Tema 1. Estadística Descriptiva (1ª parte)

Estadística

Ángel Serrano Sánchez de León

Índice

- Introducción
- Variables estadísticas
- Distribuciones de frecuencias
- Introducción a la representación gráfica de datos
- Medidas de tendencia central: media (aritmética, geométrica, armónica, cuadrática), mediana, moda, cuartiles, percentiles
- Medidas de dispersión: Rango, desviación media, desviación típica, varianza, rango intercuartil
- Momentos. Medidas de asimetría: Sesgo, curtosis
- Variables estadísticas bidimensionales

Introducción

- **Estadística** (del latín *Status* o ciencia del estado): Ciencia que se encarga de recoger, organizar e interpretar los datos.
  - Inicialmente para datos demográficos y económicos (censos de población, producciones agrícolas, riquezas, etc.), principalmente por razones fiscales.
  - Siglo XVII: cálculo de probabilidades para juegos de azar.
  - Siglo XVIII: su uso se extiende a problemas físicos (Astronomía) y actuariales (seguros marítimos).
  - Posteriormente se hace imprescindible en la investigación científica.
  - Siglo XIX, nace la Estadística como ciencia que une ambas disciplinas.
  - Hoy: importante para muchas ramas de la Ciencia desde la Medicina a la Economía.

Para qué sirve la estadística

- **Análisis de muestras** para hacer inferencias respecto a una población a partir de lo observado en la muestra (sondeos de opinión, control de calidad, etc).
- **Descripción de datos**: resumir la información contenida en un conjunto (amplio) de datos.
- **Contraste de hipótesis** o metodología de diseño experimentos que sirve para comparar las predicciones resultantes de las hipótesis con los datos observados (medicina eficaz, diferencias entre poblaciones, etc).
- **Medición de relaciones** entre variables estadísticas.
- **Predicción** de la evolución de una variable estudiando su historia y/o relación con otras variables.

Población y muestra

- **Población o universo:** conjunto completo de elementos, con alguna característica común, que es el objeto de nuestro estudio.
  - Por ejemplo, todos los sucesos en que podría concretarse un fenómeno o experimento cualesquiera.
  - Según el número de elementos: **finita** o **infinita**.
- **Muestra:** subconjunto de elementos de la población.
  - **Tamaño de la muestra:** número de elementos de la muestra.
  - Si la muestra es **significativa**, podremos extraer conclusiones sobre la población a partir del análisis de la muestra.

5

Partes de la estadística

- **Estadística Descriptiva:** se dedica a describir y analizar un grupo dado, sin sacar conclusiones sobre un grupo mayor.
  - Proceso deductivo (de lo general a lo particular).
  - Utiliza medidas de centralidad y dispersión, herramientas de visualización, etc.
- **Inferencia Estadística:** analiza una muestra para extraer conclusiones sobre la población.
  - Proceso inductivo (de lo particular a lo general).
  - Involucra cálculo de probabilidades.

Variables estadísticas

- **Variable estadística:** símbolo que representa al dato o carácter objeto de nuestro estudio de los elementos de la muestra y que puede tomar un conjunto de valores.
- Ejemplo:
  - Altura de los alumnos de la UFV.
  - Número de accidentes al mes en un tramo de la M40.
  - Temperatura de una habitación.
  - Códigos postales de los clientes de una tienda.
  - Etc.
- Los datos a los que se refieren las variables pueden ser de diversos tipos.

Variables estadísticas: Tipos

- Según los **niveles de medida** (definiciones de escala) y las operaciones permitidas con dichas variables:
  - Categóricas o cualitativas.
  - Numéricas o cuantitativas.
- Según el **número de valores** que toman:
  - Discretas.
  - Continuas.

