CAPÍTULO 1

Muestreo y distribuciones en el muestreo

1.1. INTRODUCCIÓN

Anteriormente hemos estudiado conceptos fundamentales, como eran el concepto de variable aleatoria y su distribución de probabilidades, estudiamos diferentes modelos de distribuciones tanto de tipo discreto como de tipo continuo y analizábamos sus características básicas (media, varianza, etc.). A partir de ahora estaremos interesados en saber qué modelo sigue la población, y para ello nos basaremos en la información que se obtenga de un subconjunto o parte de esa población que llamaremos **muestra**.

Cuando realizamos una introducción general de la estadística decimos que uno de los objetivos fundamentales es el obtener conclusiones basándonos en los datos que se han observado, proceso que se conoce con el nombre de **inferencia estadística**, es decir utilizando la información que nos proporciona una muestra de la población se obtienen conclusiones o se infieren valores sobre características poblacionales.

En este capítulo daremos una serie de conceptos básicos que serán fundamentales para el desarrollo posterior de la inferencia estadística.

1.2. MUESTRA ALEATORIA

Sabemos que hay diferentes métodos para investigar u observar una población (observación exhaustiva o censo, subpoblación, muestra y observación mixta), aquí nos vamos a referir a la observación parcial mediante una **muestra** y diremos que se ha investigado la población a partir de una muestra cuando los elementos que componen la muestra no reúnen ninguna característica esencial que los diferencie de los restantes, representando, por tanto, a toda la población. Las conclusiones sacadas de la muestra se pueden inferir o extender a la población total. Así por ejemplo, supongamos que deseamos conocer el precio medio o valor medio de las viviendas en una zona de Madrid en el año 2017. Para conocer la característica precio de la vivienda en esa zona, nece-

sitaríamos saber el precio de venta de cada una de las viviendas vendidas durante ese período de tiempo y el precio por el cual cada propietario vendería la suya. Esta lista completa de viviendas con sus precios, constituye la población en la que estamos interesados, cuya característica, precio medio de la vivienda o media poblacional, deseamos conocer. Pero, en ésta y en otras muchas situaciones prácticas no será posible o no será fácil, por diversas razones el obtener la población entera en la cual estamos interesados. Sin embargo, sí podemos obtener la información necesaria, precio de la vivienda, para una muestra representativa de la población y a partir de la cual inferir y obtener conclusiones para toda la población total.

La muestra debe de ser representativa de toda la población y, por tanto, tendrá características similares a las que se observarían en la población entera, de tal manera que si observando los precios de las viviendas que han sido incluidas en la muestra resulta que el precio medio de las viviendas de la muestra, media muestral $\bar{x}$, ha resultado ser 240.000 unidades monetarias podremos inferir que la media poblacional precio medio de la vivienda en toda la población o zona que estamos considerando está en torno a 240.000 unidades monetarias.

La razón principal para investigar una muestra en lugar de la población completa es que la recogida de la información para toda la población daría lugar a un coste muy elevado tanto en recursos económicos como en tiempo. Incluso en ciertos casos en que los recursos fueran suficientes para investigar la población completa, puede ser preferible el investigar sólo una muestra muy representativa, concentrando sobre ella un mayor esfuerzo para obtener medidas más precisas de las características que nos interesen. De esta forma se puede evitar lo que algunas veces ocurre en las grandes operaciones censales, por ejemplo, en el censo decenal de población de los Estados Unidos, en donde se investigó toda la población, se observó que ciertas características y grupos poblacionales estaban muy poco representados, lo cual era debido a la problemática que lleva consigo una gran operación censal, tanto por el volumen de cuestionarios como por la cantidad de información.

Cuando se selecciona una muestra de una población, un objetivo fundamental es el poder hacer inferencias sobre características poblacionales u obtener conclusiones que sean válidas para toda la población. Por tanto, es muy importante que la muestra sea representativa de la población; así pues la calidad de la inferencia o conclusión obtenida a partir de la muestra, sobre las diferentes características poblacionales estará directamente relacionada con la representatividad de la muestra. Por ejemplo, supongamos que un director comercial desea conocer la opinión sobre un nuevo producto de limpieza. No sería correcto que limitara la correspondiente encuesta a sus amigos y a las personas que viven en su barrio, pues tales personas no reflejarían la opinión de toda la población ya que la muestra no sería representativa de toda la población, ni aleatoria. Para evitar estos problemas y poder realizar una inferencia correctamente sobre toda la población a partir de una muestra es necesario que se verifique la **representatividad** y la **aleatoriedad** de la muestra.

Un objetivo básico en muestreo es seleccionar una muestra que garantice con un costo razonable una buena representatividad de la población.

El procedimiento de selección de la muestra puede conducir a diferentes tipos de muestreo, como veremos al estudiar el muestreo en poblaciones finitas. Aquí nos vamos a referir a un solo tipo de muestreo, aunque inicialmente consideremos dos:

— muestreo con reemplazamiento, y

— muestreo sin reemplazamiento.

El **muestreo con reemplazamiento** consiste en seleccionar, por mercanismos aleatorios, los elementos de la población que entran a formar parte de la muestra, pero de tal manera que cuando se observa la característica, que estamos investigando, del primer elemento seleccionado, se devuelve el elemento a la población, se selecciona el segundo elemento entre todos los elementos de la población, se anota la característica que se está investigando y se devuelve a la población, y así sucesivamente. Este procedimiento permite que un elemento de la población pueda ser seleccionado en más de una ocasión para formar parte de una muestra, pues la selección se realiza con reemplazamiento, es decir, con devolución del elemento seleccionado a la población.

