A2 - Analítica Descriptiva e Inferencial

Enunciado

Semestre 2020.2

Índice

Introducción

En esta actividad se realizará un análisis estadístico descriptivo e inferencial de los datos procesados en la actividad 1. Recordamos que el conjunto de datos usado en la actividad previa consistía en el conjunto de datos Fifa.csv, que se encuentra disponible en la plataforma Kaggle: https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global.

Este conjunto de datos contiene el estilo de juego del videojuego de consola Fifa 2017, así como estadísticas reales de los jugadores de futbol. El conjunto de datos contiene más de 17,500 registros y 53 variables.

Las principales variables que se usarán en esta actividad son:

- Name (Nombre del jugador)

- Nationality (Nacionalidad del jugador)
- National_Position (Posición de juego en equipo nacional).
- National_Kit (Número de equipación en equipo nacional)
- Club (Nombre del club)
- Club_Position (Posición de juego en club)
- Club_Kit (Número de equipación en club)
- Club_Joining (Fecha en la que empezó en el club)
- Contract_Expire (Año finalización del contrato)
- Rating (Valoración global del jugador, entre 0 y 100)
- Height (Altura)
- Weight (Peso)
- Preffered_Foot (Pie preferido)
- Birth_Date (Fecha de nacimiento)
- Age (Edad)
- Preffered_Position (Posición preferida)
- Work_Rate (valoración cualitativa en términos de ataque-defensa)
- Weak_foot (valoración de 1 a 5 de control y potencia de la pierna no preferida)
- Skill_Moves (valoración de 1 a 5 de la habilidad en movimientos del jugador)
- El resto de variables hacen referencia a atributos del jugador.

La descripción de los atributos se puede consultar en https://www.fifplay.com/encyclopedia. La descripción de las abreviaturas de la posición del jugador en el campo se puede consultar en https://www.dtgre.com/2016/10/fifa-17-position-abbreviations-acronyms.html.

Puesto que el resultado del preprocesado de los datos puede ser ligeramente distinto entre las distintas soluciones que habéis aportado, os suministramos el fichero preprocesado. Esta actividad se realizará con el fichero que os suministramos, independientemente del proceso de preprocesado que hayáis realizado en la actividad anterior. El nombre del fichero es **fifa_clean.csv**.

En esta actividad realizaremos un **análisis descriptivo e inferencial**. En especial, nos interesa investigar la puntuación del jugador (Rating) y otras variables como el control de pelota (Ball_Control) y la técnica (Dribbling). Asumimos que este conjunto de datos es una muestra representativa de los jugadores de la última década (población).

Nota importante a tener en cuenta para entregar la actividad:

- Es necesario entregar el archivo Rmd y el fichero de salida (PDF o html). El archivo de salida debe incluir: el código y el resultado de la ejecución del código (paso a paso).
- Se debe respetar la misma numeración de los apartados que el enunciado.
- No se pueden realizar listados completos del conjunto de datos en la solución. Esto generaría un documento con cientos de páginas y dificulta la revisión del texto. Para comprobar las funcionalidades del código sobre los datos, se pueden usar las funciones **head** y **tail** que sólo muestran unas líneas del fichero de datos.
- Se valora la precisión de los términos utilizados (hay que usar de manera precisa la terminología de la estadística).

- Se valora también la concisión en la respuesta. No se trata de hacer explicaciones muy largas o documentos muy extensos. Hay que explicar el resultado y argumentar la respuesta a partir de los resultados obtenidos de manera clara y concisa.
- No se puede compartir código entre compañeros ni copiar código de actividades anteriores. Cada estudiante debe encontrar su propia solución a las preguntas de la actividad.

1. Lectura del fichero

Leer el fichero **fifa_clean.csv**. Validar que los datos leídos son correctos. Si no es así, realizar las conversiones oportunas.

---

2. Rating de los jugadores

Nos interesa investigar los valores que toma la variable Rating en la población. Para ello, realizad un primer análisis visual de esta variable a partir de la muestra. Posteriormente, calculad el intervalo de confianza de la variable Rating de los jugadores. Seguid los pasos que se indican a continuación.

2.1. Análisis visual

Mostrad visualmente la distribución de la variable Rating. Usad el gráfico o gráficos que creáis más oportunos. Describid brevemente lo que se observa en los gráficos que representáis.

2.2. Intervalo de confianza

Calculad el intervalo de confianza de la variable Rating. A continuación, explicad el resultado y cómo se debe interpretar el resultado obtenido.

Nota: Los cálculos se deben realizar manualmente. No se pueden usar funciones de R que calculen directamente el intervalo de confianza. En cambio, sí se pueden usar funciones como **mean**, **sd**, **qnorm**, **pnorm**, **qt** y **pt**.

