COMPARACION DEL PROMEDIO Y LA MEDIANA COMO ESTIMADORES DE LA MEDIA PARA MUESTRAS NORMALES DEPENDIENDO DEL TAMAÑO DE MUESTRA

COMPARATION THE AVERAGE AND THE MEDIAN AS ESTIMATORS FOR THE MEAN FOR NORMAL SAMPLES DEPENDING ON THE SIZE OF SAMPLE

José W. Camero Jiménez[1], Jahaziel G. Ponce Sánchez[2]

RESUMEN

Actualmente los métodos para estimar la media son los basados en el intervalo de confianza del promedio o media muestral. Este trabajo pretende ayudar a escoger el estimador (promedio o mediana) a usar dependiendo del tamaño de muestra. Para esto se han generado, vía simulación en excel, muestras con distribución normal y sus intervalos de confianza para ambos estimadores, y mediante pruebas de hipótesis para la diferencia de proporciones se demostrará que método es mejor dependiendo del tamaño de muestra.

Palabras clave.- Tamaño de muestra, Intervalo de confianza, Promedio, Mediana.

ABSTRACT

Currently the methods for estimating the mean are those based on the confidence interval of the average or sample mean. This paper aims to help you choose the estimator (average or median) to use depending on the sample size. For this we have generated, via simulation in EXCEL, samples with normal distribution and confidence intervals for both estimators, and by hypothesis tests for the difference of proportions show that method is better depending on the sample size.

Key words.- Sampling size, Confidence interval, Average, Median.

INTRODUCCIÓN

En la práctica, el método más usual para estimar un parámetro de tendencia central, es el intervalo de confianza de la media, ya que el estimador utilizado comúnmente es el promedio o media muestral, que es el estimador máximo verosímil y cuenta con la propiedad de ser insesgado. Sin embargo, el principal problema para el estimador es cuando las muestras son pequeñas, debido a que este estimador se ve afectado por datos atípicos que alteran su valor. De esta manera el intervalo de confianza para estimar al parámetro podría verse afectado y cometer grandes errores, es por eso que se le considera al promedio como un estimador no robusto. Un buen estimador robusto es la mediana, ya que este estimador no se ve afectado por datos atípicos y para muestras pequeñas. El problema que surge ahora es calcular un intervalo de confianza basado en la mediana. Sin embargo, trabajos como la realizada por Wodruff ayudan a calcular su intervalo de confianza manera indirecta ya que se basan en la función de distribución acumulada para poder realizarlo. El objetivo del trabajo es identificar el mejor estimador de la media para diferentes tamaños de muestra.

[1] Lic. en Estadística, Catedrático Escuela Profesional de Ingeniería Estadística (EPIES)-UNI, en Post Grado de la UNALM, [2]Egresado de Ingeniería Estadística EPIES - UNI, asistente de investigación de la EPIES – UNI.

INTERVALO DE CONFIANZA PARA LA MEDIANA

Comenzaremos con la definición de un cuantil, y la construcción de su intervalo de confianza. Definimos a una población U de tamaño N, la característica y_i se conoce para cada $i = 1, N$. De esta población se extrae una muestra probabilística s de tamaño n. Se define la función de distribución para cada t ($-\infty < t < \infty$):

$$F_y(t) = \frac{1}{N}\sum_{i\epsilon U} \Delta(t - y_i)$$

Con $\Delta(a)$=1si a>0 y $\Delta(a)$=0 en otro caso.

