Estadística I

Tema 1: Introducción

Tema 1: Introducción

Contenido

- ¿Qué es la *Estadística*? - Definición.
- **Palabras clave**: población, parámetro, muestra, estadístico, tamaño poblacional, tamaño muestral, individuos, objetos.
- **Tipos de variables**: categórica (ordinal, nominal) y numérica (discreta, continua).
- ¿Por qué una muestra? Definición de muestra aleatoria simple.
- Frecuencias y distribución/tabla de frecuencias: absoluta, absoluta acumulada, relativa, relativa acumulada. Propiedades.

Tema 1: Introducción

Lecturas recomendadas

- Peña, D., Romo, J., *Introducción a la Estadística para las Ciencias Sociales*.
  - Capítulos 1, 2, 3.
- Newbold, P. *Estadística para los Negocios y la Economía* (2009).
  - Capítulo 1
  - Apartados 2.1, 2.4, 2.7. Cómo mentir con la estadística.

Definición de Estadística

Definición. La Estadística es la ciencia que trata de:

- recoger, organizar, resumir, presentar, interpretar y procesar datos para convertir los datos en información

  ⇐ Estadística Descriptiva

- predicciones, pronósticos, estimación

  ⇐ Inferencia Estadística

- ¿En qué ocasiones escuchaste/viste la palabra *estadística*?
  - Resúmenes de partidos de fútbol/tenis
  - Tasas de desempleo, número de heridos en accidentes de coche
- ¡La estadística es **mucho más** que porcentajes y números!

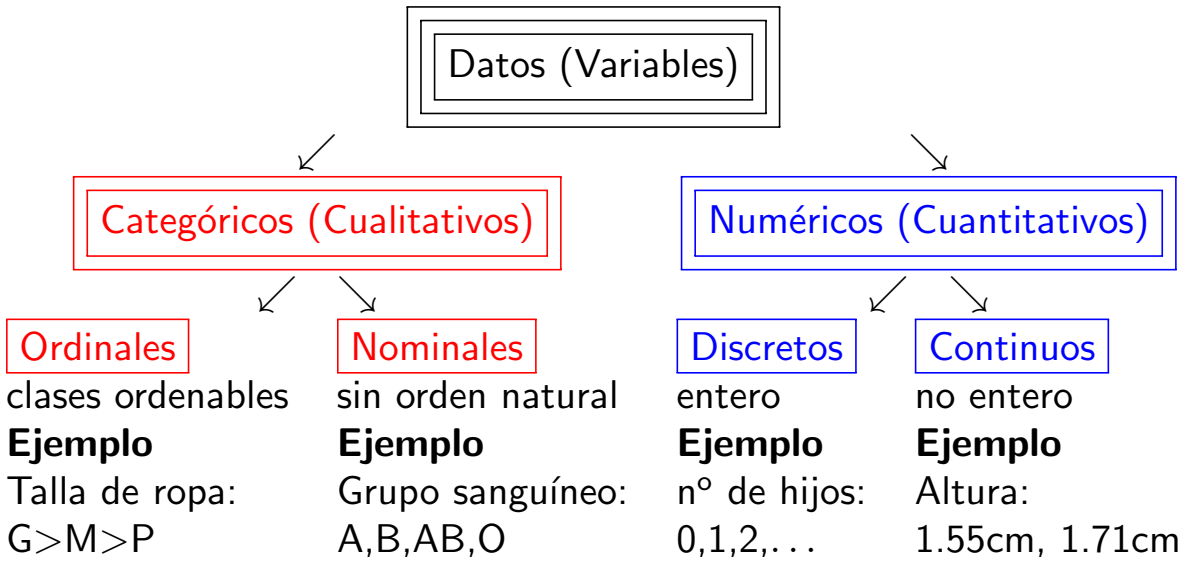
Palabras clave

- Una **población** es la colección completa de todos los ítems/individuos/objetos/sujetos de interés, o bajo investigación.
  N representa el tamaño poblacional
- Una **muestra** es un subconjunto de la población, elegida habitualmente para investigar las propiedades de la población subyacente.
  n representa el tamaño muestral
- Un **parámetro** es una característica específica de una población (fija).
- Un **estadístico** es una característica específica de una muestra (varía de muestra en muestra).
- Una **variable** es una característica de un individuo.

