

PRUEBA DE LA VARIANZA CON UNA POBLACIÓN

A veces, los analistas investigan la variabilidad de una población, en lugar de su media o proporción.

Esto es debido a que la uniformidad de la producción muchas veces es crítica en la práctica industrial.

La variabilidad excesiva es el peor enemigo de la alta calidad y la prueba de hipótesis está diseñada para determinar si la varianza de una población es igual a algún valor predeterminado.

La desviación estándar de una colección de datos se usa para describir la variabilidad en esa colección y se puede definir como la diferencia estándar entre los elementos de una colección de datos y su media.

La varianza de un conjunto de datos se define como el cuadrado de su desviación estándar; y la varianza muestral se utiliza para probar la hipótesis nula que se refiere a la variabilidad y es útil para entender el procedimiento de análisis de la varianza.

La hipótesis nula; para la prueba de la varianza, es que la varianza poblacional es igual a algún valor previamente especificado. Como el aspecto de interés, por lo general es si la varianza de la población es mayor que este valor, siempre se aplica una de una cola.

Para probar la hipótesis nula, se toma una muestra aleatoria de elementos de una población que se investiga; y a partir de esos datos, se calcula el estadístico de prueba.

Para este cálculo se utiliza la siguiente ecuación:

$$(n - 1) s^2$$
$$, = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$
$$,$$

Donde:

* $n-1$ = Grados de libertad para la prueba de tamaño n .

* s^2 = Varianza muestral.

* $,$ = Varianza poblacional si y solo si suponemos que la hipótesis nula es cierta.

EJEMPLO

1.- Averiguar si la variabilidad de edades en una comunidad local es la misma o mayor que la de todo el Estado. La desviación estándar de las edades del Estado, conocida por un estudio reciente es de 12 años. Tomamos una muestra aleatoria de 25 personas de la comunidad y determinamos sus edades. Calcular la varianza de la muestra y usar la ecuación anteriormente explicada para obtener el estadístico muestral.

Las hipótesis nula y alternativas son:

- $H_0 : \sigma^2 = 144$
- $H_1 : \sigma^2 \neq 144$

Se toma la muestra y resulta una desviación estándar muestral de 15

Años. La varianza de la muestra es entonces 225, y el estadístico ji cuadrada de la muestra es:

$$(n - 1) s^2 / \sigma_0^2 = (25 - 1)(15)^2$$

$$= \frac{225 \cdot 24}{144} = 37,5$$

$$= 37,5$$

Si la hipótesis nula es cierta, el estadístico muestral de 37,5 se obtiene de la distribución ji cuadrada teórica, en particular, la distribución con 24 grados de libertad ($25 - 1 = 24$).

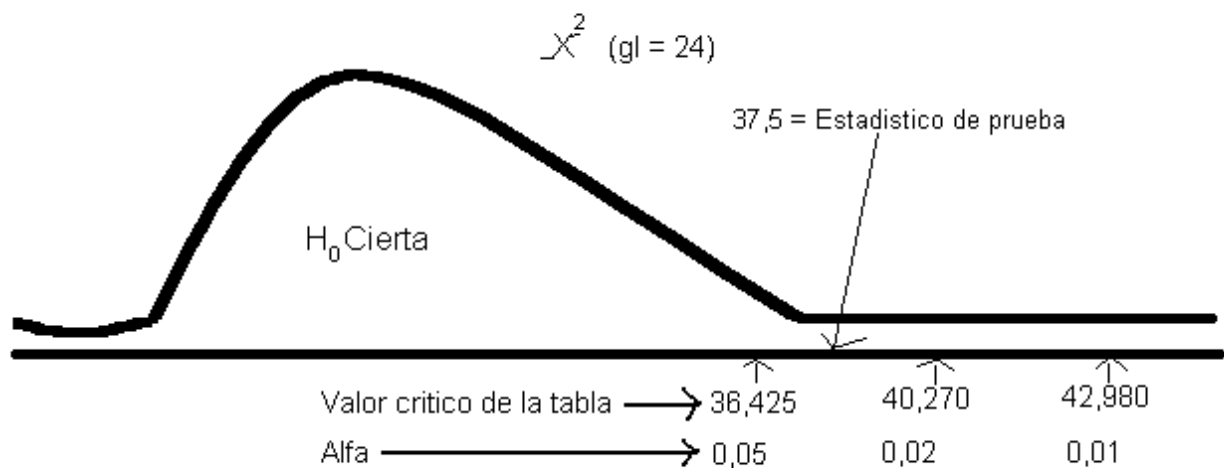
Como se puede observar en la ecuación anterior, cuanto mas grande es la varianza muestral respecto a la varianza poblacional hipotética, mas grande es el estadístico que se obtiene. Luego deducimos que de un estadístico muestral grande llevamos al rechazo de la hipótesis nula, y un estadístico muestral pequeño implicará que no se rechaze. La tabla ji cuadrada se usa para determinar si es probable o no que el valor 37,5 haya sido obtenido de la distribución muestral ji cuadrada hipotética.

Supongamos que esta prueba debe llevarse a un nivel de significancia de 0,02. En la columna 0,02 de la tabla de ji cuadrada y la fila 24, se encuentra el valor crítico de 40, 27. La regla de decisión es:

Si $\chi^2 \geq 40,27$, se rechaza la hipótesis nula de que la varianza de la población es 144 (Se rechaza H_0 si $\chi^2 \geq 40,27$).

Como estadístico de prueba calculado es 37,5, la hipótesis nula no se rechaza (con riesgo de un error de tipo II). Si en la tabla de ji cuadrada se hubiese elegido un alfa de 0,05, el valor crítico de la tabla sería 36,415, y la hipótesis nula se hubiera rechazado ($37,5 > 36,415$). En este ejemplo se ilustra la importancia de pensar con cuidado en el riesgo apropiado de un error de tipo I en una prueba de hipótesis.

Se supone que la hipótesis nula es cierta, lo que conduce a la obtención de un estadístico muestral de una distribución ji cuadrada con 24 grados de libertad.



PRUEBA DE LA VARIANZA CON DOS POBLACIONES

En ocasiones es importante comparar dos poblaciones para ver si una es mas variable que la otra en alguna medida específica. La hipótesis nula es que las dos poblaciones tienen la misma varianza, y la hipótesis alternativa es que una tiene mayor varianza que la otra. Se obtienen muestras aleatorias de cada población y se calculan las varianzas muestrales. Estos valores se usan entonces en la ecuación siguiente para calcular el estadístico de la muestra:

Cociente F

S_1^2

$F = \frac{S_1^2}{S_2^2}$

S_2^2

Donde:

- S_1^2 = Varianza de la muestra 1
- S_2^2 = Varianza de la muestra 2

Nota: Por convención, para encontrar los valores de F , por lo general se pone en el numerador la varianza muestral mas grande.

El estadístico de prueba dado por la ecuación anteriormente nombrado,

es el cociente F . Si la hipótesis nula de varianzas poblacionales iguales es

cierta, la razón de las varianzas muestrales se obtiene de la distribución F

teórica. Al consultar la tabla F se puede evaluar la probabilidad de este suceso.

Si parece probable que el cociente F pueda haberse obtenido de la distribución

muestral supuesta, la hipótesis nula no se rechaza. Si es poco probable que el cociente F se haya obtenido de la distribución supuesta, la hipótesis nula se rechaza.

La distribución F especifica que se aplica a una prueba en particular queda determinada por dos parámetros: los grados de libertad para el numerador y los grados de libertad para el denominador. Cada uno de estos valores es $n-1$. Si se conocen estos valores y se elige un valor alfa, al valor crítico de F se puede encontrar en la tabla F .

EJEMPLO

1.- Averiguar si la variabilidad del salario por hora es la misma en dos sucursales, o si la variabilidad de la sucursal 1 es mayor que la de la sucursal 2. La comparación de la variabilidad de las dos sucursales constituye el primer paso en un estudio detallado sobre ingresos.

Se toman muestras aleatorias de los salarios por hora en cada sucursal para determinar las varianzas muestrales y elegimos un nivel de significancia de 0,05. La hipótesis nula y alternativa son:

- $H_0 : \sigma_1^2 = \sigma_2^2$
- $H_1 : \sigma_1^2 > \sigma_2^2$

Los resultados de la muestra son:

$S_1 = \$3,79$ $S_2 = \$2,48$ $n_1 = 21$ (Sucursal 1)

$S_2 = \$2,48$ $S_2^2 = 6,1504$ $n_2 = 25$ (Sucursal 2)

El estadístico F se calcula mediante la ecuación anteriormente explicada:

$S_1^2 = 14,3641$

$F = \frac{S_1^2}{S_2^2} = \frac{14,3641}{6,1504} = 2,34$

$S_2^2 = 6,1504$

El cociente F indica que la varianza muestral de la población 1 es 2,34 veces la varianza muestral de la población 2. Sin embargo, dados los tamaños de las muestras ¿Es suficiente esta evidencia para rechazar la hipótesis de que las poblaciones tienen la misma varianza?. Se necesita el valor crítico de F para contestar esta pregunta. Primero, se calculan los grados de libertad para el numerador y el denominador:

$Gl(\text{numerador}) = (n_1 - 1) = (21 - 1) = 20$

$Gl(\text{denominador}) = (n_2 - 1) = (25 - 1) = 24$

Se usa la tabla F para encontrar el valor crítico. Hay dos valores de F en la tabla: uno para el nivel de significancia de 0,05 y otro para el nivel de 0,01. Al ser ésta una prueba de una cola, como sugiere la hipótesis alternativa, toda el área de 0,05 o de 0,01 estará en el extremo superior de la curva.

