INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

GOBIERNO DE CHILE
MINISTERIO DE PLANIFICACION

UNIVERSIDAD DE CHILE

DIPLOMA PREPARACIÓN Y EVALUACIÓN SOCIAL DE PROYECTOS

ESTADISTICA DESCRIPTIVA

ESTUDIOS DE CASOS

APLICACIONES DE

SPSS®

Profesora : Sara Arancibia C.
Profesor Auxiliar: Carlos Andrade.

2010

FORMULAS PARA TRIUNFAR

LA FORMULA BÁSICA. Los investigadores se han dedicado a averiguar cuál ha sido la idea, el secreto que ha llevado al triunfo a los grandes personajes de la historia. Y han encontrado una fórmula que todos los triunfadores practicaron, y sin la cual no habrían llegado a ser grandes ni famosos. Esta fórmula consiste en los siguientes cinco puntos:

a) **Dirigir el pensamiento hacia una meta fija** que se desea conseguir. Saber bien cuál es esa meta que se desea alcanzar y no desviar la atención de ella.

b) **Elaborar un plan para lograr conseguir esa meta**, un plan cuidadoso y detallado que se va siguiendo día por día, y que hace que nuestra actividad sea organizada y llena de entusiasmo.

c) **Desarrollar un sincero deseo de realizar aquello que se desea conseguir.** El deseo ardiente es el más importante motivador de las acciones. El deseo de lograr éxitos consigue la costumbre de conseguir éxitos.

d) **Adquirir una confianza grande en sí mismo**; confianza en las propias capacidades y habilidades para lograr el éxito, concediéndole muchísima mayor importancia a las cualidades positivas que se tiene que a las debilidades o a las posibilidades de derrota.

e) **Dedicarse a una acción tenaz e incansable para lograr obtener la meta que se busca conseguir**, sin desanimarse por los obstáculos, las críticas, las circunstancias adversas, o lo negativo que los demás piensen, hagan o digan. Esa energía concentrada hacia la consecución de una meta, trae enormemente las **oportunidades**, las cuales no se dejan atrapar por los que están sin hacer nada, pero se acercan generosamente a quienes se atreven a **atacar**, a trabajar fuertemente por conseguir el éxito.

Esta fórmula básica Meyer la llamó "El plan del éxito personal a base de automotivación", para desarrollar al máximo el potencial de cada uno.

Meyer resume la fórmula básica en la siguiente frase:

"Todo lo bueno que: vivamente imaginamos, ardientemente deseamos, sinceramente creamos, y entusiastamente emprendamos, de una manera impresionantemente favorable se transformará en algo placentero y beneficioso para nosotros"

(Eliécer Salesman. "100 Fórmulas para llegar al éxito")

Si una de tus metas es APRENDER aplica esta fórmula y "comienza con la mente abierta". La cualidad más importante que afectará tu éxito en el curso es tu ACTITUD. Ésta determinará lo que estés dispuesto a hacer en el curso, y la calidad de ese esfuerzo contribuirá de la manera más significativa a tu éxito.

Contenido

I **Documento introducción**

Análisis Inicial de los datos

II **Estudio de Caso.**

Encuesta Social General de 1994 (Análisis Inicial de los datos)

III **Estudio de Caso.**

Agua potable (Depuración de datos)

IV **Estudio de Caso.**

Caracterización del Mundo: Mundo 95 (Tablas y gráficos)

V **Estudio de Caso.**

Licencias Médicas (Rangos)

VI **Estudio de caso**

Premio Colegios (Sintaxis condicionales y estadística descriptiva)

VII **Estudio de caso**

Gendarmería (Agregación)

VIII **Estudio de caso**

Resultados de colegios (Fundición de archivos)

IX **Caso Encuesta laboral (Aplicación del IPC)**

I. Documento introducción :Análisis Inicial de los datos[1]

Cuando nos enfrentamos por primera vez a la realización de un análisis estadístico la máxima preocupación es profundizar en la técnica estadística seleccionada, sin .embargo, existe una etapa previa incluso más compleja y esencial que consiste en realizar un examen exhaustivo de los datos recabados.
La depuración de los datos o detección de problemas ocultos en los datos supondrá un gran avance en la consecución de resultados lógicos consistentes. Dichos problemas se pueden subsanar comenzando por una inspección visual de las representaciones gráficas de los datos, completándose con un análisis de datos ausentes o perdidos y de los casos atípicos (conocidos bajo la denominación de outliers).

Representaciones gráficas para el análisis de datos

La difusión experimentada en los últimos años por los programas estadísticos ha facilitando la incorporación de módulos específicamente diseñados para la inspección gráfica de los datos.

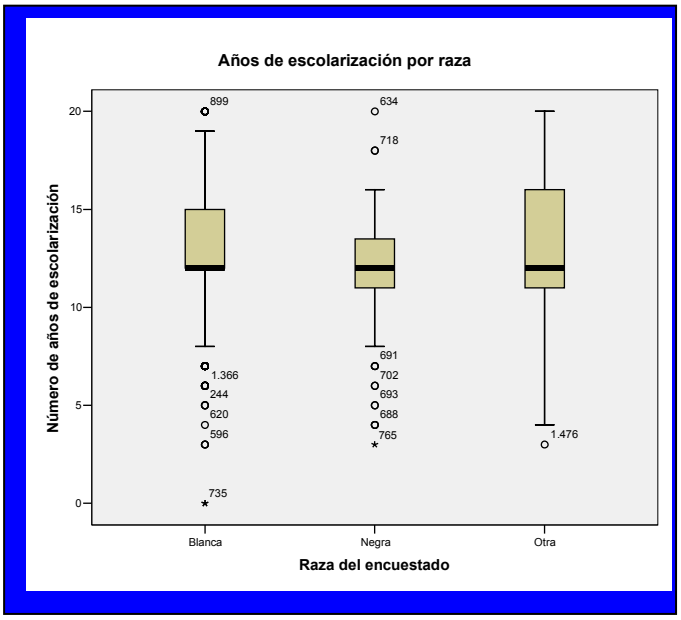
El estudio de cada variable es fundamental para conocer sus características y comprobar si es oportuna y relevante su inclusión en el análisis. Para ello se aconseja observar la forma de su distribución. Esto se consigue mediante el histograma, que representa gráficamente los datos mostrando en barras la frecuencia de los casos en cada variable. Si a su vez se pretende evaluar la normalidad de la variable, se efectuará superponiendo la curva normal sobre la distribución o realizando gráficos P-P o Q-Q.

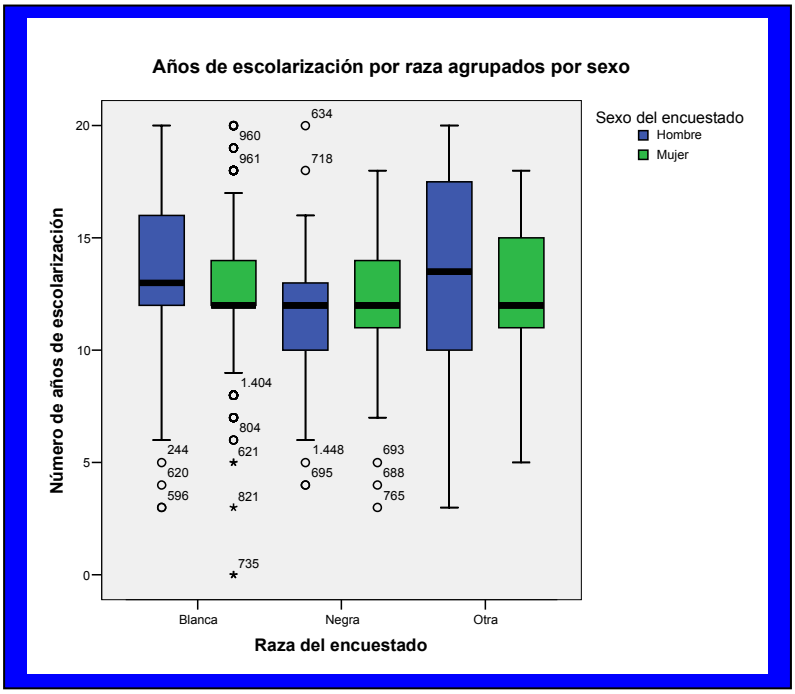
Mediante el gráfico de dispersión se podrá examinar la relación entre dos o más variables. Se trata de un gráfico de puntos de datos basados en dos variables, representadas una en el eje horizontal y la otra en el vertical. El posicionamiento de los puntos a lo largo de una línea recta se debe a la existencia de correlación lineal. Si los puntos siguen distintas formas la relación no podrá calificarse de lineal. La inexistencia de relación se podrá constatar si la nube de puntos es aleatoria y dispersa. (Mediante correlaciones bivariadas Pearson se podrá determinar mediante una prueba de hipótesis si la correlación entre dos variables de escala es significativa).

Mediante el gráfico de cajas o boxplot se puede llevar a cabo un análisis de las diferencias entre grupos, si lo que se pretende es apreciar la existencia de dos o más grupos en una variable métrica, como ocurre en el análisis discriminante o en el análisis de la varianza. Este gráfico distribuye los datos de tal forma que los límites superior e inferior de la caja marcan los cuartiles superior e inferior. La longitud de la caja es la distancia entre el primer y tercer cuartil; así, la caja contiene el 50 por ciento de los datos centrales de la distribución. La mediana se representa mediante una línea dentro de la caja. Existirá asimetría si la mediana se aproxima al final de la caja. El tamaño de la caja dependerá de la distancia entre las observaciones. También se representa la distancia entre la mayor y la menor de las observaciones mediante unas líneas que salen de la caja denominadas bigotes. En este tipo de gráfico los casos atípicos se pueden detectar por estar situados entre 1,0 Y 1,5 cuartiles fuera de la caja.

Diagrama de caja simple: Contiene un único diagrama de caja para cada categoría o variable del eje de categorías. Los diagramas de caja muestran la mediana, los cuartiles y los valores extremos para la categoría o variable.
Diagrama de caja agrupado: Tipo de gráfico en el que un grupo de diagramas de caja representa cada categoría o variable del eje de categorías. Los diagramas de caja dentro de cada agrupación vienen definidos por una variable de definición distinta.

[1] Análisis Estadístico Multivariable de Manuel Vivanco


Años de escolarización por raza
Número de años de escolarización
20
15
10
5
0
899
634
718
691
702
693
688
765
1.366
244
620
596
735
1.476
Blanca
Negra
Otra
Raza del encuestado


Años de escolarización por raza agrupados por sexo
Número de años de escolarización
20
15
10
5
0
634
960
961
718
1.404
804
621
1.448
693
695
688
765
244
620
596
821
735
Sexo del encuestado
Hombre
Mujer
Blanca
Negra
Otra
Raza del encuestado

Detección de variables con categorías mal codificadas

En muchos archivos de datos se detectan problemas en variables nominales con categorías en formato cadena sin un código asociado. Para detectar este problema es aconsejable realizar tablas de frecuencia de las variables y observar si las categorías presentan errores de digitación, como por ejemplo la variable sexo podría presentar problemas si las categorías están mal digitadas; Hombre, HOMBRE, hombre representan a la misma categoría, sin embargo en una tabla de frecuencia aparecerán como categorías diferentes. Para solucionar este problema se recomienda recodificar automáticamente asignándole a las categorías de la variable un código numérico y luego con recodificar en distinta variable asignar correctamente los códigos.

Análisis de datos ausentes

En este proceso de depuración de datos (anterior a la utilización de los métodos multivariables) el analista debe ser consciente de que se enfrenta a una información que puede no existir en determinadas observaciones y variables. Esto es lo que conocemos por datos ausentes o missing values. El porqué de la existencia de datos ausentes puede deberse a distintas razones como errores al codificar los datos e introducirlos en el computador, fallas del encuestador al completar el cuestionario, negación del encuestado a responder ciertas preguntas calificadas de controvertidas… Razones comunes y muy habituales en todo proceso investigador.

El problema de estos errores es el gran perjuicio que la inexistencia de datos ocasiona en los resultados y sus efectos en el tamaño de la muestra disponible para el análisis, dado que esta ausencia puede convertir lo que era una muestra adecuada en inadecuada. Por ello es necesario depurar esos casos y buscar soluciones. Si se puede suponer que los fundamentos teóricos de la investigación no se alteran sustancialmente, una opción sería suprimir aquellas variables y/o casos que peor se comportan respecto a los datos ausentes. En este caso el investigador deberá sopesar lo que gana con la exclusión de esta información y lo que pierde al no contar posteriormente en el análisis multivariante con la misma. Mediante este proceder se asegura de que su matriz de datos está completa y posee observaciones válidas.

Otra posibilidad sería la estimación de valores ausentes empleando relaciones conocidas entre valores válidos de otras variables y/o casos de la muestra. Por tanto, se trataría de imputar o sustituir los datos ausentes por valores estimados (bien sea la media o un valor constante) en base a otra información existente en la muestra.
Un porcentaje bajo de valores *missing* no es un problema que influya decisivamente en los resultados. Por el contrario, la falta reiterada de respuesta puede alterar seriamente el análisis. No existe una estimación respecto al porcentaje de *missing* que produce dificultades en una muestra determinada.
Según Tabachnik y Fidell (1983) más importante que el número de valores *missing* es la existencia de un patrón de comportamiento en éstos. En efecto, la presencia de *missing* que se distribuyen aleatoriamente no produce sesgos, sin embargo, la falta de respuesta sistemática asociada a ciertas variables puede generar distorsión en los resultados.

La existencia de datos ausentes nunca debe impedir la aplicación del análisis multivariable o limitar la posibilidad de generalizar los resultados de una investigación. La principal tarea del analista consistirá en identificar su presencia, y desempeñar las acciones necesarias para minimizar sus efectos.
En datos correspondientes a encuestas es habitual encontrar códigos como los siguientes.
7= No procede, 8= No sabe , 9= No contesta
97= No procede, 98= No sabe, 99= No contesta
997= No procede, 998= No sabe, 999= No contesta

Se utilizan estos códigos cuando no son parte de los posibles datos de la variable.
El SPSS tiene un menú especial para tratar los valores perdidos.
El SPSS hace diferencia para los valores perdidos por el usuario y valores perdidos por el sistema.

Detección de outliers

Al examinar los datos recabados después de un proceso muestral el investigador puede detectar la existencia de ciertas observaciones que no siguen el mismo comportamiento del resto, enfrentándose de este modo a ciertos casos que, por ser claramente diferentes de otras observaciones de la muestra, son calificados como outliers o atípicos.

El objetivo ante esta situación es identificar esa diferencia sustancial entre el valor real de la variable criterio y su valor previsto, puesto que da lugar a observaciones que no son representaciones apropiadas de la población de la cual se extrae la muestra.
Los casos atípicos se deben a errores en el procedimiento, o lo que es lo mismo, a falta al introducir los datos o al codificar. Pero también pueden ser consecuencia de un evento extraordinario que hace destacar esa observación. Este acontecimiento anormal puede tener o no una explicación. En cualquiera de estas situaciones, una vez que: los outliers el analista debe juzgar qué es lo más apropiado: si evaluar toda la incluyendo estas perturbaciones o eliminadas del análisis.

Estas decisiones han de justificarse, dado que determinados casos atípicos: aunque diferentes a la mayor parte de la muestra, pueden contener información representativa de un segmento dominante. No obstante, habrá situaciones donde lo más acertado sea su supresión porque pueden distorsionar seriamente los tests estadísticos dados los problemas que presentan.

La detección de los casos atípicos desde una perspectiva univariable pasa por la observación de aquellos casos que caigan fuera de los rangos de la distribución. Si lo que se pretende es evaluar conjuntamente pares de variables se utilizará el gráfico de dispersión. Este método bivariable permite identificar los casos atípicos al venir representado como puntos aislados. Por su parte, la detección multivariable supone evaluar cada observación a lo largo de un conjunto de variables. Esto se consigue mediante el uso de la Mahalanobis, puesto que es una medida de la distancia de cada observación en un espacio multidimensional respecto del centro medio de las observaciones.

II. Estudio de Caso[2].

Análisis Inicial de los datos (verificación): Encuesta Social General de 1994

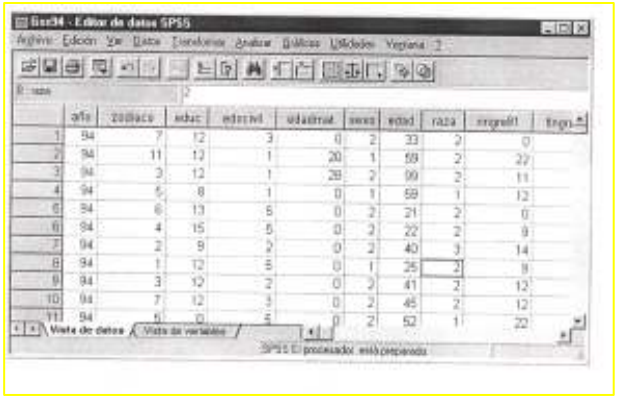
El objetivo de este estudio de caso es revisar la importancia de verificar los datos antes de realizar análisis formales y ver los métodos disponibles en SPSS.
Utilizar la ventana del Editor de Datos de SPSS y el procedimiento de Resúmenes de casos (Listados en SPSS 6.1) para ver los datos; verificar los valores fuera de rango mediante los procedimientos descriptivos; aplicar condicionales (Si) para evaluar la consistencia de las preguntas.

Archivo de trabajo: Encuesta Social General de 1994 (realizada por el National Opinion Research Center de Chicago).

Cuando trabajamos con datos es importante verificar su validez antes de proceder con el análisis. Las organizaciones de encuestas por correo aplican métodos como la verificación por
teclas, una técnica en la que dos personas independientes ingresan datos a una computadora, después de los cual los resultados son comparados y las diferencias resueltas. Las organizaciones de encuestas telefónicas suelen utilizar programas CATI (Computer Assisted Telephone Interview, en inglés) que comprueban la validez de una respuesta de manera automática (por ejemplo si la respuesta está dentro de un rango aceptable) y su consistencia con la información previa. Muchos investigadores no disponen de tales recursos, pero pueden implementar verificaciones simples de la validez y consistencia mediante SPSS. El tiempo que se invierte en examinar los datos de esta forma reducirá los falsos inicios y ahorrará trabajo. Por esta razón la verificación de los datos es un preludio crítico del análisis estadístico.

Revisión de algunos casos

Por lo general, el primer paso, aún antes de verificar los datos, es revisar las primeras observaciones y comparar sus valores con los originales de las hojas de datos o las formas de la encuesta. Esto sirve para detectar algunos errores en la definición de los datos (columnas incorrectas especificadas en el archivo ASCII de texto, lectura de caracteres alfa como campos numéricos). Ver los primeros datos se puede hacer de forma muy sencilla mediante la ventana del Editor de Datos de SPSS, mediante el procedimiento de Resúmenes de Casos, o el de Listar de las versiones anteriores (antes de SPSS 7). A continuación vemos una parte de los datos de la Encuesta social general de 1994 en SPSS.



[2]Documento de SPSS

Las primeras respuestas se pueden comparar con las hojas de datos originales o con las encuestas como una prueba primaria de la captura de los datos. Si se encuentran errores, se pueden hacer las correcciones de manera directa desde el Editor de Datos.

Note que limitamos el listado a los primeros 10 casos (la opción automática es de 100). El procedimiento de Resumir Casos también puede mostrar resúmenes de grupos.

	Años de escolaridad completados	Años de educación completados del esposo	Años de educación completados de la madre	Años de educación completados del padre
1	12	NAP	10	12
2	12	12	12	NAP
3	12	12	NS	NAP
4	8	9	5	NAP
5	13	NAP	12	NAP
6	15	NAP	20	NAP
7	9	NAP	NS	12
8	12	NAP	O	8
9	12	NAP	11	NAP
10	12	NAP	6	6
Total N	10	3	8	4

El programa mostrará, de manera automática, las etiquetas de valor de los casos en listados; esto se puede modificar dentro del cuadro de diálogo de Opciones de SPSS (haga clic en Edición... Opciones, luego muévase hacia el Navegador tab). Es importante señalar que en este caso la

elevada incidencia de "NAP" (no aplicable) en algunas variables (ver educación del padre) puede deberse al hecho de que se hicieron pocas preguntas a todos los sujetos en la Encuesta de 1994. Generalmente, la gran cantidad de datos perdidos (ver educación del padre) seria un elemento de preocupación.

Mínimo, máximo y número de casos válidos

Una segunda verificación sencilla de los datos que se puede hacer con SPSS es mediante las estadísticas descriptivas de todas las variables numéricas. El procedimiento Descriptivas reportará la media, la desviación típica, el mínimo, máximo y el número de casos válidos para cada variable numérica. La media y la desviación típica no son relevantes para las variables nominales (ver Capitulo 1), mientras que el mínimo y el máximo señalarán cualesquiera datos fuera de rango. Además, si el número de observaciones válidas es demasiado pequeño en una variable, se deberá explorar con mucho cuidado. Dado que la opción Descriptivos sólo proporciona estadísticos de resumen, no indicará cuáles observaciones contienen valores fuera de rango, pero éstas pueden ser fácilmente determinadas una vez que se conozca el valor de los datos.

El comando de sintaxis para Descriptivos que se presenta enseguida le pedirá al programa resúmenes para todas las variables (aunque éstos se registrarán sólo para las numéricas).

Sólo aparecerán variables numéricas en la lista del cuadro de diálogo. Para ejecutar el comando de sintaxis para Descriptivas sólo haga clic en el botón Aceptar para que el programa le muestre estos resúmenes.

	N	Mínimo	Máximo	Media	Desv. tip.
Año de encuesta	2992	94	94	94.00	.000
Signo zodiacal del respondiente	2971	1	12	6.56	3.408
Años de escolaridad completados	2985	O	20	13.16	2.971
Estado civil	2991	1	5	2.32	1.596
Edad cuando se caso por primera vez	1189	13	57	22.65	4.882
Sexo del respondiente	2992	1	2	1.57	.495
Edad del respondiente	2986	18	89	45.97	17.050
Raza del respondiente	2992	1	3	121	.497
Ingresos del respondiente	2075	1	22	12.91	5.518

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. típ.
Miembro de una grupo eclesiástico	509	1	2	1.67	.472
Miembro de algún otro grupo	507	1	2	1.89	.309
Cree en la existencia de Dios	1326	1	6	5.23	1.296
Años del actual matrimonio.	173	18	74	36.24	11.938
N válido (según lista)	0				

En este caso podemos ver el mínimo, máximo y el número de casos válidos para cada variable. Si examinamos variables como EDUC (grado máximo de estudios) y EDAD (años cumplidos) podemos determinar si existen valores fuera de rango. El máximo para HORASDETV parece elevado. Como ejercicio vamos a examinar las etiquetas (haga clic en Utilidades... Variables) para descubrir los rangos válidos de los valores de unas cuantas variables. Vamos a comparar éstos con los resultados que se muestran en la figura.

Para cada variable se enlista el número de observaciones válidas (N válida). Éste indica cuántas observaciones completaron los datos para todas las variables analizadas, la cual es una información útil. En este caso es cero debido a que no todas las preguntas de la GSS se aplicaron a los sujetos, ni todas fueron relevantes para ninguno. Si se descubren valores extraños en estos resúmenes, podemos ubicar las observaciones con problemas mediante la selección de datos de la función Encontrar (en el menú Edición) dentro del Editor de Datos de SPSS.

