EL DISEÑO CUANTITATIVO
UNA ESTRATEGIA METODOLOGICA PARA EL MANEJO DE LA INFORMACION CUANTITATIVA.

Abstract

The purpose of this module is to present a research process that is supported basically in a methodological design that uses a quantitative survey technique social and basic technique that uses a statistical package for the processing of the data collected, cleansed and validated.

Here is the investigative process in form of stages, which is justified in view of the pedagogical purpose of this module. This presentation is intended to provide a guide to the methodology of the research of this kind, involving all the steps and show the importance of each stage, since the results of each stage are a consequence of the precedent.

Resumen

El objetivo de este Módulo es presentar un proceso de investigación que se apoye básicamente en un diseño metodológico de tipo cuantitativo que utilice la técnica de la encuesta social como técnica básica y que se sirve de un paquete estadístico para el procesamiento de los datos recolectados, depurados y validados.

Aquí se presenta el proceso investigativo en forma etápica, lo cual se justifica teniendo en cuenta el fin pedagógico de este Módulo. Con esta presentación se pretende dar una guía de la metodología de las investigaciones de este tipo, que involucre todos los pasos y muestre

la importancia de cada etapa, dado que los resultados de cada fase son consecuencia de la precedente.

INDICE

INTRODUCCION AL MODULO.

Una buena técnica de enseñanza - aprendizaje es la asociación de la presentación de un tema con ejemplificaciones que apliquen y aclaren los conceptos vertidos en ella, más la posterior

realización de prácticas por parte del participante. Por esta razón en este Módulo se establece una dinámica de Presentación - Ejemplificación - Actividades y Autoevaluaciones.

El objetivo de este Módulo es presentar un proceso de investigación que se apoye básicamente en un diseño metodológico de tipo cuantitativo que utilice la técnica de la encuesta social como técnica básica y que se sirve de un paquete estadístico para el procesamiento de los datos recolectados, depurados y validados.

Aquí se presenta el proceso investigativo en forma etápica, lo cual se justifica teniendo en cuenta el fin pedagógico de este Módulo. Con esta presentación se pretende dar una guía de la metodología de las investigaciones de este tipo, que involucre todos los pasos y muestre la importancia de cada etapa, dado que los resultados de cada fase son consecuencia de la precedente.

¿Qué es un Diseño Cuantitativo? Puede definirse al Diseño Cuantitativo como una estrategia metodológica que permite manejar - trabajar datos cuantificables (medibles). Permite una aproximación sistemática al estudio de hechos sociales "apoyándose" preferentemente en categorías numéricas y realiza el análisis a través de diferentes formas de interrelacionar estadísticamente esas categorías numéricas.

1.1 PLANTEAMIENTO DE UN PROBLEMA DE INVESTIGACION.

Para que el ejercicio de práctica sea de fácil comprensión y, además, represente una situación real, se eligió como problema de investigación: *"El Comportamiento Lector Del Estudiante De La Facultad De Ciencias Sociales Humanas"*. Este estudio engloba la

definición del entorno socioeconómico de los estudiantes; sus opiniones respecto a su "actividad lectora" (frecuencia, duración y contenido); sus "intereses de lectura" (opiniones frente a la lectura, sus medios y fines) y finalmente los "determinantes de la lectura" (características sociodemográfica del lector, usos del tiempo libre, entorno cultural y académico).

A fin de recolectar la información se aplicó un cuestionario a los estudiantes matriculados en los 5 programas de pregrado de la Facultad de Ciencias Sociales y Humanas en el Semestre II del año 1995, de acuerdo a un diseño muestral, que permitió una selección de una población estratificada y proporcional por nivel académico.

1.2. DETERMINACIÓN DE LOS OBJETIVOS DEL ESTUDIO

En esta etapa se determina lo que se desea obtener como información por medio de la investigación. Es la de mayor importancia, dado que rige a las demás.

La importancia de esta fase del proceso esta dada por el hecho de que aquí se define qué se desea tener como resultado final a fin de establecer qué se necesita para obtener esos resultados; es decir, los datos que se almacenan en el paquete STATGRAPHICS y el tipo de proceso que se les aplique dependen de lo que se busque.

Los objetivos meteorológicos se pueden clasificar en dos tipos: **Cualitativos** (establecen cualidades) y **Cuantitativos** (proporcionan cantidades).

En la situación del estudio señalado los objetivos del estudio se pueden resumir en los siguientes apartes:

a) **Características Generales Del Entorno Sociodemográfico Y Cultural De Los Estudiantes**:
Programa académico, desagregación por nivel, sexo y estratificación social.

b) **Actividad Lectora Del Estudiante**: Averiguar sobre las temáticas preferidas, los medios de lectura utilizados, frecuencia y duración de la lectura; los objetivos de la misma y por ultimo la opinión acerca del proceso de leer, los medios y fines.

c) **Determinantes de la Lectura:** Uso del tiempo libre, el entorno cultural y académico.

1.3. El DISEÑO OPERACIONAL DEL ESTUDIO.

LA DETERMINACIÓN DE LA POBLACION MUESTRAL; LA RECOLECCION, PROCESAMIENTO Y ANÁLISIS DE LA INFORMACIÓN

1.3.1. EL DISEÑO MUESTRAL.

1.3.1. 1 La encuesta por muestreo

Muchos trabajos de investigación requieren de procesos de recolección de información ya sea de poblaciones totales o de muestras poblacionales. Aquí nos detendremos a analizar la encuesta por muestreo y, específicamente sobre el método de muestreo.

Se llama "encuesta por muestreo" cuando sólo se aplica a una fracción representativa de una población total. Al aplicar el muestreo, lo que se busca es lograr que registrando una porción relativamente pequeña de unidades se puedan obtener conclusiones semejantes a las que se lograrían si se estudiase la población total.

Cuando una muestra cumple con las condiciones de reflejar las características de la población total, se habla de "muestras representativas". En relación con el diseño de la muestra deben adoptarse dos criterios:

1) Cuál será el universo o población del estudio

2) El tamaño y el diseño de la muestra que debe extraerse (características).

Una vez decidido estos dos puntos, se cumple el proceso de obtener las "unidades" de la muestra y la preparación del proceso de aplicación del instrumento de recolección de datos.

Tradicionalmente se habla de *"Muestras Probabilísticas" y "No - Probabilísticas*". En las "probabilísticas" la característica básica es que toda unidad de observación tiene una determinada probabilidad de integrar la muestra. Otro tipo de muestra son las "intencionadas" o arbitrarias, por ejemplo "muestra por cuotas", en donde se fija a priori, la cantidad de unidades o elementos de cada categoría que habrán de integrarla: en una muestra de 100 personas, se asigna una cuota de 53 hombres y 47 mujeres.

Para que la muestra pueda ser tratada estadísticamente debe ser "aleatoria" o "aleatoria simple".

A continuación se describen brevemente los tipos más usuales de diseño muestral:

Muestreo Aleatorio Simple. Es aquel en que cada individuo dentro de una población tiene la misma probabilidad de ser seleccionado como integrante de la muestra. Así, en una población de tamaño N. cada individuo tiene un probabilidad de ser elegido de 1/N. Los elementos de la muestra se eligen en *forma aleatoria*, por ejemplo: Se desea tomar 10 estudiantes como muestra representativa de un grupo de 50. Hay dos maneras de hacerlo en forma aleatoria:

a) Por medio de un sorteo, por ejemplo, en donde se escriban papeletas con los nombres de los estudiantes, se mezclan y de entre éstos se eligen 10;

b) Escoger 10 números de una tabla de números aleatorios y buscar a Qué estudiantes corresponden utilizando la lista de asistencias.

Muestreo Sistemático. En este método los elementos se escogen estableciendo y aplicando un criterio de selección uniforme; por ejemplo: En el caso anterior (elección de 10 estudiantes de entre un grupo de 50) se pueden utilizar dos formas de proceder sistemáticamente:

a) Dado que son 50 estudiantes en el grupo, se toma uno de cada cinco según estén clasificados por puntajes o promedios de nota;

b) Escoger el 5, el 10, el 15, etc., según estén anotados en el listado de Control de Asistencia y Calificaciones del Curso.

