

CAPÍTULO 9

INTRODUCCIÓN AL ANÁLISIS DE LA VARIANZA

9.1 COMPARACIÓN DE 2 POBLACIONES

9.1.1 Ejemplos

En el tema anterior se ha estudiado el problema más sencillo de Inferencia Estadística: el de sacar conclusiones sobre **una única población** a partir de la información contenida en una muestra extraída al azar de la misma.

El siguiente problema en complejidad sería el de la **comparación de dos poblaciones**.

En estos problemas existe una variable aleatoria, cuya distribución se asume Normal, que se distribuye en dos poblaciones diferentes, y se trata de estudiar, a partir de sendas muestras obtenidas de dichas poblaciones, si existen o no diferencias entre las medias y/o entre las varianzas de ambas poblaciones.

Ejemplo 1: En el conjunto de los jóvenes universitarios valenciano se consideran las 2 poblaciones constituidas respectivamente por las chicas y por los chicos. La variable aleatoria X a estudiar es el gasto en ocio que cada estudiante realizó el pasado fin de semana. Se desea responder a las siguientes cuestiones:

- En promedio ¿gastaron más los chicos, que las chicas, las chicas que los chicos, o no hay diferencia? (Comparación de medias)
- ¿Hay mayor o menor variabilidad del gasto entre los chicos que entre las chicas? (Comparación de varianzas)

Ejemplo 2: En una factoría automovilística, las suciedades que aparecen en los capós a la salida de los hornos de secado tras la imprimación, son un problema importante. Alguien propone que, para reducir su número, se levanten los capós a la entrada de los hornos. Se plantea, en consecuencia, realizar un estudio para analizar el efecto de la medida propuesta, comparando las dos poblaciones de capós resultantes, según se levanten o no los capós, a partir del análisis de dos muestras obtenidas de las mismas mediante unas pruebas. La pregunta fundamental que se pretende contestar en el estudio es la siguiente:

- ¿Puede afirmarse que el levantar los capós reduce el número medios de suciedades? Y, si es así, ¿en cuánto los reduce? (Comparación de medias)

Ejemplo 3: En la fabricación de asientos de poliuretano para automóviles, se desea obtener asientos cuya dureza sea lo más cercana posible al valor óptimo que es 200 Nwt. La parte esencial del proceso es la inyección en los moldes, mediante una máquina, de unas determinadas cantidades de polioliol y de isocianato. Las máquinas de inyección pueden regularse para que den, en promedio, la dureza deseada (200 Nwt), pero generan una determinada variabilidad en torno a ese promedio. La empresa debe decidir entre comprar una determinada Máquina A o comprar una Máquina B más cara, pero cuyo fabricante pretende que genera menos variabilidad. Se plantea, en consecuencia, realizar un estudio comparando las dos poblaciones de asientos resultantes, según se utilice una u otra máquina, a partir del análisis de dos muestras

obtenidas de las mismas mediante unas pruebas. La pregunta fundamental que se pretende contestar en el estudio es la siguiente:

- ¿Puede afirmarse que en los asientos fabricados con la Máquina B la dureza fluctúa alrededor de 200 Nwt menos que en los fabricados con la Máquina A? Y, si es así, ¿cuánto menos? (Comparación de varianzas)

Ejemplo 4: En la fabricación de caucho SBR pueden utilizarse dos catalizadores diferentes A y B. Se plantea un estudio realizando produciendo varios “batches” (lotes) con cada catalizador, para ver si hay diferencias entre ellos en cuanto al rendimiento del proceso, que se mide por el porcentaje de conversión de butadieno en caucho. Los objetivos del estudio son contestar a las siguientes preguntas:

- ¿Hay diferencia entre los rendimientos medios obtenidos con uno u otro catalizador? (Comparación de medias)
- ¿Hay mayor o menor variabilidad del rendimiento de un batch a otro según se utilice uno u otro catalizador? (Comparación de varianzas)

9.1.2 Terminología

Vamos a introducir una terminología que utilizaremos en todo el resto del capítulo:

Variable respuesta

Es la variable aleatoria definida en las dos poblaciones estudiadas y que constituye el objeto de interés en el estudio. Así, en el primer ejemplo la variable respuesta sería “gasto en ocio de cada estudiante el pasado fin de semana”, mientras que en segundo ejemplo sería: “número de suciedades en cada capó”

Factor

Es aquella característica de la población, o del proceso que la genera, que en función de cómo se presente define las dos poblaciones en estudio. Así, en el primer ejemplo el factor cuyo efecto se pretende estudiar sería el “Sexo”, mientras que en el segundo sería la “Posición del capó a la entrada del horno”

Variantes del Factor

Son las dos alternativas consideradas para el factor y que definen las dos poblaciones. Así en el primer ejemplo las alternativas serían “Chica o Chico”, mientras que en el segundo serían “Capó Bajado o Capó Levantado”

Efectos del Factor

Se dice que el factor tiene “**efecto sobre la media**” si las medias de las dos poblaciones son distintas. Así, en el primer ejemplo diríamos que el Sexo tiene efecto sobre la media del gasto en ocio si, por ejemplo, los Chicos gastan en promedio más que las Chicas. En el segundo ejemplo, diríamos que la Posición del Capó tiene efecto sobre el número medio de suciedades si, por ejemplo, se producen menos suciedades en promedio con el Capó Levantado que con el Capó Bajado.

