Diseño Completamente Aleatorizado (DCA)

Luis Fernando Delgado Muñoz - Ing. Agroindustrial, M.Sc

Contenido

Definiciones del Diseño Experimental

- **Variable respuesta:** Característica medible de la cual el experimentador tiene la certeza de que le proporcionará en realidad información útil acerca del proceso bajo estudio.
- **Factores controlables:** Variables de entrada del sistema $X's$ sobre las cuales se generaran alteraciones predeterminadas.
- **Niveles:** Valores o atributos que toma cada uno de los factores en la fase de experimentación.
- **Tratamientos:** Combinación de los niveles de los factores a los cuales se someterán a prueba las unidades experimentales.

Definiciones del Diseño Experimental

- **Unidad Experimental:** Es la entidad física o el sujeto expuesto al tratamiento independientemente de otras unidades. La unidad experimental, una vez expuesta al tratamiento, constituye una sola replica del tratamiento.
- **Error Experimental:** Describe la variación entre las unidades experimentales tratadas de manera idéntica e independiente. Posibles orígenes del error experimental:
  - La variación natural entre unidades experimentales.
  - La imposibilidad de reproducir las condiciones del tratamiento con exactitud de una unidad a otra.
  - Cualquier otro factor externo que influya en las características medidas.

Definiciones del Diseño Experimental

- **Aleatorización:** Escogencia mediante procedimientos aleatorios de las unidades experimentales que serán sometidas a cada tratamiento.
  - Eliminar sesgos en la selección de las unidades experimentales.
  - Repartir el posible efecto que puedan tener los factores no controlables sobre la respuesta.
- **Replicación:** Consiste en realizar repeticiones de la variable respuesta por cada tratamiento.
  - Valorar la incertidumbre en el experimento.
  - Cuantificar el error experimental.

Ejemplo

Un ingeniero de desarrollo tiene interés en investigar la resistencia a la tensión de una fibra sintética nueva que se usará para hacer telas de camisa para caballero. Por experiencia se sabe que la calidad de la tela se ve afectada por la composición de algodón, el cual debe variar entre el 10 % y 40 % de la mezcla. Por tanto decide probar cinco niveles de algodón (15,20,25,30 y 35 %) y como necesita valoración de la variabilidad decide hacer 5 replicaciones por cada nivel.

1. Factor (es) de diseño
2. Niveles del factor de diseño
3. Unidad Experimental
4. Posibles Factores Perturbadores
5. Cual seria su estrategia para eliminar un posible efecto de factores perturbadores
6. Posibles Factores no controlables

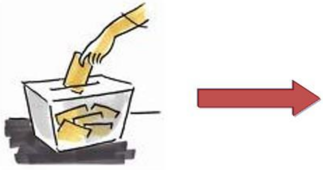
Solución

- **Factor (1):** Porcentaje de Algodón en la tela.
- **Niveles (5):** 15 % 20 % 25 % 30 % 35 %
- **Unidad Experimental:** Trozos de tela de igual tamaño tomados de forma aleatoria del proceso de producción estabilizado.
- **Posibles factores perturbadores:**
  - Alteraciones en su calibración por su uso continuo.
  - Aleatorizar el orden en que se realizan las mediciones de resistencia a la tensión de la nueva fibra sintética.
  - Se dispone de diferentes lotes de hilo y de algodón en la sección de bodega, estos deben ser aleatoriamente asignados a cada tratamiento, para eliminar sesgos por la variación en calidad de los lotes.

Solución

Aleatorización:

Porcentaje de algodón	Número de corrida experimental				
15	1	2	3	4	5
20	6	7	8	9	10
25	11	12	13	14	15
30	16	17	18	19	20
35	21	22	23	24	25



Se aleatoriza el orden en que se realizan las mediciones (aleatorización de los tratamientos).

Se selecciona un número aleatorio entre 1 y 25. Suponga que este número es 17. Entonces la observación número 17 (30% de algodón) se corre primero...

Secuencia de prueba	Corrida Experimental	Porcentaje de algodón
1	17	30
2	1	15
3	24	35
4	12	25
5	9	20
6	23	35
7	18	30
8	6	15
9	22	35
10	19	30
11	5	15
12	21	35
13	16	30
14	10	20
15	7	20
16	20	30
17	3	15
18	8	20
19	4	15
20	2	15
21	14	25
22	13	25
23	25	35
24	11	25
25	15	25

Solución

- **Factores No Controlables - Medibles:** La condición del material depende de su historia, incluyendo los procesos a los cuales ha sido sometida y los tratamientos mecánicos que ha recibido, de la cantidad de humedad que contiene y de la temperatura; todos estos datos deben especificarse en caso de que el análisis de la resistencia sea de gran importancia.

