:

- Un mundo indiferenciado;

Contrastes de hipótesis

Lo que afirmamos **versus** lo que vemos. La Ho la tomaremos como cierta hasta que no se demuestre lo contrario. Ej: juicio.

Las afirmaciones en Ho son siempre **conservadoras**:

- Un mundo indiferenciado;
- Las cosas son iguales;

Contrastes de hipótesis

Lo que afirmamos **versus** lo que vemos. La Ho la tomaremos como cierta hasta que no se demuestre lo contrario. Ej: juicio.

Las afirmaciones en Ho son siempre **conservadoras**:

- Un mundo indiferenciado;
- Las cosas son iguales;
- Variables no relacionadas;

Contrastes de hipótesis

Lo que afirmamos **versus** lo que vemos. La Ho la tomaremos como cierta hasta que no se demuestre lo contrario. Ej: juicio.

Las afirmaciones en Ho son siempre **conservadoras**:

- Un mundo indiferenciado;
- Las cosas son iguales;
- Variables no relacionadas;
- Estado primitivo en donde todo está por demostrarse.

Contrastes de hipótesis

Lo que afirmamos **versus** lo que vemos. La Ho la tomaremos como cierta hasta que no se demuestre lo contrario. Ej: juicio.

Las afirmaciones en Ho son siempre **conservadoras**:

- Un mundo indiferenciado;
- Las cosas son iguales;
- Variables no relacionadas;
- Estado primitivo en donde todo está por demostrarse.

Contrastes de hipótesis

Lo que afirmamos **versus** lo que vemos. La Ho la tomaremos como cierta hasta que no se demuestre lo contrario. Ej: juicio.

Las afirmaciones en Ho son siempre **conservadoras**:

- Un mundo indiferenciado;
- Las cosas son iguales;
- Variables no relacionadas;
- Estado primitivo en donde todo está por demostrarse.

Probabilidades	Rechazar H_0	no Rechazar H_0
H_0 verdadera	α	$1-\alpha$
H_0 falsa	$1-\beta$	β

$$\textit{p-value} = \mathbb{P}(\text{resultado tan extremo o más} \mid H_0)$$

Contrastes de hipótesis

Lo que afirmamos **versus** lo que vemos. La Ho la tomaremos como cierta hasta que no se demuestre lo contrario. Ej: juicio.

Las afirmaciones en Ho son siempre **conservadoras**:

- Un mundo indiferenciado;
- Las cosas son iguales;
- Variables no relacionadas;
- Estado primitivo en donde todo está por demostrarse.

Probabilidades	Rechazar H_0	no Rechazar H_0
H_0 verdadera	α	$1-\alpha$
H_0 falsa	$1-\beta$	β

$$\textit{p-value} = \mathbb{P}(\text{resultado tan extremo o más} \mid H_0)$$

La distribución del estadístico bajo la hipótesis nula es el valor del estadístico calculado a todas las muestras posibles de tamaño n, si fuera cierta la hipótesis nula. Imagen teórica de como deberían ser las cosas si fuera cierta la hipótesis.

Pasos para realizar una prueba de hipótesis

- Plantear la hipótesis nula y la alternativa (puede ser unilateral o bilateral) y elegir el nivel de significación (α).

Pasos para realizar una prueba de hipótesis

- Plantear la hipótesis nula y la alternativa (puede ser unilateral o bilateral) y elegir el nivel de significación (α).
- De acuerdo al (o los) parámetro (s) de que se trate, elegir un buen estimador de ese (esos) parámetro (s).

Pasos para realizar una prueba de hipótesis

- Plantear la hipótesis nula y la alternativa (puede ser unilateral o bilateral) y elegir el nivel de significación (α).
- De acuerdo al (o los) parámetro (s) de que se trate, elegir un buen estimador de ese (esos) parámetro (s).
- De acuerdo al conocimiento que se tenga de la población, al tamaño de muestra que se va a tomar y al estimador seleccionado, elegir un "estadístico" adecuado que tendrá una distribución conocida bajo el supuesto de que la hipótesis nula es verdadera (recordar que un estadístico es una función del estimador y del parámetro).

