UNIDAD DIDÁCTICA 5-1

DISTRIBUCIONES DE EN EL MUESTREO

Como se expuso en unidades anteriores, el objeto de la Inferencia Estadística es el de deducir conclusiones válidas (con un margen de error reducido y conocido) respecto a una población, a partir del análisis de una muestra obtenida al azar de dicha población.

La base para poder realizar dicha inferencia es el conocimiento de las relaciones existentes entre los parámetros de una población y las pautas de variabilidad previsibles en muestras obtenidas de la misma.

En el presente tema se introducen, en primer lugar, las ideas básicas relativas a las distribuciones de parámetros muestrales. Seguidamente se exponen unos resultados respecto a las distribuciones de $\overline{X}$ y de S^2, completamente generales, es decir que son válidos sea cual sea la distribución de la población muestreada. La tercera parte se dedica al estudio de ciertas distribuciones especiales que aparecen en el caso, particular pero extremadamente importante, de que las poblaciones muestreadas sigan una distribución normal. En este contexto se presentan de forma sucinta las distribuciones χ^2, t de Student y F de Snedecor, todas ellas ampliamente utilizadas en la Inferencia Estadística..

1. Conceptos generales

Sea una población a cuyos individuos va asociada una variable aleatoria **X**. Para obtener conclusiones sobre la distribución de esta variable se obtiene una **muestra aleatoria** constituida por N individuos de la población.

NOTA. Una muestra se dice que es aleatoria[1] si puede considerarse que todos los individuos de la población han tenido la misma probabilidad de estar incluidos en la muestra y si, además, dichos individuos han sido seleccionados independientemente unos de otros.

De una misma población es posible "a priori" obtener una gran cantidad de muestras diferentes de tamaño N (de hecho la cantidad es infinita si lo es la población original). Existe por tanto una **población de posibles muestras**, o sea una nueva población cuyos individuos son dichas muestras. A cada individuo de esta nueva población se le pueden hacer corresponder diferentes características numéricas, por ejemplo la media muestral $\overline{X}$ o la desviación típica muestral S de la muestra considerada. Estas características muestrales (X y S) serán, por tanto, nuevas variables aleatorias.

[1] En la terminología estadística se utiliza la expresión muestra aleatoria simple.

Es muy importante comprender bien la idea expresada en el párrafo anterior. Esta idea implica que cualquier característica muestral (como $\overline{X}$ o s) es una nueva variable aleatoria, variable cuya distribución dependerá de la existente en la población muestreada y del tamaño de la muestra.

Así, por ejemplo, la media muestral $\overline{X}$ es una variable aleatoria y, como tal, tiene una media y una desviación típica que dependerán, en general, de la distribución existente en la población de la que la muestra ha sido extraída.

Cualquier función de los valores muestrales, como lo son por ejemplo la media muestral $\overline{X}$ o la desviación típica muestral S, se denomina un **estadístico**. De acuerdo con lo que acaba de exponerse **todo estadístico es una variable aleatoria, cuya distribución dependerá en general de la distribución de la población y del tamaño de la muestra**.

La base de la Inferencia Estadística es, precisamente, el conocimiento de las relaciones que ligan la distribución de diferentes estadísticos muestrales (como $\overline{X}$ o s) con la distribución de la población y, en particular, con las características de dicha distribución (como la media poblacional m, o la desviación típica poblacional σ)

2. Distribución de las características muestrales

2.1. Distribución de la media muestral

Como se ha mencionado en el apartado anterior, **la media muestral $\overline{X}$ es una variable aleatoria** y **que**, por tanto, **tendrá una media y una desviación típica** que dependerá, en general, de la distribución existente en la población de la que la muestra ha sido extraída.

La media muestral se define, como sabemos, por la expresión:

$$\bar{x} = \frac{X_1 + \ldots + X_N}{N}$$

Si la población muestreada tiene de media m y de varianza σ^2, cada una de las X_i que constituye la muestra será el valor observado de una variable aleatoria con dichas media y varianza.

