CAPÍTULO VI: Muestras Aleatorias - Distribución de Muestreo

En muchos problemas de inferencia estadística, es imposible o poco práctico observar toda la población. Por ello, es necesario utilizar una muestra de observaciones tomadas de la población de interés con objeto de obtener conclusiones sobre ella. Por ejemplo, si se intenta determinar la duración promedio de una cierta marca de focos sería imposible probarlos todos si no se quiere dejar de vender ninguno y además, requeriría mucho tiempo. Los costos exorbitantes también pueden ser un factor prohibitivo al intentar estudiar la población entera. Por ejemplo, si un biólogo desea evaluar algunas características de una determinada especie de pez del mar Argentino resultaría imposible medir a toda la población, no sólo por una cuestión económica, sino también, por que difícilmente se conozca el tamaño de la población, se posean los medios suficientes para lograr acceder a toda la población, etc.
Es importante obtener muestras representativas de la población; con frecuencia se está tentado a elegir una muestra seleccionando a los miembros más convenientes de la población. Tal procedimiento puede conducir a inferencias erróneas respecto de la misma. Cualquier procedimiento de muestreo que produce inferencias que en forma consistente sobreestiman o subestiman alguna característica de la población es un procedimiento sesgado. Para eliminar cualquier posibilidad de sesgo en el procedimiento muestral es conveniente seleccionar una muestra aleatoria, en el sentido de que las observaciones se realicen independientemente y al azar.

Definición intuitiva:
Si se seleccionan *n* elementos de una población de modo tal que cada conjunto de *n* elementos de la población tenga la misma probabilidad de ser seleccionado, se dice que los *n* elementos constituyen una muestra aleatoria.

En términos estrictos esta definición corresponde a un **muestra aleatoria simple**. Existen diferentes tipos de muestras aleatorias. Por ejemplo, una **muestra aleatoria estratificada** se obtiene dividiendo a la población en grupos (estratos) y seleccionando una muestra aleatoria de cada grupo; una **muestra sistemática** se obtiene seleccionando sistemáticamente cada *k*-ésimo elemento de la población, etc.

Sea X una característica medible de interés, por ejemplo, nivel de concentración de un contaminante, la demanda de un producto o el tiempo de espera de un servicio, y f(x) su función de densidad de probabilidad. Las observaciones de una muestra se obtienen al observar la característica medible X de manera independiente bajo las mismas condiciones, *n* veces. Sea X_i la v.a. que representa la *i*-ésima observación de la v.a. X. Entonces $X_1, X_2, \ldots, X_n$ constituyen una muestra aleatoria donde los valores numéricos obtenidos son $x_1, x_2, \ldots, x_n$. Las variables aleatorias que componen una muestra aleatoria son independientes, con la misma distribución de probabilidad que la v.a. X, debido a que cada observación se obtiene bajo las mismas condiciones.

Definición formal:
Las v.a. $X_1, X_2, \ldots, X_n$ constituyen una **muestra aleatoria** de tamaño *n* de la v.a. X, si:

(1) las X_i son v.a. independientes.
(2) Todas las X_i tienen la misma distribución de probabilidad que la v.a. X.

El siguiente ejemplo sirve para ilustrar esta definición:

Se desea investigar el contenido de nicotina (mg) de una determinada marca de cigarrillos recientemente lanzada al mercado. Se sabe que, X = "contenido de nicotina de un cigarrillo de esa marca" es una v.a. con distribución normal. Se espera que las observaciones del contenido de nicotina $X_1, X_2, \ldots, X_n$ en una muestra aleatoria de *n* cigarrillos, sean v.a. independientes con la misma distribución normal que la v.a. X. Después de recopilar los datos, los valores numéricos de los contenidos de nicotina observados se denotan por: $x_1, x_2, \ldots, x_n$. Para una muestra aleatoria de $n = 6$ cigarrillos de esa marca los contenidos de nicotina fueron: $x_1 = 2.3$, $x_2 = 2.7$, $x_3 = 2.5$, $x_4 = 2.9$, $x_5 = 3.1$ y $x_6 = 1.9$ (mg).

El propósito principal de una muestra aleatoria es obtener información sobre los parámetros no conocidos de la población.

