INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

Profesor: Mattia Makovec
mmakovec@dii.uchile.cl
Auxiliar: María Paz Jofré
mjofre@ing.uchile.cl

Auxiliar N°2 IN540

29 Octubre 2008

Resumen:

(a) Definición del Problema:
  a.1) Problema de Negocio
  a.2) Problema de Investigación
  a.3) Objetivos
  a.4) Limpieza de Datos y Estadísticos Descriptivos

(b) Segmentación:
  b.1) Propiedades de una buena segmentación
  b.2) Técnicas de Segmentación

(c) Análisis Discriminante

P1. SPSS: Considere una investigación de mercados sobre el uso de máquinas tragamonedas en almacenes de barrio.

Idea: un importador de máquinas de juego quiere conocer cómo se comporta actualmente la demanda. Él está considerando realizar algunas de las siguientes acciones: importar máquinas con diseños específicos, crear una tarjeta de prepago, focalizar la distribución de sus máquinas, cambiarse de rubro, disminuir precios en ciertos sectores.

a) Exploración:

i) Realice estadísticas descriptivas de las variables.
ii) Calcule cruces entre un par de variables que le parezcan interesantes.
iii) Determine la vinculación entre estas variables.

b) Segmentación:

iv) Realice una segmentación sobre las variables R, F y M. Bautice a los segmentos y justifique la robustez de su elección.
v) Prediga su segmentación a través de una segmentación con Chaid.

c) Análisis Discriminante:

vi) ¿Es posible predecir el sexo en función de RFM?
vii) ¿Es posible predecir su segmentación de RFM a través de un análisis discriminante?

INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

Profesor: Mattia Makovec
mmakovec@dii.uchile.cl
Auxiliar: María Paz Jofré
mjofre@ing.uchile.cl

Pauta Auxiliar N°2 IN540

29 Octubre 2008

Resumen:

(a) Definición del Problema:

a.1) Problema de Negocio
a.2) Problema de Investigación
a.3) Objetivos
a.4) Limpieza de Datos y Estadísticos Descriptivos

- Media
- Mediana
- Desviación Estándar
- Mínimo
- Máximo

(b) Segmentación:

b.1) Propiedades de una buena segmentación:

- Accesibilidad
- Asociación al Problema de Negocio
- Diferenciabilidad

b.2) Técnicas de Segmentación

- Nubes Dinámicas: K-Means
- Árboles Jerárquicos: Chaid
- Heriarchical Cluster

(c) Análisis Discriminante

P1. SPSS: Considere una investigación de mercados sobre el uso de máquinas tragamonedas en almacenes de barrio.

Idea: un importador de máquinas de juego quiere conocer cómo se comporta actualmente la demanda. Él está considerando realizar algunas de las siguientes acciones: importar máquinas con diseños específicos, crear una tarjeta de prepago, focalizar la distribución de sus máquinas, cambiarse de rubro, disminuir precios en ciertos sectores.

a) Exploración:

i) Realice estadísticas descriptivas de las variables.

	N	Mínimo	Máximo	Media	Desv. Típ.	Varianza
EDAD	209	26	52	36,42	7,932	62,917
HIJOS	209	2	5	3,50	,821	,674
INGRESO	209	120000	870000	370382,78	243679,422	59379660471,108
SHARE	209	30,0%	100,0%	56,268%	22,1760%	491,774
R	209	0	19	4,49	4,414	19,482
F	209	3	19	10,60	5,100	26,010
M	209	500	10000	2212,92	2023,629	4095072,690
N válido (según lista)	209					

ii) Calcule cruces entre dos pares de variables que le parezcan interesantes.

Analyze→ Descriptive Statistics →Crosstabs

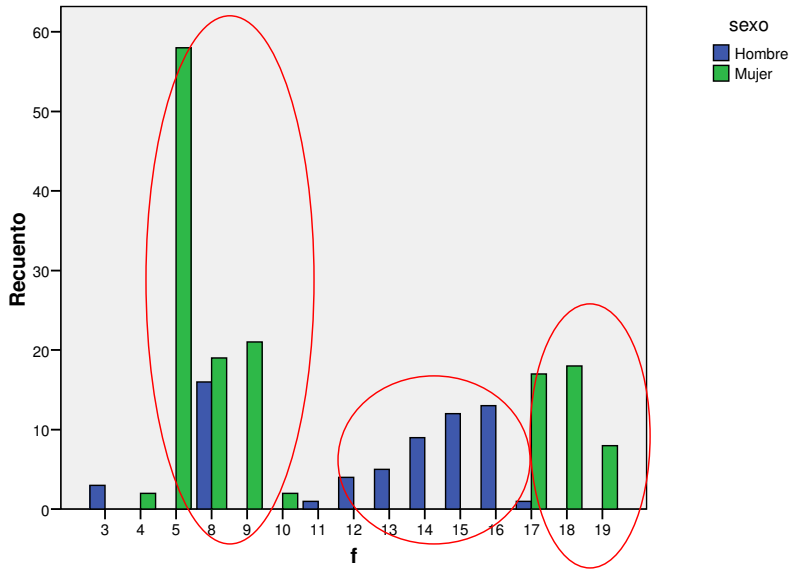
- ***Frecuencia v/s Sexo:***

Tabla de contingencia f * sexo

Recuento

		sexo		
		Hombre	Mujer	Total
f	3	3	0	3
	4	0	2	2
	5	0	58	58
	8	16	19	35
	9	0	21	21
	10	0	2	2
	11	1	0	1
	12	4	0	4
	13	5	0	5
	14	9	0	9
	15	12	0	12
	16	13	0	13
	17	1	17	18
	18	0	18	18
	19	0	8	8
Total		64	145	209