Pasos para realizar una prueba de hipótesis

- Plantear la hipótesis nula y la alternativa (puede ser unilateral o bilateral) y elegir el nivel de significación (α).
- De acuerdo al (o los) parámetro (s) de que se trate, elegir un buen estimador de ese (esos) parámetro (s).
- De acuerdo al conocimiento que se tenga de la población, al tamaño de muestra que se va a tomar y al estimador seleccionado, elegir un "estadístico" adecuado que tendrá una distribución conocida bajo el supuesto de que la hipótesis nula es verdadera (recordar que un estadístico es una función del estimador y del parámetro).
- Con el estimador o con el estadístico (es indistinto) y la distribución de probabilidad adecuada, fijar el o los puntos críticos (según la prueba sea unilateral o bilateral respectivamente), y determinar las zonas de rechazo y no rechazo de la hipótesis nula.

Pasos para realizar una prueba de hipótesis

- Plantear la hipótesis nula y la alternativa (puede ser unilateral o bilateral) y elegir el nivel de significación (α).
- De acuerdo al (o los) parámetro (s) de que se trate, elegir un buen estimador de ese (esos) parámetro (s).
- De acuerdo al conocimiento que se tenga de la población, al tamaño de muestra que se va a tomar y al estimador seleccionado, elegir un "estadístico" adecuado que tendrá una distribución conocida bajo el supuesto de que la hipótesis nula es verdadera (recordar que un estadístico es una función del estimador y del parámetro).
- Con el estimador o con el estadístico (es indistinto) y la distribución de probabilidad adecuada, fijar el o los puntos críticos (según la prueba sea unilateral o bilateral respectivamente), y determinar las zonas de rechazo y no rechazo de la hipótesis nula.
- Tomar la muestra, calcular el estimador (o el estadístico) y determinar en qué zona cae ese valor.

Pasos para realizar una prueba de hipótesis

- Plantear la hipótesis nula y la alternativa (puede ser unilateral o bilateral) y elegir el nivel de significación (α).
- De acuerdo al (o los) parámetro (s) de que se trate, elegir un buen estimador de ese (esos) parámetro (s).
- De acuerdo al conocimiento que se tenga de la población, al tamaño de muestra que se va a tomar y al estimador seleccionado, elegir un "estadístico" adecuado que tendrá una distribución conocida bajo el supuesto de que la hipótesis nula es verdadera (recordar que un estadístico es una función del estimador y del parámetro).
- Con el estimador o con el estadístico (es indistinto) y la distribución de probabilidad adecuada, fijar el o los puntos críticos (según la prueba sea unilateral o bilateral respectivamente), y determinar las zonas de rechazo y no rechazo de la hipótesis nula.
- Tomar la muestra, calcular el estimador (o el estadístico) y determinar en qué zona cae ese valor.
- De acuerdo a lo observado en el punto 5, decidir si se rechaza o no la hipótesis nula y traducir esa decisión a los términos del problema.

