CADA HECHO DE TU VIDA *Cuenta*

INEC Buenas cifras, **mejores vidas**

PRUEBAS DE SIGNIFICANCIA ESTADÍSTICA EN ENCUESTAS DE HOGARES

Junio, 2020

Dirección de Infraestructura Estadística y Muestreo
Dirección de Innovación en Métricas y Metodologías
Dirección de Estudios y Análisis de la Información

Pruebas de significancia estadística
En encuestas de hogares

2020

Quito - Ecuador, 2020

Direcciones
Dirección de Infraestructura Estadística y Muestreo
Dirección de Innovación en Métricas y Metodologías
Dirección de Estudios y Análisis de la Información

Unidades
Gestión de Diseño Muestral
Unidad de Innovación en Métricas y Metodologías
Unidad de Estudios y Análisis de la Información

Elaborado por:
William Constante
Carmen Granda
Cristina Martínez
Diego Villacreses
Diana Zambonino

Revisado por:
Christian Garcés
Lorena Moreno
María Isabel García

Aprobado por:
Markus Nabernegg

TABLA DE CONTENIDO

Lista de figuras

Lista de tablas

1. INTRODUCCIÓN

El Instituto Nacional de Estadística y Censos (INEC) es la institución pública que coordina, norma y evalúa la producción de la información estadística oficial proveniente del Sistema Estadístico Nacional, con el propósito de entregar a la sociedad y al Estado información de calidad, pertinente, veraz y oportuna, que contribuya al desarrollo nacional.

El INEC realiza varias operaciones estadísticas por muestreo, citando por ejemplo, la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU), la Encuesta Nacional Multipropósito de Hogares (Encuesta Multipropósito), la Encuesta Nacional de Salud, Salud Reproductiva y Nutrición (ENSANUT), Encuesta Nacional de Relaciones Familiares y Violencia de Género contra las Mujeres, etc.

Estas encuestas dirigidas a hogares, producen indicadores cada cierto periodo de tiempo, los cuales muestran la realidad de la sociedad ecuatoriana, por lo que se vuelve necesario que su análisis sea acompañado por pruebas de significancia estadística que permitan identificar si la diferencia en el valor de un indicador, por ejemplo de un trimestre respecto a otro[1], es estadísticamente significativa o no; este procedimiento se denomina ***pruebas de hipótesis***.

Al hablar de una muestra y no de una población, los estimadores están asociados a errores, tanto muestrales como no muestrales, los cuales conllevan a que las diferencias no sigan el mismo proceso que aquellas entre parámetros poblacionales. Por tanto, se habla de diferencias estadísticamente significativas, o no, de acuerdo al resultado de la prueba de hipótesis.

En este sentido, las pruebas de hipótesis son importantes en el marco de una encuesta probabilística para evitar asumir diferencias entre indicadores cuando no necesariamente existen. La conclusión a la que se llega a partir de la prueba de hipótesis está asociada a un grado de confiabilidad que será definido previo a la realización de la prueba.

Este documento explica los fundamentos teóricos para la realización de pruebas de hipótesis a partir de encuestas de hogares, así como su aplicación en algunos de sus principales indicadores.

2. FUNDAMENTOS TEÓRICOS DE LAS PRUEBAS DE HIPÓTESIS

A partir de una encuesta probabilística, se pueden obtener dos tipos de estimaciones: ***puntual*** y ***por intervalos***. La primera se refiere a "una estimación de un parámetro poblacional que se da mediante un solo número" (Spiegel y Stephens, 2009: pg. 228); mientras que "una estimación de un parámetro poblacional que se da mediante dos números, entre los cuales se considera que debe estar el parámetro en cuestión, se le llama estimación por intervalo del parámetro en cuestión" (Spiegel y Stephens, 2009: pg. 228).

Se puede obtener una primera intuición para saber si una diferencia entre estimaciones puntuales es estadísticamente significativa, o no, al revisar si los intervalos de confianza de los estimadores que se están comparando se traslapan o no. Sin

1 Siempre y cuando sean trimestres comparables. Por ejemplo, junio con junio o diciembre con diciembre, dado que si se comparan meses distintos, las diferencias podrían darse por efecto de estacionalidad.

embargo, esto no es evidencia suficiente y se requiere de una prueba formal de significancia estadística en base a la información muestral, llamadas pruebas de hipótesis.