---

3. Diferencias entre jugadores

Existe una creencia que los jugadores zurdos tienen mejor control de la pelota que los diestros. Vamos a comprobar qué dicen los datos al respecto. Nos preguntamos si los jugadores zurdos tienen mejor control de pelota (**Ball_Control**), valoración (**Rating**) y mejor **Dribbling** que los diestros. Para ello, primero seleccionad los jugadores que no son porteros (los porteros tienen el valor **GK** -Goal Keeper- en Club_Position). Entonces, debéis obtener dos muestras. La primera muestra contiene todos los jugadores de campo (no porteros) zurdos (**Preffered_Foot** igual a Left). La segunda muestra contiene todos los jugadores de campo (no porteros) diestros (**Preffered_Foot** Right). Usad un nivel de confianza del 95 %.

Aspectos a tener en cuenta para resolver este ejercicio:

- Se deben realizar los cálculos manualmente. No se pueden usar funciones de R que calculen directamente el contraste como 't.test' o similar. Sí se pueden usar funciones como 'mean', 'sd', 'qnorm', 'pnorm', 'qt' y 'pt'. Sí podéis usar 'var.test' si lo necesitáis.

- Debido a que se preguntan las diferencias en tres variables, es aconsejable estructurar el código con una función, a la que se pasa como parámetro la variable a analizar. No deberías escribir el mismo código tres veces.

Seguid los pasos que se especifican a continuación.

3.1. Pregunta de investigación

Formulad la/s pregunta/s de investigación que se plantea/n en este apartado.

3.2. Representación visual

Representad visualmente, mediante el gráfico que sea más apropiado el valor de estas variables en jugadores de campo (no porteros) diestros y zurdos. Se deben mostrar los valores de forma comparativa entre zurdos y diestros. Interpretad los gráficos.

3.3. Hipótesis nula y alternativa

Escribid la/s hipótesis nula/s y la/s hipótesis alternativa/s.

3.4. Método

En función de las características de la muestra, decidid qué método aplicar para validar la hipótesis planteada. Para ello, debéis especificar como mínimo: a) si es un contraste de una muestra o de dos muestras (en caso de dos muestras, si éstas son independientes o están relacionadas), b) si podéis asumir normalidad y por qué, c) si el test es paramétrico o no paramétrico, d) si el test es bilateral o unilateral, e) si se puede asumir homocedasticidad o heterocedasticidad.

Justificad vuestras elecciones.

3.5. Cálculos

Calcular el estadístico de contraste, el valor crítico y el valor p.

3.6. Tabla de resultados

Incorporad una tabla de resultados con el formato que se indica a continuación. Adjuntamos el fragmento de código que hemos usado para generar la tabla para que podáis usar este mismo formato. Necesitáis usar la librería kableExtra (si es necesario, debéis instalarla previamente a su uso).

```
library(kableExtra)

out <- data.frame( var=c("Rating", "BallControl", "Dribbling"), mean_Left=c(0,0,0),
                   mean_Right=c(0,0,0), n_Left=c(0,0,0), n_Right=c(0,0,0), obs_value=c(0,0,0),
                   critical=c(0,0,0), pvalue=c(0,0,0))
out %>% kable() %>% kable_styling()
```

var	mean_Left	mean_Right	n_Left	n_Right	obs_value	critical	pvalue
Rating	0	0	0	0	0	0	0
BallControl	0	0	0	0	0	0	0
Dribbling	0	0	0	0	0	0	0

3.7. Interpretación

A partir de los resultados obtenidos, realizad la interpretación de los mismos, dando respuesta a las preguntas formuladas.

---

4. Comparación por pares

Nos preguntamos si obtendríamos el mismo resultado si comparásemos los **jugadores de campo zurdos con aquellos jugadores de campo diestros que tienen un peso, altura y edad similar**. Para dar respuesta a esta pregunta, realizaremos un proceso similar al denominado *propensity score matching*, aunque un poco simplificado. Realizaremos lo siguiente:

- Para cada jugador de campo zurdo, localizaremos el jugador de campo diestro más similar, en cuanto a peso, altura y edad.
- Para realizar esta búsqueda, debemos implementar un algoritmo del tipo vecino más cercano.
- Para calcular la similitud entre dos jugadores, nos basaremos en la función de distancia euclídea.

La función de distancia euclídea entre dos vectores es:

```
euclidean <- function( x1, x2  ){
   return ( sqrt( sum (  (x1-x2)^2 ) ) )
}
```

El resultado de este proceso de matching serán dos muestras:

- La muestra original está compuesta por el conjunto de jugadores de campo zurdos.
- La segunda muestra tendrá el mismo tamaño que la primera. En esta muestra hay, para cada jugador zurdo de la primera muestra, el jugador diestro con características físicas similares. Es decir, existe una correspondencia entre el jugador 1 de la muestra de zurdos, con el jugador 1 de la muestra de diestros, y así para todos los jugadores.

