MATRIZ DE DATOS
(Guía de clase)

FICHA Nº 6

Con base en la bibliografía sugerida para este tema, Cortés, F. (2000) *Estadística elemental aplicada.* IFE. México, DF; y Galtung, J. (1965) *Teoría y métodos de la investigación social*, EDUDEBA, Buenos Aires; se pueden resaltar los siguientes tres aspectos principales: la estructura tripartita del dato; la noción de la matriz de datos y los principios o criterios para evaluar una matriz de datos.

A. En términos generales, el **DATO,** tal como se lo conceptualiza desde la metodología, es el valor que toma una variable en una unidad de análisis. Por esta razón se dice que su estructura es "tripartita".

✎ En primer lugar, refiere a una unidad de registro:

i) Las unidades pueden ser tanto individuos, organizaciones, territorios, años (series temporales, procesos evolutivos individuales), países, redes, textos, entrevistas, etc.

ii) Para compararse dos datos deben estar referidos a un mismo tipo de unidad. En la gran mayoría de análisis estadísticos, no se pueden combinar distintas unidades.

✎ En segundo lugar, refiere a un conjunto de variables:

i) Estas pueden ser de cualquier escala de medición (nominal, ordinal, interval, de razón).

ii) Para una misma unidad se pueden combinar las escalas de medida.

✎ Finalmente, se refiere a valores:

i) El término valor debe entenderse en el nivel lógico más bajo: como distinción dentro de un conjunto de oposiciones. No requieren ser numéricos (en el sentido de corresponder al conjunto de los números naturales, reales o racionales). Los paquetes estadísticos, como el SPSS en sus versiones más recientes admiten por ejemplo, letras.

ii) Los valores deben estar previamente definidos por las variables.

iii) Los valores deben ser coherentes con las unidades a las cuales se aplican. Si se aplica una misma variable a dos unidades distintas, conceptualmente los valores son diferentes.

MATRIZ DE DATOS
(Guía de clase)

B. La **MATRIZ DE DATOS** es un modo de ordenar los datos de manera que sea particularmente visible la estructura tripartita de los datos.

✎ Los datos se arreglan de tal forma que las unidades (U=1,2,3..... *I*) se ubican en los renglones y cada variable (V=1,2,3.... *K)* en las columnas.

i) Si se desea conocer todas las características de una unidad específica se recorre todo el renglón.

ii) Si se desea conocer como se distribuyen las unidades en las distintos valores de una variable, se recorre la columna.

✎ Las celdas están formadas por las intersecciones de los renglones y las columnas contienen los valores (r)

i) cada valor (r) es la respuesta de la *i-ésima* unidad en la *k-ésima variable.* A la inversa: toda combinación (U_i , V_K) define en la matriz un punto (R_{ik}).