Muestreo Estratificado. Para este tipo de muestreo, la población se divide en grupos (estratos) con base en algún criterio, como: estrato socioeconómico, sexo, número de créditos aprobados, etc. Los elementos muéstrales se escogen al azar dentro de cada estrato. En el caso particular de que el número de elementos que se escoja de cada estrato sea proporcional al tamaño del estrato, se tiene el caso de ***muestreo proporcional***. Este tipo de muestreo es el más utilizado, dado que las particularidades de la población pueden influir en la determinación de los elementos que integran la muestra. En el ejemplo presentado, una muestra de 10 estudiantes entre un grupo de 50, se puede establecer una estratificación proporcional en función del número de créditos aprobados; así se tendría un estrato o nivel de Estudiantes Iniciales; un segundo estrato, con los estudiantes de niveles intermedios y un tercer estrato de los estudiantes de niveles terminales. El número de estudiantes de cada estrato es de 15, 30 y 5 respectivamente; así se escogen aleatoriamente 3, 6 y 1 estudiantes de los niveles correspondientes.

Muestreo por Conglomerados. En este esquema de muestreo, también se subdivide la población en grupos, y estos se denominan *conglomerados*. La diferencia del muestreo por **conglomerados** con respecto al **estratificado** es que para el primero se deben censar todos los integrantes de los conglomerados escogidos y no sólo algunos elegidos al azar. Los conglomerados se pueden seleccionar en forma aleatoria o con probabilidad proporcional al tamaño.

CALCULO DE TAMAÑO DE MUESTRA.

Para el cálculo del tamaño de la muestra es necesario definir el nivel de confiabilidad estadístico que se desea del estudio y el error máximo de estimación de la variable que se esta dispuesto a asumir. Con respecto a la confiabilidad, la práctica común es considerarla en niveles que oscilan entre el 90 % y el 99 % (probabilidad de aceptar eventos como ciertos cuando en realidad no lo son) ; el valor seleccionado en este rango, depende del costo que se quiera asumir en términos del mayor o menor valor del tamaño de la muestra que se quiera determinar. En el siguiente anexo se presenta una simulación del impacto combinado de 3 niveles de confiabilidad y 10 niveles de error máximos de estimación, sobre el tamaño de muestra.

En relación con el error máximo de estimación, o sea, la diferencia máxima que se esta dispuesto a admitir entre el verdadero valor del promedio de la variable seleccionada y su valor estimado a través del muestreo.

EJEMPLIFICACIÓN: Con base en los datos de la muestra piloto del estudio Socioeconómico de la Transversal Intermedia se calculó una matriz de tamaños de muestra de la población a encuestar seleccionada por muestreo aleatorio estratificado[1].

Error máximo de	**NIVEL DE CONFIABILIDAD**

estimación	**90 %**	**95 %**	**99 %**
	Nº n	Nº n	Nº n
10 %	161	228	390
9 %	198	280	480
8 %	250	354	604
7 %	326	460	783
6 %	441	622	1.054
5 %	**631**	**886**	**1.489**
4 %	971	1.356	2.248
3 %	1.674	2.309	3.726
2 %	3.465	4.640	7.026
1 %	9.684	11.764	14.992

EJEMPLIFICACION.

Diseño muestral del estudio "*Comportamiento Lector de los Estudiantes de la Facultad de Ciencias Sociales y Humanas de la Universidad de Antioquia"*.

Uno debe preguntarse ¿Cual es el número mínimo de unidades de análisis (estudiantes) que se necesitan para construir una muestra (n) que nos asegure un error estándar menor de .01

[1] Botero, Luis Fernando. Análisis estadístico de los datos suministrados por la prueba piloto del estudio

(preestablecido por el equipo investigador) dado que la población estudiantil de la Facultad es de 1.330 estudiantes registrados.

EJEMPLIFICACION.

1. DELIMITACION DE LA POBLACION:

Población de estudiantes registrados: **N = 1.330**

2. ¿Cual es el nº de estudiantes de la Facultad de Ciencias Sociales y Humanas a los cuales hay que aplicarle el cuestionario para tener un error estándar menor de 0.015 , dado que la población (N) total de estudiantes es de 1.330 personas.

N = 1.330 estudiantes

Se = .015 - **1.5 %-** (**Tamaño del error de nuestras predicciones**)

V 2 = $(Se)^2$ = 0.0025

$$n' = S^2 / V^2$$

$$\mathbf{S^2} = p\,(1\text{-}p)^2 = .9\,(1 - .9) = \mathbf{.09}$$

$$\mathbf{V} = (.015)^2 = \mathbf{.000225}$$

ceo@catios.udea.edu.co
http://ceo.udea.edu.co
Ciudad Universitaria Bloque 9-252 Telefax: 2105775

n = .09 / .000225 = 36

n = n' / 1+n / N

n = 36 /1 + 36 /1.330

n' = 36 / 1.0008 = 35

n' = 225 / 1 + 225 / 1.330

n' = 225 / 1.1692 = **192**

En síntesis, el primer paso fue determinar la muestra probabilistica con base a los estimados de la población estudiantil. El segundo paso consiste en cómo y donde seleccionar a esos 192 estudiantes de los 5 programas de pregrado de la Facultad.

EJEMPLIFICACION 2.

Deteminar la población muestral para un estudio sobre integración y conflicto en la comunidad estudiantil en la Universidad de Antioquia. Para la obtención de la muestra del Ejercicio Investigativo se eligió **el muestreo por conglomerados** con los 17.000 estudiantes de la Universidad de Antioquia según programa académico. En este caso *cada programa académico es un conglomerado.*

El tamaño de la muestra del Ejercicio de Práctica se obtuvo de esta manera:

a) El tamaño de muestra para proporciones está dado por:

$$n_o = \frac{PQ}{V}$$

donde N_o = tamaño dela muestra

P = probabilidad de que se realice el evento

Q = probabilidad de que no se realice el evento

V = varianza del estimador

La varianza del estimador es:

$$V = \frac{d^2}{t}$$

donde: **d** = Error permitido en los datos

t = Valor de la abscisa en el eje X de una distribución normal tal que deje en la parte central un área igual a la confianza deseada.

Entonces:

$$n_o = \frac{t^2 P Q}{d^2}$$

b) Se establece que se requiere un 95% de confianza de que el error no sea mayor del 4%. Entonces, buscando en **La Tabla del Área Bajo la Curva Normal** se identifica que t = 1.96, o sea, aproximadamente 2.

c) Se estima que existe un 50% de probabilidad de que se efectúe el evento y , por lo tanto, un 50% de probabilidad de que no ocurra (Q = 100 - P). Se tiene al sustituir:

$$n_o = \frac{(4)\,(50)\,(50)}{16} = 625$$

c) Para poblaciones finitas se ajusta al tamaño de la muestra, pues en ocasiones resulta mayor que el de la población. El ajuste para poblaciones finitas esta dado por:

$$n = \frac{n_o}{1 + \frac{n_o}{N}}$$

donde: **n** = tamaño ajustado de la muestra

$\mathbf{n_o}$ = tamaño no ajustado de la muestra

N = tamaño de la población

Entonces:

$$n = - \frac{625}{1 + \frac{625}{119.000}} = 621.73 = 622$$

Distribucion de la población muestral en el estudio acerca del comportamiento lector del estudiante de la facultad de ciencias sociales y humanas:

Antropología	30
Historia	13
Sicología	40
Sociología	33
Trabajo Social	41

LA RECOLECCIÓN DE INFORMACIÓN.

En un proceso de investigativo hay dos etapas que se consideran cruciales con respecto a la calidad interna de los datos. Una es el diseño del instrumento de recolección y otra tiene que

ver con el "factor humano". La confiabilidad y validez de la información [2] puede verse en entredicho tanto desde el papel que juega el sujeto informador y el cuestionario como el desempeño del "encuestador". Situaciones que tradicionalmente son vista como secundarias en un trabajo investigativo. Sino que lo diga el escaso tiempo que se le dedica al trabajo de selección y formación de las personas encargadas de capturar un tipo de información.

