Caso de Estudio 1

Problema de una Fundidora

Una fundidora de acero surte de **láminas de hojalata** a tres fabricantes de latas con la siguiente especificación:

´´Peso del revestimineto de estaño debe ser al menos de $0.25\ lb$ en el fondo del envase de hojalata (**Y**).”

Algunas consideraciones:

1. La fundidora y cada uno de los fabricantes de latas tienen laboratorios donde realizan las **mediciones de los pesos de estaño** a partir de muestras de cada cargamento.
2. Se ha observado desacuerdo sobre los pesos reales.
3. Se justifica la decisión de **PLANEAR UN EXPERIMENTO** para determinar si los laboratorios están realizando mediciones consistentes.
4. Existe un elemento a considerar relacionado con el uso de elementos químicos para eliminar el estaño del metal. No se tienen las mismas condiciones en las muestras de cada laboratorio.

Caso de Estudio 1

Continuación

Una solución:

Se envían varias **muestras** (UE: discos circulares de igual área) a cada **laboratorio** bajo el supuesto que *en promedio el peso de revestimiento,* **Y***, a pesar de ser posiblemente diferente, que estas diferencias no sean muy significativas*.

Es decir, las diferencias en los pesos serán atribuibles a diferencias sistemáticas en las técnicas de medición y a la variabilidad aleatoria.

Lo anterior permite determinar si los resultados son consistentes al comparar la variabilidad de las medias de las muestras enviadas.

Caso de Estudio 1

Cont.

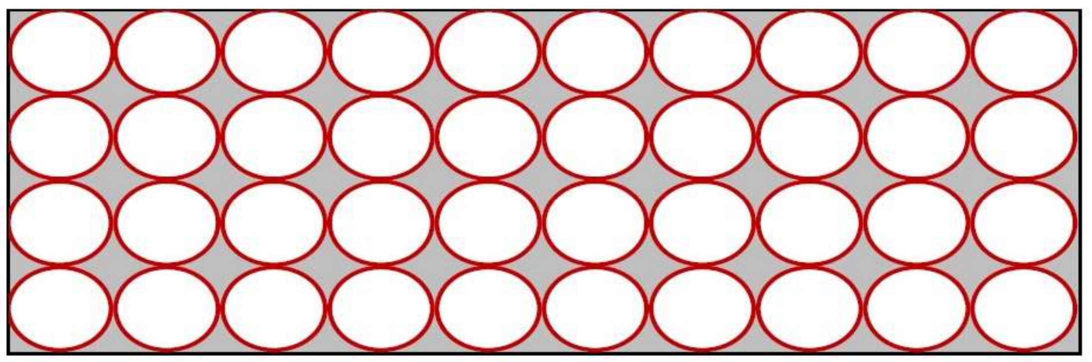
Surgen las siguientes preguntas:

1. ¿Cuántos discos enviar a cada laboratorio?.

   Esto se puede resolver a partir de expresiones de tamaño de muestra para diferencia entre medias, y a partir de un estudio piloto se calculan las varianzas.

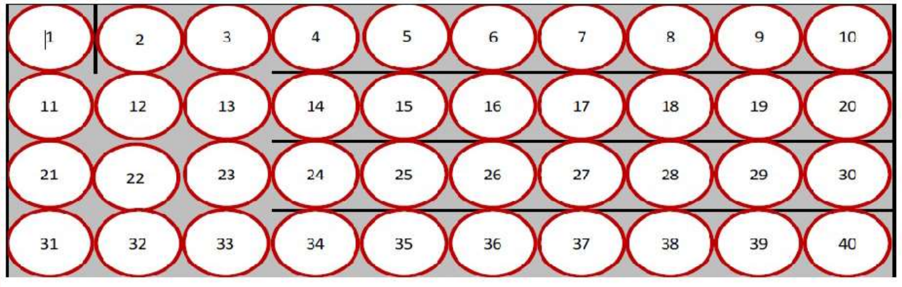
   Se supone que dicho valor se ha calculado y que este valor es de $\mathbf{n = 12}$ discos para cada laboratorio. Total de discos de $\mathbf{48}$.

2. ¿Cómo asignar los $\mathbf{12}$ discos a cada laboratorio, **de forma aleatoria**?



