ANAVA

CONTRASTE DE HIPÓTESIS PARA TRES O MAS VARIABLES

Este tema de lo que trata es de como debemos realizar una comprobación de la hipótesis nula con más de dos variables, para ello se utilizará el cálculo ANAVA que en inglés significa cálculo de análisis de varianza. A pesar de este nombre es algo prácticamente igual a la t de student dado que lo que busca es saber si se cumple la hipótesis nula con tres variables o más, teniendo en cuenta las medias aritméticas, aunque su nombre hable de las varianzas.

Este procedimiento por tanto es muy bueno para pedagogía, dado que no solo podemos ver si un tratamiento es útil o no lo es sino que también podemos ver cual es la opción más adecuada de la VI, optando por suministrar dos cantidades diferentes y comparándolas, entre si y con una muestra de control.

Es verdad que podríamos hacer para saber esto t de student de dos en dos primero mirar si hay diferencias estadisticamente significativas en la VD dependiendo de diferentes opciones de la VI de dos en dos, así por ejemplo nos permite saber si tomar el medicamento con una cantidad A y no tomar nada, despues comparar si hay diferencias entre tomar la cantidad A y la cantidad B , y mirar si la cantidad B tiene diferencias estadisticamente significativas entre B y no tomar nada. Y así con todas las opciones posibles. Esto sería un gran trabajo, ya que habría que hacer gran cantidad de comparativas, de hecho hay una fórmula para calcular el numero de comparativas que hay que hacer teniendo en cuenta el número de variables y la fórmula es esta $k(k-1)/2$) .

Además hay que tener en cuenta que cuentos más cálculos hagas, mas t de student hagas, más posibilidades de cometer un error, lo que conocemos como el error tipo I , ya que por cada cálculo habrá un aumento del alfa proporional.

Así al hacer la ANAVA muchas menos posibilidades habrá de cometer el error tipo I.

Antes de comenzar debemos saber algunos términos, la variable independiente se denominará en la ANAVA factores, pero sus opciones posibles en que se divide esta variable son llamados niveles.

A su vez dentro de los estudios ANAVA hay dos tipos de estudio, cuando sabemos exactamente los niveles del tratamiento , cual es la cantidad exacta a dar en las dos

opciones. Este caso se conoce como efectos fijos o modelos fijos.
El otro tipo de anava es cuando no conocemos exactamente las cantidades ideales a comparar del tratamiento, y lo único que nos interesa es saber si el efecto aumenta cuanto más aumentala variable dependiente, no interesa el valor exacto de esta VI. Este caso se conoce como modelo aleatorio y efecto aleatorio. En este caso también es necesario ver si el número de individuos de las diferentes muestras son iguales o no, sino lo son estaremos hablando de modelos no equilibrados y si son del mismo tamaño se llama modelo equilibrado.
En general esta prueba busca contrastar la existencia de diferencias de medias entre tres o más muestras, pero teniendo en cuenta la variabilidad de estas, de hay su nombre de analisis de varianzas.
A fin de cuentas el analisis ANAVA se centra en la varianza general que se conocerá a través de la varianza intergrupos que es la que se centra en las variables independientes que controla al investigador y sus distintas categorías,o sea varianza atribuible a los distintos niveles del factor en estudio. Pero también se centra en la varianza intragrupos, que es la varianza del posible error, cuanto puede variar el error que podemos cometer porque los grupos de partida son diferentes, o porque el propio grupo entre si son diferentes.
Entonces el cálculo de la ANAVA solo trata de relacionar ambas varianzas la intergrupos y la intragrupos, por ello lo que toca ahora es conocer como se realizan estos dos cálculos. Ya que cuando la varianza intergrupos, que se centra en las variables independientes las que manipula el investigador y sus diferencias, es mayor que la varianza intragrupos, que se centra en las diferencias de errores posibles en cada caso, se rechaza la hipótesis nula.
Y ahora a trabajar:
Un investigador desea comprobar si el tipo de personalidad influye en el número de intentos que tarda en contestar un determinado test. Para eso clasifica dos tipos diferentes de personalidad la personalidad A y la B. Su hipótesis considera que la personalidad tipo A debe dar lugar a un número superior de intentos a los originados con el tipo B. Tras haber aplicado varios test de personalidad para seleccionar a los sujetos y escoge al azar 12 sujetos de tipo A y 14 de tipo B y se les aplica dicho test Así anotamos el número de intentos que ha necesitado para responderlo. Se considera que se cumplen los supuestos de las pruebas parametricas . Su hipótesis on un nivel alfa = 0,01.
Los datos son los siguientes.