Diseños Completamente Aleatorizados (DCA)

Características

- Se investiga el efecto de un factor con a niveles (o a tratamientos) sobre una variable respuesta.
- Las unidades experimentales son homogéneas.
- Los factores perturbadores han sido satisfactoriamente controlados en el experimento.
- Se realizó un proceso de aleatorización en la asignación de unidades experimentales.

Diseños Completamente Aleatorizados (DCA)

Ejemplo

Cuadro 1: Datos ($lb/pulgada^2$) del experimento de la resistencia a la tensión.

Peso porcentual del algodón	Observaciones 1	2	3	4	5	Total	Promedio
15	7	7	15	11	9	49	9.8
20	12	17	12	18	18	77	15.4
25	14	18	18	19	19	88	17.6
30	19	25	22	19	23	108	21.6
35	7	10	11	15	11	54	10.8
Total						376	15.04

Tabla típica para un diseño de un factor

Tratamientos	Observaciones				Totales	Promedios
1	Y_{11}	Y_{12}	$\cdots$	Y_{1n_1}	$Y_{1\cdot}$	$\bar{Y}_{1\cdot}$
2	Y_{21}	Y_{22}	$\cdots$	Y_{2n_2}	$Y_{2\cdot}$	$\bar{Y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
t	Y_{t1}	Y_{t2}	$\cdots$	Y_{tn_t}	$Y_{t\cdot}$	$\bar{Y}_{t\cdot}$
					$Y_{\cdot\cdot}$	$\bar{Y}_{\cdot\cdot}$

No todos los tratamientos deben tener el mismo número de datos.

Algo de notación...

Y_{ij} = La $j - esima$ observación en el $i - esimo$ tratamiento.

$Y_{i.}$ = Suma de las observaciones del tratamiento i.

$\bar{Y}_{i.}$ = Media de las observaciones del $i - esimo$ tratamiento.

$Y_{..}$ = Suma total de las $n = n_1 + n_2 + \cdots + n_t$ mediciones.

$\bar{Y}_{..}$ = Media global o promedio general de todas las observaciones.

Algo de formulas

- $Y_{i.} = \sum_{j=1}^{n_i} Y_{ij}$
- $\bar{Y}_{i.} = \frac{\sum_{j=1}^{n_1} Y_{ij}}{n_i}$
- $Y_{..} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} Y_{ij}$
- $\bar{Y}_{..} = \frac{\sum_{i=1}^{t} \sum_{j=1}^{n_i} Y_{ij}}{n}$
- $n = \sum_{i=1}^{t} n_i$

Modelo Estadístico

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad i = 1, \cdots, t \qquad j = 1, \cdots, n_i$$

Y_{ij} Es la $j-esima$ observación de la variable de respuesta en el $i-esimo$ tratamiento.

μ Es la *media global*.

τ_i Es el efecto del $i-esimo$ tratamiento **sobre** la variable de respuesta.

ε_{ij} Es el error aleatorio de la $j-esima$ observación en el $i-esimo$ tratamiento.

Modelo Estadístico

Supuestos

1. Correcta especificación del modelo $E(\varepsilon_{ij}) = 0$.
2. Homogeneidad de Varianza (Homocedasticidad) $V(\varepsilon_{ij}) = \sigma^2$
3. Independencia $Cov(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0$
4. Normalidad $\varepsilon_{ij} \sim N(0, \sigma^2)$.

Hipótesis

Hipótesis de investigación

Parte del objetivo del estudio. Se plantea antes de realizar el experimento.

Hipótesis Estadística

Es una "traducción" de la hipótesis de investigación en términos estadísticos.

$$H_o : \tau_1 = \tau_2 = \cdots = \tau_t = 0$$
$$H_a : \tau_i \neq 0 \text{ para algún } i = 1, 2, \cdots, t.$$

Análisis de Varianza

El *Análisis de Varianza* o ANOVA (**An**alysis **of** **Va**riance) es la técnica central en el análisis de datos experimentales. Consiste en separar la **Variabilidad Total** de los datos en las diferentes fuentes de variación que están involucradas en el experimento.

$$\begin{array}{c}\text{Variabilidad}\\ \text{Total}\end{array} = \begin{array}{c}\text{Variabilidad debida}\\ \text{a los tratamientos}\end{array} + \begin{array}{c}\text{Variabilidad debida}\\ \text{al error aleatorio}\end{array}$$