Identificación de respuestas inconsistentes

En la mayoría de las bases de datos, se deben mantener ciertas relaciones entre las variables, si es que los datos se capturaron correctamente. Esto resulta particularmente verdadero con las encuestas que contienen reactivos de filtro o de tamizaje. Algunos ejemplos de la GSS son: si un sujeto nunca ha estado casado, entonces el registro de su edad al momento de casarse debería ser un código perdido; en este caso el registro debiera ser el de la edad. Estas relaciones no se pueden verificar así de fácil revisando los datos o en un simple resumen de variable, como las tablas de frecuencias. Sin embargo, sí se pueden revisar mediante las instrucciones para transformar datos que sirven para identificar los casos que violan los patrones establecidos.

Vamos a demostrar cómo probar estas relaciones entre las variables utilizando dos ejemplos mencionados en el párrafo anterior. Primero, nunca casado (EDOCIVIL=5) y edad al momento de casarse (EDADMA T) deben estar codificadas como valores perdidos. Segundo, EDADMAT debe ser menor o igual a EDAD. El objetivo es crear una variable nueva en la que se establezca un valor determinado (digamos 1) si la relación esperada no se presenta. Esto se realiza mediante las opciones Transformar... Calcular y señalando la opción SI.

En el renglón de texto del cuadro de diálogo Calcular si indicamos la condición que queremos identificar: nunca casado (EDOCIVIL=5) y tener una edad válida en el primer matrimonio (-MISSING (EDADMAT) la tilde (-) significa "NO"). Si esta condición ocurre, se creará una variable nueva (EREDOCIV) igual a uno.
Transformar /Calcular/ variable destino eredociv/expresión numérica=1

Vemos que si se mantiene la condición de error, entonces la nueva variable EREDOCIV será igual a 1. Se puede ejecutar un análisis de frecuencias para determinar si algunos casos tienen esta variable con valor 1. Se pueden seleccionar y enlistar los casos con problemas o localizarlos con la función Encontrar en la ventana del Editor de Datos.

Se puede realizar la misma operación aplicando los comandos de sintaxis SI.

```
IF (EDOCIVIL=5 AND NOT (MISSING (EDADMAT)))
EREDOCIV=1.
IF (EDADMAT > EDAD) EREDAD=1).
FREQUENCIES NARIABLES EREDOCIV EREDAD.
```

El primer Si ejecuta la misma función que los cuadro de diálogo que acabamos de ver, dándole a EREDOCIV el valor 1 si un sujeto que nunca estuvo casado reportó edad al momento del primer matrimonio. El segundo Si le asigna el valor 1 a EREDAD si la edad al momento del primer matrimonio del sujeto excede su edad actual. Luego se hace un análisis de frecuencias para ver si ocurrieron estos errores.

Ya que los datos de la GSS están muy bien revisados, no es sorprendente que no se encuentren discrepancias (valores de error iguales a 1).

Cuando se descubren errores

Si se encuentran errores, el primer paso es regresar a la hoja de registro de datos o a los cuestionarios. Los errores sencillos se pueden corregir; en algunos casos se pueden corregir errores de un sujeto con base en sus respuestas a otras preguntas. Si no se puede hacer esto, entonces se pueden codificar esos reactivos como valores perdidos y se excluirán de los análisis. Es importante mencionar que la función Valores Perdidos de SPSS puede realizar esta tarea.

Otras herramientas muy útiles para limpiar los datos

Recodificar automáticamente:

El cuadro de diálogo Recodificación automática le permite convertir los valores numéricos y de cadena en valores enteros consecutivos. Si los códigos de la categoría no son secuenciales, las casillas vacías resultantes reducen el rendimiento e incrementan los requisitos de memoria de muchos procedimientos. Además, algunos procedimientos no pueden utilizar variables de cadena y otros requieren valores enteros consecutivos para los niveles de los factores.

- La nueva variable, o variables, creadas por la recodificación automática conservan todas las etiquetas de variable y de valor definidas de la variable antigua. Para los valores que no tienen una etiqueta de valor ya definida se utiliza el valor original como etiqueta del valor recodificado. Una tabla muestra los valores antiguos, los nuevos y las etiquetas de valor.
- Los valores de cadena se recodifican por orden alfabético, con las mayúsculas antes que las minúsculas.
- Los valores perdidos se recodifican como valores perdidos mayores que cualquier valor no perdido y conservando el orden. Por ejemplo, si la variable original posee 10 valores no perdidos, el valor perdido mínimo se recodificará como 11, y el valor 11 será un valor perdido para la nueva variable.

Recodificar en la misma variable /distinta variable

El cuadro de diálogo Recodificar en las mismas variables le permite reasignar los valores de las variables existentes o agrupar rangos de valores existentes en nuevos valores. Por ejemplo, podría agrupar los salarios en categorías que sean rangos de salarios.

El cuadro de diálogo Recodificar en distintas variables le permite reasignar los valores de las variables existentes o agrupar rangos de valores existentes en nuevos valores para una variable nueva.

- Puede recodificar variables numéricas en variables de cadena y viceversa.
- Si selecciona múltiples variables, todas deben ser del mismo tipo. No se pueden recodificar juntas las variables numéricas y de cadena.

Una vez que se han limpiado los datos podemos pasar a la parte más interesante del proceso, el análisis de datos.

III. Análisis inicial de datos: Agua Potable[3]

Considere el archivo "archivo APotable (errores).sav" correspondiente a una muestra aleatoria de hogares de la región Metropolitana que contiene el consumo de agua potable del mes de Enero del 2005. Realice un análisis inicial de los datos.

Solución: Análisis inicial de datos

a) Realizar una tabla para verificar información (para esto se debe tener la información original) Analizar/Informes/resúmenes de casos. Limitar los casos a los primeros 15. Todas las variables.

Resúmenes de casos[a]

	Número de caso	Identificador del Hogar	Identificador de la comuna donde se encuentra el hogar	Consumo de Agua Potable	Ingreso del Hogar	N° de Habitantes del Hogar	Comuna donde se encuentra el hogar	m3 libres (no construidos)	m3 edificados	Longitud del frente del terreno
1	1	10807	13101	233,80	618086	5	SANTIAGO	74,54	49,92	4,99
2	2	15565	13101	207,40	348340	5	SANTIAGO	63,41	54,22	5,42
3	3	11416	13101	183,00	335000	5	SANTIAGO	54,24	61,59	6,16
4	4	11358	13101	198,80	389295	5	SANTIAGO	63,79	67,67	6,77
5	5	4626	13101	233,00	182600	5	SANTIAGO	85,99	85,42	8,54
6	6	11492	13101	199,00	481151	5	SANTIAGO	70,91	88,74	8,87
7	7	16134	13102	208,40	150750	5	INDEPENDENCIA	58,93	39,36	3,94
8	8	9456	13102	141,36	242961	3	INDEPENDENCIA	57,00	68,08	6,81
9	9	10964	13102	170,88	182970	4	INDEPENDENCIA	61,58	70,28	7,03
10	10	11461	13102	186,20	452108	5	INDEPENDENCIA	62,11	80,61	8,06
11	11	3352	13102	288,72	384921	6	INDEPENDENCIA	105,82	95,94	9,59
12	12	6357	13102	249,84	263990	9	INDEPENDENCIA	60,97	108,81	10,88
13	13	11788	13103	210,40	266167	5	CONCHALI	62,02	45,79	4,58
14	14	13035	13103	220,80	215000	5	CONCHALI	71,26	58,64	5,86
15	15	14675	13103	194,00	428000	5	CONCHALI	60,92	65,89	6,59
Total N		15	15	15	15	15	15	15	15	15

a. Limitado a los primeros 15 casos.

ii) Realizar tablas de frecuencia para las variables nominales y ordinales que Ud desea analizar

[3] Ejercicio elaborado por Sara Arancibia

Comuna donde se encuentra el hogar

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	CERRILLOS	20	5,2	5,2	5,2
	CERRO NAVIA	8	2,1	2,1	7,3
	CONCHALI	6	1,6	1,6	8,8
	EL BOSQUE	17	4,4	4,4	13,2
	ESTACION CENTRAL	4	1,0	1,0	14,3
	HUECHURABA	16	4,2	4,2	18,4
	INDEPENDENCIA	6	1,6	1,6	20,0
	LA CISTERNA	17	4,4	4,4	24,4
	LA FLORIDA	9	2,3	2,3	26,8
	LA GRANJA	6	1,6	1,6	28,3
	LA PINTANA	3	,8	,8	29,1
	LA REINA	10	2,6	2,6	31,7
	Las CONDES	1	,3	,3	31,9
	LAS CONDES	17	4,4	4,4	36,4
	LO BARNECHEA	12	3,1	3,1	39,5
	LO ESPEJO	13	3,4	3,4	42,9
	LO PRADO	11	2,9	2,9	45,7
	MACUL	6	1,6	1,6	47,3
	Maipu	1	,3	,3	47,5
	MAIPU	23	6,0	6,0	53,5
	ÑUÑOA	5	1,3	1,3	54,8
	PEDRO AGUIRRE CERDA	15	3,9	3,9	58,7
	PENALOLEN	1	,3	,3	59,0
	PEÑALOLEN	16	4,2	4,2	63,1
	PROVIDENCIA	8	2,1	2,1	65,2
	PUDAHUEL	22	5,7	5,7	70,9
	QUILICURA	15	3,9	3,9	74,8
	QUINTA NORMAL	14	3,6	3,6	78,4
	RECOLETA	12	3,1	3,1	81,6
	RENCA	18	4,7	4,7	86,2
	SAN JOAQUIN	13	3,4	3,4	89,6
	SAN MIGUEL	7	1,8	1,8	91,4
	SAN RAMON	8	2,1	2,1	93,5
	SANTIAGO	6	1,6	1,6	95,1
	VITACURA	19	4,9	4,9	100,0
	Total	385	100,0	100,0	

De la tabla se observan dos errores con las comunas Maipú y las Condes. Este error se debe a que no se digitó con un código identificador. Es aconsejable asignar un código numérico. Para solucionar este problema se debe recodificar automáticamente y luego recodificar en la misma variable.

Transformar/recodificación automática/

Variable: comuna

Variable nueva : comurec

Añadir nuevo nombre

Recodificar empezando por primer valor

Aceptar

Se crea una nueva variable comurec con código numérico.

En utilidades variables se identifican los códigos de cada etiqueta correspondiendo

13 Las Condes

14 LAS CONDES

19 Maipu

20 MAIPU

Transformar/recodificar /en la misma variable

Considere la variable comurec

Valores antiguos y nuevos

Valor antiguo:13 Valor nuevo: 14 Añadir

Valor antiguo:19 Valor nuevo: 20 Añadir

Continuar aceptar

Vuelva a realizar tablas de frecuencias de comurec

Comuna donde se encuentra el hogar

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	CERRILLOS	20	5,2	5,2	5,2
	CERRO NAVIA	8	2,1	2,1	7,3
	CONCHALI	6	1,6	1,6	8,8
	EL BOSQUE	17	4,4	4,4	13,2
	ESTACION CENTRAL	4	1,0	1,0	14,3
	HUECHURABA	16	4,2	4,2	18,4
	INDEPENDENCIA	6	1,6	1,6	20,0
	LA CISTERNA	17	4,4	4,4	24,4
	LA FLORIDA	9	2,3	2,3	26,8
	LA GRANJA	6	1,6	1,6	28,3
	LA PINTANA	3	,8	,8	29,1
	LA REINA	10	2,6	2,6	31,7
	LAS CONDES	18	4,7	4,7	36,4
	LO BARNECHEA	12	3,1	3,1	39,5
	LO ESPEJO	13	3,4	3,4	42,9
	LO PRADO	11	2,9	2,9	45,7
	MACUL	6	1,6	1,6	47,3
	MAIPU	24	6,2	6,2	53,5
	ÑUÑOA	5	1,3	1,3	54,8
	PEDRO AGUIRRE CERDA	15	3,9	3,9	58,7
	PENALOLEN	1	,3	,3	59,0
	PEÑALOLEN	16	4,2	4,2	63,1
	PROVIDENCIA	8	2,1	2,1	65,2
	PUDAHUEL	22	5,7	5,7	70,9
	QUILICURA	15	3,9	3,9	74,8
	QUINTA NORMAL	14	3,6	3,6	78,4
	RECOLETA	12	3,1	3,1	81,6
	RENCA	18	4,7	4,7	86,2
	SAN JOAQUIN	13	3,4	3,4	89,6
	SAN MIGUEL	7	1,8	1,8	91,4
	SAN RAMON	8	2,1	2,1	93,5
	SANTIAGO	6	1,6	1,6	95,1
	VITACURA	19	4,9	4,9	100,0
	Total	385	100,0	100,0	

iii) Realizar tablas con valores extremos y diagramas de caja. Esto nos permitirá verificar si los casos los valores atípicos existen o han sido mal ingresados.

Valores extremos

			Número del caso	Valor
Consumo de Agua Potable	Mayores	1	378	901,60
		2	370	571,68
		3	385	560,64
		4	46	513,48
		5	375	505,60
	Menores	1	238	125,04
		2	205	129,48
		3	330	129,76
		4	314	129,84
		5	290	135,52

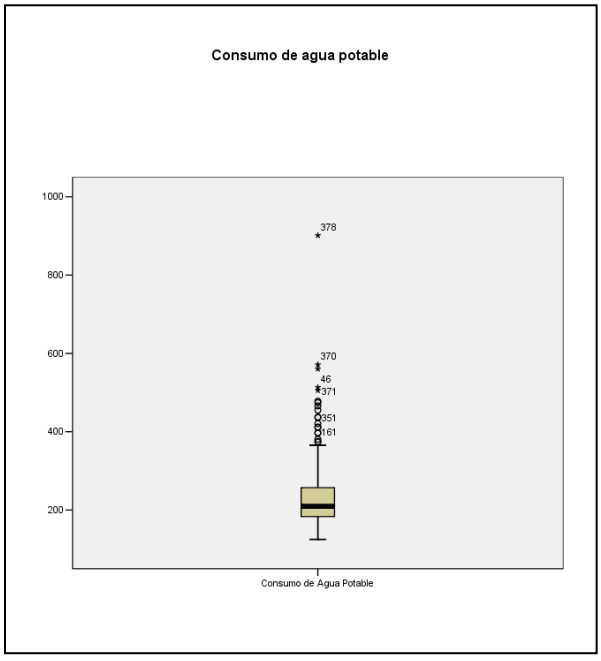
Valores extremos

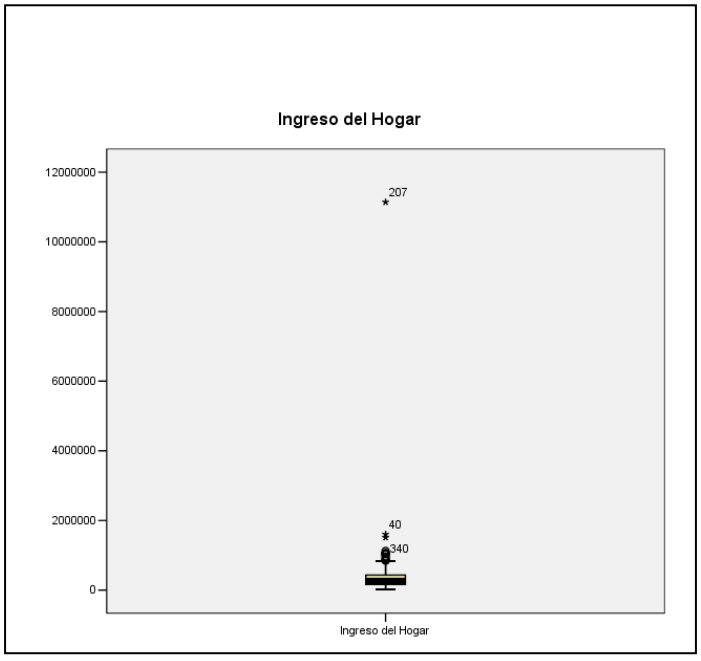
			Número del caso	Valor
Ingreso del Hoga	Mayores	1	207	1141600
		2	40	1602365
		3	146	1512608
		4	342	1126072
		5	372	1060366
	Menores	1	72	18260
		2	104	20000
		3	210	22825
		4	45	22825
		5	341	33044

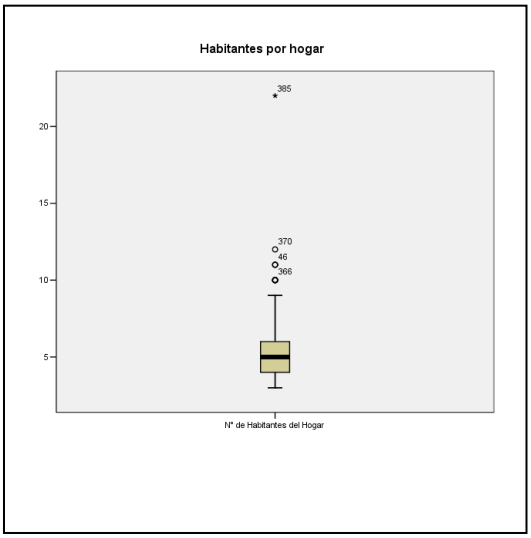
Valores extremos

			Número del caso	Valor
N° de Habitantes del Hogar	Mayores	1	385	22
		2	370	12
		3	46	11
		4	376	11
		5	378	11
	Menores	1	338	3
		2	332	3
		3	314	3
		4	303	3
		5	282	3[a]

a. En la tabla de valores extremos menores sólo se







Menu Analizar /Explorar/

Variables: Consumo agua potable, Ingreso del hogar, y No habitantes del hogar

Estadísticos/ Valores atípicos

Gráficos Diagrama de caja para cada una de las variables mencionadas.

Para cada variable se debería verificar si la información de los valores atípicos está bien registrada. Corregir en el caso que sea posible o filtrar los casos muy extremos para no sesgar los análisis.

iv) Cuando existen columnas (variables como identificador de otra variable) como el caso de la variable identificador de comuna y comurec, se debe verificar si se corresponden. Para esto puede ordenar id-comuna en forma ascendente y ver si se corresponde visualmente con la comurec.
Otra forma es crear una variable de cadena donde concatene los dos codigos (correspondiente a id-comuna y comurec). Esto le permitirá ver en una tabla de frecuencia si las variables se corresponden. En el ejemplo hay dos códigos que se corresponden con 13107 lo cual acusa error.
13107 y 23 13107 y 24 Se debe corregir.

Sintaxis del procedimiento

```
STRING concat (A13).
COMPUTE concat = CONCAT(STRING(id_comun,F11.0),STRING(comurec,F2.0)) .
EXECUTE .
```

concat

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	1310134	6	1,6	1,6	1,6
	13102 7	6	1,6	1,6	3,1
	13103 3	6	1,6	1,6	4,7
	13104 5	4	1,0	1,0	5,7
	1310529	12	3,1	3,1	8,8
	1310625	8	2,1	2,1	10,9
	1310723	1	,3	,3	11,2
	1310724	16	4,2	4,2	15,3
	1310815	12	3,1	3,1	18,4
	13109 2	8	2,1	2,1	20,5
	1311021	5	1,3	1,3	21,8
	1311112	10	2,6	2,6	24,4
	1311218	6	1,6	1,6	26,0
	13113 8	17	4,4	4,4	30,4
	13114 9	9	2,3	2,3	32,7
	1311531	14	3,6	3,6	36,4
	1311610	5	1,3	1,3	37,7
	1311711	3	,8	,8	38,4
	1311833	8	2,1	2,1	40,5
	1311932	7	1,8	1,8	42,3
	1312035	19	4,9	4,9	47,3
	13121 4	17	4,4	4,4	51,7
	1312222	15	3,9	3,9	55,6
	1312316	13	3,4	3,4	59,0
	13124 6	16	4,2	4,2	63,1
	13125 1	20	5,2	5,2	68,3
	1312620	24	6,2	6,2	74,5
	1312728	14	3,6	3,6	78,2
	1312817	11	2,9	2,9	81,0
	1312926	22	5,7	5,7	86,8
	1313014	18	4,7	4,7	91,4
	1313130	18	4,7	4,7	96,1
	1313227	15	3,9	3,9	100,0
	Total	385	100,0	100,0	

IV. Estudio de Caso: Caracterización del Mundo[4]

Considere el archivo Mundo 95, que contiene las siguientes variables de los países del Mundo en el año 1995:

Variable	Etiqueta	Etiqueta de Valor
país	País	
poblac	Población x 1000	
densidad	Habitantes x Km2	
urbana	Habitantes en ciudades (%)	
relig	Religión mayoritaria	
espvidaf	Esperanza de vida Femenina	
espvidam	Esperanza de vida Masculina	
alfabet	Alfabetización (%)	
inc_pob	Aumento de población (% anual)	
mortinf	Mortalidad infantil (Muertes por 1000 nacimientos vivos)	
pib_cap	Producto interno bruto per cápita	
región	Región Económica	1 = OCDE 2 = Europa Oriental 3 = Asia / Pacífico 4 = Africa 5 = Oriente Medio 6 = América Latina
calorías	Ingesta diaria de calorías	
sida	Casos de SIDA	
tasa_nat	Tasa de natalidad (por 1.000 habitantes)	
tasa_mor	Tasa de mortalidad (por 1.000 habitantes)	
tasasida	Casos de SIDA por 100.000 habitantes	
log_pib	Log(10) de PIB_CAP	
logtsida	Log(10) de TASASIDA	
nac_def	Tasa nacimentos/defunciones	
fertilid	Número promedio de hijos	
log_pob	Log(10) de POBLAC	
cregrano	--	
alfabmas	Hombres alfabetizados (%)	
alfabfem	Mujeres alfabetizadas (%)	
clima	Clima predominante	1 = Desierto 2 = Arido / Desierto 3 = Arido 5 = Tropical 6 = Mediterráneo 7 = Marítimo 8 = Templado 9 = Artico /

[4] Caso desarrollado por Sara Arancibia

		Templado 10 = Artico

Usted debe realizar un informe donde compare los países en al menos los siguientes aspectos: Población, densidad, % de habitantes en ciudades, esperanza de vida, alfabetización (%), tasas de natalidad y mortalidad, número promedio de hijos por familia, tasa sida, considerando las variables nominales Región, Religión mayoritaria y clima predominante.

Para su informe debe considerar al menos los siguientes puntos:

(i) Tres gráficos distintos con su interpretación.
(ii) Tablas de frecuencia
(iii) Tablas de contingencia
(iv) Outliers (Valores extremos)
(v) Medidas de tendencia central
(vi) Medidas de dispersión
(vii) Cubos OLAP
(viii) Puntuaciones z

Solución:

Comenzaremos el estudio determinando la frecuencia de las variables nominales; Región Económica, Religión Predominante y Clima Predominante de los países

Region or economic group

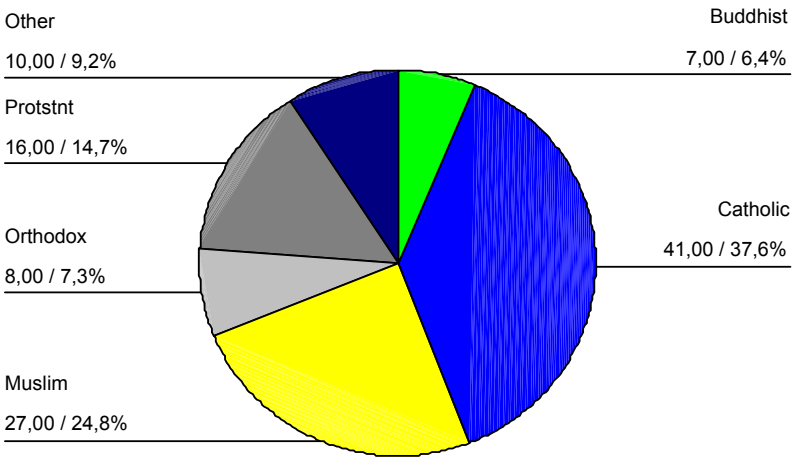
		Frequency	Percent	Valid Percent
Valid	OECD	21	19,3	19,3
	East Europe	14	12,8	12,8
	Pacific/Asia	17	15,6	15,6
	Africa	19	17,4	17,4
	Middle East	17	15,6	15,6
	Latn America	21	19,3	19,3
	Total	109	100,0	100,0

La tabla de frecuencia muestra el número de países por Región económica. Se observan dos regiones con la mayor frecuencia, (21 países) las que corresponden a la Región OECD (Organización para la Cooperación y el Desarrollo Económico) y a la Región de Latino América, correspondiendo al 19,3% del total de países. La menor frecuencia se observa en Europa del Este con 14 países de un total de 109 países.

