Tema 1

Variables estadísticas

Contenido

1.1. Introducción y conceptos básicos

La **Estadística** es la ciencia que se ocupa de los **experimentos aleatorios**, que son aquellos cuyos resultados no se pueden predecir de antemano.

El **objetivo final** de la Estadística es obtener conclusiones acerca de un experimento aleatorio cuando se conoce el resultado de ese experimento en una serie de realizaciones.

Las conclusiones que se obtienen se refieren a cierto universo o **población**, que es el conjunto de elementos sobre el que se pretende extraer nuestras conclusiones acerca de una característica relevante. Dicha característica recibe el nombre de **variable estadística** o simplemente **variable**.

Cada uno de los elementos de la población se denomina **individuo**. Un individuo no tiene que ser una persona física, puede ser una familia, un día, un negocio, etc.

La extracción de conclusiones se lleva a cabo observando el comportamiento de la variable sobre cierto subconjunto de la población que se llama **muestra**.

El valor de la variable en un individuo se llama **dato**.

En el análisis estadístico general se distinguen 3 pasos fundamentales:

Paso 1: selección de la muestra apropiada y medición de la variable estadística en ella. De esta parte se ocupa la **Teoría de muestras** y el **Diseño de experimentos**, que indican cómo se debe elegir la muestra para que represente a la población de una manera adecuada.

Paso 2: organización, descripción y resumen de los datos de la muestra. De estas cuestiones se ocupa la **Estadística descriptiva**.

Si la muestra coincide con toda la población, entonces recibe el nombre de **censo** y con el Paso 2 ya se alcanzaría el objetivo final de la Estadística. Si no es así, se pasa al Paso 3 con ayuda del **Cálculo de probabilidades**.

Paso 3: extracción de conclusiones acerca de la población. Esta parte se denomina **Inferencia estadística**.

Este curso se ocupará únicamente de la organización, la descripción y el resumen de los datos muestrales, incluyendo algunas técnicas específicas para analizar conjuntos de datos particulares de interés económico, como las variaciones de precios y ventas a lo largo del tiempo.

1.2. Tipos de variables estadísticas

Las variables estadísticas se pueden clasificar atendiendo a diversos criterios. En función de la naturaleza de los valores que pueden tomar se clasifican en escala nominal, ordinal y cardinal, y dentro de la escala cardinal se distinguen de razón (o cociente) y de intervalo.

Una variable estadística viene medida en **escala nominal** si sus valores son nombres que no admiten ningún orden real.

La **escala ordinal** se corresponde con variables cuyos valores no son números, pero se pueden ordenar.

La **escala cardinal** se refiere a variables cuyos valores son números.

Dentro de la escala cardinal, se dice que una variable está medida en **escala de razón** si el valor 0 tiene un sentido real de nulidad.

Si el valor 0 se corresponde con un valor arbitrario se dice que una variable está medida en **escala de intervalo**.

La importancia de esta distinción reside en que si una variable está medida en escala de razón, se pueden comparar los datos de dos individuos tanto por la diferencia como por el cociente. En cambio, si la variable está medida en escala de intervalo no se puede comparar por el cociente, sólo se puede hacer por la diferencia. En muchos casos es mucho más informativo utilizar el cociente que la diferencia y para verificar si es posible utilizar el cociente se debe determinar el tipo de escala.

La resta o diferencia depende de la magnitud que se está midiendo (euros, grados centígrados, etc.), es decir, es una **medida absoluta**. En cambio, el cociente no depende de las unidades, ya que indica la variación en tanto por uno, es decir, es una **medida relativa**.

Otra posible clasificación atiende a la continuidad del rango de valores que podría tomar la variable.

Son **variables discretas** aquellas que no pueden tomar ningún valor entre dos posibles.

Las **variables continuas** son aquellas para las que cualquier valor entre dos posibles es válido (aumentando la precisión si es necesario).

En la práctica, la distinción entre variables discretas y continuas es más simple: se comportan como variables discretas aquellas que toman pocos valores distintos, aunque se repitan mucho y se comportan como variables continuas aquellas que toman muchos valores distintos y prácticamente no se repiten.

Problema propuesto: Problema 1.1.

1.3. Distribuciones de frecuencias

Una vez recogidos los datos, lo primero que se debe hacer es ordenarlos. Lo más habitual es construir tablas de frecuencias, anotando los distintos valores que aparecen en la muestra y el número de veces que se repite cada uno.

El número de individuos de la muestra recibe el nombre de **tamaño muestral** y se denota habitualmente por N.

Los valores distintos de la variable ordenados (si es posible) se denotarán por x_i y se denominarán **modalidades**. El número de modalidades o valores distintos se denota por k.

El número de veces que aparece cada valor x_i se denotará genéricamente n_i y se denomina **frecuencia absoluta**.

La **tabla de frecuencias absolutas** de una variable se construye anotando las distintas modalidades que toma la variable, ordenadas (si es posible) de menor a mayor, junto con sus frecuencias absolutas. Posteriormente, se pueden añadir a la tabla otras columnas que expresan las frecuencias de diferente modo. Habitualmente se completa la tabla con una última fila en el que se anota la suma total de la columna correspondiente (si tiene sentido). La suma de la columna de frecuencias absolutas es el tamaño muestral.

Se llama **frecuencia relativa** del valor x_i a la proporción de individuos de la muestra con valor x_i y se denota por f_i, es decir, $f_i = n_i/N$.

