Distribución de Poisson para variables que no siguen la ley Normal (I)

Habitualmente se supone que las variables que recogemos, estudiamos, analizamos e intentamos modelizar dentro de una explotación porcina siguen la ley Normal o ley de Gauss. Pero puede ser que no sigan esta distribución.

Alberto Morillo Alujas[1], Daniel Villalba Mata[2] y Emilio López Cano[3]

[1]Tests and Trials SLU
[2]Universidad de Lleida
[3]Universidad Rey Juan Carlos (Madrid)

Hay que recordar que la ley Normal o de Gauss es una distribución de variable continua y que la gráfica de su función de densidad tiene una forma acampanada y es simétrica respecto de un parámetro estadístico. Esta distribución permite modelizar muchos fenómenos naturales, sociales, económicos, etc. Por ejemplo, usaremos la ley Normal para estudiar caracteres morfológicos de individuos, como el peso a determinada edad, el gasto económico por kilogramo de carne producida, la ganancia diaria de un grupo de animales, la proteína de la cebada, etc.

Otra característica muy importante de la ley Normal es que la distribución muestral de las medias muestrales sigue una distribución aproximadamente normal, aunque la distribución de la población de la que se extraen las muestras no sea normal. La forma de la distribución normal depende de su media y de su varianza. En la *figura 1* podemos ver varias funciones de densidad dependiendo de estos estadísticos.

Pero si pensamos detenidamente en las variables con las que trabajamos en una granja de cerdas, vemos que tienen otras características como:

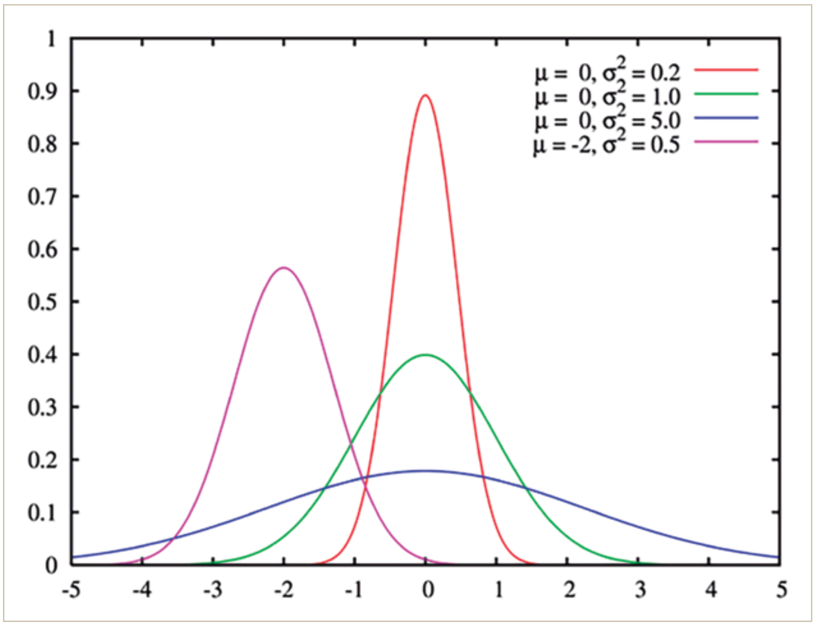

Figura 1. Distribución normal con varias medias y varianzas. La línea verde corresponde a la distribución estándar de media 0 y varianza 1.

- No son variables que tengan decimales, son valores discretos. Más concretamente, son contajes (en inglés se encuentra como "count data"), números enteros no negativos, como el número de lechones nacidos en un parto, el número de abortos en una semana, el número de repeticiones en una banda de cubrición o el número de lechones muertos por parto.
- Los sucesos que ocurren en un intervalo de tiempo determinado de tiempo son independientes.
- Estos contajes pueden ser expresados también como tasas, ya que la cantidad de veces que ocurre un evento dentro de un periodo de tiempo se puede expresar bien como un conteo sin procesar (ha habido 3 abortos hoy) o como una tasa (la tasa de abortos al día es de 3 abortos al día).

DISTRIBUCIÓN DE POISSON

Todas estas variables deben ser estudiadas bajo la distribución de Poisson, que fue propuesta por el matemático francés Siméon-Denis Poisson y dada a conocer en 1838 en su trabajo "Recherches sur la probabilité des jugements en matières criminelles et matière civile" (Investigación sobre la probabilidad de los juicios en materias criminales y civiles).