Como la media de una suma de variables es la suma de las medias, y la constante puede salir fuera al hallar esperanza matemática, se obtiene inmediatamente el siguiente importante resultado:

$$E\left(\overline{X}\right) = E\left(\frac{X_1 + \ldots + X_N}{N}\right) = \frac{E(X_1) + \ldots + E(X_N)}{N} = \frac{m + \ldots + m}{N} = m$$

Por tanto, **la media de la media muestral es la media poblacional**.

$$\boxed{E\left(\overline{X}\right) = m}$$

Por otra parte, aplicando las dos propiedades de la varianza vistas en la unidad didáctica anterior (UD4), se obtiene:

$$\mathrm{Var}\left(\overline{X}\right) = \mathrm{Var}\left(\frac{X_1 + \ldots + X_N}{N}\right) = \frac{\mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_N)}{N^2} = \frac{\sigma^2 + \ldots + \sigma^2}{N^2} = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$$

Por tanto, **la varianza de la media muestral es la varianza de la población dividida por el tamaño N de la muestra**.

$$\boxed{\sigma^2\left(\overline{X}\right) = \frac{\sigma^2}{N}}$$

En consecuencia a medida que aumenta N disminuirá $\sigma^2(\overline{X})$ y resultará más improbable que $\overline{X}$ tome valores que difieran mucho de la media poblacional m. Este resultado justifica la idea intuitiva habitual de considerar más "fiables" las conclusiones obtenidas a partir de una muestra grande que las derivadas de una muestra pequeña.

Digamos, por último, que como una consecuencia inmediata del Teorema Central del Límite visto también en la unidad didáctica anterior, la media muestral $\overline{X}$, al ser el resultado de sumar una serie de variables independientes, tiende a distribuirse normalmente, aunque la población de la que se haya extraído la muestra no siga una distribución normal.

2.2. Distribución de la varianza muestral

Análogamente a lo que se ha dicho de la media muestral, **la varianza muestral S^2 es una variable aleatoria** y **tendrá** también **una media y una desviación típica** que dependerá de la distribución existente en la población de la que la muestra ha sido extraída.

Como sabemos la expresión utilizada para calcular la varianza muestral es:

$$s^2 = \frac{\sum_{i=1}^{N}\left(X_i - \overline{x}\right)^2}{N-1}$$

Se demuestra con relativa facilidad que **la media de la varianza muestral es la varianza de la población**:

$$\boxed{E\left(s^2\right) = \sigma^2}$$

NOTA: para que se verifique el resultado anterior es necesario que, tal como se ha hecho a lo largo de este texto, s^2 se calcule dividiendo por N-1, en vez de por N como a primera vista parecería más lógico. Precisamente este hecho es el que justifica la expresión utilizada para calcular la varianza muestral

La expresión general para la varianza de s^2 es bastante complicada y no la daremos en este texto. El resultado más interesante al respecto es que, igual que sucedía con la varianza de $\overline{X}$, dicha varianza tiende a cero cuando N tiende a infinito.

$$\lim_{N\to\infty} \sigma^2(s^2) = 0$$

3. Muestreo de poblaciones normales

Los resultados expuestos anteriormente son completamente generales, en el sentido de que son válidos sea cual sea la distribución de la población muestreada.

Cuando dicha población es Normal, es posible establecer ciertos resultados adicionales de gran importancia dentro de la metodología de la Inferencia Estadística.

Así, en primer lugar, **la media muestral $\overline{X}$ se distribuirá normalmente, sea cual sea el tamaño N de la muestra**, por ser una transformada lineal de un conjunto $(X_1,...,X_N)$ de variables normales independientes. Teniendo en cuenta las expresiones generales obtenidas para la media y la varianza de $\overline{X}$, se verificará por tanto que:

$$\frac{\overline{x} - m}{\sigma/\sqrt{N}} \sim N(0,1) \qquad (1)$$

Ejercicio 1: El tiempo de transferencia de paquetes (ms) de un determinado tamaño a través de la red sigue una distribución normal con desviación típica σ=3. Si se extrae una muestra aleatoria de tamaño 16, calcular la probabilidad de que la diferencia entre la media poblacional (m) y la media muestral ($\overline{X}$) sea mayor que 2 en valor absoluto.

