Plantilla de ficha de formación

Título	Modelos lineales generales: ANOVA
Palabras clave (metaetiquetas)	Análisis multivariante, variabilidad entre grupos y dentro de grupos, prueba de hipótesis, modelos lineales
Idioma	Español
Objetivos / Metas / Resultados de aprendizaje	**El objetivo de este módulo es presentar los conceptos básicos del Análisis de Varianza (ANOVA) de un factor y de dos factores, que puede entenderse como un modelo lineal básico** **Al final de este módulo podrás:** - **Entender cómo el ANOVA puede ser útil para probar si existen diferencias entre el valor medio de una variable continua en diferentes niveles de una o varias variables categóricas.** - **Comprender e identificar las condiciones requeridas para aplicar estas técnicas.** - **Realizar Análisis de Varianza de un factor y múltiple e interpretar los resultados obtenidos.**

Curso de formación:	
Alfabetización en ciencia de datos	
Módulo de visualización de datos y análisis visual	X
Introducción a la ciencia de datos para las ciencias humanas y sociales	
Ciencia de datos para siempre	
Periodismo de datos y storytelling	

Descripción	En este módulo de capacitación, se presentará el uso de modelos lineales básicos para comprender cómo las diferencias medias se pueden atribuir o no al efecto de las variables categóricas. El análisis presentado aquí es la base de la regresión lineal, que también considera el efecto de las variables continuas. Las técnicas descritas en este módulo de formación se limitan al caso de variables categóricas (cualitativas). En este sentido, puede abordar el contenido de este módulo como una introducción al Modelado lineal general (GLM) que

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

	utiliza solo factores categóricos para explicar la variabilidad en una variable (continua) de interés. El procedimiento aquí presentado se basa en descomponer la variabilidad total medida en la muestra en diferentes fuentes: algunas son residuales (o no explicadas por los factores considerados) mientras que otras provienen de una parte sistemática que se puede atribuir a las diferentes categorías de los factores categóricos.
Contenidos dispuestos en 3 niveles	**1. INTRODUCCIÓN** Las técnicas GLM presentadas aquí en forma de Análisis de Varianza (ANOVA) permiten responder a preguntas potencialmente interesantes. Algunos ejemplos: a. ¿Los trabajadores masculinos y femeninos de una región ganan el mismo salario medio anual? b. ¿Los alumnos de un curso que siguen diferentes métodos de enseñanza obtienen la misma nota media? c. ¿El consumo semanal medio de ciertos medicamentos es diferente entre grupos de edad y/o género? El ANOVA de un factor responde las preguntas 1 y 2, mientras que la pregunta 3 requiere un ANOVA de dos factores. Nuestro objetivo es probar el efecto de una variable independiente (*factor*) clasificada en k categorías (*niveles*) sobre una variable dependiente numérica (*variable de respuesta*), y se basa en la descomposición de la variabilidad total de la muestra. Podemos abordar este problema como una prueba de hipótesis estadística de una hipótesis nula (H0; nuestro p-valorredeterminado) versus una alternativa (H1; una visión alternativa). La prueba se formula en términos de las medias poblacionales de la variable de respuesta a través de los niveles de nuestro(s) factor(es). $H0: \mu_1 = \cdots =: \mu_k$ $H1: Al\ menos\ dos\ \mu_i\ diferentes$ Los supuestos requeridos para realizar la prueba ANOVA son: - Poblaciones normales: la distribución de la variable de respuesta en todos y cada uno de los niveles debe ser normal - Igualdad de varianzas: las varianzas de la variable de respuesta entre niveles deben ser las mismas

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

- Muestras independientes: los datos de la muestra en cada nivel del factor no están correlacionados con los otros datos de la muestra (recolectados de los otros niveles)

2. ANOVA DE UN FACTOR

2.1. El procedimiento

El procedimiento ANOVA con un factor se basa en la siguiente ecuación:

$$X_{ir} = \mu + \alpha_i + u_{ir}$$

donde X_{ir} es el valor de nuestra variable de respuesta para r individual en la categoría (nivel) i. Suponemos que este valor es la suma de tres efectos:

- Un valor medio (μ), común a todos los individuos y niveles
- Un término (α_i) que captura la influencia media de pertenecer al nivel i
- Un residuo (u_{ir}), que explica las variaciones aleatorias e incontroladas. Se supone que este residual se distribuye normalmente con media cero

La prueba ANOVA es equivalente a probar si los términos α_i son idénticos en los k niveles. Si no, habrá diferencias significativas en los medios.