Gráfico de barras



Se observa que hay un grupo de mujeres que juega con mucha frecuencia y otro con poca frecuencia. En cuanto a los hombres, estos tienen una frecuencia similar.

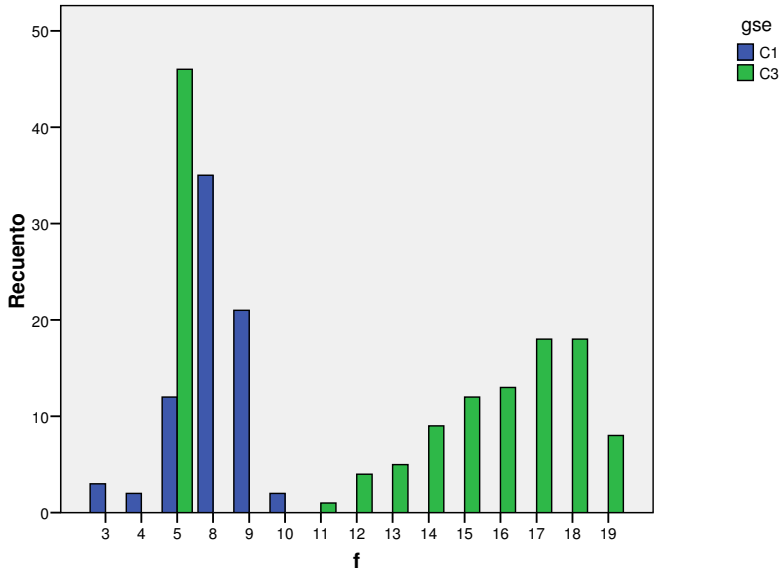
- ***Frecuencia v/s GSE:***

Tabla de contingencia f * gse

Recuento

		gse		
		C1	C3	Total
f	3	3	0	3
	4	2	0	2
	5	12	46	58
	8	35	0	35
	9	21	0	21
	10	2	0	2
	11	0	1	1
	12	0	4	4
	13	0	5	5
	14	0	9	9
	15	0	12	12
	16	0	13	13
	17	0	18	18
	18	0	18	18
	19	0	8	8
Total		75	134	209

Gráfico de barras



Se observa que el grupo C1 juega con menor frecuencia que el grupo C3.

iii) Determine la vinculación entre estas variables.

Para determinar cuan vinculadas están las variables se deben observar los estadísticos Chi-cuadrado.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1,637E2	14	,000
Likelihood Ratio	201,519	14	,000
Linear-by-Linear Association	11,138	1	,001
N of Valid Cases	209		

a. 16 cells (53,3%) have expected count less than 5. The minimum expected count is ,31.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1,676E2	14	,000
Likelihood Ratio	213,713	14	,000
N of Valid Cases	209		

a. 16 cells (53,3%) have expected count less than 5. The minimum expected count is ,36.

b) Segmentación:

iv) Realice una segmentación sobre las variables R, F y M. Bautice a los segmentos y justifique la robustez de su elección.

Analyze→ Classify → K-Means Cluster

- **Para K=2:**

Final Cluster Centers

	Cluster	
	1	2
R	3	17
F	11	5
M	$1.740	$8.333

Number of Cases in each Cluster

Cluster	1	194,000
	2	15,000
Valid		209,000
Missing		,000

De las tablas anteriores se puede ver que se definen dos segmentos bastantes diferentes entre sí, donde el primero se compone de 194 personas y el segundo sólo de 15. Es por esta razón que pareciese conveniente tratar de separar el primer segmento donde se concentra una gran masa de individuos.

- **Para K=3:**

Final Cluster Centers

	Cluster		
	1	2	3
R	1	6	18
F	16	7	5
M	$602	$2.710	$8.536

Number of Cases in each Cluster

Cluster	1	88,000
	2	107,000
	3	14,000
Valid		209,000
Missing		,000

En este caso en particular, se obtienen tres segmentos bastante diferenciados entre sí. Con esta segmentación se logró lo que se quería, es decir, ahora se tiene el primer y segundo conglomerado con un menor número de individuos que en el caso anterior. En cuanto al tercer

segmento se tiene que se comportan de manera similar al segundo grupo mencionado en la primera segmentación.