Cuando se trata de tomar una decisión sobre la población, se hacen suposiciones sobre la misma. De acuerdo a Spiegel y Stephens (2009), a estas suposiciones, que pueden ser ciertas o no, se les denomina ***hipótesis estadísticas.*** Las dos hipótesis que se plantean y que se deben verificar son:

- ***Hipótesis nula (H_0):*** esta hipótesis se plantea con la finalidad de anularla. Es una afirmación que no se rechaza a menos que los datos muestrales proporcionen evidencia suficiente para decir que es falsa. Como señala Webster (2000) nunca se puede "aceptar" la hipótesis nula como verdadera; en tal caso se habla de un no rechazo de esta hipótesis. Por ejemplo, si la hipótesis nula es que no hay diferencia entre dos estimadores, quiere decir que "cualquier diferencia que se observe se debe sólo a las fluctuaciones del muestreo de una misma población" (Spiegel y Stephens, 2009: pg. 245).
- ***Hipótesis alternativa (H_1):*** se refiere a cualquier hipótesis que difiera de la nula. Es una afirmación que se acepta si los datos muestrales proporcionan evidencia suficiente de que la hipótesis nula es falsa. Se le conoce también como la hipótesis de investigación.

Después de plantear la hipótesis nula y la alternativa, se selecciona un ***nivel de significancia,*** el cual se refiere a la probabilidad de rechazar la hipótesis nula cuando es verdadera; a lo que se le denomina un ***error de tipo I***. Esta probabilidad generalmente se denota como $\boldsymbol{\alpha}$ ***(alfa)***, que puede ser por ejemplo del 5%. Esto significa que existen cinco posibilidades en 100 de rechazar una H_0 que debería ser aceptada; es decir que existe un 95% de confianza de haber tomado la decisión correcta (Spiegel y Stephens, 2009).

Existe también el ***error de tipo II***, mediante el cual no se rechaza, es decir, se acepta una hipótesis nula que debería rechazarse. La probabilidad de este tipo de error se denota como **β *(beta)***. Se debe considerar que no se puede asumir que $\alpha + \beta = 1$. Idealmente, deberían minimizarse los dos tipos de errores, pero para cualquier tamaño de muestra, al tratar de disminuir el uno, incrementará el otro. Si rechazar una hipótesis verdadera (error tipo I) es más serio que no rechazar una hipótesis falsa, se debería seleccionar un α bajo (1% o 5%); caso contrario, α podría ser más alto (Webster, 2000). La única manera de reducir ambos errores es agrandando la muestra, lo cual no siempre es posible (Spiegel y Stephens, 2009).

En la práctica, se establece el nivel α; y para disminuir el error β, se incrementa el número de observaciones en la muestra, pues así se acortan los límites/intervalos de confianza respecto a la hipótesis planteada. La figura 1 presenta los dos tipos de error.

Figura 1. Tipos de errores

		Decisión del investigador	
		Se acepta H0	**Se rechaza H0**
Hipótesis nula	**H0 es verdadera**	Decisión correcta	Error tipo I
	H0 es falsa	Error tipo II	Decisión correcta

Fuente: (Levin y Rubin, 2004)

Con el fin de probar la hipótesis nula contra la alternativa, es necesario elegir un ***estadístico de prueba*** y un ***valor crítico***. Dado que se asume que la distribución de referencia de los estadísticos es *normal*[2], el estadístico de prueba es el estadístico **z,** que se calcula de la siguiente manera[3]:

$$z = \frac{p_t - p_{t-1}}{ee(p_t - p_{t-1})} = \frac{p_t - p_{t-1}}{\sqrt{v(p_{t-1}) + v(p_t)}},$$

Dónde:

p_{t-1} es la estimación del indicador de interés en el trimestre t del año anterior.
p_t es la estimación del indicador de interés en el trimestre t del año actual.
$v(p_t)$ es la estimación de la varianza[4] de p_t
$v(p_{t-1})$ es la estimación de la varianza de p_{t-1}

Los valores críticos se refieren a los valores de la distribución de probabilidades que están asociados al nivel de significancia dado por el investigador, en este caso, la distribución normal estándar (aquel valor del estadístico de prueba a partir del cual se rechaza la hipótesis nula). Estos valores dependen tanto del nivel de significancia α**,** como del ***tipo de prueba de hipótesis*** *(prueba de dos colas o prueba de una cola)*.

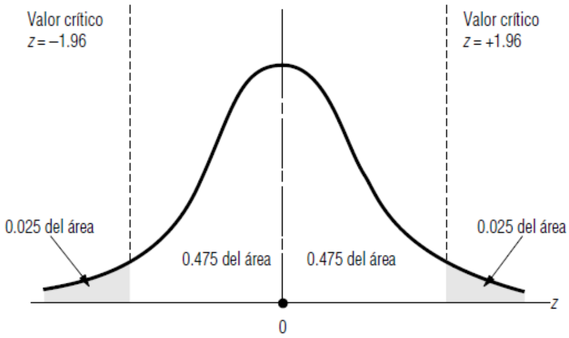
a) Prueba de dos colas: La hipótesis nula se formula en términos de igualdad estadística entre dos indicadores, y por tanto la hipótesis alternativa se refiere a que los indicadores son estadísticamente diferentes. Las hipótesis se plantean de la siguiente manera

$$H_0: p_t = p_{t-1}$$
$$H_1: p_t \neq p_{t-1}$$

Esta prueba de hipótesis puede ilustrarse de la siguiente manera:

2 Se asume normalidad dado que la muestra es lo suficientemente grande (Ley de los Grandes Números).
3 El estadístico de prueba se calcula de esta manera cuando se asume independencia entre las muestras a partir de las cuales se obtienen los indicadores que se están comparando.
4 La varianza de los dos estimadores se calcula a través de la técnica del último conglomerado, a partir de la cual sólo se toma en cuenta la varianza de los estimadores en la primera etapa, suponiendo que el muestreo en esta etapa fue realizado con reemplazo. Se usa esta técnica principalmente por dos razones. Primero, porque la primera etapa es la que proporciona el mayor porcentaje de la varianza total. En segundo lugar, resulta ser una técnica deseada por los investigadores puesto que utiliza directamente los factores de expansión que son publicados por los institutos nacionales de estadística y no requiere un conocimiento previo de las probabilidades de inclusión de la primera ni de la segunda etapa (Gutiérrez, 2018).

Figura 2. Prueba de dos colas



Fuente: (Levin y Rubin, 2004)

Se puede observar que, con un nivel de significancia de 5%, los valores críticos son de -1,96 y 1,96. Cualquier valor de **z** que sea mayor a 1,96 o menor a -1,96 indica que se debe rechazar la hipótesis nula, por lo que al área sombreada se denomina ***área de rechazo,*** y el resto es el ***área de aceptación (no rechazo)***. Si z cae en la zona de aceptación, se concluye con un 95% de confianza que los dos estimadores son estadísticamente iguales; es decir que ***no existe una diferencia estadísticamente significativa***. Es importante aclarar que con esta conclusión no se asevera que no existe diferencia entre los estimadores, si no que no existe evidencia suficiente para concluir que son diferentes (Altman y Bland, 1995). "La única forma de probar una hipótesis nula es conocer el parámetro, y eso no es posible cuando se trata de muestreo. Así, se acepta la hipótesis nula y se considera cierta, simplemente porque no podemos encontrar evidencia para rechazarla" (Levin y Rubin, 2004: pg. 330).

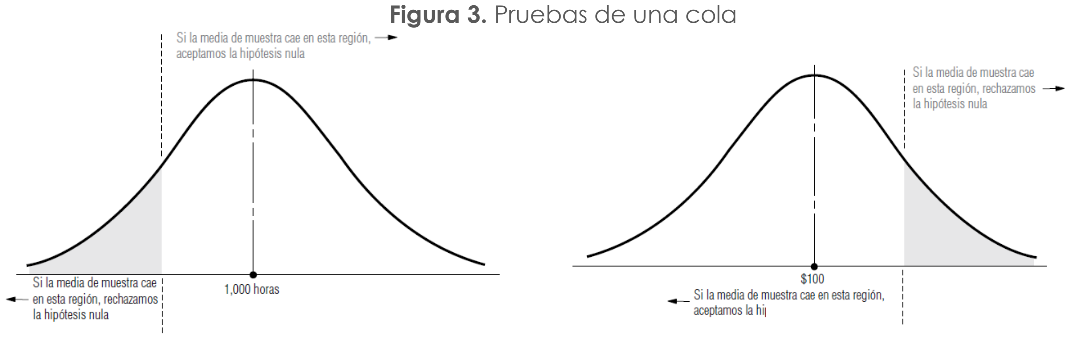
Por otro lado, si z estuviera en la zona sombreada, se rechaza la hipótesis nula y se acepta la alternativa. Es decir que ***existe diferencia estadísticamente significativa.***

b) Prueba de una cola: la hipótesis nula se formula con una relación de "mayor o igual que" o "menor o igual que". Por ejemplo:

$$\mathrm{H_0}: \mathrm{p_t} \geq \mathrm{p_{t-1}} \qquad\qquad \mathrm{H_0}: \mathrm{p_t} \leq \mathrm{p_{t-1}}$$

$$\mathrm{H_1}: \mathrm{p_t} < \mathrm{p_{t-1}} \qquad\qquad \mathrm{H_1}: \mathrm{p_t} > \mathrm{p_{t-1}}$$

Figura 3. Pruebas de una cola



Fuente: (Levin y Rubin, 2004)

En este caso, y dependiendo también del valor de α, existe un solo valor crítico que definirá la zona de rechazo.

Una vez que están definidos los valores críticos, es posible determinar la regla de decisión para rechazar o aceptar la hipótesis nula. En el caso de la prueba de dos colas, y con un nivel de significancia de 5%, la regla de decisión es la siguiente: ***"No se rechaza la hipótesis nula si -1,96 ≤ z ≤ 1,96; y se rechaza si z<-1,96 o z>1,96".***

Otra manera de evaluar significancia estadística es a través del p-valor. Éste se refiere al valor del nivel de significancia (α) más bajo al cual se rechaza la hipótesis nula (Webster, 2000). Es decir que para un α = 5%, si se obtiene un *p-valor* = 0,02, dado que α > p-valor, se rechaza la hipótesis nula. ***Para un nivel de confianza del 95%, no se rechaza la hipótesis nula si p-valor > 0,05.***

Finalmente, con la regla de decisión y el estadístico de prueba, y/o el p-valor, se puede concluir e interpretar los resultados sobre diferencias estadísticamente significativas como se indicó anteriormente.

