Diseños y análisis de experimentos: Taller Final

Andrés Felipe Palomino - David Stiven Rojas - Mateo Trochez
Códigos:1922297-1924615-1931043
Universidad del Valle

26 de julio de 2023

Universidad del Valle

Ejercicio 1: La malta cervecera se produce a partir de cebada germinada, por lo que a los cerveceros les gusta saber en qué condiciones deben germinar su cebada. Lo que sigue es parte de un experimento sobre la germinación de la cebada. Las semillas de cebada se dividieron en 30 lotes de 100 semillas, y cada lote de 100 semillas se germinó en una de las diez condiciones elegidas al azar. Las condiciones son las diez combinaciones de semanas después de la cosecha (1, 3, 6, 9 o 12 semanas) y la cantidad de agua utilizada en la germinación (4 ml u 8 ml). La respuesta es el número de semillas que germinan. Nos interesa saber si el momento y/o la cantidad de agua afectan a la germinación. Analice estos datos para determinar cómo depende la tasa de germinación de los tratamientos.

Objetivos del estudio:

- Evaluar el efecto que tiene el tiempo de germinación y cantidad de agua en el número de semillas germinadas

Hipótesis de efectos principales:

- H_0: El efecto de los niveles del factor semanas después de la cosecha es igual a cero.
- H_0: El efecto de los niveles del factor cantidad de agua en la germinación es igual a cero.
- H_1: Por lo menos uno de los niveles del factor semanas después de la cosecha tiene efecto significativo sobre el número de semillas germinadas.
- H_1: Por lo menos uno de los niveles del factor cantidad de agua en la germinación tiene efecto significativo sobre el número de semillas germinadas.

Hipótesis de la interacción:

- H_0: El efecto de la interacción de los factores es igual a cero.
- H_0: El efecto de la interacción de los factores es distinto a cero.

Modelo estadístico con estructura factorial:

$$y_{ijk} = \mu + \tau_i + \alpha_j + (\tau\alpha)_{ij} + \epsilon_{ijk} \text{ con i=1,2,3,4,5 ; j=1,2 ; k=1,2,3}$$

y_{ij}: Número de semillas germinadas
μ: Promedio global de producción
τ_i: Efecto debido al i_esimo nivel del factor semanas
α_j: Efecto debido a j_esimo nivel de la cantidad de agua
$(\tau\alpha)_{ij}$: Efecto debido a la interacción de los factores semanas después de la cosecha y cantidad de agua
ϵij: Error aleatorio debido a factores no observados.

Factores:

- Semanas después de la cosecha
- Cantidad de agua en la germinación

Niveles:

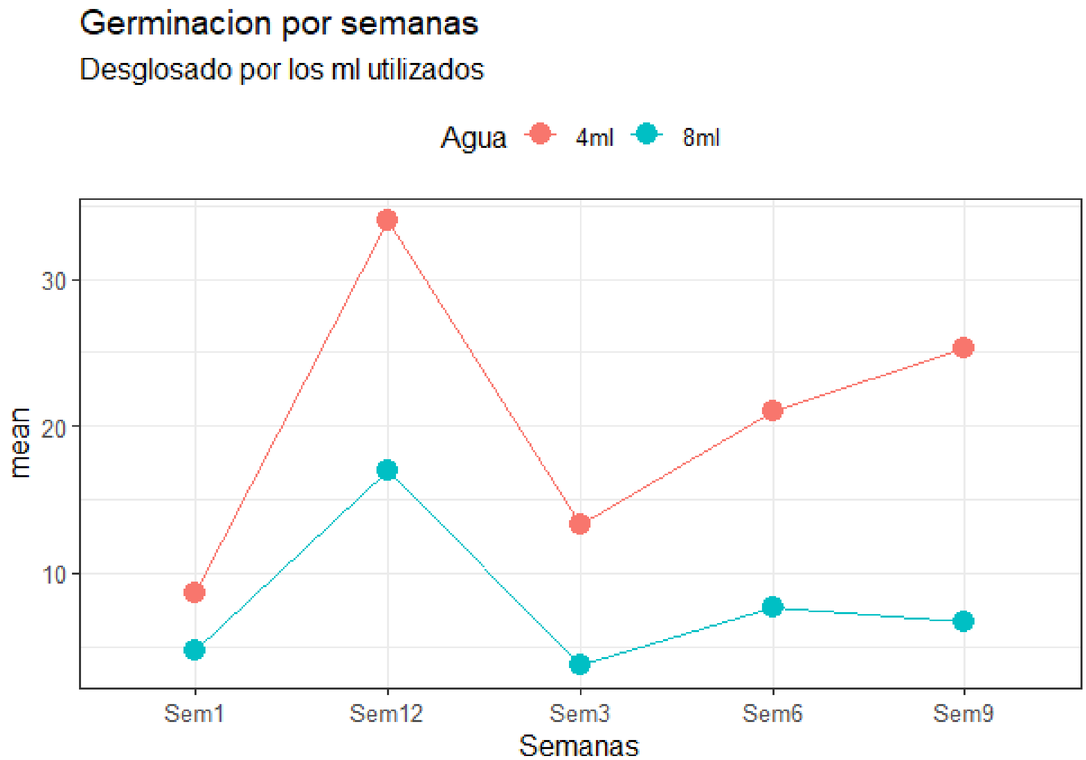
- 1 semana, 3 semanas, 6 semanas, 9 semanas, 12 semanas.
- 4 ml, 8 ml.

Tratamientos:

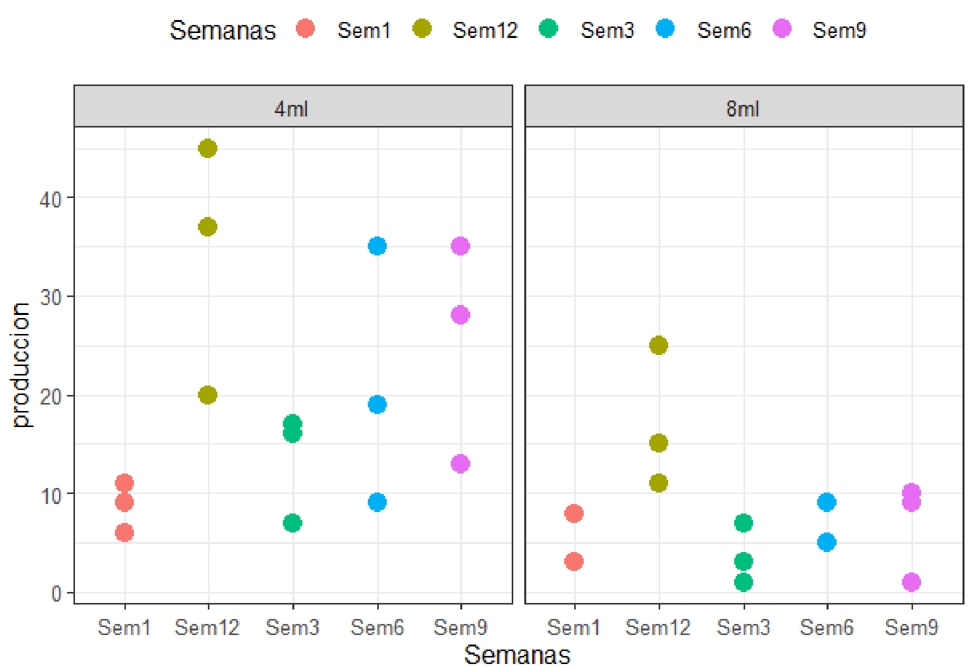
-

Tratamiento	Agua (ml)	Edad de la semilla (Semanas)
1	4	1
2	4	3
3	4	6
4	4	9
5	4	12
6	8	1
7	8	3
8	8	6
9	8	9
10	8	12