Caso de Estudio 1

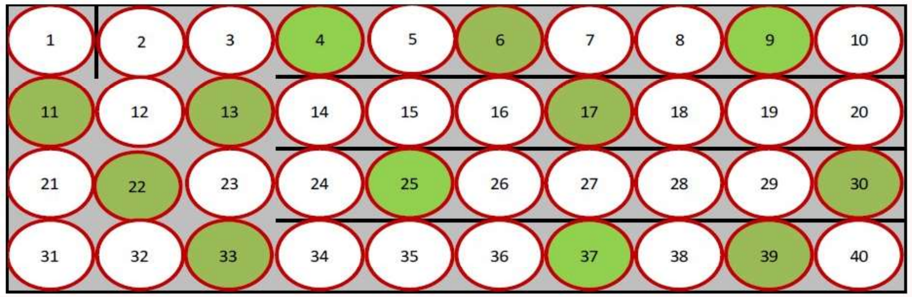
Paso 1: Enumeración de las UE



Paso 2: Selección de la UE para el laboratorio 1

En el lenguaje de programación R: *muestra< −sort(sample(1:40,12))*

4, 6, 9, 11, 13, 17, 22, 25, 30, 33, 37, 39



Caso de Estudio 1

Consideraciones Adicionales al Caso de Estudio

Suponga que se realizó el mismo procedimiento anterior para los otros tres (3) laboratorios y se obtuvieron las respectivas mediciones del peso de estaño.
Si las medias de los **pesos de los revestimientos** en los cuatro (**4**) laboratorios varían significativamente

¿Esto nos permitiría concluir que las diferencias son atribuibles a la falta de consistencia en las técnicas de medición?

Por ejemplo, suponga que otras investigaciones indican que la **cantidad de estaño** depositado electrolíticamente sobre una gran lámina de acero tiene un patrón distinto y repetido perpendicular a la dirección en que es laminado. En esta situación se concluiría que aunque los laboratorios hayan medido la cantidad de estaño consistentemente y sin error las diferencias en las determinaciones de los pesos se debe a alguna otra causa.
Además, la asignación de los discos de una tira completa a cada laboratorio es tal que las inconsistencias entre los métodos de medición de los laboratorios no puede separarse (o confundirse).

Definición y notación

Suponga que se tienen **a**-tratamientos o niveles diferentes asociados a un solo factor, con el objetivo de compararlos.

La respuesta observada de cada uno de los **a**-tratamientos es una **variable aleatoria**, denotada por **Y**.

Los datos se pueden representar con la siguiente tabla:

Tratamiento (nivel)	Observaciones				Totales	Promedios
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\cdots$	y_{an}	$y_{a.}$	$\bar{y}_{a.}$
					$y_{..}$	$\bar{y}_{..}$

Experimentos con un solo factor

Descripción de los datos de la Tabla anterior

- $\mathbf{y_{i\cdot}}$: Total de los valores de la variable respuesta en el $\mathbf{i}$-ésimo tratamiento, $\mathbf{i = 1, 2, \cdots, a}$.
- $\mathbf{\bar{y}_{i\cdot}}$: Promedio de los valores de la respuesta en el $\mathbf{i}$-ésimo tratamiento $\mathbf{i = 1, 2, \cdots a}$.
- $\mathbf{y_{\cdot\cdot}}$: Total de los valores de la variable respuesta en todas las **UE**.
- $\mathbf{\bar{y}_{\cdot\cdot}}$: Promedio de los valores de la variable respuesa en todas las **UE**.

Expresiones matemáticas

$y_{i.} = \sum_{j=1}^{n} y_{ij}$	$\bar{y}_{i.} = \frac{y_{i.}}{n}$
$y_{..} = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}$	$\bar{y}_{..} = \frac{y_{..}}{N}, N = an$

n: Número de réplicas en cada tratamiento; N: Número total de **UE**.

Experimentos con un solo factor

Ventajas y desventajas del Diseño de un Factor Completamente Aleatorizado

- En este diseño los $\mathbf{a}$-tratamientos se asignan aleatoriamente a las **UE**, las cuales deben ser **homogéneas**.
- Por regla general, éste no es el diseño más **eficiente** para ensayos de campo con plantas; sin embargo, se usa extensamente para experimentos en laboratorios e invernaderos debido a que en estos lugares es posible controlar al máximo las condiciones del experimento.
- Se puede probar cualquier **número de tratamientos**. Resulta deseable, aunque no esencial, asignar el mismo número de **UE** a cada tratamiento- **Diseño balanceado**-
- Sus principales ventajas son su sencillez y flexibilidad.
- Su principal desventaja es posible que con algún otro diseño se pueda estimar el error experimental con un mayor grado de precisión.