En el **muestreo sin reemplazamiento**, los elementos de la población que entran a formar parte de la muestra también se seleccionan aleatoriamente, pero después de observar la característica que estamos investigando no se devuelve el elemento de nuevo a la población, con lo cual no pueden volver a ser seleccionados como ocurría en el muestreo con reemplazamiento.

Así pues, si tenemos una población de N elementos y queremos seleccionar una muestra de tamaño n resulta que la probabilidad de que un elemento de la población sea seleccionado en la primera extracción para formar parte de la muestra será $\frac{1}{N}$, en ambos tipos de muestreo. Sin embargo, en la selección del segundo elemento las probabilidades son diferentes, pues en el muestreo con reemplazamiento continúa siendo $\frac{1}{N}$, ya que el número de elementos de la población sigue siendo N, pero en el muestreo sin reemplazamiento el tamaño de la población es $N-1$, pues el primer elemento seleccionado no se devuelve a la población y entonces la probabilidad de seleccionar un elemento concreto será: $\frac{1}{N-1}$. Vemos pues que en el muestreo con reemplazamiento la probabilidad de seleccionar uno a uno los n elementos de la muestra permanece constante y en el muestreo sin reemplazamiento no sucede lo mismo ya que en cada extracción no se devuelve el elemento a la población y ésta va disminuyendo a medida que se selecciona la muestra, siendo los tamaños poblacionales N, $N-1$, $N-2$, …, $N-(n-1)$.

Luego, la probabilidad de seleccionar una muestra concreta de n elementos será:

	1ª extracción	2ª extracción	…	nª extracción
Muestreo con reemplazamiento	$\frac{1}{N}$	$\frac{1}{N}$	…	$\frac{1}{N}$
Muestreo sin reemplazamiento	$\frac{1}{N}$	$\frac{1}{N-1}$	…	$\frac{1}{N-n+1}$

Si el tamaño de la población es infinito o muy grande, entonces el tamaño de la muestra *n* en comparación con ese tamaño *N* infinito o muy grande de la población es prácticamente despreciable, y entonces no existe diferencia significativa entre ambos tipos de muestreo.

En consecuencia, a partir de ahora nos vamos a referir a poblaciones de tamaño infinito o muy grandes, de tal manera que no haremos distinción ni referencia alguna a que el muestreo sea con reemplazamiento o sin reemplazamiento pues la diferencia existente entre ambos será irrelevante para nuestro estudio. No obstante hemos de tener en cuenta que si el tamaño *N* de la población es finito y realizamos un muestreo con reemplazamiento entonces le daremos el mismo tratamiento que si la población fuese de tamaño infinito, pues como hemos visto también dan lugar a un conjunto de variables aleatorias independientes e idénticamente distribuidas, es decir, a muestras aleatorias simples. Una **muestra aleatoria simple** de tamaño *n* de una población *X* está constituida por un conjunto de *n*-variables aleatorias $X_1, \ldots, X_n$ independientes e idénticamente distribuidas a la población *X*, es decir está constituida por un conjunto de observaciones muestrales independientes e idénticamente distribuidas.

Definimos a continuación de manera formal el concepto de muestra aleatoria simple con el que trabajamos en Inferencia estadística.

Definición 1.1. Muestra aleatoria simple.

> Sea *X* la variable aleatoria correspondiente a una población con función de distribución *F(x)*. Si las variables aleatorias $X_1, X_2, \ldots, X_n$ son independientes y tienen la misma función de distribución, F(x), que la de la distribución de la población, entonces las variables aleatorias $X_1, X_2, \ldots, X_n$ forman un conjunto de variables aleatorias independientes e idénticamente distribuidas que constituyen una **muestra aleatoria simple** de tamaño *n* de la población F(x).[1]

Al ser las variables aleatorias $X_1, X_2, \ldots, X_n$ independientes, resulta que la función de distribución conjunta será igual al producto de las funciones de distribución marginales, es decir:

$$\mathrm{F}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F(x_i)$$

Si la población de partida es tipo discreto entonces la función de probabilidad de la muestra será:

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X = x_i) = \prod_{i=1}^{n} P_i$$

[1] En lo sucesivo y si no indicamos lo contrario, las muestras que utilizaremos serán aleatorias simples, aunque a veces por abreviar digamos simplemente muestra aleatoria.

Si la muestra aleatoria simple procede de una población de tipo continuo con función de densidad $f(x)$, entonces la función de densidad de la muestra será:

$$f(x_1, \ \ldots, \ x_n) = \prod_{i=1}^{n} f(x_i)$$

1.3. PARÁMETROS POBLACIONALES Y ESTADÍSTICOS MUESTRALES

En general diremos que los **parámetros poblacionales** son las características numéricas de la población. En concreto, un **parámetro** es una caracterización numérica de la distribución de la población. El conocimiento del parámetro permite describir parcial o totalmente la función de probabilidad de la característica que estamos investigando. Así por ejemplo, si la característica a investigar sabemos que sigue una distribución exponencial de parámetro a, su función de densidad será:

$$f(x) = \begin{cases} a\mathrm{e}^{-ax} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

pero esta función de densidad no estará totalmente descrita hasta que no se dé el valor del parámetro a, y entonces será cuando podremos formular preguntas concretas sobre esa distribución, es decir, podremos calcular las diferentes probabilidades.

Si la característica a investigar sigue una distribución normal, $N(\mu, \sigma)$, cuya función de densidad es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}}$$

observamos que aparecen dos parámetros μ y σ, que no se han especificado, y para describir totalmente la función de densidad tendremos que dar valores a los dos parámetros μ y σ, pues si damos valor a un solo parámetro entonces diremos que está descrita parcialmente.

