El Bootstrap como Herramienta para la Enseñanza de la Distribución Muestral

María Inés Rodríguez [1] & Héctor Agnelli [2]

Resumen

La distribución muestral es un concepto básico para la adecuada comprensión de procedimientos inferenciales tales como test de hipótesis e intervalos de confianza. Existe un extenso conjunto de trabajos de investigación educativa en los que se reporta evidencia empírica acerca de que esta idea no es comprendida de manera adecuada. Creemos que la enseñanza de técnicas de remuestreo tales como bootstrap y test de permutación, además de brindar nuevas herramientas que puedan ser comprendidas y utilizadas por el futuro usuario de la estadística, contribuyen desde el punto de vista educativo a facilitar el aprendizaje de conceptos estadísticos claves tales como, distribución muestral, error estandar, intervalos de confianza y pruebas de significación. En este trabajo describimos la metodología bootstrap para el cálculo del error estandar de la mediana muestral.

Palabras clave: distribución muestral, remuestreo, error estandar, bootstrap.

Abstract

The sampling distribution is a basic concept for the proper understanding of inferential procedures such as hypothesis tests and confidence intervals. There is an extensive collection of works of educational research, reporting empirical evidence that this idea is not properly understood. We believe that teaching techniques such as bootstrap, resampling and permutation tests, in addition to providing new tools that can be understood and used by the future user of statistics, help from the educational point of view to facilitate the learning of key statistical concepts such as, sampling distribution, standard error, confidence intervals and significance tests. In this paper we describe the bootstrap methodology for calculating the standard error of the sample median.

Keywords: sampling distribution, resampling, standard error, bootstrap.

[1] Universidad Nacional de Río Cuarto, Argentina. mrodriguezbriguet@gmail.com
[2] Universidad Nacional de Rio Cuarto, Argentina. agnellih@gmail.com

1. Introducción

La prueba de hipótesis (y los valores p) es la metodología estadística más usada por los científicos experimentales aunque como muchas investigaciones lo ponen en evidencia es incorrecta su aplicación o inadecuada la interpretación del valor p o el nivel de significación (Vallecillos, 1996, Nickerson, 2000; Hubbard y Bayarri, 2003, Kline, 2004; Rodríguez y Albert, 2007). En un intento por reducir el impacto negativo de esta situación ha crecido una corriente que impulsa usar en su reemplazo los intervalos de confianza (IC). Esto ha llevado a trasladar la atención acerca del uso e interpretación que hacen los investigadores experimentales desde los tests a los IC, detectándose errores severos también en la interpretación de los mismos (Cumming, Willians, Fidler, 2004), (Belia, Fidler, Willians, Cumming, 2005). Además, en trabajos de investigación en los que se analizó la comprensión de IC por parte de los estudiantes se hallaron evidencias acerca de la incorrecta comprensión de su significado (Chance, del Mas, Garfield, 2004). Tratando de encontrar razones que expliquen estos comportamientos es conveniente analizar los conceptos que aparecen en la base constructiva de esos procedimientos inferenciales.

Esto conduce al análisis de una idea básica como es la de distribución muestral. Castro Soto y otros (2007) han recopilado un extenso conjunto de trabajos de investigación educativa en los que se reporta evidencia empírica acerca de que este concepto, no es comprendido de manera adecuada.

Por lo tanto, parece oportuno abordar el estudio de modos alternativos y a la vez complementarios de la enseñanza de la inferencia estadística y en particular, de la enseñanza de la distribución muestral. Desde esta posición creemos que las técnicas de remuestreo como el bootstrap y los tests de permutación, pueden contribuir a facilitar el aprendizaje de conceptos estadísticos claves como: distribución muestral, error estandar, intervalos de confianza y valores p. En este trabajo nos centraremos en el concepto de distribución muestral.

2. Distribuciones Muestrales

Sea F una distribución y θ un parámetro desconocido de la misma, pero fijo. Con el propósito de tener conocimiento acerca de θ, se trabaja con una muestra aleatoria $X_1,...,X_n$ de F. A partir de esta muestra se construye el estadístico $\hat{\theta} = \hat{\theta}(X_1,...,X_n)$ que será utilizado como estimador de θ. Pero para una muestra dada $X_1 = x_1,...,X_n = x_n$ ¿cuán bien aproxima $\hat{\theta} = \hat{\theta}(x_1,...,x_n)$ a θ? o en otros términos ¿cómo es la diferencia $\hat{\theta} - \theta$?. Dado que θ es desconocido la pregunta así formulada no tiene respuesta.

