HERRAMIENTAS ESTADÍSTICAS-COMPARACIÓN DE MÁS DE DOS MUESTRAS: ANOVA (PARTE I)

Módulo 13

APUNTES DE CLASE

Profesor: Arturo Ruiz-Falcó Rojas

Madrid, Mayo 2009

ÍNDICE

1. COMPARACIÓN DE MEDIAS

Supongamos que se están evaluando las características de calidad de los productos de tres proveedores. Para ello se han realizado los ensayos cuyos resultados se recogen en la Tabla 1. Si la escala de medida de la calidad es tal que cuanto mayor sea su valor, mejor es su calidad ¿qué proveedor suministra productos con mayor calidad?

	Prov. A	Prov. B	Prov. C	
Muestra 1	104,04	99,81	111,65	
Muestra 2	98,18	94,15	110,04	
Muestra 3	105,84	99,53	108,29	
Muestra 4	105,11	100,69	108,00	
Muestra 5	99,73	96,73	106,59	
Media prov.	102,58	98,18	108,91	
Media total				103,23

Tabla 1: Comparación de proveedores

Si se representan estos valores en la Figura 1, podría concluirse que los productos fabricados por el proveedor C tienen mejor calidad que los de A y B. Sin embargo, la comparación entre A y B no es tan concluyente aunque parece que los productos de A son algo mejores que los de B. Resulta pues necesario objetivar este análisis.

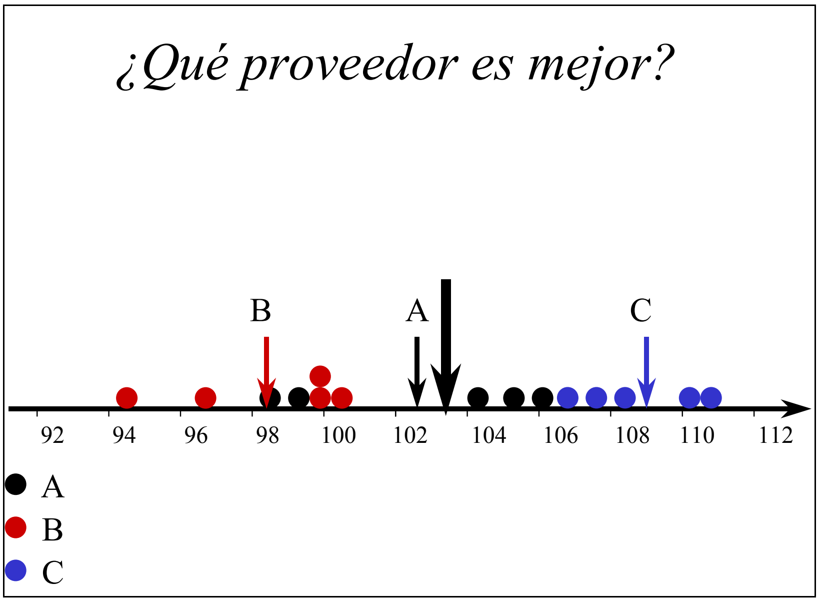

Figura 1 ¿Qué proveedor es mejor?

¿Qué razonamiento se ha seguido para sacar esta conclusión sobre la calidad de los productos de los proveedores A, B y C? En el caso de C se ha visto que cualquiera de sus muestras es superior a cualquiera de las de A ó B, de modo que la conclusión es inmediata. Sin embargo, en el caso de la comparación entre A y B, los resultados están mezclados, es decir que la variabilidad de los elementos de la misma muestra A ó B no es mucho menor que la variabilidad global de las muestras A ó B consideradas como un conjunto. Utilizando los conocimientos estadísticos adquiridos en capítulos anteriores, podrían compararse las muestras dos a dos con el contraste de la t de Student, pero esto no resulta muy práctico en problemas reales.

La herramienta estadística que sirve para resolver el problema de comparar más de dos medias es el ANÁLISIS DE LA VARIANZA, que se llama así precisamente porque compara la variabilidad de las medias muestrales (a través de la varianza muestral) con la variabilidad de los elementos dentro de la muestra.

El ANÁLISIS DE LA VARIANZA permite también descomponer la variabilidad total en componentes independientes que puedan asignarse a causas distintas (ver Tabla 2.

Volviendo al caso de los proveedores, si realizando un **ANÁLISIS DE LA VARIANZA** se puede concluir que las diferencias entre las medias de alguno de ellos es estadísticamente significativa, entonces se puede afirmar que el proveedor en cuestión suministra *"una calidad distinta"*; por el contrario, si las diferencias no son estadísticamente significativas, no se puede concluir lo anterior, siendo las fluctuaciones de los datos muestrales entre proveedores únicamente debidas al azar.

FACTOR	
VARIABILIDAD DEBIDA A LA MÁQUINA	40%
VARIABILIDAD DEBIDA A LA MATERIA PRIMA	25%
VARIABILIDAD DEBIDA A LOS TURNOS	20%
VARIABILIDAD CAUSAS COMUNES	15%
TOTAL	100%

Tabla 2: Ejemplo de Descomposición de la variabilidad

2. FUNDAMENTOS DEL ANÁLISIS DE LA VARIANZA

2.1. Modelo

Los valores de las características de calidad de las piezas de cada proveedor tendrán una variabilidad entorno a un valor medio. Si representamos como y_{ij} al valor de la muestra j del proveedor i:

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

El problema a resolver cuál de las dos situaciones siguientes es la que explican mejor los datos:

- Todos los proveedores son iguales, es decir tienen la misma media $\mu_1 = \mu_2 = \mu_3$
- La media de alguno de los proveedores es diferente a la de los demás.