Según los niveles de medida

- **Variables categóricas o cualitativas:**
  - **Nominales:** variables que difieren en cualidad más que en cantidad.
    - Operaciones permitidas: =, ≠, moda (valor más repetido).
    - Ej.: sexo, nacionalidad, código postal.
  - **Ordinales:** variables que tienen una noción de orden o rango a lo largo de un continuo.
    - Operaciones permitidas (además de las anteriores): <, > , mediana (valor central).
    - Ej.: notas de clase (sobresaliente, notable, aprobado, etc.), encuestas (totalmente de acuerdo, indiferente, en desacuerdo, etc.), puesto de los corredores de una carrera.

Según los niveles de medida (2)

- **Variables numéricas o cuantitativas:**
  - **De intervalo:** variables que pueden establecer intervalos iguales entre sus valores, lo cual permite calcular diferencias entre valores.
    - Operaciones permitidas (además de las anteriores): +, -, media aritmética.
    - Ej.: notas de clase (de 0 a 10), temperatura ºC, fechas, latitud y longitud.
  - **De proporción o razón:** variables que pueden establecer un valor 0 (ausencia total del fenómeno), lo cual permite calcular proporciones entre valores..
    - Operaciones permitidas (además de las anteriores): *, /, media geométrica.
    - Ej.: altura, peso, edad, densidad, etc.

Según el número de valores

- **Variables discretas:** cuando toman un número de valores posibles (finito o infinito) que se puede contar.
  - Representadas por números enteros.
  - Ej.: Número de alumnos en un aula, valor de un dado.
- **Variables continuas:** cuando toman cualquier valor entre dos valores dados.
  - Representadas por números reales.
  - Ej.: caudal de un río, presión atmosférica.

Distribuciones de frecuencias

- **Frecuencia:** El número de veces que aparece repetido un dato en un conjunto.
- Sea una muestra de tamaño N, de la que se mide la variable discreta x (es decir, tomará los valores $x_1, x_2, \ldots, x_N$).
- Cada uno de los valores x_i se repetirá 1 o más veces y supongamos que hay k valores diferentes.
- **Rango o recorrido:** Máximo – mínimo.
- **Distribuciones o tabla de frecuencias de una variable discreta:** representación tabular de los datos en los que se incluye la frecuencia de cada dato.
  - Varias definiciones de frecuencia.

12

Frecuencias

- **Frecuencia absoluta (n_i):** Número de veces que se repite el valor en la muestra dada.
  - Un valor dado o no está presente (vale o), será el único valor de la variable (vale *N*), o bien toma un valor intermedio.

$$0 \leq n_i \leq N$$

  - Sumando las veces que se repite cada valor obtenemos el número total de datos (tamaño de la muestra).

$$\sum_{i=1}^{k} n_i = n_1 + n_2 + \ldots + n_k = N$$

  - Problema: no nos informa de la importancia o no de cada dato.

13

Frecuencias (2)

- **Frecuencia relativa (f_i):** Cociente entre la frecuencia absoluta y el número total de observaciones N.

$$f_i = \frac{n_i}{N}$$

- Luego:

$$0 \leq f_i \leq 1 \qquad \sum_{i=1}^{k} f_i = \sum_{i=1}^{k} \frac{n_i}{N} = \frac{\sum_{i=1}^{k} n_i}{N} = \frac{N}{N} = 1$$

- La frecuencia relativa es el **tanto por uno**.
- Basta multiplicar por 100 para obtener el **tanto por ciento** (%).

Frecuencias (3)

- **Frecuencia absoluta acumulada (N_i):** Suma de las frecuencias absolutas de los valores inferiores o igual a x_i, o número de medidas por debajo o igual, que x_i.

$$N_i = \sum_{j=1}^{i} n_i$$

- Expresión recursiva:

$$N_1 = n_1$$

$$N_i = N_{i-1} + n_i,\ 1 < i \leq k$$

- De esta forma, para el último valor: $N_k = N$

15

Frecuencias (4)

- **Frecuencia relativa acumulada (F_i):** Ídem que la frecuencia absoluta acumulada, pero sumando las frecuencias relativas de los valores inferiores o iguales a x_i.
- También: cociente de la frecuencia absoluta acumulada respecto del número total de datos.

$$F_i = \frac{N_i}{N} = \frac{\sum_{j=1}^{i} n_j}{N} = \sum_{j=1}^{i} \frac{n_j}{N} = \sum_{j=1}^{i} f_j$$

- Multiplicando por 100, resulta el % acumulado.

