Capítulo 5

Estadística descriptiva

5.1. Introducción

La **estadística descriptiva** describe numéricamente, analiza y representa un conjunto de datos ordenados mediante la utilización de métodos numéricos, tablas y gráficas, simplificando y resumiendo la información. Se utiliza cuando los resultados del análisis estadístico no pretenden ir más allá del conjunto de datos investigados.

5.1.1. Muestra

Se denomina **muestra** de una población original con función $F(x)$ a la sucesión x_1, x_2, ..., x_n de los valores observables de la variable aleatoria x, que corresponden a n repeticiones independientes de un experimento aleatorio.

El **muestreo aleatorio simple** es aquel en el que cada individuo de la población tiene la misma probabilidad de ser incluido en la muestra.

Los datos que constituyen la muestra son llamados **observaciones**.

5.1.2. Caracteres estadísticos

Una observación puede ser **numérica (cuantitativa)** o **no numérica (cualitativa)**.

Los datos pueden ser **continuos** o **discretos**.

Las **modalidades** son las diferentes situaciones de un carácter.

Las observaciones obtenidas de forma aleatoria y que no se han ordenado constituyen los **datos crudos**.

Cuando los datos se ordenan según su magnitud, se obtiene una **distribución de frecuencias**. Cuando se ordenan según el tiempo de ocurrencia, se obtiene una **serie cronológica**. Cuando se ordenan según la ubicación geográfica, se obtiene una **distribución espacial**.

5.1.3. Distribución

Sea una población estadística de N individuos, descrita según una variable X, cuyas modalidades han sido agrupadas en un número n de clases. Para cada una de esas clases $i = 1, 2, ..., n$, se define:

La **frecuencia absoluta** de la clase f_i es el número de veces que se repite el valor.

La **frecuencia absoluta acumulada** de la clase F_i es el número de elementos de la muestra cuya modalidad es inferior o equivalente a las de la clase considerada.

$$F_i = f_1 + f_2 + \ldots + f_i = \sum_{j=1}^{i} f_j$$

La **frecuencia relativa** de la clase h_i es el cociente entre las frecuencia absoluta y el número total de observaciones.

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

La **frecuencia relativa acumulada** de la clase H_i es el cociente entre la frecuencia absoluta acumulada y el número total de observaciones.

$$H_i = \frac{F_i}{N}$$

5.2. Gráficos

5.2.1. Diagrama de barras

Se representan las diferentes modalidades de la variable en el eje x, y la frecuencia relativa o absoluta en el eje y. También se puede incorporar una nueva variable z para representarlo en el espacio.

5.2.2. Diagrama de sectores

Se divide un círculo en tantas porciones como modalidades existan de manera que el ángulo central de cada una es proporcional a la frecuencia.

5.2.3. Pictograma

Similares a los diagramas de barras, reemplazándolas por un mismo dibujo representado a diferentes escalas en proporción a la frecuencia. Son más imprecisos pero también más fáciles de comprender.

5.2.4. Histograma

Representan de forma sencilla una gran masa de datos, donde se puede observar las tres propiedades esenciales de una distribución: forma, tendencia central y dispersión.

En el eje x se grafican las clases. Si el ancho de la clase es constante, en el eje y se representa la frecuencia, sino, se representa la densidad de frecuencias.

A diferencia del diagrama de barras, los rectángulos verticales se representan contiguos para reflejar la idea de que **la variable es continua**.

El número de clases depende de cada situación. Se suele utilizar como parámetro el entero más próximo a la raíz del número de observaciones $\sqrt{N}$, siempre que no sea inferior a 5 ni superior a 20.

El **polígono de frecuencias** se construye una vez representado el histograma, uniendo con rectas los puntos que corresponden a las marcas de clase de cada intervalo.

5.3. Análisis exploratorio de datos

El **análisis exploratorio de datos (AED)** tiene como finalidad la examinación de los datos antes de aplicar cualquier técnica estadística. De esta forma, se consigue un entendimiento básico de los datos y las relaciones entre las variables.

