Diseño completamente al azar

Ejemplo

Suponga que tenemos 4 dietas diferentes que queremos comparar. Las dietas están etiquetadas A,B,C y D.
Estamos interesados en estudiar si las dietas afectan la tasa de coagulación en conejos. La tasa de coagulación es el tiempo en segundos que tarda una cortada en dejar de sangrar.
Tenemos 16 conejos para el experimento, por lo que usaremos 4 en cada dieta.
Los conejos están en una jaula grande hasta que se inicie el experimento, momento en que se transferirán a otras jaulas.

Cómo asignamos los conejos a los cuatro grupos tratamiento?

Método 1

Supongamos que los conejos se atrapan "al azar". Atrapamos cuatro conejos y los asignamos a la dieta A. Atrapamos otros cuatro y los asignamos a la dieta B y así sucesivamente.

Dado que los conejos fueron "atrapados al azar", esto producirá un diseño completamente al azar.

Método 1

No es necesariamente cierto.

Los primeros cuatro conejos atrapados pueden ser los más lentos y débiles, aquellos menos capaces de escapar. Esto puede sesgar los resultados.

Si los resultados del experimento dan desventaja a la dieta A, no habrá forma de determinar si los resultados son a consecuencia de la dieta A o del hecho de haber asignado los conejos más débiles a esa dieta por nuestro "proceso de aleatorización".

Método 2

Atrape a todos los conejos y etiquételos del 1 al 16.

Seleccione cuatro números aleatorios (sin reemplazo) del 1 al 16 y ponga los conejos con esa etiqueta en una jaula que recibirá la dieta A.

Entonces, seleccione otros cuatro números aleatorios y ponga los conejos correspondientes en otra jaula que recibirá la dieta B.

Así sucesivamente hasta tener cuatro jaulas con cuatro conejos en cada una.

Método 2

No hay repeticiones.

El diseño es un diseño completamente al azar pero no tiene repeticiones.

Hay 16 conejos, pero los conejos en cada jaula no son independientes. Si un conejo come mucho, los otros en la jaula tienen menos para comer.

La unidad experimental es la menor unidad de material experimental a la cual se le aplica un tratamiento en forma independiente. En este caso, las jaulas son las unidades experimentales. Para un diseño completamente al azar con repeticiones, cada conejo debe estar en su propia jaula.

Método 3

En una urna ponga las letras A,B,C y D en pedazos de papel separados.

Atrape un conejo, saque un pedazo de papel al azar de la urna y asigne el conejo a la dieta que indique el papel. No reemplace el papel. Atrape el segundo conejo y seleccione al azar otro pedazo de papel de la urna de los tres que quedan. Asigne el conejo a la dieta correspondiente.
Continue hasta que los primeros cuatro conejos sean asignados a una de las cuatro dietas. De esta manera, todos los conejos lentos tienen diferentes dietas.

Coloque otra vez los cuatro pedazos de papel en la urna y repita el procedimiento hasta que los 16 conejos estén asignados a una dieta.

Método 3

Este no es un diseño completamente al azar.

Ya que se seleccionaron los conejos en bloques de 4, y cada uno asignado a una de las dietas, el diseño es el bloques al azar.

El tratamiento es Dieta pero se ha bloqueado a través del grado de "atrapabilidad".

Método 4

Atrape a todos los conejos y márquelos del 1 al 16. Ponga 16 piezas de papel en una urna, con las letras A, B, C y D repetidas cuatro veces cada una.

Ponga otros 16 pedazos de papel numerados del 1 al 16 en otra urna. Tome un pedazo de papel de cada urna. El conejo con el número seleccionado es asignado a la dieta seleccionada.

Para hacer más fácil de recordar cuál conejo tiene cuál dieta, las jaulas se acomodan como se muestra abajo:

A	A	A	A
B	B	B	B
C	C	C	C
D	D	D	D

Método 4

El método 4 tiene algunas deficiencias. La asignación de los conejos a los tratamientos es un diseño completamente al azar. Sin embargo, el arreglo de las jaulas crea un sesgo en los resultados.

Puede haber cambios climáticos y de luz que afecten de forma diferencial a los tratamientos, de tal manera que, cualquier diferencia observada no puede ser atribuida a la dieta, sino que podría ser resultado de la posición de la jaula.

La posición de la jaula no es parte del tratamiento, pero debe ser considerada. En un diseño completamente al azar, todos los conejos tienen la misma probabilidad de recibir cualquier dieta y en cualquier posición de la jaula.

Método 5

Marque las jaulas del 1 al 16.

1	5	9	13
2	6	10	14
3	7	11	15
4	8	12	16

Ponga 16 pedazos de papel en una urna, numerados del 1 al 16. En otra urna ponga 16 pedazos de papel, marcados con las letras A, B C y D.

Atrape un conejo. Seleccione un número y una letra de cada urna. Ponga el conejo en la jaula indicada por el número escogido y asígnelo a la dieta indicada por la letra.

Repita sin reemplazo hasta que todos los conejos hayan sido asignados a una dieta y una jaula.

Método 5

Si, por ejemplo, el primer número seleccionado fué 7 y la primera letra B, entonces el primer conejo se pone en la jaula 7 y se alimenta con la dieta B.

1	5	9	13
2	6	10	14
3	7 **B**	11	15
4	8	12	16

Método 5

Un ejemplo de asignación completa es el siguiente:

1 **C**	5 **A**	9 **B**	13 **D**
2 **D**	6 **B**	10 **D**	14 **C**
3 **C**	7 **B**	11 **A**	15 **D**
4 **A**	8 **A**	12 **C**	16 **B**

Note que el diseño completamente al azar no toma en cuenta las diferencias en la altura de las jaulas. Es solamente una asignación completamente al azar.

En este ejemplo vemos que la mayoría de los conejos con la dieta A están en jaulas de la parte de abajo y los de la dieta D están en la parte superior. Un diseño completamente al azar supone que estas posiciones no producen una diferencia sistemática en la respuesta (tiempo de coagulación).

Si creemos que la posición afecta la respuesta, deberíamos usar un diseño de bloques al azar.

Ejemplo: Disminución del crecimiento de bacterias en carne almacenada.

La vida en estante de carne almacenada es el tiempo en que el corte empacado se mantiene bien, nutritivo y vendible.

El empaque estándar con aire del medio ambiente tiene una vida de 48 horas. Después se deteriora por contaminación bacterial, degradación del color y encogimiento.

El empaque al vacío detiene el crecimiento bacterial, sin embargo, se pierde calidad.

Estudios recientes sugieren que al controlar ciertos gases de la atmósfera se alarga la vida en estante.

Hipótesis de investigación: Algunas formas de gases controlados pueden mejorar la efectividad del empacamiento para carne.

Diseño de tratamientos: Un factor con 4 niveles:

1. Aire ambiental con envoltura plástica
2. Empacado al vacío
3. Mezcla de gases:
   - 1% CO (monóxido de carbono)
   - 40% O_2 (oxígeno)
   - 59% N (nitrógeno)
4. 100% CO_2 (bióxido de carbono)

Diseño experimental: Completamente al azar.

Diseño completamente al azar, un factor

Tres bisteces de res, aproximadamente del mismo tamaño (75 grs.) se asignaron aleatoriamente a cada tratamiento. Cada bistec se empaca separadamente con su condición asignada.

Variable de respuesta: Se mide el número de bacterias psichnotropicas en la carne después de 9 días de almacenamiento a $4°C$.

Estas bacterias se encuentran en la superficie de la carne y aparecen cuando la carne se echó a perder. La medición fué el logaritmo del número de bacterias por cm^2.

Cómo aleatorizar?

Se obtiene una permutación aleatoria de los números 1 a 12. Para esto se toma una secuencia de números de 2 dígitos de una tabla de números aleatorios y se les asigna el rango que les corresponda.
Por ejemplo:

# aleatorio	52	56	20	99	44	34	62	60	31	57	40	78
rango	6	7	1	12	5	3	10	9	2	8	4	11
trat	1	1	1	2	2	2	3	3	3	4	4	4
u.e.	1	2	3	4	5	6	7	8	9	10	11	12
trat	1	3	2	4	2	1	1	4	3	3	4	2

Modelo estadístico para el experimento

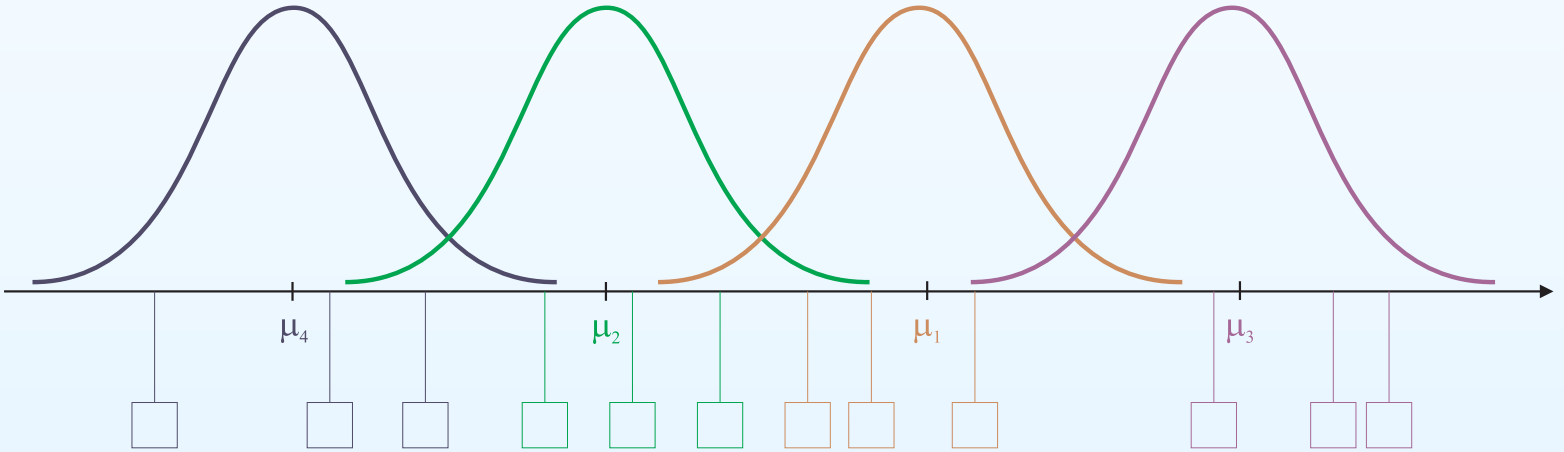
El modelo estadístico para estudios comparativos supone que hay una población de referencia de u.e. En muchos casos la población es conceptual. En el ejemplo, es posible imaginar una población de carne empacada.

