LA ESTADÍSTICA: GNOSIS DEL SER HUMANO

Alberto Rodríguez Rodríguez
Robards Javier Lima Pisco
Miguel Angel Padilla Orlando
Tania Yadira Garcia Ponce
Rosa Yesenia Vera Loor
Julio C. Pino Tarragó

Ciencias y Letras

3ciencias

LA ESTADÍSTICA:

GNOSIS DEL SER HUMANO

Alberto Rodríguez Rodríguez

Robards Javier Lima Pisco

Miguel Angel Padilla Orlando

Tania Yadira Garcia Ponce

Rosa Yesenia Vera Loor

Julio C. Pino Tarragó

Editorial Área de Innovación y Desarrollo,S.L.

Quedan todos los derechos reservados. Esta publicación no puede ser reproducida, distribuida, comunicada públicamente o utilizada, total o parcialmente, sin previa autorización.

© del texto: **los autores**

ÁREA DE INNOVACIÓN Y DESARROLLO, S.L.

C/ Els Alzamora, 17 - 03802 - ALCOY (ALICANTE) info@3ciencias.com

Primera edición: **enero 2018**

ISBN: **978-84-948074-3-5**

DOI: **http://dx.doi.org/10.17993/CcyLl.2018.14**

AUTORES

Alberto Rodríguez Rodríguez. Licenciado en Matemática, Universidad de Granma, Máster en Ciencias de la Educación y Doctor en Ciencias, Universidad de Granma. Investiga en temas relacionados con los modelos matemáticos aplicados a la Contabilidad. Actualmente profesor de la Universidad Estatal del Sur de Manabí, Ecuador.

Robards Javier Lima Pisco. Economista por la Universidad Estatal del Sur de Manabí, Magister en Administración Ambiental por la Universidad Estatal de Guayaquil. Experto en gestión y liderazgo transformador para el desarrollo del buen vivir. Universidad de Cuenca. Investiga temas relacionados con la dirección estratégica de procesos estadísticos. Actualmente profesor de la Universidad Estatal del Sur de Manabí, Ecuador.

Miguel Angel Padilla Orlando. Economista, Universidad Estatal del Sur de Manabí, Magister en Contabilidad y Auditoría, Universidad Estatal del Sur de Manabí. Investiga en temas relacionados con Estadística y Planificación Estratégica. Actualmente profesor de la Universidad Estatal del Sur de Manabí, Ecuador.

Tania Yadira Garcia Ponce. Economista, Universidad Estatal del Sur de Manabí, Magister en Contabilidad y Auditoría, Universidad Estatal del Sur de Manabí. Investiga en temas relacionados con Auditoria y Estadística. Actualmente profesor de la Universidad Estatal del Sur de Manabí, Ecuador.

Rosa Yesenia Vera Loor. Licenciada en Contabilidad y Auditoría, Universidad Técnica de Manabí, Magister en Contabilidad y Auditoría, Universidad Estatal del Sur de Manabí. Investiga en temas relacionados con la Auditoria y estadística.

Julio Cesar Pino Tarragó, Ingeniero Mecánico, Universidad de Holguín, Holguín, Cuba, Máster en Máquinas Agrícolas y Doctor en Ciencias Técnicas, Universidad Politécnica de Madrid. Investiga en temas: Evaluaciones de la explotación y mantenimiento de máquinas agrícolas; Procesamiento industrial de alimentos, Evaluación de Máquinas y Equipos Agrícolas. Actualmente profesor de la Universidad Estatal del Sur de Manabí, Ecuador.

Índice

CAPÍTULO I. VARIABLES ALEATORIAS Y FUNCIONES DE DISTRIBUCIÓN. 11

1.1. Variables aleatorias. Funciones de distribución. Variables aleatorias discretas y continuas. Niveles de medición. 11

1.2. Algunas distribuciones importantes. 19

1.3. Características numéricas de variables aleatorias. 22

1.4. Distribución conjunta de varias variables aleatorias. 26

1.5. Generación de distribuciones a partir de operaciones con variables aleatorias conocidas 28

1.6. Conclusiones 31

CAPÍTULO II. INTRODUCCIÓN AL SPSS/PC. 33

Introducción 33

2.1. Posibilidades del SPSS/PC 33

2.2. Formas de trabajo del SPSS/PC para DOS. 34

2.3. Generalidades. Comandos. Tipos de comandos. 35

2.4. Comandos de Operación 39

2.5. Comandos de definición de datos. 41

2.6. Comandos de procedimientos. Estructura típica de un programa en SPSS/PC+ para DOS. 49

2.7 Conclusiones. 52

2.8 Ejercicios del Capítulo II 52

CAPÍTULO III. INTRODUCCIÓN A LA INFERENCIA ESTADÍSTICA. 55

Introducción 55

3.3. Análisis descriptivo de datos con el SPSS/PC. Comandos FREQUENCIES, DESCRIPTIVES y MEANS 60

3.4. Pruebas de hipótesis. Verificación estadística de hipótesis estadísticas. 64

3.6 Conclusiones. 72

3.7 Ejercicios del Capítulo III 72

CAPÍTULO IV. INDEPENDENCIA DE VARIABLES Y MEDIDAS DE ASOCIACIÓN 77

Introducción 77

4.2. Otras medidas de asociación entre variables aleatorias discretas nominales y ordinales. 80

4.3. Medidas de correlación entre variables aleatorias continuas o entre continuas y discretas. 85

4.4. El procedimiento CROSSTABS del SPSS. 89

BIBLIOGRAFÍA 95

PRÓLOGO

La investigación estadística conduce siempre al estudio de varias variables simultáneamente. En este libro el lector encontrará recursos didácticos para comprender a cabalidad las relaciones que se producen entre esas variables, sustentadas en un enfoque contextualizado a procesos vivenciales, utilizando un lenguaje directo, preciso, claro, sin ambigüedades y sin perder el rigor científico para alcanzar una sólida base teórico-conceptual que permita enfrentarse a la solución de diversos problemas presentados en la vida real y en las ciencias.

La mayor parte de los problemas no se pueden resolver conociendo solo la distribución de cada variable por separado, sino la distribución de ellas en conjunto. Por esta razón debemos extender nuestros conceptos y fórmulas a variables aleatorias y multidimensionales, esto es, a pares o n-plus de variables aleatorias consideradas conjuntamente. En consecuencia, en cada capítulo se proponen suficientes ejemplos resueltos para retroalimentar y afianzar la teoría en vinculación con una práctica dinámica.

Con este libro los autores, profesores de la Universidad Estatal del Sur de Manabí (UNESUM), pretendemos que los docentes y estudiantes universitarios, encuentren la fundamentación teórica y ejercicios prácticos que le ayuden a comprender y aplicar con seguridad la estadística como ciencia cabal del conocimiento humano.

LOS AUTORES

CAPÍTULO I. VARIABLES ALEATORIAS Y FUNCIONES DE DISTRIBUCIÓN

Introducción

La aplicación de la Estadística requiere conocimientos mínimos de la Teoría de las Probabilidades y en particular de los conceptos básicos de "variable aleatoria" y "función de distribución".

Aceptaremos como válido el concepto intuitivo de probabilidad de un evento, como el de la frecuencia relativa con que ese evento aparece cuando se repite -un número suficientemente grande de veces- el "experimento" que aleatoriamente puede dar lugar a él.

Sería muy bueno que el lector estuviera familiarizado con los resultados básicos de la Teoría de Probabilidades; pero si no lo está, necesita apenas conocer que:

-La probabilidad de cualquier evento es un número entre 0 y 1 (a veces representado por un porcentaje) -Que el conjunto de todos los resultados de un experimento debe tener una probabilidad total igual a 1, y -Que la probabilidad de que ocurra al menos uno de dos eventos -que no pueden ocurrir simultáneamente- es la suma de las probabilidades de cada uno. Otras leyes del cálculo de probabilidades se manejarán directamente a partir de la intuición. Lo esencial es definir a continuación variable aleatoria y funciones de distribución.

El capítulo puede contener material más que suficiente para una primera lectura; pero quedará para consulta o guía de referencia en función de los intereses y las necesidades del lector.

1.1. Variables aleatorias. Funciones de distribución. Variables aleatorias discretas y continuas. Niveles de medición

Trataremos desde el principio con fenómenos aleatorios cuyo resultado directamente observado es uno o más números. Ello da origen al concepto de variable aleatoria, el concepto básico fundamental de la Estadística Matemática.

Llamaremos variable aleatoria, y la representaremos por X, a una cantidad que se puede determinar cuantitativamente y que tiene un carácter aleatorio, en el sentido de que en distintas observaciones de la misma categoría, puede tomar valores diferentes, no determinística mente precisados. La caracterización de una variable aleatoria supone por tanto, la determinación del conjunto de valores que puede tomar efectivamente la variable y además la probabilidad con que ellos pueden ser tomados.

Son ejemplos de variables aleatorias:

- El resultado del lanzamiento de un dado (número del 1 al 6).
- El peso y la estatura de un grupo de personas.
- El rendimiento de cierto cultivo cosechado bajo determinadas condiciones, etc.

La caracterización de estas variables supone, como se establece en la definición, la precisión de los valores posibles y sus probabilidades. Un aspecto que aparentemente separa la teoría básica de la Estadística Matemática de la Teoría General de Probabilidades es que en Estadística se parte del supuesto que tratamos exclusivamente con fenómenos aleatorios cuyos resultados son numéricos. En realidad, partimos de este supuesto para presentar el concepto de variable aleatoria; pero debe tenerse en cuenta que aunque el resultado de un fenómeno aleatorio no sea un número, podemos siempre representarlo como tal.

Así por ejemplo, si en cierta investigación estudiamos el sexo de niños al nacer, y queremos considerarlo una variable aleatoria, podemos asociar el número 1 al resultado "varón" y el 0 al resultado "hembra" o cualquier otro convenio de enumeración. Si otra variable es por ejemplo el "municipio de residencia", en lugar de los nombres de los municipios, podemos utilizar un sistema de codificación numérica de ellos.

Lo que distingue en realidad la teoría básica de la Estadística Matemática de la Teoría General de Probabilidades es que ésta pone en foco la definición estricta de una función de probabilidad en cierto espacio de resultados adecuado y rigurosamente caracterizado, mientras que la teoría de las variables aleatorias pone en foco el conjunto de valores posibles de la variable y la forma de caracterizar la probabilidad de que la variable tome ciertos valores de interés.

Variables aleatorias discretas

Se dice que una variable aleatoria es discreta si el conjunto de sus valores posibles es finito o a lo sumo infinito, pero numerable. La caracterización de tal variable se logra enumerando los valores posibles y sus probabilidades:

X: x x x ... x ...

p p p ... p ...

Aquí p = Prob [X = x], es un número real entre 0 y 1 para k=1, 2,3,... y, de acuerdo a la definición de probabilidad debe satisfacer que:

$$\mathbf{p = 1}$$

La sumatoria anterior es una suma finita para el caso en que el Conjunto de valores posibles de X sea finito o la suma de la serie.

Si el número de valores posibles es infinito numerable.

A partir del conocimiento de los valores p, se puede calcular la Probabilidad de que X tome cualquier conjunto de valores B, como la suma de las probabilidades de los valores x que están en B:

- **Ejemplo 1.** Sea X el resultado del lanzamiento del dado y supongamos primero que éste no está cargado. Entonces:

  X: 1 2 3 4 5 6 y

  p = Prob(X=k)=1/6 para todos los valores de **k=1,...,6**

Si el dado estuviera cargado, la variable aleatoria cambia completamente. Ella solo puede ser caracterizada si sabemos la probabilidad de cada uno de los resultados posibles. Supóngase

Por **ejemplo** que el dado está cargado de manera tal que el 5 y el 6 son ambos, resultados que aparecen con la doble frecuencia que cualquiera de los restantes. Entonces, sabemos al menos que:

p =p =p =p = p y **p =p = 2p**

Y teniendo presente que la suma de estas probabilidades debe ser 1, obtenemos:

p + p + p + p + 2p + 2p = 1, u 8p = 1, de donde resulta p=1/8, esto es:

X:	**1**	**2**	**3**	**4**	**5**	**6**
p:	**1/8**	**1/8**	**1/8**	**1/8**	**1/4**	**1/4**

Si quisiéramos calcular entonces cuál es la probabilidad de que al lanzar el dado tuviésemos un resultado mayor que 3, tendríamos:

Prob [X > 3] = 1/8+1/4+1/4 = 5/8.

- **Ejemplo 2.** Sea X el sexo de un niño al nacer y supongamos que es conocido que la probabilidad de un nacimiento varón es de 0.51. Entonces, si codificamos el sexo masculino como 1 y el femenino como 0, tendremos la caracterización de la variable aleatoria

X:	**0**	**1**
	0.49	**0.51**

Sea ahora Y la variable que representa el número de niños varones entre 100 nacimientos vivos. Entonces Y es otra variable aleatoria cuyos valores posibles son:

0 1 2 3 ... 100

Y se puede probar que para **k=0, 1,2,...100**:

p = **Prob (Y=k) = p q = [---------] p q**

donde **p=0.51 y q = 1-p = 0.49.**

Con cálculos prácticos puede comprobarse que los valores extremos (cercanos a 0 o a 100) son muy poco probables, mientras que los valores de p son mucho mayores alrededor de 51. Pero un hecho que nos puede llamar la atención, es que incluso p es bastante pequeño (aproximadamente 0.08). No nos dejemos engañar por una intuición falsa.

En realidad, la probabilidad de que en 100 nacimientos haya exactamente 51 varones no tiene por qué ser alta.

De mucho más interés puede ser la pregunta sobre cuál es la probabilidad de que en los 100 nacimientos aparezcan entre 45 y 55 varones:

s & 100 * k 100-k

Supóngase por último que observamos cierto día los nacimientos en un Hospital Materno Infantil y consideramos la variable Z que denota el número de partos necesarios hasta que aparezca el primer varón.

En principio, Z es una variable que tiene los valores posibles:

1 2 3 4 5 ... (sin un límite finito)

y sus probabilidades pueden ser determinadas por la fórmula:

p = Prob [Z=k] = q p, siempre con **p=0.51 y q=0.49,**

Para cada k=1,2,...

Naturalmente, que valores de Z muy grandes, son poco probables; pero en principio factibles.

Obsérvese que según la serie geométrica:

p = p 1 + q + q + q +...+ q +... = p [---] = 1

Si quisiéramos calcular por ejemplo, cuál es la probabilidad de que haya que esperar tres o más nacimientos para observar un varón, tendríamos:

Prob [Z_3] = p = 1 - p - p = 1 - p - pq = q = 0.24

Funciones de distribución

Supongamos que el conjunto de valores posibles de la variable aleatoria X no es necesariamente discreto, esto es, que X puede tomar un conjunto infinito de valores más que numerable, por ejemplo, los valores x en un intervalo dado de números reales:

a<x<b. Nos preguntamos ahora como caracterizar la ley de probabilidad asociada.

Para facilitar el problema, podemos considerar que el conjunto .de valores posibles de X es siempre el conjunto de todos los números reales, esto es el intervalo **(-_,+_)** y convenir en asignar probabilidad cero a los valores o intervalos de valores que realmente no se alcanzan. Este criterio vale incluso para variables discretas.

Para caracterizar una variable aleatoria en general, se utiliza, en lugar de la probabilidad de cada valor posible, la función:

F(x) = Prob [X _ x]

que se conoce como función de distribución de la variable aleatoria.

Para que la función F(x) defina una función de probabilidad, es necesario y suficiente que:

1ro. F(x) sea creciente en R

2do. Que exista el límite de F(x) cuando x tienda a -_ y sea cero.

3ro. Que exista el límite de F(x) cuando x tienda a +_ y sea uno.

4to. Que F(x) sea continua por la derecha, esto es, que en cada punto a e R, se tenga.

Con estas propiedades para F(x), ella sirve para calcular la probabilidad de que X tome cualquier conjunto de valores de interés. En particular:

Prob [a < X _ b] = F(b) - F(a)

Prob [a _ X _ b] = F(b) - F(a) + p(a)

Donde p(a) = F(a) - lim F(x), es el posible salto de la función F(x) en a y que representa la probabilidad de que X tome el valor puntual a. Nótese que si "a" es un punto de continuidad para F(x), la probabilidad de este valor particular es nula.

Prob [a < X < b] = F(b) - F(a) - p(b)

Prob [a _ X < b] = F(b) - F(a) + p(a) - p(b)

Cuando F(x) es continua en todo R, las probabilidades de cada punto particular son nulas y las probabilidades de los 4 intervalos: (a,b], [a,b], (a,b) y [a,b) coinciden con F(b) - F(a).

- **Ejemplo 3.** Tomemos la variable aleatoria Y del ejemplo 2 anterior, que es discreta.

Determinemos F(x) = Prob [Y _ x] para cualquier xeR. Se tiene:

La función de distribución F(x) es en este caso, una función escalonada con salto en cada valor entero **k=0,1,2**,...,100 igual a la probabilidad de este valor. Algo similar ocurre con cualquier variable aleatoria discreta.

- **Ejemplo 4.** Sea X el tiempo (en minutos) de trabajo sin fallos de un equipo (por ejemplo una máquina de tejer). El conjunto de valores posibles de esta variable es realmente el intervalo **(0,+_).** Con determinadas hipótesis se puede demostrar que la función de distribución de X tiene la forma:

F(x) = 0 si x _ 0

= 1 - e si x > 0

Donde l es cierta constante positiva que caracteriza la fiabilidad de la máquina.

Obsérvese que en este caso F(x) es continua en todos los puntos. Pierde sentido, por ejemplo, preguntarse cuál es la probabilidad de que la máquina trabaje sin fallos exactamente 2.7895 minutos; pero puede tener sentido preguntar cuál es la probabilidad de que la máquina trabaje sin fallos entre 2 y 3 minutos y ésta puede calcularse en términos de l:

Prob [2 < X < 3] = F(3) - F(2) = e - e

- **Ejemplo 5.** Una persona de 40 años de edad adquiere una póliza de seguros según la cual se le pagará cierta cantidad de dinero cuando alcance los 75 años, o a su beneficiario inmediatamente después de su muerte. Sea X el intervalo de tiempo entre el día que se contrata el seguro y el día efectivo del pago. Entonces X es una variable aleatoria que tiene cierta función de distribución igual a 0 para x<0, crece con continuidad de x=0 a x=35 y salta de modo discontinuo al valor 1 para x=35. La

precisión de la función de distribución F(x) en el intervalo 0 < x < 35 exige de cálculos de probabilidades de vivir entre 40 y 75 años bajo el supuesto que ya se alcanzaron los 40 y no van a ser precisados aquí. Lo que se trata es de mostrar un ejemplo, en el cual la función de distribución no es ni totalmente continua ni discreta, sino mixta, porque no hay por qué pensar que el hombre no pueda vivir más de 75 años.

Variables aleatorias continuas

Parece natural definir que una variable aleatoria es continua si su función de distribución F(x) es una función continua en todos los puntos de **R**. En realidad, para caracterizar una variable aleatoria continua, exigimos algo más fuerte sobre **F(x):** que ella sea continuamente diferenciable, esto es, que exista en casi todos los puntos y sea diferente de cero. La frase "en casi todos los puntos" puede entenderse, por los no matemáticos en el sentido más simple de "en todos los puntos excepto a lo sumo en un número finito".

Desde el punto de vista puramente matemático esta restricción puede ser esencial. Hay ejemplos de funciones continuas, crecientes de 0 a 1, diferenciables "en casi todos los puntos" con derivada nula; pero en la práctica tales funciones no se presentan como funciones de distribución de variables aleatorias conocidas.

En cambio, la suposición sobre la diferenciabilidad de **F(x)** y la existencia misma de f(x), nos da un instrumento más simple y manejable para el estudio de las variables aleatorias continuas.

La función f(x) se denomina función de densidad de probabilidad de la variable aleatoria continua X y a veces se denomina abreviadamente por sus siglas: fdp.

La relación entre la función de densidad de probabilidad y la función de distribución se precisa en el orden inverso por la integración:

F(x) = Prob [X _ x] = f(t)dt

y de hecho:

Para que una función f(x) sea la función de densidad de probabilidad de cierta variable aleatoria X es necesario y suficiente que:

- 1ro. f(x) esté definida y sea mayor o igual a para casi todos los x
- 2do. f(x) sea continua para casi todos los x
- 3ro. La integral: f(t)dt sea convergente y resulte 1.

- **Ejemplo 6.** Retomando el ejemplo 4) anterior, el tiempo de trabajo sin fallos de la máquina es una variable aleatoria continua cuya función de densidad de probabilidad, es:

f(x) = 0 si x < 0

= le si x > 0

- **Ejemplo 7.** Sea X el peso en kilogramos de cierto grupo de personas adultas. Si este grupo es estandar (no se trata de personas especiales por predominar en ellos enanos o deportistas por ejemplo), la distribución del peso se comporta simétricamente alrededor de un peso medio, digamos m Kg con una cierta desviación _, de manera que

Prob [m-_ < X < m+_] = 0.68,

Prob [m-2_ < X < m+2_] = 0.95 y

Prob [m-3_ < X < m+3_] > 0.99.

La función:

7 _ 8

es una **fdp** que satisface las propiedades anteriores y se dice entonces que el peso tiene una distribución normal.

Niveles de medición

La clasificación de las variables aleatorias en discretas y continuas (o en tercera instancia mixtas) debe a veces precisarse desde el punto de vista práctico con el concepto de "nivel de medición". Sin pretender dar definiciones demasiado exactas que solo dificultarían las ideas intuitivas, diremos que:

- Una variable aleatoria discreta tiene un nivel de medición "ordinal" si sus valores posibles han sido establecidos de manera que expresan cierto orden en las categorías del fenómeno aleatorio que caracterizan, aunque la magnitud absoluta de estos valores no tenga un significado específico.
- Se dice que la variable aleatoria discreta tiene un nivel de medición "nominal" si sus valores posibles han sido establecidos de manera que ellas no reflejan un orden en las categorías del fenómeno aleatorio que caracterizan.

Analicemos por ejemplo, la variable aleatoria X que representa las calificaciones en un examen y codificamos sus resultados en la forma:

5 Excelente

4 Bueno

3 Regular

2 Malo

Entonces, la variable X tiene un nivel de medición ordinal. Los valores posibles: 2, 3, 4 y 5 no tienen importancia en su magnitud absoluta pero reflejan el orden ascendente de las calificaciones.

Esta misma variable podría haberse codificado en orden inverso:

1 Excelente

2 Bueno

3 Regular

4 Mal

y mantuviera su carácter ordinal; pero si la codificáramos en la forma:

1 Excelente

2 Mal

3 Regular

4 Bien

el nivel de medición de la misma, fuera puramente nominal.

Analicemos como otro ejemplo, la variable X que representa el municipio de residencia de los individuos de un grupo. Suponiendo por ejemplo, que todos los miembros del grupo son de Guadalajara y conociendo que esta tiene 4 municipios, los valores de la variable aleatoria podrían ser codificados en la forma:

1 **Guadalajara, zona metropolitana**

2 **Tlaquepaque**

3 **Zapopan**

4 **Tonalá**

Esta variable siempre tendrá un nivel de medición nominal, porque las categorías o resultados (municipios) no tienen un orden establecido a priori con algún sentido.

En variable aleatoria continuas también se pueden distinguir dos niveles de medición: continuo por intervalos y radial.

Se dice que una variable aleatoria continua tiene un nivel de medición radial si sus valores posibles pueden relacionarse en función de su distancia a un cero absoluto.

Se dice que una variable aleatoria continua tiene un nivel de medición por intervalos si sus valores posibles pueden relacionarse por la distancia entre ellos pero no por su distancia a un cero absoluto.

En otras palabras, en variables aleatorias continuas con nivel de medición de intervalo tiene sentido conformar operaciones de suma y resta para expresar la relación entre los valores posibles. En una variable con nivel de medición radial tiene sentido además formar el cociente de sus valores para expresar la relación entre éstos.

Una longitud (digamos, la talla de un individuo) es una variable aleatoria continua con nivel de medición radial. Aunque no existen individuos de "talla 0" si existe el cero como medición absoluta de una longitud y por tanto, de la talla. Tiene sentido por ejemplo, decir que "el individuo A es 1.5 veces más alto que el individuo B", independientemente del sistema de unidades con que se mida la talla.

En cambio, la temperatura (digamos, la temperatura del medio ambiente) es una variable aleatoria con nivel de medición por intervalos y no radial, porque la "temperatura 0 absoluta", sino que depende del sistema de unidades. No tiene sentido por ejemplo, decir que "el día A fue 1.5 veces más caliente que el día B".

La determinación clara de los niveles de medición de cada variable- especialmente el nivel de medición nominal u ordinal en las discretas- es importante desde el punto de vista práctico para conocer cuáles son las características o propiedades de la variable aleatoria de interés. Por ello cualquier análisis estadístico que emprendamos en el futuro debe comenzar distinguiendo para cada variable, si se trata de:

-una variable aleatoria discreta con nivel de medición nominal

-una variable aleatoria discreta con nivel de medición ordinal

-una variable aleatoria continua.

A veces reducimos intencionalmente el nivel de medición de una variable. Esto ocurre frecuentemente en la confección de ciertas tablas demostrativas de resultados, cuando renqueamos una variable aleatoria continua en intervalos. Digamos por ejemplo, el peso en kilogramos de un niño al nacer, puede ser raqueado en las categorías:

1 **Menos de 2500 gramos**

2 Más de 2500 y menos de 4000 gramos

3 4000 gramos o más

Cuando hacemos esto, de un nivel de medición continuo, hemos pasado a un nivel de medición ordinal y por supuesto, con ello "perdemos" información. Tal reducción nunca es conveniente en el proceso mismo de captación de los datos, salvo que no sea posible precisar el dato exacto. Si tenemos el dato original con el máximo nivel de medición (de ser posible continuo) tendremos siempre una información más "fina"; y en cualquier caso siempre podremos "recodificar" o "raquear" la variable de una u otra forma que nos resulte de interés en el momento oportuno.

Hay una idea más de carácter práctico que se evidenció en algunos de los ejemplos anteriores. Si una variable aleatoria es discreta y ordinal, y el conjunto de sus valores posibles es realmente amplio, de manera que la probabilidad de cada valor particular sea relativamente pequeño, ella se comporta prácticamente como una variable aleatoria continua y puede ser tratada como tal.