Ejercicio

- Sea la variable *x* que representa el número de hijos de una serie de familias.

2	1	1	3	1	2	5	1	2	3
4	2	3	2	1	4	2	3	2	1

- Calcular:
  - Tamaño de la muestra
  - Número de valores diferentes de la variable
  - Recorrido
  - Tabla de frecuencias

Datos agrupados

- Si la variable discreta toma demasiados valores, o bien es una variable continua, los datos se **agrupan por "intervalos de clase"**.
- El recuento se hace por el número de datos que caen en cada intervalo.
- Un intervalo de clase viene delimitado por un **límite superior e inferior**.
  - **Intervalo de clase abierto:** Si el intervalo carece de límite superior o inferior.
- **Frontera de clase** (verdadero límite): promedio del límite superior de una clase y el límite inferior de la siguiente (concepto relacionado con la precisión de los datos).
- **Tamaño, anchura o amplitud del intervalo:** diferencia entre la frontera superior e inferior de un intervalo.
- **Marca de clase:** valor central o punto medio del intervalo.

Datos agrupados: Ejemplo

- Las alturas en cm de 20 alumnos de la UFV son:

180	177	178	181	175	184	180	181	179	180
184	186	167	183	177	186	181	177	173	181

- Por ejemplo, establecemos los siguientes intervalos de clase:

Intervalo	Marca de clase	Límite inferior	Límite superior	Frontera inferior	Frontera superior
[165, 170)	167	165	169	164,5	169,5
[170, 175)	172	170	174	169,5	174,5
[175, 180)	177	175	179	174,5	179,5
[180, 185)	182	180	184	179,5	184,5
[185, 190)	187	185	189	184,5	189,5

19

Datos agrupados: Pérdida de detalle

- Agrupar los datos facilita el trabajo: los datos se muestran **agregados** (agrupados, combinados, resumidos).
- Pero se pierde información (detalle), lo cual puede ser ventajoso o perjudicial, según el caso.

Datos agrupados: Elección de intervalos

- Calcular el recorrido de la variable (máximo – mínimo).
- Elegir *k*, el número de intervalos, típicamente entre 5 y 20 (dependerá del caso).
  - Truco: elegir *k* como el entero más próximo a la raíz cuadrada de *N*, el número total de medidas.
- Amplitud de intervalo: dividir el recorrido por el número de intervalos (y redondear por exceso).
- Determinar las fronteras de clase, tal que tengan una cifra decimal más que los datos (para evitar ambigüedades).
- Calcular las marcas de clase como los puntos centrales de cada intervalo.
- Realizar el conteo de datos que cae en cada intervalo.

Introducción a la representación gráfica de datos

- Para datos no agrupados:
  - Diagrama de barras
  - Polígono de frecuencias
  - Diagrama de frecuencias acumuladas
- Para datos agrupados:
  - Histograma
  - Polígono de frecuencias
  - Diagrama de frecuencias acumuladas
- Para variables cualitativas:
  - Diagrama de rectángulos
  - Diagrama de sectores (tarta)

Diagrama de barras y polígono de frecuencias

- Diagrama de barras:
  - Eje de abcisas (horizontal): valores de la variable.
  - Eje de ordenadas (vertical): barra de longitud igual a la frecuencia (absoluta o relativa) de ese valor.
  - Serán iguales en la forma pero cambiará el eje Y.
- Polígono de frecuencias:
  - Se consigue uniendo con rectas los extremos superiores de las barras del diagrama de barras