- **Para K=4:**

Final Cluster Centers

	Cluster			
	1	2	3	4
R	1	6	17	19
F	16	7	5	5
M	$602	$2.684	$7.909	$9.500

Number of Cases in each Cluster

Cluster	1	88,000
	2	106,000
	3	11,000
	4	4,000
Valid		209,000
Missing		,000

Para esta segmentación se obtienen 4 conglomerados. Los tres primeros se parecen bastante a los encontrados en la parte anterior, pero además se suma un tercer cluster muy similar al tercero, por lo que se concluye que segmentar el mercado de las máquinas tragamonedas en más de tres segmentos no aporta mayor información al estudio.

De lo obtenido en la parte anterior, se puede decir que la mejor segmentación entonces es la que separa al mercado en tres conglomerados, ya que todos los clusters se diferencian de manera clara y poseen una cantidad de individuos adecuado.

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
R	1811,922	2	2,080	206	871,289	,000
F	2350,541	2	3,442	206	682,992	,000
M	4,072E8	2	181215,436	206	2,247E3	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

La elección de estos segmentos tiene una gran robustez debido a que en el test Anova se rechaza la hipótesis nula, lo que nos indica que existe una diferencia de medias entre grupos significativa.

A continuación se bautizará los segmentos escogidos en base a las variables R, F y M, con una pequeña descripción de cada uno de ellos.

Segmento N°1: *Los Viciosos*

Este segmento lo conforman individuos que juegan con bastante frecuencia en las máquinas tragamonedas, la última vez que lo hicieron fue hace muy pocos días y destinan un monto bajo cada vez que realizan esta actividad.

Segmento N°2: *Los Recatados*

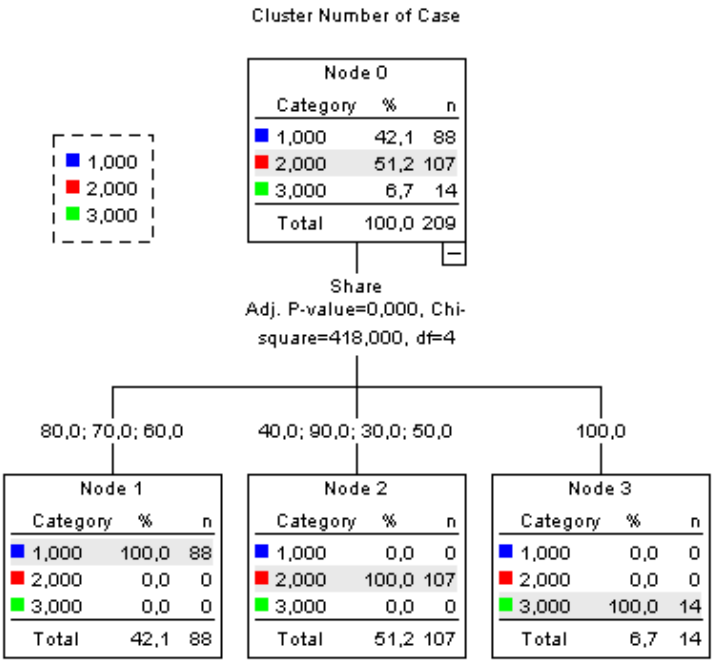
Este segmento está conformado por individuos que juegan con una frecuencia menor que los del primer segmento, pero que no deja de ser importante. Además, la última vez que jugaron en alguna máquina fue hace un tiempo no menor y destinan un monto significativo cada vez que realizan esta actividad.

Segmento N°3: *Los Despilfarradores*

Este segmento está conformado por individuos que juegan con una muy baja frecuencia y la última vez que jugaron fue hace muchos días, pero que cada vez que lo hacen destinan una cantidad significativamente alta de dinero.

v) Prediga su segmentación a través de una segmentación con Chaid.

Analyze→ Classify → Tree



Classification

Observed	Predicted			
	1	2	3	Percent Correct
1	88	0	0	100,0%
2	0	107	0	100,0%
3	0	0	14	100,0%
Overall Percentage	42,1%	51,2%	6,7%	100,0%

Growing Method: CHAID

Dependent Variable: Cluster Number of Case

Del árbol anterior se puede concluir que la variable independiente que mejor explica la nueva variable es el share.

Se puede ver en la tabla de confusión que la predicción es de un 100%, es decir, que el poder predictivo de este modelo es certero. Esto se puede deber a que la correlación entre la variable dependiente y la independiente es muy alta, lo que ayuda a predecir con mayor seguridad en que conglomerado pertenecerá cada individuo.

c) Análisis Discriminante:

vi) ¿Es posible predecir el sexo en función de RFM?

Analyze→ Classify → Discriminant

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,813	42,542	3	,000

Al tener un lambda de Wilks tan cercano a uno, se puede concluir que las variables independientes no son determinantes para inferir la variable de agrupación, por lo que la respuesta a la pregunta es negativa; no es posible predecir el sexo a partir de las variables R, F y M. Si el lambda de Wilks fuera cercano a cero, entonces sí se podría predecir el sexo a partir de las variables de consumo.

vii) ¿Es posible predecir su segmentación de RFM a través de un análisis discriminante?

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	,001	1327,465	10	,000
2	,179	350,775	4	,000

El lambda de Wilks obtenido es cero, lo cual insinúa que el poder predictivo de las funciones discriminantes es alto.