Definición:
Cualquier función de las variables aleatorias que componen una muestra aleatoria se llama **estadístico.**

Obsérvese que un estadístico es una v.a. por ser función de variables aleatorias.

Ejemplos de estadísticos:
1. Si $X_1, X_2, \ldots, X_n$ representa una muestra aleatoria de tamaño *n* de una v.a. X se define **media muestral**, y se denota: $\overline{X}$, al estadístico:

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

2. Si $X_1, X_2, \ldots, X_n$ representa una muestra aleatoria de tamaño *n* de una v.a. X, se define **varianza muestral** y se denota: S^2, al estadístico:

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

Ejemplo:
Considérese la v.a. X = "contenido de nicotina de un cigarrillo de cierta marca" y la muestra aleatoria extraída de tamaño $n = 6$, formada por los valores: $x_1 = 2.3$, $x_2 = 2.7$, $x_3 = 2.5$, $x_4 = 2.9$, $x_5 = 3.1$ y $x_6 = 1.9$ (mg). El valor del contenido medio de esa muestra es: $\overline{x} = 2.57$ mg y el valor de la varianza muestral es: $s^2 = 0.432\ mg^2$.

Puesto que un estadístico es una v.a. tiene una distribución de probabilidad, que se conoce como **distribución de muestreo.** La distribución de muestreo de un estadístico depende de la distribución de probabilidad de la población (con v.a. X), del tamaño de la muestra y del método utilizado para seleccionar ésta.

A continuación se presentan las distribuciones de muestreo más importantes:

Media Muestral

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una v.a. X con media finita μ y varianza finita σ^2. Por lo tanto, las *n* variables aleatorias son independientes y poseen la misma distribución que la v.a. X, es

decir, $E(X_i) = \mu$ y $V(X_i) = \sigma^2$. Luego el valor esperado y la varianza del estadístico media muestral, $\overline{X}$, son respectivamente:

$$E(\overline{X}) = E\left(\frac{X_1 + ... + X_n}{n}\right) = E\left(\frac{1}{n}(X_1 + ... + X_n)\right) = \frac{1}{n}(E(X_1) + ... + E(X_n)) = \frac{1}{n} n\mu = \mu$$

y

$$V(\overline{X}) = V\left(\frac{X_1 + ... + X_n}{n}\right) = V(\frac{1}{n}(X_1 + ... + X_n)) = \frac{1}{n^2}(V(X_1) + ... + V(X_n)) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

La importancia de este resultado radica en que es válido sin importar la distribución de probabilidad de la v.a. X de interés, siempre que la varianza tenga un valor finito.

De la definición de $V(\overline{X})$ se deduce que el desvío estándar de $\overline{X}$ es:

$$\sqrt{V(\overline{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sqrt{\sigma^2}}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$$

que recibe el nombre de **error estándar de la media**. Conforme el tamaño de la muestra crece, el desvío estándar, y por lo tanto la varianza, de $\overline{X}$ disminuye.

Obsérvese que si se calcula la desvío estándar de $\overline{X}$ para distintos valores de *n*, para un valor fijo de σ, este desvío estándar sufre un disminución sustancial en su valor conforme *n* toma valores cada vez más grandes, pero si *n* es mayor de 30 ó 40 este comportamiento cesa.

Proposición:
Sea X una v.a. normal con $E(X) = \mu$ y $V(X) = \sigma^2$, y sea $X_1, X_2, ..., X_n$ una muestra aleatoria de la v.a. X, que consiste en *n* variables aleatorias independientes normalmente distribuidas, con $E(X_i) = \mu$ y $V(X_i) = \sigma^2$ finita, i = 1,2, ..., *n*. Entonces la distribución de la media muestral, $\overline{X}$, es normal con media μ y varianza $\frac{\sigma^2}{n}$.

Ejemplo:
El tiempo que un pasajero invierte esperando en un punto de revisión de un aeropuerto es una v.a. con distribución normal con media 8.8 minutos y desvío estándar de 2 minutos. Si se observan al azar 25 pasajeros, encontrar la probabilidad de que el tiempo de espera promedio en la fila para estos pasajeros sea:
(a) menor que 10 minutos.
(b) Entre 7.5 y 9 minutos.
(c) Mayor de 9.5 minutos.

