ASIGNATURA FST41:Las distribuciones de $\overline{X}$ para algunas poblaciones no normales

Cirilo Alvarez Rojas

Universidad Nacional De Ingeniería
Facultad De Ingeniería Económica,Ingeniería Estadística
Y Ciencias Sociales
ESCUELA PROFESIONAL DE INGENIERÍA ESTADÍSTICA

Hasta ahora hemos considerado estadísticas que eran funciones de variables aleatorias con las mismas distribuciones normales. Ahora investigamos algunas otras estadísticas.

Sea $X_1, X_2, \ldots, X_m$ una muestra aleatoria de una población con distribución Binomial de parámetros n y θ.La función de probabilidad es

$$P(X_j = r) = \binom{n}{r} \theta^r (1-\theta)^{n-r}, \quad \begin{matrix} r = 0, 1, 2, \ldots, n \\ j = 1, 2, \ldots, m. \end{matrix}$$

donde $0 < \theta < 1$. Se desea determinar la distribución de la media muestral de estas variables aleatorias.

$$\overline{X} = \frac{1}{m} \sum_{j=1}^{m} X_j$$

Para ello utilizamos la función característica; se sabe que la función característica de la $\overline{X}$, de una muestra de tamaño n está dada por

$$\Psi_{\overline{X}}(t) = \left[\Psi_X\left(\frac{t}{n}\right)\right]^n$$

donde Ψ_X es la función característica de la población. En caso binomial, la función característica de esta población es

$$\Psi_X(t) = (1 - \theta + \theta e^{it})^{\mathrm{n}}$$

Luego, la función característica de $\overline{X}$ es

$$\Psi_{\overline{X}}(t) = (1 - \theta + \theta e^{i\frac{t}{\mathrm{m}}})^{\mathrm{nm}} \tag{1}$$

La expresión (1) es la función característica de una variable aleatoria con función de distribución binomial modificado: $\overline{X}$ puede tomar los valores

$$\overline{X} = 0, \frac{1}{\mathrm{m}}, \frac{2}{\mathrm{m}}, \ldots, \frac{\mathrm{nm}}{\mathrm{m}} = \mathrm{n},$$

y

$$P(\overline{X} = \frac{k}{m}) = \binom{nm}{k}\theta^k(1-\theta)^{nm-k}, k = 0, 1, 2, \ldots, nm$$

de donde obtiene que

$$E(\overline{X}) = E\left(\frac{k}{m}\right) = \frac{1}{m}E(k) = \frac{1}{m}nm\theta = n\theta$$

y

$$Var(\overline{X}) = Var\left(\frac{k}{m}\right) = \frac{1}{m^2}Var(k) = \frac{1}{m^2}nm\theta(1-\theta) = \frac{n\theta(1-\theta)}{m}$$

Ahora considere variables aleatorias independientes con la misma $X_j (j = 1, 2, \ldots, \mathrm{m})$ con la misma distribución de Poisson dada por

$$\mathrm{P}(X_j = r) = e^{-\lambda} \frac{\lambda^r}{r!}, \quad \begin{matrix} r = 0, 1, 2, \ldots, \\ j = 1, 2, \ldots, \mathrm{m}. \end{matrix}$$

donde $\lambda > 0$.

Para hallar la distribución del estadístico $\overline{X}$ definido en la relación (1), observe que la función característica de la población Poisson es,

$$\Psi_X(t) = e^{\lambda(e^{it} - 1)}.$$

Entonces la función característica de $\overline{X}$ es

$$\Psi_{\overline{X}}(t) = e^{\mathrm{m}\lambda(e^{i\frac{t}{m}} - 1)}$$

La expresión anterior, es la función característica de una variable aleatoria con una modificación de la distribución Poisson; $\overline{X}$ puede tomar los valores

$$\overline{X} = 0, \frac{1}{\mathrm{m}}, \frac{2}{\mathrm{m}}, \frac{3}{\mathrm{m}}, \ldots$$

y

$$\mathrm{P}\left(\overline{X} = \frac{k}{\mathrm{m}}\right) = e^{-\mathrm{m}\lambda} \frac{(\mathrm{m}\lambda)^k}{k!},\ k = 0, 1, 2, \ldots$$

