



Distribución de Poisson: gráficos de control (IV)

En el último artículo de esta serie vamos a ver dos temas no tratados hasta ahora: ¿cómo sé que mis datos siguen la distribución de Poisson?, y ¿cómo abordar mediante un control estadístico de procesos (en inglés *Statistical Process Control*, SPC) el análisis de este tipo de variables mediante los gráficos “c” y “u”?

Morillo Alujas¹, Daniel Villalba Mata² y Emilio López Cano³

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

Ya hemos visto que los datos de recuentos como las variables "nº de lechones muertos por parto", "nº de abortos a la semana" o "nº de lechones nacidos por parto" cumplen determinadas características:

- Son valores discretos, son recuentos de números enteros no negativos (no tienen decimales).
- Son sucesos que ocurren en un intervalo de tiempo determinado e independientes unos de otros.
- Pueden expresarse también como tasas, ya que la cantidad de veces que ocurre un evento dentro de un periodo de tiempo se puede expresar como un recuento sin procesar (ha habido 3 abortos hoy) o como una tasa (la tasa de abortos media es de 3 abortos al día).

Estas variables se estudian y analizan mediante la distribución de Poisson.

¿CÓMO SABER SI UNOS DATOS SIGUEN LA DISTRIBUCIÓN DE POISSON?

En la *tabla 1* vemos las características de las variables Poisson. Por tanto, si nuestra distribución cumple las anteriores características, podríamos suponer que sigue una distribución de Poisson. Pero esto puede no ser un buen argumento, por lo que podríamos realizar algún test estadístico. Una forma de hacerlo sería mediante un contraste Chi-Cuadrado. Imaginemos que recibimos de una granja la siguiente distribución del número de lechones nacidos muertos por parto en 600 partos (*tabla 2*). Estos datos se pueden encontrar en nuestro blog (<https://www.testsandtrials.com/blog/>) con el nombre “new_stillborn.xlsx” e importar en RCommander. Y podemos representarlos con un histograma con el siguiente código:

```
hist(new_stillborn$new_stillborn,breaks = - 0.5:6.5,
     main= "Número de partos por número de lechones muertos",
     xlab = "Lechones nacidos muertos por parto",
     ylab = "Frecuencia")
```

pero también podemos realizarlo yendo a *Gráficas/Histograma* y en *Opciones*, rellenar las casillas de los títulos (*figura 1*). Calculamos su media (1,19 lechones nacidos muertos por parto) y su varianza (1,05 lechones muertos por parto²), y prácticamente son iguales, (una característica de las variables Poisson). En *Estadísticos/Resúmenes/Conjunto de datos activos* podemos obtener estos estadísticos. Un test Chi-Cuadrado compara los valores esperados con los observados. Si simulamos un conjunto de valores de lechones nacidos muertos por parto con una media y varianza de 1,12 (valor medio entre la media y la varianza elegido arbitrariamente) obtenemos la *tabla 3*. Estos datos se obtienen mediante el código:

```
exp_stillborn <- rpois(600, 1.119808)
exp_stillborn
table(exp_stillborn)
```

Tabla 2. Número de partos observado según el número de lechones nacidos muertos en cada parto.

Muertos	0	1	2	3	4
Partos	170	236	115	69	10

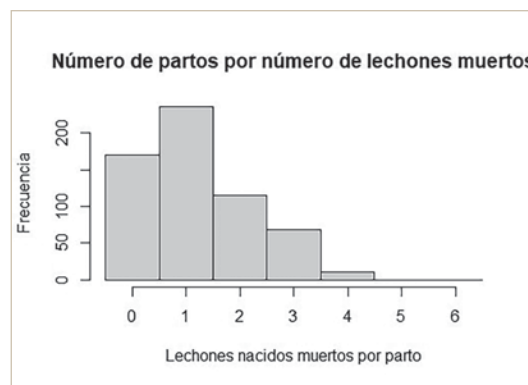


Figura 1. Simulación de 600 partos sobre el número de lechones nacidos muertos.

Tabla 1. Diferencias entre la distribución de Poisson y la distribución de Gauss o Normal.

Distribución de Poisson	Distribución de Gauss o Normal
Variables de recuentos o tasas	Variables continuas
Curva asimétrica hacia la derecha dependiendo de su media (λ)	Curva en forma de campana centrada en la media
Media = Varianza	La media y la varianza son diferentes, pero la media y la moda son iguales

Si ahora realizamos un test Chi-Cuadrado con estos valores simulados como esperados y los obtenidos de la granja como observados mediante el código:

```
observa <-c(170, 236, 115, 69, 10, 0, 0)
espera <-c(218, 214, 123, 30, 10, 4, 1)
```

y realizamos un test Chi-Cuadrado, obtenemos:

```
chisq.test(observa, espera)
Pearson's Chi-squared test
data: observa and espera
X-squared = 35, df = 30, p-value = 0.2426
```

Con lo que, con bastante probabilidad, no podemos rechazar la hipótesis de que nuestros datos siguen una distribución de Poisson.

