Clase 13: Comparación de m muestras

Matías Carrasco

6 de octubre de 2019

Índice

1. Motivación

Con frecuencia, se desea comparar más de dos tratamientos o poblaciones. A veces, las diferencias entre tratamientos se deben a cambios en el nivel de un determinado factor, como por ejemplo diferentes dosis de un medicamento determinado. Por lo tanto, el problema en el cual nos centramos en esta clase consiste en testar la igualdad de las medias de las diversas distribuciones en un experimento de un "factor".

En la clase pasada discutimos cómo comparar las medias de dos poblaciones normales. Consideremos ahora m poblaciones normales con medias desconocidas $\mu_1, \mu_2, \ldots, \mu_m$ y varianza desconocida, pero común, σ^2. Deseamos poner a prueba la igualdad de las m medias, a saber

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_m = \mu, \text{ para algún } \mu \text{ sin especificar.}$$

contra la hipótesis alternativa H_1 : no H_0.

Consideremos un muestreo aleatorio de estas poblaciones: sean

$$X_{i1}, X_{i2}, \ldots, X_{in_i}$$

un muestreo aleatorio de tamaño n_i de la distribución normal $N(\mu_i, \sigma^2)$, para $i = 1, 2, \ldots, m$. Es útil pensar en este muestreo como una tabla:

Población					Media
X_1	X_{11}	X_{12}	$\cdots$	X_{1n_1}	$\overline{X}_{1\cdot}$
X_2	X_{21}	X_{22}	$\cdots$	X_{2n_2}	$\overline{X}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_m	X_{m1}	X_{m2}	$\cdots$	X_{mn_m}	$\overline{X}_{m\cdots}$
Media general					$\overline{\overline{X}}_{\cdot\cdot}$

En la tabla hemos indicado las variables muestreadas junto con las medias de fila (medias muestrales), en donde poniendo $n = n_1 + n_2 + \cdots + n_m$ tenemos

$$\overline{X}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i} X_{ij} \quad \text{y} \quad \overline{X}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$$

El punto en la notación para las medias, $\overline{X}_{\cdot\cdot}$ y $\overline{X}_{i\cdot}$, indica el índice sobre el cual se toma el promedio. Aquí $\overline{X}_{\cdot\cdot}$ es el promedio tomado sobre ambos índices, mientras que $\overline{X}_{i\cdot}$ se toma solo sobre el índice j.

2. Partición de la varianza

Para determinar una región crítica para la prueba de H_0, primero dividiremos la suma de los cuadrados asociados con la varianza de las muestras combinadas en dos partes. Esta suma de cuadrados viene dada por

$$\begin{aligned}
\text{SS(TO)} &= \sum_{i=1}^{m}\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{\cdot\cdot})^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot} + \bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i} (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2 + 2\sum_{i=1}^{m}\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot})(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})
\end{aligned}$$

El último término del miembro derecho de esta identidad puede escribirse como

$$2\sum_{i=1}^{m}\left[(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})\right] = 2\sum_{i=1}^{m}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})(n_i\bar{X}_{i\cdot} - n_i\bar{X}_{i\cdot}) = 0$$

y el término anterior puede escribirse como

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2 = \sum_{i=1}^{m} n_i (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2$$

Entonces

$$\text{SS(TO)} = \sum_{i=1}^{m}\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot})^2 + \sum_{i=1}^{m} n_i (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2$$

La notación clásica es la siguiente:

$$\text{SS(TO)} = \sum_{i=1}^{m}\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{\cdot\cdot})^2 \qquad \text{(suma total de cuadrados)}$$

$$\text{SS(E)} = \sum_{i=1}^{m}\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i})^2 \quad \text{(suma de cuadrados dentro de los grupos, o error)}$$

$$\text{SS(T)} = \sum_{i=1}^{m} n_i (\bar{X}_{i} - \bar{X}_{\cdot\cdot})^2 \qquad \text{(suma de cuadrados entre los diferentes grupos)}$$

Luego, con esta notación, la partición de sumas de cuadrados es

$$\text{SS(TO)} = \text{SS(E)} + \text{SS(T)}$$

Cuando H_0 es cierta, podemos pensar a X_{ij}, con $i = 1,2,\ldots,m$, y $j = 1,2,\ldots,n_i$, como un muestreo aleatorio de tamaño $n = n_1 + n_2 + \cdots + n_m$ de la distribución normal $N(\mu, \sigma^2)$. Entonces $\mathrm{SS(TO)}/(n-1)$ es un estimador insesgado de σ^2. Pero un estimador insesgado de σ^2 basado solo en la muestra del i-ésimo grupo es

$$W_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot})^2 \qquad \text{para } i = 1,2,\ldots,m$$

y además $(n_i - 1)W_i/\sigma^2$ es $\chi^2(n_i - 1)$.

