ANOVA de 1 Factor: Prueba ANOVA

Clase 3

Nicolás Mejía M.
n.mejia10@uniandes.edu.co

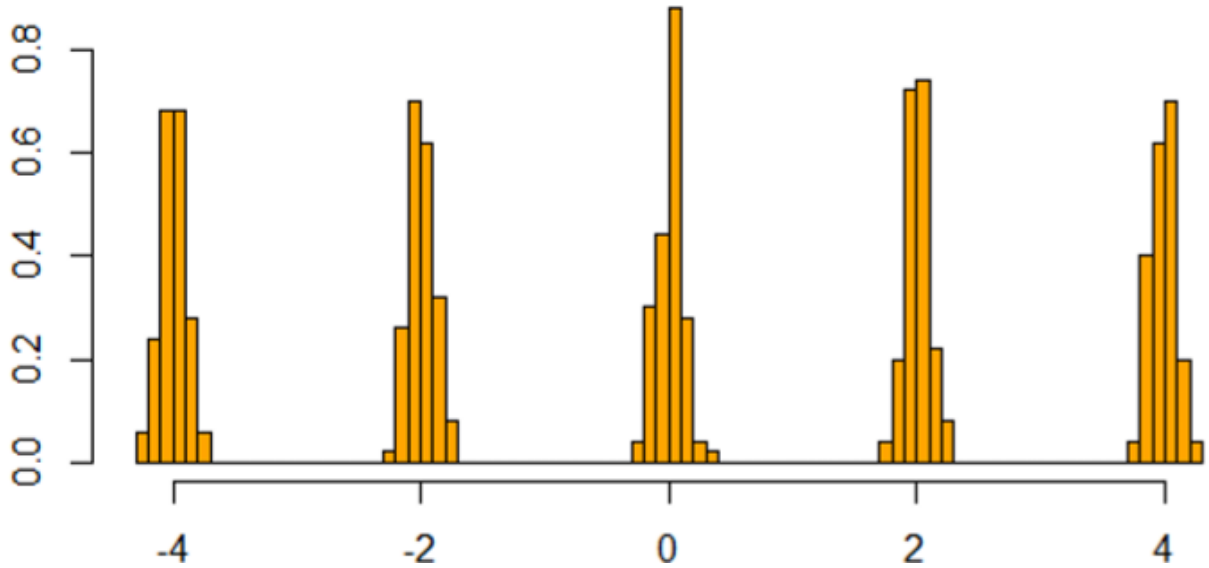
Probabilidad y Estadística II
Departamento de Ingeniería Industrial
Universidad de Los Andes, Bogotá, Colombia

2020-19

Outline

Fuentes de Variación

Cuando observamos realizaciones de un fenómeno aleatorio, usualmente atribuimos la variación en los valores que se dan a únicamente ese componte aleatorio. Pero las vairaciones pueden ser producto de otros factores.



La variación que observamos en los datos realmente se puede descomponer en varias fuentes:

$$\text{Variación Total} = \text{Var. por efecto de grupos} + \text{Var. por efecto aleatorio}$$

Terminología

Los conceptos clave:

- Factor: Es la variable que se supone afecta a nuestra Y (variable respuesta). En nuestro caso sería al estructura de grupos.
- Niveles: Corresponden a los diferentes valores que puede tomar el factor. En nuestro caso sería las distintas categorías de la agrupación, denotaremos a la cantidad de categorías como a.
- Tratamientos: Cuando hay varios factores en el estudio, los tratamientos corresponden a todas las posibles combinaciones de niveles entre los factores.
- Réplicas: Corresponde a la cantidad de veces que se repite el experimento bajo un tratamiento dado. Este número lo denotaremos como n_i y de manera informal es el número de datos por celda.

Pregunta Principal

En un experimento estadístico existen dos preguntas fundamentales:

1. ¿Influye el factor sobre la variable de respuesta? Generalmente es la pregunta científica de interés y puede ser entendida como la causalidad del cambio en la respuesta dado el cambio en el factor.
2. ¿Cuál es el tratamiento óptimo? Si el factor si influye, una pregunta práctica es cuál debe ser selecionado dado un interés particular.

