Tema 5

Població i mostra

5.1 Población y muestra. Parámetros y estadísticos

- La estadística descriptiva es útil para conjuntos de datos con un número *pequeño* de individuos.
- ¿Qué podemos hacer si el conjunto de individuos de interés es *demasiado grande*? *Grande* puede significar, por ejemplo, que examinar a todos los individuos sea costoso en tiempo y/o dinero, o incluso *de tamaño infinito*. Ejemplo: encuesta de presupuestos familiares (EPF), encuestas preelectorales, resultados de un experimento repetible infinitas veces.
- Lo que haremos será seleccionar algunos de los individuos y observar sólo estos.
- **Población.** Conjunto de individuos en cuyas características está interesado el investigador.
- **Muestra.** Subconjunto de individuos de la población que serán finalmente observados.
- Las herramientas de la estadística descriptiva aplicadas a una muestra, aportan mucha información sobre esa muestra. Pero el objetivo del investigador es llegar a hacer afirmaciones sobre toda la población, no sólo sobre la muestra concreta que observa.
- Si la muestra ha sido seleccionada *adecuadamente*, la teoría de probabilidad permite inferir características de toda la población a partir del estudio descriptivo de la muestra.

- En ocasiones, la característica poblacional de interés se identifica con un valor numérico desconocido (parámetro), y éste puede ser *inferido* a partir del cálculo de algunos estadísticos descriptivos en una muestra extraída de la población.

- La inferencia estadística se vale de herramientas del cálculo de probabilidades y de la estadística descriptiva para construir afirmaciones válidas para toda la población a partir de la observación de una muestra representativa de ésta.

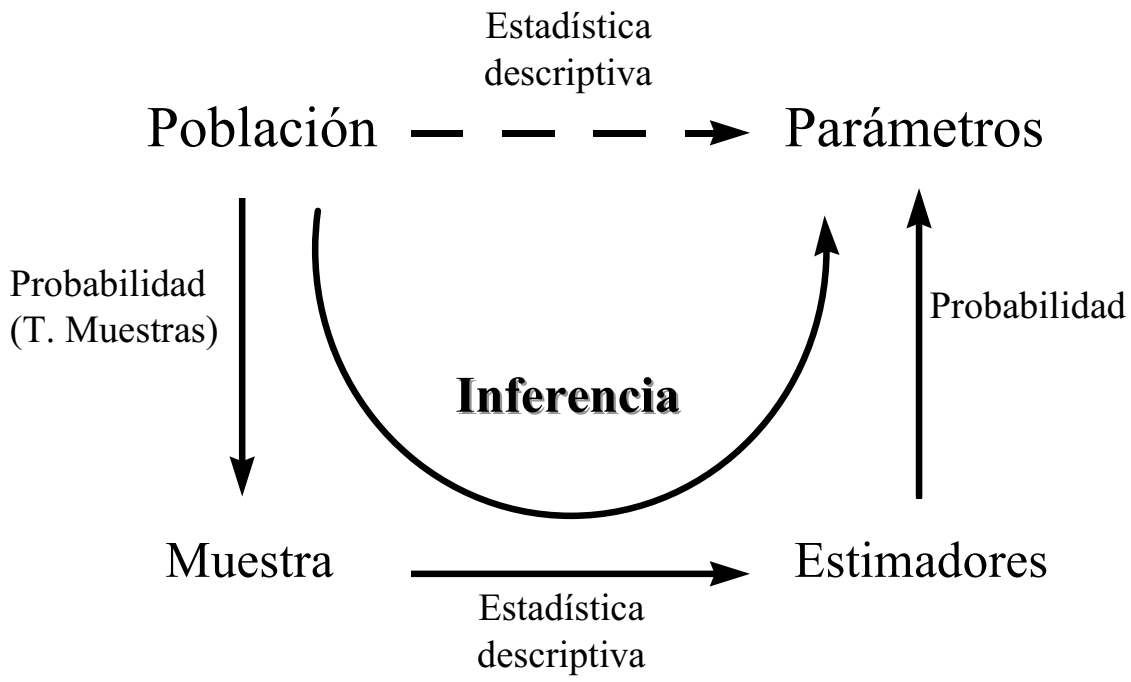


Figura 5.1: Esquema de la inferencia estadística

5.2 Objetivos de la inferencia: estimación, contraste y predicción

- **Estimación.** Una característica de interés de la población se identifica con una constante numérica (*parámetro*) cuyo valor (desconocido) se infiere a partir de la información obtenida de una muestra. Se puede dar únicamente una **estimación puntual** (un único valor numérico, como *estimador* del parámetro) o un **intervalo de confianza** (un intervalo que con una probabilidad dada incluirá el valor del parámetro).