El gráfico siguiente muestra la frecuencia y porcentaje de países por Religión predominante.

Frecuencia y porcentaje de países

por Religión Predominante



Se observa que 41 países que representan el 37,6% del total de países considerados tienen como religión predominante a la religión Católica y 27 países a la religión Musulmana representando el 24,8% del total de países considerados.

Para generar el gráfico: Graficar/Sectores/Resumen para grupos de casos/N° de casos/Religión Predominante. En el editor de gráficos se pide texto, valor y porcentaje y se colapsa los sectores a mayores del 5%.

La tabla de frecuencia para religión predominante muestra complementariamente al gráfico anterior que las religiones con menor frecuencia son las religiones Hindú, Judía, Taoísta y Tribal

Predominant religion

		Frequency	Percent	Valid Percent
Valid	Animist	4	3,7	3,7
	Buddhist	7	6,4	6,5
	Catholic	41	37,6	38,0
	Hindu	1	,9	,9
	Jewish	1	,9	,9
	Muslim	27	24,8	25,0
	Orthodox	8	7,3	7,4
	Protstnt	16	14,7	14,8
	Taoist	2	1,8	1,9
	Tribal	1	,9	,9
	Total	108	99,1	100,0
Missing		1	,9	
Total		109	100,0	

Al cruzar las variables región y religión podemos observar en la tabla de contingencia que la Religión Predominante Animista pertenece a países de África. La religión predominante Católica se encuentra en todas las regiones excepto en la Región de Oriente donde la religión predominante es la Musulmana con 15 países de un total de 17 países de la región

Predominant religion * Region or economic group Crosstabulation

Count

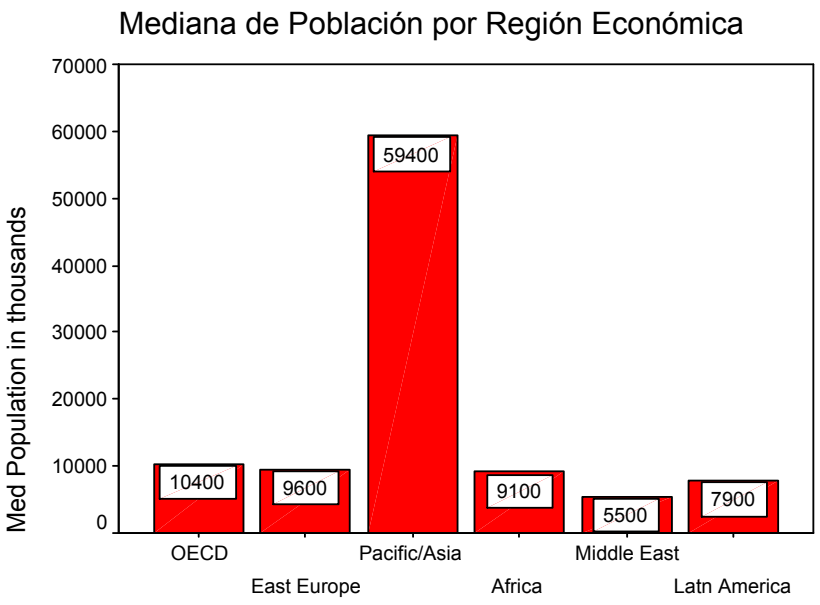
		Region or economic group						
		OECD	East Europe	Pacific/ Asia	Africa	Middle East	Latn America	Total
Predominant religion	Animist				4			4
	Buddhist			7				7
	Catholic	10	5	1	5		20	41
	Hindu			1				1
	Jewish					1		1
	Muslim		1	5	6	15		27
	Orthodox	1	6			1		8
	Protstnt	10	2	1	2		1	16
	Taoist			2				2
	Tribal				1			1
Total		21	14	17	18	17	21	108

En relación al Clima Predominante se observa que las mayores frecuencias corresponden a los climas Temperado y Tropical los que representan un 31,8% y 29,9% respectivamente, respecto al total de datos válidos.

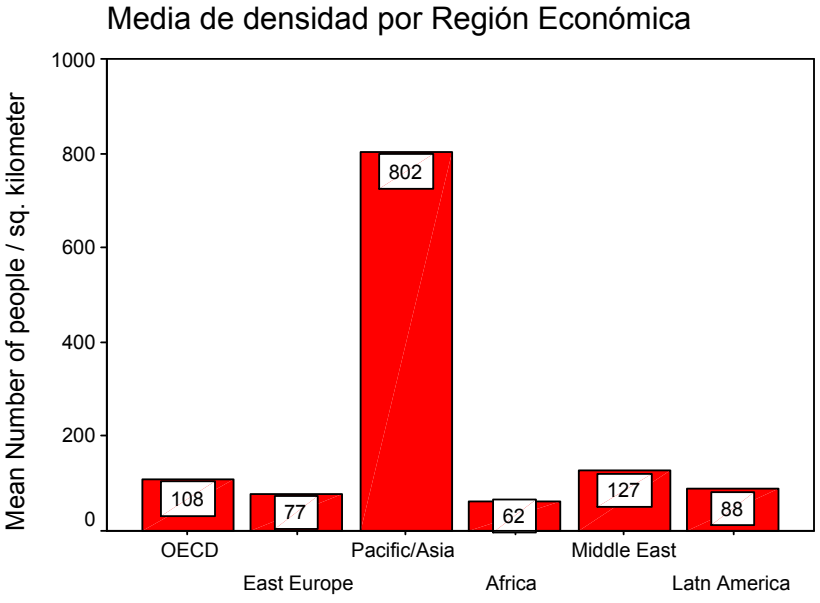
Predominant climate

		Frequency	Percent	Valid Percent
Valid	desert	7	6,4	6,5
	arid / desert	5	4,6	4,7
	arid	6	5,5	5,6
	otro	5	4,6	4,7
	tropical	32	29,4	29,9
	mediterranean	10	9,2	9,3
	maritime	4	3,7	3,7
	temperate	34	31,2	31,8
	arctic / temp	4	3,7	3,7
	Total	107	98,2	100,0
Missing	System	2	1,8	
Total		109	100,0	

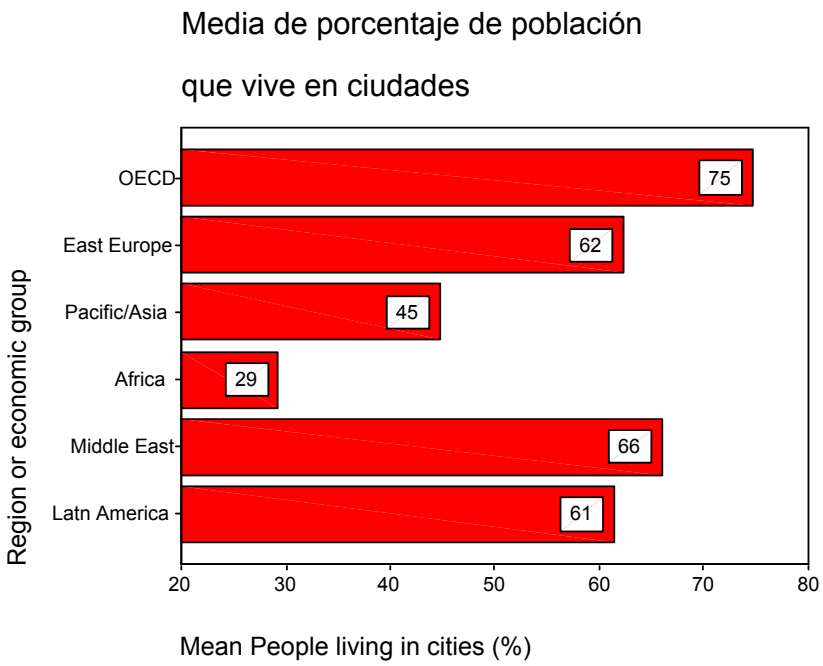
Ahora consideremos la población, densidad y habitantes que viven en ciudades. Podemos observar del gráfico correspondiente a la mediana de población por región económica que el 50% de los países del Asia/Pacífico tienen una población mayor o igual a 59.400.000 habitantes, valor notablemente alto en relación a las medianas del resto de las regiones las que oscilan entre 10.400.000 y 5.500.000 habitantes.



Coherente con lo anterior se observa que la mayor densidad por región económica corresponde a la región Asia/Pacífico con un valor promedio de 802 habitantes por kmP^{2P}, la que es considerablemente superior a la densidad promedio del resto de regiones, las que oscilan entre 127 y 62 habitantes por kmP^{2P} correspondiendo esta última a la región de África



Para generar el gráfico: Graficar/Barras/Simples/Resumen para grupos de casos/N de casos/ Región Económica.



En relación al porcentaje de personas que viven en ciudades, se observa del gráfico que el mayor porcentaje promedio corresponde a la Región OECD, con un 75% en promedio. Es considerable la diferencia con la región de África donde el promedio de población urbana es del 29%, seguido de Asia/Pacífico con un promedio del 45%.

Podemos complementar la información anterior con Cubos OLAP, los que muestran por grupos, los estadísticos que se necesiten conocer. Específicamente los Cubos siguientes muestran para las regiones OECD y África, el número de países el que corresponde a 21 y 18 países respectivamente. Se observa la media para cada una de las variables consideradas y la desviación estándar que muestra cuánto se desvían los datos, en promedio respecto a la media.

Al considerar el cubo correspondiente a la región OECD se observan los valores mínimo y máximo, es sorprendente observar que existen países con una densidad de 2,3 habitantes por kmP^{2P} y de 366 personas por kmP^{2P}. Al considerar la población, dentro de los países del OECD se puede apreciar un valor mínimo de 263.000 habitantes en oposición al valor máximo de 260.800.000 habitantes. El mayor porcentaje de población urbana corresponde al 96% y el menor corresponde al 34%.

OLAP Cubes

Region or economic group: OECD
Predominant climate: Total
Predominant religion: Total

	Population in thousands	Number of people / sq. kilometer	People living in cities (%)
N	21	21	21
Mean	33085,10	107,981	74,71
Std. Deviation	57148,25	107,936	14,89
Minimum	263	2,3	34
Maximum	260800	366,0	96
Median	10400,00	80,000	77,00

OLAP Cubes

Region or economic group: Africa
Predominant climate: Total
Predominant religion: Total

	Population in thousands	Number of people / sq. kilometer	People living in cities (%)
N	18	18	18
Mean	18415,83	63,700	28,17
Std. Deviation	24331,33	79,823	14,70
Minimum	959	2,4	5
Maximum	98100	311,0	47
Median	8900,00	39,500	24,50

Al considerar el cubo correspondiente a la región de África se observa una media de población considerablemente más baja que la media de la Región OECD y que la variabilidad en la variable población del 132% es más baja que si se compara con la región del OECD cuyo coeficiente de variabilidad es del 172,7%. Por otra parte se observa para la población urbana un mínimo de 5% siendo el porcentaje máximo del 47%, valores muy bajos si se compara con la región del OECD. Al igual que la región OECD se observa un valor mínimo de densidad de 2,4 habitantes por km^2, en oposición al máximo cuya densidad es de 311 habitantes por km^2.

Para identificar a qué países corresponden estos valores máximos y mínimos se puede solicitar los valores extremos (outliers) que muestra los cinco valores mayores y menores.

Extreme Values

Region or economic group: OECD
Number of people / sq. kilometer

		Case Number	COUNTRY	Value
Highest	1	70	Netherlands	366,0
	2	11	Belgium	329,0
	3	101	UK	237,0
	4	42	Germany	227,0
	5	56	Italy	188,0
Lowest	1	4	Australia	2,3
	2	49	Iceland	2,5
	3	21	Canada	2,8
	4	74	Norway	11,0
	5	71	New Zealand	13,0

Extreme Values

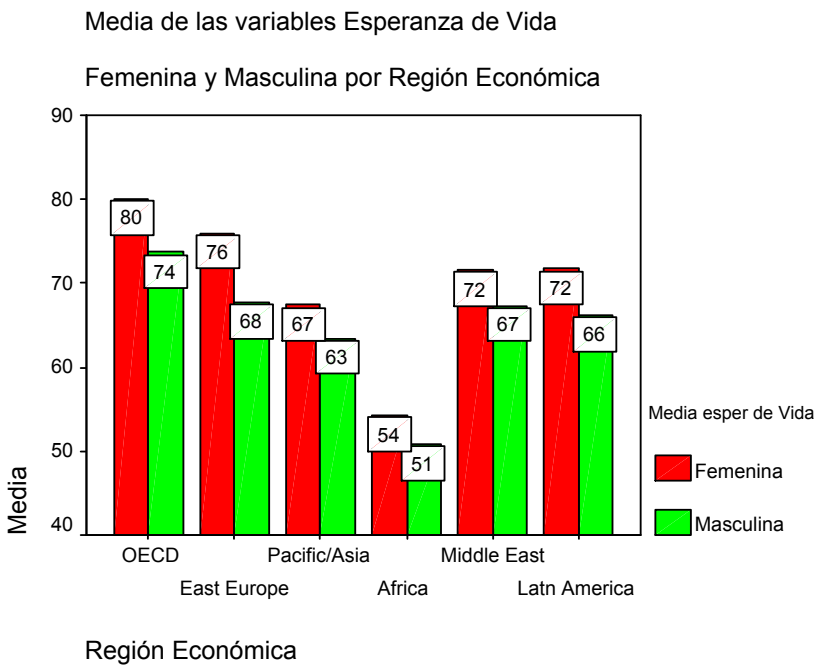
Region or economic group: Africa
Number of people / sq. kilometer

		Case Number	COUNTRY	Value
Highest	1	85	Rwanda	311,0
	2	18	Burundi	216,0
	3	73	Nigeria	102,0
	4	40	Gambia	86,0
	5	103	Uganda	76,0
Lowest	1	14	Botswana	2,4
	2	39	Gabon	4,2
	3	22	Cent. Afri.R	5,0
	4	90	Somalia	10,0
	5	109	Zambia	11,0

Para generar las tablas: Analizar/Estadísticos Descriptivos/Explorar. En Factor colocar Región económica y etiquetar por país. En Estadísticos seleccionar Valores Atípicos.

Ahora consideraremos las variables; Esperanza de vida femenina, esperanza de vida masculina, tasa de natalidad, tasa de mortalidad, tasa sida, fertilidad y alfabetización

El gráfico siguiente muestra la media de esperanza de vida femenina y masculina por Región Económica. Se observa que en todas las regiones es mayor la media de esperanza de vida femenina que masculina siendo la región del OECD, la de mayor esperanza de vida, con un promedio de 80 y 74 años para mujeres y hombres respectivamente. Es notable la diferencia con África donde se observa que el promedio de esperanza de vida es muy baja siendo la media de 54 y 51 años para mujeres y hombres respectivamente.



Para generar el gráfico: Graficar/Barras/Agrupados. Resumen para variables individuales/Media de las variables Esperanza de vida fem y masculina/eje de categorías Región Económica.

La tabla siguiente identifica los países con mayor y menor esperanza de vida

Valores Extremos (Outliers considerando todos los países)

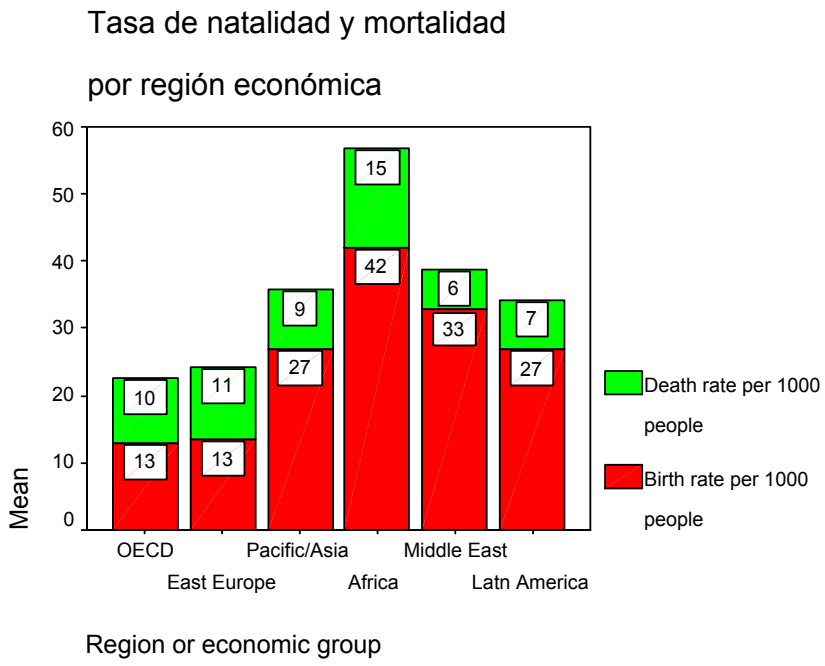
			Case Number	COUNTRY	Value
Average female life expectancy	Highest	1	94	Switzerland	82
		2	57	Japan	82
		3	38	France	82
		4	21	Canada	81
		5	56	Italy	,[a]
	Lowest	1	103	Uganda	43
		2	1	Afghanistan	44
		3	22	Cent. Afri.R	44
		4	109	Zambia	45
		5	97	Tanzania	45
Average male life expectancy	Highest	1	55	Israel	76
		2	57	Japan	76
		3	26	Costa Rica	76
		4	49	Iceland	76
		5	47	Hong Kong	,[b]
	Lowest	1	103	Uganda	41
		2	97	Tanzania	41
		3	22	Cent. Afri.R	41
		4	85	Rwanda	43
		5	45	Haiti	43

a. Only a partial list of cases with the value 81 are shown in the table of upper extremes.

b. Only a partial list of cases with the value 75 are shown in the table of upper extremes.

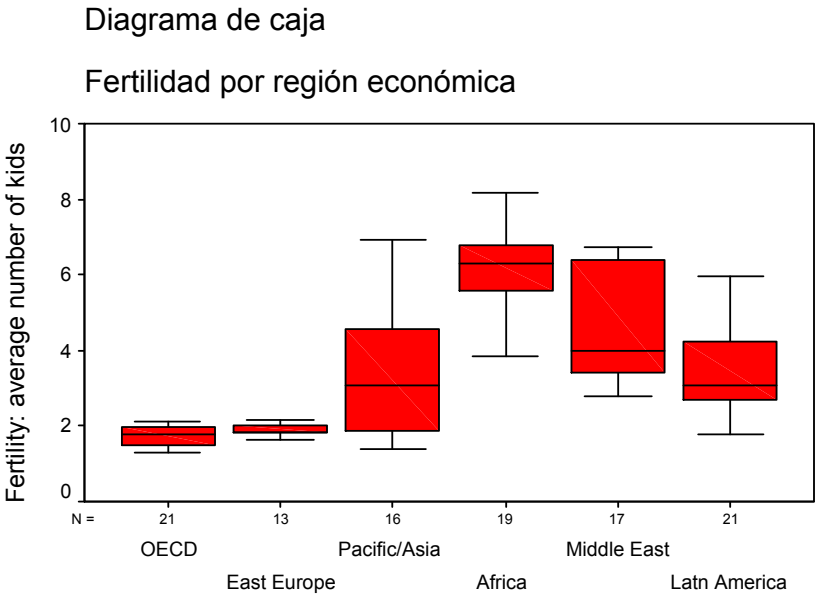
El siguiente gráfico apilado compara la tasa de natalidad y mortalidad por región económica, mostrando que las mayores tasas corresponden a la región de África, las que indican que en promedio

nacen 42 por cada 1.000 habitantes y mueren en promedio 15 por cada 1.000 habitantes. La menor tasa de natalidad en promedio corresponde a la región del OECD



Esta información está muy de acuerdo con la variable fertilidad, la que indica el promedio de hijos por familia.

El diagrama de caja muestra por región económica que las mayores tasas de fertilidad se concentran en la Región de África mostrando que la mediana representada por la línea horizontal en las cajas se aproxima al valor 6 hijos por familia en promedio. El 50% de los datos de fertilidad de los países se encuentra en la caja la que va desde el primer cuartil al tercer cuartil. La tabla de descriptivos para fertilidad por región confirma la información entregada por el diagrama de caja.



Descriptives

Fertility: average number of kids

	Region or economic group					
	OECD	East	Pacific/As	Africa	Middle	Latn
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic
Mean	1,746	1,889	3,383	6,081	4,724	3,336
5% Trimmed Mean	1,751	1,886	3,298	6,088	4,721	3,280
Median	1,800	1,840	3,065	6,290	4,000	3,080
Variance	6,150E-02	1,772E-02	3,226	1,285	2,356	1,115
Std. Deviation	,248	,133	1,796	1,134	1,535	1,056
Minimum	1,3	1,7	1,4	3,8	2,8	1,8
Maximum	2,1	2,2	6,9	8,2	6,7	5,9
Range	,8	,5	5,5	4,4	3,9	4,2
Interquartile Range	,495	,190	2,795	1,380	3,165	1,655
Skewness	-,081	,530	,791	-,586	,231	,827
Kurtosis	-1,192	,932	-,563	,119	-1,812	,332

Como complemento de la información vemos que la tabla siguiente muestra las medidas de tendencia central y de dispersión para todas las variables consideradas en este apartado.

Descriptives

	Average female life	Average male life	Birth rate per 1000	Death rate per 1000	Numbe r of aids cases /	Fertility: average number	People who read
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic
Mean	69,89	64,71	26,154	9,64	24,8271	3,558	77,95
5% Trimmed Mean	70,67	65,37	25,754	9,31	16,8072	3,475	79,74
Median	74,00	67,00	25,500	9,00	5,5512	3,065	87,50
Variance	115,241	88,926	154,112	18,400	2482,6	3,605	532,862
Std. Deviation	10,74	9,43	12,414	4,29	49,8252	1,899	23,08
Minimum	43	41	10,0	2	,00	1,3	18
Maximum	82	76	53,0	24	326,75	8,2	100
Range	39	35	43,0	22	326,75	6,9	82
Interquartile Range	12,75	12,75	21,000	4,00	23,2434	3,170	36,75
Skewness	-1,048	-1,020	,416	1,283	3,498	,665	-,955
Kurtosis	,054	,171	-1,163	1,754	15,008	-,933	-,250

Si consideramos sólo los países de las regiones OECD y África, observamos cómo cambian las medidas de tendencia central y dispersión ya que en todas las variables, los países de la región de África están con índice muy por debajo de los de la región OECD. Si queremos reconocer qué países en esas regiones tienen los cinco valores máximos y mínimos los podemos apreciar de la tabla de valores extremos.

