

Estadística

Diseño completamente al azar

Daniel Martínez Bello

Universidad de Santander
Maestría en Biotecnología

Diciembre de 2024



Diseño completamente al azar, con un factor

Idea principal

- ▶ Queremos comparar entre tratamientos que consisten en varios niveles de un factor, y la respuesta es continua.
- ▶ Estamos interesados en conocer si al comparar la media entre pares de tratamientos hay al menos una diferencia estadísticamente significativa a un nivel de significancia fijado por el investigador.
- ▶ Utilizamos el **análisis de varianza**, que corresponde a una descomposición de todas las fuentes de variación.



Diseño completamente al azar, con un factor

Ejemplo

Four catalysts that may affect the concentration of one component in a three-component liquid mixture are being investigated. The following concentrations are obtained from a completely randomized experiment:



Diseño completamente al azar, un factor

Factor			
Nivel			
1	2	3	4
y_{11}	y_{12}	y_{13}	y_{14}
y_{21}	y_{22}	y_{23}	y_{24}
y_{31}	y_{32}	y_{33}	y_{34}
y_{41}	y_{42}		y_{44}
y_{51}			

Cada letra representa un tratamiento y a su vez representa un nivel del factor bajo estudio, entonces hay cuatro tratamientos, cada tratamiento tiene entre cinco y tres replicaciones



Diseño completamente al azar, un factor

Catalizador			
Nivel			
1	2	3	4
y_{11}	y_{12}	y_{13}	y_{14}
y_{21}	y_{22}	y_{23}	y_{24}
y_{31}	y_{32}	y_{33}	y_{34}
y_{41}	y_{42}		y_{44}
y_{51}			

Cada letra representa un tratamiento y a su vez representa un nivel del factor bajo estudio, entonces hay cuatro tratamientos, cada tratamiento tiene cuatro replicaciones



Diseño completamente al azar, un factor

Pregunta científica

Existen diferencias en los promedios de concentración del componente entre los tipos de catalizador

Catalizador			
Nivel			
1	2	3	4
y_{11}	y_{12}	y_{13}	y_{14}
y_{21}	y_{22}	y_{23}	y_{24}
y_{31}	y_{32}	y_{33}	y_{34}
y_{41}	y_{42}		y_{44}
y_{51}			
<hr/>			
\bar{y}_1	\bar{y}_2	\bar{y}_3	\bar{y}_4



Diseño completamente al azar

Variable respuesta

- ▶ Concentración del componente.



Diseño completamente al azar, un factor

Hipótesis estadística

- ▶ $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$
- ▶ La hipótesis nula dice que los promedio de concentración de los componentes entre catalizadores son iguales.
- ▶ $H_1: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$
- ▶ La hipótesis alterna dice que los promedios de concentración de los componentes entre catalizadores son diferentes.



Diseño completamente al azar, un factor

Modelo estadístico

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma^2)$$

El modelo se puede interpretar como que la respuesta depende de un valor medio definido por μ mas un efecto de cada tratamiento α_i mas el componente de error de cada medición.



Diseño completamente al azar, un factor

Modelo estadístico

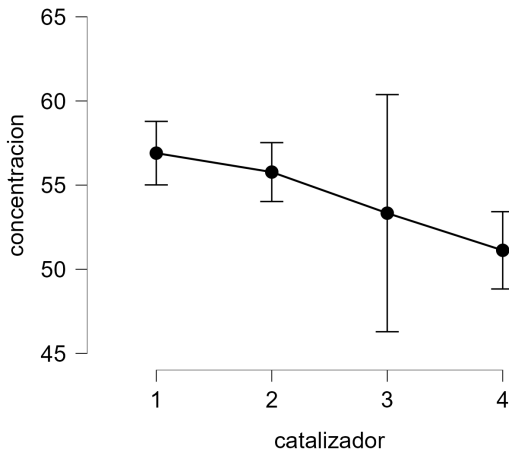
$$Y_{ij} = \bar{Y}_{\bullet} + (\bar{Y}_i - \bar{Y}_{\bullet}) + (Y_{ij} - \bar{Y}_i)$$

Donde Y_{ij} es el valor de cada observación, \bar{Y}_{\bullet} corresponde a la media total, \bar{Y}_i es la media de cada grupo. Para probar las hipótesis que nos interesan descomponemos todas las fuentes de variabilidad del modelo estadístico.



Diseño completamente al azar, un factor

Datos



Diseño completamente al azar, un factor

Análisis de varianza: fuentes de variación

- ▶ La variación de los datos dentro de cada grupo.
- ▶ La variación de los datos entre los grupos.
- ▶ La variación total de los datos



Diseño completamente al azar, un factor

Análisis de varianza: sumas de cuadrados para cada fuente de variación.

