(#410) ¿QUIÉN HA SIDO EL MEJOR ANOTADOR DE LA FASE REGULAR DE LA EUROLIGA 2018/19?

[MONOTEMA] La temporada regular de la Euroliga 2018/19 acaba de finalizar, por lo que podemos aplicar nuestro procedimiento para conocer, por ejemplo, quién ha sido el mejor anotador de la competición en estos primeros 30 partidos. Aunque es cierto que el Trofeo Alphonso Ford se otorga al jugador que mejor media de puntos tiene al final de la competición (incluyendo la segunda fase), podemos hacer una analogía a los rankings que se realizan en la NBA y que distinguen entre temporada regular y play-offs.

De la página oficial de la Euroliga obtenemos los siguientes datos de los 5 primeros clasificados en cuanto a puntos por partido:

RK	PLAYER	TEAM	G	PTS	MP	PTS/G	PTS/MP
1	JAMES, MIKE	AX Armani Exchange Olimpia Milan	30	595	1018.43	19.83	.5842
2	HIGGINS, CORY	CSKA Moscow	26	388	641.08	14.92	.6052
3	DOUGLAS, TONEY	Darussafaka Tekfen Istanbul	20	290	551.25	14.5	.5261
4	DAVIES, BRANDON	Zalgiris Kaunas	30	433	733.42	14.43	.5904
5	MICKEY, JORDAN	Khimki Moscow Region	28	398	631.47	14.21	.6303

Como puede observarse, Mike James aparece como el primero en el ranking, con una diferencia muy clara con respecto al segundo clasificado: Cory Higgins. Sin embargo, hay una diferencia ostensible en minutos jugados con lo que realmente puede que esa variable haya condicionado la anotación. Un jugador puede anotar más que otro por partido no porque sea mejor anotador, sino simplemente porque juega más. El más

productivo es el que lo hace mejor en función de los recursos empleados, que en este caso son los minutos de juego.

Ranking de puntos por partido empleando una aproximación probabilística

Tres de los cinco jugadores implicados no han disputado todos los partidos posibles, por lo que su media de puntos por partido tiene una incertidumbre asociada. Si queremos ser escrupulosos en la comparación, hemos de computar los intervalos de confianza.

Veamos qué ocurre ahora con la clasificación de máximos anotadores:

RK	PLAYER	TEAM	G	PTS	PTS/G	Iclow	Ichigh	Error Rel
1	JAMES, MIKE	AX Armani Exchange Olimpia Milan	30	595	19.83	19.83	19.83	0
2	HIGGINS, CORY	CSKA Moscow	26	388	14.92	14.303	15.543	.0415
2	DOUGLAS, TONEY	Darussafaka Tekfen Istanbul	20	290	14.5	12.695	16.305	.1245
2	DAVIES, BRANDON	Zalgiris Kaunas	30	433	14.43	14.43	14.43	0
2	MICKEY, JORDAN	Khimki Moscow Region	28	398	14.21	13.637	14.791	.0406

El primer puesto lo sigue ocupando claramente Mike James, ya que sus intervalos de confianza no se solapan con los de ningún otro jugador. Sin embargo, en cuanto al segundo puesto, todos los jugadores están en un rango estadísticamente equivalente, por lo que su media de puntos por partido es indistinguible. De este modo, lo más justo sería que todos ocuparan la segunda posición.

Probablemente habríamos de cuestionarnos seriamente la inclusión de Toney Douglas, porque su error relativo es superior al 12%, es decir, la imprecisión de la estimación es alta, y eso hace que la "calidad" del dato sobre su rendimiento no sea demasiado buena.

Ranking de puntos por minuto

Ahora sí que es el momento de aplicar lo aprendido en cuanto a

la estimación de medias ponderadas y sus intervalos de confianza. Los resultados se muestran en la siguiente tabla:

RK	PLAYER	TEAM	G	MP	PTS/MP	Iclow	Ichigh	Error Rel
1	MICKEY, JORDAN	Khimki Moscow Region	28	631.47	.6303	.6088	.6517	.0345
1	HIGGINS, CORY	CSKA Moscow	26	641.08	.6052	.5862	.6242	.0314
2;3	DAVIES, BRANDON	Zalgiris Kaunas	30	733.42	.5904	.5904	.5904	0
4	JAMES, MIKE	AX Armani Exchange Olimpia Milan	30	1018.43	.5842	.5842	.5842	0
5	DOUGLAS, TONEY	Darussafaka Tekfen Istanbul	20	551.25	.5261	.4741	.5791	.0988

Y ahora tenemos un cambio ostensible de la información sobre anotación. Mike James no es el mejor anotador por minuto, sino el 4º.

Son Jordan Mickey y Cory Higgins los que se deberían de repartir esa distinción, porque ambos superan significativamente al resto. Sin embargo, sus intervalos de confianza al 95% se solapan claramente, lo que les hace estadísticamente equivalentes, y por tanto comparten la primera posición del ranking.

