Semana 3

Comparar dos grupos *Valentín Vergara Hidd*

La primera aplicación para un problema más complejo que involucra la prueba de hipótesis, consiste en estudiar dos muestras y compararlas. ¿Podemos decir, con cierta confianza, que dos muestras cualquiera provienen de la misma población? ¿Podemos afirmar que son equivalentes?

El problema es aún más interesante cuando consideramos los casos a los que nos enfrentamos trabajando con datos reales. Por ejemplo, muestras que tienen tamaños diferentes, que tienen parámetros poblacionales μ y σ desconocidos; o que ni siquieran corresponden a dos grupos diferentes, sino que son dos muestras proporcionadas por las mismas unidades.

Para tratar de abordar todos estos escenarios, presentaremos algunas herramientas, siendo la más importante la distribución de probabilidad t.

---

1. Trabajando con dos muestras

Lo más importante al tratar de usar una prueba de hipótesis con dos muestras, es que estamos trabajando con dos *subconjuntos*. Si llamamos a las muestras s_1 y s_2, hay que tener muy claro que, en principio, no sabemos si ambas pertenecen a la misma población o no. De esta forma, el primer problema es verificar si efectivamente son dos subconjuntos de la misma población, o son independientes.

En caso que no pertenezcan a la misma población, llamaremos $\overline{x}_1$ a la media aritmética de los valores obtenidos en la muestra s_1; en tanto que $\overline{x}_2$ denomina a la media aritmética para los valores obtenidos en la muestra s_2. Así, generamos dos posibles hipótesis nulas (con sus respectivas hipótesis alternativas):

$$H_0: \quad \overline{x}_1 = \mu_1 \quad H_1: \quad \overline{x}_1 \neq \mu_1 \tag{1}$$
$$H_0: \quad \overline{x}_2 = \mu_2 \quad H_1: \quad \overline{x}_2 \neq \mu_2 \tag{2}$$

En caso que ambas muestras pertenezcan a la misma población, no sabemos si son equivalentes. Puede ser, que aún teniendo el mismo tamaño, s_1 y s_2 no sean *igual de buenos modelos* de la población. Podemos plantear una hipótesis nula y alternativa para este problema:

$$H_0: \quad \overline{x}_1 = \overline{x}_2 \quad H_1: \quad \overline{x}_1 \neq \overline{x}_2 \tag{3}$$

Finalmente, otro caso es que ambas muestras provengan exactamente de las mismas unidades (personas, instituciones, cosas, etc); pero en distintas mediciones. Esto generalmente ocurre cuando medimos la muestra en distintos momentos del tiempo. Por ejemplo, cuando se quiere probar la efectividad de una intervención. Las hipótesis, en este caso:

$$H_0: \quad \overline{x}_{1,t=1} = \overline{x}_{1,t=2} \quad H_1: \quad \overline{x}_{1,t=1} \neq \overline{x}_{1,t=2} \tag{4}$$

En este documento probaremos hipótesis para los tres casos mencionados. La lógica es similar; y la implementación en SPSS es diferente para cada caso, pero los resultados se leen exactamente de la misma forma.

2. Caso A: ¿Pertenece la muestra a la población?

Para ilustrar como funciona esta prueba, utilizaré el mismo ejemplo usado en Denis (2019). Supongan que se aplica una prueba de inteligencia a 5 personas, con resultados $\{105, 98, 110, 105, 95\}$. A partir de esta muestra $\mathcal{S}_1$ se obtiene $\overline{x} = 102,6$ y $s_x = 6,02$. Si sabemos que la prueba que se usó en el ejemplo fue diseñada para tener $\mu = 100$, la pregunta de investigación es, ¿pertenecen los puntajes de s_1 a la población? *Traduciendo* esta pregunta a una hipótesis nula (y alternativa), obtenemos:

$$\begin{aligned} H_0: &\quad \overline{x} = \mu = 100 \\ H_1: &\quad \overline{x} \neq \mu \end{aligned}$$

Siguiendo la lógica presentada en documentos anteriores, el segundo paso de la prueba de hipótesis implica fijar un valor de α y especificar la distribución de probabilidad de los resultados de *muchas* muestras. Sin embargo, no podemos utilizar la distribución normal de la misma forma que lo hemos hecho hasta ahora, ya que para ello sería necesario conocer σ. En vista que esta información no fue proporcionada, una alternativa sería hacer *bootstrapping* y obtener una aproximación a σ. Esto no es muy factible en el ejemplo, ya que la muestra es muy pequeña para obtener muestras relevantes.

