Capítulo 2

Análisis exploratorio de datos. Estadística descriptiva

1. Análisis de una variable.

    Variables nominales.

    Variables ordinales y cuantitativas discretas.

    Variables cuantitativas continuas.

    Descripción numérica basada en sumas.

    Gráficos y medidas numéricas basados en la ordenación de los datos.

2. Análisis de dos variables.

    Descripción gráfica y numérica de dos variables discretas.

    Dos variables continuas.

    Recta de regresión.

    Una variable continua y una discreta. Comparación de poblaciones.

3. Descripción de variables temporales.

    Introducción a las series temporales.

    Descomposición básica de una serie temporal.

    Números índice.

2.1 Análisis de una variable

2.1.1 Variables nominales

REFERENCIA: Peña y Romo (1997). Capítulo 2

- Una variable nominal define k clases o categorías de individuos.
- Frecuencias absolutas: $n_1, \ldots, n_k$, $\sum_{i=1}^{k} = n$.
- Frecuencias relativas (o proporciones): $f_i = n_i/n$.

- Descripción numérica:

  Distribución (o tabla) de frecuencias:

Clase i	n_i	f_i
1	n_1	f_1
$\vdots$	$\vdots$	$\vdots$
k	n_k	f_k
Total	n	1

  La moda

- Descripción gráfica:

  Diagrama de barras. Diagrama de sectores.

  Diagrama de Pareto.

  Pictograma.

2.1.2 Variables ordinales y cuantitativas discretas

REFERENCIA: Peña y Romo (1997). Capítulo 3

- Se pueden describir numérica y gráficamente con las herramientas vistas para las variables nominales (excepto el diagrama de Pareto).

- Además se pueden definir las frecuencias acumuladas: N_i, F_i.
- También se pueden hacer diagramas de barras para frecuencias acumuladas.

2.1.3 Variables cuantitativas continuas

REFERENCIA: Peña y Romo (1997). Capítulo 3

Una primera aproximación a su descripción consiste en convertir estas variables en discretas (discretización) mediante la definición de clases *artificiales*.

- Discretización: Se divide el rango de valores que toma la variable en un número finito de intervalos. Cada intervalo define una clase.
- Se han de definir las marcas de clase: puntos medios de los intervalos.
- Los diagramas de barras se transforman en histogramas: el área de cada barra del histograma es proporcional a la frecuencia (absoluta o relativa) de cada clase.
- La apariencia del histograma cambia si cambia el ancho de las clases o si cambia el punto inicial de los intervalos.
- Un histograma aporta información sobre los datos: posición, dispersión, simetría, apuntamiento, unimodalidad (o bimodalidad, o plurimodalidad).
- Alternativas al histograma: polígono de frecuencias, diagrama de tallo y hojas.
- La campana de Gauss: la distribución normal.

2.1.4 Descripción numérica basada en sumas

REFERENCIA: Peña y Romo (1997). Capítulo 4

- Descripción numérica frente a descripción gráfica.
- Una medida de posición: la media. Definición y propiedades.
- Una medida de dispersión: la desviación típica. Definición y propiedades.
- El coeficiente de variación: dispersión relativa.
- Coeficientes de asimetría y apuntamiento.
- Descripción numérica de la campana de Gauss.
- Medidas numéricas para variables dadas mediante tablas de frecuencias. Interpretación:

  Muestra artificial.

  Medias ponderadas.

2.1.5 Gráficos y medidas numéricas basados en la ordenación de los datos

REFERENCIA: Peña y Romo (1997). Capítulo 5

- Medida de posición: la mediana. Definición y propiedades.
- Medidas de dispersión:

  Rango intercuartílico.

  MEDA (mediana de las desviaciones absolutas).
- Cuartiles, quintilas, decilas, percentiles, cuantiles.
- Diagrama de caja.
- Robustez: media frente a mediana.
- Estudio de la simetría según la posición relativa de media, mediana y moda.

