

ANOVA – Parecido a diferencia de medias

Prueba de análisis de la varianza: ANOVA

Recordemos primeramente el caso cuando $K = 2$ poblaciones, para realizar la **comparación de sus medias**, calculábamos las medias de dos muestras y las comparábamos usando el **estadístico de contraste** correspondiente mediante una prueba de hipótesis de diferencia de medias.

Si se desea hacer lo mismo con $K \geq 3$, se tendrían que comparar $\binom{K}{2}$ parejas de medias. Por ejemplo, para el caso específico de $K = 3$ se tendrían las siguientes comparaciones:

$$\begin{aligned}\mu_1 &= \mu_2 \\ \mu_1 &= \mu_3 \\ \mu_2 &= \mu_3\end{aligned}$$

Lo que requeriría de 3 pruebas de hipótesis diferentes. Conforme K aumente, la cantidad de comparaciones también incrementa significativamente.

En este caso se busca un test que nos permita saber, de un solo paso, si todas las medias son iguales (**ANOVA**) y en caso de haber diferencias, se tendría que abordar la búsqueda de esas parejas que no sean iguales en una segunda fase.

ANOVA: Analysis of Variance, por sus siglas en Inglés, hace referencia al análisis de la varianza.

El problema que intentamos resolver es un **contraste de igualdad de medias** cuando tenemos **más de dos poblaciones**. Concretamente, supongamos que $k > 2$ poblaciones

EJEMPLO:

Si se desea estudiar el peso de una población de estudiantes del Tec pongamos entre 17 y 40 años, podemos segregar por tipo de sangre, A, B, O y AB ($k=4$) y preguntarnos si el peso medio de cada subpoblación segregada por el tipo de sangre es el mismo o no.

Más concretamente, sean μ_1, \dots, μ_k las medias de esta magnitud (peso en el ejemplo anterior) en cada una de las subpoblaciones o poblaciones. Nos planteamos el contraste siguiente:

$$\begin{cases} H_0: \mu_1 = \mu_2 = \dots = \mu_k \\ H_1: \exists i, j | \mu_i \neq \mu_j \end{cases}$$

La decisión final, se tomará en función de la muestra aleatoria de cada población o subpoblación.

El **test ANOVA** realiza la comparación de las medias de 3 o más poblaciones basándose en la variabilidad de los datos por grupos:

- 1. SST: Variabilidad de los datos** (respecto de la media global)
- 2. SSE: Variabilidad dentro de cada población** (respecto de la media dentro de la población),
- 3. SSA: Variabilidad de las medias por poblaciones** (respecto de la media global).

La idea del **test ANOVA** es la siguiente: si la **variabilidad total de los datos** es explicada por la **variabilidad de las medias de las poblaciones** y la **poca "variabilidad" dentro de cada población**, es indicio que las medias son diferentes.

Sea:

X: Una variable cuantitativa

Y: variable cualitativa que consta de k atributos

Así, se tienen k poblaciones o subpoblaciones, donde cada atributo $i = 1, \dots, k$ define una población o subpoblación de la siguiente manera:

X_i : variable X restringida al atributo $i = 1, \dots, k$

Así se tienen las siguientes k poblaciones o subpoblaciones, que en adelante llamaremos tratamientos: X_1, X_2, \dots, X_k

Debe entenderse que cada X_i representa un conjunto de datos, es decir, la muestra de observaciones correspondiente al tratamiento i , a manera de ejemplo, se muestra la siguiente tabla:

Datos generales para un ANOVA

X_1 $j = 1, 2, \dots, n_1$	X_2 $j = 1, 2, \dots, n_2$...	X_k $j = 1, 2, \dots, n_k$
y_{11}	y_{21}	...	y_{k1}
y_{12}	y_{22}	...	y_{k2}
:	:	...	:
y_{1j}	y_{2j}	...	y_{kj}
:	:	...	:
y_{1n_1}	y_{2n_2}		y_{kn_k}

La tabla no debe confundir al lector, pensando que cada grupo de tratamiento podría tener la misma cantidad de datos, eso dependerá de si el valor n_i es igual en cada tratamiento o no.

Supuestos

Debe tomar en cuenta para un **test ANOVA** los siguientes supuestos para X_1, X_2, \dots, X_k :

- Son normales
- Independientes
- Varianzas similares
- Medias poblacionales μ_1, \dots, μ_k

Recordemos las hipótesis para dos variables X e Y :

H_0 : X no varía según los atributos de Y , es decir se cumple $\mu_1 = \mu_2 = \dots = \mu_k$

H_1 : X varía según los atributos de Y , al menos dos medias no son iguales.

