CIMAT

PROGRAMA de Educación a Distancia

PRED
PROGRAMA DE EDUCACIÓN A DISTANCIA

# REGRESIÓN POISSON

La distribución Poisson se puede considerar un caso límite de la distribución binomial cuando el número de ensayos es grande y la probabilidad de ocurrencia de la característica de interés (éxito) es muy pequeña.

Se utiliza cuando se estudia la probabilidad de un determinado número de ocurrencias de la característica de interés.

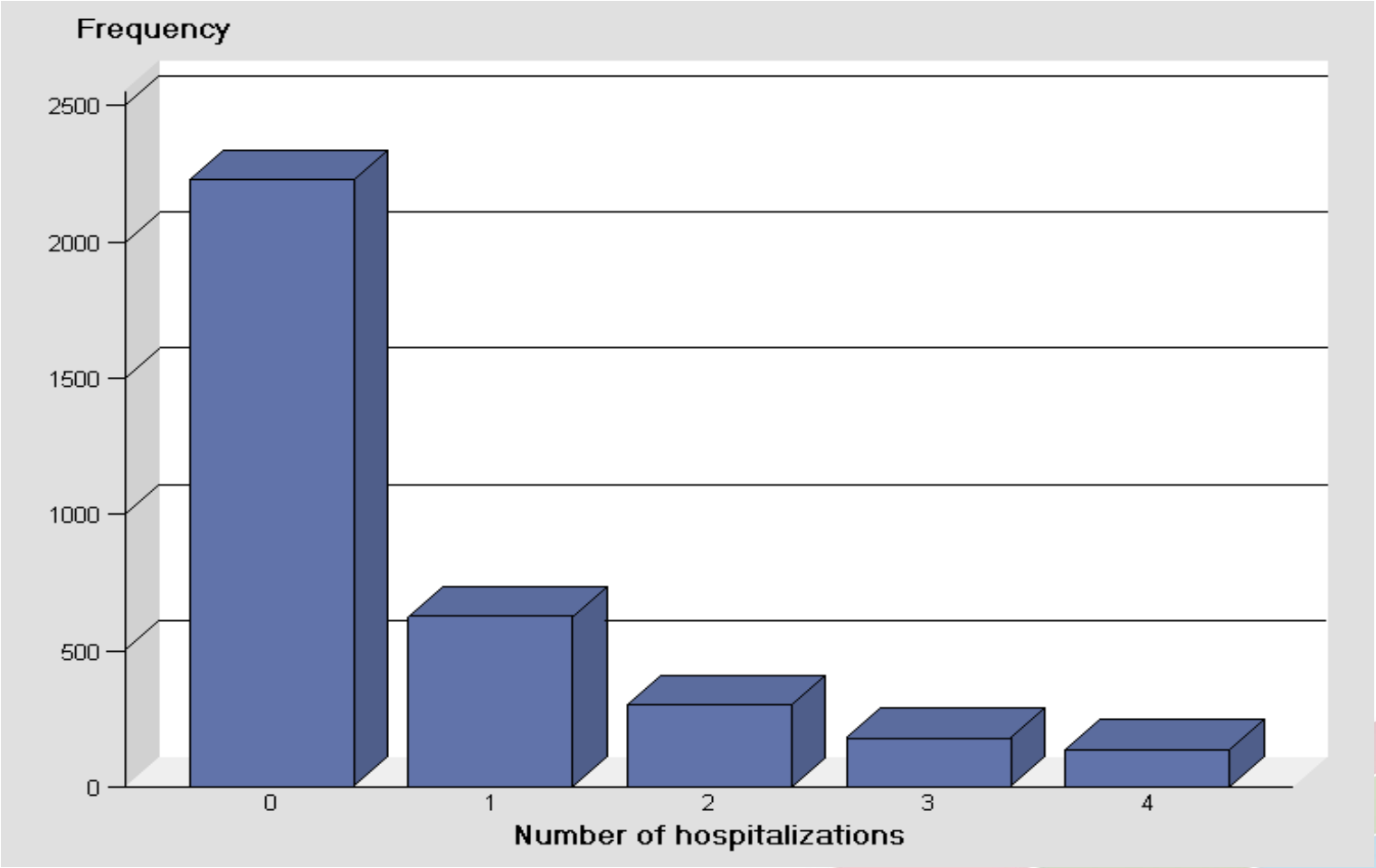
La regresión Poisson es un método para modelar variables respuesta que con conteos, es decir variables numéricas discretas que tienen valores enteros positivos (se incluye el cero):

0, 1, 2, 3, ….

Los resultados son válidos para eventos raros (ej. hospitalizaciones, complicaciones no usuales y otras condiciones que ocurren con poca probabilidad).

# Número de hospitalizaciones

(lejos de la normal)



Una característica común en datos de conteos no son solo los conteos en sí, también se incluye una unidad de referencia  con respecto a la cual ocurren los conteos, como pueden ser unidades de tiempo, distancia, área, volumen, etc.

El interés se centra en la tasa de ocurrencia y no tanto en el número absoluto de ocurrencias.