La solución, entonces, es utilizar otra distribución que no requiere conocer el valor de σ. La distribución t utiliza μ y s_x. La función de densidad de probabilidad en una distribución t es:

$$f(x) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)}\left(1 + \frac{x^2}{v}\right)^{-\frac{v+1}{2}} \tag{5}$$

donde Γ es una función cuya interpretación es muy similar a la función factorial[1]; y v son los *grados de libertad*. Inmediatamente surgen algunas preguntas:

- ¿Qué pasó con μ y s_x?
- ¿Qué son los grados de libertad?

Para responder a ambas preguntas, se debe considerar que usualmente, para simplificar los cálculos, se busca que el área bajo la curva descrita por la función de densidad de probabilidad sea igual a 1. Para ello, cuando la distribución es normal, lo que hacemos es crear una versión de la distribución con $\mu = 0$ y $\sigma = 1$ utilizando una *conversión* de cada valor x_i dada por la expresión:

$$Z_i = \frac{x_i - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{6}$$

Para la distribución t, también haremos una **conversión**, de tal forma que el área bajo la curva descrita por la ecuación 5 sea igual a 1. La conversión es:

[1] Más información aquí

$$t = \frac{\overline{x} - \mu}{\frac{s_x}{\sqrt{n}}} \tag{7}$$

Con esto respondemos la primera pregunta. La segunda es un poco más técnica, pero la versión simplificada es que corresponde al número de formas en que la muestra puede variar sin alterar el resultado de 7. Estos grados de libertad toman un valor de $n - 1$, donde n es el tamaño de la muestra.

Para el cálculo de la prueba estadística, el tercer paso en la prueba de hipótesis, simplemente tomamos el valor de la ecuación 7 reemplazando $\overline{x}$ con el promedio muestral. En este caso,

$$t = \frac{102,6 - 100}{\frac{6,02}{\sqrt{5}}} = 0,965 \tag{8}$$

Para el cuarto paso, tomar la decisión respecto de H_0, es necesario situar el valor obtenido en 8 en la distribución de probabilidad descrita en la ecuación 5. Graficando la función, se vería más o menos así:

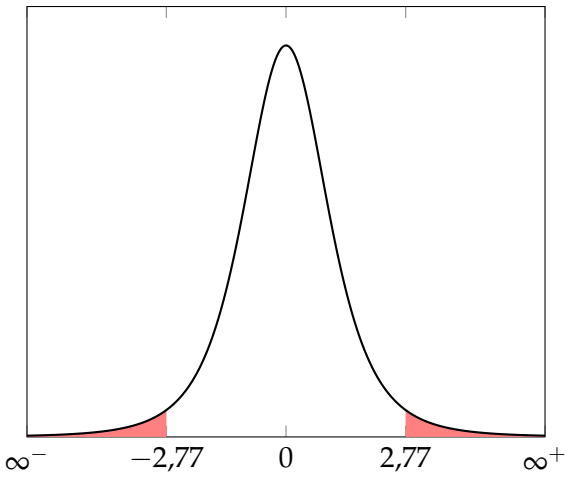

Figura 1: Función de densidad de probabilidad t con $v = 5$ y $\alpha = 0,05$

La interpretación es sencilla: si el valor calculado en 8 está en el área roja, significa que **no se acepta** H_0. Como no es el caso, aceptamos que el promedio muestral es igual al promedio poblacional. Por ende, es seguro asumir que la muestra $\mathcal{S}_1$ fue obtenida de una población con $\mu = 100$.