3. EJEMPLOS PRÁCTICOS TOMANDO COMO REFERENCIA LA ENCUESTA NACIONAL DE EMPLEO, DESEMPLEO Y SUBEMPLEO (ENEMDU)

Este proceso de pruebas de hipótesis se lleva a cabo con los indicadores que se obtienen de la ENEMDU con el propósito de hacer comparaciones estadísticas de la manera adecuada. Se asume el mismo procedimiento de la prueba de hipótesis (con el mismo estadístico de prueba) para los distintos tipos de indicadores (totales, medias, tasas, proporciones, etc.), así como los diferentes casos de comparación (entre periodos, entre grupos de interés en un mismo periodo, etc.)[5].

A continuación, se describen tres ejemplos prácticos en relación a los indicadores laborales. Se debe considerar que en el Anexo 1, se presenta la sintaxis y algunos conceptos fundamentales necesarios para el cálculo de los errores estándar de los indicadores del mercado laboral ecuatorianos y otros.

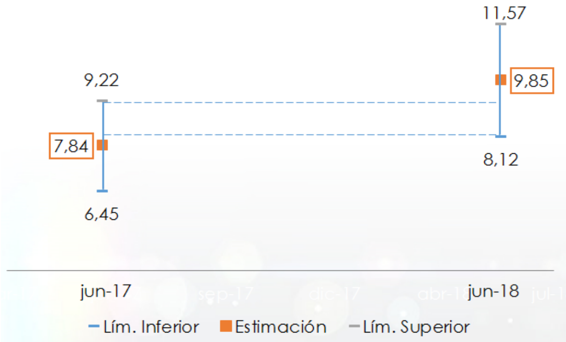
Ejemplos[6]

a) Ejemplo 1: Tasa de desempleo en Quito

Para junio del 2017, la estimación puntual de la tasa desempleo[7] en Quito fue de 7,84%, mientras que para el mismo mes del 2018, fue de 9,85%. Cada una de estas estimaciones está asociada a intervalos de confianza.

5 El cálculo de pruebas de hipótesis puede diferenciarse dependiendo del tipo de indicador o comparación que se realiza.
6 En los ejemplos que se describen a continuación, se asume independencia de las muestras de las cuales provienen los indicadores a ser comparadas.
7 Tasa de desempleo = número de desempleados / PEA

Figura 4. Tasa de desempleo en Quito



Fuente: ENEMDU, Junio-2017 y Junio-2018

Como se puede observar, los intervalos de ambas estimaciones se traslapan; lo que proporciona una primera intuición de que podría no existir una diferencia estadísticamente significativa. Sin embargo, la cuestión que interesa es tener una respuesta definitiva de sí o no, como lo señala Wooldridge (2010), por lo que se realiza la prueba de hipótesis con las siguientes hipótesis:

$$\mathrm{H_0: desempleo_{jun17} = desempleo_{jun18}} \rightarrow \text{no existe dif. estadísticamente significativa}$$
$$\mathrm{H_1: desempleo_{jun17} \neq desempleo_{jun18}} \rightarrow \text{existe dif. estadísticamente significativa}$$
$$\alpha = 0{,}05$$

Como se mencionó anteriormente, a partir de la técnica del último conglomerado se obtiene el error estándar para cada estimación.

Tabla 1. Errores estándar y varianzas de las tasas estimadas de desempleo en Quito

Periodo	Error estándar	Varianza
jun-17	0,0070	4,9519E-05
jun-18	0,0088	7,7591E-05

Fuente: ENEMDU, Junio-2017 y Junio-2018

Con esta información, se calcula el estadístico de prueba z:

$$\mathrm{z} = \frac{\mathrm{desempleo_{jun18} - desempleo_{jun17}}}{\mathrm{ee\left(desempleo_{jun18} - desempleo_{jun17}\right)}} = \frac{\mathrm{desempleo_{jun18} - desempleo_{jun17}}}{\sqrt{\mathrm{v\left(desempleo_{jun17}\right) + v\left(desempleo_{jun18}\right)}}}$$
$$= \frac{9{,}85 - 7.84}{\sqrt{4{,}9519 \times 10^{-05} + 7{,}7591 \times 10^{-05}}} = 1{,}7824$$

Dado que es una prueba de dos colas (por el signo de diferencia en la $\mathrm{H_1}$), los valores críticos son -1,96 y 1,96, por lo que z=1,7824 cae dentro de la zona de no rechazo de la hipótesis nula. Por lo tanto, la conclusión de la prueba es que ***no existe diferencia estadísticamente significativa; o que no existe evidencia suficiente para decir que la tasa de desempleo en Quito entre junio de 2017 y junio de 2018 es diferente[8].***

Tabla 2. Prueba de hipótesis de la diferencia entre jun17-jun18 de la tasa de desempleo en Quito

Jun-17		Jun-18	
Estimación	Error estándar	Estimación	Error estándar
7,84	0,0070	9,85	0,0088

8 Para el cálculo de los errores, referirse al Anexo 1, o al documento INEC (2018). Cálculo del error de muestreo y declaración de muestras complejas en la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU). Instituto Nacional de Estadística y Censos, Quito-Ecuador.

