Inferencia Estadística: Análisis de la Varianza(ANOVA)

Introducción a la Estadística

DVM

Contenidos

Introducción

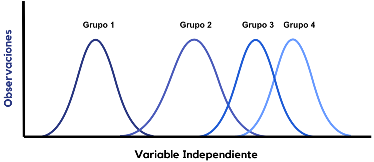
ANOVA

ANOVA por sus siglas en ingles : Analysis of Variance:

Es un procedimiento que usa las varianzas para extraer conclusiones sobre las medias poblaciones.

Supongamos que tenemos K-grupos independientes y queremos comparar sus medias como haciamos con el t-test. Si usaramos el anterior y fueramos dos a dos, acumularíamos mucho error. Por eso, hacemos un F-test a través de un ANOVA donde la hipótesis es:

$$H_o : \mu_1 = \mu_2 = \mu_3 = ... = \mu_k$$
$$H_1 : \text{Las medias de los k grupos no son iguales}$$



Introducción

ANOVA

Normalmente se usa para evaluar el efecto de un tratamiento , por ejemplo: tenemos un grupo de control y un grupo que ha recibido cierta medicación y queremos saber si la medicación ha hecho efecto, es decir, si la media de cierta v.a. de interés es diferente entre ambos grupos.

Por eso, también es habitual encontrar la siguiente formulación del modelo:

$$y_{ij} = \mu_i + E_{ij} \text{ con } E_{ij} \to N(0, \sigma^2)$$
$$\mu_i = \mu + \alpha_i$$

Donde μ_i es la media del i-nivel, μ es la media total y α_i es el efecto o la desviación inducida respecto de la media total por el nivel i.

H_0: Ningún nivel introduce un efecto sobre la media total: $\alpha_i = 0$ para toda i.

H_1 : Al menos un nivel introduce un efecto que desplaza su media: Algún $\alpha_i \neq 0$

Introducción

Anova

Nivel del Factor		Observaciones	
1	y_{11}	$\dots$	y_{1n_1}
2	y_{21}	$\dots$	y_{2n_2}
.	.	$\dots$	.
.	.	$\dots$	.
.	.	$\dots$	.
α	$y_{\alpha 1}$	$\dots$	$y_{\alpha n_\alpha}$

Introducción
ANOVA

Lo conseguimos por la descomposición de la varianza en una parte explicada por el factor(es) y otra aleatoria:
Suma de cuadrados Totales(SCT) = Suma de cuadrados Factor(SCF) + Suma Cuadrados Residuales (SCR)
$\sum_i^k \sum_j^n {}_i(y_{ij} - \bar{y})^2 = \sum_i^k n_i(\bar{y}_i - \bar{y})^2 + \sum_i^k \sum_j^n n_i(\bar{y}_{ij} - \bar{y}_i)^2$
total=entre+ intra

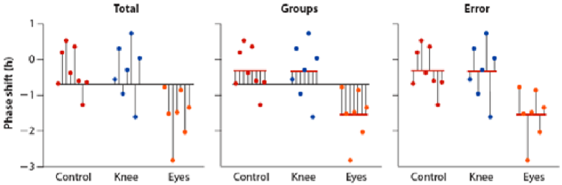


Figure 15.1-2 Depiction of the portions of variation in the circadian rhythm data (Example 15.1). Each dot is the measurement of phase shift in a single subject. The long horizontal line in black is the grand mean, $\bar{Y}$. Short horizontal lines in red are the sample means of the three light-treatment groups. Vertical lines represent deviations.

Contenidos

ANOVA de un factor

ANOVA de un factor o unidereccional evalúa la igualdad de las k-medias de la población para un resultado continuo y una única variable explicativa categórica (el factor) con cualquier número de niveles/tratamientos.

