PONENTE: GENARO SÁNCHEZ BARAJAS
PROFESOR TITULAR NIVEL C DE LA FACULTAD DE ECONOMIA DE LA UNAM Y DOCTORANTE EN LA ACADEMIA DE CIENCIAS DE RUSIA.
TEMA PARA LA MESA 3 DE ECONOMETRIA

"EVALUACIÓN ESTADÍSTICA DE ENCUESTAS MENSUALES O PERIÓDICAS": ANÁLISIS DE SESGO Y COBERTUTA ESTADISTICA

RESUMEN:

El levantamiento mensual de encuestas requiere de una supervisión estadística que permanentemente favorezca la confiabilidad de la información. Para ello es necesaria la aplicación de ciertas técnicas que detecten si existe o no relación entre el tamaño de la muestra y el valor de los indicadores.

Para ilustrar lo anterior se tomó como referencia una encuesta mensual que hace el Instituto de la Pequeña y Mediana Empresa.

La periodicidad de la encuesta requiere la aplicación de técnicas fuertes que permitan eliminar rápidamente los factores irrelevantes y retener los de gran significación en los resultados, para ello se hace el análisis de sesgo y cobertura.

A las medidas estadísticas que permitan cumplir con estos objetivos se les denominará

COEFICIENTES DE ASOCIACIÓN.

Puesto que el método de muestreo utilizado es el de proporciones correspondientes a indicadores con distribuciones fuera de cualquier curva definida por funciones matemáticas, se optó por la aplicación de pruebas de asociación no paramétrica de las variables en la pequeña y mediana empresa en conjunto.

2.- Prueba de asociación.

La escasez de recursos humanos y el limitado acceso a la computadora, en esta etapa determinaron el manejo de solo tres medidas de asociación; en la medida que se resuelvan estos problemas y que el personal se familiarice con el análisis estadístico, se aplicarán diseños muestrales y coeficientes de asociación más sofisticados.

Por otra parte, mientras el análisis estadístico no se instrumente en la computadora, mensualmente se evaluará una de las siguientes variables. Para iniciar los trabajos de julio, de la encuesta de junio se analizó:

-Personal ocupado promedio respecto al mes anterior
-Inventario de productos finales
-Fuentes de financiamiento para resolver problemas de liquidez, total industria pequeña y mediana.

<u>PERSONAL OCUPADO PROMEDIO</u>

Se recurre a la χ^2 : Ji Cuadrada basada en las tablas de contingencia para probar la hipótesis de independencia entre el tamaño de la muestra y la opinión de los empresarios. Para ello se comparan las respuestas "reales" de la muestra con las respuestas "esperadas".

Personal Ocupado

Muestra	Respuesta Real			Total
	Aumento	No Aumentó	Disminuyó	
Alimentos . . . . Otros				
Total				419
%				100

Cuando se acepta la hipótesis no es necesario modificar el tamaño de la muestra. En cambio si se rechaza la hipótesis, se identifica que si hay una relación entre el tamaño de la muestra y la opinión de los empresarios; por lo que es necesario hacer un análisis como el que se describe a continuación:

Así por ejemplo: partiendo del rechazo de la hipótesis nula basado en la χ^2, se utilizará la estadística ϕ para cuantificar la relación entre la muestra y las opiniones ; ya que si es baja quizá no valga la pena hacer las revisiones correspondientes; en cambio si es alta de inmediato se hace un análisis de sesgo y cobertura.

Phi (ϕ)

Es una medida de la fuerza de la relación que existe entre las variables descriptivas, la cuantitativa (muestra) y la cualitativa (opinión de los empresarios). Phi toma el valor de 0 cuando no hay relación y + 1 cuando las variables se relacionan a la perfección. Phi hace la corrección en el valor de χ^2 porque éste es directamente proporcional al tamaño de la muestra (n) y por ello su fórmula es :

$$\phi = \left[\frac{\chi^2}{n}\right]^{1/2}$$

V de Cramer

Cuando ϕ se obtiene de tablas de contingencia más grande a la de 2 x 2, como es el caso concreto de la encuesta, su valor no tiene límite superior, por lo que se usa V de Cramer para ajustar ϕ en términos de las columnas o de las hileras, dependiendo cual de ellas es más pequeña.

