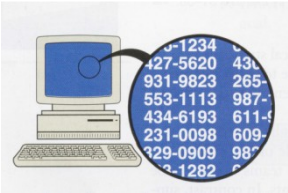
CAPÍTULO 4

Este tema es ya un tema con parte practica donde se comenzarán a ver cálculos que el alumno debe saber realizar. Antes de ponernos a trabajar con la estadística descriptiva vamos a ver como escoger la muestra para poder trabajar con ella en las páginas siguientes.
En general podemos hablar de dos grandes tipos de muestreo :

Muestreo Probabilísticos: son aquellos métodos para los que se puede calcular la probabilidad de extracción de cualquiera de las muestras posibles. Los casos seleccionados son elegidos con criterios tales que permitan la generalización a toda la población de los resultados obtenidos al estudiar la muestra.

No probabilísticos: son aquellos métodos en los que no se puede calcular la probabilidad de extracción de una determinada muestra. Este tipo de muestreo es también llamado muestreo intencional y por desgracia es muy habitual sobretodo si tenemos en cuenta que muchas veces contamos con recursos de tiempo y dinero limitados y la necesidad de poseer consentimiento de los sujetos de estudio dificulta mucho el encontrar la muestra adecuada por lo tanto en muchas ocasiones se ecoge de manera intencional a dedo.

Por otra parte a la hora del modo de escoger estas muestras podemos encontrar diversos subtipos dentro del ámbito del muestreo probabilistico entre ellos destacarán :
Muestreo Aleatorio Simple: Cada muestra de un mismo tamaño tiene las mismas posibilidades de ser seleccionada.



Muestreo Sistemático: Selecciona uno de cada k elementos de la población, al ordenarlos según algún criterio no relacionado con las características que se desean estudiar.

Muestreo estratificado: Este tipo de muestreo no es excluyente de los anteriores esto quiere decir que yo puedo realizar a la vez un muestreo sistemático y estratificado. El escoger este tipo de muestreo depende del tipo de estudio que este realizando. En el hay una característica en la población que deseo aislar a la hora de trabajar, para separar las muestras según la posesión o no de dicha característica o el grado que la posean . Este procedimiento divide la población en grupos o estratos homogéneos con respecto a la característica de interés, y toma una muestra aleatoria simple dentro de cada uno de estos estratos.

Muestreo por conglomerados: Como el anterior no es un muestreo excluyente se puede realizar a la vez que otro tipo de muestreo , y es utilizado sobretodo cuando contamos con una población muy extensa que nos dificultará incluso el introducir sus censos en la base de datos que escogerá la muestra. Este procedimiento trata de dividir la población en grupos o conglomerados heterogéneos con respecto a la característica de interés, vamos

que se busca dividir en grupos diferentes no separandoles por la característica como en el caso anterior, donde se asume que esta heterogeneidad es similar a la de la población completa. Toma una muestra aleatoria simple de conglomerados observando luego todos los elementos contenidos en este.

Una vez que ya hemos escogido la muestra y extraido toda la información se ha de organizar y después se realizará un análisis exploratorio de los datos, primero el investigador observa la realidad, y busca algún patrón alguna constante, algo que se repita y que explique el comportamiento que nos interesaba observar.
Hoy en dia la organización de datos trata de introducirlos en una base de datos informática y comprobar si he cometido algún error al introducirlos. Pero comenzaremos por el principio.

4.2 DE LA DEFINICIÓN DEL PROBLEMA Y LAS VARIABLES A LA OBSERVACIÓN Y RECOGIDA DE DATOS

Uno de los errores más frecuentes en la utilización de la estadística en el ámbito educativo es recoger datos sin seguir los pasos minunciosamente, del proceso de investigación. Una vez realizado una revisión bibliografica e identificado las variables que se investigarán se procederá a la recogida de la información o sea se le asigna un valor a las variables a estudiar. Siempre tendremos que codificar los valores o sea a una realidad no cuantificada asociarla a un número en concreto para poder trabajar con ella, en este sentido el escoger el instrumento adecuado puede ser fundamental porque en muchos casos será el instrumento quien realice este trabajo a la vez de recoger la información, por ejemplo un test para calcular el cociente intelectual.