En la mayoría de los modelos probabilísticos nos encontraremos parámetros cuyos valores tendremos que fijar para especificar completamente el modelo y poder calcular las probabilidades deseadas[2]. De manera más concreta podemos decir que uno de los problemas centrales en estadística se nos presenta cuando deseamos estudiar una población con función de distribución $F(x, \theta)$, donde la forma de la función de distribución es conocida pero depende de un parámetro θ

[2] En la Estadística clásica un parámetro se puede considerar como una constante fija cuyo valor se desconoce.

desconocido, ya que si θ fuese conocido tendríamos totalmente especificada la función de distribución. Si el parámetro θ no se conoce, entonces se selecciona una muestra aleatoria simple (X_1, …, X_n) de tamaño n de la población, y se calcula para las observaciones de la muestra el valor de alguna función $g(x_1, \ldots, x_n)$, que representa o estima el parámetro desconocido θ. El problema es determinar qué función será la mejor para estimar el parámetro θ, lo cual será resuelto en el capítulo dedicado a la estimación.

A continuación exponemos el concepto de estadístico que es fundamental para estimar los parámetros poblacionales, pues los estimaremos mediante estadísticos definidos a partir de las observaciones de una muestra aleatoria.

Definición 1.2. Estadístico.

> Un **estadístico** es cualquier función real de las variables aleatorias que integran la muestra, es decir, es una función de las observaciones muestrales, la cual no contiene ningún valor o parámetro desconocido.

Continuando con la población de función de distribución $F(x, \theta)$, en donde θ es un parámetro desconocido, y considerando una muestra aleatoria simple, (X_1, …, X_n), constituida por n variables aleatorias independientes e idénticamente distribuidas, podemos definir algunos estadísticos o funciones de esas variables aleatorias, como por ejemplo:

$$g_1(X_1, \ldots, X_n) = \frac{X_1 + \ldots + X_n}{n}$$

$$g_2(X_1, \ldots, X_n) = \frac{X_1^2 + \ldots + X_n^2}{n}$$

$$g_3(X_1, \ldots, X_n) = \frac{(X_1 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{n}$$

los cuales se determinan totalmente a partir de las observaciones muestrales.

En general un estadístico T lo representaremos como[3]:

$$T = g(X_1, \ldots, X_n)$$

es decir, como una función g de las observaciones muestrales, que a su vez será también una variable aleatoria, pues para cada muestra el estadístico T tomará una valor diferente, así pues para una muestra concreta (x_1, …, x_n) el estadístico tomará el valor:

[3] Seguiremos como norma general el utilizar letras mayúsculas para indicar las variables aleatorias, para los estadísticos, estimadores y para representar una muestra aleatoria general, y utilizaremos letras minúsculas para los valores concretos que puedan tomar las variables aleatorias, las estimaciones y la realización de una muestra o muestra concreta.

$$T = g(x_1, \ldots, x_n)$$

y a medida que vamos tomando muestras diferentes se obtienen distintos valores del estadístico, resultando que efectivamente el estadístico *T* es también una variable aleatoria y por consiguiente tendrá su correspondiente distribución, a la que llamaremos **distribución muestral del estadístico**, como veremos posteriormente.

Vemos pues que un parámetro y un estadístico son conceptos muy diferentes, pues el parámetro es una constante y cuando se conoce determina completamente el modelo probabilístico, sin embargo el estadístico es una variable aleatoria cuyo valor dependerá de las observaciones muestrales.

En diferentes ocasiones se han estudiado medidas numéricas correspondientes a conjuntos de datos, así pues estudiamos, entre otras, la media y la desviación típica. Ahora vamos a distinguir entre medidas numéricas calculadas con conjuntos de datos poblacionales y las calculadas con datos muestrales. Así pues, si la medida numérica se calcula para el conjunto de datos poblacionales le llamaremos **valor del parámetro poblacional** y si se calcula para el conjunto de datos muestrales, le llamaremos **valor del estadístico muestral**.

Definición 1.3. Parámetros media, varianza y proporción poblacional.

En una población finita de tamaño *N* los **parámetros poblacionales media, varianza y proporción poblacional** vienen dados por[4]:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad [1.1]$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (X_1 - \mu)^2 \qquad [1.2]$$

$$p = \frac{X}{N} = \frac{\textit{número de éxitos en } N \textit{ pruebas}}{\textit{número de pruebas}} \qquad [1.3]$$

[4] Si la población es infinita utilizaremos la misma notación para designar estos parámetros poblacionales, pero estos no pueden ser calculados a partir de estas sumas finitas, sino que tendremos que recurrir al cálculo de valores esperados de variables aleatorias de tipo continuo.

Definición 1.4. Estadístico media, varianza y proporción muestral.

Para una muestra aleatoria simple de tamaño n, $(X_1, \ldots, X_n)$ los **estadísticos media, varianza** y **proporción muestral** se definen como:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad [1.4]$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2 \qquad [1.5]$$

$$P_x = \frac{X}{n} = \frac{\textit{número de éxitos en n pruebas}}{\textit{número de pruebas}} \qquad [1.6]$$

El estadístico **varianza muestral**, S^2, se puede formular también mediante las siguientes expresiones algebraicas:

$$S^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right) = \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i \right)^2}{n} \right) \qquad [1.7]$$

En efecto para ver la equivalencia de la expresión [1.5] con la [1.7], consideramos el numerador de la [1.5] y tendremos:

$$\begin{aligned} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2 &= \sum_{i=1}^{n} \left(X_i^2 - 2X_i\bar{X} + \bar{X}^2 \right) \\ &= \sum_{i=1}^{n} X_i^2 - 2\bar{X} \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} \bar{X}^2 \\ &= \sum_{i=1}^{n} X_i^2 - 2\bar{X} \left(n\,\bar{X} \right) + n\,\bar{X}^2 \\ &= \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i \right)^2}{n} \end{aligned} \qquad [1.8]$$

Si en lugar de considerar las n variables aleatorias, independientes e idénticamente distribuidas $(X_1, \ldots, X_n)$, que constituyen la muestra aleatoria simple, consideramos una muestra concreta $(x_1, \ldots, x_n)$ entonces los valores de estos estadísticos muestrales son:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad [1.9]$$

$$s^2 = \frac{1}{n-1}\left(x_i - \bar{x}\right)^2 \qquad [1.10]$$

$$p = \frac{x}{n} \qquad [1.11]$$

Luego vemos que efectivamente el estadístico es una función de las observaciones muestrales, y en estos casos asigna a cada muestra observada la media de los valores, la varianza o la proporción, respectivamente[5].