Análisis descriptivo



El anterior gráfico ilustra la germinación promedio de semillas considerando las cantidades de semanas después de la cosecha y los mililitros de agua utilizados en la germinación, podemos destacar que para los 4 mililitros de agua utilizados la producción promedio de semillas es mayor con respecto al uso de 8 mililitros de agua, sin importar la cantidad de semanas, aunque esta germinación fue mucho mayor cuando fueron 12 semanas después de la cosecha



En los gráficos anteriores se observa el comportamiento de la producción de semillas germinadas en términos absolutos a lo largo de las semanas. Estos gráficos están clasificados en dos tipos de cosecha, en los cuales se usaron 4 ml y 8 ml de agua respectivamente. A partir de estos podemos evidenciar lo que ya mencionamos y es que la germinación de semillas es mucho mayor cuando utilizamos solo 4 ml en la cosecha, y el tratamiento que nos brinda mayor producción es cuando se utilizan 4 ml con 12 semanas después de la cosecha.

Analisis de varianza (ANOVA)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Agua	1	1178.13	1178.13	19.72	0.0003
Semanas	4	1321.13	330.28	5.53	0.0036
Agua:Semanas	4	208.87	52.22	0.87	0.4967
Residuals	20	1194.67	59.73		

Observando los valores P, determinamos que se rechaza la hipótesis nula para los dos factores y concluimos que son significativas a un nivel de significancia del 0.05, estos factores son la de cantidad de agua y semanas después de la cosecha, en cuanto a la interacción nos dio un valor P de 0.49 con lo cual no rechazamos la hipótesis nula y concluimos que el efecto de la interacción en el modelo es igual a cero.

Diferencia mínima significativa (LSD)

Hipotesis para el factor cantidad de agua::

- H_0: No hay diferencia significativa entre los niveles de la variable Agua.
- H_1: Existe al menos una diferencia significativa entre los niveles de la variable Agua.

Comparación	$\lvert\bar{y}_{i.} - \bar{y}_{j.}\rvert$	LSD	Conclusión
$\bar{y}_{1.} - \bar{y}_{2.}$	12.53333	5.88687	Rechazo H_0

Existencia diferencias entre la concentración de agua presentando un mayor rendimiento la de 4ml.

Hipotesis para el factor semanas despues de la cosecha::

- H_0: No hay diferencia significativa entre los niveles de la variable Semanas.
- H_1: Existe al menos una diferencia significativa entre los niveles de la variable Semanas.

Comparación	$\lvert\bar{y}_{i.} - \bar{y}_{j.}\rvert$	LSD	Conclusión
$\bar{y}_{5.} - \bar{y}_{4.}$	9.5	9.307959	Rechazo H_0
$\bar{y}_{5.} - \bar{y}_{3.}$	11.1	9.307959	Rechazo H_0
$\bar{y}_{5.} - \bar{y}_{2.}$	17	9.307959	Rechazo H_0
$\bar{y}_{5.} - \bar{y}_{1.}$	18.8	9.307959	Rechazo H_0
$\bar{y}_{4.} - \bar{y}_{3.}$	1.66	9.307959	No rechazo H_0
$\bar{y}_{4.} - \bar{y}_{2.}$	7.5	9.307959	No rechazo H_0
$\bar{y}_{4.} - \bar{y}_{1.}$	9.4	9.307959	Rechazo H_0
$\bar{y}_{3.} - \bar{y}_{2.}$	5.8	9.307959	No rechazo H_0
$\bar{y}_{3.} - \bar{y}_{2.}$	7.7	9.307959	No rechazo H_0
$\bar{y}_{2.} - \bar{y}_{1.}$	1.9	9.307959	No rechazo H_0

Observamos que el promedio asignado a la mayor cantidad de edad en las semillas (12) corresponden al mayor rendimiento de germinación en la cebada.