Las columnas de la tabla F representan los grados de libertad del numerador, por lo que se selecciona la columna 20. Las filas corresponden a los grados de libertad del denominador, así que se elige la fila 24. El

valor crítico de F a un nivel de significancia de 0,05 para 20 grados de libertad en el numerador y 24 grados de libertad en el denominador es 2,02.

El cociente F calculado a partir de los datos de la muestra es 2,34. Según este valor de prueba, la hipótesis nula se rechaza ($2,34 > 2,02$). Si acepta un riesgo del 5% de un error de tipo I, las poblaciones no tienen la misma varianza.

EJEMPLO

2.- ¿Son iguales las varianzas de dos poblaciones de edades de los artículos en inventario, o la población 2 tiene una mayor varianza? Se toman muestras aleatorias de 53 artículos de cada población de inventario y se calculan las varianzas muestrales. La prueba ha de llevarse a cabo con un nivel de significancia de 0,01. Las hipótesis nula y alternativa son:

- $H_0 : \sigma_1^2 - \sigma_2^2 = 0$
- $H_1 : \sigma_1^2 - \sigma_2^2 > 0$

Los grados de libertad del numerador y denominador son 52 ($53-1$). En

La tabla F abreviada, la fila 50 y la columna 50 se usan como aproximaciones de los grados de libertad. La regla de decisión es:

Si el cociente F calculado es mayor que 1,94, se rechaza la hipótesis nula (se rechaza H_0 si $F > 1,94$).

Los resultados de la muestra son:

$S_1^2 = 489$ $n_1 = 53$ (inventario 1)

$S_2^2 = 1,37$ $n_2 = 53$ (inventario 2)

El estadístico F se calcula mediante la ecuación anteriormente explicada:

$S_1^2 = 489$

$F = \frac{S_1^2 / n_1}{S_2^2 / n_2} = \frac{489 / 53}{1,37 / 53} = 2,8$

$S_2^2 = 1,37$

Una de las varianzas muestrales es 2,8 veces mas grande que la otra.

La hipótesis nula se rechaza ya que el estadístico de prueba (2,8) excede al valor crítico (1,94) de la tabla F . Se puede concluir que el inventario 2 tiene mas variabilidad en el tiempo que el inventario 1.

CONCEPTOS BÁSICOS PARA ANOVA

Análisis para la varianza

El procedimiento de análisis de varianza, o ANOVA, utiliza una sola

variable numérica medida en los elementos de la muestra para probar la

hipótesis nula de igualdad de medias poblacionales. Esta variable puede ser de intervalo o de escala de razón.

Esta variable algunas veces recibe el nombre de variable dependiente, en especial en programas de computadora que ejecutan ANOVA.

La hipótesis nula que se prueba en el ANOVA es que la mayoría de las poblaciones que se estudian (al menos tres) tienen el mismo valor de la media para la variable dependiente. Las hipótesis nula y alternativa en ANOVA son:

$$H_0: \mu = \mu_1 = \mu_2 = \dots = \mu_c$$

H_1 : No todas las poblaciones tienen la misma media.

En la prueba ANOVA, se reúne evidencia muestral de cada población bajo estudio y se usan estos datos para calcular un estadístico muestral. Después se consulta la distribución muestral apropiada para determinar si el estadístico muestral contradice la suposición de que la hipótesis nula es cierta. Si es así, se rechaza; de lo contrario no se rechaza.

Hemos de recordar que en la prueba de varianza con dos poblaciones se calcula el coeficiente de las varianzas muestrales y se verifica con arreglo a la distribución F . Este procedimiento también se usa en ANOVA para probar la hipótesis nula.

Se supone que todas las poblaciones bajo estudio tienen la misma varianza, sin importar si sus medias son iguales. Es decir, ya sea que las poblaciones tengan medias iguales o distintas, la variabilidad de los elementos alrededor de su respectiva media es la misma. Si esta suposición es válida, entonces se puede probar la hipótesis nula de las medias poblacionales iguales usando la distribución F .

Método dentro y método entre.

El **método dentro** para estimar la varianza de las poblaciones produce una estimación válida, sea o no cierta la hipótesis nula. El **método entre** produce una estimación válida sólo si la hipótesis nula es cierta.

El paso final en ANOVA requiere el cálculo de un cociente con la estimación del método entre en el numerador y la estimación del método dentro en el denominador.. Si la hipótesis nula de que las poblaciones tienen la misma media es cierta, esta razón consiste en dos estimaciones separadas de la misma varianza poblacional y, se puede obtener la distribución F si las medias poblacionales no son iguales. La estimación en el numerador estará inflada, y el resultado será un cociente muy grande. Al consultar la distribución F no es probable que un cociente tan grande haya sido obtenido de esta distribución, y la hipótesis nula será rechazada. La prueba de hipótesis en ANOVA es de una cola: un estadístico F grande llevará al rechazo de la hipótesis nula y un valor pequeño hará que no se rechace.

• METDO DENTRO

El método dentro de estimación de la varianza produce una estimación

válida sin importar si la hipótesis nula de las medias poblacionales iguales es cierta. Esto se debe a que la

variabilidad de los valores de la muestra se determina comparando cada elemento en los datos con la media muestral. Cada valor de la muestra obtenido de la población A se compara con la media muestral A; cada elemento obtenido de la población B se compara con la media muestral B, y así sucesivamente. La ecuación para calcular la estimación de la varianza con el método dentro es:

$$S_{w2} = \frac{\sum_{j=1}^c \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{c(n-1)}$$

Donde:

- s_{w2} = Estimación de la varianza muestral con el método entre.
- X_{ij} = i-ésimo elemento de los datos de grupo j.
- $\bar{X}_j$ = media del grupo j
- C = número de grupos
- N = número de elementos de la muestra en cada grupo.

El doble signo de suma en la ecuación, significa que primero deben

sumarse los valores indicados por el signo de la derecha, y después sumar los valores indicados por el de la izquierda. Primero, se encuentran las diferencias entre cada valor x y la media del grupo, se elevan al cuadrado y se suman. Después, se agregan estas sumas para cada grupo. El resultado es la suma del cuadrado de las desviaciones entre cada medida de la muestra y la media de su grupo. Este valor con frecuencia se llama la suma de cuadrados dentro (SCw). Esta suma se divide después entre el número adecuado de grados de libertad para poder producir una estimación de la varianza desconocida de la población.

El número adecuado de grados de libertad para el método dentro se calcula como $c(n-1)$ si el número de observaciones en cada grupo es igual. Como a cada elemento del grupo se le resta la media de ese grupo, sólo $(n-1)$ elementos de cada grupo pueden variar. Además como se tienen c grupos, c se multiplica por $(n-1)$ para obtener los grados de libertad para el método dentro.

EJEMPLO

1.- Se obtienen muestras del peso del llenado de cuatro paquetes de espinacas congeladas, a partir de tres contenedores. La pregunta es si los pesos promedio de los paquetes son iguales o diferentes entre los tres contenedores. Seguidamente se ofrecen los pesos de la muestra (en onzas), medias de grupos, media global y estimación de la varianza con el método dentro usando la ecuación correspondiente.

GRUPO 1 GRUPO 2 GRUPO 3

12,4 11,9 10,3

13,7 9,3 12,4

11,5 12,1 11,9

10,3 10,6 10,2

Media **12,00 11,00 11,2**

Media Global 11,4

$$\bullet (x_i - x_1)^2 = (12,4 - 12)^2 + (13,7 - 12)^2 + (11,5 - 12)^2 + (10,3 - 12)^2 =$$

6,19

$$\bullet (x_i - x_2)^2 = (11,9 - 11)^2 + (9,3 - 11)^2 + (12,1 - 11)^2 + (10,6 - 11)^2 =$$

5,07

$$\bullet (x_i - x_3)^2 = (10,3 - 11,2)^2 + (12,4 - 11,2)^2 + (11,9 - 11,2)^2 + (10,2 - 11,2)^2 = 3,74$$

$$(x_{IJ} - x_J)^2 \quad 6,19 + 5,07 + 3,74 \quad 15$$

$$SW^2 = \frac{\quad}{\quad} = \frac{\quad}{\quad} = \quad = 1,67$$

$$c(n-1) \quad 3(4-1) \quad 9$$

Cada valor x en la muestra se compara con la media de su propio

Grupo. Estas diferencias se elevan al cuadrado y se suman de acuerdo con la ecuación anteriormente descrita. Los valores que resultan se suman y se dividen entre los grados de libertad. El resultado, 1,67, es una estimación de la varianza común de las tres poblaciones. Con frecuencia el término SW^2 se denomina error cuadrático medio (MSE).

