Introducción a estadística. Uso de Tablas

Hugo J. Bello

Definiciones Básicas

- Una **población** es un conjunto de todos los elementos que estamos estudiando, acerca de los cuales intentamos sacar conclusiones. Debemos definir esa población de modo que quede claro cuándo cierto elemento pertenece o no a la población.
- Una **muestra** es una colección de algunos elementos de la población, no de todos.
- Una **muestra representativa** contiene las características relevantes de la población en las mismas proporciones en que están incluidas en tal población.

Ejemplo

En las elecciones generales, la **población** sería el conjunto total de votantes. Una muestra sería seleccionar a 1000 individuos para intentar predecir el resultado de las elecciones. La muestra será **representativa** si contiene la misma proporción de mujeres y hombres que la población votante, geográficamente todas las regiones están proporcionalmente representadas...

Variables estadísticas

La noción de variable estadística, es una simplificación del concepto de *variable aleatoria* que veremos en temas posteriores. De momento, definiremos *variable estadística* como:

> característica o cualidad de un individuo que está propensa a adquirir diferentes valores. Estos valores, a su vez, se caracterizan por poder medirse.

Tipos de variables estadísticas

- **Cuantitativas**: Pueden ser medidas numéricamente.
  - **Continuas**: Puede tomar cualquier valor dentro de un intervalo o intervalos
  - **Discretas**: Solo toma una cantidad discreta de valores (por ejemplo una cantidad finita de valores)
- **Cualitativas**: son aquellas características o cualidades que no pueden ser calculadas con números, sino que son clasificadas con palabras

Ejemplo

- Cualitativa: Situación laboral de una persona (empleado / estudiante / paro).
- Cuantitativa continua: IPC, IBEX
- Cuantitativa discreta: número de hermanos, edad.

Las tablas de frecuencias

Pensemos en los siguientes datos: Supongamos que hemos extraído una muestra de la producción diaria de 30 telares de alfombras

$$16.2, 15.7, 16.4, 15.4, 16.4, 15.8, 16.0, 15.2, 15.7, 16.6, 15.8,$$
$$16.2, 15.9, 15.9, 15.6, 15.8, 16.1, 15.9, 16.0, 15.6, 16.3, 16.8,$$
$$15.9, 16.3, 16.9, 15.6, 16.0, 16.8, 16.0, 16.3$$

Observamos que los datos presentan repeticiones y que por lo tanto podemos hablar de *frecuencias* de ciertos valores de datos como por ejemplo 15.6 aparece tres veces.

lo primero que vamos a hacer es **ordenar los datos** y lo segundo **apuntar cuantas veces aparece cada dato**

valor	nº de veces que aparece
15.2	1
15.4	1
15.6	3
15.7	2
15.8	3
15.9	4
16.0	4
16.1	1
16.2	2
16.3	3
16.4	2
16.6	1
16.8	2
16.9	1

Esto que acabamos de hacer es una **tabla de frecuencias** y es la manera más directa de estudiar los datos, especialmente cuando hay repeticiones.

Por último, la tabla anterior tiene quizás demasiadas columnas, una idea sería *resumir* la información agrupando los datos por intervalos. Por ejemplo podemos tomar intervalos de longitud 0.5 e ir anotando cuantos valores encontramos que pertenezcan a ese intervalo.

intervalo	nº datos en el intervalo
[15, 15.25)	1
[15.25, 15.5)	1
[15.5, 15.75)	5
[15.75, 16)	7
[16, 16.25)	7
[16.25, 16.5)	5
[16.5, 16.75)	1
[16.75, 17)	3

Definición

Una **tabla de frecuencias** (también conocida como tabla de relaciones de frecuencias) es una tabla en la que se organizan los datos en clases, es decir, en grupos de valores que escriben una característica de los datos y muestra el número de observaciones del conjunto de datos que caen en cada una de las clases.