Si las medias de las dos poblaciones son iguales, diremos que el factor no tiene efecto sobre la media

Se dice que el factor tiene “**efecto sobre la varianza**” si las varianzas de las dos poblaciones son diferentes. Así, en el primer ejemplo diríamos que el Sexo tiene efecto

sobre la varianza del gasto en ocio si, por ejemplo, hubiera más dispersión en los gastos de los Chicos que en los de las Chicas. A los efectos sobre la varianza se les llama también “**efectos de dispersión**”

Autoevaluación: precisar lo que serían la variable respuesta, el factor, las variantes y los efectos en los ejemplos 3 y 4 planteados en el Apartado 9.1.1

9.2 COMPARACIÓN DE K POBLACIONES

(Nota: Para aligerar la exposición, en todo lo que sigue nos referiremos sólo a efectos sobre las medias, aunque podrían introducirse de una forma completamente análoga los conceptos relativos a efectos sobre las varianzas).

Los problemas planteados en el Apartado 1 se generalizan de forma inmediata a los casos de comparación de K poblaciones, sin más que considerar situaciones en los que el Factor estudiado tiene un número K, mayor que 2, de variantes.

Autoevaluación: generalizar el Ejemplo 3 a un caso de comparación de 3 poblaciones y el Ejemplo 4 a un caso de comparación de 5 poblaciones.

Nota: obsérvese que aunque el número de poblaciones K pueda ser elevado sigue existiendo **un solo Factor**.

En los estudios con K poblaciones se dice que el Factor estudiado tiene efectos sobre las medias, si la media de al menos una de las K poblaciones es diferente de las de las restantes, o sea, si no es verdad que la K medias sean iguales.

Obsérvese que, a diferencia de lo que sucedía en el caso de 2 poblaciones, cuando se comparan K poblaciones no basta con saber que el factor tiene efecto, sino que hay que concretar en qué consiste dicho efecto.

Así, al comparar dos poblaciones A y B, si hay efecto del factor quiere decir que $m_A \neq m_B$.

En cambio, en la comparación de 3 poblaciones: A, B y C, si hay efecto del factor lo que se puede afirmar es que no es cierto que $m_A = m_B = m_C$. Esto puede producirse de dos formas diferentes: porque una media es diferente a las otras dos, que son iguales entre sí, o porque las 3 medias son diferentes unas de otras.

(Como veremos más adelante, al comparar 3 poblaciones el efecto del Factor tiene **2 grados de libertad**, y, en general, al comparar K poblaciones el efecto del factor tiene K-1 grados de libertad).

9.3 ESTUDIOS MULTIFACTORIALES

9.3.1 Estudio simultáneo de varios factores

En la mayor parte de los problemas reales suele haber más de un factor que puede afectar al resultado o resultados de interés.

Autoevaluación: Enunciar varios factores que pueden influir sobre el rendimiento de un proceso químico. Enunciar varios factores que pueden influir sobre la calidad organoléptica de una paella.

Tradicionalmente, el enfoque que suele adoptarse al abordar estos problemas consiste en estudiar cada factor por separado, o, en contextos de experimentación industrial, hacer pruebas modificando cada vez un sólo factor. Ello refleja la idea, que se tiene habitualmente, de que si se modifican muchos factores a la vez, no va a ser posible precisar cuáles de estos factores han sido los responsables de los cambios observados en los resultados.

Sin embargo, este enfoque, pese a parecer lógico a primera vista, es completamente desaconsejable, por resultar ineficaz e ineficiente, al exigir una gran cantidad de pruebas y no permitir estudiar las posibles interacciones (ver más adelante) entre los efectos de los factores.

La alternativa consiste en **estudiar simultáneamente** los efectos de los diferentes factores y sus posibles interacciones mediante un **estudio multifactorial**.

9.3.2 Poblaciones estudiadas

Para ilustrar las ideas que se van a exponer, plantearemos como ejemplo un posible estudio para investigar los efectos sobre el rendimiento de un proceso de obtención de cierto producto químico (variable respuesta) de dos factores:

Factor 1: Tipo de catalizador utilizado, con 2 variantes A y B

Factor 2: Temperatura en el reactor, con 3 variantes 200°C, 250°C y 300°C.