La razón por la que el método dentro produce una estimación válida de la varianza desconocida de la población, sin importar el estado de H_0 .

EJEMPLO

2.- Se pidió a cuatro personas que beben una marca determinada de café que registraran el número de tazas consumidas por día. Lo mismo se hizo con bebedores de otras tres marcas. Los resultados se muestran en la tabla. Estime la varianza poblacional común mediante el método dentro.

MARCA A MARCA B MARCA C MARCA D

3 5 2 3

2 1 10 6

5 4 5 4

6 6 7 5

Media **4 4 6 4,5**

Media Global 4,625

- $(x_i - x_1)^2 = (3 - 4)^2 + (2 - 4)^2 + (5 - 4)^2 + (6 - 4)^2 = 10$
- $(x_i - x_2)^2 = (5 - 4)^2 + (1 - 4)^2 + (4 - 4)^2 + (6 - 4)^2 = 14$
- $(x_i - x_3)^2 = (2 - 6)^2 + (10 - 6)^2 + (5 - 6)^2 + (7 - 6)^2 = 34$
- $(x_i - x_4)^2 = (3 - 4,5)^2 + (6 - 4,5)^2 + (4 - 4,5)^2 + (5 - 4,5)^2 = 9,25$

$$(x_{IJ} - x_J)^2 = 10 + 14 + 34 + 9,25 = 67,25$$

$$SW^2 = \frac{\sum (x_{IJ} - x_J)^2}{c(n-1)} = \frac{67,25}{12} = 5,60416$$

$$c(n-1) = 4(4-1) = 12$$

• METODO ENTRE

El segundo método para estimar la varianza común de la población

produce una estimación válida sólo si la hipótesis nula es cierta. Para entender el método entre recuerde el teorema del límite central. Este importante teorema en estadística establece que la distribución de las medias muestrales tiende a una distribución normal conforme crece el tamaño de la muestra, con una media μ y una desviación estándar $\sigma/\sqrt{n}$. Si el error estándar de la media es $\sigma/\sqrt{n}$, entonces la varianza de la distribución es igual al error estándar al cuadrado, σ^2/n .

Esta varianza es una medida de las diferencias entre todas las medias muestrales que puedan obtenerse de la distribución y la media de la población. La raíz cuadrada de esta varianza es el error estándar de la media, es decir, la diferencia estándar entre una media muestral y la media poblacional.

En ANOVA, para estimar la varianza de la distribución muestral de

medias, se debe estimar primero la media poblacional. La media de todos los valores muestrales proporciona esa estimación. Después, se determina la diferencia entre la media de cada grupo y esta media poblacional estimada, y estas diferencias se elevan al cuadrado y se suman. Este valor, con frecuencia se llama la suma de cuadrados entre (SCb). Esta suma se divide entonces entre el número adecuado de grados de libertad para obtener la estimación de la varianza de la distribución muestral. La ecuación siguiente da el cálculo de la estimación de la varianza de la distribución muestral de las medias:

$$(x_j - \bar{x})^2$$

j

$$Sx^2 = \sum (x_j - \bar{x})^2$$

$$c - 1$$

Donde:

- Sx^2 = Estimación de la varianza de la distribución muestral de medias.
- X_j = Media del grupo j.
- X = Media Global (media de todos los valores) usada como estimación de μ .
- C = número de grupos.

Para la distribución muestral de las medias $s^2 = Sx^2 / n$, en donde n es el

tamaño de la muestra o el número de elementos de cada grupo. Al evaluar esta ecuación para una estimación de la varianza (s^2) se obtiene:

,

- $s^2 = \frac{\sum (x_j - \bar{x})^2}{n}$

n

- $n s^2 = \sum (x_j - \bar{x})^2$,

- $\sum (x_j - \bar{x})^2 = n s^2$

Se puede calcular una estimación de s^2 , si se multiplica n por la estimación de s^2 , es decir;

$$s^2 = n s^2$$

La estimación del método entre de la varianza se puede calcular si se sustituye el valor de la ecuación anteriormente explicada por s^2 :

$$\sum_j n (x_j - \bar{x})^2$$

j

$$s^2 = \frac{\sum_j n (x_j - \bar{x})^2}{c - 1}$$

c- 1

Donde:

- s^2 = Estimación del método entre de la varianza poblacional común.
- x_j = media del grupo j.
- $\bar{x}$ = media global (media de todos los valores), usada como estimación de μ .
- c = número de grupos
- n = número de elementos de la muestra en cada grupo si el número de observaciones en cada uno es el mismo.

El valor adecuado de los grados de libertad para el método entre es c-1.

como la media global se resta de la media de cada grupo, sólo (c-1) medias pueden variar. Observe que la ecuación anteriormente descrita, supone que el número de observaciones en cada grupo, n, es el mismo.

EJEMPLO

En el ejemplo anterior al de antes, se obtuvo una muestra de los pesos de llenado de cuatro paquetes de espinacas congeladas de tres contenedores y se calculó una estimación de la varianza poblacional desconocida con el método dentro. En este ejemplo, la varianza poblacional desconocida se estimará mediante el método entre:

$$(12,0 - 11,4)^2 + (11,0 - 11,4)^2 + (11,2 - 11,4)^2$$

$$n (x_j - \bar{x})^2$$

$$j 4 (0,56)^2 2,24$$

$$s_x^2 = \frac{\sum (x_j - \bar{x})^2}{n-1} = \frac{2,24}{3-1} = 1,12$$

$$c-1 \ 3-1 \ 2$$

La estimación de la varianza poblacional, calculada con el método entre es 1,12.

• TABLA Y PRUEBA F PARA ANOVA

Una vez que se ha usado el método dentro y entre, para estimar la varianza desconocida de las poblaciones, se forma un cociente con estas dos estimaciones:

sb2 estimación de , por el método entre

$$F = \frac{\text{sb2 estimación de , por el método entre}}{\text{sb2 estimación de , por el método dentro}}$$

sb2 estimación de , por el método dentro

Si la hipótesis nula es cierta, tanto el numerador como el denominador

de la ecuación son estimaciones válidas de la varianza común de las poblaciones que se estudian. Este cociente se ajusta a la distribución F . Si la hipótesis nula es falsa el numerador de la ecuación en realidad es una estimación inflada de ; el denominador sigue siendo una estimación válida. Bajo estas condiciones, el valor F será muy grande, y se puede concluir que la hipótesis nula es falsa. La figura que mostramos a continuación presenta la distribución muestral para la prueba ANOVA junto con las regiones de aceptación y rechazo.

La siguiente figura ilustra el paso final de la prueba de hipótesis ANOVA. Si la hipótesis nula de medias poblacionales iguales es cierta, el estadístico F calculado se obtuvo de esta distribución; esto parece razonable siempre que el valor F no sea demasiado grande. De los datos muestrales resulta un valor F muy grande, se concluye que medias poblacionales diferentes son las causas de que el numerador en el cálculo de F esté inflado, y la hipótesis nula se rechaza. En la figura siguiente se puede observar que α (), la probabilidad de un error tipo I se indica en la cola superior. Si la hipótesis nula es en realidad cierta existe alguna posibilidad de que equivocadamente se declare falsa. La probabilidad de que esto ocurra es α (), es decir, el nivel de significancia de la prueba.



• TABLA ANOVA

Los resultados del análisis de varianza se presentan en una tabla ANOVA que resume los valores importantes de la prueba. Esta

tabla tiene un formato estándar que usan los libros y los problemas de computadora que ejecutan ANOVA. La siguiente tabla muestra la forma general de la tabla ANOVA.

En dicha tabla se resumen los cálculos necesarios para la prueba de igualdad de las medias poblacionales usando análisis de

varianza. Primero se usa el método dentro para estimar σ^2 . Cada valor de los datos se compara con su propia media, y la suma de las diferencias al cuadrado se divide entre los grados de libertad $c(n-1)$.

