PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Aplicaciones de diseños experimentales

Profesor: Alex de la Cruz H.

MODELOS LINEALES 1

EST631

Contenido

Introducción

1. Introducción

- Los diseños experimentales organizan de forma metódica la recolección de datos para garantizar resultados confiables y replicables.
- Los modelos estadísticos interpretan los datos, miden impactos y gestionan la incertidumbre, fortaleciendo la validez de los experimentos.

Algunas aplicaciones:

- Ensayos clínicos personalizados: Se diseñan experimentos para evaluar la eficacia de fármacos adaptados al perfil genético de pacientes, controlando la aleatorización y los factores de confusión.
- Agricultura de precisión: Se aplican diseños factoriales para analizar la interacción entre fertilizantes y métodos de riego, optimizando rendimiento y uso de recursos con sensores y análisis en tiempo real.
- Optimización industrial: Se usan diseños robustos (Taguchi) para probar combinaciones de materiales y procesos, buscando configuraciones que maximicen calidad y minimicen variabilidad.

Aplicación de diseño experimental

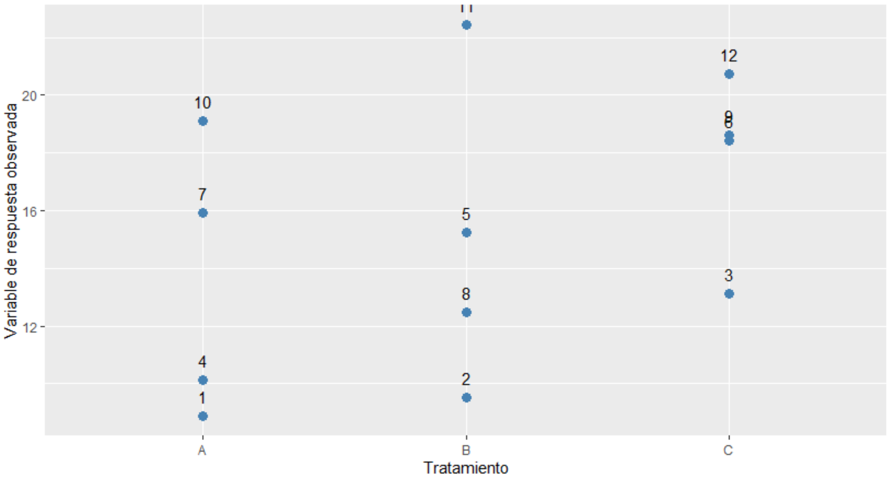
Definición 1 (diseño experimental).

Es un plan estructurado y metódico que organiza cómo se recolectarán, manipularán y analizarán datos mediante la aplicación controlada de tratamientos o intervenciones, con el fin de identificar relaciones causa-efecto, minimizar fuentes de error y garantizar resultados válidos y replicables.

Elementos principales en diseños experimentales

- Tratamientos, unidades y variables: Factores a comparar, sujetos de estudio y características medidas.
- Aleatorización y replicación: Asignación aleatoria y repetición para evitar sesgos y mejorar precisión.
- Control y modelo estadístico: Estrategias para reducir variación no deseada y marco matemático para analizar los efectos.

Elementos principales en diseños experimentales



Tipos de diseños experimentales

- Existen diferentes diseños experimentales según la pregunta de investigación y los recursos disponibles.
- Cada diseño controla la variabilidad y optimiza la precisión de las estimaciones.
- Los principales tipos de diseños se aplican ampliamente en la práctica estadística.

Tipos de diseños experimentales

Los más principales, por su forma más simple y fundamental de planificar un experimento, sirviendo como base para comprender los principios esenciales de la experimentación: aleatorización, replicación y comparación de tratamientos, son:

- Diseño completamente aleatorizado (DCA): Los tratamientos se asignan de forma totalmente aleatoria a las unidades experimentales. Es el diseño más sencillo y se usa cuando no se esperan fuentes de variación adicionales.
- Diseño en bloques completos al azar (DBCA): Las unidades se agrupan en bloques homogéneos para controlar una fuente conocida de variación, y dentro de cada bloque los tratamientos se asignan aleatoriamente.