Hay que notar, en primer lugar, que en este caso se están estudiando **6 poblaciones** distintas, resultantes de combinar las dos variantes del primer factor con las 3 variantes del segundo:

- | | |
|--------------------|--|
| Primera población: | Batches producidos usando Catalizador A a 200°C
Esta población tendrá una media $m_{A,200}$ |
| Segunda población: | Batches producidos usando Catalizador A a 250°C
Esta población tendrá una media $m_{A,250}$ |
| Tercera población: | Batches producidos usando Catalizador A a 300°C
Esta población tendrá una media $m_{A,300}$ |
| Cuarta población: | Batches producidos usando Catalizador B a 200°C
Esta población tendrá una media $m_{B,200}$ |
| Quinta población: | Batches producidos usando Catalizador B a 250°C
Esta población tendrá una media $m_{B,250}$ |
| Sexta población: | Batches producidos usando Catalizador B a 300°C
Esta población tendrá una media $m_{B,300}$ |

	Temp. 200°C	Temp. 250°C	Temp. 300°C
Catalizador A	$m_{A,200}$	$m_{A,250}$	$m_{A,300}$
Catalizador B	$m_{B,200}$	$m_{B,250}$	$m_{B,300}$

Tabla 9.1 Seis poblaciones consideradas en el estudio multifactorial 2 x 3

Se supondrá que en las 6 poblaciones la variable estudiada se distribuye normalmente con unas determinadas medias.

Hipótesis Nula Global: la Hipótesis Nula Global inicial H_0 es ninguno de los dos factores tiene efecto sobre los valores medios y que, por tanto las 6 medias son iguales: $m_{A,200} = m_{A,250} = \dots = m_{B,300}$

9.3.3 Concepto de efecto simple de un factor

El **efecto simple** de cada factor se define sobre el promedio de las condiciones estudiadas del otro factor.

Así, en nuestro ejemplo, el efecto simple del factor "Catalizador" se medirá por la diferencia entre los rendimientos medios obtenidos con uno u otro catalizador, para el promedio de las tres temperaturas estudiadas

	Temp. 200°C	Temp. 250°C	Temp. 300°C	
Catalizador A	$m_{A,200}$	$m_{A,250}$	$m_{A,300}$	m_A
Catalizador B	$m_{B,200}$	$m_{B,250}$	$m_{B,300}$	m_B
	$m_{.200}$	$m_{.250}$	$m_{.300}$	

$$\text{Así, denominado } m_A = \frac{m_{A,200} + m_{A,250} + m_{A,300}}{3} \quad \text{y} \quad m_B = \frac{m_{B,200} + m_{B,250} + m_{B,300}}{3}$$

se dirá que el factor Catalizador tiene un efecto simple sobre la media del rendimiento si $m_A \neq m_B$. La Hipótesis Nula relativa al efecto simple de este factor es que dicho efecto es nulo, o sea $m_A = m_B$

De forma análoga, el efecto simple del factor " Temperatura " se medirá por las diferencias entre los rendimientos medios obtenidos con las tres temperaturas, para el promedio de los dos catalizadores

Así, denominado:

$$m_{.200} = \frac{m_{A,200} + m_{B,200}}{2} \quad m_{.250} = \frac{m_{A,250} + m_{B,250}}{2} \quad m_{.300} = \frac{m_{A,300} + m_{B,300}}{2}$$

se dirá que el factor Temperatura tiene un efecto simple sobre la media del rendimiento si no es cierto que $m_{.200} = m_{.250} = m_{.300}$. La Hipótesis Nula relativa al efecto simple de este factor es que dicho efecto es nulo, o sea $m_{.200} = m_{.250} = m_{.300}$

9.3.4 Concepto de interacción entre dos factores

Existirá una interacción doble entre dos factores, si el efecto de uno de ellos es diferente según la variante considerada del otro factor.

Así, en nuestro estudio, existiría interacción entre los dos factores si, por ejemplo, la diferencia de los rendimientos medios entre los dos catalizadores fuera muy marcada si se operase a una Temperatura de 300°C, pero fuera pequeña o inexistente si la Temperatura fuera 200°C..

De forma simétrica, existiría interacción entre los dos factores si, por ejemplo, la diferencia de los rendimientos medios entre una temperatura de 300°C y una Temperatura de 200°C fuera mucho más marcada trabajando con el Catalizador A que con el Catalizador B.

Hay que tener en cuenta que si el efecto de un primer factor depende de la variante considerada del segundo, también necesariamente el efecto del segundo factor dependerá de la variante considerada del primero. De hecho, ambos fenómenos no son más que las dos caras de una misma moneda: la existencia de interacción entre los dos factores.

Nota: en estudios con más de dos factores, puede también plantearse la existencia de interacciones de orden superior: triples, cuádruples, etcétera... Así, existiría una interacción triple entre tres factores A, B y C si, por ejemplo, hubiera una interacción doble entre A y B cuando C está a nivel bajo, pero no existiera dicha interacción A*B cuando C está a nivel alto. En general las interacciones de orden superior a dos se presentan poco en la práctica, siendo además difíciles de interpretar, por lo que no serán consideradas en el resto de este tema.

9.4 INTRODUCCIÓN AL ANÁLISIS DE LA VARIANZA

9.4.1 Introducción

El **Análisis de la Varianza** (también conocida por sus siglas en inglés **ANOVA**) es una técnica estadística muy poderosa que permite estudiar desde los problemas más sencillos, como la comparación de dos poblaciones, hasta los más complicados, como el estudio simultáneo de los efectos que un conjunto de varios factores y sus interacciones tienen sobre una determinada variable respuesta.

