Práctica 1 Factor

Curso 2019/20

Regresión y ANOVA

Grado en Ingeniería Informática

Universidad **de** Valladolid

Grupo 10

González Caminero, Juan
Rodríguez Arroyo, Diego
Castro Caballero, Manuel
Sáenz Niño, Héctor
Valdunciel Sánchez, Pablo

Ejercicio 1

a) Estudia si este factor influye en el tiempo correspondiente.

Planteamos el siguiente contraste de hipótesis:

- Ho: media_azul = media_rojo = media_verde = u
- H1: alguna de las media (media_azul, media_rojo, media_verde) es distinta

Asumiendo las hipótesis paramétricas de aleatoriedad, normalidad y homocedasticidad, utilizamos la tabla ANOVA. Dado que el p-valor obtenido es <0.0002, podemos rechazar la hipótesis nula Ho a los niveles de confianza habituales, es decir, existen evidencias de el factor 'color' sí influye en el tiempo que los ratones tardan en salir del laberinto.

b) Compara los tiempos medios asociados a los tres colores utilizando distintos métodos y comenta los resultados.

Los tres métodos obtenidos coinciden, con un 95% de confianza, en que no existen diferencias significativas entre las medias del color azul y el color rojo, pero sí entre las medias del color azul y el color verde, y sí entre las medias del color rojo y el color verde

c) Supongamos que el propósito del experimento es comprobar si el color verde tiene algún efecto especial. Realiza un contraste adecuado para dicho objetivo.

Para ello, empleamos el test de DUNNETT, que permite realizar una comparación de todas las medias con un control.

Comparisons significant at the 0.05 level are indicated by ***.				
color Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
red - grn	-11.400	-17.350	-5.450	***
blk - grn	-14.200	-20.150	-8.250	***

Atendiendo a los resultados del test de Dunnett, existe una diferencia significativa entre las medias del color verde y el color azul, y entre las medias del color verde y el color rojo con una confianza del 95%. Estos resultados encajan con los obtenidos en el apartado b). Podemos concluir que el color verde sí tiene un efecto especial.

Práctica - Acuicultura

1. Chequeo de las hipótesis del modelo

En primer lugar, debemos determinar si se cumplen o no las siguientes hipótesis:

1. Bondad del ajuste del modelo estadístico propuesto.
2. La normalidad.
3. La homocedasticidad del error.
4. La homogeneidad de la muestra.
5. La independencia de las observaciones

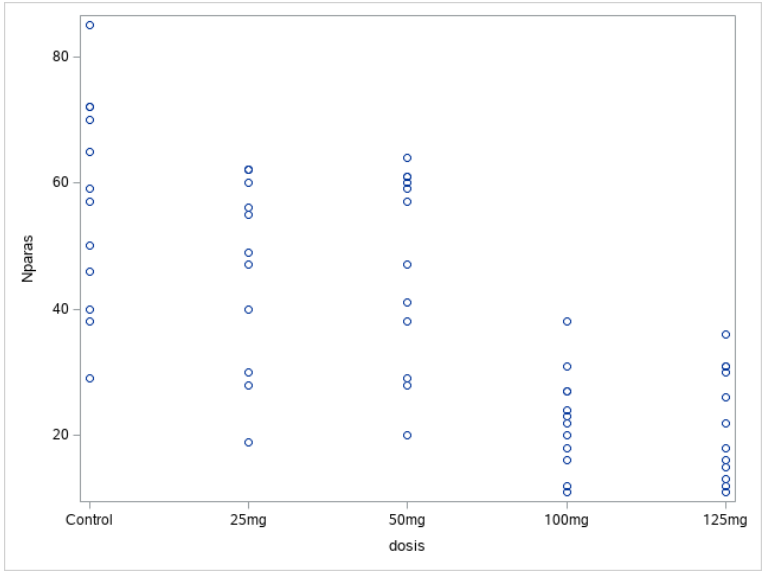
Como pasos previo para comprobar cada una de estas hipótesis realizamos un estudio descriptivo analítico y gráfico de la muestra y de los residuos. Sea la variable respuesta la ‘cantidad de parásitos ’ y sea el factor la ‘dosis de medicamento (mg)’:

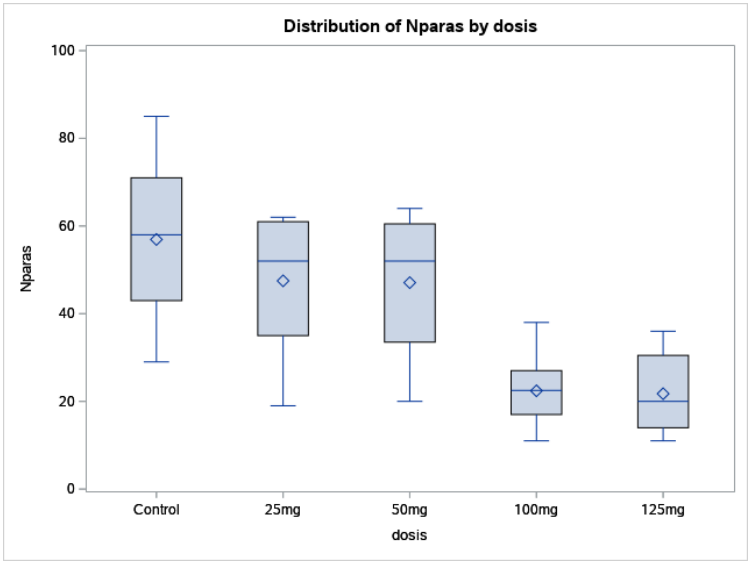
- Estadísticos básicos sobre la variable respuesta según el factor

The MEANS Procedure

Analysis Variable : Nparas						
dosis	N Obs	N	Mean	Std Dev	Minimum	Maximum
100mg	12	12	22.4166667	7.7864490	11.0000000	38.0000000
125mg	12	12	21.7500000	8.7399293	11.0000000	36.0000000
25mg	12	12	47.5000000	14.9939382	19.0000000	62.0000000
50mg	12	12	47.0833333	15.4476143	20.0000000	64.0000000
Control	12	12	56.9166667	16.7193863	29.0000000	85.0000000

- Gráfico de puntos y gráfico de cajas múltiple de la variable respuesta según el factor



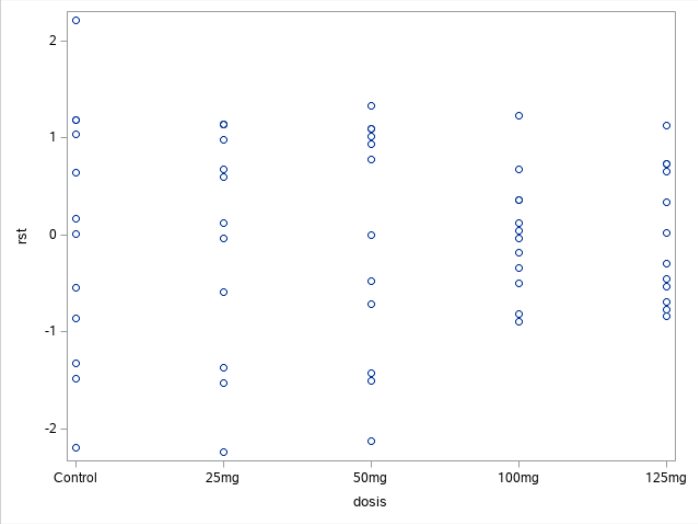


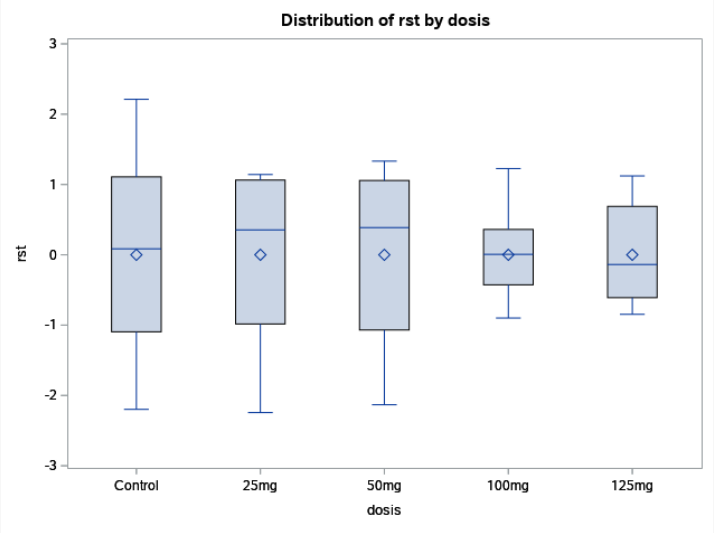
- Ajuste del modelo. Calculo de los residuos y de los residuos estandarizados

Ajustamos el modelo y almacenamos en el dataset **resal** los valores predichos (pred), los residuos (res) y los residuos estandarizados (rst). A partir de este conjunto de datos podremos evaluar si se verifican o no las distintas hipótesis del modelo citadas anteriormente.