Diseño completamente aleatorizado (DCA)

Ejemplo 1

Un investigador agrícola desea evaluar el efecto de tres tipos de fertilizantes (A, B y C) sobre el rendimiento de maíz. Para ello, dispone de 15 parcelas a las que asigna los fertilizantes de forma completamente aleatoria, con 5 parcelas por cada fertilizante.

Parcela	Fertilizante	Rendimiento (kg/ha)
1	A	3500
2	B	3650
3	C	3400
⋮	⋮	⋮
15	B	3750

Diseño completamente aleatorizado (DCA)

Ejemplo 2

Un nutricionista evalúa el efecto de tres tipos de dietas (D1, D2, D3) sobre el peso final de 30 animales. Cada animal tiene un peso inicial distinto, por lo que se incluye como covariable para ajustar el análisis.

Animal	Dieta	Peso Inicial (g)	Peso Final (g)
1	D1	1800	2200
2	D2	1750	2300
3	D3	1900	2400
⋮	⋮	⋮	⋮
30	D1	1850	2250

Diseño completamente aleatorizado (DCA)

Definición 2 (Modelo para un DCA).

Sea Y_{ij} la variable de respuesta observada para la j-ésima unidad experimental que recibe el tratamiento i, entonces el modelo estadístico del Diseño Completamente Aleatorizado se define como:

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

donde:

- μ es la media general.
- τ_i es el efecto del tratamiento i, sujeto a la restricción $\sum_{i=1}^{t} \tau_i = 0$.
- ϵ_{ij} es el término de error aleatorio, independiente e idénticamente distribuido.

Estimación del modelo para un DCA

Sea el modelo para un DCA

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2), \quad i = 1, \ldots, t, \ j = 1, \ldots, n_i, \ n = \sum_{i=1}^{t} n_i.$$

Estimación por Mínimos Cuadrados Ordinarios (MCO)

La función de suma de cuadrados es definida como:

$$S(\mu, \tau_1, \ldots, \tau_t) = \sum_{i=1}^{t} \sum_{j=1}^{n_i} [Y_{ij} - (\mu + \tau_i)]^2, \quad \text{sujeto a } \sum_{i=1}^{t} \tau_i = 0.$$

Luego:

$$\frac{\partial S}{\partial \mu} = 0 \quad \Longrightarrow \quad \hat{\mu} = \bar{Y}_{..} = \frac{1}{n} \sum_{i=1}^{t} \sum_{j=1}^{n_i} Y_{ij}.$$

$$\frac{\partial S}{\partial \tau_i} = 0 \quad \Longrightarrow \quad \hat{\tau}_i = \bar{Y}_{i.} - \bar{Y}_{..}, \quad \text{donde } \bar{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}.$$

Estimación por Máxima Verosimilitud (EMV)

Log-verosimilitud:

$$\ell(\mu, \tau, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{t}\sum_{j=1}^{n_i}[Y_{ij} - (\mu + \tau_i)]^2.$$

Maximizar ℓ para μ y τ, lleva exactamente a las mismas condiciones que los MCO

$$\frac{\partial \ell}{\partial \mu} = 0, \quad \frac{\partial \ell}{\partial \tau_i} = 0, \quad \Longrightarrow \quad \hat{\mu}_{EMV} = \bar{Y}_{..}, \quad \hat{\tau}_{i,EMV} = \bar{Y}_{i.} - \bar{Y}_{..}.$$

Estimación por Máxima Verosimilitud (EMV)

EMV de la varianza:

$$\frac{\partial \ell}{\partial \sigma^2} = 0 \quad \Longrightarrow \hat{\sigma}^2_{EMV} = \frac{1}{n} \sum_{i=1}^{t} \sum_{j=1}^{n_i} [Y_{ij} - \hat{\mu}_{EMV} - \hat{\tau}_{i,EMV}]^2.$$

Estimador de varianza (insesgado):

$$\hat{\sigma}^2 = \frac{1}{n-t} \sum_{i=1}^{t} \sum_{j=1}^{n_i} [Y_{ij} - \hat{\mu} - \hat{\tau}_i]^2.$$

ANOVA para un DCA

El modelo DCA se relaciona directamente con el ANOVA (Análisis de la Varianza) porque ANOVA es la técnica estadística que se usa para analizar este modelo.