Tomamos datos muestrales de X y descomponemos su variabilidad (dispersión alrededor de las medias muestrales) en dos partes:

a. El intragrupo (SSW) da cuenta de la variabilidad interna.
b. La variabilidad entre (SSB) explica las diferencias entre la media de cada muestra de grupo y la gran media.

La variabilidad total (SST) es la suma de SSW+SSB. Si SSB es mucho mayor que SSW, indica que hay diferencias significativas en las medias de los grupos. Por lo tanto, habrá diferencias significativas en las medias entre los niveles del factor.

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

Para comparar el peso relativo de SSB y SSW sobre la variabilidad total, los escalamos dividiéndolos por el número de grados de libertad, produciendo los valores MSB y MSW respectivamente.

$$d = \frac{MSB}{MSW} = \frac{\frac{1}{k-1}\chi^2_{k-1}}{\frac{1}{n-k}\chi^2_{n-k}} \sim F^{k-1}_{n-k}$$

Si se cumplen los supuestos requeridos, el estadístico (d) calculado como MSB⁄MSW se distribuye como un modelo F. Este estadístico permite tomar una decisión sobre la prueba: cuanto mayor sea su valor, mayor (relativamente) es la parte intermedia en comparación con la variabilidad interna.

Pero, ¿cómo podemos saber si d es alto o no? Al calcular el p-valor asociado a esta prueba: calculamos el p-valor (la probabilidad en la cola derecha de la distribución F relevante) y si este p-valor es bajo, rechazamos el valor nulo (es decir, hay diferencias significativas en la media de los niveles)

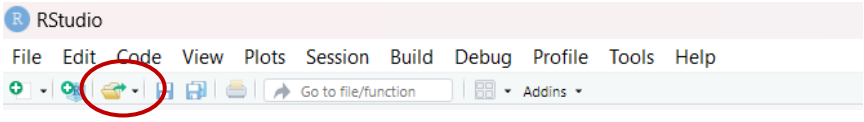
2.2. Un ejemplo

Como ejemplo ilustrativo, supongamos que queremos probar si el diseño de los envases en los que se vende una marca específica de leche tiene alguna influencia en las ventas. Con este objetivo, tomamos una muestra de 12 tiendas con características similares y, fijando el mismo precio para la leche, asignamos aleatoriamente un tipo de envase (1, 2 o 3). Luego obtenemos los datos de muestra de nuestra variable de respuesta "Ventas (*sales* en inglés), que mide cuántos miles de botellas de leche se vendieron en un mes, como se muestra a continuación:

Co-funded by the Erasmus+ Programme of the European Union
Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

	Sales	Package
1	2.2	1
2	2.5	1
3	2.4	1
4	2.6	1
5	3.1	2
6	2.8	2
7	3.2	2
8	3.3	2
9	2.5	3
10	2.8	3
11	3.2	3
12	2.5	3

Nuestros datos de muestra que se muestran arriba están contenidos en un archivo R, que podemos abrir yendo aquí (llamamos a este archivo de datos "Milk"):



Queremos probar si existen diferencias estadísticamente significativas en las ventas medias, dependiendo del diseño del envase. Estamos aplicando ANOVA con R, lo que requiere instalar paquetes específicos:

```
#install and load the relevant packages
install.packages("car")
install.packages("dplyr")
library(car)
library(dplyr)
```

Para aplicar ANOVA, primero debemos asegurarnos de que las supuestos requeridos realmente se cumplan, por lo que ejecutamos las siguientes líneas de código:

```
# test normality (by group)
Milk %>%
  group_by(Package) %>%
  summarise(statistic = shapiro.test(Sales)$statistic,
            p.value = shapiro.test(Sales)$p.value)
```

Estas líneas primero indican el conjunto de datos que se considera ("Milk"), luego agrupan los datos por los niveles del factor ("Package") y

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

finalmente ejecutan una prueba de normalidad de Spahiro en nuestra variable de respuesta ("Sales") entre grupos:

```
  Package statistic p.value
    <dbl>     <dbl>   <dbl>
1       1     0.971   0.850
2       2     0.927   0.577
3       3     0.854   0.241
```

Los altos p-valore de esta prueba de normalidad para todos los niveles nos permiten trabajar bajo el supuesto de normalidad requerido. Además, también asumimos que tenemos varianzas iguales, lo que nos lleva a ejecutar una prueba de Barlett de varianzas homogéneas como se muestra a continuación:

```
# test for homogeneous variances (by group)
bartlett.test(Milk$Sales, Milk$Package)
```

El p-valor que se muestra a continuación sugiere que esta suposición es muy realista:

```
        Bartlett test of homogeneity of variances

data:  Milk$Sales and Milk$Package
Bartlett's K-squared = 1.2076, df = 2, p-value = 0.5467
```