En el caso del Ejercicio Investigativo que aquí se utiliza como "ejemplificación" el proceso operacional fue el siguiente: Una vez seleccionada la población muestral se procedió al proceso de captura de la información. En él se involucró a un equipo de encuestadores adiestrados[3] para desarrollar las dos etapas de la recolección (Prueba Piloto y aplicación del Cuestionario Definitivo a la población muestral). La validación de la información involucró un 5 por ciento del total de cuestionarios aplicados. Depurada la información esta fue codificada de acuerdo a un Libro de Codificación para luego ser grabada en una matriz de datos diseñada especialmente en el paquete STATGRAPHICS. La información tabulada, fue nuevamente depurada y presentada previamente en listados de distribución de frecuencias de valores del total de respuestas para ser analizada por el equipo de trabajo. A continuación se realizaron unos análisis de algunos cruces de variables que se consideraron pertinentes de acuerdo a los objetivos del estudio.

1.4 NIVELES DE MEDICION EN LAS CIENCIAS SOCIALES.

Un aspecto fundamental en los estudios de opinión que utilizan como estrategia metodológica los diseños cuantitativos es la decisión de cómo medir actitudes y

[2] En la practica es casi imposible una "medición perfecta" Generalmente se tiene un grado de error. Lo ideal es lograr minimizar este error tratando el cualificar el proceso de recolección mediante un instrumento y captura validado,

comportamientos de los sujetos. La idea es obtener un resultado de medición en la variable dependiente como consecuencia de cada variación en la variable independiente, por ejemplo, mejores rendimientos académicos según estrato socioeconómico.

Usualmente se considera esencial que la medición de la variable dependiente sea *cuantitativa*; además debe ser *objetiva* en el sentido de que los resultados de la medición sean siempre los mismos, independientemente de quién la haga. Esta insistencia en la medición objetiva y cuantitativa no necesariamente significa que sólo sea posible estudiar aspectos cuantificables y medibles del comportamiento humano. Una medición cuantitativa podría abarcar una amplia gama de situaciones cualitativas que por razón de los mismos objetivos del estudio debe ser generalizados con un alto grado de confiabilidad y validez[4].

No siempre es fácil usar puntajes estrictamente numéricos para medir el opiniones, actitudes o comportamientos. Imagine un estudio de corte cuasiexperimental para analizar si las personas que obtienen una buena calificación en un curso son más o menos dadas a utilizar su tiempo libre en actividades de tipo extraacademico en la Universidad. En este caso, **la variable independiente** es si los estudiantes obtienen buenas o malas calificaciones (evaluaciones) y la **variable dependiente** es el grado de participación en actividades extraacadémicas. Un método para medir la variable dependiente es asignar a los estudiantes en categorías, en este caso, grado de participación en actividades extraacadémicas: Si participan o no. Este tipo de medición se llama ***nominal*** porque sólo clasifica o asigna etiquetas o *nombre*, a distintas categorías de personas.

[3] Estudiantes registrados en el curso Diseño Cuantitativo I del segundo semestre de 1995.
[4] Confiabilidad es la ausencia de errores de medición en "mediciones sucesivas". Se entiende por validez el grado en que las mediciones empíricas representan la dimensión conceptual que se quiere medir (La Recolección de información. Mayor Alberto y Rodríguez Humberto. Modulo 3. Bogotá 1987.

Pero suponga que usted quiere que los estudiantes evalúen los cursos que han tomado según los consideren: *"muy buenos, buenos, regulares, malos o muy malos"*. Aquí se está introduciendo una relación entre los datos, calificando unos de mejores y otros de peores. Esto se conoce como **medición *ordinal*** porque implica un *ordenamiento por rangos* o puntajes. Es importante recordar que la medición ordinal no implica la existencia de intervalos iguales entre los juicios de bueno, mejor, el mejor, etc. Puede estar seguro de que "muy bueno" es "mejor" que "bueno" es "mejor" que "malo", es decir, se pueden ordenar por rango. Sin embargo, ¿sería posible afirmar que hay exactamente el mismo intervalo entre "muy bien" y "bien", que entre "malo" y "muy malo"?

Cuando se trata de puntajes numéricos, como el número de años cumplidos, o el tiempo dedicado a la lectura, se está tratando con números y se puede suponer que las distancias entre los puntajes son las mismas. Este tipo de medición se llama ***de intervalo***, porque supone ***intervalos iguales*** **entre los datos en una escala numérica continua**. Se supone que declarar 20 años y 25 años cumplidos de edad representa el mismo intervalo que el que existe entre 40 y 45 años cumplidos de edad. Como consecuencia, podemos realizar operaciones numéricas con este tipo de datos. Esto permite realizar tipos de análisis estadísticos más complejos que aprovechan las relaciones numéricas que se pueden establecer entre los puntajes.

Finalmente, hay otro tipo de medición que se menciona en la mayoría de los textos de técnicas de investigación, y es el llamado de *razón o proporción*. Supone una escala numérica que tiene un cero absoluto (es decir, casos en los que en verdad se da una situación de cero, como ausencia total de ingresos o de una característica).

Hay un aspecto que se debe tener muy en cuenta en la medición de variables de tipo cualitativas, y es el siguiente: es muy frecuente que a escalar *ordinales*, tales como una

escala de notas de cinco puntos entre Excelente, bueno, regular, deficiente y malo, se les asignen números del 1 al 5 y se las trate como si representaran *intervalos numéricos iguales*. Con frecuencia, esto se justifica plenamente, pero siempre se debe considerar lo que implica la conversión de una escala ordinal a una de intervalo, y si esta conversión es apropiada.

RESUMEN:

Los cuatro tipos de medición se pueden resumir así:

Nominal	Asigna o clasifica personas a categorías (sexo, nivel)
Ordinal	Permite ordenar los datos por rango (estatura, peso)
Intervalo	Permite colocar los datos en una escala numérica continua en la cual los intervalos entre los puntos son iguales (grupos etarios quinquenales).
Razón	Es igual que la de intervalo, pero tiene un cero absoluto

AUTOEVALUACION

¿Qué tipos de medición serían apropiados para medir:

(a) Aprobación o reprobación en un examen de capacidad de comprensión de una lengua moderna.

(b) Evaluación de los docentes con respecto a la utilización de recursos pedagógicos innovadores en el salón de clase.

(c) Puntajes de los estudiantes en una prueba de estadística, con rangos de 0 a 100%?

1.5. DISEÑO DEL INSTRUMENTO DE RECOLECCIÓN DE INFORMACION.

El instrumento primario de recolección de datos es **el cuestionario** que se utiliza para recolectar los datos que permitan el cumplimiento de los objetivos preestablecidos. Es importante hacer hincapié en esto, aunque resulte obvio que las respuestas estarán en función de las preguntas y que las preguntas deben plantearse en concordancia con los objetivos que se persiguen en el estudio.

Un *cuestionario* es un formulario con preguntas y en ocasiones con las posibles respuestas. Se puede entregar a la persona para que lo llene, usualmente auxiliándose con algunas instrucciones, y después recogerlo (Autoadministrado). Otra posibilidad es llenarlo en colaboración con el sujeto interrogado durante la entrevista. Da mejores resultados la segunda opción; sin embargo, resulta claro que requiere más encuestadores. Una mezcla de ambas situaciones puede resultar conveniente cuando existe la posibilidad de "validar" la información mediante medios directos o indirectos que permitan acceder a las fuentes de información: Usual mente reunir un grupo de personas en una sala y aclarar pregunta por pregunta, al mismo tiempo que los encuestados llenan sus cuestionarios. Otro método muy utilizado actualmente es la encuesta telefónica. Pero sabemos que son baratas pero poco confiables. Si lo que queremos es averiguar lo que piensa la gente que tiene teléfono, entonces debemos encuestar por teléfono. Pero si lo queremos es averiguar lo que piensa la

población en general, entonces no debemos encuestar por teléfono porque no todos los estratos tienen tiene teléfono

Los cuestionarios pueden ser de dos tipos: Abiertos o Cerrados, o bien una mezcla de ambos. Los cuestionarios abiertos plantean la pregunta y dejan abierta la respuesta; es decir, no proponen respuestas posibles; por ejemplo:

¿Cuál es tu estado civil? _____

Como se observa, se deja una línea para la respuesta pero no se plantean posibles respuestas. La ventaja de los cuestionarios con **preguntas abiertas** es que son útiles para explorar un proceso o un problema, sin limitar las respuestas. Este tipo de formato da la posibilidad al sujeto de describir la razón de sus ideas. La desventaja es que se dificulta el procesamiento de las respuestas del cuestionario, pues primero deben cerrarse tipificando las respuestas y después vaciarse éstas en hojas de codificación antes de su almacenamiento o grabación en un paquete especializado.