De ello se deduce que la suma de estas m variables chi-cuadrado independientes, a saber,

$$\sum_{i=1}^{m} \frac{(n_i - 1)\, W_i}{\sigma^2} = \frac{\mathrm{SS(E)}}{\sigma^2}$$

también es chi-cuadrado con $(n_1 - 1) + (n_2 - 1) + \cdots + (n_m - 1) = n - m$ grados de libertad. Por lo tanto, $\mathrm{SS(E)}/(n-m)$ es un estimador insesgado de σ^2. Ahora tenemos

$$\frac{\mathrm{SS(TO)}}{\sigma^2} = \frac{\mathrm{SS(E)}}{\sigma^2} + \frac{\mathrm{SS(T)}}{\sigma^2}$$

donde

$$\frac{\mathrm{SS(TO)}}{\sigma^2} \text{ es } \chi^2(n-1) \quad \text{y} \quad \frac{\mathrm{SS(E)}}{\sigma^2} \text{ es } \chi^2(n-m)$$

Usando argumentos geométricos sobre proyecciones ortogonales, idénticos a los que hemos usado en las clases anteriores, se puede ver que SS(E) y SS(T) son independientes y la distribución de $\mathrm{SS(T)}/\sigma^2$ es $\chi^2(m-1)$.

3. Análisis de varianza

Como bajo H_0, la suma $\mathrm{SS(T)}/\sigma^2$ es $\chi^2(m-1)$, tenemos $\mathbf{E}\left(\mathrm{SS(T)}/\sigma^2\right) = m - 1$ y se deduce que $\mathbf{E}\left(\mathrm{SS(T)}/(m-1)\right) = \sigma^2$. Ahora, el estimador $\mathrm{SS(E)}/(n-m)$ de σ^2 siempre es insesgado, ya sea que H_0 sea cierta o no. Sin embargo, si las medias $\mu_1, \mu_2, \ldots, \mu_m$ no son iguales, el valor esperado del estimador de σ^2 basado en SS(T) será mayor que σ^2. En efecto, tenemos

$$\begin{aligned}
\mathbf{E}\left(\mathrm{SS(T)}\right) &= \mathbf{E}\left(\sum_{i=1}^{m} n_i \left(\bar{X}_i - \bar{X}_{\cdot\cdot}\right)^2\right) = \mathbf{E}\left(\sum_{i=1}^{m} n_i \bar{X}_{i\cdot}^2 - n\bar{X}_{\cdot\cdot}^2\right) \\
&= \sum_{i=1}^{m} n_i \left\{ \mathrm{Var}\left(\bar{X}_i\right) + \left[\mathbf{E}\left(\bar{X}_i\right)\right]^2 \right\} - n\left\{ \mathrm{Var}(\bar{X}_{\cdot\cdot}) + \left[\mathbf{E}\left(\bar{X}_{\cdot\cdot}\right)\right]^2 \right\} \\
&= \sum_{i=1}^{m} n_i \left\{ \frac{\sigma^2}{n_i} + \mu_i^2 \right\} - n\left\{ \frac{\sigma^2}{n} + \bar{\mu}^2 \right\} \\
&= (m-1)\sigma^2 + \sum_{i=1}^{m} n_i \left(\mu_i - \bar{\mu}\right)^2
\end{aligned}$$

en donde $\bar{\mu} = (1/n)\sum_{i=1}^{m} n_i \mu_i$. Si las medias son todas iguales, entonces

$$\mathbf{E}\left(\frac{\mathrm{SS(T)}}{m-1}\right) = \sigma^2$$

Si las medias no son todas iguales, entonces

$$\mathbf{E}\left(\frac{\mathrm{SS(T)}}{m-1}\right) = \sigma^2 + \sum_{i=1}^{m} n_i \frac{(\mu_i - \bar{\mu})^2}{m-1} > \sigma^2$$

Podemos basar la prueba de H_0 en el cociente de $\mathrm{SS(T)}/(m-1)$ y $\mathrm{SS(E)}/(n-m)$, los cuales son estimadores insesgados de σ^2, siempre que $H_0 : \mu_1 = \mu_2 = \cdots = \mu_m$ sea cierta. Bajo H_0, el cociente debería tomat valores cercanos a 1. Sin embargo, en el caso de que las medias sean diferentes, el cociente debería ser grande, ya que $\mathbf{E}\left(\mathrm{SS(T)}/(m-1)\right)$ es mayor a σ^2.

Bajo H_0, el cociente

$$F = \frac{\mathrm{SS(T)}/(m-1)}{\mathrm{SS(E)}/(n-m)} = \frac{\left[\mathrm{SS(T)}/\sigma^2\right]/(m-1)}{\left[\mathrm{SS(E)}/\sigma^2\right]/(n-m)}$$

tiene distribución de Fisher (también llamada F) con $m-1$ y $n-m$ grados de libertad porque $\mathrm{SS(T)}/\sigma^2$ y $\mathrm{SS(E)}/\sigma^2$ son variables chi-cuadrado independientes. Rechazaríamos H_0 si el valor observado de F_{obs} es demasiado grande, porque esto indicaría que tenemos un $\mathrm{SS(T)}$ relativamente grande, lo que sugiere que las medias son desiguales. Por lo tanto, la región crítica tiene la forma $F \geq F_{m-1,n-m}(\alpha)$.