Cada unidad de la población tiene un valor de la variable de respuesta, y, la cual tiene media μ y varianza σ^2.

Se supone una población de referencia para cada tratamiento considerado en el estudio, y las variables en el experimento se suponen seleccionadas aleatoriamente de dicha población de referencia, como resultado de la aleatorización.

Nota. Para estudios observacionales, suponemos que las unidades observadas se seleccionaron aleatoriamente de cada una de las poblaciones.

Diseño completamente al azar, un factor



Diseño completamente al azar, un factor

Modelo estadístico lineal para un diseño completamente al azar.

Modelo de medias:

$$y_{ij} = \mu_i + \epsilon_{ij} \quad i = 1, 2, \ldots, t \quad j = 1, 2, \ldots, r$$

donde
y_{ij} es la observación de la j-ésima u.e. del i-ésimo tratamiento,
μ_i es la media del i-ésimo tratamiento,
ϵ_{ij} es el error experimental de la unidad ij.
Suponemos que hay t tratamientos y r repeticiones en cada uno.

En el ejemplo de la carne empacada, tenemos:

Diseño completamente al azar, un factor

bistec	tratamiento	observación	log (conteo/cm^2)	y_{ij}	Modelo
6	1	1	7.66	y_{11}	$\mu_1 + \epsilon_{11}$
7	1	2	6.98	y_{12}	$\mu_1 + \epsilon_{12}$
1	1	3	7.80	y_{13}	$\mu_1 + \epsilon_{13}$
12	2	1	5.26	y_{21}	$\mu_2 + \epsilon_{21}$
5	2	2	5.44	y_{22}	$\mu_2 + \epsilon_{22}$
3	2	3	5.80	y_{23}	$\mu_2 + \epsilon_{23}$
10	3	1	7.41	y_{31}	$\mu_3 + \epsilon_{31}$
9	3	2	7.33	y_{32}	$\mu_3 + \epsilon_{32}$
2	3	3	7.04	y_{33}	$\mu_3 + \epsilon_{33}$
8	4	1	3.51	y_{41}	$\mu_4 + \epsilon_{41}$
4	4	2	2.91	y_{42}	$\mu_4 + \epsilon_{42}$
11	4	3	3.66	y_{43}	$\mu_4 + \epsilon_{43}$

Diseño completamente al azar, un factor

El modelo:

$$y_{ij} = \mu_i + \epsilon_{ij}$$

lo llamaremos **modelo completo** ya que incluye una media separada para cada una de las poblaciones definidas por los tratamientos.

Si no hay diferencia entre las medias de las poblaciones, es decir,

$$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$$

se genera el **modelo reducido**

$$y_{ij} = \mu + \epsilon_{ij}$$

que establece que las observaciones provienen de la misma población con media μ.

Diseño completamente al azar, un factor

El modelo reducido representa la hipótesis de no diferencia entre las medias

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$$

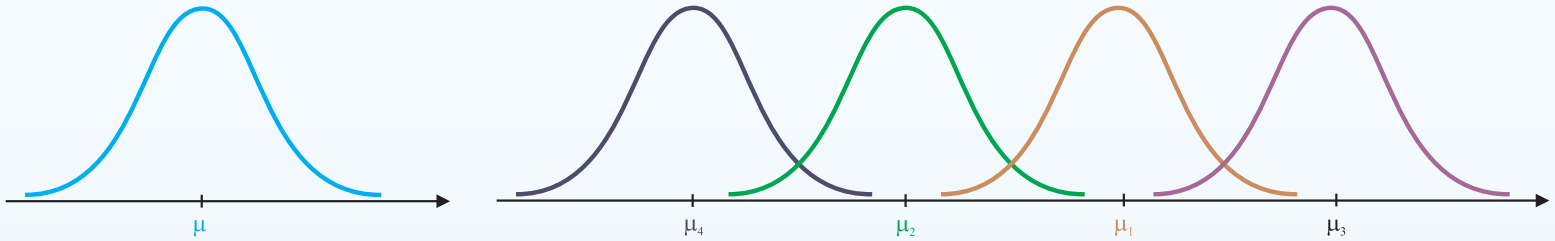
El modelo completo representa la hipótesis alternativa:

$$H_a : \mu_i \neq \mu_k \quad i \neq k$$

El investigador debe determinar cuál de los dos modelos describe mejor a los datos en el experimento.

Diseño completamente al azar, un factor

una población o varias poblaciones



$$y_{ij} = \mu + \epsilon_{ij}$$

$$y_{ij} = \mu_i + \epsilon_{ij}$$

Diseño completamente al azar, un factor

Pregunta de investigación: Hay más crecimiento bacterial con algunos métodos de empacado que con otros?

Pregunta estadística: Cuál modelo describe mejor los resultados del experimento?

Se requiere un método para estimar los parámetros de los dos modelos y con base en algun criterio objetivo determinar cuál modelo o hipótesis estadística se ajusta mejor a los datos del experimento.

Diseño completamente el azar, un factor

Los estimadores de mínimos cuadrados son aquellos que resultan de minimizar la suma de cuadrados de los errores experimentales.

Si los errores experimentales son independientes con media cero y varianzas homogéneas, los estimadores de mínimos cuadrados son insesgados y tienen varianza mínima.

Nota. El muestreo aleatorio en los estudios observacionales y la aleatorización en los experimentales aseguran la suposición de independencia.

Estimadores para el modelo completo

$$
\begin{aligned}
y_{ij} &= \mu_i + \epsilon_{ij} \quad i = 1, \ldots, t \quad j = 1, \ldots, r \\
\epsilon_{ij} &= y_{ij} - \mu_i \\
SSE_c &= \sum_{i=1}^{t} \sum_{j=1}^{r} \epsilon_{ij}^2 = \sum_{i=1}^{t} \sum_{j=1}^{r} (y_{ij} - \mu_i)^2
\end{aligned}
$$

La SSE_c es una medida de qué tan bien se ajusta el modelo a los datos.

Queremos determinar los estimadores $\hat{\mu}_i$ tales que se minimice esta SSE_c.

Vamos a tener t ecuaciones normales, una para cada tratamiento, encontradas a partir de derivar la SSE_c con respecto a cada μ_i e igualarlas a cero.

Estimadores para el modelo completo

Para una i:

$$\frac{\partial}{\partial \mu_i} \sum_{j=1}^{r} (y_{ij} - \mu_i)^2 = -2 \sum_{j=1}^{r} (y_{ij} - \mu_i)$$

igualando a cero

$$-2 \sum_{j=1}^{r} (y_{ij} - \hat{\mu}_i) = 0$$

$$\sum_{j=1}^{r} y_{ij} - r\hat{\mu}_i = 0$$

$$\hat{\mu}_i = \frac{\sum_{j=1}^{r} y_{ij}}{r} = \bar{y}_{i.}$$

Estimadores para el modelo completo

Por lo tanto,

$$\hat{\mu}_i = \bar{y}_i \quad i = 1, \ldots, t$$

Entonces,

$$\begin{aligned} SSE_c &= \sum_{i=1}^{t} \sum_{j=1}^{r} (y_{ij} - \hat{\mu}_i)^2 \\ &= \sum_{i=1}^{t} \sum_{j=1}^{r} (y_{ij} - \bar{y}_{i.})^2 \\ &= \sum_{i=1}^{t} \left[\sum_{j=1}^{r} (y_{ij} - \bar{y}_{i.})^2 \right] \end{aligned}$$

Estimadores para el modelo completo

La varianza muestral del i-ésimo tratamiento es:

$$S_i^2 = \frac{\sum_{j=1}^{r} (y_{ij} - \bar{y}_{i.})^2}{r-1}$$

es una estimador de σ^2 de los datos del i-ésimo grupo.

$$S^2 = \frac{\sum_{i=1}^{t} \left[\sum_{j=1}^{r} (y_{ij} - \bar{y}_{i.})^2\right]}{t(r-1)} = \frac{SSE_c}{t(r-1)}$$

es un **estimador combinado** (pooled) de σ^2 de todos los datos del experimento.

Es un buen estimador si podemos hacer la suposición de que σ^2 es homogénea en todos los grupos.