La media y varianza se puede hallar utilizando directamente la función característica. Así

$$\Psi'_{\overline{X}}(t) = \lambda i e^{\mathrm{m}\lambda(e^{i\frac{t}{m}}-1)} \left[e^{i\frac{t}{\mathrm{m}}}\right]$$

y

$$\Psi'_{\overline{X}}(0) = \lambda i = i\mathrm{E}(\overline{X}) \Rightarrow \mathrm{E}(\overline{X}) = \lambda$$

Ejercicio 1.

Encuentre la distribución de la media muestral $\overline{X}$ de una muestra aleatoria simple de tamaño n extraída de una población en la que característica X tiene la distribución Gamma.

Ejercicio 2.

Sea $\overline{X}$ la media muestral de una muestra aleatoria simple de tamaño n extraída de una población en la cual la característica X tiene la distribución uniforme dada por

$$f(x) = \begin{cases} 1 & \text{si} \quad x \in [0,1] \\ 0 & \text{si} \quad x \notin [0,1] \end{cases}$$

Pruebe la función de densidad $g(x)$ de $\overline{X}$ es para $j = 0, 1, 2, \ldots, n-1$ es de la forma

$$g_{\overline{X}}(x) = \frac{n^n}{(n-1)!} \sum_{k=0}^{j} (-1)^k \binom{n}{k} \left(x - \frac{k}{n} \right)^{n-1}, \left(\frac{j}{n} \leq x \leq \frac{j+1}{n} \right)$$

Muestreo en Poblaciones Normales:

Distribución de la diferencia de medias muestrales con varianza conocidas

En muchas situaciones surge la necesidad de comparar las medias muestrales de dos poblaciones distintas. Por ejemplo supongamos que estamos interesados en comparar los tiempos medios de duración de dos tipos de tubos fluorescente. La fabricación de ambos tipos de tubos de fluorescentes se realiza por empresas distintas y con diferentes procesos de fabricación. Por lo tanto, los tubos producidos por cada empresa tendrán una distribución diferente, una de la otra, de los tiempos de duración de los tubos.

Teorema 1

Sean $X_1, X_2, \ldots, X_m$ y $Y_1, Y_2, \ldots, Y_n$ dos muestras aleatorias, cada una de distribuciones normales independientes con medias desconocidas μ_x y μ_y, pero varianzas conocidas σ_x^2 y σ_y^2, respectivamente. Entonces

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_x - \mu_y, \frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}\right)$$

Es decir, la diferencia de medias muestrales, $\overline{X} - \overline{Y}$, se distribuye en forma normal con $E(\overline{X} - \overline{Y}) = \mu_x - \mu_y$ y varianza.
$Var(\overline{X} - \overline{Y}) = \sigma^2_{\overline{x}-\overline{y}} = \frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}$.

Además,

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}} \sim \mathcal{N}(0, 1).$$

Prueba

Como las muestras son independientes, también lo son las medias muestrales $\overline{X}, \overline{Y}$. luego por la propiedad del función característica resulta

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

y como $\Psi_{\overline{X}}(t) = e^{it\mu_x + \frac{\sigma_x^2}{m}\frac{t^2}{2}}$ y $\overline{\Psi_{\overline{Y}}(t)} = e^{-it\mu_y + \frac{\sigma_y^2}{n}\frac{t^2}{2}}$. Entonces la función característica de la diferencia de medias muestrales resulta

$$\Psi_{\overline{X}-\overline{Y}}(t) = e^{it\mu_x + \frac{\sigma_x^2}{m}\frac{t^2}{2}} e^{-it\mu_y + \frac{\sigma_y^2}{n}\frac{t^2}{2}} = e^{it(\mu_x - \mu_y) + \left(\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}\right)\frac{t^2}{2}}.$$