1.1. Bondad del ajuste del modelo

Para comprobar la bondad del ajuste representamos los residuos estandarizados según los niveles del factor. Para su representación utilizamos tanto el gráfico de de puntos como el gráfico de cajas múltiples.





Observando la nube de puntos y las cajas, podemos observar como hay tal vez un exceso de valores negativos para los niveles *25mg* y *50mg* del factor, lo cual puede suponer que el modelo no se ajusta muy bien.

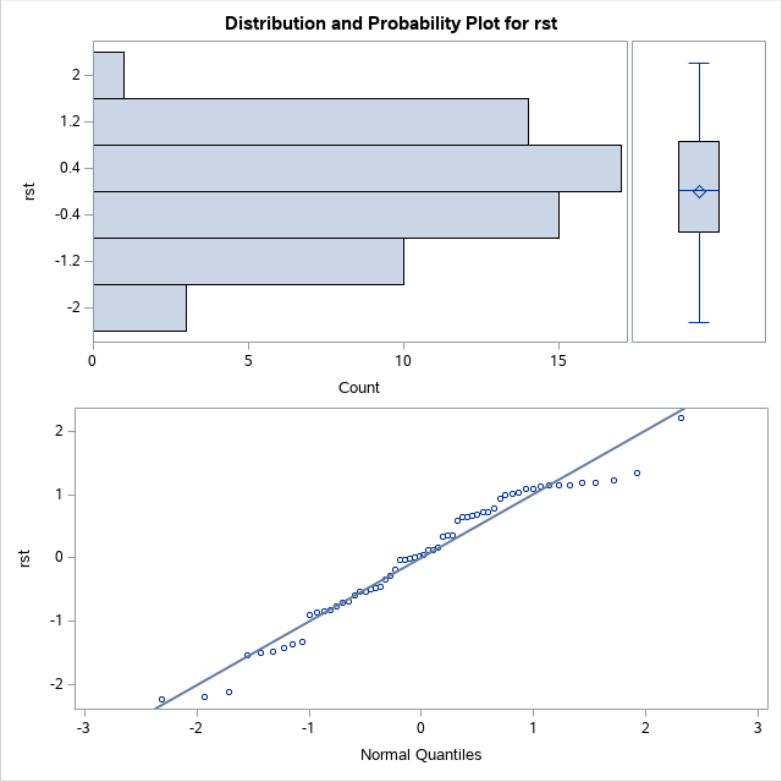
1.2. Normalidad de los errores

Realizamos un análisis de normalidad de los residuos estandarizados.

Variable: rst

Moments			
N	60	Sum Weights	60
Mean	0	Sum Observations	0
Std Deviation	1.00843897	Variance	1.01694915
Skewness	-0.3563463	Kurtosis	-0.476039
Uncorrected SS	60	Corrected SS	60
Coeff Variation	.	Std Error Mean	0.13018891

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.963561	Pr < W	0.0703
Kolmogorov-Smirnov	D	0.104238	Pr > D	0.1013
Cramer-von Mises	W-Sq	0.096905	Pr > W-Sq	0.1242
Anderson-Darling	A-Sq	0.707668	Pr > A-Sq	0.0645



Atendiendo a los resultados, podemos afirmar que sí se verifica la hipótesis de normalidad de los residuos.