Fuente de Variación	SC	GL	Estimación de σ^2	Coefficiente F
Grupos Entre	$n \sum (x_j - \bar{x})^2$	$c - 1$	SSb / glb	S^2_{SSb} / S^2_{Sw}
Grupos Dentro	$\sum (x_{ij} - \bar{x}_j)^2$	$c(n-1)$	SSw / glb	
Total	$\sum (x_{ij} - \bar{x})^2$	$nc - 1$		

Donde:

- j = Número de la columna
- i = Número de la fila
- c = Número de columnas (grupos)
- n = Número de elementos en cada grupo (tamaño de la muestra)

La tabla ANOVA contiene columnas con las fuentes de variación, las sumas de cuadrados, los grados de libertad, las

estimaciones de la varianza y el valor F para el procedimiento de análisis de varianza.

EJEMPLO

Una analista de una cadena de supermercados, quiere saber si las tres tiendas tienen el mismo promedio en

dólares por compra. Se elige una muestra aleatoria de seis compras en cada tienda. La tabla número 1 presenta los datos recolectados de esta muestra junto con las medias maestras para cada tienda y la media global de todos los datos. Hará una prueba con un nivel de significancia de 0,01.

La hipótesis nula que se quiere probar es que todas las poblaciones de las que se obtuvieron los datos maestras tienen la misma media. La hipótesis alternativa es que las poblaciones no tienen la misma media. Las primeras dos medias maestras en la tabla número 1 sugieren que la hipótesis nula es cierta, ya que son muy cercanas. La tercera media muestral, es considerablemente mas pequeña que las otras dos. Pero, ¿Se debe esta diferencia a la aleatoriedad del muestreo o al hecho de que las poblaciones tienen medias distintas? Esta es la pregunta que vamos a responder con el procedimiento de ANOVA.

Tabla número 1 Datos maestras para ANOVA (en dólares) para el ejemplo

Tienda 1 Tienda 2 Tienda 3

12,05 15,17 9,48

23,94 18,52 6,92

14,63 19,57 10,47

25,78 21,40 7,63

17,52 13,59 11,90

18,45 20,57 5,92

Media 18,73 18,14 8.72

Media global : $\bar{x} = 15,20$, $c=3$, $n=6$

Se usan ambos métodos, dentro y entre, para estimar la varianza de las tres poblaciones. Recuerde la suposición fundamental de ANOVA : todas las poblaciones tienen la misma varianza sin importar si tienen la misma media. La tabla número 2 contiene los cálculos para el método dentro, y la tabla número 3 da los cálculos para el método entre.

Tabla número 2 Cálculos del método dentro para el ejemplo.

Tienda 1 $(12,05 - 18,73)^2 + (23,94 - 18,73)^2 + (14,63 - 18,73)^2 + (25,78 - 18,73)^2 + (17,52 - 18,73)^2 + (18,45 - 18,73)^2 =$

139,82

Tienda 2 $(15,17 - 18,14)^2 + (18,52 - 18,14)^2 + (19,57 - 18,14)^2 + (21,40 - 18,14)^2 + (13,59 - 18,14)^2 + (20,57 - 18,14)^2 =$

48,25

Tienda 3 $(9,48 - 8,72)^2 + (6,92 - 8,72)^2 + (10,47 - 8,72)^2 + (7,63 - 8,72)^2 + (11,90 - 8,72)^2 + (5,92 - 8,72)^2 =$

26,02

Suma de cuadrados dentro (SCw) = $139,82 + 48,25 + 26,02 = 214,09$

Tabla número 3 Cálculos del método entre para el ejemplo.

$$(18,73 - 15,20)^2 + (18,14 - 15,20)^2 + (8,72 - 15,20)^2 = 63,09$$

Suma de los cuadrados entre (SCb) = $6(63,09) = 378,54$

Los valores calculados en las tablas 2 y 3 se usan para rellenar la tabla ANOVA. Como se tienen tres poblaciones en la prueba,

$c = 3$. Se obtuvo una muestra de seis valores de cada población, así que $n = 6$. La tabla número cuatro presenta la tabla ANOVA para este ejemplo.

Tabla número 4 Tabla ANOVA para el ejemplo.

Fuente de

Variación SC gl Estimación de , Coeficiente F

Grupos entre 378,54 2 189,27 13,26

Grupos dentro 214,09 15 14,27

592,63 17

Los grados de libertad se calcularon como sigue:

- $c - 1 = 3 - 1 = 2$ (Grupos entre)
- $c (n - 1) = 3 (6 - 1) = 15$ (Grupos dentro)

Como se puede ver en la tabla número 4, el método entre para estimar μ , produce un valor de 189,27, mientras que la estimación

del método dentro es de 14,27. El cociente F indica que la estimación del método entre es 13,26 veces el valor del método dentro. ¿Se debe esta diferencia al error de muestreo, o se debe a que la hipótesis nula es falsa?. Para contestar a esta pregunta se consulta la tabla F y se determina un valor crítico.

Dos grados de libertad están asociados con el numerador del cociente de F , y se asocian quince grados de libertad con el denominador. De la tabla F el valor crítico es 6,36 para estos grados de libertad a un nivel de significancia de 0,01. El valor F calculado de 13,26 es mayor que el valor crítico, lo que significa que se tiene suficiente evidencia muestral para rechazar la hipótesis nula de medias poblacionales iguales.

EJEMPLO 2

Se pide a cuatro personas que beben una marca determinada de café que registren el número de tazas que consumen durante un día. Se hace lo mismo con bebedores de otras marcas. Los resultados se muestran a

continuación. Construya la tabla ANOVA para probar si existe alguna diferencia en el número promedio de tazas consumidas, para cada marca.

Marca A Marca B Marca C Marca D

3 5 2 3

2 1 10 6

5 4 5 4

6 6 7 3

4 4 6 3

Media global 4.25

$n = 4$

$c = 4$

METODO DENTRO

- Marca A $(3 - 4)^2 + (2 - 4)^2 + (5 - 4)^2 + (6 - 4)^2 = 10$
- Marca B $(5 - 4)^2 + (1 - 4)^2 + (5 - 4)^2 + (6 - 4)^2 = 17$
- Marca C $(2 - 6)^2 + (10 - 6)^2 + (5 - 6)^2 + (7 - 6)^2 = 34$
- Marca D $(3 - 3)^2 + (6 - 3)^2 + (4 - 3)^2 + (5 - 3)^2 = 14$

METODO ENTRE

$$(4 - 4,25)^2 + (4 - 4,25)^2 + (6 - 4,25)^2 + (3 - 4,25)^2 = 6,75$$

TABLA ANOVA

Fuente de

Variación SC GL Estimación , Coeficiente F

$$\text{Metodo } 6,75 \quad 3 \quad 6,75 / 3 \quad 2,25 / 6,25 = 0,36$$

Entre

$$\text{Metodo } 75 \quad 12 \quad 75 / 12$$

Dentro

TOTAL 81,75 15

ANALISIS DE LA VARIANZA CON DOS CRITERIOS DE CLASIFICACIÓN

Anova con dos criterios

En ocasiones, es deseable identificar dos causas posibles para las diferencias en la variable dependiente. Si es el caso, se lleva a cabo un programa ANOVA con dos criterios de clasificación, donde se identifican dos causas posibles para la variabilidad de la variable dependiente. Se toman al azar dos muestras de la población de interés y se usan los resultados muestrales para probar la hipótesis nula relevante.

EJEMPLO

Hace un par de ejemplos, el analista intentó determinar si había alguna diferencia en el promedio en dólares por compra entre tres tiendas. ¿ Qué ocurre si también quiere determinar si existe alguna diferencia en el promedio de compra debida a los efectos de dos campañas distintas de publicidad ?

Los datos de la tabla número 1 del ejemplo del que estamos tratando se vuelven a disponer de manera que se puedan examinar usando dos criterios de clasificación para el análisis de varianza. Hay tres grupos en el factor 1 (tiendas) y dos grupos en el factor 2 (campañas de publicidad). Se tomó una muestra de tres elementos ($n=3$) y se tomaron medidas para cada una de las seis celdas de la tabla ($3 * 2 = 6$).

Tabla número 5 Datos muestrales (dólares) de ANOVA para el ejemplo.

Campaña de

Publicidad TIENDA 1 TIENDA 2 TIENDA 3 **MEDIAS**

12,05 15,17 9,48

A (16,87) 23,94 (17,75) 18,52 (8,96) 6,92 **14,53**

14,63 19,57 10,47

25,78 21,4 7,63

B (20,58) 17,52 (18,52) 13,59 (8,48) 11,90 **15,86**

18,45 20,57 5,92

18,73 18,14 8,72 15,20

Media global 15,20 $r = 2$ $c=3$ $n=3$

- Media tienda 1 18,73
- Media tienda 2 18,14
- Media tienda 3 8,72
- Media campaña A 14,53
- Media campaña B 15,86
- Media tienda 1 y campaña A 16,87

- Media tienda 2 y campaña A 17,75
- Media tienda 3 y campaña A 8,96
- Media tienda 1 y campaña B 20,58
- Media tienda 2 y campaña B 18,52
- Media tienda 3 y campaña B 8,48

Interacción

Una razón más importante para aplicar el procedimiento de ANOVA con dos criterios de clasificación es que permite probar una tercera hipótesis. Como se obtendrá una muestra de cada combinación entre las variables en filas y columnas, se puede examinar el efecto de interacción de estas variables.

