Analysis of Variance (ANOVA)

Aniel Nieves-González

Introducción

- Recordemos que la prueba t para dos muestras (two sample t-test) queremos examinar si $\mu_1 = \mu_2$.

Introducción

- Recordemos que la prueba t para dos muestras (two sample t-test) queremos examinar si $\mu_1 = \mu_2$.
- La prueba t para dos muestras (two sample t-test) puede generalizarse para el caso en que se quieren comparar más de dos μ's.

Introducción

- Recordemos que la prueba t para dos muestras (two sample t-test) queremos examinar si $\mu_1 = \mu_2$.
- La prueba t para dos muestras (two sample t-test) puede generalizarse para el caso en que se quieren comparar más de dos μ's.
- En ANOVA se analiza la variabilidad con el objetivo de examinar la igualdad de las medias (μ's) (note la aparente paradoja).

Introducción

- Recordemos que la prueba t para dos muestras (two sample t-test) queremos examinar si $\mu_1 = \mu_2$.
- La prueba t para dos muestras (two sample t-test) puede generalizarse para el caso en que se quieren comparar más de dos μ's.
- En ANOVA se analiza la variabilidad con el objetivo de examinar la igualdad de las medias (μ's) (note la aparente paradoja).
- EN ANOVA, la hipótesis nula es: $H_0 : \mu_1 = \cdots = \mu_k$.

Introducción

- Recordemos que la prueba t para dos muestras (two sample t-test) queremos examinar si $\mu_1 = \mu_2$.
- La prueba t para dos muestras (two sample t-test) puede generalizarse para el caso en que se quieren comparar más de dos μ's.
- En ANOVA se analiza la variabilidad con el objetivo de examinar la igualdad de las medias (μ's) (note la aparente paradoja).
- EN ANOVA, la hipótesis nula es: $H_0 : \mu_1 = \cdots = \mu_k$.
- En ANOVA consideramos dos variables aleatorias:

Introducción

- Recordemos que la prueba t para dos muestras (two sample t-test) queremos examinar si $\mu_1 = \mu_2$.
- La prueba t para dos muestras (two sample t-test) puede generalizarse para el caso en que se quieren comparar más de dos μ's.
- En ANOVA se analiza la variabilidad con el objetivo de examinar la igualdad de las medias (μ's) (note la aparente paradoja).
- EN ANOVA, la hipótesis nula es: $H_0 : \mu_1 = \cdots = \mu_k$.
- En ANOVA consideramos dos variables aleatorias:
  - Variable de respuesta (dependiente)

Introducción

- Recordemos que la prueba t para dos muestras (two sample t-test) queremos examinar si $\mu_1 = \mu_2$.
- La prueba t para dos muestras (two sample t-test) puede generalizarse para el caso en que se quieren comparar más de dos μ's.
- En ANOVA se analiza la variabilidad con el objetivo de examinar la igualdad de las medias (μ's) (note la aparente paradoja).
- EN ANOVA, la hipótesis nula es: $H_0 : \mu_1 = \cdots = \mu_k$.
- En ANOVA consideramos dos variables aleatorias:
  - Variable de respuesta (dependiente)
  - Variables explicaticas o factores. Esta es de tipo categórico.

Introducción

- Recordemos que la prueba t para dos muestras (two sample t-test) queremos examinar si $\mu_1 = \mu_2$.
- La prueba t para dos muestras (two sample t-test) puede generalizarse para el caso en que se quieren comparar más de dos μ's.
- En ANOVA se analiza la variabilidad con el objetivo de examinar la igualdad de las medias (μ's) (note la aparente paradoja).
- EN ANOVA, la hipótesis nula es: $H_0 : \mu_1 = \cdots = \mu_k$.
- En ANOVA consideramos dos variables aleatorias:
  - Variable de respuesta (dependiente)
  - Variables explicaticas o factores. Esta es de tipo categórico.
- Cuando se tiene un factor, usamos el ANOVA de una vía (one-way ANOVA), cuando se consideran dos factores tenemos el "two-way" ANOVA, etc.

