Muestras aleatorias y distribuciones de muestreo.

Definición 1.

Si las variables aleatorias X_1, X_2, …, X_n tienen la misma función de probabilidad que la distribución de la población y su distribución de probabilidad conjunta es igual al producto de las marginales, entonces X_1, X_2, …, X_n forma un conjunto de n variables aleatorias independientes e idénticamente distribuidas (IID) que constituye una muestra aleatoria de la población.

La función (densidad) conjunta de probabilidad de X_1, X_2, …, X_n es la función de verosimilitud de la muestra dada por:

$$L(\underline{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

en donde

$$\underline{x} = \left(x_1, x_2, \ldots x_n\right)$$

denota los datos muestreados.

Cuando las realizaciones de

$$\underline{x} = \left(x_1, x_2, \ldots x_n\right)$$

se conocen, la función

$$L(\underline{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

depende sólo del parámetro desconocido θ.

Definición 2

Un parámetro es una característica numérica de la distribución de la población de manera que cuando se conoce esta, la distribución, queda descrita, sino total, al menos parcialmente.

Los parámetros o funciones de los parámetros se estiman a partir de la información contenida en una muestra.

Definición 3.

Un estadístico es cualquier función de las variables aleatorias que se observaron en la muestra de manera que esta función no contiene cantidades desconocidas.

Si se utiliza un estadístico T para estimar un parámetro desconocido θ, entonces T recibe el nombre de estimador de θ y el valor específico que tome T, por ejemplo t, se denomina estimación de θ.

Definición 4

La distribución de muestreo de un estadístico T es la distribución de probabilidad de T que puede obtenerse como resultado de un número infinito de muestras aleatorias independientes, cada una de tamaño n, provenientes de la población de interés.

Teorema 1

Sea $X_1, X_2, \ldots, X_n$ un conjunto de n variables aleatorias independientes cada una con función generadora de momentos

$$m_{x_1}(t), m_{x_2}(t), \ldots, m_{x_n}(t)$$

Si $Y=a_1X_1 + a_2X_2+ \ldots + a_nX_n$ en donde $a_1, a_2, \ldots, a_n$ son constantes, entonces:

$$m_y(t) = m_{x_1}(a_1 t).m_{x_2}(a_2 t)\ldots m_{x_n}(a_n t)$$

Teorema 2

Sea $X_1, X_2, \ldots, X_n$ un conjunto de variables aleatorias independientes y normalmente distribuidas con medias $E(X_i)$ y varianzas $Var(Xi)=\sigma_i^2$ para i = 1, 2 , …, n. Si $Y=a_1X_1 + a_2X_2+ \ldots + a_nX_n$ en donde $a_1, a_2, \ldots, a_n$ son constantes, entonces Y es una variable aleatoria normal con media

$$E(Y) =a_1\mu_1+ a_2\mu_2+ \ldots + a_n\mu_n$$

y con varianza

$$var(Y)= a_1^2\sigma^2_1+ a_2^2\sigma^2_2+ \ldots + a_n^2\sigma^2_n$$

(La hipótesis de normalidad puede quitarse)

La distribución de muestreo de la media muestral $\overline{x}$.

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria que consiste en un conjunto de variables aleatorias independientes e idénticamente distribuidas (v.a IID) tales que $E(X_i)=\mu$ y $Var(X_i)=\sigma^2$ para todo i.

Entonces el estadístico

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

se define como la media de las n v.a IID.

Aplicando el teorema 2 se tiene :

$$(a_i = \frac{1}{n}, \ \forall i)$$

$$E(\overline{X}) = \frac{1}{n}\sum_i E(X_i) = \frac{1}{n}\sum_i \mu = \frac{1}{n}(n\mu) = \mu$$

$$Var(\overline{X}) = \sum_i \frac{1}{n^2}\sigma^2 = \frac{\sigma^2}{n}$$

de donde

$$d.e.(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

Teorema 3

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria que consiste en un conjunto de variables aleatorias independientes y normalmente distribuidas tales que $E(X_i)=\mu$ y $Var(X_i)=\sigma^2$ para i = 1, 2, …, n.
Entonces la distribución de la media muestral $\overline{X}$ es normal con media μ y varianza σ^2/n.

$$\overline{X} \ \rightarrow \ N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Un problema de máximo interés consiste en saber lo que ocurre si no se especifica la distribución de probabilidad de la población a partir de la cual se extrae la muestra.

