Regresión y ANOVA: Pulgones [*]

García Prado, Sergio
sergio@garciparedes.me

23 de octubre de 2017

1. Descripción del conjunto de datos

El conjunto de datos sobre el cual se va a realizar el análisis de la varianza se refiere a una serie de mediciones sobre el número de pulgones por planta de trigo. El experimento fue realizado recogiendo 40 plantas (muestras aleatorias que supondremos independientes) de trigo, durante un periodo de 6 semanas.

Para la realización de este análisis se ha utilizado la plataforma SAS [2], en concreto la *University Edition*. En este caso, el conjunto de datos ha sido suministrado en forma de fragmento de código, el cual se incluye en la figura 28. El conjunto de datos sigue una estructura tabular de 240 filas (referidas a cada observación) y 3 columnas (referidas a la `semana`, identificador de `muestra` en esa semana y `recuento` de pulgones en dicha observación) tal y como se muestra en la figura 1. El código *SAS* utilizado en este caso se muestra en la figura 29.

Obs	semana	repet	recuento
1	1	1	12
2	1	2	1
3	1	3	6
4	1	4	1
5	1	5	5

Figura 1: Visión preliminar del conjunto de datos `pulgones`

2. Cuestiones

El objetivo general del estudio es el siguiente: **"Se trata de analizar si existen diferencias en el número de pulgones por planta entre las diferentes semanas"**, para lo cual se proponen una seria de sub-objetivos que se tratarán de responder en las siguientes secciones.

2.1. ¿Es adecuado utilizar un modelo de un factor para ello? Haz un análisis descriptivo de los datos por semanas y valora las hipótesis que se asumen en el modelo.

Para poder responder a la pregunta sobre si es adecuado utilizar un modelo de un factor para la comparación del número de pulgones por planta entre las distintas semananas, es necesario plantearse cómo han sido recogidas las observaciones, así como estudiar la distribución de la variable respuesta (`recuento` en este caso)

[*] URL: `https://github.com/garciparedes/anova-pulgones`

En el primer caso, se presupone que las 40 observaciones referidas a cada semana han sido elegidas siguiendo algún procedimiento aleatorio. Por tanto, asumiremos como válida la hipótesis de independencia de las observaciones para la realización del análisis de un factor.

En cuanto a la hipótesis de normalidad, es decir, el estudio acerca de que los valores de la variable recuento siguen una distribución aproximadamente normal, se ha realizado un estudio descriptivo sobre los mismos. Para dicha tarea se ha utilizado el fragmento de código incluido en la figura 30. A partir de esta sentencia se han obtenido los resultados mostrados en la tabla de las figuras 2, 3, 4, 5, 6 y 7, así como los *Histogramas* incluidos en las figuras 8, 9 y 10 y los *Gráficos de Normalidad* incluidos en las figuras 11, 12 y 13.

Tras el análisis de los mismos podemos empezar a sospechar acerca de una cierta falta de normalidad en los datos, concretamente en los referidos a las de la categoria `semana` 2 y 3. Esto se ve reflejado en los distintos valores de los tests de normalidad, así como en la apariencia de los histogramas y la marcada desviación respecto de la recta de la distirbución normal en los diagramas de *cuantil-cuantil* respecto de la normal

A continuación, se ha realizado un *box-plot* de la variable `recuento` divida en las categorías marcadas por `semana`, la cual se puede apreciar en la figura 14. En este caso, se puede apreciar una clara diferencia entre las medias de las semanas 2 y 3 con respecto del resto, lo cual también sucede a nivel de la varianza.

Por tanto, se ve reforzada la idea comentada anteriormente de falta de normalidad en el conjunto de datos. Esto hace que los contrastes de hipótesis acerca de la igualdad de medias (que asumen una distribución de varianzas igual para todos los factores) se vea gravemente comprometida. Esto hace que los resultados que se obtengan en futuras secciones tengan que ser tomados con gran prudencia.

Otro suceso a remarcar es el carácter temporal de los datos, que a pesar de no estar siendo estudiado en este trabajo, podría tener una alta influencia en los resultados, ya el particionamiento en categorías se está llevando a cabo sobre un soporte temporal (`semana`). Sin embargo, la descripción del conjunto de datos no especifica que estas hayan quedado descritas de una manera secuencial, es decir, no se ha determinado que existe un orden fijado por el valor de `semana`. Por tanto, es necesario tomar conclusiones prudentes en este sentido.

2.2. Realiza el contraste de igualdad de medias y analiza los residuos. ¿Qué conclusiones sacas?

Para la realización de un contraste de hipótesis referido a la igualdad de medias, se ha decidido utilizar el fragmento de código incluido en la figura 32. Este fragmento de código realiza un contraste de igualdad de medias, y además suministra una serie de gráficos relacionados con los residuos generados por el contraste. Esto es equivalente a lo descrito en la siguiente ecuación:

$$\begin{aligned} H_0: &\quad \forall i,j \quad \mu_i = \mu_j \\ H_1: &\quad \exists i,j \quad \mu_i \neq \mu_j \end{aligned}$$
$$i,j \in \{semana_1, ..., semana_6\} \wedge i \neq j$$

El contrate de igualdad de medias ha generado los resultados obtenidos en la figura 15, que tal y como se puede apreciar a partir del resultado del *p-valor*, la hipótesis nula (de igualdad de medias) es rechazada. Un paso intuitivo en este paso es realizar algún test múltiple como *Bonferroni* o *Tukey*.