Se llama **frecuencia absoluta acumulada** del valor x_i al número de individuos de la muestra con valor menor o igual a x_i y se denota por N_i. Se calcula sumando las frecuencias absolutas hasta el valor x_i. El último valor N_k tiene que ser justo el tamaño muestral, porque al final se acumulan **todos** los valores de la muestra.

Se llama **frecuencia relativa acumulada** del valor x_i a la proporción de individuos de la muestra con valor menor o igual a x_i y se denota genéricamente por F_i. Se puede calcular de dos formas: o bien sumando las frecuencias relativas hasta x_i, o bien calculando la proporción de frecuencias absolutas acumuladas respecto al total, es decir, N_i/N.

Se llama **distribución de frecuencias** a la columna de los valores x_i acompañada de cualquiera de las columnas de frecuencias anteriores. Conociendo una sola de las columnas de frecuencias y el tamaño muestral se puede calcular el resto.

Si la variable es **nominal**, **no** tiene sentido calcular las **frecuencias acumuladas**, porque al no poder ordenar los datos realmente, no se puede hablar de un valor menor o igual que otro.

Problemas propuestos: Problemas 1.2 y 1.3.

1.4. Datos agrupados

En muchas ocasiones el número de valores distintos que aparecen en una muestra es tan grande que es muy difícil (y poco informativo) representarlos en una tabla de frecuencias. En estos casos se suelen agrupar los datos para conseguir una representación adecuada.

A la hora de agrupar en intervalos, lo primero que hay que decidir es cuántas clases se quieren. No puede ser un número muy alto porque no se ganaría nada en el resumen, ni un número muy bajo, porque se perdería demasiada información.

Una forma de dividir los datos es agruparlos en intervalos de igual tamaño. Si al hacer esto quedan demasiados valores en un grupo y muy pocos en el resto, se pueden hacer más divisiones en el grupo más frecuente y unir grupos consecutivos poco frecuentes.

Para determinar los intervalos de igual tamaño hay que encontrar el rango de valores que se pretende dividir, es decir, hay que localizar los valores máximo y mínimo, observar cuánto distan y dividir ese rango entre el número de clases elegido para saber cuánto medirán las clases de agrupación. Para localizar cuales serían los intervalos, se comenzaría por el valor mínimo y sumando la amplitud de cada clase se irían calculando los extremos.

Este criterio no es el único que se puede utilizar, cualquier otra agrupación lógica es válida, por ejemplo agrupar en intervalos con números redondos, etc.

Para que no haya dudas, en las clases obtenidas se considerará que el extremo inferior está excluido y el superior está incluido, salvo para el primero, en el que se considera ambos incluidos. Matemáticamente se escribe $(a, b]$ (los paréntesis significan "excluido" y los corchetes "incluido"). La forma de elegir los extremos de las clases no influye más que cuando se hace la agrupación y se cuentan las frecuencias. En el tratamiento estadístico posterior se seguirá habitualmente esta notación ya que aunque técnicamente a veces se ajuste mejor otra, los resultados no cambian.

Una vez que se dispone de los intervalos de agrupación se pueden calcular todas las frecuencias de forma análoga a como se hacia anteriormente, por ejemplo, $n_i =$ número de individuos con valor dentro de la clase i, etc. Si se necesita resumir toda una clase $(a, b]$ por un sólo número, se puede elegir el centro de la clase, que se denomina **marca de clase** y se denota por $x_i = (a + b)/2$.

Con la agrupación por intervalos se consigue **resumir la muestra para representarla** más fácilmente, **pero se pierde información**. Si no se trata de representar la información fácilmente, sino que se **necesita hacer cálculos con el ordenador**, no se debe agrupar, sino **introducir toda la información** de la que se dispone.

Problemas propuestos: Problemas 1.4 y 1.5.

1.5. Apéndice

Interpretación de proporciones, porcentajes y variaciones

Para comparar relativamente un dato con respecto a otro se utiliza el cociente o proporción.

Ejemplo 1.1. Si un pantalón cuesta 36€, una chaqueta cuesta 30 y unos zapatos 40€, se puede decir que el precio del pantalón es $36/30 = 1{,}2$ veces el precio de la chaqueta, mientras que es $36/40 = 0{,}9$ veces el precio de los zapatos.

Lo habitual es hablar en porcentajes, no en tanto por uno, por lo que se suele multiplicar por 100 esos cocientes y se diría que el precio del pantalón representa el $1{,}2 \times 100 = 120\,\%$ el precio de la chaqueta, mientras que representa el $0{,}9 \times 100 = 90\,\%$ del precio de los zapatos. □

Al calcular una proporción todo lo que esté por encima de 1 significa un incremento y por debajo una disminución.

Ejemplo 1.2. En el Ejemplo 1.1, como 1,2 está por encima de 1, significa que el pantalón es más caro que la chaqueta. En concreto, $1{,}2 - 1 = 0{,}2$, por lo que el precio del pantalón es 0,2 veces superior al precio de la chaqueta.

En tanto por ciento, se diría que el pantalón cuesta un $20\,\%$ más que la chaqueta. En relación con los zapatos, como 0,9 está por debajo de 1, el pantalón es más barato que los zapatos. En concreto, como $1 - 0{,}9 = 0{,}1$, el precio del pantalón es 0,1 veces inferior al precio de los zapatos, o lo que es lo mismo, el pantalón es un $10\,\%$ más barato que los zapatos. □