El cuantil $Q_y(\beta)$ (0< β<1) se define así por:

$$Q_y(\beta) = \inf\{t: F_y(t) \geq \beta\}$$

Para el caso de la mediana ($\beta = 0.5$), se define así:

$$Q_y(0.5) = \inf\{t: F_y(t) \geq 0.5\}$$

El problema surge para estimar $Q_y(0.5)$, es decir, la mediana, a partir de los datos de la muestra de tamaño "n". El procedimiento general se realiza de la siguiente forma: en primer lugar se estima la función de distribución $F_y(t)$ según la muestra y luego se calcula la mediana mediante $Q_y(0.5) = \hat{F}_y^{-1}(t)$ donde $\hat{F}_y^{-1}$es la inversa de la función de distribución muestral. Para ellos usaremos el aporte de Wodruff [1]:

Sean dos constantes d_1 y d_2, y para cada valor de $Q_y(\beta)$,

$$P\left\{d_1 \leq \hat{F}_y\left(Q_y(\beta)\right) \leq d_2\right\} \cong$$

$$P\left\{\hat{F}_y^{-1}(d_1) \leq Q_y(\beta) \leq \hat{F}_y^{-1}(d_2)\right\}$$

Así, se sigue que para cada constante d_1 y d_2 tales que:

$$P\left\{d_1 \leq \hat{F}_y\left(Q_y(\beta)\right) \leq d_2\right\} = 1 - \alpha$$

Un intervalo aproximado del $100(1-\alpha)\%$ de confianza para $Q_y(\beta)$ es:

$$\left[\hat{F}_y^{-1}(d_1), \hat{F}_y^{-1}(d_2)\right]$$

Si el tamaño de muestra n es suficientemente grande, entonces se considera $\hat{F}_y\left(Q_y(\beta)\right)$ aproximadamente normal, y se pueden elegir:

$$d_1 = \beta - Z_{\alpha/2}\left\{V(\hat{F}_y\left(Q_y(\beta)\right))\right\}^{1/2}$$
$$d_2 = \beta + Z_{\alpha/2}\left\{V(\hat{F}_y\left(Q_y(\beta)\right))\right\}^{1/2}$$

Donde $Z_{\alpha/2}$ denota la cola superior de área $1-\alpha/2$ de la distribución normal estándar. Para el caso de la mediana ($\beta = 0.5$), la ecuación se puede reducir a:

$$d_1 = \beta - Z_{\alpha/2}\left\{\frac{\beta(1-\beta)}{n}\right\}^{1/2}$$
$$d_2 = \beta + Z_{\alpha/2}\left\{\frac{\beta(1-\beta)}{n}\right\}^{1/2}$$

Ejemplo:

A continuación realizaremos un ejemplo para conocer cómo se aplica este intervalo de confianza para la mediana. Se tiene la siguiente muestra de tamaño $n = 30$.

2.463	-2.276	0.194	3.285	0.698
-0.348	5.464	0.555	2.325	8.912
4.661	3.153	-0.079	2.27	4.656
2.329	0.878	7.255	1.049	0.971
0.436	-0.432	5.039	6.682	2.501
2.783	5.749	5.291	5.527	1.076

Calcularemos su intervalo de confianza del 95% basado en la mediana. Sabemos que su intervalo de confianza viene dado por:

$$[\hat{F}_y^{-1}(d_1), \hat{F}_y^{-1}(d_2)]$$

Dónde:

$$d_1 = \beta - Z_{\alpha/2}\left\{\frac{\beta(1-\beta)}{n}\right\}^{1/2}$$
$$d_2 = \beta + Z_{\alpha/2}\left\{\frac{\beta(1-\beta)}{n}\right\}^{1/2}$$

Sabiendo esto, $\beta = 0.5$ y $Z_{\alpha/2} = 1.96$

$$d_1 = 0.5 - 1.96\left\{\frac{0.5(1-0.5)}{30}\right\}^{1/2} = 0.32108$$

$$d_1 = 0.5 + 1.96\left\{\frac{0.5(1-0.5)}{30}\right\}^{1/2} = 0.67892$$

Para poder calcular los valores del intervalo $[\hat{F}_y^{-1}(0.32108), \hat{F}_y^{-1}(0.67892)]$, recordemos que para determinar dado a, $\hat{F}_y^{-1}(a)$, el procedimiento es el habitual $\hat{F}_y^{-1}(a) = \inf\{y_i, i \in s: \hat{F}_y(y_i) \geq a\}$, por lo que se tiene que ordenar los datos. Ver tabla 1.