La interacción ocurre cuando los niveles de un factor se interrelacionan de manera significativa con los niveles del segundo factor ejerciendo influencia sobre la variable dependiente.

En el procedimiento de ANOVA con dos criterios de clasificación, la primera hipótesis nula para probar se refiere a la presencia de interacción:

H_0 : No hay interacción entre los factores en las filas y los factores en las columnas para la población bajo prueba.

Si se encuentra interacción, hay que determinar porqué ciertos niveles de un factor interactúan con ciertos niveles del segundo factor. Esto se hace examinando las diferentes medias en las celdas. La interacción se encuentra pocas veces, pero cuando la hay, el analista no puede interesarse por probar otras hipótesis.

Si no se encuentra interacción, las variables en filas y columnas se examinan para buscar diferencias en la variable independiente. Las hipótesis nulas bajo prueba son:

- H_0 : No existe diferencia en el valor promedio de la variable dependiente para las poblaciones en las filas.
- H_1 : No existe diferencia en el valor promedio de la variable dependiente para las poblaciones en las columnas.

EJEMPLO

Un analista de personal, quiere llevar a cabo un análisis de varianza con

dos criterios de clasificación para determinar si la variable tiempo con la compañía está afectada por uno de dos factores: La localización del empleado en las áreas de trabajo de la compañía y el nivel de salario del empleado. Cada empleado se asigna a una de las cuatro áreas de trabajo de la compañía en diferentes partes de la ciudad. Existen tres tipos de empleados según el método de pago: por horas, por mes y por año. El conjunto de datos consiste en muestras de cada combinación posible entre la localización del empleado y su método de pago. Esto da como resultado una tabla de datos con 12 celdas. Las hipótesis nulas que se van a probar son:

- H_0 : No hay interacción entre la localización y el método de pago.
- H_0 : No hay diferencia en el tiempo con la compañía debida a la localización.
- H_0 : No hay diferencia en el tiempo con la compañía debida al método de pago.

EJEMPLO

Un estudio investigó los efectos de la información respecto al precio, la

marca y la tienda sobre la evaluación del producto por los consumidores. Supongamos que la variable dependiente es la percepción del consumidor sobre la calidad del producto, medida en una escala numérica. Dos de los factores estudiados fueron precio y marca. Se tenían cinco precios para calculadoras (\$17, \$28, \$39, \$50 y la ausencia de precio) y tres marcas (Hewlett Packard, Royal y Sony). Los datos consistieron en muestras de todas las combinaciones posibles de precio y marca. Esto dio como resultado una tabla de datos con 15 celdas. Las hipótesis nulas bajo prueba son:

- H_0 : No hay interacción entre el precio y la marca.
- H_0 : No hay diferencia en la percepción de la calidad del producto debida a los diferentes niveles de precio.
- H_0 : No hay diferencia en la percepción de la calidad del producto debida a las diferentes marcas.

Son muchos los cálculos requeridos por un análisis de varianza con dos

criterios de clasificación. La disponibilidad generalizada de paquetes de computadora que realizan ANOVA ha eliminado prácticamente los cálculos manuales para esta técnica. Sin embargo, es importante saber que se está haciendo con los datos para lograr una interpretación y un entendimiento apropiados. Los cálculos específicos para un procedimiento de ANOVA con dos criterios de clasificación no se presentarán aquí, pero se describirá la naturaleza general del análisis y se interpretará una salida de computadora.

La suposición clave que fundamenta el ANOVA con dos criterios de clasificación es la misma que para el ANOVA con un criterio: Se supone que todas las poblaciones bajo estudio tienen la misma varianza. Si se tiene tres filas en la tabla de datos y cinco columnas, hay quince celdas y quince poblaciones que deben mostrarse. Independientemente de si las medidas de estas quince poblaciones son las mismas, debe suponerse que varían en el mismo grado. Todas deben tener la misma varianza para que el procedimiento de ANOVA funcione correctamente.

Existen cuatro formas de estimar la varianza común de las poblaciones en el procedimiento de ANOVA con dos criterios de clasificación. Una de estas formas, el método dentro, produce una estimación fiable de esta varianza independientemente de que cualquiera de las tres hipótesis nulas sean ciertas. Igual que en el procedimiento de ANOVA con un criterio, el método dentro mide la variabilidad de cada valor muestral alrededor de su propia medida de la celda. Aún cuando varias de las celdas en la tabla de datos tengan medias diferentes, esto no influirá en los cálculos de la varianza estimada con el método dentro. Al calcular la suma de cuadros usando el método entre se compara el primer dato con la media de la celda en la que está. La diferencia se eleva al cuadrado y se suma a los cuadrados de las diferencias entre todos los otros valores de la muestra y las medias de sus propias celdas. El valor que se obtiene se divide entre el número apropiado de grados de libertad, $rc(n - 1)$. Como la media de la celda se resta de cada uno de los n elementos en la celda, uno de estos elementos no tiene libertad para variar. Cada celda tiene entonces $(n - 1)$ grados de libertad, y hay r (el número de filas) multiplicado por c (el número de columnas) celdas. Esta estimación dentro de la varianza es el denominador de cada cociente F .

El segundo método para estimar la varianza es válido solo si no hay interacción entre las poblaciones. Si la hay, este método produce una estimación inflada. El valor de gl se calcula de la misma manera que para la prueba de la tabla de contingencia: $(r - 1)(c - 1)$.

El tercer método para estimar la varianza produce una estimación válida sólo si la hipótesis nula sobre la igualdad de la media de columnas es cierta. Si esta hipótesis es falsa, se obtendrá una estimación inflada. Esto es lo mismo que usar el método entre para estimar la varianza en un procedimiento de ANOVA con un criterio.

Los grados de libertad son el número de columnas menos uno, $(c - 1)$.

El último método para estimar la varianza es válido sólo si la hipótesis sobre medias iguales en las filas es cierta. Si no lo es, se obtiene una estimación inflada. De nuevo, el procedimiento es similar al método entre para estimar la varianza en un ANOVA con un criterio. Los grados de libertad son el número de filas menos uno, $(r - 1)$. La tabla de continuación contiene las fórmulas para el procedimiento de ANOVA con dos criterios de clasificación.

Tabla de análisis de varianza con dos criterios de clasificación.

Fuente de Coeficiente

variación SC gl Estimación de F

Filas $cn'' (x_{ri} - \bar{x})^2 r - 1$ SC_r / gl_r S^2_r / S^2_w

I

Columnas $cn'' (x_{cj} - \bar{x})^2 c - 1$ SC_c / gl_c S^2_c / S^2_w

j

Interacción $n [\text{''''} (x_{ij} - \bar{x}_i - (r - 1)(c - 1) SC_i / gl_i S^2_i / S^2_w$

i j

$- x_{cj} - \bar{x})^2]$

Grupos $\text{''''''} (x_{ijk} - \bar{x}_{ij})^2 rc (n - 1) SC_w / gl_w$

dentro i j k

TOTAL $\text{''''''} (x_{ijk} - \bar{x}_{ij})^2 nrc - 1$

I j k

j = Número de la columna.

i = Número del renglón.

k = Número de la observación dentro de una celda.

r = Número de filas.

c = Número de columnas.

n = Número de observaciones en cada celda.

El resultado final de un procedimiento de ANOVA con dos criterios es el cálculo de tres cocientes F . El denominador para cada uno

de estos cocientes es la estimación del método dentro para la varianza desconocida de la población. Los

numeradores de los cocientes

son las estimaciones obtenidas bajo la suposición de que cada una de las tres hipótesis es cierta. Cada cociente F se examina para ver

si es muy grande. Cualquier cociente F que sea mas grande que el valor de la tabla F da como resultado el rechazo de la hipótesis nula

correspondiente. La forma general de cada uno de los tres cocientes F es:

si2 (estimación de interacción de)

$$\bullet F = \frac{\text{---}}{\text{---}} \quad \text{(interacción)}$$

Sw2 (estimación por el método dentro de)

Sc2 (estimación por el método entre por columna de)

$$\bullet F = \frac{\text{---}}{\text{---}} \quad \text{---(columna)}$$

Sw2 (estimación por el método dentro de)

Sr2 (estimación por el método entre por filas de)

$$\bullet F = \frac{\text{---}}{\text{---}} \quad \text{(filas)}$$

Sw2 (estimación por el método dentro de)

Si las tres hipótesis nulas son ciertas, los cálculos para los numeradores y denominadores de estos tres cocientes F serán estimaciones válidas de la misma varianza poblacional desconocida. Como se ha visto, una razón de este tipo se obtiene de la distribución F . Sin embargo, si cualquiera de las tres hipótesis nulas es falsa, el numerador de la razón correspondiente estará inflado y dará un valor grande de F que llevará el rechazo de la hipótesis nula.