Sea X = "tiempo que un pasajero invierte esperando en un punto de revisión de un aeropuerto" la v.a., $X \sim N(\mu = 8.8, \sigma^2 = 2^2)$ y sea $\overline{X}$ = "tiempo promedio que un pasajero invierte esperando en un punto de revisión de un aeropuerto entre 25 pasajeros seleccionados al azar", por la proposición anterior, $\overline{X} \sim N(\mu = 8.8, \frac{\sigma^2}{n} = \frac{2^2}{25})$

(a) $P(\overline{X} < 10) = 0.9987$.

(b) $P(7.5 \leq \overline{X} \leq 9) = P(Z \leq 9) - P(Z \leq 7.5)$

$= 0.6915 - 0.0006 = 0.6909$.

(c) $P(\overline{X} > 9.5) = 0.0401$

Si se muestrea una población que tiene una distribución de probabilidad desconocida la distribución de muestreo de la media muestral seguirá siendo aproximadamente normal, con media μ y varianza σ^2/n, si el tamaño de la muestra, n, es grande. Este es uno de los teoremas más útiles en estadística, se le conoce como **Teorema Central del Límite**, y se enuncia de la siguiente manera:

Teorema Central del Límite:
Si $X_1, X_2, \ldots, X_n$ es una muestra aleatoria de tamaño n de una v.a. X, con media μ y varianza finita σ^2, y si $\overline{X}$ es la media muestral , entonces la forma límite de la distribución de

$$Z = \frac{\overline{X} - \mu}{\sqrt{\sigma^2 / n}}$$

cuando $n \to \infty$, es la distribución normal con media 0 y varianza 1 ($Z \sim N(0,1)$).

La esencia de Teorema Central del Límite recae en el hecho de que para n grande, la distribución de $(\overline{X} - \mu) / \sqrt{\sigma^2/n}$ es, en forma aproximada, normal con media 0 y desvío estándar 1, sin importar cual sea el modelo de probabilidad a partir del cual se obtuvo la muestra. Si el modelo de probabilidad de la v.a. X, es semejante a una distribución normal (esto es, simétrico y existe una concentración relativamente alta alrededor del eje de simetría), la aproximación normal será buena aún para pequeñas muestras. Por otro lado, si el modelo de la v.a X se parece muy poco a una distribución normal (por ejemplo, existe una alta asimetría) la aproximación normal sólo será adecuada para valores relativamente grandes de n. En general, para $n > 30$.

Ejemplo:
Un contratista piensa comprar una gran cantidad de lámparas de alta intensidad a cierto fabricante. Éste asegura al contratista que la duración promedio de la lámparas es de 1000 hs. con un desvío estándar igual a 80hs. El contratista decide comprar las lámparas sólo si una muestra aleatoria de 64 de éstas da como resultado una vida útil promedio de por lo menos 1010 hs. ¿Cuál es la probabilidad de que el contratista adquiera las lámparas?

Sea la v.a. X = "vida útil de una lámpara", con media $\mu = 1000$ hs. y varianza $\sigma^2 = 80^2$ hs^2, y sea $\overline{X}$ = "vida útil promedio de una lámpara entre las 64 lámparas seleccionadas al azar". Por el Teorema Central del Límite $Z = \frac{\overline{X} - 1000}{80/\sqrt{64}} \sim N(0,1)$.

La probabilidad de que el contratista adquiera las lámparas equivale a calcular:

$$P(\overline{X} \geq 1010) = 0.1587.$$

Varianza Muestral

Si $X_1, X_2, ..., X_n$ representa una muestra aleatoria de tamaño *n* de una v.a. X, se define **varianza muestral** y se denota: S^2, al estadístico:

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

Al ser una v.a., es natural preguntarse, ¿cuál es su distribución?
Si $X \sim N(\mu, \sigma^2)$ y $X_1, X_2, ..., X_n$ representa una muestra aleatoria de tamaño *n* de una v.a. X, luego

$$\frac{(n-1)S^2}{\sigma^2} = \chi_\nu^2$$

es una v.a. Chicuadrado con $\nu = n - 1$ grados de libertad.