- **Contraste de hipótesis.** Se realiza una afirmación sobre la población y se infiere su veracidad en base al estudio de dicha afirmación restringida a la muestra.

- **Predicción.** Se desea obtener información acerca del valor que toma una cierta característica en un individuo no observado de la población. Para ello se observa dicha característica (y posiblemente otras) en los individuos de la muestra.

5.3 Obtención de datos: muestras y experimentos

- El control estadístico del proceso de obtención de datos es un requerimiento fundamental para el posterior proceso de inferencia.

- El proceso de recogida de datos debe garantizar que la muestra sea **representativa** de la población de interés:

  1. Todos los individuos de la población tienen probabilidad positiva de pertenecer a la muestra.
  2. Se conoce la probabilidad de que cada individuo forme parte de la muestra (o al menos se puede estimar esa probabilidad).

- **Diseño de experimentos.** Se intentan controlar uno o varios factores (generalmente cualitativos) simultáneamente para estudiar su efecto en la característica de interés.

  Ejemplos: Evaluación de nuevos fármacos (grupo de control); experimentos en marketing (efecto de distintas presentaciones de un producto en las ventas).

- **Muestreo.** Errores muestrales y errores no muestrales.

  ¿La población que se muestrea es la que realmente interesa?

  ¿Es adecuada la forma en la que se plantea la pregunta? Efecto de los encuestadores. Preguntas comprometedoras.

  ¿Qué hacemos si hay muchos individuos que no responden a la pregunta?

- **Tipos de muestreo:**

  Muestreo aleatorio simple.

  Muestreo sistemático.

  Muestreo aleatorio estratificado

  Muestreo por conglomerados.

  Otros.

5.4 Muestreo aleatorio simple

- **Definición** Se desea extraer una muestra de tamaño n de una población de tamaño N. El muestreo aleatorio simple es un procedimientos aleatorio de selección de la muestra que garantiza que todos los subconjuntos de n elementos de la población tienen la misma probabilidad de ser elegidos como muestra.

- Muestreos con y sin reemplazamiento.

- **Estudio probabilístico de una m.a.s** (Aquí estableceremos una conexión clara entre las tres partes en las que dividimos el temario de estadística y probabilidad: la estadística descriptiva, la probabilidad y la inferencia estadística.)

  - La población en cuya característica x estamos interesados puede representarse por los N valores que esta característica toma en los individuos que la integran:

    $$\mathcal{X} = \{x^1, \ldots, x^N\}.$$

  - Supongamos, para simplificar la exposición, que la característica x es discreta y que puede tomar los k valores $c_1, \ldots, c_k$. Entonces, las técnicas de **estadística descriptiva** nos dicen que las características esenciales de la población $\mathcal{X}$ pueden expresarse (sin

pérdida de información) en la tabla de frecuencias (absolutas N_i y relativas f_i)

Valores de x	N_i	$f_i = N_i/N$
c_1	N_1	f_1
$\vdots$	$\vdots$	$\vdots$
c_k	N_k	f_k
Total	N	1

- Si conociésemos esa distribución de frecuencias podríamos calcular cualquier característica (*parámetro*) de la población $\mathcal{X}$.
- Examinemos los elementos que forman la m.a.s. de la población $\mathcal{X}$. Empecemos por la primera observación:

  > *Valor que toma la característica x en el individuo de la población que aleatoriamente es elegido como primer integrante de la m.a.s. extraída de la población $\mathcal{X}$*

  Este *valor* es una **variable aleatoria** (porque el primer individuo se elige aleatoriamente) que puede tomar cualquiera de los valores $x^1, \ldots, x^N$ (cualquier individuo puede ser elegido) con probabilidades $1/N$ (el m.a.s. garantiza equiprobabilidad).
  Dado que la característica x es discreta, realmente ese *primer valor de la m.a.s* sólo podrá tomar los k valores que puede tomar x, es decir, $c_1, \ldots, c_k$, y los tomará con probabilidades dadas por la regla de Laplace (*casos favorables partidos por casos posibles*):

  $$\Pr(\textit{primer valor en la m.a.s} = c_j) = \frac{N_j}{N} = f_j, \quad j = 1, \ldots, k.$$

  Así, la variable aleatoria *primer valor en la m.a.s*, a la que llamaremos X_1, es una v.a. discreta cuya **distribución de probabilidades** viene dada por

X_1	c_1	$\ldots$	c_k
p_j	f_1	$\ldots$	f_k

- El resto de los valores de la m.a.s., a los que llamaremos $X_2, \ldots, X_n$ son también variables aleatorias con idéntica distribución que X_1 (suponemos muestreo con reemplazamiento). Además estas n variables aleatorias son mutuamente independientes (de nuevo, debido al esquema de muestreo con reemplazamiento).