Experimentos con un solo factor

El Modelo Estadistico

Los datos se describen mediante el siguiente modelo estadístico lineal llamado modelo de análisis de varianza o de un solo factor (ANOVA)

$$\mathbf{Y_{ij}} = \mu + \tau_{\mathbf{i}} + \varepsilon_{\mathbf{ij}}, \tag{1}$$

- $\mathbf{i} = \mathbf{1}, \mathbf{2}, \cdots, \mathbf{a}, \ \mathbf{j} = \mathbf{1}, \mathbf{2}, \cdots, \mathbf{n}.$
- μ: Parámetro común a todos los tratamientos (media global)
- $\tau_{\mathbf{i}}$: Parámetro único del $\mathbf{i}$-ésimo tratamiento (efecto del $\mathbf{i}$-ésimo tratamiento)
- $\varepsilon_{\mathbf{ij}}$ Error aleatorio.

Experimentos con un solo factor

Objetivos

El objetivo del ANOVA es probar hipótesis con respecto a los **efectos de los tratamientos** y hacer una estimación de ellos.

Supuestos

Los errores ε_{ij} son variables aleatorias (v.a.) independientes e idénticamente distribuídas **Normal** con media **cero** y varianza σ^2-constante para todos los niveles del factor, es decir:

$$\varepsilon_{ij} \ \mathbf{i.i.d.} \sim \mathbf{N}(\mathbf{0}, \sigma^2)$$

o equiv.

$$\varepsilon_{ij} \sim \mathbf{NID}(\mathbf{0}, \sigma^2).$$

Experimentos con un solo factor

Supuesto Adicional

El experimento se debe ejecutar o realizar en orden ALEATORIO de manera que el medio ambiente donde se aplican o usan los tratamientos sea lo más uniforme posible. Es denominado Diseño Experimental Completamente Aleatorizado (DECA).

Nota: En el modelo (1) se pueden presentar dos situaciones con respecto al efecto de los tratamientos, visto a continuación.

Experimentos con un solo factor

Modelo de Efectos Fijos

1. Los $\mathbf{a}$-tratamientos son seleccionados (fijados) por el experimentador. Se prueban hipótesis acerca de las medias $\mu_i = \mu + \tau_i$ de los tratamientos.

   Las conclusiones se aplican sólo a los niveles del factor considerados en el análisis. Éstas NO se pueden extender a tratamientos similares que no hayan sido considerados.

   Se denomina Modelo de Efectos Fijos. Parámetros a estimar: μ, τ_i y σ^2.

Modelo de Efectos Aleatorios

2. Considerar los $\mathbf{a}$-tratamientos como una muestra aleatoria(m.a.) de una población mayor de tratamientos.

   En este caso se pueden generalizar las conclusiones (basadas en la muestra de tratamientos), a todos los tratamientos de la población considerados o no considerados en el análisis.

   En este caso, los $\boldsymbol{\tau_i}$ son **variables aleatorias** y **NO** tiene sentido estimar sus valores para los niveles de tratamientos considerados.

   Se hacen pruebas de hipótesis con respecto a la variabilidad de los τ_i.

   Denominado: Modelo de Efectos Aleatorios o modelo de componentes de varianza.

Mod. Efectos FIJOS: Prueba de igualdad de medias de tratamientos

Para realizar la prueba de igualdad de las $\mathbf{a}$-medias de los tratamientos, se platean las hipótesis:

$$\mathbf{H_0} \;:\; \mathbf{E}[\mathbf{Y_{ij}}] = \mu + \tau_\mathbf{i} = \mu \;,\; \text{ para } \mathbf{i} = \mathbf{1}, \mathbf{2}, \cdots, \mathbf{a}.$$
$$\mathbf{H_1} \;:\; \mathbf{E}[\mathbf{Y_{ij}}] = \mu + \tau_\mathbf{i} \neq \mu_\mathbf{j} = \mu + \tau_\mathbf{j} \text{ para al menos un par } (\mathbf{i}, \mathbf{j})$$

La media del $\mathbf{i}$-ésimo tratamiento, $\mu_\mathbf{i}$, consta de la suma de la media general μ y del efecto del $\mathbf{i}$-ésimo tratamiento $\tau_\mathbf{i}$.
Se trata de probar la igualdad de las $\mathbf{a}$-medias asociadas a los $\mathbf{a}$-tratamientos, i.e:

$$\mathbf{H_0} \;: \mu_1 = \mu_2 = \cdots = \mu_\mathbf{a}$$
$$\mathbf{H_1} \;:\; \mu_\mathbf{i} \neq \mu_\mathbf{j} \text{ para al menos un par } (i, j)$$

Si $\mathbf{H_0}$-es verdadera, todos los tratamientos tienen la media común μ.

