Análisis al dataframe kobe_basket

Alejandro Zavala

2023-09-10

Contents

Descripción del problema

Los jugadores de baloncesto que encestan varias canastas seguidas se describen como tener una *mano caliente*. Los aficionados y los jugadores han creído durante mucho tiempo en la mano caliente, fenómeno que refuta la suposición de que cada disparo es independiente del próximo.

Sin embargo, [un artículo de 1985] (http://www.sciencedirect.com/science/article/pii/0010028585900106) de Gilovich, Vallone y Tversky recopiló evidencia que contradijo esta creencia y demostró que los disparos sucesivos son independientes eventos. Este artículo inició una gran controversia que continúa hasta el día de hoy, como se puede ver buscando en Google *baloncesto de mano caliente*.

En este analisis se aplicara un enfoque para responder preguntas como esta. Los objetivos son: 1. Pensar en los efectos de eventos independientes y dependientes 2. Simular rachas de disparos en R 3. Comparar una simulación con la real datos para determinar si el fenómeno de la mano caliente parece ser real.

Nuestra investigación se centrará en el desempeño de un jugador: Kobe Bryant de los Lakers de Los Ángeles. Su actuación ante los Orlando Magic en el 2009. Las finales de la NBA le valieron el título de *Jugador Más Valioso* y muchos espectadores coentaron cómo parecía mostrar una mano caliente.

Exploración al dataframe

```
kable(head(kobe_basket))
```

vs	game	quarter	time	description	shot
ORL	1	1	9:47	Kobe Bryant makes 4-foot two point shot	H
ORL	1	1	9:07	Kobe Bryant misses jumper	M
ORL	1	1	8:11	Kobe Bryant misses 7-foot jumper	M
ORL	1	1	7:41	Kobe Bryant makes 16-foot jumper (Derek Fisher assists)	H
ORL	1	1	7:03	Kobe Bryant makes driving layup	H
ORL	1	1	6:01	Kobe Bryant misses jumper	M

Viendo las variables que contien el dataframe

```
names(kobe_basket)
```

```
## [1] "vs"          "game"        "quarter"     "time"        "description"
## [6] "shot"
```

Para ver la estructura de un dataframe en R

```
str(kobe_basket)
```

```
## tibble [133 x 6] (S3: tbl_df/tbl/data.frame)
##  $ vs         : Factor w/ 1 level "ORL": 1 1 1 1 1 1 1 1 1 1 ...
##  $ game       : int [1:133] 1 1 1 1 1 1 1 1 1 1 ...
##  $ quarter    : Factor w/ 5 levels "1","1OT","2",..: 1 1 1 1 1 1 1 1 1 3 ...
##  $ time       : Factor w/ 116 levels "00:00.0","00:00.5",..: 114 109 102 100 96 85 64 21 11 91 ...
##  $ description: Factor w/ 80 levels "Bryant 3pt Shot: Made (16 PTS) Assist: Bynum (1 AST)  ",..: 40 1
##  $ shot       : chr [1:133] "H" "M" "M" "H" ...
```

Adicional si queremos conocer mas acerca del dataframe

```
?kobe_basket
```

```
## starting httpd help server ... done
```

Ahora, por ejemplo, en el Juego 1 Kobe tuvo la siguiente secuencia de aciertos y errores de sus nueve intentos de tiro en el primer cuarto:

