Tema 1: Distribuciones en el muestreo

(transparencias de A. Jach `http://www.est.uc3m.es/ajach/`)

- Muestras aleatorias
- Estadísticos
- Concepto de distribución muestral
- Media muestral
  - Distribución muestral en el caso normal
  - Distribución muestral para muestras grandes: TCL
- Cuasi-varianza muestral
  - Distribución muestral en el caso normal: χ^2 y Lema de Fisher

Muestras aleatorias

Definición 1.

Se llama población al conjunto de individuos de interés.
Por ejemplo:

- *Todos los estudiantes de la UC3M*
- *Todos los peces del mar Mediterráneo*
- *Todos los pacientes del hospital de Getafe*
- *Todas las bombillas de bajo consumo producidas por Philips*

Definición 2.

Se llama muestra a un subconjunto de la población, elegido para estudiar las propiedades de la población original.

Muestras aleatorias

¿Por qué utilizamos muestras?

- Una muestra es mucho más fácil de conseguir que un censo (censo: información sobre todos los miembros de una población), es decir, es mucho más barato y requiere mucho menos tiempo.

- El censo de una población es muchas veces imposible de realizar. Por ejemplo: todos los peces del mar Mediterráneo.

Muestras aleatorias

Definición 3.

Una muestra aleatoria simple (M. A. S.) es una colección de n individuos donde cada uno de ellos es elegido al azar y todos tienen la misma probabilidad de ser elegidos.

De forma rigurosa, sea X la v. a. de estudio en la población, con distribución F. Una muestra aleatoria simple de tamaño n es un conjunto de n v. a. $X_1, X_2, \ldots, X_n$ tales que:

- *$X_1, X_2, \ldots, X_n$ tienen todas distribución F ($X_i \sim F \quad \forall i$).*
- *$X_1, X_2, \ldots, X_n$ son independientes entre sí.*

Cada valor concreto $(x_1, x_2, \ldots, x_n)$ de dicha M. A. S. se denomina muestra particular.

Muestras aleatorias

NOTACIÓN:

- $X \sim F \quad \Leftrightarrow \quad X$ tiene distribución F .
- $X_i \sim F \quad i.i.d. \quad \Leftrightarrow \quad$ todas las $X_i, i = 1, \ldots, n$ son **i**ndependientes entre sí y con **i**déntica **d**istribución $F \Leftrightarrow X_1, X_2, \ldots, X_n$ es una m. a. s.

i. i. d. = independientes e idénticamente distribuídas

Observación: F representa la distribución de X y puede denotar indistintamente:

- La función de probabilidad de X (para v. a. discretas); p. ej. : $Bernoulli(p), Binomial(n, p), Poisson(\lambda), \ldots$
- La función de densidad de X (para v. a. continuas); p. ej. : $N(\mu, \sigma), Exp(\lambda), U(a, b), \ldots$
- La función de distribución de X (para v. a. discretas o continuas): $F(x) = P(X \leq x)$

Estadísticos

Definición 4.
Un parámetro (poblacional) es un número (FIJO, NO ALEATORIO, CONSTANTE, ÚNICO) que describe alguna característica de la población (generalmente es desconocido).

Ejemplo 1.
Si queremos estudiar la altura de los estudiantes de la UC3M, un parámetro de interés es la media poblacional μ (desconocida).

Definición 5.
Un estadístico (muestral) es un número que se calcula a partir de los datos de una muestra (SU VALOR CAMBIA DE UNA MUESTRA A OTRA).
De forma rigurosa, un estadístico es una función real de la muestra aleatoria $X_1, X_2, \ldots, X_n$. Por tanto, un estadístico es una variable aleatoria.

Ejemplo 1 (cont.) Para aproximar μ usamos el estadístico $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
Para una muestra particular $(x_1, x_2, \ldots, x_n)$, calculamos $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

¡OJO! : $\overline{X} \neq \overline{x}$

Distribución Muestral

- Supongamos que en una población seleccionamos todas las posibles muestras de tamaño n.
- Para cada muestra particular calculamos el mismo estadístico muestral (por ejemplo la media muestral).
- La distribución de probabilidad de todos los posibles valores del estadístico muestral se llama distribución muestral.