Aunque en principio el ANOVA se desarrolló como un procedimiento para estudiar posibles efectos de los factores sobre las medias de las poblaciones implicadas, es sencillo, mediante un procedimiento original que se desarrollará en el curso,

generalizarlo para permitir estudiar también los posibles efectos sobre las varianzas de dichas poblaciones.

En el Apartado 9.4.2 se da una idea intuitiva del funcionamiento del ANOVA, mientras que en el Apartado 9.4.3 se exponen los pasos a seguir para obtener el Cuadro Resumen del Análisis de la Varianza así como su interpretación general.

9.4.2 Idea intuitiva del ANOVA

La idea básica del ANOVA consiste en descomponer la variabilidad total observada en unos datos en una serie de términos, asociados a los efectos de cada factor estudiado y a sus posibles interacciones, más una parte residual con la que después se compararán los primeros.

Con el fin de precisar estas ideas, vamos a utilizar el ejemplo sencillo que se expuso en el Apartado 3.

Ejemplo:

Se desea estudiar los efectos que el catalizador y la temperatura tienen sobre el rendimiento de un proceso químico. Se van a comparar dos catalizadores (A y B) y tres temperaturas (200, 250 y 300). Se han realizado 2 batches (lotes) con cada una de las 6 combinaciones posibles de catalizador y temperatura, o sea un total de 12 lotes

Veamos unos posibles resultados (rendimientos obtenidos en cada lote expresados, por ejemplo, en Tms) en algunos casos hipotéticos extremos:

	Temperatura		
	200	250	300
Cat: A	20 20	20 20	20 20
Cat: B	20 20	20 20	20 20

Caso 1: los 12 valores x_{ijk} han resultado idénticos. ¡No hay variabilidad en los datos! La variabilidad total, medida por la Suma de Cuadrados Total es nula:

$$SC_{total} = \sum_{i,j,k} (x_{ijk} - \bar{x})^2 = 0$$

Consecuencia: ni el catalizador ni la

temperatura influyen en el rendimiento

	Temperatura		
	200	250	300
Cat: A	20 20	20 20	20 20
Cat: B	30 30	30 30	30 30

Caso 2: ¡Sí que hay variabilidad en los datos!

$$SC_{total} = \sum_{i,j,k} (x_{ijk} - \bar{x})^2 = 300$$

Cuando se analiza esta variabilidad (de ahí el nombre de ANOVA) se constata que se debe exclusivamente al efecto de la catalizador

sobre la media (el B da valores sistemáticamente más altos que el A), no habiendo ningún efecto de la temperatura.

En la tabla resumen del ANOVA se obtendría, aplicando unas fórmulas que se verán más adelante: $SC_{total}=300$ $SC_{cat}= 300$ $SC_{temp}=0$

	Temperatura		
	200	250	300
Cat: A	20 20	25 25	30 30
Cat: B	30 30	35 35	40 40

Caso 3: la variabilidad total es ahora

$$SC_{total} = \sum_{i,j,k} (x_{ijk} - \bar{x})^2 = 500$$

Cuando se analiza esta variabilidad se constata que parte de ella se debe al efecto de la catalizador sobre la media (el B da valores

10 unidades más altos que el A), pero que también hay una parte debido al efecto de la temperatura (los valores aumentan 5 unidades al pasar de 200 a 250 y otras 5 unidades al pasar de 250 a 300).

Por otra parte se constata que no hay interacción entre ambos factores, porque la diferencia entre A y B es la misma sea cual sea la temperatura, y el efecto de aumentar la temperatura es el mismo en ambos catalizadores.

Calculando, mediante las fórmulas que se verán más adelante, las Sumas de Cuadrados asociadas a cada efecto se obtendría:

$$SC_{total}=500 \quad SC_{var}= 300 \quad SC_{temp}=200 \quad SC_{interaccion}=0$$

	Temperatura		
	200	250	300
Cat: A	20 20	25 25	30 30
Cat: B	30 30	40 40	50 50

Caso 4: la variabilidad total es ahora

$$SC_{total} = \sum_{i,j,k} (x_{ijk} - \bar{x})^2 = 1175$$

Cuando se analiza esta variabilidad se constata que una parte de ella se debe al efecto de la catalizador sobre la media (el B da

valores más altos que el A), y otra parte se debe al efecto de la temperatura (los valores aumentan al aumentar la temperatura).

Sin embargo, adicionalmente, se constata además que hay una interacción entre ambos factores: la diferencia entre B y A es mayor a temperatura altas que bajas, y el efecto de aumentar la temperatura es más marcado con el catalizador B que con el A.

Calculando, mediante las fórmulas que se verán más adelante, las Sumas de Cuadrados asociadas a cada efecto se obtendría:

$$SC_{total}=1175 \quad SC_{var}= 675 \quad SC_{temp}=450 \quad SC_{interaccion}=50$$

	Temperatura		
	200	250	300
Cat: A	19 21	26 24	28 32
Cat: B	30 31	39 43	49 52

Caso 5: (este caso es el único realista) los efectos de los factores vienen parcialmente enmascarados por la variabilidad residual originada por factores no controlados.