Diagrama de barras y polígono de frecuencias

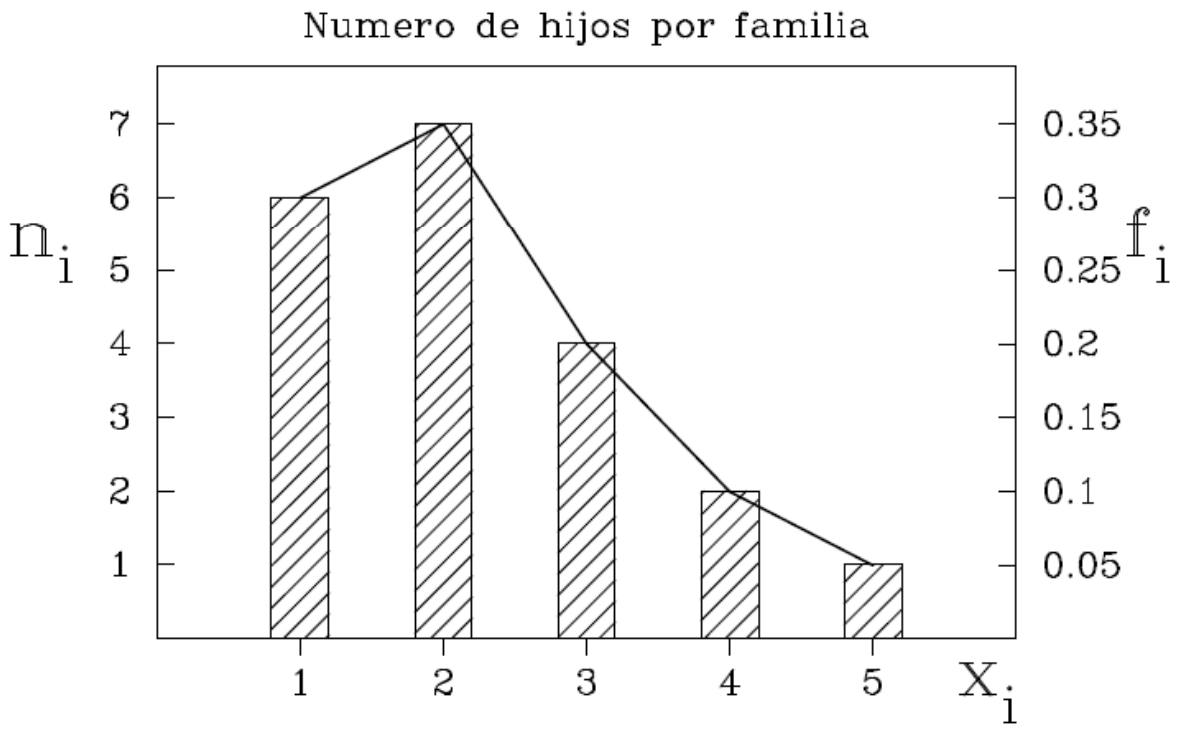
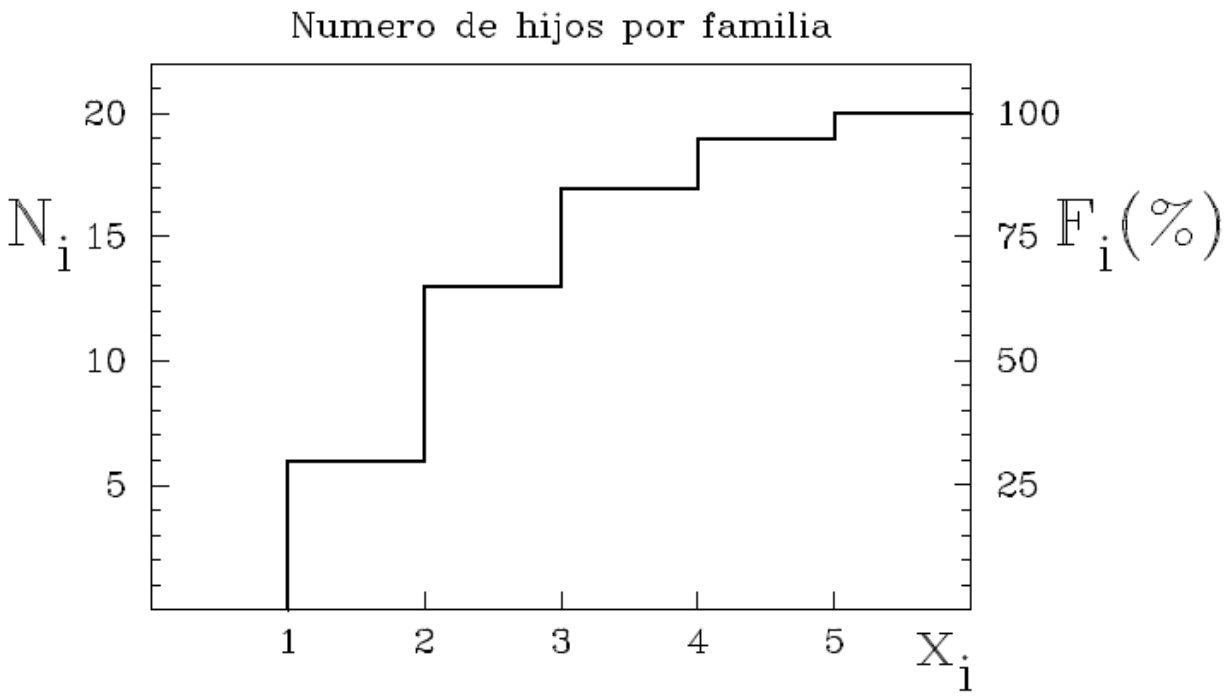