2.2. Hipótesis requeridas

Para que se pueda aplicar el **ANÁLISIS DE LA VARIANZA** es preciso que se cumplan estas tres hipótesis:

- Los datos han de ser independientes. Para asegurar esto, las muestras cuyas medias se desea comparar han de extraerse de manera aleatoria.
- Las poblaciones base de donde proceden las muestras han de ser normales.
- Las poblaciones base de donde proceden las muestras han de tener la misma varianza (heterocedsticidad).

Estas hipótesis implican que las perturbaciones se distribuyan según una N(0, σ^2).

2.3. Contraste ANOVA

Podremos estimar la varianza de la población σ^2 a través de los siguientes estimadores:

- **Estimar la varianza de la población σ^2 a través de la varianza de cada una de las muestras**. Esta estimación se hace ponderando las varianzas muestrales. Si k es el número de muestras (en adelante denominaremos a cada *"muestra" "tratamiento"*, n_i es el tamaño de la muestra correspondiente a tratamiento i-ésimo y N es el número total de datos disponible en las distintas muestras, el estimador denominado *"varianza residual"* se define:

$$\hat{\sigma}^2 = s_R^2 = \frac{\sum_i^k (n_i - 1)s_i^2}{\sum_i^k (n_i - 1)} = \frac{\sum_i^k (n_i - 1)s_i^2}{N - k}$$

- **Estimar la varianza de la población σ^2 suponiendo que los tratamientos no tienen ningún efecto** (es decir que todos tienen la misma media). En estas condiciones podremos estimar σ^2 a través de la varianza de las medias muestrales:

$$\hat{\sigma}^2 = s_T^2 = \frac{\sum_i^k n_i\left(\bar{y}_i - \bar{y}\right)^2}{k-1}$$

De este modo, si se verifican la hipótesis de que los tratamientos no tienen ningún efecto, ambas estimaciones no podrán diferir mucho. En efecto, si la hipótesis es cierta, el estadístico cociente de ambas varianzas muestrales se distribuye según una F. Es decir:

$$\frac{s_T^2}{s_R^2} \propto F_{k-1,n-k}$$

La metodología para realizar el **ANÁLISIS DE LA VARIANZA** puede resumirse como sigue:

- Fijar el nivel de significación para el contraste, por ejemplo α=95%. Establecer el contraste de hipótesis:

  $\Rightarrow H_0$: Los tratamientos son todos iguales: $\mu_1=\mu_2=\mu_3=\ldots=\mu_k$.

  $\Rightarrow H_1$: Alguno de los tratamientos es diferente.

- Calcular los estimadores s_R^2 y s_T^2.

- Calcular el valor del estadístico $\frac{s_T^2}{s_R^2}$

- Calcular el valor de $F_{k-1,\,n-k}$ para el nivel de significación prefijado. Si:

  $\Rightarrow \frac{s_T^2}{s_R^2} > F_{k-1,n-k}$ La diferencia entre los tratamientos es estadísticamente significativa con un nivel de significación α.

$\Rightarrow \dfrac{s_T^2}{s_R^2} < F_{k-1,n-k}$ La diferencia entre los tratamientos no es estadísticamente significativa con un nivel de significación α.

3. TABLA ANOVA

Denominando S a la suma de los cuadrados, se tiene:

$$s_R^2 = \frac{S_R}{N-k}$$

$$s_T^2 = \frac{S_T}{k-1}$$

Si S_D es la suma de los cuadrados con respecto a la media global, el estadístico s_D^2 es también un estimado de σ^2 si se cumplen las hipótesis de igualdad de medias:

$$s_D^2 = \frac{\sum_{i}^{k}\sum_{j}^{n_i}\left(y_{ij}-\overline{y}\right)^2}{N-1} = \frac{S_D}{N-1}$$

Es fácil comprobar que se verifica la siguiente igualdad:

$$S_D = S_R + S_T$$

En cuanto a los grados de libertad:

$\nu_D = \nu_R + \nu_T$, es decir $N-1 = (N-k)+(k-1)$

S_D se denomina también *"suma corregida de cuadrados"* y se calcula fácilmente mediante la siguiente ecuación:

$$S_D = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - N\overline{y}^2$$

En general lo más cómodo es calcular S_D y S_T, calculando S_R por diferencia. Es costumbre presentar el **ANÁLISIS DE LA VARIANZA** en forma de tabla:

FUENTE DE VARIACIÓN	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	CUADRADO MEDIO	CONTRASTE
ENTRE TRATAMIENTOS (VE)	$S_T = \sum_i^k n_i\left(\bar{y}_{t_i} - \bar{y}\right)^2$	$\nu_T = k - 1$	$s_T^2 = \frac{S_T}{k-1}$	$\frac{s_T^2}{s_R^2}$
DENTRO DE TRATAMIENTOS (VNE)	$S_R = \sum_i^k \sum_j^{n_i} \left(y_{ij} - \bar{y}_i\right)^2$	$\nu_R = N - k$	$s_R^2 = \frac{S_R}{N-k}$	
TOTAL EN RELACIÓN A LA MEDIA GENERAL (VT)	$S_D = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}^2 - N\bar{y}^2$	$\nu_D = N - 1$	$s_D^2 = \frac{S_D}{N-1}$	

