Estadística descriptiva. Exploración de los datos

Conceptos Preliminares

La *Población* es el conjunto completo de individuos a los cuales se referirán las conclusiones de su estudio. Tamaño de la población N.

La *Muestra* es un reducido grupo representativo de individuos de la población. A partir de ésta, el investigador, con técnicas estadísticas puede *inferir* las características y relaciones existentes en una población. Tamaño de la muestra n.

Los *Sujetos* o *Individuos* son los elementos que integran la población o muestra.

Los *Parámetros poblacionales* son los diferentes índices estadísticos descriptivos de toda una población. Se simbolizan con letras griegas. Por ejemplo, la media $\mu = (\sum x_i)/N$.

Función estadístico: cada parámetro de la población puede ser estimado a partir de los datos observados de una muestra extraída al azar. Las funciones que proporcionan estas estimaciones son los estadísticos. Por ejemplo, el estadístico que estima la media $\overline{x} = (\sum x_i)/n$.

La *variable* es cada uno de los caracteres o aspectos que se van a estudiar en los individuos.

Los datos se recogen en la matriz de datos que es una matriz cuyas filas representan los individuos y las columnas las diferentes variables.

Datos *missing* son aquellos valores que no se han registrado en la matriz.

Las variables se clasifican en:

- Variables categóricas, que son variables no métricas (binarias, o con más de dos categorías) y que a su vez pueden ser,
  - Nominales: sexo, grupo sanguíneo, tratamiento recibido, etc.
  - Ordinales: condición general de salud (buena, regular, mala), etc.
- Variables cuantitativas, que son variables métricas y que pueden ser,
  - Discretas: nº de hijos, edad en años, etc.
  - Continuas: peso, altura, presión arterial sistólica, etc.

Exploración de datos univariantes

Datos categóricos

Los datos categóricos se examinan con la correspondiente tabla de frecuencias y la representación de los porcentajes con diagramas de sectores o barras.

1. La *frecuencia absoluta* que es el número de veces que observamos el mismo valor de la variable (n_i).
2. La *frecuencia relativa* que es el cociente entre la frecuencia absoluta y el número total de repeticiones del experimento (f_i).
3. La *frecuencia acumulada (absoluta o relativa)* que es la suma de frecuencias absolutas (o relativas) anteriores con la del valor de la variable actual ($N_i = n_1 + \cdots + n_i$, $F_i = f_1 + \cdots + f_i$).
4. El *porcentaje* es la frecuencia relativa multiplicada por 100 ($100 f_i$).
5. El *porcentaje acumulado* es la frecuencia relativa acumulada multiplicada por 100 ($100 F_i$).

- el *diagrama de barras*, con el cual colocamos en el eje de abcisas los distintos valores discretos de la variable y en el eje de ordenadas las frecuencias absolutas o relativas,
- el *diagrama de sectores*, a cada valor se le asigna un sector cuyo ángulo sea proporcional a la frecuencia.

Datos numéricos

Los datos numéricos son más ricos en información que los categóricos. Las medidas descriptivas

Descripción basada en momentos

Representan la posición, dispersión, asimetría y apuntamiento de la distribución.

Ventajas:

- Utilizan todos los datos de la distribución.

- Fáciles de obtener: sumas, sumas de cuadrados, sumas de cubos, y sumas de potencias cuartas.

Inconvenientes:

- Difícil interpretación práctica en algunos casos.
- Los principales se ven afectados por valores anormales (outliers). Son medidas poco robustas.

Para sintetizar una distribución de datos cuantitativos es necesario dar las medidas que representen los 4 aspectos fundamentales de distribuciones de variables cuantitativas.

- Medidas de tendencia central. Resumen la posición central de la distribución. El estimador de la media poblacional (μ) es la *media* $\bar{x}$ que se obtiene calculando el promedio de los datos.

  Interpretación física: centro de gravedad.

- Medidas de dispersión. Permiten evaluar la separación de un conjunto de datos respecto a la media. El estimador de la *varianza* (σ^2) se denota por s^2 que se obtiene calculando la suma promediada del cuadrado de cada dato menos la media.

  Interpretación física: momento de inercia.

  La *desviación típica o estándar* σ es más útil, su estimador se simboliza por s: $\sigma = +\sqrt{\sigma^2}$, $\quad s = +\sqrt{s^2}$ y caracteriza la dispersión o grado de homogeneidad de una distribución.

  Notemos que en el caso particular de una distribución normal la desviación estándar sí tiene una interpretación más práctica.

  ¡Ojo!, ¡hay que tener cuidado! la media y varianza sólo deberían emplearse en distribuciones simétricas.....

- Medidas de forma: *asimetría*. Calculamos ahora momentos de orden 3 y obtenemos valores positivos (asimetría positiva $\Gamma_1 > 0$), negativos (asimetría negativa $\Gamma_1 < 0$), y nulos (simetría $\Gamma_1 = 0$).

- Medidas de forma: apuntamiento/*curtosis*. Calculamos ahora momentos de orden 4. Diremos que es platicúrtica ($\Gamma_2 < 0$) si es más aplanada que la normal, leptocúrtica ($\Gamma_2 > 0$) si es más apuntada, y mesocúrtica ($\Gamma_2 = 0$) si la forma coincide con la de la ley normal.

Descripción basada en ordenaciones

Estas medidas tienen la ventaja de ser más robustas, pues los valores extremos no afectan tanto al valor del índice.
- El *percentil* de orden k corresponde al valor de la variable que deja por debajo el $k\,\%$ de los sujetos de la población. El P_{75} deja por debajo al $75\,\%$ de la población.
- Los *deciles* dividen el conjunto ordenado de datos en 10 partes iguales.
- Los *cuartiles* dividen el conjunto ordenado de datos en 4 partes iguales.
- La *Mediana* es el valor de la variable que divide la distribución en dos partes iguales. Es el percentil 50, el decil 5 y el cuartil 2. Es una medida central que podemos sustituir en lugar de la media en el caso de distribuciones muy asimétricas.

La representación más habitual en el caso de estos datos es el *histograma*, con el cual dibujamos un rectángulo con área igual a la frecuencia absoluta (o relativa) correspondiente: colocaremos en el eje de abcisas los límites de los intervalos y sobre la ordenada el cociente entre la frecuencia y la amplitud (longitud) del intervalo.

Algunas tablas y fórmulas

Regla d'Sturges para escoger el número de intervalos:

Tamaño de la muestra	6 a 10	11 a 22	23 a 44	45 a 90	91 a 181	...
Número de intervalos	4	5	6	7	8	...

- Media y varianza muestrales (tamaño de muestra n)

$$\overline{x} = \frac{1}{n}\sum x_i = \frac{1}{n}\sum_{k=1}^{p} x_k n_k, \qquad s^2 = \frac{1}{n-1}\sum_{k=1}^{p}(x_k - \bar{x})^2 n_k.$$

- Coeficientes de asimetría y curtosis muestrales

$$G_1 = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s}\right)^3 n_i, \qquad G_2 = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s}\right)^4 n_i - 3.$$

- Medidas basadas en ordenaciones muestrales: mediana, percentiles, deciles, cuartiles y la moda. Salvo para la moda, que la miramos en las frecuencias, las demás medidas las señalaremos en el polígono de porcentajes acumulados.