La variable Y del ejemplo 2 sirve para una buena ilustración. Y toma como valores posibles, cualquiera de los enteros entre 0 y 100 y la probabilidad de cada uno no es tan interesante porque es demasiado pequeña. En la Teoría de Probabilidades se demuestra que

Prob [0 _ Y _ k] puede aproximarse incluso por una fdp como la formulada para el ejemplo 7, y en la práctica, esta variable se trabaja como una variable aleatoria continua con distribución normal.

Otro ejemplo típico, es la edad en un grupo de estudios. En principio, la edad (medida en años) es una variable aleatoria discreta con nivel de medición ordinal. Si el rango de edades de tal grupo es amplio, la edad puede tratarse prácticamente como una variable aleatoria continua. Si estamos tratando en cambio con un grupo de niños de edad escolar, de manera que la edad está entre 5 y 10 años, no es admisible tal consideración.

Probablemente la primera tarea práctica a entrenar en el aprendizaje de la Estadística es la del reconocimiento del carácter continuo o discreto de las variables aleatorias y de su nivel de medición.

1.2. Algunas distribuciones importantes

Ciertas funciones de distribución surgen tan frecuentemente en la práctica y asociadas a fenómenos de naturaleza diferente, que han merecido un nombre particular y ser objeto de estudio específico.

Es importante familiarizarse con estas leyes de distribución.

Ellas son formuladas lo más generalmente posible utilizando ciertos parámetros que en problemas concretos adquieren valores de constantes particulares.

Ley de Bernoulli

Se dice que una variable aleatoria X se ajusta a la ley de distribución de Bernoulli, si ella es dicotómica, esto es tiene dos valores posibles, que convenimos sean 0 y 1, y se cumple que:

Prob [X = 1] = p para cierto valor conocido p _ (0,1), y

Prob [X = 0] =q = 1 - p

La variable que representa el resultado "varón" en el nacimiento de un niño se ajusta a la ley de Bernoulli con p=0.51. Cualquier otra variable que represente el exito o fracaso de un suceso con probabilidad constante y conocida, se ajusta también a esta ley.

Ley binomial

Se dice que una variable aleatoria X tiene una ley de distribución binomial con parámetros n y p, n _ Z y p _ (0,1), si sus valores posibles son los (n+1) enteros positivos

El nombre de binomial para esta distribución se debe a que las probabilidades p se determinan en última instancia por los coeficientes del desarrollo del binomio de Newton **(p + q)**

El ejemplo más importante de variable aleatoria con distribución binomial se obtiene como el número de éxitos en n intentos independientes repetidos de un fenómeno de Bernoulli. Digamos para concretar, el número de "varones" en n nacimientos vivos, o el número de "seis" en n lanzamientos de un dado, o el número de asmáticos en un muestra de volumen n tomada con repetición de una población cuya proporción de asmáticos es conocida.

Ley de Poisson

Se dice que una variable aleatoria X tiene una distribución de Poisson con parámetro **l, l > 0,** si admite como valores posibles los enteros positivos:

k

La distribución de **Poisson** surge por ejemplo, como caso límite o aproximado de la distribución binomial cuando el valor de n es grande y p es pequeño, de manera que el producto **"np"** es prácticamente constante. El parámetro l es en este caso, precisamente, **np.**

La distribución de Poisson aparece en muchos otros fenómenos. Supóngase por ejemplo que la variable aleatoria **X** representa el número de llamadas a una central telefónica en el intervalo de tiempo **t** y admitamos que las llamadas son independientes y que la probabilidad de una llamada en el tiempo **dt** es proporcional a **dt: µdt**

Con estas hipótesis se puede demostrar matemáticamente que

Prob [X = k] = e [------- , k=0,1,2,.... y t>0

y entonces **X** tiene distribución de **Poisson** con parámetro l**=µt.**

Este resultado es importante en el diseño de centrales telefónicas.

Las hipótesis de las llamadas se pueden hacer en muchos otros casos (denominados por lo general "flujo elemental de sucesos") y resulta que obedecen también a la Ley de Poisson, por **ejemplo:**

- La probabilidad de que **k** átomos radioactivos se desintegren en el intervalo de tiempo **t.**
- La probabilidad de que **k** rayos cósmicos atraviesen un contador Geiger-Muller en el intervalo de tiempo **t.**
- La probabilidad de que un almacén venda k piezas de un cierto tipo de mercancía en el intervalo de tiempo **t.**

- La probabilidad de que **k** automóviles pasen por cierta calle en el intervalo de tiempo **t.**
- La probabilidad de que **ocurran k** suicidios en el intervalo de tiempo **t.**
- La probabilidad de **encontrar k** corpúsculos en un campo de área t en un microscopio, etc.

Ley de distribución uniforme

Sea X una variable aleatoria continua con valores posibles en un intervalo **(a,b).** Se dice que ella tiene una distribución uniforme en ese intervalo si su **fdp** es constante en ese intervalo.

Más precisamente:

f(x) = 0 si x < a

= 1/(b-a) si a < x < b

= 0 si x > b

Nótese que el valor constante en **(a,b)** queda únicamente determinado como **1/(b-a)** por la condición:

f(x) dx = 1

La distribución uniforme es la extensión continua de los eventos **"igualmente probables"** con carácter discreto.

Esta distribución es por ejemplo, la del error de lectura en la medición de un equipo cuya escala tiene una pasó **h.** En efecto, en tal caso, el error de lectura (por aproximación a la marca de la escala más cercana) está uniformemente distribuido en el intervalo **(-h/2,h/2).** La distribución uniforme es la que se usa además, en los conocidos problemas de cálculo de probabilidad geométrica.

Ley de distribución exponencial

Se dice que una variable aleatoria X tiene distribución exponencial si admite como valores posibles los reales positivos y la función de densidad de probabilidad en cada punto depende exponencialmente del mismo. Más precisamente:

f(x) = l e si x > 0

= 0 si x < 0

Habíamos visto como ejemplo, que el tiempo de trabajo sin fallos en un equipo se ajusta a una distribución exponencial con cierto parámetro l, que según veremos más tarde, puede interpretarse como el inverso del tiempo medio de trabajo sin fallos. Esta distribución aparece con mucha frecuencia en Teoría de la Fiabilidad.

Ley de distribución normal

La distribución más importante en la práctica es la distribución normal, dada por:

f(x) = -----] e -_ < x < _

Esta distribución tiene dos parámetros: **m y _.**

El ejemplo 7 muestra una variable que se ajusta a esta distribución y algunas propiedades que tiene. En el último epígrafe veremos en definitiva que si una variable aleatoria X es resultado o suma de un conjunto bastante grande de variables aleatorias, ninguna de las cuales tiene, por separado una influencia determinante sobre el resultado final, la variable X tiene aproximadamente distribución normal.

En lo sucesivo, para abreviar que X tiene distribución normal con parámetros m y _, escribiremos **X e N (m, _).** Si en particular, **m=0 y _=1**, la distribución se llama normal normalizada o estandarizada.

La función de distribución correspondiente:

5--6

Es la que aparece tabulada normalmente en los libros de Estadística. Cualquier cálculo con una variable **X e N (m, _)** no estandarizada, puede reducirse a ésta por la fórmula simple:

Prob [x _ X _ x] = _ [-----] - _ [-----]

Otras distribuciones

En el punto 4 veremos otras distribuciones continuas importantes que surgen como consecuencia de operaciones entre variables aleatorias, particularmente, la distribución Chi-cuadrado, la distribución de **Student** y la distribución de Fischer.

Existen otras distribuciones que surgen frecuentemente en la práctica y que merecen observarlas al menos una vez para familiarizarse con ellas.

Entre las discretas, vale la pena citar la distribución de Pascal o distribución de **Furry** (con este último nombre se le conoce en la teoría delos rayos cósmicos). la distribución de Polya o distribución binomial negativa (que generaliza de hecho a la distribución de **Poisson** y la de Pascal), la ley de distribución **multinomial** (que generaliza la binomial a variables con más de 2 valores posibles), las distribuciones geométrica e **hipergeométrica**, etc.

Entre las continuas, vale la pena citar la distribución de **Weybull**, que surge también en teoría de fiabilidad y las distribuciones de **Cauchy**, de **Maxwel**l, de **Rayleigh** y las distribuciones Beta y Gamma.

1.3. Características numéricas de variables aleatorias

Asociadas con las variable aleatorias y su distribución existen una serie de características numéricas que explican tendencias generales de la distribución y cuyo conocimiento permite en algún sentido precisar la misma. Para una caracterización completa de la variable aleatoria, es necesario por supuesto, conocer toda la función de distribución; pero en muchas aplicaciones esto puede ser difícil y sin embargo, es conocido al menos, que los valores más probables de X yacen en un intervalo relativamente pequeño.

Por tanto, el conocimiento de tal intervalo es importante para obtener una idea aproximada de la distribución completa. Entre las características más importantes que ayudan a caracterizar este intervalo se encuentran la esperanza matemática y la dispersión.

Esperanza matemática y dispersión

Suponga primero que X es una variable aleatoria discreta:

X: x x x x ...

p p p p ...

Se llama esperanza matemática de X a: M [X] = x p si es esta suma existe

La existencia de la suma es trivial cuando la variable toma un número finito de valores posibles; pero si el número de valores es infinito enumerable equivale a exigir que la serie sea convergente. En la teoría es preferible exigir que la serie sea absolutamente convergente, para que la existencia de la esperanza matemática no dependa de los signos de codificación de los valores de la variable.

Se llama dispersión o varianza de X a:

D [X] = x - M(x) p

Si la suma es absolutamente convergente.

Si pensamos en p como una medida teórica de la frecuencia relativa **f =n /n** del valor **X=x** en n repeticiones de **X**, queda claro que la esperanza matemática es el valor medio teórico de la variable **X,** o como también se le llama, su valor esperado. La dispersión de X es una medida de cuánto los valores de la variable se separan o dispersan de su valor esperado **M[X].**

Suponga ahora que X es una variable aleatoria continua con **fdp f(x).** De manera análoga al caso discreto, se define la esperanza matemática de X como y se define la dispersión de X por la expresión:

M [X] = x f(x) dx

D [X] = (x - M[x]) f(x) dx

Cuando estas integrales convergen absolutamente.

Si X es una variable aleatoria en general, caracterizada por una función de distribución **F(x),** la esperanza matemática y la dispersión de X se definen por las integrales de **Riemann-Stieltjes**

- **Ejemplo 8.** Suponga que **X** es una variable aleatoria con distribución binomial y parámetros **n y p**. Entonces, el valor esperado de X es np y la dispersión de X es **npq**.

En efecto utilizando el binomio:

= np p (1 - p)

Con cálculos análogos se demuestra que **D [X] = npq**.

Así, en el ejemplo concreto de la variable aleatoria que mide el número de varones en 100 nacidos vivos, podemos afirmar que el número esperado de varones es **np=51** y la dispersión es **npq=25.**

- **Ejemplo 9.** Suponga que X tiene distribución exponencial con parámetro l. Entonces

M[X] = l x e dx = l -x e /l + e dx = [-]

- **Ejemplo 10**. Suponga que X e N(m,_). Se deja como ejercicio para el lector probar que la esperanza de X es precisamente **m** y su **dispersión** _

Mediana y moda.

La esperanza o valor esperado es una de las características de una variable aleatoria más frecuentemente usada. Ella puede ser definida para cualquier variable aleatoria (discreta o continua), con la única condición que la serie o la integral que la determinan sea absolutamente convergente. Pero no en todas las variables aleatorias, ellas tienen la misma importancia.

Si por **ejemplo**, la variable **aleatoria** es discreta, con nivel de medición ordinal, el valor esperado, depende en buena medida, de la forma de codificación de la variable y su interpretación debe manejarse con cuidado y en dependencia de esta codificación.

Supóngase por ejemplo que queremos medir con X el nivel de escolaridad en un grupo de personas. Una alternativa de medir X pudiera ser "el grado terminado": 0,1,2,3,...,12,13,14,...,17, indicando de 0 a 6 la enseñanza primaria (de acuerdo con el sistema de grados escolares en Cuba), de 7 a 12 la enseñanza media general y superior y de 13 a 17 la enseñanza universitaria.

Otra alternativa de medir X mucho "menos fina" sería por categorías de niveles terminados:

0 'Primaria sin terminar'

1 'Primaria terminada'

2 'Secundaria'

3 'Preuniversitario'

4 'Universitario'

El valor esperado de X en una determinada población puede ser apenas un indicativo del nivel medio alcanzado en la población; pero es fácil comprender que puede variar muy significativamente en dependencia del nivel de medición.

Si la variable fuese discreta nominal (piénsese de nuevo en el ejemplo de los municipios de Guadalajara), el valor esperado de X carece completamente de cualquier sentido intuitivo.

En variables aleatorias discretas con nivel de medición ordinal, la mediana puede ser una mejor medida de la tendencia central y en variables aleatorias discretas con nivel de medición nominal, la moda puede ser la mejor y única medida de tal tendencia.

Supongamos que **X** es una variable aleatoria discreta y ordenamos los valores de **X** en forma ascendente o descendente:

x _ x _ x _ ... _ x _ ...

Llamamos mediana de **X** a un número real μ que divide a la muestra en dos grupos igualmente probables, esto es, al valor **μ** tal que:

Si X es una variable aleatoria continua con fdp f(x), la mediana de X será análogamente el valor μ para el cual:

$$\int_{-\infty}^{\mu} f(x)\,dx = \int_{\mu}^{+\infty} f(x)\,dx = \frac{1}{2}$$

y para una variable general cualquiera con una función de distribución F(x) se define análogamente con la integral de **Riemann-Stieltjes.**

Puede probarse que si una variable aleatoria es simétrica respecto a un valor m, entonces la mediana y la media son iguales a m; pero para distribuciones no simétricas, la mediana y la media no coinciden necesariamente. La mediana de X se usa muchas veces como alternativa

de la esperanza matemática en la medición de la tendencia central porque ella es menos sensible a los valores extremos de la variable.

Sea X una variable aleatoria discreta. Llamaremos moda de X, al valor más probable de X, esto es, la moda es el valor x entre

p = max { p }

Si X es continua con **fpd f(x**) se llama moda de X al valor x para el cual se alcanza el máximo de **f(x).**

Si X es una variable simétrica alrededor de m y tiene un solo máximo, la moda coincide con m (la esperanza y la mediana); pero la moda es una característica de variables aleatorias que puede definirse incluso en variables aleatorias con un nivel de medición nominal.

Desviación media, rango o amplitud

De manera similar, existen además de la dispersión o varianza, otras medidas de la forma en que se agrupan los valores de la variable alrededor de la media.

Supóngase que X es discreta. Se llama desviación media de X a:

D = ? x - M[x] ? p

Si X es continua, se llma desviación media de X a:

D = ? x - M[x] ? f(x) dx

Cuando estas series o integrales convergen.

De manera análoga, se trabaja frecuentemente con la desviación estándar media de X, definida por:

ó

Obsérvese en general que

S = D[X]

Esto hace que S tenga el mismo sistema de unidades de X (mientras que D[X] se mide en unidades de X al cuadrado).

Para variables ordinales es útil considerar como medida de dispersión el rango o amplitud. Se llama amplitud de una variable aleatoria X a la diferencia entre el valor más probable y el menos probable.

Otras características numéricas

Se pueden usar otras características numéricas de una variable aleatoria, definidas a partir del siguiente concepto general. Supóngase que g(x) es una función real de variable real, y definamos el valor esperado de una variable aleatoria X respecto a la función g(x) por la expresión:

O

Estas características se definen, como siempre, en el caso en que la serie o la integral sean absolutamente convergentes.

Como puede apreciarse, las características antes definidas: M[X], D[X] y D [X] son todas casos particulares de esta definición con funciones g(x) adecuadas: g(x)=x, g(x)=(x-M[x]) , g(x)=?x-M[x]? respectivamente.

Se llama momento de la variable X de orden k, k=1,2, 3...a la característica numérica μ que resulta de tomar **g(x)=x.**

Así, el momento de orden 1 es el valor esperado de X, el momento de orden 2 el valor esperado del cuadrado de X, etc.

Se llama momento central de la variable X de orden k, k=1,2,3,... a la característica numérica n que resulta de tomar:

g(x)=(x- M[x]) .

Así, el momento central de orden 2 es la dispersión. Después veremos que los momentos centrales de orden impar son siempre nulos.

Tomando otras funciones g(x) se obtienen características numéricas que luego comentaremos en el curso.

1.4. Distribución conjunta de varias variables aleatorias

Concepto de independencia.

Como todas las fórmulas y conceptos se pueden generalizar inmediatamente de dos a más dimensiones, formularemos solo las ideas con pares de variables aleatorias **(X,Y)** Sean **X** y **Y** variables aleatorias. Se llama distribución conjunta de **X e Y** a la función

F(x,y) = Prob [X _ x , Y _ y]

Se puede demostrar que si F(x,y) es una función de distribución se cumple que:

b) lim F(x,y) = 0 para cualquier y fijado

Además, para cada y fijado, **F(x,y)** es una función creciente de la variable x, y para cada x F(x ,y) es una función creciente de la variable y.

Como en el caso unidimensional, podemos hablar en particular de pares de funciones aleatorias discretas y continuas. Diremos que **(X,Y)** es discreta, si puede tomar a lo sumo un número finito o enumerable de valores **(x ,y)** para los cuales están definidas las probabilidades **p** , de manera que:

t t ij

Diremos que el par de variables aleatorias **(X,Y)** es continua si existe una función f(x,y), que llamaremos función de densidad de probabilidad conjunta de X e Y, tal que:

F(x,y) = f(t,u)dtdu,

Se demuestra en análisis que:y está claro que la definición de probabilidad exige que:

f(x,y)dxdy = 1,

Concepto de independencia

En general, el conocimiento de la distribución de X y la de Y no es suficiente para conocer la distribución del par **(X,Y).** Esto se logra fácilmente sólo en el caso en que las variables X e Y son independientes.

Se dice que dos variables aleatorias X e Y son independientes si la función de distribución del par es el producto de las funciones de distribución de cada variable por separado.

Si **F (x**) es la función de distribución de **X, F (y),** la función de distribución de Y y F **(x,y),** la función de distribución conjunta, la noción de independencia exige que:

En el caso particular de variables aleatorias discretas, si

p = Prob [X=x] p = Prob[Y=y]

y p = Prob[X=x ,Y=y],

la definición de independencia exige que p = p p para todos los i y j posibles.

En el caso particular de variables aleatorias continuas, si **f (x) es la fdp de X, f (y), la fdp de Y y f (x,y), la fdp** conjunta, la noción de independencia exige que:

- **Ejemplo 11)** Consideremos el disparo de un cañón y supongamos que la desviación acimutal X y la desviación de altura Y, para un blanco de coordenadas son variables independientes y normalmente distribuidas con parámetros **m , _ , m , _** respectivamente. En tal caso, la distribución conjunta del par **(X,Y)** está dada por la **fdp:**

  1 # x y $

Pero esta no es la forma de la distribución normal bidimensional en general. Cuando X e Y no son independientes aparece un término más en la exponencial y se corrige el factor para garantizar que la integral sobre el plano de 1:

Aquí r es el nuevo parámetro que se encuentra entre -1 y 1 y que expresa en algún sentido la dependencia de X e Y. Por ejemplo, para el caso en que r = ±1, las variables X e Y son proporcionales y la masa total de probabilidad está concentrada en la recta:

y = m ± [---] (x - m)

Como veremos después en el estudio de regresión.

A menos que tengamos criterios de antemano para conocer si las variables son independientes, la distribución conjunta de un par de variables tiene más información que las distribuciones por separado.

En efecto, si tenemos la **fdp f (x,y)** conjunta del par de variables **(X,Y),** podremos obtener las **fdp f (x) y f (y) de X** e Y respectivamente por las integrales:que se denominan distribuciones marginales de **X e Y** respectivamente para enfatizar que fueron obtenidas a partir de la distribución del par.

En el caso discreto las definiciones e ideas son similares sustituyendo integrales por sumas.

- **Ejemplo 12)** Supongamos que (X,Y) tiene una distribución normal bidimensional como definimos en el **ejemplo 11**, con parámetros **m ,_ ,m ,_ ,r.** No resulta difícil demostrar que independientemente del valor de r, la distribuciones marginales de X e Y resultan normales con parámetros **m ,_ y m ,_** respectivamente.

Este solo ejemplo basta para ilustrar que pueden existir infinitas distribuciones bidimensionales con las mismas distribuciones marginales.

Dado el par de variables aleatorias (X,Y), se llama distribución condicional de X respecto a Y a la función:

XY

Desde el punto de vista de estas definiciones, **X y Y** son independientes si y solo si:esto es, si la distribución condicional de X a Y coincide con la distribución marginal de **X** o la distribución condicional de **Y a X** coincide con la distribución marginal de **Y**.

1.5. Generación de distribuciones a partir de operaciones con variables aleatorias conocidas

El desarrollo de la teoría y la práctica estadística exige constantemente formar operaciones con variables aleatorias y plantea el problema de buscar las distribuciones de la variable resultado a partir de las distribuciones de las variables originales, o por ejemplo, establecer relaciones entre las características numéricas de la variable resultante con las características numéricas de las variables originales.

La teoría de la distribución conjunta da una potente arma matemática para obtener este tipo de resultados. De acuerdo a los objetivos del curso, nos conformamos con formular y ejemplificar algunos de ellos sin pretender demostraciones.

Un ejemplo típico es el siguiente problema: Supóngase que **X** tiene una distribución conocida, digamos continua, con **fdp f (x)** y consideramos la variable aleatoria

Y = a X + b

Esto es, la variable aleatoria que resulta de una transformación lineal de X. ¿Cuál es entonces la distribución de Y?, o bien ¿Qué relación existe entre la distribución de X y la de Y?

En este caso es fácil ver que:

Prob [Y_y] = Prob X _ [----]

y - b

Por tanto:

f (y) = --] f ([-----])

Así resulta en particular, que si X tiene una distribución normal, Y también tendrá una distribución normal, con esperanza trasladada y dispersión corregida en función de los valores **a y b.**

Si X e N(m,_) y se define Y=[-----], entonces resultará Y e N(0,1)

Pero también debe observarse que esto no es cierto para cualquier distribución. Si por ejemplo, X tiene distribución exponencial, la distribución de Y no es exponencial, a menos que b sea 0. De hecho, el conjunto de valores posibles de Y no serán los reales positivos sino trasladados.

La respuesta se complica si en lugar de una simple transformación lineal de X hacemos una transformación más compleja.

Otra pregunta típica. Supóngase que (X,Y) es un par de variables aleatorias con **fdp** conjunta **f(x,y)** conocida y formamos:

Z = X + Y, ó Z = X - Y, ó Z = X*Y ó Z = X/Y

Suponiendo en este último caso que Y no toma el valor 0. ¿Cuál es en cada caso la fdp de Z?

Resulta por ejemplo, para la suma que la **fdp** de **Z** es:y en el caso particular que **X e Y** sean independientes,

Con este resultado se puede demostrar en particular que la suma de dos variables aleatorias con distribución normal tiene también distribución normal. Más precisamente: dónde _ puede expresarse en términos de _ y _, en una forma no trivial que precisaremos más tarde.

Observe sin embargo, que la suma de dos variables con distribución exponencial, por ejemplo, no es una variable con distribución exponencial.

Uno de los resultados más frecuentemente usados en Estadística Inferencial es que la media de un conjunto de n variables aleatorias distribuidas normalmente tiene también distribución normal. Más precisamente, si **X , X , ...,X** son variables aleatorias con distribuciones respectivas **N(m ,_), i=1,2,...,n y**

X = (X + X + ... + X) / n,

Entonces, X tiene distribución **N(m,_)** donde m es precisamente:

m= (m +m + ... +m)/n

y _ se determina a partir de _ ,_ ,...,_ y las correlaciones entre las variables X . En particular si **X ,X ,...,X** son independientes, resulta

_ = r _ + _ + ...+ _ / n.

Si como situación singular, pero bastante frecuente, todas las X tienen la misma distribución: **X e N(m,_),** resultará que la media:

X e N(m,_/r n).

Uno de los resultados más fuertes de la Teoría de Probabilidades o de la teoría de las variables aleatorias, generaliza el resultado anterior para demostrar, como antes comentamos, que: Si tenemos n variables aleatorias, si **X , X , ..., X** con cualquier distribución y X es la suma de estas variables, n es suficientemente grande y ninguna de las X tiene -en cierto sentido que no vamos a precisar, pero que intuitivamente puede ser claro- una incidencia o papel determinante en la suma, entonces ésta tiene una distribución aproximada a la normal. Utilizaremos este resultado para sospechar que una variable pueda tener distribución normal.

Resultados análogos; pero por supuesto muy diferentes en forma, se encuentran en otras operaciones además de la suma y trabajando con distintas distribuciones. Particularmente importantes, por la frecuencia con que se encuentran en la práctica son los siguientes:

Distribución c (Chi-cuadrado)

Supóngase que X , X , ..., X son todas variables aleatorias independientes, normalmente distribuidas con esperanza 0 y la misma dispersión _ y formamos la variable aleatoria:

2 2

Entonces, se puede demostrar que la distribución de Z tiene la forma:

2 _ _(n/2)

Que aparecerá muy frecuentemente en la práctica. Esta distribución se conoce como distribución Chi-cuadrado con parámetros n y _ y es usual denotar las variables que tienen esta distribución en la forma c. En la práctica recordaremos que la suma de cuadrados de variables aleatorias independientes, con distribución normal centrada y con idéntica dispersión, resulta en una distribución chi-cuadrado.

La función _(x) que aparece en la expresión explícita de la fdp es la función especial gamma, generalización del factorial. Como por suerte, los valores de la distribución c están tabulados o incorporados a paquetes de programación, este detalle no llega a ser importante para los no matemáticos.

Distribución de Student

Supongamos que **X** es una variable aleatoria con distribución normal con esperanza 0 y dispersión _ , y sea **Y** una variable aleatoria con distribución **c** con parámetros **n** y _. En la práctica estadística se forma frecuentemente el cociente:

5-----

y resulta que este cociente tiene una distribución estadística bien determinada y útil, dada por:

f(t) = [------] [------------ 1 + [---]

y que se conoce como distribución de Student con parámetro n o con "n grados de libertad". Es muy interesante que la distribución de Student resulte independiente de la _ inicial. También esta función está tabulada y aparece incorporada en los paquetes. En la práctica, lo que debemos recordar es que el cociente de una distribución normal por la raíz cuadrada de una distribución chi-cuadrado resulta en una distribución de Student.