Ejercicio 2: Una compañía textil fabrica un tejido en un gran número de telas. Le gustaría que las telas fueran homogéneas a fin de obtener un tejido con resistencia uniforme. El ingeniero del proceso sospecha que, además de la variación usual de la resistencia dentro de la muestra de tejido de la misma tela, puede haber también variaciones en la resistencia entre una tela y otra. Para investigar esta posibilidad, el ingeniero selecciona cuatro telas al azar y hace cuatro determinaciones de la resistencia del tejido fabricado en cada tela. Este experimento se corre de manera aleatoria y los datos obtenidos se muestran en la siguiente tabla.

Telas	Observaciones			
	1	2	3	4
1	98	97	99	96
2	91	90	93	92
3	96	95	97	95
4	95	96	99	98

Para este caso seleccionamos un diseño con factores aleatorios debido a qué aunque nuestro interés radica en determinar si se tiene una resistencia uniforme en el gran número de telas. Por lo que se realiza una selección aleatoria para poder hacer inferencias de la población.

Modelo estadístico (Modelo de efectos aleatorios):

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \text{ con i=1,2,3,4 ; j=1,2,3,4}$$

Donde:
y_{ij} = Resistencia de la i_esima tela y j_esima observación.
μ = Promedio global de resistencia.
τ_i= es el efecto aleatorio de la i_ésima tela
ϵ_{ij}= error aleatorio debido a factores no observados.

$$\tau_i \sim N(0, \sigma_\tau^2) \text{ y } \epsilon_{ij} \sim N(0, \sigma^2)$$

Hipótesis de interés

- H_0 : No hay variedad en los efectos de las telas.
- H_1 : Existe variabilidad entre las telas.

$$H_0 : \sigma_\tau^2 = 0 \text{ vs } H_{1:}\sigma_\tau^2 > 0$$

Fuentes de Variación	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Entre grupos (Telas)	3	89.19	29.73	15.68	0.0002
Dentro de grupos (Error)	12	22.75	1.90		
Total	15	112.24			

Tabla: Análisis de varianza.

Como se observa en la Tabla de Análisis de varianza, se tiene un valor p de 0.0002, por lo que a un nivel de significancia del 5 %, se concluye que se rechaza la hipótesis nula, por lo que hay variaciones en la resistencia entre las telas y por ende se necesita de un proceso de homogeneización en las telas para obtener un tejido con resistencia uniforme.

Componentes de Varianza:

$\hat{\sigma}^2 = CM_{Error}$=1.9

$\hat{\sigma}_\tau^2 = \frac{CM_{Trat}-CM_{Error}}{r} = \frac{29{,}73-1{,}9}{4} = 6{,}9575$

$\hat{\sigma}_{Total}^2 = 6{,}9575 + 1{,}9 = 8{,}8575$

La varianza total (8.8575) se descompone en una parte atribuible a la diferencia entre las telas (1.9) y otra procedente de la variabilidad existente dentro de ellos (6.9575)

Ejercicio 3:

Una empresa automotriz busca determinar cuál marca de neumáticos tiene una mayor durabilidad. Para lograrlo, llevarán a cabo un experimento en el que compararán cuatro marcas de neumáticos (A, B, C, D). Estas marcas serán sometidas a una prueba de recorrido de 50,000 km, utilizando cuatro tipos diferentes de automóviles y colocando los neumáticos en las cuatro posiciones posibles en cada automóvil. La medida que se utilizará es el desgaste de los neumáticos, donde valores más altos indicarán un mayor nivel de desgaste.

	Automóvil			
Posición	1	2	3	4
1	$C = 12$	$D = 11$	$A = 13$	$B = 8$
2	$B = 14$	$C = 12$	$D = 11$	$A = 13$
3	$A = 17$	$B = 14$	$C = 10$	$D = 9$
4	$D = 13$	$A = 14$	$B = 13$	$C = 9$

Para este caso seleccionamos un Diseño cuadrado latino, ya que estamos interesados en cuál marca de neumáticos tiene una mayor durabilidad, sin embargo, es necesario controlar fuentes de variabilidad como lo puede ser el automóvil que se utilice y además la posición de las mismas también genera un desgaste diferente en los neumáticos.