El proceso de ***tipificación*** consiste en encontrar cuáles son las respuestas más comunes, asignarles un código y después, con esos códigos, codificar las respuestas de todos los cuestionarios. Los *cuestionarios con preguntas cerradas* proponen las posibles respuestas a las preguntas planteadas. Aunque limitan las posibles respuestas, son de gran utilidad para estudios masivos en los cuales se pueden preestablecer las posibles categorías de las respuestas.

Las categorías de respuestas típicas para este tipo de preguntas pueden ser como las siguientes:: si/no, de acuerdo/desacuerdo, cierto/falso, o respuestas de tipo escalar, tales

como por ejemplo: excelente, bueno, regular, malo, muy malo. Veamos un ejemplo de pregunta cerrada:

¿Con quién vive? *Anote el código respuesta: _____*

	código
Con sus padres	1
Con familiares	2
Hermanos	3
Amigos	4
Esposa(o)	5
Solo	6
Otros	7

VENTAJAS DE LA PREGUNTA CERRADAS: Una de las ventajas de los cuestionarios cerrados es que muchas personas encuentran más fácil contestar este tipo de formularios, razón por la que se muestran mas complacientes a llenarlos. Otra ventaja de este tipo de preguntas es que pueden elaborarse en un formato que posibilite la grabación o captación de los datos sin necesidad de hacer un vaciado de las respuestas a hojas de codificación.

En el Apéndice A se muestra el cuestionario del Ejercicio de Trabajo, el cual corresponde a este tipo de cuestionarios con preguntas cerradas. Obsérvese en él que del lado izquierdo de cada página se encuentra la pregunta y las posibles respuestas, mientras que en el lado derecho de cada posible categoría de respuesta están los códigos numéricos de las respuestas correspondientes. Esta disposición facilita la captación de los datos, pues el personal de digitación sólo tendrá que digitar el número de las categorías de respuestas del

lado derecho de las páginas, teniendo como guía los números preimpresos que le indican en qué posición deben ir los datos[5].

En resumen, por un lado los cuestionarios con preguntas cerradas son útiles para capturar y homogeneizar respuestas; pero por otro lado, los cuestionarios con preguntas abiertas proporcionan más información que los anteriores, aunque su procesamiento requiere la tipificación de las respuestas y, por lo tanto, cuando no es posible plantear una pregunta completamente cerrada, es frecuente hacer un mezcla y proponer varias respuestas, aunque dejando la posibilidad de una respuesta abierta, lo que da origen a cuestionarios mixtos ; por ejemplo:

Ocupación de la persona que más aporta o contribuye al sostenimiento de tu casa.

1. Empleado
2. 2. Obrero
3. Patrón
4. Comerciante
5. Profesional
6. Otro. _____

En el caso en que el número de respuestas a la opción de respuesta abierta (en el ejemplo es la opción 6) sea significativo o que el sujeto conteste con respuestas importantes, valdrá la pena tipificar estas nuevas respuestas.

[5] En caso de no poder atender en forma individual a la(s) persona(s) que responderá(n) los cuestionarios es recomendable la elaboración de un cuadernillo de instrucciones en donde se clarifiquen las preguntas y se ejemplifique su solución.

Otro aspecto importante, que adquiere gran relevancia en el momento de **depurar los datos** captados, es establecer **identificadores** para numerar o reconocer los cuestionarios. Se sugiere dejar casillas en el formulario para el número de clasificación y para identificar el encuestador.

El numero asignado a cada formulario sirve para saber en qué cuestionario está el error y recurrir a éste con objeto de corregirlo y, en caso de que el error esté en el cuestionario que completó el encuestador X, recurrir a éste para que corrija la respuesta.

1.6. MANEJO DEL PROGRAMA STATGRAPHIC

El manejo del paquete estadístico STATGRAPHICS, versión 5.0, tiene como objetivo generar información tabulada, estadísticos y gráficos de un conjunto de datos previamente almacenados en un archivo. Este paquete se maneja con un sistema o estructura de menús y submenús que facilita y garantiza su operación rápida y amigable.

El STATGRAPHICS es un paquete estadístico que consiste en un sistema integrado de procedimientos estadísticos y gráficos que genera en la pantalla uno o dos cuadros con una serie de opciones que permiten el ordenamiento, clasificación y procesamiento de datos. El programa STATGRAPHICS está constituido por módulos que reciben el nombre de subprogramas; cada uno de ellos realiza un tipo de proceso de sistematización y gestión de archivos de datos.

Para almacenar datos de cuestionarios en el paquete Statgraphics se deben seguir algunos procedimientos previos como los que aquí se mencionan:

Para empezar, hay que definir la estructura de la matriz de datos. Con un cuestionario vacío en la mano se procede a determinar cuántas variables tiene. Una variable es cada propiedad o aspecto que se esté observando de un sujeto o de un proceso.

En el caso del Ejercicio de Práctica se considera como primera variable por analizar es el Programa académico; la siguiente variable es el Nº de créditos cursados del estudiante, la tercera el estrato socioeconómico, etc. En realidad, se tiene una variable por cada respuesta que se espere por cada pregunta.

En un cuestionario se plantean generalmente dos situaciones:

a) El caso más usual es que se espere una respuesta a una pregunta; por ejemplo, en la tercera pregunta del Ejercicio de Práctica que es la variable Estrato, hay una pregunta que espera una sola respuesta. Se pregunta el estrato en que está registrado la vivienda del estudiante encuestado. La respuesta puede ser Estrato 3 o 4 (**a una pregunta una respuesta**)

b) Otro caso es cuando se tienen múltiples respuestas para la misma pregunta ; por ejemplo, cuando se pregunta ¿Qué aparatos electrodomésticos tiene el estudiante en su casa?. Como se esperan hasta ocho respuestas, **se deben considerar como ocho preguntas independientes (variables) para los efectos de la codificación y almacenamiento de los datos**.

Una vez establecido el número de preguntas se procede al "Diseño de la Matriz de almacenamiento de los datos teniendo en cuenta sus características (Type) de medición (Numéricas, nominales u ordinales) y el ancho de columna (width) a utilizar.

1.7. GRABACIÓN DE DATOS.

Con la Hoja de Codificación en mano se puede realizar un primer proceso de depuración de los datos, observando por columna, la existencia de algún **código incompatible o inválidos** con los asignados a esa pregunta. Con esta revisión se detectan los errores más obvios y se dejan al computador sólo los que no lo son tanto. Para depurar los datos en el programa STATG, es recomendable correr esos datos, sin ninguna selección o modificación, con el subprograma **FRECUENCY TABULATION**. Esto permite la detección de *códigos inválidos*; es decir, de valores no esperados; por ejemplo, se pregunta el sexo del estudiante; la respuesta debe ser uno de los dos valores posibles 1 (masculino) o 2 (femenino), si no corresponde a ninguno, se tiene un código inválido que hay que buscar en el listado de los datos y corregirlo en el archivo en donde se hayan almacenado.

1.8. EJECUCIÓN DEL PROGRAMA STATGRAPHIC

Una vez grabados y depurados los datos en el programa, se procede a la ejecución de éste, para generar listados con información univariada o multivariada de tipo descriptivo o inferencial para analizar el comportamiento de esos datos.

1.9. ANÁLISIS DE LA INFORMACIÓN

En esta última etapa se analiza la información que aparece en los listados generados por el STATGRAPHICS durante la ejecución del programa. Cabe destacar que lo importante no es la ejecución del programa, sino entender los resultados. Para esto es necesario saber algo de estadística; por lo tanto, es recomendable tener a mano un manual de esta materia. ¿Pero qué es lo que debemos básicamente saber de estadística? En primera instancia saber qué tipo de

estadístico debemos utilizar y segundo saber leer los resultados o parámetros estadísticos que la máquina nos entrega.

