PROYECTO DEL CURSO DE ESTADÍSTICA INFERENCIAL

Prof.: MSc. Julio R. Vargas A.

I. INTRODUCCION

El presente trabajo está orientado a aplicar los conocimientos de estadística inferencial a un caso práctico o del mundo profesional, con el propósito de mostrar su aplicabilidad e importancia en el desarrollo de las actividades del mundo empresarial.

Lo primero que se hizo fue definir el planteamiento del problema, el cual debe ser bien claro, preciso y delimitado, luego se procedió calcular el tamaño de la muestra teniendo claro el error y el nivel de significancia. Posteriormente se obtuvo la muestra o los elementos muéstrales. Con los datos muéstrales se procedió a su organización, clasificación y procesamiento. Se aplicó el análisis de varianza como estadístico de prueba para contrastar la hipótesis nula.
Las conclusiones se presentan al final del documento.

II. PLANTEAMIENTO DEL PROBLEMA.

Se aplicada una prueba a diferentes Personas con diferentes grados de escolaridad, se está interesado en probar si existe o no diferencia entre el grado de escolaridad (variable nominal) y el promedio de la calificación obtenida por cada persona (variable numérica) en la prueba.

El análisis de varianza (ANOVA) es una prueba que nos permite medir la variación de las respuestas numéricas como valores de evaluación de diferentes variables nominales.

Tabla del Modelo Anova

Fuentes de variación	Grados de libertad	Suma de cuadrados	Media cuadrado	F
Tratamiento (intergrupo)	c-1	SCC (suma de cuadrados de los tratamientos)	CMC(cuadros medios de los tratamientos)	F= CMC/CME
Error (intragrupo)	n-c	SCE(suma de cuadrados del error)	CME(cuadrados medios del error)	
Total	n-1	SCT(suma de cuadrados totales		

Para lo cual se requerirán las calificación de la prueba a cada persona y sus diferentes grados de escolaridad, con lo que se intentará probar si existe o no diferencia entre el grado escolar (variable nominal) y el promedio de la calificación (variable numérica).

Para analizar si existe diferencia en los promedios se procede a realizar una prueba **F** que se explica posteriormente.

III. CALCULO DEL TAMAÑO DE LA MUESTRA.

La muestra estará formada el grupo de individuos que realmente se estudiarán, es un subconjunto de la población. Para que se puedan generalizar a la población los resultados obtenidos en la muestra, ésta ha de ser «representativa» de dicha población. Para ello, se han de definir con claridad los criterios de inclusión y exclusión y, sobre todo, se han de utilizar las técnicas de muestreo apropiadas para garantizar dicha representatividad.

¿Cuántas personas será necesario estudiar para estimar la media de las calificaciones que han las personas de acuerdo al grupo de escolaridad?. Para lo cual establecemos un nivel de confianza del 95% y se admite un error de 2 puntos como máximo; faltaría por conocer la desviación estándar de la población, la cual no conocemos. Por lo que la podemos obtener a partir de la prueba piloto y que es de 20. Por lo que el número mínimo de personas que ha de estudiarse será de:

$$n = \frac{Z_{\alpha/2}{}^2 \sigma^2}{e^2} = 64 = \frac{1.96^2 * 66.796}{e^2}$$

$$e^2 = \frac{256.6035}{64} = 4.0094$$

$$e = \sqrt{4.00943} = 2.0024$$

Como hemos visto en el cálculo, estaremos admitiendo que con una muestra de n=64 individuos con escolaridades de Doctorados, Maestrías y Estudiantes Universitarios. El error será menor a 2.0024 puntos. (Aproximadamente 2 puntos).

n = tamaño de la muestra

σ^2 = varianza de la población en estudio (en este caso no la conocemos)

Z= El valor normalizado del nivel de confianza fijado que es 0.05

Hemos hecho una prueba piloto de treinta individuos para obtener la varianza muestral la cual es suficiente por ser muestra grande y aleatoria para estimar al parámetro desconocido σ^2.

Con la prueba piloto de 30 calificaciones obtuvimos la varianza muestral que usamos en la ecuación anterior.

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. típ.	Varianza
Calificación obtenida en la prueba	30	32.18399860	67.15204678	49.42302357	8.172899957	66.796
N válido (según lista)	30					

IV. Obtención de la muestra aleatoria del tamaño que sea definido para poder hacer inferencias válidas sobre la población en estudio.