Luego por el teorema de la unicidad de la función característica se concluye que

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_x - \mu_y, \frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}\right).$$

Ejemplo

Analizando los sueldos de los empleados de dos empresas se deduce que en la empresa A el salario medio mensual es de 2900 nuevos soles con una varianza de 250 (nuevos soles)^2, y en la empresa B el salario medio es de 2508 nuevos soles con una con una varianza de 300 (nuevos soles)^2. Si se toma una muestra aleatoria de 25 personas de la empresa A y de 36 personas de la empresa B. Determinar la probabilidad de que la muestra procedente de la empresa A tenga un salario medio que sea al menos 400 nuevos soles superior al salario medio de la empresa B.

Solución:

La información que se tiene es la siguiente:

Población A: $\mu_x = 2900 \quad \sigma_x^2 = 250 \quad m = 25$

Población B: $\mu_y = 2508 \quad \sigma_y^2 = 300 \quad m = 36$

continuación

POr Teorema 1, se sabe que

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(2900 - 2508, \frac{250}{25} + \frac{300}{36}\right)$$

$$\overline{X} - \overline{Y} \sim \mathcal{N}(392, 18{,}33)$$

Luego, tenemos

$$P(\overline{X} - \overline{Y} > 400) = P\left(\frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{\mathrm{m}} + \frac{\sigma_y^2}{\mathrm{n}}}} > \frac{400 - 392}{\sqrt{18{,}33}}\right)$$