The UNIVARIATE Procedure
semana = 1
Fitted Normal Distribution for recuento

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	5.225
Std Dev	Sigma	5.993533

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.22973908	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.47080889	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	2.78848176	Pr > A-Sq	<0.005

Figura 2

The UNIVARIATE Procedure
semana = 2
Fitted Normal Distribution for recuento

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	31.35
Std Dev	Sigma	20.2732

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.10994832	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.10627489	Pr > W-Sq	0.092
Anderson-Darling	A-Sq	0.76500683	Pr > A-Sq	0.044

Figura 3

The UNIVARIATE Procedure
semana = 3
Fitted Normal Distribution for recuento

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	31.95
Std Dev	Sigma	24.476

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.13175010	Pr > D	0.080
Cramer-von Mises	W-Sq	0.13592100	Pr > W-Sq	0.037
Anderson-Darling	A-Sq	0.87715030	Pr > A-Sq	0.023

Figura 4

The UNIVARIATE Procedure
semana = 4
Fitted Normal Distribution for recuento

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	4.825
Std Dev	Sigma	8.224066

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.29570976	Pr > D	<0.010
Cramer-von Mises	W-Sq	1.17631596	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	6.12733807	Pr > A-Sq	<0.005

Figura 5

The UNIVARIATE Procedure
semana = 5
Fitted Normal Distribution for recuento

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	1.5
Std Dev	Sigma	4.157169

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.35911607	Pr > D	<0.010
Cramer-von Mises	W-Sq	1.61245216	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	8.19021656	Pr > A-Sq	<0.005

Figura 6

The UNIVARIATE Procedure
semana = 6
Fitted Normal Distribution for recuento

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	0.525
Std Dev	Sigma	1.484752

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.4631786	Pr > D	<0.010
Cramer-von Mises	W-Sq	2.1361712	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	10.3687426	Pr > A-Sq	<0.005