Otro resultado muy importante, cuya demostración desborda los límites de este texto, es que **en el muestreo de poblaciones normales $\overline{X}$ y S^2 son independientes**.

En el estudio de las pautas de variabilidad de estadísticos que aparecen en el muestreo de poblaciones normales, aparecen tres nuevas distribuciones de probabilidad: la χ^2, la **t de Student** y la **F de Snedecor**, de enorme importancia en Inferencia Estadística.

En los siguientes apartados presentamos sucintamente dichas distribuciones, limitándonos a dar su definición, propiedades más importantes, manejo de sus tablas, y su principal aplicación en contextos inferenciales.

3.1. Distribución χ^2 de Pearson[2]

La distribución χ^2 aparece en el estudio de la distribución de la varianza S^2 de una muestra de una población normal. La denominación de "Pearson" se debe al científico y matemático británico Karl Pearson que la estudió a finales del siglo XIX.

Por definición una variable aleatoria **Y** sigue una distribución χ^2 con ν grados de libertad si es la suma de ν variables N(0,1) independientes.

Así, si $X_1,...,X_\nu$ son variables N(0,1) independientes, la nueva variable $\mathbf{Y} = X_1^2 + ... + X_\nu^2$ sigue una distribución χ^2 con ν grados de libertad.

Si $X_1,...,X_\nu$ son variables N(0,1) independientes:

$$Y = \sum_{i=1}^{\nu} X_i^2 \sim \chi_\nu^2$$

La siguiente figura refleja la forma de la función de densidad de una variable χ^2 con 10 grados de libertad.

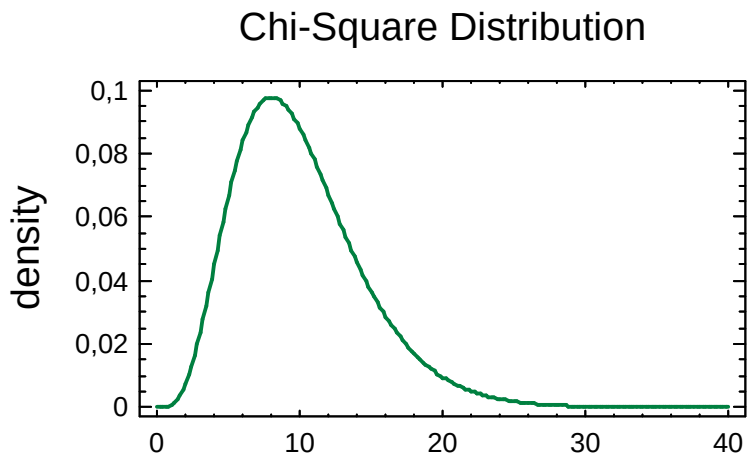

Figura 1. Función de densidad de una variable χ^2_{10}

Como puede apreciarse la distribución sólo toma valores positivos (lo que era obvio por tratarse de una suma de cuadrados) y es asimétrica positiva. Dicha asimetría, sin embargo, decrece a medida que aumentan los grados de libertad de la variable.

Se demuestra que la media de una variable χ^2 con ν grados de libertad es precisamente igual a ν, y la varianza es igual a 2ν.

La tabla de la χ^2 del Anejo disponible en PoliformaT da, para diferentes valores de α y de ν, el valor x_α tal que la probabilidad de que una χ^2 con ν grados de libertad sea mayor que x_α es igual a α.

[2] La distribución χ^2 también puede verse escrita como Gi-dos o Ji-dos.

La gran importancia de la distribución χ^2 en Inferencia Estadística, deriva de la siguiente propiedad fundamental, cuya demostración excede de los límites de este texto: si s^2 es la varianza en una muestra de tamaño N extraída de una población normal de varianza σ^2, se demuestra que $(N-1)\frac{S^2}{\sigma^2}$ sigue un distribución χ^2 con (N-1) grados de libertad.