El análisis del comportamiento de los procedimientos inferenciales de la estadística clásica (a diferencia de la estadística Bayesiana) se basa en dar respuesta a esta otra pregunta ¿cómo será a la larga el comportamiento de este procedimiento? Es decir, basamos nuestra confianza en el funcionamiento de un procedimiento, no en una aplicación particular del mismo sino por su comportamiento en el muestreo repetido.

Como $\hat{\theta} = \hat{\theta}(X_1,...,X_n)$ es una función de variables aleatorias esto significa que en el muestreo repetido $\hat{\theta}$ presenta variabilidad entre muestras. Si establecido el tamaño n de la muestra pudiésemos tomar todas las posibles muestras de ese tamaño, podríamos calcular para cada una de ellas el valor del estimador $\hat{\theta} = \hat{\theta}(x_1,...,x_n)$ y en consecuencia tener la distribución de $\hat{\theta}$ como variable aleatoria, es decir tendríamos la llamada *distribución muestral* de $\hat{\theta}$.

Conocida esta distribución podríamos preguntarnos como será en promedio el comportamiento de la diferencia $\hat{\theta} - \theta$, y esta cantidad es llamada el *sesgo del estimador*:

$$Sesgo = E_\theta\left(\hat{\theta} - \theta\right)$$

También podemos analizar la variación de $\hat{\theta}$ alrededor de su esperanza y podríamos obtener el llamado *error estandar* de la estimación. De aquí la importancia de conocer la distribución muestral de los estimadores.

En general esperamos que la distribución muestral de $\hat{\theta}$ dependa del parámetro desconocido θ, por lo tanto al estudiar las propiedades de la distribución muestral del estimador obtendremos información acerca del comportamiento de $\hat{\theta}$ como estimador en el muestreo repetido: ¿Cuál será su valor esperado?, ¿Cuál será su dispersión? ¿Cómo será la forma de su distribución? Pero ahora surge otra pregunta cómo hallamos $F_{\hat{\theta}}$, es decir, la distribución de $\hat{\theta}$. En general podemos se mencionan tres maneras distintas de arribar al conocimiento de la distribución de $\hat{\theta}$.

El primero de ellos, es el método analítico o teórico. A manera de ejemplo, en los cursos tradicionales de estadística se enseña cuál es la distribución muestral de la media cuando se tienen muestras que proviene de una distribución normal o se enuncia la distribución aproximada de la media acudiendo al Teorema Central de Límite. Pero esta metodología no se puede extender de una manera obvia a otros estimadores, como por ejemplo la mediana muestral. De hecho en los cursos básicos de estadística y en los textos destinados a estos cursos en general no se habla de la precisión de la mediana, aunque si se insiste en su uso como medida descriptiva valiosa para valores centrales de distribuciones asimétricas.

Un segundo método, lo constituyen las distribuciones obtenidas por simulación Monte Carlo. Volviendo al caso de la mediana se asume conocida la población a muestrear, se simulan muestras de esta población y se estudia entonces el comportamiento en el muestreo repetido del estadístico de interés, en este caso la mediana. La restricción que tiene este método es la suposición acerca del conocimiento de la población.

Una tercera opción para estudiar la distribución muestral, es utilizar el remuestreo, metodología que será bosquejada más adelante.

3. Problemas reportados Acerca de la Comprensión de la Distribución Muestral

La comprensión del proceso de muestreo requiere distinguir entre dos propiedades básicas que tienen las muestras: la representatividad y la variabilidad. La primera de ellas referida a la similitud entre la muestra y la población queda asociada al diseño de la experiencia y la segunda es una característica vinculada con la variación entre muestras de una misma población extraídas en condiciones similares. En términos distribucionales y denominando F_θ a la distribución de la variable en la población, $\hat{F}_\theta$ a la distribución empírica de la variable en la muestra y $F_{\hat{\theta}}$ a la distribución muestral del estadístico algunos de los errores conceptuales vinculados con la comprensión de estas distribuciones son los siguientes:

a) Considerar a la distribución empírica de la variable en la muestra como una cabal representación de la distribución de la variable en la población sin importar el tamaño de la muestra

b) Confundir la distribución poblacional F_θ o a $\hat{F}_\theta$ con la distribución muestral del estadístico. Sobre esta última creencia cuando en las clases para desarrollar el concepto de distribución muestral se trabaja casi exclusivamente con muestras aleatorias de una distribución normal y se analiza el comportamiento de la media como estadístico, por cierto la media tiene también distribución normal- con la misma media aunque con varianza más pequeña- y por lo tanto involuntariamente se refuerza la idea del parecido necesario entre la forma de F_θ y la forma de $F_{\hat{\theta}}$, sin que quede claro que muestreando una población no normal, bajo las hipótesis del teorema central del límite, la distribución de la media es aproximadamente normal y por lo tanto distinta de la distribución original.

c) No distinguir claramente la distribución real del estadístico $F_{\hat{\theta}}$ con una eventual distribución asintótica utilizada como aproximación sin importar las condiciones de validez de esta aproximación -situación muy común para los estadísticos de los test de hipótesis-. Los errores señalados se convierten en obstáculos para la comprensión de conceptos fundamentales para la inferencia tales como errores estandar o valores p.

Así a manera de síntesis podemos decir que para asimilar el concepto de distribución muestral los alumnos deben estar familiarizados con las ideas de variabilidad, distribución y muestreo aleatorio y ser capaces, al menos, de

- distinguir entre la distribución de las observaciones de una muestra y la distribución de un estadístico en muchas muestras seleccionadas aleatoriamente.
- describir como podría ser una distribución muestral para diferentes poblaciones y diferentes tamaños de muestra (basándose en el centro, la dispersión y donde se encontrarían la mayor parte de los valores)
- apreciar que cuando aumenta el tamaño de muestra menor es la variabilidad del estadístico muestral.
- Comprender que el error estandar es una medida de la variabilidad del estadístico.

- Asumir que algunos valores del estadístico serán más o menos verosímiles que otros.

Estos conceptos son fundamentales para la adecuada comprensión e interpretación de IC, valores p, potencia y replicabilidad. Desde el punto de vista computacional están disponibles en la web, distintos simuladores que permiten al estudiante familiarizarse con estos conceptos seleccionando una entre varias distribuciones poblacionales, como así también para trabajar con diferentes tamaños de muestras, que contribuyen a comprender las ideas centrales de la inferencia estadística.

Presentada entonces la situación anterior, nos parece oportuno desde el punto de vista de la enseñanza, plantear en los cursos básicos la posibilidad de utilizar otras herramientas complementarias para la formación en inferencia estadística, como son los procedimientos de remuestreo. Se sabe que ellos constituyen herramientas útiles en situaciones complejas, que sin duda son enseñados a los futuros profesionales de la estadística. Estos métodos que no son presentados, salvo excepciones, en los cursos iniciales, prescinden del conocimiento las distribuciones muestrales teóricas y se basan en el muestreo repetido (remuestreo) de la misma muestra original. Nosotros creemos que por su esencia brindan la oportunidad de facilitar el aprendizaje de conceptos estadísticos claves.

4. El Remuestreo

Los test paramétricos tradicionales al igual que la construcción de intervalos de confianza están basados en el conocimiento de distribuciones muestrales ya sea del estadístico del test o de los estimadores puntuales en el caso de los intervalos de confianza. En cambio. Al generar este proceso de remuestreo varias veces se genera una distribución muestral que estima la distribución muestral teórica desconocida. (Yu, C.H.,2008).

Este proceso demanda un uso intensivo de la computadora, de allí que su actual vigencia práctica está dada por la disponibilidad de recursos computacionales veloces y baratos. En una etapa del proceso se utiliza la simulación Monte Carlo, pero en lugar de generar las simulaciones a partir de un modelo distribucional hipotético o postulado, tal cual es el uso habitual de las simulaciones Monte Carlo, en el remuestreo las simulaciones se hacen a partir de los datos reales que pertenecen a la muestra original disponible. Existen distintas técnicas de remuestreo, como ser: Validación cruzada, Jackknife, Test de permutaciones, Bootstrap, entre otras. En este trabajo nos centraremos en esta última metodología.