1.4. FUNCIÓN DE DISTRIBUCIÓN EMPÍRICA

Sabemos que la función de distribución de una variable aleatoria *X* estaba definida como:

$$F\left(x\right) = P\left(X \leq x\right)$$

y puede representar la proporción de valores que son menores o iguales que *x*.

De manera similar podemos definir la función de distribución empírica para una muestra.

Definición 1.5. Función de distribución empírica de la muestra.

Consideremos una población con función de distribución $F(x)$ y sean $(x_1, \ldots, x_n)$ los valores observados correspondientes a una muestra aleatoria simple procedente de esa población, y designamos por $N(x)$ el número de valores observados que son menores o iguales que x. Entonces definimos la **función de distribución empírica de la muestra**, que la notaremos por $F_n(x)$, como:

$$F_n\left(x\right) = \frac{N\left(x\right)}{n} \qquad [1.12]$$

EJEMPLO 1.1

Dada una muestra aleatoria formada por las observaciones muestrales (3, 8, 5, 4, 5). Obtener la función de distribución empírica y su correspondiente representación gráfica.

---

[5] Se observa que al definir el estadístico varianza muestral se divide por $n - 1$ en lugar de por n, la razón la veremos con más detalle después, pero aquí ya adelantamos que se ha definido así la varianza muestral s^2, para que esta s^2 sea un estimador insesgado de la varianza poblacional σ^2, pues si hubiéramos dividido por n entonces el estadístico no sería un estimador insesgado.

Solución:

Utilizando la expresión [1.12] podemos obtener la función de distribución empírica que aparece en la Tabla 1.1.

Tabla 1.1. *Función de distribución empírica.*

Observaciones muestrales x	$N(x)$	$F_5(x)$
—	$< 3, 0$	0,0
3	$\leq 3, 1$	0,2
4	$\leq 4, 2$	0,4
5	$\leq 5, 4$	0,8
8	$\leq 8, 5$	1,0

La representación gráfica de esta función de distribución la tenemos en el Gráfico 1.1.

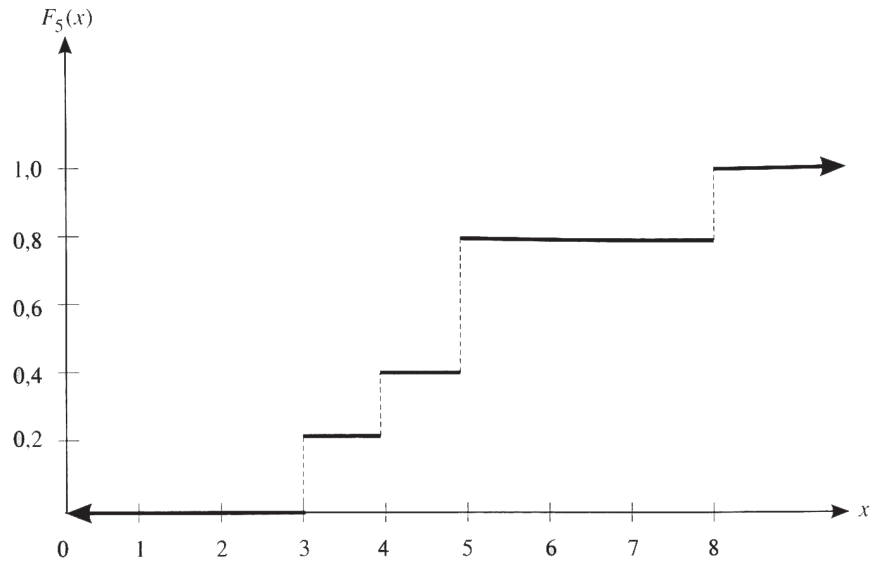

Gráfico 1.1. *Función de distribución empírica.*

La función de distribución empírica tiene las mismas propiedades que la función de distribución de la variable aleatoria, y, se puede demostrar, utilizando el **teorema de Glivenko-Cantelli**[6], que $F_n(x)$ converge en probabilidad a $F(x)$. Lo cual, a efectos prácticos, implica que cuando el

[6] El teorema de Glivenko-Cantelli, llamado también Teorema fundamental de la Estadística, por su papel fundamental en la inferencia estadística, indica que la función de distribución empírica de la muestra $F_n(x)$ converge en probabilidad a la función de distribución de la población $F(x)$. Es decir, para $\varepsilon > 0$, se verifica:

tamaño de la muestra crece la gráfica de la función de distribución empírica se aproxima bastante a la de la función de distribución de la población, y se puede utilizar como estimador de la misma.

De todo esto se deduce que la función de distribución empírica o su gráfica se puede utilizar para determinar la forma general de la distribución poblacional. También es fácil y muy frecuente el reconocer la forma de la distribución observando el histograma correspondiente que nos daría idea de la función de densidad.