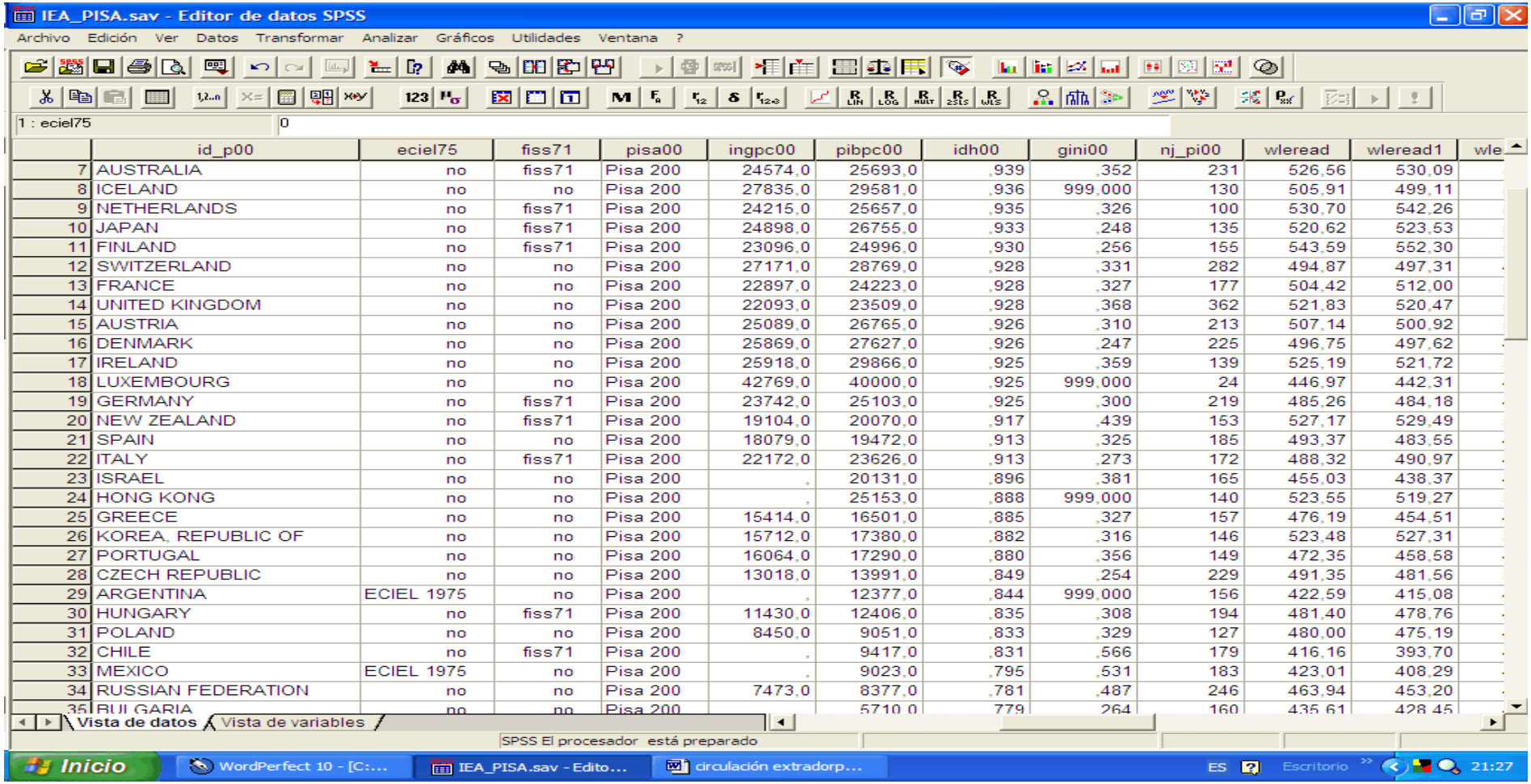
ii) La falta de valor (de un valor de los predeterminados)en una celda es denominado "sin datos" o "missing values". En ocasiones, la ausencia de valor se representa mediante un código (99, 999, etc). Es una situación frecuente en las matrices de datos. Una forma de valorar una matriz es por la cantidad de "sin datos" que tiene.

✎ En el cuadro siguiente se presenta un segmento de una matriz de datos elaborada con el objetivo de analizar cuáles son las regularidades en la estructura explicativa de los niveles de aprendizajes que se han presentado durante los últimos treinta años en las evaluaciones internacionales.

i) Las unidades son países que al menos han participado en alguna de las siguientes evaluaciones: First International Science Study (FISS,1971), el estudio ECIEL (1975), Third International Mathematics and Science Study (TIMSS), Programme for International Student Assessment (PISA). A estos países se le añade la matriz de datos elaborada por Heyneman & Loxley para su estudio de 1982.

ii) Las variables que se presentan son: participación en ECIEL, participación en FISS, participación en PISA, ingreso per cápita en el 2000, PIB per cápita en el 2000, índice de desarrollo humano 2000, índice de Gini 2000, número de escuelas en PISA 2000, puntaje global en lectura en PISA 2000; etc.