Ejemplo 2.

Ejercicio con la Tabla de Números Aleatorios (disponible en la página web) Entrega puntuable.

Definición 6.

Sea T un estadístico basado en la muestra aleatoria $X_1, X_2, \ldots, X_n$. La distribución de la v. a. $T(X_1, X_2, \ldots, X_n)$ se denomina distribución muestral del estadístico.

Distribución Muestral

Ejemplo 3.

Distribución muestral del estadístico "proporción de chicas" en una m.a.s. de tamaño 5 para la población "alumnos en clase".

- *El parámetro poblacional de interés es* $p = \frac{\sharp \text{ chicas}}{\sharp \text{ total de alumnos}}$

  (En este caso podríamos calcular directamente p y no necesitaríamos estimarlo a partir de una muestra. **Es un ejemplo poco realista***)*

- *Sea* $X_1, \ldots, X_5$ *una m.a.s. de tamaño* 5*:* $X_i = 1$ *si chica y* $X_i = 0$ *si no.*

- *El estadístico es:*

$$T(X_1, \ldots, X_5) = \frac{\sum_{i=1}^{5} X_i}{5}$$

- *¿Qué tipo de variable aleatoria es T?*

- *¿Qué valores toma y con qué probabilidades?* $\Leftrightarrow$ *¿Cuál es su distribución?*

La media muestral y su distribución

Definición 7.

Sea $X_1, \ldots, X_n$ una m.a.s. de una población con media poblacional μ y varianza poblacional σ^2. Se define la media muestral como el estadístico

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

¿Cuál es la distribución muestral de la media muestral?

Empecemos calculando su esperanza: $X_1, \ldots, X_n$ m.a.s. $\Leftrightarrow$ $X_i \sim F$ i.i.d.

$$\begin{aligned}
E[\overline{X}] &= E[\tfrac{1}{n} \textstyle\sum_{i=1}^{n} X_i] && \text{definición} \\
&= E[\tfrac{1}{n}(X_1 + X_2 + \cdots + X_n)] && \text{desplegamos la suma} \\
&= \tfrac{1}{n}(E[X_1] + E[X_2] + \cdots + E[X_n]) && \text{linearidad de E} \\
&= \tfrac{1}{n}(\mu + \mu + \cdots + \mu) && E[X_i] = \mu \\
&= \tfrac{1}{n}(n \cdot \mu) = \mu
\end{aligned}$$

La media muestral y su distribución

Hemos demostrado que

$$E[\overline{X}] = \mu$$

Observaciones:

- La distribución del estadístico media muestral está centrada en la media poblacional μ
- Una media muestral particular $\overline{x}$ (obtenida a partir de una muestra particular $x_1, \ldots, x_n$) puede ser mayor o menor que μ, sin embargo, en media, la media muestral estará cerca de la media poblacional. (Ejemplo Números Aleatorios)

La media muestral y su distribución

Calculemos ahora la varianza de $\overline{X}$:

$$
\begin{aligned}
Var[\overline{X}] &= Var[\tfrac{1}{n}\sum_{i=1}^{n} X_i] && \text{definición} \\
&= Var[\tfrac{1}{n}(X_1 + X_2 + \cdots + X_n)] && \text{desplegamos la suma} \\
&= \tfrac{1}{n^2} Var[(X_1 + X_2 + \cdots + X_n)] && \text{propiedad de Var} \\
&= \tfrac{1}{n^2}(Var[X_1] + Var[X_2] + \cdots + Var[X_n]) && \text{INDEPENDENCIA} \\
&= \tfrac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots + \sigma^2) && Var[X_i] = \sigma^2 \\
&= \tfrac{1}{n^2}(n \cdot \sigma^2) = \tfrac{\sigma^2}{n}
\end{aligned}
$$

La media muestral y su distribución

Hemos demostrado que

$$Var[\overline{X}] = \frac{\sigma^2}{n} \quad \Leftrightarrow \quad DT[\overline{X}] = \frac{\sigma}{\sqrt{n}}$$

Observaciones:

- En general para medir la dispersión de cualquier estadístico, y en particular la dispersión de la media muestral, se suele usar la Desviación Típica (DT) y no la Varianza (Var).
- La distribución de la media muestral tiene menor dispersión que la variable original (en una proporción de $\sqrt{n}$ a 1).