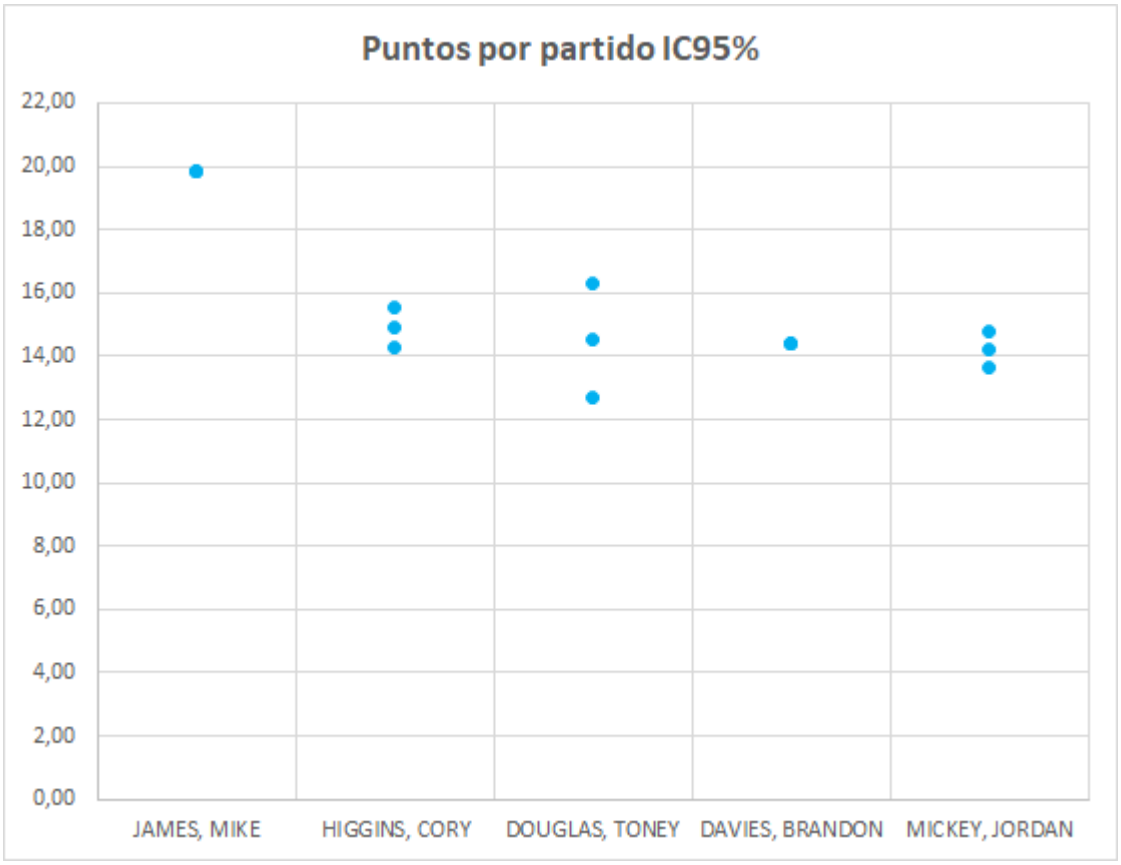
Ayuda gráfica

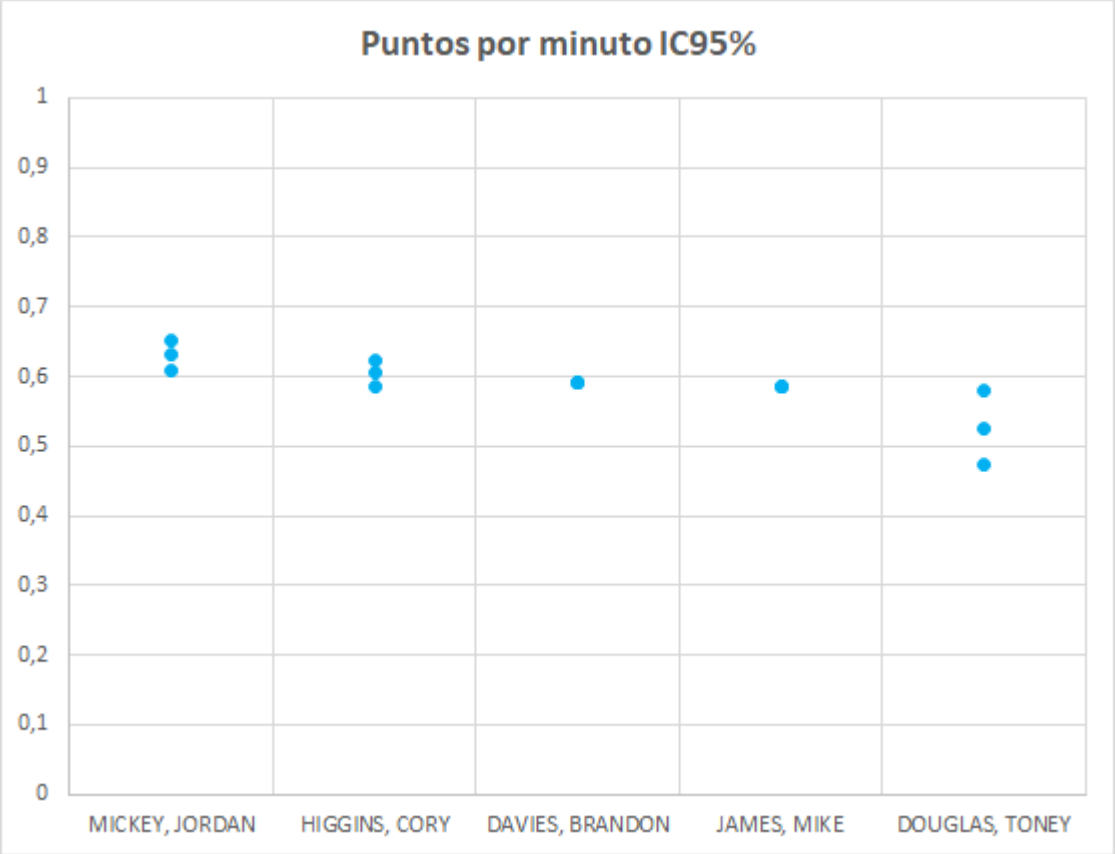
Una forma de representar los datos que nos puede ayudar es dibujar las estimaciones puntuales junto con los intervalos de confianza, con puntos en la misma vertical para cada jugador. Así, vemos más claramente qué estimaciones son más precisas y cuáles "llegan" a un determinado umbral de rendimiento.