Estimadores para el modelo completo

Para los datos del ejemplo:

tratamiento	comercial	vacío	mezcla	CO2
	7.66	5.26	7.41	3.51
	6.98	5.44	7.33	2.91
	7.80	5.80	7.04	3.66
$\hat{\mu}_i = \bar{y}_{i.}$	7.48	5.50	7.26	3.36
$\sum_{j=1}^{r} (y_{ij} - \bar{y}_{i.})^2$	0.3848	0.1512	0.0758	0.3150

$$SSE_c = 0.3848 + 0.1512 + 0.0758 + 0.3150 = 0.9268$$

$$S^2 = \frac{SSE_c}{t(r-1)} = \frac{0.9268}{4(2)} = 0.11585$$

Estimadores para el modelo reducido

$$
\begin{aligned}
y_{ij} &= \mu + \epsilon_{ij} \\
\epsilon_{ij} &= y_{ij} - \mu \\
SSE_r &= \sum_{i=1}^{t}\sum_{j=1}^{r} \epsilon_{ij}^2 = \sum_{i=1}^{t}\sum_{j=1}^{r} (y_{ij} - \mu)^2
\end{aligned}
$$

$$
\frac{\partial}{\partial \mu} \sum_{i=1}^{t}\sum_{j=1}^{r} (y_{ij} - \mu)^2 = -2 \sum_{i=1}^{t}\sum_{j=1}^{r} (y_{ij} - \mu)
$$

igualando a cero

$$
\begin{aligned}
\sum_{i=1}^{t}\sum_{j=1}^{r} \hat{\mu} &= \sum_{i=1}^{t}\sum_{j=1}^{r} y_{ij} \\
rt\mu &= y_{..} \\
\hat{\mu} &= \frac{y_{..}}{rt} = \bar{y}_{..}
\end{aligned}
$$

Estimadores para el modelo reducido

Entonces,

$$SSE_r = \sum_{i=1}^{t}\sum_{j=1}^{r}(y_{ij} - \hat{\mu})^2 = \sum_{i=1}^{t}\sum_{j=1}^{r}(y_{ij} - \bar{y}_{..})^2$$

Para el ejemplo,

$$\hat{\mu} = \bar{y}_{..} = \frac{70.80}{12} = 5.90$$

		Modelo reducido $y_{ij} = \mu + \epsilon_{ij}$		Modelo completo $y_{ij} = \mu_i + \epsilon_{ij}$	
	Observado	Estimado	Diferencia	Estimado	Diferencia
Tratamiento	y	$\hat{\mu}$	$(y_{ij} - \hat{\mu})$	$\hat{\mu}_i$	$(y_{ij} - \hat{\mu}_i)$
Comercial	7.66	5.90	1.76	7.48	0.18
	6.98	5.90	1.08	7.48	-0.50
	7.80	5.90	1.90	7.48	0.32
Vacío	5.26	5.90	-0.64	5.50	-0.24
	5.44	5.90	-0.46	5.50	-0.06
	5.80	5.90	-0.10	5.50	0.30
Mezcla	7.41	5.90	1.51	7.26	0.15
	7.33	5.90	1.43	7.26	0.07
	7.04	5.90	1.14	7.26	-0.22
CO2	3.51	5.90	-2.39	3.36	0.15
	2.91	5.90	-2.99	3.36	-0.45
	3.66	5.90	-2.24	3.36	0.30
		$SSE_r = 33.7996$		$SSE_c = 0.9268$	

Diseño completamente al azar, un factor

Siguiendo con el ejemplo:

Modelo completo $y_{ij} = \mu_i + \epsilon_{ij}$ $SSE_c = \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2 = 0.9268$

Modelo reducido $y_{ij} = \mu + \epsilon_{ij}$ $SSE_r = \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 = 33.7996$

Diferencia:

$$SSE_r - SSE_c = \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 - \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$$

haciendo álgebra

$$= \sum_i \sum_j (\bar{y}_{i.} - \bar{y}_{..})^2 = r \sum_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

En el ejemplo: $SSE_r - SSE_c = 32.8728$

Diseño completamente al azar, un factor

$SSE_r - SSE_c = SS_t$ suma de cuadrados de tratamientos.

Representa la **reducción** en SSE al haber incluido tratamientos en el modelo, también se le conoce como **reducción en suma de cuadrados debida a tratamientos.**

Llamaremos $SS_{total} = SSE_r$ ya que es la suma de cuadrados de las diferencias de cada observación y la media general $\bar{y}_{..}$

Entonces, tenemos la partición:

$$\begin{aligned} SS_{total} &= SS_t + SSE_c \\ \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 &= \sum_i \sum_j (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2 \end{aligned}$$

desviación de la observación ij con respecto a la media general	desviación de la media del grupo con respecto a la media general	desviación de la observación ij con respecto a la media de su grupo

Diseño completamente al azar, un factor

$$\begin{aligned}\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 &= \sum_i \sum_j \left[(y_{ij} - \bar{y}_{i.}) + (\bar{y}_{i.} - \bar{y}_{..})\right]^2 \\ &= \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2 + \sum_i \sum_j (\bar{y}_{i.} - \bar{y}_{..})^2 \\ &\quad +2 \sum_i \sum_j (y_{ij} - \bar{y}_{i.})(\bar{y}_{i.} - \bar{y}_{..})\end{aligned}$$

$$\begin{aligned}\sum_i \sum_j (y_{ij} - \bar{y}_{i.})(\bar{y}_{i.} - \bar{y}_{..}) &= \sum_i (\bar{y}_{i.} - \bar{y}_{..}) \sum_j (y_{ij} - \bar{y}_{i.}) \\ &= \sum_i (\bar{y}_{i.} - \bar{y}_{..})(y_{i.} - r\bar{y}_{i.}) = 0\end{aligned}$$

Grados de libertad. Representan el número de piezas de información independientes en las sumas de cuadrados.

En general, es el número de observaciones menos el número de parámetros estimados de los datos.

Sea $n = rt$, el tamaño de muestra total.

Así, $SS_{total} = \sum_i^t \sum_j^r (y_{ij} - \bar{y}_{..})^2$ donde $\bar{y}_{..}$ es el estimador de μ, tiene $n - 1$ g.l.

$SSE = \sum_i^t \sum_j^r (y_{ij} - \bar{y}_{i.})^2$ se estimaron t parámetros $(\mu_1, \mu_2, \ldots, \mu_t)$ por lo tanto tiene $n - t$ g.l.

$SS_t = SS_{total} - SSE = (n-1) - (n-t) = t - 1$ g.l.

Tabla de Análisis de Varianza

ANOVA

F.V.	g.l.	SS	CM
Tratamientos	$t-1$	SS_t	$CM_t = SS_t/t-1$
Error	$n-t$	SSE	$CME = SSE/n-t = \hat{\sigma}^2$
Total	$n-1$	SS_{total}	

Se puede demostrar que:

$$E(CME) = \sigma^2$$

$$E(CM_t) = \sigma^2 + \frac{1}{t-1}\sum_{i=1}^{t} r(\mu_i - \bar{\mu})^2; \quad \bar{\mu} = \sum_i \mu_i/t$$

Tabla de Análisis de Varianza

Si suponemos $\epsilon_{ij} \sim NID(0, \sigma^2) \quad i = 1, \ldots, t \quad j = 1, \ldots, r$
en el modelo completo $y_{ij} = \mu_i + \epsilon_{ij}$

Entonces, $y_{ij} \sim NID(\mu_i, \sigma^2)$.

Se puede demostrar que:

$$\frac{SS_{total}}{\sigma^2} = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$\frac{SSE}{\sigma^2} = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2}{\sigma^2} \sim \chi^2_{n-t}$$

Cuando $H_0 : \mu_1 = \mu_2 = \ldots = \mu_t$ es cierta

$$\frac{SS_t}{\sigma^2} = \frac{\sum_i r(\bar{y}_{i.} - \bar{y}_{..})^2}{\sigma^2} \sim \chi^2_{t-1}$$

Tabla de Análisis de Varianza

Por el Teorema de Cochran (Montgomery, 2001, pág. 69), SS_t y SSE son independientes, por lo tanto cuando H_0 es cierta,

$$F_0 = \frac{SS_t/\sigma^2(t-1)}{SSE/\sigma^2(n-t)} = \frac{CM_t}{CME} \sim F_{t-1,n-t}$$

Además, $E(CM_t) = \sigma^2 + \theta_t^2 = \sigma^2$ cuando $\theta_t^2 = 0$ que es cuando H_0 es cierta. Es decir,

$$\begin{aligned} E(CM_t) &= E(CME) \text{ cuando } H_0 \text{ es cierta} \\ E(CM_t) &> E(CME) \text{ cuando } H_0 \text{ no es cierta} \end{aligned}$$

Entonces, si $CM_t > CME$, o sea, valores grandes de F_0 llevan a rechazar la hipótesis nula $H_0 : \mu_1 = \mu_2 = \ldots = \mu_t$. Por lo tanto, la región de rechazo es:

$$F_0 > F^{\alpha}_{t-1,n-t}$$

Tabla de Análisis de Varianza

ANOVA

F.V.	g.l.	SS	CM	F	$E(CM)$
Tratamientos	$t-1$	SS_t	$CM_t = \frac{SS_t}{t-1}$	$\frac{CM_t}{CME}$	$\sigma^2 + \theta_t^2$
Error	$n-t$	SSE	$CME = \frac{SSE}{n-t}$		σ^2
Total	$n-1$	SS_{total}			

$$SS_t = \sum_{i=1}^{t} r\left(\bar{y}_{i.} - \bar{y}_{..}\right)^2$$

$$SSE = \sum_{i=1}^{t}\sum_{j=1}^{r}\left(y_{ij} - \bar{y}_{i.}\right)^2$$

$$SS_{total} = \sum_{i=1}^{t}\sum_{j=1}^{r}\left(y_{ij} - \bar{y}_{..}\right)^2$$

Tabla de Análisis de Varianza

En el ejemplo de empacado de carne:

F.V.	g.l.	SS	CM	F	$Pr > F$
trat	3	32.8728	10.958	94.55	0.000
error	8	0.9268	0.1159		
total	11	33.7996			

Por lo tanto, se rechaza la hipótesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_4$, es decir, hay algún método de empaque que tiene diferente comportamiento en promedio.