Factores:

- Tipo de Automóvil
- Posición de neumático
- Marca de neumático

Niveles:

- Automóvil (1), Automóvil (2), Automóvil (3), Automóvil (4).
- Posición (1), Posición (2), Posición (3), Posición (4).
- A, B, C, D.

Tratamientos:

Los tratamientos deben cumplir el diseño, por lo que en primer lugar cada Marca de neumático debe encontrarse en todas las posiciones y cada automóvil, para ello por filas y columnas se debe encontrar solo una repetición por marca de neumático.

Hipótesis de interés:

- H_0 : El efecto de las marcas de neumáticos es igual.
- H_1 : El efecto de alguna de las marcas presenta diferencias.

$$H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0 \text{ vs } H_1 : \tau_i \neq 0 \text{ para algún i donde i=1,2,3,4}$$

Modelo Estadístico (DCL)

$$y_{ijk} = \mu + \tau_i + \alpha_j + \gamma_k + \epsilon_{ijk} \text{ con i=1,2,3,4 ; j=1,2,3,4 ; k=1,2,3,4}$$

Donde:
y_{ij} = es el valor de la Variable de respuesta con el j_esima posición y k_esimo automóvil para la i_ésima marca de neumático.
μ = Promedio global de desgaste.
τ_i= es el efecto debido al i_ésima marca de neumático.
α_j= es el efecto debido al j_ésimo posición del neumático.
γ_k= es el efecto debido al k_ésimo automóvil.
ϵ_{ijk}= error aleatorio debido a factores no observados.

Tabla: Análisis de varianza.

Fuente de Variación	G.L	Sum_Sq	Mean_Sq	F_value	Pr(>F)
Tratamiento	3	30.69	10.23	11.42	0.006
Automóvil	3	38.69	12.90	14.40	0.003
Posición	3	6.19	2.06	2.30	0.176
Error	6	5.38	0.90		
Total	15	80.94	5.40		

Observando el valor P de la Tabla del Análisis de varianza, concluimos que rechazamos la hipótesis nula, Por ende el efecto de alguna de las marcas presenta diferencias, es decir que por lo menos una presenta una marca de neumático, presenta una durabilidad distinta a las demás.

LSD

Se desea probar $H_0 : \mu_i = \mu_j$ vs $H_1 : \mu_i \neq \mu_j$ para todo $i \neq j$.

Estadistico de prueba:

$$|\bar{y}_{i.} - \bar{y}_{j.}|$$

LSD= $t_{\frac{\alpha}{2},(t-1)(t-2)} * \sqrt{\frac{2CM_{Error}}{t}}$
LSD= $2{,}446912 * \sqrt{\frac{2*0{,}895}{4}} = 1{,}637634$
Donde $\mu_i = \mu_j$ son significativamente diferentes a nivel α si $|\bar{y}_{i..} - \bar{y}_{j..}| > LSD$

Comparación	$\lvert\bar{y}_{i..} - \bar{y}_{j..}\rvert$	LSD	Conclusión
$\bar{y}_{1..} - \bar{y}_{2..}$	2	1.6376	Rechazo H_0
$\bar{y}_{1..} - \bar{y}_{3..}$	3.5	1.6376	Rechazo H_0
$\bar{y}_{1..} - \bar{y}_{4..}$	3.25	1.6376	Rechazo H_0
$\bar{y}_{2..} - \bar{y}_{3..}$	1.5	1.6376	No rechazo H_0
$\bar{y}_{2..} - \bar{y}_{4..}$	1.25	1.6376	No rechazo H_0
$\bar{y}_{3..} - \bar{y}_{4..}$	0.25	1.6376	No rechazo H_0

Con base en la prueba LSD se observa que la Marca A presenta diferencias significativas entre la Marca B, C y D. Además, esta es la que presenta mayor media, por lo que quiere decir que la marca A de neumáticos es la peor, ya que presenta un mayor nivel de desgaste medio. Con respecto a las demás marcas, entre ellas se observa que la media son significativamente iguales, por lo que son iguales de buenas.