Descriptives

Region or economic group: OECD

	Average female	Average male life	Birth rate per 1000	Death rate per	Number of aids	Fertility: average	People who read
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic
Mean	80,10	73,71	12,952	9,63	29,1052	1,746	97,67
5% Trimmed Mean	80,11	73,74	12,944	9,65	23,6322	1,751	98,22
Median	80,00	74,00	13,000	10,00	15,8713	1,800	99,00
Variance	1,390	1,314	2,748	1,633	1131,049	6,150E-02	11,333
Std. Deviation	1,18	1,15	1,658	1,28	33,6311	,248	3,37
Minimum	78	71	10,0	7	3,10	1,3	85
Maximum	82	76	16,0	12	157,94	2,1	100
Range	4	5	6,0	5	154,84	,8	15
Interquartile Range	2,00	1,50	2,000	2,00	24,2397	,495	2,00
Skewness	-,201	-,256	,302	-,169	3,090	-,081	-3,027
Kurtosis	-,827	,519	-,512	-,492	11,201	-1,192	10,370

Descriptives

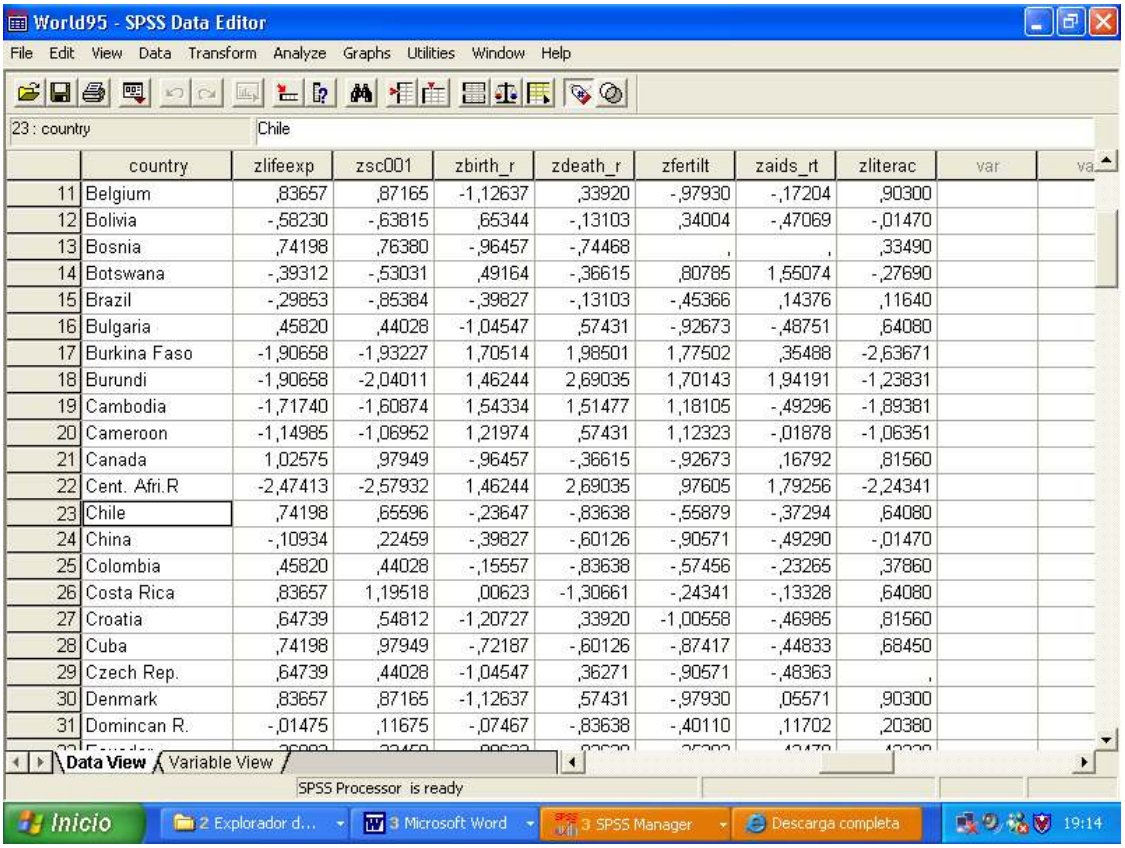
Region or economic group: Africa

	Average female	Average male life	Birth rate per 1000	Death rate per	Number of aids	Fertility: average	People who read
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic
Mean	54,26	50,79	42,000	14,74	75,7491	6,081	47,26
5% Trimmed Mean	54,01	50,49	42,389	14,71	66,0056	6,088	47,29
Median	55,00	51,00	44,000	14,00	36,3077	6,290	50,00
Variance	63,649	52,731	41,111	25,538	7641,570	1,285	319,094
Std. Deviation	7,98	7,26	6,412	5,05	87,4161	1,134	17,86
Minimum	43	41	28,0	6	,13	3,8	18
Maximum	70	66	49,0	24	326,75	8,2	76
Range	27	25	21,0	18	326,61	4,4	58
Interquartile Range	12,00	11,00	5,000	7,00	112,6254	1,380	34,00
Skewness	,425	,352	-1,256	,126	1,562	-,586	,012
Kurtosis	-,434	-,458	,452	-,847	2,587	,119	-,964

Para generar la tabla: Analizar/Estadísticos Descriptivos/Explorar: esperanza de vida femenina y masculina, tasa de natalidad y mortalidad, promedio de hijos por familia etc. Factor: Región. Etiquetar por: país. Estadísticos: Valores Atípicos. Al editar la gráfica se borra lo que no se quiere mostrar.

Si queremos comparar Chile en esperanza de vida femenina y masculina, tasa de natalidad, tasa de mortalidad, fertilidad, tasa sida y alfabetización respecto al resto de países de la base de datos consideramos las puntuaciones z de cada una de ellas, las que nos muestran que:

- esperanza de vida femenina en Chile está sobre la media en 0,74 desviaciones estándares.
- esperanza de vida masculina en Chile está sobre la media en 0,65 desviaciones estándares.
- tasa de natalidad en Chile está bajo la media en 0,23 desviaciones estándares.
- tasa de mortalidad en Chile está bajo la media en 0,83 desviaciones estándares.
- fertilidad (promedio de hijos por familia) en Chile está bajo la media en 0,55 desviaciones estándares.
- tasa sida en Chile está bajo la media en 0,37 desviaciones estándares
- alfabetización (% de personas que saben leer) en Chile está sobre la media en 0,64 desviaciones estándares.



	country	zlifeexp	zsc001	zbirth_r	zdeath_r	zfertilt	zaids_rt	zliterac	var	va
11	Belgium	,83657	,87165	-1,12637	,33920	-,97930	-,17204	,90300		
12	Bolivia	-,58230	-,63815	,65344	-,13103	,34004	-,47069	-,01470		
13	Bosnia	,74198	,76380	-,96457	-,74468	,	,	,33490		
14	Botswana	-,39312	-,53031	,49164	-,36615	,80785	1,55074	-,27690		
15	Brazil	-,29853	-,85384	-,39827	-,13103	-,45366	,14376	,11640		
16	Bulgaria	,45820	,44028	-1,04547	,57431	-,92673	-,48751	,64080		
17	Burkina Faso	-1,90658	-1,93227	1,70514	1,98501	1,77502	,35488	-2,63671		
18	Burundi	-1,90658	-2,04011	1,46244	2,69035	1,70143	1,94191	-1,23831		
19	Cambodia	-1,71740	-1,60874	1,54334	1,51477	1,18105	-,49296	-1,89381		
20	Cameroon	-1,14985	-1,06952	1,21974	,57431	1,12323	-,01878	-1,06351		
21	Canada	1,02575	,97949	-,96457	-,36615	-,92673	,16792	,81560		
22	Cent. Afri.R	-2,47413	-2,57932	1,46244	2,69035	,97605	1,79256	-2,24341		
23	Chile	,74198	,65596	-,23647	-,83638	-,55879	-,37294	,64080		
24	China	-,10934	,22459	-,39827	-,60126	-,90571	-,49290	-,01470		
25	Colombia	,45820	,44028	-,15557	-,83638	-,57456	-,23265	,37860		
26	Costa Rica	,83657	1,19518	,00623	-1,30661	-,24341	-,13328	,64080		
27	Croatia	,64739	,54812	-1,20727	,33920	-1,00558	-,46985	,81560		
28	Cuba	,74198	,97949	-,72187	-,60126	-,87417	-,44833	,68450		
29	Czech Rep.	,64739	,44028	-1,04547	,36271	-,90571	-,48363	,		
30	Denmark	,83657	,87165	-1,12637	,57431	-,97930	,05571	,90300		
31	Domincan R.	-,01475	,11675	-,07467	-,83638	-,40110	,11702	,20380		

V. Caso de Estudio: Licencias Médicas[5]

Suponga que Ud es Director de Recursos Humanos de una empresa y entre sus múltiples tareas debe realizar un informe respecto a las licencias médicas otorgadas durante el año. En su documento debe contemplar por lo menos la siguiente información. (Considere el archivo Licencias Médicas.sav correspondiente a datos de un año).

a) Muestre un gráfico que permita visualizar número de licencias por ocupación agrupado por sexo. Comente.

b) Identifique las actividades económicas que presentan un porcentaje superior al 20% de licencias médicas. (Muestre tablas o gráficos).

c) Determine cuáles son los diagnósticos (identifique 2) que más se presentan en los hombres, y cuáles son los diagnósticos (identifique 2) que más se presentan en las mujeres. ¿En qué categoría de edades se encuentran esos diagnósticos con mayor frecuencia? Considere las siguientes categorías de edad.
Hasta 24 años
25 a 40 años
41 a 50 años
51 a 60 años
Mayor a 60 años

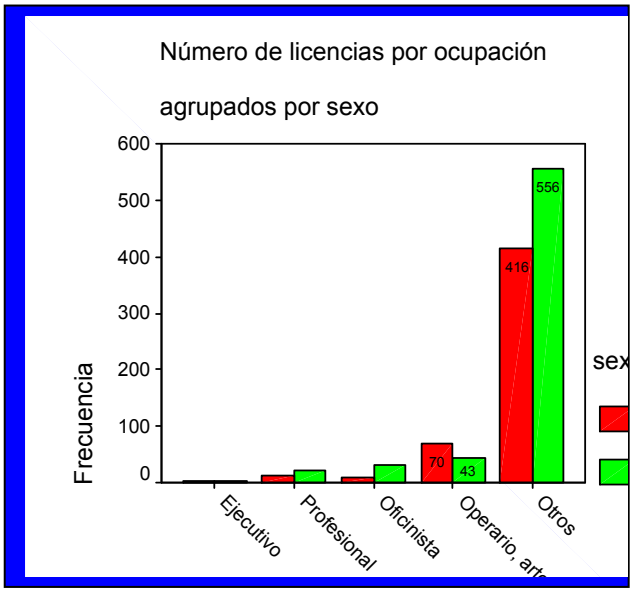
d) Calcular una variable que muestre los días de licencia médica del empleado (Explique el procedimiento o muestre la sintaxis). (Ayuda debe considerar la fecha de inicio y la fecha de término)

e) Realice una tabla, que muestre número de licencias y porcentaje respecto al total, por rangos quintiles de edad, versus rangos cuartiles de los días de licencia médica (variable calculada en el punto anterior). Explique el procedimiento (Pegar la sintaxis).
¿Cuántas mujeres del primer rango quintil de edad presentan hasta 7 días de licencia?

f) Determine para hombres y mujeres los estadísticos; número de casos (N), media, mediana, desviación estándar, mínimo, máximo de la variable edad. Grafique un histograma de edad para hombres y mujeres. Comente los resultados
¿Qué puede decir de la variabilidad de las edades para el grupo de hombres en comparación con el grupo de las mujeres?

Solución
a) Muestre un gráfico que permita visualizar número de licencias por ocupación agrupado por sexo. Comente.

[5]Caso elaborado por Sara Arancibia



Se observa tanto para hombres como para mujeres una despreciable cantidad de licencias en la categoría "Ejecutivos" Se puede apreciar un bajo número de licencias tanto en las categorías Profesional como Oficinista presentándose en ambas categorías mayor número de licencias en mujeres. La categoría de ocupación "operarios, artesanos" presenta mayor cantidad de licencias, observándose 70 licencias en hombres y 43 licencias en mujeres. Con una gran diferencia se presenta la categoría "Otros" que agrupa todo el resto de ocupaciones no mencionadas anteriormente mostrando mayor cantidad en mujeres (556) que en hombres (416).

b) Identifique las actividades económicas que presentan un porcentaje superior al 20% de licencias médicas. (Muestre tablas o gráficos).

Actividad económica

		Frecuencia	Porcentaje	Porcentaje válido
Válidos	,00	20	1,7	1,7
	1,00	1	,1	,1
	2,00	139	11,9	11,9
	3,00	116	9,9	9,9
	4,00	4	,3	,3
	5,00	203	17,4	17,4
	6,00	44	3,8	3,8
	7,00	306	26,2	26,2
	8,00	274	23,5	23,5
	9,00	60	5,1	5,1
	Total	1167	100,0	100,0

Las actividades económicas que presentan el mayor número de licencias médicas con un porcentaje superior al 20% son "Servicios "(código 7) y "Servicio doméstico "(código 8) representando el 26,2 % y 23,5% del total de licencias médicas respectivamente.

c) Determine cuáles son los diagnósticos (identifique 2) que más se presentan en los hombres, y cuáles son los diagnósticos (identifique 2) que más se presentan en las mujeres. ¿En qué categoría de edades se encuentran esos diagnósticos con mayor frecuencia? Considere las siguientes categorías de edad.
Hasta 24 años
25 a 40 años
41 a 50 años
51 a 60 años
Mayor a 60 años

De la tabla de contingencia se puede apreciar que los diagnósticos de mayor frecuencia en las mujeres son la gripe (Código 1) con 214 observaciones y Neurosis (Código 3) con 118 observaciones. El diagnóstico de mayor frecuencia en los hombres también es la gripe (código 1) con 88 observaciones seguido de Lumbago con 76 observaciones.

abla de contingencia Diagnóstico * sex

Recuento

	sexo		
	hombre	mujer	Total
Diagnósti 1,00	88	214	302
2,00	5	5	10
3,00	70	118	188
5,00	13	16	29
6,00	76	56	132
7,00	70	20	90
8,00	43	29	72
9,00	6	3	9
10,00	4	5	9
11,00	3	0	3
12,00	33	114	147
13,00	2	4	6
16,00	5	3	8
17,00	15	6	21
19,00	10	4	14
26,00	7	0	7
27,00	13	10	23
28,00	8	2	10
29,00	6	5	11
33,00	3	1	4
34,00	12	15	27
36,00	4	1	5
38,00	6	13	19
40,00	1	0	1
52,00	0	1	1
58,00	0	5	5
71,00	8	6	14
Total	511	656	1167

Para gripe (Hombres y mujeres)

De la tabla de contingencia se observa que el diagnóstico "gripe" se presenta en todas las categorías de edad tanto para hombres como mujeres , observándose las mayores frecuencias en la categoría 25-40 años.

Tabla de contingencia rangos de edad * sexo * Diagnóstico

Diagnóstico: 1

			sexo		
			hombre	mujer	Total
rangos de edad	Hasta 24	Recuento	14	49	63
		% del total	4,6%	16,2%	20,9%
	25-40	Recuento	45	144	189
		% del total	14,9%	47,7%	62,6%
	41-50	Recuento	13	12	25
		% del total	4,3%	4,0%	8,3%
	51-60	Recuento	7	6	13
		% del total	2,3%	2,0%	4,3%
	mayores a 60	Recuento	9	3	12
		% del total	3,0%	1,0%	4,0%
Total		Recuento	88	214	302
		% del total	29,1%	70,9%	100,0%

Para Neurosis (Mujeres)

El segundo diagnóstico mas frecuente en mujeres, se presenta en todas las categorías de edad aunque sólo se presenta un caso en la categoría mayores de 60 años, alcanzando el mayor porcentaje de licencias (39,4%) la categoría 25 a 40 años.

Tabla de contingencia rangos de edad * sexo * Diagnóstico

Diagnóstico: 3

			sexo	
			mujer	Total
rangos de edad	Hasta 24	Recuento	14	18
		% del total	7,4%	9,6%
	25-40	Recuento	74	119
		% del total	39,4%	63,3%
	41-50	Recuento	23	34
		% del total	12,2%	18,1%
	51-60	Recuento	6	13
		% del total	3,2%	6,9%
	mayores a 60	Recuento	1	4
		% del total	,5%	2,1%
Total		Recuento	118	188
		% del total	62,8%	100,0%

Para Lumbago (Hombres)
El segundo diagnóstico mas frecuente en hombres se presenta en todas las categorías de edad, alcanzando el mayor porcentaje de licencias (28%) la categoría 25 a 40 años.

Tabla de contingencia rangos de edad * sexo * Diagnóstico

Diagnóstico: 6

			sexo	
			hombre	Total
rangos de edad	Hasta 24	Recuento	8	13
		% del total	6,1%	9,8%
	25-40	Recuento	37	71
		% del total	28,0%	53,8%
	41-50	Recuento	16	27
		% del total	12,1%	20,5%
	51-60	Recuento	8	13
		% del total	6,1%	9,8%
	mayores a 60	Recuento	7	8
		% del total	5,3%	6,1%
Total		Recuento	76	132
		% del total	57,6%	100,0%

d) Calcular una variable que muestre los días de licencia médica del empleado (Explique el procedimiento o muestre la sintaxis). (Ayuda debe considerar la fecha de inicio y la fecha de término)

La variable "dias" Dias de licencia se calculó según la sintaxis

```
COMPUTE dias = CTIME.DAYS(fecha_t2 - fecha_i2)+1 .
VARIABLE LABELS dias 'dias de licencia' .
EXECUTE .
```

e) Realice una tabla, que muestre número de empleados y porcentaje respecto al total por rangos quintiles de edad, versus rangos cuartiles de los días de licencias médica (variable calculada en el punto anterior). Explique el procedimiento (Pegar la sintaxis).
¿Cuántas mujeres del primer rango quintil de edad presentan hasta 7 días de licencia?

Solución:
Se calculan los rangos quintiles según la sintaxis:

```
RANK
  VARIABLES=edad  (A) /NTILES (5) /PRINT=YES
  /TIES=MEAN .
```

SE estiman los valores quintiles

Estadísticos

Edad

N	Válidos	1167
	Perdidos	0
Media		36,19
Mediana		34,00
Percentiles	20	26,00
	40	31,00
	60	37,00
	80	45,00

En definición de variables se asigna a cada rango quintil la etiqueta respectiva:

1 <=26
2 26-31
3 31-37
4 37-45
5 >=45

Se calculan los rangos cuartiles de la variable "dias" (Dias de licencia médica) según la sintaxis

Estadísticos

dias de licencia

N	Válidos	1167
	Perdidos	0
Media		13,47
Mediana		9,00
Percentiles	25	7,00
	50	9,00
	75	15,00

```
RANK
  VARIABLES=dias (A) /NTILES (4) /PRINT=YES
  /TIES=MEAN .
```

En definición de variables se asigna a cada rango cuartil la etiqueta respectiva:

1 <=7
2 7-9
3 9-15
4 >=15

Tabla de contingencia NTILES of EDAD * NTILES of DIAS

			Cuartiles de días (Días de licencia médica)				
			<=7	7-9	9-15	>=15	Total
Quintiles de edad	<=26	Recuento	60	79	52	36	227
		% del total	5,1%	6,8%	4,5%	3,1%	19,5%
	26-31	Recuento	58	75	43	51	227
		% del total	5,0%	6,4%	3,7%	4,4%	19,5%
	31-37	Recuento	56	81	53	54	244
		% del total	4,8%	6,9%	4,5%	4,6%	20,9%
	37-45	Recuento	50	49	75	65	239
		% del total	4,3%	4,2%	6,4%	5,6%	20,5%
	>=45	Recuento	42	39	72	77	230
		% del total	3,6%	3,3%	6,2%	6,6%	19,7%
Total		Recuento	266	323	295	283	1167
		% del total	22,8%	27,7%	25,3%	24,3%	100,0%

Al agregar la capa "sexo" a la tabla de contingencia anterior y usando pivotes se obtiene la tabla siguiente

Tabla de contingencia NTILES of EDAD * NTILES of DIAS * sexo

sexo: mujer

			Cuartiles de días (Dias de licencia médica)				
			<=7	7-9	9-15	>=15	Total
Quintiles de edad	<=26	Recuento	28	56	35	25	144
		% del total	4,3%	8,5%	5,3%	3,8%	22,0%
	26-31	Recuento	30	60	34	35	159
		% del total	4,6%	9,1%	5,2%	5,3%	24,2%
	31-37	Recuento	32	51	32	27	142
		% del total	4,9%	7,8%	4,9%	4,1%	21,6%
	37-45	Recuento	24	21	51	32	128
		% del total	3,7%	3,2%	7,8%	4,9%	19,5%
	>=45	Recuento	18	14	27	24	83
		% del total	2,7%	2,1%	4,1%	3,7%	12,7%
Total		Recuento	132	202	179	143	656
		% del total	20,1%	30,8%	27,3%	21,8%	100,0%

Existen 28 mujeres del primer rango quintil de edad que presentan hasta 7 días de licencia médica (4,3% del total).

f) Determine para hombres y mujeres los estadísticos; número de casos (N), media, mediana, desviación estándar, mínimo, máximo de las variables edad. Grafique un histograma de edad para hombres y mujeres. Comente los resultados
¿Qué puede decir de la variabilidad de las edades para el grupo de hombres en comparación con el grupo de las mujeres?

Cubos OLAP

sexo: hombre

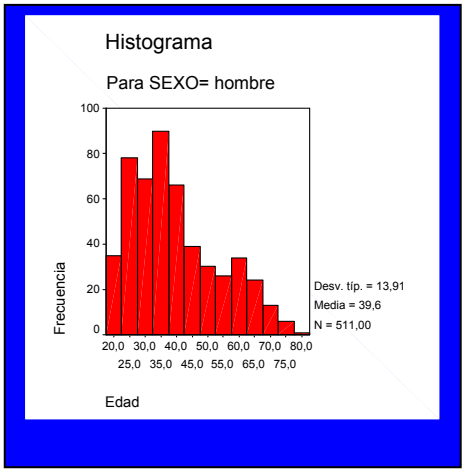
	N	Media	Mediana	Desv. típ.	Mínimo	Máximo
Edad	511	39,59	37,00	13,907	18	79

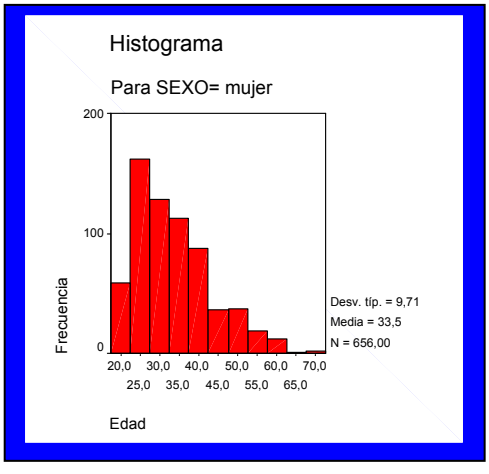
Cubos OLAP

sexo: mujer

	N	Media	Mediana	Desv. típ.	Mínimo	Máximo
Edad	656	33,55	32,00	9,707	18	69

Para Hombres:
La base considera 511 licencias de hombres cuya edad promedio es 39,59 años con una desviación promedio respecto a la media de 13,9 años. La edades oscilan entre 18 y 79 años correspondiendo el 50% de licencias a hombres menores o iguales a 37 años.
Para Mujeres:
La base considera 656 licencias de mujeres cuya edad promedio es 33,55 años con una desviación promedio respecto a la media de 9,7 años. La edades oscilan en un rango menor al de hombres entre 18 y 69 años correspondiendo el 50% de licencias a mujeres menores o iguales a 32 años. Este sesgo hacia edades menores respecto al de hombres se puede observar en los histogramas.
Ambos histogramas muestran asimetría positiva.