Ejemplos

- Pob todos los estudiantes de la UC3M. Variable: altura $\in (0, \infty)$
  Param: Altura media de todos los estudiantes. Estadístico: Altura media de los estudiantes muestreados.
- Pob: todos los peces de un lago. Variable: tamaño $\in \{G, M, P\}$
  Param: Número de peces pequeños en todo el lago. Estadístico: Número de peces pequeños capturados.
- Pob: todos los pacientes del Hospital de Getafe. Variable: grupo sanguíneo $\in \{\text{A, B, AB, O}\}$
  Param: Porcentaje de grupo sanguíneo AB entre todos los pacientes. Estadístico: porcentaje de grupo sanguíneo AB entre los pacientes muestreados.
- Pob: todas las bombillas de la marca *Acme*. Variable: tiempo de vida en días $\in \{0, 1, 2, \ldots\}$.
  Param: Variación en el tiempo de vida de todas las bombillas.
  Estadístico: Variación en el tiempo de vida de las bombillas muestreadas.

Tipos de datos



Notación: Se usan en general las letras X, Y, Z. Ejemplo:

X = altura en cm (letras mayúsculas en definición)

$x = 1.55$ (letras minúsculas para valores específicos)

$x_1 = 1.55,\ x_2 = 1.71$ (con más de uno, se añaden subíndices)

¿Por qué se usa una muestra?

En la práctica no estudiamos la población porque:

- Podemos destruir la población (ej. tiempo de vida de una bombilla).
- La población puede existir como concepto pero no en la realidad (ej. población de ítems defectuosos).
- Imposible de realizar (ej. población de todos los peces del mar).
- Demasiado caro.
- Tiempo de ejecución excesivo.

Definición de muestra aleatoria simple (*m.a.s.*)

Definición. Una muestra aleatoria simple es una parte de la población obtenida de forma que,

- cada miembro de la población se elige estrictamente al azar,
- cada miembro tiene la misma probabilidad de ser elegido, y
- cada posible muestra de n objetos es igualmente probable de ser elegida.

Notación: Una muestra de tamaño n obtenida de una variable X significa que:

- Tenemos n individuos seleccionados aleatoriamente de una población.
- Para cada uno de los individuos conocemos el valor de la variable X.
- Si X es categórica o discreta, es conveniente escribir los diferentes valores muestrales que toma X como $x_1, x_2, \ldots, x_k$, $k \leq n$ (ordenados desde el menor al mayor, salvo que X sea nominal).

Frecuencias y distribuciones de frecuencias

Definición. Una distribución de frecuencias es

- una lista o una tabla ...
- conteniendo agrupaciones de clases (categorías o intervalos donde toman valor los datos) ...
- y las correspondientes frecuencias mediante las cuales los datos toman valor dentro de cada clase o categoría.

Frecuencias:

- frecuencia absoluta es el (número de veces que el valor aparece en la muestra).
- frecuencia relativa es el (proporción de veces que el valor aparece en la muestra).

¿Por qué usar distribuciones de frecuencias?

- Una distribución de frecuencias es una forma de resumir los datos.
- La distribución condensa los datos primarios en una forma más útil ...
- y permite una interpretación visual rápida de los datos.

Agrupaciones por clases: datos categóricos y discretos

Clase, x_i	Frec. Absol., n_i	Frec. Relat., f_i	Frec. Absol. Acumul., N_i	Frec. Relat. Acumul., F_i
x_1	n_1	$f_1 = \frac{n_1}{n}$	$N_1 = n_1$	$F_1 = f_1$
x_2	n_2	$f_2 = \frac{n_2}{n}$	$N_2 = N_1 + n_2$	$F_2 = F_1 + f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	$f_k = \frac{n_k}{n}$	$N_k = n$	$F_k = 1$
Total	n	1	vacío	vacío

Nota:

- n_i = número de x_i en la muestra, $f_i = \frac{\text{número de } x_i}{n}$
- $N_i = N_{i-1} + n_i$, $F_i = F_{i-1} + f_i$
- $0 \leq f_i, F_i \leq 1$
- F_i y N_i no tienen sentido para variables categóricas nominales

Agrupaciones por clases

Ejemplo 1: Los datos inferiores muestran el grupo sanguíneo al que pertenecen los 40 individuos de una muestra.

$$AB,\ A,\ B,\ O,\ A,\ A,\ A,\ B,\ O,\ AB,$$
$$B,\ O,\ B,\ B,\ B,\ A,\ A,\ A,\ AB,\ B,$$
$$O,\ A,\ A,\ A,\ AB,\ AB,O,\ B,\ B,\ AB,$$
$$O,\ B,\ O,\ O,\ A,\ A,\ O,\ B,\ AB,\ AB$$

- ¿Qué tipo de variable es *grupo sanguíneo*? Obtén la distribución de frecuencias de los datos.
- ¿Qué porcentaje de la gente de la muestra pertenece al grupo sanguíneo A?
- ¿Qué porcentaje de la gente de la muestra pertenece a un grupo sanguíneo diferente de O?