Distribución de Fischer

Supongamos que **X y X** son variables aleatorias independientes, ambas con distribución c y parámetros **n y _ y n y _** respectivamente. En las aplicaciones se conforma frecuentemente el cociente: y su distribución resulta también independiente de _ y dada por:

1 ²

Para x positivo y 0 para **x<0**. Esta función se conoce como distribución de Fischer con n y n grados de libertad. Para los no especialistas se pide recordar apenas que el cociente de dos distribuciones c independientes conduce a una distribución de Student.

El otro tipo de problemas planteado, es expresar por ejemplo:

M[X+Y], M[X-Y], M[XY], D[X+Y], D[X-Y], etc., en términos de **M[X], M[Y], D[X], D[Y]**

En este sentido se tienen resultados como los siguientes:

M[X + Y] = M[X] + M[Y]

M[cX] = cM[X] para cualquier constante c

M[XY] = M[X]M[Y] si las variables son independientes.

Así, la esperanza matemática es lineal en las variables; pero la esperanza de un producto de variables aleatorias no siempre es el producto de las esperanzas matemáticas de cada variable. Ello ocurre si ellas son independientes.

Para la dispersión, las propiedades son ligeramente diferentes.

En general:

D[X+Y] = D[X] + D[Y] si las variables X e Y son independientes.

D[cX] = c D[X] para cualquier constante c.

D[X+c] = D[X] para cualquier constante c.

La última propiedad significa que la dispersión no se altera por una traslación simple.

Todas estas propiedades pueden demostrarse a partir de la definición de esperanza matemática y dispersión, o las de esta última, pueden lograrse teniendo presente que la definición de dispersión de X es equivalente a:

D[X] = M (x-M[x]) = M[X] - M[X]

Propiedades similares pueden investigarse para otras características numéricas como los momentos, teniendo presente que el momento de orden k, es en esencia:

μ = M [X]

y el momento central de orden k, es

$n = M\ (X - M[X])$ para cualquier k=1,2,....

1.6. Conclusiones

En este capítulo hacemos un resumen de los conocimientos que podemos requerir de la teoría básica de variables aleatorias.

Realmente se incluyen muchos más conceptos de los que pueden ser estrictamente necesarios para lograr aplicaciones en diferentes ramas; pero en todo caso quedan como material de consulta.

Para los no especialistas en matemáticas o ingenierías, probablemente lo más importante es el concepto intuitivo de función de distribución; la distinción de variables aleatorias discretas y continuas y sus niveles de medición; la comprensión de cuáles características de una variable aleatoria reflejan mejor la tendencia central o la dispersión en dependencia del nivel de medición y el concepto probabilístico de independencia.

CAPÍTULO II. INTRODUCCIÓN AL SPSS/PC

Introducción

El SPSS/PC es un paquete de análisis de información general diseñado para el ambiente interactivo de la IBM PC. Sus primeras siglas (SPSS) significan Statistical Package for the Social Sciences, porque fué concebido como un paquete para el análisis estadístico especialmente rico en métodos no paramétricos y desarrollado para máquinas grandes IBM compatibles. Con el tiempo, el paquete fue enriqueciéndose en métodos de procesamiento estadístico general y especialmente de análisis multivariado; pero conservó su nombre original. Las últimas siglas (/PC) especifican que estamos hablando ahora de su versión para Personal Computers.

Los diseñadores del SPSS se convirtieron en una Compañía de Software (SPSS Inc) y hoy en día, el SPSS está ranqueado como uno de los paquetes de procesamiento estadístico más rigurosos y completos.

Disponemos aquí, las siguientes versiones para DOS:

- SPSS/PC, versión 1.0, desde 1985
- SPSS/PC+, versión 1.0 de aproximadamente 1986
- SPSS/PC+, versión 3.x, de aproximadamente 1990
- SPSS/PC+, versión 4.x, de 1992.

y trabajaremos prácticamente con la versión 3.1 o superior, aunque la mayoría de las ideas son válidas para versiones anteriores.

Ya salió al mercado la versión del SPSS/PC para Windows y debe constituir una revolución desde el punto de vista de las formas de trabajo y del aprovechamiento de las posibilidades gráficas del Microsoft Windows.

Desde antes se disponía una versión del SPSS para máquinas grandes (por ejemplo EC-1022 o EC-1035 del sistema SUMCE, compatibles con la IBM-360), y haremos referencia a ella bajo el nombre SPSS para explicar su posible enlace con el SPSS/PC.

Este capítulo pretende sólo mostrar las posibilidades y formas generales de trabajo del SPSS/PC para DOS y debe ser complementado con el Manual de Usuarios correspondiente.

2.1. Posibilidades del SPSS/PC

El SPSS/PC+ facilita la escritura de reportes, la construcción de tablas y las pruebas estadísticas que permitía el SPSS , ahora en una microcomputadora IBM compatible. Estas posibilidades permiten usar el SPSS/PC para aplicaciones en:

- Estadística Técnica
- Estadística Agropecuaria
- Ciencias Médicas
- Ciencias Sociales
- Ciencias de la Conducta
- Control de la Calidad
- Análisis de Encuestas

El SPSS/PC brinda entre otras las siguientes facilidades:

- Entrada y ejecución interactiva de los comandos

- Ayuda mientras se trabaja
- Salida hacia pantalla, impresora, fichero en disco o una combinación de las anteriores.
- Salida formateada para una pantalla de 80 columnas o una página impresa.
- Ficheros y procedimientos compatibles con el SPSS .

Con el SPSS/PC, es posible efectuar las siguientes operaciones:

- Entrar datos interactivamente o desde un fichero.
- Recodificar variables y calcular nuevas variables.
- Salvar datos hacia ficheros recuperables.
- Transferir ficheros entre micros y mainframes
- Tabular variables y combinaciones de variables.
- Efectuar una serie de análisis estadísticos.
- Producir reportes formateados.
- Entrar los datos originales, matrices o ficheros con su propia definición, al estilo del BASE.
- Incluir las observaciones selectivamente en reportes o análisis**.**

Entre los análisis estadísticos se incluyen:

- Análisis descriptivo de datos
- Tablas de contingencia
- Test paramétricos y no paramétricos de comparación de dos poblaciones.
- Análisis de varianza paramétricos y no paramétricos.

Análisis de varianza paramétrico multifactorial y multivariado

- Análisis de regresión simple y múltiple. Análisis de correlación.
- Análisis discriminante
- Análisis de factor.
- Análisis de clusters.
- Series de tiempo.

2.2. Formas de trabajo del SPSS/PC para DOS

El SPSS/PC opera vía comandos y tiene en general dos formas de trabajo: una interactiva y otra de procesamiento en lotes. En la primera, el usuario va introduciendo los comandos y observando los resultados uno a uno. En la segunda, escribe los comandos en forma de un programa (puede ser incluso fuera del **SPSS/PC)** y luego los ejecuta en lote.

La primera vía parece más atractiva; pero exige de mucho más tiempo en una máquina que tenga concretamente el paquete. La segunda vía parece más clásica o cercana al procesamiento en mainframes; pero es más económica. Las versiones del **SPSS/PC** se han distinguido -además de, por el potencial de los métodos estadísticos que contiene- por ir incrementando las facilidades de tratamiento interactivo; pero aun así, dentro del sistema de versiones para DOS, el procesamiento en lotes, o en todo caso combinado, sigue siendo más eficiente. En la versión del **SPSS/PC** para Windows, las posibilidades del procesamiento interactivo llevan éste a un plano cualitativamente superior.

En el **SPSS/PC+** versión 1.0, se pueden entrar interactivamente los datos, y con la ayuda de un HELP, dar todos los comandos; pero hay que teclear exactamente cada uno de estos. En el **SPSS/PC**, versión **3.x**, aparece un sistema de Menú que ayuda a la construcción como tal de los comandos, además de que se perfecciona el sistema de captación interactiva de datos; pero en cualquier caso, aprender a usar el **SPSS/PC** supone familiarizarse con su lenguaje de

comandos y el procesamiento el lotes facilita este aprendizaje porque torna las ideas generales más nítida.

Utilizaremos por ello la variante más clásica y económica.

Supondremos que prepararemos nuestros datos en forma de un fichero externo con ayuda de un editor de textos, un tabulador electrónico o un sistema gestor de bases de datos cualquiera. Prepararemos también el "programa" o secuencia de comandos con un editor de textos externo. Programaremos generalmente las salidas dirigidas hacia un fichero en disco y acudiremos al **SPSS/PC** (o incluso a la máquina que contiene este paquete) sólo para para correr el programa, corregirlo o perfeccionarlo -utilizando las facilidades interactivas- y obtener los resultados en disco. Estos resultados podrán ser analizados también en cualquier otro equipo o fuera del **SPSS/PC.**

2.3. Generalidades. Comandos. Tipos de comandos

Como todo lenguaje de programación, el lenguaje de comandos del **SPSS/PC** está sujeto a ciertas reglas sintácticas generales que aquí queremos esbozar. Una descripción detallada de estas reglas se debe consultar en el Manual de Usuario [1] o en el resumen que al respecto se elaboró en el Centro de Estudios de Informática **[2].**

Diagramas sintácticos

Todos los comandos del SPSS/PC serán descritos con ayuda de un diagrama sintáctico utilizando los convenios de la FNB: las palabras claves se escriben con mayúsculas para indicar que deben ser tecleadas exactamente; con minúsculas se escriben parámetros que deben ser sustituidos por el usuario; elementos entre corchetes son opcionales; etc.

Tipos de comandos

Existen tres tipos de comandos en el SPSS

- Comandos de operación: que fijan el modo de trabajo del SPSS/PC
- Comandos de definición de datos: que describen, etiquetan, y manipulan los datos
- Comandos de procedimientos: que realizan los procesamientos estadísticos en sí.

Orden de los comandos

Generalmente los comandos se ejecutan en el mismo orden que aparecen y este no es sintácticamente importante; pero debe tenerse en cuenta que los comandos que definen las variables deben preceder a los que la utilizan. Además, los comandos que realizan transformaciones de datos no se ejecutan hasta que no se leen los datos y la lectura de los datos no se hace hasta cada comando procedimiento.

Características generales de los comandos

Comienzan siempre con una palabra clave que es el nombre del comando y terminan con un punto. La longitud máxima de la línea de comandos es de 80 caracteres; pero los comandos pueden utilizar más de una línea. Se recomienda mucho utilizar un formato estructurado para claridad. Ejemplo de un comando **SET:**

```
SET
   DISK=ON /
   DISK='salida.lst' /
   EJECT=OFF /
```

SCREEN=OFF /

ECHO=OFF .

Subcomandos

Muchos comandos tienen opciones para definir datos, manipular datos y definir el formato de las respuestas. Estas especificaciones opcionales se llaman subcomandos y comienzan con una palabra clave y se separan con slash (/). Algunos subcomandos son opcionales y otros obligatorios. El orden puede o no ser importante. Ambas cosas se precisan en la sintaxis de cada comando. En el ejemplo anterior, **DISK, EJECT, SCREEN y ECHO son** subcomandos de **SET.**

Palabras claves

Las palabras claves son palabras definidas especialmente por el SPSS/PC para identificar comandos, subcomandos, funciones, operadores y otras especificaciones. En el ejemplo anterior, **SET, DISK, EJECT, SCREEN, ECHO, ON y OFF** son palabras claves. Todas las palabras claves pueden abreviarse con las tres primeras letras, excepto WITH. Casi todas las palabras claves son reservadas y no se pueden utilizar como nombres de variables. La palabra clave ALL se usa para referirse a todas las variables definidas por el usuario. La palabra reservada TO se usa entre dos nombres de variables para especificar una lista de estas que incluye a ambas.

Valores

Los valores se refieren a especificaciones en comandos o datos que se procesan con el SPSS. Cuando se especifica un número puede escribirse con o sin ceros a la izquierda. Si un valor de variable es un texto y se usa en un comando tiene que situarse entre apóstrofes o comillas.

Delimitadores

Los delimitadores se usan para separar valores de datos, palabras claves, argumentos y especificaciones. El espacio es el delimitador más usado. Los operadores **aritméticos (+,- ,* y /),** algunos símbolos especiales como paréntesis, apóstrofes, comillas, el slash, el signo igual y otros son también delimitadores y pueden estar precedidos o seguidos por espacios en forma opcional.

Las comas se usan para separar los argumentos de una función.

Cadenas

La palabra cadena se usa para referirse a datos alfanuméricos o especificaciones como títulos y etiquetas. Las variables de cadena que contiene 8 o menos caracteres se denominan variables de cadenas cortas. Si tiene más (**hasta 255 caracteres**) se denominan cadenas largas. En los comandos de transformación solo se pueden utilizar variables de cadena corta o valores de cadena.

Ficheros

Los ficheros que se pueden usar durante una sesión de trabajo con el SPSS/PC son:

- **Fichero activo:** Es el fichero en memoria más recientemente definido en una sesión por los comandos de definición y transformación de datos. Contiene los datos y el diccionario de datos (etiquetas de variables, etiquetas de valores de variables, formatos, definición de valores perdidos, etc). Se puede modificar este fichero utilizando transformaciones datos o salvarlo usando los comandos SAVE o EXPORT. El

fichero activo como tal no tiene nombre pues se mantiene solo en memoria. Cuando se graba, con SAVE lo hacemos normalmente con extensión .SYS y le llamamos fichero sistema.

- **Fichero de comandos**: Es un fichero texto que contiene comandos del SPSS/PC y que denominamos programa. Se utiliza el comando INCLUDE para procesar los comandos de un fichero de este tipo como vía de procesamiento en lotes. Es un buen hábito nombrar estos ficheros con una extensión fijada, por ejemplo: .SPS
- **Fichero de datos**: Es un fichero texto que contiene solamente datos de entrada. Es un buen hábito nombrar este fichero con una extensión fijada, por ejemplo: .**DAT**. En principio el fichero de datos para el SPSS/PC debe considerarse como una matriz rectangular de datos, cada columna representa una variable y cada fila un caso o artículo. Las excepciones de este formato serán tratadas posteriormente.
- **Fichero en disco:** Es un fichero texto que contiene respuestas de procedimientos del SPSS. Por defecto se denomina **SPSS.LIS** pero es buen hábito ponerle nombre propio con extensión fija **(.LIS o .LST) para que no se sobreescriban resultados.**
- **Fichero de ocurrencias:** Es un fichero texto que crea el **SPSS/PC** y que contiene todos los subcomandos que se entran por el usuario y que se procesan durante una sesión de trabajo. Puede servir para dejar registrada una sesión de trabajo interactivo en forma de un programa. Tiene por defecto el nombre **SPSS.LOG**.
- **Fichero de resultados especiales:** Es un fichero texto que contiene un conjunto rectangular de datos producido por el comando WRITE o ciertas matrices de resultados producidas por los comandos **CLUSTER, CORRELATION, FACTOR, ONEWAY o REGRESSION**. Por defecto este fichero se denomina SPSS.PRC.
- **Fichero de sistema:** Es un fichero binario creado por el comando SAVE del SPSS/PC. Contienen los datos y el diccionario del fichero activo en el momento de la grabación.
- Se recarga con el comando GET y se convierte entonces en un fichero activo. Normalmente se identifican por la extensión SYS.
- **Fichero portable**: Es un fichero texto que contiene datos y un diccionario para intercambiar con el SPSS o el SAS. Se genera en el SPSS con el comando EXPORT o en el SAS con el usando PROC TOSPSS. Se carga en el SPSS con el comando IMPORT y se convierte entonces en un fichero activo.

Variables

Las reglas para la definición de variables son las siguientes

- Sobre el nombre de las variables: Tienen un máximo de 8 caracteres y el primer carácter debe ser una letra. Los nombres de las variables se establecen en el comando DATA LIST, **COMPUTE, COUNT o IF. En el DATA LIST** se pueden definir variables a partir de un nombre raíz y un número final utilizando la palabra TO (ejemplo CALIF1 TO CALIF10). Después cualquier referencia de lista de variables con TO se refiere al orden en el fichero activo.
- Sobre las variables de sistema: El SPSS crea variables sistemas que comienzan con el signo **$: $DATE, $CASENUM**, **$WEIGHT**. Tales variables no se pueden nombrar en el **DATA LIST** ni se pueden redefinir con COUNT, COMPUTE o IF.
- Formato de las variables: Los valores se almacenan internamente y se presentan o imprimen de acuerdo a un formato específico. La especificación de formato no afecta la precisión de los datos almacenados en memoria.
- Tipos de variables: Las variables pueden ser de dos tipos: numéricas o de cadenas. Por defecto, el **DATA LIST** supone que son numéricas. Para especificar una variable de cadena hay que hacerlo explícitamente.

Expresiones de transformación.

Las expresiones de transformación se usan en los comandos **COMPUTE, IF, PROCESS IF y SELECT IF.** En ellas intervienen los operadores **aritméticos (+,-,*,/ y **),** constantes y funciones, con la sintaxis aritmética usual.

Funciones.

Las funciones se utilizan en las expresiones de los comandos **COMPUTE, IF y SELECT IF**. Todas las funciones se pueden usar con variables numéricas. Solo la función LAG se puede utilizar con variables de cadena corta. El argumento de las funciones puede ser en general una expresión. Las principales funciones son:

Funciones numéricas:

- ABS(arg) Valor absoluto del argumento
- RND(arg) Parte entera del argumento: RND(-4.7)=-5
- TRUNC(arg) Trunca a parte entera: TRUNC(-4.7)=-4
- MOD10(arg) Resto del argumento dividido por 10
- SQRT(arg) Raiz cuadrada del argumento
- EXP(arg) Exponencial del argumento
- LG10(arg) Logaritmo base 10
- LN(arg) Logaritmo natural o neperiano (base e)
- SIN(arg) Seno del argumento (éste en radianes)
- COS(arg) Coseno del argumento (éste en radianes)

Funciones de valores perdidos

- **VALUE(arg)** Incluye todos los valores excepto el omitido por el sistema. El argumento tiene que ser un solo nombre de variable. La función VALUE se usa con un operador relacional en las expresiones lógicas de los comandos **IF y SELECT IF.**
- **MISSING(arg) Retorna 1** (verdadero) si el valor es perdido y 0 (falso) en caso contrario. El argumento es una variable.
- **SYSMIS(arg)** Retorna 1 (verdadero) si el valor es perdido u omitido por el sistema y 0 (falso) en caso contrario. El argumento es una variable.

Función cross-case

- **LAG(arg)** El valor de la variable un caso u observación antes. El argumento es una variable.

Funciones de números aleatorios

- **UNIFORM (arg)** Genera un número pseudo-aleatorio entre 0 y el valor del argumento. El argumento tiene que ser un número real.
- **NORMAL (arg)** Genera un número pseudo-aleatorio norma con media 0 y desviación estándar igual al argumento. Este tiene que ser un número real positivo.

Función de fecha

- **YRMODA(yy,mm,dd)** Convierte año, mes y día a un número de día, que es el número de días desde el 15 de octubre de 1582 (primer día del calendario Gregoriano), para realizar operaciones con fechas.

Expresiones lógicas

Una expresión lógica es una expresión que se puede evaluar como verdadera, falsa u omitida (perdida) de acuerdo a condiciones que cumplen los datos. Se utilizan como condicionales en la definición de variables o la selección de casos en los comandos

IF, PROCESS IF y SELECT IF. Las expresiones lógicas pueden ser simples relaciones lógicas entre variables con operadores relacionales: **EQ o =; NE o <>; LT o <; GT o >; LE o <=; GE o >=; o** pueden ser expresiones compuestas con los operadores lógicos AND, OR o NOT, con las reglas del algebra booleana.

2.4. Comandos de Operación

Los comandos de operación no afectan los datos que se procesan sino que fijan ciertos parámetros de operación del SPSS/PC. Puede usarlos para obtener información del estado de esos parámetros en un momento dado o para cambiar cualquiera de ellos. Permiten en general:

a) Seleccionar opciones de operación del SPSS/PC.

b) Obtener ayuda durante la ejecución del SPSS.

c) Ejecutar comandos del SPSS/PC desde un fichero externo.

d) Finalizar la sesión de trabajo con el SPSS/PC.

Selección de opciones de operación del SPSS/PC

El SPSS/PC utiliza un solo comando para definir sus parámetros de operación: destino de los listados, formato de salida, etc. Este es el comando SET. La sintaxis del comando SET incluye varios subcomandos. Escrita en forma simplificada con los subcomandos más usuales es:

 # & NARROW * $

Así, el subcomando DISK controla el envío de la salida hacia un fichero en disco (por defecto el SPSS.LIS). El subcomando SCREEN controla el envío de la salida hacia pantalla (por defecto en ON).

El comando **PRINTER** determina por su parte si la salida se envía o no hacia impresora (por defecto en OFF). Es un buen hábito enviar la salida sólo hacia fichero en disco. La ejecución será así mucho más rápida. Posteriomente se podrá siempre examinar los resultados por pantalla y la impresión se diferirá hasta que estas salidas sean definitivas.

Los subcomandos **LENGHT y WIDTH** controlan respectivamente la longitud y el ancho de la página, mientras que **EJECT** determina si se hace o no un cambio duro de página en el fichero de salida al terminar el procesamiento de cada comando. Se recomienda mantener el valor por defecto de **PRINTER en OFF** porque de lo contrario la impresión de la salida requeriría mucho papel.

El subcomando **RESULTS** permite definir el nombre del fichero de resultados especiales (ver definición explícita de este). **ECHO** controla la visualización o no del enunciado de los comandos en la salida. La salida puede abreviarse si ECHO está en OFF; pero ECHO en ON ayuda a organizar las respuestas.

El comando **SEED** permite fijar la semilla para la generación de números aleatorios. Por defecto, esta semilla se fija en forma pseudo-aleatoria a partir del reloj de la máquina; pero si así se hace, el mismo programa puede proporcionar resultados diferentes cuando se

ejecuta en dos momentos distintos. Si se desea evitar esto se puede fijar la semilla con el comando **SET SEED.**

Hay otros subcomandos del comando SET que permiten cambiar los caracteres usados para imprimir las tablas o los histogramas, para cambiar el nombre del fichero de ocurrencias, para cambiar el indicador de comandos (SPSS/PC:) o el finalizador del mismo (.) y muchos otros. Ellos pueden consultarse en la ayuda.

Normalmente un comando SET como el siguiente, puede ser suficiente para la mayoría de los programas prácticos:

SET

DISK=ON /

DISK='salida.lst' /

EJECT=OFF /

SCREEN=OFF /

ECHO=OFF .

El nombre del fichero en disco 'salida.lst' se variará en cada programa. Normalmente este comando encabeza un programa en **SPSS/PC**; pero puede incluirse otros comandos **SET** en un mismo programa (por ejemplo para activar o desactivar temporalmente el **ECHO)**

Comandos de obtención de ayuda

Estos comandos se utilizan siempre en modo interactivo. El comando **HELP** es propio de las primeras versiones del **SPSS/PC** y después quedó practicamente sustituído por el **REVIEW** como sistema de Menús-Ayuda-Edición. Otros comandos de ayuda son el **SHOW** y el **DISPLAY.**

SHOW es un comando que resulta en una tabla por pantalla con todos los parámetros del comando SET. El lista cada parámetro del **SET** activo, incluyendo los que no han sido alterados y conservan el valor por omisión. Se especifica con solo el nombre del comando sin parámetros ni subcomandos.

DISPLAY es un comando que puede ejecutarse una vez que haya un fichero activo. Visualiza información acerca de las variables solicitadas. La sintaxis es:

DISPLAY |{ }|

El comando **DISPLAY** envía la información hacia las salidas especificadas por el comando **SET (pantalla, disco o impresora).**

Ejecución de comandos de un fichero externo

El comando **INCLUDE** permite la ejecución de un fichero de comandos escrito como un fichero texto externo. La sintaxis incluye sólo como parámetro el nombre del fichero de comandos a ejecutar:

INCLUDE 'fichero'.

El nombre del fichero puede incluir nombre de torre y camino precediendo al fichero propiamente dicho. Se pueden anidar ficheros INCLUDE en varios niveles (normalmente hasta 5 niveles).

Finalización de la sesión de trabajo con el SPSS

El comando **FINISH** permite terminar la sesión de trabajo con el **SPSS/PC** y regresar al prompt del **DOS**. Su sintaxis no tiene especificaciones. Atención que para usar varios ficheros de comandos sucesivamente o en forma anidada, ninguno de ellos puede contener el comando FINISH.

2.5. Comandos de definición de datos

Los comandos de definición de datos describen, etiquetan y manipulan datos. Es posible con ellos:

- Declarar las variables
- Definir los datos
- Etiquetar las variables y valores de variables
- Titular y formatear las salidas
- Identificar valores como omitidos
- Calcular nuevas variables
- Cambiar los valores de variables existentes
- Seleccionar observaciones en forma temporal para usarlas en un análisis o en forma permanente para todos los análisis sucesivos.

Comando DATA LIST

El comando **DATA LIST** es el que se utilizar para definir por primera vez un conjunto de datos a analizar. Presenta tres variantes de acuerdo al formato de datos de entrada: **FIXED**, **FREE o MATRIX** y estas palabras representan subcomandos.

El **DATA LIST FIXED** se utiliza para datos con un formato fijo, esto es que comienzan y terminan en columnas fijadas y que deben ser cuidadosamente especificadas como en FORTRAN. No utilizaremos normalmente este formato.

El **DATA LIST FREE** se utiliza para datos con un formato libre. En este caso los datos se leen apenas en el orden en que son declaradas las variables, independientemente de su posición en el fichero. Este formato es mucho más flexible y cómodo; pero nótese que debemos asegurarnos tener una matriz exactamente rectangular de datos, sin omisiones, para que no se produzcan corrimientos.

En los casos en que un dato no sea disponible, su lugar tiene que ser llenado por un valor convencional **(por ejemplo: 0, -1 o 9999)** de manera que pueda ser identificado posteriormente como un valor perdido.

El **DATA LIST MATRIX** es una forma especial de entrada de datos para algunos comandos particulares del **SPSS/PC** y será analizada con éstos.