Respecto la variabilidad de la edad consideraremos el coeficiente de variación dado que se trata de dos distribuciones cuyas medias son distintas.
Para hombres:

$$CV = \frac{\sigma}{\mu} * 100 = \frac{13,907}{39,59} * 100 = 35,12\%$$

Para mujeres:

$$CV = \frac{\sigma}{\mu} * 100 = \frac{9,707}{33,57} * 100 = 28,91\%$$

Del cálculo anterior se aprecia mayor variación relativa en la distribución de edades de los hombres respecto al de las mujeres.

VI. Estudio de caso: Premio Colegios [6]

Enunciado

Suponga que usted es un asesor del Ministerio de Educación y debe preparar un informe en relación a los rendimientos de los estudiantes de enseñanza media del año 2006. Entre los diversos informes que debe realizar se le ha pedido que sugiera qué colegios premiar con un estímulo por los resultados de la prueba SIMCE de los segundos medios.

El SIMCE es el sistema nacional de medición de resultados de aprendizaje del Ministerio de Educación de Chile. Su propósito principal es construir al mejoramiento de la calidad y equidad de la educación, informando sobre el desempeño de los alumnos y alumnas en algunas áreas del curriculum nacional y relacionándolos con el contexto escolar y social en el que ellos aprenden.

Las pruebas SIMCE evalúan el logro de los Objetivos Fundamentales y Contenidos Mínimos Obligatorios del Marco Curricular en diferentes subsectores de aprendizaje, a través de una prueba común que se aplica a nivel nacional, una vez al año, a los estudiantes que cursan un determinado nivel educacional. Hasta el 2005 la aplicación de las pruebas se alternaron entre 4° Básico, 8° Básico y 2° Medio. Desde el 2006, las pruebas evalúan todos los años el nivel del 4° Básico y se alternan los niveles de 8° Básico y 2° Medio. (Fuente: Resultados nacionales SIMCE 2006. MINEDUC)

Se dispone de un archivo con los datos de los 2319 colegios evaluados en la prueba SIMCE 2° Medio del 2006. Algunas de las variables de interés son:

VARIABLE	ETIQUETA DE VARIABLE	ETIQUETA DEVALOR
Idest	Identificador del establecimiento	
Región	Nombre de la Región	
Comuna	Nombre de la comuna	
ddca	Dependencia	CP: Corporación Privada MC; Corporación Municipal MD; DAEM (Departamento de Administración de Educación Municipal) PP: Particular Pagado PS: Particular Subvencionado
ruralida	Caracterización del establecimiento	1= Rural 2=Urbano
prom_len	Promedio puntaje de lenguaje	
prom_mat	Promedio puntaje de matemáticas	

Después de múltiples reuniones con expertos en educación, usted ha llegado a definir junto con los expertos un criterio para premiar a las escuelas; crear grupos homogéneos de escuelas y definir puntajes de corte para cada grupo. De esta forma se estará distinguiendo a los colegios que se destacan entre colegios con similares características. El premio se otorgará a los colegios con puntajes promedios mayores o iguales al percentil 75 (para cada grupo). Los grupos homogéneos se definieron en base a dos criterios: la

[6] Caso elaborado por Sara Arancibia

dependencia del establecimiento definido como Municipal, Privado y Subvencionado y la caracterización del establecimiento Rural y Urbano

Los grupos homogéneos definidos por el grupo experto son,

1: Municipal y Rural
2: Municipal y Urbano
3: Privado y Rural
4: Privado y Urbano
5: Subvencionado y Rural
6: Subvencionado y Urbano

Usted como asesor del Ministerio de Educación debe aplicar los criterios definidos con los expertos para crear los grupos de colegios homogéneos e identificar cuáles son los establecimientos premiados realizando distintas comparaciones por dependencia, caracterización y zona (Norte, Central y Sur). Además debe determinar si existen diferencias significativas para los puntajes promedios de la SIMCE por caracterización y por dependencia

Para realizar su análisis deberá lograr los siguientes objetivos específicos desglosados en tareas elementales

1. Limpiar y ordenar la base de interés para el análisis

a) Crear la variable "Dependencia" considerando sólo tres categorías:
Municipalizado, Privado y Subvencionado

b) Crear la variable zona considerando Zona Norte, Centro y Sur

c) Crear la variable "puntprom" correspondiente al puntaje promedio entre matemática y lenguaje

d) Crear la variable "grupo" correspondiente a cada grupo homogéneo. Para esto deberá crear con sintaxis (sintaxisgrupo) la variable solicitada asignando los códigos 1 al 6 según corresponda.

e) Determinar para cada grupo el percentil 75.

f) Con otra sintaxis (sintaxispremio) crear la variable "premio" donde según el criterio mencionado 1=SI recibe premio y 0=NO recibe premio.

2. Realizar un análisis descriptivo de los datos

a) ¿Cuántos establecimientos rurales y urbanos existen en el archivo de datos y qué porcentaje representan del total? ¿Qué tipo de dependencia se observa con mayor y menor frecuencia? ¿Qué grupo homogéneo de establecimientos presenta mayor frecuencia?.

b) ¿Cuántos colegios obtuvieron puntajes promedio en Matemáticas inferior o igual a 250 puntos; entre 251 y 300 puntos y superior a 300 puntos?.

c) ¿Qué porcentaje de colegios obtuvieron puntajes promedio en Lenguaje superior a 300 puntos? ¿Cuántos de ellos son Municipalizados y Urbanos? ¿Qué puede decir de los Municipalizados y Rurales?

d) ¿Qué porcentaje representa el total de colegios premiados respecto al total de colegios? ¿Qué porcentaje de los colegios Municipalizados resultaron premiados? ¿Qué porcentaje de los colegios premiados son Subvencionados? ¿Qué porcentaje de los premiados son de la zona Norte,Centro y Sur? ¿Qué porcentaje de la zona Centro son premiados? ¿Qué porcentaje del total son premiados y del Sur?.

e) ¿Qué porcentaje de los colegios premiados son urbanos? ¿Qué porcentaje de los colegios rurales son premiados? ¿Qué porcentaje de colegios resultaron premiados por grupo?

f) ¿A qué zona pertenecen los cinco mayores puntajes promedios SIMCE por tipo rural y urbana?. Realice un gráfico que permita observar la forma de la distribución de los puntajes promedio SIMCE para los colegios rurales y los urbanos y muestre además un diagrama de caja (boxplot) por tipo para el puntaje promedio ¿Qué puede observar?

g) Determine mediante una gráfica si hay diferencias entre las medias de los puntajes en lenguaje y en matemáticas por dependencia para el grupo de colegios en estudio. ¿Existen diferencias en los puntajes de lenguaje por dependencia, agrupados por tipo rural y urbano? Determine si el comportamiento de los resultados de puntajes de lenguaje y matemáticas es similar si se compara los segmentos rural y urbano

h) Determine los estadísticos básicos de tendencia central, de dispersión y de forma de la distribución de los puntajes promedios SIMCE para los distintos grupos homogéneos, y muestre gráficamente la media de los puntajes promedios SIMCE por grupo homogéneo.

i) Realice un gráfico considerando a todos los colegios en estudio y otro gráfico considerando sólo el segmento de premiados, que permitan observar la posición del grupo en relación al resto de los grupos en cuanto a los descriptivos básicos. Interprete.

j) Compare la variabilidad entre los distintos grupos para el puntaje promedio SIMCE.

SOLUCION:

1. Limpiar y ordenar la base de interés para el análisis

a) Crear la variable "Dependencia" considerando sólo tres categorías:
Municipalizado, Privado y Subvencionado
En primer lugar observamos que la variable de dependencia "ddcia" de la base de datos viene con formato cadena o string. Recodificaremos automáticamente y luego llevaremos las cinco categorías sólo a tres categorías.
Para esto ir al menú Transformar/ recodificación automática/

```
AUTORECODE
  VARIABLES=ddcia  /INTO depend
  /PRINT.
```

Old Value New Value Value Label

CP	1	Corporación Privada
MC	2	Corporación Municipal
MD	3	DAEM
PP	4	Particular Pagado
PS	5	Particular Subencionado

Para crear tres categorías juntaremos las categorías Corporación Municipal y DAEM en Municipalizado y las categorías Corporación Privada y Particular pagado en Privado
Para esto ir al menú Transformar/Recodificar en distinta variable

```
RECODE
  depend
  (4=2)  (5=3)  (1=2)  (2 thru 3=1)  INTO  dependencia .
VARIABLE LABELS dependencia 'Tipo de dependencia'.
EXECUTE .
```

En definición de la variable
Asignar etiquetas de valor a los códigos 1 al 3
1= Municipalizado
2=Privado

3=Subvencionado

b) Crear la variable zona considerando Zona; Norte, Centro y Sur

En primer lugar se observa que la variable Región viene en formato de cadena. Se recodificará automáticamente.

```
AUTORECODE
  VARIABLES=region  /INTO reg
  /PRINT
```

En el visor de resultados se puede observar los códigos de cada categoría

Old Value	New Value	Value Label
Región de Aisén del General Carlos Ibañez del Campo	1	Región de Aisén del General Carlos Ibañez del Campo
Región de Antofagasta	2	Región de Antofagasta
Región de Atacama	3	Región de Atacama
Región de Coquimbo	4	Región de Coquimbo
Región de la Araucanía	5	Región de la Araucanía
Región de Los Lagos	6	Región de Los Lagos
Región de Magallanes y de la Antártica Chilena	7	Región de Magallanes y de la Antártica Chilena
Región de Tarapacá	8	Región de Tarapacá
Región de Valparaíso	9	Región de Valparaíso
Región del Biobío	10	Región del Biobío
Región del Libertador General Bernardo O' Higgins	11	Región del Libertador General Bernardo O' Higgins
Región del Maule	12	Región del Maule
Región Metropolitana	13	Región Metropolitana

Para crear las categorías de zona se recodificará en distintas variables

```
RECODE
  reg
  (1=3)  (8=1)  (9=2)  (10=3)  (2 thru 4=1)  (5 thru 7=3)  (11 thru 13=2)  INTO  zona .
VARIABLE LABELS zona 'zona'.
EXECUTE .
```

En definición de variables

1= Norte

2= Centro

3= Sur

c) Crear la variable "puntprom" correspondiente al puntaje promedio entre matemática y lenguaje

Al ver el formato de las variables prom_len y prom_mat se observa que viene con tipo: String o cadena y medida nominal. Lo primero que debemos hacer antes de sacar el promedio es cambiar en vista de variables el tipo String a numérico.

Para crear la variable puntprom seleccione Transformar/Calcular

Variable destino: puntprom

Tipo: numérico

Etiqueta: Promedio de Matemáticas y Lenguaje
Expresión: MEAN(prom_len,prom_mat)
Sintaxis de puntprom

```
COMPUTE puntprom = MEAN(prom_len,prom_mat) .
VARIABLE LABELS puntprom 'puntaje promedio entre lenguaje y matemáticas'.
EXECUTE .
```

d) Crear la variable "grupo" correspondiente a cada grupo homogéneo. Para esto deberá crear con sintaxis (sintaxisgrupo) la variable solicitada asignando los códigos 1 al 6 según corresponda.

Para crear la variable de grupo primero recodificaremos automáticamente la variables ruralida a código numérico con nombre caract
Donde
caract=1 Rural
caract=2 Urbano

Sintaxis

```
AUTORECODE
  VARIABLES=ruralida  /INTO caract
  /PRINT.
```

Ahora formamos los seis grupos según criterio dado

Creación de la variable grupo

```
*** Sintaxis Grupo ***.
IF (dependencia = 1 & caract = 1) grupo = 1 .
IF (dependencia = 1 & caract = 2) grupo = 2 .
IF (dependencia = 2 & caract = 1) grupo = 3 .
IF (dependencia = 2 & caract = 2) grupo = 4 .
IF (dependencia = 3 & caract = 1) grupo = 5 .
IF (dependencia = 3 & caract = 2) grupo = 6 .
EXECUTE .
```

Luego en la definición de variables en valores se define:

1: Municipal y Rural
2: Municipal y Urbano
3: Privado y Rural
4: Privado y Urbano
5: Subvencionado y Rural
6: Subvencionado y Urbano

e) Determinar para cada grupo el percentil 75.
Para el cálculo de los percentiles por grupo: Datos/Segmentar, variable: grupo. Luego Analizar/Frecuencias [Estadísticos]: Percentil 75

Sintaxis

```
SORT CASES BY grupo .
SPLIT FILE
  LAYERED BY grupo .
```

```
FREQUENCIES
  VARIABLES=puntprom  /FORMAT=NOTABLE
  /PERCENTILES= 75
  /ORDER=  ANALYSIS .
```

No olvide volver a Datos/ Segmentar archivo/ Analizar todos los casos.

f) Con otra sintaxis (sintaxispremio) crear la variable "premio" donde según el criterio mencionado 1=SI recibe premio y 0=NO recibe premio.

Se consideró el siguiente criterio para premiar a los colegios (donde 1=SI, 0=NO)

Estadísticos

puntaje promedio entre lenguaje y matemáticas

Municipalizado y Rural	N	Válidos	68
		Perdidos	0
	Percentiles	75	225,6250
Municipalizado y Urbano	N	Válidos	587
		Perdidos	0
	Percentiles	75	246,0000
Privado y Rural	N	Válidos	19
		Perdidos	0
	Percentiles	75	322,0000
Privado y Urbano	N	Válidos	388
		Perdidos	0
	Percentiles	75	324,0000
Subvencionado y Rural	N	Válidos	86
		Perdidos	0
	Percentiles	75	246,3750
Subvencionado y Urbano	N	Válidos	1171
		Perdidos	0
	Percentiles	75	286,0000

```
*** Sintaxis Premio ***.
IF (grupo = 1 & puntprom >= 225.625) premio = 1 .
IF (grupo = 1 & puntprom <  225.625) premio = 0 .
IF (grupo = 2 & puntprom >= 246) premio = 1 .
IF (grupo = 2 & puntprom <  246) premio = 0 .
IF (grupo = 3 & puntprom >= 322.5)  premio = 1 .
IF (grupo = 3 & puntprom <  322.5)  premio = 0 .
IF (grupo = 4 & puntprom >= 324)  premio = 1 .
IF (grupo = 4 & puntprom <  324)  premio = 0 .
IF (grupo = 5 & puntprom >= 246.375)  premio = 1 .
IF (grupo = 5 & puntprom <  246.375)  premio = 0 .
IF (grupo = 6 & puntprom >= 286) premio = 1 .
IF (grupo = 6 & puntprom <  286) premio = 0 .

VARIABLE LABELS premio 'premio (SI=1, NO=0)' .
EXECUTE .
```

En definición de variables se agrega la etiqueta de valor
1=SI
0=NO

2. Realizar un análisis descriptivo de los datos

a) ¿Cuántos establecimientos rurales y urbanos existen en el archivo de datos y qué porcentaje representan del total? ¿Qué tipo de dependencia se observa con mayor y menor frecuencia? ¿Qué grupo homogéneo de establecimientos presenta mayor frecuencia?

Se debe realizar una tabla de frecuencias de la variable caract, dependencia y grupo.
Analizar/ frecuencias.
Sintaxis del procedimiento:

```
FREQUENCIES
  VARIABLES=dependencia tipo grupo
  /ORDER= ANALYSIS .
```

Caracterización del establecimiento

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Rural	173	7,5	7,5	7,5
	Urbano	2146	92,5	92,5	100,0
	Total	2319	100,0	100,0	

Tipo de dependencia

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Municipalizado	655	28,2	28,2	28,2
	Privado	407	17,6	17,6	45,8
	Subvencionado	1257	54,2	54,2	100,0
	Total	2319	100,0	100,0	

grupo

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Municipalizado y Rural	68	2,9	2,9	2,9
	Municipalizado y Urbano	587	25,3	25,3	28,2
	Privado y Rural	19	,8	,8	29,1
	Privado y Urbano	388	16,7	16,7	45,8
	Subvencionado y Rural	86	3,7	3,7	49,5
	Subvencionado y Urbano	1171	50,5	50,5	100,0
	Total	2319	100,0	100,0	

De la tabla de frecuencia, se observa que existen 2146 colegios de tipo urbano y 173 colegios de tipo Rural representando el 92,5% y 7,5 % respectivamente sobre el total de colegios considerados en la base de datos.
Por otra parte de la tabla de frecuencia de dependencia se tiene que la mayor frecuencia se presenta en los establecimientos subvencionados representando el 54,2% del total y la menor frecuencia en los establecimientos Privados representando el 17,6% del total.
Respecto a los grupos homogéneos el de mayor frecuencia es el grupo de Subvencionado y Urbano representando aproximadamente la mitad de los colegios considerados en estudio, le sigue el grupo de Municipalizados y Urbanos representando un cuarto de los colegios en estudio.

b) ¿Cuántos colegios obtuvieron puntajes promedio en Matemáticas inferior o igual a 250 puntos; entre 251 y 300 puntos y superior a 300 puntos?

Para responder esta pregunta se debe crear rangos de puntajes en base al puntaje de Matemáticas.
Transformar/Recodificar/en distinta variable

Ingresar la variable prom_mat y definir variable nueva rangmat (notar que la variable prom_mat es una variable de números enteros)

Sintaxis del procedimiento:

```
RECODE
  prom_mat
  (Lowest thru 250=1)  (251 thru 300=2)  (301 thru Highest=3)  INTO  rangmat.
  VARIABLE LABELS rangmat 'rangos de puntajes en matematicas'.
EXECUTE .
```

.

En la ventana de definición de variables considerar la variable rangmat y en valores definir cada rango como:

1 =Hasta 250
2 = 251-300
3= superior a 300

Luego realizar una tabla de frecuencias de la variable rangmat

rangos de puntajes en matematicas

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Hasta 250	1169	50,4	50,4	50,4
	251-300	658	28,4	28,4	78,8
	Superior a 300	492	21,2	21,2	100,0
	Total	2319	100,0	100,0	

En la tabla de frecuencia se puede apreciar la cantidad de colegios por rangos de puntajes en matemáticas Se puede observar que aproximadamente la mitad de los colegios en estudio obtuvieron un puntaje promedio en matemáticas menor o igual a 250 puntos. Un poco más de la quinta parte de los colegios obtiene un puntaje superior a 300 puntos.

c) ¿Qué porcentaje de colegios obtuvieron puntajes promedio en Lenguaje superior a 300 puntos? ¿Cuántos de ellos son Municipalizados y Urbanos? ¿Qué puede decir de los Municipalizados y Rurales?

En primer lugar se debe crear dos rangos para la variable prom_len; Hasta 300 puntos y Superior a 300 puntos.
Transformar/Recodificar/en distinta variable
Ingresar la variable prom_len y definir variable nueva rangleng (notar que la variable prom_len es una variable de números enteros)

Sintaxis del procedimiento

```
RECODE
  prom_len
  (Lowest thru 300=1)  (301 thru Highest=2)  INTO  rangleng .
VARIABLE LABELS rangleng 'rangos de puntajes en lenguaje'.
EXECUTE .
```

En la ventana de definición de variables considerar la variable rangleng y en valores definir cada rango como:

1 =Hasta 300
2 = superior a 300

Luego realizar una tabla de frecuencias de la variable rangleng

De la tabla de frecuencias se puede observar que el 13,8% de los colegios obtuvieron un puntaje superior a

rangos de puntajes en lenguaje

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Hasta 300	1999	86,2	86,2	86,2
	Superior a 300	320	13,8	13,8	100,0
	Total	2319	100,0	100,0	

300 puntos.
Para responder cuántos de estos colegios son Municipalizados y Urbanos, se puede seleccionar a rangleng igual a 2 que corresponde a los puntajes superiores a 300 puntos y luego pedir una tabla de contingencia para las variables dependencia y tipo.
Datos/Seleccionar casos

Sintaxis del procedimiento

```
USE ALL.
COMPUTE filter_$=(rangleng = 2).
VARIABLE LABEL filter_$ 'rangleng = 2 (FILTER)'.
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .

CROSSTABS
  /TABLES=dependencia  BY caract
  /FORMAT= AVALUE TABLES
  /CELLS= COUNT
  /COUNT ROUND CELL .
```

Tabla de contingencia Tipo de dependencia * Caracterización del establecimiento

Recuento

		Caracterización del establecimiento		
		Rural	Urbano	Total
Tipo de dependencia	Municipalizado	0	10	10
	Privado	7	193	200
	Subvencionado	2	108	110
Total		9	311	320

Se ha considerado los puntajes superiores a 300 puntos

De la tabla de contingencia se puede apreciar que de los colegios con puntajes superiores a 300 puntos, sólo 10 corresponden a Municipalizado y Urbano y ninguno a Municipalizado y Rural

Otra forma de responder a esta pregunta podría ser solicitando una tabla de contingencia para dependencia y tipo con una capa dada por la variable rangleng (para esto seleccionar todos los casos)

Sintaxis del procedimiento

```
CROSSTABS
  /TABLES=dependencia  BY tipo  BY rangleng
  /FORMAT= AVALUE TABLES
  /CELLS= COUNT
  /COUNT ROUND CELL .
```

'abla de contingencia Tipo de dependencia * Caracterización del establecimiento * rangos de puntajes en lenguaje

Recuento

rangos de puntajes en lenguaje			Caracterización del establecimiento		Total
			Rural	Urbano	
Hasta 300	Tipo de dependencia	Municipalizado	68	577	645
		Privado	12	195	207
		Subvencionado	84	1063	1147
	Total		164	1835	1999
Superior a 300	Tipo de dependencia	Municipalizado	0	10	10
		Privado	7	193	200
		Subvencionado	2	108	110
	Total		9	311	320

De esta forma se obtiene el mismo resultado. Si se quiere mostrar una tabla focalizada a la respuesta, se puede pivotar editando la tabla y moviendo al pivote de rangos de puntajes al extremo superior izquierdo. De esta forma se puede obtener la siguiente tabla.

Tabla de contingencia Tipo de dependencia * Caracterización del establecimiento * rangos de puntajes en lenguaje

Recuento

rangos de puntajes en lenguaje: Superior a 300

		Caracterización del establecimiento		Total
		Rural	Urbano	
Tipo de dependencia	Municipalizado	0	10	10
	Privado	7	193	200
	Subvencionado	2	108	110
Total		9	311	320

d) ¿Qué porcentaje representa el total de colegios premiados respecto al total de colegios? ¿Qué porcentaje de los colegios Municipalizados resultaron premiados? ¿Qué porcentaje de los colegios premiados son Subvencionados? ¿Qué porcentaje de los premiados son de la zona Norte, Centro y Sur? ¿Qué porcentaje de la zona Centro son premiados? ¿Qué porcentaje del total son premiados y del Sur?
Para responder a estas preguntas se puede considerar tablas de frecuencia y de contingencia.
En primer lugar solicitaremos una tabla de frecuencia de premio

premio (SI=1, NO=0)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	NO	1730	74,6	74,6	74,6
	SI	589	25,4	25,4	100,0
	Total	2319	100,0	100,0	

De la tabla de frecuencias se obtiene que el 25,4% del total de colegios resulta premiado.
Para saber qué porcentaje de los colegios Municipalizados resultaron premiados, y qué porcentaje de los colegios premiados son Subvencionados podemos realizar una tabla de contingencia de la variable dependencia versus premio solicitando el porcentaje fila y columna.

```
CROSSTABS
 /TABLES=dependencia BY premio
 /FORMAT= AVALUE TABLES
 /CELLS= COUNT ROW COLUMN TOTAL
 /COUNT ROUND CELL .
```

De la tabla se observa que de los colegios Municipalizados el 25, 3% resultaron premiados y que del total de premiados el 54% corresponde a Subvencionados

Tabla de contingencia Tipo de dependencia * premio (SI=1, NO=0)

			premio (SI=1, NO=0)		
			NO	SI	Total
Tipo de dependencia	Municipalizado	Recuento	489	166	655
		% de Tipo de dependencia	74,7%	25,3%	100,0%
		% de premio (SI=1, NO=0)	28,3%	28,2%	28,2%
	Privado	Recuento	302	105	407
		% de Tipo de dependencia	74,2%	25,8%	100,0%
		% de premio (SI=1, NO=0)	17,5%	17,8%	17,6%
	Subvencionado	Recuento	939	318	1257
		% de Tipo de dependencia	74,7%	25,3%	100,0%
		% de premio (SI=1, NO=0)	54,3%	54,0%	54,2%
Total		Recuento	1730	589	2319
		% de Tipo de dependencia	74,6%	25,4%	100,0%
		% de premio (SI=1, NO=0)	100,0%	100,0%	100,0%

De la misma forma para saber los porcentajes en relación a la zona podemos realizar una tabla de contingencia de zona versus premio solicitando los porcentajes fila, columna y total

```
CROSSTABS
 /TABLES=zona BY premio
```

/FORMAT= AVALUE TABLES
/CELLS= COUNT ROW COLUMN TOTAL
/COUNT ROUND CELL .

