ESTIMACION DEL TAMAÑO DE LA MUESTRA PARA EL ESTUDIO DE VARIEDADES "CONJOINT ANALYSIS"

Jairo Castaño

Enero de 1989

El objetivo principal del estudio de preferencia de 15 variedades es detectar y cuantificar los criterios de esas preferencias de los agricultores. Mediante encuesta se decidió alcanzar este objetivo. Como primera medida una encuesta debe tener definida la población y el marco muestral en donde se va a desarrollar.

La población de interés está conformada por las fincas tradicionalmente frijoleras, con disponibilidad de lote para siembra en el segundo semestre de 1988 y que estan ubicadas en los municipios de El Tambo, Darien-Restrepo y Caldono. El marco muestral tiene como unidades de muestreo al agricultor dueño de cada finca y a sus vecinos.

Identificado nuestro objeto de estudio debemos determinar la variable más importante en la información de las encuestas, para que a partir de ello hallemos el tamaño de la muestra. En este caso esa variable es: **Preferencia**, o sea, la calificación que le asigna el agricultor a cada una de las 15 variedades; la medición más útil que obtendremos de esta variable es su Media, de tal forma que cada variedad tenderá a situarse en una calificación promedia de $\bar{Y}i$ puntos dentro de una escala de 1 a 15.

Ahora bién, los diferentes métodos de cálculo de un tamaño de muestra requieren de ciertos elementos como la **Varianza** estimada en la población (S^2), la **Confianza** ($1-\alpha$) asociada a la precisión de la estimación y el **Error** (E) inherente al muestreo.

El nivel de confianza y el error involucrado son asignados por el investigador con base en consideraciones de precisión deseada, de disponibilidad de recursos económicos, humanos, y de tiempo, de tal forma que pueden ser elementos rígidos o flexibles según el caso. La

varianza requiere de un conocimiento previo de la heterogeneidad de la población. En el caso del conjoint sólo se disponía de un estudio realizado en Pescador con 10 agricultores y 8 variedades, del que se podía obtener una variación aproximada de preferencia por cada variedad.

Métodos para el Cálculo de "n"

Se citarán 4 métodos más apropiados para cálculo de n, con una explicación breve de cada uno:

1. Estimación de n para la diferencia de Medias: Y1-Y2.
2. Estimación de n con un error relativo E = XY.
3. Estimación de n con un error absoluto d.
4. Estimación de n mediante el teorema Chebyshev.

Para los 3 primeros métodos se parte del supuesto que la variable Y (preferencia) tiene una distribución normal o aproximadamente normal (Teorema central del límite). Mediante el procedimiento UNIVARIATE de SAS se chequeó esta normalidad obteniendose en resumen el siguiente cuadro:

Cuadro 1. Prueba de normalidad sobre la preferencia en 8 variedades (Pescador)

No.	Variedad	W	Prob.< |W|	
1	ASR205	0.806	0.02	
2	A486	0.928	0.44	***
3	A36	0.798	0.017	
4	A66	0.85	0.195	***
5	PVA1261	0.85	0.66	
6	BAT1297	0.85	0.076	
7	AT40	0.89	0.268	***
8	CALIMA	0.96	0.876	***

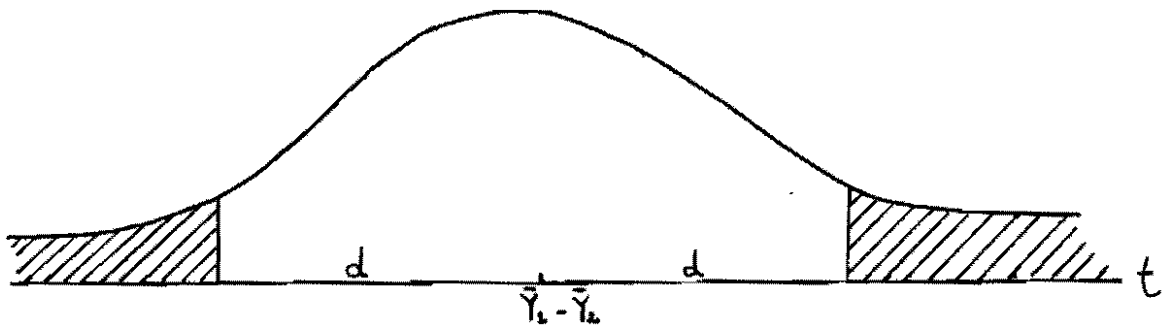
***: Acepta la hipótesis nula de que la preferencia por la variedad es distribuida normalmente.