Filtrando información

```
game1_q1 <- kobe_basket %>% filter(game == 1 & quarter == 1) %>% arrange(time)
game1_q1
```

```
## # A tibble: 9 x 6
##   vs     game quarter time  description                                  shot
##   <fct> <int> <fct>   <fct> <fct>                                        <chr>
## 1 ORL       1 1       0:00  Kobe Bryant misses layup                     M
## 2 ORL       1 1       0:52  Kobe Bryant misses 19-foot jumper            M
## 3 ORL       1 1       4:07  Kobe Bryant misses 12-foot jumper            M
## 4 ORL       1 1       6:01  Kobe Bryant misses jumper                    M
## 5 ORL       1 1       7:03  Kobe Bryant makes driving layup              H
## 6 ORL       1 1       7:41  Kobe Bryant makes 16-foot jumper (Derek Fishe~ H
## 7 ORL       1 1       8:11  Kobe Bryant misses 7-foot jumper             M
## 8 ORL       1 1       9:07  Kobe Bryant misses jumper                    M
## 9 ORL       1 1       9:47  Kobe Bryant makes 4-foot two point shot      H
```

Observando la salida, podemos suponer que cuando el tiro (shot) es M signidica que perdio el tiro (miss) y si es H que lo encesto (hit). calc_streak calcula las rachas de hits.

```
kobe_streak <- calc_streak(kobe_basket$shot)
kobe_streak
```

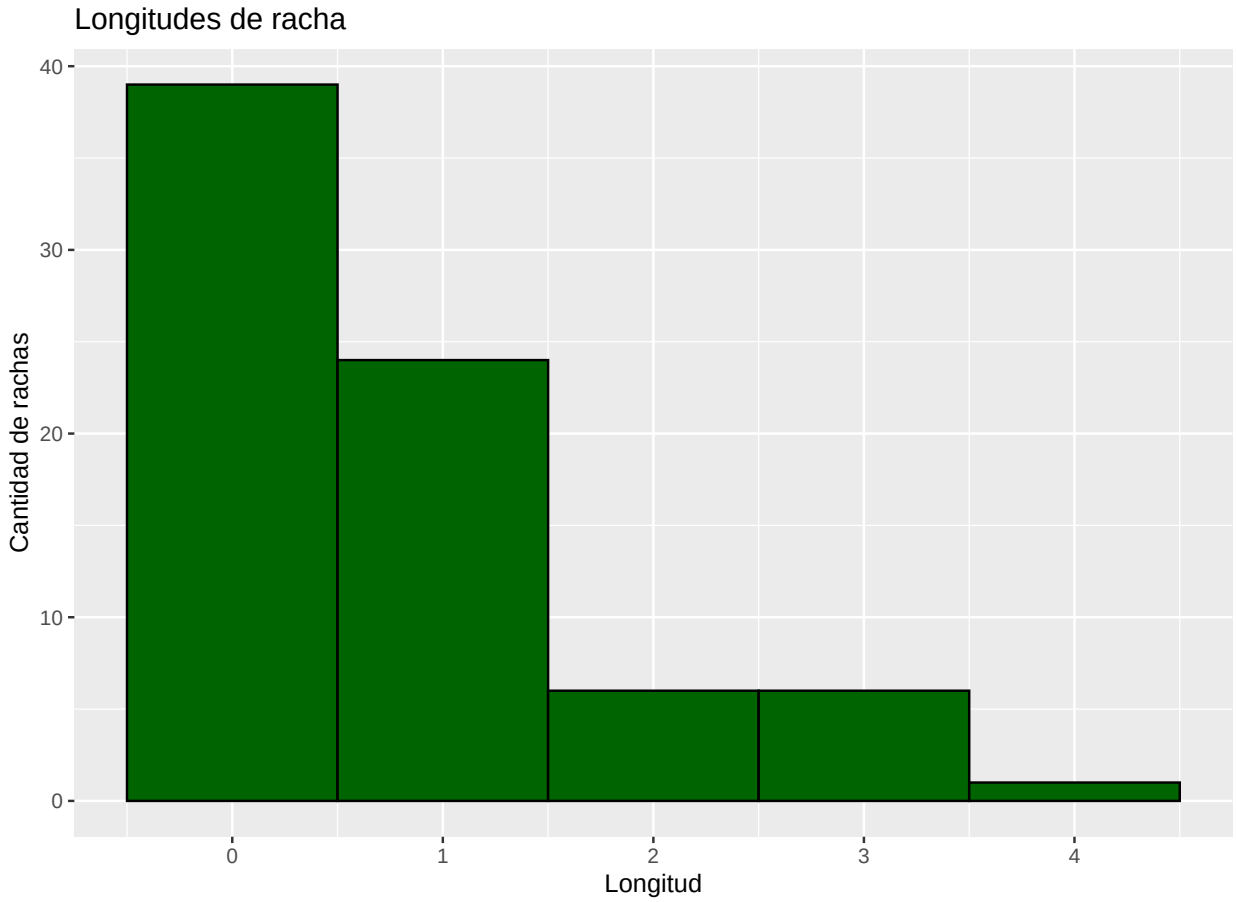
```
##    length
## 1       1
## 2       0
## 3       2
## 4       0
## 5       0
## 6       0
## 7       3
## 8       2
## 9       0
## 10      3
## 11      0
## 12      1
## 13      3
## 14      0
## 15      0
## 16      0
```

```
## 17      0
## 18      0
## 19      1
## 20      1
## 21      0
## 22      4
## 23      1
## 24      0
## 25      1
## 26      0
## 27      1
## 28      0
## 29      1
## 30      2
## 31      0
## 32      1
## 33      2
## 34      1
## 35      0
## 36      0
## 37      1
## 38      0
## 39      0
## 40      0
## 41      1
## 42      1
## 43      0
## 44      1
## 45      0
## 46      2
## 47      0
## 48      0
## 49      0
## 50      3
## 51      0
## 52      1
## 53      0
## 54      1
## 55      2
## 56      1
## 57      0
## 58      1
## 59      0
## 60      0
## 61      1
## 62      3
## 63      3
## 64      1
## 65      1
## 66      0
## 67      0
## 68      0
## 69      0
## 70      0
```

```
## 71      1
## 72      1
## 73      0
## 74      0
## 75      0
## 76      1
```