Los resultados de cada subprograma o submenús son diferentes, por lo que la interpretación también lo es. En la exposición de cada subprograma incluido en este Módulo se hace un análisis de los resultados obtenido con el ejemplo.

1.10. AUTOEVALUACION.

1. Si se desea aplicar un cuestionario con preguntas para las que se conocen todas las posibles respuestas ¿qué tipo de cuestionario conviene elaborar? ¿Cuál en caso de no tener idea de qué respuestas se obtendrán?

2.Sabemos que la característica básica que debe reunir una pregunta es su carácter excluyente y exhaustivo. Justifique su decisión con respecto a las siguientes preguntas:

¿Cuánto tiempo dedica a estudiar y a trabajar (indique horas y minutos)?

¿Cuáles actividades extraacadémicas son sus favoritas?

¿Con qué frecuencia y duración realiza este tipo de actividades?

3. Si se tiene un cuestionario con 40 preguntas y cada una requiere una respuesta que ocupe dos posiciones, ¿Cuántas columnas serán necesarias para grabar cada cuestionario?

2. CAPITULO: MANEJO DEL PROGRAMA STATGRAPHICS

El paquete STATGRAPHICS no requiere instrucciones exhaustivas que indiquen paso a paso lo que el computador tiene que hacer. En su lugar se pueden elegir entre las 22 opciones de los subprogramas que el paquete ofrece.

En este Módulo se explican los procedimientos básicos para "arrancar" la ejecución del programa y como "navegar" en el Menú Principal. Además, se establece la forma en que deben almacenar los datos y la secuencia de las instrucciones.

2.1 Unidades de Análisis y Variables.

La unidad de análisis es el sujeto, hecho, objeto, etc. cuyas características o datos están en análisis: Los estudiantes matriculados, un ventero, una madre comunitaria, un propietario de la Comuna 9 o de la 10, etc.. Las variables son los aspectos o propiedades en estudio; por ejemplo, en una encuesta sobre el perfil socioeconómico de los estudiantes de Sociología de la Universidad de Antioquia, las variables de cada caso (persona que contesta el cuestionario) son , entre otras: sexo, edad, nº de créditos aprobados, lugar habitual de residencia, nivel educativo familiar, posición ocupacional de los aportantes de ingresos familiares etc..

3. CAPITULO: OTRAS INSTRUCCIONES ÚTILES

En este Módulo se estudiaran aquellas que son básicas para el análisis de una información cuantificada. Con las instrucciones analizadas en este Módulo se pueden lograr la descripción de los valores de unas variables (Frecuency Tabulation) o la determinación de una variable definida como independiente sobre otra(s) considerada dependientes. Finalmente se revisará en este Módulo, algunas técnicas estadísticas que permiten observar el comportamiento de los datos asociados o correlacionados.

3.1. REVISIÓN Y CONTROL DE IMPRESIÓN DEL PROGRAMA

En esta sección se exponen instrucciones que permiten revisar un

3.1.1 Revisión de los datos almacenados

A pesar de la facilidad en la elaboración de una matriz de datos en STATGRAPHICS, no deja de ser factible que cometan errores de sintaxis, ya sea por error de grabación, o de codificación. Se puede hacer una corrida de verificación de los datos utilizando la opción Frecuency Tabulation del Submenu Descriptive Methods. Esto permite asegurarse de la corrección de la grabación.

3.1.2 Formato de diseño de la matriz de datos

En esta sección se estudian las instrucciones que permiten, la definición sobre el tipo y ancho la variable.

3.1.7. Recodificación de valores

La recodificación de valores significa cambiar su valor actual por uno nuevo designado por el usuario. Para efectuar recodificaciones se utiliza la instrucción RECODE. Por ejemplo, la variable Nº de créditos aprobados del Ejercicio de Práctica, puede utilizar dos páginas para tan sólo la impresión de la tabla de frecuencias; por otra parte, es difícil su interpretación dada la extensión de la Tabla. En este caso se decide recodificar los valores en intervalos de 20 créditos. La recodificación se solicita utilizando la misma variable para guardar los datos recodificados.

4. DESCRIPTIVE METHODS: SUBPROGRAMAS DE ESTADÍSTICA DESCRIPTIVA

En esta sección se presentan dos opciones de estadística descriptiva más usuales: aunque **SUMMARY STATISTICS Y FRECUENCY TABULATION**. Se da una descripción de cada subprograma, el procedimiento para su ejecución y un ejemplo con las variables del Ejercicio de Práctica; además se interpretan sus resultados y se describen sus estadísticas, opciones y limitaciones.

Téngase en cuenta que aunque **SUMMARY STATISTICS Y FRECUENCY TABULATION** calcula estadísticas para todas las variables, éstas serán totalmente válidas sólo para las variables cuantitativas, y que muchas no tendrán sentido para las variables cualitativas.

4.1. SUMMARY STATISTICS (Resumen de Estadísticos)

4.1.1 PRESENTACIÓN

El subprograma **SUMMARY STATISTICS** tiene una función muy semejante al de **Frecuency Tabulations**: calcula las estadísticas de tendencia central y de dispersión de las variables que se seleccionen; sin embargo, no elabora tablas de frecuencia ni histogramas. Adicionalmente, se le puede solicitar la impresión de la suma de los valores de cada variable[6]. Dado que su única función es calcular estadísticas, su uso se restringe a variables cuantitativas, ya que para variables cualitativas no tienen sentido medidas como la desviación estándar.

RESULTADOS E INTERPRETACIÓN

Analicemos el Ejercicio de Práctica:

Se puede plantear una pregunta con base en estos resultados

4.2. FRECUENCY TABULATION (Distribución de Frecuencias)

4.2.1. PRESENTACIÓN

El subprograma FRECUENCY TABULATION es el más utilizado del paquete STATGRAPHICS , dado que no sólo sirve para el cálculo de estadísticas, sino también para **depurar datos.**

[6] Se pueden seleccionar hasta 12 variables para la impresión. Se recomienda utilizar el comando F5 (Propmt).

Este subprograma elabora **una tabla de frecuencias** en donde indica cada código que aparece como respuesta, el número de veces que aparece y el porcentaje que representan esas apariciones con respecto al total de casos. Igualmente calcula las estadísticas de tendencia de central y de dispersión de las variables que se proporcionen en el listado de variables.

A este tipo de opciones u subprogramas se le llama de estadística univariada, dado que sólo analiza una variable a la vez. También se le llama de **estadística descriptiva** ya que las estadísticas que calcula describen la distribución de los datos.

ESTADÍSTICAS, OPCIONES Y LIMITACIONES

ESTADÍSTICAS

1. Media (promedio)
2. Error estándar
3. Mediana
4. Moda
5. Desviación estándar
6. Varianza
7. KURTOSIS, medida de afilamiento de la distribución
8. Sesgo
9. Intervalo

10 Mínimum

11. Máximum

5. CATEGORICAL DATA ANALYSIS : SUBPROGRAMAS PARA ANÁLISIS DE VARIABLES CUALITATIVAS.

En esta sección se presentan dos opciones para el análisis de datos no numéricos: CROSSTABULATIONS (Cruces de variables) y CHI - CUADRADO. Se da una descripción de estos subprogramas, el procedimiento para su ejecución y un ejemplo con las variables del Ejercicio de Práctica; además se interpretan sus resultados y se describen sus estadísticas, opciones y limitaciones.

Téngase en cuenta que aunque esos estadísticos calculan estadísticas para todas las variables éstas serán totalmente válidas sólo para las variables cualitativas.

5.1. CROSSTABULATION (Cruce de Variables)

5.1.2. PRESENTACIÓN

El subprograma CROSSTABULATION elabora *tablas de contingencia* y calcula sus estadísticas.