Prueba de hipótesis de la diferencia entre jun17-jun18				
Diferencia	**Valor z**	**Normal**	**p-valor**	**Conclusión**
2,01	1,7824	0,9627	0,0747	No rechazo H_0

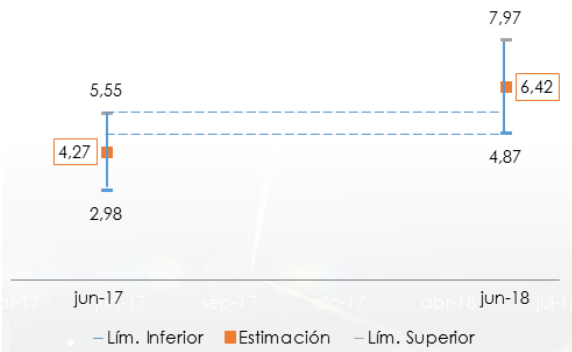
Fuente: ENEMDU, Junio-2017 y Junio-2018

De igual manera, en la tabla 2 se observa el p-valor, el cual es 0,0747; al ser mayor a 0,05, no rechazo la H_0; y se llega a la misma conclusión que con el estadístico de prueba.

b) Ejemplo 2: Tasa de desempleo en Machala

Para junio del 2017, la estimación puntual de la tasa desempleo en Machala fue de 4,27%, mientras que para el mismo mes del 2018, fue de 6,42%. Cada una de estas estimaciones está asociada a intervalos de confianza, como se muestra en el Gráfico 5.

Figura 5. Tasa de desempleo en Machala



Fuente: ENEMDU, Junio-2017 y Junio-2018

Como se puede observar, los intervalos de ambas estimaciones se traslapan; lo cual señala una primera intuición de que podría no existir una diferencia estadísticamente significativa. Sin embargo, se realiza la prueba de hipótesis para llegar a una conclusión certera, donde:

$$H_0\text{: desempleo}_{\text{jun17}} = \text{desempleo}_{\text{jun18}} \rightarrow \text{no existe dif. estadísticamente significativa}$$
$$H_1\text{: desempleo}_{\text{jun17}} \neq \text{desempleo}_{\text{jun18}} \rightarrow \text{existe dif. estadísticamente significativa}$$
$$\alpha = 0{,}05$$

A partir de la técnica del último conglomerado se obtiene el error estándar para cada estimación.

Tabla 3. Errores estándar y varianzas de las tasas estimadas de desempleo en Machala

Periodo	Error estándar	Varianza
jun-17	0,0065	4,2136E-05
jun-18	0,0079	6,2124E-05

Fuente: ENEMDU, Junio-2017 y Junio-2018

Con esta información, se calcula el estadístico de prueba z:

$$z = \frac{\text{desempleo}_{\text{jun18}} - \text{desempleo}_{\text{jun17}}}{\text{ee}(\text{desempleo}_{\text{jun18}} - \text{desempleo}_{\text{jun17}})} = \frac{\text{desempleo}_{\text{jun18}} - \text{desempleo}_{\text{jun17}}}{\sqrt{\text{v}(\text{desempleo}_{\text{jun17}}) + \text{v}(\text{desempleo}_{\text{jun18}})}}$$

$$= \frac{6{,}42 - 4{,}27}{\sqrt{4{,}2136 \times 10^{-05} + 6{,}2124 \times 10^{-05}}} = 2{,}1094$$

Dado que es un prueba de dos colas (por el signo de diferencia en la H_1), los valores críticos son -1,96 y 1,96, por lo que z=2,1094 cae dentro de la zona de rechazo de la hipótesis nula. Por lo tanto, acepto la hipótesis alternativa, y la conclusión de la prueba es que ***sí existe diferencia estadísticamente significativa***; o ***que existe evidencia suficiente para decir que la tasa de desempleo en Machala entre junio2017 y junio 2018 es diferente***. Como se observa, en este caso, la prueba de hipótesis contradice la intuición del cruce de intervalos de confianza, por lo que se corrobora la necesidad de realizar la prueba de hipótesis para concluir con certeza.

