Regresión y ANOVA: Pulgones [*]

García Prado, Sergio
sergio@garciparedes.me

22 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 14. 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 15.

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

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

recogidas las observaciones, así como estudiar la distribución de la variable respuesta (`recuento` en este caso)

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 16. A partir de esta sentencia se han obtenido los resultados mostrados en la tabla de las figuras **??**, 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

[TODO]

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

[TODO]

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

[TODO]

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

[TODO]

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

[TODO]

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

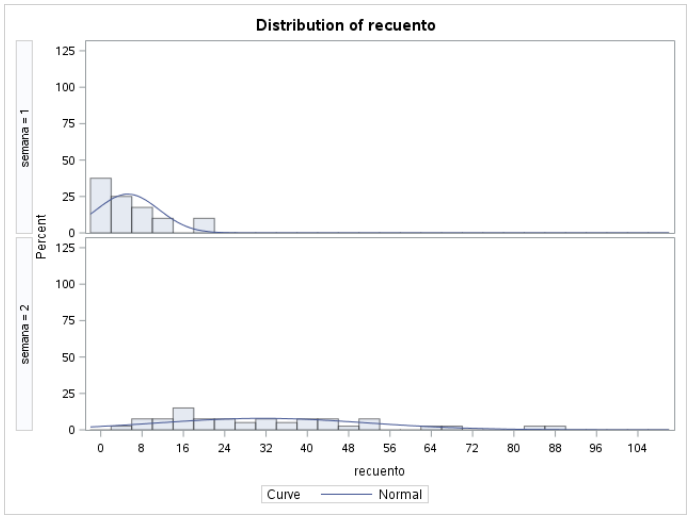