Tabla 3 ANÁLISIS DE LA VARIANZA DE 1 FACTOR

A esta tabla se le suele denominar "tabla ANOVA", (del inglés **An**alysis **o**f **Va**riance). De amera análoga al análisis de regresión, al cociente de la variabilidad explicada por los tratamientos respectos de la variabilidad total, se denomina coeficiente de determinación:

$$R^2 = \frac{VE}{VT} = \frac{S_T}{S_D}$$

Construyendo la tabla ANOVA correspondiente al caso de los proveedores, resulta:

	Suma de	g. de l.	Cuadrado

	Cuadrados		Medio
Entre trat.	291	2	146
Dentro trat.	91	12	8
Suma corregida	382	14	27

Tabla 4: Tabla ANOVA de los proveedores

$S_T^2/S_R^2=$ 19,16

$F_{2,12}(0,95)=$ 3,8853

Como 19,16 > 3,88 se rechaza la hipótesis de que todos los proveedores son iguales. No obstante, para poder dar por bueno el resultado es preciso comprobar que se satisfacen las hipótesis de partida. Para ello se realiza un análisis de los residuos (ver Figura 2) sin que se aprecie en él ningún aspecto que haga dudar de la normalidad de sus distribución.

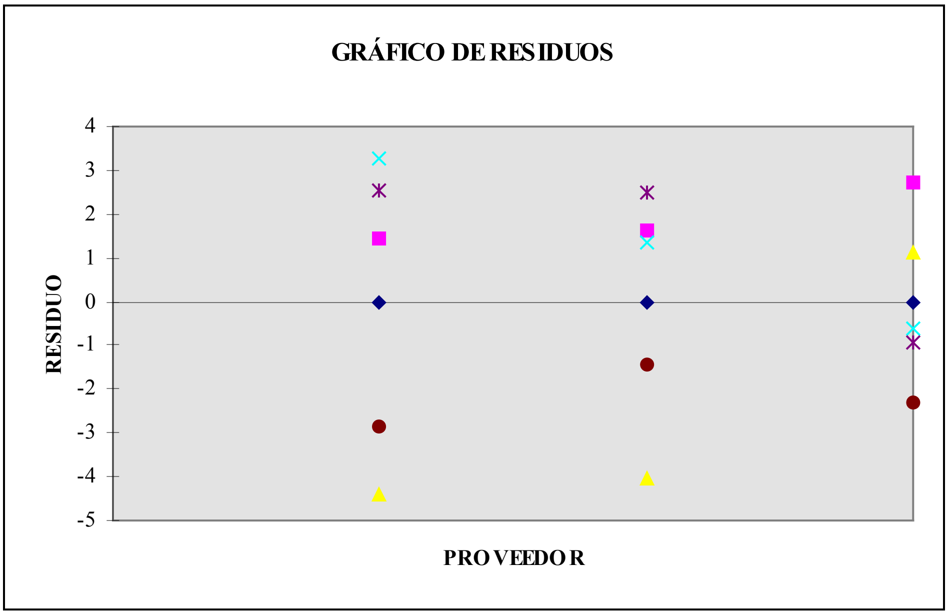

Figura 2: Análisis de los residuos

4. ANÁLISIS DE LA DIFERENCIA ENTRE LAS MEDIAS

El análisis de la varianza nos indica si alguno de los proveedores es distinto, pero no indica cual es. Para resolver esto se puede hacer lo siguiente:

1. Construir una distribución de referencia con la t de Student para cada uno de los proveedores para ver si solapa o no a los otros.
2. Contrastar las diferencias de las medias de todos los pares posibles de medias utilizando la distribución de Student. Esto presenta el inconveniente de que si cuantos más niveles se analicen (proveedores distintos, en este caso) la probabilidad de cometer un error de tipo I aumenta . En efecto, si el nivel de confianza es 0,95 y tenemos 3 proveedores, el número de comparaciones es 3; entonces la probabilidad de concluir que un grupo es diferente sin que lo sea es $1\text{-}0.95^3$= 0,143.
3. Método de Bonferroni. Es útil cuando el número de grupos es grande porque corrige en parte el efecto anterior.
4. Realizar comparaciones múltiples. Proporciona intervalos de confianza para las diferencias de las medias de todos los pares de grupos. Los más utilizados son:
    a. Dunnet. Se utiliza cuando se toma uno de los grupos como referencia.
    b. MCB (Multiple Comparation with the Best) de Hsu. Compara con el grupo “bueno” (el más alto o el más bajo.
    c. Fisher LSD (least significant difference)
    d. Tukey.

5. PROCEDIMIENTO DE APLICACIÓN

En general, el procedimiento de aplicación del análisis de la varianza consta de los siguientes pasos (ver esquema en Figura 3).