Tabla 1. *Sobre la tabla se realiza la siguiente inversión:*

DATOS	$\hat{F}_y(y_i)$	DATOS	$\hat{F}_y(y_i)$
-2.276	0.0333	2.463	0.5333
-0.432	0.0667	2.501	0.5667
-0.348	0.1000	2.783	0.6000
-0.079	0.1333	3.153	0.6333
0.194	0.1667	3.285	0.6667
0.436	0.2000	4.656	0.7000
0.555	0.2333	4.661	0.7333
0.698	0.2667	5.039	0.7667
0.878	0.3000	5.291	0.8000
0.971	0.3333	5.464	0.8333
1.049	0.3667	5.527	0.8667
1.076	0.4000	5.749	0.9000
2.27	0.4333	6.682	0.9333
2.325	0.4667	7.255	0.9667
2.329	0.5000	8.912	1.0000

Por lo que el intervalo de confianza basado en la mediana viene dado por:

$$\left[\hat{F}_y^{-1}(0.32108), \hat{F}_y^{-1}(0.67892)\right] = [0.971, 4.656]$$

INTERVALO DE CONFIANZA PARA LA MEDIA

Supongamos que disponemos de la siguiente información:

Una muestra aleatoria de tamaño n $(X_1, \ldots, X_n)$ extraída de una población normal $N(\mu, \sigma^2)$ con σ^2 desconocida.

El estimador puntual $\hat{\theta}$ del parámetro μ es el promedio muestral $\overline{X}$.

El estadístico pivote que usaremos en este caso, que será:

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$

Donde S es la desviación estándar muestral. Recordemos que T tiene distribución t de Student con $v = n - 1$ grados de libertad.

El nivel de confianza$(1 - \alpha)$establecido a priori por el investigador (los usuales son 0.95, 0.90, 0.99)

Dada la distribución del estadístico y el nivel de confianza, se tiene la siguiente igualdad probabilística de acuerdo a Mendehall et al [3]:

$$P\left(-t_{\alpha/2} \leq \frac{\overline{X} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2}\right) = 1 - \alpha$$

Donde $t_{\alpha/2}$ es el valor característico de la variable T de Student verificando que $P(T \geq t_{\alpha/2}) = \frac{\alpha}{2}$

La expresión anterior es equivalente a:

$$P\left(\overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

Por lo que el intervalo de confianza para la media μ con una probabilidad $(1 - \alpha)$ viene dado por:

$$\left[\overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}; \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right]$$

METODO

Procedimiento de comparación

La comparación se realizó mediante simulación en Macros de EXCEL teniendo como ejemplo lo desarrollado por Peró Cebollero et al [2].

Para ello se generó 5,000 muestras de tamaño variable "n" (20, 25, 30,…, 200) de una población con distribución normal con media $\mu = 0$ y varianza $\sigma^2 = 1$. Luego de esto, se calculó el intervalo de confianza de la mediana y media para cada muestra y se contabilizó la cantidad de intervalos que contienen al parámetro de interés (recordemos que para muestras normales tanto la media como la mediana son iguales), luego se calculó la proporción de intervalos de confianza que contienen al parámetro, entre la mediana y la media ($p_{med,n}$ y $p_{prom,n}$ respectivamente), que contienen al parámetro de interés. Y por último se realizó la prueba de hipótesis para comparar que la proporción de intervalos de confianza que contiene el parámetro según la mediana es mejor que la proporción de intervalos de confianza que contiene al parámetro según la media (es decir que $p_{med,n} > p_{prom,n}$), dependiendo del tamaño de muestra.

ANALISIS DE DATOS

Primero se realizó con un tamaño de muestra pequeño ($n = 20$) y luego mostraremos para los distintos tamaños de muestra.