Variation among groups vs. variation within groups.

Two-sample t-test

Considere el two-sample t-test para comparar μ_1 y μ_2 de dos poblaciones que obedecen $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$ respectivamente y para las cuales $\sigma_1 = \sigma = \sigma_2$. Suponga que obtiene dos SRS's independientes. Considere el test

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2.$$

Sabemos que

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{donde} \quad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Variation among groups vs. variation within groups.

Si $n_1 = n = n_2$ entonces

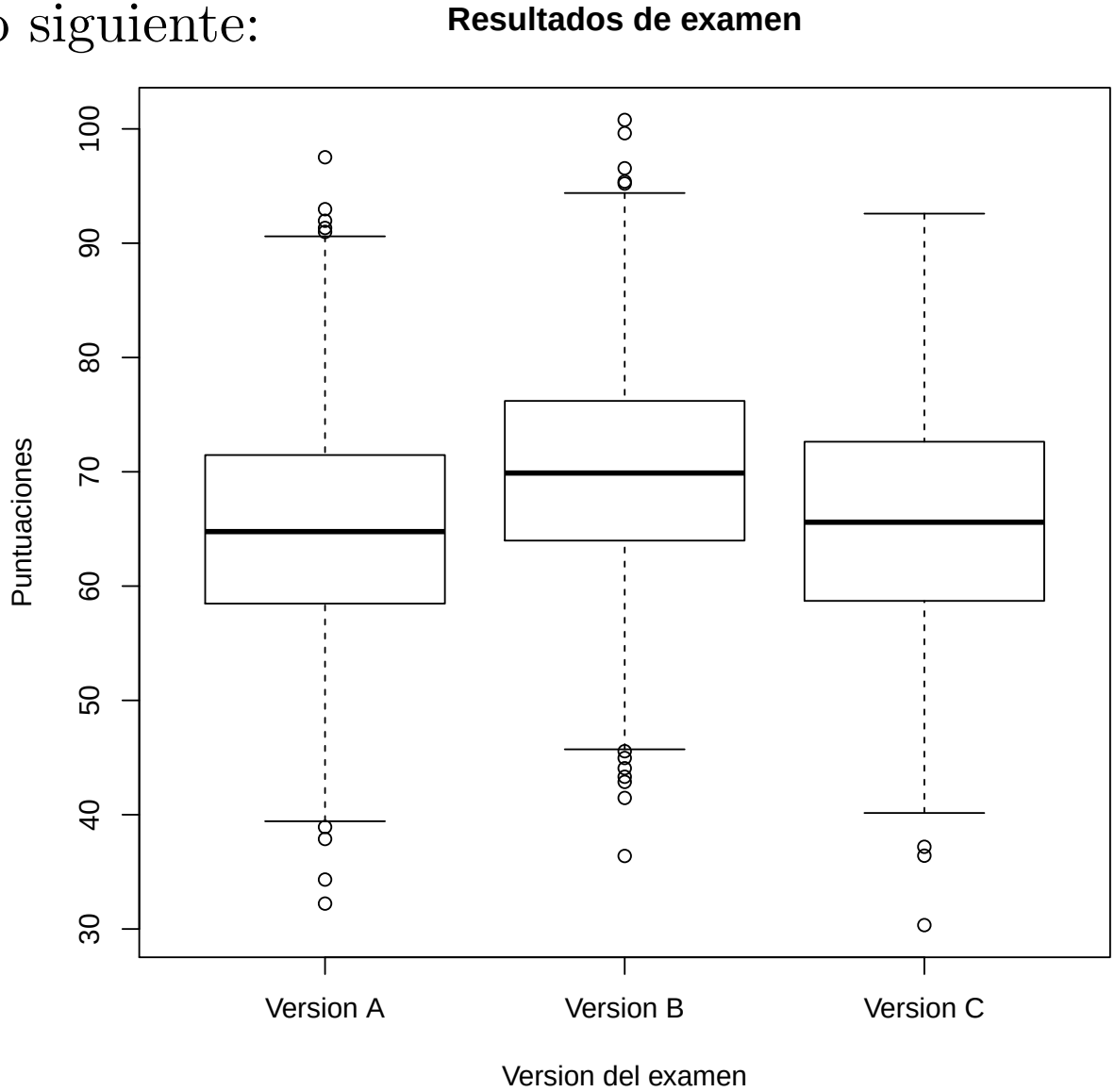
$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{2}{n}}} = \frac{\sqrt{\frac{n}{2}}(\bar{X}_1 - \bar{X}_2)}{s_p}$$