El valor de la estadística V también oscila entre 0 y +1. Así, un valor alto de V significa que hay un alto grado de asociación.
Su fórmula es:

$$V = \left(\frac{\phi^2}{min(r-1, c-1)} \right)^{1/2}$$

En resumen, si una vez aplicadas las estadísticas χ^2, ϕ y V, se encuentra que el valor de V es alto, entonces se toma la decisión de hacer el análisis de sesgo y cobertura, para lo cual se analiza la información a fin de validarla y determinar si los resultados pueden atribuirse a relaciones o asociaciones legítimas o a la selección aleatoria de la muestra.

Si es ésta última habrá que hacer lo siguiente:

1) recalcular el tamaño de la muestra (cobertura) en los grupos industriales afectados y,
2) mantener el porciento dentro de ciertos limites de control (sesgo).

PROCEDIMIENTO

A continuación se expone un ejemplo completo con datos del mes de junio, empezando por la χ^2, ϕ y V , hasta el análisis de cobertura para el caso extremo en que tuviéramos que recalcular toda la muestra, aplicando el **muestreo simple aleatorio**; así como para el cálculo especifico para algunos grupos industriales, **usando el muestreo estratificado proporcional**.

TABLA DE CONTINGENCIA						
GRUPO INDUSTRIAL	A_1	A_2	A_3	TOTAL		
B_1	R_{11}	S_{21}	T_{31}	V_1	=	$R_{11} + S_{21} + T_{13}$
B_2	R_{12}	S_{22}	T_{32}	V_2	=	$R_{12} + S_{22} + T_{32}$
B_3	R_{13}	S_{23}	T_{33}	V_3	=	
B_4	R_{14}			V_4	=	
B_5	R_{15}			V_5	=	
B_6	R_{16}			V_6	=	
B_7	R_{17}			V_7	=	
B_8	R_{18}	S_{28}		V_8	=	
B_9	R_{19}			V_9	=	$R_{19} + S_{29} + T_{39}$
B_{10}	R_{110}		T_{310}	V_{10}	=	
B_{11}	R_{111}	S_{211}		V_{11}	=	
B_{12}	R_{112}			V_{12}	=	
B_{13}	R_{113}			V_{13}	=	
B_{14}	R_{114}		T_{314}	V_{14}	=	
B_{15}	R_{115}	S_{215}	T_{315}	V_{15}	=	
B_{16}	R_{116}	S_{216}	T_{316}	V_{16}	=	$R_{116} + S_{216} + T_{316}$

TOTAL	R	S	T	V = R + S +T

Construyendo la tabla de contingencia con los resultados observados para el Personal Ocupado en junio, se obtiene la tabla 3X16 que aparece a continuación para las dos variables descriptivas Bi (cuantitativa y A : Cualitativa: opinión de los empresarios).

Donde:

B1: grupo industrial

A_1 : Aumentó

A_2 : No aumentó

A_3 : Disminuyó

$R = \Sigma\ R_i$

$S = \Sigma\ S_i$

$T = \Sigma\ T_i$

$V = \Sigma\ V_i = R + S + T$

$V_i = \Sigma\ (R_i + S_i + T_i)$

i = 1, 2, 3, ..., 16

Personal Ocupado Promedio				
	Respuesta Real			
Muestra	Aumento	No vario	Disminuyo	Total
Fab. de alimentos	10	61	13	84
Industria Textil	3	22	3	28
Fab. de Prendas de Vestir	4	27	9	40
Fab. de Calzado e Ind. del Cuero	5	25	7	37
Ind. y Prod. de Madera y Corcho				0
Excepto Muebles	1	9	5	15
Fab. y Rep. de Muebles de Madera	1	11	9	21
Ind. Editorial de Impresión y Conexas	6	13	1	20
Industria Química	3	11	2	16
Fab. de Prod. de Hule y Plástico	4	19	2	25
Fab. de Productos Minerales no Metalicos	3	24	9	36
Industrias Metalicas Básicas	-	4	1	5
Fab. de Prod. Metalicos	2	27	12	41
Fab. de Maq. y Equipo Excepto los Electricos	9	13	2	24
Fab. de Maq. y Equipo y Aparatos Electricos	-	4	3	7
Construcción de Equipo de Transporte	3	6	5	14
Otras Industrias Menufactureras	2	3	1	6
TOTAL	56 R	279 S	84 T	419 V

Cálculo de las frecuencias esperadas

GRUPO INDUSTRIAL	X_1	X_2	X_3	TOTAL
B_1	V_1 (R/V) = 11	V_1 (S/V) = 56	V_1 (T/V) = 17	V_1 = 84
B_2	V_2 (R/V) = 4	V_2 (S/V) = 17	V_2 (T/V) = 6	V_2 = 28
B_3	V_3 (R/V) = 5	V_3 (S/V) = 27	V_3 (T/V) = 9	V_3 = 40
B_4	V_4 (R/V) = 5	V_4 (S/V) = 25	V_4 (T/V) = 7	V_4 = 37
B_5	V_5 (R/V) = 2	V_5 (S/V) = 10	V_5 (T/V) = 3	V_5 = 15
B_6	V_6 (R/V) = 3	V_6 (S/V) = 14	V_6 (T/V) = 4	V_6 = 21
B_7	V_7 (R/V) = 3	V_7 (S/V) = 13	V_7 (T/V) = 4	V_7 = 20
B_8	V_8 (R/V) = 2	V_8 (S/V) = 11	V_8 (T/V) = 3	V_8 = 16
B_9	V_9 (R/V) = 3	V_9 (S/V) = 17	V_9 (T/V) = 5	V_9 = 25
B_{10}	V_{10} (R/V) = 5	V_{10} (S/V) = 24	V_{10} (T/V) = 7	V_{10} = 36
B_{11}	V_{11} (R/V) = 1	V_{11} (S/V) = 3	V_{11} (T/V) = 1	V_{11} = 5
B_{12}	V_{12} (R/V) = 5	V_{12} (S/V) = 28	V_{12} (T/V) = 8	V_{12} = 41
B_{13}	V_{13} (R/V) = 3	V_{13} (S/V) = 16	V_{13} (T/V) = 5	V_{13} = 24
B_{14}	V_{14} (R/V) = 1	V_{14} (S/V) = 5	V_{14} (T/V) = 1	V_{14} = 7
B_{15}	V_{15} (R/V) = 2	V_{15} (S/V) = 9	V_{15} (T/V) = 3	V_{15} = 14
B_{16}	V_{16} (R/V) = 1	V_{16} (S/V) = 4	V_{16} (T/V) = 1	V_{16} = 6
TOTAL	R=56	S=279	T=84	V=419

Agrupándolos por celda, tendremos:

Celda			fr	Fe	fr-fe
1	-	1	10	11	-1
1	-	2	61	56	5
1	-	3	13	17	-4
2	-	1	3	5	-2
2	-	2	22	17	5
2	-	3	3	6	-3
3	-	1	4	4	0
3	-	2	27	27	0
3	-	3	9	9	0
4	-	1	5	5	0
4	-	2	25	25	0
4	-	3	7	7	0
5	-	1	1	2	-1
5	-	2	9	10	-1
5	-	3	5	3	2
6	-	1	1	3	-2
6	-	2	11	14	-3
6	-	3	9	4	5
7	-	1	6	3	3
7	-	2	13	13	0
7	-	3	1	4	-3
8	-	1	3	2	1
8	-	2	11	11	0
8	-	3	2	3	-1
9	-	1	4	3	1
9	-	2	19	17	2
9	-	3	2	5	-3
10	-	1	3	5	-2
10	-	2	24	24	0
10	-	3	9	7	2
11	-	1	0	1	-1