Figura 7

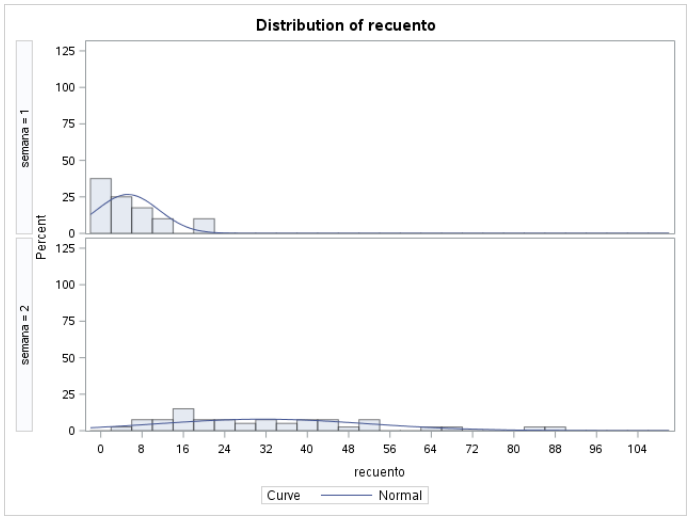

Figura 8: Histograma: Semanas 1 y 2

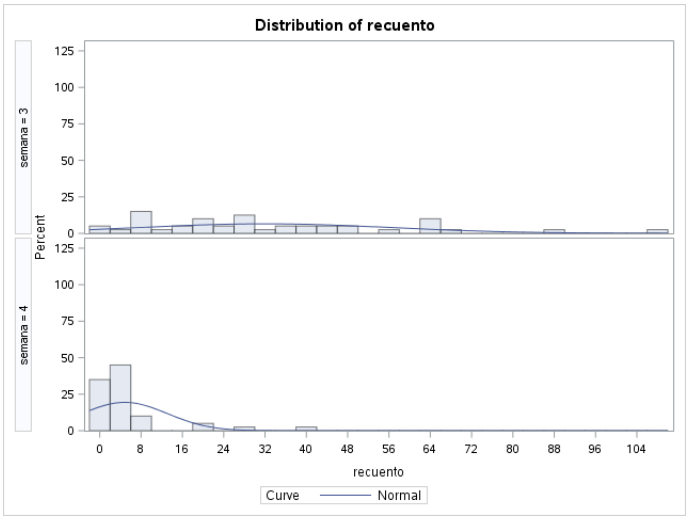

Figura 9: Histograma: Semanas 3 y 4

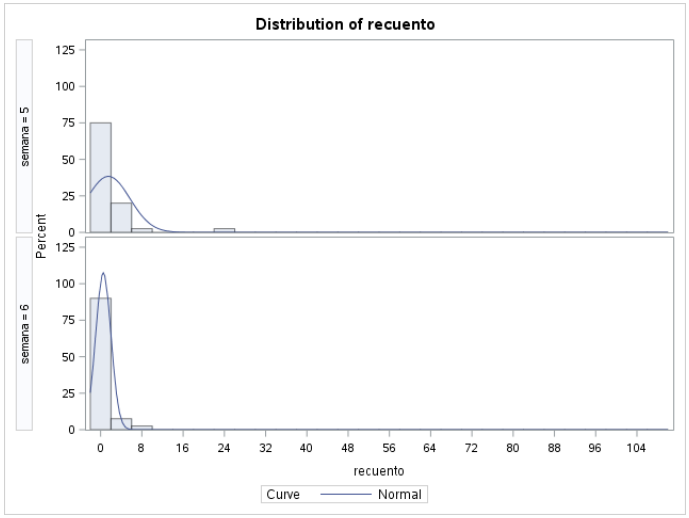

Figura 10: Histograma: Semanas 5 y 6

Sin embargo, en este caso tras examinar los gráficos de residuos nos damos cuenta de que la hipótesis de igualdad de medias podría haber sido rechazada debido a la falta de normalidad de los datos. Esto se puede apreciar a través de los gráficos de residuos y residuos Studentizados de las figuras 16 y 17. En los estos se aprecia la falta de normalidad, ya que los errores (en ambos casos) están claramente sesgados hacia valores positivos (algo que no debería ocurrir). En el caso de los residuos studentizados, estos superan el valor 2 en un gran número de casos, algo que no debería ocurrir en bajo la presunción de normalidad de residuos. Otro fenómeno que se puede apreciar en estos datos es la existencia de heterocedasticidad entre las distintas categorías, que surge entre las semanas $\{1, 4, 5, 6\}$ y las semanas $\{2, 3\}$.

Para confirmar la falta de normalidad en los residuos, se puede visualizar el gráfico *cuantil-cuantil* de normalidad que se muestra en la figura 18. En este se puede apreciar una distribución que una vez más hace pensar en la inexistencia de normalidad.

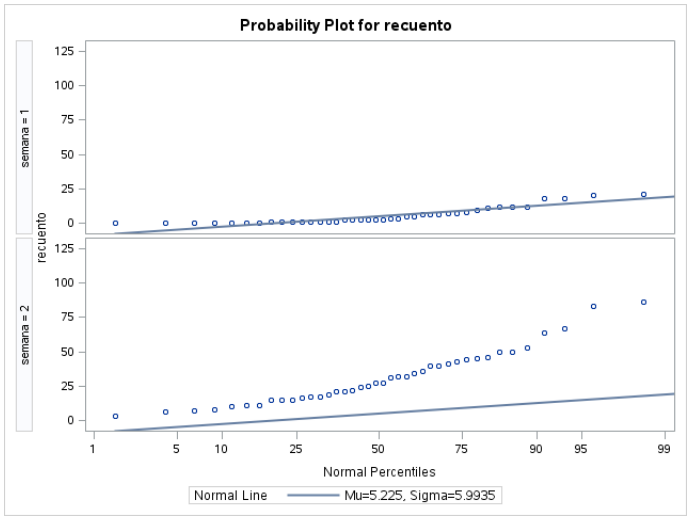

Figura 11: Gráfico de Normalidad: Semanas 1 y 2

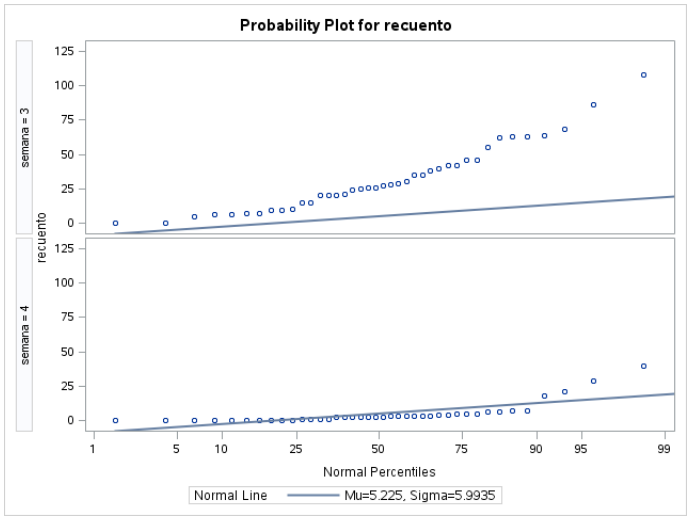

Figura 12: Gráfico de Normalidad: Semanas 3 y 4

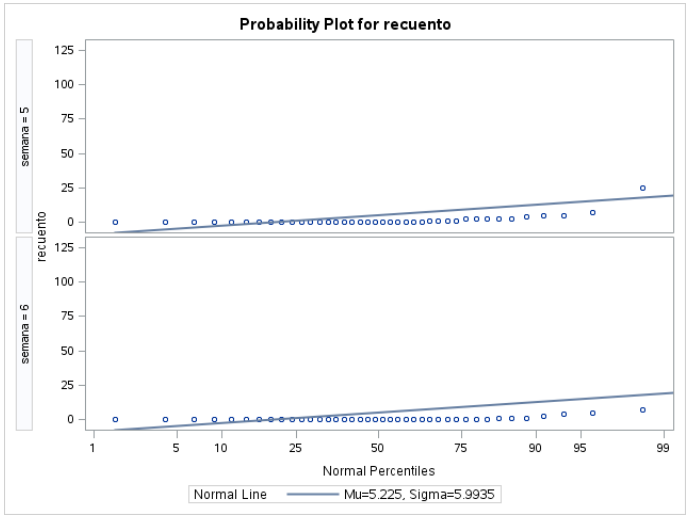

Figura 13: Gráfico de Normalidad: Semanas 5 y 6

2.3. Realiza el test de *Levene*. ¿Te sorprende el resultado?

El código utilizado para la realización del test de *Levene* se muestra en la figura 33. Este test se refiere a la realización del contrate de hipótesis de igualdad de varianzas entre distintas poblaciones. Por tanto, se puede modelizar como:

$$\begin{aligned} H_0: &\quad \forall i,j \quad \sigma_i^2 = \sigma_j^2 \\ H_1: &\quad \exists i,j \quad \sigma_i^2 \neq \sigma_j^2 \end{aligned}$$

$$i, j \in \{semana_1, ..., semana_6\} \wedge i \neq j$$

Los resultados obtenidos se muestran en la figura 19. Tras el valores, se puede apreciar que debido al reducido valor que toma el *p-valor*, tenemos que rechazar la hipótesis de igualdad de varianzas.

Estos resultados no sorprenden debido a las conclusiones que se habían obtenido previamente, tanto a partir de los residuos, como del *box-plot* de la figura 14, que se realizó al principio del estudio para la obtención de

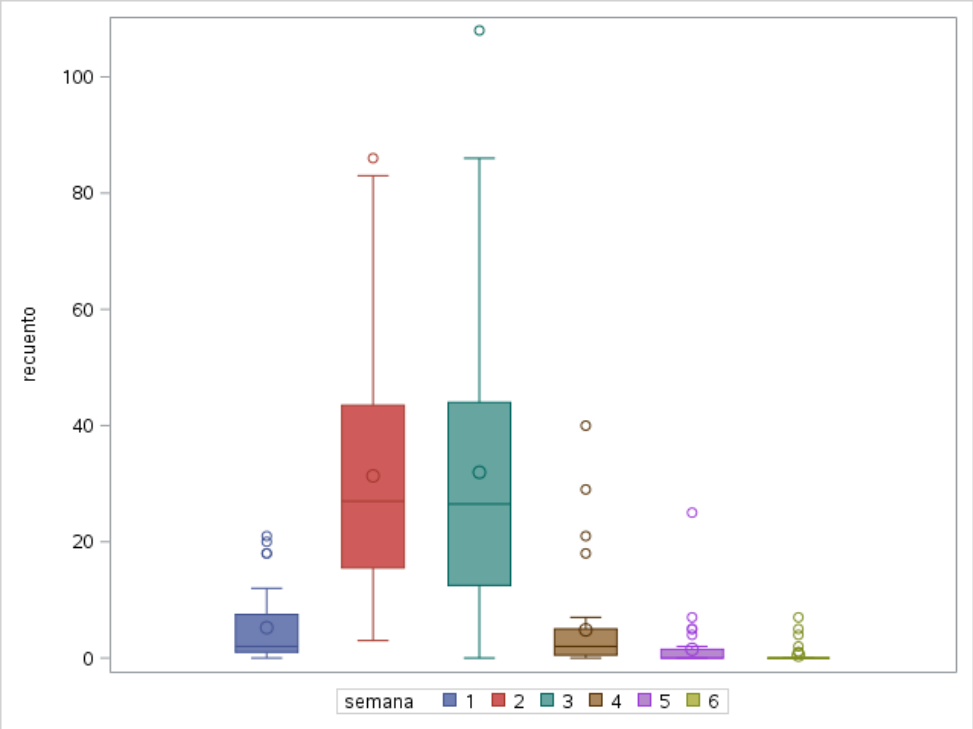

Figura 14: *Box-plot*: de la variable `recuento` clasificada por `semana`

The GLM Procedure
Dependent Variable: recuento

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	44393.33750	8878.66750	47.01	<.0001
Error	234	44191.72500	188.85353		
Corrected Total	239	88585.06250			

R-Square	Coeff Var	Root MSE	recuento Mean
0.501138	109.3922	13.74240	12.56250

Source	DF	Type I SS	Mean Square	F Value	Pr > F
semana	5	44393.33750	8878.66750	47.01	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
semana	5	44393.33750	8878.66750	47.01	<.0001

Figura 15: Resultados contraste de hipótesis de igualdad de medias

una visión global de los datos. A partir de estos se podía intuir la falta de igualdad de varianzas, que tras el test de *Levene* se ha demostrado de manera analítica.

2.4. Transforma la respuesta mediante $log(recuento + 1)$ y repite el apartado 2.2. ¿Qué cambios observas?

Para la realización de la transformación del conjunto de datos se ha utilizado el framento de código incluido en la figura 34, que además genera un *box-plot* a través del cual se obtiene una visión global acerca de la distribución de los datos. Este se incluye en la figura 20.