Para contrastar la hipótesis, se toma una muestra de tamaño N :

$$N = \sum_{i=1}^k n_i$$

Dónde:

n_i : número de observaciones del tratamiento i – ésmo

y_{ij} : la j – ésima observación del tratamiento i – ésmo

$T_i = \sum_{j=1}^{n_i} y_{ij}$: suma de las observaciones en el tratamiento i – ésmo

$\bar{y}_i = \frac{T_i}{n_i}$: promedio de las observaciones en el tratamiento i – ésmo

$T = \sum_{i=1}^k T_i$: suma total de las observaciones

$\bar{y} = \frac{T}{n}$: promedio total o general observado

Para probar $H_0: \mu_1 = \dots = \mu_k$

Estadístico de prueba: $F = \frac{S_1^2}{S^2}$ con $(k-1, N-k)$ g.l.

donde

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}^2 - \frac{T^2}{N}$$

$$SSA = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^k \frac{T_i^2}{n_i} - \frac{T^2}{N}$$

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = SST - SSA$$

$$s_1^2 = \frac{SSA}{k-1} \text{ y } s^2 = \frac{SSE}{N-k}$$

Reglas de decisión: Las pruebas de ANOVA son de cola derecha

	RA	RR
0	$1 - \alpha$	f_c

Dónde $f_c = f_{1-\alpha, v_1, v_2} = f_{1-\alpha, k-1, N-k}$ y $ValorP = P(F > f_{obs})$

También se puede meter $F_{\chi^2, n-1, N-1}$ con
cola derecha

Ejemplo 5:

Se quiere determinar si la dosis de determinado tratamiento (Baja, Media, Alta) influye en el tiempo de sueño (en minutos) de los que la consumen.

Se plantea el problema de determinar si los tiempos de sueño varían o no, según la dosis del medicamento. ¿Puede afirmarse que los tiempos de sueño no varían según la dosis de medicamento?, use un nivel de significancia de 5%.

Para responder al problema planteado, se cuenta con los siguientes datos tomados de varios voluntarios expuestos a la aplicación de 3 dosis de medicamento (Alta, Media, Baja), registrando el número de minutos que duerme, una vez injerido la dosis:

Sea

X_1 : Tiempo de sueños de los que usan tratamiento

μ : Dosis de medicamento

Sea u_1, u_2, u_3 los tiempos promedio de sueño en cada dosis (Baja, media, Alta)

$H_0: u_1 = u_2 = u_3$ Todas iguales

$H_1: \exists i, j | u_i \neq u_j$ Al menos 2 diferentes

Tabla inicial

Dosis

Baja	Media	Alta	$\rightarrow k=3$
67	96	79	
69	98	29	
72	130	15	
79	65	33	
		17	

$$\frac{96 + 98 + 130 + 65}{4}$$

$$\frac{79 + 29 + 15 + 33 + 17}{5}$$

$$79 + 29 + 15 + 33 + 17 = 163$$

Dosis

Baja	Media	Alta
67	96	79
69	98	29
72	130	15
79	65	33
		17

Para contrastar la hipótesis, se toma una muestra de tamaño N :

$$N = \sum_{i=1}^k n_i$$

Dónde:

- n_i : número de observaciones del tratamiento i - ésmo
- y_{ij} : la j - ésmia observación del tratamiento i - ésmo
- $T_i = \sum_{j=1}^{n_i} y_{ij}$: suma de las observaciones en el tratamiento i - ésmo
- $\bar{y}_i = \frac{T_i}{n_i}$: promedio de las observaciones en el tratamiento i - ésmo
- $T = \sum_{i=1}^k T_i$: suma total de las observaciones
- $\bar{y} = \frac{T}{n}$: promedio total o general observado

