Análisis ANOVA y primeras conclusiones

María Isabel Ruiz Martínez

February 5, 2023

List of Tables

List of Figures

1 Descripción del dataset

Los datasets que se van a usar han sido recopilados por Luis Castillo Vidal y corresponden a la actividad de sus alumnos en la asignatura Desarrollo Basado en Agentes.

El primer dataset, tras haber sido filtrados los registros erróneos, consta de 47828 filas correspondientes a los diferentes acciones de unos drones en una serie de mundos virtuales. En cada registro se detallan los siguientes atributos:

- *year*: identifica el curso académico en el que se realizó dicha acción.
- *group*: grupo de prácticas que ha prograda al dron que acomete la acción.
- *date*: fecha en la que se lleva a cabo la acción.
- *map*: mundo virtual en el que se ha realizado la acción.
- *action*: indica el tipo de acción realizada.

En la Tabla 1 se presentan los primeros seis registros del dataset. Además, en la Tabla 2 puede apreciarse un resumen de los datos que tenemos.

	year	group	date	map	action
1	1516	Achernar	17/10/2015 19:41:45	0	0
2	1516	Bellatrix	17/10/2015 19:41:45	0	0
3	1516	Cerastes	17/10/2015 19:41:45	0	0
4	1516	Denebola	17/10/2015 19:41:45	0	0
5	1516	Elnath	17/10/2015 19:41:45	0	0
6	1516	Furud	17/10/2015 19:41:45	0	0

Table 1: Muestra del primer dataset

year	group	date	map	action
Min. :1516	Length:47828	Length:47828	Min. :0.000	Min. :0.000
1st Qu.:1516	Class :character	Class :character	1st Qu.:1.000	1st Qu.:1.000
Median :1617	Mode :character	Mode :character	Median :3.000	Median :2.000
Mean :1700			Mean :3.834	Mean :2.325
3rd Qu.:1920			3rd Qu.:6.000	3rd Qu.:3.000
Max. :1920			Max. :9.000	Max. :5.000

Table 2: Resumen del primer dataset que tenemos

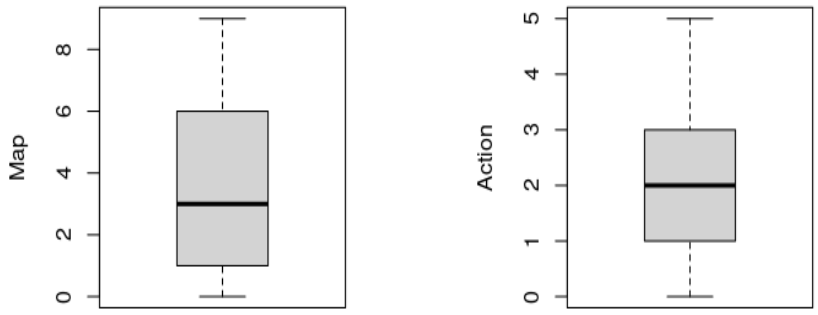
La Figuras 1a y 1b muestran, respectivamente, los gráficos de caja y bigotes de las variables *map* y *action*.

Además, podemos ver la distribución de los diferentes niveles del factor *map* y de los diferentes niveles del factor *action* en las Figuras 2a y 2b.

Por último, también podemos ver la distribución de las variables *map* y *action* en función del año (variable *year*) en las Figuras 3a y 3b.

El segundo dataset, tras haber sido eliminados algunas columnas que no eran interesantes para nuestro estudio, consta de 118 filas correspondientes a las diferentes calificaciones de los equipos en las dos prácticas realizadas en la asignatura Desarrollo Basado en Agentes. En cada registro se detallan los siguientes atributos:

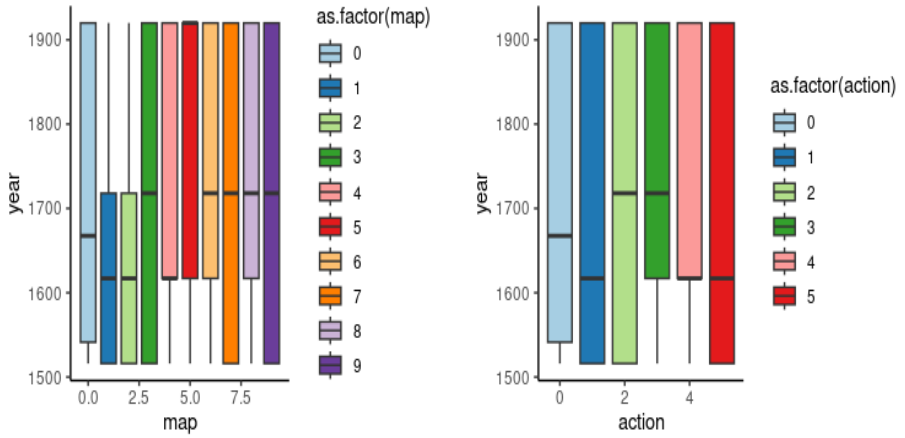
- *Group*: grupo de prácticas que ha prograda al dron.
- *Team*: cadena de texto que identifica el curso académico en el que se realizó dicha acción, la práctica realizada y el grupo de prácticas conjuntamente.