Si X~N(m, σ^2) y S^2 es la varianza en una muestra de tamaño N

$$(N-1)\frac{s^2}{\sigma^2} \sim \chi^2_{N-1} \qquad (2)$$

Ejercicio 2: Obtener la probabilidad de obtener una varianza muestral superior a 10 al sacar una muestra de tamaño 20 de una población normal de varianza igual a 5

3.2. Distribución t de Student

En la inferencia respecto a medias en poblaciones normales desempeña un papel fundamental la distribución t de Student. El nombre procede del seudónimo con el que firmaba sus trabajos el matemático inglés W. S. Gosset, que la desarrolló en el curso de sus trabajos centrados en la industria cervecera).

Siendo **X** una variable N(0,1) e **Y** una variable χ^2 con ν grados de libertad independiente de la anterior, se dice que **t** sigue una distribución t de Student con ν grados de libertad si $t = \frac{X}{\sqrt{Y/\nu}}$

Si X ~ N(0,1), Y ~ χ^2_ν y X,Y independientes:

$$\frac{X}{\sqrt{Y/\nu}} \sim t_\nu$$

La **Figura 2** muestra la forma de la función de densidad de una variable t de Student con 5 grados de libertad.

Como puede apreciarse la distribución es simétrica respecto al origen, que es la media de la variable, recordando mucho a una N(0,1). De hecho cuando ν tiende a ∞ la variable t de Student tiende a distribuirse como una N(0,1).

La tabla de la t del Anejo disponible en PoliformaT hoja refleja, para diferentes valores de α y ν, los valores t_α tales que la probabilidad de que una variable t de Student con ν grados de libertad sea superior a t_α es igual a α.

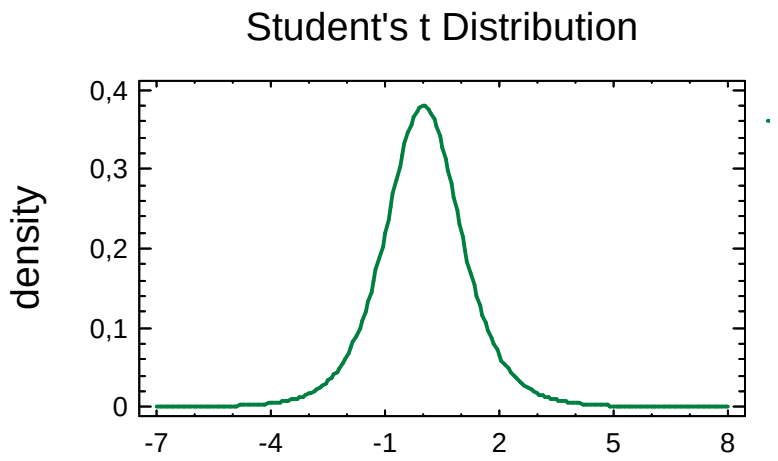

Figura 2. Función de densidad de una variable t_5

Ejercicio 3: Obtener un valor x tal que la probabilidad de que una t de Student con 10 grados de libertad sea en valor absoluto mayor que x, sea igual al 5%.

La importancia de la distribución t de Student en la teoría del muestreo radica en el siguiente resultado: siendo $\bar{X}$ y S la media y la desviación típica de una muestra de tamaño N extraída de una población normal de media m, el estadístico:

Si X~N(m, σ^2) y $\bar{X}$ y S^2 son la media y la varianza en una muestra de tamaño N

$$t = \frac{\bar{x} - m}{s/\sqrt{N}} \sim t_{N-1} \quad (3)$$

NOTA: Apréciese la analogía entre la expresión (3) y la (1) vista en el apartado 3. El hecho de sustituir en (1) una desviación típica poblacional σ por su estimación muestral S, se traduce en que el estadístico pasa de tener una distribución N(0,1) a una t de Student con N-1 grados de libertad (que son los grados de libertad con que está estimada σ).