Celda			fr	fe	fr-fe
13	-	1	9	3	6
13	-	2	13	16	-3
13	-	3	2	5	-3
14	-	1	0	1	-1
14	-	2	4	5	-1
14	-	3	3	1	2
15	-	1	3	2	1
15	-	2	6	9	-3
15	-	3	5	3	2
16	-	1	2	1	1
16	-	2	3	4	-1
16	-	3	1	1	0

11	-	2	4	3	1
11	-	3	1	1	0
12	-	1	2	5	-3
12	-	2	27	28	-1
12	-	3	12	8	4

Donde: fr = frecuencia real
fe = frecuencia esperada

Haciendo las comparaciones, entre fr, fe para sustituirlas en la fórmula, se obtiene:

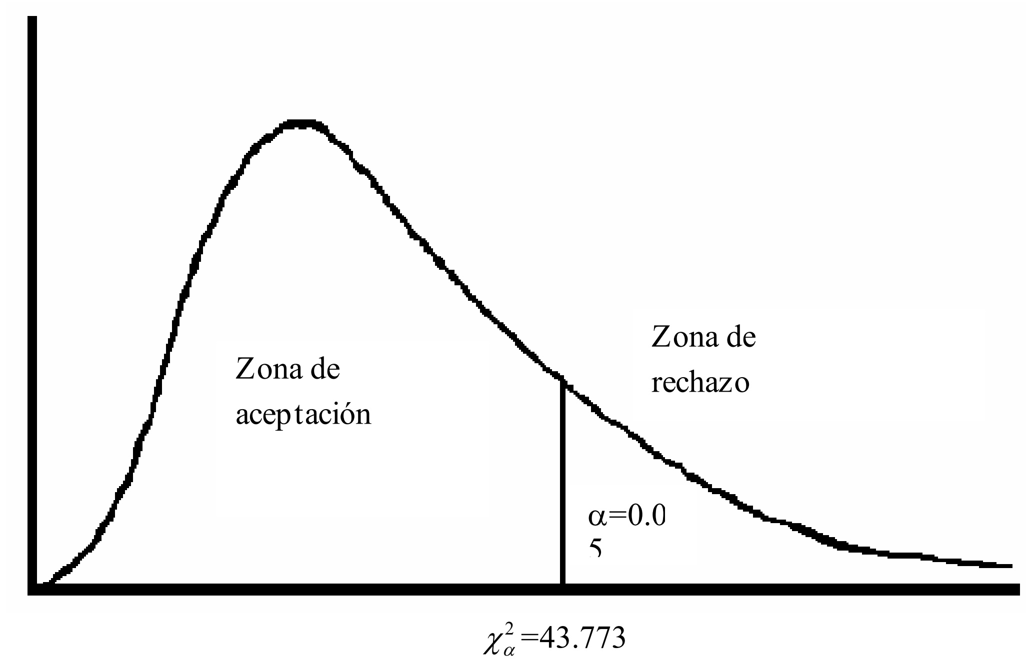
$(fr-fe)^2$	$\frac{(fr-fe)^2}{fe}$
1	0.0909
25	0.4464
16	0.9412
4	0.8000
25	1.4706
9	1.5000
0	0.0000
0	0.0000
0	0.0000
0	0.0000
0	0.0000
0	0.0000
1	0.5000
1	0.1000
4	1.3333
4	1.3333
9	0.6429
29	7.2500
9	3.0000
0	0.0000
9	2.2500
1	0.5000
0	0.0000
1	0.3333
1	0.3333