Contrastes

$$K_1 = \tau_1 - \tau_2$$
$$K_2 = \tau_1 + \tau_2 - 2\tau_3$$
$$K_3 = \tau_1 + \tau_2 + \tau_3 - 3\tau_4$$

Hipótesis según el contraste K:
K1:

$$H_0 : \tau_1 - \tau_2 = 0 \text{ vs } H_1 : \tau_1 - \tau_2 \neq 0$$

K2:

$$H_0 : \tau_1 + \tau_2 - 2\tau_3 = 0 \text{ vs } H_1 : \tau_1 + \tau_2 - 2\tau_3 \neq 0$$

K3:

$$H_0 : \tau_1 + \tau_2 + \tau_3 - 3\tau_4 = 0 \text{ vs } H_1 : \tau_1 + \tau_2 + \tau_3 - 3\tau_4 \neq 0$$

De los contrastes anteriores tenemos lo siguiente:

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: aov(formula = tiempo ~ metodo + automovil + orden, data = datos)

Linear Hypotheses:
        Estimate Std. Error t value Pr(>|t|)
k1 == 0   2.0000     0.6693   2.988   0.0643 .
k2 == 0   5.0000     1.1592   4.313   0.0141 *
k3 == 0   4.2500     1.6394   2.592   0.1067
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

Con base en los valores p, concluimos que se rechaza la hipótesis nula para el contraste K2, lo que quiere decir que lo que quiere decir que el efecto del tratamiento 3 es distinto al promedio de los efectos del tratamiento 1 y 2.

Validación de supuestos

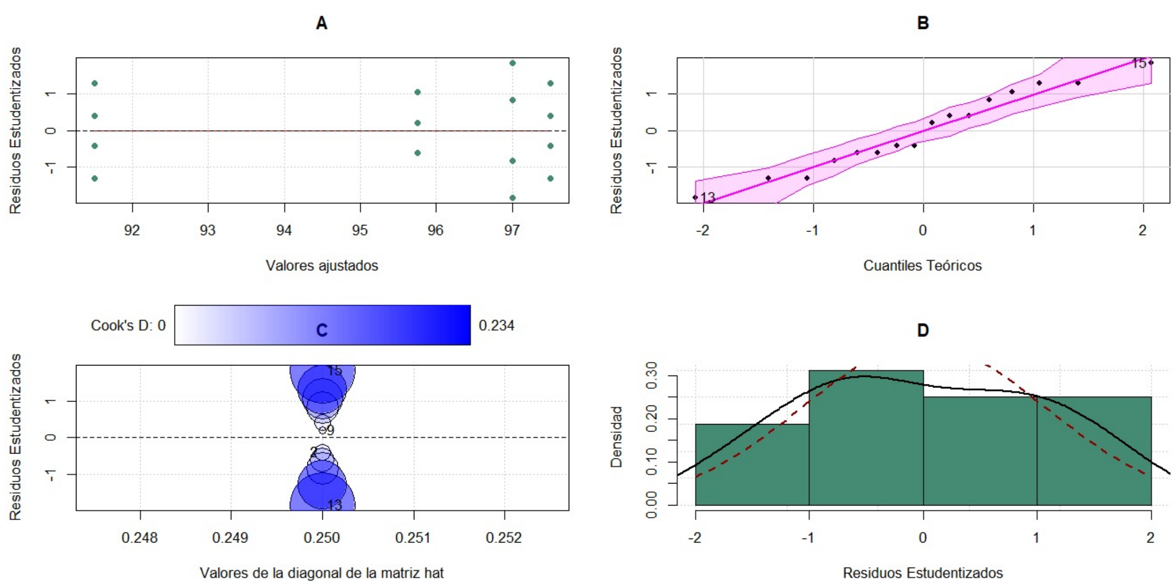


Figura : Residuales (A) , QQ-norm (B) , Puntos Influyentes (C), Histograma Residuales (D)