$n = 8$

$n = 9$

$$96 + 98 + 130 + 65 = 389$$

$$67 + 69 + 72 + 79 = 287$$

7

n_i	4	4	5
T_i	287	389	163
T_i^2	$287^2 = 20592,25$	$389^2 = 37830,25$	$163^2 = 5313,8$
n_i	4	4	5
\bar{y}_i	71,75	97,25	32,6

$$N = 4 + 4 + 5 = 13$$

$$T = 287 + 389 + 163 = 839$$

$$\bar{y} = \frac{T}{N} = \frac{839}{13} = 64,5385$$

Baja

$$y_{ij}^2 = 67^2 + 69^2 + 72^2 + 79^2 +$$

media

$$96^2 + 98^2 + 130^2 + 65^2 +$$

Alta

$$74^2 + 29^2 + 15^2 + 33^2 + 17^2 = 68275$$

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}^2 - \frac{T^2}{N}$$

$$SSA = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^k \frac{T_i^2}{n_i} - \frac{T^2}{N}$$

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = SST - SSA$$

$$s_1^2 = \frac{SSA}{k-1} \text{ y } s^2 = \frac{SSE}{N-k}$$

$$SST = \frac{\sum y_{ij}^2 - T^2}{N} = \frac{68275 - 839^2}{13} = 19127,2308$$

$$SSA = \sum_{i=1}^k \frac{T_i^2}{n_i} - \frac{T^2}{N} = (20592,25 + 37830,25 + 5313,8) - \frac{839^2}{13}$$

$$SSA = 9588,53077$$

$$SSE = SST - SSA = 19127,2308 - 9588,53077$$

$$SSE = 7538,7$$

Estadístico de prueba: $F = \frac{S_1^2}{S_2^2}$ con $(k-1, N-k)$ g.l.

$$s_1^2 = \frac{SSA}{k-1} \text{ y } s^2 = \frac{SSE}{N-k}$$
$$\begin{array}{l} SSA = 98853077 \\ SSE = 75387 \\ N = 13 \\ k = 3 \end{array}$$

$$S_1^2 = \frac{98853077}{3-1} \quad S^2 = \frac{75387}{13-3}$$

$$F_{obs} = \frac{S_1^2}{S^2} = \frac{\frac{98853077}{3-1}}{\frac{75387}{13-3}} = 10,5631$$

$$F_c = f_{0.05, 2, 10} = 7.10282$$

$$V = k-1, N-k$$

$$V = 3-1, 13-3$$

$$V = 2, 10$$

II Como $f_{0.05} = 10,5631 > f_c = 7,10282$
Se rechaza H_0 , No son independientes

Valor P

$$P(F > 10,5631) = 0,00372 \quad \alpha = 0,05$$

II Como $P = 0,00372 < \alpha = 0,05$
Se rechaza H_0 , No son independientes

Cada año, los miembros del equipo de atletismo de una universidad se dividen al azar en tres grupos que entran con métodos diferentes. El primer grupo realiza largos recorridos a ritmo pausado, el segundo grupo realiza series cortas de alta intensidad y el tercero trabaja en el gimnasio con pesas y en bicicleta estacionaria con pedaleo de alta frecuencia. Despues de un mes de entrenamiento se realiza un test de rendimiento de la prueba de 110m con vallas. Seguidamente se muestran los tiempos obtenidos en el test en una muestra de 11 miembros:

<i>MetodoI</i>	<i>MetodoII</i>	<i>MetodoIII</i>
15	14	13
16	13	12
14	15	14
15	16	

A un nivel de significancia del 5%, ¿Puede considerarse que los tres métodos producen resultados equivalentes? (8 puntos)

$$H_0: u_1 = u_2 = u_3 \quad \text{metodo}_i = M_i, i = 1, 2, 3$$

$$H_1: \exists i, j | u_i \neq u_j$$

<i>M₁</i>	<i>M₂</i>	<i>M₃</i>	<i>k = 3</i>
15	14	13	
16	13	12	
19	15	17	
15	16		

<i>h_i</i>	4	4	3	<i>N = 11</i>
<i>T_i</i>	60	58	39	<i>T = 157</i>
<i>T_i²</i>	900	841	507	<i>Ȳ = 17,27</i>
<i>h_i</i>				
<i>Ȳ_i</i>	15	17,5	13	

$$\chi^2_{ij} = 2257$$

$$SST = 2257 - \frac{157^2}{11} = \frac{178}{11}$$

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}^2 - \frac{T^2}{N}$$

$$SSA = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^k \frac{T_i^2}{n_i} - \frac{T^2}{N}$$

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = SST - SSA$$

$$s_1^2 = \frac{SSA}{k-1} \text{ y } s^2 = \frac{SSE}{N-k}$$

$$SSA = (900 + 841 + 507) - \frac{157^2}{11} = \frac{79}{11}$$

$$SSE = \frac{178}{11} - \frac{79}{11} = 9$$

Estadístico de prueba: $F = \frac{S_1^2}{S^2}$ con $(k-1, N-k)$ g.l.