1.5. DISTRIBUCIÓN MUESTRAL DE ESTADÍSTICOS

Como veremos posteriormente los estadísticos muestrales (proporción, media y varianza muestral) se pueden utilizar para estimar los correspondientes parámetros poblacionales. Así pues, para estudiar propiedades de estos estadísticos, como estimadores de los parámetros poblacionales, será necesario estudiar las características de la distribución de probabilidad de estos estadísticos.

Sabemos que los estadísticos muestrales se calculan a partir de los valores $(X_1, \ldots, X_n)$ de una muestra aleatoria, y estos estadísticos son también variables aleatorias. Como tales variables aleatorias tienen su distribución de probabilidad, así pues los estadísticos muestrales: proporción, media, varianza, etc., tendrán su correspondiente distribución de probabilidad. Si tales distribuciones de probabilidad se pueden obtener, entonces será posible establecer afirmaciones probabilísticas sobre esos estadísticos.

La distribución exacta de los estadísticos dependerá del tamaño muestral *n*. Así, en muchas situaciones, encontrar la distribución de probabilidad exacta del estadístico media muestral $\overline{X}$, incluso para *n* pequeño y variables aleatorias discretas, será bastante pesado, pero sin grandes dificultades teóricas. En muchos casos esto será relativamente sencillo, mientras que en otros lo mejor que se puede hacer es tomar una muestra grande y utilizar la distribución límite apropiada.

El término distribución muestral se utiliza para poner de manifiesto que hay diferencia entre la distribución de la población de la cual se ha extraído la muestra y la distribución de alguna función de esa muestra.

Conceptualmente, la distribución muestral de un estadístico puede ser obtenida tomando todas las posibles muestras de un tamaño fijado *n*, calculando el valor del estadístico para cada muestra y construyendo la distribución de estos valores.

En esta sección estamos interesados en determinar las distribuciones de probabilidad de algunos estadísticos muestrales, en concreto, para la media $\overline{X}$ y varianza S^2 muestral, que serán de bastante utilidad en diferentes aplicaciones estadísticas.

$$\lim_{n \to \infty} P\left[\sup_{-\infty < x < +\infty} | F_n(x) - F(x) | \geq \varepsilon \right] = 0$$

Lo cual significa que si la muestra es suficientemente grande y se verifica el teorema, entonces la muestra puede proporcionar información casi exacta sobre la distribución de la población.

Así, por ejemplo, si el estadístico es la media muestral $\overline{X}$, la distribución muestral de $\overline{X}$ puede construirse tomando todas las muestras posibles de tamaño n, calculando el valor del estadístico $\overline{X}$ para cada muestra, que lo notaremos por $\bar{x}$, y formando la distribución de los valores $\bar{x}$.

EJEMPLO 1.2

Sea una empresa dedicada al transporte y distribución de mercancías, la cual tiene una plantilla de 50 trabajadores. Durante el último año se ha observado que 25 trabajadores han faltado un solo día al trabajo, 20 trabajadores han faltado dos días y 5 trabajadores han faltado tres días. Si se toma una muestra aleatoria, con reemplazamiento, de tamaño dos (X_1, X_2) del total de la plantilla, obtener:

1. La distribución de probabilidad del número de días que ha faltado al trabajo un empleado, su media y su varianza.
2. Distribución de probabilidad del estadístico media muestral $\overline{X}$.
3. La distribución de probabilidad del estadístico varianza muestral, S^2.
4. La media y varianza del estadístico media muestral.
5. La probabilidad de que el estadístico media muestral, $\overline{X}$, sea menor que 2.
6. La media y varianza del estadístico varianza muestral.
7. La probabilidad de que el estadístico varianza muestral, S^2, sea menor o igual que 0,5.

Solución:

1. Empezaremos obteniendo la distribución de probabilidad de la variable aleatoria:

X: número de días que ha faltado al trabajo un empleado elegido aleatoriamente de la plantilla total.

La variable aleatoria $\overline{X}$, puede tomar los valores 1, 2 o 3, y como la selección se hace de manera aleatoria, todos los trabajadores tendrán la misma probabilidad de ser seleccionados, luego la distribución de probabilidad de la variable aleatoria X viene dada en la Tabla 1.2, y será la distribución de probabilidad de la población.

Tabla 1.2. *Distribución de probabilidad de la variable aleatoria X.*

Observaciones muestrales X x	**Probabilidades $P(X=x)=P(x)$**
1	$P(X=1)=P(1)=\frac{25}{50}=0,5$
2	$P(X=2)=P(2)=\frac{20}{50}=0,4$
3	$P(X=3)=P(3)=\frac{5}{50}=0,1$

A partir de esta distribución de probabilidad tenemos que la media será:

$$\mu = E[X] = \sum_{i=1}^{n} X_i P(X = x_i) = 1(0,5) + 2(0,4) + 3(0,1) = 1,6$$

y la varianza:

$$\begin{aligned}\sigma^2 = Var\ (X) = E[(X - \mu)] &= \sum_i (x_1 - \mu)^2 \cdot P(X = x_i) \\ &= (1 - 1,6)^2 (0,5) + (2 - 1,6)^2 (0,4) + (3 - 1,6)^2 (0,1) \\ &= 0,44\end{aligned}$$

Observamos que si sumamos el número total de faltas al trabajo que se han producido en la población de los 50 empleados y dividimos por los 50 empleados tenemos la media.

$$\frac{25 \cdot 1 + 20 \cdot 2 + 5 \cdot 3}{50} = \frac{80}{50} = 1,6$$

Análogamente sucede con la varianza.

Por esto, en lo sucesivo μ y σ^2 serán consideradas como la media y la varianza poblacional, respectivamente.