Dado que los supuestos necesarios parecen cumplirse, llevamos a cabo la metodología ANOVA ejecutando las siguientes líneas de código:

```
# run the ANOVA
anova(lm(Sales ~ Package, Milk))
```

Lo que produce la siguiente salida:

```
Analysis of Variance Table

Response: Sales
          Df  Sum Sq Mean Sq F value Pr(>F)
Package    1 0.21125 0.21125  1.6794 0.2241
Residuals 10 1.25792 0.12579
> 
```

Los resultados de la prueba ANOVA indican que los diferentes diseños de los envases parecen no tener impacto en las ventas medias: la parte de variabilidad explicada por los diferentes niveles del factor “Paquete” (variabilidad entre grupos) no es significativamente mayor que la parte residual (variaciones internas). Como consecuencia, el valor de p

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

asociado a esta prueba es alto y nos dice que no hay razones para rechazar la hipótesis nula de ventas medias iguales entre diseños.

3. ANOVA de dos factores

3.1 El procedimiento

Las ideas explicadas para el caso de ANOVA de un factor pueden extenderse para acomodar problemas en los que más de un factor puede estar afectando mi variable de respuesta. Ahora, la prueba ANOVA ahora se amplía para tener en cuenta un segundo factor más una posible interacción como:

$$X_{ijr} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + u_{ijr}$$

donde X_{ijr} es el valor de nuestra variable de respuesta para el individuo r en la categoría (nivel) i del factor α y el nivel j del factor β. Suponemos que estos valores se alejan de la media global (μ), como la suma de cuatro efectos:

a. Un desplazamiento (α_i) que captura la influencia media de pertenecer al nivel i del factor α
b. Un segundo desplazamiento (β_j) que captura la influencia media de pertenecer al nivel $j$de factor β
c. Un término de interacción entre estos dos factores $(\alpha\beta)_{ij}$
d. Un residual u_{ir}, que da cuenta de las variaciones aleatorias e incontroladas. Se supone que este residual se distribuye normalmente con media cero

Ahora las comparaciones entre las distintas partes de la variabilidad son más complejas. Cada fuente de variación se compara (convenientemente escalada por el número de grados de libertad) con la varianza residual. La intuición es la misma que en el ANOVA de un factor, pero hay tres pruebas diferentes, como se resume en la siguiente tabla:

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

SOURCE OF VARIATION	SUM OF SQUARES	d.f.	MEAN OF SQUARES	F
Factor α	SS_α	k-1	MS_α	MS_α/MSR
Factor β	SS_β	h-1	MS_β	MS_β/MSR
Interaction $(\alpha\beta)$	$SS_{\alpha\beta}$	(k-1) (h-1)	$MS_{\alpha\beta}$	$MS_{\alpha\beta}/MSR$
Residual	SSR	n-hk	MSR	
Total	SST	n-1		

3 .2. Un ejemplo

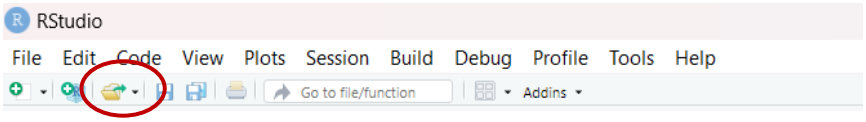
Vamos a ilustrar empíricamente el ANOVA de dos factores suponiendo que tenemos el siguiente problema: Un centro de salud quiere analizar la posible influencia de la edad y el sexo en el uso de un medicamento. Para ello se realiza una encuesta por muestreo y se agrupan los usuarios por edad en cuatro categorías (niños, adolescentes, adultos, adultos mayores) y género. Se extrae una muestra de 24 individuos, seleccionándose de forma independiente 3 individuos por sexo y grupo de edad. La variable de respuesta es el consumo mensual de este medicamento (en €), y tenemos el siguiente conjunto de datos:

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

	consumption	sex	age
1	3.0	Male	Child
2	4.0	Male	Child
3	2.8	Male	Child
4	3.2	Female	Child
5	3.0	Female	Child
6	4.1	Female	Child
7	1.8	Male	Teenager
8	1.0	Male	Teenager
9	1.5	Male	Teenager
10	2.1	Female	Teenager
11	1.2	Female	Teenager
12	1.7	Female	Teenager
13	2.5	Male	Adult
14	2.8	Male	Adult
15	3.0	Male	Adult
16	3.0	Female	Adult
17	4.0	Female	Adult
18	2.9	Female	Adult
19	5.0	Male	Senior
20	5.2	Male	Senior
21	6.0	Male	Senior
22	4.9	Female	Senior
23	5.1	Female	Senior
24	6.2	Female	Senior

