

Semana 4

i

Comparar k grupos *Valentín Vergara Hidd*

Con las herramientas que hemos visto hasta ahora, podemos entender la /lógica/ detrás de las pruebas de hipótesis, de amplio uso en estadística descriptiva. Seguiremos en ese contexto, estudiando algunas pruebas estadísticas que se utilizan en distintos casos. Para ello, en este documento extenderemos la comparación a k grupos. Se vuelve necesario además, presentar otra distribución de probabilidad: F .

1. El problema

Muchas veces no basta sólo con comparar dos grupos, como en el caso de la prueba t para muestras independientes. Existen situaciones en las que es pertinente comparar un número finito de grupos, pero mayor que dos. Retomando un ejemplo del video anterior, imaginen que buscamos comparar el ingreso por hogar para cada una de las quince regiones del país. En ese caso, si quisiésemos obtener información sobre comparaciones entre pares de regiones, tendríamos que encontrar todos los posibles pares que se pueden formar utilizando las quince regiones. Esto es una tarea bastante trivial, simplemente utilizamos $\binom{15}{2} = 105$. Se puede ver, que tratar de entender la información recogida en 105 pruebas de hipótesis puede ser algo bastante tedioso y poco práctico.

Otra alternativa sería formular el problema en términos más generales y que a la vez nos permitan responder a la pregunta: ¿es el ingreso de los hogares igual en las quince regiones del país? Obviamente, sabemos que esto no es cierto, al mirar los resultados presentados en el video anterior. Sin embargo, es un buen ejemplo para ilustrar que basta con una sola región con ingresos distintos al resto, para responder la pregunta. Antes de continuar desarrollando los cuatro pasos de la prueba de hipótesis, es pertinente revisar una herramienta que utilizaremos más adelante: la distribución F .

2. Distribución F

Consideren dos poblaciones con distribución normal y con varianzas iguales ($\sigma_1^2 = \sigma_2^2$). La razón de sus varianzas muestrales $F = \frac{s_1^2}{s_2^2}$ es a su vez una variable con distribución F , que recibe su nombre por [quien la investó/descubrió](#) y se define para una variable aleatoria X con parámetros d_1 y d_2 como ([Triola, 2018](#)):

$$f(x; d_1, d_2) = \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2}x\right)^{-\frac{d_1+d_2}{2}} \quad (1)$$

donde B es la [función Beta](#), relacionada con la función Γ , que a su vez tiene bastante parecido conceptual a la función factorial.

Siendo más explícito en la forma en que se logra esta distribución. Si tomamos dos poblaciones con igual varianza y con distribución normal; para luego tomar un número elevado de muestras $n \rightarrow \infty$, la razón entre las varianzas muestrales sigue una distribución F .

¿Cómo se ve esta distribución? En seguida la vamos a graficar, pero antes hay algunas cosas que podemos presuponer de ella. Puesto que se construye como la razón entre dos varianzas, sabemos que su dominio corresponde a números igual o mayor que cero. Utilizando la ecuación 1:

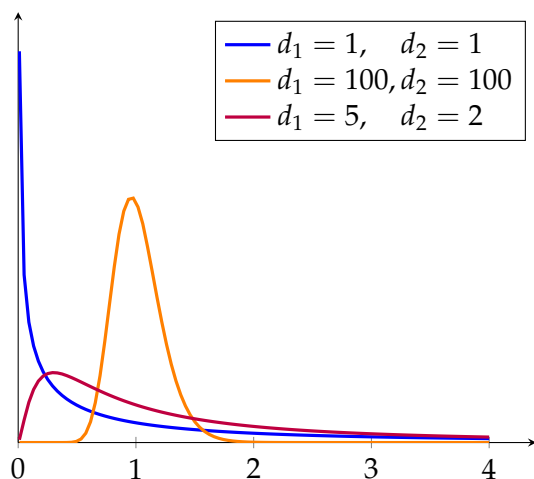


Figura 1: Algunos ejemplos de distribución F , utilizando distintos parámetros

Entonces, podemos ver cómo distintos valores de los parámetros de la distribución afectan la manera en que ésta se ve, pero la interpretación sigue siendo la misma. Además, ahora podemos volver a otro aspecto que había quedado pendiente en un documento anterior: uno de los pasos al comparar dos medias utilizando una distribución t busca establecer si las varianzas de ambos grupos son iguales. La forma en que se ejecuta esta comparación es simplemente calculando F con las varianzas de ambas muestras; y contrastando el resultado con la función de probabilidad dada por la ecuación 1 normalizada. Al igual que en casos anteriores, no es necesario hacer este largo cálculo, dado que hay valores que se repiten y que por tanto, son fáciles de encontrar en cualquier libro de estadística en forma de una [tabla de valores de \$F\$](#) .