2. Seleccionamos una muestra aleatoria, con reemplazamiento, de tamaño dos (X_1, X_2), siendo:

X_1: variable aleatoria correspondiente al número de días que falta el primer trabajador seleccionado.

X_2: variable aleatoria correspondiente al número de días que falta el segundo trabajador seleccionado.

Ambas variables aleatorias X_1 y X_2 tienen la misma distribución de probabilidad que la de la variable aleatoria X, correspondiente a la población.

Pero como nos interesa obtener la distribución de probabilidad del estadístico media muestral:

$$\bar{X} = \frac{1}{2}(X_1 + X_2)$$

ésta estará relacionada con la distribución de probabilidad de las variables aleatorias X_1 y X_2.

Para tener las distribuciones de probabilidad de los estadísticos media $\bar{X}$ y varianza S^2 muestral necesitaremos tener los diferentes valores que puede tomar y sus probabilidades. Para ello empezaremos obteniendo las posibles muestras, con reemplazamiento, de tamaño dos, sus probabilidades y los valores correspondientes de los estadísticos media y varianza muestral, que vienen dados en la Tabla 1.3.

Tabla 1.3. *Muestras de tamaño dos y valores obtenidos para las distribuciones de probabilidad de $\bar{X}$ y S^2.*

Muestras de tamaño dos (x_1, x_2)	$\bar{X}$	S^2	$P(X_1=x_1, X_2=x_2)$
(1, 1)	1,0	0,0	0,25
(1, 2)	1,5	0,5	0,20
(1, 3)	2,0	2,0	0,05
(2, 1)	1,5	0,5	0,20
(2, 2)	2,0	0,0	0,16
(2, 3)	2,5	0,5	0,04
(3, 1)	2,0	2,0	0,05
(3, 2)	2,5	0,5	0,04
(3, 3)	3,0	0,0	0,01

Para obtener las probabilidades correspondientes a los diferentes valores muestrales, tendremos en cuenta que las variables x_1 y x_2 son independientes, pues el muestreo se ha realizado con reemplazamiento. Luego:

$$\begin{aligned} P\left(\bar{X}=1\right) &= P\left(X_1=1, X_2=1\right) \\ &= P\left(X_1=1\right)\cdot P\left(X_2=1\right) \\ &= (0{,}5)\,(0{,}5) = 0{,}25 \end{aligned}$$

$$\begin{aligned} P\left(\bar{X}=1{,}5\right) &= P\left[\left(X_1=1, X_2=2\right) \text{ o } \left(X_1=2, X_2=1\right)\right] \\ &= P\left(X_1=1, X_2=2\right) + P\left(X_1=2, X_2=1\right) \\ &= P\left(X_1=1\right)\cdot P\left(X_2=2\right) + P\left(X_1=2\right)\cdot P\left(X_2=1\right) \\ &= (0{,}5)\,(0{,}4) + (0{,}4)\,(0{,}5) \\ &= 0{,}20 + 0{,}20 = 0{,}40 \end{aligned}$$

Análogamente obtendremos las restantes probabilidades.

La información que nos proporciona la Tabla 1.3 la utilizaremos para obtener la distribución de probabilidad del estadístico media muestral $\bar{X}$, así pues:

$$\begin{aligned} &P\left(\bar{X}=1\right) = 0{,}25 \\ &P\left(\bar{X}=1{,}5\right) = 0{,}20 + 0{,}20 = 0{,}40 \\ &P\left(\bar{X}=2\right) = 0{,}05 + 0{,}16 + 0{,}05 = 0{,}26 \\ &P\left(\bar{X}=2{,}5\right) = 0{,}04 + 0{,}04 = 0{,}08 \\ &P\left(\bar{X}=3\right) = 0{,}01 \end{aligned}$$

Luego la distribución de probabilidad del estadístico media muestral $\bar{X}$ la tenemos en la Tabla 1.4.

Tabla 1.4. *Distribución de probabilidad del estadístico media muestral* $\bar{X}$.

Valores del estadístico $\bar{X}$ $\bar{x}$	**Probabilidades** $P(\bar{X}=\bar{x})=P(\bar{x})$
1	0,25
1,5	0,40
2	0,26
2,5	0,08
3	0,01

3. Análogamente podemos obtener la distribución de probabilidad del estadístico varianza muestral S^2. Los diferentes valores del estadístico S^2 aparecen en la tercera columna de la Tabla 1.3, así pues, para la primera muestra tenemos:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

$$= \frac{1}{2-1}\left[(1-1)^2 + (1-1)^2\right] = 0$$

Para la segunda muestra será:

$$S^2 = \frac{1}{2-1}\left[(1-1,5)^2 + (2-1,5)^2\right] = 0,5$$

y de manera análoga tendríamos los restantes valores.

Las probabilidades correspondientes a los diferentes valores del estadístico S^2, las obtenemos a partir de la Tabla 1.3, así pues:

$$P\left(S^2 = 0,0\right) = 0,25 + 0,16 + 0,01 = 0,42$$
$$P\left(S^2 = 0,5\right) = 0,20 + 0,20 + 0,04 + 0,04 = 0,48$$
$$P\left(S^2 = 2,0\right) = 0,05 + 0,05 = 0,10$$

Y la distribución de probabilidad del estadístico varianza muestral S^2 viene dada en la Tabla 1.5.

Tabla 1.5. *Distribución de probabilidad del estadístico varianza muestral S^2.*

Valores del estadístico S^2 s^2	**Probabilidades $P(S^2 = s^2) = P(s^2)$**
0,0	0,42
0,5	0,48
2,0	0,10

4. Para el cálculo de la media y varianza del estadístico media muestral tendremos en cuenta su distribución de probabilidad dada en la Tabla 1.4.

