1 Muestras "independientes" vs "apareadas"

¿Las personas son más altas a la mañana al despertarse que a la noche al acostarse?. En realidad sí, y esto es debido a que la columna se estira un poco al estar varias horas en posición horizontal.

Suponga ahora que se quiere estudiar este "crecimiento" δ, al dormir. Más concretamente se quiere un IC al 95% para δ. Se proponen dos diseños(**ambos válidos**) para investigar este asunto:

1.1 Diseño con muestras independientes

Se eligen al azar e independientemente n personas, se las visita bién temprano a la mañana, se las despierta, y se las mide. Se tendrá entonces la muestra:

$$\overset{independientes}{Y_1, Y_2, \cdot\cdot, Y_n} \sim N(\mu_y; \sigma) \tag{1}$$

Luego, se eligen al azar e independientemente **otras** n personas, se las visita bién tarde a la noche, antes de que se acuesten, y se las mide. Se tendrá entonces la muestra:

$$\overset{independientes}{X_1, X_2, \cdot\cdot, X_n} \sim N(\mu_x; \sigma) \tag{2}$$

Notar que estas dos muestras son también independientes **entre sí**, o sea

$$Y_1, Y_2, \cdot\cdot, Y_n \; independiente \; X_1, X_2, \cdot\cdot, X_n \tag{3}$$

Como se quiere un IC para $\delta = \mu_y - \mu_x$, y $\overline{Y}$, $\overline{X}$ son buenos estimadores insesgados de μ_y y μ_x respectivamente, se propone como estimador de δ, a $\widehat{\delta} = \overline{Y} - \overline{X}$. Usando el teorema de combinación lineal de normales

$$\widehat{\delta} = \overline{Y} - \overline{X} \sim N(\mu_y - \mu_x; \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}}) = N(\delta; \sqrt{\frac{2\sigma^2}{n}}) \tag{4}$$

Recordando que σ^2 se lo estima mediante el pool $S_p^2 = \frac{(n-1)S_y^2 + (n-1)S_x^2}{(n-1)+(n-1)} = \frac{S_y^2 + S_x^2}{2}$, entonces $\widehat{\delta}$ estima al δ de interés, con un error de estimación $\sqrt{\frac{2S_p^2}{n}}$.

Si se quiere un IC para δ, estandarizando(4) resulta $\frac{(\overline{Y}-\overline{X})-\delta}{\sqrt{\frac{2\sigma^2}{n}}} \sim N(0;1)$, y finalmente:

$$\text{IC}_\delta : \left[(\overline{Y} - \overline{X}) \pm t_{2(n-1);\frac{\alpha}{2}} \sqrt{\frac{2S_p^2}{n}} \right] \tag{5}$$

1.2 Diseño con muestras apareadas

Se eligen al azar e independientemente n personas, se las visita bién temprano a la mañana, se las despierta, y se las mide. Se tendrá entonces la muestra:

$$\overset{independientes}{Y_1, Y_2, \cdot\cdot, Y_n} \sim N(\mu_y; \sigma)$$

Luego, **a estas mismas** n personas, se las visita bién tarde a la noche, antes de que se acuesten, y se las mide. Se tendrá entonces la muestra:

$$\overset{independientes}{X_1, X_2, \cdot\cdot, X_n} \sim N(\mu_x; \sigma)$$

Notar que ahora estas dos muestras si bién son independientes "dentro", **NO son independientes "entre"**, ya que al tratarse de alturas de las mismas personas, X_i será muy dependiente de Y_i (obviamente: si una persona es 'alta' a la mañana, con toda seguridad será 'alta' a la noche y viceversa). De otra manera, la correlación ρ entre X_i y Y_i será bastante cercana a 1(por supuesto vale cero en el caso de un diseño con muestras independientes).

Como se quiere un IC para $\delta = \mu_y - \mu_x$, igual que antes se propone como estimador de δ, a $\widehat{\delta} = \overline{Y} - \overline{X}$. Pero como ahora $\overline{Y}$ y $\overline{X}$ son **dependientes**, no se puede usar el teorema de combinación lineal de normales para estudiar la distribución de $\widehat{\delta}$ vista en (4), (ya que se necesitaria conocer ρ). Se recurrirá a un "artificio".