EJEMPLO

La siguiente tabla presenta la tabla ANOVA con dos criterios de

clasificación para el ejemplo con el que comenzamos este tema (tres tiendas). Se calcularon cuatro estimaciones de la varianza común de todas las poblaciones. No obstante, solo el método dentro produce una estimación válida sin importar el estado de ninguna hipótesis nula. La evidencia muestral ha producido el valor 16,019 como la estimación del método dentro para .

Fuente de

variación SC gl Estimación de Coeficiente F

Filas 8,013 1 8,013 0,50

Columnas 378,381 2 189,90 11,81

Interacción 13,851 2 6,925 0,43

Dentro 192,223 12 16,019

TOTAL 592,468 17

Las tres hipótesis nulas son:

- H_0 : No hay interacción entre tienda y campaña publicitaria en la población.
- H_0 : Las poblaciones en las filas (campaña publicitaria) tienen ambas la misma medida.
- H_0 : Las poblaciones en las columnas (tiendas) tienen todas las mismas medidas.

Se calculan los cocientes F de la tabla anterior dividiendo cada una de

las estimaciones de correspondientes a las tres hipótesis nulas entre 16,019, la estimación válida de .
Estos cálculos son:

COCIENTE F	GRADOS DE LIBERTAD	F crítica (.05)
8.013/16.019 = 0.500	1.112	4.75
189.190/16.019 = 11.810	2.12	3.88
6.925/16.019=0.432	2.12	3.88

Observe que la estimación dentro de ($S_2w = 16,019$) se usa en todos los denominadores. Además se dan los grados de libertad para el numerador y el denominador para cada prueba de hipótesis. Estos valores aparecen en la tabla anterior para cada fila de la tabla.

A continuación se encuentran las razones críticas de la tabla F calculando para interacción (0,432) es menor que el valor crítico (3,88), de manera que la hipótesis nula no se rechaza. El valor F calculado para las filas (0,500) es menor que el valor crítico (4,75), por lo que no se rechaza la hipótesis nula para las filas. El valor F calculado para las columnas (11,810) es mayor que el valor crítico (3,88) y la hipótesis nula para las columnas se rechaza.

Las conclusiones para el análisis de varianza con dos criterios de clasificación son:

- No hay interacción entre tiendas y campañas publicitarias en la población.
- Las campañas publicitarias tienen ambas la misma media.
- Las tiendas tienen diferentes medias.

Para las primeras dos conclusiones existe la posibilidad de un error tipo II. En cualquier caso, la hipótesis nula puede de hecho ser falsa. Para la última conclusión se tiene la posibilidad de haber rechazado una hipótesis nula que en realidad es cierta. Un paso final muy importante es la evaluación de los riesgos de error y las penalizaciones asociadas.

La tercera conclusión establece que las diferentes tiendas no tienen la misma media en la población.

Vale la pena hacer algunos comentarios finales sobre los procedimientos del análisis de varianza con uno o dos criterios de clasificación. Como se ha comentado varias veces a lo largo del tema, la suposición clave de ANOVA es que todas las poblaciones tengan la misma varianza. En realidad, son tres las suposiciones que

deben cumplir para que el procedimiento ANOVA arroje resultados adecuados.

- Todas las poblaciones que se prueban deben tener la misma varianza para la variable dependiente.
- Todas las poblaciones que se prueban deben seguir una distribución normal para la variable dependiente.
- Las muestras tomadas de las poblaciones que se prueban deben ser aleatorias.

Deben verificarse estas tres suposiciones para asegurar un análisis válido.

En ocasiones, estas suposiciones se ignoran, en particular porque los paquetes de computadora que ejecutan ANOVA no preguntan al analista si se han tenido en cuenta. El analista debe, por lo menos, tener una idea intuitiva de que se cumplen las dos primeras y que se están usando muestras aleatorias.

OTROS DISEÑOS ANOVA

Los diseños con uno y dos criterios de clasificación descritos en este capítulo constituyen los procedimientos básicos de ANOVA que se usan en la mayor parte de las aplicaciones. En ocasiones se emplean modificaciones a estos procedimientos al examinar los efectos de diferentes factores sobre una variable de interés. En esta sección se describen las cuatro variaciones básicas más comunes.

- Los diseños básicos de ANOVA suponen que los tamaños de las muestras son iguales. Para el ANOVA con un criterio de clasificación se usa el mismo tamaño de muestra en cada tratamiento. En el diseño de dos criterios se usa el mismo tamaño de muestra en cada celda de la tabla de datos. Algunos diseños más complejos pueden manejar muestras desiguales. Esto puede ser conveniente cuando hay diferentes proporciones de los elementos de una población y el analista quiere reflejar estas diferencias en la muestra.
- En los diseños descritos en este tema, se muestrearon todas las poblaciones de interés. Suponga que un ANOVA con dos criterios de clasificación básicos se obtiene una muestra de todas las poblaciones. Un diseño alternativo de ANOVA elige al azar esas poblaciones para la muestra y extiende los resultados a esas poblaciones.
- El procedimiento básico de ANOVA se puede ampliar para cubrir tres factores o más. Una alternativa es elegir al azar operadores (3 factores) para la muestra y extender los resultados del ANOVA a todas las poblaciones.

El estudio factorial de ANOVA con dos criterios estudiaba estos efectos. De hecho, este diseño factorial de tres criterios de clasificación obtiene una muestra de cinco niveles de precio, tres nombres de marca y tres nombres de tienda.

- En el diseño básico de ANOVA con un criterio de clasificación, los sujetos se asignan al azar a los tratamientos. Esta asignación aleatoria proporciona alguna seguridad de que los sujetos en cada tratamiento son más o menos los mismos, y elimina así los efectos de sujetos diferentes.

El diseño de bloques aleatorizados elimina el efecto de las diferencias entre los sujetos de los tratamientos, al someter a cada sujeto a todos estos tratamientos.

Los distintos métodos de llevar a cabo el procedimiento de análisis de varianza a veces reciben el nombre de diseños de experimentos. El término experimento sugiere una aplicación científica más que una de negocios, y esto es cierto en general. Las aplicaciones de negocios, con frecuencia incluyen la observación más que la manipulación de variables, así que los procedimientos avanzados de diseño o experimentos no son muy comunes. Sin embargo hay ocasiones en que las situaciones de negocios se pueden controlar para examinar el efecto de los distintos factores sobre la variable de interés. En estos casos, los procedimientos avanzados pueden ser muy útiles y, de hecho, el control de calidad y la investigación para el diseño de productos son dos áreas de importancia en este sentido.

LA PRUEBA DE KRUSKAL-WALLIS

La estructura de los datos

Consideremos un diseño con un solo factor, completamente aleatorizado. La prueba que estudiamos es una extensión de la prueba U de Mann – Whitney para el caso de k muestras o k niveles de un factor, mutuamente independientes. Y es una alternativa al ANOVA de un factor, efectos fijos, complementos al azar, cuando no se cumplen los supuestos paramétricos del mismo.

Transformemos las $n_1 + n_2 + \dots + n_k = N$ observaciones en rangos u ordenes O_{ij} , de modo que la menor de las Y_{ij} reciba el valor 1, la siguiente el valor 2, ..., y la mayor el valor N. Si suponemos que la variable es continua no deberían existir empates.

Llamamos O_i a la suma de los ordenes que han correspondido a la muestra i:

$$O_i = \sum_j O_{ij}$$

$$O_i$$

$$\text{Por tanto: } O_i = \frac{\sum_j O_{ij}}{n_i}$$

$$n_i$$

será la correspondiente media aritmética. La suma de los N valores valdrá

$$N \frac{N(N+1)}{2}$$

$$\sum_j O_{ij} = \frac{N(N+1)}{2} \quad (i' = 1, 2, \dots, N)$$

$$\sum_j O_{ij}^2$$

Y su promedio:

$$\sum_j O_{ij}^2 / (N+1)$$

$$O = \frac{\sum_j O_{ij}^2}{N} = \frac{\sum_j O_{ij}^2}{N}$$

$$N^2$$

Estadísticos de contraste

Podemos considerar que las k muestras son aleatorias e independientes extraídas de un población finita de tamaño N, en la cual:

$$\sum_j O_{ij}^2 - \frac{(\sum_j O_{ij})^2}{N} = \sum_j O_{ij}^2 - \frac{O_i^2}{N}$$

$$= \sum_j O_{ij}^2 - \frac{O_i^2}{N} = \sum_j O_{ij}^2 - \frac{O_i^2}{N}$$

$$\sum_j O_{ij}^2 - \frac{O_i^2}{N}$$

$$(N+1)$$

Entonces, $H = \frac{1}{N(N+1)} \sum_{i=1}^k n_i(n_i+1) - \frac{3}{4}(N+1)$

$N(N+1) n_i$

1

$E(H) = \frac{1}{N(N+1)} \sum_{i=1}^k n_i(n_i+1) - \frac{3}{4}(N+1)$

$N(N+1)/12$

El estadístico de contraste tiene un valor esperado que no depende del número de observaciones; solo depende del número de muestras.