Diseño completamente al azar, un factor

Se quieren comparar t niveles de un factor, lo que implica t tratamientos y se dispone de n_i u.e. para el tratamiento i, $i = 1, \ldots, t$. Hay dos situaciones:

1. Los t tratamientos son escogidos específicamente por el investigador. En esta situación deseamos probar hipótesis acerca de las medias de los tratamientos y nuestras conclusiones se aplicarán solamente a los niveles del factor considerados en el análisis. Las conclusiones **no** se pueden extender a tratamientos similares que no fueron explícitamente considerados. Este es el **modelo de efectos fijos**.
2. Los t tratamientos son una muestra aleatoria de una población de tratamientos. En esta situación nos gustaría poder extender las conclusiones (las cuales están basadas en la muestra de tratamientos considerada) a todos los tratamientos de la población. Este es el **modelo de efectos aleatorios**.

Diseño completamente al azar, un factor

A las cantidades $n_1, n_2, \ldots, n_t$ se les llama **repeticiones** de cada tratamiento.

Si $n_i = r \; \forall i$ se dice que el diseño es **balanceado**.

y_{ij} es la respuesta de la u.e. j del tratamiento i,
$i = 1, \ldots, t \quad j = 1, \ldots, n_i$.

Diseño completamente al azar

Estructura de los datos.

tratamientos					
1	2	3	...	t	
y_{11}	y_{21}	y_{31}	...	y_{t1}	
y_{12}	y_{22}	y_{32}	...	y_{t2}	
y_{13}	y_{23}	y_{33}	...	y_{t3}	
.	.	.	...	.	
.	.	.	...	.	
.	.	.	...	.	
y_{1n_1}	y_{2n_2}	y_{3n_3}	...	y_{tn_t}	
$y_{1.}$	$y_{2.}$	$y_{3.}$	...	$y_{t.}$	totales
$\bar{y}_{1.}$	$\bar{y}_{2.}$	$\bar{y}_{3.}$	...	$\bar{y}_{t.}$	medias

Diseño completamente al azar

$$
\begin{aligned}
n &= \sum_{i=1}^{t} n_i \\
y_{i.} &= \sum_{j=1}^{n_i} y_{ij} \quad i = 1, \ldots, t \text{ total tratamiento } i \\
\bar{y}_{i.} &= \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} \quad i = 1, \ldots, t \text{ media tratamiento } i \\
y_{..} &= \sum_{i=1}^{t} \sum_{j=1}^{n_i} y_{ij} = \sum_{i=1}^{t} y_{i.} \text{ total de las observaciones} \\
\bar{y}_{..} &= \frac{y_{..}}{n} \text{ media general}
\end{aligned}
$$

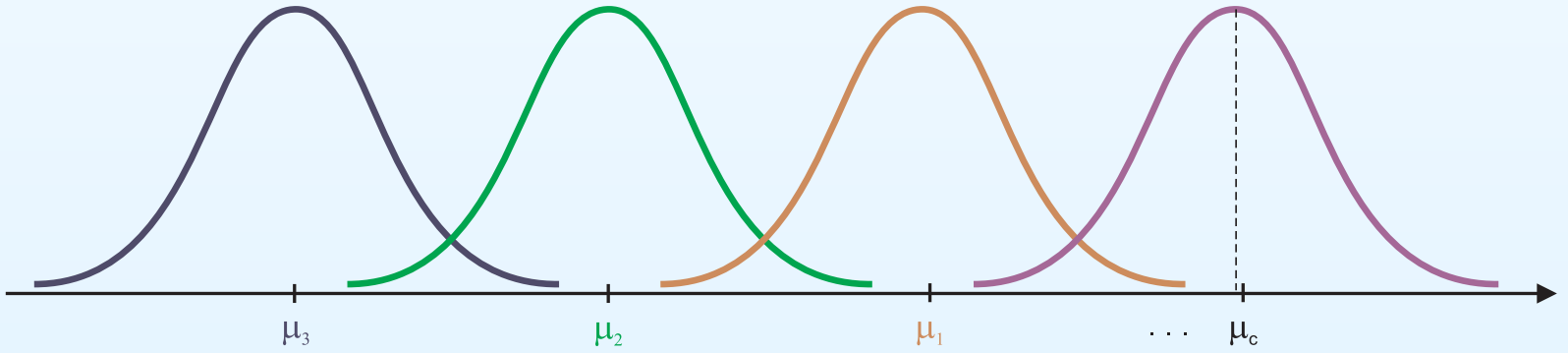
Diseño completamente al azar

Se tienen t muestras aleatorias independientes de tamaños $n_1, n_2, \ldots, n_t$ respectivamente.

$y_{11}, y_{12}, \ldots, y_{1n_1}$ es una muestra aleatoria de $N(\mu_1, \sigma^2)$

$y_{21}, y_{22}, \ldots, y_{2n_2}$ es una muestra aleatoria de $N(\mu_2, \sigma^2)$

$y_{t1}, y_{t2}, \ldots, y_{tn_t}$ es una muestra aleatoria de $N(\mu_t, \sigma^2)$



Diseño completamente al azar

Las observaciones en cada una de estas muestras se pueden representar por el modelo lineal simple

$$y_{ij} = \mu_i + \epsilon_{ij} \quad i = 1, \ldots, t \quad j = 1, \ldots, n_i$$

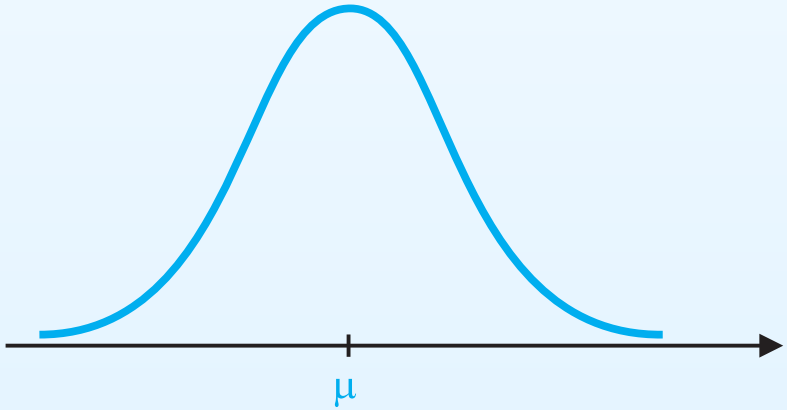
con ϵ_{ij} error experimental en la observación j-ésima del tratamiento i-ésimo.

Estamos suponiendo independencia entre y dentro de las muestras, es decir, ϵ_{ij} son independientes y $\epsilon_{ij} \sim N(0, \sigma^2)$.

Otra forma de verlo

Como suponemos que las u.e. son homogéneas, es decir, el promedio de respuesta de todas las u.e. es el mismo (μ) **antes** de aplicar los tratamientos, y si se observan en condiciones similares, las respuestas las podemos modelar como

$$y_{ij} = \mu + \epsilon_{ij}$$



Entonces al aplicar el tratamiento i-ésimo a un grupo (de tamaño n_i) de u.e. se introduce un efecto (τ_i) de ese tratamiento en las variables por observar.

El modelo se puede escribir como:

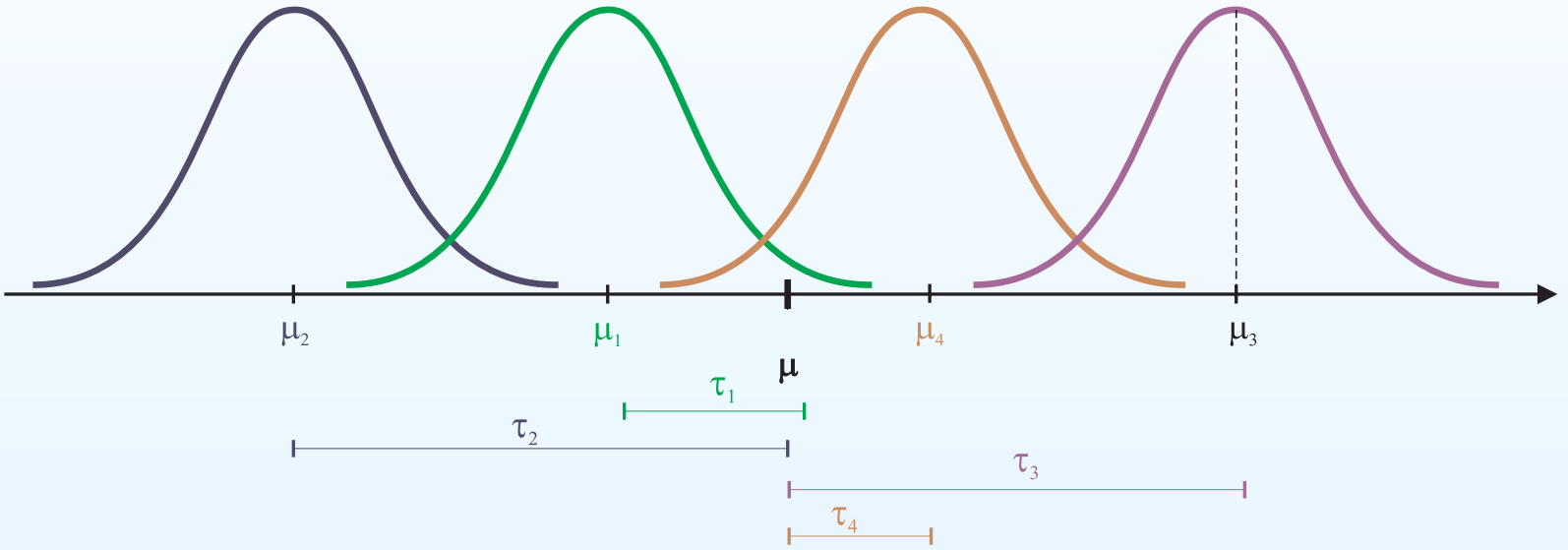
Modelo de efectos

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \quad i = 1, \ldots, t \quad j = 1, \ldots, n_i$$

donde

μ es la media general, común a todas las u.e.
τ_i es el efecto del tratamiento i-ésimo

Modelo de efectos



Modelo de efectos

El modelo de efectos implica que se empieza el experimento con u.e. con la misma capacidad de respuesta (μ) y con la misma varianza (σ^2).