Estadísticos para realizar pruebas de hipótesis

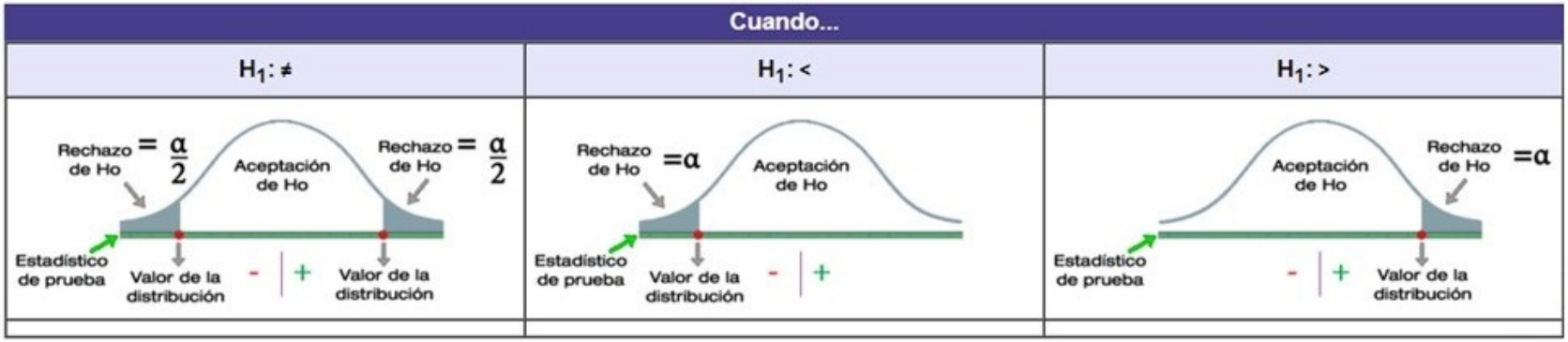
Parámetro involucrado en la prueba	Descripción de la población	Estadístico de test o medida de discrepancia
Media poblacional μ	Población N(μ, σ), σ conocido o n > 30	$z = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}}$
Media poblacional μ	Población N(μ, σ), σ desconocido	$t_{n-1} = \frac{\overline{x} - \mu_0}{s / \sqrt{n}}$
Varianza poblacional σ^2	Población normal	$\chi_{n-1}^{\ 2} = \frac{(n-1)s^2}{\sigma_0^{\ 2}}$
Proporción poblacional P	Población dicotómica, nP > 5 y nQ > 5	$z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}}$
Diferencia de medias poblacionales μ_1 - μ_2	Ambas poblaciones normales, σ_1 y σ_2 conocidas	$z = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\left(\frac{\sigma_1^2}{n_1}\right) + \left(\frac{\sigma_2^2}{n_2}\right)}}$

Estadísticos para realizar pruebas de hipótesis

Parámetro involucrado en la prueba	Descripción de la población	Estadístico de test o medida de discrepancia
Diferencia de medias poblacionales $\mu_1 - \mu_2$	Ambas poblaciones normales, σ_1 y σ_2 desconocidas pero supuestas iguales	$t_{n_1+n_2-2} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$
Diferencia de medias poblacionales (muestras dependientes)	Población normal	$t_{n-1} = \frac{\bar{d}}{s_d / \sqrt{n}}$
Diferencia de proporciones poblacionales $P_1 - P_2$	Poblaciones dicotómicas, $nP_1 > 5$, $nP_2 > 5$, $nQ_1 > 5$, $nQ_2 > 5$	$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\overline{pq}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$
Cociente de varianzas poblacionales	Ambas poblaciones normales	$F_{(n_1-1),(n_2-1)} = \frac{s_1^2}{s_2^2}$

Pruebas de una y dos colas

- Se desconoce la dirección en que se sospecha la falsedad de la hipótesis nula y se especifica la hipótesis alternativa como $P1 \neq P2$ se dice entonces que la prueba de hipótesis es bilateral.
- Se conoce de antemano que la hipótesis nula, si se rechaza, tiene una dirección determinada y, en ese caso se plantea que $P1 > P2$ o que $P1 < P2$. La prueba es entonces unilateral (derecha en el primer caso, izquierda en el segundo).



Test de hipótesis en R

```
X<-rnorm(100)
t.test(X)

##
##  One Sample t-test
##
## data:  X
## t = -2, df = 99, p-value = 0.07
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.395  0.015
## sample estimates:
## mean of x
##     -0.19
```

Comprobación de t, p-valor e IC

```
(t<-(mean(X)-0)/(sd(X)/sqrt(length(X))))
```

```
## [1] -1.8
```

```
(p_valor<-2*pt(-abs(t), df=length(X)-1))
```

```
## [1] 0.069
```