Es fácil conocer su distribución muestral. Bajo la hipótesis nula, los $1, 2, \dots, N$ ordenes estarán distribuidos al azar en las k muestras, con la única restricción de que haya n_i de ellos en cada una. Las distintas maneras como N elementos se pueden distribuir en k muestras de tamaño n_i son:

$N!$

$n_1! n_2! \dots n_k!$

Si para cada una de esas posibilidades calculamos el valor de H y llamamos a las que coincidan $t(h)$, la posibilidad de aparición de este valor viene dada por

$n_1! n_2! \dots n_k!$

$f(h) = \frac{t(h)}{N!}$

$N!$

Averiguar las probabilidades exactas para cada caso es bastante engorroso. Es más útil servirse de una aproximación si los n_i son moderadamente grandes. Para cada muestra tenemos:

$[O_i - (N+1)/2]$

$Z_i = \frac{[O_i - (N+1)/2]}{\sqrt{N(N+1)/12}}$

$\text{Var}(O_i)$

Cuya distribución tiende a $N(0,1)$. Por tanto si Z_i^2 es $\hat{E}$. De modo que :

$N - n_i [O_i - (N+1)/2]^2$

$H = \frac{1}{N(N+1)} \sum_{i=1}^k n_i(n_i+1) - \frac{3}{4}(N+1)$

$N \text{ var}(O_i)$

Es la suma de k términos Z_i^2 ponderados, deberá seguir una distribución . Como se debe cumplir que:

$N(N+1)$

$$n_i O = \text{-----}$$

2

no todos estos términos serán independientes, por lo que los grados de libertad serán $k - 1$.

H es un estadístico de contraste que la sigue la distribución con $k - 1$ grados de libertad.

Si hay empates es las observaciones y se promedian los rangos, la distribución muestral de H se verá afectada. Cuantos mas empates hay, mas conservadora se hace la prueba, lo que significa que se torna mas difícil considerar H como significativo.

Se precisa una corrección que depende exclusivamente del numero de observaciones empatadas. Si llamamos t_k al numero de valores empatados en el orden k , la corrección consiste en dividir H por el siguiente termino:

r

$$(t_k^3 - t_k)$$

k

$$C = 1 - \text{-----}$$

$$N^3 - N$$

La prueba es fundamentalmente una prueba respecto a los promedios, muy poco sensible a desigualdad de las poblaciones subyacentes en sesgo, curtosis o amplitud de los datos. Es decir, que bajo H_0 de igualdad de promedios, el estadístico H tiende a ser pequeño aunque las poblaciones difieran en la forma o en las escala, lo cual implica que la probabilidad de rechazar H_0 es verdadera sigue siendo próxima a alfa.

EJEMPLO:

Un psicólogo esta interesado en el efecto que determinado tipo de castigo y determinado tipo de premio tienen sobre la conducta agresiva de niños de 8 a 10 años. Con este fin, prepara un experimento con 22 niños de esas edades tomados al azar de un colegio de EGB. Al azar los asigna a las dos condiciones experimentales y al grupo control. Al cabo de unos meses les pasa una prueba de agresividad y los obtiene los siguientes datos.

Se desea saber si hay evidencias suficientes para concluir que diferencia entre las medianas de los tratamientos, a nivel de significación de 0,01.

- Hipótesis

H_0 = Las tres poblaciones subyacentes son idénticas.

H_1 = Las tres poblaciones no tienen la misma mediana.

- Supuestos:

- Las k muestras de tamaños n_1 , n_2 , n_3 son aleatorias
- Las N observaciones son mutuamente independientes
- La variable dependiente es continua.
- El nivel de medida es al menos ordinal.
- Las poblaciones son idénticas excepto posiblemente en los promedios.

- Estadístico de contraste

Transformemos los datos en ordenes:

Oi

Control: 13.5, 21, 19, 1.5, 13.5 68.5

Premio: 16, 22, 19, 16, 10.5, 8.5, 7, 4.5 103.5

Castigo: 16, 19, 8.5, 10.5, 4.5, 3, 1.5, 6, 12 81

Calculo H:

12 68.52 103.5 812

$H_k = \frac{1}{n(n-1)} \left(\sum_{i=1}^n R_i^2 - \frac{n(n+1)^2}{4} \right) - 3(23) =$

$\frac{(22)(23) + 5 + 8 + 9}{12 \cdot 11} - 69 =$

$= 0.0237(938.45 + 1339.03 + 729) - 69 = 71.3 + 69 = 2.3$

Hago la corrección por empates:

(13.5, dos veces): $23 - 2 = 21$

(19, tres veces): $33 - 3 = 30$

(1.5, dos veces): $23 - 2 = 21$

(16, tres veces): $33 - 3 = 30$

(10.5, dos veces): $23 - 2 = 21$

(8.5, dos veces): $23 - 2 = 21$

(4.5, dos veces): $23 - 2 = 21$

$(6 + 24 + 6 + 24 + 6 + 6 + 6) = 78$

$C = 1 - \frac{78}{223 - 22} = 1 - \frac{78}{201} = 0.9993$

$223 - 22 = 201$

23

$h^*k = \frac{2.3}{2.316} = 0.993$

0.993

- Zona Crítica

Puesto que un grupo tiene tamaño superior a 8 , utilizamos la distribución de t con 2 g.l. La zona critica estará formada por todos los valores iguales o superiores a $t_{0.99, 2} = 9.21$

- Decisión:

Puesto que $2.316 < 9.21$, no podemos rechazar la hipótesis nula ($p > 0.05$).

- Conclusión

No hay evidencia suficiente en los datos para afirmar que los tratamientos tienen efecto diferencial.

RESUMEN

Además de medias y proporciones, muchas veces interesa la variabilidad de las poblaciones. En este tema se presentaron métodos para la viabilidad de una sola población y para comparar las viabilidades de dos poblaciones.

También se presento en este tema la manera de examinar los efectos de los diferentes factores sobre la variable de interés (Variable independiente). En el análisis de varianza con un criterio, las medidas de la variable dependiente se hacen para cada nivel del factor que se piensa que afecta a esta variable. Se pueden examinar dos factores relevantes al mismo tiempo en el procedimiento de ANOVA con dos criterios de clasificación, y estudiar los efectos de tres o mas factores sobre la variable dependiente a través de procedimientos mas avanzados.

El análisis de varianza es un buen ejemplo de una técnica estadística que resulta muy practica debido al uso generalizado de las computadoras. El volumen de cálculos es tal que es muy difícil realizar un diseño de cualquier tamaño útil solo con cálculos manuales. Los programas de computadora que ejecutan ANOVA están disponibles para computadoras personales al igual que para las mas grandes. Estos programas, por lo general, realizan análisis con uno y dos criterios de clasificación y algunas veces también ofrecen técnicas mas avanzadas.

APLICACIONES DE CONCEPTO ESTADÍSTICO AL MUNDO DE LOS NEGOCIOS

Existen muchas aplicaciones de las técnicas de ANOVA presentadas a lo largo de este tema que son importantes para el mundo de los negocios. Cuando el valor promedio de alguna variable se compara con tres o mas poblaciones, las conclusiones que resultan de un estudio de ANOVA pueden ser muy útiles para el administrador. Con frecuencia se modifican las variables de producción para determinar que combinación lleva al proceso de manufactura optimo.

EJERCICIOS

Numero 1. – Prueba de varianza con una población.

Los instrumentos científicos de medición como el altímetro de un avión, deben proporcionar lecturas correctas y con errores de medición muy pequeños. El gerente de producción esta preocupado por el índice de variación en las lecturas producidas por los altímetros de su compañía. Los altímetros están diseñados para tener una desviación estándar de 200 pies. El gerente decide probar si la variabilidad de estos instrumentos es mayor que 200 pies. Selecciona una muestra de siete altímetros y calcula una desviación estándar de 250 pies.

- Establezca las hipótesis nula y alternativa.

- Calcule los grados de libertad.
- Establezca las reglas de decisión para un nivel de significancia de 0,05
- Pruebe si la variabilidad de los altímetros de la compañía es mayor que 200 pies.
- La hipótesis nula y alternativa son:

$H_0 : < 40,000$ menor e igual

$H_1 : > 40,000$ mayor o igual

- $gl = (n - 1) = (7 - 1) = 6$
- La regla de decisión es:

Si $> 12,59$, se rechaza la hipótesis nula de que la varianza de la población es 40,000 (se rechaza H_0 si $> 12,59$).