La aplicación de los tratamientos tiene el efecto de alterar las medias, que ahora son $\mu_i = \mu + \tau_i$, pero supone que no se modifican las varianzas.

En este caso, la hipótesis a probar es:

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_t = 0$$

$$H_a : \tau_i \neq 0 \text{ para al menos una } i$$

Modelo de efectos

Estimadores de mínimos cuadrados:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \quad i = 1, \ldots, t \quad j = 1, \ldots, n_i$$

$$SSE = \sum_{i=1}^{t} \sum_{j=1}^{n_i} \epsilon_{ij}^2 = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \mu - \tau_i)^2$$

$$\frac{\partial}{\partial \mu} \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \mu - \tau_i)^2 = -2 \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \mu - \tau_i)$$

$$\frac{\partial}{\partial \tau_i} \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \mu - \tau_i)^2 = -2 \sum_{j=1}^{n_i} (y_{ij} - \mu - \tau_i) \quad i = 1, \ldots, t$$

Modelo de efectos

Igualando a cero:

$$\begin{aligned}
\sum_{i=1}^{t}\sum_{j=1}^{n_i} y_{ij} &= n\hat{\mu} + \sum_{i=1}^{t} n_i\hat{\tau}_i \\
\sum_{j=1}^{n_1} y_{1j} &= n_1\hat{\mu} + n_1\hat{\tau}_1 \\
\sum_{j=1}^{n_2} y_{2j} &= n_2\hat{\mu} + n_2\hat{\tau}_2 \\
\cdots &\quad \cdots \\
\sum_{j=1}^{n_t} y_{tj} &= n_t\hat{\mu} + n_t\hat{\tau}_t
\end{aligned}$$

Las ecuaciones normales no son linealmente independientes, por lo tanto **no** hay una solución única. Esto ocurre porque el modelo de efectos está sobreparametrizado.

Modelo de efectos

Se añade una ecuación linealmente independiente:

a) $\sum_{i=1}^{t} \hat{\tau}_i = 0$

$$\begin{aligned} \hat{\mu} &= \bar{y}_{..} \\ \hat{\tau}_i &= \bar{y}_{i.} - \bar{y}_{..} \quad i = 1, \ldots, t \end{aligned}$$

b) $\hat{\mu} = 0$

$$\begin{aligned} \hat{\mu} &= 0 \\ \hat{\tau}_i &= \bar{y}_{i.} \quad i = 1, \ldots, t \end{aligned}$$

c) $\hat{\tau}_1 = 0$

$$\begin{aligned} \hat{\mu} &= \bar{y}_{1.} \\ \hat{\tau}_i &= \bar{y}_{i.} - \bar{y}_{1.} \quad i = 2, \ldots, t \end{aligned}$$

Modelo de efectos

Hay un número infinito de posibles restricciones que se pueden usar para resolver las ecuaciones normales. Entonces

Cuál usar?

No importa ya que en cualquier caso

$$\widehat{\mu + \tau_i} = \bar{y}_{i.}$$

Aunque no podemos obtener estimadores únicos de los parámetros del modelo de efectos, podemos obtener estimadores únicos de **funciones** de estos parámetros.

A estas funciones se les llama **funciones lineales linealmente estimables.**

Diseño completamente al azar, Tabla de ANOVA

F.V.	g.l.	SS	CM	F	$E(CM)$
Tratamientos	$t-1$	SS_t	$CM_t = \frac{SS_t}{t-1}$	$\frac{CM_t}{CME}$	$\sigma^2 + \frac{\sum_i n_i(\tau_i - \bar{\tau})^2}{t-1}$
Error	$n-t$	SSE	$CME = \frac{SSE}{n-t}$		σ^2
Total	$n-1$	SS_{total}			

$$SS_t = \sum_{i=1}^{t} n_i \left(\bar{y}_{i.} - \bar{y}_{..}\right)^2 = \sum_{i=1}^{t} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{n}$$

$$SSE = \sum_{i=1}^{t}\sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i.}\right)^2 = \sum_{i=1}^{t}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{t} \frac{y_{i.}^2}{n_i}$$

$$SS_{total} = \sum_{i=1}^{t}\sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{..}\right)^2 = \sum_{i=1}^{t}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{n}$$

$$n = \sum_{i=1}^{t} n_i$$

Intervalos de confianza

$$\hat{\mu}_i = \bar{y}_{i.} \quad S^2_{\bar{y}_{i.}} = \frac{S^2}{n_i} \quad \text{con } S^2 = CME = \hat{\sigma}^2 \quad S_{\bar{y}_{i.}} = \sqrt{\frac{CME}{n_i}}$$

Como suponemos que

$$y_{ij} \sim N\left(\mu_i, \sigma^2\right)$$

entonces

$$\bar{y}_{i.} \sim N\left(\mu_i, \sigma^2/n_i\right)$$

como estimamos la varianza:

$$\frac{\bar{y}_{i.} - \mu_i}{S_{\bar{y}_{i.}}} \sim t_{n-t}$$

Por lo tanto, un intervalo del $(1-\alpha)100\%$ de confianza para μ_i es

$$\bar{y}_{i.} \pm t_{n-t}^{1-\alpha/2}\left(S_{\bar{y}_{i.}}\right)$$

Contrastes

En el ejemplo del empacado de carne teníamos:

	Comercial	Al vacío	CO,O2,N	CO2
$\hat{\mu}_i = \bar{y}_{i.}$	7.48	5.50	7.26	3.36

$S^2 = CME = 0.116$ con 8 g.l.

Una vez que rechazamos la hipótesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

Qué sigue?

Contrastes

Se podrían contestar preguntas como:

- Es más efectiva la creación de una atmósfera artificial que el aire ambiente con plástico para reducir el crecimiento de bacterias?
- Son más efectivos los gases que el vacío?
- Es más efectivo el tratamiento de CO2 puro que la mezcla CO,O2 y N?

Un contraste es una función lineal de los parámetros μ_i definido como

$$C = \sum_{i=1}^{t} k_i \mu_i = k_1 \mu_1 + k_2 \mu_2 + \ldots + k_t \mu_t$$

donde $\sum_{i=1}^{t} k_i = 0$.

Contrastes

Los contrastes para las preguntas anteriores son:

- comercial vs. atmósfera artificial

$$C_1 = \mu_1 - \frac{1}{3}(\mu_2 + \mu_3 + \mu_4)$$

- vacío vs. gases

$$C_2 = \mu_2 - \frac{1}{2}(\mu_3 + \mu_4)$$

- mezcla de gases vs. CO2 puro

$$C_3 = \mu_3 - \mu_4$$

Contrastes

El estimador del contraste

$$C = \sum_{i=1}^{t} k_i \mu_i \quad \text{es} \quad \hat{C} = \sum_{i=1}^{t} k_i \hat{\mu}_i = \sum_{i=1}^{t} k_i \bar{y}_{i.}$$

Si suponemos que

$$y_{ij} \sim N\left(\mu_i, \sigma^2\right)$$

entonces

$$\bar{y}_{i.} \sim N\left(\mu_i, \sigma^2/n_i\right)$$

Por lo tanto,

$$\hat{C} = \sum_{i=1}^{t} k_i \bar{y}_{i.} \sim N\left(\sum_{i=1}^{t} k_i \mu_i, \sigma^2 \sum_{i=1}^{t} \frac{k_i^2}{n_i}\right)$$

Contrastes

Ya que:

$$E\left(\sum_{i=1}^{t} k_i \bar{y}_{i.}\right) = \sum_{i=1}^{t} k_i E\left(\bar{y}_{i.}\right) = \sum_{i=1}^{t} k_i \mu_i$$

$$V\left(\sum_{i=1}^{t} k_i \bar{y}_{i.}\right) \underbrace{=}_{m.indep} \sum_{i=1}^{t} k_i^2 V\left(\bar{y}_{i.}\right) = \sum_{i=1}^{t} k_i^2 \frac{\sigma^2}{n_i} = \sigma^2 \sum_{i=1}^{t} \frac{k_i^2}{n_i}$$

$$\hat{V}\left(\hat{C}\right) = \hat{\sigma}^2 \sum_{i=1}^{t} \frac{k_i^2}{n_i} = CME \sum_{i=1}^{t} \frac{k_i^2}{n_i}$$

Contrastes

Entonces,

$$\frac{\sum_{i=1}^{t} k_i \bar{y}_{i.} - \sum_{i=1}^{t} k_i \mu_i}{\sqrt{CME \sum_{i=1}^{t} k_i^2/n_i}} \sim t_{g.l.error}$$

De aquí un intervalo del $100(1-\alpha)\%$ de confianza para el contraste C es:

$$\hat{C} \pm t_{g.l.error}^{1-\alpha/2} \sqrt{CME \sum_{i=1}^{t} k_i^2/n_i}$$

Contrastes

Además,

$$\frac{\hat{C} - C}{\sqrt{\sigma^2 \sum_{i=1}^{t} k_i^2/n_i}} \sim N(0,1)$$

Si $H_0 : \sum_{i=1}^{t} k_i \mu_i = 0$, es decir, $H_0 : C = 0$ es cierta, entonces,

$$\frac{\hat{C}^2}{\sigma^2 \sum_{i=1}^{t} k_i^2/n_i} \sim \chi_1^2$$

Sea

$$SS_c = \frac{\hat{C}^2}{\sum_{i=1}^{t} k_i^2/n_i}$$

entonces

$$\frac{SS_c/\sigma^2}{SSE/\sigma^2(n-t)} = \frac{\hat{C}^2/\sum_{i=1}^{t} k_i^2/n_i}{CME} \sim F_{1,n-t}$$

Por lo tanto, para probar $H_0 : C = 0$ se rechaza si $F_c > F_{1,n-t}^{\alpha}$

Contrastes

El número de contrastes que se pueden hacer es muy grande, sin embargo, esta técnica tiene su mayor utilidad cuando se aplica a comparaciones planeadas antes de realizar el experimento.