MATRIZ DE DATOS
(Guía de clase)



	id_p00	eciel75	fiss71	pisa00	ingpc00	pibpc00	idh00	gini00	nj_pi00	wleread	wleread1	wle
7	AUSTRALIA	no	fiss71	Pisa 200	24574,0	25693,0	,939	,352	231	526,56	530,09	
8	ICELAND	no	no	Pisa 200	27835,0	29581,0	,936	999,000	130	505,91	499,11	
9	NETHERLANDS	no	fiss71	Pisa 200	24215,0	25657,0	,935	,326	100	530,70	542,26	
10	JAPAN	no	fiss71	Pisa 200	24898,0	26755,0	,933	,248	135	520,62	523,53	
11	FINLAND	no	fiss71	Pisa 200	23096,0	24996,0	,930	,256	155	543,59	552,30	
12	SWITZERLAND	no	no	Pisa 200	27171,0	28769,0	,928	,331	282	494,87	497,31	
13	FRANCE	no	no	Pisa 200	22897,0	24223,0	,928	,327	177	504,42	512,00	
14	UNITED KINGDOM	no	no	Pisa 200	22093,0	23509,0	,928	,368	362	521,83	520,47	
15	AUSTRIA	no	no	Pisa 200	25089,0	26765,0	,926	,310	213	507,14	500,92	
16	DENMARK	no	no	Pisa 200	25869,0	27627,0	,926	,247	225	496,75	497,62	
17	IRELAND	no	no	Pisa 200	25918,0	29866,0	,925	,359	139	525,19	521,72	
18	LUXEMBOURG	no	no	Pisa 200	42769,0	40000,0	,925	999,000	24	446,97	442,31	
19	GERMANY	no	fiss71	Pisa 200	23742,0	25103,0	,925	,300	219	485,26	484,18	
20	NEW ZEALAND	no	fiss71	Pisa 200	19104,0	20070,0	,917	,439	153	527,17	529,49	
21	SPAIN	no	no	Pisa 200	18079,0	19472,0	,913	,325	185	493,37	483,55	
22	ITALY	no	fiss71	Pisa 200	22172,0	23626,0	,913	,273	172	488,32	490,97	
23	ISRAEL	no	no	Pisa 200	.	20131,0	,896	,381	165	455,03	438,37	
24	HONG KONG	no	no	Pisa 200	.	25153,0	,888	999,000	140	523,55	519,27	
25	GREECE	no	no	Pisa 200	15414,0	16501,0	,885	,327	157	476,19	454,51	
26	KOREA, REPUBLIC OF	no	no	Pisa 200	15712,0	17380,0	,882	,316	146	523,48	527,31	
27	PORTUGAL	no	no	Pisa 200	16064,0	17290,0	,880	,356	149	472,35	458,58	
28	CZECH REPUBLIC	no	no	Pisa 200	13018,0	13991,0	,849	,254	229	491,35	481,56	
29	ARGENTINA	ECIEL 1975	no	Pisa 200	.	12377,0	,844	999,000	156	422,59	415,08	
30	HUNGARY	no	fiss71	Pisa 200	11430,0	12406,0	,835	,308	194	481,40	478,76	
31	POLAND	no	no	Pisa 200	8450,0	9051,0	,833	,329	127	480,00	475,19	
32	CHILE	no	fiss71	Pisa 200	.	9417,0	,831	,566	179	416,16	393,70	
33	MEXICO	ECIEL 1975	no	Pisa 200	.	9023,0	,795	,531	183	423,01	408,29	
34	RUSSIAN FEDERATION	no	no	Pisa 200	7473,0	8377,0	,781	,487	246	463,94	453,20	
35	BULGARIA	no	no	Pisa 200		5710,0	,779	,264	160	435,61	428,45	

C. Galtung propone tres principios para realizar una **EVALUACIÓN DE LA MATRIZ** de datos que se derivan lógicamente de las nociones de "estructura tripartita del dato" y de la noción de "matriz de datos".