$(fr-fe)^2$	$\frac{(fr-fe)^2}{fe}$
2	0.1176
9	1.8000
4	0.8000
0	0.0000
4	0.5714
1	1.0000
1	0.3333
0	0.0000
9	1.8000
1	0.0357
16	2.0000
36	12.0000
9	0.5625
9	1.8000
1	1.0000
1	0.2000
4	4.0000
1	0.5000
9	1.0000
4	1.3333
1	1.0000
1	0.2500
0	0.0000
Total	53.90

$$\chi^2 = \frac{(fr-fe)^2}{fe} \qquad \chi^2 = 53.90$$

con $\alpha = 0.05$ y (c-1)(R-1) = (3-1)(16-1) = 30 grados de libertad.

el valor critico de $\chi^2_\alpha = 43.773$ tenemos que



Como $\chi^2 = 53.90 > \chi^2_\alpha = 43.773$ se rechaza la hipótesis nula de que no hay diferencia entre el tamaño de la muestra y la opinión de los empresarios.

Luego se inicia la prueba Phi (ϕ) para cuantificar el grado de asociación entre las dos variables descriptivas.

$$\phi = \left[\frac{\chi^2}{n}\right]^{1/2} = \left(\frac{53.90}{419}\right)^{1/2} = (0.12864)^{1/2} = 0.358$$

La interpretación es que hay una relación sensiblemente significativa.

Como la tabla de contingencia es más grande que una de dos por dos, se aplica la V Cramer para corregir el valor de ϕ.

$$V = \left[\frac{\phi^2}{C-1}\right]^{1/2} = \left[\frac{(0.358)^2}{2}\right]^{1/2} = \left[\frac{0.128164}{2}\right]^{1/2}$$

$V = (0.064082)^{1/2}$
$V = 0.25$

Puesto que el valor de V oscila entre cero y más uno, no se modifica el tamaño de la muestra para el mes de junio porque la asociación no es fuerte.

Si se hubiera tomado la decisión de hacer el análisis de cobertura y sesgo, el procedimiento seria:

Cobertura: Nuevo Tamaño de la Muestra

A) Muestreo simple aleatorio

Antecedentes

Obtener el tamaño de la muestra adecuado para asegurar con una probabilidad igual a 95%, que el error en la estimación del número medio de empresas necesarias no sea mayor del 6%.

Para ella se tomó la muestra aleatoria del mes de junio, la cual fue de 419 empresas distribuidas en 16 grupos industriales de la siguiente manera:

	Concepto	Nº de Empresas (X_i)
	Total	419
1 .-	Fab. de alimentos	84
2 .-	Industria Textil	28
3 .-	Fab. de Prendas de Vestir	40
4 .-	Fab. de Calzado e Ind. del Cuero	37
5 .-	Ind. y Prod. de Madera y Corcho Excepto Muebles	15
6 .-	Fab. y Rep. de Muebles de Madera	21
7 .-	Ind. Editorial de Impresión y Conexas	20
8 .-	Industria Química	16
9 .-	Fab. de Prod. de Hule y Plástico	25
10 .-	Fab. de Productos Minerales no Metalicos	36
11 .-	Industrias Metalicas Básicas	5
12 .-	Fab. de Prod. Metalicos	41
13 .-	Fab. de Maq. y Equipo Excepto los Electricos	24
14 .-	Fab. de Maq. y Equipo y Aparatos Electricos	7
15 .-	Construcción de Equipo de Transporte	14
16 .-	Otras Industrias Menufactureras	6

2 Cálculo

Como no se conocen los valores de los parámetros poblacionales μ y σ^2, es necesario estimarlos a partir de las estadísticas $\bar{x}$ y S^2 de la muestra. Así;