Agrupaciones por clases

Ejemplo 1 cont.:

- Categórica, nominal con 4 clases diferentes. La distribución de frecuencias es:

Clase	Frecuencia Absoluta	Frecuencia Relativa
A	12	0.300
B	11	0.275
AB	8	0.200
O	9	0.225
Total	40	1

- 30 %
- $100\,\% - 22.5\,\% = 77.5\,\%$

Agrupaciones por clases

Ejemplo 2: La tabla inferior muestra diferentes niveles de satisfacción (I=insatisfecho, M=muy, S=satisfecho) en relación a 901 empleados.

Clase	Frecuencia Absoluta
MI	62
I	108
S	319
MS	412
Total	901

- ¿Qué tipo de variable se está estudiando? Obtén la distribución de frecuencias de los datos.
- ¿Qué porcentaje de la gente muestreada está satisfecha?
- ¿Cuántos individuos están insatisfechos o peor? ¿En %?
- ¿Cuántos individuos están al menos satisfechos? ¿En %?

Agrupaciones por clases

Ejemplo 2 cont.:

- Categórica, ordinal con 4 clases diferentes. La distribución de frecuencias es:

Clase	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Absoluta Acumulada	Frecuencia Relativa Acumulada
MI	62	0.07	62	0.07
I	108	0.12	170	0.19
S	319	0.35	489	0.54
MS	412	0.46	901	1
Total	901	1		

- 35 %
- 170, 19 %
- $319 + 412 = 731$ ó $901 - 170 = 731$, $35\,\% + 46\,\% = 81\,\%$ ó $100\,\% - 19\,\% = 81\,\%$

Agrupaciones por clases

Ejemplo 3: De entre las plantas que han sido tratadas con un nuevo pesticida, se seleccionaron 50 para evaluar el comportamiento del nuevo pesticida. En cada una de las plantas muestreadas se contó el número de hojas atacadas por un hongo. El resultado se muestra a continuación.

x_i	Frecuencia Absoluta
0	6
1	10
2	12
3	8
4	5
5	4
6	3
8	1
10	1
Total	50

Agrupaciones por clases

Ejemplo 3 cont.:

- ¿Qué puedes decir acerca de la variable en estudio? Obtén su distribución de frecuencias.
- ¿Qué porcentaje de las plantas muestreadas tuvo sólo 3 hojas atacadas?
- ¿Cuántas plantas muestreadas tuvieron no más de 3 hojas atacadas?
- ¿Cuántas plantas muestreadas tuvieron al menos 6 hojas atacadas?
- ¿Qué porcentaje de las plantas muestreadas tuvo entre 3 y 5 hojas atacadas?
- ¿Qué porcentaje de las plantas muestreadas tuvo al menos 8 hojas atacadas?
- ¿Qué porcentaje de las plantas muestreadas tuvo a lo sumo 2 hojas atacadas?

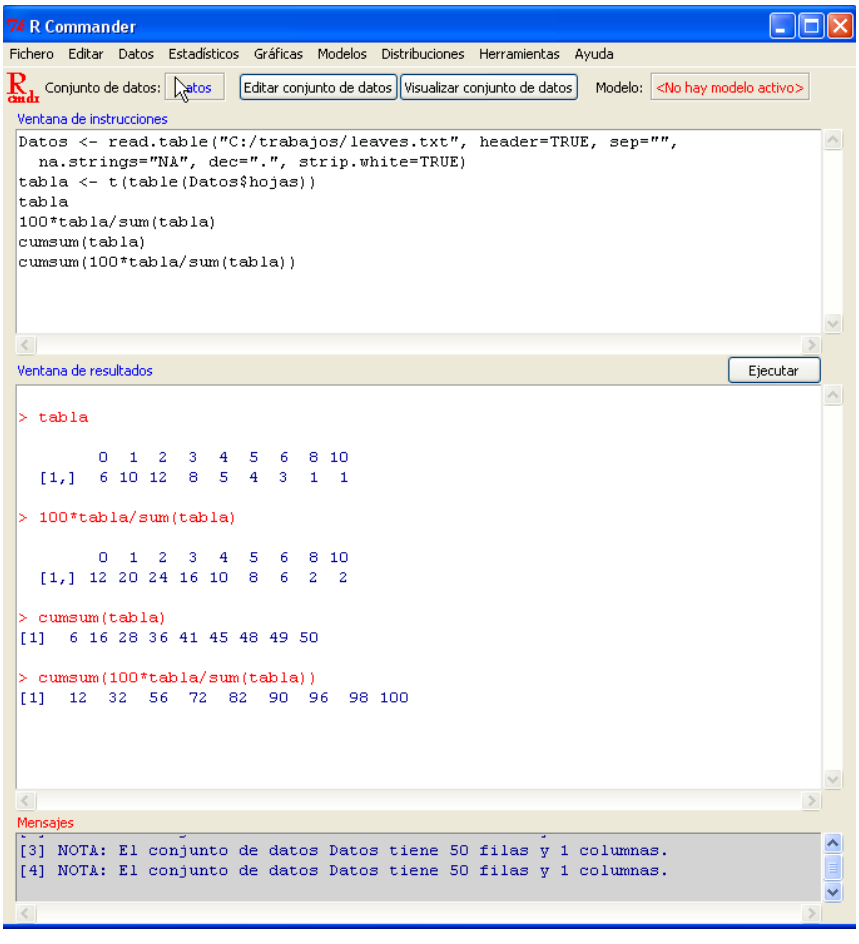
Agrupaciones por clases

Ejemplo 3 cont.:

- Numérica, discreta con 9 valores diferentes. La distribución de frecuencias es:

x_i	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Absoluta Acumulada	Frecuencia Relativa Acumulada
0	6	0.12	6	0.12
1	10	0.20	16	0.32
2	12	0.24	28	0.56
3	8	0.16	36	0.72
4	5	0.10	41	0.82
5	4	0.08	45	0.90
6	3	0.06	48	0.96
8	1	0.02	49	0.98
10	1	0.02	50	1
Total	50	1		

Tabla de frecuencias en R Commander



Agrupaciones por clases

Ejemplo 3 cont.:

- $16\,\%$
- 36
- $3+1+1$ ó $50-45=5$
- $16\,\%+10\,\%+8\,\%=34\,\%$ ó $(8+5+4)/50=34\,\%$
- $2\,\%+2\,\%=4\,\%$ ó $100\,\%-96\,\%=4\,\%$
- $56\,\%$

Agrupaciones por clases que son intervalos: datos continuos (y discretos)

Intervalo $[l_{i-1}, l_i)$	Marca de clase $x_i = \frac{l_i+l_{i-1}}{2}$	n_i	f_i	N_i	F_i
$[l_0, l_1)$	x_1	n_1	f_1	N_1	F_1
$[l_1, l_2)$	x_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[l_{k-1}, l_k)$	x_k	n_k	f_k	n	1
Total		n	1	vacío	vacío

Nota:

- Se incluye el extremo izquierdo, pero se excluye el extremo derecho (convención típica).
- Es posible aplicar la convención en sentido opuesto - verifica su definición en el software.
- Útil para tabular datos discretos si X toma muchos valores diferentes.

Agrupaciones por clases que son intervalos: datos continuos (y discretos)

- Muy frecuentemente los intervalos tomados como clases poseen la misma amplitud.
- Determinar la amplitud a para cada intervalo mediante

$$a = \frac{\text{número mayor - número menor}}{\text{número de intervalos deseados}}$$

- ¿Cuántos intervalos? Aproximadamente entre 5 y 20. Más concretamente:
  - $k \approx \sqrt{n}$ si n es *pequeño*.
  - $k \approx 1 + 3.22 \log(n)$ si n es *grande*.
- Los intervalos nunca se solapan.
- Redondea la amplitud del intervalo para obtener los extremos de los intervalos deseados.

Agrupaciones por clases que son intervalos: datos continuos (y discretos)

Ejemplo 4: Un fabricante de aislantes selecciona al azar 20 días de invierno y anota la temperatura más elevada del día (en grados Fahrenheit)

$$24, 35, 17, 21, 24, 37, 26, 46, 58, 30,$$
$$32, 13, 12, 38, 41, 43, 44, 27, 53, 27$$

Obtén la distribución de frecuencias de los datos.

- Se ordenan los datos primarios en orden ascendente: 12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58
- Se obtiene el rango (valor mayor – valor menor): $58 - 12 = 46$
- Se selecciona el número de clases: es decir $k = 5$
- Se calcula la amplitud de los intervalos: 10 ($46/5 \Rightarrow$ redondeo).
- Se determinan los extremos: 10 pero menor que 20, 20 pero menor que 30, etc.
- Se cuentan las observaciones que corresponden a cada clase.

Agrupaciones por clases que son intervalos: datos continuos (y discretos)

Ejemplo 4 cont.:

Intervalo	Marca de clase	n_i	f_i	N_i	F_i
$[10, 20)$	15	3	0.15	3	0.15
$[20, 30)$	25	6	0.30	9	0.45
$[30, 40)$	35	5	0.25	14	0.70
$[40, 50)$	45	4	0.20	18	0.90
$[50, 60)$	55	2	0.10	20	1
Total		20	1		

- ¿En cuántos días la temperatura se encontraba por debajo de 30°F? ¿En %?
  $(3 + 6 = 9$, que es el 45 %)
- ¿En cuántos días la temperatura se encontraba en al menos 45°F? ¿En %?
  $(2 + 4\frac{45-40}{50-40} = 4$, que es el 20 %)

Tabla de frecuencias en RCommander