Las variables de interés para este estudio son:

- La calificación de las personas (CALIFICACIÓN)
- La escolaridad (GRADO ESCOLAR)

Por lo que para efectos del estudio nos limitaremos a ellos a continuación se muestran los datos obtenidos en forma aleatoria.

TABLA 1:

CALIFICACIÓN	GRADO ESCOLAR
67.15204678	DOCTORADO
64.36842105	DOCTORADO
60.91130604	ESTUDIANTE
55.38986355	ESTUDIANTE
53.917154	ESTUDIANTE
53.3460039	MAESTRÍA
52.15984405	ESTUDIANTE
51.86842105	ESTUDIANTE
51.12768031	DOCTORADO
50.63060429	ESTUDIANTE
50.35477583	MAESTRÍA
48.38596491	MAESTRÍA
47.07407407	DOCTORADO
44.09454191	MAESTRÍA
43.41520468	ESTUDIANTE

CALIFICACIÓN	GRADO ESCOLAR
39.5662768	ESTUDIANTE
39.07309942	ESTUDIANTE
38.71247563	DOCTORADO
34.95321637	ESTUDIANTE
34.27777778	ESTUDIANTE
34.27192982	ESTUDIANTE
67.63611386	DOCTORADO
62.77020467	DOCTORADO
60.88483775	ESTUDIANTE
56.50144025	ESTUDIANTE
51.76861802	ESTUDIANTE
53.63085832	MAESTRÍA
50.77179452	ESTUDIANTE
50.89056506	ESTUDIANTE
48.66061841	DOCTORADO

43.23781676	MAESTRÍA
41.82066277	ESTUDIANTE
41.57212476	ESTUDIANTE
41.21539961	MAESTRÍA
40.8245614	ESTUDIANTE
40.79824561	ESTUDIANTE
33.09835159	ESTUDIANTE
32.1839986	ESTUDIANTE
58.49961104	ESTUDIANTE
56.18983249	ESTUDIANTE
51.46872891	ESTUDIANTE
53.4198814	MAESTRÍA
53.7674174	ESTUDIANTE
50.90286877	ESTUDIANTE
49.49529961	DOCTORADO
50.07639845	ESTUDIANTE
48.55589372	MAESTRÍA

52.67230843	ESTUDIANTE
47.98778555	MAESTRÍA
48.23106247	MAESTRÍA
46.83381069	DOCTORADO
45.52452004	MAESTRÍA
43.28708589	ESTUDIANTE
41.03983895	MAESTRÍA
41.53716416	ESTUDIANTE
43.38891669	ESTUDIANTE
39.98564149	MAESTRÍA
39.42669945	ESTUDIANTE
38.45267793	ESTUDIANTE
39.80270585	ESTUDIANTE
37.09940719	ESTUDIANTE
41.13772888	DOCTORADO
34.4219837	ESTUDIANTE
40.28758583	ESTUDIANTE

V. FORMULACION DE LA HIPOTESIS.

Queremos confirmar que el grado académico no tiene ningún efecto en la calificación de la prueba, queremos por lo tanto comparar las medias μ_1 con μ_2 y μ_3 considerando que μ_1 representa la media de las calificaciones de las personas con grado de Doctor y μ_2 representa la media de las calificaciones de las personas con grados de Master y μ_3 Es la media de las personas que son Estudiantes es decir no tiene grado académico. Este contrate lo queremos hacer contra la Hipótesis alternativa de que las medias son diferentes, para lo cual elegimos un nivel de significación de α=0.05 (es del 5%) Para lo cual formulamos formalmente la hipótesis.

$$H_o: \mu_1 = \mu_2 = \mu_3$$

H_a: μ_1 , μ_2 , μ_3 son diferentes

VI. ANALISIS DE LOS DATOS Y USO DEL ESTADÍSTICO DE PRUEBA PARA CONTRASTAR LA HIPOTESIS.

ANALISIS DE VARIANZA (ANOVA)

El primer paso es ordenar los datos de acuerdo al valor nominal que le corresponde para así obtener:
El número de datos, el promedio y la desviación estándar de cada uno de los valores nominales.
De la TABLA 1 obtenemos los tres valores nominales que toma la variable GRADO ESCOLAR, estos tres valores son: 1 DOCTORADO, 2 ESTUDIANTE, 3 MAESTRÍA.

La siguiente tabla nos muestra estos resultados.