Tabla de contingencia zona * premio (SI=1, NO=0)

			premio (SI=1, NO=0)		
			NO	SI	Total
zona	Norte	Recuento	209	72	281
		% de zona	74,4%	25,6%	100,0%
		% de premio (SI=1, NO=0)	12,1%	12,2%	12,1%
		% del total	9,0%	3,1%	12,1%
	Centro	Recuento	1076	341	1417
		% de zona	75,9%	24,1%	100,0%
		% de premio (SI=1, NO=0)	62,2%	57,9%	61,1%
		% del total	46,4%	14,7%	61,1%
	Sur	Recuento	445	176	621
		% de zona	71,7%	28,3%	100,0%
		% de premio (SI=1, NO=0)	25,7%	29,9%	26,8%
		% del total	19,2%	7,6%	26,8%
Total		Recuento	1730	589	2319
		% de zona	74,6%	25,4%	100,0%
		% de premio (SI=1, NO=0)	100,0%	100,0%	100,0%
		% del total	74,6%	25,4%	100,0%

Se obtiene que de los colegios premiados el 12,2% corresponde a la zona Norte, el 57,9% corresponde a la zona Centro y el 29,9% corresponde a la zona Sur. Ahora respecto a los colegios de la zona Centro el 24,1% resulta premiado y respecto al total de colegios el 7,6% son premiados y de la zona Sur.

e) ¿Qué porcentaje de los colegios premiados son urbanos? ¿Qué porcentaje de los colegios rurales son premiados? ¿Qué porcentaje de colegios resultaron premiados por grupo?

En forma análoga a la pregunta anterior se puede resolver con una tabla de contingencia de premio versus tipo y premio versus grupo

Sintaxis del procedimiento
CROSSTABS
/TABLES=tipo grupo BY premio
/FORMAT= AVALUE TABLES
/CELLS= COUNT ROW COLUMN TOTAL
/COUNT ROUND CELL .

Tabla de contingencia Caracterización del establecimiento * premio (SI=1, NO=0)

			premio (SI=1, NO=0)		
			NO	SI	Total
Caracterización del establecimiento	Rural	Recuento	131	42	173
		% de Caracterización del establecimiento	75,7%	24,3%	100,0%
		% de premio (SI=1, NO=0)	7,6%	7,1%	7,5%
		% del total	5,6%	1,8%	7,5%
	Urbano	Recuento	1599	547	2146
		% de Caracterización del establecimiento	74,5%	25,5%	100,0%
		% de premio (SI=1, NO=0)	92,4%	92,9%	92,5%
		% del total	69,0%	23,6%	92,5%
Total		Recuento	1730	589	2319
		% de Caracterización del establecimiento	74,6%	25,4%	100,0%
		% de premio (SI=1, NO=0)	100,0%	100,0%	100,0%
		% del total	74,6%	25,4%	100,0%

De la tabla se obtiene que de los colegios premiados, el 92,9% son urbanos y del total de colegios rurales el 24,3 % son premiados
De la tabla de contingencia de grupo versus premio se puede ver que el criterio utilizado para premiar a los mejores colegios según su rendimiento SIMCE otorgó casi equitativamente el 25% de colegios premiados por grupo.

Tabla de contingencia grupo * premio (SI=1, NO=0)

			premio (SI=1, NO=0)		
			NO	SI	Total
grupo	Municipalizado y Rural	Recuento	51	17	68
		% de grupo	75,0%	25,0%	100,0%
		% de premio (SI=1, NO=0)	2,9%	2,9%	2,9%
		% del total	2,2%	,7%	2,9%
	Municipalizado y Urbano	Recuento	438	149	587
		% de grupo	74,6%	25,4%	100,0%
		% de premio (SI=1, NO=0)	25,3%	25,3%	25,3%
		% del total	18,9%	6,4%	25,3%
	Privado y Rural	Recuento	15	4	19
		% de grupo	78,9%	21,1%	100,0%
		% de premio (SI=1, NO=0)	,9%	,7%	,8%
		% del total	,6%	,2%	,8%
	Privado y Urbano	Recuento	287	101	388
		% de grupo	74,0%	26,0%	100,0%
		% de premio (SI=1, NO=0)	16,6%	17,1%	16,7%
		% del total	12,4%	4,4%	16,7%
	Subvencionado y Rural	Recuento	65	21	86
		% de grupo	75,6%	24,4%	100,0%
		% de premio (SI=1, NO=0)	3,8%	3,6%	3,7%
		% del total	2,8%	,9%	3,7%
	Subvencionado y Urbano	Recuento	874	297	1171
		% de grupo	74,6%	25,4%	100,0%
		% de premio (SI=1, NO=0)	50,5%	50,4%	50,5%
		% del total	37,7%	12,8%	50,5%
Total		Recuento	1730	589	2319
		% de grupo	74,6%	25,4%	100,0%
		% de premio (SI=1, NO=0)	100,0%	100,0%	100,0%
		% del total	74,6%	25,4%	100,0%

f) ¿A qué zona pertenecen los cinco mayores puntajes promedios SIMCE por tipo rural y urbana? Realice un gráfico que permita observar la forma de la distribución de los puntajes promedio SIMCE para los colegios rurales y los urbanos y muestre además un diagrama de caja (boxplot) por tipo para el puntaje promedio ¿Qué puede observar?

Para responder a la pregunta se puede solicitar en explorar una tabla de valores extremos de puntaje promedio por tipo, identificando por zona, y en gráficos pedir el histograma y diagrama de cajas.

```
EXAMINE
  VARIABLES=puntprom BY caract /ID= zona
  /PLOT BOXPLOT HISTOGRAM
  /COMPARE GROUP
  /STATISTICS EXTREME
  /MISSING LISTWISE
  /NOTOTAL.
```

Valores extremos

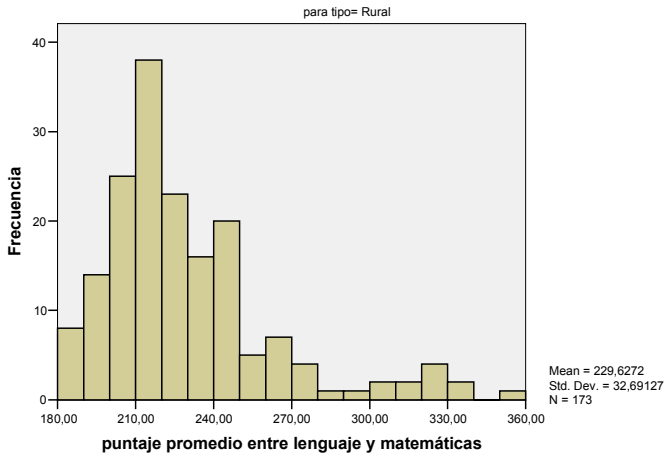
Mayores

	Caracterización del establecimiento		Número del caso	zona	Valor
puntaje promedio entre lenguaje y matemáticas	Rural	1	2011	Centro	351,50
		2	2016	Centro	334,50
		3	2010	Centro	330,50
		4	2015	Centro	324,50
		5	2012	Sur	322,00[a]
	Urbano	1	2090	Centro	357,50
		2	2052	Centro	352,50
		3	2170	Centro	352,50
		4	2203	Centro	350,50
		5	2093	Centro	348,50

a. En la tabla de valores extremos mayores sólo se muestra una lista parcial de los casos con el valor 322,00.

De esta forma se obtiene que los colegios de mayor puntaje promedio tanto en los colegios de tipo rural como de tipo urbano se encuentran en la zona Centro, excepto el colegio con el quinto mejor puntaje de tipo rural que recae en la zona Sur (aunque existen otros puntajes con 322 puntos en promedio)
En los histogramas se puede observar la forma de la distribución de los puntajes promedios para los colegios de tipo rural y los de tipo rrbano
En el histograma correspondiente a los colegios de tipo rural se puede apreciar una asimetría positiva con puntajes mas sesgados hacia puntajes bajos y con varios colegios con puntajes en el extremo superior (casos extremos y atípicos). Claramente no es una distribución simétrica, y además algo levantada denotando que es leptocurtica, por tanto no se asemeja a una distribución normal.

Histograma



En cambio la distribución de los puntajes de los colegios de tipo urbano se observa bastante simétrica sin puntajes claramente extremos y/o atípicos, pero no es clara la forma de una curva normal.

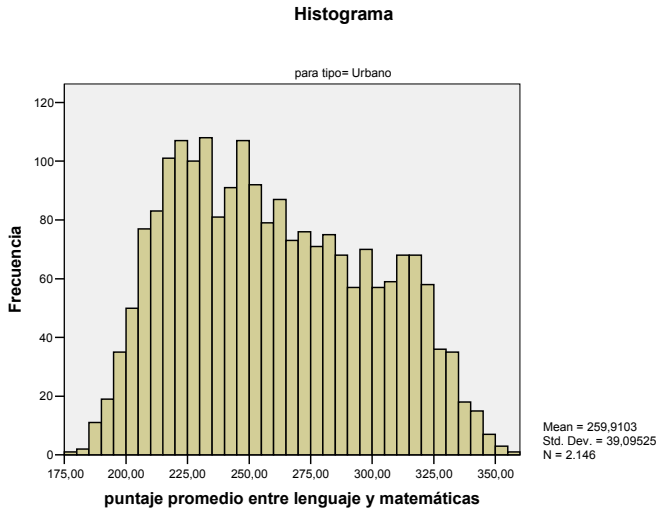
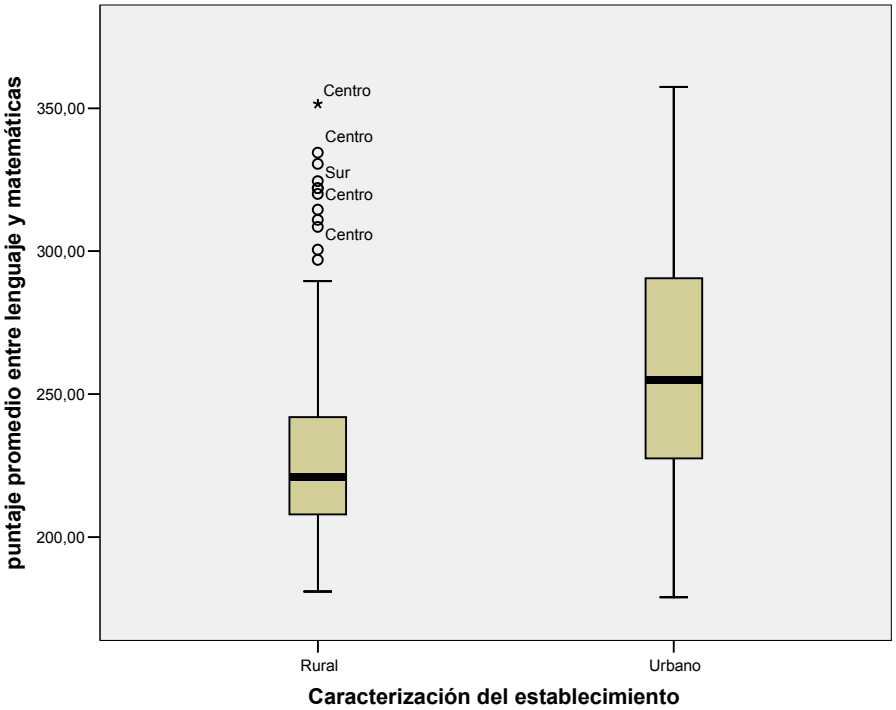


Diagrama de caja para puntaje promedio por tipo de colegio Rural y Urbano



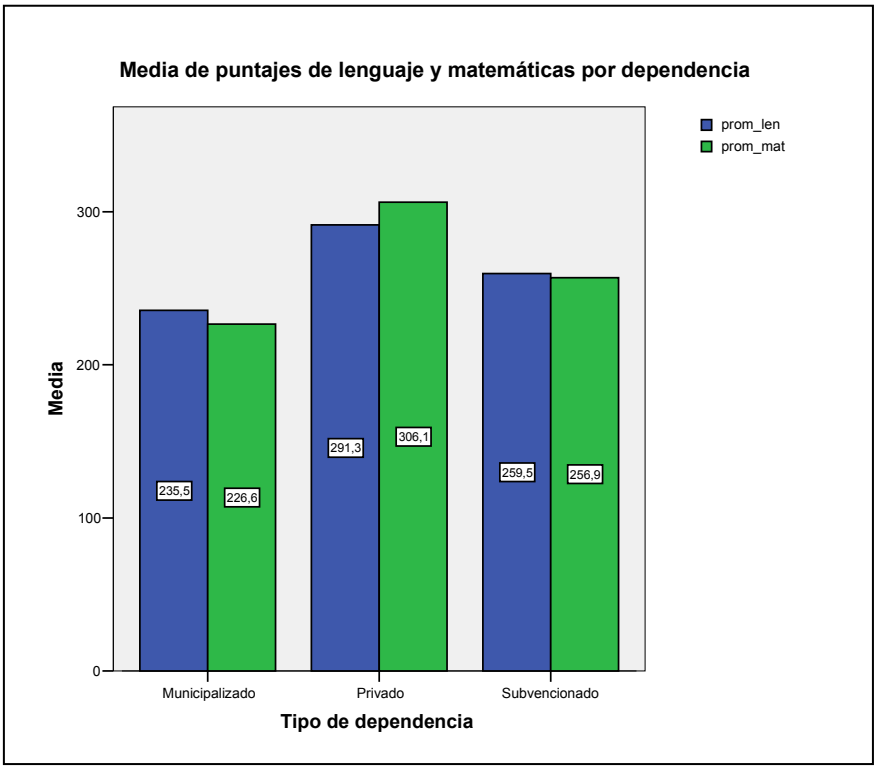
El diagrama de cajas por tipo rural y urbano nos muestra claramente las diferencias en ambas distribuciones. Se observa que la mediana de puntajes de los colegios rurales (línea horizontal negra) está cerca de los 220 puntos lo que indica que la mitad de este tipo de colegios tiene un puntaje promedio inferior o igual al valor de la mediana que en este caso es 221 puntos. Se observan varios valores atípicos y un valor extremo en la parte superior de puntajes. Los puntajes de los colegios rurales en general están

más abajo que si comparamos con los puntajes de colegios urbanos. Se observa que la mediana de los colegios urbanos está sobre los 250 puntos. Específicamente la mediana es 255 puntos lo que indica que el 50% de los colegios urbanos tienen puntajes inferiores o iguales a 255 puntos. No se observan valores extremos ni atípicos.

g) Determine mediante una gráfica si hay diferencias entre las medias de los puntajes en lenguaje y en matemáticas por dependencia para el grupo de colegios en estudio. ¿Existen diferencias en los puntajes de lenguaje por dependencia, agrupados por tipo rural y urbano? Determine si el comportamiento de los resultados de puntajes de lenguaje y matemáticas es similar si se compara los segmentos rural y urbano

Para realizar esta gráfica se selecciona Grafico/barras/Para distintas variables/Agrupado

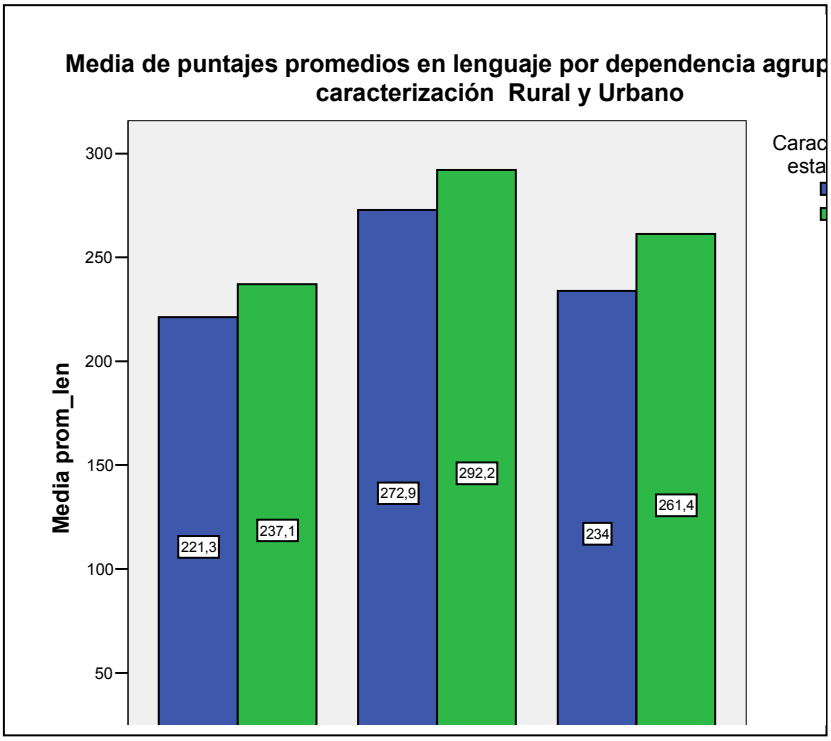
```
GRAPH
  /BAR(GROUPED)=MEAN(prom_len) MEAN(prom_mat) BY dependencia
  /MISSING=LISTWISE .
```



Del gráfico se observa que tanto para lenguaje como para matemáticas los colegios privados tiene mejores resultados observándose sin embargo diferencias en casi 15 puntos a favor del promedio de matemáticas. Los colegios subvencionados presentan puntajes intermedios si comparamos los privados y los municipalizados con puntajes en torno a 257 puntos en matemáticas y con una diferencia de solo 2 puntos a favor del puntaje promedio en lenguaje. Los colegios municipalizados presentan los puntajes promedios mas bajos en ambas pruebas con una diferencia de cerca de 9 puntos a favor de lenguaje.

Para responder a la pregunta si existen diferencias en los puntajes de lenguaje por dependencia, agrupados por tipo rural y urbano hacemos un gráfico de barras agrupados seleccionando resúmenes para grupos de casos

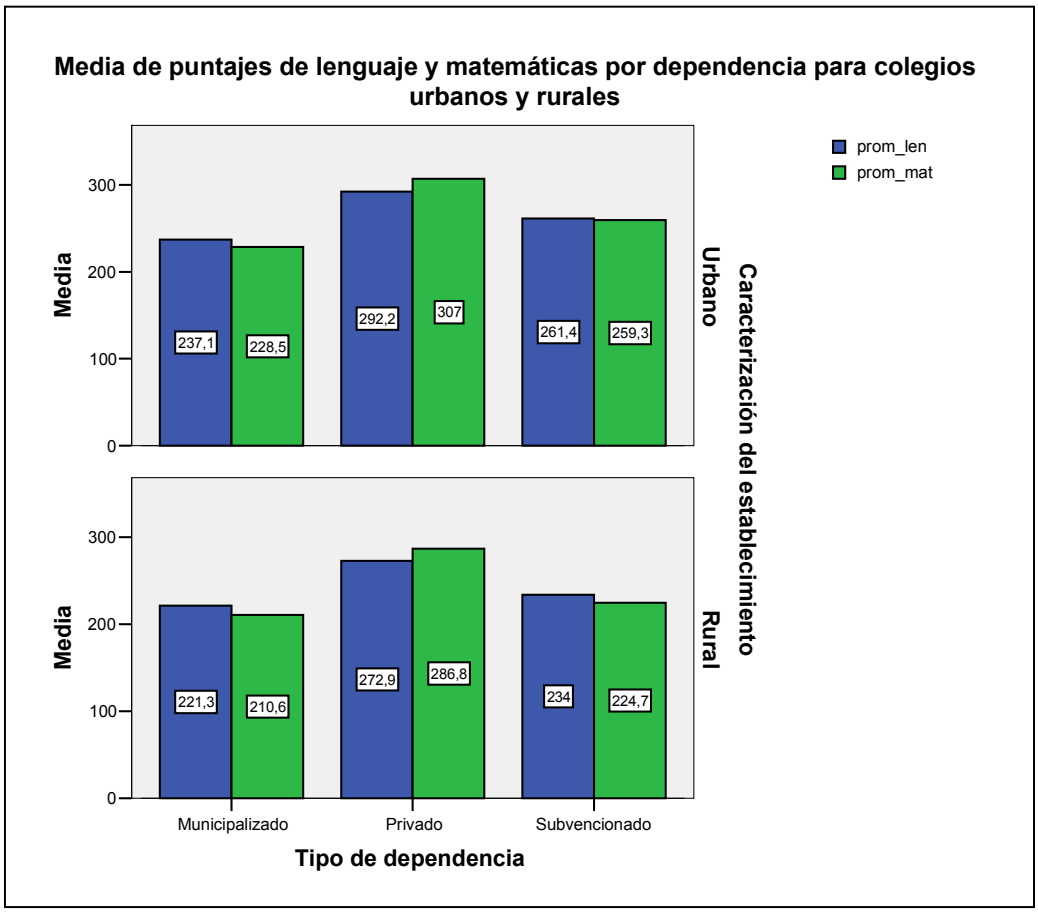
```
GRAPH
  /BAR(GROUPED)=MEAN(prom_len) BY dependencia BY caract
```



Se observa una clara diferencia entre los grupos con caracterización rural y urbana, notándose en los tres grupos de dependencia un promedio en lenguaje bastante mas bajo en los colegios rurales en relación a los colegios urbanos.

Para comparar el comportamiento en los resultados de ambas pruebas por dependencia entre la caracterización de Rural y Urbano es conveniente hacer un gráfico en dos paneles. Para esto vamos a gráficos de barras agrupados y seleccionar resumen para distintas variables

```
GRAPH
  /BAR(GROUPED)=MEAN(prom_len) MEAN(prom_mat) BY dependencia
  /PANEL ROWVAR=caract ROWOP=CROSS
  /MISSING=LISTWISE .
```



El gráfico nos muestra un comportamiento muy similar por dependencia entre los urbanos y rurales en cuanto a que los puntajes promedios más altos se observan en los colegios privados, los puntajes intermedios en los subvencionados y los puntajes más bajos en los municipalizados.
En todos los casos los urbanos presentan mayor puntaje promedio si se compara con los puntajes de los colegios rurales.

h) Determine los estadísticos básicos de tendencia central, de dispersión y de forma de la distribución de los puntajes promedios SIMCE para los distintos grupos homogéneos, y muestre gráficamente la media de los puntajes promedios SIMCE por grupo homogéneo. Interprete los estadísticos para el grupo de municipalizados y rurales.