ANOVA de un factor

Tabla

Origen variación	Suma Cuadrados	g.l	Cuadrados Medios	Estadístico F
Tratamientos	$SC_F = \sum n_i(\bar{x}_i - \bar{x})^2$	n-r	$CM_F = SC_F/n - r$	$F = CM_F/CM_R$
Residual	$SC_R = \sum_i \sum_j n_i(\bar{x}_i j - \bar{x}_i)^2$	r-1	$CM_R = SC_R/r - 1$	
Total	$SCT = \sum_i \sum_j n_i(x_{ij} - \bar{x})^2$	n-1		

- $\bar{x}_i$: media de las observaciones del nivel i del factor (el tratamiento i)
- $\bar{x}$: media de todas las observaciones
- x_{ij}: valor de la observacion j del nivel i del factor
- n: numero total de observaciones
- r: numero de niveles del factor

ANOVA de un factor

Ejemplo1

Se realiza un ensayo clínico para comparar programas de pérdida de peso y los participantes son asignados aleatoriamente a uno de los programas de comparación. Los participantes siguen el programa asignado durante 8 semanas. El resultado de interés es la pérdida de peso (X), definida como la diferencia en el peso medido al comienzo del estudio el peso medido al final del estudio (8 semanas), medido en kgs.

ANOVA de un factor
Ejemplo1

Se consideran tres programas:

1. Dieta baja en calorias
2. Dieta baja en grasas
3. Dieta baja en carbohidratos
4. Grupo Control (evaluar el efecto placebo: la pérdida de peso debido simplemente a participar en el estudio)

Un total de veinte pacientes aceptan participar en el estudio y son asignados aleatoriamente a uno de los cuatro grupos de dieta. Los pesos se miden al inicio y se aconseja a los pacientes sobre la implementación adecuada de la dieta asignada. Las pérdidas de peso observadas despues de las ocho semanas se muestran a continuación.

ANOVA de un factor

Ejemplo1

Baja calorías	Baja grasa	Baja carbohidratos	Control
8	2	3	2
9	4	5	2
6	3	4	-1
7	5	2	0
3	1	3	3

¿Existe una diferencia estadísticamente significativa en la pérdida de peso promedio entre las cuatro dietas?

ANOVA de un factor

Ejemplo1

Ejecutaremos el ANOVA :

- Nuestra hipótesis y el nivel de significación:
  $H_o : \mu 1 = \mu 2 = \mu 3 = \mu 4$
  H_1: Las medias no son todas iguales al $\alpha = 0.05$
- El estadístico de prueba es el estadístico F para ANOVA, $F = CM_T/CM_R$ y el valor crítico de F vendría dado por $F_{alpha(3,16)} = 3,24$ así que rechzaríamos nuestra H_o si $F > 3,24$
- Tenemos que calcular nuestro estadístico F computando los valores de la tabla ANOVA:

ANOVA de un factor

Ejemplo1

Para calcular las sumas de cuadrados primero debemos calcular las medias muestrales para cada grupo y la media general basada en la muestra total.

	Baja calorías	Baja grasa	Baja carbohidratos	Control
	8	2	3	2
	9	4	5	2
	6	3	4	-1
	7	5	2	0
	3	1	3	3
$\bar{x}_i$	6,6	3,0	3,4	1,2

Media global=$\bar{x}$=3,6

Ahora, podemos calcular:

$SC_F = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + n_3(\bar{x}_3 - \bar{x})^2 + n_4(\bar{x}_4 - \bar{x})^2$

$= 5(6,6 - 3,6)^2 + 5(3 - 3,6)^2 + 5(3,4 - 3,6)^2 + 5(1,2 - 3,6)^2 = 75,8$

$SC_R = SCT - SC_F$

$SCT = (8 - 3,6)^2 + (9 - 3,6)^2 + ... + (0 - 3,6)^2 + (3 - 3,6)^2 = 123,2$

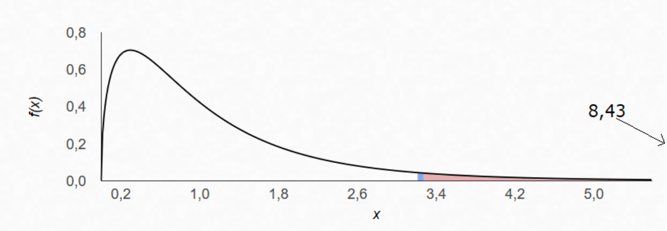
$SC_R = SCT - SC_F = 123,2 - 75,8 = 47,4$

ANOVA de un factor

Ejemplo1

Origen variación	Suma Cuadrados	g.l	Cuadrados Medios	Estadístico F
Tratamientos	$SC_F = 75,8$	16	$CM_F = 75,8/16$	$F = \frac{75,8/16}{47,4/3} = 8,43$
Residual	$SC_R = 47,4$	3	$CM_R = 47,4/3$	
Total	SCT=123,2	19		

Rechazamos H_0 porque $8,43 > 3,24$. Tenemos evidencia estadísticamente significativa con $\alpha = 0,05$ para mostrar que hay una diferencia en la pérdida de peso promedio entre las cuatro dietas.