TABLA 2

		GRADO ESCOLAR		
CALIFICACIÓN	TOTALES	DOCTORADO	ESTUDIANTE	MAESTRÍA
Cuadrados medios	7133.522799	2828.000955	2089.735312	2215.78653
Desviación estándar	8.477715089	10.44722904	8.383858417	4.69392914
media	47.29390233	53.17895218	45.71362283	47.0721418
n	64	11	39	14
SUMA DE CUADRADOS	147677.5588	32199.45645	84170.66229	31307.4401

1	2	3
67.15204678	60.91130604	53.3460039
64.36842105	55.38986355	50.3547758
51.12768031	53.917154	48.3859649
47.07407407	52.15984405	44.0945419
49.49529961	51.86842105	43.2378168
38.71247563	50.63060429	41.2153996
67.63611386	43.41520468	53.4198814
62.77020467	41.82066277	48.5558937
48.66061841	41.57212476	53.6308583
46.83381069	40.8245614	47.9877855
41.13772888	40.79824561	48.2310625
	33.09835159	45.52452
	32.1839986	41.0398389
	58.49961104	39.9856415
	56.18983249	
	51.46872891	
	53.7674174	
	50.90286877	
	50.07639845	
	39.5662768	
	39.07309942	
	34.95321637	
	34.27777778	
	34.27192982	
	60.88483775	
	56.50144025	
	51.76861802	
	50.77179452	
	50.89056506	
	52.67230843	
	43.28708589	
	41.53716416	
	43.38891669	
	39.42669945	
	38.45267793	
	39.80270585	
	37.09940719	
	34.4219837	
	40.28758583	

Y1(Doctorado)	y^2	
67.15204678	4509.397387	COMO ESTA TABLA
64.36842105	4143.293628	TAMBIÉN SE HACEN
51.12768031	2614.039694	LOS CALCULOS
47.07407407	2215.96845	PARA LOS DE OTROS
49.49529961	2449.784683	
38.71247563	1498.655769	
67.63611386	4574.643898	
62.77020467	3940.098594	
48.66061841	2367.855784	
46.83381069	2193.405824	
41.13772888	1692.312737	

$\bar{y}$	Suma de C.	Cuadrados Medios	Desv. Est.
53.17895218	32199.45645	2828.000955	10.44722904

$$S = \frac{n\sum y_i^2 - (\sum y_i)^2}{n(n-1)}$$

formula de la desviación estándar

Y2(MASTER)	y^2
53.3460039	2845.79613
50.3547758	2535.60345
48.3859649	2341.2016
44.0945419	1944.32863
43.2378168	1869.5088
41.2153996	1698.70917
53.4198814	2853.68373
48.5558937	2357.67481
47.9877856	2302.82756
48.2310625	2326.23539
45.52452	2072.48192
53.6308583	2876.26896
41.039839	1684.26838
39.9856415	1598.85153
659.009985	**31307.4401**
$\bar{y}$ =	**47.0721418**

NOTACIÓN

$c =$ número de valores nominales

$n =$ total de datos

$n_j =$ total de datos de la j - esima columna

$y =$ promedio total

$y_j =$ promedio de la j.esima columna

$y_{ij} =$ dato número i de la columna j

$CM =$ Corrección de la media

$SCC =$ Suma del cuadrado de los tratamientos

$SCT =$ Suma de los cuadrados totales

$SCE =$ Suma de los cuadrados del error

$gl1 =$ grados de libertad uno

$gl2 =$ grados de libertad dos

$CMC =$ Cuadrado medio de los tratamientos

$CME =$ Cuadrado medio del error

$F =$ Valor para la prueba F

Obtenemos:

$$CM = ny^2$$

$$SCC = \sum_{j=1}^{c} n_j y_j{}^2 - CM$$

$$SCT = \sum_{j=1}^{c} \sum_{i=1}^{n_j} y_{ij}^2 - CM$$

$$SCE = SCT - SCC$$

$$gl1 = c - 1$$

$$gl2 = n - c$$

$$CMC = \frac{SCC}{gl1}$$

$$CME = \frac{SCE}{gl2}$$

$$F = \frac{CMC}{CME}$$

Para nuestro ejemplo:

c = 3 número de columnas (número de valores nominales DOCTORADO, ESTUDIANTE Y MAESTRÍA)
n = 64 total de datos o muestra.
n1 =11 DOCTORADO
n2 = 39 ESTUDIANTE
n3 =14 MAESTRÍA

$y = 47.29390233$ promedio total
$y_1 = 53.17895218$ promedio DOCTORADO
$y_2 = 45.71362283$ promedio ESTUDIANTE
$y_3 = 47.0721418$ promedio MAESTRÍA
$y_1^2 = 2828.000955$ cuadrado del promedio DOCTORADO
$y_2^2 = 2089.735312$ cuadrado del promedio ESTUDIANTE
$y_3^2 = 2215.78653$ cuadrado del promedio MAESTRÍA