- ▶ La suma de cuadrados dentro de cada grupo.

$$SC_{dentro} = \sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2$$

- ▶ La suma de cuadrados entre los grupos.

$$SC_{entre} = \sum_i \sum_j (Y_i - \bar{Y}_{\bullet})^2 = \sum_j n_j (Y_j - \bar{Y}_{\bullet})^2$$

- ▶ La suma de cuadrados total de los datos

$$SC_{dentro} + SC_{entre} = SC_{total} = \sum_i \sum_j (Y_{ij} - \bar{Y}_{\bullet})^2$$



Diseño completamente al azar, un factor

Análisis de varianza: sumas de cuadrados para cada fuente de variación.

Fuente de Variacion	G.L.	Sumas de Cuadrados (SC)	Cuadrados Medios (CM)
Entre grupos	r-1	$SC_{entre} = \sum_j n_i (\bar{Y}_i - \bar{Y}_{\bullet})^2$	$\frac{SC_{entre}}{r-1}$
Error (dentro de tratamientos)	n-r	$SC_{dentro} = \sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2$	$\frac{SC_{dentro}}{n-r}$
Total	n-1	$SC_{total} = \sum_i \sum_j (Y_{ij} - \bar{Y}_{\bullet})^2$	



Diseño completamente al azar, un factor

Análisis de varianza: prueba de F

El objetivo de la prueba de F en el análisis de varianza es determinar si hay al menos una diferencia entre las medias de los grupos de tratamiento

$$F = \frac{SC_{dentro}}{SC_{entre}} = \frac{\frac{SC_{entre}}{r-1}}{\frac{SC_{dentro}}{n-r}}$$

La prueba es a una cola y la regla de decisión es

Concluir H_0 Si $F^* \leq F(1 - \alpha; r - 1, n - r)$

Concluir H_1 Si $F^* > F(1 - \alpha; r - 1, n - r)$



Diseño completamente al azar, un factor

Table: Descriptives - concentracion

catalizador	N	Mean	SD	SE	Coefficient of variation
1	5	56.900	1.520	0.680	0.027
2	4	55.775	1.100	0.550	0.020
3	3	53.333	2.836	1.637	0.053
4	4	51.125	1.443	0.722	0.028



Diseño completamente al azar, un factor

Table: ANOVA - concentracion

Cases	Sum of Squares	df	Mean Square	F	p	η^2
catalizador	84.948	3	28.316	9.653	0.002	0.707
Residuals	35.202	12	2.933			



Diseño completamente al azar, un factor

Comparaciones múltiples

- ▶ Cuando la prueba de F nos revela que hay diferencias significativas entre los grupos de tratamiento, nos interesa saber entre cuales grupos encontramos la diferencia.
- ▶ Existen varios procedimientos para realizar estas comparaciones entre grupos tambien conocidas como COMPARACIONES MULTIPLES.
- ▶ Una forma es realizar **multiples pruebas de t entre pares de grupos**, y otra forma es realizar un metodo de comparaciones multiples como puede ser el **metodo de Tukey de Diferencias Honestas Significativas**.



Diseño completamente al azar, un factor

Comparaciones múltiples

El problema de las comparaciones múltiples

Cuando se comparan múltiples pares de grupos de tratamiento, se asume que hay una probabilidad de cometer un error al aceptar la hipótesis nula, y este error lo conocemos como el nivel de significancia α .

Esto implica que el p-valor hallado en cada comparación puede estar sobrevalorado, y esta sobrevaloración se puede corregir ajustando el p-valor utilizando un metodo de ajuste tal como el metodo de Bonferroni, o el metodo de Holm, o el metodo de Hochberg o muchas mas opciones.



Diseño completamente al azar, un factor

Comparaciones múltiples, Bonferroni

El metodo de Bonferroni consiste en multiplicar el p-valor por el numero de comparaciones multiples, cuando el p-valor ajustado da un valor mayor que 1.0, el p-valor final es igual a 1.0.



Diseño completamente al azar

Comparaciones múltiples, prueba de Tukey

Se puede hacer un intervalo de confianza que si incluye a 0 significa que la diferencia entre dos grupos de tratamiento no es significativa

$$\hat{D} \pm T_s\{\hat{D}\}$$

$$\hat{D} = \bar{Y}_i - \bar{Y}_{i'}$$

$$s^2\{\hat{D}\} = s^2\{\bar{Y}_i\} + s^2\{\bar{Y}_{i'}\} = CM_{ERROR}\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)$$

$$T = \frac{1}{\sqrt{2}}q(1 - \alpha; r, n - r)$$

Donde q corresponde a los valores tabulados de la “distribucion del rango studentizado”.