3.3. Distribución F de Fisher (o de Snedecor)

En el estudio de los modelos de Regresión Lineal y de Análisis de la Varianza (que se desarrollan en posteriores partes de esta unidad didáctica) desempeña un papel fundamental la distribución F, denominada así por el estadístico estadounidense G. W. Snedecor que la propuso en honor del famoso estadístico inglés R.A. Fisher creador de los modelos anteriormente mencionados.

Siendo $\mathbf{X}_1$ y $\mathbf{X}_2$ dos variables χ^2 independientes con ν_1 y ν_2 grados de libertad, se dice que una variable **F** sigue una distribución **F** con ν_1 y ν_2 grados de libertad si $F = \frac{X_1/\nu_1}{X_2/\nu_2}$

Si $X_1 \sim \chi^2_{\nu_1}$, $X_2 \sim \chi^2_{\nu_2}$ y X_1, X_2 independientes:

$$F = \frac{X_1/\nu_1}{X_2/\nu_2} \sim F_{\nu_1,\nu_2}$$

La **Figura 3** refleja la forma de la función de densidad de una F con 5 grados de libertad en el numerador y 10 grados de libertad en el denominador.

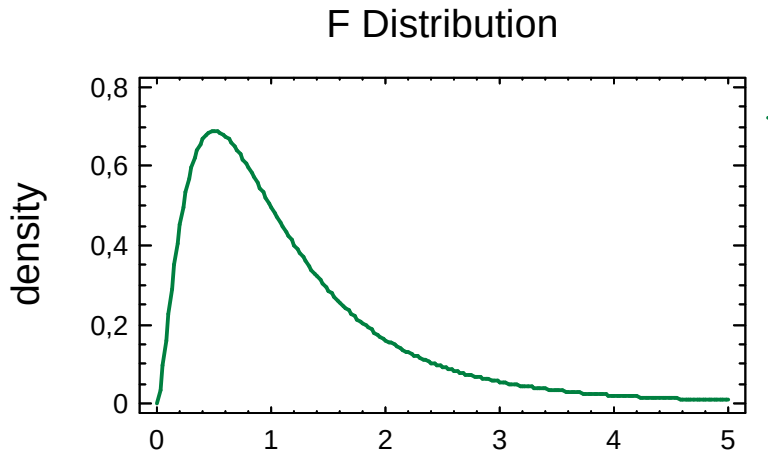

Figura 3. Función de densidad de una variable $F_{5,10}$

Como puede apreciarse la distribución es marcadamente asimétrica, si bien el coeficiente de asimetría decrece ligeramente a medida que aumentan los grados de libertad del numerador y del denominador.

La media de una distribución F es cercana a 1.

En general las tablas de la distribución F son muy prolijas, por exigir tres entradas: una para v_1, otra para v_2 y otra para los niveles α de probabilidad. En este texto nos hemos limitado a la sucinta tabla del Anejo disponible en PoliformaT, que refleja para un conjunto de valores seleccionados de v_1 y v_2, los valores críticos correspondientes a unas probabilidades de ser superados del 5% y del 1%.

La distribución F se utiliza en la práctica con el objetivo de comparar la variabilidad debida a diferentes fuentes. En concreto si S_1^2 es la varianza en una muestra de tamaño N_1 extraída de una población normal de varianza σ_1^2, y S_2^2 es la varianza de una muestra de tamaño N_2 extraída de una población normal de varianza σ_2^2, y ambas muestras son independientes, el cociente $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ se distribuye como una variable F de Snedecor con (N_1-1) y (N_2-1) grados de libertad.

$$\text{Si } X_1 \sim N(m_1, \sigma_1^2),\ X_2 \sim N(m_2, \sigma_2^2) \text{ son independientes y}$$
$$S_1^2 \text{ y } S_1^2 \text{ son las varianzas muestrales de } X_1 \text{ y } X_2 \text{ (tamaños } N_1 \text{ y } N_2\text{)} \qquad (4)$$
$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{N_1-1,N2-1}$$

En particular si las dos varianzas poblacionales son iguales, el cociente entre las dos varianzas muestrales seguirá una distribución F.

$$\text{Si} \quad \sigma_1^2 = \sigma_2^2 \quad \Rightarrow \quad \frac{S_1^2}{S_2^2} \sim F_{N_1-1,N2-1}$$

Ejercicio 4: Calcular aproximadamente la probabilidad de que al extraer dos muestras de tamaño 25 de una misma población normal, la segunda varianza muestral resulta más del doble que la primera.