Formulación de Hipótesis Equivalente

Una forma equivalente de expresar las hipótesis anteriores es en términos de los efectos de los tratamientos $\tau_{\mathbf{i}}$, como sigue:

$$\begin{aligned} \mathbf{H_0} &: \tau_1 = \tau_2 = \cdots = \tau_a = \mathbf{0} \\ \mathbf{H_1} &: \ \tau_i \neq \mathbf{0}, \ \text{ para al menos un } \ \mathbf{i}, \end{aligned}$$

es decir, probar la igualdad de las medias de los tratamientos, es equivalente a probar la hipótesis de que los efectos de cada uno de los tratamientos, τ_i, son cero.

Para realizar dicha prueba de hipótesis se utiliza el análisis de varianza.

Análisis de Varianza

El análisis de varianza se basa en la descomposición de la variabilidad total de los datos en dos componentes, i.e. la descomposición de la suma total de cuadrados corregida, $\mathbf{SST}$, dada por:

$$\mathbf{SST} = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2,$$

$\mathbf{SST}$ se puede escribir como:

$$\begin{aligned}\mathbf{SST} = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 &= \sum_{i=1}^{a}\sum_{j=1}^{n}[(\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})]^2 \\ &= \sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2\end{aligned}$$

Experimentos con un solo factor

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2}_{SST} = \underbrace{n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2}_{SS_{Trat}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2}_{SS_E}$$

...

La Ecuación anterior muestra que la **variabilidad total** de los datos observados, medida por la suma total de cuadrados corregida (SST), se descompone en la suma de cuadrados de las diferencias entre los promedios de los tratamientos (SS_{Trat}) y el promedio general y en la suma de cuadrados de las diferencias entre las observaciones dentro de los tratamientos y el promedio de dichos tratamientos(SS_E).

Los **grados de libertad** asociados a cada una de las **sumas de cuadrados** (**SS**) son:

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2}_{SST} = \underbrace{n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2}_{SS_{Trat}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2}_{SS_E}$$

$$N-1 \quad = \quad a-1 \quad + \quad N-a.$$

Experimentos con un solo factor

Tabla resumen del análisis de varianza (o ANOVA) para el modelo del DECA.

Tabla Anova

Fuente de Variación	Suma Cuad.	gl	Media Cuad.	Est. F
Entre los tratamientos	SS_{Trat}	a-1	MS_{Trat}	$F = \frac{MS_{Trat}}{MSE}$
Error: Dentro de los tratamientos	SSE	N-a	MSE	
Total	SST	N-1		

$$SS_{Trat} = \frac{1}{n}\sum_{i=1}^{a} y_{i\cdot}^2 - \frac{y_{\cdot\cdot}^2}{N} \quad \text{y} \quad SST = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{\cdot\cdot}^2}{N}$$

$$SSE = SST - SS_{Trat}.$$

Experimentos con un solo factor

Ejemplo Láminas de Hojalata en R:

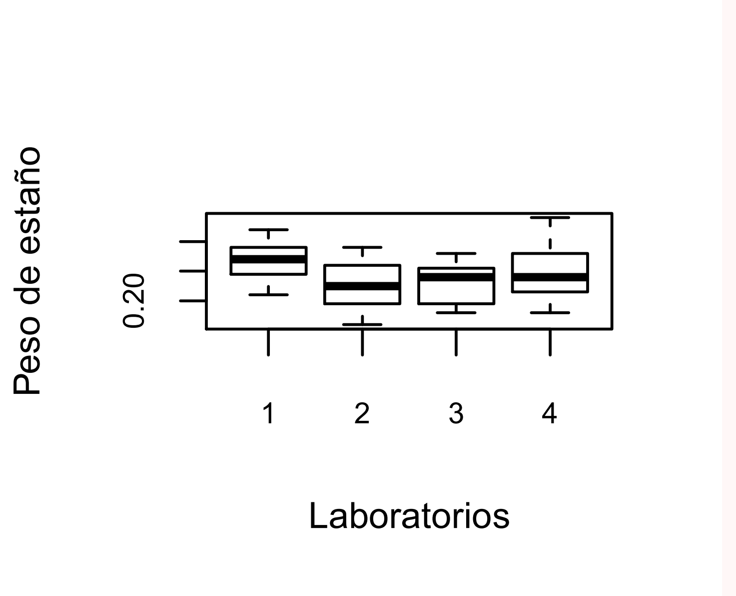
Retomando el **Caso de Estudio**. Después de hacer la asignación aleatoria de cada uno de las **48** láminas circulares, se envían las **12** láminas correspondientes a los priemros **12** números aleatorios al primer laboratorio, y de manera análoga se hace la asignación para los demás laboratorios: Los tres(3) laboratorios de los fabricantes y el laboratorio de la fundidora.