Desde un punto de vista práctico es probablemente más interesante valorar a los jugadores en función de si llegan a una determinada cota superior de rendimiento, ya sea por partido o por minuto. Las líneas horizontales que marcan cada punto del eje vertical nos sirven como referencia.

Como hemos mencionado antes, la imprecisión de Toney Douglas para los puntos por partido es quizá inasumible para sacar conclusiones claras sobre su rendimiento, aunque el error

relativo en los puntos por minuto es menor del 10% y, por tanto, algo más preciso.



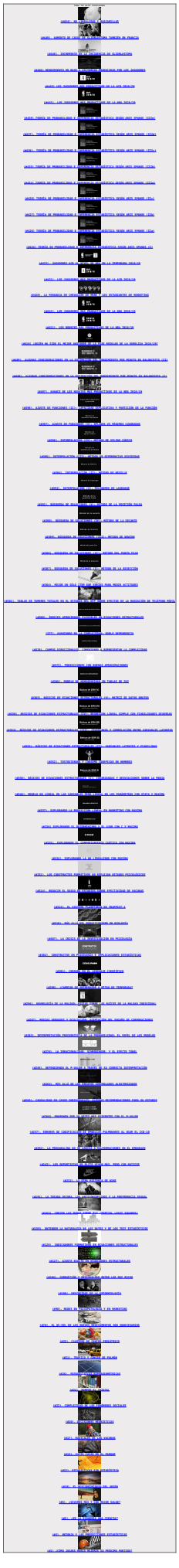


Conclusión

¿Ha sido Mike James el mejor anotador de la fase regular de la Euroliga 2018/19? Pues quizá depende de lo que se entienda por "mejor". Desde luego es el que más puntos ha conseguido por partido (significativamente mejor que el resto una vez tenido en cuenta el error de estimación). Sin embargo, el rendimiento por minuto de Jordan Mickey y Cory Higgins ha sido significativamente mejor, por lo que estos jugadores han aportado relativamente más a sus respectivos equipos.

Evidentemente el análisis de anotación no nos dice nada sobre los fallos cometidos ni sobre otras acciones de juego, por lo que es algo pobre para evaluar la capacidad ofensiva o la productividad global de los jugadores. Para ello necesitamos incluir más variables, lo que complica técnicamente las estimaciones de los errores.

En posteriores artículos intentaremos ofrecer soluciones para ello.



(#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:

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

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

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

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-α)% 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-α)% es el nivel de confianza. De este modo, para un nivel de confianza del 95% tenemos:

Unbalanced Eqn

Aproximación de Cochran

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

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

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

Unbalanced Eqn

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 bootrstrapping.

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:

$$r_3 = 0.5734; IC95 = (0.5373; 0.6428)$$

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:

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

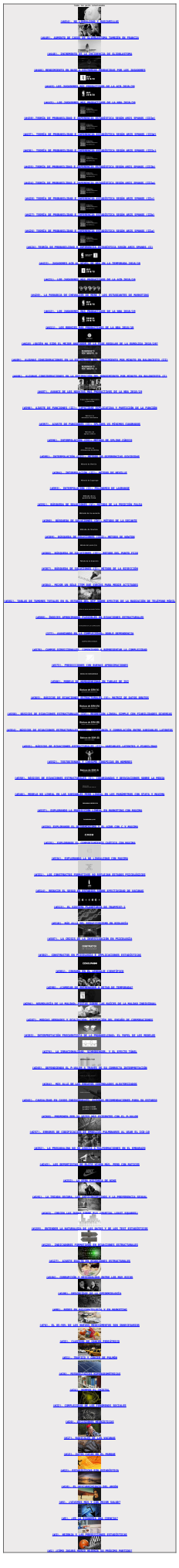
En nuestro caso, ese valor era de 0.0079, por lo que se cumple

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 impreciones 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 intuititva, 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