La información utilizada para el test de igualdad de medias a menudo se resume en una tabla de análisis de varianza, o tabla ANOVA, como la que se muestra en la Tabla 1, donde el cuadrado medio (MS) es la suma de cuadrados (SS) dividida por sus grados de libertad.

Fuente	Suma de cuadrados	Grados de libertad	MS	F
Tratamiento	SS(T)	$m-1$	$\mathrm{MS(T)} = \frac{\mathrm{SS(T)}}{m-1}$	$\frac{\mathrm{MS(T)}}{\mathrm{MS(E)}}$
Error	SS(E)	$n-m$	$\mathrm{MS(E)} = \frac{\mathrm{SS(E)}}{n-m}$	
Total	SS(TO)	$n-1$		

Tabla 1: Tabla ANOVA.

Las siguientes fórmulas a veces simplifican los cálculos de SS(TO), SS(T) y SS(E) (y también reducen los errores de redondeo creados por restar los promedios de las observaciones):

$$\mathrm{SS(TO)} = \sum_{i=1}^{m}\sum_{j=1}^{n_i} X_{ij}^2 - \frac{1}{n}\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i} X_{ij}\right]^2$$

$$\mathrm{SS(T)} = \sum_{i=1}^{m}\frac{1}{n_i}\left[\sum_{j=1}^{n_i} X_{ij}\right]^2 - \frac{1}{n}\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i} X_{ij}\right]^2$$

y

$$\mathrm{SS(E)} = \mathrm{SS(TO)} - \mathrm{SS(T)}$$

Ejemplo 1

Una ventana que se fabrica para un automóvil tiene cinco pernos para fijarla. Una compañía que fabrica estas ventanas realiza "pruebas de extracción" para determinar la fuerza necesaria para sacar un perno de la ventana. Sea X_i, $i = 1,2,3,4,5$, igual a la fuerza requerida en la posición i, y suponga que la distribución de X_i es $N(\mu_i, \sigma^2)$. Queremos probar la hipótesis nula $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$, utilizando siete observaciones independientes en cada

posición. A un nivel de significancia $\alpha = 0.01$, H_0 se rechaza si

$$F = \frac{\text{SS(T)}/(5-1)}{\text{SS(E)}/(35-5)} \geq 4.02 = F_{0.01}(4,30)$$

Los datos observados, junto con ciertas sumas, se muestran en las tablas siguientes:

	Observaciones							$\sum_{j=1}^{7} x_{ij}$	$\sum_{j=1}^{7} x_{ij}^2$
X_1:	92	90	87	105	86	83	102	645	59847
X_2:	100	108	98	110	114	97	94	721	74609
X_3:	143	149	138	136	139	120	145	970	134936
X_4:	147	144	160	149	152	131	134	1017	148367
X_5:	142	155	119	134	133	146	152	981	138415
Total								4334	556174

Para estos datos

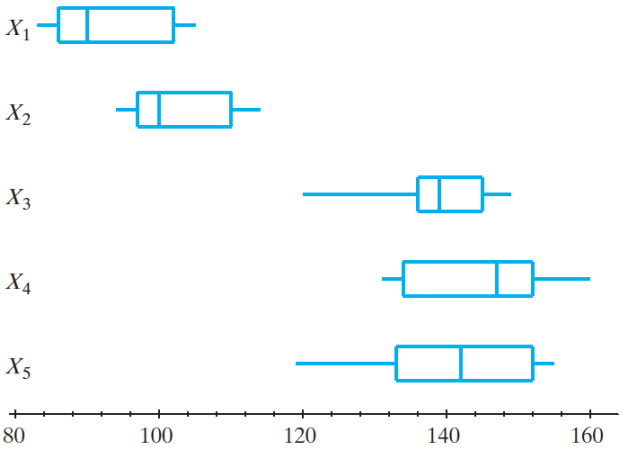
$$\text{SS(TO)} = 556174 - \frac{1}{35}(4334)^2 = 19500.97$$
$$\text{SS(T)} = \frac{1}{7}\left(645^2 + 721^2 + 970^2 + 1017^2 + 981^2\right) - \frac{1}{35}(4334)^2 = 16672.11$$
$$\text{SS(E)} = 19500.97 - 16672.11 = 2828.86$$

Como el valor observado de F es entonces

$$F_{\text{obs}} = \frac{16672.11/4}{282886/30} = 44.20$$

la hipótesis nula es claramente rechazada.

Fuente	Suma de cuadrados	Grados de libertad	MS	F
Tratamiento	16672.11	4	4168.03	44.20
Error	2828.86	30	94.30	
Total	19500.97	34		



Pero, ¿por qué se rechaza H_0? Los diagramas de caja (boxplot) que se muestran en la figura ayudan a responder esta pregunta. Parece que las fuerzas requeridas para sacar los pernos en las posiciones 1 y 2 son similares, y las de las posiciones 3, 4 y 5 son bastante similares, pero diferentes de las posiciones 1 y 2. ■