Nos quedamos de esta forma por el momento con el DATA LIST FREE.

Un subcomando opcional que generalmente seguirá al subcomando FREE es la definición de un fichero de datos externo, cuando este existe. Después de un "/" se anotan los nombres de las variables como en el siguiente ejemplo.

```
DATA LIST FREE FILE='ALUMNOS.DAT' /
   GRUPO
   INICIALE (A3)
   EDAD
```

SEXO

CALIF1 TO CALIF10 .

Aquí 'ALUMNOS.DAT' es el fichero externo que contiene los datos.

De acuerdo a la definición de las variables, este fichero tiene varias columnas: una primera que significa el número del grupo; la segunda que es una variable cadena de 3 caracteres y que brinda las iniciales de cada alumno; la tercera es la edad; la cuarta es un número que identifica el sexo y a continuación 10 variables que representan las calificaciones en el mismo número de asignaturas.

En general, el fichero **ALUMNOS.DAT** debe tener una apariencia rectangular como ésta:

1 ANE 27 1 3 4 3 4 5 4 2 3 4 2

1 DFE 28 2 2 3 2 4 3 3 4 3 3 2

1 PLE 32 1 3 2 4 3 3 4 3 3 2 3

2 APE 25 1 3 2 3 2 4 3 2 4 3 3

2 PKI 19 1 3 4 -1 4 3 5 5 5 4 3

y disponer tantas filas como casos (aquí alumnos) vayan a ser procesados. Observé como el **DATA LIST** asume las variables numéricas por omisión y como se especifica una variable de cadena: en general escribiendo entre paréntesis **(An)** donde n es la longitud máxima de la cadena. Observe además el uso del **TO** para declarar un arreglo.

Nótese por último que ninguno de los nombres de variables puede exceder 8 caracteres.

Comando BEGIN DATA-END DATA

Como alternativa del trabajo con un fichero externo, el comando **DATA LIST FREE** puede hacer referencia a un conjunto de datos incluídos en el propio programa entre las palabras claves **BEGIN DATA y END DATA**, después del **DATA LIST** y en la siguiente forma:

DATA LIST FREE /

GRUPO

INICIALE (A3)

EDAD

SEXO

CALIF1 TO CALIF10 .

BEGIN DATA

1 ANE 27 1 3 4 3 4 5 4 2 3 4 2

1 DFE 28 2 2 3 2 4 3 3 4 3 3 2

1 PLE 32 1 3 2 4 3 3 4 3 3 2 3

2 APE 25 1 3 2 3 2 4 3 2 4 3 3

2 PKI 19 1 3 4 -1 4 3 5 5 5 4 3

(Aquí continúan los datos hasta completar la matriz)

END DATA.

A menos que el juego de datos sea suficientemente pequeño utilizaremos siempre la alternativa de un fichero externo.

BEGIN DATA y END DATA se usan obligatoriamente al especificar datos matriciales (DATA **LIST MATRIX**) como veremos posteriormente.

Comando VARIABLE LABELS

El comando **VARIABLE LABELS** permite especificar etiquetas para las variables. Estas etiquetas acompañarán al nombre de las variables en la mayoría de las salidas del SPSS/PC y sirven por tanto para esclarecer estas. La sintaxis se ilustra claramente en el siguiente ejemplo:

```
VARIABLE LABELS
   GRUPO    'Grupo del estudiante'
   INICIALE 'Iniciales del nombre y apellidos'
   CALIF1   'Calificación en A. Matemático I'
   CALIF2   'Calificación en Algebra I'
   CALIF3   'Calificación en Lógica'
   CALIF4   'Calificación en Programación'
   CALIF5   'Calificación en Inglés I'
   CALIF6   'Calificación en A. Matemático II'
   CALIF7   'Calificación en Algebra II'
   CALIF8   'Calificación en Matemática Discreta'
   CALIF9   'Laboratorio de Programación'
   CALIF10  'Calificación en Inglés II' .
```

Las etiquetas de variable son cadenas constantes y se anotan entrecomillas o apóstrofes. Ellas no se pueden extender más de una línea ni tener más de 40 caracteres. Todas las variables tienen que haber estado definidas anteriormente, en este ejemplo, por el **DATA LIST** anterior. No es obligatorio que todas las variables tengan etiquetas (por ejemplo EDAD y SEXO no la tienen por ser obvias). Se permiten varios comandos **VARIABLE LABEL** dentro del mismo programa o sesión de trabajo con el SPSS/PC.

Comando VALUE LABELS

El comando **VALUE LABELS** define etiquetas descriptivas para valores individuales de una variable. Tiene sentido para variables discretas con nivel de medición nominal u ordinal. La sintaxis se ilustra claramente en el siguiente ejemplo:

```
VALUE LABEL
   GRUPO
      1 'Grupo A-1'
      2 'Grupo B-2'
      3 'Grupo C-1' /
   SEXO
      1 'Femenino'
      2 'Masculino' /
   CALIF1 TO CALIF10
```

2 'Mal'

3 'Regular'

4 'Bien'

5 'Excelente'.

Observe que para cada variable se definen las etiquetas de sus valores posibles. Variables como la EDAD (prácticamente continua no tienen etiquetas de valores). Las variables tienen que estar anteriormente definidas, por ejemplo con el **DATA LIST** anterior. Las etiquetas de cada valor de una variable son cadenas y por tanto se encierran entre comillas. Las variables se separan con **slash** en la especificación de este comando.

Las mismas etiquetas se pueden asignar a los valores de varias variables similares (por ejemplo **CALIF1 a CALIF10)**

Por razones prácticas, se recomienda mucho utilizar cadenas cortas en las etiquetas de valores de las variables. Hay comandos que sólo trabajarán posteriormente con los primeros caracteres de la etiqueta de valores.

Comando FORMAT

FORMAT es un comando que permite definir el formato de las variables. Aunque se puede utilizar para definir formatos con comas o precedidos del signo dólar, lo más importante es su uso para definir el número de cifras totales y decimales de cada variable. Ello se logra con la sintaxis

FORMAT

Lista de variables (Fn.m) /

Lista de variables...

El formato (Fn,m) especifica que la variable debe ser entendida con ancho total n (incluyendo una posición para el signo y una para el punto si no es entera) y m cifras decimales. Se sobreentiende que n es mayor o igual que m+2. Por ejemplo:

FORMAT

GRUPO SEXO CALIF1 TO CALIF10 (F2.0) /

EDAD (F3.0) /

PROMEDIO (F4.1) .

Todas las variables deben estar definidas previamente. En este ejemplo, se supone que después de haber declarado GRUPO, SEXO,

EDAD CALIF1 TO CALIF10, la variable PROMEDIO, fué calculada como media de las 10 calificaciones y tiene una cifra entera y una decimal (el punto y el signo completan el ancho de 4).

Observe que **FORMAT** no se puede utilizar para definir formatos en variables cadenas como **INICIALE.**

Comando MISSING VALUES

El comando MISSING VALUES sirve para definir valores omitidos (perdidos) por el usuario. Estos son los valores que fueron incluidos en el fichero convencionalmente para "llenar

datos faltantes" y respetar la estructura rectangular. El SPSS/PC realizará determinado tratamiento con este tipo de valor según se especifique en cada comando procedimiento.

La sintaxis se explica claramente en el siguiente ejemplo:

```
MISSING VALUES
   EDAD (0) /
   CALIF1 TO CALIF10 (-1) .
```

Otros comandos de definición de datos

Como alternativas del **DATA LIST**, los datos pueden ser inicialmente definidos con el comando **GET** o el comando **IMPORT** si existen previamente definidos un fichero sistema o un fichero portable.

Entre ellos, el más importante por el momento es GET.

Su uso se ilustra en el siguiente ejemplo:

```
GET FILE='ALUMNOS.SYS' /
   DROP=INICIALE PROMEDIO.
```

El fichero **ALUMNOS.**SYS tuvo que ser previamente creado con el comando **SAVE FILE**. El contiene los datos y el diccionario de los datos. El subcomando opcional DROP deja de cargar una variable o listado de variables que se especifiquen. Aunque ellas estaban en el fichero sistema, no forman parte del fichero activo.

Otros comandos de transformación de datos

Otros comandos que permiten transformaciones de datos son **RECODE, COUNT, COMPUTE** e IF. Veamos las formas de uso más importantes de cada uno y sus sintaxis.

RECODE permite recodificar una variable aleatoria. Típicamente, se utiliza para cambiar los códigos de una variable aleatoria discreta o para ranquear una variable aleatoria continua en forma discreta. Ambos usos se indican en los siguientes ejemplos.

```
RECODE SEXO (1=2)(2=1) /
    CALIF1 TO CALIF10 (3,4,5=3) .
```

En el proceso de recodificación el valor que precede al signo de igualdad es el valor anterior y el valor que sigue es el nuevo valor. En el ejemplo anterior, el comando RECODE intercambia los valores 1 y 2 de la variable SEXO. Para las variables CALIF1 a

CALIF10 se reducen a tres sus categorías: lo que originalmente era 3, 4 o 5 pasa simplemente a ser 3.

```
RECODE
  EDAD (LO THRU 15=1)(15 THRU 20=2)(20 THRU 30=3)(3O THRU HI=4).
```

En este comando se ranquea la edad en las siguientes categorías:

1 Menores o iguales a 15

2 Más de 15 y menores o iguales a 20

3 Más de 20 y menores o iguales a 30

4 Más de 30.

Observe el uso de la palabra reservada TRHU para delimitar los rangos. Observe el uso de las palabras reservadas **LO (LOWEST) Y HI (HIGHEST)** para hablar de los valores más bajos y más altos posibles respectivamente sin tener que precisarlos.

Un detalle importante es que el orden del ranqueo al escribir el comando RECODE es significativo. Si un valor terminal de rangos se repite, el no se vuelve a alterar. Recibe el rango del primer ranqueo. Por eso, en el ejemplo, la categoría 1 abarca el 15, la categoría 2 no abarca el 15 y abarca el 20, etc. Si se hubiese requerido un ranqueo como el siguiente:

1 Menores de 15

2 Entre 15 y 30

3 Mayores de 30

lo correcto hubiera sido escribir el RECODE así:

RECODE EDAD (15 THRU 30=2)(LO THRU 15=1)(30 THRU HI=3) .

Como EDAD es un valor entero, este comando produce idéntico resultado al siguiente

RECODE EDAD (LO THRU 14=1)(15 THRU 30=2)(31 THRU HI=3) .

Efectivamente, este último parece más claro; pero es una solución menos general basada en qué EDAD toma solo valores enteros.

Imagínese que la edad se hubiese medido en años y fracciones, o en lugar de la EDAD hubiesemos tenido una variable realmente continúa. El orden de escritura de la recodificación en el comando

RECODE, es en general, la solución más simple.

Se puede utilizar en el RECODE la palabra reservada ELSE para identificar los restantes valores no especificados antes. Por ejemplo:

RECODE EDAD (LO THRU 15=1)(15 THRU 20=2)(20 THRU 30=3)(ELSE=4).

RECODE se puede utilizar también para variables cadenas cortas. Los valores originales y recodificados deben escribirseentre apóstrofes o comillas. Se puede utilizar también para recodificar valores perdidos (MISSING), en valores válidos:

RECODE CALIF1 TO CALIF10 (MISSING=2).

COMPUTE es un comando que permite la definición de una nueva variable para cada caso a partir de una expresión. Equivale a una instrucción de asignación en la forma:

COMPUTE variable=expresión.

La variable de la izquierda del signo igual recibe o modifica su valor anterior para cada caso de acuerdo con la expresión. En la expresión pueden intervenir variables, constantes, operadores aritméticos y funciones. Ejemplo:

COMPUTE PROMEDIO=(CALIF1+CALIF2+CALIF3)/3 .

Otro ejemplo:

COMPUTE EVOLAMAT=ABS(CALIF1-CALIF6).

El comando IF define una instrucción de asignación con una condicional. La sintaxis es:

IF (condición) variable=expresión.

La condición es una expresión lógica simple o compuesta. La instrucción de asignación: variable=expresión es similar a la del **COMPUTE.**

Ejemplos:

IF (SEXO=1) Y=10+X .

IF (SEXO=2) Y=10+SIN(X) .

es una pareja de comandos que define el valor de Y en función de X en dependencia del valor de la variable SEXO.

Otro ejemplo

COMPUTE VSELEC=0 .

IF ((SEXO=1) AND (PROMEDIO>4)) VSELEC=1 .

Esto es una pareja de comandos que define la variable VSELEC de la siguiente forma. En principio **VSELEC** vale la constante 0 para todos los casos. Después del IF, la variable **VSELEC** vale la constante 1 para las personas de SEXO igual 1 y promedio mayor que

4. Esta puede ser una forma de "marcar" las "mujeres de alto índice académico" para realizar después algún procesamiento particular con ellas.

Comandos de selección de casos y peso.

Estos comandos permiten seleccionar temporalmente o definitivamente casos o atribuirles peso diferente a algunos casos para ampliar o reducir su valor relativo en ciertos procesamientos.

PROCESS IF (condición simple)

Es la sintaxis del comando selector de casos **PROCESS IF**. Este hace que el siguiente comando procedimiento se ejecute solo para los casos que satisfacen la condición. Ejemplo:

PROCESS IF (PROMEDIO > 4) .

LIST VARIABLES=INICIALE.

o

PROCESS IF (SELEC=1) .

LIST VARIABLES=INICIALE.

Son parejas de comandos que hacen listados selectivos. En el primer caso, por la condición de ser el promedio superior a 4. En el segundo caso por la condición de ser la variable SELEC igual a:

1. Recuérdese que la definición de SELEC fue ejemplificada. Esta es la única forma de incluir en un PROCESS IF una condición compuesta como **(SEXO=1)AND(PROMEDIO>4).**

SELECT IF es un comando bastante similar a PROCESS IF. Su sintaxis:

SELECT IF (condicion) especifica que en este comando la condición puede ser directamente simple o compuesta. Semánticamente, y lo más importante es que **SELECT IF** hace una selección permanente de casos para toda la sesión de trabajo (mientras que PROCESS IF solo actuaba sobre el comando procedimiento siguiente).

Ejemplo:

SELECT IF (SEXO=1) . Selecciona sólo las mujeres para todos los procesamientos siguientes. Después de un tal comando no se podría dar otro como

SELECT IF (SEXO=2) .porque el fichero activo se quedaría sin casos.

Otro comandos de selección de casos son **N y SAMPLE** . Ejemplos de su uso son:

N 50 .selecciona los primeros 50 casos del fichero de datos. Este comando se utiliza frecuentemente en la "fase de puesta a punto" de un programa en SPSS para no demorarnos con la lectura de un **gran fichero de datos**.

Se utiliza también:

N 1200 ESTIMATED para dar al SPSS/PC un número estimado de casos y pueda optimizar el espacio reservado a éstos en memoria. Esto es importante cuando esta memoria es escasa para ciertos procedimientos.

SAMPLE .25 selecciona una muestra aleatoria de aproximadamente 1/4 del volumen total del fichero de datos para la ejecución.

SAMPLE 500 FROM 3420 selecciona una muestra aleatoria de 500 casos cuando el fichero activo tiene 3420 casos o más.

En ambos comandos, la semilla de los números aleatorios se busca a partir del valor del SEED en el **SET.WEIGHT** es un comando que permite pesar o sopesar casos. Si por ejemplo queremos que todos los artículos de sexo igual 1 sean considerados "dobles" en la muestra podemos hacer lo siguiente.

COMPUTE PESO=1.

IF (SEXO=1) PESO=2.

WEIGHT BY PESO.

La transformación de ponderación permanece activa hasta aparecer el comando WEIGHT OFF o un nuevo comando de ponderación por otra variable.

Téngase presente que si la cantidad de observaciones ponderadas excede el tamaño general de la muestra, las pruebas de significación resultan infladas.

Comandos para formateo de las salidas del SPSS.

Además de **VARIABLE LABEL y VALUE LABEL**, sirven para formatear las salidas los comandos **TITLE, SUBTITLE** y *.

TITLE permite definir un título de hasta 58 caracteres que aparecerá en el encabezamiento de todo el fichero de salidas.

Ejemplo:

TITLE='INVESTIGACION CON ESTUDIANTES DE PRIMER AÑO' .

SUBTIITLE inserta un subtítulo en la segunda línea de cada página del fichero de salidas. Es independiente de **TITLE** y se puede utilizar repetidamente para separar partes de un programa. Cada subtitulo sustituye el anterior y surte efecto a partir de la próxima página del listado.

El comando "*" facilita insertar textos aclaratorios dentro de un fichero de comandos del SPSS/PC, como un comentario del programador. Puede tener varias líneas. Se considera que finaliza en el punto. ES por eso importante que un comentario largo no tenga puntos intermedios**. Ejemplo**:

PROCESS IF (VSELEC=0) .IF ((SEXO=1)AND(PROMEDIO>4)) VSELEC=1 .

* Los comandos anteriores seleccionan las mujeres con índice académico alto; para los casos con esta condición la variable **VSELEC resultará 1.**

En el grupo de tres comandos anteriores, los últimos tres renglones precedidos de **"*"** y finalizados con un **"."** son un comentario del programador.

Cuando una primera parte de cierto programa está ya probada y se puede y quiere pasar directamente a probar otra parte, los comandos de la primera parte pueden hacerse preceder provisionalmente de un "*" (un asterisco en cada comando). Durante la ejecución del programa ellos se "saltarán" como comentarios.

Una vez probada la segunda parte, bastará eliminar los asteriscos provisionales para obtener y ejecutar el programa definitivo.

2.6. Comandos de procedimientos. Estructura típica de un programa en SPSS/PC+ para DOS

Los comandos de procedimientos son aquellos que permiten en sí la presentación y el análisis estadístico de los datos. Ellos son los que van a ser objeto de estudio principal en la siguiente conferencia.

Las generalidades y los comandos de operación y definición de datos pueden ayudarnos a pensar en una estructura más o menos **"típica"** de un programa en SPSS, como la siguiente:

```
SET
   DISK=ON /
   DISK='diskfile.lst' /
   EJECT=OFF /
   SCREEN=OFF /
   ECHO=OFF .
DATA LIST FREE FILE='datfile.dat' /
   VAR1
  ...
   (Declaración de variables; no olvidar el punto final) .
VARIABLE LABEL
   VAR1 'etiqueta de la variable'
   ...
VALUE LABEL
   VAR1
      1 'etiqueta1'
      2 'etiqueta2'
      ...
      n 'etiquetan'  /
   VAR2
      ...
FORMAT
   VAR1 (Fn.m) /
   VAR2 ...
```

MISSING VALUES

VAR1(value) /

VAR2 ...

TITLE='aquí va el título del trabajo'.

SUBTITLE='aquí va el primer subtitulo'.

* Aquí van comandos de transformación de datos y/o comandos procedimientos.

SUBTITLE='aquí va el segundo subtítulo'.

* Aquí van comandos de transformación de datos y comandos procedi- mientos asociados a la segunda parte o "subtítulo", y así sucesivamente .

FINISH.

Un ejemplo simple y completo podría ser el siguiente:

SET

DISK=ON /

DISK='ALUMNOS.LST' /

EJECT=OFF /

SCREEN=OFF /

ECHO=OFF .

DATA LIST FREE FILE='ALUMNOS.DAT' /

GRUPO

INICIALE (A3)

EDAD

SEXO

CALIF1 TO CALIF10 .

VARIABLE LABELS

GRUPO 'Grupo del estudiante'

INICIALE 'Iniciales del nombre y apellidos'

CALIF1 'Calificación en A. Matemático I'

CALIF2 'Calificación en Algebra I'

CALIF3 'Calificación en Lógica'

CALIF4 'Calificación en Programación'

CALIF5 'Calificación en Inglés I'

CALIF6 'Calificación en A. Matemático II'

CALIF7 'Calificación en Algebra II'

CALIF8 'Calificación en Matemática Discreta'

CALIF9 'Laboratorio de Programación'

CALIF10 'Calificación en Inglés II' .

VALUE LABEL

GRUPO

1 'Grupo A-1'

2 'Grupo B-2'

3 'Grupo C-1' /

SEXO

1 'Femenino'

2 'Masculino' /

CALIF1 TO CALIF10

2 'Mal'

3 'Regular'

4 'Bien'

5 ‘Excelente’.

FORMAT

GRUPO SEXO CALIF1 TO CALIF10 (F2.0) /

EDAD (F3.0) .

MISSING VALUES

EDAD (0) /

CALIF1 TO CALIF10 (-1) .

TITLE='INVESTIGACION DE ALUMNOS DE PRIMER AÑO’.

SUBTITLE='CALCULO DE INDICE ACADEMICO’.

COMPUTE PROM1= (CALIF1+CALIF2+CALIF3+CALIF4+CALIF5)/ 5.

COMPUTE PROM2= (CALIF6+CALIF7+CALIF8+CALIF9+CALIF10)/ 5.

VARIABLE LABEL

PROM1 'PROMEDIO DEL 1ER SEMESTRE'

PROM2 'PROMEDIO DEL 2DO SEMESTRE’.

FORMAT

PROM1 PROM2 (F4.1) .

DESCRIPTIVES

PROM1 PROM2.

* El comando procedimiento DESCRIPTIVES calcula las estadísticas descriptivas de las variables PROM1 y PROM2; será objeto de estudio en la próxima conferencia .

SUBTITLE='SELECCION DE MUJERES DE ALTO RENDIMIENTO’.

COMPUTE VSELEC=0.

IF ((SEXO=1)AND(PROM1>4)AND(PROM2>4)) VSELEC=1.

PROCESS IF (VSELEC=1) .

LIST VARIABLES=GRUPO INICIALE.

* El comando procedimiento **LIST** lista los valores de las variables especificadas para todos los casos; puede estudiarse por la ayuda.

FINISH.

Cuando con un mismo juego de datos se espera hacer muchos procesamientos estadísticos, es recomendable dividirlos en **"módulos" o "programas independientes"** para facilitar su puesta a punto. En tal caso es recomendable además hacer un primer programa que defina y conforme los datos y que finalmente los grabe como un fichero sistema con un **SAVE FILE**. Los próximos programas no tienen que comenzar de nuevo con la definición de los datos. Estos se cargan con su diccionario con el comando **GET FILE.**

2.7 Conclusiones

Como **Sistema de Análisis** de Información el **SPSS/PC** permite entrar datos interactivamente o desde un fichero, recodificar variables y calcular nuevas variables, salvar datos hacia ficheros recuperables, transferir ficheros entre micros y mainframes, tabular variables y combinaciones de variables, efectuar una serie de análisis estadísticos, producir reportes formateados, con o sin selección de casos, etc.; pero el **SPSS/PC** es sobre todo, un paquete de procesamiento estadístico y sus méritos fundamentales radican en la amplitud y el rigor de las pruebas estadísticas que contiene.

Entre las pruebas estadísticas posibles a realizar con el **SPSS/PC**, se destacan el análisis descriptivo de datos, las tablas de contingencia, los tests paramétricos y no paramétricos de comparación de dos poblaciones, los análisis de varianza paramétricos y no paramétricos, los análisis de varianza paramétricos multifactorial y multivariado, el análisis de regresión simple y múltiple, el análisis discriminante, el análisis de factor, la formación de **clusters,y** el estudio de series de tiempo.

El **SPSS/PC** puede trabajarse en forma interactiva con ayuda en línea, y esta forma es recomendada si se dispone sin restricción de una máquina con el paquete; pero su utilización puede ser también optimizada utilizando procedimientos en lotes de comandos, escribiendo "programas" con una estructura general bastante uniforme y descrita en este capítulo.

En cualquier caso es importante conocer la sintaxis general y las posibilidades de los comandos de operación y de definición de datos y además las potencialidades de los comandos procedimientos que realizan los procesamientos en sí.

En este **Capítulo II** se han visto los principales comandos de los dos primeros tipos y en los capítulos siguientes se irán estudiando los comandos procedimientos asociados a los procesamientos estadísticos objeto de estudio.

2.8 Ejercicios del Capítulo II

1. Utilice los ficheros ALUMNOS.DAT y ALUMNOS.SPS.

Cópielos hacia el disco rígido y examínelos con un editor de textos. El contenido de estos ficheros debe estar claro para usted. "ALUMNOS.DAT" contiene los datos del ejemplo del capítulo y "ALUMNOS.SPS" es el programa final elaborado. Comience añadiendo algunos datos al fichero.

2. Ejecute en lotes el fichero programa con el SPSS/PC.

Una vez que termine analice el fichero de respuestas "ALUMNOS.LST". Asegúrese que puede interpretar todos los resultados.

3. Examine los ficheros SPSS.LOG y SPSS.LIS que crea el SPSS.

4. Intente ahora reproducir interactivamente cada uno de los comandos del fichero ALUMNOS.SPS. Auxíliese del sistema de Menús y Ayudas. Analice cada error posible.

5. Se desea hacer una investigación sobre la calidad de diferentes vinos caseros.

El objetivo del resto de la práctica es preparar un fichero de datos y un fichero programa para ser ejecutado con el SPSS. La mayor parte de las actividades siguientes puede hacerla en una máquina cualquiera con ayuda de un editor de textos. Acuda al SPSS/PC solo en los momentos de prueba.

a) Comience a preparar su fichero programa fijando las opciones de operación con el comando SET.

b) Utilice el comando DATA LIST par declarar las siguientes variables en formato libre: CALIDAD, VINO, PROVINCI, DISP, PRECIO, COSTO, AZUCAR, PH, SODIO, ALCOHOL, PREMIO y CLARIDAD. Todas las variables son numéricas excepto VINO que es alfabética con 20 caracteres. Todas las variables restantes aparecerán en el fichero externo de datos "VINOS.DAT" como numéricas enteras, excepto PRECIO y COSTO que deben ser leídas con dos cifras decimales, PH y ALCOHOL con una.

c) Introduzca las siguientes etiquetas para variables que necesitan ser más especificadas:

PROVINCI Provincia de origen

DISP Disponibilidad de venta

PRECIO Precio de un litro

COSTO Costo de cinco litros

AZUCAR Azucar (lbs) necesario para 20 litros

PH Nivel ácido resultante

SODIO Cantidad de sales (mg/litro)

PREMIO Reconocimiento alcanzado

CLARIDAD Claridad lograda

d) Introduzca las siguientes etiquetas de valores para las variables que son especificadas.