Notación

Cuando se tiene un solo factor Y_{ij} representa la variable de interés para la j-ésima observación del i-ésimo nivel. Se asume que $Y_{ij} \sim Normal(\mu_i, \sigma^2)$.

Así mismo, se utilizará la siguiente notación para los promedios:

$$N = \sum_{i=1}^{a} n_i$$

$$\bar{Y}_{..} = \frac{1}{N} \sum_{i=1}^{a} \sum_{j=1}^{n_i} Y_{ij}$$

$$\bar{Y}_{i.} = \hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}, \quad i \in \{1, \ldots, a\}$$

Notación

Al organizar la información en una tabla:

	Factor A				
Obs (j)	Nivel 1	...	Nivel i	...	Nivel a
1	Y_{11}	...	Y_{i1}	...	Y_{a1}
2	Y_{12}	...	Y_{i2}	...	Y_{a2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	Y_{1j}	...	Y_{ij}	...	Y_{aj}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	Y_{1n_1}	...	Y_{in_i}	...	Y_{an_a}
Promedios columna	$\bar{Y}_{1.}$	...	$\bar{Y}_{i.}$	...	$\bar{Y}_{a.}$
Desv. Est.	S_1	...	S_i	...	S_a

Prueba ANOVA

La pregunta de interés se expresa con la prueba:

$$H_0: \quad \text{El Factor NO influye sobre } Y \Leftrightarrow \mu_1 = \mu_2 = \cdots = \mu_a$$
$$H_1: \quad \text{El Factor SI influye sobre } Y \Leftrightarrow \text{Algún par } \mu_i \neq \mu_j$$

En la clase anterior vimos que en el caso en que $a = 2$, la prueba se puede resolver como una diferencia de medias. Vamos a ver ahora el caso general cuando se tienen más de dos niveles ($a \geq 3$). Quermos formalizar numericamente la descomposicion de la variación:

$$\text{Variación Total} = \text{Var. por efecto de grupos} + \text{Var. por efecto aleatorio}$$

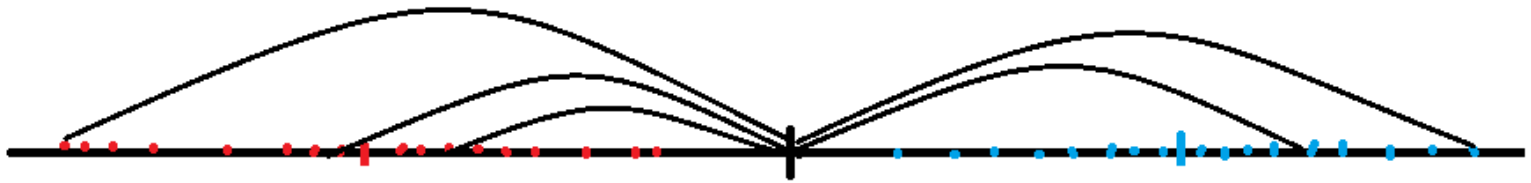
y después comprobar estadísticamente si la varianza por grupos es la mayor en magnitud.

Outline

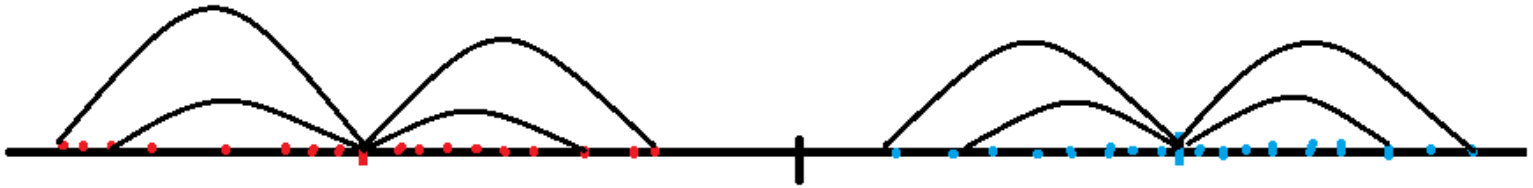
Prueba ANOVA

Para determinar si las medias de *a* poblaciones normales e independientes son iguales o no, se mira la variabilidad de los datos!!!

