Medwave. Año XI, No. 2, Febrero 2011.

ESTADÍSTICA APLICADA A LA INVESTIGACIÓN EN SALUD

Construcción de una Base de Datos

Autor: Fernando Quevedo Ricardi.

Filiación: Departamento de Educación en Ciencias de la Salud, Facultad de Medicina, Universidad de Chile

Correspondencia: fquevedo@med.uchile.cl

Fecha de Aceptación: 23 de diciembre, 2010.

Citación: Quevedo F. Construcción de una Base de Datos. *Medwave* 2011;11(2).

Resumen

En la sección Series, Medwave publica artículos relacionados con el desarrollo y discusión de herramientas metodológicas para la investigación clínica, la gestión en salud, la gesión de la calidad y otros temas de interés. En esta edición se presentan dos artículos que forman parte del programa de formación en Medicina Basada en Evidencias que se dicta por e-Campus de Medwave. El artículo siguiente pertenece a la Serie **"Estadística Aplicada a la Investigación en Salud"**.

Palabras Claves: bioestadística, base de datos

Presentación de Resultados

Una vez que los datos han sido recogidos y registrados, comienza el procesamiento de datos. En la actualidad este procesamiento contempla el uso de computador, de manera que el primer paso será disponer los datos en una base de datos. Esto puede hacerse usando el programa Excel u otro equivalente, asignando a cada caso una fila en la planilla Excel, y a cada variable, una columna.

Número	Nombre	Edad	Sexo	Peso
1	Angélica	25	F	49,50
2	Marcos	18	M	42,12
3	Sonia	12	F	27,09

Tabla 1: Presentación de Resultados

En el ejemplo tenemos una base de datos con 3 casos y 4 variables. Es importante que los códigos y valores asignados a las variables respondan a criterios estables y estén expresados en el mismo formato, porque eso permitirá aplicar los filtros cuando se agrupen los casos. Es decir, que todas las edades estén expresadas en años cumplidos, por ejemplo, que todas las letras que indican el sexo sean mayúsculas, que los valores de peso estén expresados en kilos y siempre con dos decimales, por ejemplo, etc.

Resumen de Datos: Tablas y Gráficos

Las formas más usadas para presentar los datos son las tablas y los gráficos.

Anotaremos algunos conceptos sobre tipos de gráficos que le permitirán, por una parte, evaluar si un gráfico presentado en algún estudio está bien elegido, y por la otra, seleccionar los tipos más adecuado para sus propias investigaciones.

Tipos de Gráficos

El tipo de gráfico a utilizar dependerá principalmente del tipo de la variable y de los objetivos del estudio.

Así, para una variable cualitativa o cuantitativa-discreta, se puede utilizar un gráfico de barras simples o un gráfico de sectores.

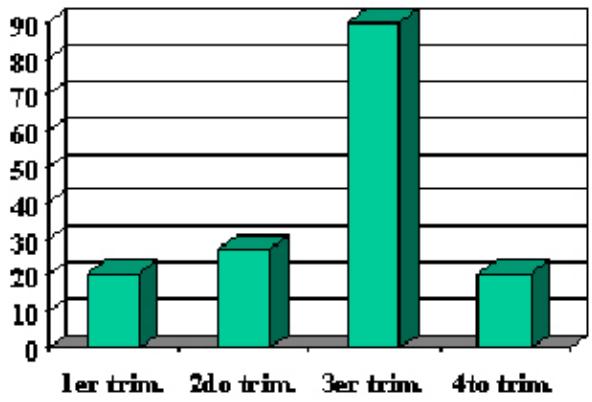

Gráfico 1: Barras Simples

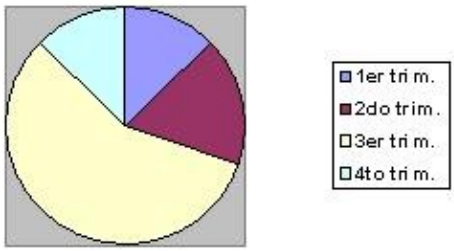

Gráfico 2: Sectores

Si se trata de dos variables de tipo cualitativa o cuantitativa-discreta, por ejemplo: estado civil y sexo, entonces es útil un gráfico de barras agrupadas.

