

(#409) . ALGUNAS CONSIDERACIONES EN LA ESTIMACIÓN DEL RENDIMIENTO POR MINUTO EN BALONCESTO (II)

[MONOTEMA] Una vez expuesta la [necesidad de gestionar adecuadamente las estadísticas por minuto en baloncesto, y tras explicar cómo ha de computarse la media](#), el siguiente paso es discutir las opciones para el cálculo de la varianza de esa media, y por ende, del error estándar necesario para conocer la imprecisión de la estimación.

Aproximación Normal

[Levy & Lemeshow \(1999\)](#) proponen una aproximación a la estimación del error estándar de la media de la siguiente forma:

$$SE(r_2) = \frac{r_2}{\sqrt{n}} (V_x^2 + V_y^2 - 2\rho_{xy}(V_x)(V_y))^2 \sqrt{\frac{N-n}{N-1}}$$

donde:

$$V_x^2 = \left(\frac{N-1}{N} \right) \left(\frac{S_x^2}{(\bar{x})^2} \right)$$

$$V_y^2 = \left(\frac{N-1}{N} \right) \left(\frac{S_y^2}{(\bar{y})^2} \right)$$

$$\rho_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x S_y}$$

Como se puede apreciar, esta expresión cuenta además con la consideración de un factor de finitud que hace que el error estándar sea cero cuando $N=n$, es decir, cuando tenemos que la

muestra es en sí toda la población. Es una fórmula a la que se llega a través del desarrollo de Taylor de la función ratio en el entorno de la media.

El intervalo de confianza $100(1-\alpha)\%$ bajo la aproximación Normal es el siguiente:

$$r_2 - z_{1-(\alpha/2)}SE(r_2) \leq R \leq r_2 + z_{1-(\alpha/2)}SE(r_2)$$

donde α es el "tamaño del test" y $100(1-\alpha)\%$ es el nivel de confianza. De este modo, para un nivel de confianza del 95% tenemos:

$$r_2 - 1.96SE(r_2) \leq R \leq r_2 + 1.96SE(r_2)$$

Aproximación de Cochran

[Gatz & Smith \(1995\)](#), basándose en el trabajo de [Cochran \(1977\)](#), proponen el siguiente estimador:

$$SE(r_3) = \frac{n}{(n-1)(\sum_{i=1}^n y_i)^2} \left[\sum_{i=1}^n (y_i r_i - \bar{y} r_3)^2 - 2r_3 \sum_{i=1}^n (y_i - \bar{y})(y_i r_i - \bar{y} r_3) + (\bar{r}_3)^2 \sum_{i=1}^n (y_i - \bar{y})^2 \right]$$

Al igual que en el caso anterior, se puede construir un intervalo de confianza usando la aproximación Normal:

$$r_3 - 1.96SE(r_3) \leq R \leq r_3 + 1.96SE(r_3)$$

No obstante, [Gatz & Smith \(1995\)](#) son prudentes en advertir que no siempre sería correcto asumir la aproximación Normal, sobre todo para muestras pequeñas. Así, la estimación de los errores estándar por remuestreo y el establecimiento de puntos de corte de la distribución empírica remuestrada sería una alternativa a considerar.

Bootstrapping

[Gatz & Smith \(1995\)](#), muestran que la aproximación de Cochran proporciona errores estándar que no difieren estadísticamente de los obtenidos por bootstrapping.

El método de remuestro consiste básicamente en la extracción de muestras con repetición de la muestra original, y la construcción de una distribución empírica de la media ponderada, donde se puede calcular su error estándar (también empírico). La implementación de intervalos de confianza puede realizarse de varias maneras, también empleando la aproximación normal, o los percentiles de la distribución empírica, que en el caso de 2 colas sería el percentil 2.5% y el 97.5% de la distribución.