Diseño completamente al azar

Comparaciones múltiples, Tukey test

Se puede también calcular un estadístico para cada comparación entre dos grupos de tratamiento.

$$\begin{aligned}\hat{D} &= \bar{Y}_i - \bar{Y}_{i'} \\ s^2\{\hat{D}\} &= s^2\{\bar{Y}_i\} + s^2\{\bar{Y}_{i'}\} = CM_{ERROR}\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right) \\ q^* &= \frac{\sqrt{2}\hat{D}}{s\{\hat{D}\}}\end{aligned}$$

Donde q^* corresponde al estadístico que se compara contra los valores tabulados de la "distribución del rango studentizado", y la regla de decisión es: aceptar H_0 si $|q^*| \leq q(1 - \alpha, r, n - r)$, de otra forma se acepta la H_1 .



Diseño completamente al azar, un factor

Table: Post Hoc Comparisons - catalizador

		Mean Diff	95% CI for Mean Difference		SE	t	P _{tukey}	P _{bonf}
			Lower	Upper				
1	2	1.125	−2.286	4.536	1.149	0.979	0.764	1.000
	3	3.567	−0.147	7.280	1.251	2.851	0.061	0.087
	4	5.775	2.364	9.186	1.149	5.026	0.001	0.002
2	3	2.442	−1.442	6.325	1.308	1.867	0.292	0.519
	4	4.650	1.054	8.246	1.211	3.840	0.011	0.014
3	4	2.208	−1.675	6.092	1.308	1.688	0.371	0.703



Diseño completamente al azar, un factor

Ejemplo

Una empresa de productos lácteos utiliza seis máquinas de llenado de cajas de leche UHT, de la misma marca y modelo. Control de calidad se ha quejado que las seis máquinas no aplican la misma cantidad de llenado entre los cartones. Un consultor requiere que 20 cajas de cada máquina sean tomadas al azar de un turno de producción, y el contenido de cada caja sea pesado cuidadosamente.



Diseño completamente al azar

Variable respuesta

- ▶ La diferencia de peso en gramos entre el peso de interes que es 1000 gramos de leche y el peso obtenido con una bascula de precisión hasta cuatro dígitos decimales.

$$\text{DIFERENCIA} = 1000 - \text{PESO DEL LIQUIDO}$$



Diseño completamente al azar, un factor

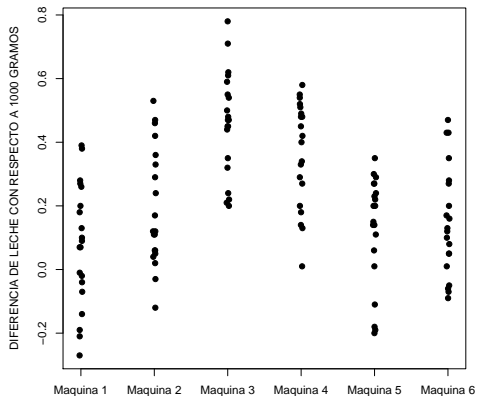
Hipótesis estadística

- ▶ $H_0: \mu_{Maquina1} = \mu_{Maquina2} = \mu_{Maquina3} = \mu_{Maquina4} = \mu_{Maquina5} = \mu_{Maquina6}$
- ▶ La hipótesis nula dice que la media de la máquina 1 es igual a la media de la máquina 2, igual a la media de la máquina 3, igual a la media de la máquina 4, igual a la media de la máquina 5, igual a la media de la máquina 6.
- ▶ $H_A: \mu_{Maquina1} \neq \mu_{Maquina2} \neq \mu_{Maquina3} \neq \mu_{Maquina4} \neq \mu_{Maquina5} \neq \mu_{Maquina6}$
- ▶ La hipótesis nula dice que la media de la máquina 1 es diferente a la media de la máquina 2, diferente a la media de la máquina 3, diferente a la media de la máquina 4, diferente a la media de la máquina 5, diferente a la media de la máquina 6.