- Podemos decir entonces que $X_1, \ldots, X_n$ *son* n *copias independientes de una variable aleatoria* X *cuya distribución de probabilidad coincide con la distribución de frecuencias relativas de la característica* x *en la población* $\mathcal{X}$
- Identificaremos la población $\mathcal{X}$ con la variable aleatoria X, y diremos que $X_1, \ldots, X_n$ *es una muestra aleatoria simple de* X. Obsérvese que las *medidas descriptivas* de la población $\mathcal{X}$ (por ejemplo, la media aritmética de la característica x, $\bar{x} = \sum_{j=1}^{k} c_j f_j$) coinciden con los parámetros de la distribución de probabilidad de X (en el ejemplo, $E(X) = \sum_{j=1}^{k} c_j f_j$).
- En ocasiones no haremos referencia explícita a la población física en una de cuyas características estamos interesados, sino que directamente diremos que nos interesa estudiar la variable aleatoria X y para ello tomamos una m.a.s. de X.
- **Notación**

  $X_1, \ldots, X_n$ son variables aleatorias: los valores que tomará una cierta variable o característica X en los n individuos que acaben conformando una muestra aleatoria.

  $x_1, \ldots, x_n$ son los valores concretos de la característica X en los n individuos que finalmente forman parte de la muestra. Son constantes, valores concretos, dado que una vez realizado el muestreo ya no hay incertidumbre alguna sobre qué individuos son observados ni, como consecuencia, sobre los n valores de x que son medidos.

- Normalmente se hará la hipótesis de que N es tan grande en relación a n que, en la práctica, se puede suponer que la población tiene tamaño infinito y que, por lo tanto, los muestreos con y sin reemplazamiento son equivalentes.

- En el resto del curso se supondrá que las muestras de tamaño n son obtenidas mediante muestreo aleatorio simple de una población de tamaño infinito.

5.5 La función de distribución empírica

Sea la variable aleatoria X con función de distribución F. Consideramos una muestra aleatoria simple de tamaño n de X, es decir, $X_1, \ldots, X_n$ v.a.i.i.d. con distribución dada por F. Sea $x_1, \ldots, x_n$ una realización de esa m.a.s.

Se llama FUNCIÓN DE DISTRIBUCIÓN EMPÍRICA a la función

$$F_n(x) = \frac{1}{n}\#\{x_i \leq x : i = 1 \dots n\} = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(x_i),$$

donde

$$I_{(-\infty,x]}(x_i) = \begin{cases} 1, & \text{si } x_i \leq x \\ 0, & \text{si } x_i > x, \end{cases}$$

que a cada número real x le asigna la proporción de valores observados que son menores o iguales que x.

Es inmediato comprobar que la función F_n así definida es una función de distribución:

1. $F_n(x) \in [0, 1]$ para todo $x \in \mathbb{R}$.
2. F_n es continua por la derecha.
3. F_n es no decreciente.
4. $\lim_{x \Rightarrow -\infty} F_n(x) = 0$.
5. $\lim_{x \Rightarrow \infty} F_n(x) = 1$.

Concretamente, F_n es la función de distribución de una variable aleatoria discreta (que podemos llamar X_e) que pone masa $1/n$ en cada uno de los n puntos x_i observados:

x_i	x_1	x_2	$\cdots$	x_n
$p_i = \Pr(X_e = x_i)$	$1/n$	$1/n$	$\cdots$	$1/n$

A la distribución de X_e se le llama DISTRIBUCIÓN EMPÍRICA asociada al conjunto de valores $\{x_1, \dots, x_n\}$.

Obsérvese que si fijamos el valor de x y dejamos variar la muestra, lo que obtenemos es una variable aleatoria. En efecto, se tiene entonces que

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i),$$

donde

$$I_{(-\infty,x]}(X_i) = \begin{cases} 1, & \text{si } X_i \leq x \\ 0, & \text{si } X_i > x \end{cases}$$

y, por lo tanto, cada término $I_{(-\infty,x]}(X_i)$ es una variable aleatoria de Bernoulli con probabilidad de éxito

$$p = \Pr(I_{(-\infty,x]}(X_i) = 1) = \Pr(X_i \leq x) = F(x).$$

De ahí se deduce que F_n es una variable aleatoria y que $nF_n(x)$ tiene distribución binomial con parámetros n y $p = F(x)$.

De lo anterior se sigue que la función de distribución empírica es una *función de distribución aleatoria*.

El siguiente teorema recoge algunas de las propiedades de la función de distribución empírica .