Las parejas de lotes con idéntico catalizador y temperatura no han resultado exactamente iguales debido al efecto de otros factores no controlados (pureza de las materias primas, pequeña variabilidad en presiones, ...) y, por tanto, no dan exactamente los mismos rendimientos.

La variabilidad total se debe ahora, no sólo al efecto del catalizador, la temperatura y su interacción, sino además a una variabilidad residual asociada a todos los restantes factores no controlados que influyen sobre los rendimientos.

Calculando, mediante las fórmulas que se verán más adelante, las Sumas de Cuadrados asociadas a cada efecto se obtendría:

$$SC_{total}=1261.67 \quad SC_{var}= 736.33 \quad SC_{temp}=450.17 \quad SC_{interac}=50.17 \quad SC_{resid}=25.0$$

Intuitivamente, la forma de estudiar si un determinado efecto simple o una interacción son significativos es constatando si la variabilidad explicada por dicho efecto es significativamente mayor que la variabilidad residual.

Ello se lleva a cabo construyendo el Cuadro Resumen del Análisis de la Varianza, tal como se expone a continuación.

9.4.3 Cuadro resumen del Análisis de la Varianza

Los resultados de un ANOVA se sintetizan en una tabla, el Cuadro Resumen del Análisis de la Varianza, que tiene siempre 6 columnas y una fila por cada factor y cada posible interacción doble, además de una fila para la variabilidad residual y otra para la variabilidad total.

La primera columna de la tabla, titulada “origen de la variabilidad”, recoge las diferentes fuentes de variabilidad, tal como se indica a continuación para nuestro ejemplo

Origen de la variabilidad					
CATALIZADOR					
TEMPERATURA					
CATAL*TEMPER					
RESIDUAL					
TOTAL					

Sumas de Cuadrados

La segunda columna de la tabla recoge la variabilidad asociada a cada término, cuantificada mediante sus respectivas Sumas de Cuadrados. El cálculo de estas Sumas de Cuadrados, que se lleva a cabo mediante unas fórmulas que se explicarán más adelante, es la única parte laboriosa del ANOVA

En la siguiente tabla se ha rellenado ya esa segunda columna

Origen de la variabilidad	Suma de cuadrados				
CATALIZADOR	736.333				
TEMPERATURA	450.167				
CATAL*TEMPER	50.167				
RESIDUAL	25.000				
TOTAL	1261.667				

Grados de libertad

Paralelamente a esta descomposición de la SC_{total} en sus componentes, se realiza una descomposición de los "grados de libertad" totales, que son siempre el número de datos menos 1 ($12-1=11$ en el ejemplo), en los grados de libertad asociados a cada término.

Los grados de libertad asociados al efecto simple de un factor son siempre el número de variantes del factor menos 1 ($gl_{catal}=2-1=1$ y $gl_{temper}=3-1=2$), mientras que los de una interacción son el producto de los grados de libertad de los factores correspondientes ($gl_{cat*temp}=1 \times 2=2$).

Los grados de libertad residuales se obtienen por diferencia entre los totales y la suma de los asociados a los diferentes efectos ($gl_{resid}=11-1-2-2=6$)

Origen de la variabilidad	Suma de cuadrados	grados de libertad			
CATALIZADOR	736.333	$2-1=1$			
TEMPERATURA	450.167	$3-1=2$			
CATAL*TEMPER	50.167	$1 \times 2=2$			
RESIDUAL	25.000	$11-1-2-2=6$			
TOTAL	1261.667	$12-1=11$			

Cuadrados Medios

La comparación de la "varianza" asociada a cada efecto con la varianza residual permite estudiar si dicho efecto es o no significativo. Dichas "varianzas" se estiman dividiendo cada Suma de Cuadrados por sus correspondientes grados de libertad, obteniéndose unos estadísticos a los que se denomina Cuadrados Medios (El CM_{total} , que no es más que la varianza de los datos, no acostumbra a calcularse).

Origen de la variabilidad	Suma de cuadrados	grados de libertad	Cuadrado Medio		
CATALIZADOR	736.333	1	736.333		
TEMPERATURA	450.167	2	225.083		
CATAL*TEMPER	50.167	2	25.0833		
RESIDUAL	25.000	6	4.1667		
TOTAL	1261.667	11			

F_{ratios}

El $CM_{residual}$ es una estimación de la σ^2 existente en las poblaciones muestreadas, asumiendo que dichas poblaciones tienen todas la misma σ^2 (o del promedio de dichas varianzas en el caso de que difieran de unas poblaciones a otras).

Por otra parte, El CM asociado a cada efecto es también una estimación (independiente de la anterior) de dicha σ^2 si dicho efecto no existe en la población, pero tiende a ser mayor que σ^2 en el caso de que exista un efecto real poblacional.

Por lo tanto, para ver si un efecto es significativo, hay que analizar si su CM es significativamente mayor que el $CM_{residual}$ o, lo que es lo mismo, si el cociente $CM_{efecto}/CM_{residual}$ (al que se denomina **F-ratio**) es significativamente mayor que 1.