Las tablas de contingencia son la representación conjunta de la distribución de frecuencias de dos o más variables. Si se tienen dos variables entonces cada posible combinación entre sus valores crea una *celdilla*. Cada par de procesados, uno por variable, incrementa el conteo en la celdilla que le corresponde. Así, al final del proceso de elaboración de tablas se tiene una que indica las posibles combinaciones entre las dos variables y el número de veces

que se encontró de cada combinación. Las estadísticas se calculan tomando como base la tabla así formada.

En este subprograma se puede utilizar cualquier tipo de variables: *Cuantitativas Discretas, Cuantitativas Continuas, Cualitativas Nominales, Cualitativas Ordinales.* Por supuesto, las estadísticas que se calculan tienen validez sólo para las variables adecuadas; por ejemplo, el coeficiente de correlación entre sexo y edad no tiene ningún sentido, ya que sexo es cualitativa nominal y edad es cuantitativa continua; pero el coeficiente de correlación entre estatura y peso si tiene validez, porque ambas son cuantitativas, y con ese coeficiente se puede determinar si existe o no relación entre esas variables, con un nivel de significancia dado.

Todas las estadísticas que calcula CROSSTABULATION son *medidas de asociación* o, en su defecto, de independencia; pero la negación de una lleva la afirmación de la otra; si dos variables no están asociadas, son independientes, o llevará si dos variables no son independientes están asociadas..

Las tablas de contingencia también se conocen como ***TABLAS CRUZADAS***. Cuando se forman con dos variables reciben el nombre de *tablas de dos entradas*; cuando son más de dos las variables involucradas, son tablas de **n** entradas, donde **n** es el número de variables involucradas. A cada tabla obtenida del cruce de dos o más variables se le conoce como *subtabla*, dado que en realidad es parte de la tabla principal formada por las dos primeras variables; por ejemplo, supóngase una ejecución con el subprograma CROSSTABULATION con un cruce entre **las variables Edad y Nº de Créditos, teniendo como variable de control a la variable sexo**. El resultado son dos tablas cruzadas entre edad y escolaridad, una para sexo masculino y otra para femenino. En este ejemplo, la variable de control divide la tabla principal en dos subtablas, una por sexo.

5.1.2. EJECUCIÓN DEL SUBPROGRAMA

El subprograma CROSSTABULATION tiene dos formatos de ejecución:

el modo entero y el modo general. Estos modos son los mismos que los de Frecuency Tabulation. El modo entero sólo procesa variables numéricas con valores enteros; mientras que el modo general procesa tanto variables numéricas como alfanuméricas; además, las variables numéricas pueden tener fracción decimal.

5.1.3. RESULTADOS E INTERPRETACIÓN

Se analizan dos casos: uno en que sólo se estudia la tabla de contingencia y otro en el que se hace una prueba de hipótesis acerca de la independencia entre las variables de la tabla de contingencia.

a) La figura 5.4.2 presenta la tabla de contingencia resultante del cruce de la variable sexo (V7) con la variable turno (V2). En esta se observa que hay 276 mujeres en el turno matutino y 289 en el vespertino, mientras que hay 485 hombres en el turno matutino y 371 en el vespertino. En los totales de renglón (ROW TOTAL) se tiene que hay 565 mujeres y 856 hombres, mientras que los totales de la columna (COLUMN TOTAL) indican que hay 761 estudiantes en el turno matutino y 663 en el vespertino. El total de respuestas es de 1424.

Las cifras que se encuentran en cada celdilla corresponden a: conteo (número de casos que cumplen la combinación de valores de las variables), porcentaje que representa ese conteo

con respecto al total del renglón, porcentaje con respecto al total de la columna y porcentaje con respecto al gran total; por ejemplo, en el cruce de sexo femenino con turno matutino se tienen 276 casos. Los porcentajes de esas celdillas se obtienen como sigue:

$$\% \text{ respecto al renglón} = \frac{\text{Conteo}}{\text{tot. rengl.}} * 100 = \frac{276}{565} * 100 = 48.8\ \%$$

$$\% \text{ respecto a la columna} = \frac{\text{Conteo}}{\text{tot. col}} * 100 = \frac{276}{761} * 100 = 36.3\ \%$$

$$\% \text{ respecto al gran total} = \frac{\text{Conteo}}{\text{gran. tot.}} * 100 = \frac{276}{1424} * 100 = 19.4\%$$

b) El cruce de las variables sexo (V7) y pesos recodificado (V8R) produce las tablas y estadísticas que se muestran en la figura 5.4.3. Como de costumbre, los valores de la variable recodificada son convencionales; también se pudieron asignar como valores recodificados los valores de *marca de clase* (promedio entre los valores máximo y mínimo de la clase).

En la búsqueda de los mayores conteos en las celdillas se encuentra que para el sexo femenino el máximo conteo corresponde a la clase que abarca entre 46 y 50 Kg.; mientras que para el sexo masculino ese conteo máximo corresponde a la clase de 56 a 60 Kg. Esto

no resulta sorprendente, pues en los resultados de AGGREGATE se observó que el promedio de los hombres es mayor que el de las mujeres.

Con base en esta tabla y estadísticas se puede plantear la pregunta siguiente: ¿las variables sexo y peso son independientes entre sí ? Para contestar esta pregunta se hace uso de la prueba de hipótesis basada en la ji-cuadrada; la aceptación de Ho (hipótesis nula) indica la independencia entre las variables (si se rechaza Ho, se acepta la hipótesis alternativa H1, o sea, la no independencia entre las variables).

La prueba de hipótesis se plantea como:

Ho: independencia entre las variables
H1: no independencia entre las variables

Para determinar el valor crítico que separa las zonas de aceptación y de rechazo en una gráfica de la distribución ji-cuadrada, se calculan los grados de libertad y se fija la confianza deseada. Los grados de libertad de una tabla de contingencia se calculan como:

$gl = (r - 1) (c - 1)$ donde: gl = grados de libertad
r = número de renglones
c = número de columnas de la tabla.

Se tiene al sustituir:

gl = (13 - l) (3-1) = (12) (2) = 24 grados de libertad

Este valor también se encuentra en el listado de resultados (24 DEGREES O FREEDOM); pero se calculó para mostrar cómo se obtiene.

Con una confianza del 95% (a = 0.05) y 24 grados de libertad se busca el valor crítico correspondiente en una tabla de ji-cuadrada. El valor en tablas es de 36.4151. El valor calculado (CHI- SQUARE) es de 268.07898, por lo que este cae en la región crítica (figura siguiente); entonces se acepta H1 (se rechaza Ho) y se concluye que no hay independencia entre sexo y peso. Esto confirma la observación elaborada a partir de los resultados de AGGREGATE . Se observa que hay 77 valores faltantes (NUMBER OF MISSING OBSERVATIONS = 77). Las otras estadísticas corresponden en su mayor parte a medidas normalizadas basadas en la ji-cuadrada y resultan inútiles cuando se efectúa alguna comparación entre varias tablas de contingencia. Obsérvese que no se calculó el coeficiente de correlación (PERSON'S R) . Esto se debe a que la variable sexo es alfanumérica y esta medida necesita que los valores sean cuantitativos.

Si las variables son numéricas y se solicita el cálculo del coeficiente de correlación (estadística 11), entonces se calcula este; pero sólo tiene significado si las variables además de ser numéricas son cuantitativas.

----------GRÁFICA---------- 5.4.4

6. SUBPROGRAMAS DE CORRELACIÓN Y REGRESIÓN

En esta sección se presentan los subprogramas de correlación y regresión de la misma manera en que se estudian los subprogramas de estadística descriptiva en la sección anterior. Se describe su función y el formato de la ejecución; se hace un estudio de caso con las variables del Ejercicio de Práctica , se analizan los resultados y se describen las estadísticas, opciones y limitaciones.

6.1. CORRELATION ANALYSIS

6.1.1. Presentación

Este subprograma calcula la correlación entre variables, ya sea entre los casos que se indiquen o en forma de una matriz de correlación, según se solicite la ejecución a CORRELATION ANALYSIS. También indica el número de valores con que calcula el coeficiente, así como la significancia de éste.