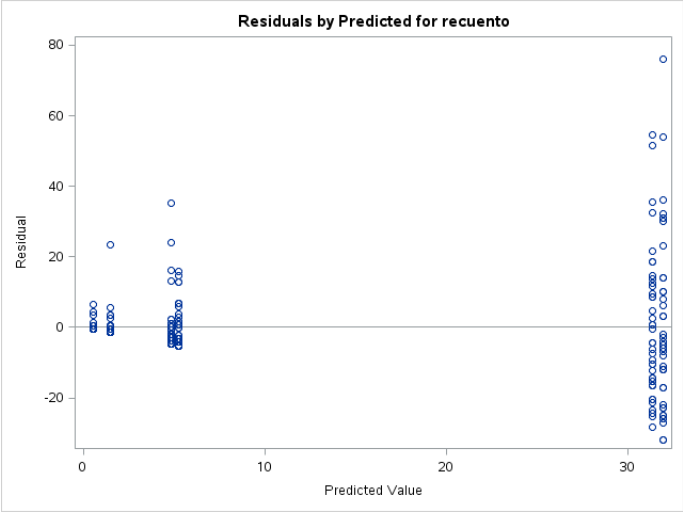

Figura 16: Scatter Plot: Residuos del contraste de hipótesis de igualdad de medias

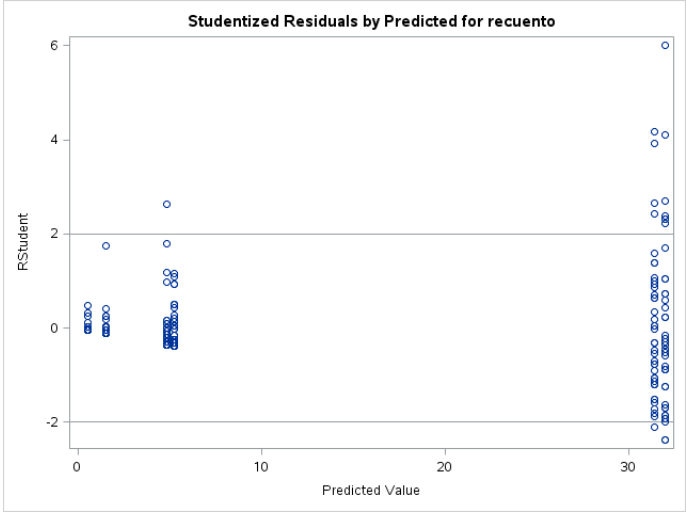

Figura 17: Scatter Plot: Residuos Studentizados del contraste de hipótesis de igualdad de medias

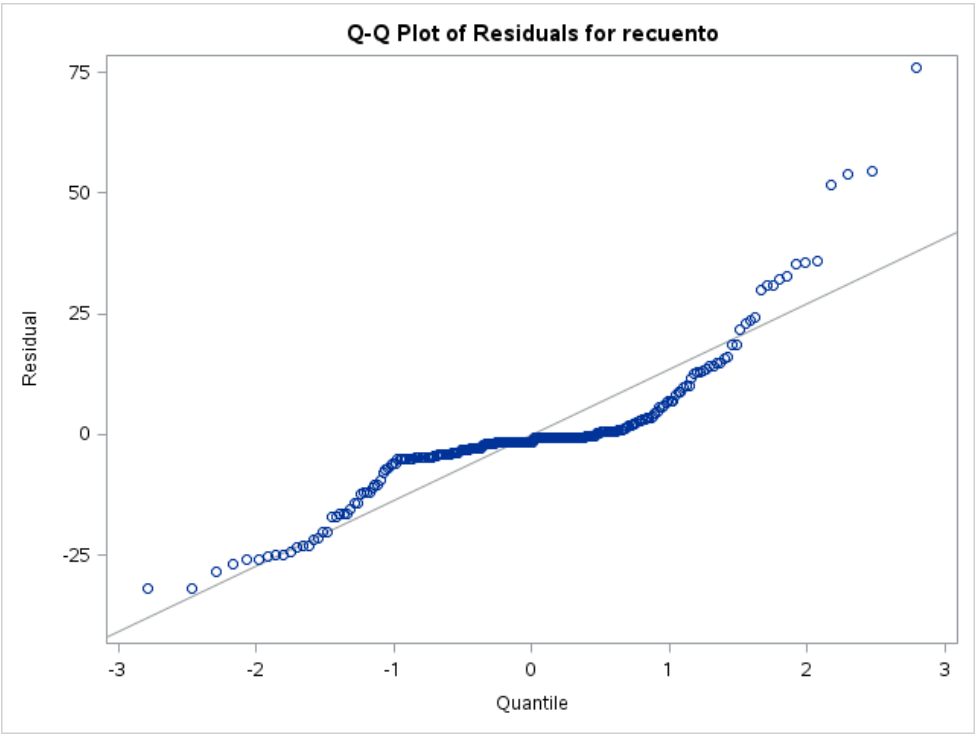

Figura 18: Gráfico de Normalidad: Residuos del contraste de hipótesis de igualdad de medias

Levene's Test for Homogeneity of recuento Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
semana	5	12167534	2433507	9.93	<.0001
Error	234	57348832	245080		

Figura 19: Resultados del test de Levene de la variable `recuento` particionada por `semana`