2.2 Análisis de dos variables

2.2.1 Descripción gráfica y numérica de dos variables discretas

REFERENCIA: Peña y Romo (1997). Capítulo 7

- Descripción numérica básica: tablas de frecuencias.
  - La población dividida en clases según dos criterios.
  - Frecuencias conjuntas absolutas (n_{ij}) y relativas (f_{ij}).
  - Frecuencias marginales: $n_{i\cdot}, n_{\cdot j}, f_{i\cdot}, f_{\cdot j}$
  - Frecuencias condicionadas.
  - El concepto de independencia entre variables.
  - Para variables cuantitativas podemos calcular medidas numéricas condicionadas: media condicionada, desviación típica condicionada, etc.
- Descripción gráfica.
  - Diagrama de barras tridimensional (poco operativo).
  - Diagramas de barras múltiples.

2.2.2 Dos variables continuas

REFERENCIA: Peña y Romo (1997). Capítulo 8

Siempre podemos discretizar las dos variables. Así reducimos nuestro problema al caso anterior, aunque perdemos información en esta operación. Hay métodos específicos para variables continuas.

- Descripción gráfica.

  - Histograma tridimensional (poco operativo).
  - Diagramas de dispersión o nube de puntos.

    Dependencia e independencia.

    Linealidad de la relación.

- Medidas numéricas.

  Medidas de la relación lineal: covarianza y correlación. Propiedades.

  Un gráfico de dispersión siempre es más informativo que un coeficiente: datos atípicos, heterogeneidad.

  Correlación no implica relación causa-efecto: correlación espuria.

2.2.3 Recta de regresión

REFERENCIA: Peña y Romo (1997). Capítulos 9

- Si existe relación lineal entre dos variables entonces la nube de puntos se sitúa en torno a una línea recta: la recta de regresión.
- Buscar los parámetros de esta recta (pendiente y ordenada en el origen) es interesante (al menos) por dos motivos:
  - La recta es un buen resumen de la nube de puntos: la recta de regresión *es una medida de posición de la nube de puntos* (estadística descriptiva).
  - La ecuación de la recta puede ayudar a predecir valores de la variable colocada en el eje de ordenadas (y) a partir del conocimiento de los valores de la variable situada en el eje de abscisas (x) (predicción).
- La recta de regresión como recta de las medias condicionadas.
- Medida de dispersión de la nube de puntos en torno a la recta.
- Recta de regresión de mínimos cuadrados ordinarios.
- Cálculo de los coeficientes.

- Desviación típica residual. Coeficiente de determinación
- Predicción de un nuevo valor de y.
- Predicción fuera de la muestra.

2.2.4 Una variable continua y una discreta. Comparación de poblaciones

REFERENCIA: Peña y Romo (1997). Capítulos 10

- Descripción gráfica

  Gráficos de medias y desviaciones.

  Diagrama de caja múltiple.
- Descripción numérica para dos poblaciones.

  Datos apareados.

  Dos muestras independientes.

2.3 Descripción de variables temporales

2.3.1 Introducción a las series temporales

REFERENCIA: Peña y Romo (1997). Capítulo 11

- ¿Qué es una serie temporal? Gráfico de una serie temporal.
- Estacionariedad. Implicaciones en el cálculo de estadísticos descriptivos.
- Estacionalidad.
- Correlograma de una serie temporal. Interpretación.

2.3.2 Descomposición básica de una serie temporal

Referencia: Peña y Romo (1997). Capítulo 12

- Tendencia, componente estacional y componente irregular.
- Análisis de la tendencia: regresión frente al tiempo, medias móviles, diferenciación.
- Análisis de la estacionalidad: desestacionalización de la serie restando las medias mensuales.

2.3.3 Números índice

Referencia: Peña y Romo (1997). Capítulos 7, 8, 9 y 10

- Definición y aplicaciones de los números índices.
- Agregación simple y ponderada. Índices de Laspeyres, Paasche y Marshall-Edgeworth.
- El Índice de Precios al Consumo.