Para $n = 20$, las hipótesis a comparar son las siguientes:

$$Ho: p_{med,20} \leq p_{prom,20}$$
$$Ha: p_{med,20} > p_{prom,20}$$

Los resultados de la simulación muestran:

Según el método para la mediana, se encontraron 4789 intervalos de confianza que contienen al parámetro, por lo que $p_{med,20} = \frac{4789}{5000} = 0.9578$.

De igual manera, según el método para la media muestral, se encontraron 4754 intervalos de confianza que contienen al parámetro, por lo que $p_{prom,20} = \frac{4754}{5000} = 0.9508$.

El estadístico para probar la diferencia de proporciones es la siguiente:

$$Z = \frac{p_{med,20} - p_{prom,20}}{\sqrt{\frac{p_{med,20}(1-p_{med,20})}{5000} + \frac{p_{prom,20}(1-p_{prom,20})}{5000}}}$$

$$= \frac{0.9578 - 0.9508}{\sqrt{\frac{0.9578(1-0.9578)}{5000} + \frac{0.9508(1-0.9508)}{5000}}}$$

$$= 1.6762$$

Luego de esto compararemos con el valor de $Z_{o,0.05} = 1.645$ y al ser mayor se concluye que para una muestra de tamaño 20, el intervalo de confianza de la mediana estima mejor al parámetro que el intervalo de confianza de la media.

En la Tabla 2 se muestra la proporción de intervalos de confianza que contienen al parámetro de interés, su diferencia, el valor de $Z_{o,0.05}$ y si se acepta o rechaza la hipótesis nula.

Tabla 2. *Comparación de intervalos de confianza para la mediana y media para cada tamaño de muestra.*

(Se compara la hipótesis $Ho: p_{med,n} \leq p_{prom,n}$ *vs.* $Ha: p_{med,n} > p_{prom,n}$*)*

n	Mediana	Media	z	z_tabla	Conclusión
20	0.9578	0.9508	1.6762	1.6450	**Se rechaza Ho**
25	0.9550	0.9506	1.0375	1.6450	Se acepta Ho
30	0.9564	0.9494	1.6523	1.6450	**Se rechaza Ho**
35	0.9606	0.9510	2.3360	1.6450	**Se rechaza Ho**
40	0.9668	0.9506	4.0741	1.6450	**Se rechaza Ho**
45	0.9680	0.9488	4.8134	1.6450	**Se rechaza Ho**
50	0.9376	0.9488	-2.4201	1.6450	Se acepta Ho
55	0.9484	0.9524	-0.9212	1.6450	Se acepta Ho
60	0.9498	0.9524	-0.6028	1.6450	Se acepta Ho
65	0.9534	0.9494	0.9301	1.6450	Se acepta Ho
70	0.9588	0.9514	1.7870	1.6450	**Se rechaza Ho**
75	0.9344	0.9512	-3.6196	1.6450	Se acepta Ho
80	0.9410	0.9506	-2.1205	1.6450	Se acepta Ho
85	0.9480	0.9492	-0.2717	1.6450	Se acepta Ho
90	0.9538	0.9482	1.2972	1.6450	Se acepta Ho
95	0.9594	0.9504	2.1689	1.6450	**Se rechaza Ho**
100	0.9464	0.9480	0.3577	1.6450	Se acepta Ho
105	0.9498	0.9486	0.2732	1.6450	Se acepta Ho
110	0.9532	0.9480	1.1999	1.6450	Se acepta Ho
115	0.9588	0.9470	2.7860	1.6450	**Se rechaza Ho**
120	0.9458	0.9490	-0.7168	1.6450	Se acepta Ho
125	0.9504	0.9486	0.4110	1.6450	Se acepta Ho
130	0.9582	0.9470	2.6363	1.6450	**Se rechaza Ho**
135	0.9416	0.9466	-1.0883	1.6450	Se acepta Ho
140	0.9434	0.9470	-0.7909	1.6450	Se acepta Ho
145	0.9528	0.9466	1.4185	1.6450	Se acepta Ho
150	0.9550	0.9450	2.2948	1.6450	**Se rechaza Ho**
155	0.9434	0.9458	-0.5246	1.6450	Se acepta Ho
160	0.9492	0.9480	0.2717	1.6450	Se acepta Ho
165	0.9526	0.9474	1.1930	1.6450	Se acepta Ho
170	0.9404	0.9500	-2.1095	1.6450	Se acepta Ho
175	0.9496	0.9470	0.5871	1.6450	Se acepta Ho
180	0.9562	0.9482	1.8752	1.6450	**Se rechaza Ho**
185	0.9442	0.9476	-0.7515	1.6450	Se acepta Ho
190	0.9482	0.9468	0.3139	1.6450	Se acepta Ho
195	0.9530	0.9458	1.6427	1.6450	Se acepta Ho
200	0.9426	0.9492	-1.4589	1.6450	Se acepta Ho