Notación

Si estamos ante una tabla de frecuencias

- A cada observación (habitualmente de la muestra ordenada) de la muestra la llamaremos x_i
- Al *nº de veces que aparece el dato* x_i lo llamaremos **frecuencia absoluta** y lo denotaremos por n_i
- Al número total de datos lo llamaremos N
- A la suma de la frecuencia n_i más todas las anteriores le llamamos **frecuencia absoluta acumulada** y la denotamos por N_i.

Si si la tabla está agrupada por intervalos

- A cada intervalo llamaremos $I_i = [l_i, l_{i+1})$
- al valor medio del intervalo lo llamaremos **marca de clase** $x_i = \frac{l_i + l_{i+1}}{2}$
- las frecuencias absuluta y absoluta acumulada se calculan igual pero en vez de contar el número de veces que aparece el dato contamos el *número de datos encontrados en el intervalo*.

Juntemos todo esto en el ejemplo anterior tenemos para la tabla sin agrupar

x_i	n_i	N_i
15.2	1	1
15.4	1	2
15.6	3	5
15.7	2	7
15.8	3	10
15.9	4	14
16.0	4	18
16.1	1	19
16.2	2	21
16.3	3	24
16.4	2	26
16.6	1	27
16.8	2	29
16.9	1	30

Y para la tabla agrupada por intervalos tenemos

intervalo	x_i	n_i	N_i
[15, 15.25)	15.125	1	1
[15.25, 15.5)	15.375	1	2
[15.5, 15.75)	15.625	5	7
[15.75, 16)	15.875	7	14
[16, 16.25)	16.125	7	21
[16.25, 16.5)	16.375	5	26
[16.5, 16.75)	16.625	1	27
[16.75, 17)	16.875	3	30

Medidas de concentración

Las medidas de concentración proporcionan información de los *valores centrales* en torno a los cuales se distribuyen los datos.

Son cálculos que realizaremos usando distintas estrategias sobre las tablas de frecuencias que hemos visto

Media

En general la media artimética de un conjunto de números

$$x_1, x_2, x_3, ..., x_n$$

se obtiene sumando todos valores y dividiendo por el número de sumando es decir

$$\frac{x_1 + x_2 + x_3 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

a este valor se le denota $\overline{x}$

Idea geométrica

Dados dos puntos en el espacio x e y, el punto medio entre ambos es justamente $(x+y)/2$. Esto ocurre tanto en la recta como en el espacio

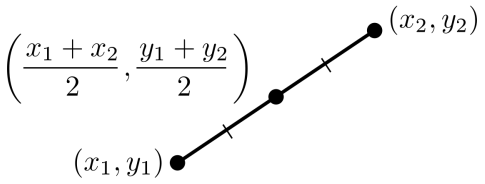

Figure 1: punto medio

Cómo calcularla

Para calcularlo, si tenemos una tabla de frecuencias

x_i	n_i
x_1	n_1
x_2	n_2
x_3	n_3
$\vdots$	$\vdots$
x_N	n_N

puesto que cada valor x_i se repite n_i veces, calcularemos la media de la siguiente manera

$$\overline{x} = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + x_3 \cdot n_3 + ... + x_N \cdot n_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i \cdot n_i$$

en el caso de que tengamos una tabla de frecuencias agrupadas por intervalos

I_i	$x_i = \frac{x_i + x_{i+1}}{2}$	n_i
$[l_1, l_2)$	x_1	n_1
$[l_2, l_3)$	x_2	n_2
$[l_3, l_4)$	x_3	n_3
$\vdots$	$\vdots$	$\vdots$
$[l_K, l_{K+1})$	x_K	n_K

haremos lo mismo pero usando la marca de clase $x_i = \frac{l_i + l_{i+1}}{2}$

Ejemplos

La siguiente tabla recoge las ventas de 100 sucursales de una empresa

ventas (miles)	x_i	n_i	$x_i \cdot n_i$
[700, 800)	750	4	3000
[800, 900)	850	7	5950
[900, 1000)	950	8	7600
[1000, 1100)	1050	10	10500
[1100, 1200)	1150	12	13800
[1200, 1300)	1250	17	21250
[1300, 1400)	1350	13	17550
[1400, 1500)	1450	10	14500
[1500, 1600)	1550	9	13950
[1600, 1700)	1650	7	11550
[1700, 1800)	1750	2	3500
[1800, 1900)	1850	1	1850