3. Caso B: Dadas dos muestras, ¿pertenecen a la misma distribución?

En este caso, es necesario comparar dos muestras. Para ello, asumimos que $\overline{x}_1 \rightarrow \mu_1$ y que $\overline{x} \rightarrow \mu_2$. Por tanto, el primer paso define la hipótesis nula y alternativa como:

$$\begin{aligned} H_0 &: \quad \mu_1 = \mu_2 \\ H_1 &: \quad \mu_1 \neq \mu_2 \end{aligned}$$

Noten que al escribir las hipótesis de esta forma, permite además de responder la preguntas que encabeza esta sección, responder la pregunta sobre la **igualdad** de ambas muestras. Otro detalle interesante es que las muestras pueden tener distinto tamaño y aún así el procedimiento que detallaré a continuación funciona.

El segundo paso implica definir α y v. En este caso, considerando una muestra 1 con tamaño n_1 y una muestra 2 con tamaño n_2, $v = (n_1 - 1) + (n_2 - 1)$.

La prueba estadística se calcula de forma diferente, en caso de tener igual número de elementos en ambas muestras, o si $n_1 \neq n_2$.

$$t = \begin{cases} \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} & , \text{si } n_1 = n_2 \\ \\ \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} & , \text{si } n_1 \neq n_2 \end{cases} \tag{9}$$

Noten que el término s_p^2 aún no ha sido definido; y que corresponde a una corrección que toma en cuenta el tamaño de cada muestra y su varianza. Formalmente,

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{10}$$

Por ahora, **asumiremos que las varianzas de ambos grupos son estadísticamente iguales**. Esto es un supuesto importante, que podremos levantar una vez que veamos otra distribución de probabilidad que permite probar dicho supuesto.

Finalmente, se toma la decisión al posicionar el valor obtenido en la ecuación 9 en la distribución de probabilidad t, dado α y v.

Ejemplo 1: Calificaciones

Este ejemplo es de Denis (2019). Se midió el tiempo de estudio de un grtupo de estudiantes y luego se observó quienes aprobaron y reprobaron. Los resultados son:

Resultado	Tiempo de estudio (minutos)
Reprueba	30
Reprueba	25
Reprueba	59
Reprueba	42
Reprueba	31
Aprueba	140
Aprueba	90
Aprueba	95
Aprueba	170
Aprueba	120

Estos datos muestran que $\bar{x}_A = 123$; $s_A = 13,57$ y $\bar{x}_R = 37,4$; $s_R = 33,09$. Luego, es

posible definir el primer paso de la prueba de hipótesis:

$$H_0: \quad \overline{x}_A = \overline{x}_R$$
$$H_1: \quad \overline{x}_A \neq \overline{x}_R$$

Noten que construímos las hipótesis de esta forma, porque si ambas muestras provienen de la misma población $\overline{x}_A = \overline{x}_R = \mu$. Existe la posibilidad de que ambas muestras no provengan de la distribución poblacional, pero para efectos de esta prueba, es un riego que podemos asumir[a].
Para el segundo paso, utilizaremos $\alpha = 0,05$ y $v = (5-1) + (5-1) = 8$. Con estos valores, se obtiene la siguiente distribución t:

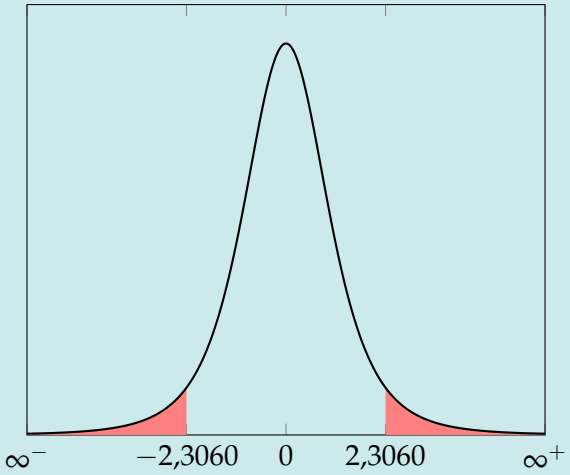

Figura 2: Función de densidad de probabilidad t con $v = 8$ y $\alpha = 0,05$

Para obtener los valores que delimitan el área de rechazo de H_0 (área roja), se puede integrar la función en un intervalo definido, o se puede consultar alguna Tabla como esta.
Dado que ambos grupos tienen la misma cantidad de elementos, la prueba estadística se calcula con:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{123 - 37,4}{\sqrt{\frac{33,09^2}{5} + \frac{13,57^2}{5}}} = 5,351 \tag{11}$$

Finalmente, en el cuarto paso, se posiciona el valor obtenido en la ecuación 11 en la distribución de la Figura 2. Como el valor se encuentra en el área roja, se rechaza H_0, lo que implica que ambos grupos no son iguales. En el contexto del problema, significa que el tiempo de estudio es estadísticamente diferente para quienes aprobaron y quienes reprobaron.