Tabla 4. Prueba de hipótesis de la diferencia entre jun17-jun18 de la tasa de desempleo en Machala

Jun-17		Jun-18	
Estimación	Error estándar	Estimación	Error estándar
4,27	0,0065	6,42	0,0079

Prueba de hipótesis de la diferencia entre jun17-jun18				
Diferencia	Valor z	Normal	Valor-p	Conclusión
2,15	2,1094	0,9825	0,0349	Rechazo H_0

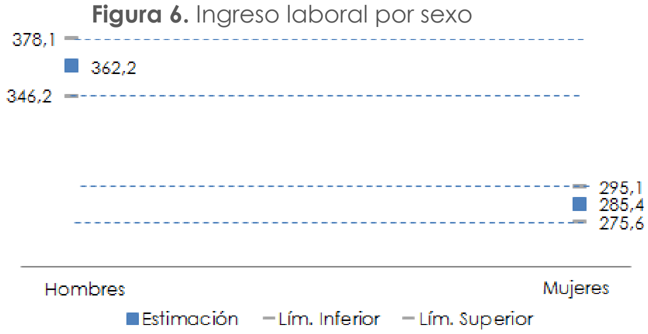
Fuente: ENEMDU, Junio-2017 y Junio-2018

De igual manera, como se observa en la tabla 4, el p-valor es de 0,0349; al ser menor a 0,05, se rechaza la H_0; y se llega a la misma conclusión que con el estadístico de prueba.

c) Ejemplo 3: Ingreso laboral por sexo

Como se mencionó anteriormente, otra aplicación de las pruebas de hipótesis en la ENEMDU es la ***comparación de estimadores de un mismo periodo***. En este caso se va a probar si existe, o no, diferencia estadísticamente significativa entre la media del ingreso de los hombres y el de las mujeres en junio del 2018 a nivel nacional.

Figura 6. Ingreso laboral por sexo



Fuente: ENEMDU, Junio-2018

Como se puede observar, los intervalos de ambas estimaciones no se traslapan; lo que proporciona una primera intuición de que podría existir una diferencia estadísticamente significativa. Sin embargo, se realiza la prueba de hipótesis para llegar a una conclusión certera, donde:

$$\mathrm{H_0: ingreso_{hombres} = ingreso_{mujeres} \rightarrow no\ existe\ dif.\ estadísticamente\ significativa}$$
$$\mathrm{H_1: ingreso_{hombres} \neq ingreso_{mujeres} \rightarrow existe\ dif.\ estadísticamente\ significativa}$$
$$\alpha = 0{,}05$$

A partir de la técnica del último conglomerado se obtiene el error estándar para cada estimación.

Tabla 5. Errores estándar y varianzas del ingreso laboral promedio estimado, por sexo

Ingreso laboral	Error estándar	Varianza
Ingreso laboral promedio hombres	8,1398	24,6944
Ingreso laboral promedio mujeres	4,9693	66,2565

Fuente: ENEMDU, Junio-2018

Con esta información, se calcula el estadístico de prueba z:

$$\mathrm{z} = \frac{\mathrm{ingreso_{hombres} - ingreso_{mujeres}}}{\mathrm{ee}\left(\mathrm{ingreso_{hombres} - ingreso_{mujeres}}\right)} = \frac{\mathrm{ingreso_{hombres} - ingreso_{mujeres}}}{\sqrt{\mathrm{v}(\mathrm{ingreso_{hombres}}) + \mathrm{v}(\mathrm{ingreso_{mujeres}})}}$$
$$= \frac{362{,}2 - 285{,}4}{\sqrt{24{,}6944 + 66{,}2565}} = 8{,}0530$$

Dado que es un prueba de dos colas (por el signo de diferencia en la $\mathrm{H_1}$), los valores críticos son -1,96 y 1,96, por lo que z=8,0530 cae dentro de la zona de rechazo de la hipótesis nula. Por lo tanto, se acepta la hipótesis alternativa, y la conclusión de la prueba es que ***sí existe diferencia estadísticamente significativa***; o que ***existe evidencia suficiente para decir que el ingreso laboral promedio entre hombres y mujeres en junio 2018 es diferente***.

Tabla 6. Prueba de hipótesis de la diferencia entre ingreso promedio de hombres y mujeres

Hombres		Mujeres	
Estimación	Error estándar	Estimación	Error estándar
362,2	8,1398	285,4	4,9693

Prueba de hipótesis de la diferencia				
Diferencia	Valor z	Normal	Valor-p	Conclusión
76,8	8,0532	1	8,8818E-16	Rechazo $\mathrm{H_0}$

Fuente: ENEMDU, Junio-2018

De igual manera, se observa el p-valor en la tabla 6, el cual es 8,88e-16; al ser menor a 0,05, rechazo la H_0; y se llega a la misma conclusión que con el estadístico de prueba.

