Taxis en una ciudad

El problema.

Se trata de estimar el número de taxis en una gran ciudad.

Supondremos que los *taxis* tiene asignado un número o licencia que empieza en *uno* y que las licencias son correlativas: $i = 1, 2, ..., N$. El parámetro de interés es N.

Para hacer la estiamción de N tomamos una muestra de n taxis y nos fijamos en su licencia $i_1, i_2, ..., i_n$ y especialmente en el máximo $x = max(i_i)$. Supondremos que no tenemos ninguna información previa sobre N.

De acuerdo con el teorema de *Bayes*:

$$p(N|D\ I) = \frac{p(N|I)\ p(D|N\ I)}{\sum_{N'} p(N'|I)\ p(D|N'\ I)}$$

La función de verosimilitud es fácil de calcular si el dato D relevante es el máximo de la muestra extraida $x = max(D)$:

$$p(x\ |N\ n) = \frac{\binom{x-1}{n-1}}{\binom{N}{n}}$$

Si el prior es no informativo, el resultado final será:

$$p(N\ |x\ n) = \frac{n-1}{x} \frac{\binom{x}{n}}{\binom{N}{n}}$$

Para llegar a esta expresión se utiliza el siguiente teorema:

$$\sum_{N' \geq x}^{\infty} \binom{N'}{n}^{-1} = \frac{x}{n-1} \binom{x}{n}^{-1}$$

Aunque no conociéramos este teorema, siempre podemos utilizar la forma proporcioanl del teorema de Bayes y normalizar después.

$$p(N\ |x\ n) \propto p(x\ |N\ n) = \frac{\binom{x-1}{n-1}}{\binom{N}{n}}$$

Comprobación.

```
#Número verdadero de taxis
Nv <- 100

licencias <- 1:Nv

# veces que cojo el taxi (número de la muestra)
n <- 10

# el valor más alto de la muestra
x <- max( sample( licencias, n ) )

#Rango para el parámetro N
Nmax <- 2 * Nv

N <- 1:Nmax
```

```r
# probabilidad de los posibles valores de N
# teoría
pN <- rep( 0, Nmax )

# calculada numéricamente
pNnum <- rep( 0, Nmax )

for ( i in x:Nmax ) {

  pN[ i ] <- ( ( n -  1 ) / x ) * choose( x, n ) / choose( N[ i ], n )

  pNnum[ i ] <- choose( x - 1, n - 1 ) / choose( N[ i ], n )

}

#Normalizo la distribución calculada numéricamente
pNnum <- pNnum / sum( pNnum )
```

Estadísticos.

Conocida la función de distribución $p(N \mid x\ n)$ se puede calcular la media:

$$< N > = \frac{n-1}{n-2}\,(x-1)\ , \quad n > 2$$

y la varianza:

$$\sigma_N^2 = \frac{(n-1)(x-1)(x-n+1)}{(n-2)^2(n-3)}\ , \quad n > 3$$

Del mismo modo, si no sabemos hacer estos cálculos analíticamente podemos hacerlos numéricamente.

```r
Nmedia <- ( n - 1) * ( x - 1) / ( n - 2 )

varN <- ( n - 1 ) * ( x - 1 ) * ( x - n + 1 ) / ( ( n - 2)^2 * ( n - 3 ) )

#Media calculada a partir de la distribución numérica
Nest <- sum( N * pNnum )

#Varianza estimada (calculada) a partir de la distribución numérica
varNest <- sum( ( N - Nest )^2 * pNnum )

#Redondeo
Nest <- round( Nest )

eNest <- round( sqrt( varNest ) )

Nmedia <- round( Nmedia )

eN <- round( sqrt( varN ) )

# gráfica
plot( N, pN,

      cex = 0.2,
```

```r
      ylab = "p(N | x n)",

      main = paste( "Nv=", Nv,", n=", n,", x=", x,", Ne=", Nmedia,"+-", eN )

    )

# comprobación
points( N, pNnum, col = 2 )

# resaltamos el valor esperado
abline( v = Nmedia, col = 2 )
```