[a] Si existen dudas, siempre se puede hacer una prueba de hipótesis como la del caso A

4. Caso C: ¿Existen variaciones en la muestra en distintas mediciones?

Este es un caso particular, pero no por ello menos utilizado en ciencia. Si contamos con información de distintas mediciones de una variable a **las mismas unidades de análisis** (las mismas personas, instituciones, cosas, cualquier objeto de estudio), ¿podemos afirmar que estas mediciones, son diferentes **considerándolas de forma grupal**?

En términos generales, si medimos la variable aleatoria X en un grupo de tamaño n, obtendremos un conjunto de medidas $\{x_1, x_2, x_3, \ldots, x_n\}$. Si además, consideramos que podemos efectuar la misma medición en diferentes momentos, estamos introduciendo un nuevo indicador al que llamaremos j. Siguiendo el ejemplo, la primera vez que medimos la variable obtendremos un conjunto $\{x_{1,1}, x_{2,1}, x_{3,1}, \ldots, x_{n,1}\}$, la segunda medición tendrá al conjunto $\{x_{1,2}, x_{2,2}, x_{3,2}, \ldots, x_{n,2}\}$. Podemos agregar nuevas mediciones, de manera que en términos generales, cada medición para un elemento determinado de la muestra es $x_{i,j}; i \in \{1, 2, 3, \ldots, n\}; j \in \{1, 2, 3, \ldots, k\}$. Consideremos un caso simple, en el que $k = 2$. Para responder a la pregunta del encabezado, utilizaremos una prueba de hipótesis, siguiendo la misma lógica que se ha presentado hasta ahora.

Así, el primer paso consiste en identificar las hipótesis:

$$\begin{aligned} H_0 &: \quad \overline{x}_{\cdot,1} = \overline{x}_{\cdot,2} \\ H_1 &: \quad \overline{x}_{\cdot,1} \neq \overline{x}_{\cdot,2} \end{aligned}$$

Para luego, en el segundo paso, establecer valores para α y los grados de libertad, que para este problema corresponden a $v = n - 1$. Con esta información se puede obtener la distribución de probabilidad quse se utilizará en los siguientes pasos.

El tercer paso consiste en calcular la prueba estadística, que para este problema definiremos como:

$$t = \frac{\overline{D}}{s_D / \sqrt{n}} \tag{12}$$

Donde D es una medida agregada que se obtiene a partir de las diferencias individuales para cada medición. Es decir,

$$\overline{D} = \frac{1}{n} \sum_{i=1}^{n} d_i \tag{13}$$

$$s_D = \frac{1}{n} \sum_{i=1}^{n} (d_i - \overline{D})^2 \tag{14}$$

$$d_i = x_{i,j} - x_{i,h}; \quad j, h \in \{1, 2, 3, \ldots, k\} \tag{15}$$

El último paso consiste en tomar el valor de t y posicionarlo en la distribución obtenida en el paso 2. Depediendo de la zona dentro de la distribución de probabilidad en la que se ubique t, se decide si aceptar o no H_0.

Ejemplo 2: Mejor actor/actriz

Un ejemplo bastante ilustrativo se puede encontrar en Triola (2018). Es un buen ejemplo, porque muestra que no necesariamente medimos a las mismas unidades en distintos puntos de tiempo (que es la manera en que usualmente se presentan las pruebas t para muestras

dependientes), sino que se trabaja con distintas unidades cuyas unidades en cada punto de tiempo son comparables.
Retomando el ejemplo, consiste en una lista con una muestra de 5 años de los ganadores de premios Oscar a mejor actor y mejor actriz. Para cada uno de ellos, se reporta su edad.

Referencias

Denis, Daniel. 2019. *SPSS data analysis for univariate, bivariate, and multivariate statistics*. Hoboken, NJ: Wiley.

Triola, Mario. 2018. *Elementary statistics*. United States: Pearson.