Diagrama de frecuencias acumuladas

- Eje de abcisas (horizontal): valores de la variable.
- Eje de ordenadas (vertical): sobre cada x_i una perpendicular cuya longitud será la frecuencia acumulada (N_i o F_i) de ese valor.
- Los puntos se unen con tramos horizontales y verticales a modo de escalera ascendente.

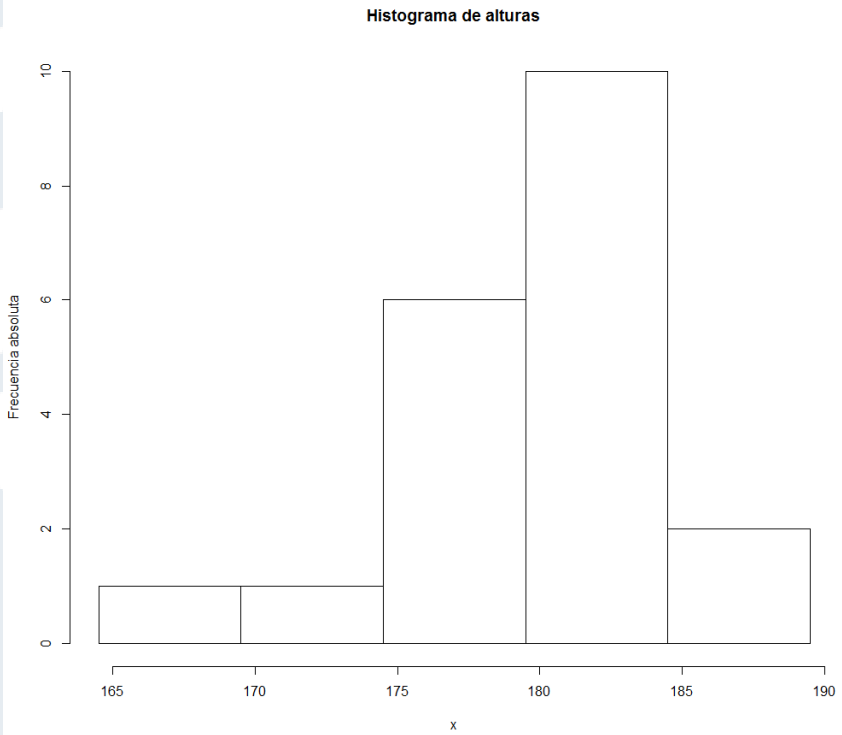
Diagrama de frecuencias acumuladas



Histograma

- Para datos agrupados.
- Conjunto de rectángulos adyacentes, cada uno de los cuales representa un intervalo de clase.
- La base de cada rectángulo es proporcional a la amplitud del intervalo y está centrada en una marca de clase.
- La altura se suele determinar para que el área de cada rectángulo sea igual a la frecuencia de la marca de clase correspondiente.
- En el caso de que la amplitud de los intervalos sea constante, la altura coincide con la frecuencia de cada marca de clase.