5.3.1. Diagrama de tallo y hojas

Es un híbrido que combina los aspectos visuales del histograma con la información numérica que proporciona una tabla de distribución de frecuencias.

5.3.2. Gráfico de caja y bigote

Es una gráfica basada en cuartiles. La caja encierra el rango entre el primer cuartil Q_1 y el tercer cuartil Q_3, y se traza una línea a través de ella que representa la mediana Q_2. Los bigotes representan el valor mínimo y el valor máximo del rango.

Capítulo 6

Regresión y correlación

6.1. Introducción

6.1.1. Relaciones entre variables

La **relación funcional** entre dos o más variables se expresa por una fórmula matemática. Es una relación **determinística**.

La **relación estadística**, al ajustar una curva a observaciones, existe variación de los puntos en relación a la curva.

6.1.2. Regresión y correlación

El **análisis de regresión** permite encontrar el modelo que vincula las variables en cuestión, brindando un mecanismo de pronóstico.

El **análisis de correlación** determina la medida del grado de exactitud entre variables.

6.2. Regresión

El **modelo de regresión** expresa formalmente los aspectos esenciales de la relación estadística entre variables: la tendencia de la variable y a variar con la variable x de manera sistémica y la dispersión de los puntos en torno a la curva que relaciona las variables.

Existe una distribución de probabilidades de Y para cada valor de X: las variables X son fijas y las variables Y son aleatorias. Existen grupos de valores de Y para cada valor de la variable X llamados **subpoblaciones**.

6.2.1. Modelo de regresión bivariado

En ciertos diagramas de dispersión, no existe una curva que pase exactamente por todos los puntos, pero hay un fuerte indicio que los puntos se encuentran dispersos al azar en torno a una línea recta. Por lo tanto, se asume que la media de la variable aleatoria Y

está relacionada con x por la siguiente función:

$$E(Y|x) = \mu_{Y|x} = \alpha = \beta X_i$$

Para generalizar un modelo lineal probabilístico, se incorpora un término de error aleatorio ε_i, haciendo que Y no pueda ser pronosticado exactamente.

$$Y_i = \frac{\alpha + \beta X_i}{I} + \frac{\varepsilon_i}{II}$$

Si la media y la varianza de ε son 0 y σ^2 respectivamente, entonces la esperanza es

$$E(Y|x) = \alpha + \beta x$$

y la varianza es

$$Var(Y|x) = \sigma^2$$

Se dice que el modelo es de **regresión simple** cuando hay dos variables asociadas. Si esto no ocurre, el modelo es de **regresión múltiple**.

6.2.2. Estimación de los parámetros

Las siguientes ecuaciones normales conducen a obtener los estimadores a y b:

$$\sum_{i=1}^{n} Y_i - na - b\sum_{i=1}^{n} X_i = 0$$

$$\sum_{i=1}^{n} X_i Y_i - a\sum_{i=1}^{n} X_i - b\sum_{i=1}^{n} X_i^2 = 0$$

6.2.3. Varianza de la regresión

Se suele llamar **ecuación predictiva** a la ecuación de regresión muestral, ya que su principal objetivo es predecir valores medios de la variable dependiente asociados con un valor dado de la variable independiente.

La variación se da por la diferencia de los valores observados con los valores ajustados o estimados.

$$S_{y|x}^2 = \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{n-2}$$

6.3. Predicción y pronóstico

6.3.1. Predicción

La **predicción** es la estimación del valor medio de Y dado un valor particular de X.

$$\widehat{Y}_h = a + bX_h$$

La varianza de la regresión es

$$\sigma^2(\widehat{Y}_H) = \sigma^2\left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right)$$

6.3.2. Pronóstico

El **pronóstico** es la proyección de un solo valor de Y correspondiente a un valor de X particular.

$$\widetilde{Y} = a + bX_h$$

La varianza es

$$\sigma^2(Y_i - \widehat{Y}_H) = \sigma^2\left(1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right)$$