La aplicabilidad de la distribución de Poisson es muy variada. Como anécdota, decir que a finales del siglo XIX se utilizó por parte del alto mando del ejército prusiano para estudiar el número de soldados muertos por coz de caballo. Conocido que la media era de 0,61 soldados muertos por coz de caballo en cada regimiento, se determinó que una tasa mayor de 4 soldados muertos en un año y regimiento por esta causa constituía una negligencia de los mandos.

Esta distribución se aplica a variables discretas de la naturaleza en fenómenos que ocurren un determinado número de veces en un periodo determinado de tiempo o de volumen, o de superficie, o en nuestro caso, por ejemplo, en un corral, un parto, etc. La probabilidad de ocurrencia del fenómeno es constante en el tiempo o espacio. Ejemplos de distribución de Poisson en otros ámbitos diferentes de la producción porcina podrían ser:

- Número de coches que pasan por delante de nuestra granja de 9 a 12 de la mañana,

- Número de errores ortográficos por página escrita,
- Número de llamadas de teléfono recibidas al día,
- Número de animales muertos que encontramos en la carretera desde nuestra casa a nuestra granja al día, etc.

También nos ayuda a analizar y modelizar las tasas al permitirnos determinar qué variables explicativas (valores X) tienen un efecto en una variable de respuesta dada (valor Y, el conteo o una tasa). Por ejemplo, una granja podría aplicar la regresión de Poisson para comprender y predecir el número de abortos a la semana.

Vemos que no hay variables Poisson que contengan decimales: no pasan 3,4 coches delante de nuestra granja de 9 a 12 de la mañana (pasan o 3 o 4), aunque "de media", a la semana sí pasen 3,4 coches. Si estuviéramos interesados en modelizar en vez de la frecuencia de ocurrencia diaria, la media semanal, entonces, con mucha probabilidad deberíamos usar la distribución normal.

¿Qué nos aporta la distribución de Poisson?

Primero, nos ayuda a estudiar correctamente las variables que provienen de contajes. Veremos más adelante en esta serie de artículos que vamos a dedicarle, cómo, aunque quizá la media pudiera coincidir usando la ley Normal, los errores estándar no coincidirán, y eso es muy importante a la hora de modelizar una variable o un suceso.

Segundo, la distribución de Poisson nos ayuda a responder preguntas del tipo: si en nuestra explotación tenemos 3 abortos a la semana de media, ¿cuál es la probabilidad de que se produzcan 7 abortos a la semana? ¿Y de que se produzcan menos de 3 abortos a la semana?

Veamos la resolución de esta pregunta que, aunque *a priori* parece compleja en su cálculo, no lo es, y además nos servirá para conocer matemáticamente un poco la ley de Poisson.

La probabilidad de nuestra variable aleatoria, en este caso el número de abortos viene dada por la siguiente expresión[1]:

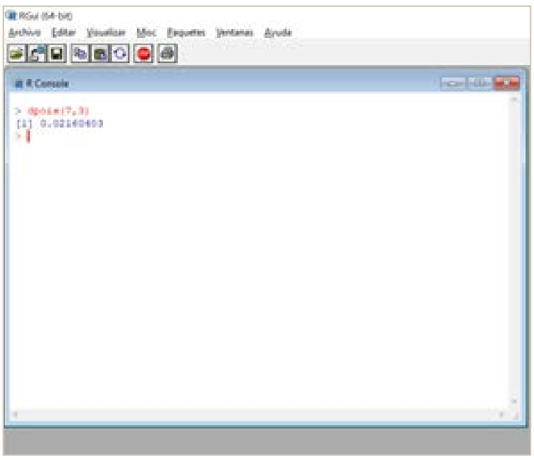

Figura 2. Consola de R con el cálculo de las probabilidades de Poisson del ejemplo.

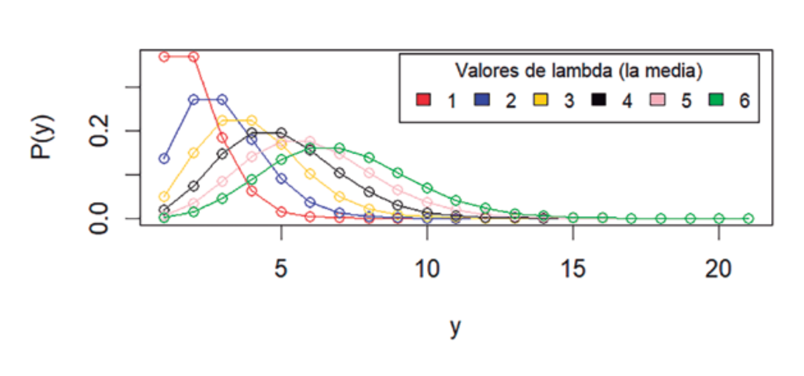

Figura 3. Distribuciones de densidad de Poisson para las medias señaladas.

$$P(x) = \frac{\lambda^k e^{-\lambda}}{k!}$$

que es la función de densidad de la ley de Poisson, donde en nuestro caso:

- "k" es el número de abortos a la semana en nuestra granja, que es una variable aleatoria discreta que toma los valores 0, 1, 2, 3 y en nuestro caso, k = 7.
- "λ" es la media del número de abortos a la semana. Aquí, λ = 3.
- "e" es un número irracional (tiene infinitos decimales, pero utilizaremos el valor 2,71828).