Distribución muestral de $\overline{X}$ en el caso normal

Teorema 1.

Sea $X_1, \ldots, X_n$ una m.a.s. de una población con distribución normal de media μ y varianza σ^2, es decir $X_i \sim N(\mu, \sigma)$ i.i.d. Entonces

$$\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Esto se cumple porque la combinación lineal de variables aleatorias con distribución normal tiene también distribución normal.

En la práctica:

Tipificamos $\overline{X}$ para obtener una v.a. con distribución $N(0,1)$:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

y utilizamos la Tabla de la Normal para calcular probabilidades.

Distribución muestral de $\overline{X}$ en el caso normal

Ejemplo 4.

En una fábrica de embalaje de cereales, la cantidad de cereal que se coloca en cada caja se distribuye normalmente con media 500 gr y desviación típica 20 gr. Frecuentemente se realizan controles de calidad, durante los cuales se eligen 15 cajas al azar y de forma independiente para pesar su contenido. Si el peso medio de las 15 cajas es superior a 510 gr o inferior a 490 gr se detiene la cadena de embalaje para proceder al equilibrado de las máquinas. ¿Cuál es la probabilidad de que se detenga la cadena durante uno de estos controles?

(P. Gil y S. Montes. Univ. de Oviedo.)

$X =$ peso en gramos de una caja de cereales; $X \sim N(500, 20)$

$X_1, X_2, \ldots, X_{15}$ es una m.a.s. de $X \Leftrightarrow X_i \sim N(500, 20)$ i.i.d.

Por lo tanto $\overline{X} = \frac{1}{15} \sum_{i=1}^{15} X_i \sim N(500, \frac{20}{\sqrt{15}})$

Distribución muestral de $\overline{X}$ en el caso normal

Ejemplo 5.

Supongamos que $X \sim N(\mu, \sigma)$ *y que* $X_1, \ldots, X_n$ *es una m.a.s. proveniente de* X*.*

- *¿*$P(X < \mu)$ *es mayor, menor o igual que* $P(\overline{X} < \mu)$*?*
- *¿*$P(X < \mu + 1)$ *es mayor, menor o igual que* $P(\overline{X} < \mu + 1)$*?*
- *¿*$P(X > \mu + 1)$ *es mayor, menor o igual que* $P(\overline{X} > \mu + 1)$*?*

Distribución muestral de $\overline{X}$ para muestras grandes

Teorema 2.

Sean $X_1, X_2, \ldots, X_n$ v.a. i.i.d. con media μ y varianza σ^2. Si n es suficientemente grande, entonces $\overline{X}$ es aproximadamente normal con media μ y varianza σ^2/n. Escribimos

$$\overline{X} \overset{A}{\sim} N(\mu, \sigma/\sqrt{n})$$

Teorema Central del Límite (TCL)

Observaciones:

- El TCL se aplica para cualquier tipo de distribución, no hace falta normalidad. De hecho, si la distribución de origen es normal, sabemos que $\overline{X} \sim N(\mu, \sigma/\sqrt{n})$ de manera exacta independientemente del tamaño de muestra.
- En la práctica aplicamos el TCL para $n > 30$ (aunque este corte varía en la bibliografía)
- Si la distribución de origen es muy diferente de la normal, se suele exigir un tamaño de muestra mayor para aplicar el TCL

Distribución muestral de $\overline{X}$ para muestras grandes

Ejemplo 6.

Supongamos que el aumento anual del salario de los altos cargos en España se distribuye normalmente con media 12.2 % y desviación típica 3.6 %. Se toma una muestra aleatoria simple de nueve observaciones.

- *¿Cuál es la distribución de $\overline{X}$? ¿Es una distribución exacta o aproximada?*
- *¿Cuál es la probabilidad de que la media muestral sea menor del 10 %?*
- *¿Serían los cálculos válidos sin la hipótesis de normalidad? ¿Y sin la hipótesis de m.a.s.?*

Distribución muestral de $\overline{X}$ para muestras grandes

Definición 8.