CALIDAD 1: Muy bueno 2: Bueno 3: Regular 4: Malo

PROVINCI 1:P.del Río 2:C.Habana..., etc.

DISP 1: Local 2: Provincial 3: Nacional 4: Exportación

PREMIO 0: Sin premio 1: Popular 2: Provincial 3: Nacional

CLARIDAD 0: Regular 1: Claro 2: Excelente

e) Declare los formatos de las variables de acuerdo con las notas introductorias. Declare como valor perdido el valor 0 para PREMIO. Deténgase un momento para precisar el tipo de cada variable numérica (continua o discreta y dentro de estas el nivel de medición ordinal).

f) Introduzca un comando procedimiento que liste todas las variables para comprobar.

g) Introduzca un comando para grabar los datos y el directorio como un fichero sistema: "VINOS.SYS" .

h) Introduzca el comando FINISH y grabe el fichero de comandos como VINOS0.SPS.

6. Prepare un fichero de datos 'VINOS.DAT' que se ajuste a las especificaciones anteriores y ejecute con él su programa. Analice el fichero de respuestas.

7. Inicie la preparación de un nuevo fichero de comandos "VINOS1.SPS" con las siguientes características

a) Deberá arrojar los resultados hacia "VINOS1.LST" y después de fijados los parámetros de operación, realizará como primera tarea la carga del fichero sistema "VINOS.SYS"

b) Recodifique los valores de la variable CALIDAD en la forma siguiente:

1: Muy Bueno en 1: Especial

2 y 3: Bueno y Regular en 2: Normal

4: Malo en 3: Malo

Haga los cambios de etiquetas consecuentes.

c) Establezca rangos para la variable que define el % de alcohol en el vino, de manera que quede:

1: Bajo para valores menores de 7.5

2: Aceptable para valores entre 7.5 y 12

3: Alto para valores superiores a 12

d) Haga uso del comando DESCRIPTIVES para obtener los estadísticos descriptivos de todas aquellas variables que sean continuas.

e) Intente hacer uso del comando **FREQUENCIES** para obtener la distribución de aquellas variables que son discretas. La sintaxis más simple de FREQUENCIES es:

FREQUENCIES lista de variables.

f) Escriba los comandos pertinentes que le permitan seleccionar los vinos que tienen CALIDAD Normal o Especial, PREMIO al menos provincial y grados de alcohol Bajo o Aceptable. Se pretende en primer lugar listar estos vinos y hacer las operaciones g) y h).

g) Para los vinos que cumplen las condiciones anteriores se desea obtener el COSTO y el PRECIO promedio. Introduzca los comandos pertinentes.

h) Se desea obtener también el COSTO y el PRECIO promedio para los vinos que no cumplen las condiciones anteriores. Intro- duzca los comandos pertinentes.

i) Ponga títulos y subtítulos a su programa. Introduzca el comando FINISH.

8. Ejecute VINOS1.SPS, corrija los errores y analice los resultados. Conserve sus programas para la próxima actividad.

Capítulo III. Introducción a la inferencia estadística

Introducción

En este capítulo pretendemos dar una visión general de la inferencia estadística, más concretamente, de la teoría de la estimación y la teoría de las dócimas de hipótesis. Ellas constituyen el fundamento del análisis descriptivo de datos, también llamada estadística descriptiva. Estudiaremos además la implementación de ésta con el paquete SPSS.

La teoría de las dócimas de hipótesis es especialmente importante porque es también el fundamento de muchos de los métodos estadísticos de más frecuente uso y que serán objeto fundamental de estudio en el curso.

La inferencia estadística se ocupa de sacar conclusiones sobre las variables que caracterizan los fenómenos aleatorios a partir de los datos de una muestra. Comenzamos por tanto precisando el concepto de muestra.

3.1. Concepto de muestra

Supongamos que se quiere estudiar un conjunto de objetos homogéneos respecto a cierto índice cualitativo o cuantitativo que caracteriza esos objetos. Por ejemplo si se tiene un grupo de estudiantes, el índice cualitativo puede ser el sexo y el índice cuantitativo puede ser el promedio académico.

A veces se realiza una investigación total, es decir, se examina cada uno de los objetos del conjunto respecto al índice que interesa. En otros casos esto no es posible, por ejemplo, porque el conjunto es muy grande o infinito o por ejemplo, porque el estudio lleva implícito un conjunto considerable de gastos materiales o la destrucción del objeto de estudio. En estos casos se escogen fortuitamente del total un número limitado de objetos y se someten éstos al estudio. A este subconjunto se le llama conjunto muestral o simplemente muestra y se asocia con el conjunto general o población del cual se extrae.

Los conceptos de muestra y población son relativos. Por ejemplo, si se decide trabajar con todos los estudiantes de un grupo, puede considerarse todavía que esto es una muestra de un conjunto más general, digamos, los resultados posibles de un grupo abstracto de estudiantes con condiciones similares.

Para el análisis estadístico, las muestras son normalmente finitas y se llama volumen de la muestra al número de objetos del subconjunto muestral. Si estamos estudiando una variable aleatoria X, una muestra de volumen n consiste simplemente de cualquier subconjunto de n valores

x x x . . . x

de X.

Existen distintos métodos de selección de una muestra. A la selección de muestras se dedica de hecho, toda una rama de la estadística. Aquí solo veremos los conceptos básicos.

Se dice que una muestra ha sido seleccionada con reposición cuando al seleccionar cada objeto del conjunto quedan establecidas las condiciones para que el mismo objeto pueda ser seleccionado otra vez con la misma probabilidad (independientemente de los resultados anteriores).

Si se extraen cartas de una baraja, por ejemplo, esto significa que cada carta sea reintegrada y el juego sea barajeado antes de sacar la próxima.

Si la elección de un objeto implica que el mismo no pueda ser elegido otra vez, la muestra se dice seleccionada en forma única.

Si el volumen del conjunto general o población es suficientemente grande, y la muestra constituye solamente una parte ínfima del conjunto, la diferencia entre muestras con repetición y única tiende a eliminarse. En el caso límite, cuando se analiza un conjunto general infinito, esa diferencia desaparece.

Para que por los datos de una muestra se pueda juzgar con bastante certeza el índice que nos interesa de la población, es necesario que el objeto de la muestra lo represente correctamente, esto es que la muestra sea representativa. Intuitivamente hablando, la Ley de los Grandes Números nos permite garantizar que si la muestra se escoge aleatoriamente y tiene un volumen suficiente, ella será representativa. Posteriormente precisaremos el concepto de representatividad.

La determinación del volumen mínimo de una muestra, para realizar un análisis estadístico de cierta fiabilidad es uno de los objetivos de la Estadística Matemática y depende concretamente del análisis estadístico. Veremos en esta misma conferencia y a lo largo del curso, algunos ejemplos de criterios para determinar tal volumen.

Para la fundamentación matemática de los métodos estadísticos que estudiaremos necesitamos un concepto más abstracto y general de muestra de una variable aleatoria que es el siguiente:

p 1 ² n p

Desde este punto de vista, como muestra de volumen n, entendemos matemáticamente no una muestra particular, sino realmente todo el conjunto de muestras posibles de volumen n que se pueden seleccionar con X. La muestra es así, una variable aleatoria, realmente, un n-plus de variables aleatorias, cada una de las cuales tiene la misma distribución de la variable original. Un valor concreto de la muestra **(X ,X ,...,X),** es prácticamente hablando, una muestra concreta **(x ,x ,...,x).**

Esta definición nos permite hablar de las características de una muestra como variable aleatoria, por ejemplo, de la media:

a+

resumiendo en una sola variable aleatoria, todos los valores posibles de esa característica:

x = x /n

en cualquier muestra particular.

En concordancia con los conceptos del capítulo 1, si la variable aleatoria **X** es discreta y la muestra es extraída con reposición, o si la variable aleatoria X es continua, las componentes de la muestra **X i=1,2,...,n** serán independientes y la distribución conjunta de la muestra como **n-plus (X ,X ,...,X),** será la potencia **n-ésima** de la distribución de **X**.

Trabajaremos siempre con muestras independientes o muestras con reposición, por las razones que antes explicamos.

3.2. Estimación estadística de los parámetros de una distribución a partir de los datos de una muestra

Sea X una variable aleatoria con cierta distribución F(x) dependiendo de ciertos parámetros (los parámetros pueden ser la esperanza matemática, la dispersión, u otros).

Una de las técnicas de la Estadística de aplicación más frecuente tiene como objetivo determinar los parámetros de la distribución a partir del análisis de muestras. Esta rama de la Estadística se conoce como Teoría de la Estimación.

Supongamos por ejemplo, que queremos estimar la esperanza matemática de **X, M[X]=a,** que en principio es desconocida. La teoría de la estimación estadística argumenta que la media muestral constituye un "buen estimador" de la esperanza matemática. ¿En qué sentido?

Sea **(X ,X ,...,X)** una muestra de volumen **n de X**. Cada una de las

X tiene entonces la misma distribución **F(x)** y en particular

X

Entonces:

i=1 n a

De esta forma, el valor esperado de X como variable aleatoria, es precisamente la esperanza matemática buscada.

Enfatizamos que trabajamos con la media muestral como una **v.a.** Para cada muestra concreta: **(x ,x ,...,x)** la media adquiere un valor concreto x, y al reunir todas las posibles muestras, recorremos todos los valores posibles x de la variable X. Los valores posibles de la media muestral no coinciden con la esperanza matemática o valor esperado de X. De hecho, la esperanza matemática es un valor teórico constante y la media muestral tiene un carácter aleatorio. Lo que hemos demostrado es que el valor esperado de esta media muestral es la esperanza matemática. Observe además que este resultado es independiente del volumen de la muestra.

Formulemos en general algunos conceptos:

p p

Así, por ejemplo, el resultado anteriormente demostrado puede formularse diciendo que la media muestral es una estimativa no desviada de la esperanza matemática.

Veamos otro ejemplo. Supongamos que queremos buscar una estimativa de la varianza o dispersión de X:

D[X] = M (X - M[X]) = M X - M[x]

Parece natural considerar como estimativa "la dispersión muestral":

t i t i p t i p

²

M[_] = -----] D[X] < D[X]

En otras palabras, _ tiende a subestimar la dispersión de la población y en este sentido es una estimativa desviada o sesgada.

Se puede construir una estimativa no desviada de D[X] corrigiendo

(X - X)

² n ² n n-1

de manera que S es una estimativa no desviada de la dispersión **D[X].** Así, para estimar la dispersión de una variable aleatoria resulta mejor utilizar la estimativa S que la dispersión muestral. S se conoce como dispersión muestral corregida, y su raíz cuadrada:

s ²

se conoce como desviación estándar corregida.

La diferencia entre S y _ es despreciable si n es grande; pero es importante cuando el volumen de la muestra es pequeño.

Durante el curso veremos ejemplos de otros estimadores desviados de un cierto parámetro (por ejemplo, el coeficiente de correlación muestral de dos variables) es un estimador que tiende a sobrevalorar el coeficiente de correlación poblacional, esto es, una estimativa desviada por exceso y por tanto debe ser corregido.

Hay otros criterios para determinar cuál entre un conjunto de estimativas puede ser mejor. Por ejemplo, los valores posibles de una estimativa **D** de cierta parámetro **D** pueden ser más o menos dispersos alrededor de su valor esperado. Otro criterio muy utilizado es el del "estimador de máxima verosimilitud**" [3].**

Veamos aquí sólo algunos conceptos que permiten tener una idea más precisa de la eficiencia de una estimativa, basados en el conocimiento de la probabilidad g de que la estimativa D esté cercana a D en una magnitud dada, esto es, el valor

g = Prob (¦ D - D ¦ < _)

Este valor (g) se llama fiabilidad de la estimativa con un error _. Por supuesto, que para conocer g, es necesario estudiar con detalle la distribución de la estimativa. Veamos un ejemplo importante:

Supongamos que X tiene distribución normal con esperanza a conocida y utilizamos X como estimativa de a. Ya vimos en el capítulo 1 que la variable aleatoria:

t i

tiene también distribución normal con esperanza a y dispersión **_ /n. Entonces, para cada _ > 0,**

Este valor, puede ser buscado si se fijan _ y _. En particular, **Para _=[--], resulta g = Prob(¦X-a¦<[--]) = 0.9544**

de manera que la fiabilidad de la estimación de la esperanza por la media con un error menor o igual a 3 veces la raíz cuadrada de la dispersión (de la media) es superior al 99%. De manera general para

t_ t_ 2 ²

El concepto de fiabilidad da origen a otro tipo de estimación por los llamados intervalos de confianza. En lugar de estimar el parámetro **D** por un valor **D** , estimamos el parámetro D por un intervalo **(D ,D),** dentro del cual, y con cierta probabilidad, se encontrará D.

Si el intervalo se busca simétrico, se conforma a partir de la expresión:

g = Prob(¦ D - D ¦ < _) = Prob (D - _ < D < D + _)

y se leerá **"el intervalo (D -_,D +_) abarca el parámetro D con una confianza g".**

Así por ejemplo, para la esperanza matemática de una distribución normal, con dispersión conocida _ , buscamos un intervalo de confianza en la forma:

X - _ < a < X + _determinando el ancho _ necesario para una fiabilidad g.

Por ejemplo, para **g=0.9544, tenemos _ = 2_/rn,** por tanto, si la dispersión es conocida,

X - ---] , X + ---]

es un intervalo de confianza 95.44% de la esperanza matemática de una distribución normal.

Lo más importante es comprender la interpretación clara de este resultado. Aquí lo aleatorio (porque se saca de la muestra) es el intervalo. Por ello decimos, por ejemplo "con fiabilidad

g=95.44%, el intervalo X - [---], X + [---] abarca la esperanza matemática a".

Si tenemos la **v.a. X** con distribución normal y dispersión conocida _, tomamos una muestra y calculamos

& ²_ ²_ *

Obtenemos un intervalo concreto. No es correcto interpretar que "con el 95.44% de confianza la esperanza matemática está en ese intervalo". La fiabilidad significa que si hacemos el cálculo de un intervalo concreto como éste un número grande de veces, aproximadamente en el 95% de los casos ese intervalo abarcará la esperanza matemática.

Para una fiabilidad dada g, podemos dar el intervalo de confianza de la esperanza matemática a de una distribución normal cuya dispersión _ es conocida en la forma:

donde t se determina por la relación:

2 ²

Es muy importante saber que este intervalo de confianza es válido

si X tiene distribución normal y la dispersión es conocida. Si la dispersión fuera desconocida, parece natural que el intervalo de confianza tuviera la forma:

X - [---] , X + [---

que se calcularía utilizando la estimativa S de _ . Pero ahora no podríamos afirmar que:

& tS tS * 2 ²

_

porque no solamente X, sino S son aleatorios y por tanto t se determina de forma distinta.

Una forma de calcular esta probabilidad puede ser la siguiente:

X - a

g = Prob ¦T¦ < t = 2 | S (x) dx

y si se fija de antemano g, el valor t debe ser determinado atendiendo a la fdp de Student S (x) en lugar de la normal.

A manera de conclusión, si la dispersión es desconocida, el intervalo de confianza g para la esperanza matemática de una distribución normal se determina por:

7 r n r n 8

donde t se calcula a partir de la fdp de Student por la relación

2 | S (x) dx = g

Haciendo razonamientos similares se determinan teóricamente también las formas de los intervalos de confianza para dispersión de una distribución normal, o para otros parámetros de una distribución dada.

Los cálculos teóricos hechos pueden ilustrar además como se utilizaron los razonamientos probabilísticos para estimar volumen de muestras a priori.

Supongamos por ejemplo que tenemos una variable aleatoria con distribución normal con esperanza a desconocida y dispersión _ conocida, y queremos estimar la esperanza matemática a partir de la media en una muestra de cierto volumen n. ¿Cuál es el volumen mínimo de la muestra para estimar a por un intervalo de confianza **"g" y con error "_"** prefijado?

Sabemos que:

Prob(¦X-a¦< --]) = [---] | e dx = g

El número t se determina por la fdp normal a partir de "g" y debemos tener

[--] _ _, donde _ es el error prefijado.

p t _ p

3.3. Análisis descriptivo de datos con el SPSS/PC. Comandos FREQUENCIES, DESCRIPTIVES y MEANS

El SPSS/PC+ tiene tres procedimientos esenciales para el análisis descriptivo de datos: **FREQUENCIES, DESCRIPTIVES y MEANS.**

FREQUENCIES es típicamente un procedimiento para el estudio de frecuencias y cálculo de características numéricas de variables aleatorias discretas; pero puede utilizarse para hacer histogramas (omitiendo las tablas de valores) y calcular las estadísticas descriptivas de variables aleatorias continuas.

La sintaxis más simple de este comando se ilustra en el siguiente ejemplo. Supongamos que en nuestro fichero activo tenemos una variable DIA que representa el día de la semana en que acuden pacientes al Cuerpo de Guardia de cierta unidad hospitalaria con infartos del miocardio. DIA es una variable aleatoria discreta con valores posibles: 1.Domingo 2.Lunes 3.Martes 4.Miercoles 5.Jueves 6.Viernes 7.Sabado y por tanto, con un nivel de medición ordinal. Supongamos además que hay varios casos en el fichero que no ingresaron y que se les ha anotado en esta variable el valor 0 como valor perdido.

El resultado del comando:

FREQUENCIES

VARIABLES=DIA .será una salida como la siguiente:

DIA DIA DEL INGRESO

				Valid	Cum
Value label	**Value**	**Frequencies**	**Percent**	**Percent**	**Percent**
Domingo	**1.00**	**19**	**7.9**	**17.3**	**17.3**
Lunes	**2.00**	**11**	**4.8**	**10.0**	**27.3**
Martes	**3.00**	**19**	**7.9**	**17.3**	**44.5**
Miércoles	**4.00**	**17**	**7.1**	**15.5**	**60.0**

Jueves	**5.00**	**15**	**6.3**	**13.6**	**73.6**
Viernes	**6.00**	**13**	**5.4**	**11.8**	**85.5**
Sábado	**7.00**	**16**	**6.7**	**14.5**	**100.0**
MISSING	**0.00**	**130**	**34.2**	**MISSING**	
	------	**-----**	**------**		
TOTAL	**240**	**100.0**	**100.0**		

La tabla tiene una interpretación obvia. Cada fila se corresponde con uno de los valores posibles de la variable. Observe que por defecto la tabla está ordenada por valores ascendentes de la variable (excepto los valores perdidos que ocupan la última fila). Para cada valor posible se dá la frecuencia de casos, el por ciento que representa respecto al total de casos (incluyendo aquellos con valores perdidos), el por ciento válido que se calcula sobre la base del total de la muestra restando de antemano los casos perdidos y el porciento acumulado.

La **sintaxis** de **FREQUENCIES** incluye varios subcomandos opcionales que controlan lo siguiente:

FORMAT Controla el formato de la tabla. En particular permite definir un orden alternativo al ascendente de valores **(AVALUE): DVALUE, AFREQ y DFREQ**. Permite además suprimir las tablas, por ejemplo, para el trabajo con variables aleatorias continuas con la opción NOTABLE y controla otras opciones de formato que se pueden ver en la ayuda.

Ejemplo: FREQUENCIES

ALCOHOL/

FORMAT=NOTABLE /

HISTOGRAM=NORMAL /

STATISTICS=ALL.

es un comando que omite la tabla para la variable aleatoria continua **ALCOHOL** y que dibuja un histograma y extrae todos los estadísticos de esta variable.

BARCHART Que dibuja un gráfico de barras en una variable aleatoria discreta. En este gráfico, producido a lo largo de la hoja es posible dominar la escala del eje y (eje de frecuencias o por cientos). Ver sintaxis detallada en ayuda

HISTOGRAM Que produce un histograma para variables aleatorias continuas. Tal histograma se produce ranqueando la variable en intervalos. Es controlable el criterio de ranqueo y se puede además superponer al histograma una curva normal para el análisis visual del ajuste, como muestra el ejemplo de **ALCOHOL.**

HBAR Produce para cada variable listada un gráfico de barras o un histograma en dependencia de si le cabe o no en una hoja (esto es una manera de "distinguir automáticamente" variables aleatorias discretas y continuas en un listado de variables dadas en el comando.

NTILES Arroja los valores de las variables que dividen la muestra en grupos de igual número de casos. Por ejemplo,

FREQUENCIES

VARIABLES=ALCOHOL /

FORMAT=NOTABLE / NTILES=4 .

Arrojará tres valores de la variable que dividen a la muestra en 4 partes iguales (el del medio es precisamente la mediana y los dos extremos, las medianas de los intervalos a la derecha e izquierda del intervalo principal.

PERCENTILES Que refleja los valores de las variables que se corresponden con los percentiles. En el ejemplo anterior, se podría obtener el mismo resultado, dando en lugar de **NTILES**, el subcomando

PERCENTILES=0.25 0.5 0.75.

STATISTICS Que imprime estadísticos descriptivos de las variables.

Los estadísticos se expresan por palabras claves y están disponibles:

MEAN Media

KURTOSIS Curtosis

SEMEAN Error estándar de la media

MEDIAN Mediana

RANGE Rango (diferencia de máximo y mínimo)

MODE Moda

MINIMUN Mínimo

STDDEV Desviación estándar

MAXIMUM Máximo

VARIANCE Varianza

SUM Suma

SKEWNESS Sesgo o asimetría

SESKEW Error estándar del sesgo

SEKURT Error estándar de la curtosis

ALL Todos los estadísticos

Se hace notar que **FREQUENCIES** es el único comando del **SPSS/PC** que permite seleccionar los estadísticos por su nombre. Los restantes procedimientos exigen normalmente un código u opción.

MISSING Es un subcomando de **FREQUENCIES** que permite controlar el tratamiento de valores perdidos. Observe que por defecto **FREQUENCIES** distingue los valores perdidos, no los incluye en el cálculo de los porcientos válidos como tampoco en los estadísticos. El subcomando **MISSING** hace que el **SPSS/PC** trate los valores omitidos por el usuario como si ellos no lo fueran.

La **sintaxis** general del comando **FREQUENCIES** puede verse en el Manual [1] o sintetizada en [2] y no es requerido aprenderla de memoria pues siempre está disponible en la ayuda.

El segundo comando de análisis descriptivo de datos con el SPSS/PC es el comando **DESCRIPTIVES**. Él es propiamente un comando para variables aleatorias continuas y produce una tabla resumen de los estadísticos que se soliciten para cada una de las variables solicitadas.

La sintaxis general de DESCIPTIVES es:

DESCRIPTIVES

VARIABLES=lista de variables /

STATISTICS= statistic numbers /

OPTIONS=options numbers .

Los números que especifican los estadísticos son:

1: Media

2: Error estándar de la media

5: Desviación estándar

6: Varianza

7: Curtosis

8: Sesgo o asimetría

9: Rango

10: Mínimo

11: Máximo

12: Suma

13: Media, desviación estándar, mínimo y máximo (estadísticos por defecto)

Ejemplo:

DESCRIPTIVES

VARIABLES=CALIF1 TO CALIF10 EDAD .

El subcomando OPTIONS permite controlar el tratamiento de valores perdidos y algunos aspectos del formato de la tabla de salida con ciertos números de opciones. Como opción especial (número de opción 3), calcula las variables en forma estandarizada. Esto es si en la lista de variables especificada hay una variable denominada VAR, genera una nueva variable con el nombre ZVAR que se determina por la relación:

ZVAR = [----------

XVAR

Por ejemplo, el comando

DESCRIPTIVES

ALCOHOL AZUCAR SODIO PH /

OPTIONS=3 .genera en el fichero activo cuatro nuevas variables que denomina ZALCOHOL, **ZAZUCAR, ZSODIO y ZPH** y que representan los valores estandarizados de las 4 variables de la lista.

Esta es una facilidad importante porque algunos procesamientos estadísticos (por ejemplo el análisis de **clusters**) recomiendan comenzar estandarizando las variables para que todas tengan media 0 y desviación estándar 1 y **DESCRIPTIVES** brinda la posibilidad de conseguir estas variables sin que previamente usted tenga que calcular la media y la desviación estándar de cada variable ni hacer el **COMPUTE** correspondiente

El comando **MEANS** del **SPSS/PC** es un procedimiento que permite calcular los principales estadísticos descriptivos de una variable aleatoria continua (especialmente la media y la desviación estándar) por niveles de ruptura de una variable aleatoria discreta. Esta variable representa típicamente una variable de agrupación.

Ejemplos típicos pueden ser los siguientes:

MEANS

ALCOHOL BY CALIDAD.

Produce una tabla con la media y la desviación estándar de la variable ALCOHOL para la población general y para cada subgrupo determinado por la variable discreta CALIDAD.

MEANS

INDICE BY SEXO BY RAZA .

produce una tabla con la media y la desviación estándar de la variable INDICE para la población general, para cada uno de los sexos y dentro de estos para cada una de las razas. Cada BY adicional produce un nuevo nivel de ruptura y en total se admiten 6 dimensiones (5 variables de ruptura o frases claves BY).

La sintaxis general de MEANS tiene la forma:

MEANS

variable_continua BY var_discreta [BY var_discreta...]

STATISTICS= statistic numbers /

OPTIONS=options numbers .

Los números de opciones permiten aquí controlar algunos aspectos del formato de la tabla, sustituir la desviación estándar por la varianza o eliminar uno de los dos estadísticos fundamentales (media o desviación estándar). El subcomando STATISTICS permite realizar un test de linealidad o un análisis de varianza elemental respecto a la variable que representa el primer (y en este caso único) nivel de ruptura; pero sería sólo con fines descriptivos, en el sentido de obtener una idea inicial de la dependencia de la media. El SPSS/PC incluye otros comandos para hacer tales tipos de análisis con todo rigor.

Observe por último que ninguno de los tres comandos proporciona intervalos de confianza. Los intervalos de confianza para la media de una distribución normal se estiman en el SPSS/PC con un comando que veremos posteriormente (ONEWAY).

3.4. Pruebas de hipótesis. Verificación estadística de hipótesis estadísticas

El análisis de una investigación en cualquier rama de las Ciencias conduce muchas veces a formular una hipótesis en términos estadísticos que debe ser verificada.