Origen de la variabilidad	Suma de cuadrados	grados de libertad	Cuadrado Medio	F _{ratio}	
CATALIZADOR	736.333	1	736.333	176.72	
TEMPERATURA	450.167	2	225.083	54.02	
CATAL*TEMPER	50.167	2	25.0833	6.02	
RESIDUAL	25.000	6	4.1667		
TOTAL	1261.667	11			

P-values

¿Cómo se comprueba si una F-ratio es, o no, significativamente mayor que 1? Para responder a esta pregunta hay que introducir una nueva distribución: la **F de Fisher**.

Se demuestra que el cociente s_1^2/s_2^2 de dos estimaciones independientes de una misma varianza (respectivamente con n_1 y n_2 grados de libertad) sigue una determinada distribución estadística denominada **distribución F de Fisher**. Las

probabilidades de esta distribución vienen tabuladas en función de sus grados de libertad n_1 y n_2 , tal como se verá en el tema siguiente.

Si un efecto es inexistente en las poblaciones estudiadas, es decir, si la H_0 asociada a este efecto es cierta, el cociente $F_{ratio} = CM_{efecto}/CM_{residual}$ se distribuye como una F de Fisher, con los correspondientes grados de libertad. Por el contrario, si existe realmente un efecto, la F_{ratio} tenderá a tomar valores más elevados que los que cabría esperar para una F de Fisher.

En consecuencia, para ver si el CM de un efecto es significativamente mayor que el $CM_{residual}$, lo que implicaría la existencia de un efecto real a nivel poblacional, se comprueba si la correspondiente F_{ratio} es demasiado elevada para ser una F de Fisher con los grados de libertad correspondientes.

Para ello se calcula (en la última columna) la probabilidad de que una F de Fisher, con los grados de libertad correspondientes, tome valores tan elevados o más como el valor obtenido para la F_{ratio} . Cuanto menor sea esa probabilidad, a la que se denomina **P-value**, más fuerte será la evidencia respecto a la existencia poblacional del efecto correspondiente.

En general se considera que un efecto es significativo (y a veces se indica con un *) si su P-value es menor que 0.05, y que es muy significativo (y a veces se indica con dos *) si resulta menor que 0.01

Origen de la variabilidad	Suma de cuadrados	grados de libertad	Cuadrado Medio	F_{ratio}	P-value
CATALIZADOR	736.333	1	736.333	176.72	0.0000**
TEMPERATURA	450.167	2	225.083	54.02	0.0001**
CATAL*TEMPER	50.167	2	25.0833	6.02	0.0368*
RESIDUAL	25.000	6	4.1667		
TOTAL	1261.667	11			

Se concluye que los efectos simples de CATALIZADOR y de la TEMPERATURA son muy significativos ($p\text{-value} < 0.01$) y que la Interacción CATAL*TEMP es significativa ($0.01 < p\text{-value} < 0.05$)

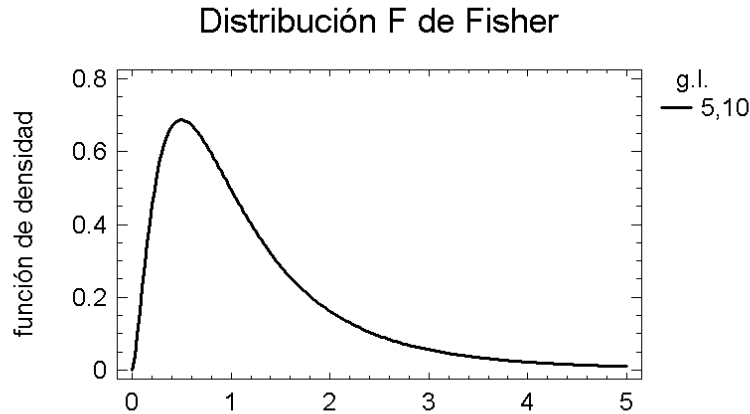
9.5 LA DISTRIBUCIÓN F

Como acabamos de indicar, el estudio de la significación de los diferentes efectos en un Análisis de la Varianza exige el manejo de una nueva distribución: la distribución F a la que su creador, Snedecor, bautizó con este nombre en honor del estadístico inglés R.A.Fisher.

Siendo X_1 y X_2 dos variables Gi-dos independientes con v_1 y v_2 grados de libertad, se dice que una variable **F** sigue una distribución **F** con v_1 y v_2 grados de libertad si:

$$F = \frac{X_1/v_1}{X_2/v_2}$$

La siguiente figura refleja la forma de la función de densidad de una F con 5 grados de libertad en el numerador y 10 grados de libertad en el denominador.



Como puede apreciarse la distribución es marcadamente asimétrica, si bien el coeficiente de asimetría decrece ligeramente a medida que aumentan los grados de libertad del numerador y del denominador.

La media de una distribución es cercana a 1. (Más concretamente su valor es $v_2/(v_2-2)$ para $v_2 > 2$)

Autoevaluación: Justificar intuitivamente el que la media de una distribución F sea cercana a 1.