Figura 8: Histograma: Semanas 1 y 2

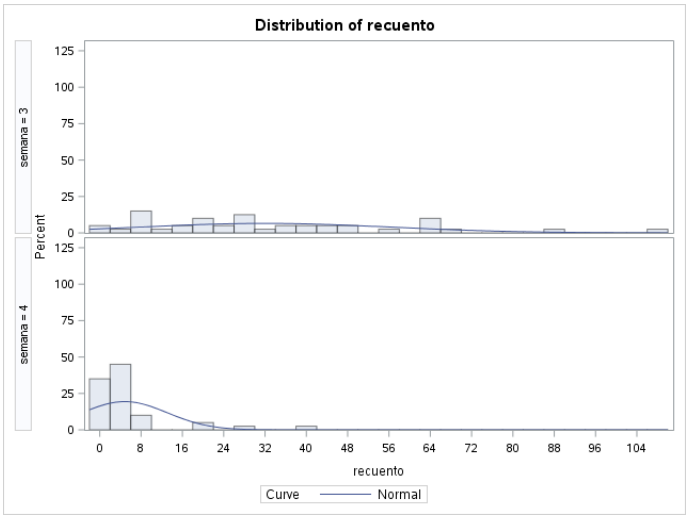

Figura 9: Histograma: Semanas 3 y 4

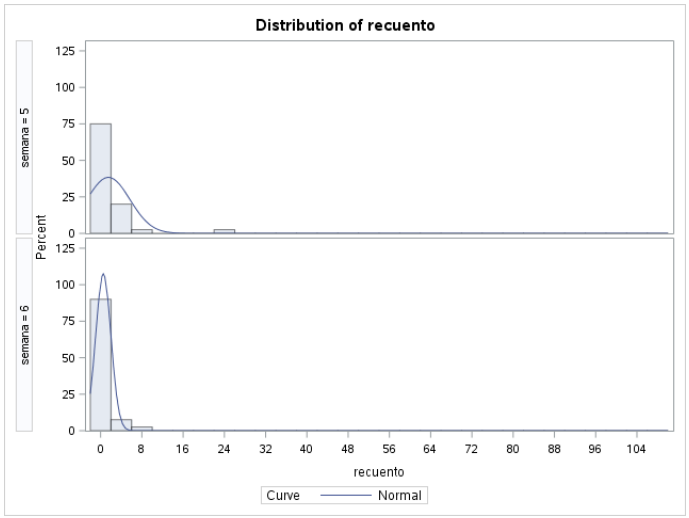

Figura 10: Histograma: Semanas 5 y 6

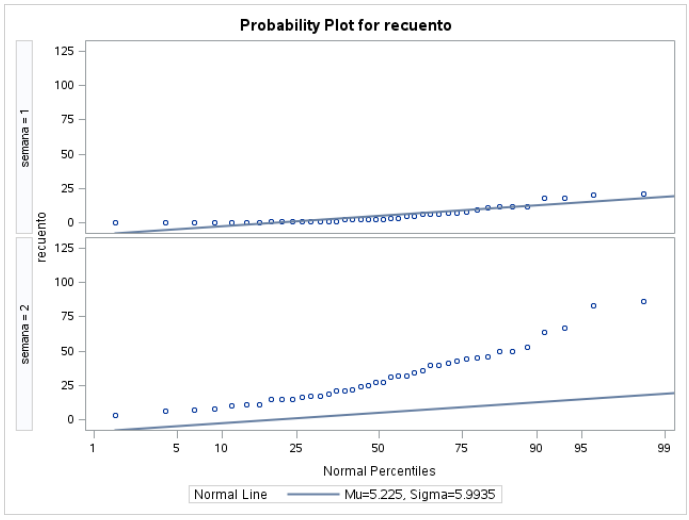

Figura 11: Gráfico de Normalidad: Semanas 1 y 2

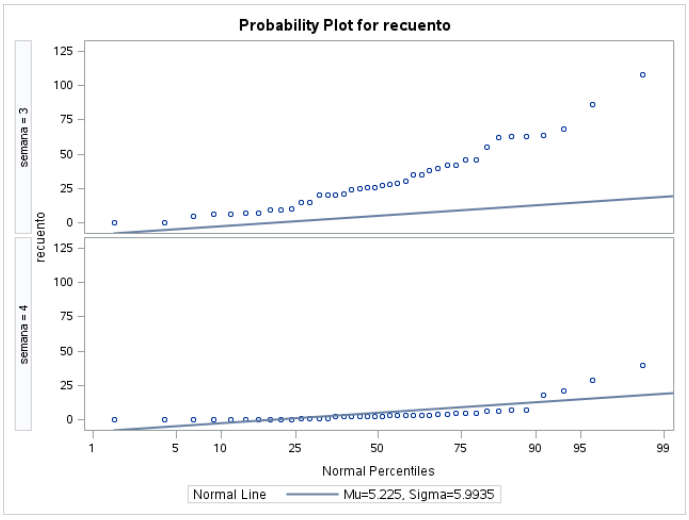

Figura 12: Gráfico de Normalidad: Semanas 3 y 4

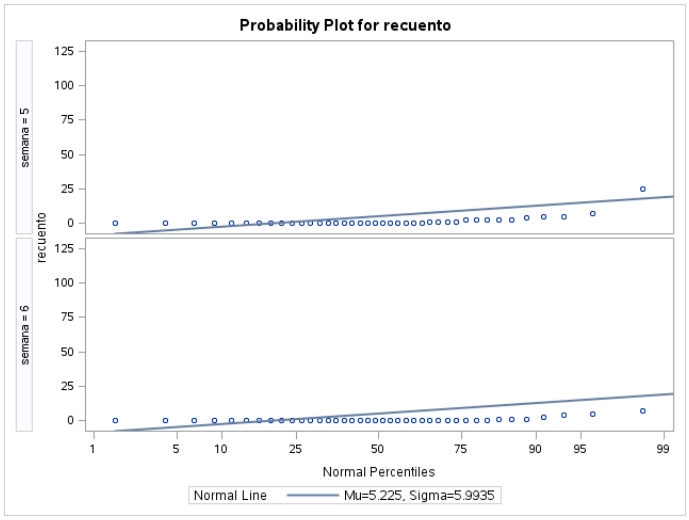

Figura 13: Gráfico de Normalidad: Semanas 5 y 6

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 14: *Código SAS:* Lectura del conjunto de datos.

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

Figura 15: *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 16: *Código SAS:* Estudio Descriptivo de la variable `recuento` particionada por `semana`.

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

Figura 17: *Código SAS: Box-Plot* de la variable `recuento` particionada por `semana`.

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/.