Utilizando la definición de valor esperado de una variable aleatoria de tipo discreto tenemos:

$$\mu_{\bar{x}} = E\left[\bar{X}\right] = \sum_i \bar{x}_i \cdot P\left(\bar{X} = \bar{x}_i\right)$$

$$= 1(0{,}25) + 1{,}5(0{,}40) + 2(0{,}26) + 2{,}5(0{,}08) + 3(0{,}01)$$
$$= 1{,}60$$

$$\sigma^2_{\bar{X}} = Var\left(\bar{X}\right) = E\left[\left(\bar{X} - E\left(\bar{X}\right)\right)^2\right]$$
$$= \sum_i \left(\bar{x}_i - 1{,}60\right)^2 \cdot P\left(\bar{X} = \bar{x}_i\right)$$
$$= (1 - 1{,}60)^2 (0{,}25) + \ \ldots\ + (3 - 1{,}60)^2 (0{,}01)$$
$$= 0{,}09 + \ \ldots\ + 0{,}019$$
$$= 0{,}22$$

5. Teniendo en cuenta la distribución de probabilidad del estadístico media muestral $\bar{X}$, Tabla 1.4, se tiene:

$$P\left(\bar{X} < 2\right) = P\left(\bar{X} = 1\right) + P\left(\bar{X} = 1{,}5\right)$$
$$= 0{,}25 + 0{,}40$$
$$= 0{,}65$$

6. Teniendo en cuenta la distribución de probabilidad del estadístico varianza muestral, S^2, dada en la Tabla 1.5, y procediendo de manera análoga a como lo hemos hecho para el estadístico media muestral, tendremos:

$$\mu_{s^2} = E\left[S^2\right] = \sum_i s_i^2 \cdot P\left(S^2 = s_i^2\right)$$

$$= 0{,}0(0{,}42) + 0{,}5(0{,}48) + 2{,}0(0{,}10)$$
$$= 0{,}44$$

$$\sigma_s^2 = Var(S^2) = E\left[\left(S^2 - E\left[S^2\right]\right)^2\right]$$
$$= \sum_i \left(s_i^2 - 0,44\right) P\left(S^2 = s_i^2\right)$$
$$= (0,0-0,44)^2\,(0,42) + (0,5-0,44)^2\,(0,48) + (2,0-0,44)^2\,(0,10)$$
$$= 0,0813 + 0,0017 + 0,2434$$
$$= 0,32$$

7. Basándonos en la distribución de probabilidad del estadístico varianza muestral S^2, Tabla 1.5, se tiene:

$$P\left(S^2 \leq 0,5\right) = P\left(S = 0,0\right) + P\left(S^2 = 0,5\right)$$
$$= 0,42 + 0,48$$
$$= 0,90$$

Con este ejemplo, se pone de manifiesto que incluso para muestras de tamaño pequeño y estadísticos con pocos valores posibles se hace pesado el obtener la distribución de probabilidad de los estadísticos muestrales. Para evitar esto en los siguientes apartados daremos algunos resultados que simplifican estos problemas.

1.6. MEDIA Y VARIANZA DE ALGUNOS ESTADÍSTICOS

En el Ejemplo 1.2 hemos obtenido:

— La media, μ, y varianza, σ^2, poblacional.

— Los estadísticos media $\bar{X}$ y varianza S^2 muestral.

— La media y varianza de los estadísticos media muestral, $\bar{X}$, y varianza muestral, S^2 para una muestra de tamaño $n = 2$.

Estos resultados se recogen en la Tabla 1.6, en donde se observa:

1.º Que $E\left[\bar{X}\right] = E\left[X\right]$,
es decir, que la media del estadístico media muestral es igual a la media de la población.

2.º Que $E\left[S^2\right] = Var\left(X\right)$,
es decir, que la media del estadístico varianza muestral es igual a la varianza de la población.

3.º Que $Var\left(\bar{X}\right) = \dfrac{Var(X)}{n}$,
es decir, que la varianza del estadístico media muestral es igual a la varianza de la población dividida por el tamaño de la muestra, n.

Tabla 1.6. *Media y varianza poblacional de los estadísticos media y varianza muestral del Ejemplo 1.3, para n = 2.*

	Población X	**Estadístico media muestral** $\bar{X}$	**Estadístico varianza muestral** S^2
Media	$\mu = E\left[X\right] = 1{,}6$	$\mu_{\bar{x}} = E\left[\bar{X}\right] = 1{,}6$	$\mu_{s^2} = E\left[S^2\right] = 0{,}44$
Varianza	$\sigma = Var\left(X\right) = 0{,}44$	$\sigma^2_{\bar{x}} = Var\left(\bar{X}\right) = 0{,}22$	$\sigma^2_{s^2} = Var\left(S^2\right) = 0{,}32$

Estos resultados no sólo se verifican para este ejemplo sino que se verifican en general, como veremos en los siguientes teoremas.