X_a	X_b	X^2a	X^2b
38	33	1444	1089
36	32	1296	1021
32	31	1024	961
30	30	900	900
29	3260	841	900

29	26	841	676
28	25	784	676
27	23	729	625
27	22	729	529
25	22	625	484
22	21	484	484
19	20	361	400
	18		321
	16		256

Con estos datos podremos hacer los sumatorios de cada uno porque con estos sumatorios podremos calcular las varianzas , que a fin y al cabo son lo que se necesita para trabajar esta anava.

grupo	n	ΣX	ΣX^2	$(\Sigma X)^2$	media	Desviacion
A	12	342	10058	116964	28,5	5,32
B	14	354	9328	125316	25,29	5,38
Total	26	696	19386	484416	26,77	5,49

Con estos datos se comienza a calcular las sumas de cuadrados, de tres grupos , la del total SCt, la suma de cuadrados SCe, la de dentro del grupo o intragrupos SCd.
La primera es

$$S.C_T = \sum_{i=1}^{N} X_1^2 - \frac{\left(\sum_{i=1}^{N} X_1\right)^2}{N}$$

a la segunda parte de esta formula , la del sumatorio de x puesto al cuadrado y dividido entre n, se le denomina término de correción C .
Con todo esto la formula la podemos rellenar con los datos que aparecen arriba:

$$S.C._T = 19.386 - \frac{696^2}{26} = 19.386 - 18.631,38 = 754,62.$$

Una vez calculado esto pasaremos a calcular el SCe, la suma de los cuadrados entre grupos, o sea la varianza de los diferentes grupos o niveles:

Suma de cuadrados entre grupos (SC_E)

$$S.C._E = \sum_{i=1}^{G} \frac{\left(\sum_{i=1}^{n} X_1\right)^2}{n} - \frac{\left(\sum_{i=1}^{N} X_1\right)^2}{N}$$

sustituyendo los valores lo unico que teneis que tener claro es que la primera parte de la formula, es el sumatorio de x que yo pondría al cuadrado dividido entre el número de gente , tanto de un grupo como del otro grupo, si hubiese más grupos, tendríamos que hacer lo mismo en vez de dos veces en tres veces o cuatro …

$$S.C._E = \left(\frac{342^2}{12} + \frac{354^2}{14}\right) - \frac{696^2}{26} = 9.747 + 8.951,14 - 18.631,38 = 66,76$$

La última de las varianzas a calcular son la suma de cuadrado intragrupos. SCd.

Suma de cuadrado dentro (SC_D), cuyos términos ya han sido explicados:

$$S.C._D = \sum_{i=1}^{N} X_1^2 - \sum_{i=1}^{G} \frac{\left(\sum_{i=1}^{n} X_1\right)^2}{n}$$

La SCd= 19,386 -(9,747+8,951,14)= 687,86
Esto nos sirve para comprobarlo:
SCd=SCt – SCe = 754,62-66,76= 687, 86
y por lo tanto SCe+ SCd= SCt
Ahora pasaremos a mirar el cuadro de la ANAVA donde colocaremos estos mismos datos que hemos calculado ahora. Al que sumaremos las MC las medias cuadraticas, dividiendo las varianzas respectivas con sus propios g.l. . Cuando tengamos reyenada la tabla podemos hacer la F , dividiendo la mediacuadratica entre grupos entre la media cuadratica intragrupos.
 Ahora miraremos los respectivos grados de libertad. Para la varianza total SCt su g.l. es N-1 o el numero total de sujetos menos uno, para el SCe= G- 1 o sea el numero de grupos menos uno 2-1 en este caso, para el SCd se puede obtener por la resta del numero de gente menos el número de grupos. Así tendrémos esta tabla.

anava	S.C	g.l.	M.C.	F.
Entre grupos	66,76	1	66,76	2,33
Dentro de los grupos	687,86	24	28,66	
Total	754,62	25		

Para interpretar este valor debemos encontrar en la tabla de la f para 99 de nivel de confianza, si se rechaza la hipótesis nula o se acepta.
Para saber el balor teórico debería coger la tabla del percentil 99 y alli en la primera fila buscar el primer grado de libertad el de entre grupos y en la primera columna el de dentro del grupo

Percentil 99

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10
1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022	6056
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
6	13.75	10.92	9.78	9.15	8.73	8.47	8.26	8.10	7.98	7.87
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
15	[illegible]	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16	[illegible]	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	[illegible]	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	[illegible]	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	[illegible]	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17

El valor teórico es de 7,82 y el valor empírico es de 2,33, por lo tanto 2,33<7,82 eso significa que se cumple la hipótesis nula