Como se observa en la figura (A), los residuales no presentan algún patrón de comportamiento que indique problemas de heteroscedasticidad, ya que estos no presentan formas de cono. Los residuales encuentran alrededor de dos desviaciones estándar, lo que indica que no hay evidencia de puntos atípicos, lo cual se rectifica con la Figura (C), ya que las distancias de Cook encontradas no son mayores al punto límite ($\frac{4}{n} = 0{,}25$). En la Figura (B) se observa que todos los puntos se encuentran dentro de las bandas de confianza y en la Figura (D) se presenta el histograma que no presenta comportamientos asimétricos marcados, lo cual indica normalidad en los errores. Rectificando las conclusiones anteriores se realizó la prueba de shapiro-wilk para evaluar normalidad y se encontró un valor p de 0.7698, lo que indica que se cumple el supuesto de normalidad, al igual se realizó la prueba de Breusch Pagan para evaluar varianza constante, se encontró con un valor p de 0.2153 lo que indica que se cumple el supuesto de homoscedasticidad. Por lo tanto, no encontramos ante un diseño experimental bien especificado y respaldado teóricamente.

Ejercicio 4:
Se llevó a cabo un experimento con el objetivo de analizar el desempeño de cuatro marcas de detergentes. Utilizando un equipo especializado, se registraron las mediciones de "blancura" para un total de 12 cargas de lavado, las cuales fueron distribuidas en tres modelos de lavadoras distintos:

Detergente	Lavadora 1	Lavadora 2	Lavadora 3
A	45	43	51
B	47	44	52
C	50	49	57
D	42	37	49

Para este caso seleccionamos un diseño en bloques debido a qué aunque nuestro interés radica en la blancura por tipo de detergente, se ha realizado un bloqueo por lavadora para poder determinar este efecto de diferencias significativas.

Modelo estadístico DCBA:

$$y_{ij} = \mu + \tau_i + \alpha_j + \epsilon ij \text{ con i=1,2,3,4 y j=1,2,3}$$

y_{ij}: Desempeño del detergente i y en la lavadora j
μ: Promedio global del desempeño en "blancura"
τ_i: Efecto debido al i_esimo detergente
α_j: Efecto debido a la j_esima lavadora
ϵij: Error aleatorio debido a factores no observados.
Hipótesis Estadística para DBCA

- $H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$
- $H_1 : \tau_i \neq 0$ para algún i donde i=1,2,3,4

Análisis de varianza:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Detergente	3	133.67	44.56	34.13	0.0004
Lavadora	2	170.17	85.08	65.17	0.0001
Residuals	6	7.83	1.31		

Observando el valor del p concluimos que rechazamos la hipótesis nula, independiente de la significancia, lo que nos indica que el efecto de los detergentes no es igual 0, lo que indicaría que el rendimiento en “blancura” cambia según la marca de detergente. A su vez, el efecto del bloque lavadora tiene un aporte significativo

Intervalos de confianza:

	Estimación Puntual	2.5 %	97.5 %
Detergente A	46.33	44.93	47.73
Detergente B	47.67	46.27	49.07
Detergente C	52.00	50.60	53.40
Detergente D	42.67	41.27	44.07

- En promedio, se espera que la “Blancura” del detergente A se encuentre con un 95 % entre 44.93 y 47.73, con un valor medio de 46.33.
- En promedio, se espera que la “Blancura” del detergente B se encuentre con un 95 % entre 46.27 y 49.07, con un valor medio de 47.67.
- En promedio, se espera que la “Blancura” del detergente C se encuentre con un 95 % entre 50.60 y 53.40, con un valor medio de 52.00.
- En promedio, se espera que la “Blancura” del detergente C se encuentre con un 95 % entre 41.27 y 44.07, con un valor medio de 42.67.

Obteniendo así los mayores niveles de rendimiento, el detergente C, seguido del B y A y por último el de menor blancura, el cual es el D.