✎ El principio de comparabilidad requiere que toda proposición (U_i , V_k) determina un valor (R_{ik}) que debe ser verdadero o falso para cada i ,k.

i) En una misma matriz de datos, la misma variable V_k debe ser (teórica y metodológicamente) la misma para todas las U_i unidades. Se deriva de que una variable se define para un mismo tipo de unidades.

ii) El mismo tipo de unidad de análisis es el referente de todas las variables U_i expuesto en la matriz. Se deriva de que una matriz tiene un solo tipo de unidades.

iii) Su mayor aplicación es en la investigación comparativa (internacional, inter-étnica) donde se integran datos provistos por distintos contextos culturales. Aquí se debe garantizar la comparabilidad "semántica" entre las preguntas de un cuestionario por ejemplo, más que la comparabilidad "morfosintáctica". Al respecto de la comparación de indicadores internacionales véase: PRWZEWORSKI, Adam & TEUNE, Henry (1970) *The logic of comparative social inquire*. John Wiley. NY.

✎ El principio de clasificación supone que las categorías de respuestas R_{ik} debe producir una clasificación de todos los pares (U_i, V_k).

i) Se requiere que todas las unidades sean en principio lógicamente clasificables en alguno de los valores (R) dispuestos para las *k* variables. Es una consecuencia del principio de exhaustividad integrado en el concepto de variable.

ii) Solamente debe lógicamente existir un único valor (R) para cada punto (U_i , V_k), lo cual se deriva del principio de exclusión del concepto de variable.

iii) Si lógicamente se supone que habrán distintas proposiciones (R_{ik}) para una misma unidad (U_i), en la matriz deberán existir tantas "columnas" como clasificaciones se pretendan hacer de las unidades.

✎ El principio de integridad de la matriz de datos exige que para todo valor (U_i, V_k) debe existir empíricamente un valor (R_{ik}).

i) La ausencia de uno de los valores (R_{ik}) se denomina "sin datos" o "missing

value" y da lugar a un complejo trabajo metodológico y estadístico para examinar qué características pueden tener las unidades que carecen de valores.

ii) Con este principio se evalúa concretamente el trabajo empírico de producción del dato y del "llenado de la matriz", y no como en los restantes casos, el trabajo lógico de definición de su estructura.

iii) Es la forma de evaluar el trabajo de campo (control de tasas de rechazo totales o parciales en una encuesta, disponibilidad de la información secundaria, etc).

iv) Hay dos formas de examinar la integridad de la matriz: por filas o por columnas.

- Por filas, se examina para cada unidad qué número de celdas carecen de datos. Puede concluirse en la eliminación total de casos en algunas técnicas.

- Por columnas, se examina para cada variable cuál es la frecuencia con que se presenta la ausencia de valores. Puede concluirse en la eliminación de la variable de los análisis subsiguientes.

v) Se acostumbra a definir un estándar de "admisibilidad máxima " para la ausencia de valores. Galtung señaló la conveniencia de que aquella no superase el 5% para cada variables. En la práctica el estándar depende del tipo de instrumento que se haya utilizado en la recolección. Por ejemplo, en los censos se toleran frecuencias más altas. Las variables de ingreso suelen tener altos niveles.

D. Un trabajo fundamental previo de todo análisis estadístico es establecer y fundamentar las **DECISIONES** que orientarán el tratamiento de los casos"sin datos". Si el analista no toma explícitamente decisiones debe saber que todos los paquetes estadísticos trabajan sobre determinados supuestos relativos a la ausencia de información.