La idea fundamental es tomar el estimador de la varianza ($\hat{\sigma}^2 = S_p^2$) y compararlo con la varianza total de los datos: S^2_{Total}.



Variabilidad Total



Variabilidad Ponderada (estimador de la varianza)

Prueba ANOVA

Cuando H_0 es cierto (el factor no influye), entonces:

$$E(S_p^2) = E(S_{Total}^2) = \sigma^2$$

Por otro lado, cuando H_1 es cierto (el factor si influye),

$$E(S_p^2) = \sigma^2 < E(S_{Total}^2)$$

La idea de la prueba ANOVA es entonces comparar estos dos estimadores de la varianza. Si cuando se realizan los datos se tiene que

$$S_p^2 << S_{Total}^2$$

Entonces se rechaza H_0 en favor de H_1.

Prueba ANOVA

El estadístico de prueba para contrastar las hipótesis de interés debe comparar los valores de S_p^2 y de S_{Total}^2.

Dificultad Técnica: Los estimadores S_p^2 y S_{Total}^2 no son independientes entre sí, lo cual dificulta encontrar un estadístico que posea una distribución conocida.
Es mejor descomponer la información en las sumas y las medias cuadráticas.

Outline

La variación total

La varianza muestral sobre todos los datos es una medida de la variación total. De acuerdo a la notación se calcula como:

$$SST = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2$$

El término usual $\frac{1}{(N-1)}$ que acompaña a la suma solo la estandariza para hacer un promedio. Enfoquémonos en entender SOLO la suma de cuadrados.

Allí se contempla la resta $(Y_{ij} - \bar{Y}_{..})$, es decir, la diferencia entre un dato particular y el promedio total . Aquí se mide TODA la variación de un dato con respecto a su comportamiento común dado por el promedio. Por eso denominaremos esta sumatoria como la suma de cuadrados TOTAL.

La variación por efecto del factor

Si el factor NO es significativo, entonces los promedios por niveles serían iguales entre ellos, y en particular al promedio total (i.e. $\bar{Y}_{i.} \approx \bar{Y}_{..}$). Luego la resta $(\bar{Y}_{i.} - \bar{Y}_{..})$, que es la diferencia entre el promedio de un nivel y el promedio general, se podría ver como una medida de discrepancia entre la homogeneidad de las medias.

Unificando con el marco anterior, se define entonces la suma de cuadrados del FACTOR.

$$SSA = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (\bar{Y}_{i.} - \bar{Y}_{..})^2 = \sum_{i=1}^{a} n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2$$

La variación por efecto aleatorio

Finalmente, la variación dada por el efecto aleatorio explica la desviación de un dato con respecto al promedio de su nivel. Luego es natural utilizar la diferencia $(Y_{ij} - \bar{Y}_{i.})$ como dicha medida de discrepancia.

Nuevamente, unificando con el marco anterior, se define la suma de cuadrados del ERROR como:

$$SSE = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$$

Así como el SSA se interpreta como la variación explicada por el factor, el SSE es la variación que NO se puede explicar!