Para llevar a cabo las pruebas de hipótesis, se operacionaliza este proceso en un software estadístico, dependiendo del cual variarán las especificaciones y la sintaxis para calcular los errores estándar como se menciona en el Anexo 1.

d) Ejemplo 4: Pruebas de hipótesis para los indicadores de pobreza y desigualdad

Para el cálculo de los indicadores de pobreza y desigualdad, se emplea el paquete DASP (Distributive Analysis Stata Package) para Stata, en su versión 2.3. Esta herramienta de libre distribución se emplea en el análisis de las condiciones de vida de la población. Fue desarrollado por: PEP (Partnership for Economic Policy), CIRPÉE (Centre Interuniversitaire sur le Risque, les Politiques Économiques et l'Emploi), UNDP (United Nations Development Programme) y el Banco Mundial.

Entre las utilidades del paquete se encuentra la estimación de estadísticas utilizadas para el análisis de la pobreza, la desigualdad, el bienestar social y la equidad, así como la estimación de las diferencias en dichas estadísticas, sus errores estándar e intervalos de confianza –tomando en cuenta el diseño de la encuesta–, realizar descomposiciones distributivas, entre otros.

- **Instalación del paquete**

Para instalar este paquete en Stata se usa la siguiente sintaxis:

```
set more off
net from http://dasp.ecn.ulaval.ca/modules/DASP_V2.3/dasp
net install dasp_p1, force
net install dasp_p2, force
net install dasp_p3, force
net install dasp_p4, force
```

- **Diseño de la muestra**

Para el uso correcto de los comandos del paquete, se debe declarar el diseño de muestra de la ENEMDU. Este es el mismo que se emplea en los indicadores de empleo:

```
svyset id_upm [iw=fexp], strata(plan_muestreo) vce(linearized) singleunit(certainty)
```

- **Variable utilizada en el cálculo**

La variable que da cuenta del estándar de vida en estos cálculos es el ingreso per cápita familiar (ipcf).

- **Sintaxis en STATA**

Los comandos detallados a continuación permiten comparar dos periodos (t y t-1) para obtener las estimaciones de los indicadores requeridos, la diferencia entre los periodos, el error estándar, el intervalo de confianza y el p-valor.

Indicador	Sintaxis
Pobreza y pobreza extrema (Incidencia, brecha y severidad)	difgt ipcf ipcf, alpha(\`a') cond1(ipcf!=. & \`x'==\`niv') cond2(ipcf!=. & \`x'==\`niv') pline1(\`pov') pline2(\`pov') file1(\`base1') file2(\`base2')
Gini	digini ipcf ipcf, cond1(ipcf!=. & \`x'==\`niv') cond2(ipcf!=. & \`x'==\`niv') file1(\`base1') file2(\`base2')

Theil	dientropy ipcf ipcf, cond1(ipcf!=. & `x'==`niv') cond2(ipcf!=. & `x'==`niv') file1(`base1') file2(`base2')
Atkinson	diatkinson ipcf ipcf, cond1(ipcf!=. & `x'==`niv') cond2(ipcf!=. & `x'==`niv') file1(`base1') file2(`base2') epsilon(`ep')

- Alpha: Se puede elegir el parámetro alpha, que puede ser 0, 1 o 2, dando cuenta del FGT 0, FGT 1 y FGT 2, respectivamente. Se usa la local (a) para hacer la estimación de los 3 casos en un bucle.
- Epsilon: Se puede elegir el parámetro épsilon, que puede ser 0.5, 1, 1.5 o 2. Se utiliza la local (ep) para hacer la estimación de los cuatro parámetros en un bucle.
- Condición: Se puede condicionar la estimación para determinados subgrupos poblacionales o categorías. Lo hacemos por área, dominio y a veces provincia[9].
- Errores estándar e intervalos de confianza: Se obtienen al 95% de confianza. Esto puede cambiarse si es requerido.

9 Esto va a depender del mes de levantamiento de información. En el mes de diciembre la representatividad de la encuesta incluye dominios provinciales.

ANEXO 1. SINTAXIS EN STATA, SPSS Y R PARA CÁLCULO DE ERRORES ESTÁNDAR

1. Sintaxis en STATA

Para llevar a cabo la prueba de hipótesis, se utiliza el comando *svyset*, mediante el cual se declara el diseño de la encuesta:

svyset id_upm [iw=fexp], strata(estrato) vce(linearized) singleunit(certainty)

Donde:

- id_upm es el identificador de la unidad primaria de muestreo[10]
- fexp es el factor de expansión
- strata(estrato) declara la variable que identifica el estrato, que en este caso es plan_muestreo, dado que esta variable es la concatenación de dominio, área y estrato socio-económico[11]
- vce(linearized) es para aplicar la linealización de Taylor al algoritmo de la varianza antes de usar la técnica del último conglomerado, que STATA usa por defecto
- singleunit(certainty) especifica cómo manejar los estratos con una unidad de muestreo. "certainty" hace que los estratos con conglomerados individuales sean tratados como unidades de certeza, es decir éstas unidades no contribuyen para el cálculo del error estándar

Una vez que se ha declarado el diseño de la encuesta, se calculan los estadísticos. Por ejemplo, para la ***tasa*** de desempleo en los distintos dominios en jun-18 se utiliza el comando svy: ratio, con el cual se obtienen las estimaciones por los diferentes dominios, los errores estándar y los intervalos de confianza.

```
svy: ratio desempleo/pea, over(dominio),
estat cv
estat effects, deff
```

A partir de los errores estándar obtenidos, y haciendo el mismo proceso para jun-17, se puede calcular el estadístico de prueba y el valor-p.