Luego podemos observar la distribución de estas longitudes de racha.

```
ggplot(data = kobe_streak, aes(x = length)) +
  geom_histogram(color="black", fill="darkgreen",binwidth = 1) +
  ggtitle(label="Longitudes de racha") +
  xlab("Longitud") +
  ylab("Cantidad de rachas")
```



¿Comparado con que?

Hemos demostrado que Kobe tuvo algunas rachas largas de tiros, pero ¿son lo suficientemente largas como para respaldar la creencia de que tenía las manos calientes? ¿Con qué podemos compararlos?

Para responder a estas preguntas, volvamos a la idea de *independencia*. Dos procesos son independientes si el resultado de uno no afecta el resultado del segundo. Si cada tiro que realiza un jugador es un proceso

independiente, haber acertado o fallado el primer tiro no afectará la probabilidad de acertar o fallar el segundo.

Un tirador con una mano caliente tendrá tiros que *no* son independientes entre sí. Específicamente, si el tirador hace su primer tiro, el modelo de mano caliente dice que tendrá una probabilidad *mayor* de hacer su segundo tiro.

Supongamos por un momento que el modelo de mano caliente es válido para Kobe. Durante su carrera, el porcentaje de veces que Kobe hace una canasta (es decir, su porcentaje de tiro) es aproximadamente del 45%, o en notación de probabilidad.

$P(\text{shot } 1 = \text{H}) = 0.45$

Verificando y redondeando a opinión personal

```
agg_shot <- kobe_basket %>%
  group_by(shot) %>%
  summarise(shot_p = n()/nrow(kobe_basket))

kable(agg_shot)
```

shot	shot_p
H	0.4360902
M	0.5639098

Si hace el primer tiro y tiene una mano caliente (*no* tiros independientes), entonces la probabilidad de que acierte su segundo tiro ascendería a, digamos, 60%

$ P(shot 2 = H , | , shot 1 = H) = 0.60 $

Como resultado de estas mayores probabilidades, se esperaría que Kobe tuviera rachas más largas. Compare esto con la perspectiva escéptica en la que Kobe *no* tiene la mano caliente, donde cada tiro es independiente del siguiente. Si acierta su primer tiro, la probabilidad de que acierte el segundo sigue siendo 0.45.