Si asumimos esta última opción, los intervalos de confianza al 95% serían:

$$r_3 - P_b(2.5) \leq R \leq r_3 + P_b(97.5)$$

Ilustración práctica

Vamos a emplear de nuevo los [datos de Mike James](#), que nos van a permitir calcular la imprecisión de su media de puntos por minuto de las 3 formas que acabamos de explicar.

Para ello, suponemos que James ha jugado sólo 25 de los 30 partidos posibles (los 25 primeros), por lo que la estimación de los puntos por minuto tendrá una imprecisión asociada.

Los resultados, con el error estándar y al 95% de confianza son los siguientes:

Aproximación Normal:

$$r_2 = 0.5734; SE(r_2) = 0.0117; IC95 = (0.5504; 0.5965)$$

Aproximación de Cochran:

$$r_3 = 0.5734; SE(r_3) = 0.0288; IC95 = (0.5171; 0.6298)$$

Bootstrapping Normal:

$$r_3 = 0.5734; SE(r_3) = 0.0275; IC95 = (0.5196; 0.6273)$$

Bootstrapping percentil:

$$r_3 = 0.5734; IC_{95} = (0.5373; 0.6428)$$

Como puede apreciarse, todos los intervalos de confianza contienen al parámetro poblacional, que conocemos (recordemos que sabíamos el rendimiento en los 30 partidos), y que es $R = 0.5842$

De entre todos los procedimientos explicados, el primer de ellos es el que proporciona estimaciones más precisas, porque el error estándar es bastante más pequeño. La clave está en la inclusión de este factor de finitud:

$$\sqrt{\frac{N-n}{N-1}}$$

Si ese factor no se tiene en cuenta, entonces el valor del error estándar sería de 0.0283, es decir, muy similar al obtenido con el método de Cochran y el de remuestreo.

Creemos, sin embargo, que si se entiende que todos los partidos de una competición forman una población finita, y que si el jugador participa en todos ellos entonces su rendimiento no tiene imprecisión, entonces sería conveniente introducir factores de finitud en las estimaciones, que corrijan los errores estándar en muestras finitas (y pequeñas), y así obtener mayor fiabilidad.

[Levy & Lemeshow \(1999\)](#) recomiendan que sólo se use esa aproximación si:

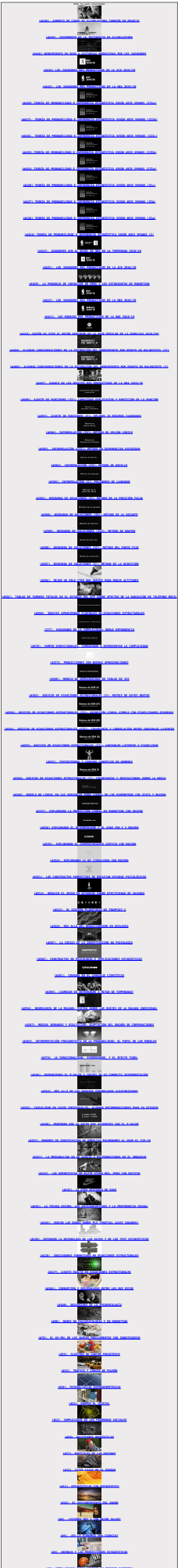
$$\frac{S_y}{\sqrt{ny}} \sqrt{\frac{N-n}{N}} \leq 0.05$$

En nuestro caso, ese valor era de 0.0079, por lo que se cumple esa condición.

Conclusión

Ya tenemos un poco más claras algunas de las opciones que tenemos para analizar rendimientos por minuto en baloncesto. Tras explicar cómo se puede calcular la media, hemos presentado varias alternativas para el cálculo de los errores estándar y el intervalo de confianza asociado

El error relativo cometido en el primer caso es del 4.01%, lo que se antoja aceptable para seguir confiando en lo que nos diga un rendimiento de 25 partidos sobre 30 posibles. Hay que tener cuidado cuando el tamaño de la muestra comienza a bajar con respecto al de la población, porque el error relativo se incrementa, y entonces habremos de buscar un criterio de inclusión en el ranking de final de temporada, ya que aquellos jugadores con un rendimiento demasiado impreciso no deberían aparecer en él.