TEOREMA 1.1

Si $(X_1, ..., X_n)$ es una muestra aleatoria simple de tamaño n procedente de una población, descrita por la variable aleatoria X, con media $E[X] = \mu$ y varianza $Var\,(X) = \sigma^2$, entonces la **esperanza de la media muestral** es igual a la media de la población, μ, y la **varianza de la media muestral** es igual a la varianza poblacional, σ^2, dividida por n, es decir:

$$E\left[\bar{X}\right] = \mu \qquad \text{y} \qquad \text{Var}\left(\bar{X}\right) = \frac{\sigma^2}{n} \qquad [1.13]$$

Demostración:

Teniendo en cuenta la definición de muestra aleatoria simple, resulta que las variables aleatorias X_1, …, X_n son independientes, todas tienen la misma distribución de probabilidad que la población X y en consecuencia todas tienen la misma media y la misma varianza que la población X, es decir:

$$E\left[X_1\right] = \ \ldots\ = E\left[X_n\right] = E\left[X\right] = \mu$$

$$\text{Var}(X_1) = \ \ldots\ = \text{Var}(X_n) = \text{Var}(X) = \sigma^2$$

Luego si tenemos en cuenta las propiedades de los valores esperados, resulta que la media o esperanza matemática del estadístico media muestral será:

$$E\left[\bar{X}\right] = E\left[\frac{X_1 + \ \cdots\ + X_n}{n}\right]$$

$$= \frac{1}{n} E\left[X_1 + \ \cdots\ + X_n\right]$$

$$= \frac{1}{n}\left(E[X_1] + \cdots + E[X_n]\right)$$
$$= \frac{1}{n}\left(\mu + \cdots + \mu\right) = \frac{n\mu}{n} = \mu$$

Análogamente para la varianza, y dado que las variables aleatorias X_1, …, X_n son independientes, resulta:

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{X_1 + \ldots + X_n}{n}\right)$$
$$= \frac{1}{n^2}\text{Var}\left(X_1 + \ldots + X_n\right)$$
$$= \frac{1}{n^2}\left(\text{Var}(X_1) + \ldots + \text{Var}(X_n)\right)$$
$$= \frac{1}{n^2}\left(\sigma^2 + \ldots + \sigma^2\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

A la correspondiente **desviación típica** del estadístico $\bar{X}$ se le llama **error estándar de la media** y viene dado por:

$$\text{error estándar de la media muestral } \bar{X} = \frac{\sigma}{\sqrt{n}} \qquad [1.14]$$

Observando los resultados de la expresión [1.13] se pone de manifiesto que el valor central del estadístico media muestral es la media poblacional μ, y como la dispersión del estadístico media muestral $\bar{X}$ en torno a su media μ es:

$$\text{Var}(\bar{X}) = E\left[\left(\bar{X} - \mu^2\right)\right] = \frac{\sigma^2}{n}$$

resulta que cuanto mayor sea el tamaño muestral n menor será la Var($\bar{X}$), es decir, menor será la dispersión de $\bar{X}$ en torno a la media poblacional μ, y el valor observado del estadístico $\bar{X}$ estará más próximo a μ, lo cual nos permite decir que el estadístico media muestral $\bar{X}$ puede ser considerado como un buen estimador de la media poblacional μ.

En el Gráfico 1.2 se indica la distribución muestral del estadístico media muestral, $\bar{X}$, para muestras de tamaño $n = 25$ y $n = 110$ procedentes de una población normal $N(100, 6)$, en donde se observa que cada distribución muestral está centrada sobre la media poblacional, pero cuando el tamaño muestral aumenta la distribución muestral del estadístico media muestral está más concentrada en torno a la media de la población. En consecuencia el error estándar de la media muestral es una función decreciente del tamaño n de la muestra, y la probabilidad de que la media muestral difiera de la media poblacional en una cantidad fija, disminuye cuando el tamaño de la muestra crece.

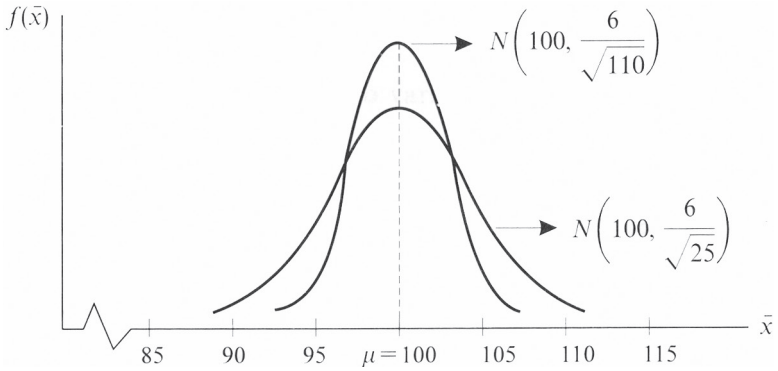

Gráfico 1.2. *Representación gráfica de las funciones de densidad del estadístico media muestral para muestras de tamaño n = 25 y n = 110, de una población N(100, 6).*

El aumento de la muestra tiene un límite, pues llega un momento que aunque el tamaño de la muestra siga aumentando la precisión prácticamente no aumenta. En efecto, supongamos una población con $\sigma = 12$ y calculamos la desviación estándar del estadístico $\bar{X}$ para diferentes valores de n, obteniéndose la Tabla 1.7.

Tabla 1.7. *Diferentes valores de la desviación estándar de $\bar{X}$ cuando $\sigma = 12$ para n = 5, 10, 20, 30, …*

Valores de n	5	10	20	30	40	50	60	70	80	90	100
Desviación estándar $\frac{\sigma}{\sqrt{n}}$	5,38	3,79	2,68	2,19	1,89	1,69	1,55	1,43	1,34	1,26	1,20

Observando los valores de la Tabla 1.7 y su correspondiente representación gráfica, Gráfico 1.2, se observa que la desviación estándar de $\bar{X}$ disminuye sustancialmente a medida que n aumenta, pero cuando n pasa de 40 esta disminución se reduce hasta tal extremo que cuando n sigue creciendo y toma valores superiores a 80 o 90 la desviación estándar de $\bar{X}$ prácticmente no disminuye. En consecuencia, podemos decir que si utilizamos el estadístico media muestral $\bar{X}$ para tener conocimiento o hacer inferencias sobre el parámetro media poblacional μ no es conveniente tomar muestras de tamaño demasiado grande pues el aumento del coste no compensa con la escasa disminución de la precisión.