Se tienen 2 muestras, pero se construye una nueva muestra, llamada "apareada", definida así $Z_i = Y_i - X_i$. O sea la nueva muestra será $Z_1, Z_2, \cdot\cdot, Z_n$. ¿De que población viene esta muestra? Como $Z_i = Y_i - X_i$ y recordando resultados sobre combinación lineal:

$$\begin{aligned}
\mu_z &= \mu_y - \mu_x = \delta \qquad (6)\\
\sigma_z^2 &= Var(Y_i) + Var(X_i) - 2Cov(Y_i; X_i) \text{ y como } \rho = \frac{Cov(Y_i; X_i)}{\sigma_y \sigma_x} = \frac{Cov(Y_i; X_i)}{\sigma^2}\\
\sigma_z^2 &= \sigma^2 + \sigma^2 - 2\rho\sigma^2 = 2\sigma^2(1-\rho)
\end{aligned}$$

Luego se tiene que

$$\overset{independientes}{Z_1, Z_2, \cdot\cdot, Z_n} \sim N(\delta; \sqrt{2\sigma^2(1-\rho)}) \qquad (7)$$

No se conoce ρ y tampoco σ, pero que tal si agrupamos estas ignorancias llamando $\sqrt{2\sigma^2(1-\rho)} = \sigma_z$, entonces queda

$$\overset{independientes}{Z_1, Z_2, \cdot\cdot, Z_n} \sim N(\delta; \sigma_z) \qquad (8)$$

y ahora sí, como estimador de δ se usará $\widehat{\delta} = \overline{Z}$ y entonces

$$\widehat{\delta} = \overline{Z} \sim N(\delta; \sqrt{\frac{\sigma_z^2}{n}}) = N(\delta; \sqrt{\frac{2\sigma^2(1-\rho)}{n}}) \qquad (9)$$

En este caso σ_z^2 se lo estima mediante S_z^2, y entonces $\widehat{\delta}$ estima al δ de interés, con un error de estimación $\sqrt{\frac{S_z^2}{n}}$.

Si se quiere un IC para δ, estandarizando(9) resulta $\frac{\overline{Z}-\delta}{\frac{\sigma_z}{\sqrt{n}}} \sim N(0;1)$, y entonces se tendrá finalmente:

$$\text{IC}_\delta : \left[\overline{Z} \pm t_{(n-1);\frac{\alpha}{2}} \sqrt{\frac{S_z^2}{n}}\right] \qquad (10)$$

Example 1 *Supóngase que usando muestras independientes con* $n = 10$ *se tiene para las alturas a la mañana y a la noche(son 20 personas en total)*

Y	*171.16*	*181.77*	*171.09*	*166.67*	*186.71*	*185.83*	*174.97*	*166.34*	*162.99*	*172.79*
X	*175.34*	*178.24*	*169.75*	*176.12*	*161.61*	*167.73*	*181.96*	*183.50*	*164.77*	*175.04*

de (4), $\widehat{\delta} = \overline{Y} - \overline{X} = 174.03 - 173.41 = 0.62cm$, y como

$$S_p^2 = \frac{S_y^2 + S_x^2}{2} = \frac{68.32 + 52.41}{2} = 60.37$$

resulta que el desvío del estimador es (usando también 4) $\sqrt{\frac{2S_p^2}{n}} = \sqrt{\frac{260.37}{10}} = 5.10cm$. Notar que este valor es muy grande ya que es el desvío correspondiente a un crecimiento estimado de $\widehat{\delta} = 0.62cm$. Además usando la expresión (5) se tiene para un nivel de confianza 95% que el crecimiento δ estará entre$[0.62 \pm 2.101 * 5.1] = [-10.1; 11.34]\, cm$, realmente muy ancho.