$$\Rightarrow t^2 = \frac{\frac{n}{2}(\bar{X}_1 - \bar{X}_2)^2}{s_p^2}$$

Note que la estadística t (t^2) puede interpretarse como variación entre grupos (numerador) versus variación dentro del grupo (denominador). En inglés: "variation among groups vs. variation within groups".

- El estadístico F que usaremos en ANOVA guarda relación con como se definió el estadístico F para el test de igualdad de σ's.

Considere lo siguiente:



¿Serán los μ's iguales? El comparar la variabilidad entre grupos y dentro del grupo nos contesta la pregunta.

Variation among groups vs. variation within groups.

Suponga que obtiene k SRS's independientes de tamaño n_i, $i = 1, \ldots, k$. Sean $x_{i1}, \ldots, x_{i2}, \ldots, x_{in_i}$ las observaciones del i-ésimo grupo. Algunas definiciones importantes:

- Media de la muestra de la i-ésima muestra:

$$\bar{X}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}, \quad i = 1, \ldots, k$$

- Sea $n = \sum_{i=1}^{k} n_i$.
- Grand (sample) Mean:

$$\bar{\bar{X}} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}}{n}$$

- Sum of Squares Total. Este es el "total variation":

$$\text{SST} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{\bar{X}})^2$$

Variation among groups vs. variation within groups.

- Sum of Squares among groups (suma de cuadrados entre grupos y se denota SSA ó SSG). Este es el "among group variation":

$$\text{SSG} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{\bar{X}})^2$$

- Sum of squares within groups or sum of squares error (suma de cuadrados dentro de los grupos, y se denota SSW ó SSE):

$$\text{SSE} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{X}_i)^2 = \sum_{i=1}^{k}(n_i - 1)s_i^2$$

- Note que:

$$\text{SST} = \text{SSG} + \text{SSE}.$$

Esto es:

total variation = variation among groups+variation within groups.

Variation among groups vs. variation within groups.

Los grados de libertad (degrees of freedom, df) son:

- Asociados a SSE tenemos que $\text{df}_{\text{SSE}} = n - k$.
- Asociados a SST tenemos que $\text{df}_{\text{SST}} = n - 1$.
- Asociados a SSG tenemos que $\text{df}_{\text{SSG}} = k - 1$.
- Note que los df de SST son

$$\text{df}_{\text{SST}} = n - 1 = \text{df}_{\text{SSG}} + \text{df}_{\text{SSE}} = k - 1 + n - k.$$

- La razón de las sumas de cuadrados y los correspondientes grados de libertad nos da los "mean square terms", esto es:

$$\text{MSG} = \frac{\text{SSG}}{k-1} = \text{mean square among groups}$$

$$\text{MSE} = \frac{\text{SSE}}{n-k} = \text{mean square error}$$

$$\text{MST} = \frac{\text{SST}}{n-1} = \text{mean square total}$$

- Note que $s_p^2 = \text{MSE}$ (más de esto adelante).

La prueba de hipótesis

- En ANOVA se usará una forma del estadistico F para comparar la variación entre grupos con la variación en un grupo.
- Los calculos se organizaran en una tabla (tabla ANOVA) y luego se interpretaran los resultados.