1.3. Homocedasticidad de los errores

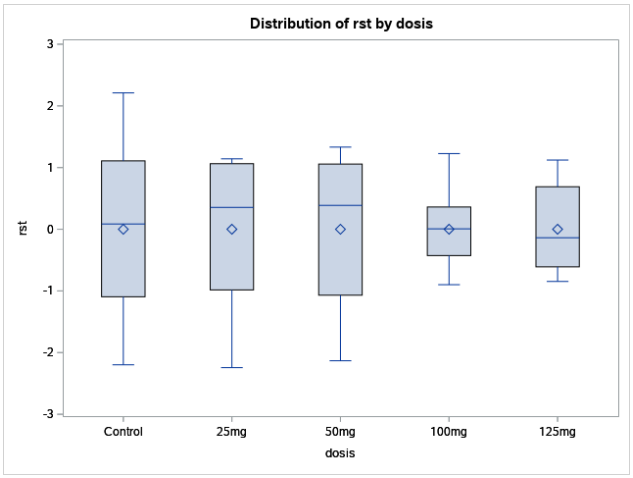
Para determinar si se verifica la homocedasticidad de los errores, esto es, si la varianza de los residuos es constante y no varía en los diferentes niveles del factor, podemos calcular la varianza de los residuos según los niveles del factor. No obstante, los gráficos de puntos y de cajas realizados para determinar la bondad del ajuste, ya nos indicaban que la varianza de los residuos sí que varía según el factor.

Nivel del factor	Varianza de los residuos
Control	279,5379
25mg	224,8182
50mg	238,6288
100mg	60,6288
125mg	76,3864

Comprobamos así que NO se verifica la hipótesis de homocedasticidad de los errores. Dado que los tamaños de los grupos son iguales, la heterocedasticidad no afecta al F-test ni a los distintos métodos de comparaciones múltiples siempre que la razón entre la varianza máxima y la mínima sea menor que 3. En este caso la razón entre la varianza máxima y la mínima es 4.217, por lo que la falta de homocedasticidad si afectará al F-test.

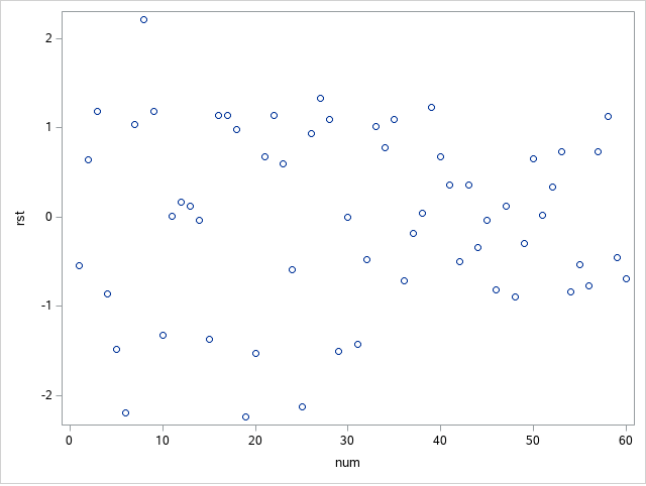
1.4. Homogeneidad de la muestra

Observando el gráfico de cajas de los residuos estandarizados según los valores del factor generado anteriormente, podemos concluir que no existen datos atípicos y que los errores son homogéneos.



1.5. Independencia de los errores

Para comprobar la independencia de los errores representamos los residuos estandarizados frente al índice.



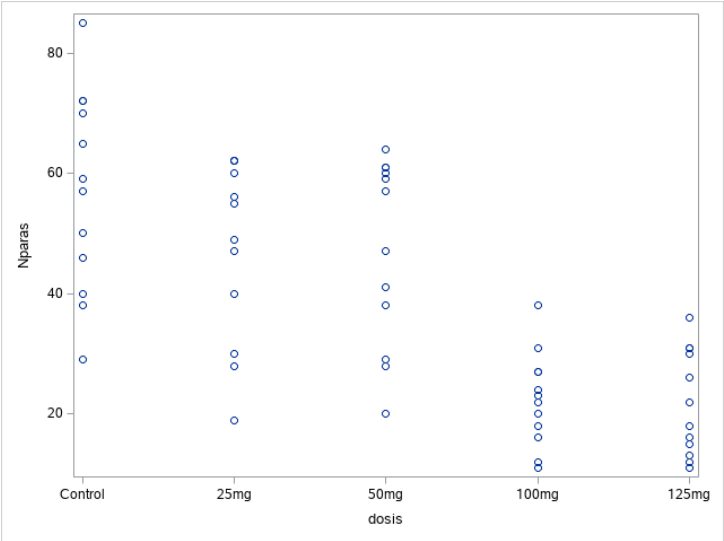
Vemos que no hay ninguna dependencia de las observaciones, por lo que podemos afirmar que SÍ se verifica la hipótesis de independencia.