$ P(shot 2 = H , | , shot 1 = H) = 0.45 $

En otras palabras, hacer el primer tiro no afectó la probabilidad de que hiciera el segundo tiro. Si los tiros de Kobe son independientes, entonces tendría la misma probabilidad de acertar cada tiro independientemente de sus tiros anteriores: 45%.

Ahora que hemos planteado la situación en términos de tiros independientes, volvamos a la pregunta: ¿cómo sabemos si las rachas de tiros de Kobe son lo suficientemente largas como para indicar que tiene las manos calientes? Podemos comparar la duración de su racha con la de alguien sin manos calientes: un tirador independiente.

Simulemos ahora el resultado, en R el comando sample require de 3 parametros, el primero son las opciones a simular, la cantidad de elementos a esperados y si se permite con reemplazamiento. Simulando una moneda con 100 lanzamientos con reemplazamiento

```
coin_outcomes <- c("heads", "tails")
sim_fair_coin <- sample(coin_outcomes, size = 100, replace = TRUE)
table(sim_fair_coin)
```

```
## sim_fair_coin
## heads tails
##    51    49
```

```
sim_fair_coin
```

```
##   [1] "tails" "tails" "heads" "heads" "heads" "heads" "tails" "tails" "tails"
##  [10] "tails" "heads" "heads" "tails" "heads" "tails" "tails" "tails" "tails"
##  [19] "tails" "heads" "tails" "tails" "tails" "tails" "heads" "tails" "tails"
##  [28] "heads" "heads" "tails" "heads" "heads" "heads" "tails" "heads" "tails"
##  [37] "heads" "heads" "heads" "tails" "heads" "heads" "tails" "heads" "heads"
##  [46] "tails" "heads" "heads" "heads" "heads" "heads" "tails" "tails" "tails"
##  [55] "tails" "tails" "tails" "tails" "heads" "heads" "heads" "heads" "heads"
##  [64] "heads" "tails" "heads" "heads" "tails" "tails" "tails" "heads" "heads"
##  [73] "heads" "tails" "heads" "tails" "tails" "tails" "tails" "heads" "heads"
##  [82] "heads" "heads" "tails" "tails" "tails" "heads" "heads" "tails" "tails"
##  [91] "heads" "tails" "heads" "tails" "heads" "heads" "tails" "heads" "heads"
## [100] "tails"
```