Una clase de contrastes, conocida como **Contrastes ortogonales** (como son los del ejemplo anterior) tienen propiedades especiales con respecto a la partición de sumas de cuadrados y grados de libertad y con respecto a su relación entre ellos. La ortogonalidad implica que un contraste no aporta información acerca de otro.

Dos contrastes, con coeficientes $\{k_i\}, \{l_i\}$ son **ortogonales** si

$$\sum_{i=1}^{t} \frac{k_i l_i}{n_i} = 0$$

Contrastes

Para t tratamientos existe un conjunto de $t-1$ contrastes ortogonales, los cuales hacen una partición de la suma de cuadrados de tratamientos en $t-1$ componentes independientes, cada uno con 1 g.l. Por lo tanto las pruebas realizadas con contrastes ortogonales son independientes.

En el ejemplo anterior, los contrastes son ortogonales.

	k_1	k_2	k_3	k_4
C_1	1	-1/3	-1/3	-1/3
C_2	0	1	-1/2	-1/2
C_3	0	0	1	-1

ANOVA

La tabla de ANOVA incorporando las pruebas de hipótesis de los 3 contrastes es:

F.V.	g.l.	SS	CM	F	$Pr > F$
trat	3	32.8728	10.958	94.55	0.000
C_1	1	10.01	10.01	86.29	0.000
C_2	1	0.07	0.07	0.62	0.453
C_3	1	22.82	22.82	196.94	0.000
error	8	0.9268	0.1159		
total	11	33.7996			

Se rechaza $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
Se rechaza $H_{01} : \mu_1 = \frac{1}{3}(\mu_2 + \mu_3 + \mu_4)$
No se rechaza $H_{02} : \mu_2 = \frac{1}{2}(\mu_3 + \mu_4)$
Se rechaza $H_{03} : \mu_3 = \mu_4$

$$SS_{C1} = \frac{\hat{C_1}^2}{\frac{1}{r}\sum_{i=1}^{4} k_i^2} = \frac{(2.11)^2}{\frac{1^2+3(-1/3)^2}{3}} = \frac{4.4521}{0.4444} = 10.01$$

Otro ejemplo

Los siguientes datos son los tiempos de coagulación de sangre para 24 animales que fueron aleatoriamente asignados a una de cuatro dietas (A,B,C,D)

Dieta A	Dieta B	Dieta C	Dieta D
62	63	68	56
60	67	66	62
63	71	71	60
59	64	67	61
	65	68	63
	66	68	64
			63
			59

Otro ejemplo

- Pruebe la hipótesis de igualdad de medias
  $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$.
- Pruebe el siguiente contraste: (pendiente)

  El promedio de la dieta A y B es igual al promedio de la C y D

El análisis en R:

- Los datos están en el archivo *coag.txt*
- El programa está en *anova_coag.txt*

Comparaciones múltiples

En muchas situaciones prácticas, se desea comparar pares de medias. Podemos determinar cuáles medias difieren probando las diferencias entre **todos** los pares de medias de tratamientos.

Es decir, estamos interesados en contrastes de la forma

$$\Gamma = \mu_i - \mu_j \quad \forall i \neq j$$

Lo primero que se nos viene a la mente es hacer una prueba t para cada par de medias, es decir, probar

$$\begin{aligned} H_0 : \mu_i &= \mu_j \\ H_a : \mu_i &\neq \mu_j \quad \forall i \neq j \end{aligned}$$

Comparaciones múltiples

Si suponemos varianzas iguales, se tiene la estadística de prueba

$$t_c = \frac{\bar{y}_{i.} - \bar{y}_{j.}}{s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}$$

y se rechaza H_0 al nivel de significancia α si

$$t_c \leq t^{\alpha/2}_{n_i+n_j-2} \quad \text{ó} \quad t_c \geq t^{1-\alpha/2}_{n_i+n_j-2}$$

Esto es equivalente a decir que se rechaza H_0 si

$$|t_c| = \frac{|\bar{y}_{i.} - \bar{y}_{j.}|}{s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} > t^{1-\alpha/2}_{n_i+n_j-2}$$

o equivalente a

$$|\bar{y}_{i.} - \bar{y}_{j.}| > t^{1-\alpha/2}_{n_i+n_j-2}\ s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

Comparaciones múltiples

Esta prueba conocida como Diferencia Mínima Significativa (DMS ó LSD) en el contexto de ANOVA, lo que hace es comparar el valor absoluto de la diferencia de cada par de medias con DMS:
Si

$$|\bar{y}_{i.} - \bar{y}_{j.}| > DMS = t_{glerror}^{1-\alpha/2} \sqrt{CME\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

se rechaza $H_0 : \mu_i = \mu_j$.

CME es el cuadrado medio del error que es una estimación ponderada de la varianza basada en t estimaciones de la varianza.

El utilizar este procedimiento no es conveniente por que el nivel de significancia global, es decir, para el conjunto de todas las pruebas, resulta muy superior al nivel de significancia (α) planeado.

Comparaciones múltiples

Por ejemplo, si se tienen 4 medias de tratamientos, entonces se tienen

$$\binom{4}{2} = \frac{4!}{2!2!} = 6$$

pares a comparar, es decir, 6 pruebas de hipótesis a realizar, con lo que se pueden cometer $0, 1, 2, 3, 4, 5,$ ó 6 errores Tipo I, si todas las medias son iguales.

Se define otra forma de error tipo I basado en los riesgos acumulados asociados a la familia de pruebas bajo consideración.

Este es el **error tipo I del experimento** α_E que es el riesgo de cometer el error tipo I al menos una vez.

La probabilidad de error tipo I del experimento puede evaluarse para una familia de pruebas independientes.

Comparaciones múltiples

Sin embargo, todas las pruebas a pares usando la DMS no son independientes, puesto que el CME es el mismo en cada una de las estadísticas de prueba y el numerador contiene las mismas medias en varias de las estadísticas de prueba.

Aún así, se puede evaluar el límite superior de la probabilidad de error tipo I del experimento, suponiendo n pruebas independientes.

Suponga que la H_0 es cierta para cada una de las $n = \binom{t}{2}$ pruebas y que son independientes.

Sea $\alpha_c = P(\text{error tipo I})$ en una sola prueba (comparación) con $(1 - \alpha_c) = P(\text{decisión correcta})$.

Comparaciones múltiples

La probabilidad de cometer x errores tipo I está dada por la distribución binomial como:

$$P(X = x) = \binom{n}{x} \alpha_c^x (1 - \alpha_c)^{n-x}$$

$$P(X = x) = \frac{n!}{(n-x)!x!} \alpha_c^x (1 - \alpha_c)^{n-x} \quad x = 0, 1, 2, \ldots, n$$

La probabilidad de no cometer ningún error tipo I es

$$P(X = 0) = (1 - \alpha_c)^n$$

Comparaciones múltiples

La probabilidad de cometer al menos 1 error tipo I es

$$P(X \geq 1) = 1 - P(X = 0) = 1 - (1 - \alpha_c)^n$$

es decir, la máxima probabilidad de cometer al menos un error tipo I entre las n comparaciones es:

$$\begin{aligned} \alpha_E &= 1 - (1 - \alpha_c)^n \quad \text{de aquí} \\ \alpha_c &= 1 - (1 - \alpha_E)^{1/n} \end{aligned}$$

Comparaciones múltiples

# de pruebas indep. n	α_E cuando $\alpha_c = 0.05$	α_c cuando $\alpha_E = 0.05$
1	0.05	0.05
2	0.098	0.025
3	0.143	0.017
4	0.185	0.013
5	0.226	0.010
10	0.401	0.005

Por el razonamiento anterior es que han surgido una serie de pruebas de diferentes autores para hacer comparaciones múltiples tratando de mantener la

$$P(\text{error tipo I del experimento}) = \alpha$$

Bonferroni

$$\alpha_E \leq n\alpha_c$$

n comparaciones, la igualdad se dá cuando las pruebas son independientes.

Entonces,

$$\alpha_c = \alpha_E/n$$

Si queremos $\alpha_E = 0.05$ entonces, $\alpha_c = 0.05/n$ y se hacen las pruebas t para los pares de medias con un nivel de significancia α_c en cada una de ellas.

Tukey

Conocida como la prueba de la Diferencia Mínima Significativa Honesta (DMSH)

$$DMSH = q^{\alpha}_{t,glerror}\sqrt{\frac{CME}{r}} \quad \text{si } n_i = r \quad \forall i$$

$$DMSH = q^{\alpha}_{t,glerror}\sqrt{\frac{CME}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

Si $|\bar{y}_{i.} - \bar{y}_{j.}| > DMSH$ se rechaza $H_0 : \mu_i = \mu_j$.

$q^{\alpha}_{\nu_1,\nu_2}$ se obtiene de las "tablas de rangos estudentizados".