La prueba de hipótesis

Suponga que tiene k poblaciones de las cuales se hacen $n_1, n_2, \ldots, n_i, \ldots, n_k$ observaciones independientes. O sea tenemos k SRS's independientes. Suponga que las k poblaciones se distribuyen normalmente con medias $\mu_1, \mu_2, \ldots, \mu_k$ respectivamente y varianza común σ^2. El test de hipótesis es:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_a : \mu_l \neq \mu_m \quad \text{para algún} \quad \begin{array}{c} 1 \leq l \leq k, \\ 1 \leq m \leq k, \quad l \neq m \end{array}$$

H_a puede escribirse como: no todas los μ's son iguales. El estadístico F se calcula como sigue:

$$F = \frac{MSG}{MSE}$$

Cuando H_0 es cierto el estadístico se distribuye con la distribución $F(k-1, n-k)$.

La prueba de hipótesis

Observe lo siguiente:

1. Cuando H_0 es falsa (H_a cierta) el estadistico F tiende a ser un número grande. Rechazamos H_0 cuando F es suficientemente grande.
2. Para un nivel de significancia α calcularemos el "upper tail critical value" F_u y rechazamos H_0 si $F > F_u$.
3. El P-value del F-test es la probabilidad de que la variable aleatoria con ditribución F es mayor o igual que el valor calculado (observado) de F. Esto es:

$$P_{\text{value}} = Pr(F > F_{\text{obs}}).$$

One-way ANOVA model

Suponga que tiene k poblaciones de las cuales se hacen $n_1, n_2, \ldots, n_i, \ldots, n_k$ observaciones independientes. O sea tenemos k SRS's independientes. Sean $x_{i1}, \ldots, x_{i2}, \ldots, x_{in_i}$ las observaciones del i-ésimo grupo. Suponga que las k poblaciones se distribuyen normalmente con medias $\mu_1, \mu_2, \ldots, \mu_k$ respectivamente y varianza común σ^2. El modelo ANOVA es

$$x_{ij} = \mu_i + \epsilon_{ij} \quad \begin{matrix} i = 1, \ldots k \\ j = 1 \ldots, n_i \end{matrix}$$

donde $\epsilon_{ij} \sim N(0, \sigma)$ y son independientes. Observe que el modelos ANOVA sigue el esquema

$$\text{datos} = \text{modelo} + \text{residual}$$

One-way ANOVA model

- El modelo ANOVA tiene $k+1$ parámetros desconocidos: σ y μ_i, $i = 1, \ldots, k$. Para estimar μ_i tenemos

$$\bar{X}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}, \quad i = 1, \ldots, k$$

  El residual es $e_{ij} = x_{ij} - \bar{X}_i$ refleja la variación relativo a la media arithmética del k-ésimo grupo.
- En ANOVA se supone que σ es igual para todas las poblaciones que estamos comparando. De no ser así se procede a transformar los datos (logaritmo u otra transformación) de manera que las desviaciones estandar de las muestras sean parecidas (s_i's).
- *Rule of thumb*: Si $\max_{1 \leq i \leq k}\{s_i\} < 2 \min_{1 \leq i \leq k}\{s_i\}$, entonces puede suponerse que σ es el mismo para todas las poblaciones bajo consideración y los resultados serán aproximadamente correctos. (ANOVA no es extremadamente sensitivo a una violación de esta condición).

One-way ANOVA model, contd.

- Si las poblaciones comparten el mismo σ entonces podemos usar el estimador ponderado ("pooled estimator") de σ. Suponga que calcula $s_1^2, \ldots, s_k^2$ (varianzas de las muestras) partiendo de k SRS's independientes y de tamaño $n_1, \ldots, n_k$ respectivamente. El "pooled sample variance" es:

$$s_p^2 = \frac{\sum_{j=1}^{k}(n_j - 1)s_j^2}{\sum_{j=1}^{k}(n_j - 1)} = \frac{(n_1 - 1)s_1^2 + \cdots + (n_k - 1)s_k^2}{(n_1 - 1) + \cdots + (n_k - 1)}$$