Dado que sólo hay dos elementos en los "resultados", la probabilidad de que "lancemos" una moneda y salga cara es 0,5. Digamos que estamos tratando de simular una moneda injusta que sabemos que solo sale cara el 20% de las veces. Podemos ajustar esto agregando un argumento llamado "prob", que proporciona un vector de dos ponderaciones de probabilidad.

```
sim_unfair_coin <- sample(coin_outcomes, size = 100, replace = TRUE, prob = c(0.2, 0.8))
sim_unfair_coin
```

```
##   [1] "tails" "tails" "tails" "tails" "tails" "tails" "heads" "tails" "tails"
##  [10] "tails" "tails" "tails" "tails" "tails" "heads" "heads" "tails" "tails"
##  [19] "tails" "tails" "tails" "tails" "tails" "tails" "heads" "tails" "tails"
##  [28] "tails" "tails" "tails" "tails" "tails" "tails" "tails" "tails" "tails"
##  [37] "tails" "tails" "tails" "tails" "tails" "tails" "tails" "heads" "heads"
##  [46] "tails" "tails" "tails" "tails" "tails" "heads" "tails" "tails" "tails"
##  [55] "tails" "tails" "tails" "tails" "tails" "tails" "tails" "tails" "tails"
##  [64] "tails" "tails" "tails" "heads" "tails" "tails" "tails" "tails" "tails"
##  [73] "tails" "tails" "heads" "tails" "heads" "tails" "tails" "tails" "tails"
##  [82] "tails" "heads" "heads" "tails" "tails" "tails" "heads" "tails" "heads"
##  [91] "tails" "tails" "tails" "heads" "tails" "tails" "tails" "tails" "tails"
## [100] "tails"
```

```
table(sim_unfair_coin)
```

```
## sim_unfair_coin
## heads tails
##    15    85
```

```
sim_fair_coin
```

```
##   [1] "tails" "tails" "heads" "heads" "heads" "heads" "tails" "tails" "tails"
##  [10] "tails" "heads" "heads" "tails" "heads" "tails" "tails" "tails" "tails"
##  [19] "tails" "heads" "tails" "tails" "tails" "tails" "heads" "tails" "tails"
##  [28] "heads" "heads" "tails" "heads" "heads" "heads" "tails" "heads" "tails"
##  [37] "heads" "heads" "heads" "tails" "heads" "heads" "tails" "heads" "heads"
##  [46] "tails" "heads" "heads" "heads" "heads" "heads" "tails" "tails" "tails"
##  [55] "tails" "tails" "tails" "tails" "heads" "heads" "heads" "heads" "heads"
```

```
##  [64] "heads" "tails" "heads" "heads" "tails" "tails" "tails" "heads" "heads"
##  [73] "heads" "tails" "heads" "tails" "tails" "tails" "tails" "heads" "heads"
##  [82] "heads" "heads" "tails" "tails" "tails" "heads" "heads" "tails" "tails"
##  [91] "heads" "tails" "heads" "tails" "heads" "heads" "tails" "heads" "heads"
## [100] "tails"
```

Simular a un jugador de baloncesto que realiza tiros independientes utiliza el mismo mecanismo que utilizamos para simular un lanzamiento de moneda. Para simular un solo disparo desde un tirador independiente con un porcentaje de tiro del 50% escribimos,

```
shot_outcomes <- c("H", "M")
sim_basket <- sample(shot_outcomes, size = 133, replace = TRUE)
table(sim_basket)
```

```
## sim_basket
##  H  M
## 74 59
```

```
sim_basket
```

```
##   [1] "H" "H" "H" "H" "M" "H" "H" "M" "H" "H" "H" "H" "M" "M" "H" "M" "H" "M"
##  [19] "M" "M" "H" "M" "M" "M" "H" "M" "H" "M" "H" "M" "H" "H" "M" "M" "H" "M"
##  [37] "H" "M" "M" "M" "H" "M" "M" "H" "M" "H" "H" "M" "H" "M" "H" "H" "M" "M"
##  [55] "M" "M" "M" "H" "H" "H" "H" "M" "H" "H" "H" "H" "M" "H" "H" "M" "H" "H"
##  [73] "M" "M" "M" "M" "M" "M" "M" "H" "H" "H" "H" "H" "H" "H" "M" "M" "M" "M"
##  [91] "H" "M" "M" "H" "H" "H" "H" "M" "M" "M" "H" "H" "H" "M" "H" "H" "M" "H"
## [109] "H" "M" "M" "H" "H" "H" "H" "M" "M" "M" "H" "H" "M" "H" "H" "H" "H" "M"
## [127] "M" "H" "H" "H" "H" "H" "H"
```