La *correlación* es una medida de asociación entre variables. muestra cuanta relación existe entre los valores de una con respecto a los de la otra (correlación simple) o de las otras (correlación múltiple). El coeficiente de correlación simple, r, de las variables X y Y se define como el coeficiente de la covarianza de X y Y, entre el producto de las desviaciones estándar de X y de Y, o sea,

$r = Cov(x,y)/ S(x)S(y)$

donde r = coeficiente de correlación

Cov(x,y)= covarianza de x y y

S(x) = desviación estándar de x

S(y) = desviación estándar de y

El coeficiente de correlación es una medida estandarizada; es decir, sólo puede tomar ciertos valores dentro de un intervalo dado. En este caso el intervalo es de -1 a 1. Los valores negativos del coeficiente de correlación indican una relación inversa entre las variables X y Y; es decir, mientras que X crece, Y decrece. Los valores positivos señalan una relación directa: Cuando X crece, Y también crece.

Tanto el valor 1 como -1 indican una relación perfecta (directa o inversa) entre las variables. Esta situación es difícil de encontrar en los fenómenos reales con más frecuencia, el coeficiente r toma valores que pueden acercarse a esos extremos, aunque casi nunca iguala. un valor de r igual o cercano a cero, indica que no hay relación entre las variables. Así se tiene que valores iguales o cercanos a 1 o a -1 señalan una alta relación entre las variables, mientras que valores iguales o cercanos a cero indican que no hay relación entre estas. Entonces, en una recta numérica para el coeficiente de correlación se tiene:

1 0 +1 valores de r

l ---------- l ---------- l

Correlación negativa Correlación positiva

(relación inversa) (relación directa)

La pregunta siguiente es: **¿Cuándo un valor de r indica que hay relación y cuándo no?** Esto depende de las condiciones que fije el investigador; en ocasiones un valor de r = 0.5404 señala relación, mientras que para otro tipo de fenómeno un valor de r = 0.9666 indica que

no hay suficiente relación entre las variables. La comprobación de esto se hace mediante **una prueba de hipótesis** en que se determina si el coeficiente de correlación es significativamente igual o diferente de cero (hipótesis nula e hipótesis alternativa).

6.1.2. EJECUCIÓN DEL SUBPROGRAMA

6.1.3. RESULTADOS E INTERPRETACIÓN

a) Tabla de correlación

Fig. 6.1.2 . Resultados de la primera ejecución al subprograma PEARSON CORR.

La pregunta que se impone es ¿el coeficiente de correlación 0.3078 indica relación significativa entre las variables V30 y V35? Se puede responder haciendo la prueba de hipótesis relativa al coeficiente de relación, con base en los valores f (distribución f de Fisher). La prueba de hipótesis es:

Ho: p = 0 p = coeficiente de correlación poblacional
H1: p dif 0

Las etapas para la prueba de hipótesis acerca del coeficiente de correlación, son semejantes a las otras pruebas: se busca un valor en tablas que divida la zona de aceptación y de rechazo,

se compara el valor calculado con el de la cota y se acepta la hipótesis nula (Ho) y la hipótesis alternativa (H1).

El valor calculado se obtiene con la fórmula :

$$F = \frac{r^2 / (k - 1)}{(1-r^2)/(N-k)}$$

donde: r = coeficiente de correlación
k = número total de variables
N = número total de valores
k-1 = grados de libertad del numerador
N-k = grados de libertad del denominador

Si r2 = 0.3058, k = 2 variables y N = 577 valores, entonces:

$$F = \frac{0.3078/(2-1)}{(1-3078)/577-2)} = \frac{0.0947408}{0.0015743} = 60.179635$$

Para una confiabilidad del 95% (a = 0.005), 1 grado de libertad en numerador y 575 grados de libertad en el denominador se tiene que el valor en tablas es 3855. El valor F calculado cae en la región de rechazo, por lo que se acepta H1 (véase fig. 6.1.3). El coeficiente de correlación obtenido es significativamente diferente de cero; por lo tanto, hay relación significativa entre las calificaciones de secundaria y las de primer semestre de bachillerato; relación que intuitivamente se esperaba.

----------GRÁFICA----------6.1.3

b) los resultados de la segunda ejecución a PEARSON CORR se encuentran en la figura 6.1.4

El número de casos, el valor promedio y la desviación estándar de las variables edad (V4), estatura (V6) y peso (V8) se indican el la figura 6.1.4a. Obsérvese que el número de casos para estas variables es mayor que para las variables procesadas en la primera ejecución. Esto se debe a que la selección temporal deja de tener efecto en cuanto se ejecuta la siguiente tarea, en este caso, la segunda ejecución a PEARSON CORR.

En la figura 6.1.4b se muestra la matriz de correlación formada con las variables V4, V6 Y V8. La diagonal principal de esta matriz contiene los coeficientes de correlación de las variables consigo mismas, por lo que estos coeficientes valen 1; pero el número de casos para el cálculo de estos coeficientes es cero , advertencia de que no son útiles. Si se resta de 1 el valor de la significancia de estos coeficientes y se multiplica por 100 la diferencia, se obtiene la confiabilidad del coeficiente; por ejemplo: la significancia del coeficiente de correlación entre V4 (edad) y V6 (estatura) es de 0.108, por lo que su confianza es :

Confianza = (1 - 108) * 100 = 89.2%

6.2. PARTIAL CORR

5.2.1. Presentación

Este subprograma calcula los coeficientes de correlación parcial con una más variables de control.

La *correlación parcial* es, en cierta medida, semejante a las tablas de contingencia (subprograma CROSSTABULATION) cuando, además de la variable dependiente e independiente, se incluyen variables de control que desglosan las tablas de contingencia en subtablas de acuerdo con los valores de las variables de control. En el caso de las tablas de contingencia la subdivisión es física; es decir, se separan físicamente los valores y se crean subtablas que contienen menos valores (**casos).** En la correlación parcial el efecto es matemático: se elimina la influencia de la variable de control en el cálculo de la correlación entre las otras variables al suponer que hay una relación lineal entre las variables de control y las variables dependientes e independientes.

6.2.2. EJECUCIÓN DEL SUBPROGRAMA

Ejemplificación

Supóngase que se busca el cálculo de la correlación parcial entre las variables peso y estatura, con la edad como variable de control; también se quiere la impresión de los valores de la media (promedio), desviación estándar y número de casos válidos para cada variable (estadística 2). Entonces la ejecución es :

ceo@catios.udea.edu.co
http://ceo.udea.edu.co
Ciudad Universitaria Bloque 9-252 Telefax: 2105775

PARTIAL CORR	V8 WITH V6 BY V4
STATISTICS	2

5.2.3. Estadísticas

a) Estadísticas:

1. Coeficiente de correlación de PEARSON, r
2. Coeficiente de determinación, r2
3. Significancia de r
4. Error estándar de la estimación
5. Valor de la ordenada al origen
6. Pendiente

7. REGRESSION ANALYSIS (Análisis de Regresión)

7.1. PRESENTACIÓN

Este subprograma efectúa *análisis de regresión*. La *regresión* es una técnica basada en los mínimos cuadrados que permite analizar la relación entre dos variables (*regresión simple*), o entre una variable dependiente y varias independientes (*regresión múltiple*).

La regresión sirve para analizar relaciones funcionales; es decir, relaciones en las que los valores de una variable están en función de los de otra u otras. Esta relación se expresa como y = f(x) cuando es una relación simple (se lee: *y* es una función de *x)*, y como *y* = f(*x1, x2, x3 .....xn*) cuando es múltiple.

Como uno de los resultados del análisis de regresión se obtiene una *resta de ajuste*, o *resta de regresión*. La recta que minimiza la suma de los cuadrados de los errores es la recta de regresión. Los errores, o residuales, son las distancias entre los valores definidos por la recta (predichos) y los valores observados (reales). Uno de los métodos para obtener los parámetros que definen tal recta es el de los mínimos cuadrados. Esta recta sirve para estimar valores. En el caso de *la regresión simple*, dado un valor de la variable independiente se calcula el que corresponda a la variable dependiente. O al revés, dado un valor de la variable dependiente se estima el valor correspondiente a la variable independiente. *El caso de la regresión múltiple es diferente*; con valores de las variables independientes se puede estimar el valor de la variable dependiente; pero un valor de la variable dependiente no sirve para estimar los valores de las variables independientes. En todo caso sólo se puede estimar uno, el resto tendría que declararse. También se obtiene el coeficiente de correlación (simple o múltiple) que indica la magnitud de la relación que guardan las variables incluidas en el análisis.