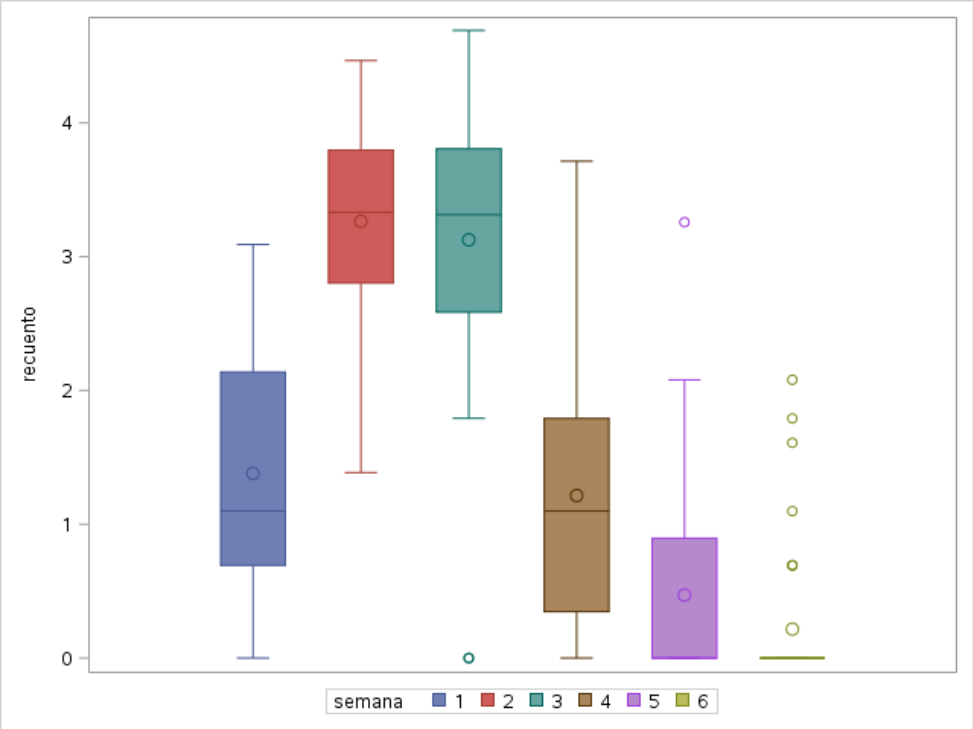

Figura 20: *Box-plot*: de la variable transformada.

Una vez transformada la variable, el siguiente paso a realizar es realizar el test de igualdad de medias sobre los datos transformados siguiendo la misma metodología que en la sección 2.2. En este caso, además del test de igualdad de medias, se ha realizado el test de *Levene* para comprobar si la transformación ha hecho que las medias de los distintos grupos sean más semejante. Por último, dado que esta hipótesis no ha podido rechazarse, como se verá a continuación, se ha realizado un test de *Tukey* para estudiar qué grupos son diferentes entre sí. Se ha preferido este test puesto que en este caso era posible su utilización (todos los grupos tienen el mismo número de observaciones) además de ser un método potente desde el punto de vista de las diferencias que es capaz de detectar. Estas tareas se han realizado utilizando el bloque de código incluido en la figura 36.

Los resultados del test de igualdad de medias conjunto sobre la variable transformada se han incluido en la figura 27. Tal y como se puede apreciar, en este caso también se rechaza la hipótesis de igualdad de medias debido a que el *p-valor* es muy reducido.

Sin embargo, en este caso debemos tomar los resultados con un mayor grado de seguridad, ya que tal y como se puede apreciar en los gráficos de residuos y residuos studentizados de las figuras 22 y 23, en este caso estos siguen una distribución más uniforme. A pesar de no estar distribuidos de manera uniforme, se puede apreciar una clara corrección con respecto de los valores sin transformar.

Esto también se ve reflejado en el gráfico de normalidad de los errores, el cual se muestra en la figura 24. En este caso ocurre lo mismo que en el caso de los gráficos de residuos: a pesar de que la corrección de normalidad no es ideal, esta diferencia se ve corregida en gran medida. Estos resultados incitan a pensar que ya se puede realizar interpretaciones con un mayor grado de seguridad. Sin embargo, para estar más seguros se ha realizado un test de *Levene*.

The GLM Procedure
Dependent Variable: recuento

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	339.4187667	67.8837533	93.18	<.0001
Error	234	170.4676155	0.7284941		
Corrected Total	239	509.8863823			

R-Square	Coeff Var	Root MSE	recuento Mean
0.665675	52.95501	0.853519	1.611781

Source	DF	Type I SS	Mean Square	F Value	Pr > F
semana	5	339.4187667	67.8837533	93.18	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
semana	5	339.4187667	67.8837533	93.18	<.0001

Figura 21: Resultados contraste de hipótesis de igualdad de medias de la variable transformada.

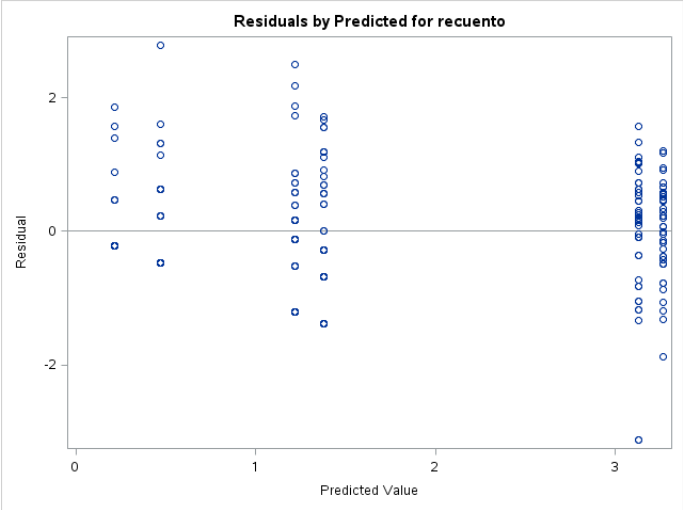

Figura 22: Scatter Plot: Residuos del contraste de hipótesis de igualdad de medias de la variable transformada.