(a) Gráfico de caja y bigotes de la variable *map*

(b) Gráfico de caja y bigotes de la variable *action*

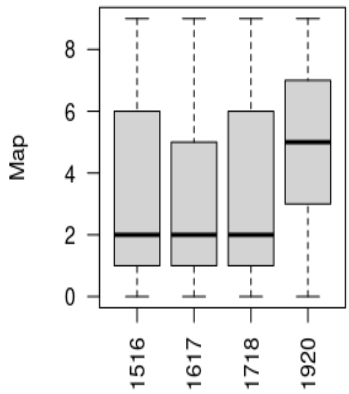
Figure 1: Diagramas de caja y bigotes de las variables *map* y *action*



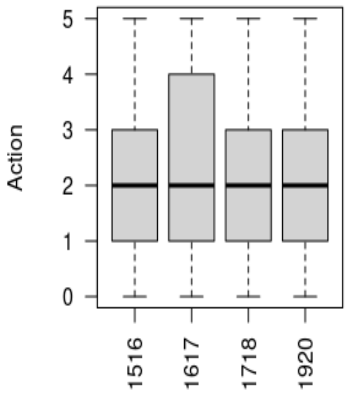
(a) Distribución de los diferentes niveles del factor *map*

(b) Distribución de los diferentes niveles del factor *action*

Figure 2: Distribuciones para diferentes niveles de los factores



(a) Distribución de la variable *map* en función de *year*



(b) Distribución de la variable *action* en función de *year*

Figure 3: Distribuciones de las variables *map* y *action* dependiendo del valor de *year*

- *Size*: tamaño del grupo de práctcas.
- *Year*: identifica el curso académico en el que se realizó dicha acción.
- *Grade*: calificación obtenida por el grupo de prácticas.

En la Tabla 3 se presentan los primeros seis registros del dataset. Además, en la Tabla 4 puede apreciarse un resumen de los datos que tenemos.

	Group	Team	Size	Year	Grade
1	G1	DBA 1819 P3 GL	4	1819	10,00
2	G2	DBA 1920 P3 GJ	4	1920	4,01
3	G3	DBA 1819 P2 GH	4	1819	7,96
4	G4	DBA 1920 P2 GE	4	1920	8,95
5	G5	DBA 1920 P3 GK	4	1920	4,51
6	G6	DBA 1415 P3 G6	6	1415	7,20

Table 3: Muestra del segundo dataset

2 Introducción

En este estudio inicial se desarrollará un modelo estadístico para determinar el efecto de los parámetros *map* y *action* (dos varibles explicativas) en la variable respuesta *year*.

La relevancia de cada una de las variables en el modelo se determinará por el test *two way ANOVA* con un 5% de nivel de significancia y se empleará la técnica de los *mínimos cuadrados* para estimar los coeficientes del modelo considerado.

Group	Team	Size	Year	Grade
Length:118	Length:118	Min. :3.000	Min. :1314	Length:118
Class :character	Class :character	1st Qu.:4.000	1st Qu.:1516	Class :character
Mode :character	Mode :character	Median :5.000	Median :1718	Mode :character
		Mean :4.831	Mean :1662	
		3rd Qu.:6.000	3rd Qu.:1819	
		Max. :6.000	Max. :1920	

Table 4: Resumen del segundo dataset que tenemos

3 Two way ANOVA

Un resumen de los resultados obtenidos al realizar el test two way ANOVA se muestra en la Tabla 5. Puede observarse que la variable *map* es significante al nivel 0, que la variable *action* es significante al nivel 0.01 y que la variable *map:action* (el término de interacción) no es significante. Así pues, puede concluirse que el dataset es homógeneo, es decir, las combinaciones *map:action* son estadísticamente iguales en todos los años considerados.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
map	1	62918413.58	62918413.58	2689.15	0.0000
action	1	101244.69	101244.69	4.33	0.0375
map:action	1	8797.03	8797.03	0.38	0.5398
Residuals	47824	1118946173.13	23397.17		

Table 5: Resultados del test two way ANOVA

La notación escalar del modelo ajustado al aplicar el test tiene la siguiente estructura:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_1 \cdot x_2 + \epsilon \tag{1}$$

donde β_0 es el intercepto, β_1 y β_2 son los coeficientes de los efectos principales, β_3 es el coeficiente del término de interacción, x_1 y x_2 son los parámetros sometidos a investigación (en este caso, x_1 representa el parámetro mapa y x_2 representa la acción), y representa el año y ϵ es el *término error*.