<u>Representación de los datos</u>

Como en la mayor parte de los procedimientos estadísticos debe comenzarse por representar gráficamente los datos. Si el número de datos por grupo es menor que 10, se recomienda emplear el diagrama de puntos; si es superior a 10 se recomiendan histogramas o diagramas de caja. En esta representación deben buscarse valores atípicos. Si estos valores atípicos no se deben a un error o una causa subsanable (por ejemplo, error de transcripción de datos) debe pensarse en la necesidad de transformar los datos para que cumplan las hipótesis de normalidad. En la Tabla 5 se dan algunas indicaciones de transformaciones recomendadas. De manera general se pueden emplear las transformaciones de Box Cox.

Relación media-varianza	λ	Transformación estabilizadora de la varianza	Ejemplos típicos de aplicación
$\sigma \propto \eta^2$	-1	Inversa	
$\sigma \propto \eta^{1,5}$	-1/2	Inversa de la raíz cuadrada	
$\sigma \propto \eta$	0	Logaritmo	Análisis de varianzas muestrales, s
$\sigma \propto \eta^{0,5}$	½	Raíz cuadrada	Datos procedentes de una distribución de Poisson
$\sigma \propto const$	1	No se transforma	

Tabla 5: Transformaciones de datos para eliminar la heterocedasticidad

Si los datos proceden de un fenómeno de tipo binomial, por ejemplo porcentaje de unidades rechazadas, la transformación adecuada es y = arcsin(p). Si proceden de un fenómeno de tipo Poisson, por ejemplo número de defectos, la transformación adecuada es y = $\sqrt{c}$.

Construcción de la tabla ANOVA y realización del contraste

Esto puede completarse con la construcción de intervalos de confianza para las medias de cada grupo y los contraste múltiples.

Validación de las hipótesis

Es muy importante tener en cuenta en todo momento que la validez de las conclusiones está supeditada a que las hipótesis realizadas sean ciertas. Estas comprobaciones pueden hacerse analizando los residuos, es decir las diferencias que existen entre lo explicado por el modelo y los valores obtenidos.

$$y_{ij} = \bar{y}_i + e_{ij}$$

Por lo tanto es preciso realizar las siguientes comprobaciones:

- Independencia de los datos. En caso de que los datos se hayan producido según patrones temporales, etc. se deben representar los residuos en la secuencia que se obtuvieron y no deben observarse tendencias, rachas, etc.

- Normalidad de las perturbaciones. Los residuos deben distribuirse normalmente. Debe representarse en un papel probabilístico.
- Heterocedasticidad. Se representan los residuos por grupos tener una dispersión parecida. Ver por ejemplo la Figura 2. Si el número de datos es al mismo para todos los grupos, el ANOVA es bastante robusto frente a esta hipótesis.

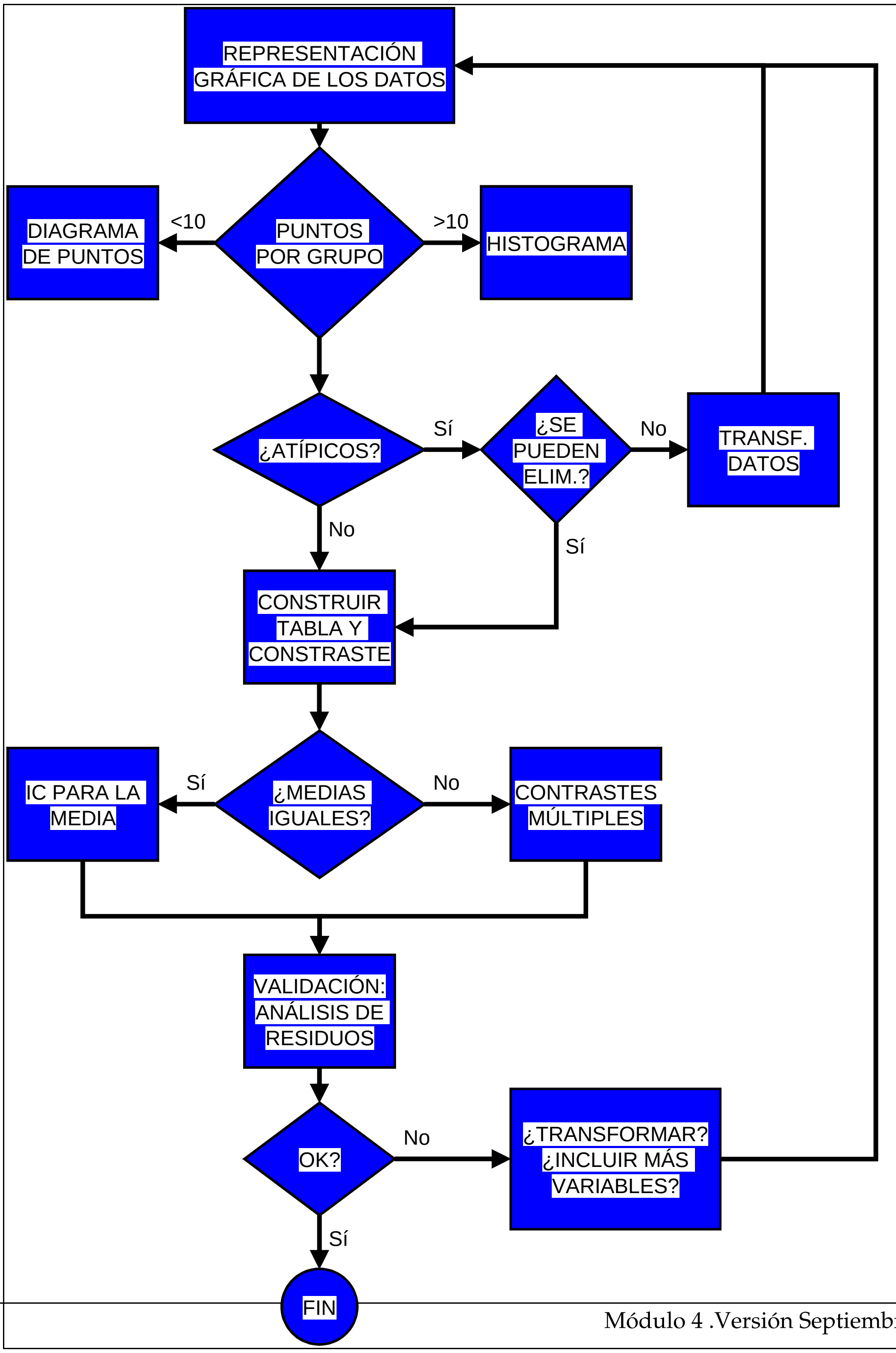
REPRESENTACIÓN GRÁFICA DE LOS DATOS
DIAGRAMA DE PUNTOS
<10
PUNTOS POR GRUPO
>10
HISTOGRAMA
¿ATÍPICOS?
Sí
¿SE PUEDEN ELIM.?
No
TRANSF. DATOS
No
Sí
CONSTRUIR TABLA Y CONSTRASTE
IC PARA LA MEDIA
Sí
¿MEDIAS IGUALES?
No
CONTRASTES MÚLTIPLES
VALIDACIÓN: ANÁLISIS DE RESIDUOS
OK?
No
¿TRANSFORMAR? ¿INCLUIR MÁS VARIABLES?
Sí
FIN