Note que:

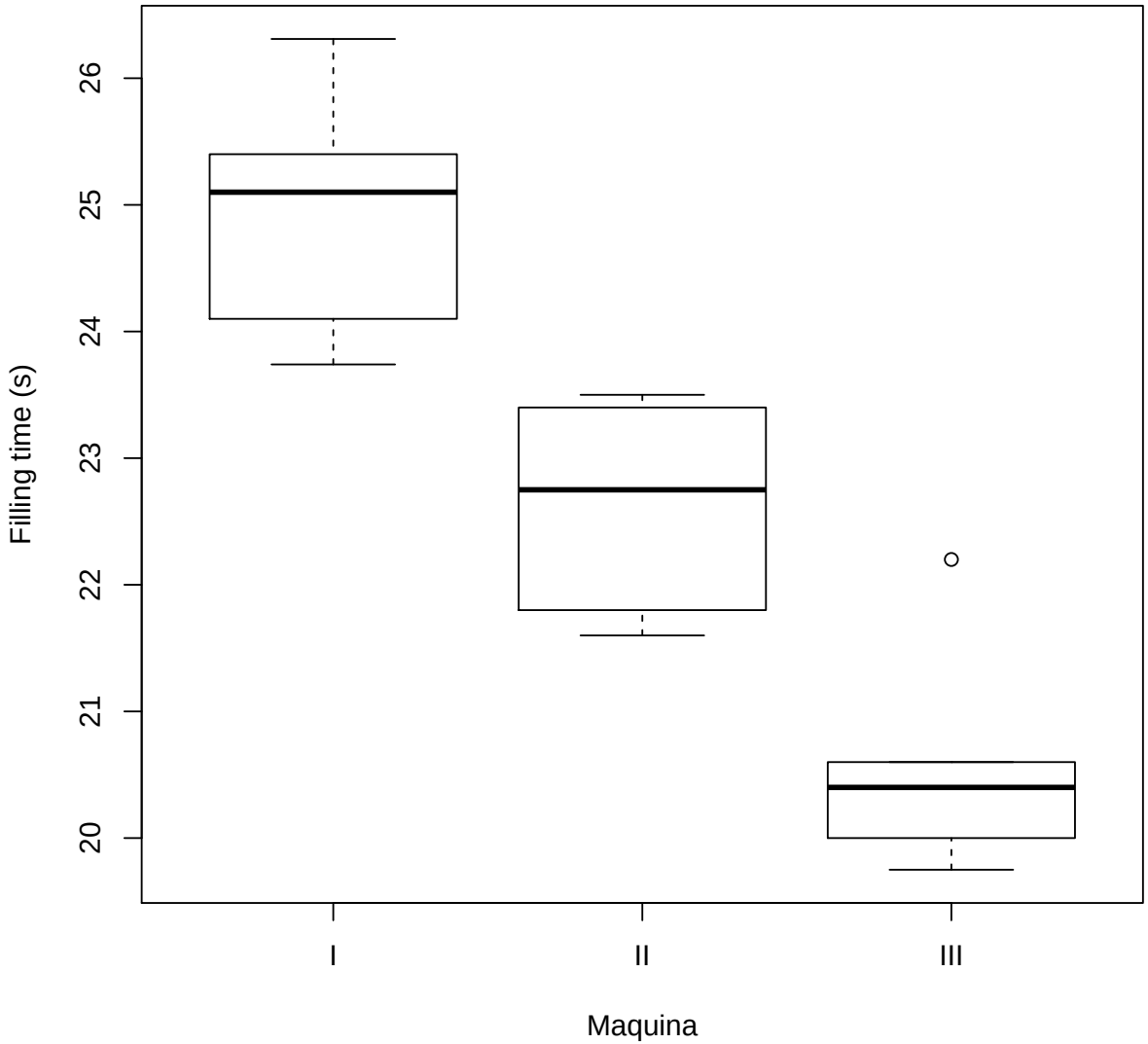
$$s_p^2 = \frac{\text{SSE}}{n - k} = \text{MSE}$$