La Tabla 6 muestra los valores de los coeficientes de la fórmula que se han obtenido tras ajustar el modelo de regresión a los datos.

	x
(Intercept)	1655.03
map	12.31
action	-1.51
map:action	0.12

Table 6: Coeficientes del modelo

La Figura 4 muestra la interacción entre los parámetros mapa y acción. Así pues, puede observarse que todas las líneras de la gráfica siguen más o menos el mismo patrón, lo que evidencia que no hay una gran interacción entre ambos.

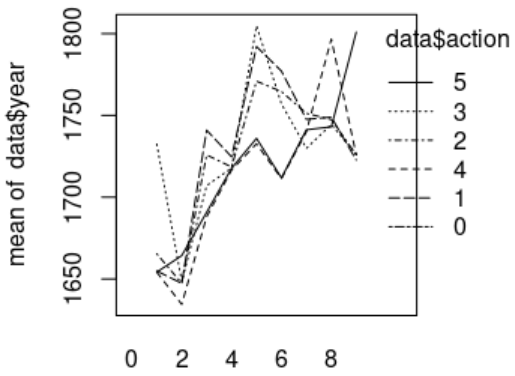


Figure 4: Interacción entre las variables *map* y *action*

La Figura 5 muestra que no se violan las suposiciones que hemos realizado sobre el modelo. La media y la varianza de los residuos no parece que varíe respecto de los valores ajustados. Como consecuencia, conluiré que podemos suponer la homocedasticidad. Además, si nos fijamos en el *Normal Q-Q plot*, puede observarse que los residuos son gaussianos.

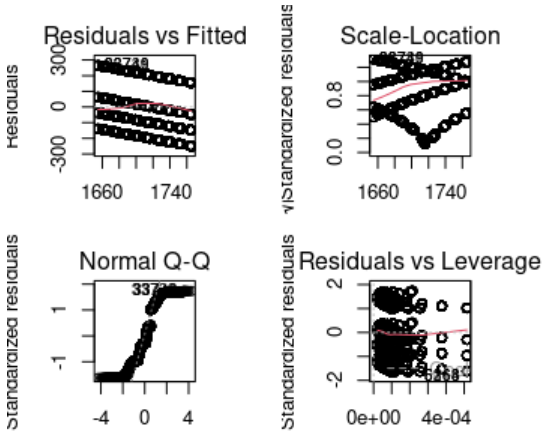


Figure 5: Gráficas diagnósticas del modelo ANOVA

4 Segmentación de los datos

Se ha realizado una segmentación de los registros del primer dataset en función de la calificación obtenida por los grupos que han realizado las acciones en los mapas.

Para ello, se ha sustituido la columna *Team* del segundo dataset por el nombre del grupo en cada caso y se ha eliminado la columna *Group*. Después de realizar este proceso el dataset tiene la forma que se muestra en la Tabla 7

	group	size	year	grade
1	Lesath	4	1819	10,00
2	Jabbah	4	1920	4,01
3	Haldus	4	1819	7,96
4	Elnath	4	1920	8,95
5	Keid	4	1920	4,51
6	Furud	6	1415	7,20

Table 7: Dataset 2 tras ser preparado para hacer join

A continuación, se transformarán los valores de la columna *grade* a tipo `double` y se agruparán los datos de este dataset por grupo y año, calculando la calificación media de las prácticas para cada par grupo-año. El resultado de realizar esta operación puede apreciarse ne la Tabla 8.

	group	year	size	mean_grade
1	Achernar	1314	6	9.40
2	Achernar	1415	5	9.00
3	Achernar	1516	5	8.20
4	Achernar	1617	5	10.00
5	Achernar	1718	5	8.33
6	Bellatrix	1516	5	8.20

Table 8: Dataset 2 tras agrupar por grupo y año

Ahora, se juntan ambos datasets usando la función `inner_join` por grupo y año. El resultado de esta operación puede observarse en la Tabla 9.