En general las tablas de la distribución F son muy prolijas, por exigir tres entradas: una para v_1 , otra para v_2 y otra para los niveles α de probabilidad. En este texto nos hemos limitado a la sucinta tabla de la hoja siguiente, que refleja para un conjunto de valores seleccionados de v_1 y v_2 , los valores críticos correspondientes a unas probabilidades de ser superados del 5% y del 1%.

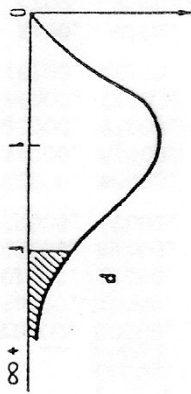
Aunque la tabla sólo da probabilidades correspondientes a las colas de la derecha de la distribución, que son las que interesan en general en la práctica, también es posible calcular a partir de ellas probabilidades en la cola de la izquierda, sin más que aplicar la definición de la variable F.

La distribución F se utiliza en la práctica con el objetivo de comparar la variabilidad debida a diferentes fuentes. En concreto si s_1^2 es la varianza en una muestra de tamaño N_1 extraída de una población normal de varianza σ_1^2 y s_2^2 es la varianza de una muestra de tamaño N_2 extraída de una población normal de varianza σ_2^2 , y ambas muestras son independientes,

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

se distribuye como una variable F de Snedecor con (N_1-1) y (N_2-1) grados de libertad.

TABLA
DE LA DISTRIBUCION DE F
(LEY DE SNEDECOR)