W es la prueba de normalidad Shaphiro Wilk (para n < 51), que examina la hipótesis nula Ho: La variedad es normal. Si Prob < |W| es mayor que 0.05 se rechaza la Ho. En este caso se ha detectado que 4 de las 8 variedades tienen su preferencia distribuida aproximadamente Normal, a pesar de contar con sólo 10 observaciones de cada una.

Consecuentemente, podemos pasar entonces a calcular n en cada uno de los métodos, teniendo en cuenta que utilizaremos el estadígrafo t (usado en el cálculo de muestras pequeñas) en lugar de el estadígrafo z.

Método 1: <u>n para Diferencias de Medias.</u>

Para el caso de la comparación de dos medias con el fin de detectar diferencias significativas entre ellas, se cuenta con:



Donde $d = t_{n1+n2-2,\ \alpha/2} \sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}$ (t para el caso de muestras pequeñas)

dado que n1 = n2 = n, entonces $n = \frac{(t_{2n-2,\ \alpha/2})^2 \times (S_1^2 + S_2^2)}{d^2}$

n aumentará a medida que crezcan las varianzas y el nivel de confianza deseado.

Ahora, deseamos una mínima diferencia entre 2 variedades con lo cual la diferencia **d** es significativa con cierta probabilidad **t**, esto significa que una diferencia mayor a un valor **d** es inmediatamente detectada como significativa, o sea, determinará que una variedad tiene mayor preferencia que la otra; de los datos del cuadro 2.

Cuadro 2: Media y Varianza de preferencia en 8 Variedades (Pescador)

Variedad	X Med	X Var.
1	3.30	0.78889
2	3.25	3.68056
3	5.80	5.95556
4	5.60	4.98889
5	4.70	7.28889
6	4.10	6.76667
7	5.20	3.51111
8	4.05	4.85833

Para el cálculo de **n** escogeremos las variedades 5 y 6, que tienen las varianzas más altas (así obtendremos un **n** más grande) y variaremos **d** y **t** (cuadro 3):

Cuadro 3: Cálculo de n para Diferencias de Medias de Preferencia

t / d	t18, 0.025 = 2.1 95%	t18, 0.05 = 1.73 90%	t18, 0.075 = 1.53* 85%
1	62	42	33
1.2	43	30	23
1.4	30	22	17

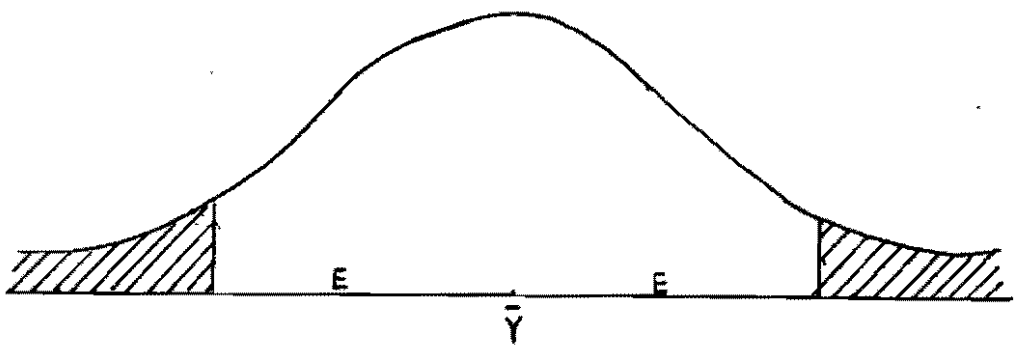
* valor calculado por interpolación.

Para explicar mejor el cuadro 3, supongamos que la variedad 3 tuviera una preferencia promedio de 7 puntos y la variedad 7, 8.1 puntos, entonces con una muestra de 62 agricultores estas dos variedades serían diferentes al 95% de confiabilidad pues la diferencia de sus promedios es mayor que 1. Caso contrario sucedería si esa diferencia fuera 1 o menos.