6.4. Correlación

6.4.1. Coeficiente de correlación

La medida del grado de relación entre dos variables se denomina **coeficiente de correlación** ρ.

$$\rho = \frac{E(X - \mu_x)E(Y - \mu_y)}{\sqrt{E(X - \mu_x)^2 E(Y - \mu_y)^2}} = \frac{\text{cov}}{\sigma_x \sigma_y}$$

6.4.2. Coeficiente de determinación

El ajuste es más preciso cuando el coeficiente se acerca más a 1.

$$R^2 = 1 - \frac{SCE}{SCT} = \frac{SCR}{SCT}$$

$$SCE = \sum (y_i - \hat{y}_i)^2, \qquad SCR = \sum (\hat{y}_i - \overline{y})^2, \qquad SCT = \sum (y_i - \overline{y})^2$$

6.4.3. Correlación muestral

El estimador de ρ se obtiene considerando los momentos muestrales.

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (x_i - \overline{x})^2}} = \frac{m_{1,1}}{S_x S_y}$$

Capítulo 7

Series cronológicas

7.1. Series cronológicas

Las **series cronológicas** son aquellas que están formadas por valores de una variable observada a intervalos regulares de tiempo.

Su análisis permite controlar las operaciones presentes y llevar adelante el planeamiento futuro a través de la aplicación de técnicas estadísticas. También sirve para predecir acontecimientos futuros. Se determina el patrón actual y se obtienen claves del patrón futuro.

7.2. Componentes de una serie de tiempo

Una serie de tiempo se puede considerar como una combinación de 4 elementos:

- Tendencia a largo plazo (T): el movimiento gradual de acuerdo a una curva. Pueden deberse a cambios en la riqueza, tecnología, población.
- Variación estacional (S): la tendencia a variar hacia arriba y abajo durante épocas específicas del año. Puede deberse al clima, costumbres, u otros factores.
- Variación cíclica (C): la tendencia a variar hacia arriba y abajo por períodos de tiempo más extendidos.
- Variación aleatoria (I): variación por acontecimientos imprevistos. Son breves y no repetitivas.

El **modelo aditivo** supone que los cuatro componentes son independientes unos de otros:

$$Y_i = T_i + S_i + C_i + I_i$$

El **modelo multiplicativo** supone que los cuatro componentes se relacionan entre sí.

$$Y_i = T_i \cdot S_i \cdot C_i \cdot I_i$$

7.3. Suavización de series de tiempo

Antes de tratar de modelar una serie de tiempo, es útil graficarla para determinar la naturaleza de los componentes, si es que existen.

Se utiliza una expresión lineal que transforma la serie $X(t)$ en una serie suavizada

$$Z(t) = F(X(t)),\ t = 1,\ 2,\ ...,\ n-1,\ n$$

La función F es el **filtro lineal**. El más utilizado es el **promedio móvil**.

7.3.1. Promedios móviles

Dado un conjunto de números Y_1, Y_2, ..., Y_N, se define un movimiento medio $\overline{Y}$ de orden N que viene dado por la sucesión de medias aritméticas.

$$\overline{Y}_{n_1} = \frac{Y_1 + ... + Y_N}{N}, \qquad \overline{Y}_{n_1+1} = \frac{Y_2 + ... + Y_{N+1}}{N},$$

Esta técnica auxilia en la identificación de la tendencia a largo plazo en una serie de tiempo ya que amortigua las fluctuaciones a corto plazo. Como inconvenientes, tiene que se pierden los períodos al comienzo y a la finalización de la serie, se pueden generar componentes que los datos originales no tenían y están fuertemente afectados por los valores extremos.

7.3.2. Suavización exponencial

La **suavización exponencial** es una clase especial de promedio móvil ponderado. Es útil en pronóstico a corto plazo y proporciona una impresión de los movimientos globales a largo plazo de los datos. Está dado por:

$$S_i = W\,Y_i + (1 - W)S_{i-1}$$

donde $S_1 = Y_1$, S_i es el valor de la serie exponencialmente suavizada calculada en el periodo i, Y_i es el valor observado de la serie en el periodo i, y W es el coeficiente de suavización asignado en forma subjetiva ($0 < W < 1$).