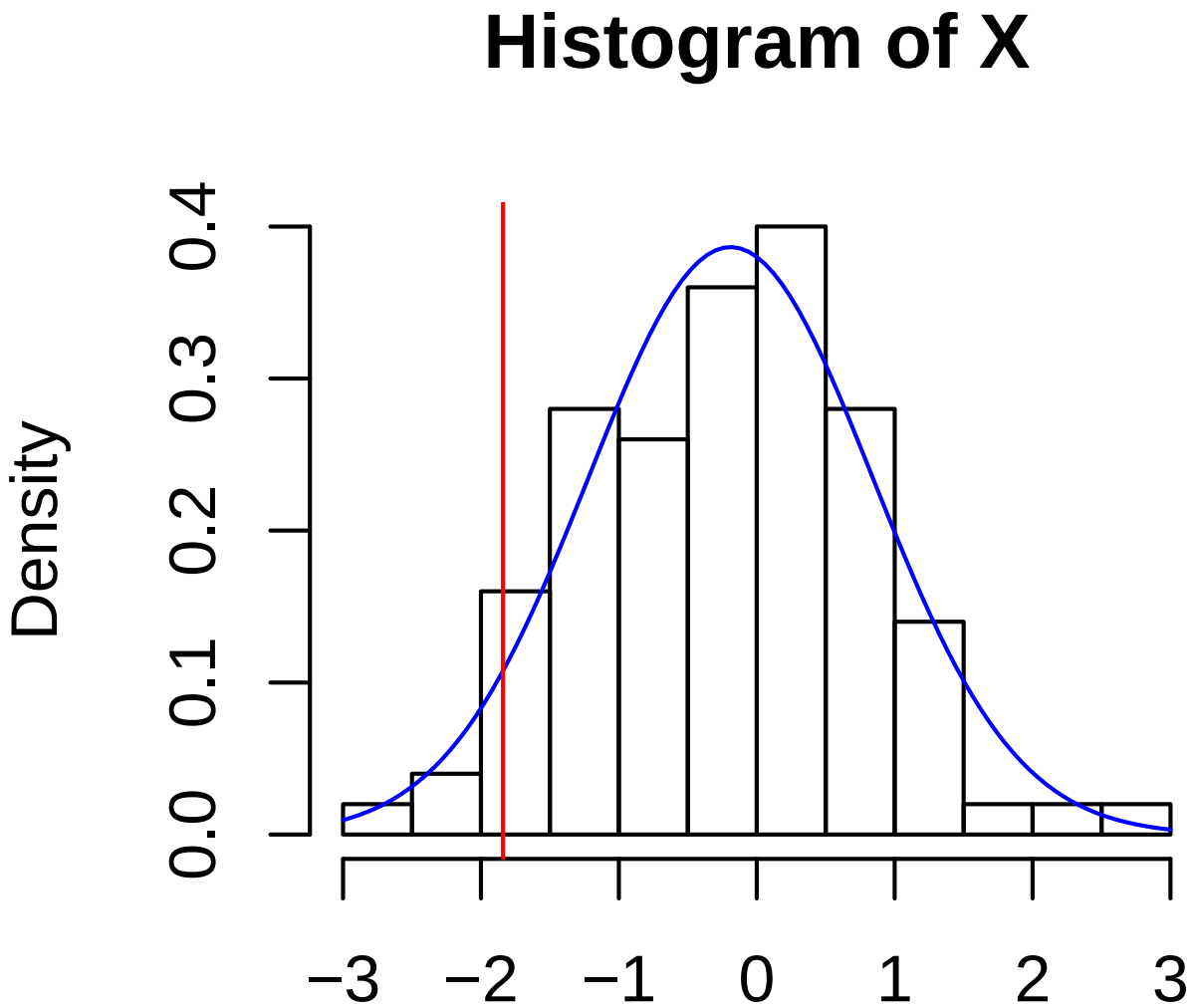
```
#Intervalo de confianza
mean(X)-qt(0.975, 99)*(sd(X)/sqrt(length(X)))
```

```
## [1] -0.39
```

```
mean(X)+qt(0.975, 99)*(sd(X)/sqrt(length(X)))
```

```
## [1] 0.015
```

Test de hipótesis en R



Test de hipótesis en R

```
X<-rnorm(100, 1, 1)
t.test(X)

##
##  One Sample t-test
##
## data:  X
## t = 10, df = 99, p-value <0.0000000000000002
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.85 1.26
## sample estimates:
## mean of x
##       1.1
```

SOCIEDAD ECUATORIANA ESTADÍSTICA

Gracias!!!