$$= P\left(Z > \frac{8}{4{,}2814}\right)$$

$$= P(z > 1{,}8685) = 1 - P(z > 1{,}8685) = 1 - \Phi(1{,}8685)$$

$$= 0{,}03085$$

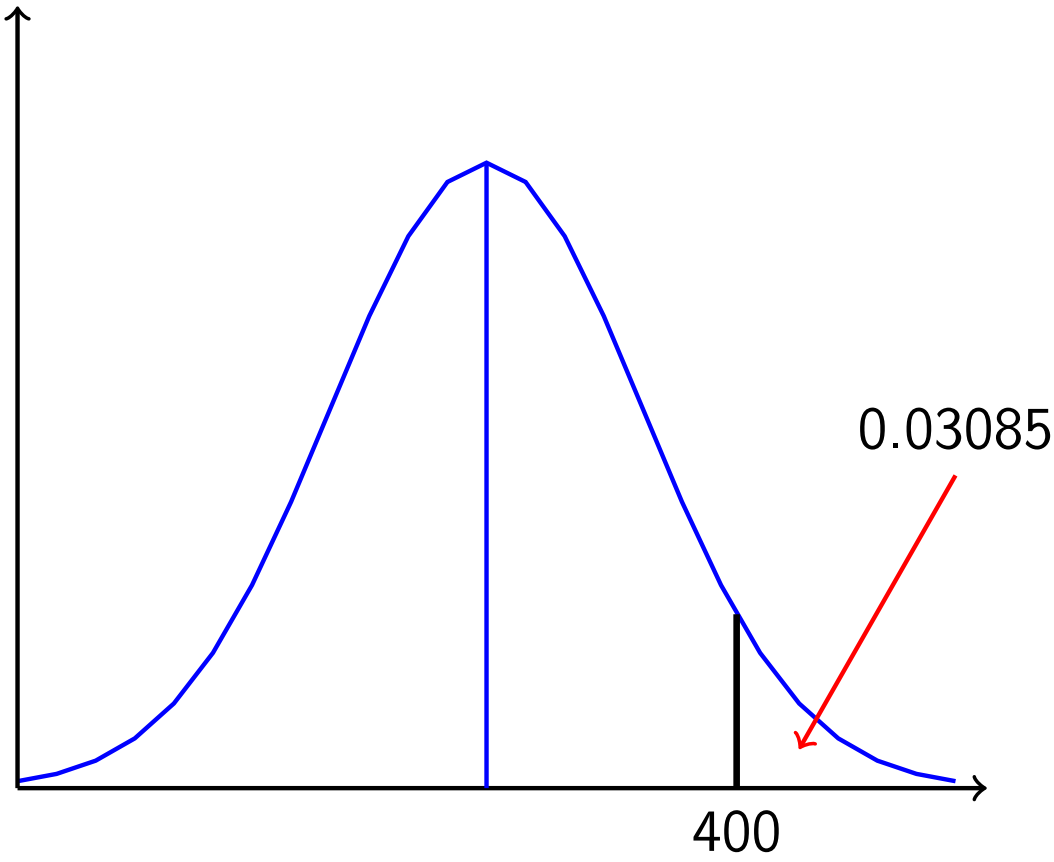

Figura 1: Representación gráfica de la muestral de la diferencia de medias muestrales.

$$P(\overline{X} - \overline{Y} > 400) = 0{,}03085$$

Distribución de la diferencia de medias muestrales en Poblaciones normalmente distribuidas con varianzas desconocidas

Ahora se desea encontrar la distribución la diferencia de dos medias muestrales cuando las varianzas poblacionales son desconocidas. Sean $X_1, X_2, \ldots, X_m$ y $Y_1, Y_2, \ldots, Y_n$ dos muestras aleatorias, cada una de distribuciones normales independientes con medias desconocidas μ_x y μ_y y varianzas desconocidas σ_x^2 y σ_y^2, respectivamente.

Se puede considerar dos casos, y a continuación tomar cada uno por turno:

(a) Ambas varianzas poblacionales son iguales, $\sigma_x^2 = \sigma_y^2 = \sigma^2$. En este caso, se asume que ambas muestras provienen de poblaciones que pueden tener varianzas iguales. Esto significa que se puede usar ambas muestras combinadas para estimar σ^2.

(b) Ambas varianzas poblacionales son desiguales, $\sigma_x^2 \neq \sigma_y^2$. En este caso, se asumimos que ambas muestras provienen de poblaciones que no tienen varianzas iguales, por lo que se debe estimar σ_x^2 y σ_y^2 por separado.

Caso 1:$\sigma_x^2 = \sigma_y^2 = \sigma^2$

Sea S_X^2 la varianza muestral de $X_1, X_2, \ldots, X_m$, y sea S_Y^2 la varianza muestral de $Y_1, Y_2, \ldots, Y_n$. Sabemos que S_X^2 es un estimador insesgado para σ_x^2 y que S_Y^2 es un estimador insesgado para σ_y^2. Como se asume que $\sigma_x^2 = \sigma_y^2 = \sigma^2$, entonces tanto S_X^2 como S_Y^2 son estimadores insesgados de σ^2.

La razón por la que este caso es interesante radica principalmente en la siguiente idea. Tenemos dos estimadores insesgados de σ^2. ¿Podríamos combinar estos estimadores de alguna manera para obtener otro estimador insesgado, que sea mejor que cualquiera de los dos estimadores insesgados originales? La respuesta a ésta pregunta es sí. Pero, ¿a qué nos referimos con mejor?

En esta situación, nos referimos a obtener otro estimador insesgado que tenga una varianza menor que cualquiera de los dos estimadores insesgados originales.