Para hacer una comparación válida entre Kobe y nuestro tirador independiente simulado, necesitamos alinear tanto su porcentaje de tiros como el número de tiros intentados.

```
shot_outcomes <- c("H", "M")
sim_basket <- sample(shot_outcomes, size = 133, replace = TRUE, prob = c(0.45, 0.55))
table(sim_basket)
```

```
## sim_basket
##  H  M
## 46 87
```

```
sim_basket
```

```
##   [1] "H" "H" "M" "M" "M" "H" "M" "M" "M" "H" "M" "H" "H" "M" "H" "M" "M" "M"
##  [19] "M" "H" "M" "H" "M" "H" "M" "M" "M" "M" "M" "M" "M" "M" "H" "M" "H" "M" "M"
##  [37] "M" "M" "M" "M" "H" "M" "M" "H" "H" "M" "M" "H" "M" "M" "M" "H" "M" "M" "M"
##  [55] "M" "M" "M" "H" "M" "M" "H" "M" "M" "M" "M" "H" "H" "M" "H" "M" "M" "M" "H"
##  [73] "M" "H" "H" "M" "M" "M" "M" "H" "M" "M" "M" "M" "M" "M" "H" "M" "M" "M" "M"
##  [91] "H" "M" "H" "M" "M" "H" "M" "M" "M" "M" "M" "H" "H" "H" "M" "H" "M" "M" "M"
## [109] "H" "H" "H" "M" "H" "M" "M" "M" "M" "M" "M" "H" "H" "M" "M" "H" "H" "H" "M"
## [127] "M" "M" "M" "H" "H" "H" "M"
```

Tenga en cuenta que hemos llamado al nuevo vector "sim_basket", el mismo nombre que le dimos al vector anterior que refleja un porcentaje de disparo del 50%. En esta situación, R sobrescribe el objeto antiguo con el nuevo, así que asegúrese siempre de no necesitar la información en un vector antiguo antes de reasignar su nombre.

Con los resultados de la simulación guardados como `sim_basket`, tenemos los datos necesarios para comparar a Kobe con nuestro tirador independiente.

Ambos conjuntos de datos representan los resultados de 133 intentos de tiro, cada uno con el mismo porcentaje de tiro del 45%. Sabemos que nuestros datos simulados son de un tirador que tiene disparos independientes. Es decir, sabemos que el tirador simulado no tiene

```
sim_streak <- calc_streak(sim_basket)
sim_streak
```

```
##    length
## 1       2
## 2       0
## 3       0
## 4       1
## 5       0
## 6       0
## 7       1
## 8       2
## 9       1
## 10      0
## 11      0
## 12      0
## 13      1
## 14      1
## 15      1
## 16      0
## 17      0
## 18      0
## 19      0
## 20      0
## 21      0
## 22      1
## 23      1
## 24      0
## 25      0
## 26      0
## 27      0
## 28      0
## 29      1
## 30      0
## 31      2
## 32      0
## 33      1
## 34      0
## 35      1
## 36      0
## 37      0
## 38      0
## 39      0
```

```
## 40      0
## 41      1
## 42      0
## 43      1
## 44      0
## 45      0
## 46      2
## 47      1
## 48      0
## 49      0
## 50      1
## 51      2
## 52      0
## 53      0
## 54      0
## 55      1
## 56      0
## 57      0
## 58      0
## 59      0
## 60      1
## 61      0
## 62      0
## 63      0
## 64      1
## 65      1
## 66      0
## 67      1
## 68      0
## 69      0
## 70      0
## 71      3
## 72      1
## 73      0
## 74      0
## 75      3
## 76      1
## 77      0
## 78      0
## 79      0
## 80      0
## 81      2
## 82      0
## 83      3
## 84      0
## 85      0
## 86      0
## 87      3
## 88      0
```

Haga un gráfico de la distribución de longitudes de racha simuladas del tirador independiente. ¿Cuál es la duración típica de la racha de este tirador independiente simulado con un porcentaje de tiro del 45%? ¿Cuánto dura la racha más larga de canastas del jugador en 133 tiros?

```r
ggplot(data = sim_streak, aes(x = length)) +
  geom_histogram(color="black", fill="darkred",binwidth = 1) +
  ggtitle(label="Longitudes de racha simuladas") +
  xlab("Longitud") +
  ylab("Cantidad de rachas")
```