A partir de estas muestras, nos preguntaremos si los jugadores zurdos son mejores en Rating que los jugadores diestros relacionados.

Para poder realizar este análisis, calcularemos primero el jugador diestro más similar a cada jugador zurdo. Seguid el esquema que se especifica a continuación.

4.1. Jugador más similar

En primer lugar, implementad una función **my.nn** que, dado un jugador calcule el jugador más similar en términos de edad, peso y altura. La función debe ser:

```
  my.nn <- function( x, sample ){

  }
```

donde **x** es el jugador zurdo, de tipo vector que guarda la edad, peso y altura. Y **sample** es la muestra de diestros con valores en edad, peso y altura. Para calcular el ejemplo de **sample** más similar a **x**, usad la función de distancia euclídea que os hemos suministrado. La función **my.nn** devuelve el índice (posición) del jugador de la muestra **sample** que se parece más a **x**.

Una vez realizada esta función, podemos usarla para calcular, para cada jugador de la muestra de zurdos, el jugador diestro más similar. Para ello, implementad la función siguiente:

```
my.nn.sample <- function( sample1, sample2 ){
}
```

donde **sample1** es la muestra de jugadores zurdos y **sample2** es la muestra de jugadores diestros. Esta función devuelve la muestra **Right.paired**, que contiene el listado de jugadores diestros más similares a los jugadores zurdos de la muestra 1. En definitiva, esta función realiza una iteración sobre la muestra de zurdos. Para cada jugador de la muestra de zurdos, llama a la función **my.nn**.

Recomendamos que probéis la función con una muestra pequeña de datos para validar que el código es correcto.

4.2. Muestras

Llegados a este punto, tenemos dos muestras: **Left.sample**, con los jugadores de campo zurdos. Y **Right.paired**, que contiene los jugadores diestros más similares a los jugadores de la muestra **Left.sample**.

Mostrad las primeras filas de las dos muestras.

Nota: Puede que el algoritmo del vecino más cercano necesite bastante tiempo en computarse, ya que las muestras contienen muchos datos. Si es así, os recomendamos que apliquéis el cálculo sobre una muestra de 100 jugadores zurdos. Podéis también tomar una muestra de 200 jugadores diestros. Así evitaréis el exceso de tiempo computacional requerido y la actividad se considera igualmente válida.

4.3. Hipótesis nula y alternativa

A partir de las dos muestras, ¿podemos afirmar que los zurdos son mejores en Rating que los correspondientes jugadores diestros? Escribid la hipótesis nula y alternativa.

4.4. Método

Explicad el método que aplicaréis y el por qué de su elección.

4.5. Cálculos

Realizad los cálculos de: estadístico de contraste, valor crítico y valor p. Usad un nivel de confianza del 95 %.

Nota: Los cálculos se deben realizar manualmente. No se pueden usar funciones de R que calculen directamente el contraste. En cambio, sí se pueden usar funciones como **qnorm**, **pnorm**, **qt** y **pt**.

4.6. Interpretación

Interpretad el resultado obtenido.

4.7. Reflexión

Comparad el análisis realizado en el apartado 3, con el realizado en este apartado. Explicad también qué sentido puede tener realizar una comparación de este tipo en relación al contraste de la pregunta 3.

---

5. Comparación entre clubes

Es bien conocida la rivalidad entre los clubes de Barcelona y Madrid. Se desea calcular si el porcentaje de jugadores con un Rating superior a 90 es diferente en Barcelona y Madrid con un nivel de confianza del 97 %.

Para ello, seguid los pasos que se especifican a continuación.

5.1. Hipótesis nula y alternativa

Escribid la hipótesis nula y alternativa.

5.2. Método

Explicad qué método aplicaréis para dar respuesta a la pregunta formulada. Justificad vuestra elección.

5.3. Cálculos

Preparad las muestras y realizad los cálculos oportunos. Al igual que anteriormente, no podéis usar funciones que ya realicen este contraste automáticamente. Sí podéis usar **pnorm**, **qnorm**, **pt**, **qt**.

5.4. Resultados e interpretación

Escribid la interpretación de los resultados y dad respuesta a la pregunta formulada.

---

6. Resumen ejecutivo

Enumerad brevemente las conclusiones obtenidas en este análisis inferencial. Explicad qué resultados se han obtenido y con qué nivel de confianza podéis extraer las conclusiones.

Puntuación de la actividad

- Apartado 1 (10 %)
- Apartado 2 (10 %)
- Apartado 3 (20 %)
- Apartado 4 (20 %)
- Apartado 5 (20 %)

- Apartado 6 (10 %)
- Calidad del informe dinámico (10 %)