1.6. Resumen

1. Bondad del ajuste del modelo estadístico propuesto. - No muy buena
2. La normalidad. - Sí
3. La homocedasticidad del error. - No
4. La homogeneidad de la muestra. - Sí
5. La independencia de las observaciones - Sí

2. Cuestiones de la práctica, a-f

a) Representa gráficamente los datos. ¿Te parece que el medicamento es efectivo contra los parásitos? Realiza un contraste de hipótesis para verificarlo.



Fijándonos en el gráfico de puntos obtenido en la sección 1.1, podemos ver como no parece haber una disminución significativa en el número de parásitos hasta que la dosis del medicamento es de 100 ó 125 mg. Podríamos decir que el medicamento es efectivo si se aplica una dosis elevada.

Para determinar si el medicamento es efectivo podemos plantear el siguiente contraste de hipótesis:

- **Ho**: todos los tratamientos tienen el mismo efecto en el número de parásitos, lo cual implica que no tratar con ningún medicamento tiene el mismo efecto que tratar con diferentes dosis.
- **H1**: no todos los tratamientos tienen el mismo efecto en el número de parásitos.

Lo cual se traduce en:

- **Ho**: media_control = media_25mg = media_50mg = media_100mg = media_125mg
- **H1:** alguna de las medias es diferente.

Aplicamos el test ANOVA para realizar el contraste de hipótesis planteado.

Dependent Variable: Nparas

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	12372.93333	3093.23333	17.58	<.0001
Error	55	9680.00000	176.00000		
Corrected Total	59	22052.93333			

Dado que el p-valor obtenido es <0,0001, podemos rechazar la hipótesis nula a casi cualquier nivel de confianza, es decir, existen evidencias para afirmar que no todos los tratamientos tienen el mismo efecto en el número de parásitos.

b) Obtén estimadores puntuales e intervalos de confianza para las medias de los tratamientos, utilizando dos métodos distintos, y explica las ventajas e inconvenientes de cada uno de ellos.

Utilizamos el test T-Student y el test de Bonferroni para calcular los intervalos de las medias de los tratamientos, alpha igual a 0,05.

- T-Student

dosis	N	Mean	95% Confidence Limits	
Control	12	56.917	49.242	64.592
25mg	12	47.500	39.825	55.175
50mg	12	47.083	39.408	54.758
100mg	12	22.417	14.742	30.092
125mg	12	21.750	14.075	29.425

- Bonferroni

dosis	N	Mean	Simultaneous 95% Confidence Limits	
Control	12	56.917	46.698	67.135
25mg	12	47.500	37.282	57.718
50mg	12	47.083	36.865	57.302
100mg	12	22.417	12.198	32.635
125mg	12	21.750	11.532	31.968

Observamos como los intervalos proporcionados por Bonferroni tienen una menor amplitud.

c) Analiza las diferencias significativas entre pares de medias de los 5 grupos utilizando diferentes métodos, comentando y comparando los resultados. ¿Cambian las conclusiones si utilizas alfa= 0.1?

Los métodos **LSD, Bonferroni, Tukey** y **Scheffe** coinciden en que existen diferencias significativas entre las siguientes medias con una confianza del 95%:

- Control y 100mg
- Control y 125mg
- 25mg y 100mg
- 25mg y 125mg
- 50mg y 100mg
- 50mg y 125mg.

Al cambiar el alpha de 0.05 a 0.10, el método LSD determina que también existen diferencias significativas entre Control y 25mg, y entre Control y 50mg. El resto de métodos no varían sus resultados.

d) Realiza el test de Dunnett para comparar los efectos de las distintas dosis con el grupo “Control “ y comenta el resultado.