Tabla del ANOVA

$$H_o : \tau_1 = \tau_2 = \cdots = \tau_t = 0$$

Fuente de Variabilidad	Grados de Libertad	Suma de Cuadrados	Cuadrados Medios	F_c	Valor p
Tratamientos	$t-1$	$\sum_{i=1}^{t} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$	$CM_\tau = \frac{SC_\tau}{t-1}$	$\frac{CM_\tau}{CM_\varepsilon}$	$\Pr(F > F_c)$
Error	$n-t$	$\sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2$	$CM_\varepsilon = \frac{SC_\varepsilon}{n-t}$		
Total	$n-1$	$\sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{..})^2$			

Modelo Estadístico

Supuestos

1. Correcta especificación del modelo $E(\varepsilon_{ij}) = 0$.
2. Homogeneidad de Varianza (Homocedasticidad) $V(\varepsilon_{ij}) = \sigma^2$
3. Independencia $Cov(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0$
4. Normalidad $\varepsilon_{ij} \sim N(0, \sigma^2)$.

- Si no se cumple el supuesto de **correcta especificación**, los estimadores pierden las propiedades de insesgamiento.
- Si no se cumplen los supuestos de **homogeneidad** e **independencia**, los estimadores pierden su condición de optimalidad.
- Si no se cumple el supuesto de **normalidad**, se pierde la eficiencia y la potencia en las pruebas de hipótesis.

Verificación de supuestos

La verificación de los supuestos puede realizarse fácilmente mediante los residuales

$$e_{ij} = Y_{ij} - \hat{Y}_{ij}$$

donde Y_{ij} es el valor observado y $\hat{Y}_{ij}$ es el valor estimado con el modelo.

Correcta especificación

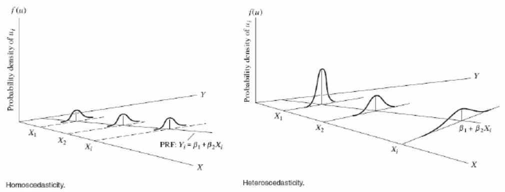
Este supuesto es difícil de validar, debido a que se requiere que en el modelo planteado, estemos involucrando todas las fuentes de variación que están influyendo en el comportamiento de la variable de respuesta.

Aquí, toma importancia la adecuada selección del tipo de diseño que vamos a implementar, pues si seleccionamos una metodología muy alejada de la situación real, no podemos obtener buenas conclusiones respecto a nuestro experimento.

Homogeneidad de Varianza

Homocedasticidad se refiere al supuesto de que la variable dependiente (Y) presenta una distribución con igual varianza en todo el rango de valores de la variable independiente (X).

Homocedasticidad vs Heterocedasticidad



Si no se cumple este supuesto los estimadores dejan de ser óptimos y las pruebas estadísticas (ANOVA, pruebas t) e intervalos de confianza pierden validez (altera el nivel de confianza)

Independencia -> (no correlación de los errores)

$$Cov(\varepsilon_i, \varepsilon_j) \neq 0$$

Si no se cumple este supuesto los estimadores dejan de ser óptimos y las pruebas estadísticas (ANOVA, pruebas t) e intervalos de confianza pierden validez (altera el nivel de confianza).

En las situaciones en las que se pueda garantizar que las observaciones y_i, constituyen una muestra aleatoria (independientes e idénticamente distribuidas), no existirá correlación de los errores, es decir, que es posible controlar este aspecto, algunas ocasiones, con base en el procedimiento de selección de la muestra (Behar, 2003).

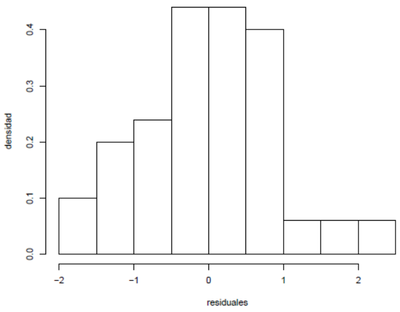
Normalidad

La normalidad de los errores permite la estimación por intervalos de confianza no sólo para los coeficientes de regresión, sino también para la predicción. Permite el planteamiento de pruebas de hipótesis sobre los parámetros del modelo. Cuando los errores no son normales, los intervalos y las pruebas de hipótesis no son exactas y pueden llegar a ser inválidas (Behar, 2003).