7.2. EJECUCIÓN DEL SUBPROGRAMA

Ejemplificación

El Ejercicio de Práctica no proporcionan un caso de regresión múltiple claro, aunque sí uno para regresión simple. Basta con señalar que para la regresión múltiple se deben anotar más

de una variable independiente, mientras que para la regresión simple debe haber una sola variable independiente; así pues, la generalización de regresión simple a múltiple es muy sencilla.

Como ejemplo se solicita la regresión entre el peso (V8) y la estatura (V6). Se indica un nivel de inclusión impar (1), por lo que REGRESSION genera la tabla con la inclusión de V6 y también la tabla resumen o final.

ANEXO A:

GLOSARIO DE TÉRMINOS

Contiene la definición de los términos usados en el Módulo. Como nombre de la entrada se encuentra el concepto operativo o comando y entre paréntesis su traducción.

Conviene consultar esta sección cuando se encuentre en el Módulo un término de desconocido, dado que buena parte de las dificultades para familiarizarse con paquetes informáticos se debe generalmente al lenguaje especializado (anglicismos) que se emplea en los manuales y revistas especializadas.

Archivo (file). Conjunto de registros o datos relacionados entre sí, los cuales se pueden almacenar en memoria RAM o ROM. Para distinguir los archivos y facilita su clasificación, se les debe asignar un nombre (filename) hasta de 11 caracteres, de los cuales, los últimos

tres se conocen como extensión y se separan de los básicos mediante un punto. Ejemplos de extensiones: DOC; BAK; BAT; EXE[7].

Archivo "asqui"(ASCII Files). Es un archivo de texto (con extensión txt) que contiene solamente letras, números, signos de puntuación y códigos de control pertenecientes a la tabla ASCII.[8]

Backup (backup). Comando externo del DOS que permite hacer copias de seguridad.

Bit (bit). Dígito binario. Cada uno de los dos posibles valores que puede utilizar una computador digital. Los valores son el cero y el uno.

Byte (byte). Grupo de 8 bits. Es la unidad mínima que puede manejar una computador; es decir, opera con paquetes de 8 bits. Un byte es la cantidad de bits necesaria para almacenar una letra, un dígito o un carácter especial. Sus múltiplos son KB (kilobyte = 1024 bytes), MB(megabytes = 1024 KB) y GB (gigabyte = 1024 MB).

Carácter (character). Dato que ocupa una posición (1 byte). Los caracteres pueden ser letras, números o caracteres especiales como: @ # % $ * ? = \ , etc.

INDICATIVO (prompt). Carácter o caracteres que el sistema operativo o el paquete de aplicación exhibe en pantalla solicitando la intervención del usuario; por ejemplo: el sistema operativo DOS presenta la letra de la unidad que está en función (**C:\ >**) y un signo de mayor (**C:>),** el SPSS la palabra SPSS/PC y dos puntos (SPSS/PC:). El paquete STATGRAPHICS presenta la letra de la unidad C **:\STATG>.**

[7] Mejía, M. Aurelio, Diccionario Técnico Actualizado. Edit. Divulgación Técnica Electrónica. Medellín, 1991. p. 183
[8] Ibid. p. 37

Caso (case). En el STATGRAPHICS es la unidad de análisis, es el sujeto, es el cuestionario que se aplicó a partir de una muestra o de una población; ejemplo: la aplicación de un cuestionario a un grupo de 205 estudiantes implica la existencia de 50 casos. Aquí, un caso es cada estudiante.

Código (code). Convención para representar letras, números y caracteres especiales con el fin de identificar características o categorías de respuestas.

Código ASCII(ASCII code). Convención del ASCII (Instituto Estadounidense de Normas) que representa letras, dígitos y caracteres especiales que facilitan la comunicación datos.

Configuración (configuration). Dispositivos, y sus capacidades, que constituyen el hardware de un computador.

Copia de respaldo (backup). Copia de seguridad que se hace de la información de un disco o disquete a fin de poder recuperarla en caso que se dañe el medio original.

CPU (CPU) Generalmente se llama así a la caja o a la torre que contiene todos los componentes como los discos, la motherboards, las tarjetas de entrada y salida, de vídeo y comunicación, etc..

Create (crear). Comando del subprograma DATA MAGEMENTS del STAGRAPHICS que se utiliza para crear un archivo nuevo.

Data Management (Administración de Datos). Concepto que abarca todos los subprogramas del STATGRAPHICS que permiten diseñar, almacenar y editar archivos de datos.

Datos (data). Valores alfanuméricos que procesa el computador para convertirlos en información.

Dígito (dígito): Cada uno de los números que integran un dato numérico; por ejemplo: 1995 tiene cuatro dígitos, o sea, cuatro números.

Edit (editar) Comando del subprograma DATA MAGEMENTS del STAGRAPHICS que se utiliza para .editar un archivo ya creado.

Grabación De Datos. (data entry). Almacenamiento o **digitación** de los datos de la hoja d codificación o del cuestionario al programa; por ejemplo, digitación en la matriz de datos del STATGRAPHICS

Hoja de Cálculo o electrónica (electronic spreadsheet). Denominación que reciben todos los paquetes de software que presentan en pantalla una tabla que el usuario llena con datos. También, mediante ecuaciones, el usuario describe operaciones a nivel de columna, renglón y dato que crean nuevos datos, o bien actualizan los ya existentes.

Label (label). Membrete, nombre o etiqueta de una categoría de una variable.

Listado (printout) Papel impreso que sale de una impresora.

Memoria (memory). RAM: memoria de trabajo, Dispositivo de almacenamiento temporal. ROM memoria de configuración.

Microprocesador (microprocesor). Circuito integrado (chip) que contiene toda la información necesaria para que un computador trabaje. Se les llama también CPU. Ejemplos de ellos son : un procesador 486 DX4 de 100 Mghz o un Pentium de 75 Mghz.

Monitor (monitor). Pantalla en la que el sistema de computación exhibe las peticiones de datos y comandos y los resultados del proceso.

MS-DOS (MS-DOS). Sistema operativo para computadores desarrollado por Microsoft (MS). Las letras DOS significan Disk Operating System (sistema operativo del disco)

Name (nombre). Nombre de la variable. En el paquete STATGRAPHICS se utiliza este promt para indicar el nombre de la variable.

Procesador de Palabras (word processor). Programa que sirve para crear y editar archivos de dos tipos: documentos y no documentos. Los documentos son textos que requieren edición avanzada, como márgenes, encabezados, pies de página, tabulación, etc. Ejemplos de documentos son las cartas, los informes y la documentación en general. Los más comunes son el *Winword o Word 6* o el Wordperfect.

Programa (program). Conjunto de instrucciones que indican a la computador qué hacer y cómo hacerlo. También es el conjunto de instrucciones que llevan a cabo un atarea específica.

Sistema Operativo (operating system). Software básico de un sistema de computación. El sistema operativo también se considera como la interfaz entre el hardware y el software del usuario.

System Profile (Perfil del Sistema operativo del Paquete). Conjunto de subprogramas del paquete STATGRAPHICS que permiten manejar y controlar el paquete.

Type (tipo). Tipo de variable. En el paquete STATGRAPHICS, hay que definir si la variable es numérica, cualitativa, etc...

Unidad Central de Procesamiento (central processor unit). La **CPU** es la parte "pensante" de un computador, ya que se encarga del procesamiento y del control de los demás dispositivos.

Variable (data). En relación con STATGRAPHICS es la característica que varía que se mide u observa de un elemento, una muestra o la población; ejemplos: edad, escolaridad, residencia, opinión, etc. Es un comando de entrada que pide que se señale el nombre de la(s) variable(s) que se va(n) a trabajar.

Width (ancho). Ancho en nº de columnas que ocupa la variable en la matriz de datos del STATGRAPHICS.