Vemos que la distribución de Poisson solo necesita un parámetro para ser definida: el número medio constante de sucesos aparecidos, en nuestro caso "λ", y así para responder a nuestra primera pregunta tendremos que usar la fórmula anterior con los valores correspondientes:

$$P(x = 7) = \frac{3^7 e^{-3}}{7!} = 0{,}0216;\ 2{,}16\ \%$$

Es decir, para una granja como la nuestra, con una media de 3 abortos a la semana, hay un 2,16 % de probabilidad de que en una semana haya 7 abortos. Si queremos hacer estos cálculos en *Excel* debemos usar la función *=POISSON.DIST(7;3;0)*. Hay que usar 2 datos: la media (en este caso, 3) y el valor del cual queremos calcular la probabilidad en nuestro caso 7. El valor 0 es un valor lógico que indica que no es una probabilidad acumulada. En *R* todavía es más fácil, basta con escribir la función *dpois(7,3)*, sin activar *RCommander* y nos dará la salida de la *figura 2*.

Para responder a la segunda pregunta, probabilidad de tener menos de 3 abortos, debemos pensar que esta probabilidad es la suma de las probabilidades de tener 0 abortos, 1 aborto y 2 abortos por semana. Así calcularemos igual que antes la probabilidad de tener 0, 1 y 2 abortos y las sumaremos:

$$P(x < 3) = P(x = 0) + P(x = 1) + P(x = 2)$$

$$P(x = 0) = \frac{3^0 e^{-3}}{0!} = 0{,}0498 = 4{,}98\ \%$$

$$P(x = 1) = \frac{3^1 e^{-3}}{1!} = 0{,}1494 = 14{,}94\ \%$$

$$P(x = 2) = \frac{3^2 e^{-3}}{2!} = 0{,}2240 = 22{,}40\ \%$$

$$P(x < 3) = 4{,}98 + 14{,}94 + 22{,}40 = 42{,}32\ \%$$

Por lo que la probabilidad de tener menos de 3 abortos durante una semana es del 42,32 % en nuestro caso. En *Excel* escribiremos *=POISSON.DIST(2;3;1)*. El valor 1 indica que es la probabilidad acumulada hasta el valor 2. Si escribimos en la consola de *R* *ppois(2,3)* nos devolverá el valor:

```
> ppois(2,3)
[1] 0.4231901
```

que es igual a lo calculado hasta ahora, bien con la calculadora a mano, en *Excel* o en *R*.

Otra propiedad curiosa que tiene la ley de Poisson es que su varianza es igual a la media y a la moda. Por lo tanto, su desviación típica será la raíz cuadrada de la media. Dado su coeficiente de asimetría, esta distribución es siempre asimétrica positiva con cola hacia la derecha, aunque a medida que la media crece, el coeficiente de asimetría[2] se acerca más a cero como podemos apreciar en la *figura 3*[3]. Y, además, como consecuencia del teorema central del límite[4], para valores grandes de "λ", una variable aleatoria de Poisson puede aproximarse a una distribución normal de media 0 y de varianza 1.

En los próximos números seguiremos estudiando ejemplos y utilidades de la ley de Poisson en nuestras granjas.

NOTAS

[1]En los textos estadísticos de la ley de Poisson habitualmente se denota la media con la letra griega lambda (λ).

[2]El coeficiente de asimetría (Y_1) de una distribución de Poisson se calcula mediante la fórmula $Y_1=1/(1/\sqrt{\lambda})$, siendo "λ" la media de la distribución de Poisson.

[3]El lector interesado encontrará el código de *R* para generar la *figura 3* en nuestro blog *https://www.testsandtrials.com/blog/*

[4]El teorema central del límite señala que, en condiciones muy generales, si Sn es la suma de n variables aleatorias independientes, con media y varianza finitas, entonces la función de distribución de Sn «se aproxima bien» a una distribución normal. Así pues, el teorema asegura que esto ocurre cuando la suma de estas variables aleatorias e independientes es lo suficientemente grande (*https://es.wikipedia.org/wiki/Teorema_del_límite_central*).