$CM = ny^2$

CM = 64 (47.29390233)^2 =64(2236.713198) = **143149.6446**

$$SCC = \sum_{j=1}^{c} n_j y_j{}^2 - CM = (n_1 y_1{}^2 + n_2 y_2{}^2 + n_3 y_3{}^2) - CM$$

SCC = (11(2828.000955)+39(2089.735312)+14(2215.78653))-143149.6446

SCC = (31108.01051+81499.67717+31021.01142)))-143149.6446
SCC = 143628.6991-143149.6446

$$SCT = \sum_{j=1}^{c}\sum_{i=1}^{n_j} y_{ij}^2 - CM$$

$$SCT = 147677.5588 - 143149.6447$$

$$SCT = 4527.914147$$

SCC= **479.0544662**

SCE = SCT – SCC =4527.914147 – 479.0544662 = 4048.859681
g.l. 1 = c-1 = 3-1 = 2
g.l. 2 = n-c = 64 – 3 = 61

$$CMC = \frac{SCC}{g.l.1} = \frac{479.0544662}{2} = 239.5272331$$

$$CME = \frac{SCE}{g.l.2} = \frac{4048.859681}{61} = 66.37474886$$

$$F = \frac{CMC}{CME} = \frac{239.5272331}{66.37474886} = 3.608710198$$

ANOVA DE LOS RESULTADOS

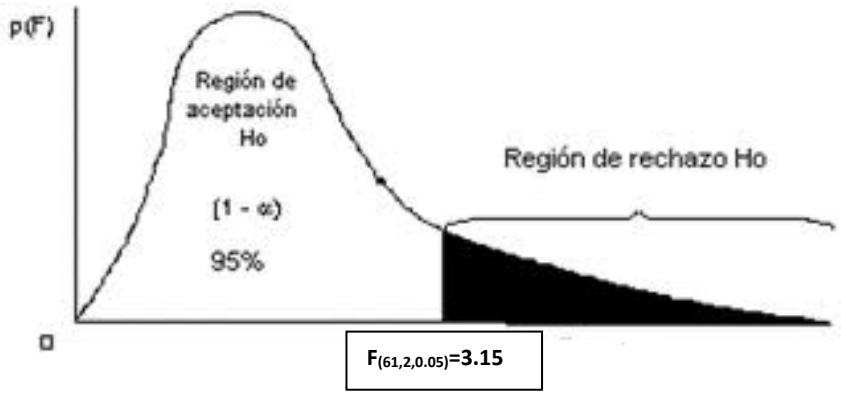
Fuentes de variación	Grados de libertad	Suma de cuadrados	Media cuadrado	F
Tratamiento (intergrupo)	2	SCC=479.0543	CMC=SCC/c-1 CMC=239.5272	F= CMC/CME F=3.6087
Error (intragrupo)	61	SCE=4048.8600	CME=SCE/n-c CME=66.3748	
Total	63	SCT=4529.9140		

Para ello necesitamos F, gl.1 y gl. 2

Estos valores los buscamos es la Tabla F con v1=2 y v2=61 y α=0.05

$F_{(61,2),0.05}$ =3. 15

Por lo tanto F > $F_{(61,2),0.05}$ Por lo tanto RECHAZAMOS Ho. (3.60 > 3.15)



VII. CONCLUSIONES

Hemos utilizado el l método de análisis de varianza para comparación de promedios, que parte del supuesto inicial de que no existe diferencia entre los promedios y que los resultados de la muestra son producto exclusivamente del azar.

A este supuesto inicial es lo que conocemos como la hipótesis nula y se le designa con $\mathbf{H_0}$.

El procedimiento seguido con el Análisis de Varianza (ANOVA) nos conduce a establecer si entre las medias de los grupos hay o no diferencias significativas.

Que como hemos visto en los cálculos finales de la prueba F hay evidencia que las medias parecen ser diferentes. No sabemos entre quienes hay diferencias es decir si entre las tres o solo en dos de ellas. Pero la prueba nos ha indicado que el supuesto que las medias eran iguales no se confirma, es decir parece que el grado académico tiene efecto en las calificaciones de las pruebas.

Lo que implica que la hipótesis alternativa H_a se acepta, estos es existe al menos una pareja de valores nominales cuyos promedios son diferentes.