Ecuación Fundamental

De esta forma, como se describió intuitivamente, se puede demostrar matemáticamente la siguiente relación:

Ecuación Fundamental de la Suma de Cuadrados

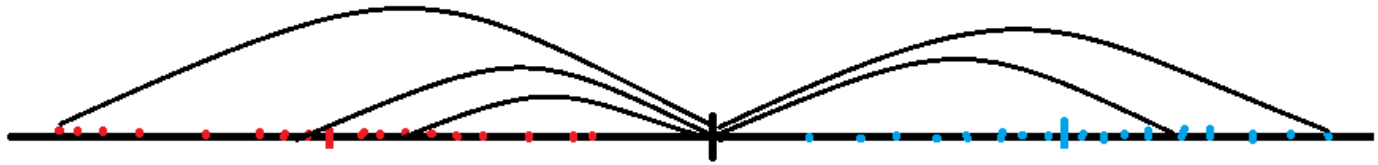
$$SST = SSA + SSE$$

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(\bar{Y}_{i.} - \bar{Y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2$$

La propiedad interesante (y útil) de esta descomposición es que SSA y SSE son variables aleatorias indepenientes !!!

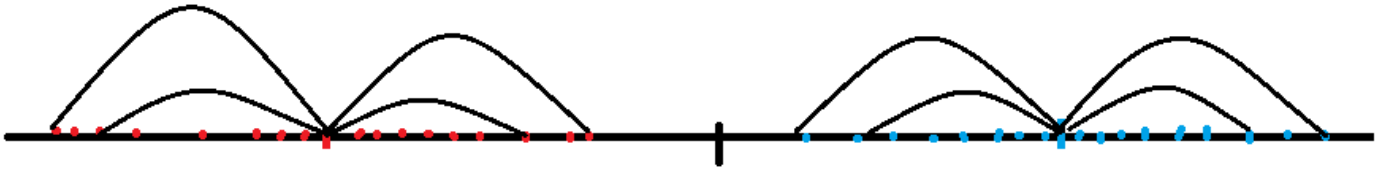
Visualmente

Graficamente se ve así



Variabilidad Total

SST =



Variabilidad Ponderada (estimador de la varianza)

SSE +

Variabilidad entre los promedios (diferencia de medias)

SSA

Medias Cuadráticas

Las sumas de cuadrados están sumando sobre todos los datos, luego si queremos ver en promedio como es esa variabilidad, debemos dividir por la cantidad de términos efectivamente utilizados:

Los grados de libertad

Son el número de variables que pueden variar libremente en un estadístico.

En este caso, piensen que los datos son como un recurso que se va gastando a medida que hacemos estimaciones, luego al hacer una cuenta (como las sumas de cuadrados), realmente no se están utilizando todos los datos, sino los que van quedando.

Por eso decía que términos efectivamente utilizados=grados de libertad.

Grados de libertad totales

Hay una forma informal para hallar los grados de libertad: Simplemente se suman la cantidad de términos positivos y se restan los negativos. Simbólicamente, los términos en negativo hacen alusión a los " recursos que se van gastando".

Por ejemplo considere el $SST = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2$

- ¿Cuantos Y_{ij} hay? Estos son los datos, y en nuestra notación, N es el total de datos. Como los Y_{ij} van sumando en la expresión, pues se suma N en los grados de libertad.
- ¿Cuantos $\bar{Y}_{..}$ hay? Este es el promedio total de los datos, y solo hay uno! Como $\bar{Y}_{..}$ aparece restando en la expresión, restamos uno en los grados de libertad.

Agregando resultados, se tiene que los grados de libertad totales son $N - 1$.

Grados de libertad del factor y del error

Usando la misma lógica, los grados de libertad para el SSA y el SSE son:

$$SSA = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (\bar{Y}_{i.} - \bar{Y}_{..})^2$$

¿Cuantos $\bar{Y}_{i.}$ hay? Hay uno para cada nivel, es decir hay a. ¿Cuantos $\bar{Y}_{..}$ hay? Solo hay uno, y va restando. Agregando, se tiene entonces que los grados de libertad del SSA son $a - 1$.

$$SSE = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$$

¿Cuantos Y_{ij} hay? N. ¿Cuantos $\bar{Y}_{i.}$ hay? a, y van restando. Agregando, se tiene entonces que los grados de libertad del SSE son $N - a$.