En la Figura 1 se muestra el intervalo de confianza para la diferencia de proporciones, $p_{med,n} - p_{prom,n}$ dependiendo de cada tamaño de muestra.

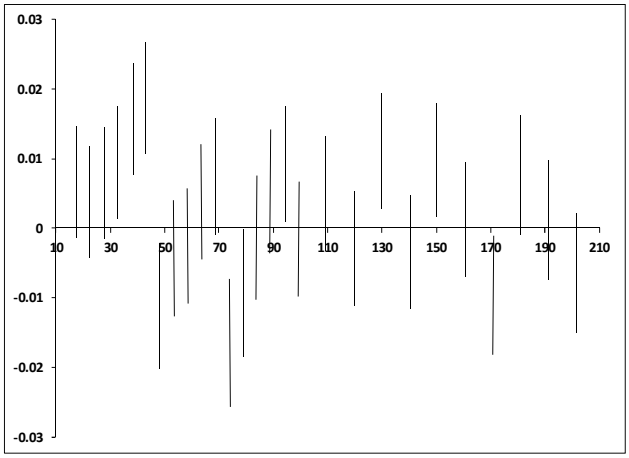

Fig.1 *Gráfica de intervalos de confianza para la diferencia de proporciones de intervalos de confianza de la mediana y la media para cada tamaño de muestra.*

CONCLUSIONES

Se puede ver que, según la tabla y la gráfica, para tamaños de muestra pequeños (entre 20 y 50), el intervalo de confianza de la mediana estima mejor al parámetro de interés, que el intervalo de confianza de la media.

Se puede ver que a medida que aumenta el tamaño de muestra, ambos métodos estiman por igual al parámetro de interés.

Aunque se ha probado que para muestras pequeñas, el intervalo de confianza de la mediana estima mejor que el intervalo de confianza de la media, el principal problema que surge es poder calcular este intervalo de confianza. Por lo cual, es recomendable utilizar el estimador de intervalo de confianza de la mediana desarrollado por Wodruff.

REFERENCIAS

1. **Arcos Cebrián A., Rueda García M., Martínez Miranda M., Gonzales Aguilera S**, "Intervalos de confianza alternativos para los cuantiles de una población finita", Estadística Española, Vol. 44, Núm. 149, pp.69 a 88 (1), 2002.

2. **Peró Cebollero M., Guardía Olmos J., Freixa Blanxart M., Turbany J., "**Técnicas basadas en la mediana como alternativas a pruebas clásicas de decisión, Psicothema, Vol. 20, Nº 004, 2008.
3. **Mendenhall, William. Scheaffer, Richard l. Wackerly, Dennis D,** "Estadística matemática con aplicaciones". Grupo Editorial Iberoamericana.

Correspondencia: jcameroj@uni.edu.pe

Recepción de originales: marzo 2013
Aceptación de originales: mayo 2013