Figura 3: Esquema de los pasos de aplicación del ANOVA

EJEMPLO 1. COMPARACIÓN DE TERMÓMETROS

Se está realizando una comparación de cuatro termómetros. Con cada uno de ellos se ha realizado tres ensayos de medida del punto de fusión de un compuesto químico.

Los datos obtenidos son los de la tabla siguiente:

TERMÓMETRO A	TERMÓMETRO B	TERMÓMETRO C	TERMÓMETRO D
174,0	173,0	171,5	173,5
173,0	172,0	171,0	171,0
173,5	173,0	173,0	172,5

Tabla 6: Comparación de termómetros

Representación de los datos

Como solo se dispone de 3 datos por termómetro, se representará un diagrama de puntos.

APLICACIÓN MINITAB

`Minitab dispone de la opción` ***`Graf->DotPlot.`*** *`Se obtiene:`*

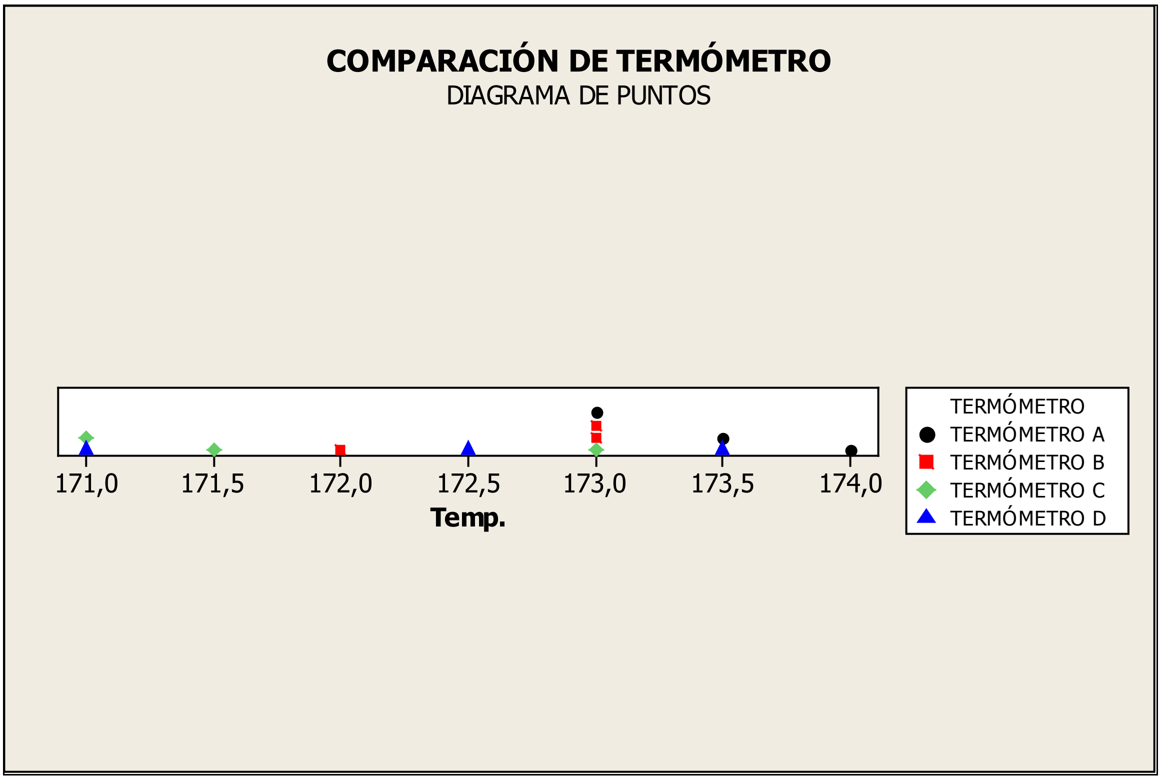

Figura 4: Diagrama de puntos de los datos de los termómetros

En este diagrama no se observan diferencias entre termómetros. Si se calculan los estadísticos, los datos anteriores podrían encajar en una normal de media 172,58 ^{o}C y desviación 0,996 ^{o}C.

```
Variable    N  N*    Mean  SE Mean  StDev  Minimum      Q1  Median      Q3
Temp.      12   0  172,58    0,288  0,996   171,00  171,63  173,00  173,38

Variable   Maximum

Temp.       174,00
```