Para responder a esta pregunta hay varios posibles procedimientos.
Una forma posible es realizar un cubo OLAP para la variable puntprom con variable de agrupación grupo.

Analizar/ Reporte/ Cubo OLAP
Sintaxis del procedimiento

```
OLAP CUBES
  puntprom   BY grupo
  /CELLS=COUNT MEAN MEDIAN MIN MAX STDDEV SKEW KURT NPCT
  /TITLE='OLAP Cubes'.
```

Para dejar la forma de la tabla siguiente se puede editar el cubo, mover el pivote de grupo, al lado superior derecho, el pivote de estadísticos, al lado inferior izquierdo y el pivote variable al lado superior izquierdo.

Estadísticos descriptivos para el puntaje promedio entre matemáticas y lenguaje de la prueba SIMCE por grupo

puntaje promedio entre lenguaje y matemáticas

grupo	N	Media	Mediana	Mínimo	Máximo	Desv. típ.	Asimetría	Curtosis	% del total de N
Municipalizado y Rural	68	215,9265	215,7500	181,00	257,00	15,49103	,216	-,111	2,9%
Municipalizado y Urbano	587	232,8169	228,0000	185,00	348,00	27,32073	1,156	1,845	25,3%
Privado y Rural	19	279,8947	277,5000	213,00	351,50	43,41933	,129	-1,616	,8%
Privado y Urbano	388	299,6224	311,0000	187,50	357,50	34,06601	-,892	,017	16,7%
Subvencionado y Rural	86	229,3547	225,2500	181,00	322,00	29,31110	,971	1,015	3,7%
Subvencionado y Urbano	1171	260,3335	260,0000	179,00	337,00	33,95561	,026	-,856	50,5%
Total	2319	257,6511	252,0000	179,00	357,50	39,45891	,320	-,920	100,0%

- Una segunda forma de lograr esta tabla de forma inmediata es utilizando el menú Medias.

Analizar/ Comparar medias/ medias

Sintaxis del procedimiento

```
MEANS
  TABLES=puntprom  BY grupo
 /CELLS COUNT MEAN MEDIAN MIN MAX STDDEV SKEW KURT NPCT
```

Para dejar la misma forma de la tabla anterior se puede transponer filas y columnas desde el menú Pivotar una vez editada la tabla.

Informe

puntaje promedio entre lenguaje y matemáticas

grupo	N	Media	Mediana	Mínimo	Máximo	Desv. típ.	Asimetría	Curtosis	% del total de N
Municipalizado y Rural	68	215,9265	215,7500	181,00	257,00	15,49103	,216	-,111	2,9%
Municipalizado y Urbano	587	232,8169	228,0000	185,00	348,00	27,32073	1,156	1,845	25,3%
Privado y Rural	19	279,8947	277,5000	213,00	351,50	43,41933	,129	-1,616	,8%
Privado y Urbano	388	299,6224	311,0000	187,50	357,50	34,06601	-,892	,017	16,7%
Subvencionado y Rural	86	229,3547	225,2500	181,00	322,00	29,31110	,971	1,015	3,7%
Subvencionado y Urbano	1171	260,3335	260,0000	179,00	337,00	33,95561	,026	-,856	50,5%
Total	2319	257,6511	252,0000	179,00	357,50	39,45891	,320	-,920	100,0%

Una tercera forma sería segmentar el archivo previamente por la variable grupo y luego pedir en el menú frecuencias los estadísticos solicitados.

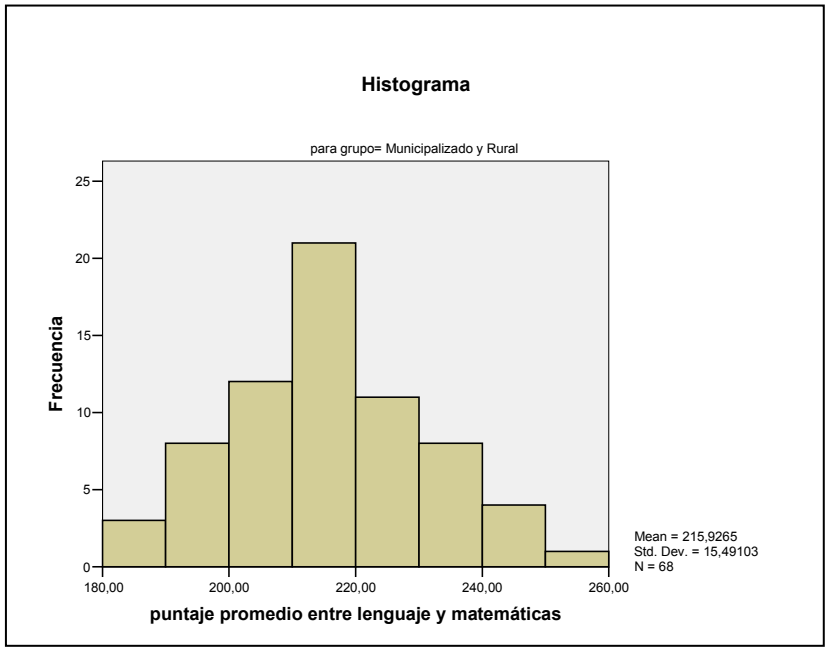
Sintaxis procedimiento 3

```
SORT CASES BY grupo .
SPLIT FILE
  LAYERED BY grupo .

SPLIT FILE
  LAYERED BY grupo .
FREQUENCIES
  VARIABLES=puntprom
  /STATISTICS=STDDEV MINIMUM MAXIMUM MEAN MEDIAN SKEW KURT
  /ORDER=  ANALYSIS .
```

De la tabla de estadísticos realizada en el punto anterior se puede observar que en promedio los puntajes de los colegios del grupo de municipalizados y rurales alcanzan un valor de 215,92 puntos, con una

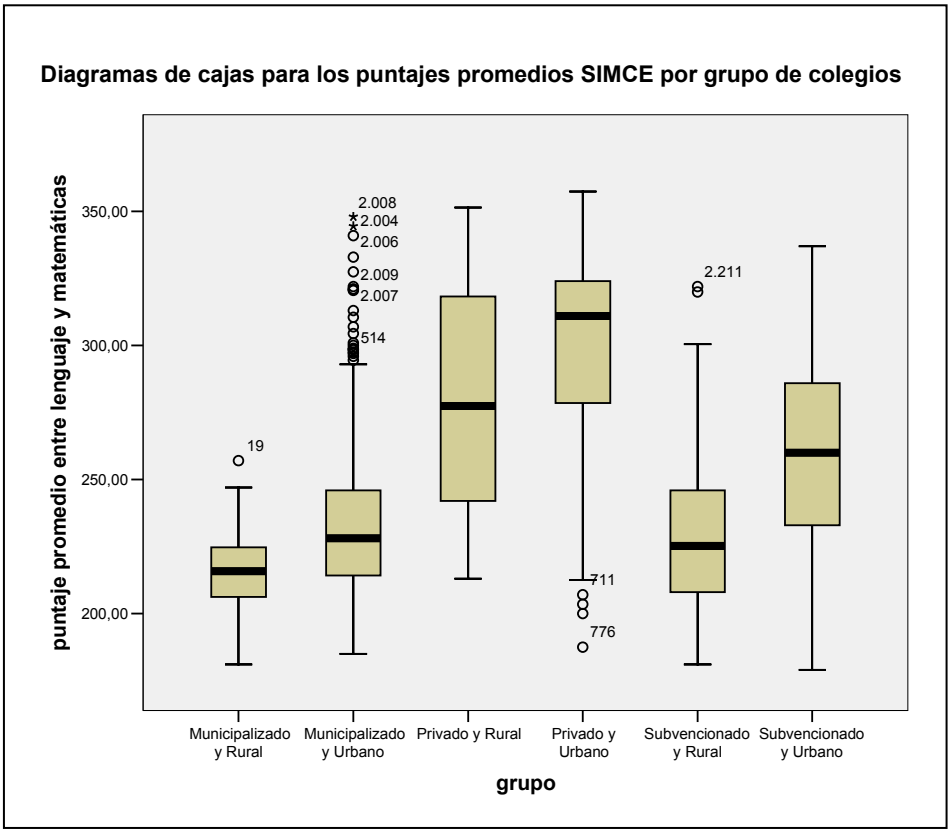
variabilidad relativamente baja de 15,49 puntos, es decir los puntajes se desvían en promedio respecto a la media en 15,49 puntos, calculado en base a 68 colegios, los que representan un 2,9% del total de colegios considerados en el estudio. Se observa que la mediana está cercana a la media de los datos (característica de distribuciones cercanas a distribuciones simétricas), e indica que el 50% de los colegios obtiene un puntaje menor o igual a 215,75 puntos, con puntajes que oscilan entre 181 puntos y 257 puntos, puntajes muy bajos si se compara con los demás grupos de colegios. Se observa la curtosis negativa y muy cercana a cero, esto significa que la curva es muy similar a la normal en cuanto a que se asemeja mucho a una curva mesocurtica (característica de la curva normal). La asimetría es positiva muy cercana a cero por tanto se asemeja bastante a una curva simétrica. En síntesis se puede apreciar una distribución con un buen ajuste a una normal. Esto se puede apreciar en el histograma correspondiente. (Para verificar el ajuste a una normal se puede ver el estadístico de Kolmogorov- Smirnov).



i) Realice un gráfico considerando a todos los colegios en estudio y otro gráfico considerando sólo el segmento de premiados, que permitan observar la posición del grupo en relación al resto de los grupos en cuanto a los descriptivos básicos. Interprete.

Para esto realizaremos un diagrama de cajas por grupo.

```
EXAMINE
  VARIABLES=puntprom BY grupo
  /PLOT BOXPLOT HISTOGRAM
  /COMPARE GROUP
  /STATISTICS NONE
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

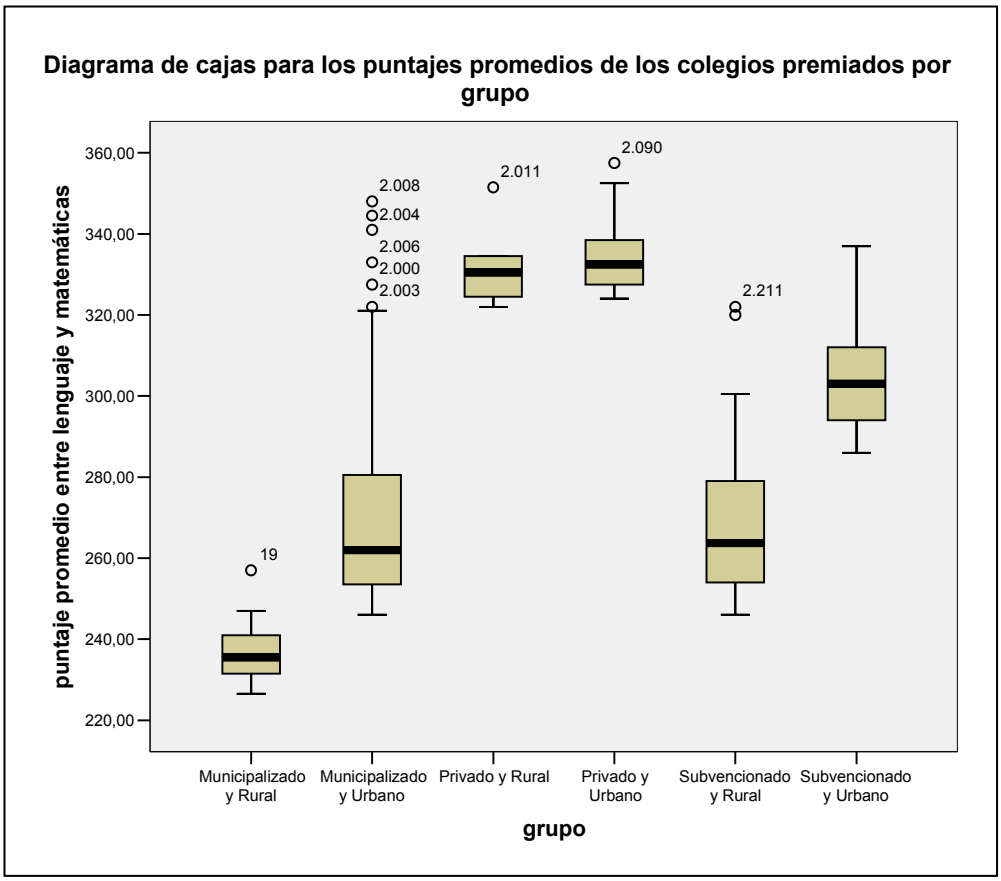


Del gráfico se puede apreciar que los mayores puntajes se concentran en los colegios privados, alcanzando la mayor mediana los colegios privados urbanos, aunque se observa que existen algunos colegios con puntajes extremos muy bajos. Por otra parte se puede ver que los puntajes más bajos se concentran en los colegios municipalizados y rurales presentando la menor de las medianas y es el grupo que presenta la menor variabilidad en los puntajes, es decir son bastante homogéneos en cuanto al resultado promedio SIMCE, mostrando sólo un caso sobre los 250 puntos en promedio. Se observa que el grupo de municipalizados urbanos a pesar de tener la mayoría de los colegios concentrados en puntajes bajos es el único grupo que presenta varios colegios con puntajes atípicos y un extremo en la parte alta de puntajes. Respecto a los subvencionados se ve clara diferencia entre los urbanos y rurales donde hay una marcada diferencia a puntajes más bajos en los colegios rurales.
Para obtener el diagrama de cajas para los premiados, primero se selecciona a los colegios que obtuvieron premio y una vez filtrada la base se procede a realizar la gráfica.

```
USE ALL.
COMPUTE filter_$=(premio = 1).
VARIABLE LABEL filter_$ 'premio = 1 (FILTER)'.
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .
```

Es claro que la posición de cada grupo en general se mantiene en relación a la gráfica analizada anteriormente. Los puntajes de los premiados con mayor puntaje se concentran en los privados, observándose muy concentrados, bastante homogéneos entre si. Sin embargo los puntajes de los premiados

con menor puntaje se concentran en los municipalizados rurales con un puntaje extremo pero que no pasa a la mediana del grupo de municipalizados urbanos. El grupo de los municipalizados urbanos premiados es el que presenta mas variabilidad en los puntajes, mostrando seis colegios con puntajes atípicos superiores en el rango de puntajes del grupo de los privados. El 50% de los colegios que cae en la caja del rango intercuartil (entre el percentil 25 y el percentil 75) del grupo de municipalizados urbanos se comporta de una manera muy similar a los colegios que se encuentran en la caja del rango intercuartil del grupo de subvencionados rurales. Se observa mas notoriamente la diferencia en los puntajes de los colegios premiados que pertenecen al grupo de los subvencionados rurales y los que pertenecen a los subvencionados urbanos.



Si consideramos la base de datos como si fuera la población de colegios que rinde la SIMCE, podríamos decir que claramente se observa diferencias significativas entre las medias de puntajes entre los diferentes grupos de colegios.

j) Compare la variabilidad entre los distintos grupos para el puntaje promedio SIMCE considerando todos los colegios en estudio.

Para comparar dispersión o variabilidad entre distribuciones diferentes se calcula el coeficiente de variación que se calcula en base a la desviación estándar y la media.
Indica la magnitud relativa de la desviación estándar en comparación con la media de la distribución, expresada como porcentaje.

Se calcula como.

$$C.V = \frac{\sigma}{\mu}.100$$

Para esto se puede pedir una tabla con los estadísticos básicos media y desviación estándar y luego en otra columna se puede calcular el coeficiente de variación en base a la fórmula anterior.

Informe

puntaje promedio entre lenguaje y matemáticas

grupo	N	Media	Desv. típ.	Coeficiente de Variación
Municipalizado y Rural	68	215,9265	15,49103	7,17
Municipalizado y Urbano	587	232,8169	27,32073	11,73
Privado y Rural	19	279,8947	43,41933	15,51
Privado y Urbano	388	299,6224	34,06601	11,37
Subvencionado y Rural	86	229,3547	29,31110	12,78
Subvencionado y Urbano	1171	260,3335	33,95561	13,04
Total	2319	257,6511	39,45891	15,31

De la tabla podemos observar que el grupo de colegios municipalizados y rurales es el que tiene menor variabilidad relativa, concepto que se pudo apreciar en los diagramas de caja. El grupo de los colegios que presentan mayor variabilidad en puntajes son los grupos "Privado y Rural" y el grupo "Subvencionado y Urbano".

VII. Gendarmería (Uso de Agregar y Sintaxis)[7]

Considere el archivo "Gendarmería.sav" correspondiente a datos de una muestra de internos del país. Se le pide realice un informe que describa información básica de esta muestra de internos, donde por lo menos debe responder las preguntas siguientes.

g) Identifique qué delito es el más frecuente en las mujeres y cuál en los hombres. Comente qué problema aprecia con los datos y explique cómo lo solucionaría.

h) Muestre una tabla y una sintaxis del procedimiento que le permite obtener la respuesta.

i) Crear un archivo que permita obtener por código de unidad penal y sexo; el promedio de edad de internos al iniciar la condena, región a la que pertenece la unidad penal, nombre de la unidad penal y número de internos. Muestre la sintaxis correspondiente.

j) Muestre una tabla que permita identificar por sexo las cinco unidades penales que más internos registran en la base de datos. Muestre la sintaxis

k) Crear un archivo que muestre sólo las unidades penales (por código de unidad penal) que registran hombres y mujeres. Muestre la sintaxis.

l) Crear un archivo que permita visualizar por nacionalidad y sexo; número de internos, pena máxima, pena mínima, y el promedio de pena en años. Mostrar la sintaxis.

m) Determine los estadísticos descriptivos para las variables edad al iniciar la condena y pena en años. ¿Cuál es el promedio de edad de internos hombres al iniciar la condena? ¿Cuál es el promedio de edad de internos mujeres de la Región Metropolitana al iniciar la condena? ¿Cuál es el mínimo y máximo de pena en años para hombres, y para mujeres? Muestre una tabla de donde se desprenda la respuesta.

Solución:

Si se realiza una tabla de contingencia de delito versus sexo o una tabla de frecuencia segmentada por sexo, se puede apreciar la falta de código por delito. Al digitar el delito en formato cadena (texto) se produce el problema de digitar de maneras distintas el mismo delito y al calcular las frecuencias no aparece correctamente el valor, dado que el SPSS los toma como delitos distintos. Lo que se debe realizar es una recodificación automática y luego limpiar los datos de la variable con recodificar en la misma variable (o en distinta variable) quedando la nueva variable con código numérico e identificando este a un único delito

Dada la gran cantidad de categorías de delito no es adecuado mostrar la tabla de contingencia o de frecuencias para responder a la pregunta. Una forma más eficiente es considerar Agregar por delito y sexo y luego pedir una tabla de valores extremos

```
AGGREGATE
 /OUTFILE='D:\SPSS-MGPP 2004\GUIAS 2004\AGRdelito sexo.sav'
 /BREAK=delito_1 sexo
 /N_BREAK=N.
```

[7]Caso elaborado por Sara Arancibia

```
EXAMINE
 VARIABLES=n_break BY sexo /ID= delito_1
 /PLOT BOXPLOT STEMLEAF
 /COMPARE GROUP
 /STATISTICS EXTREME
 /MISSING LISTWISE
 /NOTOTAL.
```

Extreme Values[b]

N_BREAK

Highest

Sexo del interno		Case Number	Primera mención de delito	Value
F	1	69	TRAFICO ILEGAL DE ESTUPEFACIENTES	24
	2	58	ROBO CON VIOLENCIA	5
	3	55	ROBO CON INTIMIDACION	4
M	1	56	ROBO CON INTIMIDACION	200
	2	70	TRAFICO ILEGAL DE ESTUPEFACIENTES	133
	3	53	ROBO CON FUERZA	114
	4	59	ROBO CON VIOLENCIA	64
	5	51	ROBO	55

b. The requested number of extreme values exceeds the number of data points. A smaller number of extremes is displayed.

```
AUTORECODE
 VARIABLES=u_.penal /INTO upenal
 /PRINT.
AGGREGATE
 /OUTFILE='D:\MGPP2004\SPSS\AGR1.sav'
 /BREAK=cod._uni sexo
 /edad_c_1 = MEAN(edad_cum) /region_1 = FIRST(region) /upenal_1 =
 FIRST(upenal)
 /casos=N.

SORT CASES BY sexo .
SPLIT FILE
 LAYERED BY sexo .
EXAMINE
 VARIABLES=casos /ID= upenal_1
 /PLOT BOXPLOT STEMLEAF
 /COMPARE GROUP
 /STATISTICS EXTREME
 /MISSING LISTWISE
 /NOTOTAL.
```

Valores extremos

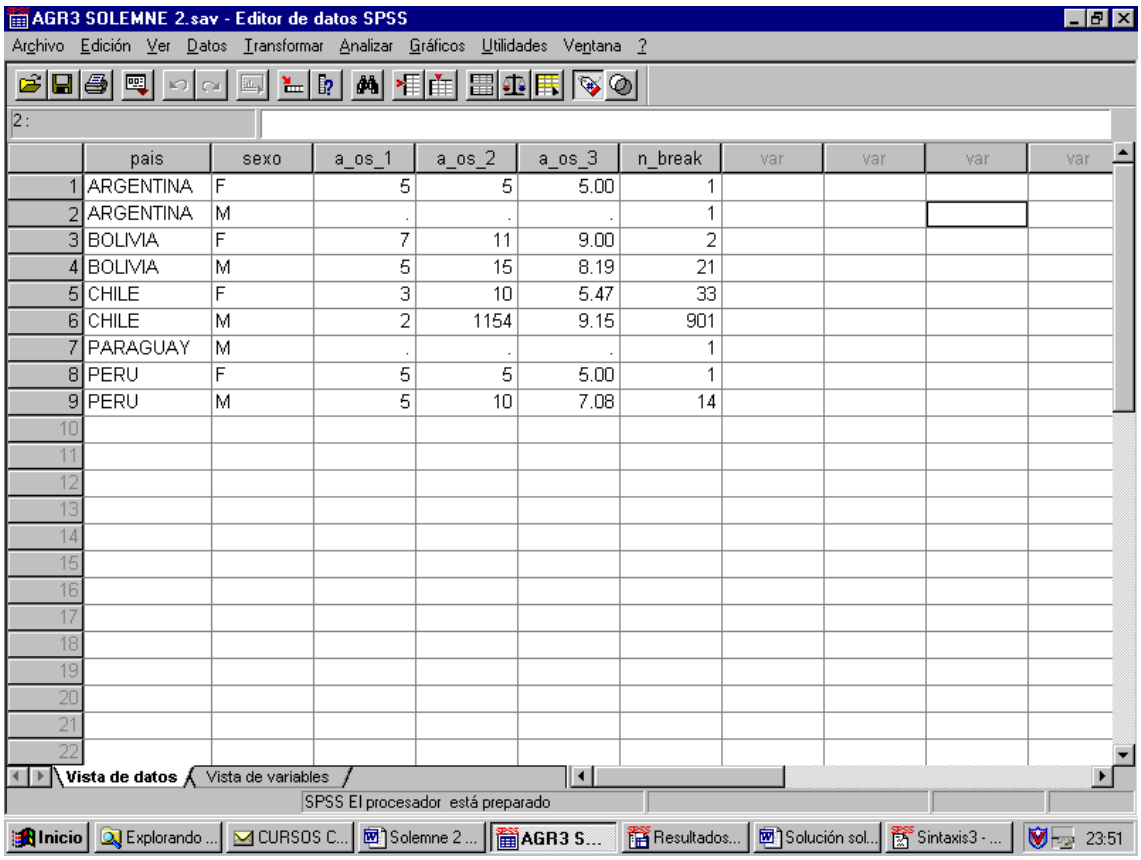
Sexo del interno				Número del caso	Nombre de la unidad	Valor
F	CASOS	Mayores	1	17	CPF. SANTIAGO	11
			2	1	CP. ARICA	5
			3	2	CCP. IQUIQUE	5
			4	11	CCP. CONCEPCION	3
			5	6	CDP. OVALLE	.[a]
		Menores	1	14	CDP. CASTRO	1
			2	7	CDP. ILLAPEL	1
			3	5	CCP. COPIAPO	1
			4	3	CDP. CALAMA	1
			5	10	CDP. LOS ANGELES	.[b]
M	CASOS	Mayores	1	94	CDP. SANTIAGO SUR	130
			2	18	CP. ARICA	63
			3	97	CCP. COLINA I	63
			4	39	CP. VALPARAISO	54
			5	101	CCP. COLINA II	49
		Menores	1	92	CDP. PORVENIR	1
			2	67	CET. CONCEPCION	1
			3	89	CDP. PUERTO AYSEN	1
			4	102	CET. METROPOLITANO	1
			5	46	CCP. SAN FERNANDO	.[b]

a. En la tabla de valores extremos mayores sólo se muestra una lista parcial de los casos con el valor 1.

b. En la tabla de valores extremos menores sólo se muestra una lista parcial de los casos con el valor 1.