Con este criterio, el tamaño de 30 para la muestra parece el más conveniente contando con un 90% de confianza y una diferencia máxima permitida de 1.2

Método 2: <u>n con un Error (E) Relativo</u>

Algunas veces, podríamos desear por ejemplo, estimar $\bar{Y}$ con un error no mayor al 10%, o sea, queremos definir un valor E = porcentaje del estimador, en este caso el estimador es $\bar{Y}$, que variará para cada media.



$$\text{Asi, } n = \frac{s^2 t^2}{(\% Y)^2} = \frac{s^2 (t_{9,\alpha/2})^2}{E^2}$$

Y con los datos del cuadro 2 obtendremos:

Cuadro 4: Cálculo de n para un error relativo

	85%				90%				95%			
Variedad	10% X	15% X	20% X	30% X	10% X	15% X	20% X	30% X	10% X	15% X	20% X	30% X
1	18.54	8.24	4.63	2.06	24.26	10.78	6.06	2.69	37.00	16.44	9.25	4.11
2	89.20	39.64	22.30	9.91	116.69	51.86	29.17	12,96	177.97	79.10	44.49	19.77
3	45.32	20.14	11.33	5.03	59.28	26.35	14.82	6.58	90.42	40.18	22.60	10.04
4	40.72	18.10	10.18	4.52	53.27	23.67	13.31	5.91	81.25	36.11	20.31	9.02
5	84.47	37.54	21.11	9.38	110.50	49.11	27.62	12.27	168.53	74.90	42.13	18.72
6	103.05	45.80	25.76	11.45	134.80	59.91	33.70	14.97	205.60	91.37	51.40	22.84
7	33.24	14.77	8.31	3.69	43.48	19.32	10.87	4.83	66.32	29.47	16.58	7.36
8	75.82	33.70	18.95	8.42	99.19	44.08	24.79	11.02	151.28	67.23	37.82	16.80

Como vemos la variedad 6 es la que presenta los n más altos, así que nos guiaremos por ella. De acuerdo a nuestras expectativas, los tamaños 25.76 (85% de confianza), 33.7 (90%) y 22.8 (95%) son los más apropiados. En el caso de querer redondear ese tamaño a un entero, podríamos optar por un error porcentual del 21.2% sobre la media (0.87) y obtendríamos un tamaño n = 30 con una confiabilidad del 90%. Esto significa que con 30 agricultores podríamos estimar en la variedad 6, por ejemplo, una calificación promedia de 4.1 puntos con un error de magnitud maximo de 0.87 en 27 de los 30 agricultores, y de más de 0.87 en el resto de ellos.

Método 3: <u>Estimación de n con un error absoluto d</u>

Si en el caso anterior, en lugar de controlar el error relativo queremos controlar el error absoluto d en $\bar{Y}$, constante para cualquier $\bar{Y}$, obtendremos:

$$n = \frac{t^2_{9,\alpha} S^2}{d^2}$$

Cuadro 5: Cálculo de n para un error absoluto

	85%				90%				95%			
Variedad	0.5	1.0	1.2	1.4	0.5	1.0	1.2	1.4	0.5	1.0	1.2	1.4
1	8.07	2.01	1.40	1.34	10.56	2.64	1.83	1.34	16.11	4.02	2.79	2.05
2	37.68	9.42	6.54	6.28	49.30	12.32	8.55	6.28	75.19	18.79	13.05	9.59
3	60.98	15.24	10.58	10.17	79.77	19.94	13.85	10.17	121.67	30.41	21.12	15.51
4	51.08	12.77	8.86	8.52	66.82	16.70	11.60	8.52	101.92	25.48	17.69	13.00
5	74.63	18.65	12.95	12.45	97.63	24.40	16.95	12.45	148.91	37.22	25.85	18.99
6	69.29	17.32	12.02	11.56	90.64	22.66	15.73	11.56	138.24	34.56	24.00	17.63
7	35.95	8.98	6.24	5.99	47.03	11.75	8.16	5.99	71.73	17.93	12.45	9.14
8	49.74	12.43	8.63	8.30	65.08	16.27	11.29	8.30	99.25	24.81	17.23	12.66