4.3 PERMISOS Y ÉTICA EN LA INVESTIGACIÓN Y RECOGIDA DE DATOS.

Hoy en día y sobretodo en nuestro ámbito educativo es fundamental no solo realizar los cálculos bien sino cubrir unos mínimos éticos. Sobretodo por comportamientos como los estudios realizados sobre las personas por la alemania nazi. En este sentido os aconsejo ver una película llamada “the experiment” del 2010. Estos principios éticos fueron creados por la ciencia médica tras la segunda guerra mundial y el ámbito educativo comparte estos principios éticos que deben seguirse, entre los que destacan la participación voluntaria sea del ámbito que sea y el informe de consentimiento. Se debe evitar todo riesgo de daño físico o psíquico , guardando la confidencialidad y el anonimato.
Los aspectos éticos deben ser tenidos en cuenta al comienzo de la investigación y los estandares éticos que se deben seguir se encontrarán en el AERA American Educational Research Association.

4.4. DE LOS INSTUMENTOS A LOS DATOS: ELECCIÓN DEL PROGRAMA LA MATRIZ DE DATOS Y EL LIBRO DE CÓDIGOS

Una vez que hemos recogido la información debemos trasladar los datos a una hoja de cálculo o un software para trabajarlo. A este proceso se le llama tabular datos. Nosotros trabajaremos con una tablita de toda la vida. Donde los valores a estudiar se llamarán puntuaciones directas Xi .

La codificación de datos se hará en ese mismo momento y consiste en darle un valor numérico a una característica que no la tiene. Este proceso es esencial dado que dependiendo del nivel de medida de nuestra hoja de cálculos se realizará un cálculo u otro. Este hecho puede llegar a confundir a una persona que mire la tabla y no esté familiarizado con la codificación empleada , por lo tanto para solucionar este inconveniente se ha creado el libro de códigos.
El libro de códigos es una tabla donde se explican las variables del estudio el orden en el que se introducen en la matriz, vamos que a la hora de colocarlas no puedo hacerlo al tun tun , tengo que poner cada fila dependiento del orden de las columnas de la matriz. Este libro de códigos poseerá tres columnas En la primera columna o columna llamada etiqueta , se debe identificar el item y el nombre de la variable y su etiqueta, por ejemplo si hablasemos del agua hablaríamos de H2O. Despues se le asigna una columna nueva que es la columna etiqueta descripción . En la tercera columna la del código se le asignará una etiqueta de valor a dicho valor. El libro de códigos busca evitar errores de tabulación,

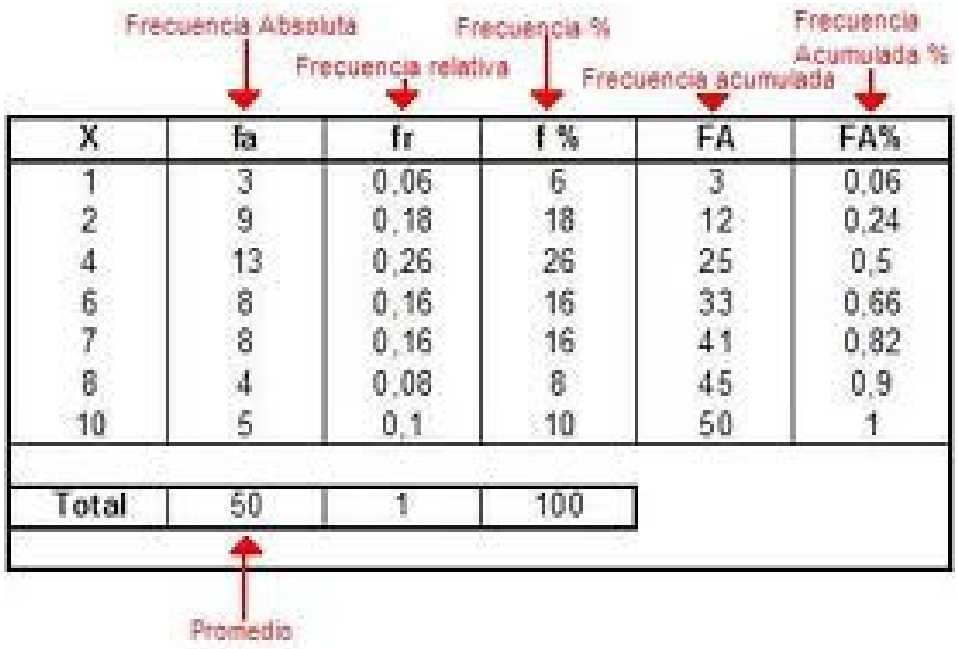
Ítem	Variable	Etiqueta variable	Código	Etiqueta valores
C.1	C1Ident	Identificación	Asignar un valor numérico a cada sujeto (por ejemplo, 01; 02; 03, ...).	—
C.2	C2Sexo	Sexo	0 1	Hombre Mujer
C.3	C3C_aut	Comunidad Autónoma	1 2 3 4 5	Galicia Extremadura Andalucía Madrid Castilla y León
C.4	C4CI	Cociente intelectual	Cualquier valor entre 50-150	—
C.5	C5Rmat	Rendimiento matemático	Cualquier valor entre 0-100	—
C.6	C6Satis	Satisfacción con el curso	1 2 3 4 5	Muy insatisfecho Muy satisfecho