$$s_1^2 = \frac{SSA}{k-1} \text{ y } s^2 = \frac{SSE}{N-k}$$
$$\begin{array}{rcl} SSA & = & \frac{79}{11} \\ SSE & = & 9 \\ N & = & 11 \\ k & = & 3 \end{array}$$

$$S_1^2 = \frac{\frac{79}{11}}{3-1} = \frac{79}{22} \quad S^2 = \frac{9}{11-3} = \frac{9}{8}$$

$$F_{0.05} = \frac{\frac{79}{22}}{\frac{9}{8}} \approx 3.19$$

$$V = k-1, N-k$$

$$V = 3-1, 11-3$$

$$F_C = F_{0.05, 2, 8} = 4.95897 \quad V = 2, 8$$

R/ Como $F_{0.05} = 3.19 < F_C = 4.95897$
NO se rechaza H_0 .

Con valor P

$$P(F > 3.19) = 0.09579 \quad \alpha = 0.05$$

R/ Como $P = 3.19 < \alpha = 0.05$
NO se rechaza H_0 .

Ejemplo

(Anova) Con el anuncio del cierre parcial de la carretera de Circunvalación los usuarios de esta vía se han visto en la necesidad de tomar rutas alternas. Los conductores pueden optar por seguir utilizando la ruta Circunvalación (CI), cruzar el centro de San José (SJ) o utilizar algún camino rural (CR). Para tener un panorama sobre los nuevos tiempos (en horas) de traslado desde Alajuela hasta Cartago se realizó un muestreo y se obtuvieron los siguientes datos

CI	2.2	2.1	1.8	1.5	1.9	1.7
SJ	1.3	1.7	2.1	1.3	1.6	2.0
CR	2.6	2.4	1.9	2.5	2.1	

Con una significancia de 0.1, ¿puede asegurarse que los tiempos medios de traslado a Cartago desde Alajuela varían según la ruta que se utilice? (6 puntos)

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_1: \exists i, j: \mu_i \neq \mu_j$$

CI	SJ	CR
2.2	1.3	2.6
2.1	1.7	2.7
1.8	2.1	1.9
1.5	1.3	2.5
1.9	1.6	2.1
1.7	2.0	
	1.5	

$$k = 3$$

n_i	6	7	5	$N = 18$
T_i	11.2	11.5	11.5	$\bar{T} = 39.2$
T_{i^2}	1568	529	529	$\bar{y} = 1.9$
n_i	75	28	20	
y_i	1.87	1.68	2.3	$\chi^2_{\text{obs}} = 67.52$

$$SST = 67.52 - \frac{39.2^2}{18} = 2.54$$

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}^2 - \frac{\bar{T}^2}{N}$$

$$SSA = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^k \frac{T_i^2}{n_i} - \frac{\bar{T}^2}{N}$$

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = SST - SSA$$

$$s_1^2 = \frac{SSA}{k-1} \quad y \quad s^2 = \frac{SSE}{N-k}$$

$$SSA = \left(\frac{1568}{75} + \frac{529}{28} + \frac{529}{20} \right) - \frac{39.2^2}{18} = \frac{1333}{1050}$$

$$SSE = 2.54 - \frac{1333}{1050} = \frac{667}{525}$$

$$s_1^2 = \frac{SSA}{k-1} \text{ y } s^2 = \frac{SSE}{N-k} \quad F = \frac{s_1^2}{s^2} \text{ con } (k-1, N-k) \text{ g.l.}$$

$$SSA = \frac{1333}{1050}$$

$$SSE = \frac{667}{525}$$

$$k=3$$

$$N=18$$

$$v=2, 15$$

$$S_{12}^2 = \frac{\frac{1333}{1050}}{3-1}$$

$$S^2 = \frac{\frac{667}{525}}{18-3}$$

$$\alpha = 0.01$$

$$F_{0.05} = \frac{\frac{1333}{1050}}{\frac{667}{525}} = 7.49$$

$$F_C = f_{0.01, 2, 15} = 6.36$$

R/ Como $F_{0.05} = 7.49 > F_C = 6.36$
NO se rechaza H_0 .

Con valor P

$$P(F > 7.49) = 0.00555 \quad \alpha = 0.01$$

R/ Como $P = 0.00555 < \alpha = 0.01$
NO se rechaza H_0 .