```
AGGREGATE
 /OUTFILE='D:\MGPP2004\SPSS\AGR2 .sav'
 /BREAK=cod._uni
 /N_BREAK=N.
USE ALL.
COMPUTE filter_$=(n_break = 2).
VARIABLE LABEL filter_$ 'n_break = 2 (FILTER)'.
VALUE LABELS filter_$ 0 'No seleccionado' 1 'Seleccionado'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .

AGGREGATE
 /OUTFILE='D:\MGPP 2004\SPSS\AGR3 .sav'
 /BREAK=pais sexo
 /a_os_1 = MIN(a_os) /a_os_2 = MAX(a_os) /a_os_3 = MEAN(a_os)
 /N_BREAK=N.
```




```
*** Generación de Cubos OLAP *** .
OLAP CUBES
 edad_cum a_os BY region BY sexo BY pais
 /CELLS=COUNT MEAN STDDEV MEDIAN MIN MAX NPCT
 /TITLE='OLAP Cubes'.
```

OLAP Cubes

Región de la unidad: Total
Sexo del interno: Total
NOmbre del país: Total

	Edad al iniciar condena	Pena en años
N	973	649
Mean	21,79	8,91
Std. Deviation	16,22	45,15
Median	24,00	5,00
Minimum	0	2
Maximum	70	1154
% of Total N	100,0%	100,0%

OLAP Cubes

Región de la unidad: Total
Sexo del interno: M
NOmbre del país: Total

	Edad al iniciar condena	Pena en años
N	936	615
Mean	21,51	9,09
Std. Deviation	16,09	46,38
Median	24,00	5,00
Minimum	0	2
Maximum	67	1154
% of Total N	96,2%	94,8%

OLAP Cubes

Región de la unidad: 13
Sexo del interno: M
NOmbre del país: Total

	Edad al iniciar condena	Pena en años
N	313	161
Mean	13,81	7,06
Std. Deviation	15,68	3,12
Median	,00	5,00
Minimum	0	3
Maximum	64	20
% of Total N	32,2%	24,8%

OLAP Cubes

Región de la unidad: Total
Sexo del interno: F
NOmbre del país: Total

	Edad al iniciar condena	Pena en años
N	37	34
Mean	28,95	5,65
Std. Deviation	18,13	2,04
Median	27,00	5,00
Minimum	0	3
Maximum	70	11
% of Total N	3,8%	5,2%

OLAP Cubes

Región de la unidad: 13
Sexo del interno: F
NOmbre del país: CHILE

	Edad al iniciar condena	Pena en años
N	11	9
Mean	32,55	5,33
Std. Deviation	18,53	1,87
Median	27,00	5,00
Minimum	0	3
Maximum	70	10
% of Total N	1,1%	1,4%

OLAP Cubes

Región de la unidad: 13
Sexo del interno: M
NOmbre del país: CHILE

	Edad al iniciar condena	Pena en años
N	302	158
Mean	14,04	7,08
Std. Deviation	15,74	3,15
Median	,00	5,00
Minimum	0	3
Maximum	64	20
% of Total N	31,0%	24,3%

OLAP Cubes

Región de la unidad: Total
Sexo del interno: M
NOmbre del país: CHILE

	Edad al iniciar condena	Pena en años
N	899	587
Mean	21,41	9,15
Std. Deviation	16,07	47,47
Median	23,00	5,00
Minimum	0	2
Maximum	67	1154
% of Total N	92,4%	90,4%

OLAP Cubes

Región de la unidad: Total
Sexo del interno: F
NOmbre del país: CHILE

	Edad al iniciar condena	Pena en años
N	33	30
Mean	27,39	5,47
Std. Deviation	18,51	1,91
Median	26,00	5,00
Minimum	0	3
Maximum	70	10
% of Total N	3,4%	4,6%

VIII Resultados de Colegios (Fundición de Archivos)[8]

Considere los "*colegios arica.sav*" y "*alumnos arica.sav*" que tienen las siguientes variables:

arica.sav
código del colegio,
nombre del colegio,
región,
comuna,
modalidad,
dependencia,
número de alumnos.

alumnos arica.sav
Identificación alumno,
código colegio,
puntaje mat 98,
puntaje leng 98.

Responda las siguientes preguntas

a) ¿Qué porcentaje de alumnos representa el colegio con mayor número de alumnos?

b) ¿Cuántos alumnos de la muestra están en colegios Municipal, P Subvencionado y P. Pagado?

c) ¿A qué colegios pertenecen los alumnos con los mayores puntajes promedio(mat y leng 98)?

d) ¿A qué colegios pertenecen los alumnos con los mayores puntajes en matemática segmentado por dependencia?

e) ¿A qué colegio municipalizado pertenece el alumno con mayor puntaje en lenguaje?

f) Determine la media, mediana, desv típica, mínimo, máximo de los puntajes de matemática y lenguaje 98.

   i. Determine por modalidad de colegio la media, mediana, desv típica, mínimo, máximo de los puntajes de matemática y lenguaje 98

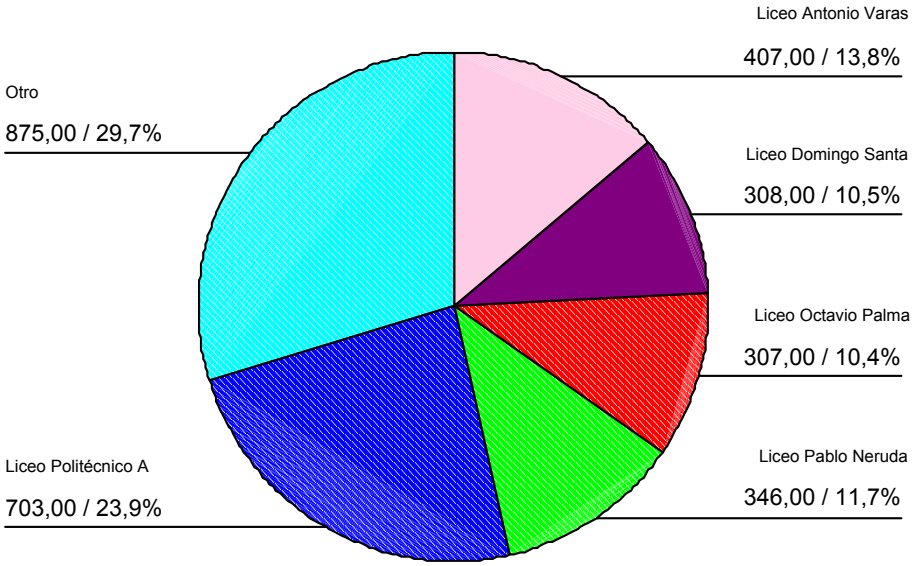
   ii. Determine por dependencia de colegio la media, mediana, desv típica, mínimo, máximo de los puntajes promedio de matemática y lenguaje 98

Solución:

a) A partir del archivo *colegios arica.sav* podemos hacer un gráfico de sectores donde los datos del gráfico son valores individuales de los casos, los sectores representan número de alumnos (nalum) y en las etiquetas de los sectores se indica el nombre del colegio.

Al editar el gráfico y agrupar los sectores menores que 10% se obtiene:

[8] Caso elaborado por Sara Arancibia



iii. Para responder esta pregunta debemos fundir los dos archivos.

Para esto, abrir el archivo "alumnos arica.sav" y seleccionar Datos/fundir archivos/añadir variables/leer archivo: "colegios arica".

Activar [Emparejar los casos en las variables clave para los archivos ordenados] y [El archivo externo es una tabla de claves].

Variable clave: Código.

Nota: Previamente hay que preocuparse de ordenar la variable clave en forma ascendente en ambos archivos

A partir del archivo nuevo se pide una tabla de frecuencia de la variable dependencia

Dependencia

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Municipal	60	39,0	39,0	39,0
	P.Subvencionado	55	35,7	35,7	74,7
	P.Pagado	39	25,3	25,3	100,0
	Total	154	100,0	100,0	

iv. Para determinar a qué colegio pertenecen los alumnos con mayores puntajes en matemáticas podemos pedir en explorar, variable; puntaje de mat 98, etiquetar por nombre del colegio. Estadísticos valores extremos.

v. A h)TAREA

IX Estudio de caso: Encuesta laboral (Aplicación IPC)

Se realizó una encuesta dirigida a personas activas en el ámbito laboral, con el objeto de conocer algunas características de ésta y conocer qué ha sucedido con la evolución de los sueldos. Para tal efecto, se tomó una muestra de 470 personas activas. Los datos entregados por los entrevistados se depositaron en el archivo caso encuesta laboral (aplicación IPC).
Usando los antecedentes que residen en la base de datos antes señalada y los datos de IPC anual (base Dic 1998=100) responda las siguientes preguntas. Considere

Año	IPC
1993	71,68
1994	78,09
1995	84,49
1996	90,10
1997	95,54
1998	100,00
1999	102,31
2000	106,94
2001	109,76
2002	112,86
2003	114,07
2004	116,84

a) Elaborar un cuadro que muestre los siguientes estadísticos descriptivos: Mínimo, Máximo, Media, Mediana y desviación estándar, del sueldo promedio inicial, del sueldo promedio inicial expresado en pesos del año 2004 y sueldo promedio actual por categoría laboral

```
***Sintaxis IPC***.
IF (contrato = 93) IPC = 71.68 .
IF (contrato = 94) IPC = 78.09 .
IF (contrato = 95) IPC = 84.49 .
IF (contrato = 96) IPC = 90.10 .
IF (contrato = 97) IPC = 95.54 .
IF (contrato = 98) IPC = 100 .
IF (contrato = 99) IPC = 102.31 .
IF (contrato = 2000) IPC = 106.94 .
IF (contrato = 2001) IPC = 106.94 .
IF (contrato = 2002) IPC = 112.86 .
IF (contrato = 2003) IPC = 114.07 .
IF (contrato = 2004) IPC = 116.84 .
EXECUTE .
```

Sintaxis Utilizada para Actualización del sueldo Inicial en moneda de dic del 2004

```
COMPUTE suel2004 = (sueldini / IPC) * 116.84 .
EXECUTE .
```

```
MEANS
```

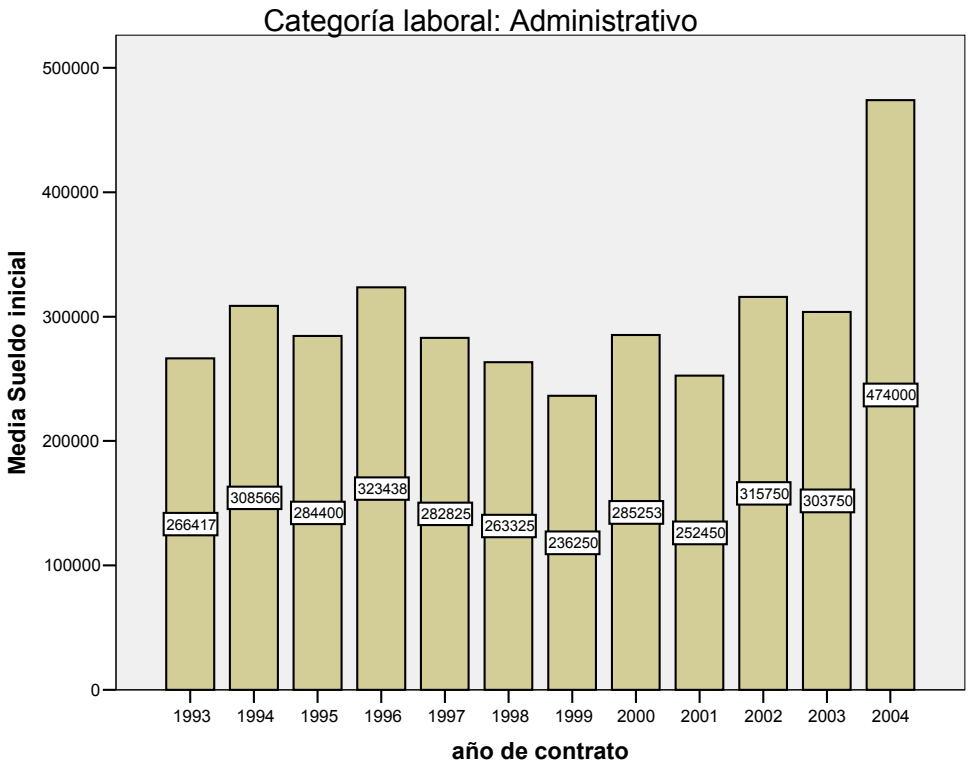
```
TABLES=sueldini sueld suel2004  BY catlab
/CELLS COUNT MIN MAX MEAN MEDIAN STDDEV  .
```

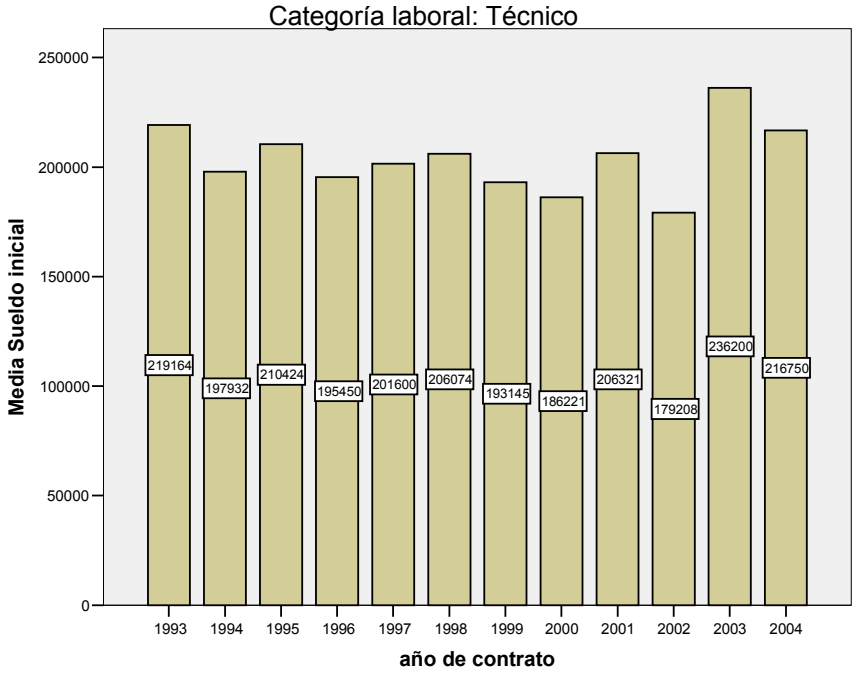
Informe

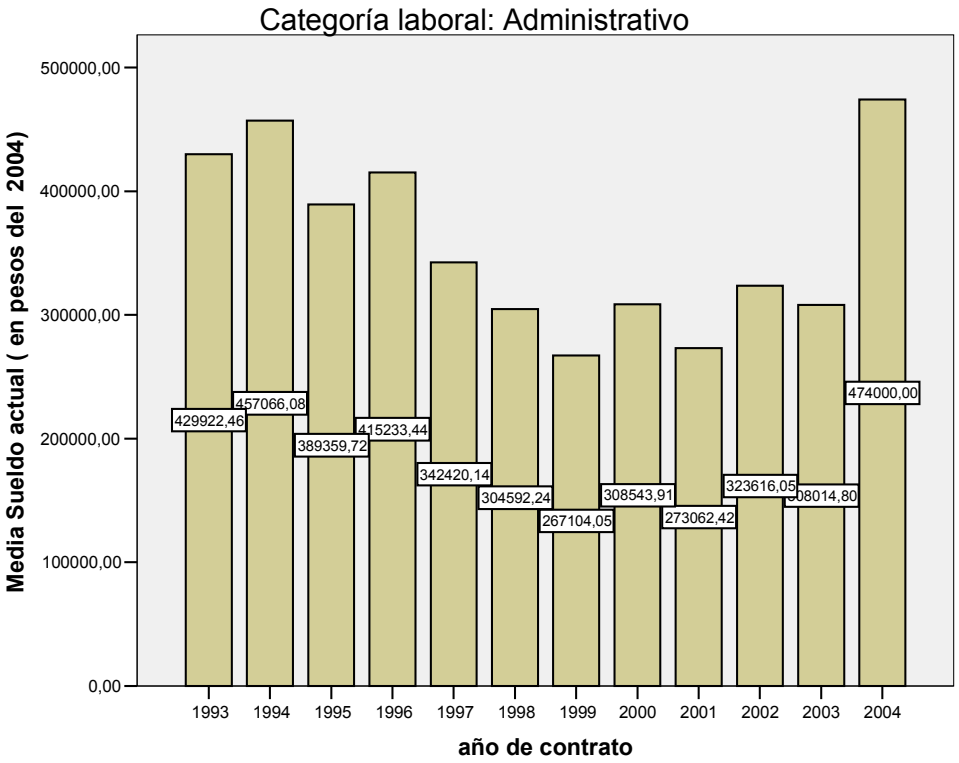
Categoría laboral		Sueldo inicial	Sueldo actual (en pesos del 2004)	sueldo inicial en pesos del 2004
Administrativo	N	142	142	142
	Mínimo	166500	227146,40	229440,81
	Máximo	570000	846719,98	855272,71
	Media	287285,56	380863,4484	384609,4015
	Mediana	268000,00	366611,8709	370315,0211
	Desv. típ.	78070,001	120759,16120	121893,4700
Técnico	N	285	285	285
	Mínimo	135000	136965,55	139760,77
	Máximo	300000	467245,90	476781,53
	Media	201469,74	272755,6075	278291,0066
	Mediana	202500,00	263933,6151	269320,0154
	Desv. típ.	34432,401	61964,72683	63258,53212
Directivo	N	43	43	43
	Mínimo	236250	336507,91	326706,71
	Máximo	1199700	1386998,08	1346600,08
	Media	535590,70	765540,9758	743243,6659
	Mediana	495000,00	751059,4599	729183,9416
	Desv. típ.	163393,602	237648,82763	230727,0171
Total	N	470	470	470
	Mínimo	135000	136965,55	139760,77
	Máximo	1199700	1386998,08	1346600,08
	Media	257965,59	350502,5952	352950,9564
	Mediana	225000,00	304926,2635	311149,2484
	Desv. típ.	119195,691	177517,26027	171214,0115

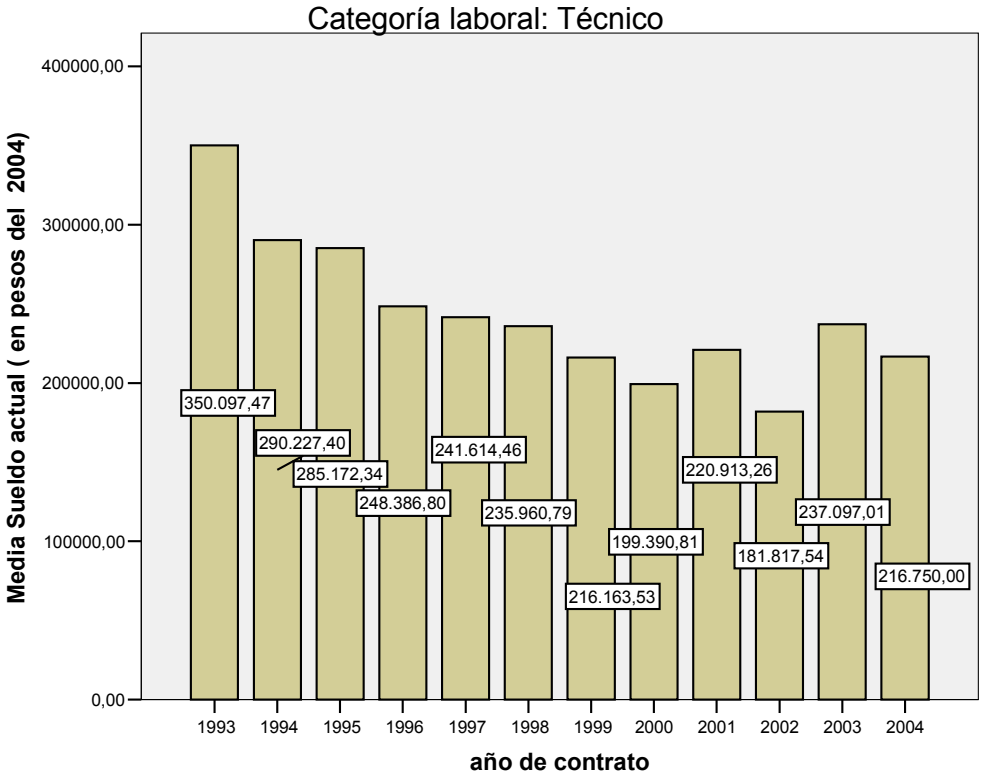
B) Considere por separado cada categoría laboral . Realice un gráfico para la media de los sueldos iniciales (moneda nominal) por año de contrato. Luego realice un gráfico para la media de los sueldos actuales .

```
IF (contrato < 2000) contrat = contrato + 1900 .
VARIABLE LABELS contrat 'año de contrato' .
EXECUTE .
IF (contrato  >=  2000) contrat = contrato  .
VARIABLE LABELS contrat 'año de contrato' .
EXECUTE .
```


Categoría laboral: Administrativo
Media Sueldo inicial
500000
400000
300000
200000
100000
0
266417
308566
284400
323438
282825
263325
236250
285253
252450
315750
303750
474000
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
año de contrato


Categoría laboral: Técnico
Media Sueldo inicial
250000
200000
150000
100000
50000
0
219164
197932
210424
195450
201600
206074
193145
186221
206321
179208
236200
216750
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
año de contrato


Categoría laboral: Administrativo
Media Sueldo actual (en pesos del 2004)
500000,00
400000,00
300000,00
200000,00
100000,00
0,00
429922,46
457066,08
389359,72
415233,44
342420,14
304592,24
267104,05
308543,91
273062,42
323616,05
308014,80
474000,00
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
año de contrato


Categoría laboral: Técnico
Media Sueldo actual (en pesos del 2004)
400000,00
300000,00
200000,00
100000,00
0,00
350.097,47
290.227,40
285.172,34
248.386,80
241.614,46
235.960,79
216.163,53
199.390,81
220.913,26
181.817,54
237.097,01
216.750,00
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
año de contrato