4.1.- Bootstrap

Efron (1979) introdujo el bootstrap como un método general para estimar la distribución muestral de un estadístico basándose en los datos observados. Un problema típico de la estadística involucra la estimación de un parámetro desconocido. Ante esta situación emergen dos preguntas básicas 1) ¿Qué estimador debe ser usado? 2) habiendo elegido el estimador ¿Cuán preciso es este estimador?

El bootstrap es una metodología general destinada a responder a la segunda pregunta. Es una técnica basada en la potencia computacional como un *sustituto* del análisis teórico. El bootstrap permite determinar la distribución muestral de, virtualmente, cualquier estadístico computado a partir de una muestra. En los cursos tradicionales de estadística se enseña que si se quiere estimar la esperanza de una distribución se toma una muestra aleatoria y se calcula la media muestral como estimador del parámetro. Pero también los datos proporcionan para la precisión del estimador una estimación de su error estandar y a partir de éste se pueden construir intervalos de confianza. Como antes se ha señalado, en general no se habla de la precisión de la mediana y mucho menos de sus intervalos de confianza. El bootstrap permite, obviando ya sea deducciones matemáticas complejas, o los comportamientos asintóticos de los estimadores, superar estos obstáculos. Para esto el bootstrap aproxima la distribución muestral de la mediana sin apelar al conocimiento de la distribución poblacional, sino, considerando la distribución empírica de los datos. Es decir trata a los datos disponibles como una población y saca muestras aleatorias con reemplazo y de igual tamaño a la original, de dicha población. Por eso, este procedimiento se denomina remuestreo. Se puede notar que en estas muestras aparecerán valores repetidos de los datos originales como consecuencia del muestreo con reemplazo.

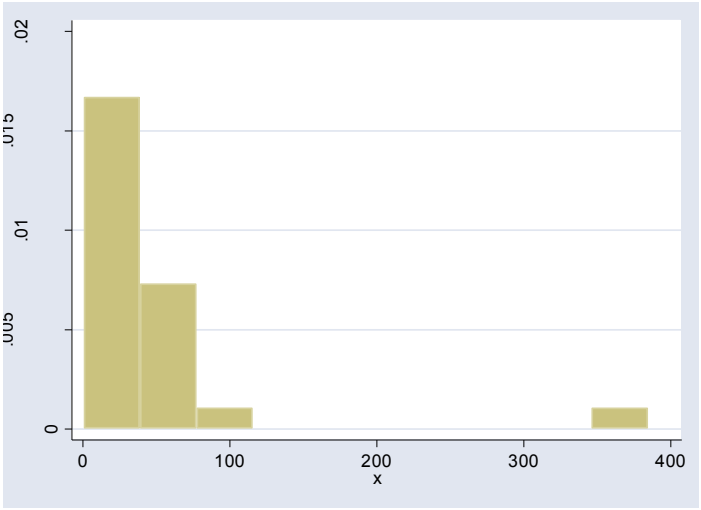
Para cada muestra se calcula la mediana y considerando una gran cantidad de estas muestras, por ejemplo 1000, se obtiene la distribución muestral aproximada de la mediana muestral y entonces se puede obtener una estimación de su error estandar. Esto que se ha hecho para la mediana se puede hacer para otros estimadores no importa cuan complicados sean. La técnica bootstrap permite también estimar el sesgo del estimador.

Por otra parte a partir de la obtención del error estandar se construyen intervalos de confianza. (DiCiccio, Efron,1996; Efron, 1993). El único requisito para alcanzar estos resultados es disponer de un software que permita calcular el estimador y generar el proceso del remuestreo (Boos, Stefanski, 2010).

Ejemplo: La siguiente tabla brinda los datos de una muestra aleatoria de tamaño *n=25*

1	4	6	12	13	14	18	19	20	22	23	24	26
31	34	37	46	47	56	61	63	65	70	97	385	-

El siguiente histograma revela la distribución asimétrica de esta muestra



Histograma de los valores muestrales

La media es $\bar{x} = 47{,}76$ y la mediana $m = 26$. Por otra parte el error estandar de la media es E.E (media) = 14.8, este error estandar es fácilmente calculado a partir del conocimiento del desvío estándar S de los datos como $S/\sqrt{n}$. Sin embargo, no existe una fórmula tan simple que entregue el error estandar de la mediana. Utilizando el software Stata se puede aplicar la metodología bootstrap aplicando el siguiente comando:

- ***bootstrap "sum x, detail" r(p50), reps(1000) saving(rem)***

Esta instrucción indica que a partir de la muestra original se tomaran 1000 muestras con reemplazo todas de tamaño n=25 y para cada una de ellas se calculará la mediana. La salida que brinda el software mencionado, es la siguiente:

Variable	Reps	Observed	sesgo	Std. Err.	[95% I.C]	
m	1000	26	2.67	**7.57**	19	46

Se observa que el error estandar de la mediana es E.E (m) = 7.57. Este valor señala que para un conjunto de datos asimétricos como el que estamos analizando, la mediana es menos variable que la media.

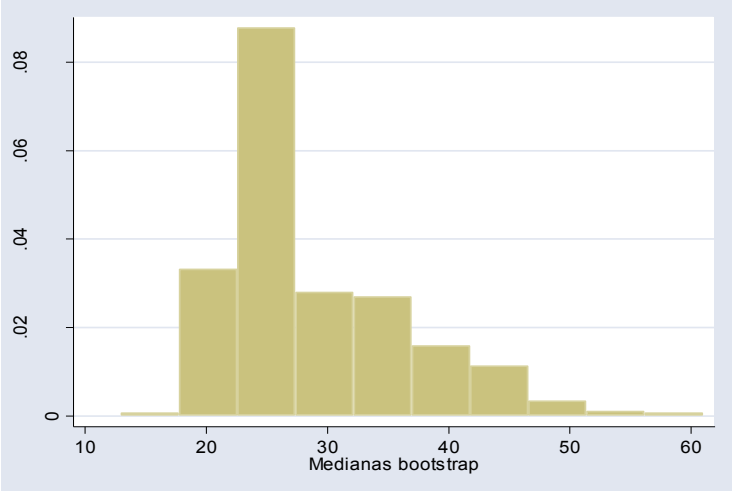
El software también almacena en el archivo llamado rem los valores de las 1000 medianas y se puede solicitar tanto, un resumen de estos valores como un histograma de los mismos para tener una visión aproximada de cómo es la distribución muestral de la mediana. El resumen está presentado en la siguiente tabla

Variable	Obs	Mean	Std. Dev.	Min	Max
m	1000	28.67	7.57	13	61

Aquí se observa que lo consignado como desvío estandar de la distribución muestral bootstrap es una estimación del error estandar de la mediana tal como había sido indicado en la salida exhibida anteriormente. Por otra parte la diferencia entre la media de estos 1000 valores (28.67) y el valor original de la mediana (26) es llamado el sesgo y es el valor 2.67, también antes consignado. Cabe señalar que menor sea el sesgo más centrada estará la distribución en el valor original del estadístico.

La metodología bootstrap también entrega un I.C para la mediana muestral. Ordenadas las medianas de menor a mayor se toman los valores que están en la posición 25 y 975 y estos valores determinan un I. C del 95%, el I.C obtenido en nuestro ejemplo es [19-46] señalado en la salida. Cabe acotar que existen también otros procedimientos, aquí no mostrados, para calcular I.C utilizando bootstrap.

A continuación se presenta el histograma de la distribución empírica de las 1000 medianas obtenidas por bootstrap.



Histograma de los valores de las medianas

5. Consideraciones

La distribución muestral es un concepto básico y fundamental para la comprensión de la inferencia estadística, pero como se ha señalado su enseñanza y aprendizaje presenta dificultades. La construcción de la distribución de un estadístico a partir de un conjunto de datos es de difícil visualización, pues se la debe obtener ya sea mediante fórmulas o mediante simulaciones, pero en ambos casos haciendo supuestos distribucionales acerca de la población muestreada. El uso de la metodología boostrap, basada en el remuestreo, permite apreciar, prescindiendo de fórmulas, cómo varía un estadístico de muestra a muestra y por lo tanto, cómo se va construyendo la distribución muestral.

Por tal motivo, estimamos relevante continuar indagando acerca de cuáles son las maneras más adecuadas para introducir esta metodología basada en el uso intensivo del recurso computacional, como herramienta complementaria a la enseñanza tradicional de la estadística inferencial en los cursos básicos o introductorios de estadística.