Comparisons significant at the 0.05 level are indicated by ***.

dosis Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
25mg - Control	-9.417	-23.036	4.202	
50mg - Control	-9.833	-23.452	3.786	
100mg - Control	-34.500	-48.119	-20.881	***
125mg - Control	-35.167	-48.786	-21.548	***

El resultado del test de Dunnet es coherente con los gráficos obtenidos previamente. Con una confianza del 95%, existen diferencias significativas entre no aplicar ningún tratamiento (Control) y aplicar un tratamiento de 100 ó 125 mg; sin embargo, no existen diferencias significativas entre no aplicar ningún tratamiento y aplicar un tratamiento de 25 ó 50 mg.

e) Se desea comparar el efecto de las dosis bajas del medicamento, -25-50 mg- con el de las dosis altas, -100-125 mg.- Construye un intervalo de confianza para la diferencia entre ambos efectos e interpreta el resultado obtenido.

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
25-50 vs. 100-125	-17.3750000	3.82970843	-4.54	<.0001	-25.0499072	-9.7000928

Dado que el intervalo de confianza obtenido no contiene al 0, podemos afirmar que existen evidencias para decir que los efectos de los tratamientos de 25 y 50 mg son distintos a los efectos de los tratamientos de 100 y 125 mg con una confianza del 95%.

f) Verifica las hipótesis del modelo con ayuda de gráficos de residuos.

Realizados en la sección 1. Chequeo de las hipótesis del modelo.

Código SAS

```
*------------ Ejercicio 2 --------------------;

* 1. Importación de los datos;

proc import
datafile='/folders/myfolders/Datos/Datos_ejercicios_ANOVA_1F/datatab_6_29.xls'
        out=mice replace
        DBMS=xls;
        sheet="mice";
        getnames=yes;
run;

proc print; run;


* a) Estudia si este factor influye en el tiempo correspondiente;



proc anova data=mice;
class color;
model time=color;
run;




* b);

means color/DUNCAN BON TUKEY;
run;


*c) ;

means color/DUNNETT ('grn');
run;


* ---------------------  Acuicultura ----------------------------;

* Importar los datos;

data peces;
input dosis $;
do i=1 to 12;
input Nparas;
output;end;
cards;
Control
50
```

65
72
46
38
29
70
85
72
40
57
59
25mg
49
47
30
62
62
60
19
28
56
62
55
40
50mg
20
59
64
61
28
47
29
41
60
57
61
38
100mg
20
23
38
31
27
16
27
18
22
12
24
11
125mg
18
30
22
26
31

```
11
15
12
31
36
16
13
;
run;

* 1.Chequeo de las hipótesis del modelo;

** 1.1. Estadísticos básicos de la variable respuesta según el factor;

proc means data=peces;
var Nparas;
class dosis;
run;

*** Gráfico de puntos;
proc sgplot data=peces;
scatter y=Nparas x=dosis;
run;

*** Gráfico de cajas múltiple;
proc boxplot data=peces;
plot Nparas*dosis;
run;


** 1.2. Ajuste del modelo.

*** Cálculo de los valores predichos, los residuos y de los residuos estandarizados;
proc glm data=peces;
class dosis;
model Nparas=dosis;
output out=resal p=pred r=res student=rst;
proc print data=resal;
run;

*** Representación de los residuos estandarizados según los niveles del factor;
proc sgplot data=resal;
scatter y=rst x=dosis;
run;

proc boxplot data=resal;
plot rst*dosis;
run;



** 1.3. Normalidad de los errores;

proc univariate plot normal data=resal;
var rst;
run;
```

```
** 1.4. Homocedasticidad de los errores;

*** Obtenemos las varianzas de los errores según el nivel del factor;
proc univariate data=resal;
var res;
class dosis;
run;



** 1.5. Independencia de los errores;

data a;
set resal;
num=_n_;
run;
proc sgplot;
*scatter  y=Nparas x=num;
scatter  y=rst x=num;
run;quit;

*-------------------------------------------------------------------------------------;

* 2. Cuestiones;
** a);

*Gráfico de puntos obtenido en la sección 1.1;

proc anova data=peces;
class dosis;
model Nparas=dosis;
run;

** b);

means dosis/ T  BON CLM;
run;


** c);
means dosis/ LSD BON TUKEY SCHEFFE CLDIFF;
run;


means dosis/ LSD BON TUKEY SCHEFFE CLDIFF alpha=0.1;
run;

** d);

means dosis/DUNNETT ('Control');
run;
```

```
** e);
proc glm data =peces;
class dosis;
model Nparas=dosis/solution clparm;
estimate '25-50 vs. 100-125' dosis 0 0.5 .5 -.5 -.5;
run;
```