Construcción de la tabla ANOVA y realización del contraste
APLICACIÓN MINITAB

Minitab dispone de la opción ***Stat->ANOVA->One Way.*** *Se obtiene la salida siguiente (ver comentarios de interprectación en mayúsculas:*

One-way ANOVA: Temp. versus TERMÓMETRO

```
Source       DF      SS     MS     F      P
TERMÓMETRO    3   4,417  1,472  1,81  0,223
Error         8   6,500  0,812
Total        11  10,917

S = 0,9014   R-Sq = 40,46%   R-Sq(adj) = 18,13%
```

P= 0,223 LUEGO NO SE PUEDE RECHAZAR LA IGUALDAD DE NINGÚN TERMÓMETRO

```
                                 Individual 95% CIs For Mean Based on
                                 Pooled StDev
Level         N     Mean  StDev  --------+---------+---------+---------+-
TERMÓMETRO A  3  173,500  0,500                (---------*---------)
TERMÓMETRO B  3  172,667  0,577         (---------*---------)
TERMÓMETRO C  3  171,833  1,041  (---------*---------)
TERMÓMETRO D  3  172,333  1,258      (---------*---------)
                                 --------+---------+---------+---------+-
                                     171,6     172,8     174,0     175,2

Pooled StDev = 0,901
```

LOS INTERVALOS DE CONFIANZA ESTÁN TODOS SOLAPADOS. ES COHERENTE CON LO ANTERIOR

```
Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of TERMÓMETRO

Individual confidence level = 98,74%


TERMÓMETRO = TERMÓMETRO A subtracted from:

TERMÓMETRO      Lower   Center   Upper
TERMÓMETRO B  -3,1908  -0,8333  1,5242
TERMÓMETRO C  -4,0242  -1,6667  0,6908
TERMÓMETRO D  -3,5242  -1,1667  1,1908

TERMÓMETRO      +---------+---------+---------+---------
TERMÓMETRO B        (-----------*----------)
TERMÓMETRO C    (-----------*----------)
TERMÓMETRO D      (-----------*-----------)
                +---------+---------+---------+---------
             -4,0      -2,0       0,0       2,0
```

```
TERMÓMETRO = TERMÓMETRO B subtracted from:

TERMÓMETRO      Lower   Center   Upper
TERMÓMETRO C  -3,1908  -0,8333  1,5242
TERMÓMETRO D  -2,6908  -0,3333  2,0242

TERMÓMETRO       +---------+---------+---------+---------
TERMÓMETRO C         (-----------*-----------)
TERMÓMETRO D            (----------*-----------)
                 +---------+---------+---------+---------
              -4,0      -2,0       0,0       2,0


TERMÓMETRO = TERMÓMETRO C subtracted from:

TERMÓMETRO      Lower  Center   Upper
TERMÓMETRO D  -1,8575  0,5000  2,8575

TERMÓMETRO       +---------+---------+---------+---------
TERMÓMETRO D                (----------*-----------)
                 +---------+---------+---------+---------
              -4,0      -2,0       0,0       2,0
```

<u>EN EL TEST DE TUKEY LOS INTERVALOS DE CONFIANZA DE LAS DIFERENCIAS ENTRE TERMÓMETROS ESTÁN TODOS SOLAPADOS. ES COHERENTE CON LO ANTERIOR</u>

```
Fisher 95% Individual Confidence Intervals
All Pairwise Comparisons among Levels of TERMÓMETRO

Simultaneous confidence level = 82,43%


TERMÓMETRO = TERMÓMETRO A subtracted from:

TERMÓMETRO      Lower   Center   Upper
TERMÓMETRO B  -2,5305  -0,8333  0,8638
TERMÓMETRO C  -3,3638  -1,6667  0,0305
TERMÓMETRO D  -2,8638  -1,1667  0,5305

TERMÓMETRO     --+---------+---------+---------+-------
TERMÓMETRO B        (----------*-----------)
TERMÓMETRO C   (----------*----------)
TERMÓMETRO D      (----------*-----------)
               --+---------+---------+---------+-------
              -3,0      -1,5       0,0       1,5


TERMÓMETRO = TERMÓMETRO B subtracted from:

TERMÓMETRO      Lower   Center   Upper
TERMÓMETRO C  -2,5305  -0,8333  0,8638
TERMÓMETRO D  -2,0305  -0,3333  1,3638
```

```
TERMÓMETRO     --+---------+---------+---------+-------
TERMÓMETRO C       (----------*-----------)
TERMÓMETRO D          (-----------*----------)
               --+---------+---------+---------+-------
              -3,0      -1,5       0,0       1,5

TERMÓMETRO = TERMÓMETRO C subtracted from:

TERMÓMETRO      Lower  Center   Upper
TERMÓMETRO D  -1,1972  0,5000  2,1972

TERMÓMETRO     --+---------+---------+---------+-------
TERMÓMETRO D                (----------*-----------)
               --+---------+---------+---------+-------
              -3,0      -1,5       0,0       1,5
```