calculemos la media

$$\begin{aligned}\overline{x} &= \frac{1}{N}\sum_{i=1}^{N} x_i \cdot n_i \\ &= \frac{750 \cdot 4 + 850 \cdot 7 + 950 \cdot 8 + ... + 1850 \cdot 1}{100} \\ &= \frac{3000 + 5950 + 7600 + 10500 + 13800 + 21250 + 17550 + 14500 + 13950 + 11550 + 3500 + 1850}{100} \\ &= \frac{125000}{100} = 1250\end{aligned}$$

Moda

La **moda** (Mo) es el valor que aparece con mayor frecuencia en un conjunto de datos. Si hubiera varios datos con frecuencia máxima (no hay un único dato con mayor frecuencia sino varios) la moda serían todos ellos. Por lo tanto la moda puede no ser única.

Cómo calcularla

Simplemente buscamos el valor o valores con mayor frecuencia absoluta.

Ejemplo

La siguente muestra es el resultado realizar una muestra de renta per capita en un barrio concreto de madrid (en miles de euros anuales brutos)

$$14, 23, 15, 17, 12, 0.5, 30, 12, 23, 18, 25, 30, 15, 12, 23$$

renta	n_i
0.5	1
12	3
14	1
17	1
18	1
23	3
30	2

Luego la moda es:

$$Mo = 12, 23$$

Mediana

La mediana Me representa el valor de la variable de posición central en un conjunto de datos ordenados.

Si tenemos un número impar de datos, tomamos como mediana simplemente el valor que ocupe la posición central al ordenarlos, por ejemplo para los datos

$$1, 2, 3, 4, 5$$

El dato central es 3 así que $Me = 3$.

Si tenemos un número par de datos, al ordenar los datos no encontraremos un dato central, sino dos, con lo cual tomaremos como mediana la media entre estos dos datos centrales. Por ejemplo para los datos

$$7, 8, 9, 10, 11, 12$$

No hay un dato central, podríamos decir que los valores 9 y 10 son *centrales* así que tomamos $Me = (9 + 10)/2$

Cómo calcularla

Si los datos no están en una tabla de frecuencias, haremos lo que hemos comentado en los párrafos anteriores. Si hemos organizado los datos por frecuencias, haremos lo siguiente:

Dada una tabla (en la que calculamos tambíen las frecuencias acumuladas)

x_i	n_i	N_i
x_1	n_1	N_1
x_2	n_2	N_2
x_3	n_3	N_3
$\vdots$	$\vdots$	$\vdots$
x_N	n_N	N_N

1. Buscamos el valor que ocupa la *posición central* mirando en la tabla cual es el primer dato x_i cuya frecuencia acumulada supera o iguala $N/2$.
2. Si encontramos un dato cuya frecuencia acumulada N_i **iguala** $N/2$ tomamos como mediana la media de ese dato x_i y el siguiente x_{i+1}. $Me = \frac{x_i + x_{i+1}}{2}$
3. Si no encontramos uno cuya N_i iguale a $N/2$ sino uno que la supere directamente, tomamos ese dato como mediana. $Me = x_i$.

Bibliografía

- John A. Rice. Mathematical Statistics and Data Analysis
- F. M. Dekking, C. Kraailkamp, H. P. Lopuhaa, L. E. Meester. A Modern Introduction to Probability and Statistics. Understanding Why and How.
- https://es.wikipedia.org/wiki/Distribuci%C3%B3n_Bernoulli
- https://es.wikipedia.org/wiki/Distribuci%C3%B3n_binomial
- https://es.wikipedia.org/wiki/Distribuci%C3%B3n_geom%C3%A9trica