Otro punto clave son los datos perdidos o missing data, son aquellos datos que faltan ya sea porque un individuo no ha contestado o por el contrario que se haya perdido una información . En estos casos se debe dejar en blanco la matriz de datos, porque si se pusiese un cero estaríamos tomandolo como un valor real. Otra opción es definir el missing data por medio de un numero fuera de rango, esta opción se utiliza cuando nos interesa saber cual es el motivo que ha llevado a que el valor esté en blanco. .
La forma habitual de introducir los datos es por medio de una matriz de datos. Es una tabla donde las filas representan a un sujeto y las columnas a las variables a estudiar. .
El problema de la matriz es que a simple vista es dificil de entender sobretodo cuantas mas variables y sujetos trabajemos.

4.5. ORGANIZACIÓN DE LOS DATOS : DEPURACIÓN DE DATOS Y DISTRIBUCCIONES DE FRECUENCIAS.

El problema que nos proboca la matriz es que es dificil de entender, por lo tanto en muchas ocasiones tenemos el problema de clasificación , por lo que antes de empezsr a trabajar con ellos debemos hacer una depuración de datos. Esta depuración de datos tiene dos etapas , el control de calidad de la tabulación y la depuración de datos propiamente dicha.
Es un control de calidad , para comprobar que todos los valores se han colocado de manera correcta.
El primer paso es el control de calidad que se basa en comprobar que hemos metido bien los datos en la tabla , para ello seleccionamos al azar unos pocos test o cuestionarios y

comprobamos si hay errores de tabulación . Si encontramos muchos errores se debe repetir toda la tabulación .
La depuración de datos propiamente dicha trata de verificar si hay valores fuere de rango, o sea que hay alores que no estan dentro de los valores posibles. Para ello debemos conocer los valores mínimos y máximos si lo que está tabulado no coincide eso nos indica que hay un error y debemos solucionarlo.
Una vez que ya hemos depurado podemos comenzar a trabajar con ellos para lo cual lo introduciré o en un programa informático o en una tabla de distribución de frecuencias . En donde además de trabajar con ellas y adivinar cosas sobre la muestra nos permitirá hacer otros tipos de depuraciones. . Se ha de realizar una tabla de frecuencas por cada variable (siempre y cuando no utilices un software informático por lo tanto en el trabajo no hará falta realizarlos solo serán necesarias a la hora del exámen para poder realizar los cálculos). En una distribución de frecuencias colocas en una tabla todos los valores obtenidos en una característica las puntuaciones directas. Aunque seamos sinceros estos cálculos cambiarán dependiendo del tipo de variable que trabajemos . Si trabaamos con una variable de intervalo se realizara de una forma y si trabajamos con una variable dicotómica de otra, también se diferenciará si trabajamos con una variable donde los valores se repiten mucho por ejemplo número de hijos o donde no se repiten horas de estudio semanales. Esto lo iremos viendo en las clases prácticas dado que explicarlo de manera teórica sería dificil de entender. .
Solo explicaré la idea general . En la primera columna colocaría todos los valores posibles desde el máximo al mínimo , mientras que justo a su derecha colocaré otra columna que se llamará frecuencia absoluta , en esta columna tendre que decir cuantas veces se repite el valor que esta delante por ejemplo en las notas tenemos 0,1,1,1, en este caso la frecuencia absoluta de 0 es 1 porque se repetirá una vez , pero la frecuencia absoluta de 1 será 3 porque se repetirá tres veces. . También son muy utiizadas las frecuencias relativas esta frecuencia me indica la proporción, y se calcula dividiendo la frecuencia absoluta dividida entre el total de gente. Una vez hecha esta colúmna se hará una columna más que es el porcentaje que trata de multipriccar por cien la columna anterior la de ffrecuenca relativa. Además se calculan las frecuencas acumuladas que tratan de ir sumando los valores de las columnas anteriores , asi la frecuencia absoluta acumulada, se trata de ir sumando los valores dados en la frecuencia absoluta , mientras la frecuencia relativa acumulada , busca ir acumulando todos los valores de la frecuencia relativa. Como se ve en este ejemplo.