EN EL TEST DE FISHER LOS INTERVALOS DE CONFIANZA DE LAS DIFERENCIAS ENTRE TERMÓMETROS ESTÁN TODOS SOLAPADOS. ES COHERENTE CON LO ANTERIOR

Validación de las hipótesis

Independencia de los datos

No se conoce en el orden en el que se han tomado los datos, por lo que no se puede estudiar esta tendencia. En la Figura 5 se ha representado el residuo en función de la temperatura, sin que se aprecien tendencias.

Normalidad de las perturbaciones

En la Figura 6 puede verse la normalidad de los residuos.

Heterocedasticidad

En la Figura 7 se aprecia que en los termómetros C Y D los datos están algo más dispersos. No obstante solo son tres datos y además al tener el mismo número de datos por termómetro, no se considera importante esta indicación.

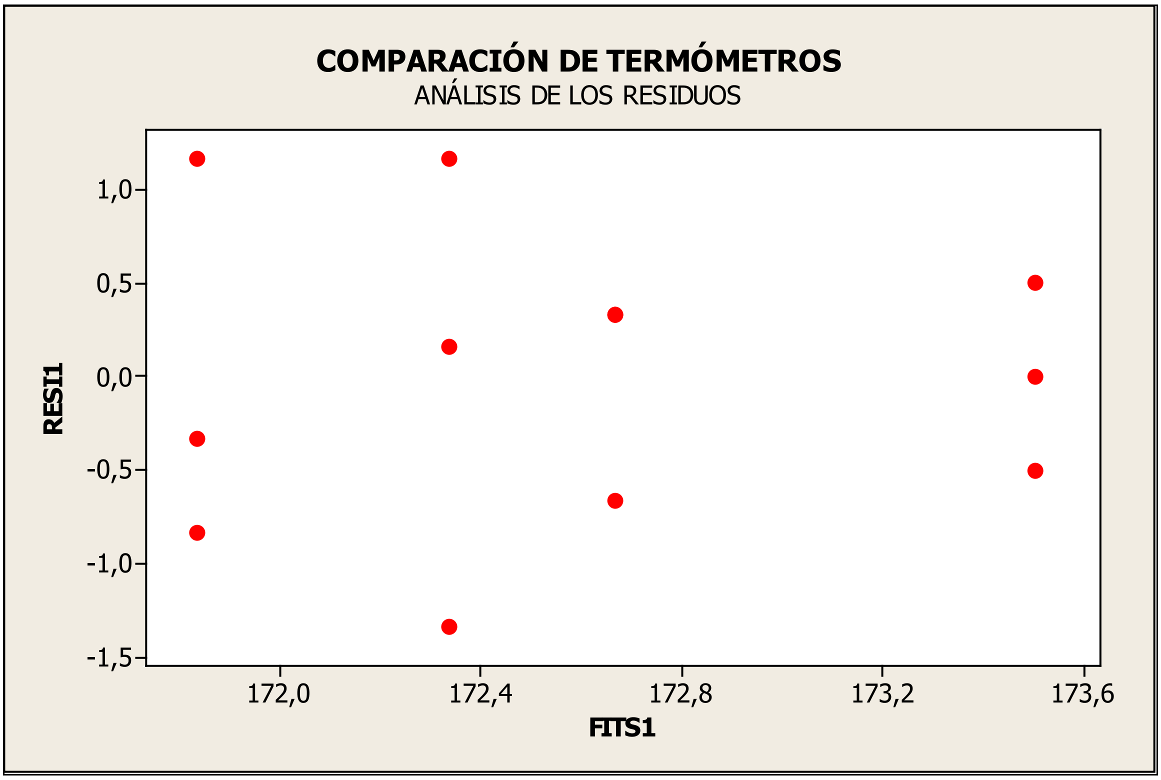

Figura 5: Análisis de tendencias de los residuos en función de la temperatura

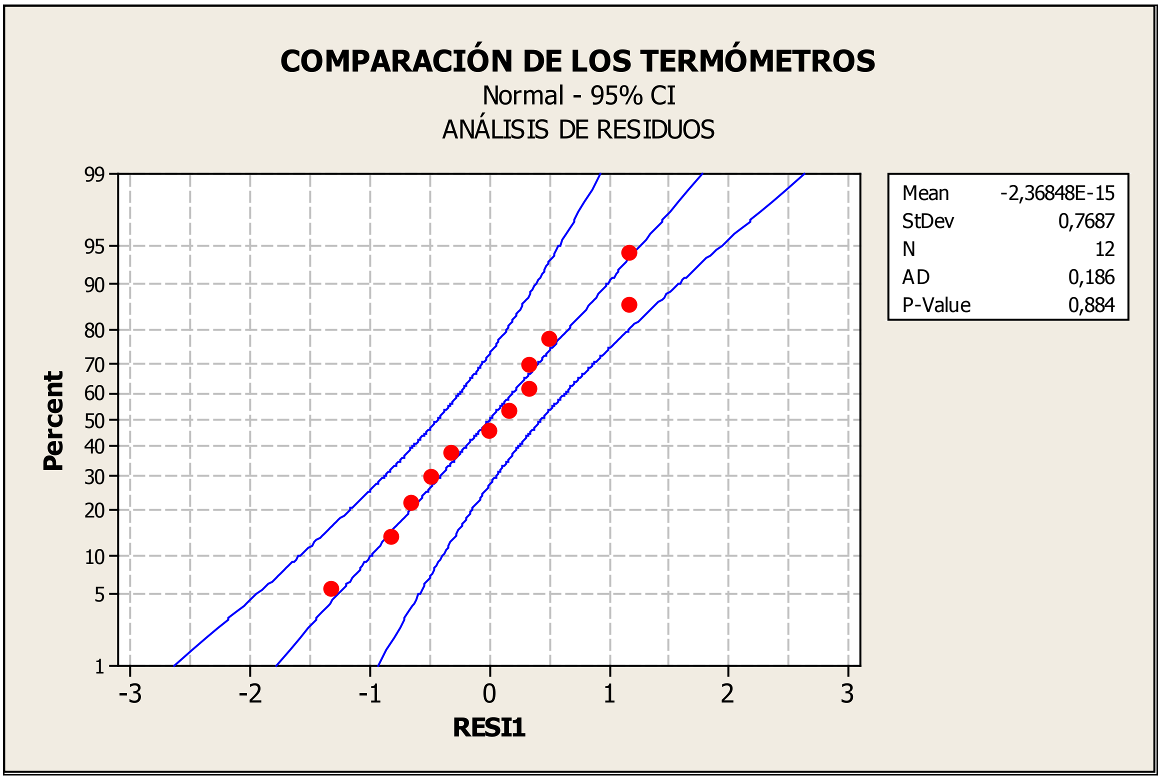

Figura 6: Comprobación de la normalidad de los residuos

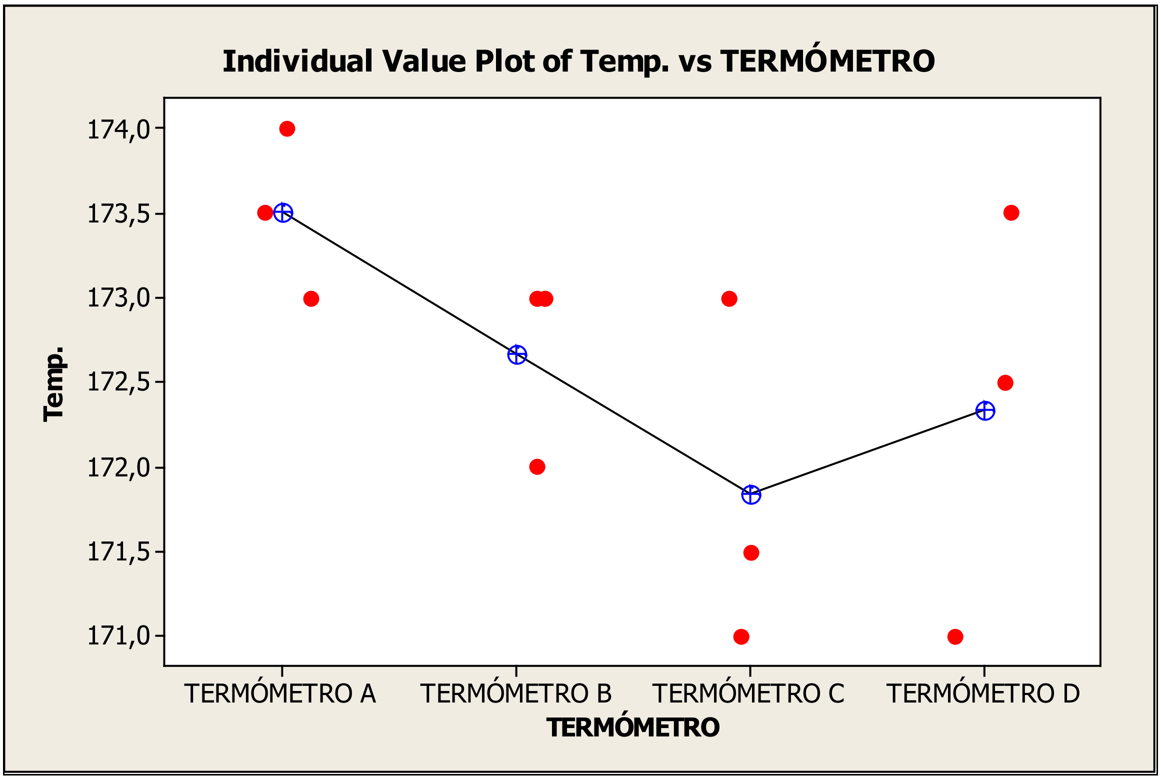

Figura 7: Variabilidad de los datos en cada termómetro

Tras este análisis de residuos quedan validadas las hipótesis y las conclusiones del estudio.

6. EJERCICIOS

6.1. Estudio de durabilidad de alfombras

Con los datos del fichero Exh_aov.MTW, estudie si alguno de los cuatro tipos de alfombras tiene mayor durabilidad.