Tukey

Para el ejemplo del empaque de carne:

	Comercial	Al vacío	CO,O2,N	CO2
$\bar{y}_{i.}$	7.48	5.50	7.26	3.36

$S^2 = CME = 0.116$ con $8 g.l.$ $t = 4$, $r = 3$

$$DMSH = q_{4,8}^{0.05}\sqrt{\frac{0.116}{3}} = (4.53)(0.197) = 0.891$$

$$\begin{aligned}
|\bar{y}_{1.} - \bar{y}_{2.}| &= 1.98^{**} \\
|\bar{y}_{1.} - \bar{y}_{3.}| &= 0.22 \\
|\bar{y}_{1.} - \bar{y}_{4.}| &= 4.12^{**} \\
|\bar{y}_{2.} - \bar{y}_{3.}| &= 1.76^{**} \\
|\bar{y}_{2.} - \bar{y}_{4.}| &= 2.14^{**} \\
|\bar{y}_{3.} - \bar{y}_{4.}| &= 3.90^{**}
\end{aligned}$$

Student-Newman-Keuls (SNK)

Se calcula un conjunto de valores críticos

$$k_p = q^{\alpha}_{p,f} S_{\bar{y}_{i.}} \quad p = 2, 3, \ldots, t$$

donde $q^{\alpha}_{p,f}$ es el percentil $1 - \alpha$ de la distribución del rango estudentizado para el número p de medias involucradas en la comparación y f g.l. del error, y $S_{\bar{y}_{i.}} = \sqrt{\frac{CME}{r}}$

Para el ejemplo de la carne empacada:

p	2	3	4
$q^{.05}_{p,8}$	3.26	4.04	4.53
k_p	0.642	0.796	0.892

Student-Newman-Keuls (SNK)

	Comercial	Al vacío	CO,O2,N	CO2
$\bar{y}_{i.}$	7.48	5.50	7.26	3.36

Medias ordenadas:

$$\bar{y}_{4.} = 3.36 \quad \bar{y}_{2.} = 5.50 \quad \bar{y}_{3.} = 7.26 \quad \bar{y}_{1.} = 7.48$$

$$\begin{aligned}
|\bar{y}_{4.} - \bar{y}_{1.}| &= 4.12 > k_4^{**} \\
|\bar{y}_{4.} - \bar{y}_{3.}| &= 3.90 > k_3^{**} \\
|\bar{y}_{4.} - \bar{y}_{2.}| &= 2.14 > k_2^{**} \\
|\bar{y}_{2.} - \bar{y}_{1.}| &= 1.98 > k_3^{**} \\
|\bar{y}_{2.} - \bar{y}_{3.}| &= 1.76 > k_2^{**} \\
|\bar{y}_{3.} - \bar{y}_{1.}| &= 0.22 < k_2(N.S.)
\end{aligned}$$

Duncan

Es similar a la de SNK. Los promedios de los t tratamientos se ordenan en forma ascendente y el error estándar de cada promedio se determina con

$$S_{\bar{y}_{i.}} = \sqrt{\frac{CME}{r}} \quad \text{si } n_i = r \quad \forall i$$

Para muestras de diferente tamaño, se reemplaza la r por la media armónica (n_h) de los $\{n_i\}$

$$n_h = \frac{t}{\sum_{i=1}^{t} \left(\frac{1}{n_i}\right)}$$

Duncan

De las tablas de Duncan de rangos significativos se obtienen los valores de $r^{\alpha}_{p,f}$ para $p = 2, 3, \ldots, t$.
p es el número de medias involucradas en la comparación, α es el nivel de significancia y f los grados de libertad del error.

Se calculan

$$R_p = r^{\alpha}_{p,f} S_{\bar{y}_{i.}} \quad p = 2, 3, \ldots, t$$

Para el ejemplo de la carne empacada:

p	2	3	4
$r^{.05}_{p,8}$	3.26	3.39	3.47
R_p	0.642	0.668	0.684

Duncan

	Comercial	Al vacío	CO,O2,N	CO2
$\bar{y}_{i.}$	7.48	5.50	7.26	3.36

Medias ordenadas:

$$\bar{y}_{4.} = 3.36 \quad \bar{y}_{2.} = 5.50 \quad \bar{y}_{3.} = 7.26 \quad \bar{y}_{1.} = 7.48$$

$$\begin{aligned}
|\bar{y}_{4.} - \bar{y}_{1.}| &= 4.12 > R_4^{**} \\
|\bar{y}_{4.} - \bar{y}_{3.}| &= 3.90 > R_3^{**} \\
|\bar{y}_{4.} - \bar{y}_{2.}| &= 2.14 > R_2^{**} \\
|\bar{y}_{2.} - \bar{y}_{1.}| &= 1.98 > R_3^{**} \\
|\bar{y}_{2.} - \bar{y}_{3.}| &= 1.76 > R_2^{**} \\
|\bar{y}_{3.} - \bar{y}_{1.}| &= 0.22 < R_2(N.S.)
\end{aligned}$$

Dunnett

Para comparar las medias de los tratamientos con la media del tratamiento control.
Suponga que el tratamiento t es el control, queremos probar las hipótesis

$$\begin{aligned} H_0 : \mu_i &= \mu_t \\ H_a : \mu_i &\neq \mu_t \quad i = 1, 2, \ldots, t-1 \end{aligned}$$

$H_0 : \mu_i = \mu_t$ se rechaza si

$$|\bar{y}_{i.} - \bar{y}_{t.}| > D = d_\alpha(t-1, glerror)\sqrt{\frac{CME}{r}}$$

con $d_\alpha(k, \nu)$ es el percentil $1-\alpha$ de las tablas de Dunnett.
Para el ejemplo de la carne empacada, el tratamiento 1 es el control.

	Comercial	Al vacío	CO,O2,N	CO2
$\bar{y}_{i.}$	7.48	5.50	7.26	3.36

Dunnett

$$d_{0.05,3,8} = 2.42$$

$$D = 2.42\left(\sqrt{\frac{CME}{r}}\right) = 0.477$$

$$\begin{aligned} |\bar{y}_{2.} - \bar{y}_{1.}| &= 1.98 > D^{**} \\ |\bar{y}_{3.} - \bar{y}_{1.}| &= 0.22 < D(N.S.) \\ |\bar{y}_{4.} - \bar{y}_{1.}| &= 4.12 > D^{**} \end{aligned}$$

Scheffé

Scheffé (1953) propuso un método para probar **todos** los posibles contrastes.

Considere cualquier contraste

$$C = \sum_{i=1}^{t} k_i \mu_i \quad \text{estimado con } \hat{C} = \sum_{i=1}^{t} k_i \bar{y}_{i.}$$

con error estándar

$$S_C = \sqrt{CME \left[\sum_{i=1}^{t} \frac{k_i^2}{n_i} \right]}$$

La hipótesis nula pra el contraste $H_0 : C = 0$ se rechaza si

$$|C| > S(\alpha_E)$$

donde

$$S(\alpha_E) = S_C \sqrt{(t-1) F^{\alpha_E}_{t-1, g.l.error}}$$

Análisis de residuales

Tenemos el modelo

$$y_{ij} = \mu_i + \epsilon_{ij} \quad \text{ó} \quad y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

$$\epsilon_{ij} \sim NID\left(0, \sigma^2\right)$$

Suposiciones:

- errores normales
- independientes
- varianza constante

La prueba F del análisis de varianza es robusta a falta de normalidad.

Análisis de residuales

Si los errores experimentales están correlacionados, el error estándar estará mal estimado. La independencia se justifica aleatorizando las u.e. a los tratamientos en experimentos y seleccionando muestras aleatorias en estudios observacionales.

Si no hay homogeneidad de varianzas el estimador de σ^2 es malo, aunque se ha visto en estudios que si el diseño es balanceado no efecta mucho. También si los tamaños de muestra mayores corresponden a las poblaciones con mayor varianza.

Análisis de residuales, Normalidad

Residuales

$$\begin{aligned} e_{ij} &= y_{ij} - \hat{y}_{ij} \\ \hat{y}_{ij} &= \widehat{\mu + \tau_i} = \hat{\mu}_i = \bar{y}_{i.} \\ e_{ij} &= y_{ij} - \bar{y}_{i.} \end{aligned}$$

- Prueba no parámetrica (Kolmogorov-Smirnov)
- Histograma (muestras grandes)
- gráfica en papel normal
- análisis de residuales estandarizados para detectar outliers.
  Si $\epsilon_{ij} \sim N(0, \sigma^2)$ entonces $\frac{\epsilon_{ij}-0}{\sigma} \sim N(0,1)$. Sean
  $d_{ij} = \frac{e_{ij}}{\sqrt{CME}}$, esperamos que:
  68% de los residuales estandarizados estén entre -1 y 1
  95 % estén entre -2 y 2
  Virtualmente todos estén entre -3 y 3.

Prueba de Bartlett

$$H_0: \quad \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_t^2$$
$$H_a: \quad \text{no } H_0$$

Estadística de Prueba:

$$U = \frac{1}{C}\left[(n-t)ln(\hat{\sigma}^2) - \sum_i (n_i - 1)ln(\hat{\sigma}_i^2)\right]$$

donde $\hat{\sigma}^2 = \sum_i \frac{(n_i-1)\hat{\sigma}_i^2}{n-t}$ $\hat{\sigma}_i^2 = \sum_j \frac{(y_{ij} - \bar{y}_{i.})^2}{n_i - 1}$

$$C = 1 + \frac{1}{3(t-1)}\left(\sum_i \frac{1}{n_i - 1} - \frac{1}{n-t}\right)$$

H_0 se rechaza si $U > \chi^2_{\alpha,t-1}$ (prueba sensible a falta de normalidad)

Prueba de Levene

Se calcula

$$d_{ij} = |y_{ij} - \tilde{y}_{i.}| \quad i = 1, \ldots, t \quad j = 1, \ldots, n_i$$

donde $\tilde{y}_{i.}$ es la mediana de las observaciones en el tratamiento i.