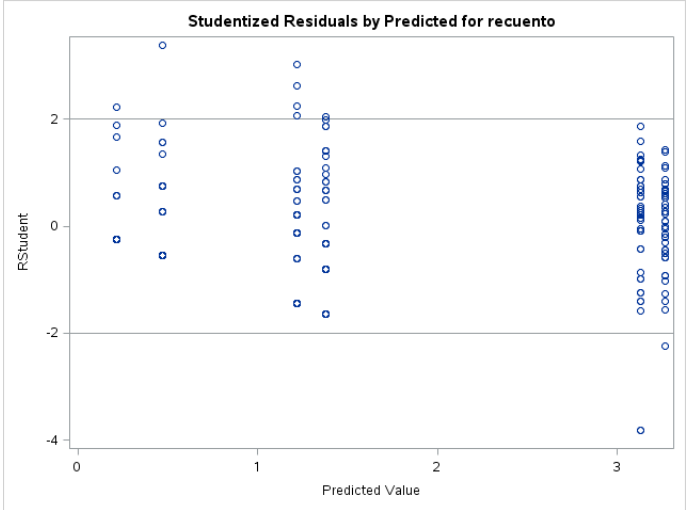

Figura 23: Scatter Plot: Residuos Studentizados del contraste de hipótesis de igualdad de medias de la variable transformada.

Tras la realización del test de *Levene* se puede comprobar que ahora la hipótesis de igualdad de medias toma un valor marcadamente mayor que en el caso de los datos sin transformar. A pesar de que este no sea muy elevado, nos da buenos indicios acerca de la existencia de normalidad. Los resultados del test se incluyen en la figura 25.

Por último, se ha realizado un test de *Tukey* para conocer qué grupos de medias son diferentes respecto del resto. Tras realizar dicho experimento se han obtenido los resultados de la figura 26, que indican que se puede afirmar con un grado de confianza del 95 % que existen tres grupos de medias, formados por las semanas $\{1, 4\}$, $\{2, 3\}$ y $\{5, 6\}$. Esta idea ya se podía apreciar a partir del *box-plot*, sin embargo, ahora ha sido confirmada de manera analítica.

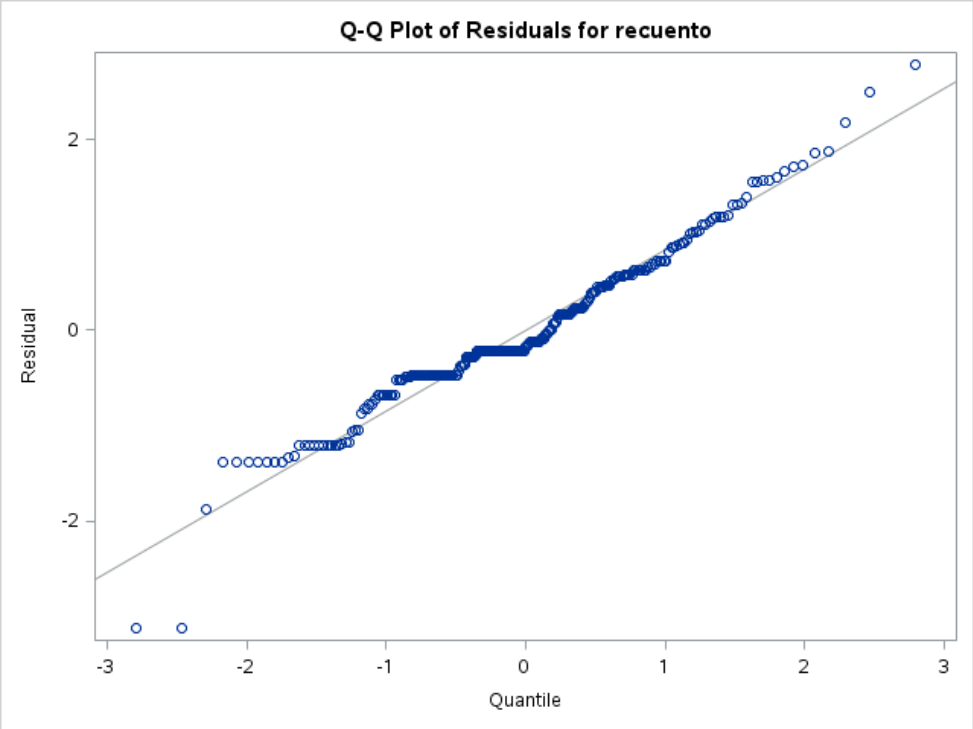

Figura 24: Gráfico de Normalidad: Residuos del contraste de hipótesis de igualdad de medias de la variable transformada.

Levene's Test for Homogeneity of recuento Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
semana	5	19.3589	3.8718	2.40	0.0379
Error	234	377.3	1.6124		

Figura 25: Resultados del test de Levene de la variable transformada.