Por otro lado si se utiliza el diseño con apareo, suponiendo la misma muestra a la mañana, y a la noche las alturas de estas mismas personas(10 personas en total)

Y	171.16	181.77	171.09	166.67	186.71	185.83	174.97	166.34	162.99	172.79
X	170.70	181.22	170.67	166.13	186.33	185.26	174.41	165.87	162.53	172.38

la muestra "apareada" será

Z	0.4557	0.5526	0.41995	0.5354	0.3830	0.5709	0.5595	0.4714	0.4425	0.4132

notar que ahora Z contiene los crecimientos de cada persona. En este caso $\widehat{\delta} = \overline{Z} = 0.48cm$, y el desvío de ese estimador es $\sqrt{\frac{S_z^2}{n}} = \sqrt{\frac{0.004721}{10}} = 0.02173cm$ realmente muy razonable. Y el intervalo de confianza será $[0.48 \pm 2.26 * 0.02173] = [0.43; 0.53]\, cm$.

1.3 Comentarios

- En este ejemplo ¿Cuál diseño es mejor? . Decididamente el con apareo.
- Como se vió, si las muestras se tomaron bajo un diseño apareado(con dependencias entre X_i y Y_i), no se pueden hallar IC(ni hacer PH) usando las expresiones para muestras independientes, ya que no lo son. Sin embargo si las muestras se tomaron con un diseño independiente, sí se puede construir la muestra apareada, y usar sus fórmulas. Pero esto no es conveniente pues si se observa (9), la varianza del estimador apareado(como en este caso $\rho = 0$), será la misma que (4), con el estimador bajo independencia. O sea habrá igualdad de varianzas del estimador, pero de (10) la t de Student tendrá solo n-1 grados de libertad en lugar de los 2(n-1) si se hubiesen usado las expresiones para muestras independientes.

- ¿Cuando es más eficiente usar un diseño con apareo? La respuesta está en (9), o sea cuando el criterio de apareamiento de las muestras hace que ρ sea positivo y lo mayor posible, ya que así se reduce la varianza del estimador. Es el caso del problema analizado: apareando por individuo, como la altura de una misma persona al acostarse y al levantarse están fuertemente correlacionados, es conveniente un diseño con apareo.

Example 2 *Problema 25 de la página 341 del libro de Walpole y Myers. Una compañia de taxis está tratando de decidir si el uso de neumáticos radiales en lugar de los comunes mejora la economía de combustible. Se equiparon 12 automóviles con neumáticos radiales y recorrieron una trayectoria de prueba fija.* ***Sin cambiar conductores****, se equipó a esos mismos vehículos con neumáticos comunes, y de nuevo recorrieron el mismo trayecto. Se registró el consumo de gasolina en Km/litro como sigue:*

Auto	***Radial***	***Común***
1	*4.2*	*4.1*
2	*4.7*	*4.9*
3	*6.6*	*6.2*
4	*7.0*	*6.9*
5	*6.7*	*6.8*
6	*4.5*	*4.4*
7	*5.7*	*5.7*
8	*6.0*	*5.8*
9	*7.4*	*6.9*
10	*4.9*	*4.7*
11	*6.1*	*6.0*
12	*5.2*	*4.9*

(En este ejemplo, está explicitamente señalado en el enunciado que es un diseño con apareo, debido al texto **"Sin cambiar conductores").**

Sin embargo notar que aunque esta frase no estuviese, es razonable que convenga un diseño con apareo por conductor, ya que, cada conductor tiene una forma de frenar, acelerar, etc. que le es propia. Entonces es esperable que si un conductor gasta mucho con cubiertas radiales, también gastará mucho con las comunes; y al revés, si gasta poco con las radiales seguramente gastará poco con las comunes. Esto quiere decir que la correlación ρ entre Radiales y Comunes será positiva y alta, y de aquí la conveniencia del diseño con apareo.

De otra manera: al detectar una "conducta propia" de cada conductor, que estará presente ya sea manejando con radiales o con comunes, se deduce la conveniencia de aparear. Notar que al restar para obtener la muestra apareada, los efectos de las "conductas propias" sobre el consumo de combustible se cancelan, y queda la muestra Z, con información más clara respecto de la relación entre tipo de cubiertas y consumo.