Teorema 5.5.1. *Sea $\{X_n\}_{n\geq 1}$, sucesión de variables aleatorias independientes e idénticamente distribuidas definidas en el espacio de probabilidad $(\Omega, \mathcal{A}, P)$ con función de distribución común F. Se denota por F_n la función de distribución empírica obtenida de las n primeras variables aleatorias $X_1, \ldots, X_n$. Sea $x \in \mathbb{R}$. Se verifica lo siguiente:*

(a) $\Pr(F_n(x) = \frac{j}{n}) = \binom{n}{j} F(x)^j (1 - F(x))^{n-j}$, $j = 0, \ldots, n$.

(b) $E(F_n(x)) = F(x)$, $Var(F_n(x)) = (1/n)F(x)(1 - F(x))$.

(c) $F_n(x) \to F(x)$ *casi seguro.*

(d)

$$\frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F(x)(1 - F(x))}} \to_D Z,$$

donde Z es una variable aleatoria con distribución normal estándar y la convergencia es en distribución.

Demostració: Los apartados (a) y (b) son consecuencia inmediata del hecho de que $nF_n(x) \sim B(n, p = F(x))$. Por otro lado, si definimos $Y_i = I_{(-\infty,x]}(X_i)$, se tiene que $F_n(x) = \bar{Y}_n$, la media aritmética de las variables aleatorias $Y_1, \ldots, Y_n$. Así, el apartado (c) es una aplicación inmediata de la ley fuerte de los grandes números y el apartado (d) es consecuencia del teorema central de límite. □

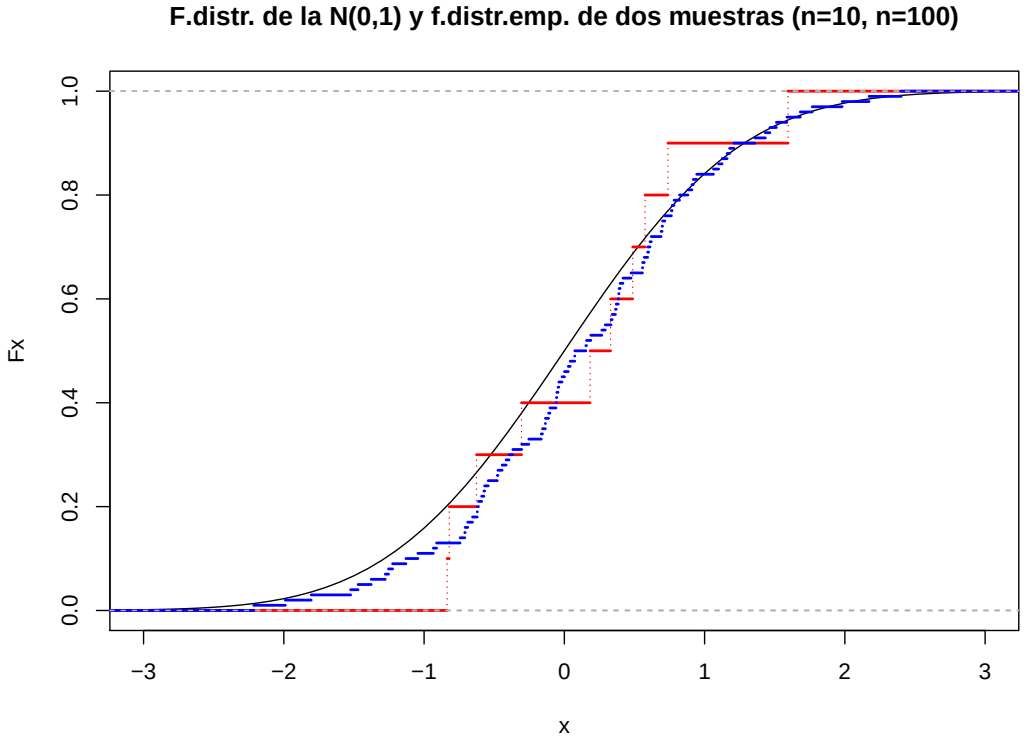
El siguiente teorema refuerza el resultado (c) anterior, puesto que afirma que la convergencia de $F_n(x)$ a $F(x)$ se da uniformemente.

Teorema 5.5.2 (Teorema de Glivenko-Cantelli)**.** *Sea $\{X_n\}_{n\geq 1}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas definidas en el espacio de probabilidad $(\Omega, \mathcal{A}, P)$ con función de distribución común F. Se denota por F_n la función de distribución empírica obtenida de las n primeras variables aleatorias $X_1, \ldots, X_n$. Entonces,*

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \Rightarrow 0 \text{ casi seguro.}$$

Exemple 5.5.1.

En la figura siguiente se muestra la función de distribución de una variable aleatoria $N(0,1)$ y la función de distribución empírica de dos muestras de esa variable aleatoria una de tamaño $n = 10$ (la más alejada de la teórica) y la otra de tamaño $n = 100$. Se aprecia que cuando n crece la proximidad entre la función de distribución empírica y la teórica es cada vez mayor.



□