# Tratamiento del tiempo en estudios de seguimiento

La regresión Poisson puede modelar eventos raros cuando se tienen seguimientos con intervalos de tiempo iguales.

Sin embargo, también se pueden modelar eventos raros cuando los intervalos de tiempo son desiguales.

- Intervalos de seguimiento iguales por unidad.
- Cinco eventos (X) por 100 intervalos de tiempo, 10 por unidad. (Tasa de ocurrencia= 5/100)

OOOOOOOOOO
OOOOXOOXOO
OOOOOOOOOO
OOOOXOOOOO
OOOOOOOOOO
OOOOOOOOOO
OOOOOOOOOO
OOOOOOOOOO
OOOOOOOOOO
OXOOOXOOOO

- Intervalos de seguimiento desiguales por unidad.
- Cinco eventos (X) por 100 intervalos de tiempo (Tasa de ocurrencia= 5/100)

OOOOOOOOOOOO
OOOOXOOXOO
OOOOO
OOOOXOOOOOO
OOO
OOOOOOOOO
OOOOOOOOOOOO
OOOOOOOOO
OOOOOOOOOO
OXOOOXOOOOOOO

Supongamos que Y es la respuesta (número de personas hospitalizadas por apendicitis en 1,000 pacientes por año), y que $\lambda$ es la tasa de ocurrencia por unidad de tiempo (# de casos hospitalizados por cada 1,000 personas por año).

El modelo Poisson se puede escribir como:

$$\Pr(Y = y) = \frac{e^{-\lambda U}(\lambda U)^{y}}{y!}$$

donde U son las unidades consideradas (en éste caso los meses).

En éste modelo $E(Y) = \lambda U$ y se tiene que:

$$Log(E(Y)) = Log(\lambda U) = Log(\lambda) + Log(U)$$
$$= \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + Log(U)$$

Por lo que tenemos:

$$Log(\lambda) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

Por lo tanto, la representación matemática del modelo de regresión Poisson es:

$$\log(y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

$$y = e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k}$$

$$y = e^{\beta_0} \cdot e^{\beta_1 X_1} \cdots e^{\beta_k X_k}$$

Nótese que la función liga en éste caso es el logaritmo natural.

Para éste caso, los coeficientes de regresión exponenciados representan tasas relativas (rr) por unidad.

$$rr = e^{\beta_1}$$

Ejemplo:

- El coeficiente de regresión para el número de hospitalizaciones comparando pacientes con placebo respecto a pacientes con tratamiento es: $\beta_1 = 0.634$
- La tasa relativa de hospitalización es entonces:

$$e^{0.634} = 1.89$$

# Tasas y el concepto «Offset»

Cuando los eventos ocurren sobre un periodo de tiempo, espacio u otro índice de tamaño de unidades, el modelo se enfoca en la taza de ocurrencia de los eventos.

Por ejemplo, si analizamos el número de asesinatos en 2009 para una muestra de ciudades, es recomendable establecer una tasa dividiendo el número de asesinatos entre tamaño de la población de cada ciudad.

# Tasas y el concepto «Offset»

Cuando la respuesta Y tiene como índice el tamaño de población = $t$, la tasa es $Y/t$. El valor esperado de la tasa será entonces $\mu/t$, donde $\mu = E(Y)$.

Un modelo para la tasa esperada será:

$$\log\left(y/t\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

La representación equivalente será entonces:

$$\log(\mu) - \log(t) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

$$\log(\mu) = \log(t) + \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

Si usamos la función inversa (exponencial) todos lo términos entonces:

$$\mu = t \cdot e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k}$$

El ajuste $-\log(t)$ al logaritmo del valor esperado $\mu$ se llama *Offset*.

# Supuesto de la Regresión Poisson (comúnmente poco realista)

La regresión Poisson supone que el valor esperado es igual a la varianza.

Ejemplo: El número promedio de hospitalizaciones es igual a la varianza de la taza de hospitalizaciones:

$$E(Y) \approx V(Y)$$

$$\lambda = n \cdot p \approx n \cdot p \cdot (1 - p)$$

donde p es la probabilidad de hospitalización y n es el número de ensayos.

# Sobredispersión

- Si la varianza es mayor a la media se dice que los datos tienen *sobredispersión*.

- Datos con sobredispersión tienen errores estándar y valores de significancia $p$ muy pequeños.

# Regresión Binomial Negativa

- Siempre permite corregir por *sobredispersión*.
- Cuando no hay sobredispersión coincide con el modelo Poisson.
- Permite estimar un parámetro de sobredispersión con el cual se corrigen las predicciones e sus correspondientes errores de estimación.

El modelo basado en la Binomial Negativa, a diferencia de la Poisson, supone que la varianza es mayor al valor esperado, donde:

$$E(Y) = \lambda$$

$$V(Y) = \lambda + D\lambda^2$$

Al índice $D$ se le conoce como parámetro de dispersión. Conforme $D$ se aproxima a cero, la varianza se aproxima al valor esperado. Si $D = 0$, la distribución corresponde a una Poisson.