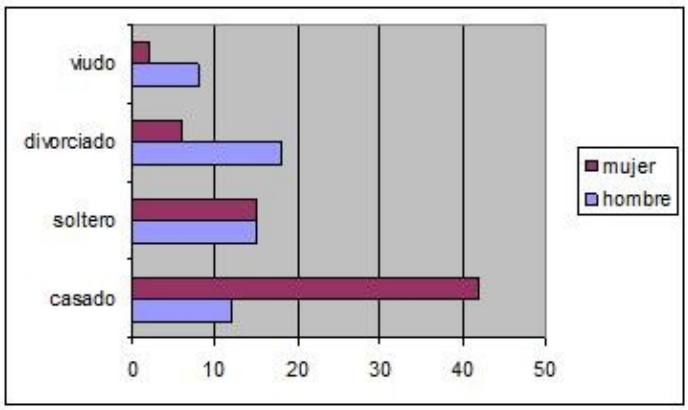

Gráfico 3: barras Agrupadas

En el caso de una variable cuantitativa en escala continua, se puede utilizar polígono de frecuencia.

Postulantes clasificados según puntaje, separados por sexo.

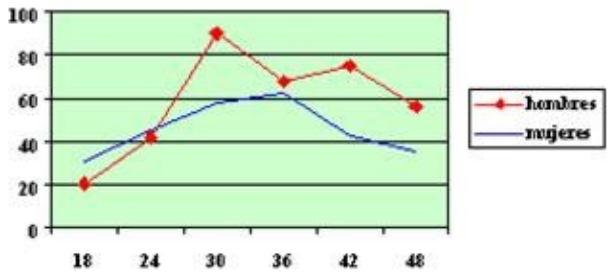

Gráfico 4 : Polígono de Frecuencia

El polígono de frecuencia se construye localizando el punto medio de cada intervalo de clase y marcando un punto a la altura de la frecuencia correspondiente al intervalo. Estos puntos se unen luego con líneas rectas.

El polígono de frecuencias es particularmente útil para comparar la distribución de una variable entre muestras distintas.

Tipos de Tablas

Distribuciones de Frecuencia

Tanto los datos cualitativos como los cuantitativos pueden resumirse en distribuciones de frecuencia. En el caso de los datos cualitativos la construcción de una tabla de frecuencia es relativamente simple: consiste en asignar a cada categoría de la escala un número (frecuencia) que corresponde a la cantidad de veces que se repite dicha categoría entre las unidades observadas.

Ejemplo: La distribución de frecuencia de la variable estado civil, en una muestra de 25 personas se presenta en la siguiente tabla:

Estado civil	Frecuencia
Casado	12
Soltero	8
Viudo	5
Total	25

Tabla 2: Distribución de Frecuencia

Para variables discretas tales como el número de camas de un hospital, o el número de caras observadas al lanzar cinco monedas, los valores de clase que se han de utilizar son obvios en general. Así, una tabla de frecuencia del número de caras que se presentan al lanzar cinco monedas 100 veces, sería:

Valor de clase	Frecuencia
5	4
4	15
3	29
2	30
1	17
0	5
Total	100

Tabla 3: Variables Discretas

Cuando el recorrido de la variable es pequeño y el tamaño de la muestra es grande hay valores de la variable que se repiten, por ejemplo si preguntamos por el número de personas que trabajan en cada familia, en una muestra de 50 familias, tendremos los siguientes resultados:

Personas que trabajan

2	1	2	2	1	2	4	2	1	1
2	3	2	1	1	1	3	4	2	2
2	2	1	2	1	1	1	3	2	2
3	2	3	1	2	4	2	1	4	1
1	3	4	3	2	2	2	1	3	3

Se puede observar que el recorrido de la variable va de 1 a 4, por lo tanto al hacer un conteo de la variable se tiene la siguiente tabla:

Personas que trabajan	Nº de familias
1	16
2	20
3	9
4	5
Total	50

Cuando el tamaño de la muestra y el recorrido de la variable son grandes, será necesario agrupar en intervalos de clases. Por ejemplo si a un grupo de 50 familias se le consulta por sus ingresos semanales (en miles de pesos).