Cómo elegir *W*

Si sólo se quiere suavizar una serie mediante la eliminación de variaciones cíclicas e irregulares que no se desean, debe seleccionarse un valor pequeño de W (cercano a 0), pero si se quieren hacer pronósticos, se elegirá un valor grande de W (cercano a 1).

7.4. Análisis de tendencia

La **tendencia** es la componente más estudiada, con fines de pronóstico a largo y mediano plazo.

7.4.1. Tendencia lineal

El método de mínimos cuadrados permite ajustar una línea recta de la forma:

$$Y = a + bx$$

7.4.2. Tendencia polinómica

Si el modelo a ajustar a la tendencia es una curva, es un polinomio de segundo grado:

$$\hat{y}_i = B_0 + B_1 X_i + B_2 X_i^2$$

7.4.3. Tendencia potencial

Cuando una serie parece estar incrementando con rapidez cada vez mayor tal que la diferencia porcentual de una observación a otra es constante, se puede considerar una ecuación de tendencia potencial de la forma:

$$Y_i = B_0 \cdot B_1^{X_i}$$

7.4.4. Aislamiento y eliminación de la tendencia en datos anuales y mensuales

Si los pronósticos que se desean elaborar son a corto plazo, se debe eliminar el efecto del componente tendencia, ya que esta es la componente a largo plazo.

$$\frac{Y_i}{\hat{y}_i} = C_i I_i$$

Este cociente se denomina **relativas cíclicas irregulares**.

7.5. Variación cíclica

Es muy difícil encontrar un modelo regular, en promedio, que permita la proyección mecánica hacia el futuro en corto plazo. Además la identificación del estado actual de los movimientos cíclicos se ve obstaculizada por la presencia de movimientos irregulares. Por lo tanto, con datos anuales, con la descomposición se llega hasta la obtención de las relativas cíclicas-irregulares.

7.6. Variación estacional

Un conjunto de números mostrando los valores relativos de una variable durante los meses del año se llama **índice estacional** de la variable.

El valor de cada mes expresa la actividad de ese mes en particular como porcentaje de la actividad del mes promedio.

7.7. Resumen de los pasos en el análisis de series de tiempo

1. Coleccionar los datos.
2. Representar la serie, anotando cualitativamente la presencia de la tendencia de larga duración, variaciones cíclicas y variaciones estacionales.
3. Construir la curva o recta de tendencia de larga duración y obtener los valores de tendencia apropiados.
4. Si están presentes variaciones estacionales, obtener un índice estacional y ajustar los datos a estas variaciones estacionales.
5. Ajustar los datos desestacionalizados a la tendencia. Los datos resultantes contienen solamente las variaciones cíclicas e irregulares.
6. Representar las variaciones cíclicas obtenidas anteriormente, anotando cualquier periodicidad que pueda aparecer.
7. Combinando los resultados con cualquier otro tipo de información útil, hacer una predicción. Si es posible, discutir las fuentes de error y su magnitud.

Capítulo 8

Inferencia estadística

La **inferencia estadística** se ocupa de analizar la información con el fin de obtener conclusiones respecto a la población estudiada, dándole a esas conclusiones una medida de probabilidad.

Las características muestrales, o **estadísticos**, se obtienen a partir de observaciones aleatorias. Los estadísticos son también variables aleatorias, ya que dependen de datos que son valores aleatorios.

8.1. Estadísticos tratados como variables aleatorias

8.1.1. Valor medio, varianza y distribución

Los valores muestrales x_i son independientes e idénticamente distribuidos, con esperanza común μ y varianza σ^2 ya que provienen de la misma población.