Para pensar

Cuestión 1: de una misma población se extraen muestras pequeñas (N=5) y muestras grandes (N=50). En general, ¿la varianza muestral será mayor en las muestras grandes o en las muestras pequeñas?

Las varianzas muestrales serán del mismo orden en muestras grandes que en muestras pequeñas, dado que en ambos casos el valor medio esperado de dichas varianzas muestrales es idéntico, al coincidir con la varianza poblacional.

Aunque la conclusión anterior es una consecuencia obvia de la propiedad $E(S^2)=\sigma^2$, es muy frecuente que uno piense que será mayor la varianza en muestras grandes que en muestras pequeñas, posiblemente por identificar inconscientemente la varianza con una medida de dispersión de los datos del tipo del recorrido.

Cuestión 2: intuitivamente, ¿qué quiere decir que en el muestreo de poblaciones normales $\bar{X}$ y S^2 son independientes? ¿Lo serían también en el muestreo de una variable Uniforme(0,1)?

La independencia de $\bar{X}$ y S^2, implica que el hecho de que en una muestra la media muestral $\bar{X}$ haya resultado, por ejemplo, muy grande, no indica nada acerca de si la varianza muestral habrá resultado en dicha muestra grande o pequeña.

Esto no sucedería al muestrear una variable U(0,1) pues, por ejemplo, si $\bar{X}$ resultara muy cercana a 1 ello sólo puede suceder si todos los valores son muy próximos a 1 (dado que no puede haber valores mayores que 1) lo que implicaría que la varianza muestral sería en este caso menor que la que cabría esperar si $\bar{X}$ hubiera resultado, por ejemplo, próxima a 0,5.

Cuestión 3: justificar la afirmación de que la asimetría de una χ^2 decrece a medida que aumentan los grados de libertad de la variable como una consecuencia del Teorema Central del Límite.

Como una χ^2 es una suma de μ variables independientes, su distribución tenderá a la normal (y por tanto a hacerse simétrica) al aumentar el número de sumandos (μ), o sea los grados de libertad de la χ^2

Cuestión 4: justificar intuitivamente el que la media de una distribución F sea cercana a 1.

Como $E(\chi^2_\nu)=\nu$, las dos expresiones en el numerador y en el denominador de la expresión que define la F tienen media igual a 1. Una F resulta, por tanto, igual al cociente de dos variables independientes, ambas con la misma media, por lo que es intuitivo que su valor medio será cercano a 1 (Nota: decimos sólo "cercano" porque la media de un cociente no coincide exactamente con el cociente de las medias)

Ejercicios resueltos

Anejos del Capítulo 8 del libro de R. Romero y L.R. Zúnica "Métodos Estadísticos en Ingeniería" SPUPV 637

Ver boletín correspondiente en PoliformaT (EST GII: Recursos / 04 | Ejercicios)

Para saber más

- Descartes. Instituto de Tecnologías Educativas (ITE) del Ministerio de Educación: http://recursostic.educacion.es/descartes/web/materiales_didacticos/inferencia_estadistica/distrib_muestrales.htm

- VESTAC. Internet Scout Project. Distribution of mean. http://lstat.kuleuven.be/java/index.htm

- The sample mean experiment. Virtual Laboratories. http://www.math.uah.edu/stat/applets/SampleMeanExperiment.xhtml

Fuentes

Métodos Estadísticos en Ingeniería (Romero Villafranca, Rafael)