Para obtener una ***media*** se usa el comando svy: mean, con el cual se obtienen las estimaciones por los diferentes dominios, los errores estándar y los intervalos de confianza.

```
svy: mean ilareal if empleo==1, over(p02)
```

Con los errores estándar obtenidos para las medias que se desea comparar, se puede calcular el estadístico de prueba y el valor-p.

10 A partir del 2018, la UPM son conglomerados, antes eran sectores censales.

11 Por ejemplo, si plan_muestreo=1011, significa que la observación pertenece al área 1 (urbano), dominio 01 (Quito), estrato 1 (alto).

2. Sintaxis en SPSS

- Asistente de preparación del plan para el análisis[12].

```
CSPLAN ANALYSIS
  /PLAN FILE='C:\(dirección donde se encuentra el documento)\(nombre del plan.csaplan' [13]
  /PLANVARS ANALYSISWEIGHT=fexp
  /SRSESTIMATOR TYPE=WOR
  /PRINT PLAN
  /DESIGN STRATA=estrato CLUSTER=id_upm
  /ESTIMATOR TYPE=WR.
EXECUTE.
```

- Descriptivos de muestras complejas:

```
CSDESCRIPTIVES
  /PLAN FILE='C:\Users\(dirección donde se encuentra el documento)\plan1.csaplan' [14]
  /SUMMARY VARIABLES=tdeso[15]
  /SUBPOP TABLE=NDOMINIO DISPLAY=LAYERED
  /MEAN
  /STATISTICS SE CV DEFF CIN(95)
  /MISSING SCOPE=ANALYSIS CLASSMISSING=EXCLUDE.
EXECUTE.
```

3. Sintaxis en R

- Librerías necesarios:

```
library(foreign)
library(survey)
library(dplyr)
library(srvyr)
```

- Declaración del diseño muestral de la encuesta:

```
d1 <- enemdu %>% as_survey_design[16](ids = id_upm
                       strata = estrato,
                       weights = fexp,
                        nest = T)
options(survey.lonely.psu="certainty")
```

- Cálculo de estadísticos:

```
tabla1 <- d1 %>%
summarise(tdesem = survey_mean(tdeso, vartype=c("se","ci","cv"), na.rm=T, deff = T))
```

12 Aquí se declara el diseño muestral de la encuesta
13 Aquí se especifica la dirección donde se quiere guardar el plan
14 Aquí se especifica la dirección donde se quiere guardar el plan
15 tdeso: tasa de desempleo
16 Este comando se utiliza para especificar los estratos y los conglomerados que se presentaron en el diseño muestral de la encuesta.

```
tabla <- d1 %>%
group_by(NDOMINIO)%>%
summarise(tdesem = survey_mean(tdeso, vartype=c("se","ci","cv"), na.rm=T, deff = T))
```

REFERENCIAS

Abdelkrim, Araar y Duclos, Jean-Yves (2007). ***DASP: Distributive Analysis Stata Package***. PEP, World Bank, UNDP y Université Laval.

Abdelkrim, Araar y Duclos, Jean-Yves (2013). ***User Manual for Stata Package DASP: Version 2.3***. PEP, World Bank, UNDP y Université Laval.

Altman, D. y Bland, M. (1995). ***Statistics notes: Absence of evidence is not evidence of absence.*** BMJ; 311:485.

Gutiérrez, A. (2018). ***Revisión del diseño de muestreo y esquema de análisis de la Encuesta Nacional de Empleo, Desempleo y Subempleo-ENEMDU***. Misión de Asistencia Técnica. Quito.

INEGI. ***Encuesta Nacional de Ocupación y Empleo (ENOE). Pruebas de significancia. Pruebas de hipótesis***. Recuperado de: http://www.beta.inegi.org.mx/contenidos/proyectos/enchogares/regulares/enoe/doc/enoe_significancia.pdf

Levin, R. y Rubin. D. (2004). ***Estadística para administración y economía***, séptima edición. Pearson: México.

Spiegel, M. y Stephens, R. (2009). ***Estadística***, cuarta edición. McGrawHill: México D.F.

Webster, A. (2000). ***Estadística aplicada a los negocios y la economía***, tercera edición. McGrawHill: Bogotá.

Wooldridge, J. (2010). ***Introducción a la econometría: Un enfoque moderno***, cuarta edición. Cengage Learning: México D.F.


CADA
HECHO
DE TU
VIDA
Cuenta
@ecuadorencifras
INEC/Ecuador
@InecEcuador
INECEcuador
t.me/ecuadorencifras
INEC Ecuador