LOS GRÁFICOS DE CONTROL “C” Y “U” O GRÁFICOS DE ATRIBUTOS

El gráfico de control “c” se utiliza para controlar el número total de eventos en un determinado proceso en soporte continuo, es decir, eventos por unidad, que siguen una distribución de Poisson teóricamente con infinitos valores posibles, por lo que no hay un tamaño de grupo para calcular proporciones.

El uso más habitual es la monitorización del número total de no conformidades medidas en una serie de k muestras del mismo tamaño. Por ejemplo, se toman muestras de 10 sacos de pienso, y se cuentan dentro de cada saco los defectos, por ejemplo, cambios de tamaño de partícula. Contamos, por tanto, el número total de defectos para cada 10 sacos, independientemente de en qué unidad estén. Cuando en la situación anterior tenemos tamaños de muestra variables, entonces puede interesar monitorizar el promedio de defectos por unidad. En este caso utilizamos el gráfico “u”.

Un ejemplo sería el número de repeticiones semanales en una granja de cerdas que, al tener tamaños de muestra variables (nº de cubriciones por semana), lo analizamos mediante el gráfico “u”. Para crear un gráfico “u” con RCommander, cargamos la librería “qcc”. En *Herramientas/Cargar plugin(s)* de *Rcmdr* elegimos la librería *RcmdrPlugin.qcc*. También podemos usar este código para activar la librería “qcc”:

```
library(qcc)
```

Tabla 3. Número de partos esperados según el número de lechones nacidos muertos en cada parto.

Muertos	0	1	2	3	4	5	6
Partos	218	214	123	30	10	4	1

Cargaremos el archivo de datos “rep_Poisson.xlsx” desde <https://www.testsandtrials.com/blog/> que contiene los datos de las repeticiones en una granja de cerdas durante 44 semanas. Tiene 3 variables:

- Repiten. Nº de cerdas que repiten celo esa semana.
- Cubiertas. Nº de cerdas cubiertas en esa semana.
- Opera. Operario que realiza la cubrición.

Para generar el gráfico “u”, iremos a *Control Charts/U/Graph*. Debemos elegir primero la variable que nos interesa (*Select the sample variable*) que en este caso es “repiten” y al lado debemos seleccionar la variable que nos da el tamaño de la muestra (*Select the variable to samples sizes(s)*) y en nuestro caso elegiremos la variable “cubiertas” que hemos creado anteriormente para este efecto. Aceptaremos y R nos genera la *figura 2*.

El gráfico generado nos informa de que la aparición media de repeticiones en este tiempo estudiado es de 9,33 % (*Center*). Como el número de cubriciones no es constante, no podemos calcular un valor fijo para los límites de control, pero las líneas punteadas nos indican que nuestros datos están dentro de los límites de control, es decir, los límites naturales del proceso. Vemos también que hay un punto amarillo que nos indica una violación de las normas de control. Este aparece dado que es el séptimo punto consecutivo a un lado de la media, y deberíamos prestar atención al mismo para ver a qué se debió ya que la probabilidad de que aparezcan 7 o más datos a uno u otro lado de la media es igual a 0,0078125 ($1/2^7$).

Es importante resaltar la importancia de utilizar el gráfico adecuado, según la naturaleza de los datos y su distribución de probabilidad. Un error común es realizar gráficos de control de la media para este tipo de variables de recuentos de eventos o de número medio de eventos por unidad. Los resultados pueden ser muy distintos, provocando, o bien falsas alarmas, o bien la falta de detección de situaciones fuera de control.

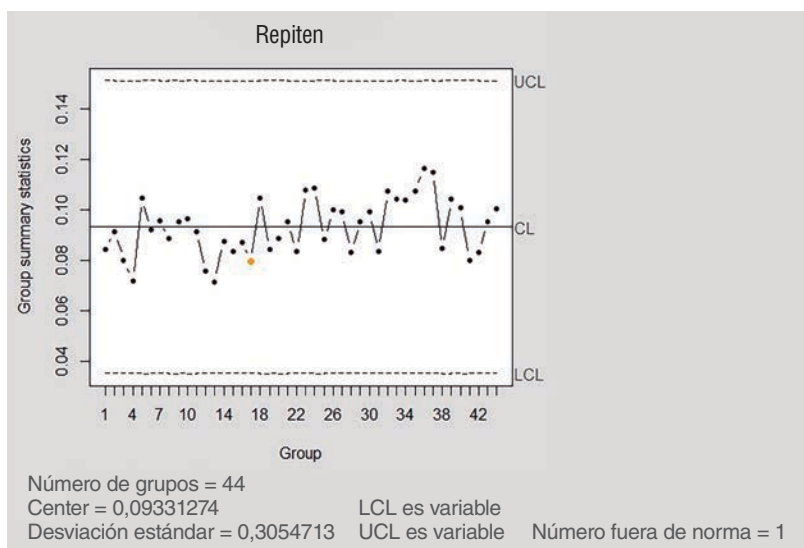


Figura 2. Gráfico “u” con los datos de repeticiones de celo.