El modelo describe cómo se descompone la variabilidad total de los datos en:

- una parte explicada por los tratamientos (τ_i), y
- una parte no explicada (los errores ϵ_{ij})

El modelo de Diseño Completamente Aleatorizado (DCA) es la base del Análisis de la Varianza (ANOVA) de un factor.

ANOVA para un DCA

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2), \quad i = 1, \ldots, t, \; j = 1, \ldots, n_i.$$

ANOVA asociado es una descomposición de varianzas. El objetivo principal es descomponer la variación total observada en:

$$\underbrace{SS_{Total}}_{\text{Total}} = \underbrace{SS_{Trat}}_{\text{Tratamientos}} + \underbrace{SS_{Error}}_{\text{Error}},$$

ANOVA para un DCA

- $SS_{Total} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2$
- $SS_{Trat} = \sum_{i=1}^{t} n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2$
- $SS_{Error} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$
- $\bar{Y}_{i.}$ es la media de las observaciones del tratamiento i.
- $\bar{Y}_{..}$ es la media general de todas las observaciones.

ANOVA para un DCA

Esta descomposición refleja la estructura del modelo:

$$Y_{ij} = \underbrace{\bar{Y}_{..}}_{\text{Media general}} + \underbrace{(\bar{Y}_{i.} - \bar{Y}_{..})}_{\text{Efecto del tratamiento}} + \underbrace{(Y_{ij} - \bar{Y}_{i.})}_{\text{Error}}.$$

ANOVA para un DCA

El ANOVA se usa para probar si existen diferencias significativas entre tratamientos:

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_t = 0.$$

El estadístico F es:

$$F_0 = \frac{CM_{Trat}}{CM_{Error}} = \frac{SC_{Trat}/(t-1)}{SC_{Error}/(n-t)}.$$

Por tanto, aplicar ANOVA es equivalente a estimar y contrastar los efectos del modelo DCA.

ANOVA para un DCA

El ANOVA se usa para probar si existen diferencias significativas entre tratamientos:

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_t = 0.$$

El estadístico F es:

$$F_0 = \frac{CM_{Trat}}{CM_{Error}} = \frac{SC_{Trat}/(t-1)}{SC_{Error}/(n-t)}.$$

Por tanto, aplicar ANOVA es equivalente a estimar y contrastar los efectos del modelo DCA.

Modelo para un DCA con covariable

Un modelo DCA con covariable se expresa como:

$$Y_{ij} = \mu + \tau_i + \beta(X_{ij} - \bar{X}_{..}) + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2), \quad i = 1, \ldots, t, \; j = 1, \ldots, n_i.$$

- μ: media general ajustada.
- τ_i: efecto del i-ésimo tratamiento, con restricción $\sum_{i=1}^{t} \tau_i = 0$.
- β: coeficiente de la covariable.
- X_{ij}: covariable observada y $\bar{X}_{..}$: media global de la covariable.

Modelo para un DCA con covariable

- **¿Por qué se centra la covariable?** Centrar la covariable alrededor de su media mantiene μ como la media ajustada de la respuesta cuando la covariable está en su valor promedio y reduce la correlación entre efectos de tratamiento y covariable.
- **Propósito del modelo** Incluir una covariable ajusta la comparación de tratamientos al eliminar variabilidad atribuible a diferencias sistemáticas, aumentando así la precisión de los estimadores.