Para usar la Tabla, las columnas representan los grados de libertad de la primera muestra (de donde se obtiene s_1^2) y las filas son los grados de libertad de la segunda muestra. En ambos casos, los grados de libertad se obtienen con $gl = n - 1$. Además, la primera muestra (cuya varianza va en el numerador), siempre es la que tiene el mayor valor de varianza.

3. Pasos de la prueba de hipótesis

Los pasos de la prueba de hipótesis no varían respecto a lo que ya hemos revisado. Sin embargo, la forma en que se fijan algunos valores o se formulan hipótesis presenta algunas diferencias, por lo que es bueno revisarlas en detalle.

A modo de ejemplo, utilizaremos un caso propuesto en [Triola \(2018\)](#):

Ejemplo 1: Exposición al plomo

En minería, se suele utilizar una instalación especializada para extraer los metales a partir de las rocas. En el proceso se producen varios contaminantes, siendo uno de ellos el plomo. Los datos que se presentarán a continuación provienen de un estudio que midió los niveles de plomo en la sangre de niños residentes de una localidad a 7 km de una de las instalaciones mencionadas. Se midió el CI de los participantes del estudio utilizando la escala de inteligencia de Wechsler y se separó a los niños entre aquellos con niveles bajos, medios y altos de plomo en la sangre.

- Bajo nivel de plomo: {85, 90, 107, 85, 100, 97, 101, 64, 111, 100, 76, 136, 100, 90, 135, 104, 149, 99, 107, 99, 113, 104, 101, 111, 118, 99, 122, 87, 118, 113, 128, 121, 111, 104, 51, 100, 113, 82, 146, 107, 83, 108, 93, 114, 113, 94, 106, 92, 79, 129, 114, 99, 110, 90, 85, 94, 127, 101, 99, 113, 80, 115, 85, 112}
- Nivel medio de plomo: {78, 97, 107, 80, 90, 83, 101, 121, 108, 100, 110, 111, 97, 51, 94, 80}
- Alto nivel de plomo: {93, 100, 97, 79, 97, 71, 111, 99, 85, 99, 97, 111, 104, 93, 90, 107}

Como se puede ver, el tamaño de los tres grupos no es igual. La pregunta de investigación que estos datos sugieren tiene que ver con la homogeneidad de los tres grupos. Es decir, ¿existe evidencia que respalde que los tres grupos provienen de la misma población? Si esto fuese cierto, las medias de los tres grupos deberían ser estadísticamente equivalentes. Aunque suene un poco extraño, para verificar si las medias de los grupos son iguales, utilizaremos una prueba de /análisis de varianza/, que nos va a permitir utilizar la distribución F para estudiar este indicador en cada grupo.

Para facilitar el proceso, a continuación se presenta el resumen de los datos, por grupo.

	Nivel Bajo	Nivel Medio	Nivel Alto
n	78	22	21
\bar{x}	102.7	94.1	94.2
s_x	16.8	15.5	11.4
Distribución	Aprox. normal	Aprox. normal	Aprox. normal

Tabla 1: Resultados descriptivos por grupo

3.1. Hipótesis nula y alternativa

Nos interesa saber si existen diferencias en el CI de niños que presentan niveles **Altos**, **Medios** y **Bajos** de plomo. Por tanto, estableceremos como hipótesis nula un escenario en el que los tres grupos tienen, en promedio, el mismo CI.

$$H_0 : \mu_A = \mu_M = \mu_B$$

$$H_1 : \mu_i \neq \mu_j, \quad i, j \in \{A, M, B\}, \quad i \neq j$$

Para la hipótesis alternativa, basta con que uno de los grupos sea diferente.

3.2. Establecer α y definir los valores críticos de F .

Como usualmente lo hemos presentado, pensemos en un valor $\alpha = 0,05$

Referencias

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