Sea $X \sim B(p)$ la v.a. que mide la presencia de una característica de interés en los individuos de una población, donde p es la proporción de individuos con esa característica (p es un parámetro poblacional).
Sea $X_1, X_2, \ldots, X_n$ una m.a.s. a partir de $X \Leftrightarrow X_1, X_2, \ldots, X_n \sim B(p)$ i.i.d.

El estadístico proporción muestral se define como $\hat{p} = \dfrac{\sum_{i=1}^{n} X_i}{n}$ **(Ejemplo 3)**

Observaciones:

- $\hat{p}$ no es más que el estadístico $\overline{X}$ cuando la distribución de X es $B(p)$.
- Si $X \sim B(p) \Rightarrow E(X) = p$ y $V(X) = p(1-p)$.

  Entonces, al igual que para $\overline{X}$ tenemos que: $E(\hat{p}) = p \quad V(\hat{p}) = \dfrac{p(1-p)}{n}$
- Si n es grande podemos aplicar el TCL: $\hat{p} \overset{A}{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

Distribución muestral de $\overline{X}$ para muestras grandes

Ejemplo 7.

Expertos de La Sexta calculan que un 65% *de los hombres de edades comprendidas entre los 30 y los 70 años ven el partido de liga del sábado por la noche. Se toma una muestra aleatoria simple de 200 individuos de esa población. Si la hipótesis es cierta, ¿cuál es la probabilidad de que más del* 66% *de los encuestados vea el partido?*

$$X = \begin{cases} 1 & \text{si el individuo ve el partido} \\ 0 & \text{si no} \end{cases} \quad ; \quad X \sim B(0.7)$$

$X_1, X_2, \ldots, X_{200}$ es una m.a.s. de $X \Leftrightarrow X_i \sim B(0.7)$ i.i.d.

Sea $\hat{p} = \overline{X} = \dfrac{1}{200}\displaystyle\sum_{i=1}^{200} X_i$: $\qquad E(\hat{p}) = 0.7$ y $DT(\hat{p}) = \sqrt{\dfrac{0.7 \cdot 0.3}{200}} = 0.0324.$

Como $n = 200 >> 30$, podemos aplicar el TCL y tenemos que

$$\hat{p} \overset{A}{\sim} N\left(0.7, 0.0324\right)$$

La cuasi-varianza muestral y su distribución

Definición 9.

Sea $X_1, \ldots, X_n$ una m.a.s. de una población con media poblacional μ y varianza poblacional σ^2. Se define la cuasivarianza muestral como el estadístico

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

¿Cuál es la distribución muestral de la cuasi-varianza muestral?
¿Por qué utilizamos la cuasi-varianza y no la varianza?

Empecemos con la esperanza: $X_1, \ldots, X_n$ m.a.s. $\Leftrightarrow$ $X_i \sim F$ i.i.d.

Se puede demostrar que: $E[S^2] = \sigma^2$

Si utilizásemos la varianza muestral $V^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{n-1}{n} S^2$, su esperanza sería $E[V^2] = \frac{n-1}{n} \sigma^2$, y no sería un estadístico centrado.

La cuasi-varianza muestral y su distribución

Por lo tanto, independientemente de cual sea la distribución de X,

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

Observación:

- La distribución del estadístico cuasi-varianza muestral está centrada en la varianza poblacional σ^2
- Una cuasi-varianza muestral particular s^2 (obtenida a partir de una muestra particular $x_1, \ldots, x_n$) puede ser mayor o menor que σ^2, sin embargo, en media, la cuasi-varianza muestral estará cerca de la varianza poblacional.

El valor de la varianza de S^2 sí depende del tipo de distribución que estemos considerando. Nosotros la obtendremos sólo para el caso en el que la distribución de origen sea normal.

Distribución muestral de S^2 en el caso normal

Definición 10.