Algunos ejemplos:

- Determinar si las frecuencias de calificación de un determinado examen (2,3,4 y 5) se corresponden con lo esperado.
- Determinar si la precisón de dos equipos de medición es la misma.
- Determinar si los valores medios de las mediciones de dos magnitudes coinciden.
- Determinar si la frecuencia observada de aparición de un suceso coincide en la probabilidad hipotética del mismo.
- Determinar entre un conjunto de niveles de un determinado factor cuál es el que más influye en el valor de cierta variable, etc.
- Determinar si dos variables cualitativas tienen o no la misma distribución.

Se considera que la hipótesis ha sido formulada en términos estadísticos si ha sido precisada de forma tal que investigue las características de la distribución de una variable aleatoria.

Así, hay hipótesis que hablan del tipo de distribución supuesta. En el ejemplo de las calificaciones se pretende comprobar si las frecuencias de las muestras "se ajustan" a cierta distribución esperada.

En otros casos se conoce la ley de distribución y las hipótesis se perfilan en términos de precisar sus parámetros. En el ejemplo del método de experimentación se supone que la variable X que éste mide tiene una distribución normal y se desea conocer si su dispersión o varianza coincide con un valor prefijado.

Se puede formular hipótesis sobre la igualdad de los parámetros de dos o varias distribuciones. En el ejemplo de los equipos de medición el resultado de la medición de cada equipo: X y X se admite distribuido normalmente y se desea verificar si tienen la misma dispersión.

Son posibles también hipótesis sobre independencia de las muestras y otros aspectos que puedan formularse en última instancia en términos de una distribución o de los valores de ésta.

Hay pruebas estadísticas destinadas especialmente a comprobar ciertos parámetros de una distribución sin conocer previamente esta, pero en cualquier caso se trata siempre de una hipótesis formulada en términos de una distribución que se busca o de los parámetros de ésta.

La hipótesis "en el año 1990 no habrá guerra" no es estadística, o al menos no está formulada en términos estadísticos. Por ello tiene un papel fundamental en las investigaciones, conocer cuál es el tipo de pruebas estadísticas más frecuente de aplicación en la rama o especialidad y saber cómo se formulan estas hipótesis en términos estadísticos.

Algunos conocimientos básicos generales son necesarios. A la par de la hipótesis presentada (que llamamos hipótesis fundamental o H) se analiza la hipótesis que la contradice (que llamaremos hipótesis alternativa o H).

Por ejemplo, si la hipótesis fundamental consiste en suponer que la precisión de las mediciones X y X de dos equipos es igual, tendremos:

H : D(X) $ D(X) hipótesis alternativa

Si se sabe con anterioridad que D(X) _ D(X) podríamos tener

H : D(X) = D(X)

H : D(X) > D(X)

La hipótesis presentada puede ser cierta o falsa y la verificación de ella se realiza por métodos estadísticos. Esto justifica la redundancia "verificación estadística de hipótesis estadísticas" que hemos utilizado en el título del epígrafe.

La verificación estadística consiste esencialmente en comprobar si una determinada variable estadística (estadígrafo) tiene para la muestra los valores que pudieran ser esperados si la hipótesis fundamental fuera cierta.

La teoría de inferencia estadística o teoría de los test de hipótesis se ocupa pues de buscar para hipótesis típicas, una variable aleatoria, especialmente escogida, cuya distribución exacta o aproximada es conocida cuando la hipótesis fundamental es cierta. Esta variable estadística, denominada estadístico, estadígrafo, criterio estadístico, etc., y su distribución condicionada a la hipótesis fundamental proporcionan el criterio de decisión.

Si el valor concreto que tiene el estadígrafo en la muestra sale fuera de los rangos probables, la hipótesis fundamental se rechaza y se acepta entonces la hipótesis alternativa.

Si el valor concreto que tiene el estadígrafo en la muestra está dentro de los rangos probables de los valores de la variable conocida a partir de la distribución condicionada a H . no existen razones para rechazar esta hipótesis y entonces se acepta.

Así en nuestro ejemplo de comparación de dispersiones está demostrado que un buen estadígrafo es el cociente de sus estimados muéstrales:

F =

No resulta difícil ver que si X y X tienen distribución normal entonces **D(X) y D(X)** tienen distribución c y si la hipótesis fundamental es cierta, de acuerdo con lo que vimos en el Capítulo 1, entonces F tiene distribución conocida (Distribución de Fisher- Snedecor) y sus valores no pueden estar demasiado lejanos de uno. Más precisamente, si la hipótesis H es cierta, se conocen los valores probables de esta variable para una muestra de volumen dado.

Si al evaluar este estadígrafo para la muestra:

S

F =

S

Obtenemos un valor fuera del rango probable (digamos mucho más grande que 1), se rechazará la hipótesis fundamental. En caso contrario hay que aceptarla. Conociendo la función de distribución de **Fisher-Snedecor** podremos precisar cuán grande debe ser el valor de F observado para tener suficientes criterios para rechazar la hipótesis fundamental.

Otro concepto importante al formular las hipótesis y seleccionar el estadígrafo es el de los tipos de errores.

Al hacer verificaciones estadísticas se pueden producir errores de dos tipos.

Errores del tipo I: Rechazar una hipótesis verdadera, digamos rechazar H cuando H es cierta.

Errores del tipo II: No rechazar una hipótesis falsa, digamos digamos no rechazar H cuando realmente H es falsa.

Al formular las hipótesis y seleccionar los estadísticos estos tipos de errores resultan en cierto sentido complementarios aunque su relación no es tan simple y por tanto son difíciles de controlar simultáneamente.

Está claro que la consecuencia de estos errores puede ser muy diferente. Supongamos por ejemplo que los equipos de qué hablamos son instrumentos diferentes de medición de resistencia de determinada estructura constructiva. Rechazar la solución correcta "utilizarlos simultáneamente" sería un error de primer tipo que entrañaría probablemente pérdida de tiempo en la construcción. Por el contrario la resolución "utilizar simultáneamente" los equipos que realmente no son igualmente precisos puede traer consecuencias, más o menos graves que el error anterior, según los criterios técnicos del especialista.

Para un mismo par de hipótesis puede haber más de un estadígrafo más o menos sensible a la probabilidad de error de uno u otro tipo. El mismo par de hipótesis verificados con estadígrafos diferentes, resulta en pruebas estadísticas diferentes y es aplicado a diferentes tipos de situaciones. De allí que el investigador deba saber no solo formular el par de hipótesis contrarias, sino comprender cabalmente como seleccionar entre ellas cuál

es la hipótesis fundamental y cuál la alternativa y además cuál es el criterio o estadígrafo que mejor conviene a su problemática.

Llamaremos en general "significación del test" a la probabilidad del error del tipo I con el estadígrafo seleccionado:

_ = Prob(Rechazar H con el estadígrafo seleccionado ¦ **H** es cierta) y pondremos especial énfasis en controlar este tipo de error.

En nuestro ejemplo, el test se termina de formular con el cálculo de la significación

Prob (F > F) = 7

Si obtenemos **_ < 0.01** se puede rechazar la hipótesis fundamental por considerar que las precisiones de ambos equipos difieren en forma "altamente significativa".

Si este valor _ está entre **0.01 y 0.05** se puede rechazar la hipótesis y considerar que existen diferencias "significativas" en la precisión de los equipos.

Si el valor **_ > 0.05** no hay razones para rechazar la hipótesis y debemos considerar que la precisión de los equipos es la misma. En algunas investigaciones nos conformamos con valores de _ menores que **0.05** y es usual comentar en este caso **(0.05<_<0.10**) que las diferencias son "medianamente significativas"

Si denotamos ß a la probabilidad de un error de tipo II.

ß = Prob (Aceptar H con el estadígrafo seleccionado? H es falsa) está claro conceptualmente, que para un volumen dado de muestra, no es posible reducir simultáneamente _ y ß. Si se deducen las exigencias de _, entonces **ß** crecerá. Por ejemplo si tomara _ muy pequeña, la probabilidad de un error de tipo I será infima; pero realmente en pocos casos se rechazará la hipótesis, y por tanto tiende a aumentar la probabilidad de aceptarla aunque sea falsa, esto es, la probabilidad **ß** del error de tipo II.

?¿Cómo elegir más conveniente el nivel de significación o comparación de _: **0.01, 0.05**, **0.10, etc**?. La respuesta depende de la gravedad de las consecuencias de los errores para cada problema concreto. Por ejemplo si un error de tipo I da lugar a grandes pérdidas y uno de tipo II a pequeñas pérdidas hay que tomar _ lo más pequeño posible, digamos _<0.01. Como regla general se trabaja con una significación del 5% (valor comparativo de _ igual a 0.05). Cuando se desea imponer fuerza a la conclusión de rechazo de la hipótesis se trabaja con significaciones del 1% y por el contrario, cuando se trata de investigaciones preliminares cuyo objetivo es encontrar cualquier aspecto que sea potencialmente significativo, se trabaja con niveles de significación del 10% (comparación de _ con 0.10).

El valor

1- ß = Prob (no cometer error del tipo II) = = Prob (Rechazar H con el estadígrafo dado ¦ H es falsa) se llama potencia del criterio.

La teoría de inferencia estadística se ocupa pues de buscar estadígrafos cuya distribución exacta o aproximada es conocida cuando la hipótesis fundamental es cierta; entre estos escoge para formular los test, aquellos que sean "más sensibles" al incumplimiento de la hipótesis fundamental de forma tal que permite trabajar con bajos niveles de significación y obtener la máxima potencia del criterio.

Resultados de **J. Neyman** y **E. Pearson** que pueden estudiarse más profundamente después, permiten construir estadígrafos para un _ prefijado para los cuales ß es mínimo.

El único método para disminuir simultáneamente las probabilidades de errores de tipo I y II consiste en aumentar el volumen de las muestras.

En el curso veremos entre otros, tests para verificar:

1.- H : fdp(X) = f (X)

H : fdp(X) $ f (X) utilizando un estadígrafo c

2.- H : D(X) = D(X)

H : D(X) $ D(X) utilizando un estadígrafo F

3.- H : M(X) = M(X)

H : M(X) $ M(X) utilizando un estadígrafo con distribución normal o con distribución de Student.

4.- H : P = P

H : P $ P utilizando un estadígrafo con distribución normal

5.- H : M(X) = M(X) = ...= M(X)

H : Ei,j : M(X) @ M(X) con un estadígrafo F

6.- H : fdp(X) = fdp(X)

H : fdp(X) $ fdp(X) utilizando ciertos estadígrafos los cuales permiten dar respuestas a los problemas ejemplos formulados al principio de este epígrafe.

Los tests 2 y 3 son típicamente tests de comparación de poblaciones normales (si X y X) tienen distribución normal, su posible relación se determina comparando sus valores esperados y sus dispersiones. El test 5 es una generalización del test 3 de comparación de medias de distribuciones normales y por su importancia fundamental, sus múltiples aplicaciones e implicaciones dedicaremos un capítulo aparte. El conduce a los llamados Análisis de Varianza.

El test 4 mide el parámetro de una distribución normal. En conjunto, los tests 2,3,4, y 5 tienen de común que son "paramétricos", en el sentido de que parten del conocimiento de la distribución (digamos normal o binomial) y verifican hipótesis sobre los parámetros de la distribución.

Los tests 1 y 6 son "no paramétricos" porque se refieren a determinar la distribución de una variable que no se conoce de antemano o a la comparación de la distribución de dos variables sin ser previamente conocida. De hecho, el test 6 es una "alternativa no paramétrica" de los tests 3 y 4 de comparación de poblaciones normales. También hay alternativas no paramétricas del análisis de varianza.

En general, la docimasia de hipótesis no paramétrica es de considerable importancia para las Ciencias Sociales y para cualquier investigación en que se trabaje con niveles cualitativos, más que valores numéricos; pero adquiere también una relevancia fundamental en todas las ramas de la ciencia y la técnica cuando se desconocen a priori las formas de la distribución o hay razones para pensar que no se ajustan a la distribución normal.

Como parte del análisis descriptivo de datos veremos en particular dos tests del tipo 1 para verificar si una distribución se ajusta o no a una esperada.

3.5.Pruebas de bondad de ajuste a una distribución

Procedimientos no paramétricos **CHISQUARE y KOLMOGOROV-SMIRNOV.**

Supongamos en principio que tenemos una variable aleatoria discreta X que puede tomar un número finito de valores **x , x , ... ,x** y a partir de los datos de una muestra queremos verificar si la distribución de X se ajusta a cierta distribución, esto es, si

Prob(X=x) = E ,

Donde E es cierto valor esperado. Sea O la frecuencia muestral observada del valor x . Parece natural construir un test sobre la base del estadígrafo:

No resulta difícil demostrar que si el volumen n de la muestra es bastante grande, y la hipótesis fundamental es cierta este estadígrafo tiene aproximadamente distribución c con **n-1** grados de libertad y sus valores no deben ser demasiados lejanos de **0**.

Esto es, el estadígrafo resultará en un valor observado V grande si las frecuencias observadas y esperadas difieren sustancialmente. Por tanto, con ayuda de la distribución **c** podemos determinar cuán grande debe ser el valor **V** para rechazar el ajuste a la distribución con una probabilidad de **error _.**

En la práctica, **si _ = Prob(V > V)** es menor que el nivel de significación comparativo escogido (por ejemplo **_<0.05)** se rechaza el ajuste a la distribución esperada. Si _ es mayor que el valor prefijado, no hay razones para rechazarla.

La potencia de este test depende como en todos del volumen de la muestra. Hay que lograr además que las frecuencias esperadas E para cada valor x no sean demasiado pequeñas, pues en tal caso crece el valor de cada sumando en V y se exagera la significación del chi-cuadrado. En la práctica se considera que la aplicación del tests es cuestionable si hay frecuencias esperadas menores que 5 en alguno de los valores posibles. Cuando esto ocurre, es recomendable "unir" categorías o valores posibles poco frecuentes en una única categoría que tenga una frecuencia esperada menor que ese mínimo.

El SPSS/PC tiene un procedimiento no paramétrico para la dócima del test chi-cuadrado de bondad de ajuste. La sintaxis es:

NPAR TESTS

CHISQUARE=variable (rango mínimo, rango máximo) /

EXPECTED=number list .

En esencia, la sintaxis exige especificar el nombre de la variable (o de las variables, pues pueden ser más de una en el mismo subcomando **CHISQUARE)** seguidas opcionalmente de un rango mínimo y máximo de sus valores posibles. Las variables probadas tienen que ser enteras. Para especificar las frecuencias esperadas, porcentajes o proporciones, se utiliza una lista de valores sobre el subcomando **EXPECTED**. Estos valores tienen que ser positivos y más grandes que 0. Es importante dar un valor esperado para cada uno de los valores posibles en orden. Se recomienda utilizar proporciones, comprobando que su suma dé uno. Si no se especifican valores esperados, el **SPSS/PC** asume por omisión que se espera igual frecuencia de valores para todas las categorías. Si de acuerdo a las proporciones esperadas y a los volúmenes de la muestra, alguna de las frecuencias esperadas resultara menor que

5, el **SPSS/PC** hace una advertencia reportando que el test es cuestionable en este ejemplo por tener frecuencias esperadas bajas.

Ejemplos:

- Si en un fichero activo tenemos una variable **CALIF** que representa las calificaciones (en categorías 2, 3,4 y 5) y queremos verificar si ella se ajusta a la distribución clásica que se establece en Pedagogía, debemos escribir:

NPAR TESTS

CHISQUARE=CALIF /

EXPECTED=0.25 0.6 0.1 0.05

- Si en un fichero activo tenemos una variable DIA que representa como antes se ilustró, el día de ingreso por infarto del miocardio en cierto Cuerpo de Guardia Hospitalario y nos interesa determinar si todos estos días son "igualmente probables o no", deberemos utilizar el comando:

NPAR TEST

CHISQUARE=DIA.

En este caso, la probabilidad esperada por omisión es 1/7 para cada día de la semana.

En cualquier caso, la salida del SPSS/PC incluirá una tabla de valores con las frecuencias observadas y esperadas en la muestra, el residual (diferencia entre valores observados y esperados), el valor del estadígrafo chi-cuadrado, sus grados de libertad (número de valores posibles menos 1) y su significación. Mirando esta última decidimos simplemente rechazar o no el ajuste a la distribución propuesta. El examen de la tabla nos permite detallar la conclusión. Por ejemplo, si el test del segundo ejemplo resultara significativo, en la tabla podríamos ver cuál es el día de la semana en que el número de infartos observados es mucho más o mucho menos frecuente que lo esperado

Véase una salida típica del SPSS/PC para este test:

----------Chi-square Test

DIA DIA DE INGRESO

Cases

Category	Observed	Expected	Residual
Domingo 1.00	19	15.71	3.29
Lunes 2.00	11	15.71	-4.71
Martes 3.00	19	15.71	3.29
Miércoles 4.00	17	15.71	1.29
Jueves 5.00	15	15.71	-.71
Viernes 6.00	13	15.71	-2.71
Sábado 7.00	16	15.71	.29

Total	110		

Chi-Square	**D.F.**	**Significance**
3.400	**6**	**0.757**

Como en este caso resultó la significación **0.757>0.05,** no hay razones para suponer que haya un día de la semana en que hay signicativamente más o menos ingresos, o en otras palabras, de los datos de la muestra debemos aceptar que todos los días son igualmente probables.

Supongamos ahora que la variable X no es continua y queremos analizar si su distribución se ajusta a cierta ley dada. Podemos utilizar el test chi-cuadrado, ranqueando previamente el

valor de la variable X en intervalos y determinando la frecuencia esperada en cada uno bajo el supuesto de que la distribución es la hipotética; pero esto último debemos hacerlo **"fuera del SPSS/PC"** y utilizando concretamente la función de distribución.

Comoquiera que este proceso puede ser laborioso y además la potencia del criterio depende mucho de los intervalos de ranqueo de nuestra variable, está formulado e implementado en el SPSS/PC un test, conocido como Test de Kolmogorov-Smirnov para la prueba de bondad de ajuste a algunas distribuciones continuas importantes, concretamente, la distribución normal, la distribución uniforme y la distribución de Poison.

El test de Kolmogorov-Smirnov verifica en definitiva el ajuste a la distribución hipotética determinando también las diferencias más extremas en valores observados y esperados en intervalos óptimamente adecuados y convierte esta diferencia en un estadístico Z que tiene aproximadamente distribución normal normalizada. Si la hipótesis fundamental sobre el buen ajuste a la distribución es cierta este estadígrafo no puede resultar demasiado grande.

La sintaxis de este test es:

& UNIFORM [min,max]*

En el caso de la distribución uniforme los parámetros opcionales son los extremos mínimo y máximo del intervalo. En el caso de la distribución normal son la media y la desviación estándar. En el caso de la distribución de Poisson es la media (1/μ) según el capítulo 1. Si no se dan algunos de estos parámetros, el test toma por omisión las estimaciones muestrales de estos. Puede darse uno de los dos parámetros, cuando hay dos, pero si es el segundo debe precederse por una coma para significar que el primero debe tomarse de la muestra.

Por ejemplo, si en el fichero activo tenemos las variables **ALCOHOL y AZUCAR** y queremos probar su ajuste a la distribución normal, sin prefijar parámetros, podemos escribir:

NPAR TESTS

K-S(NORMAL)=ALCOHOL AZUCAR.

La salida del SPSS/PC tendrá el aspecto típico siguiente:

----------Kolmogorov -Smirnov Goodness of Fit Test

ALCOHOL GRADOS PROOF DE ALCOHOL

Test Distribution - Normal Mean: 4.577

Standard Deviation: .603

Cases: 35

Most Extremes Differences

Absolute	Positive	Negative	K-S Z	2-tailed P
0.22941	0.155585	-0.22941	1.357	0.050

y claro está, realizará una tabla similar para la variable AZUCAR.

En un caso como el del **ALCOHOL**, la significación es lo suficientemente pequeña como para arrojar dudas sobre la normalidad.

Debe tenerse presente en general que el test de K-S para la prueba de normalidad es bastante potente con volúmenes de muestras superiores a 100. Esto es, en muestras pequeñas, el rechazo de la bondad de ajuste se producirá más raramente. Como en el

ejemplo, la muestra es de volumen **35,** este criterio adicional nos ayuda a concluir que la variable **ALCOHOL** no se ajusta a una distribución normal.

Si en el fichero activo tenemos una variable ERROR que significa el error de una medición y queremos probar su ajuste a la distribución uniforme en el intervalo (-1,1), debemos escribir

NPAR TESTS

K-S(UNIFORM,-1,1)=ERROR.

El test de Kolmogorov-Smirnov tendrá otros usos, como veremos posteriormente.

3.6 Conclusiones

Cualquier procesamiento estadístico conlleva -y muchas veces comienza- con una caracterización de las variables desde el punto de vista de su distribución y las características fundamentales de ella. Ello se logra con el análisis descriptivo de datos.

La estimación de características fundamentales de la variables a partir de los datos de una muestra se hace sobre la base de una teoría conocida como Teoría de la Estimación Estadística y el SPSS/PC cuenta con tres procedimientos básicos para hacer tales estimativas: FREQUENCIES, DESCRIPTIVES y MEANS.

La determinación del ajuste a una distribución, se hace sobre la base de la teoría de Tests de hipótesis y el SPSS/PC cuenta con dos procedimientos básicos para tal análisis: el test CHI-CUADRADO de ajuste a una distribución cualquiera y el test de KOLMOGOROV-SMIRNOV para prueba de bondad de ajuste a la distribución normal, uniforme o de Poisson.

Hemos aprovechado este capítulo para dar un concepto riguroso de muestra, ideas sobre lo que son las estimativas adecuadas y el fundamento de los tests de hipótesis, todo lo cual será utilizado en el resto del material.

3.7 Ejercicios del Capítulo III

1. Utilice el fichero de datos y el programa ALUMNOS de los ejercicios del Capítulo anterior.

- Analice la composición por sexos en el grupo de alumnos.
- Analice la composición por edades. Determine la edad media y su variabilidad y haga un estudio de frecuencia por edades en rangos, por ejemplo: menores de 17, entre 17 y 19 y mayores de 19 años.
- Haga un análisis descriptivo de las calificaciones en cada asignatura. Determine en particular si estas calificaciones se ajustan a las frecuencias esperadas pedagógicamente y cuál es la asignatura cuyas calificaciones difieren más significativamente de la distribución esperada.
- Determine si el índice académico del alumno en el año se ajusta en el grupo a una distribución normal. Precise sus principales características numéricas.
- Suponga que al definir un alumno de alto rendimiento en computación, se tiene en cuenta que tenga un índice académico superior a 4 y calificaciones de 5 en las asignaturas de computación. Conforme una variable dicotómica que informe si el alumno es o no de alto aprovechamiento y calcule la proporción de alumnos con tal condición en su grupo.

- Suponga que históricamente usted tiene un 5% de alumnos de alto aprovechamiento en computación en una especialidad como esta.

Determine si la proporción en este grupo de alumnos se corresponde o difiere significativamente del dato histórico.

Ejecute su programa. Analice cualquier advertencia (WARNING) o error (ERROR) hasta eliminarlos.

2. Utilice en esta pregunta el ejemplo y los datos de los vinos.

Recuerde que en el capítulo anterior usted terminó generando un fichero sistema 'VINOS.SYS' que puede utilizar ahora.

- Determine la frequencia de las provincias de origen de los vinos. Obtenga un gráfico de barras. Determine la moda.
- Determine la distribución de frecuencias de las variables ordinales: CALIDAD, DISP, PREMIO y CLARIDAD. Obtenga también gráficos de barras y como estadísticos la moda y la mediana. ¿Existe alguna calidad que predomine significativamente entre los vinos?. ¿Se ajustan los premios a una distribución similar a la de las calificaciones de estudiantes?
- Determinar un resumen estadístico de las principales características de las variables aleatorias continuas con el procedimiento DESCRIPTIVES. ¿Cuáles estadísticos son más interesantes? ¿Cuáles de estas variables se ajustan a una distribución normal?

3. Cierto productor fabrica piezas cuya calidad se cataloga en forma descendente por las categorías A, B, C, D, y E. La calidad de cada pieza resulta aleatoria en la producción y la ganancia depende de la calidad según los lotes especificados abajo. Un lote de 1000 piezas arrojó los siguientes resultados que se exponen en la misma tabla.

Haga un estudio descriptivo de la calidad y la ganancia. Utilice todos los estadísticos posibles pero elimine aquellos que no aportan información esencial. El estudio debe incluir gráficos de distribución y un estimado de la ganancia total por lotes de 1000 piezas.

Calidad	Ganancia en dependencia de calidad	Número de piezas de esa calidad en el lote
A	$ 1.00	310
B	0.80	250
C	0.60	244
D	0.00	71
E	-0.60	125

Sugerencia: En este ejemplo vale la pena incluir los datos dentro del fichero programa y utilizar adecuadamente variables de ponderación (Ver comando WEIGHT en Capítulo anterior).

4. Se observa una fuente radioactiva durante 100 intervalos de tiempo de 5 segundos cada uno y se cuenta el número de partículas emitidas. En la práctica, 4 mediciones no son

confiables y por tanto se dispone del número de partículas emitidas en 96 intervalos de tiempo de 5 segundos. Teóricamente, el número de partículas que se emiten en un tiempo t (segundos) debe obedecer una ley de probabilidad de Poisson con parámetro μt donde μ es la razón de emisiones por segundo.

-Pruebe la hipótesis de ajuste a partir de los datos:

Número de partículas emitidas = k	**Número de intervalos en los cuales hubo k emisiones**
0	**59**
1	**27**
2	**9**
3	**1**
Más de 3	**0**

-¿Cuál es el parámetro μ para esta sustancia?. Determine otras características numéricas de esta distribución. Grafíquela.

5. Teóricamente, el número de bacterias de cierto tipo presentes en t cm de agua, obedece o bien:

a) A una ley de probabilidad de Poisson con parámetro μt donde μ es la razón de bacterias por cm de agua

b) O a una ley de distribución normal con media μt y dispersión aproximadamente igual a este valor.

Para decidir el mejor ajuste a una de estas dos distribuciones se contabilizaron las bacterias existentes en 576 pequeñas muestras de 1/4 cm de agua resultando en los datos de la tabla de abajo.