Valor de F con la
probabilidad P de ser superado

v ₂	v ₁ = 1		v ₁ = 2		v ₁ = 3		v ₁ = 4		v ₁ = 5		v ₂	v ₁ = 6		v ₁ = 8		v ₁ = 12		v ₁ = 24		v ₁ = ∞	
	P = 0.05	P = 0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01		P = 0.05	P = 0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
1	161.4	4052	109.5	4999	215.7	5403	224.6	5625	230.2	5764	1	236.0	5859	238.9	5981	243.9	6106	249.0	6236	251.3	6366
2	18.51	384.9	19.00	89.00	19.16	39.17	19.25	39.25	19.30	39.30	2	19.33	39.33	19.37	39.36	19.41	39.42	19.45	39.46	19.50	39.50
3	10.13	21.12	9.55	20.81	9.28	20.46	9.12	20.71	9.01	20.24	3	8.94	20.23	8.84	20.19	8.74	20.05	8.64	20.00	8.53	20.00
4	7.71	15.20	6.94	18.00	6.59	16.69	6.39	15.98	6.26	15.52	4	6.16	15.21	6.04	14.80	5.91	14.37	5.77	14.33	5.63	14.46
5	6.61	10.26	5.79	13.27	5.41	12.06	5.19	11.39	5.05	10.97	5	4.95	10.67	4.82	10.27	4.68	9.89	4.53	9.47	4.36	9.02
6	5.99	8.74	5.14	10.91	4.76	9.78	4.53	9.15	4.39	8.75	6	4.28	8.47	4.15	8.10	4.00	7.72	3.84	7.31	3.67	6.88
7	5.59	8.25	4.76	9.55	4.38	8.45	4.12	7.85	3.97	7.45	7	3.87	7.19	3.73	6.84	3.57	7.17	3.44	3.57	3.23	6.53
8	5.32	8.02	4.46	8.65	4.07	7.59	3.84	7.01	3.69	6.63	8	3.58	6.37	3.44	6.03	3.28	6.47	3.12	3.28	2.93	6.35
9	5.12	7.85	4.26	8.02	3.86	7.39	3.63	6.42	3.48	6.06	9	3.37	5.80	3.23	5.47	3.07	5.11	2.90	2.71	2.71	4.81
10	4.96	7.71	4.10	7.56	3.71	7.16	3.48	6.55	3.33	5.64	10	3.22	5.59	3.07	5.06	2.91	4.71	2.74	2.54	2.54	4.91
11	4.84	7.65	3.98	7.20	3.59	6.82	3.36	6.22	3.20	5.32	11	3.09	5.07	2.95	4.74	2.79	4.40	2.61	2.40	2.40	3.60
12	4.75	7.61	3.88	6.93	3.49	6.53	3.26	5.93	3.11	5.06	12	3.00	4.82	2.85	4.50	2.69	4.16	2.50	2.30	2.30	3.86
13	4.67	7.57	3.80	6.70	3.41	6.24	3.18	5.74	3.02	4.86	13	2.92	4.62	2.77	4.30	2.60	3.96	2.42	2.18	2.18	3.86
14	4.60	7.54	3.74	6.51	3.34	5.95	3.11	5.56	2.96	4.69	14	2.85	4.46	2.70	4.14	2.53	3.80	2.35	2.13	2.13	3.80
15	4.54	7.51	3.68	6.30	3.29	5.42	3.06	5.29	2.85	4.56	15	2.79	4.32	2.64	4.00	2.48	3.67	2.29	2.07	2.07	3.87
16	4.49	7.48	3.63	6.23	3.24	5.29	3.01	5.18	2.81	4.44	16	2.74	4.20	2.59	3.89	2.42	3.55	2.24	2.01	2.01	3.87
17	4.43	7.45	3.59	6.11	3.20	5.18	2.96	5.09	2.77	4.34	17	2.70	4.10	2.51	3.79	2.38	3.45	2.19	1.96	1.96	3.65
18	4.41	7.43	3.55	6.01	3.16	5.09	2.93	4.58	2.72	4.25	18	2.66	4.01	2.51	3.71	2.34	3.37	2.15	1.92	1.92	3.57
19	4.38	7.41	3.52	5.93	3.13	5.01	2.90	4.50	2.74	4.17	19	2.63	3.94	2.48	3.63	2.31	3.30	2.11	1.88	1.88	3.49
20	4.35	7.39	3.49	5.85	3.10	4.94	2.87	4.43	2.71	4.10	20	2.60	3.87	2.45	3.56	2.28	3.23	2.08	1.84	1.84	3.43
21	4.32	7.37	3.47	5.78	3.07	4.87	2.84	4.37	2.68	4.04	21	2.57	3.81	2.42	3.51	2.25	3.17	2.05	1.81	1.81	3.36
22	4.30	7.35	3.44	5.72	3.05	4.82	2.82	4.31	2.66	3.99	22	2.55	3.76	2.40	3.45	2.23	3.12	2.03	1.78	1.78	3.31
23	4.28	7.33	3.42	5.66	3.03	4.76	2.80	4.26	2.64	3.94	23	2.53	3.71	2.38	3.41	2.20	3.07	2.00	1.76	1.76	3.28
24	4.26	7.31	3.40	5.61	3.01	4.72	2.78	4.22	2.62	3.90	24	2.51	3.66	2.36	3.36	2.18	3.03	1.98	1.73	1.73	3.24
25	4.24	7.29	3.38	5.57	2.99	4.68	2.76	4.18	2.60	3.86	25	2.49	3.63	2.34	3.32	2.16	2.99	1.96	1.71	1.71	3.18
26	4.22	7.27	3.37	5.53	2.98	4.64	2.74	4.14	2.59	3.82	26	2.47	3.59	2.32	3.29	2.15	2.96	1.95	1.69	1.69	3.13
27	4.21	7.25	3.35	5.49	2.96	4.60	2.72	4.11	2.57	3.78	27	2.46	3.56	2.30	3.26	2.13	2.93	1.93	1.67	1.67	3.10
28	4.20	7.23	3.34	5.45	2.95	4.57	2.71	4.07	2.56	3.75	28	2.44	3.53	2.28	3.23	2.12	2.90	1.91	1.65	1.65	3.06
29	4.18	7.21	3.33	5.42	2.93	4.54	2.70	4.04	2.54	3.73	29	2.43	3.50	2.27	3.20	2.10	2.87	1.90	1.64	1.64	3.03
30	4.17	7.19	3.32	5.39	2.92	4.51	2.69	4.02	2.53	3.70	30	2.42	3.47	2.27	3.17	2.09	2.84	1.89	1.62	1.62	3.01
40	4.08	7.08	3.15	5.18	2.76	4.13	2.61	3.83	2.45	3.51	40	2.25	3.29	2.18	2.91	2.00	2.66	1.79	1.51	1.51	2.80
60	4.00	7.02	3.15	4.98	2.76	4.13	2.52	3.65	2.37	3.34	60	2.25	3.19	2.10	2.82	1.92	2.50	1.70	1.42	1.42	2.60
120	3.92	6.98	3.07	4.79	2.68	4.05	2.45	3.48	2.29	3.17	120	2.17	3.06	2.01	2.66	1.83	2.34	1.61	1.39	1.39	2.40
∞	3.94	6.84	2.99	4.60	2.60	3.98	2.37	3.33	2.21	3.02	∞	2.09	2.80	1.94	2.51	1.75	2.18	1.52	1.30	1.30	2.00

En particular si las dos varianzas poblacionales son iguales, el cociente entre las dos varianzas muestrales seguirá una distribución F.

Autoevaluación: ¿Calcular aproximadamente la probabilidad de que al extraer dos muestras de tamaño 25 de una misma población normal, la segunda varianza muestral resulta más del doble que la primera?

Siguiente Capítulo

En el siguiente Capítulo se profundizará en el tema del Análisis de la Varianza desarrollando las siguientes cuestiones:

- forma de realizar el cálculo de las Sumas de Cuadrados.
- importancia del análisis de los residuos
- análisis de las diferencias entre las diferentes variantes de un factor con más de 2 variantes, cuando su efecto ha resultado significativo
- estudio de efectos sobre las varianzas
- Planes Factoriales Equilibrados
- interpretación de las interacciones

Al completar este capítulo, el lector estará capacitado para llevar a cabo complicados estudios estadísticos, en los que se aborde la investigación de los posibles efectos que un conjunto de factores y sus interacciones tienen sobre el valor medio y sobre la varianza de las variables respuesta de interés.