Modelo para un DCA con covariable

Ejemplo:

Un ejemplo típico es comparar el rendimiento de varios fertilizantes (τ_i) ajustando por la humedad inicial del suelo (X_{ij}) que podría influir en la respuesta.

Su relación con regresión lineal

Para estimar este modelo en contexto de regresión lineal, se expresa como un *modelo lineal general*:

$$Y_{ij} = \beta_0 + \beta_1 D_{i1} + \beta_2 D_{i2} + \cdots + \epsilon_{ij}.$$

- β_0 representa la media del grupo de referencia (por ejemplo, tratamiento 1).
- β_k representa la diferencia entre el tratamiento $k+1$ y el tratamiento de referencia.
- D_{ik} es la variable indicadora (*dummy*) para el tratamiento $k+1$.

Su relación con regresión lineal

El vínculo entre la forma teórica y la forma estimable es:

$$\begin{cases} \text{Grupo base:} & \mu + \tau_1 = \beta_0 \\ \tau_2 & = \tau_1 + \beta_1 \\ \tau_3 & = \tau_1 + \beta_2 \\ \vdots & \end{cases} \implies \quad \text{cumple que } \sum_{i=1}^{t} \tau_i = 0.$$

Así, el MCO estima diferencias de tratamiento respecto a la categoría base, y juntos codifican la misma información que los τ_i del modelo teórico.

Ejemplo

Supongamos un experimento con $t = 3$ tratamientos. El modelo teórico es:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad \text{con} \quad \sum_{i=1}^{3} \tau_i = 0.$$

Supongamos que los valores reales son:

$$\mu = 50, \quad \tau_1 = -2, \quad \tau_2 = 1, \quad \tau_3 = 1. \quad (\text{Verifica:} \ -2 + 1 + 1 = 0)$$

Ejemplo

Entonces, las medias poblacionales por tratamiento son:

$$\begin{cases} E[Y|\text{tratamiento } 1] = 50 + (-2) = 48 \\ E[Y|\text{tratamiento } 2] = 50 + 1 = 51 \\ E[Y|\text{tratamiento } 3] = 50 + 1 = 51. \end{cases}$$

Ejemplo

En regresión lineal, se re-parametriza así:

$$Y_{ij} = \beta_0 + \beta_1 D_{i2} + \beta_2 D_{i3} + \epsilon_{ij},$$

donde:

- β_0 = media del grupo de referencia (tratamiento 1)
- β_1 = diferencia (tratamiento 2) − (tratamiento 1)
- β_2 = diferencia (tratamiento 3) − (tratamiento 1)

Ejemplo

$$
\begin{cases}
\beta_0 = E[Y|\text{tratamiento } 1] = 48 \\
\beta_1 = E[Y|\text{tratamiento } 2] - E[Y|\text{tratamiento } 1] = 51 - 48 = 3 \\
\beta_2 = E[Y|\text{tratamiento } 3] - E[Y|\text{tratamiento } 1] = 51 - 48 = 3.
\end{cases}
$$

Los efectos τ_i se relacionan con los β así:

$$
\begin{cases}
\tau_1 = \beta_0 - \mu = 48 - 50 = -2 \\
\tau_2 = \tau_1 + \beta_1 = -2 + 3 = 1 \\
\tau_3 = \tau_1 + \beta_2 = -2 + 3 = 1.
\end{cases}
$$

Ejemplo de aplicación

Para ilustrar un modelo para DCA, usaremos la base de datos `InsectSprays`, que contiene información sobre el número de insectos eliminados al aplicar seis tipos diferentes de insecticidas. Cada observación corresponde a un conteo de insectos tratados con uno de los sprays, asignado de forma aleatoria a unidades experimentales independientes.

- Existen un total de 72 observaciones.
- El objetivo del análisis es determinar si existen diferencias significativas entre los tipos de insecticidas en cuanto a su efectividad para eliminar insectos.

Desarrollo en clase