X	fa	fr	f %	FA	FA%
1	3	0,06	6	3	0,06
2	9	0,18	18	12	0,24
4	13	0,26	26	25	0,5
6	8	0,16	16	33	0,66
7	8	0,16	16	41	0,82
8	4	0,08	8	45	0,9
10	5	0,1	10	50	1
Total	50	1	100		

EDAD	fi	Fi	hi	hi x 100%	Hi	Hi x 100%
11	4	4	0,13	13	0,13	13
12	14	18	0,47	47	0,60	60
13	10	28	0,33	33	0,93	93
14	2	30	0,07	7	1	100
	30		1	100		

Estas últimas columnas se usan mucho para la construcción de baremos , para interpretar las puntuaciones de los test, donde se reciben el nombre de percentiles. El percentil es el porcentaje de sujetos que deja por debajo de si una puntuaciÓn determinada. Percentil es Pk , es una medida muy Útil para describir una población , nos dice como esta posicionado un valor respecto al total de una muestra $Pk = \frac{(Kn)}{100}$ .Entendemos k como el valor determinado del que queremos conocer el percentil , por eso Pk es el percentil de k y kn es la multiplicación del valor k por el número total de veces que aparece un valor

O sea es el porcentaje de sujetos que deja por debajo de si una puntuación determinada.

4.5.2 SINTESIS

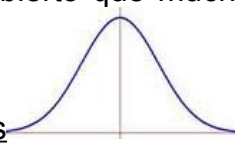
La distribución de frecuencias es una forma rápida y fácil de transformar una columna de datos en algo rapidamente comprensible. Una vez hecha la depuración de datos

pararemos a hacer las correciones correspondientes. Podemos o corregir valores tratándolos como si fuesen valores perdidos (dejar el espacio erroneo en blanco). O la segunda opción, mejor y más costosa identifica a cada sujeto con error , identificar el instrumento de medida pro ejemplo coger su encuesta y corregir el dato. Tras realizar las correciones volvemos a realizar la distribución de frecuencias , para verificar que se han corregido todos los errores. Y entonces estamos listos para comenzar a analizar los datos.

4.6 APROXIMACIÓN INTUITIVA A LAS REPRESENTACIONES GRÁFICAS Y LA CURVA NORMAL

Una representación gráfica es una forma atractiva de ordenar la información disponible en la matriz y comprenderla a simple vista. Para realizar un gráfico necesitamos una distribucion de frecuencias. Y podemos escoger un gráfico u otro dependiendo de las variables.

Al ver la representación gráfica de una distribución de frecuencias normal, se ha descubierto que muchas tinen la misma distribución (forma) que llaman campana de Guass



Es una distribución teórica simetrica, o sea si doblamos por la mitad la gráfica la forma de ambos lados coinciden y es asintótica , que la media, la mediana y la moda coinciden . Muchas variables educativas psicológicas sobre todo en muestras grandes porque la mayoría de los sujetos esta en el centro de los valores centrales o no se alejan demasiado. La curva normal y otras distribuciones teóricas binomial t, F, x, son fundamentales en el campo de la inferencia estadística. Donde reciben el nombre de distribuciones muestrales. Nos permite ver cuando una diferencia es o no estadisticamente significativa , es si el azar no puede explicar una diferencia de tal magnitud o si puede.

Curtosis (esto no aparece en el libro pero hay muchas preguntas del examen que lo preguntan)

El Coeficiente de Curtosis analiza el grado de concentración que presentan los valores alrededor de la zona central de la distribución. O sea lo alta que es la curva , porque si es muy alta hay mucha gente en ese valor central.

Se definen 3 tipos de distribuciones según su grado de curtosis:

Distribución mesocúrtica: presenta un grado de concentración medio alrededor de los valores centrales de la variable (el mismo que presenta una distribución normal).

Distribución leptocúrtica : presenta un elevado grado de concentración alrededor de los valores centrales de la variable.

Distribución platicúrtica: presenta un reducido grado de concentración alrededor de los valores centrales de la variable.