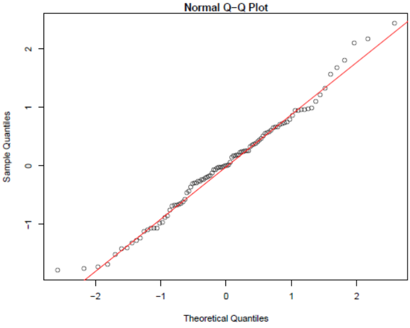
Normalidad

Caso donde se cumple el supuesto:

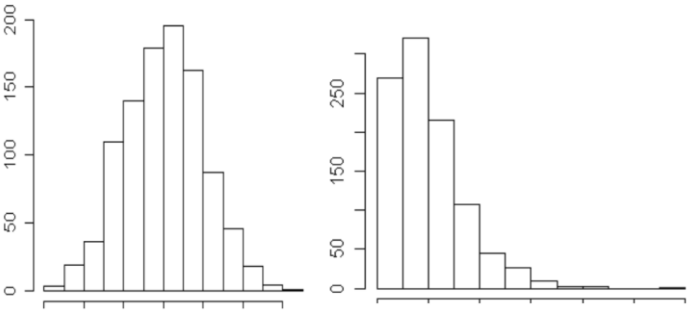
Histograma de los residuales



qq-plot de los residuales



Normalidad



Normalidad **No Normalidad**

Pruebas Formales

Homocedasticidad:
Prueba de Goldfeld-Quant, prueba de White.

Incorrelación de los errores: (correlación temporal)
Prueba de Durbin-Watson, prueba de rachas.

Normalidad de los errores:
Prueba de Shapiro-Wilks, prueba de Anderson-Darling.

Algunos de estos supuestos se pueden corregir por medio de transformaciones en algunas de las variables (y o x).

Comparaciones Múltiples

Con el ANOVA solo podemos concluir que existen diferencias estadísticamente significativas entre los tratamientos. Sin embargo, no es posible conocer cuáles son los tratamientos que difieren.

Comparación de parejas de medias de tratamientos

$$H_0 : \mu_i = \mu_j \qquad \text{para todo } i \neq j$$

Mínima Diferencia Significativa (LSD)

$$LSD = t_{\left(\frac{\alpha}{2};n-t\right)}\sqrt{CM_{\varepsilon}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

Si $|\bar{Y}_{i.} - \bar{Y}_{j.}| > LSD$ se rechaza H_0.

Prueba de Tukey (DSH)

$$T_\alpha = q_{\alpha(p,f)}\sqrt{\frac{CM_\varepsilon}{r}}$$

- α Nivel de significancia.
- q Valor de la tabla de Tukey.
- p Es el número de tratamientos en el experimento.
- f Grados de libertad asociados al error.
- CM_ε Cuadrado Medio del Error
- r Número de observaciones en cada tratamiento.

Si $|\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}| > T_\alpha$ se rechaza H_0.

Prueba de Duncan

1. Se ordenan las medias en forma ascendente.
2. Se determina el error estándar en cada promedio

$$S_{\bar{Y}_{i.}} = \sqrt{CM_{\varepsilon}/n_h}$$

   - Si el diseño es balanceado, entonces $n_h = r$.
   - Si el diseño es desbalanceado, entonces $n_h = \frac{t}{\sum_{i=1}^{t} 1/n_i}$
3. Usando la tabla de de intervalos significativos de Duncan, se obtienen los valores $r_{\alpha(p,f)}$ donde $p = 2, 3, \cdots, t$, α es el nivel de significancia y f son los grados de libertad del error.

$$R_p = r_{\alpha(p,f)} * S_{\bar{Y}_{i.}}$$

Prueba de Duncan

- Se prueban las diferencias observadas entre las medias, comenzando con el valor más alto contra el más pequeño (R_t).
- Después se calcula la diferencia entre el valor más alto contra el segundo más pequeño y se compara con el intervalo mínimo significativo R_{t-1} y así sucesivamente.
- Si $|\bar{Y}_{i.} - \bar{Y}_{j.}| > R_p$ se rechaza H_0.

Prueba de Dunnett

$$H_0 : \mu_i = \mu_c \qquad \text{para todo } i = 1, 2, \cdots, t$$

$$D = d_{\alpha(t-1,f)} \sqrt{CM_\varepsilon \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

Si $|\bar{Y}_{i.} - \bar{Y}_{j.}| > D$ se rechaza H_0.

MUCHAS GRACIAS