El cuadro 5 presenta los diferentes tamaños obtenidos y podemos de allí extraer los más idóneos para nuestro caso, y esos son: **n** = 24.4 con **d** = 1 y 90% de confianza y **n** = 25.8 con **d** = 1.2 y 95% confianza para la variedad 5. Si quisieramos redondear n a 30 con los nismos níveles de confianza en ambos casos, tendríamos para el primero un error **d** = 0.9, y para el segundo un error **d** = 1.1; lo que interpretaremos en el último caso, que de 30 agricultores, 28 darían un puntaje promedio de 4.7 puntos con un error no mayor a 1.1 puntos, mientras que los otros dos tendrían un error mayor.

Método 4: <u>Teorema de Chebyshev</u>

El Teorema de Chebyshev se utiliza cuando se desconoce todo acerca de la población de estudio, es decir, se ignora a que función de distribución pertenece, y se define:

$$n = \frac{S^2}{d^2 \times \alpha}$$

donde α es el nivel de confiabilidad y **d** la diferencia máxima permitida de la estimación de Y.

Cuadro 6: Cálculo de n por el teorema de Chebyshev

	95%				90%				85%			
Variedad	0.5	1.0	1.2	1.4	0.5	1.0	1.2	1.4	0.5	1.0	1.2	1.4
1	63.11	15.77	10.95	8.04	31.55	7.88	5.47	4.02	15.77	3.94	2.73	2.01
2	294.44	73.61	51.11	37.55	147.22	36.80	25.55	18.77	73.61	18.40	12.77	9.38
3	476.44	119.11	82.71	60.77	238.22	59.55	41.35	30.38	119.11	29.77	20.67	15.19
4	399.11	99.77	69.29	50.90	199.55	49.88	34.64	25.45	99.77	24.94	17.32	12.72
5	583.11	145.77	101.23	74.37	291.55	72.88	50.61	37.18	145.77	36.44	25.30	18.59
6	541.33	135.33	93.98	69.04	270.66	67.66	46.99	34.52	135.33	33.83	23.49	17.26
7	280.88	70.22	48.76	35.82	140.44	35.11	24.38	17.91	70.22	17.55	12.19	8.95
8	388.66	97.16	67.47	49.57	194.33	48.58	33.73	24.78	97.16	24.29	16.86	12.39

Como se observa, este método es el que arroja los n más elevados y esto por la sencilla razón de que al no conocerse la distribución de las muestras, no se puede utilizar ningún estadígrafo como t o z que optimice el cálculo.

Sin embargo, si nuestro caso fuera el de desconocer la distribución de las preferencias, optaríamos posiblemente por la muestra de 30 agricultores con un error máximo de 1.1 puntos sobre la preferencia media (observar variedad 5) y con una confianza del 85%.

CONCLUSION:

Resumiendo, y como lo podemos apreciar en el cuadro 7, los diferentes métodos arrojan suficiente luz como para recomendar un tamaño de muestra de 30 agricultores, basados en un estupendo nivel de confianza y un error de muestreo tolerable.

Cuadro 7: Tamaños óptimos de muestras

	Diferencia de Medias			Error Relativo			Error Absoluto				Chebyshev			
Método	d			E			d				d			
	1	1.2	1.4	20%	21%	30%	0.9	1	1.1	1.2	1	1.1	1.2	1.4
Nivel de Confianza														
85%	33			25.7							36.4	30	25.3	
90%		30		33.7	30		30	24.4						37
95%			30			22.8			30	25.8				

Bibliografía

Cochran, G. William. 1977. Técnicas de Muestreo. South Orleans, Massachusetts.

Servín, de Abad Adela, Servín Anorade Luis A. 1978. Introducción al Muestreo. México. Editorial Limusa.

Teoría de Pequeñas Muestras. Artículo sobre Muestreo.

S.A.S. User's Guide: Statistics Version 5 Edition. SAS Institute Inc. 1985. pp 956.

Pérez E., Jaime E. 1983. Inferencia Estadística. Cali, Colombia. Universidad del Valle, Departamento de Matemáticas.