Los resultados de los pesos de revestimiento de estaño se listan a continuación:

```
% Programa para diseños de experimentos de un solo factor
laboratorio<-c(rep(1,12),rep(2,12),rep(3,12),rep(4,12))
laboratorio<-as.factor(laboratorio)
peso<-scan()
0.25 0.27 0.22 0.30 0.27 0.28 0.32 0.24 0.31 0.26 0.21 0.28
0.18 0.28 0.21 0.23 0.25 0.20 0.27 0.19 0.24 0.22 0.29 0.16
0.19 0.25 0.27 0.24 0.18 0.26 0.28 0.24 0.25 0.20 0.21 0.19
0.23 0.30 0.28 0.28 0.24 0.34 0.20 0.18 0.24 0.28 0.22 0.21

boxplot(split(peso,laboratorio),xlab='Laboratorios',
        ylab='Peso de estaño',cex.lab=0.6,cex.axis=0.5)
anovapeso<-aov(peso~ laboratorio); summary(aov(peso~laboratorio))
```

Boxplot y ANOVA Usando el R



	Df	SS	MS	F_o	Pr(>F): Valor P
Laboratorio	3	0.01301	0.004335	2.81	0.0504
Residuales	44	0.06789	0.001543		
Total	47	0.08090			

Estimación de los Parámetros del modelo

Considere el modelo:

$$\mathbf{Y_{ij}} = \mu + \tau_{\mathbf{i}} + \varepsilon_{\mathbf{ij}} \text{ para } \mathbf{i} = \mathbf{1}, \mathbf{2}, \cdots, \mathbf{a} \quad \text{y} \quad \mathbf{j} = \mathbf{1}, \mathbf{2}, \cdots, \mathbf{n}.$$

Para la estimación de los parámetros se usa el método de **Mínimos Cuadrados**, *sin necesidad de suponer la normalidad* de los errores $\varepsilon_{\mathbf{ij}}$, con el fin de estimar a μ y a los $\tau_{\mathbf{i}}$.

Para ello, se hallan los valores de μ y $\tau_{\mathbf{i}}$ tales que se MINIMICE la suma de cuadrados de errores dada por:

$$\mathbf{L} = \sum_{\mathbf{i=1}}^{\mathbf{a}} \sum_{\mathbf{j=1}}^{\mathbf{n}} \varepsilon_{\mathbf{ij}}^{\mathbf{2}} = \sum_{\mathbf{i=1}}^{\mathbf{a}} \sum_{\mathbf{j=1}}^{\mathbf{n}} (\mathbf{y_{ij}} - \mu - \tau_{\mathbf{i}})^{\mathbf{2}}$$

Experimentos con un solo factor

Continuación

Los valores apropiados de μ y τ_i que minimizan a L son las soluciones a las $(a+1)$-ecuaciones simultáneas:

$$\frac{\partial L}{\partial \mu}\Big|_{\hat{\mu},\hat{\tau}_i} = 0$$

$$\frac{\partial L}{\partial \tau_i}\Big|_{\hat{\mu},\hat{\tau}_i} = 0\,, \quad i = 1, 2, \cdots, a$$

equivale a:

$$-2\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \hat{\mu} - \hat{\tau}_i) = 0$$

$$-2\sum_{j=1}^{n}(y_{ij} - \hat{\mu} - \hat{\tau}_i) = 0\,, \quad i = 1, 2, \cdots, a$$

Continuación

Y después de simplificar se obtiene:

$$
\begin{aligned}
N\hat{\mu}+n\hat{\tau}_1+n\hat{\tau}_2+\cdots+n\hat{\tau}_a &= y_{..} \\
n\hat{\mu}+n\hat{\tau}_1 &= y_{1.} \\
n\hat{\mu}+n\hat{\tau}_2 &= y_{2.} \\
\vdots \qquad\qquad &\vdots \\
n\hat{\mu}+n\hat{\tau}_a &= y_{a.}
\end{aligned}
$$

A las anteriores $(a+1)$-ecuaciones con $(a+1)$-incógnitas se le denominan Ecuaciones Normales de Mínimos Cuadrados.
Sumando las últimas a-ecuaciones se obtiene la primera ecuación, de donde las ecuaciones normales no son LI y por tanto no hay solución única para $\mu, \tau_1, \cdots, \tau_a$.

Continuación

Usando la restricción: $\sum_{i=1}^{a} \hat{\tau}_i = 0$, se halla como solución al sistema de ecuaciones:

$$\hat{\mu} = \bar{y}_{..}$$
$$\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..}\ , \ \text{ para } \ i = 1, 2, \cdots, a.$$

Conocidos como estimadores de mínimos cuadrados de los parámetros del modelo.