93	74	86	107	77	92	77	87	100	77	91	90	73
80	94	105	88	66	107	95	69	80	83	87	89	94
105	78	79	98	86	97	112	97	79	96	92	86	
103	82	86	89	87	93	104	77	87	115	87	96	

Evidentemente, el recorrido de la variable es grande, por lo tanto necesitamos tabular con intervalos de clases. Para decidir sobre la cantidad de intervalos se debe tener en cuenta las siguientes consideraciones:

- Al tomar pocos intervalos aumenta la pérdida de información.
- Normalmente se trabaja con un máximo de 10 intervalos.

Para tabular los datos del ejemplo en cinco intervalos de clases, debemos considerar que el recorrido de la variable va de 66 a 115 y por lo tanto el rango de variación de la variable es de 50 mil pesos (115 menos 66). Como el objetivo es construir una tabla con 5 intervalos, cada intervalo deberá tener una amplitud de 10 (50/5) Así entonces conseguiremos una tabla como la siguiente (donde LS significa límite superior y LI límite inferior).

Intervalos LI - LS	Frecuencia
66 - 76	4
76 - 86	11
86 - 96	20
96 - 106	9
106 - 115	6
Total	50

Tipos de Frecuencia

Uno de los primeros pasos que se realizan en cualquier estudio estadístico es la tabulación de datos, es decir, recoger la información de la muestra resumida en una tabla en la que a cada valor de la variable se le asocian determinados números que representan el número de veces que ha aparecido.

Estos números se denominan **frecuencias**. Así se tienen las frecuencias que se enumeran a continuación.

Frecuencia absoluta

Esta frecuencia la denotaremos por ni y la definiremos como el número de veces que se repite un determinado valor de la variable. La suma de todas las frecuencias absolutas es igual al tamaño de la muestra.

Frecuencia relativa

Esta frecuencia la denotaremos por **hi** y la definiremos como el cuociente entre la frecuencia absoluta y el tamaño de la muestra. Donde n es el tamaño de la muestra y el recorrido de esta frecuencia es:

$$h_i = \frac{n_i}{n}$$
$$0 \leq h_i \leq 1$$

Figura 1. : Frecuencia hi

La frecuencia relativa es un tanto por uno, sin embargo también se puede escribir en tanto por ciento. La suma de todas las frecuencias relativas debe ser igual a uno (1,0) o 100 si se está amplificando tanto por 100.

Frecuencia absoluta acumulada (Ni)

Para poder calcular este tipo de frecuencia hay que tener en cuenta que la variable ha de ser cuantitativa o cualitativa ordinal. La frecuencia absoluta acumulada es el número de observaciones que hay desde el valor menor de la variable hasta un valor determinado de ella. Esta frecuencia tiene dos propiedades:

- La primera frecuencia absoluta acumulada es igual a la primera frecuencia absoluta.
- La última frecuencia absoluta acumulada es igual al tamaño de la muestra.

Frecuencia relativa acumulada (Hi)

Es la proporción (o porcentaje si se amplifica por 100) de observaciones que hay desde el valor menor de la variable hasta un valor determinado de ella. Esta frecuencia tiene dos propiedades:

- La primera frecuencia relativa acumulada es igual a la primera frecuencia relativa.
- La última frecuencia relativa acumulada es igual a uno (1,0) o 100.