Nuevamente, los datos de muestra que se muestran arriba (contenidos en un archivo R llamado "medicine") se pueden cargar en Rstudio yendo aquí:



Ahora, estamos aplicando un ANOVA de dos factores (edad y sexo) con R, que requiere instalar y cargar paquetes específicos:

```
#install and load the relevant packages
install.packages("car")
install.packages("dplyr")
library(car)
library(dplyr)
```

Para aplicar ANOVA, primero probamos si los supuestos requeridos realmente se cumplen, ejecutando pruebas de normalidad y de igualdad

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

de varianzas. Las pruebas de normalidad (en todos los grupos de edad y los dos géneros) se realizan ejecutando:

```
# we test normality by group first
Medicine %>%
  group_by(age,sex) %>%
  summarise(statistic = shapiro.test(consumption)$statistic,
            p.value = shapiro.test(consumption)$p.value)
```

Primero indicamos el conjunto de datos que se considera ("Medicine"), luego agrupamos los datos por los niveles de los dos factores considerados en nuestro análisis ("age" y "sex") y finalmente realizamos una prueba de normalidad de Spahiro en la variable "consumption". en todos los grupos:

```
  age      sex    statistic p.value
  <fct>    <fct>      <dbl>   <dbl>
1 Child    Male       0.871   0.298
2 Child    Female     0.881   0.328
3 Teenager Male       0.980   0.726
4 Teenager Female     0.996   0.878
5 Adult    Male       0.987   0.780
6 Adult    Female     0.818   0.157
7 Senior   Male       0.893   0.363
8 Senior   Female     0.862   0.274
```

Debe tenerse en cuenta que ahora, cuando nos referimos a los niveles de los dos factores, debemos considerar todos los pares de categorías posibles entre ellos. Nuevamente encontramos p-valores altos para esta prueba de normalidad en todos los casos, lo que nos permite trabajar bajo el supuesto de normalidad requerido. Además, también se requieren varianzas homogéneas y, en este caso, esta suposición se prueba realizando una prueba de Levene como:

```
#testing for equal variances
leveneTest(consumption ~ age*sex, data=Medicine, center="mean")
```

El p-valor encontrado indica que tampoco tenemos evidencia empírica en la muestra en contra de este supuesto:

```
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group  7  0.9575 0.4926
```

Dado que los supuestos necesarios para realizar un proceso ANOVA de dos factores parecen ser válidos, lo hacemos ejecutando las siguientes líneas de código:

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

```
# two factor ANOVA analysis
anova(lm(consumption ~ age*sex, Medicine))
```

El resultado del análisis viene en forma de la siguiente tabla ANOVA múltiple:

```
Analysis of Variance Table

Response: consumption
          Df Sum Sq Mean Sq F value    Pr(>F)
age        3 45.250 15.0833 51.8625 1.827e-08 ***
sex        1  0.327  0.3267  1.1232     0.305
age:sex    3  0.223  0.0744  0.2560     0.856
Residuals 16  4.653  0.2908
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los resultados de este ANOVA de dos factores proporcionan información muy útil que permite dar una respuesta basada en datos a nuestra pregunta de investigación. Las pruebas realizadas indican que los valores medios del consumo del medicamento son significativamente diferentes entre los cuatro niveles del factor "edad" (nótese que es el único caso en el que tenemos un p-valor bajo, lo que lleva a rechazar la hipótesis nula de medias iguales). Sin embargo, tampoco encontramos diferencias significativas en el consumo medio por sexo ni entre las interacciones entre grupo de edad y sexo.

Autoevaluación (preguntas y respuestas de opción múltiple)

En ANOVA de un factor, los residuos:

a) Se supone que están correlacionados
b) Se supone que son normales
c) No necesitamos ninguna suposición sobre los residuos.

La hipótesis nula en un ANOVA de un factor establece que:

a) Todos los medios son los mismos en todos los niveles.
b) Solo hay dos medios que son iguales
c) Todos los medios son diferentes.

El estadístico ANOVA de dos factores para probar la significación del factor α tiene una distribución:

a) Chi-cuadrado
b) t de Student
c) F de Snedecor

Recursos (videos, enlaces a referencias)

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

material relacionado	
PPT relacionado	
Bibliografía	NEWBOLD, P. et al. (2008): Statistics for Management and Economics, (6th edition) Ed. Prentice Hall. Capítulo 17, págs. 635-661.
Proporcionado por	[Unioví]

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.