- El polígono de frecuencias y el de frecuencias acumuladas se haría de manera similar a los datos no agrupados.

Histograma

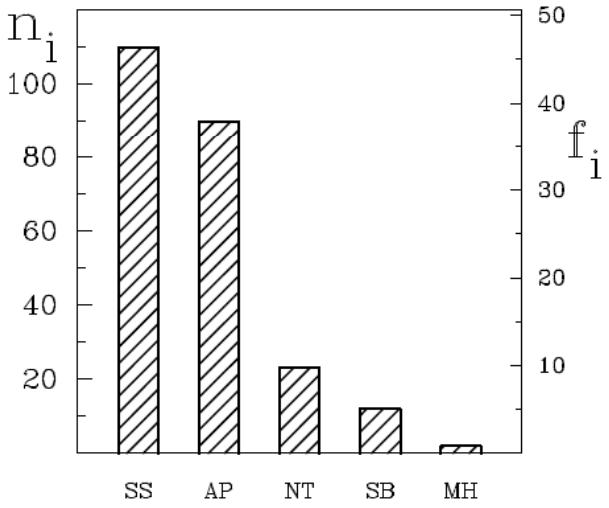

Histograma de alturas
Frecuencia absoluta
0
2
4
6
8
10
165
170
175
180
185
190
x

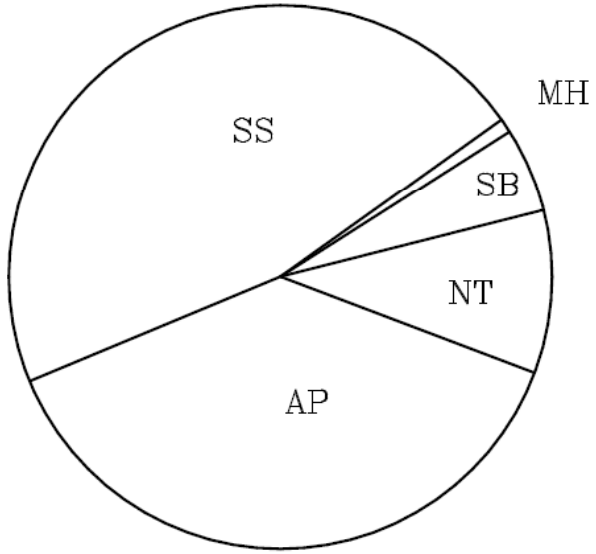
Datos cualitativos

- Diagrama de rectángulos: similar al de barras para datos numéricos no agrupados.
- Diagrama de sectores (tarta): diagrama circular donde cada sector tiene un área proporcional a la frecuencia relativa de la variable.
- Ejemplo: notas de clase.

Nota	n_i	f_i	N_i	F_i	α_i
Suspenso (SS)	110	0.46	110	0.46	165.6
Aprobado (AP)	90	0.38	200	0.84	136.8
Notable (NT)	23	0.10	223	0.94	36.0
Sobresaliente (SB)	12	0.05	235	0.99	18.0
Matrícula de Honor (MH)	2	0.01	237	1.00	3.6

Datos cualitativos





30