Esto se entenderá mejor con un ejemplo. Tomaremos los datos de las personas que trabajan en cada familia, cuyas frecuencias absolutas ya anotamos en una tabla:

Personas que trabajan (Xi)	Nº de familias (ni)	hi	hi %	Ni	Hi	Hi%
1	16	16/50 = 0,32	32	16	16/50=0,32	32
2	*20*	*20/50 = 0,40*	*40*	*36*	*36/50=0,72*	*72*
3	9	9/50 = 0,18	18	45	45/50=0,90	90
4	5	5/50 = 0,10	10	50	50/50=1,00	100
TOTAL	50	1,0	100			

Interpretación: para efectos didácticos se interpretarán los valores de la segunda fila:

- Hay 20 familias que tienen 2 personas que trabajan.
- Un 40% de las familias tiene 2 personas que trabajan.
- 36 familias tienen 2 personas o menos, que trabajan.
- El 72% de las familias tiene 2 personas o menos que trabajan.

Tabla de Asociación

En una tabla de asociación los datos están clasificados según dos o más variables o criterios.

Ejemplo:
La siguiente tabla muestra la asociación entre edad y estado civil.

EDAD	CASADA Nº	SOLTERA Nº	CONVIVIENTE Nº	TOTAL Nº
17- 20	5	2	7	14
21 - 25	18	6	5	29
26 - 30	23	17	12	52
31 – 35	15	15	4	34
36 – 40	25	8	2	35
TOTAL	86	48	30	164

En este tipo de tablas se pueden identificar tres frecuencias:

- La frecuencia marginal por estado civil (86 casadas, 48 solteras, 30 convivientes).
- La frecuencia marginal por edad (14, 29, 52, 34 y 35).
- La frecuencia conjunta (5,2,7,18,6,5,23,17,12,15,15,4,25,8,2).

Cada frecuencia tiene su interpretación particular. Así, la frecuencia por estado civil indica que hay 86 casadas, la frecuencia por edad indica que hay 29 mujeres que tienen entre 21 y 25 años y finalmente, la frecuencia conjunta indica que son 18 las mujeres casadas que tienen entre 21 y 25 años.

En una tabla de asociación los porcentajes se pueden calcular de acuerdo a tres criterios, que en este ejemplo serían:

- Usando como referencia los totales por estado civil.
- Usando como referencia los totales por edad.
- Usando como referencia el total general.

El criterio a utilizar dependerá de la pregunta que se quiera responder. Para ilustrar esto tomemos como ejemplo la frecuencia conjunta 18. Usando este valor se pueden responder tres preguntas que se enumeran a continuación.

¿Qué porcentaje de las casadas tiene entre 21 y 25 años? Para responder a esta pregunta usaremos el total por estado civil.

$$\frac{18 \times 100}{86} = 20{,}9\%$$

Respuesta: el 20,9% de las casadas tiene entre 21 y 25 años.

Cabe hacer notar que en esta pregunta la condición de casada antecede al hecho de tener entre 21 y 25 años.

¿Qué porcentaje de las mujeres que tienen entre 21 y 25 años, son casadas? En este caso responderemos usando el total por edad.

$$\frac{18 \times 100}{29} = 62\%$$

Respuesta: El 62% de las mujeres que tienen entre 21 y 25 años de edad, son casadas.

Cabe hacer notar que en esta pregunta la condición de tener entre 21 y 25 años antecede al hecho de ser casada.

¿Qué porcentaje de las mujeres son casadas y tienen entre 21 y 25 años de edad?

Ahora, como la pregunta no especifica una condición antes que la otra, sino más bien exige que las dos condiciones se den en forma simultánea, procede entonces utilizar como referencia el total general.

$$\frac{18 \times 100}{164} = 11\%$$

Respuesta: El 11% de las mujeres son casadas y tienen entre 21 y 25 años de edad.

> Los artículos de la Serie "Estadística Aplicada a la Investigación en Salud" provienen del curso *Estadística Aplicada a la Investigación en Salud*. Si le interesa ahondar en estos contenidos, le invitamos tomar el curso en el siguiente link.