Decida cuál de las distribuciones en más aceptable. Argumente su respuesta

Número de bacterias presentes = k	**Número de muestras con k bacterias**
0	**229**
1	**211**
2	**93**
3	**35**
4	**7**
5 o más	**1**

6. El tiempo de trabajo sin fallos (por roturas de hilo), medido en minutos, de una máquina textil es una variable aleatoria. Para aplicar técnicas basadas en teoría de colas

relacionadas por ejemplo con la optimización de los servicios de atención del operario, es imprescindible demostrar primero que se cumplen las hipótesis necesarias sobre el ajuste a la distribución exponencial del tiempo de trabajo sin fallos. El parámetro l de esta distribución exponencial debe ser interpretado como la razón de fallos por minuto, o lo que es equivalente, el inverso algebraico del tiempo medio de trabajo sin fallos (l = 1/T)

Se realizaron una serie de mediciones en la fábrica que se resumen en la siguiente tabla:

Tiempos sin fallos	**Número de observaciones en ese rango**
0 - 5 mtos	**710**
5 - 10	**210**
10- 15	**55**
15- 20	**19**
Mas de 20	**6**

- Estime el tiempo medio de trabajo sin fallos y el valor de l (para esto puede no valer la pena cargar el SPSS/PC)
- Determine cuál debería ser la frecuencia esperada sin fallos en cada uno de los intervalos de tiempo que aparecen en la tabla (estos cálculos no puede hacerlos con el SPSS/PC)
- Determine con el SPSS/PC si el tiempo medio de trabajo sin fallos se ajusta a la distribución exponencial.

Capítulo IV. Independencia de variables y medidas de asociación

Introducción

Para el examen de dos variables aleatorias discretas (nominales u ordinales) se utilizan elementalmente tablas de contingencia y el estadístico Chi Cuadrado. Para el análisis de independencia de variables aleatorias continuas se utiliza popularmente el coeficiente de correlación. Queremos analizar éstas y otras formas de medir la asociación entre variables y en particular esclarecer las ventajas y desventajas de cada una.

4.1. Independencia de variables aleatorias discretas. El test Chi-cuadrado.

Una tabla de contingencia (producida por ejemplo por el procedimiento CROSSTABS del SPSS/PC +) tiene el aspecto siguiente:

Crosstabulation : HIPNOSIS Nivel logrado

By MIGRANA Nivel de cura alcanzado

Count p p p p

p Curada p Mejoria p Sin camb p Row

MIGRANA -> Tot Pctp 1.00 p 2.00 p 3.00 p Total

HIPNOSIS[---------+---------+---------+----------¦

1.00 p 13 p 5 p p 18

Profunda p 72.2 p 27.8 p p 18.0

p 56.5 p 15.6 p p

p 13.0 p 5.0 p p

+----------+

2.00 p 10 p 26 p 17 p 53

Media p 18.9 p 49.1 p 32.1 p 53.0

p 43.5 p 81.3 p 37.8 p

p 10.0 p 26.0 p 17.0 p

+----------+

3.00 p p 1 p 28 p 29

Ligera p p 3.4 p 96.6 p 29.0

p p 3.1 p 62.2 p

p p 1.0 p 28.0 p

[----------

Column 23 32 45 100

Total 23.0 32.0 45.0 100.0

Number of Missing Observations : 0

Esta tabla se origina en una investigación en la cual se está utilizando la hipnosis para el tratamiento de la migraña. Se desea determinar si hay alguna dependencia entre el nivel de hipnosis que se logra en los sujetos y el resultado final del tratamiento.

Aunque el examen de los diferentes porcentajes de filas y columnas en una tabla de contingencia es útil como un primer paso en el estudio de la relación entre las dos variables, estos porcentajes no permiten cuantificar o probar esa relación. Para este propósito se consideran varios índices que miden la extensión de la asociación y/o algunos tests estadísticos que prueban la no existencia de asociación.

El test Chi-Cuadrado de Independencia

Supongamos que tratamos con dos variables aleatorias discretas (nominales u ordinales) para la cual se ha realizado una tabla de contingencia. Las dos variables serán independientes si la probabilidad de que un caso quede en una celda dada de la tabla es igual al producto de las probabilidades marginales de las dos categorías que definen la celda.

Para construir un estadístico que mide la independencia se repite el cálculo anterior para cada celda de la tabla. Si las variables son independientes, la probabilidad de que una observación caiga en la celda (i,j) se estima por:

P(row=i and column=j) = ---------- * ----------

Para obtener la frecuencia esperada E multiplicamos por el volumen total de la muestra:

E = ----------

Estas frecuencias esperadas deben ser comparadas con las frecuencias observadas O en la tabla.

Las diferencias **E - O** se llaman residuales, se elevan al cuadrado para evitar la compensación de diferencias positivas y negativas y se dividen por las frecuencias esperadas **E** para establecer magnitudes relativas. Resulta en el estadístico:

² s s

Si la hipótesis fundamental de independencia es cierta, este estadístico tiene distribución aproximadamente igual a la Chi Cuadrado, con grados de libertad determinado por el producto (C-1)*(R-1), donde C y R son el número de filas y columnas de la tabla. La idea de los grados de libertad es aquí clara pues este es efectivamente, el número de celdas de la tabla que "podrían llenarse libremente" si están fijados los totales marginales de filas y columnas.

El valor del estadístico anterior es comparado con los valores teóricos de la distribución Chi-cuadrado, lo que determina la significación del valor y por tanto un criterio para rechazar o no la hipótesis de independencia. Si en ejemplo de partida se pide al SPSS el estadístico k , obtendríamos:

Chi-Square D.F. Significance Min_E.F. Cells__with__E.F.<5

65.52526 4 .0000 4.140 1 OF 9 (11.1%)

Aquí el valor del Chi-Cuadrado observado es 65.52526, el número grados de libertad 4 (2*2), la mínima frecuencia esperada: 4.140 y el número de celdas con frecuencia esperada menor que 5 es 1 y la significación del Chi-cuadrado, lo más importante, es 0.0000, menor que 0.05, e incluso menor que 0.001. Podemos rechazar entonces la hipótesis de independencia con nivel de significación del 1%. El hecho de que haya celdas con frecuencia

esperada menor que 5 puede ser, sin embargo un indicador de que el Chi Cuadrado esté deformado por exceso, como veremos más tarde.

El Test Chi-cuadrado tiene realmente muchas limitaciones y los principales detractores llegan incluso a decir que el único caso en que él puede ser aplicado con fiabilidad, es el caso de las tablas **2*2**.¿Cuáles son realmente las limitaciones?.

1. El test Chi-cuadrado es una prueba de independencia. Suministra muy poca información sobre la forma y sobre la estrechez de la asociación entre las variables. Por ello junto con el Chi-cuadrado o en lugar de él se utilizan otros estadísticos o pruebas que hablan de la forma o fortaleza del enlace.

2. La magnitud del Chi-cuadrado observado en una tabla de contingencia depende no solamente de la bondad del ajuste al modelo de independencia sino también del volumen de la muestra. En cambio la significación del test, que se determina por los valores del Chi-cuadrado teórico no depende del volumen de la muestra sino de las dimensiones de la tabla. De esta forma, si el volumen de una muestra crece, sin variar las dimensiones de la tabla, el

Chi-cuadrado observado crece y puede hacerse más significativo aún cuando los residuales se mantengan relativamente pequeños en relación con las frecuencias esperadas. Algunos de los estadísticos que veremos intentan corregir el Chi-cuadrado de este defecto incluyendo en su cálculo, el volumen de la muestra junto con las dimensiones de la tabla y/o estandarizando su valor, por **ejemplo entre 0 y 1.**

3. La distribución Chi-cuadrado para el estadístico es sólo aproximada y esta aproximación se sustenta en ciertas condiciones.

Los datos deben provenir de muestras aleatorias con distribuciones multinomiales y las frecuencias esperadas en cada celda no deben ser excesivamente pequeñas. Tradicionalmente se ha recomendado que las frecuencias esperadas deben ser mayores o iguales que 5, aunque estudios más recientes (Everitt, 1977) demuestran que esto es quizás demasiado exigente. Por ello, junto al cálculo del Chi- cuadrado los paquetes estadísticos reflejan normalmente si hay celdas con frecuencias esperadas menor que 5. A veces la presencia de celdas esperadas con frecuencia menor de 5 puede resolverse "reagrupando", lo que conduce a una reducción de las dimensiones de la tabla; pero en otros casos esto no es conveniente. En el ejemplo de la hipnosis, hay una celda con frecuencia esperada menor que 5 y esa celda es precisamente la (1,1): Hipnosis Profunda y Nivel de cura total. Veáse que:

E = ---------- = --------- = 4.1

Algunos paquetes **(incluidos el SPSS/PC)** permiten incluir en cada celda la frecuencia esperada para hacer este análisis. Lo importante es cómo evitar en general este problema. En esencia, las tablas de contingencia no pueden tener dimensiones demasiado grandes. Si se quiere evitar frecuencias esperadas bajas, debemos reducir las dimensiones de la tabla. En nuestro ejemplo, el problema no parece tan grave, pero si quisiéramos eliminar las celdas con frecuencia esperada menor que 5 podríamos lograrlo, considerando sólo dos categorias de hipnosis ("Profunda o Media" y "Ligera") o dos niveles de efectividad en el tratamiento ("Curado total o parcial" y "Sin cambios"); pero con cualquiera de las dos variantes perdemos información y no parece tan imprescindible.

Para mejorar la aproximación de la distribución en el caso de una tabla 2x2, se utiliza frecuentemente la "corrección por continuidad de Yates". Esta corrección consiste en reducir en 0.5 el valor de los residuales antes de elevar al cuadrado. Conover y Mantel (1974) han desarrollado una controversia sobre los méritos o insuficiencias de esta corrección, pero pese a todo, es bastante usada.

Se puede utilizar, en lugar del test Chi-cuadrado un test basado en la distribución hipergeométrica, conocido como test exacto de Fisher. Agunos paquetes (como el SPSS/PC) lo incluyen como alternativa, al menos para casos de tablas 2x2. En tales casos la significación que se obtiene es exacta y nos liberamos de los errores de aproximación del Chi-cuadrado; pero para tablas de dimensiones mayores, el cálculo de la probabilidad exacta se dificulta incluso en computadoras, por los cálculos de los factoriales grandes. En el Banco de Software del Centro de Estudios de Informática hay un programa realizado por D. Izada y L. García (denominado Tablas) que a partir de los datos de frecuencias observadas en una tabla de contingencia de cualquier dimensión calcula las probabilidades exactas. Sólo esto es una alternativa definitiva para corregir el problema del carácter aproximado del test Chi-cuadrado.

4.2. Otras medidas de asociación entre variables aleatorias discretas nominales y ordinales

En muchas investigaciones, más que discernir la dependencia de dos variables nos interesa la naturaleza y fortaleza de la asociación. Los indicadores que miden esto se llaman medidas de asociación. Ninguna medida resume adecuadamente todos los tipos de asociación posibles. Las medidas se diferencian por su interpretación, y por la forma en que ellas pueden reflejar asociaciones perfectas o parciales. Las medidas difieren además por la forma en que ellas se afectan por otros factores, por ejemplo los totales marginales. Así, hay medidas que son muy "sensibles a los marginales", en el sentido que su valor está muy influenciado por la distribución de los totales marginales de filas y columnas. Tales medidas reflejan por tanto información sobre los marginales, además de sobre la asociación.

Una medida de asociación determinada puede tener un valor bajo para una tabla dada, pero para ello no significa que las variables objeto de estudio no estén relacionadas, sino que ellas no están relacionadas en la forma a la que es sensible dicha medida. Por ello ninguna medida individual es la mejor para todas las situaciones. Al seleccionar una, debe tenerse en cuenta el tipo de datos, la hipótesis de interés así como las propiedades de cada medida. A veces se calcula un gran número de medidas y luego se referencian aquellas que más nos "ayudan a respaldar" nuestra hipótesis, como si fueran las únicas que se hubieran calculado; pero esto no es correcto a menos que haya un verdadero análisis de cuáles medidas son las que necesitamos observar.

Veamos primero algunas medidas de asociación para variables discretas en dependencia del tipo de datos a tratar.

Medidas nominales

Consideramos aquí medidas que asumen apenas que las dos variables de la tabla están medidas nominalmente. Como tales, estas medidas pueden suministrar solamente alguna indicación sobre la estrechez de la asociación y no pueden indicar nada sobre la dirección o cualquier otra cosa de la naturaleza de la relación. Las medidas se pueden clasificar en dos tipos: aquellas que están basadas en el Chi-cuadrado y aquellas que se fundamentan en la lógica de reducción proporcional del error (**PRE, Propotional Reduction Error, como se le dice clásicamente).**

Medidas nominales basadas en Chi-cuadrado

Como se explicó antes, el test Chi-cuadrado en sí, no proporciona una buena medida del grado de asociación entre las dos variables; pero como está tan expandido el uso del Chi-cuadrado

en la dócima de independencia, se ha estimulado la definición de medidas de asociación basadas en el Chi-cuadrado tratando de minimizar la influencia del volumen de la muestra y de los grados de libertad, así como restringir el rango de los valores de la medida al intervalo [0,1].

Estas medidas ayudan entonces a comparar los resultados del Chi- cuadrado en tablas diferentes cuando hay variación de las dimensiones y de los volúmenes de las muestras. Sin estas correcciones, es absolutamente inadmisible comparar con el Chi-cuadrado tales tablas.

El llamado coeficiente Phi modifica el Chi-cuadrado dividiéndolo por el volumen de la muestra y extrayendo la raíz cuadrada del resultado:

2

Para tablas en las cuales una dimensión es mayor que 2, phi no yace necesariamente entre 0 y 1, ya que el valor del Chi-cuadrado puede ser más grande que el volumen de la muestra. Por tanto Phi sólo queda estandarizado en el intervalo [0,1] en tablas en las cuales R=2 o C=2 (R y C denotan siempre el número de filas y columnas).

Para obtener una medida que debe hacer siempre entre 0 y 1 Pearson sugirió el uso de:

5----------6

que es conocido como coeficiente de contingencia. El valor de este coeficiente está siempre entre 0 y 1; pero generalmente no puede llegar a ser 1. De hecho, el máximo valor de C depende sobre el número de filas y columnas. Por ejemplo para una tabla 4x4, el máximo valor de C es 0.87. **Crammer** introdujo la siguiente variante:

2

Este estadístico, conocido como V de **Crammer**, puede alcanzar el máximo de 1 en tablas de cualquier dimensión y además coincide con Phi si una de las dimensiones es igual a 2.

Si se piden al SPSS, los estadísticos anteriores en la tabla de contingencia del ejemplo inicial, aparecerá:

Statistic	**Value**	**Significance**
Crammer's V	**.57239**	
Contingency Coefficient	**.82918**	

El **SPSS** normalmente presenta la V de Crammer en tablas de dimensión superior y la Phi en tablas 2x2.

Las medidas nominales basadas en el test Chi-cuadrado son difíciles de interpretar. Entre las tres mencionadas recomendamos la V de Crammer. Su carácter estandarizado permite al menos comparar la "estrechez de la asociación" entre tablas diferentes; pero esta estrechez o fortaleza de la asociación que estamos comparando, no responde a ningún concepto intuitivo claro de asociación. Así, en el ejemplo, el valor absoluto de la V de Crammer poco nos dice; pero si tenemos otra tabla similar que pretenda medir la posible influencia del nivel de la Hipnosis en el resultado del tratamiento de otra enfermedad, digamos Alcoholismo, y disponemos también de la V de Crammer en la tabla de Contingencia HIPNOSIS- ALCOHOLISMO, podemos compararlas para determinar en cuál de las dos enfermedades hay "mayor relación" entre el nivel de cura y el grado de profundidad de la hipnosis.

Medidas nominales de asociación basadas en la lógica de PRE.

La idea de la Reducción Proporcional en el error fue introducida por Goodman y Kruskal (1954). Con este tipo de medidas, el significado de la asociación resulta más claro.

En esencia estas medidas indican relación entre:

- La medida del error al predecir los valores de una variable basándonos en el conocimiento sólo de esa variable.
- La medida del error al predecir los valores de esa variable basándonos en el conocimiento de otra variable adicional.

Para ilustrarlo con un ejemplo vease la Tabla de Contingencia del tratamiento de la migraña por la hipnosis. Si queremos pronosticar el resultado del tratamiento atendiendo sólo a la variable MIGRANA y no tenemos ninguna información adicional (sobre la profundidad de la hipnosis), tendremos que guiarnos por la categoría modal y afirmar que:

- Lo más frecuente es que el tratamiento de hipnosis no produzca cambios en la migraña.

La probabilidad de error de esta conclusión es uno menos la probabilidad de la categoría modal:

P(1) = 1-0.45 = 0.55

Si usamos ahora el conocimiento sobre la profundidad de la hipnosis podemos mejorar este pronóstico. Para cada categoría (nivel) de la hipnosis podemos predecir el resultado del tratamiento por la categoría modal de la varible MIGRANA en esa fila y entonces concluir que:

- La mayoría de los sujetos que se someten a una hipnosis profunda se curan de la migraña.
- Lo más frecuente en aquellos casos que alcanzan una hipnosis media, es lograr cierta mejoría
- Aquellos en los que la hipnosis es ligera, no mejoran nada, en su mayoría.

La probabilidad del error cuando usamos la variable HIPNOSIS para pronosticar MIGRANA es la suma de las probabilidades de todas las celdas que no resultaron modales por filas:

P(2) = 0.05+0.10+0.17+0.01 = 0.33

El estadístico lambda de Goodman y Kruskal con MIGRANA como variable dependiente se calcula como:

l = ---------- = ---------- = 0.40

Ella significa que cuando se usa la profundidad de la hipnosis para predecir el resultado del tratamiento de la migraña se produce una disminución de un 40% en el error de predicción.

Lambda siempre toma valores entre 0 y 1. Un valor de 0 significa que la variable independiente no ayuda en la predicción de la variable dependiente. Un valor de 1 significa que la variable independiente permite pronosticar exactamente las categorías de la dependiente (la perfección ocurre solamente cuando en cada fila hay solamente una celda diferente de cero). Cuando dos variables son independientes, lambda es 0; pero un valor 0 para lambda, no significa que las variables tengan independencia estadística. Como todas las medidas de asociación, lambda se construye para medir un tipo específico de dependencia: la reducción del error cuando una variable es usada para predecir los valores de la otra. En el caso particular que no exista este tipo de asociación, lambda resultará cero. Recuérdese siempre que no existe ninguna medida sensible a todos los tipos imaginables de asociación.

En cada tabla se pueden calcular dos valores de lambda, uno asociado con cada una de las variables como variable dependiente y estos valores no tiene por qué coincidir. en algunas

aplicaciones no es factible o conveniente distinguir cuál de las dos variables es la dependiente o la independiente y de hecho, si se determina la presencia de este tipo de asociación resulta deseable utilizar indistintamente una variable para predecir la otra o viceversa.

Con este objetivo, se puede dar una definición de una lambda "simétrica" para el análisis de estas tablas.

Cuando se le solicita el estadístico lambda al SPSS/PC, el paquete brinda el valor simétrico de lambda así como los dos valores asimétricos correspondientes, y el usuario selecciona cuál o cuáles son de su interés. Véase como saldría en el ejemplo dado.

		With HIPNOSIS	With MIGRANA
Statistic	**Symmetric**	**Dependent**	**Dependent**
Lambda	**.35294**	**.29787**	**.40000**

Medidas Ordinales

Lógicamente, las variables con nivel de medición ordinal, pueden ser analizadas en tablas de contingencia con las medidas de asociación descritas anteriormente para variables simplemente nominales; pero otras medidas reflejan información adicional sobre la base del ranqueo.

Al considerar la clase de relaciones que pueden existir entre dos variables ordinales, arribamos a la noción de dirección de la relación y al concepto de "correlación generalizada". Las variables están positivamente correlacionadas cuando los casos con valores bajos de una de las variables tienen también valores bajos de la otra y los casos con valores altos de una de las variables los tiene también altos en la otra. Dicho en otra palabras, la correlación es positiva cuando la "relación de dependencia tiende estadísticamente a ser creciente". Las variables correlacionadas negativamente muestran el comportamiento inverso: a más altos valores de la primera variable corresponden los más bajos valores de la segunda.

Para tablas con dos variables ordinales hay varias medidas de asociación que se basan en la comparación de los valores de ambas variables en todos los pares posibles de casos u observaciones. Un par de casos se dice "concordante" si los valores de ambas variables para un caso son más altos (o ambos más bajos) que los valores para el otro caso. El par de casos se dice "discordante" si el valor de una variable para un caso es mayor que el correspondiente valor para el otro caso y la dirección está invertida en la segunda variable. Cuando los casos tienen valores id idéntico sobre una o las dos variables se dicen "ligados". De esta forma, para cualquier par de casos con medidas de las variables X e Y, el par puede ser concordante o discordante o ligado, esto último en tres formas posibles: ligado respecto a X pero no de Y, ligado respecto a Y pero no de X y ligado respecto a **X e Y.**

En una tabla de contingencia.

- Son concordantes cualquier par de casos que yazcan sobre las celdas (i,j) y (k,l) si se cumple que: (i<k) y (j<l) o si se cumple que: (i>k) y (j>l). Los pares de casos concordantes tienden a reflejar correlación positiva.
- Son discordantes cualquier par de casos que yazcan sobre las celdas (i,j) y (k,l) si se cumple que: (i<k) y (j>l) o si se cumple que: (i>k) y (j<l). Los pares de casos discordantes tienden a reflejar correlación negativa.
- Son ligados respecto a la primera variable los casos que yazcan sobre las celdas (i,j) y (i,l), esto es, cuando k=i (celdas sobre la misma fila). Ellos tienden a reflejar independencia de la segunda variable respecto a la primera.

- Son ligados respecto a la segunda variable los casos que yazcan sobre las celdas (i,j) y (k,j), esto es, cuando l=j (celdas sobre la misma columna). Ellos tienden a reflejar independencia de la primera variable respecto a la segunda.
- Son ligados completamente los casos que estén en la misma celda. Ellos no reflejan ningún tipo de dependencia.

De esta forma se tiene una medida de la tendencia a la correlación positiva o negativa contando los pares concordantes **(P)** y los pares discordantes **(Q).** Los pares ligados: **T , T , T** pueden ayudar a corregir estas medidas teniendo en cuenta lo que ellos reflejan sobre independencia.

Las siguientes medidas de asociación tienen todas el mismo numerador: P-Q, esto es, la diferencia entre el número de pares concordantes y discordantes, determinado con todos los pares posibles de observaciones. Estas medidas difieren en principio en la forma en la cual se normaliza la diferencia P-Q.

La más simple de las medidas se obtiene dividiendo **P-Q** por el número total de pares. Si no hay pares con ligaduras, esta medida, conocida como Tau-a de Kendall, está en el rango de -1 a 1; pero si hay ligaduras, el rango es más estrecho y depende realmente del número de pares ligados. Como en principio, todas las observaciones sobre la misma fila o la misma columna están ligadas, el estadístico **Tau-a de Kendall** resulta difícil de interpretar.

La medida Tau-b intenta normalizar P-Q considerando las ligas en un par sobre cada variable por separado; pero no las ligas sobre ambas variables:

b

Si ninguna frecuencia marginal es 0, Tau-b puede alcanzar los valores +1 y -1 pero solamente en una tabla cuadrada. Una medida que puede alcanzar estos valores extremos en cualquier tabla de dimensiones **RxC, es la llamada Tau-c:**

_ = ---------- , donde K=min(R,C).

Puede demostrarse que los valores de los coeficientes Tau-b y Tau-c no difieren mucho si las frecuencias de los marginales son aproximadamente iguales. Por ello decimos que la Tau de Kendall son "sensibles a la distribución de los marginales". En general la Tau-c de Kendall es la medida más utilizada para reflejar el concepto de "correlación generalizada". Podría concebirse como una alternativa no paramétrica de las medidas de correlación de variables aleatorias continuas.

Hay otra medida de asociación: la Gamma de Foodman y Kruskal, que está estrechamente ligada a las Tau de Kendall:

g = ---------

Gamma puede ser interpretado como la probabilidad de que un par aleatorio de observaciones sea concordante menos la probabilidad de que el par sea discordante, asumiendo la ausencia de ligaduras. El valor absoluto tiene una interpretación además en la lógica del PRE relacionada con la reducción proporcional en el error en el pronóstico de concordancia o discordancia. **Gamma es 1** si todas las observaciones se concentran en la diagonal principal de una tabla.

En el caso de independencia, Gamma es 0. Sin embargo, lo recíproco, que Gamma sea 0 implique independencia, no es cierto, salvo en **tablas 2x2.**

En el cálculo de Gamma, no se hace distinción entre las variables dependiente o independiente. Somers (1962) propuso una variante antisimétrica de Gamma que difiere

solamente de ella en que incluye el número de pares ligados por la variable dependiente y no por la variable independiente. La D de Somers se define por:

d = ----------

Este coeficiente indica el exceso de proporción entre pares concordantes y discordantes, en relación con pares no ligados sobre la variable independiente. La D de **Somers** se puede calcular considerando cada una de las variables como variable independiente y puede darse además una variante simétrica, que utiliza en el denominador el valor medio de los denominadores de los dos coeficientes asimétricos.

Cuando se solicitan al SPSS/PC estos estadísticos para el ejemplo de la **HIPNOSIS** contra **MIGRANA** aparece algo así:

		With HIPNOSIS	With MIGRANA
Statistic	**Symmetri**	**Dependent**	**Dependent**
Sommers' D	**.67886**	**.65774**	**.70096**

Statistic	**Value**	**Significance**
Kendall's Tau B	**.67901**	**.00000**
Kendall's Tau C	**.63360**	**.00000**
Gamma	**.94034**	

Aquí un par de casos es concordante si siendo diferentes los niveles de hipnosis logrados, la efectividad del tratamiento resulta en diferencias consecuentes. Un par de sujetos es discordante si siendo los niveles de hipnosis diferentes, la efectividad del tratamiento resulta en diferencias inversas a lo esperado. Están ligados respecto a la variable dependiente MIGRAÑA y no respecto a HIPNOSIS aquellos pares de sujetos en los cuales se logra el mismo resultado del tratamiento a pesar de que los niveles de hipnosis logrados fueron diferentes, etc. Pueden considerarse pares de sujetos no ligados respecto a HIPNOSIS todos aquellos que alcanzaron diferentes niveles de hipnosis en el tratamiento, independientemente del resultado.