Grupo Industrial	X_i	%	X_i^2
1	84	20	7,056
2	28	7	784
3	40	10	1,600
4	37	9	1,369
5	15	4	225
6	21	5	441
7	20	5	400
8	16	4	256
9	25	6	625
10	36	9	1,296
11	5	1	25
12	41	10	1,681
13	24	6	576
14	7	2	49
15	14	3	196
16	6	1	36
Suma	419	102	16,615

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{16}(419) = 26 \text{ empresas}$$

$$S^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2 = \frac{1}{16}(16{,}615) - (26)^2 = 1038 - 676 = 362 \text{ empresas}$$

Considerando que el error en la estimación (e) del promedio de empresas no debe ser superior al 6%, y recordando que el estimador de μ = $\bar{x}$ = 26 empresas, se observa que e = 26 (0.06) = 1.56 empresas.

Igualmente, como se desconoce el valor de σ^2 y tomando en cuenta que su estimador proviene de una muestra mayor de 30 empresas, la distribución teórica a la cual se aproxima la distribución de muestreo es a la normal.

En este caso se estima μ de la población con variable aleatoria asociada X mediante el empleo de $\bar{x}$, proveniente de n = 419 con e = 6 % y un nivel de confianza ξ = 95 %, donde Z= desviación correspondiente al nivel de confianza de ξ en la distribución normal; en este caso la probabilidad ξ le corresponde Z_{α} = ± 1.96.

Considerando a K $\sigma_{\bar{x}_x}$ como = $Z\alpha(\sigma_{\bar{x}})$ este razonamiento para obtener el tamaño de la muestra se basa en el hecho de que:

$$P(\bar{x} - k\sigma \le \mu \le \bar{x} + k\sigma) = Pk = 1 - \alpha = 95\%$$

α = nivel de significación = 5%

En otras palabras $P\left[\left|\hat{p}-p\right| \geq 0.06p\right] = 1-0.95 = 5\%$

Ello significa que el error en la estimación del valor de μ en valores absolutos es:
|error en la estimación de μ| = $k\sigma$, por lo que
|error máximo admisible| = |error en la estimación de μ| = e
Derivado de lo anterior se puede escribir.
$e = k\sigma_{\bar{x}} = Z_{\alpha}\sigma_{\bar{x}}$ donde Z_{α} = variable estandarizada.

donde $\sigma\bar{x} = \dfrac{\sigma}{\sqrt{n}}$,para una población infinita.

Sabiendo que K= Z

Cuando la población es finita $e = k\sqrt{\dfrac{\sigma^2}{n}\dfrac{N-n}{N-1}}$

Como no se conoce σ^2, la estima S^2 y sabiendo que K=Z

$e = Z\sigma_{\bar{x}} = Z\sqrt{\dfrac{S^2}{n}\dfrac{N-n}{N-1}}$

Para obtener el tamaño de la muestra (n), se despeja de la ecuación anterior elevando al cuadrado ambos miembros.

$$e^2 = Z^2\frac{S^2}{n}\frac{N-n}{N-1}$$

Así: $n = \dfrac{Z^2 * S^2 * N}{e^2N - e^2 + Z^2S^2}$

Con e = 6% ; en absolutos e = 26(0.06) = 1.56 empresas;
α = 5 %
ξ = 95 %

Z = $\pm$ 1.96
S^2 = 362
N = 8,966

$$n = \frac{Z^2 * S^2 * N}{e^2N - e^2 + Z^2S^2} = \frac{(1.96)^2(362)(8,966)}{(1.56)^2(8,966) + (362)(1.96) - (1.56)^2} = \frac{12,468,650}{21,820 + 1,391 - 2} = \frac{12,468,650}{23,209}$$

n = 537 empresas.

Comprobación del valor de (e)

$$e^2 = Z^2\frac{S^2}{n}\frac{N-n}{N-1} = (3.84)(0.6741)(0.94)$$

e^2 = 2.43 luego e = 1.56 empresas= error permitido= error de muestreo.