2.5. Realiza el test de *Kruskal-Wallis* sobre los datos originales para contrastar la igualdad de medias

En esta sección se realiza un test no paramétrico de igualdad de medias sobre el conjunto de datos originales. Los tests no paramétricos se basan en el estudio de la distribución sin el apoyo en parámetros de distribuciones adicionales. En este caso el test utilizado es el de *Kruskal-Wallis*. Para ello, se ha utilizado el bloque de código incluido en la figura 36.

Los resultados obtenidos se han incluido en la figura 27. A partir de estos se llega a la misma conclusión que se hacía anteriormente: la hipótesis de igualdad de medias debe ser rechazada.

The GLM Procedure
Bonferroni (Dunn) t Tests for recuento

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	234
Error Mean Square	0.728494
Critical Value of t	2.96564
Minimum Significant Difference	0.566

Means with the same letter are not significantly different.

Bon Grouping	Mean	N	semana
A	3.2641	40	2
A			
A	3.1259	40	3
B	1.3798	40	1
B			
B	1.2144	40	4
C	0.4699	40	5
C			
C	0.2165	40	6

Figura 26: Resultados test de *Tukey* de la variable transformada.

Wilcoxon Scores (Rank Sums) for Variable recuento Classified by Variable semana

semana	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	40	4568.00	4820.0	393.642469	114.20000
2	40	7821.00	4820.0	393.642469	195.52500
3	40	7541.50	4820.0	393.642469	188.53750
4	40	4248.00	4820.0	393.642469	106.20000
5	40	2654.50	4820.0	393.642469	66.36250
6	40	2087.00	4820.0	393.642469	52.17500

Average scores were used for ties.

Kruskal-Wallis Test	
Chi-Square	155.7551
DF	5
Pr > Chi-Square	<.0001

Figura 27: Resultados test de *Kruskal-Wallis* la variable `recuento` particionada por `semana`

3. Código fuente

```
data pulgones;
  do semana=1 to 6;
    do repet=1 to 40;
      input recuento @@;
      output;
    end;
  end;
  datalines;
  12 1 6 1 5 7 1 1 2 1 20 0 9 7 0 12 2 0 0 2 8 0 11 2 21 0 3 18 2 2 6 6
  5 1 12 0 3 1 1 18 40 16 32 15 44 41 43 53 67 21 6 31 15 11 21 40 15 50
  17 32 24 7 25 11 64 22 50 27 3 46 45 10 8 27 34 19 86 83 17 36 86 63
  20 68 55 42 24 29 20 27 26 63 40 46 7 15 10 30 46 26 15 42 6 28 7 9 5
  35 6 9 108 38 35 64 21 20 62 25 0 0 29 2 3 0 4 2 6 7 5 4 6 0 0 5 1 3 2
  2 2 5 0 1 1 0 3 1 2 0 3 3 18 7 21 0 0 0 2 3 0 40 5 7 0 0 0 1 1 2 1 0
  25 1 0 0 0 0 0 0 0 5 0 2 0 0 0 2 0 0 0 4 0 0 0 0 2 0 0 0 0 2 1 0 0 1 7
  0 0 0 4 1 5 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
;
run;
```

Figura 28: *Código SAS:* Lectura del conjunto de datos.

```
proc print data=pulgones (obs=5) n;
run;
```

Figura 29: *Código SAS:* Vista preliminar del conjunto de datos.

```
proc univariate data=pulgones;
   class semana;
   var recuento;
   probplot recuento / normal
                     (mu=est sigma=est color=blue w=1);
run;
```

Figura 30: *Código SAS:* Estudio Descriptivo de la variable `recuento` particionada por `semana`.

```
proc sgplot data=pulgones;
  vbox recuento /  group=semana;
run;
```

Figura 31: *Código SAS: Box-Plot* de la variable `recuento` particionada por `semana`.

```
proc glm data=pulgones PLOTS(UNPACK)=DIAGNOSTICS;
  class semana;
  model recuento=semana;
run;
```

Figura 32: *Código SAS:* Contraste de hipótesis de medias (con obtención de gráficos referidos a residuos).

```
proc glm data=pulgones;
  class semana;
  model recuento=semana;
  means semana / hovtest=levene;
run;
```

Figura 33: *Código SAS:* Test de *Levene*

```
data pulgones_log;
  set pulgones;
  recuento=log(recuento + 1);
run;

proc sgplot data=pulgones_log;
  vbox recuento /  group=semana;
run;
```

Figura 34: *Código SAS:* Transformación del conjunto de datos mediante $log(recuento + 1)$.

```
proc glm data=pulgones_log PLOTS(UNPACK)=DIAGNOSTICS;
  class semana;
  model recuento=semana;
  means semana / hovtest=levene tukey;
run;
```

Figura 35: *Código SAS:* Test de igualdad de medias sobre datos transformados.

```
proc npar1way data=pulgones wilcoxon;
  class semana;
  var recuento;
run;
```

Figura 36: *Código SAS:* Test de *Kruskal-Wallis*.

Referencias

[1] Barba Escribá, L. Regresión y ANOVA, 2017/18. Facultad de Ciencias: Departamento de Estadística.

[2] SAS® Software Institute. Sas. https://www.sas.com/.