$(n-1)s^2 / (7 - 1) = (250)^2$

- $= \frac{250^2}{6} = 9,375$

2002

Como el estadístico de prueba calculado (9,375) es menor al valor crítico de la tabla (12,59), la hipótesis nula no se puede rechazar a un nivel de significancia de 0,05. No existe suficiente evidencia muestral para concluir que la desviación estándar poblacional es mas de 200 pies.

Número 2. – Prueba de varianza con dos poblaciones.

Carla Mitchell, analista de los laboratorios Abbott, un fabricante nacional de medicamentos, esta preocupada por la calidad de uno de sus productos. Abbott compra el material para fabricar este producto a dos proveedores. El nivel de defectos en la materia prima es aproximadamente el mismo entre los dos proveedores, pero Carla esta preocupada por la variabilidad que existe de un embarque a otro. Si el nivel de defectos tiende a variar en forma excesiva para uno proveedor, puede afectar la calidad del medicamento. Para comparar la variación relativa de los dos proveedores, Carla selecciona 11 embarques de cada uno y mide los porcentajes de defectos en la materia prima, junto con la desviación estándar. Los resultados son:

$S_1 = 0,61$ $n_1 = 11$ (proveedor 1)

$S_2 = 0,29$ $n_2 = 11$ (proveedor 2)

- Establezca la hipótesis nula y alternativa
- Calcule los grados de libertad
- Establezca la regla de decisión para un nivel de significancia de 0,05.
- Pruebe si la variabilidad del nivel de defectos por embarque de un proveedor 1 es mayor que para el proveedor 2.
- Las hipótesis nula y alternativa son:

$H_0 : \hat{\sigma}^2 - \sigma^2 < 0$

$H_1 : \hat{\sigma}^2 - \sigma^2 > 0$

- $gl_1 = (n_1 - 1) = (11 - 1) = 10$

$$gl_2 = (n_2 - 1) = (11 - 1) = 10$$

- El valor crítico F es 2,97. La regla de decisión es

Si el cociente F calculado es mayor que 2,97, se rechaza H_0 (se rechaza H_0 si $F > 2,97$)

$$S_{12} (0.61)^2$$

$$\bullet \text{-----} = \text{-----} = 4.42$$

$$S_{22} (0.29)^2$$

Una de las varianzas muestras es 4,42 veces la otra. La hipótesis nula se rechaza porque el estadístico (2,97). Carla debe concluir que la variabilidad en los niveles de defectos de los embarques para el proveedor 1 es mayor que para los del proveedor 2.

Número 3.- Análisis de la varianza con un criterio de clasificación.

La dueña de la corporación LUZ COLOR decide reemplazar varias pinturas de aerosol. Después de investigar la situación, concluye que 4 marcas parecen comparables en términos de coste y vida útil proyectada, ella determina que el factor decisivo entre las cuatro marcas es la cantidad de pintura que se usa en la operación normal. Mide entonces el espesor de la pintura, en milímetros, para varias pruebas , con los siguientes resultados.

	AEROSOL 1	AEROSOL 2	AEROSOL 3	AEROSOL 4
	5,4	6,1	8,2	7,2
	5,9	5,9	8,5	6,5
	6,2	6,3	6,9	6,8
	7,0	6,5	9,4	7,1
	5,1	7,2	7,9	7,4
	5,5	6,9	8,6	6,7
MEDIA =	5,85	6,48	8,25	6,95

$$\text{Media Global} = x = 6.88$$

- Establezca las hipótesis nula y alternativa.
- Calcule los grados de libertad
- Establezca la regla de decisión si se prueba la hipótesis nula al 0,01 de nivel de significancia.
- ¿ A que conclusión se llega?
- Las hipótesis nula y alternativa son:

$$H_0 : = = =$$

H_1 : No todas las poblaciones tienen la misma media.

$$\bullet G_{ln} = c(n - 1) = 4(6 - 1) = 20$$

$$gl_2 = (c - 1) = (4 - 1) = 3$$

- Encuentra el elemento de la tabla F para la columna 3 y la fila 20. Para un nivel de significancia de 0,01 este valor crítico es 4,94. La regla de decisión es:

Si el cociente F calculado es mayor que 4,94, se rechaza la hipótesis nula (se rechaza H_0 si $F > 4,94$)

- La siguiente tabla contiene la tabla ANOVA para este problema.

Fuente de variación	SC	gl	Estimación de	Cociente F
Grupos entre	18,61	3	6,203	16,37
Grupos dentro	7,57	20	0,379	
Total	26,18	23		

Las sumas de cuadrados en esta tabla son:

$$(5,4 - 5,85)^2 + (5,9 - 5,85)^2 + (6,2 - 5,85)^2 +$$

$$(7,0 - 5,85)^2 + (5,1 - 5,85)^2 + (5,5 - 5,85)^2 +$$

$$(6,1 - 6,48)^2 + (5,9 - 6,48)^2 + (6,3 - 6,48)^2 +$$

$$(6,5 - 6,48)^2 + (7,2 - 6,48)^2 + (6,9 - 6,48)^2 +$$

$$(8,2 - 8,25)^2 + (8,5 - 8,25)^2 + (6,9 - 8,25)^2 +$$

$$(9,4 - 8,25)^2 + (7,9 - 8,25)^2 + (8,6 - 8,25)^2 +$$

$$(7,2 - 6,95)^2 + (6,5 - 6,95)^2 + (6,8 - 6,95)^2 +$$

$$(7,1 - 6,95)^2 + (7,4 - 6,95)^2 + (6,7 - 6,95)^2 +$$

$$SC_n = 7,57$$

$$(5,85 - 6,88)^2 + (6,48 - 6,88)^2 + (8,25 - 6,88)^2 + (6,95 - 6,88)^2 = 3,1$$

$$SC_b = 6(3,1) = 18,6$$

$$Sb^2 / gl_b (18,6/3) 6,2$$

$$F = \frac{6,2}{0,379} = 16,36$$

$$Sn^2 / gl_w (7,57/20) 0,379$$

La hipótesis nula se rechaza ya que el estadístico de prueba (16,36) es mayor que el valor crítico (4,94). La conclusión es que el espesor de la pintura difiere entre todas estas cuatro marcas de aerosol.

Número 4. –

Pedro Martínez, analista de la compañía de investigaciones de mercado profesional marketing, esta llevando a cabo de un estudio para un cliente a fin de determinar si la edad y la escolaridad afectan a los ingresos percibidos. La tabla siguiente da los resultados del conjuntos de datos de Julie. ¿Cual será la conclusión de Julie si hace la prueba con un nivel de significancia de 0,05?

Grupo de Edad	Preparatoria	Universidad	Posgrado
---------------	--------------	-------------	----------

18 a < 30	\$25,000	\$36,250	\$42,500
	31,450	39,400	46,000
	27,500	35,450	47,250
30 a < 50	28,000	46,250	52,600
	30,950	44,400	56,700
	26,250	48,450	57,750
50 +	35,000	46,250	62,800
	38,250	49,400	66,700
	37,700	55,450	70,250

Los datos anteriores se ejecutan en un programa de computadora

Los cocientes F calculados se comparan con los valores F críticos.

gl	F crítica ($\alpha = 0,05$)	Estadístico de prueba F calculado
2,18	3,55	55,87
2, 18	3,55	161,24
4,18	2,93	4,78

De acuerdo con los valores críticos apropiados y los cocientes críticos de F de la anterior tabla , las tres hipótesis nulas se rechazan. Con esto como fundamento, Julie concluye que:

- Existe interacción entre las celdas y las escolaridad. Parece haber diferencias inesperadas cuando ciertos campos de edad se comparan con ciertos niveles de escolaridad . El rechazo de la hipótesis nula de que no hay interacción hace que Julie regrese a los datos muestrales para buscar las combinaciones de edad / escolaridad que produjeron los resultados inesperados.
- Los grupos de edad tienen diferentes niveles de ingresos
- Los grupos de escolaridad tienen diferentes niveles de ingresos.

EL ANÁLISIS DE LA VARIANZA (ANOVA) ES UN PROCEDIMIENTO ESTADÍSTICO PARA DETERMINAR SI LAS MEDIAS DE TRES O MAS POBLACIONES IGUALES.

A1 O11 O12 ... O1j ... O1n1 O1 O1

A2 O21 O22 ... O2j ... O2n2 O2 O2

.-----

.-----

.-----

.Ai Oi1 Oi2 ... Oij ... Oini Oi Oi

.-----

.-----

Ak Ok1 Ok2 ... Okj ... Oknk Ok Ok