ANOVA de un factor

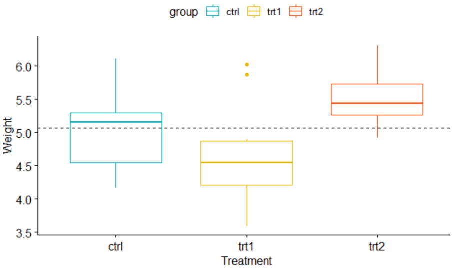
- ANOVA es una prueba que proporciona una evaluación global de una diferencia estadística en más de dos medios independientes.
- En este ejemplo, encontramos que hay una diferencia estadísticamente significativa en la pérdida de peso promedio entre las cuatro dietas consideradas.
- Además de informar sobre el resultadod el contraste es buena práctica informar sobre las mesdias muestrales de los distintos niveles para facilitar la interpretación de los resultads (esto es, los de la dieta baja en calorias bajaron en promedio 6,6 , los de una baja en grasas un 3,0 ...etc)

ANOVA de un factor

Ejemplo2

Vamos a usar el dataset **PlantGrowth** en R:
"Resultados de un experimento para comparar los rendimientos (medidos por el peso seco de las plantas) obtenidos bajo un control y dos condiciones de tratamiento diferentes."

	weight	group
1	6.11	ctrl
2	4.32	trt1
3	4.89	trt1
4	6.03	trt1
5	5.54	trt2
6	4.92	trt2
7	4.17	ctrl
8	4.61	ctrl
9	5.87	trt1
10	4.81	trt1



ANOVA de un factor

Práctica

Consideremos la propagación de un patógeno a través de los naranjos en Sevilla. Obtenemos una muestra aleatoria independiente de 3 naranjas de 4 naranjos y puntúamos el número de infecciones fúngicas por naranja. Los datos son los siguientes:

Naranjo a	11	10	9
Naranjo b	9	8	7
Naranjo c	5	7	6
Naranjo d	5	3	4

Con estos datos realiza un ANOVA, interpreta y, en caso de rechazar H_o, explica qué parejas son estadísticamente diferentes mediante un test de Bonferroni o Tukey.

Contenidos

ANOVA de dos factores

- A veces estamos interesados y disponemos de datos de dso factores y queresmos ver como estos afectan a una variable respuesta continua no sólo a nivel individual si no que también en su interacción.
  - Por ejemplo: queremos saber cómo el género, el gremio en el que se trabaja y la interacción de ambos afecta al salario.
- Se trata de testear simultaneamenet tres hipótesis:

H_o: Las medias de las observaciones agrupadas por el Factor A son iguales;
Las medias de las observaciones agrupadas por el Factor B son iguales;
No hay interacción entre los dos factores.
H_1: Al menos una de las medias del Factor A es diferente.
Al menos una de las medias del factor B es diferente
Si hay interacción entre los factores.

ANOVA de dos factores

- Estructura de los datos:

		Factor B			
		1	2	...	b
Factor A	1	$y_{111}, y_{112}, \ldots, y_{11n}$	$y_{121}, y_{122}, \ldots, y_{12n}$	$\cdots$	$y_{1b1}, y_{1b2}, \ldots, y_{1bn}$
	2	$y_{211}, y_{212}, \ldots, y_{21n}$	$y_{221}, y_{222}, \ldots, y_{22n}$	$\cdots$	$y_{2b1}, y_{2b2}, \ldots, y_{2bn}$
	.	.	.		.
	.	.	.		.
	.	.	.		.
	a	$y_{a11}, ya12, \ldots, y_{a1n}$	$y_{a21}, y_{a22}, \ldots, y_{a2n}$	$\cdots$	$y_{ab1}, y_{ab2}, \ldots, y_{abn}$