Ejemplo (Máquina de cereal)

Suponga que el jefe de producción (manager) le interesa cambiar una máquina para llenar cajas de cereal. Tres fabricantas de máquinas le permiten probarlas antes de comprarla. Las mismas, en terminos de especificaciones y contratos de servicio son las mismas. Al manager le interesa saber si el tiempo de llenado medio de las 3 máquinas son significaticamente distintos. Se procede a entrenar operadores para las máquinas y luego del entrenamiento se obtiene lo siguiente.

Máquina		
I	II	III
25.40	23.40	20.00
26.31	21.80	22.20
24.10	23.50	19.75
23.74	22.75	20.60
25.10	21.60	20.40
$\bar{X}_1 = 24.93$	$\bar{X}_2 = 22.61$	$\bar{X}_3 = 20.59$

Ejemplo (Máquina de cereal)



Ejemplo (Máquina de cereal)

El test de hipótesis es:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_a : \mu_l \neq \mu_m \quad \text{para algún} \quad \begin{array}{c} 1 \leq l \leq 3, \\ 1 \leq m \leq 3, \quad l \neq m \end{array}$$

Ahora calculamos:

$$\bar{\bar{X}} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}}{n} = 22.71$$

$$\text{SSG} = \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{\bar{X}})^2 = 47.164$$

$$\text{SSE} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{X}_i)^2 = 11.0532$$

$$\text{SST} = \text{SSE} + \text{SSG} = 58.2172$$

Ejemplo (Máquina de cereal)

Luego tenemos lo siguiente:

$$\text{MSG} = \frac{SSG}{k-1} = \frac{47.164}{3-1} = 23.582$$
$$\text{MSE} = \frac{SSE}{n-k} = \frac{47.164}{15-3} = 0.9211$$
$$F_{\text{obs}} = \frac{MSG}{MSE} = \frac{23.582}{0.9211} = 25.60$$

Y ahora calculamos el P-value, usando la tabla o algún software como R (en R es el comando: 1-pf(F_{obs},$df1 = k-1$,$df2 = n-k$)):

$$P_{\text{value}} = Pr(F > F_{\text{obs}}) = 0.00004684037$$

Para un nivel de significancia $\alpha = 0.05$ buscamos en la tabla que $F_u(2, 12) = 3.89$. Luego, $F_{\text{obs}} > F_u(2, 12)$. Por lo tanto H_0 se rechaza.

Ejemplo (Máquina de cereal)

Generalmente se utilizará algún software para hacer pruebas ANOVA. Dicho software presentará como output una tabla como la siguiente:

Source	df	SS	Mean square
Among groups	$3-1=2$	47.164	23.5820
Within groups (error)	$15-3=12$	11.0532	0.9211
Total	$15-1=14$	58.2172	
$F = 26.60$			
$P_{\text{value}} = 4.684037 \times 10^{-5}$			

- Note que sabiendo lo que significa cada elemento en la tabla y como se relacionan unos con otros, usted puede llenar una tabla ANOVA que esté parcialmente llena.

Observe lo siguiente:

- Aunque H_0 sea cierto uno no espera que $\bar{X}_1 = \cdots = \bar{X}_k$ debido a la variación en las muestras. Si H_0 es cierto uno espera que $\bar{X}_1, \ldots, \bar{X}_k$ estén "cerca". La varianza de las muestras provee información sobre que tan cerca están.
- Si H_0 es cierta esperamos que la varianza entre grupos sea "pequeña".
- Si H_0 es cierta entonces SSG medirá la variabilidad de la población tan bien como SSE.
- Rechazar H_0 nos lleva a concluir que los factores o el tratamiento que distingue a cada grupo afecta las medias de las poblaciones (o sea las hace distintas).
- Rechazar H_0 no nos dice cual(es) μ's se diferencia(n) unos de otros.