Se sabe que

$$\frac{(m-1)S_X^2}{\sigma_x^2} \sim \chi^2(m-1) \quad \text{y} \quad \frac{(n-1)S_Y^2}{\sigma_y^2} \sim \chi^2(n-1)$$

y por la propiedad de la distribución χ^2 se tiene que,

$$\frac{(m-1)S_X^2}{\sigma_x^2} + \frac{(n-1)S_Y^2}{\sigma_y^2} \sim \chi^2(n-1) \sim \chi^2(m+n-2)$$

y como las varianzas son iguales resulta

$$\frac{(m-1)S_X^2}{\sigma^2} + \frac{(n-1)S_Y^2}{\sigma^2} \sim \chi^2(n-1) \sim \chi^2(m+n-2) \qquad (*)$$

Por otro lado, se sabe que

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}} \sim \mathcal{N}(0,1).$$

Nuevamente cono las varianzas son iguales resulta,

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim \mathcal{N}(0,1). \qquad (**)$$

y teniendo en cuenta, que las medias muestrales son independientes de las varianzas muestrales, por definición del estadístico "t" student, efectuando la división del de relación de $(**)$ entre la relación de $(*)$ se obtiene

$$T = \frac{\frac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}}}{\sqrt{\frac{\frac{(m-1)S_X^2}{\sigma^2}+\frac{(n-1)S_Y^2}{\sigma^2}}{m+n-2}}} \sim t(m+n-2)$$

$$= \frac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{S\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim t(m+n-2)$$

donde

$$S^2 = \frac{(m-1)S_X^2+(n-1)S_Y^2}{m+n-2}$$

se llama **varianza ponderada** y es un estimador insesgado de σ^2.

ejemplo

Se midió la proporción de consumo de oxígeno de dos grupos de hombres. Un grupo (X) entrenó regularmente durante un período de tiempo, y el otro (Y) entrenó de forma intermitente. Las estadísticas calculadas a partir de los datos registrados se dan a continuación: Suponiendo

Grupo X :	m $= 9$	$\mu_x = 43{,}71$	$s_X^2 = 34{,}5744$
Grupo Y :	n $= 7$	$\mu_Y = 39{,}63$	$s_Y^2 = 58{,}9824$

distribuciones normales independientes para las lecturas, y que las varianzas de las poblaciones son iguales, encuentre la $P(\overline{X} > \overline{Y})$.

Solución:

$$
\begin{aligned}
P(\overline{X} > \overline{Y}) &= P(\overline{X} - \overline{Y} > 0) \\
&= P\left(T > \frac{0 - (43{,}71 - 39{,}63)}{\sqrt{\frac{(9-1)(34{,}5744)+(7-1)(58{,}9824)}{9+7-2}}} \right) \\
&= P\left(T > \frac{-4{,}08}{6{,}71{,}08} \right) \\
&= P(T > -0{,}60796) \\
&= 0{,}72328
\end{aligned}
$$

La probabilidad de que la media muestral del grupo mX sea mayor a la media muestral del grupo Y es 0.72328.

Caso 2: $\sigma_x^2 \neq \sigma_y^2$

Sea S_X^2 la varianza muestral de $X_1, X_2, \ldots, X_m$ y sea S_Y^2la varianza muestral de $Y_1, Y_2, \ldots, Y_n$. Si $\sigma_x^2 \neq \sigma_y^2$ entonces

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{m} + \frac{\S_Y^2}{n}}} \tag{2}$$

se puede demostrar que tiene una distribución "t" de Student con algunos grados de libertad, que se denota por ν. Actualmente, no existe una fórmula exacta para los grados de libertad ν en el caso de que ambas varianzas poblacionales sean desconocidas y no se puedan asumir iguales.

La denominada ecuación de Welch-Satterthwaite se utiliza para calcular una aproximación a los grados de libertad efectivos de una combinación lineal de varianzas muestrales independientes, también conocidas como grados de libertad agrupados. Esta ecuación está dada por

$$\nu = \frac{\left(\frac{S_X^2}{m} + \frac{S_Y^2}{n}\right)^2}{\frac{\left(\frac{S_X^2}{m}\right)^2}{m-1} + \frac{\left(\frac{S_Y^2}{n}\right)^2}{n-1}}$$

y se usa comúnmente en escenarios en los que se desconocen dos variaciones de población y no hay evidencia para suponer que son iguales.