ANOVA de dos factores

$$y_{ijk} = \mu + \alpha_i + \beta_i + \gamma_{ij} + E_{ijk} \text{ con } E_{ijk} \to N(0, \sigma^2)$$

i: nivel i del Factor A

j: nivel j del Factor B

k: nivel k de las n observaciones por celda

α_i: efecto Factor A

β_i: efecto Factor B

γ_{ij}: efecto interacción de AB

ANOVA de dos factores

Tabla

	Origen de la variación	Suma de Cuadrados	df	Cuadrados medios	F
Factor A	Entre Filas (SSR)	$nc\sum_i^r(\bar{y}_i - \bar{y})^2$	r-1	MSR=SSR/(r-1)	$F_R = MSR/MS$
Factor B	Entre columnsa (SSC)	$nr\sum_j^c(\bar{y}_j - \bar{y})^2$	c-1	MSC= SSC/(c-1)	$F_C = MSC/MS$
AB	Interacción (SSI)	$n\sum_i^r\sum_j^c(\bar{y}_{ij} - \bar{y}_i - \bar{y}_j - \bar{y})^2$	(r-1)(c-1)	MSI=SSI/(c-1)(r-1)	$F_I = MSI/MSE$
	Error (SSE)	$\sum_k^n\sum_i^r\sum_j^c(y_{ijk} - \bar{y}_{ij})^2$	rc(n-1)	MSE=SSE/rc(n-1)	
	Total (SST)	$\sum_k^n\sum_i^r\sum_j^c(y_{ijk} - \bar{y})^2$	nrc-1		

n: número de observaciones por cada celda
r: número total de filas (de niveles para Factor A)
c: número total de columnas (de niveles para Factor B)
i: fila i de las r
j: columna j de las c
k: observación k de las n
y_{ijk} : *observaciónindividual*
$_{ij}$: media de la celda
$\bar{y}_i$: media de la fila i
$\bar{y}_j$:media de la columna j
$\bar{y}$: media total

ANOVA de dos factores

Ejemplo1 [1]

Tenemos los siguiente datos donde la variable categórica 1 toma los niveles 1,2,3 y la variable categórica 2 toma los niveles A,B y C. Para la variable respuesta tenemos una variable continua en la que recogemos 3 observaciones por cada celda:
Imaginemos que la variable 1, 2 y 3 son cualquier variable categoricas y continuas de interés, respectivamente.

	A	B	C	D
1		Group 1	Group 2	Group 3
2	Group A	10	10	9
3		7	2	6
4		10	10	9
5	Group B	1	4	4
6		1	2	1
7		9	4	9
8	Group C	1	5	10
9		4	7	10
10		4	5	10

[1] gracias a https://mattchoward.com

ANOVA de dos factores

Ejemplo1

	A	B	C	D	E	F	G
28							
29	ANÁLISIS DE VARIANZA						
30	*Origen de las variaciones*	*Suma de cuadrados*	*Grados de libertad*	*Promedio de los cuadrados*	*F*	*Probabilidad*	*Valor crítico para F*
31	Muestra	80,5185185	2	40,2592593	5,12735849	0,01728359	3,55455715
32	Columnas	29,8518519	2	14,9259259	1,9009434	0,17824027	3,55455715
33	Interacción	52,1481481	4	13,037037	1,66037736	0,20286148	2,92774417
34	Dentro del gru|	141,333333	18	7,85185185			
35							
36	Total	303,851852	26				
37							

a) $.017 < .05$: La Variable 1 (filas) tiene un efecto significativo y hay una diferencia notable entre los grupos A,B,C.

b) $.178 > .05$: La Variable 2 (columnas) no tiene un efecto significativo y no hay una diferencia notable entre los grupos 1,2,3.

c) $.203 > .05$ No hay interacción significativa entre las Variables 1 y 2.

En conclusión: sabemos que hay un efecto significativo de la Variable 1 pero no de la Variable 2 y de la interacción. Por esto, sabemos que hay alguna diferencia entre los grupos A,B y C pero no sabemos cuál.

ANOVA de dos factores

Ejemplo1

En R:

```
              Df Sum Sq Mean Sq F value Pr(>F)  
var1           2  80.52   40.26   5.127 0.0173 *
var2           2  29.85   14.93   1.901 0.1782  
var1:var2      4  52.15   13.04   1.660 0.2029  
Residuals     18 141.33    7.85                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA de dos factores

Ejemplo1

Y de ahí podemos ir más allá:

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Obs ~ var1 + var2 + var1 * var2, data = datos)

$var1
                      diff       lwr        upr     p adj
Group B-Group A -4.222222 -7.593451 -0.8509934 0.0131948
Group C-Group A -1.888889 -5.260118  1.4823399 0.3471533
Group C-Group B  2.333333 -1.037895  5.7045621 0.2089430
```