Teorema central del límite.

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria que consiste en un conjunto de variables aleatorias independientes e idénticamente distribuidas (v.a IID) en una distribución de probabilidad no especificada y tales que $E(X_i)=\mu$ y $Var(X_i)=\sigma^2$ para todo i. Entonces el promedio muestral

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

tiene una distribución de media μ y varianza σ^2/n que tiende a una normal conforme n tiende a ∞

Debe de notarse el hecho de que si el modelo de probabilidad de la población es semejante a una distribución normal, la aproximación normal será buena aun para muestras pequeñas.

En general, para n>30, la aproximación normal será relativamente buena y puede emplearse

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

para hacer inferencias sobre μ cuando se conoce el valor de la varianza poblacional σ^2.

Distribución en el muestreo de S^2

Supongamos que la población se encuentra normalmente distribuida con μ conocida y σ^2 desconocida. Se define S^2 como

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

donde para cada i

$$X_i \rightarrow N(\mu, \sigma)$$

Teorema 5

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución normal de media μ y varianza σ^2. La distribución de la variable aleatoria

$$Y = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma^2}$$

es del tipo de la Chi-cuadrado con n grados de libertad.

Desde un punto de vista práctico, la varianza muestral tal y como se encuentra definida anteriormente tiene poco uso, pues rara vez se conoce la media poblacional μ. En su lugar se emplea la varianza muestral, definida por:

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

Más adelante se verá porqué se emplea el divisor n-1 en lugar de n.

El reemplazo de la media desconocida m por la media muestral $\bar{x}$ da origen a la presencia de otro estadístico en la definición de S^2. Como consecuencia se tiene que la distribución de muestreo

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

Teorema 6

Sea X_1 y X_2 son variables aleatorias independientes y cada una tiene una distribución Chi-cuadrado con ν_1 y ν_2 grados de libertad, entonces $Y = X_1 + X_2$ tiene también una

Chi-cuadrado con $\nu_1 + \nu_2$ grados de libertad.

Teorema 7

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

donde

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

Demostración

$$(n-1)S^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2 = \sum_{i=1}^{n} \left[(X_i - \mu) + (\mu - \overline{X})\right]^2 = \sum_{i=1}^{n} \left[(X_i - \mu) - (\overline{X} - \mu)\right]^2 = \ldots =$$

$$\sum_{i=1}^{n} \left[(X_i - \mu)\right]^2 - n(\overline{X} - \mu)^2$$

de donde

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

Dividiendo los dos miembros por σ^2

$$\frac{(n-1)S^2}{\sigma^2} + \left(\frac{(\overline{X} - \mu)}{\sigma/\sqrt{n}}\right)^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma^2}$$

Por el teorema 5

$$\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma^2}$$

sigue una Chi-cuadrado con n grados de libertad y

$$\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)^2$$

de manera similar sigue una chi-cuadrado con 1 grado de libertad, dado que

$$\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right) \rightarrow N(0, 1)$$

En vitud del teorema 6 se sigue que

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

La distribución t de Student

Se sabe, cuando una muestra proviene de una distribución normal con desviación estándar conocida σ, que la distribución de

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

Generalmente el valor de σ no se conoce y lo que se hace es reemplazar σ por un estimador s, que es el valor de la desviación estándar muestral S.

Desafortunadamente

$$\left(\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \right)$$

ya no es N(0; 1) aún cuando la muestra provenga de una distribución normal.

Sin embargo, es posible determinar la distribución muestral exacta de

$$\left(\frac{\overline{X} - \mu}{S / \sqrt{n}} \right)$$

cuando la población es normal N(μ, σ) con μ y σ desconocidos.