Medias Cuadráticas

Observe que la ecuación fundamental también aplica para los grados de libertad:

$$gl_T = gl_A + gl_E$$

$$N - 1 = (a - 1) + (N - a)$$

Con esto se definen las medias cuadráticas:

$$MST = \frac{SST}{gl_T} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{..})^2}{N - 1}$$

$$MSA = \frac{SSA}{gl_A} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(\bar{Y}_{i.} - \bar{Y}_{..})^2}{a - 1}$$

$$MSE = \frac{SSE}{gl_E} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2}{N - a}$$

Outline

Idea

Recodemos que la pregunta de interes es:

$$H_0: \quad \text{El Factor NO influye sobre } Y \Leftrightarrow \mu_1 = \mu_2 = \cdots = \mu_a$$
$$H_1: \quad \text{El Factor SI influye sobre } Y \Leftrightarrow \text{Algún par } \mu_i \neq \mu_j$$

Sabemos que $SST = SSA + SSE$, donde el SST es un valor que no cambia.

- Si el factor SI es significativo, el SSA debería llevarse una gran parte de la variabilidad y el SSE se vería reducido.
- Si el factor NO es significativo, el SSA debería ser pequeño comparado con el SSE, pues este último sería el que se lleve la variabilidad.

Por lo tanto, solo basta comparar el tamaño relativo del SSA con respecto al SSE, y con eso sabremos si el factor afecta o no.

Prueba F

Bajo la validez de la hipótesis nula, donde las medias de los niveles son iguales, se cumple lo siguiente:

Aplicación del teorema de Cochran

Las sumas de cuadrados anteriores, divididas por la varianza σ^2, se distribuyen χ^2 con sus respectivos grados de libertad:

$$\frac{SST}{\sigma^2} = \frac{SSA}{\sigma^2} + \frac{SSE}{\sigma^2}$$

$$\chi^2_{N-1} = \chi^2_{a-1} + \chi^2_{N-a}$$

Ahora bien, de Proba 1 sabemos que $\frac{\frac{\chi^2_{gl1}}{gl_1}}{\frac{\chi^2_{gl2}}{gl_2}} \sim F_{gl_1,gl_2}$ luego se tiene que:

Prueba F

Bajo la hipótesis nula, el estadístico dado por:

El estadístico F

$$F = \frac{\frac{\frac{SSA}{\sigma^2}}{a-1}}{\frac{\frac{SSE}{\sigma^2}}{N-a}} = \frac{\frac{SSA}{a-1}}{\frac{SSE}{N-a}} = \frac{MSA}{MSE} \sim F_{a-1,N-a}$$

Valores grandes del estadístico F están a favor de que el factor SI es significativo, mientras que valores pequeños son evidencia de que el factor es NO significativo. Luego estadísticamente tenemos la siguiente región de rechazo:

$$RHO \Leftrightarrow F \geq F_{1-\alpha,a-1,N-a}$$

Donde α es el nivel de significiancia.

La tabla ANOVA

Toda esta información se puede organizar en forma de tabla de la siguiente manera:

Fuente	**SS**	**gl**	**MS**	**F**
Factor A	$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(\bar{Y}_{i.}-\bar{Y}_{..})^2$	$a-1$	$SSA/(a-1)$	MSA/MSE
Error	$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i.})^2$	$N-a$	$SSE/(N-a)$	-
Total	$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{..})^2$	$N-1$	$SST/(N-1)$	-

Donde $SST = SSA + SSE$ y $gl_T = gl_A + gl_E$

Ejemplo

A continuación se muestra una tabla con información sobre las calificaciones obtenidas en una prueba de estadística para diferentes tipos de examen. ¿El tipo de examen afecta la nota obtenida?

	Diseño 1	**Diseño 2**	**Diseño 3**	**Diseño 4**
Tamaño de la Muestra	10	6	12	8
Promedio	3.44	3.81	3.95	4.10
Varianza	0.18	0.12	0.06	0.09

Respuesta: La F calculada tiene un valor de 6.912 con un pvalor de 0.001