Sean $X_1, \ldots, X_k$ k v.a. i.i.d. con distribución $N(0,1)$. Se define la distribución χ^2 (ki-dos o ki-cuadrado) con k grados de libertad (χ^2_k) como la distribución de la v.a.

$$W = \sum_{i=1}^{k} X_i^2$$

La esperanza y la varianza de esta distribución son:

$$E[W] = k \quad Var[W] = 2k$$

Ejemplo 8.

Sea W una v.a. con distribución χ^2_{10}. A partir de la tabla de la χ^2, tenemos:

- $P(W > 4.87) = 0.9$
- $P(W < 3.94) = 1 - P(W > 3.94) = 1 - 0.95 = 0.05$
- $P(3.25 < W < 18.31) = P(W > 3.25) - P(W > 18.31) = 0.975 - 0.05 = 0.925$

Distribución muestral de S^2 en el caso normal

Lema 1.

Sean $X_1, X_2, \ldots, X_n$ v.a. i.i.d. $N(\mu, \sigma)$. Entonces:

- $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ *(ya lo sabíamos)* **Lema de Fisher**
- $\frac{n-1}{\sigma^2} S^2 \left(= \sum_{i=1}^{n} (\frac{X_i - \overline{X}}{\sigma})^2 \right) \sim \chi^2_{n-1}$
- $\overline{X}$ *y* S^2 *son independientes.*

Como corolario podemos calcular $Var[S^2]$ en el caso normal ($\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$):

$$
\begin{aligned}
Var[S^2] &= Var[\tfrac{\sigma^2}{n-1} \tfrac{n-1}{\sigma^2} S^2] \\
&= \tfrac{\sigma^4}{(n-1)^2} Var[\tfrac{n-1}{\sigma^2} S^2] && \text{propiedad de Var} \\
&= \tfrac{\sigma^4}{(n-1)^2} 2(n-1) && Var[\tfrac{n-1}{\sigma^2} S^2] = 2(n-1) \\
&= \tfrac{2\sigma^4}{(n-1)}
\end{aligned}
$$

Distribución muestral de S^2 en el caso normal

Ejemplo 9.

Un economista opina que el incremento anual (en porcentaje) del salario de los trabajadores de un banco sigue una distribución normal con desviación típica 3.37. Se toma una muestra aleatoria simple de 16 trabajadores del banco. Si la hipótesis del economista es cierta, calcula la probabilidad de que la cuasi-desviación típica sea mayor que 2.89.

$X =$ incremento salarial anual (en porcentaje); $X \sim N(\mu, 3.37)$

$X_1, X_2, \ldots, X_{16}$ es una m.a.s. de $X \Leftrightarrow X_i \sim N(\mu, 3.37)$ i.i.d.

Lema de Fisher: $\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1} \Rightarrow \frac{15}{3.37^2} S^2 \sim \chi^2_{15}$

$$P(S > 2.89) = P(S^2 > 2.89^2) = P\left(\frac{15}{3.37^2} S^2 > \frac{15}{3.37^2} 2.89^2\right)$$
$$\overset{W \sim \chi^2_{15}}{=} P(W > 11.02) \in (0.75, 0.9) \quad (\text{ó} \sim 0.75)$$

Distribución muestral de $\overline{X}$: Resumen

Objetivo: estimar la media poblacional μ utilizando el estadístico $\overline{X}$ a partir de n observaciones $X_1, \ldots, X_n$.

$X_1, \ldots, X_n$ ¿son i.i.d.?

NO ↗ ⊗

SÍ ↘ $E[\overline{X}] = \mu$, $Var[\overline{X}] = \frac{\sigma^2}{n}$

⊗ ¿n grande?

NO ↗ **TCL:** $\overline{X} \overset{A}{\sim} N(\mu, \frac{\sigma}{\sqrt{n}})$

SÍ ↘ ¿$X \sim N$?

NO ↗ ⊗

SÍ ↘ $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

Caso particular:

$$X_1, \ldots, X_n \sim B(p) \text{ } i.i.d. \Rightarrow \hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}; \quad \mu = p, \ \sigma^2 = p(1-p).$$

Distribución muestral de S^2: Resumen

Objetivo: estimar la varianza poblacional σ^2 utilizando el estadístico S^2 a partir de n observaciones $X_1, \ldots, X_n$.