✎ Una primera primera decisión de análisis es tratar la ausencia de información como una situación aleatoria en la producción del dato. (Principio de ignorabilidad fuerte)

i) Se supone que la falta del valor no obedeció a ningún problema sistemático en la generación del dato. Por ejemplo, problemas de formulación de una pregunta en un cuestionario, o problemas de comunicación con un determinado grupo social.

ii) Se supone que de haber problemas de levantamiento de la información, estos se compensaron entre sí. Son tratables dentro del marco más general de los errores de medición.

iii) La falta de valores no está concentrada en un mismo bloque de variables, metodológicamente derivadas de un mismo concepto. Es decir, se supone que no hay problemas de operacionalización del concepto.

✎ Puede suponerse que si la ausencia de datos se concentra en algunas variables y tiene una magnitud muy baja, se trata la falta de valores como una categoría residual que se agrega en todos los análisis. (Principio de ignorabilidad débil).

i) Se supone que la variable no cumplía con ser exhaustiva y de ahí deriva el no cumplimiento del principio de clasificación de la matriz.

ii) Se supone que la categoría residual (por ejemplo, "otros no considerados") no altera la operacionalización del concepto involucrado. Esta situación genera problemas con las escalas ordinales, intervales y de razón, donde lógicamente no se puede interpretar una categoría "otros".

iii) Se recomienda que la categoría residual no supere como máximo el 20% .

✎ Si la ausencia de información se concentra en muchas variables para un mismo conjunto de unidades, se puede suponer que existe una razón sistemática teóricamente sustantiva que la produjo.

i) Se supone que las unidades que carecen de información en varias o en todas las variables conforman una situación de "rechazo" del cuestionario, que puede ser parcial o total.

ii) Ignorar esta situación conduce a un sesgo en todos los estadísticos calculados y en todas las estimaciones poblaciones realizadas.

iii) Se supone que las unidades comparten un mismo conjunto de atributos. Por lo general, puede ser directamente proporcionado por las variables de control que se disponen en la misma matriz de datos.

iv) Cuando no se observa directamente en la matriz un patrón regular hipotéticamente causante del rechazo, es necesario realizar inferencias sobre cuál es la causa.

- El rechazo puede ser el resultado de un evento circunstancial que afectó a una sub-población de encuestados
- Puede ser el resultado de un mal encuestador
- Puede ser el resultado de una encuesta que no consideró formulaciones específicas para poblaciones particulares (por ejemplo, traducciones apropiadas).
- Puede ser el resultado de una toma de postura política frente a la investigación en curso.
- Cuando se trabaja integrando información secundaria producida para distintos países, puede ser el resultado de falta de análisis en algunos países debido a condiciones estructurales (por ejemplo, el subdesarrollo).

✎ Si la ausencia de información para una variable tiene una frecuencia importante (mayor al 10%, por ejemplo) y se presenta ***sin ningún patrón de regularidad para un conjunto específico de unidades*** (por ejemplo, en el caso anterior), se puede realizar una imputación de datos faltantes.

i) La imputación opera lógicamente identificando un patrón de regularidad para las unidades que sí dieron respuestas a esa variable. (Principio de intrapolación de datos).

ii) Identificado dicho patrón mediante un modelo estadístico, se le asigna el valor a las unidades que carecen de información.

iii) Se requiere un modelo estadísticamente satisfactorio, es decir con un "buen ajuste" a los datos.

✎ Podría resultar el caso de que en una matriz de datos estuviera ausente toda una columna que resulta fundamental para el tipo de análisis que se desea hacer. Por ejemplo, la estimación de la pobreza sea por el método de los recursos o por un método combinado, requiere contar con información sobre el ingreso no-monetario del hogar. Los censos por lo regular no aportan esta variable aunque sí lo hacen las encuestas de hogares. Es posible formular un modelo de imputación del ingreso no monetario para una encuesta de hogares y extrapolar la imputación al censo. Véase Cortés et al (2003) *Perfiles de la Pobreza en Chiapas. Línea de Base al año 2000.* El Colegio de México.