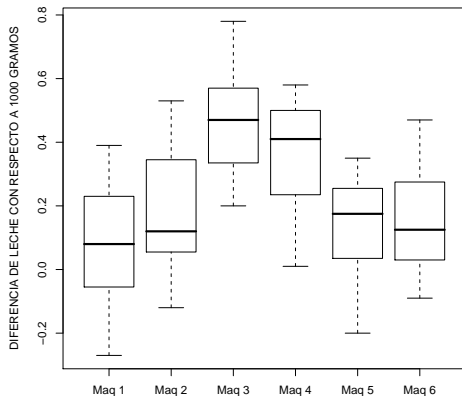
Diseño completamente al azar, un factor

LOS DATOS



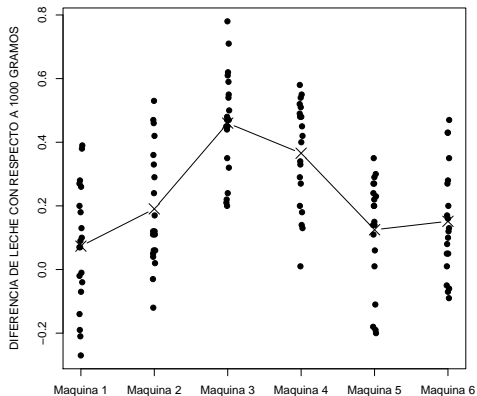
Diseño completamente al azar, un factor

Datos



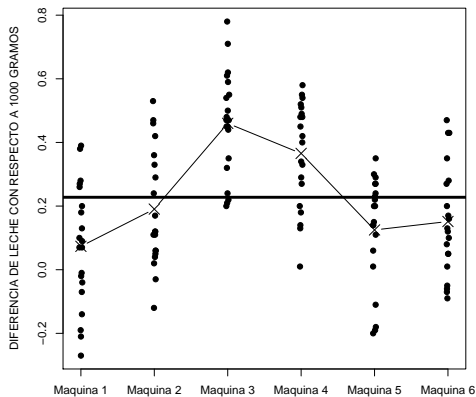
Diseño completamente al azar, un factor

Datos



Diseño completamente al azar, un factor

Existen diferencia en el llenado promedio



Diseño completamente al azar, un factor

Análisis de varianza, salida de software

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
M\'aquina	5	2.2893	0.4579	14.784	3.636e-11
Residuals	114	3.5306	0.0310		

Se encuentran diferencias significativas entre las máquinas, a los niveles de significancia $\alpha=0.05$ y $\alpha=0.01$.



Diseño completamente al azar, un factor

Comparaciones múltiples

Máquina 1	0.0735
Máquina 2	0.1905
Máquina 3	0.4600
Máquina 4	0.3655
Máquina 5	0.1250
Máquina 6	0.1515



Diseño completamente al azar, un factor

Comparaciones múltiples, pvalor sin ajustar

Pairwise comparisons using t tests with pooled SD

	1	2	3	4	5
2	0.0377	-	-	-	-
3	2.5e-10	4.1e-06	-	-	-
4	7.2e-07	0.0021	0.0922	-	-
5	0.3567	0.2417	2.2e-08	3.3e-05	-
6	0.1638	0.4849	1.9e-07	0.0002	0.6349

Hay diferencias entre las máquinas 1 y 3, 1 y 4, 2 y 3, 2 y 4, 3 y 5, 3 y 6, 4 y 5, y 4 y 6, al nivel de significancia $\alpha=0.01$.



Diseño completamente al azar, un factor

Comparaciones múltiples, ajustadas por Bonferroni

Pairwise comparisons using t tests with pooled SD

	1	2	3	4	5
1					
2	0.5658				
3	3.7e-09	6.1e-05			
4	1.1e-05	0.0318	1.0000		
5	1.0000	1.0000	3.3e-07	0.0005	
6	1.0000	1.0000	2.9e-06	0.0030	1.0000

P value adjustment method: Bonferroni

Hay diferencias entre las máquinas 1 y 3, 1 y 4, 2 y 3, 3 y 5, 3 y 6, 4 y 5, y 4 y 6, al nivel de significancia $\alpha=0.05$.



Diseño completamente al azar, un factor

Comparaciones múltiples, Tukey

```
$f.m\'aquina
diff      lwr      upr    p adj
2-1  0.117 -0.044   0.278 0.293
3-1  0.386  0.225   0.547 0.000
4-1  0.292  0.130   0.453 0.000
5-1  0.051 -0.109   0.212 0.939
6-1  0.078 -0.083   0.239 0.726
3-2  0.269  0.108   0.430 0.000
4-2  0.175  0.013   0.336 0.025
5-2 -0.065 -0.226   0.095 0.846
```

Hay diferencias entre las máquinas 1 y 3, 1 y 4, 2 y 3, 4 y 2 al nivel de significancia $\alpha=0.05$.



Diseño completamente al azar, un factor

Comparaciones múltiples, Tukey

\$f.maquina				
	diff	lwr	upr	p adj
6-2	-0.0390	-0.200	0.122	0.981
4-3	-0.0945	-0.255	0.066	0.536
5-3	-0.3350	-0.496	-0.173	0.000
6-3	-0.3085	-0.469	-0.147	0.000
5-4	-0.2405	-0.401	-0.079	0.001
6-4	-0.2140	-0.375	-0.052	0.002
6-5	0.0265	-0.134	0.187	0.996

Hay diferencias entre las máquinas 3 y 5, 3 y 6, 4 y 5, y 4 y 6, al nivel de significancia $\alpha=0.05$.