(#408) . ALGUNAS CONSIDERACIONES EN LA ESTIMACIÓN DEL RENDIMIENTO POR MINUTO EN BALONCESTO (I)

[MONOTEMA] Hace unos años incidí en la necesidad de tomar una aproximación probabilística a la construcción de rankings para valorar el rendimiento de jugadores de baloncesto. Como se puede leer en [este post y en el artículo que publiqué en RICYDE](#), es necesario considerar las imprecisiones en las estimaciones de los valores medios que caracterizan el rendimiento de los jugadores: puntos, rebotest, asistencias, etc.

Cuando un jugador no juega todos los partidos de la temporada, su valor medio de puntos es un estimador del valor medio de puntos que habría obtenido si los jugara todos, si consideramos que todos los partidos componen la población, y asumimos (con cierto riesgo) que los partidos en los que realmente juega son una muestra aleatoria de esa población.

En el [artículo](#), se explica paso a paso un método para hacerlo, incluyendo a aquellos jugadores en los que la precisión sea admisible, es decir, no tengan un intervalo de confianza demasiado grande. De este modo, sólo sería posible la comparación rigurosa entre jugadores que hubieran jugado un número mínimo de partidos. De manera intuitiva, eso es lo que realmente se suele hacer en la valoración de los rankings en las competiciones profesionales, aunque esos criterios de inclusión no sean del todo precisos y justificados estadísticamente.

Sin embargo, las variables del box-score no están ponderadas por los minutos de juego, y esto propicia que se puedan obtener mejor (o peor) rendimiento bruto en función del número de minutos jugados, y no de la habilidad subyacente del jugador. Por tanto, es muy recomendable comparar el rendimiento de los jugadores por minuto jugado, en aras de obtener índices de “productividad”, o capacidad de aportar al rendimiento del equipo en función de los recursos empleados, que en este caso son los minutos que se está en pista.

Pero al construir una variable de rendimiento por minuto jugado, nos encontramos con ciertas dificultades estadísticas que merecen ser discutidas, ya que ni la estimación de la media, ni de la varianza, ni del error típico son tan sencillas como las de una variable sin ponderar. El objetivo de este post, es comentar algunas de esas opciones que los analistas tenemos para realizar nuestro trabajo, centrándonos en el cálculo de la media. Dejaremos para más adelante el cómputo del error.

El cálculo de la media

Partamos de un ejemplo práctico para ilustrar el problema; la estimación de los puntos por minuto del máximo anotador de la fase regular de la Euroliga 2018/19: [Mike James](#). El jugador del Olimpia Milan ha anotado 595 puntos (X) en 30 partidos, es decir, una media de 19.8.

Pero ha jugado 1018 minutos y 26 segundos, es decir, 1018.433 minutos (Y), por lo que los puntos por minuto ($R=X/Y$) han sido: 0.5842. Sin embargo, la media de todos los puntos por minuto de los 30 partidos es 0.5839, es decir, difiere (en este caso ligeramente) de lo obtenido cuando se divide 595 entre 1018.433. ¿Cómo es esto posible?

Recordemos que una de las primeras cosas que aprendemos en estadística es que la esperanza matemática de la media muestral es la media poblacional. Es decir, la media de todas

las medias muestrales es la media poblacional, dicho de otro modo, la media muestral es un estimador insesgado de la media poblacional.

Pero no ocurre así en este caso, y la razón es que precisamente tenemos una variable de "razón", o un ratio entre dos variables aleatorias: los puntos y los minutos. Cuando se tiene ese ratio, la media muestral no es un estimador insesgado de la media poblacional.

En su recomendable libro, [Levy & Lemeshow \(1999\)](#), admiten en la página 191 que en la práctica ese error es muy pequeño en la mayoría de ocasiones, y que se suele despreciar.

Sin embargo, tal y como demuestran van