Distribuciones Continuas utilizadas en Inferencia Estadística

Distribución Chi- cuadrado

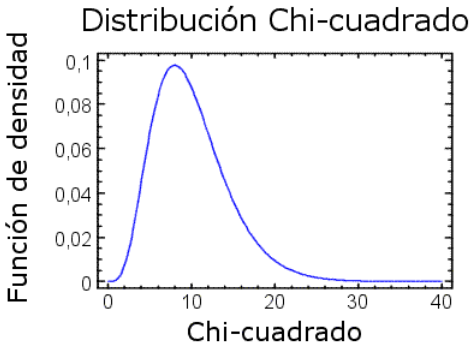
Es una de las distribuciones de muestreo de mayor utilidad.

Definición:

La v.a. X se dice que tiene una **distribución chi-cuadrado** de parámetro ν, y se denota $X \sim \chi^2_\nu$, si su función de densidad de probabilidad está dada por:

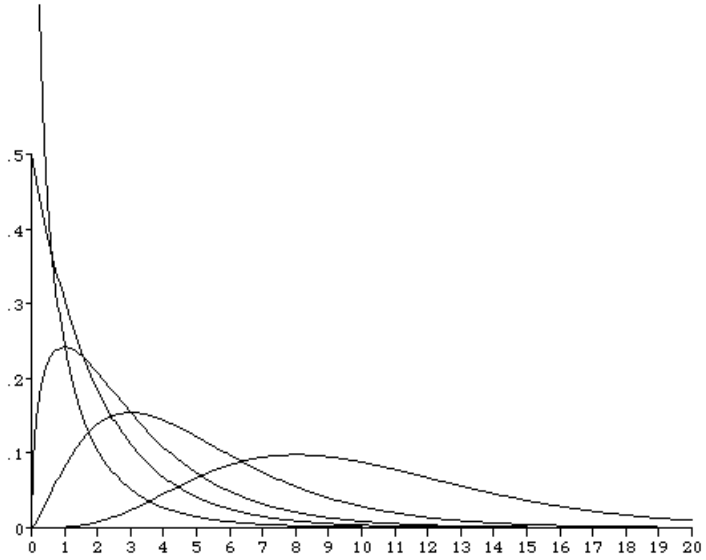
$$f(x) = \begin{cases} \dfrac{1}{2^{\nu/2}\,\Gamma(\dfrac{\nu}{2})} x^{\nu/2 - 1} e^{-x/2} & \text{si } x > 0 \\ 0 & \text{en el resto} \end{cases}$$

Obsérvese que esta distribución es un caso particular de la distribución Gamma, donde $\beta = 2$ y $\alpha = \nu/2$.



Características de la función de densidad:

- Esta caracterizada por un sólo parámetro ν, que recibe el nombre de *grados de libertad*.
- Es una curva asimétrica, con sesgo positivo (hacia la derecha).
- Posee un pico mayor que el de la distribución normal.
- Conforme $\nu \to \infty$ la distribución Chi-cuadrado se aproxima a la distribución normal.



Variables chi-cuadrado con valores de ν progresivamente crecientes.

La función de distribución acumulada, $F_{\chi^2\nu}(x) = P(\chi^2_\nu \leq x)$, se encuentra tabulada. La tabla proporciona probabilidades de la forma $P(\chi^2_\nu \leq \chi^2_{1-\alpha,\nu}) = 1-\alpha$, $0 \leq \alpha \leq 1$. Las probabilidades que no tienen la forma $P(\chi^2_\nu \leq \chi^2_{1-\alpha,\nu})$ se obtienen con el empleo de las reglas básicas de probabilidad.

La **esperanza y varianza** de una v.a. Chi-cuadrado con parámetro ν, son respectivamente:

$$E(X) = \nu \quad y \quad V(X) = 2\nu.$$

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño *n* de la v.a. $X \sim N(\mu,\sigma^2)$, entonces la v.a.

$$\frac{(n-1)S^2}{\sigma^2}$$

es una variable χ^2 con *n* - 1 grados de libertad.
Esta distribución interviene de manera especial al hacer inferencias respecto a la varianza de una distribución.