Si deseamos distribuir la muestra de 537 empresas por grupo industrial, se hace con el procedimiento llamado de **afijación proporcional de la muestra, de conformidad con la importancia que tenga cada estrato (Ni) dentro del universo (N).**

Grupo Industrial	Ni/N %	n= 537	n_i
1			
2			
3			
4			
5			

Donde i = 1, 2, 3, 4, 5, ..., 16
por lo que $n_1 + n_2 + n_3 + ...+n_{16} = n = 537$

B. muestreo estratificado

Tomando como referencia los datos de este diseño muestral que aplicamos en el inciso en que hablamos de la precisión, donde indicamos que el error de muestreo se mide con el error estándar, entonces digamos ahora que si el error estándar de la proporción proveniente de una distribución de muestreo estratificada finita es:

$$\sigma_p = \sqrt{\sum_1^k W_i^2 S_i^2 \frac{N_i - n_i}{N_i * n_i}}$$

$$\sigma_p^2 = \sum_1^k W_i^2 S_i^2 \frac{N_i - n_i}{N_i * n_i} = \frac{\sum^k W_i^2 S_i^2 N_i - n_i \sum^k W_i^2 S_i^2}{N_i * n_i}$$

$$\sigma_p^2 (N_i * n_i) = N_i \sum_1^k W_i^2 S_i^2 - n_i \sum_1^k W_i^2 S_i^2$$

$$\sigma_p^2 (N_i * n_i) + n_i \sum_1^k W_i^2 S_i^2 = N_i \sum_1^k W_i^2 S_i^2$$

Entonces :

$$n_i (\sigma_p^2 N_i + \sum_1^k W_i^2 S_i^2) = N_i \sum_1^k W_i^2 S_i^2$$

$$n_i = \frac{N_i \sum^k W_i^2 S_i^2}{\sigma_p^2 N_i + \sum_1^k W_i^2 S_i^2} ; como S^2 = pq$$

$$n_i = \frac{N_i \sum^{k} W_i^2 pq}{\sigma_p^2 N_i + \sum_1^{k} W_i^2 pq}$$

Ejemplo:

Estratos	N_i	W_i	n_i muestra	Empresas de la muestra que contestaron	P_i
1	7,000	0.7	200	160	160 ÷ 200 = 0.8
2	1,000	0.1	100	40	40 ÷ 100 = 0.4
3	2,000	0.2	100	60	60 ÷ 100 = 0.6
	10,000	1	400	260	

Con $\sigma p = 0.025$

Como $\sum_1^{k} W_i^2 S_1^2$ =(0.49)(0.16)+(0.01)(0.24)+(0.04)(0.24)=0.0784 + 0.0024 + 0.0096 = 0.904

La muestra para cada estrato se va obteniendo así:

$$n_1 = \frac{7{,}000(0.0904)}{(0.025)^2 7{,}000 + 0.0904}$$

$$n_1 = \frac{633}{4{,}465} = 142;$$

$$n_2 = \frac{1{,}000(0.0904)}{(0.000625)1{,}000 + 0.0904} = \frac{90.4}{0.715} = 126$$

$$n_3 = \frac{2{,}000(0.0904)}{(0.000625)2{,}000 + 0.09041} = 135$$

$n_1 + n_2 + n_3 = n = 402$

<u>Sesgo : Limites de Central</u>

Para el análisis de sesgo se definen limites de control (o de confianza) donde con cierta probabilidad se mantendrá el valor del porciento con un tamaño dado de muestra.

Así limites de control = $p \pm Z\ \sigma p$

Cuando se salga de esos limites de control nuevamente se hará la prueba de X^2 ; si se rechaza la hipótesis nula, nuevamente se revisará la muestra en el grupo y se determinará si el porciento es legitimo o se debe a errores de muestreo, de tal manera que el proceso se vuelve interactivo, en el sentido de que se harán ajustes cuantas veces sea necesario hasta llegar a muestras satisfactorias.