Teorema 8

Sea Z normal N(0, 1) y X una Chi-cuadrado con ν grados de libertad. Si Z y X son independientes, entonces la variable aleatoria

$$T = \frac{Z}{\sqrt{X / \nu}}$$

sigue una t-Student con ν grados de libertad.

La similitud de la t-Student y la N(0,1) es alta para valores grandes de ν, sobre todo para ν ≥30.

Teorema 9

Cuando se muestrea una población normal N(μ,σ) , el estadístico

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

sigue una t-Student con n-1 grados de libertad.

Distribución de la diferencia de dos medias muestrales

Supongamos que X sigue una $N(\mu_X, \sigma)$ e Y una $N(\mu_Y, \sigma)$ donde X e Y son variables aleatorias independientes con **varianzas iguales conocidas**.

Sabemos que

$$\overline{X} \rightarrow N(\mu_x; \sigma/\sqrt{n_x}) \quad e \quad \overline{Y} \rightarrow N(\mu_y, \sigma/\sqrt{n_y})$$

Entonces

$$\overline{X} - \overline{Y} = \overline{X} + (-1)\overline{Y}$$

sigue una distribución normal de media μ_X-μ_Y y varianza $1^2 . \sigma^2/n_X + (-1)^2.\sigma^2/n_Y$

Por tanto si se conoce el valor de σ^2, el estadístico

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sigma\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

se distribuye según una normal de N(0; 1)

Se ha supuesto que σ es conocido. Sin embargo, es poco probable que esto suceda. Por tanto para el caso en el que el muestreo se lleve a cabo sobre dos poblaciones normales independientes con **varianzas iguales pero desconocidas**, para cada una de las muestras obtenidas pueden definirse las varianzas muestrales definidas S^2_X y S^2_Y y dado que

$$\frac{(n_x - 1)S_x^{\,2}}{\sigma^2} \rightarrow \chi^2_{n_X - 1}$$

$$\frac{(n_y - 1)S_y^{\,2}}{\sigma^2} \rightarrow \chi^2_{n_Y - 1}$$

y teniendo en cuente el teorema 6

$$W = \frac{(n_x - 1)S_x^{\,2}}{\sigma^2} + \frac{(n_y - 1)S_y^{\,2}}{\sigma^2}$$

sigue también una chi-cuadrado con n_X+n_Y-2 grados de libertad

Luego

$$T = \frac{Z}{\sqrt{\frac{W}{(n_x + n_y - 2)}}}$$

sigue una t-Student con $n_X + n_Y - 2$ grados de libertad.

Por tanto

$$T = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \Big/ \frac{1}{\sigma}\sqrt{\frac{(n_X - 1)S_X^{\,2} + (n_Y - 1)S_Y^{\,2}}{n_X + n_Y - 2}}$$

simplificando

$$T = \frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{(n_x - 1)S_x^{\,2} + (n_y - 1)S_y^{\,2}}{n_x + n_y - 2}\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} = \frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{S_P\sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$$

donde

$$S_P = \sqrt{\frac{(n_x - 1)S_x^{\,2} + (n_y - 1)S_y^{\,2}}{n_x + n_y - 2}}$$

es un estimador combinado de la varianza.

Si suponemos que las **varianzas poblacionales son distintas pero conocidas**, entonces se tiene:

$$\overline{X} \quad \rightarrow N(\mu_x; \frac{\sigma_x}{\sqrt{n_x}})$$

$$\overline{Y} \quad \rightarrow N(\mu_y; \frac{\sigma_y}{\sqrt{n_y}})$$

De donde

$$\overline{X} - \overline{Y} \quad \rightarrow N\left(\mu_x - \mu_y; \sqrt{\frac{\sigma_x^{\,2}}{n_x} + \frac{\sigma_y^{\,2}}{n_y}}\right)$$

y por tanto

$$Z = \frac{\overline{X} - \overline{Y} - \mu_x - \mu_y}{\sqrt{\frac{\sigma_x^{\,2}}{n_x} + \frac{\sigma_y^{\,2}}{n_y}}} \quad \rightarrow N(0,1)$$