Ejemplo:
La autoridad sanitaria de un país decide llevar a cabo una investigación sobre los residuos que producen las empresas de un determinado sector. Seleccionada una muestra aleatoria simple de 9 empresas y suponiendo que los residuos se distribuyen normalmente con media 23 Tn y desviación estándar de 6 Tn., calcular la probabilidad de que la varianza de la cantidad de residuos que producen las empresas muestreadas sea superior a 60,12Tn^2.

X = "cantidad de residuos que produce una empresa"(Tn), **$X \sim N(\mu = 23, \sigma^2 = 6^2)$,** se tomó una muestra aleatoria de 9 empresas,

$P(S^2 > 60{,}12) = ???$

$$P(S^2 > 60.12) = P(\frac{\chi_8^2 6^2}{9-1} > 60.12) = P(\chi_8^2 > \frac{60{,}12 \quad 8}{36}) = P(\chi_8^2 > 13.36) =$$
$$= 1 - P(\chi_8^2 \leq 13.36) = 1 - 0.90 = 0.1$$

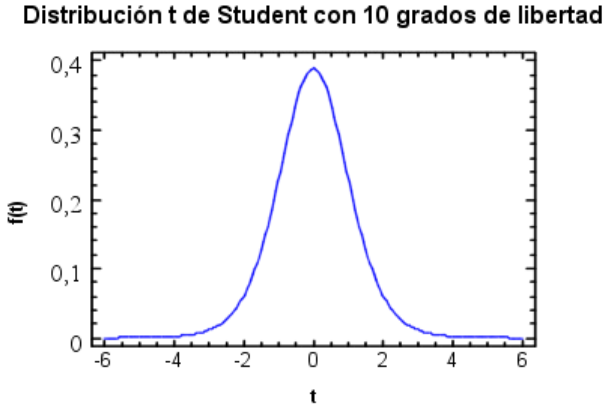
Distribución t de Student

Definición:
Sea Z una variable aleatoria normal estándar, Z ~ N(0, 1), y H una variable aleatoria chi-cuadrado de parámetro ν, H ~ χ^2_ν. Si Z y H son variables aleatorias independientes entonces la variable aleatoria

$$T = \frac{Z}{\sqrt{\frac{H}{\nu}}}$$

tiene una **distribución t de Student** de parámetro ν, se denota T ~ t_ν y una función de densidad está dada por:

$$f(t) = \frac{\Gamma(\frac{(\nu+1)}{2})}{(\pi\,\nu)^{1/2}\ \Gamma(\frac{\nu}{2})}\ \frac{1}{\left[\frac{t^2}{\nu}+1\right]^{\frac{\nu+1}{2}}} \qquad \text{si } -\infty < t < +\infty .$$



Características de la función de densidad:

- Esta caracterizada por un sólo parámetro ν, que recibe el nombre de *grados de libertad*.
- Es simétrica respecto del eje de ordenadas (recta T = 0).
- El valor máximo de la función se alcanza en μ = 0.
- Está definida para toda la recta real.
- Conforme aumenta el valor de ν la distribución se aproxima a la distribución normal estándar.

La $P(t_\nu \geq t)$, se encuentra tabulada. La tabla proporciona probabilidades de la forma $P(t_\nu \geq t_{1\alpha,\nu}) = \alpha$, $0 \leq \alpha \leq 1$. Las probabilidades que no tienen la forma $P(t_\nu \geq t_{\alpha,\nu})$ se obtienen con el empleo de las reglas básicas de probabilidad y de la simetría de la distribución t de Student.

La **esperanza y varianza** de una v.a. t de Student con parámetro ν, son respectivamente:

$$E(T) = 0 \quad y \quad V(T) = \nu / (\nu-2).$$

Sea $X_1, \ldots, X_n$ una muestra aleatoria de tamaño *n* de una v.a. $X \sim N(\mu, \sigma^2)$ y sean las variables aleatorias

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad y \qquad H = \frac{(n-1)\, S^2}{\sigma^2}$$

entonces, la variable

$$T = \frac{Z}{\sqrt{\frac{H}{n-1}}} = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$

tiene una distribución t de student con *n*-1 grados de libertad, t_{n-1}, puesto que $Z \sim N(0,1)$ y $H \sim \chi^2_{n-1}$.

Fin Capítulo V