En este ejemplo, las Taus de Kendall reflejan que predominan los pares concordantes significativamente (el test de hipótesis correspondiente mide la significación del la diferencia con cero de estos coeficientes). Podemos afirmar sin lugar a dudas de ser significativo, que a mayor profundidad en el nivel de hipnosis, corresponde mayor efectividad en el tratamiento. Gamma confirma lo anterior asegurando que el comportamiento de la mayoría de los sujetos los ubica sobre la diagonal de la tabla y la D de Somers con MIGRAÑA como variable dependiente, significa que el exceso de proporción de pares con niveles concordantes de hipnosis- efectividad sobre pares discordantes en hipnosis-efectividad en relación con los pares que no tuvieron el mismo nivel de hipnosis es el 70% .

4.3. Medidas de correlación entre variables aleatorias continuas o entre continuas y discretas

En el caso de variables aleatorias continuas no tiene sentido el análisis de tablas de contingencia ni las pruebas Chi-Cuadrado. Claro que las variables se pueden discretizar en rangos, y entonces hacer para ellas una tabla de contingencia; pero comoquiera que estamos perdiendo mucha información, las medidas de asociación propias de variables

aleatorias discretas, brindan menos información que las que podemos obtener directamente de las variables sin ranquear. El resto del curso trata de formalizar formas de estudio de la dependencia en niveles superiores al discreto. Aquí sólo hacemos una presentación para relacionar con los conceptos de tablas de contingencia.

Medidas de correlación entre variables aleatorias continúas

En estos casos es necesario prestar atención a otras magnitudes que reflejan la fortaleza de la correlación atendiendo al hecho mismo de que las variables tienen nivel de medición continuo. Un estadístico asimétrico que es útil para medir en particular la fortaleza de un enlace de tipo lineal entre las variables es el coeficiente de correlación lineal de Pearson. Su interpretación, o mejor sus diferentes interpretaciones, están ligadas a la teoría de regresión lineal que será objeto de estudio detallado en otro tema. Aquí sólo esbozamos lo esencial.

Suponga que tenemos una muestra **(x ,y) , i=1,2,....,n** del par de variables aleatorias **X, Y,** y suponga que se investiga la existencia de una dependencia lineal de la forma:

Y = B +B X

en el sentido en que cada valor de la muestra se pueda representar:

y = B + B x + e

Aquí e denota un error aleatorio que se supone sujeto a ciertas condiciones (e depende de x pero para cada i tiene una distribución normal con esperanza 0 y cierta dispersión independiente de x).

Una formulación un poco más general supone una muestra **(x ,y), i= 1,2,...,n, j=1,2,....,m** en la que a cada valor de X pueden corresponder varios valores de Y y entonces se busca la dependencia lineal de la forma:

M(Y | X) = B + B X

y = B + B x + e

i

donde **M(Y ¦ X)** es el valor esperado de la variable Y para cada valor fijado de X y y denota la media muestral de los valores y que corresponden al valor **X=x .**

El método de los mínimos cuadrados sirve para determinar **B y B** a partir de los datos de la muestra de manera que sea mínima la suma de los cuadrados de las distancias de los valores observados a los pronosticados por la ecuación lineal de la regresión y en particular resulta como coeficiente de la regresión:

i j

Si B = 0, esto significa que no hay una dependencia lineal de Y respecto a X; pero si B es diferente de cero, su valor no es, en modo alguno, una medida de la estrechez de la relación. De hecho, geométricamente hablando, **B** apenas mide la pendiente de la recta de regresión, y físicamente hablando, el valor de **B** depende del sistema de unidades escogido para **X e Y.**

Para resolver estos inconvenientes, se utiliza escribir la recta de regresión en la forma estandarizada:

x

Donde

R = --------- B = ----------

Que se conoce como coeficiente de correlación de Pearson.

A diferencia de B , coeficiente de Y respecto a X, el coeficiente de correlación R es simétrico y adimensional, esto es, llegamos al mismo resultado si en lugar de partir de una dependencia lineal de Y respecto a X partimos de una dependencia lineal de X respecto a Y o si hacemos cualquier transformación lineal de las variables simulando un cambio de escala.

La principal interpretación de R, resulta de lo siguiente. Puede demostrarse que el coeficiente de correlación es una magnitud que siempre está en el intervalo [-1,1]. Si R=±1, todos los puntos de la muestra se encuentran exactamente sobre una recta. Si R>0 la recta es creciente, si R<0, la recta es decreciente y en la medida en que el valor absoluto de R se acerque a 1, la dispersión de los valores de y alrededor de las medias condicionales para cada x es menor. De esta forma el coeficiente de correlación nos da una medida de la fortaleza del enlace lineal entre X e Y.

Normalmente, junto con el cálculo de R a partir de la muestra, se testa la hipótesis:

H : R = 0 contra

H : R $ 0 (o contra H : R > 0)

Para ello, se utiliza un estadígrafo **T** que se define por la relación entre el coeficiente **B** y su dispersión y que tiene distribución **T de Student con N-2** grados de libertad.

En la teoría de la regresión, se da otra interpretación de R como medida de la bondad del ajuste de la regresión. Puede ser probado en definitiva que la variabilidad de variable dependiente puede ser descompuesta en dos sumas: aquellas atribuibles a la regresión y aquellas que no lo son:

$$(y - y) = (y - y) + (y - y)$$

Aquí y = B + B x es el valor pronosticado por la regresión, y su diferencia con el valor observado y es el error residual, en tanto su diferencia con la media y puede considerarse el error de la regresión.

Resulta:

^{2} **suma de cuad residuales suma de cuad regresión**

Por tanto, R es una medida de la parte "explicada" de la variación total por el modelo de regresión.

Se suele utilizar un test de Fischer Snedecor para medir cuán significativamente pequeño o grande es este tipo de relaciones; pero ello será objeto de estudio en el tema de análisis de varianza. Baste decir por el momento que en el caso del estudio de correlación de dos variables (una "independiente" y una "dependiente", el test es equivalente al T de Student para medirla significación del coeficiente de correlación y en casos de más de una variable independiente (regresión lineal múltiple), se pueden medir coeficientes de correlación "parciales" respecto a cada una de las variables y para cada uno de ellos, su significación; pero el análisis de varianza anterior es la única forma de medir el ajuste del modelo de regresión en general, o la dependencia lineal de la variable dependiente respecto al conjunto de variables independientes.

Uno de los errores más frecuentes por la interpretación simplista del coeficiente de correlación, es la de afirmar que las variables son independientes cuando el coeficiente de correlación es cero o cercano a cero. El coeficiente R mide la fortaleza o estrechez del enlace lineal entre variables. El hecho de que R sea 0 o cercano a este valor significa que no hay dependencia lineal entre las variables; pero no significa que no haya cualquier otro tipo de dependencia. Sólo en el caso en que sepamos que el par de variables (X,Y) tiene una distribución normal bi-variada, la independencia lineal y la independencia en general significan una misma cosa. Pero si esperamos una relación por ejemplo, de la forma

Y = X

y X tiene distribución normal, no podemos esperar distribución normal para Y. De hecho, si tal relación existe entre las variables, el coeficiente de correlación será cero, porque no hay dependencia lineal entre las variables, sino cuadrática.

El análisis de varianza que utilizamos para estudiar que parte de la dispersión de la variable dependiente era explicada por el modelo de regresión lineal y que dió origen a la otra interpretación de R, puede ser extendido a cualquier otro tipo de regresión no necesariamente lineal, con sólo cambiar la fórmula para el cálculo de los estimados y . En lugar de la expresión **B +B x** , podremos determinar y por la expresión concreta de la

regresión buscada y de esta forma obtenemos un test para medir la bondad del ajuste, aun cuando no tengamos un "coeficiente de correlación para este tipo de regresión en particular".

Surge naturalmente la pregunta de si será posible definir un coeficiente que mida en algún sentido si hay cualquier tipo de dependencia funcional (no necesariamente lineal) entre X e **Y**.

Realmente, si en el análisis anterior, sustituimos el valor estimado y por la medida condicionada a x : y , obtenemos una expresión similar:

t i t x t i x

Obtenemos así un coeficiente:

² disp dentro de grupos disp entre grupos

Si existe una dependencia de correlación de Y respecto a **X,** la dispersión de las réplicas (dispersión entre grupos) debe ser mucho menor que la dispersión entre los grupos y el valor de n debe acercarse a **1.** Si n se acerca a **0,** la dispersión de los grupos es significativamente grande y ello es una medida de que no hay ninguna dependencia de correlación.

El coeficiente n (sería mejor denotarlo **n** para enfatizar que estamos hablando de **X** como variable independiente y Y como variable dependiente) se determina siempre en forma positiva (la raíz cuadrada positiva de n), se denomina Relación de correlación y es una medida de si hay alguna dependencia funcional de Y respecto a X.

Puede verificarse que n es mayor o igual que el módulo de **R.** Por tanto si **R=±1,** resultará **n=1. Si n=0**, resultará **R=0.** Los recíprocos de estas afirmaciones no son ciertos. la relación de correlación tiene la ventaja de que es una medida de la estrechez de cualquier tipo de dependencia, en tanto R mide solamente la fortaleza de los enlaces de correlación lineal. La gran desventaja de n es que nada dice sobre la forma de la dependencia, sólo, que la hay en el sentido estadístico.

Atiéndase al hecho particular de que si no se disponen de réplicas, esto es si en la muestra, para cada valor de X hay sólo valor de Y, n resultará 1; pero realmente esto poco nos ayuda.

Todos los paquetes estadísticos traen forma de determinar el coeficiente de correlación y su significación, y muchos de ellos incluyen también la posibilidad de calcular la relación de correlación **n.** El procedimiento **CROSSTABS** del **SPSS** puede utilizarse también con este objetivo. En este caso (se supone que estamos tratando variables aleatorias continuas) puede pedirse la omisión de las tablas y no solicitar ninguno de los estadísticos propios de variables aleatorias discretas, limitándonos al cálculo del coeficiente de correlación de Pearson y la relación **n**.

El **SPSS** tiene además un procedimiento: **CORRELATIONS** que permite confeccionar una "matriz de correlaciones" entre pares de variables de una lista bastante amplia. Se conforma así una tabla de doble entrada con los nombres de las variables y en cada celda se ilustra el coeficiente de correlación entre la variable especificada por esa fila y esa columna y su significación.

CORRELATIONS es un buen procedimiento para ejecutar antes de hacer cualquier análisis (análisis de regresión, análisis de varianza, análisis discriminante) con un alto número de variables.

Medidas de asociación o dependencia de variables continuas respecto a variables discretas.

La relación de correlación n suele usarse también para determinar posible dependencia de una variable aleatoria continua respecto a una ordinal, con tal que ésta tenga un número no demasiado pequeño de valores. Si la variable ordinal tiene un número bastante grande de valores, de hecho puede tratarse estadísticamente como continua y ser sometida prácticamente a toda la teoría de regresión .

Para medir la asociación entre variables aleatorias continuas y variables aleatorias discretas, nominales u ordinales con un número reducido de valores se utiliza el Análisis de Varianza que como hemos prometido, será objeto de estudio en otra conferencia.

El análisis de regresión puede incluir también el tratamiento de variables aleatorias discretas, en una forma completamente equivalente al análisis de varianza. Por ello en muchos sentidos, esta conferencia es un antecedente de estos dos tópicos: el Análisis de Varianza y el Análisis de Regresión .

4.4. El procedimiento CROSSTABS del SPSS

El procedimiento CROSSTABS del SPSS/PC sirve para obtener tablas de contingencia y las diferentes medidas de asociación de que hemos hablado. Es un procedimiento propio de análisis de independencia y de cálculo de medidas de asociación entre variables aleatorias discretas, de nivel nominal u ordinal; aunque por razones de generalidad incluye las medidas de asociación propias de variables continuas como el coeficiente de correlación y la relación de correlación.

Su sintaxis más usual es:

CROSSTABS

varlist BY varlist /

OPTIONS = number /

STATISTICS =number .

Este comando producirá una tabla de contingencia para cada variable incluída en la lista antes del **"BY"** con cada variable de la lista después del **"BY"** . Las variables que aparecen a la izquierda, conformarán las filas de la tabla y las de la derecha, la columnas de la tabla.

OPTIONS es un subcomando que determina el contenido de las celdas y algunos aspectos del formato. Por defecto la tabla se produce sólo con las cantidades. Si en OPTIONS se dán por ejemplo, los números 3, 4 y 5, las celdas incluirán además los porcentajes de fila, los porcentajes de columna y los porcentajes totales. Otras opciones permiten incluir en la celda las frecuencias esperadas, los residuales, los residuales cuadráticos y relativos, etc. Otro grupo de opciones permite modificar el formato como tal de la tabla (evitar los valores de

las variables, por ejemplo). Todas estas opciones pueden consultarse prácticamente en la ayuda.

STATISTICS es un subcomando que permite añadir al final de la tabla los estadísticos (medidas de asociación) deseados, y en algunos casos, se incluye su significación. Por defecto, se produce sólo la tabla. Si se incluye por ejemplo los números 1 2 y 4, además del Chi-cuadrado se incluirán la V de Crammer y la l de Godman y Kruskal. Otros estadísticos pueden consultarse en la ayuda.

Por ejemplo, la tabla del inicio del capítulo y otra similar de Hipnosis vs Alcoholismo se podrían producir con un comando como este:

```
CROSSTABS
   HIPNOSIS BY MIGRANA ALCOHOL /
   OPTIONS=3 4 5 /
   STATISTICS=1 2 4 .
```

El comando CROSSTABS permite especificar más de una tabla con diferentes subcomandos de especificación de variables. Permite además rupturas de tablas por variables de "control" o "estratos", especificadas después de un BY adicional. Por ejemplo:

```
CROSSTABS
   HIPNOSIS BY MIGRANA /
   HIPNOSIS BY MIGRANA BY SEXO /
   OPTIONS=3 4 5 /
   STATISTICS=1 2 4 .
```

Este comando producirá tres tablas de contingencia con el mismo formato y los mismos estadísticos. La primera tabla será una tabla con todos los datos de la muestra. El segundo subcomando produce dos nuevas tablas, supuestamente una para el sexo 1 y otra para el sexo **2.**

Por último, la práctica hace que a veces tengamos los datos no a nivel de sujetos, sino ya tabulados. Normalmente esto constituye una pérdida de considerable información, porque nos dificulta hacer estudios de correlación de las variables que aparecen en tablas diferentes. No obstante, podemos aun así, calcular el chi- cuadrado y las restantes medidas de asociación correspondientes a cada tabla por separado; Para calcular los estadísticos con el SPSS/PC hay que "introducir la tabla" con ayuda del comando WEIGHT. El siguiente programa muestra cómo hacerlo en el caso de nuestra tabla concreta.

```
SET
   DISK=ON /
   DISK='MIGRANA.LST' /
   ECHO=OFF /
   SCREEN=OFF /
   EJECT=OFF .
DATA LIST FREE /
   HIPNOSIS
   MIGRANA
```

```
    CASOS .
VARIABLE LABEL
    HIPNOSIS 'NIVEL LOGRADO'
    MIGRANA 'NIVEL DE CURA ALCANZADO'.
VALUE LABEL
    HIPNOSIS
        1 'Profunda'
        2 'Media'
        3 'Ligera' /
    MIGRANA
        1 'Curada'
        2 'Mejoría'
        3 'Sin cambios' .
* Observe la introducción de la tabla.
BEGIN DATA
1   1   13
1   2   5
1   3   0
2   1   10
2   2   26
2   3   17
3   1   0
3   2   1
3   3   28
END DATA .
WEIGHT BY CASOS.
CROSSTABS
    HIPNOSIS BY MIGRANA /
    OPTIONS=3 4 5 /
    STATISTICS=1 2 4 .
FINISH.
```

4.5 Conclusiones

Las tablas de contingencia constituyen una forma de medir la independencia entre dos variables aleatorias discretas y calcular ciertas medidas de asociación. El estadístico básico para la prueba de independencia es el test Chi-cuadrado; pero tiene tantas limitaciones que

debe ser utilizado con mucho rigor y completado con otras medidas de asociación. Existen múltiples medidas que miden la asociación en forma diferente y ninguna de estas medidas es por sí sola buena para todos los casos.

El procedimiento CROSSTABS del SPSS/PC permite hacer tablas de contingencia y calcular los estadígrafos más universales de asociación entre variables aleatorias discretas.

Los capítulos siguientes tienen como objetivo el estudio de la dependencia entre variables aleatorias cuando una o varias de ellas tiene un nivel de medición superior al discreto.

4.6 Ejercicios del Capítulo IV

1. Recuerde el ejemplo de la investigación sobre vinos y realice tablas de contingencia entre las variables CALIDAD PREMIO y DISPONIBILIDAD. Determine en particular si el PREMIO depende más fuertemente de la popularidad (dada por la disponibilidad) o de la calidad de los vinos.

2. En una investigación psicológica se analizaron dos grupos de niños: uno compuesto de casos con disfunción cerebral mínima y otro grupo de niños sanos (grupo de control). En ambos grupos se midió la capacidad de concentración del niño y la organicidad arrojando los siguientes resultados:

ENFERMOS 83 49 30 ENFERMOS 80 5² 30

Determine si hay diferencias entre los grupos desde el punto de vista de estas variables y cuál de ellas está más fuertemente asociada a la disfunción.

3. En un experimento para probar el efecto de seis drogas diferentes, se contaminó el alimento de 6 grupos de animales independientes y se observaron los sobrevivientes. El resultado fue el siguiente:

SOBREVIVIENTES ²0 43 46 ²4 ²4 35

¿Son las drogas igualmente letales? Cuestiónese el test.

4. En una investigación de la posible influencia del hábito de fumar en las embarazadas sobre el peso del niño, resultó que de 59 madres fumadoras, 17 tuvieron niños de bajo peso, mientras que de 87 madres no fumadoras, 7 tuvieron niños de bajo peso. Verifique si tal diferencia puede apreciarse por casualidad en menos de una vez en 100 veces.

5. Una serie de observaciones sobre la frecuencia de aparición de la Hepatitis B en cuatro áreas independientes da los siguientes resultados:

FRECUENCIA DE HEPATITIS 19 10 8 5

Pruebe que la diferencia en frecuencia entre las cuatro áreas es significativa al nivel del 1%. ¿Es cuestionable el test?

6. En una investigación sobre acúfenos (ruidos en el oído) se someten grupos de pacientes a tratamientos diferentes y se quiere confirmar ante todo que la combinación del tratamiento médico y psicológico produce las mayores esperanzas de mejoría. Los datos fueron los siguientes:

Mejoría del acúfeno.

+----------+

Grupo	p	Ninguna	Alguna	Bastante	p
p Control	p	30	4	0	p
p Psicológico	p	3	13	4	p
p Médico	p	14	24	20	p
p Psic+Médico	p	7	11	45	p

Demuestre la hipótesis.

7. En la misma investigación del punto 6 quiso investigarse la posible relación entre la mejoría y varios resultados del análisis pisológico: concretamente la existencia de alucinaciones auditivas, la tristeza del paciente, y la sobrecarga laboral (en los casos en que estos trabajaban). Las proporciones de cada una de las variables fueron las siguientes.

	Mejoría del acúfeno.		
	Ninguna	Alguna	Bastante
Casos con:			
Alucinaciones	3	4	0
Tristeza	29	24	19
Total de casos	54	54	72
Casos con:			
Stress por sobre-carga laboral	12	10	5
Total de casos que trabajan	44	36	55

Determine si la esperanza de mejoría del acúfeno es independiente de la presencia de alucinaciones auditivas en el paciente.

Analice la dependencia (o independencia) de la esperanza de mejoría respecto a la presencia de tristeza y respecto al stress por sobrecarga laboral. Determine en última instancia, entre estos tres factores psicológicos, cuáles tienen mayor repercusión en la posible mejoría del acúfeno y cuáles tienen mayores posibilidades de servir para su pronóstico.

8. Regrese al ejemplo de los vinos y suponga que usted quiere sacar alguna conclusión sobre la diferencia de calidad por provincias; pero una tabla de contingencia de CALIDAD vs PROVINCI tendría una dimensión excesivamnente alta. ¿Cómo se resuelve este problema?

9. Se quiere demostrar que el peso del recién nacido está influído por el nivel de hemoglobina de la madre durante el embarazo y para ello se muestra la siguiente tabla de contingencia:

Peso del recién nacido

p p p

Demuestre que la asociación en estos niveles es altamente significativa. Argumente incluso tendencias a la linealidad del peso del recién nacido (en categorías) respecto a niveles de hemoglobina; pero analice finalmente el nivel de medición de las variables y critique el procesamiento estadístico.

10. Suponga que se ensaya el tratamiento con acupuntura ante un tipo de enfermedad reumática; pero es necesario determinar la frecuencia e intensidad de las descargas eléctricas. Estas variables, calificadas en tres niveles clásicos arrojaron los siguientes resultados durante la experimentación:

+----------+----------¦ +----------+----------¦

Extraiga las conclusiones que usted pueda de los resultados así tabulados. Incluya análisis de posible tendencia lineal entre la cura y los niveles de intensidad y frecuencia. Critique rigurosamente este análisis estadístico en función de los objetivos.

REFERENCIAS BIBLIOGRÁFICAS

Amón., J. (1980). *Estadística para Psicólogos: 1 Estadística Descriptiva.* Madrid: Pirámide.

Badii, M., Guillen, A., & Abreu, J. (Agosto 2014). *Tamaño Óptimo de Muestra en Ciencias Sociales y Naturales Optimal Simple Size (OSS) in Social and Natural Sciences. International Journal of Good Conscience. 9(2)41-51.* ISSN 1870-557X.

Badii, M., Guillen, A., Lugo Serrato, O., & Aguilar Garnica, J. (s.f.). *Correlación no-paramétrica y su aplicación en la investigaciones científica non-parametric correlation and its application in scientific research.* International Journal of Good Conscience.

Beltrán, J., & Peris, M. J. (2013). *Introducció a l'estadística aplicada a les ciències socials. Servei de Publicacions de la UJI .* Collecció Sapientia.

Biosca, A., Espinet, M. J., Fandos, M. J., Jimeno, M., & Villagrà, J. (1999). *Matemáticas aplicadas a las Ciencias Sociales II.* Barcelona: Edebé.

Cazau, P. (2006). *Introducción a la investigación en ciencias sociales. Tercera Edición.* Buenos Aires: Marzo.

Colera, J., García, R., & Oliveira, M. J. (2003). *Matemàtiques aplicades a les Ciències Socials. .* Madrid: Anaya.

Elorza, H. (2007). *Estadística para las ciencias sociales, del comportamiento y de la salud.* México: CENGAGE Learning.

Escudero Vallés, R. (1994). *Métodos estadísticos aplicados a la economía.* Ariel Economía.

Field, A. (2009). *Discovering Statistics Using SPSS for Windows.* Sage, London: Third Edition.

García Ferrando, M. (1997). *Socioestadística.* Madrid: Alianza Editorial.

Glass, G., & J., S. (1980). *Métodos Estadísticos aplicados a las Ciencias Sociales.* Madrid: Prentice Hall.

Gracia, F., Mateu, J., & Vindel, P. (1997). *Problemas de Probabilidad y Estadística.* Valencia: Tilde.

Ibáñez, M. V., & Simó, A. (2002). *Apuntes de Estadística para Ciencias Empresariales.* Castellón: UJI.

kim, J. O., & Mueller, C. W. (1978). *Factor Analysis: Statistical Methods and Practical Issues from the Series: Quantitative Applications in the Social Sciences.* London: Sage University Paper.

Krathwohl, D. R. (1998). *Methods of Educational and Social Science Research: An Integrated Approach: M Waveland Press, Inc. (Second Edition).*

Marascuilo, L., & Serlin, R. (1988). *Statistical Methods for the Social and Behavioral Sciences. W.H. Freeman and Company.* Nueva York.

Marini, Elisabetha, Racugno, Walter, Tarli, B., & M., S. (1999). *Univariate estimates of sexual dimorphism: the effects of intrasexual variability.* American Journal of Physical Anthropology.

Martín Pliego, J. (2004). *Introducción a la Estadística Económica y Empresarial.* AC. Colección Plan Nuevo.

Pérez Juste, R., García Llamas, J., Gil Pascual, J., & Galán González, A. (2009). *Estadística aplicada a la Educación.* Madrid.: UNED - Pearson.

Pérez López, C. (200). (2004). Técnicas de análisis multivariante de datos con SPSS. Madrid. Pearson.

Pérez López, C. (2004). *Técnicas de análisis multivariante de datos con SPSS.* Madrid. : Pearson.

Pett, M. (1997). *Nonparametric statistics for health care research.* . Thousand Oaks: Sage Publications Inc.

Pick, S., & López, A. L. (1994). *Cómo investigar en ciencias sociales. 5ª ed.* México.: Trillas.

Pino, J. (2017). *Conceptos básicos de estadística para las ciencias sociales.* Ecuador: Charito.

Quivy, R., & Van Campenhoudt, L. (2000). *Manual de investigación en ciencias sociales.* México: Noriega.

Raquel., B. (2004). La hipótesis en estudios cualitativos. El caso de la inducción analítica en una investigación sobre adolescencia. *Pilquen.*

Ritzer, F. J. (2003). *Estadística para las Ciencias Sociales.* México: McGraw-Hill.

Solanas, A. (2002). *La Enseñanza de la Estadística en las Ciencias del Comportamiento a Inicios del Siglo XXI. Metodología de las Ciencias del Comportamiento 4.* no. 2, 157- 183.

Spinelli, J., & Stephens., M. (1997). *Cramér-von mises tests of fit for the Poisson distribution.* Can. Jour. Statisti.

T., R. M., Mateo, M., F., R. D., & M., R. (1991). *Estadística Aplicada a las Ciencias Sociales: Teoría y Ejercicios (EAC).* Málaga: Secretariado de Publicaciones de la Universidad de Málaga.

Valles, M. S. (2000). *Técnicas cualitativas de investigación social.* Madrid.: Síntesis.

Wackerly, D. D., III, W. M., & Sheaffer., R. L. (2000). *Estadística matemáticas con aplicaciones, 6a ed.* México.: Biblioteca de Matemáticas, Thomson.

Ciencias y Letras

3ciencias