En el **supuesto de que σ^2_X y σ^2_Y sean desconocidas** y haya que estimarlas a partir de S^2_X y S^2_Y el problema se complica, en tal caso se tiene:

$$t = \frac{(\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^{\ 2}}{n_x} + \frac{S_y^{\ 2}}{n_y}}}$$

sigue una distribución t-Student con f grados de libertad, en donde f es la aproximación de Welch:

$$f = \frac{\left(\frac{S_x^{\ 2}}{n_x} + \frac{S_y^{\ 2}}{n_y}\right)^2}{\frac{\left(S_x^{\ 2}/n_x\right)^2}{n_x + 1} + \frac{\left(S_y^{\ 2}/n_y\right)^2}{n_y + 1}}$$

expresado en número entero.

La distribución F

La inferencias con respecto a la varianza σ^2 cuando se muestrea una población normal se formula con base a

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

En esta sección se formularán inferencias con respecto a la varianza de dos distribuciones normales independientes con base en las muestras aleatorias de cada una.

Teorema 10

Sea X una variable aleatoria que se distribuye según una Chi-cuadrado con ν_1 grados de libertad e Y otra variable aleatoria independiente de X, que se distribuye según una Chi-cuadrado con ν_2 grados de libertad. Entonces la variable aleatoria

$$F = \frac{X/\nu_1}{Y/\nu_2}$$

tiene una función de densidad de probabilidad dada por

$$g(f;\nu_1,\nu_2) = \begin{cases} \frac{\Gamma[(\nu_1+\nu_2)/2]\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} f^{(\nu_1-2)/2}(\nu_2+\nu_1 f)^{-(\nu_1+\nu_2)/2} & Si \quad f > 0 \\ 0 & f \le 0 \end{cases}$$

Es fácil ver que si

$$F = \frac{X/\nu_1}{Y/\nu_2}$$

sigue una distribución F con ν_1 ,ν_2 grados de libertad, entonces F´=1/F sigue una F con ν_2 ,ν_1 grados de libertad

$$F' = \frac{1}{F} = \frac{Y/\nu_1}{X/\nu_2}$$

Es por esto que en las tablas sólo aparecen los valores cuantiles $f_{1-\alpha;v1;v2}$ para α<0.5. Si se desean los valores cuantiles para α>0.5

$$P(F \le f_{1-\alpha;\nu_1,\nu_2}) = P(\frac{1}{F} > \frac{1}{f_{1-\alpha;\nu_1,\nu_2}}) = 1-\alpha$$

o bien

$$P(F' \le f'_{\alpha;\nu_2,\nu_1}) = \alpha$$

siendo

$$f'_{\alpha;\nu_2,\nu_1} = \frac{1}{f_{1-\alpha;\nu_1,\nu_2}}$$

Volviendo al problema de desarrollar estadísticos para formular inferencias con respecto a las varianzas de dos distribuciones normales independientes.

Sean X_1, X_2, …., X_{nx} variables aleatorias independientes N(μ_X, σ_X)

Sean Y_1, Y_2, …, Y_{ny} variables aleatorias independientes N(μ_Y, σ_Y)

Si X e Y son independientes

$$\frac{(n_x-1)S_x^{\,2}}{\sigma_x^{\,2}} \to \chi^2_{n_x-1}$$

$$\frac{(n_y-1)S_y^{\,2}}{\sigma_y^{\,2}} \to \chi^2_{n_y-1}$$

entonces por el teorema 10

$$\frac{\dfrac{(n_x-1)S_x^{\,2}}{\sigma_x^{\,2}}\Big/(n_x-1)}{\dfrac{(n_y-1)S_y^{\,2}}{\sigma_y^{\,2}}\Big/(n_y-1)} = \frac{S_x^{\,2}/\sigma_x^{\,2}}{S_y^{\,2}/\sigma_y^{\,2}} \to F_{nx-1,ny-1}$$