Se evalúa si el promedio de estas observaciones d_{ij} es igual para todos los tratamientos, es decir, se hace un ANOVA para probar igualdad de medias de d_{ij}.

Prueba de Welch

La prueba F usual es robusta ante heteroscedasticidad (varianzas diferentes) si los tamaños de muestra son muy parecidos o, si los tamaños de muestra más grandes corresponden a las poblaciones con varianzas más grandes.

Sin embargo, se han construído algunas procedimientos de prueba de igualdad de medias $(H_0 : \mu_1 = \mu_2 = \ldots = \mu_t)$ como por ejemplo el desarrollado por Welch, conocido como la prueba de Welch, utilizada cuando no hay homoscedasticidad.

Sean $W_i = n_i/\hat{\sigma}_i^2 \quad \bar{y}^* = \sum_i W_i \bar{y}_{i.} / \sum_i W_i$ y

$$\Lambda = \sum_i \frac{(1 - W_i/W_.)^2}{n_i - 1}$$

donde $W_. = \sum_i W_i$.

Prueba de Welch

Entonces

$$F_c = \frac{\sum_i W_i \frac{(\bar{y}_{i.} - \bar{y}^*)^2}{t-1}}{1 + 2(t-2)\Lambda/(t^2-1)}$$

tiene aproximadamente una distribución F con
$\nu_1 = t - 1 \quad y \quad \nu_2 = (t^2 - 1)/3\Lambda$ grados de libertad.

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_t$ se rechaza al nivel de significancia α si

$$F_c > F^{\alpha}_{\nu_1, \nu_2}.$$

Transformaciones

Se utilizan las transformaciones para cambiar la escala de las observaciones para que se cumplan las suposiciones del modelo lineal y dar inferencias válidas del análisis de varianza.

Cuando las transformaciones son necesarias, se hace el análisis y se hacen las inferencias en la escala transformada pero se presentan tablas de medias en la escala de medición original.

1. Distribución Poisson. Mediciones que son conteos (número de plantas en cierta área, insectos en plantas, accidentes por unidad de tiempo) tienen distribución Poisson.

La transformación $x = \sqrt{y + a}, \quad a \in \Re$ es la adecuada.

2. Distribución binomial. Observaciones del número de éxitos en n ensayos independientes tiene distribución binomial (proporción de semillas germinadas, proporción de plantas con flores en un transecto). $\hat{\pi} = y/n$

La transformación $x = sin^{-1}\sqrt{\hat{\pi}}$ es la adecuada.

Las **transformaciones del tipo potencia** alteran la simetría o asimetría de las distribuciones de las observaciones.

Si suponemos que la desviación estándar de y es proporcional a alguna potencia de la media, es decir,

$$\sigma_y \propto \mu^{\beta}$$

Una transformación de las observaciones, del estilo:

$$x = y^p$$

Transformaciones

Da una relación

$$\sigma_x \propto \mu^{p+\beta-1}$$

Si $p = 1 - \beta$ entonces la desviación estándar de la variable transformada x será constante, ya que $p + \beta - 1 = 0$ y $\sigma_x \propto \mu^0$.

La transformación de Box-Cox

$$x = (y^p - 1)/p \quad p \neq 1$$

$$x = log_e y \quad p = 1$$

El estimador de p se encuentra maximizando

$$L(p) = -\frac{1}{2} log_e \left[CME(p)\right]$$

donde $CME(p)$ es el cuadrado medio del error del análisis de varianza usando la transformación $x = (y^p - 1)/p$ para el valor dado p.

Transformaciones

Se determina $CME(p)$ para un conjunto de valores de p, se grafica $CME(p)$ vs. p y se toma el valor de p que corresponde al valor mínimo de $CME(p)$.

JMP calcula la transformación de Box-Cox, da una gráfica de p vs. CME y da la opción de guardar los datos transformados en el archivo.

La dificultad de utilizar esta transformación es la interpretación.

Ejemplo

Los siguientes datos son el número de errores en un examen de sujetos bajo la influencia de dos drogas. El grupo 1 es un grupo control (sin droga), a los sujetos del grupo 2 se les dió la droga 1, a los del grupo 3 la droga 2 y a los del grupo 4 las dos drogas.

Grupo 1 (sin droga)	Grupo 2 (droga 1)	Grupo 3 (droga 2)	Grupo 4 (dos drogas)
1	12	12	13
8	10	4	14
9	13	11	14
9	13	7	17
4	12	8	11
1	10	10	14
1		12	13
		5	14

Ejemplo

Correr el ejemplo con SPSS y JMP.

1. Probar homogeneidad de varianzas. (Bartlett y Levene)
2. Hacer prueba de Welch
3. Probar con algunas transformaciones, checando normalidad y homogeneidad de varianzas

ej2_1_messy.sav
ej2_1_messy.jmp
ej2_1_messy.txt

Relación entre Regresión y ANOVA

Cualquier modelo de ANOVA se puede escribir como un modelo de regresión lineal.

Suponga el ejemplo de la carne empacada

tratamiento	comercial	vacío	mezcla	CO2
	7.66	5.26	7.41	3.51
	6.98	5.44	7.33	2.91
	7.80	5.80	7.04	3.66

Un diseño completamente al azar con un solo factor (método de empacado) con 4 niveles (4 tratamientos) y 3 repeticiones en cada tratamiento (diseño balanceado).

Relación entre Regresión y ANOVA

Modelo ANOVA completamente al azar un solo factor balanceado:

$$y_{ij} = \mu_i + \epsilon_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, 2, 3, 4 \\ j = 1, 2, 3 \end{cases}$$

El modelo de regresión equivalente es:

$$y_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \epsilon_{ij} \begin{cases} i = 1, 2, 3, 4 \\ j = 1, 2, 3 \end{cases}$$

Relación entre Regresión y ANOVA

Donde las variables x_{1j}, x_{2j}, x_{3j} están definidas como:

$$x_{1j} = \begin{cases} 1 & \text{si la observación } j \text{ es del tratamiento 1} \\ 0 & \text{en otro caso} \end{cases}$$

$$x_{2j} = \begin{cases} 1 & \text{si la observación } j \text{ es del tratamiento 2} \\ 0 & \text{en otro caso} \end{cases}$$

$$x_{3j} = \begin{cases} 1 & \text{si la observación } j \text{ es del tratamiento 3} \\ 0 & \text{en otro caso} \end{cases}$$

Relación entre Regresión y ANOVA

La relación entre los parámetros del modelo ANOVA y el modelo de regresión es:

Si la observación viene del tratamiento 1, entonces $x_{1j} = 1, x_{2j} = 0, x_{3j} = 0$ y el modelo de regresión es

$$\begin{aligned} y_{1j} &= \beta_0 + \beta_1(1) + \beta_2(0) + \beta_3(0) + \epsilon_{1j} \\ &= \beta_0 + \beta_1 + \epsilon_{1j} \end{aligned}$$

y el modelo ANOVA es:

$$y_{1j} = \mu_1 + \epsilon_{1j} = \mu + \tau_1 + \epsilon_{1j}$$

Por lo tanto:

$$\beta_0 + \beta_1 = \mu_1 = \mu + \tau_1$$

Relación entre Regresión y ANOVA

Similarmente, para las observaciones del tratamiento 2

$$\begin{aligned} y_{2j} &= \beta_0 + \beta_1(0) + \beta_2(1) + \beta_3(0) + \epsilon_{2j} \\ &= \beta_0 + \beta_2 + \epsilon_{2j} \end{aligned}$$

y la relación entre los parámetros es:

$$\beta_o + \beta_2 = \mu_2 = \mu + \tau_2$$

Lo mismo para las observaciones del tratamiento 3

$$\begin{aligned} y_{3j} &= \beta_0 + \beta_1(0) + \beta_2(0) + \beta_3(1) + \epsilon_{3j} \\ &= \beta_0 + \beta_3 + \epsilon_{3j} \end{aligned}$$

y la relación entre los parámetros es:

$$\beta_o + \beta_3 = \mu_3 = \mu + \tau_3$$

Relación entre Regresión y ANOVA

Finalmente, considere las observaciones del tratamiento 4, para las cuales el modelo de regresión es:

$$\begin{aligned} y_{4j} &= \beta_0 + \beta_1(0) + \beta_2(0) + \beta_3(0) + \epsilon_{4j} \\ &= \beta_0 + \epsilon_{4j} \end{aligned}$$

entonces $\beta_0 = \mu_4 = \mu + \tau_4$

Por lo tanto,

$$\begin{aligned} \beta_0 &= \mu_4 \\ \beta_1 &= \mu_1 - \mu_4 \\ \beta_2 &= \mu_2 - \mu_4 \\ \beta_3 &= \mu_3 - \mu_4 \end{aligned}$$

Relación entre Regresión y ANOVA

Entonces, para probar la hipótesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ tendríamos que probar $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, lo cual se puede hacer con cualquier paquete de cómputo estadístico.
Para el ejemplo de la carne empacada:

tratamiento	y	x_1	x_2	x_3
1	7.66	1	0	0
1	6.98	1	0	0
1	7.80	1	0	0
2	5.26	0	1	0
2	5.44	0	1	0
2	5.80	0	1	0
3	7.41	0	0	1
3	7.33	0	0	1
3	7.04	0	0	1
4	3.51	0	0	0
4	2.91	0	0	0
4	3.66	0	0	0

Relación entre Regresión y ANOVA

Si pedimos una regresión $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$ y pedimos una tabla de análisis de varianza del modelo $y_{ij} = \mu + \tau_i + \epsilon_{ij}$ las dos tablas ANOVA son idénticas.