El valor medio de la media muestral $E(\overline{x})$ es igual al valor medio de la población.

$$E(\overline{x}) = \mu$$

La varianza de la media muestral $Var(\overline{x})$ se obtiene por hallar la varianza del promedio de n valores independientes o idénticamente distribuidos.

$$Var(\overline{x}) = \frac{\sigma^2}{n}$$

El error estándar de la media muestral $\sigma(\overline{x})$ mide la variabilidad casual en medias de muestras.

$$\sigma(\overline{x}) = \sqrt{Var(\overline{x})} = \frac{\sigma}{\sqrt{n}}$$

Cuando n tiende a infinito, el desvío de la media muestral tiende a cero. Esto significa que cuanto mayor es la extensión de la muestra, menor será el error o fluctuación de las medias de una muestra a otra.

Si las muestras son extraídas de una población finita y el muestreo se realiza sin reposición, se debe introducir un factor de corrección por población finita en el error de la media:

$$\sigma(\overline{x}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

donde N es la extensión de la población y n es la extensión de la muestra.

8.1.2. Teorema del Límite Central

Sí se considera la suma de n variables aleatorias x, independientes e idénticamente distribuidas, cada una con media y varianza finita, a mayor número de variables involucradas, la distribución de la suma se aproxima a una distribución normal.

Se concluye, por lo tanto, que la variable aleatoria media muestral se distribuye normalmente con parámetros $E(\overline{x})$ y $\sigma(\overline{x})$.

En muchos casos de interés práctico, si $n \geq 30$, la aproximación normal será satisfactoria, independientemente de la forma de la distribución de la población.

8.1.3. Estimadores

Un estimador insesgado de σ^2 es la varianza muestral corregida:

$$S'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Con lo cual:

$$E[S'^2] = \sigma^2$$

8.2. Casos de distribución por muestreo

Frecuentemente se necesita hacer estimación relacionada a la media $\overline{X}$ y varianza S^2 de la población, la proporción p de elementos en una población que pertenecen a una clase de interés, la diferencia entre las medias de dos poblaciones y la diferencia entre las proporciones de dos poblaciones.

8.3. Distribución por muestreo de medias

8.3.1. Población normal con desvío σ conocido

Estandarizando la variable aleatoria media muestral se obtiene una variable normal estándar.

$$z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

8.3.2. Población normal, σ desconocido, muestra chica ($n < 30$)

Si la varianza se desconoce será reemplazada por la varianza muestral. La variable resultante se distribuye como t de Student.

$$t_{n-1} = \frac{\overline{x} - \mu}{\frac{S'}{\sqrt{n}}}$$

8.3.3. Población normal, σ desconocido, muestra grande $(n \geq 30)$

Cuando la muestra es grande, se puede considerar la varianza poblacional desconocida reemplazada por la varianza muestral y la distribución de la variable resultante sigue siendo normal:

$$z = \frac{\overline{x} - \mu}{\frac{S}{\sqrt{n}}}$$

8.4. Distribución muestral de proporciones

En la población, una proporción se define como:

$$\pi = \frac{K}{N}$$

siendo K el número de elementos que tienen una característica deseada y N el total de elementos de la población. En la muestra, se define como:

$$p = \frac{x}{n}$$

siendo p la proporción muestral, x la cantidad de elementos que poseen la categoría deseada y n la extensión de la muestra. Suele considerarse a p como la proporción de éxitos, y por esto se la asocia a la distribución binomial.

La esperanza y la varianza de la proporción son:

$$E[p] = E\left[\frac{x}{n} = n\frac{\pi}{n} = \pi\right]$$

$$Var[p] = Var\left[\frac{x}{n}\right] = \frac{n\pi(1-\pi)}{n^2} = \frac{\pi(1-\pi)}{n}$$

El error estándar de p mide las variaciones casuales de proporciones muestrales de una muestra a otra:

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Este error debe ajustarse por un factor de corrección por población finita si el muestreo se hace sin reposición:

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}\sqrt{\frac{N-n}{N-1}}$$

Luego, la distribución muestral es la siguiente:

$$p \sim N\left(\pi;\ \sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

$$z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0,1)$$