	group	year	size	mean_grade	date	map	action
1	Achernar	1516	5	8.2	17/10/2015 19:41:45	0	0
2	Achernar	1516	5	8.2	22/10/2015 17:29:21	1	1
3	Achernar	1516	5	8.2	22/10/2015 17:29:22	1	2
4	Achernar	1516	5	8.2	22/10/2015 17:29:39	1	3
5	Achernar	1516	5	8.2	22/10/2015 17:34:09	1	1
6	Achernar	1516	5	8.2	22/10/2015 17:34:10	1	2

Table 9: Unión de ambos datasets

Por último, se separarán los datos de esta última tabla en función de las calificaciones obtenidas por los diferentes grupos:

- **Suspenso:** calificación mayor o igual que 0 y menor que 5.
- **Aprobado:** calificación mayor o igual que 5 y menor que 7.
- **Notable:** calificación mayor o igual que 7 y menor que 9.
- **Sobresaliente:** calificación mayor o igual qye 9 y menor que 10.
- **Matrícula de Honor:** calificación igual a 10.

Una muestra de los datasets generados tras la segmentación puede apreciarse en las tablas 10, 11, 12, 13 y 14.

	group	year	size	mean_grade	date	map	action
1	Girtab	1617	6	4.67	15/11/2016 7:31:31	1	1
2	Girtab	1617	6	4.67	15/11/2016 7:31:53	1	1
3	Girtab	1617	6	4.67	15/11/2016 12:50:03	1	1
4	Girtab	1617	6	4.67	15/11/2016 12:50:28	1	1
5	Girtab	1617	6	4.67	15/11/2016 13:27:27	1	1
6	Girtab	1617	6	4.67	15/11/2016 13:27:58	1	1

Table 10: Muestra del dataset de los grupos **suspensos**. Consta de 2397 registros con 7 campos cada uno.

	group	year	size	mean_grade	date	map	action
1	Bellatrix	1920	4	6.835	05/11/2019 10:24:51	1	1
2	Bellatrix	1920	4	6.835	05/11/2019 10:24:51	1	2
3	Bellatrix	1920	4	6.835	05/11/2019 10:25:04	1	1
4	Bellatrix	1920	4	6.835	05/11/2019 10:25:26	1	2
5	Bellatrix	1920	4	6.835	05/11/2019 10:25:39	1	1
6	Bellatrix	1920	4	6.835	05/11/2019 10:25:45	1	2

Table 11: Muestra del dataset de los grupos **aprobados**. Consta de 7053 registros con 7 campos cada uno.

	group	year	size	mean_grade	date	map	action
1	Achernar	1516	5	8.2	17/10/2015 19:41:45	0	0
2	Achernar	1516	5	8.2	22/10/2015 17:29:21	1	1
3	Achernar	1516	5	8.2	22/10/2015 17:29:22	1	2
4	Achernar	1516	5	8.2	22/10/2015 17:29:39	1	3
5	Achernar	1516	5	8.2	22/10/2015 17:34:09	1	1
6	Achernar	1516	5	8.2	22/10/2015 17:34:10	1	2

Table 12: Muestra del dataset de los grupos con **notable**. Consta de 18840 registros con 7 campos cada uno.

	group	year	size	mean_grade	date	map	action
1	Bellatrix	1617	5	9.72	07/11/2016 11:13:31	1	1
2	Bellatrix	1617	5	9.72	07/11/2016 11:36:59	1	1
3	Bellatrix	1617	5	9.72	07/11/2016 11:41:07	1	1
4	Bellatrix	1617	5	9.72	07/11/2016 11:42:22	1	1
5	Bellatrix	1617	5	9.72	07/11/2016 11:45:37	1	1
6	Bellatrix	1617	5	9.72	07/11/2016 11:49:23	1	1

Table 13: Muestra del dataset de los grupos con **sobresaliente**. Consta de 17891 registros con 7 campos cada uno.

	group	year	size	mean_grade	date	map	action
1	Achernar	1617	5	10	09/11/2016 21:23:22	1	1
2	Achernar	1617	5	10	09/11/2016 21:25:11	1	1
3	Achernar	1617	5	10	10/11/2016 0:21:37	1	1
4	Achernar	1617	5	10	10/11/2016 0:22:50	1	1
5	Achernar	1617	5	10	10/11/2016 0:23:12	1	1
6	Achernar	1617	5	10	10/11/2016 0:25:53	1	1

Table 14: Muestra del dataset de los grupos con **matrícula de honor**. Consta de 1646 registros con 7 campos cada uno.