Universidad Nacional de Colombia

Especialización en Analítica

Módulo Simulación y optimización

AJUSTE DE DISTRIBUCIONES A PARTIR DE DATOS EN R

Basado en:

https://cran.r-project.org/doc/contrib/Ricci-distributions-en.pdf

Discrete Event System Simulation. Jerry Banks, John S. Carson II, Barry Nelson, David Nicol. Ed. Pearson Prentice Hall 4th edition. NJ 2005

Pocos fenómenos del mundo real pueden predecirse completamente. Para el analista, dichos fenómenos son probabilísticos antes que determinísticos.

En simulación de sistemas, el analista propone modelos apropiados para muestrear el problema de interés.

Pasos (Banks et al., 2005):

1. Obtener datos del sistema real de interés.
2. Seleccionar una forma de distribución de probabilidad conocida
3. Estimar los parámetros de la distribución
4. Probar la bondad del ajuste con el fin de definir si se puede aceptar el modelo.

Una de las principales tareas en la solución de un sistema real es la recolección de datos. Aún cuando haya datos disponibles, es posible que estos no tengan una forma útil para construir un modelo con ellos.

Recomendaciones (Banks et al., 2005):

1. Planear. Puede iniciar con una sesión de observación de práctica en la cual se recolecten algunos datos en formatos que se diseñan para tal fin.
2. Analice los datos a medida que se recogen. Defina si los datos son adecuados para ajustar las distribuciones que se necesitan como insumo de la simulación, no recoja datos superfluos!

3. Trate de combinar conjuntos de datos homogéneos. Pruebe la homogeneidad de los datos en periodos sucesivos de tiempo y durante el mismo periodo en días sucesivos. Puede usar una prueba t de diferencia de medias.
4. Existe la posibilidad de que una cantidad de interés no se observe en su totalidad. Esto puede ocurrir cuando no se observan las actividades previas y posteriores al periodo de interés.
5. Para descubrir si hay una relación entre dos variables, haga un diagrama
6. Considere la posibilidad de que las observaciones sean autocorrelacionadas.
7. Tenga presente la diferencia entre datos de entrada y datos de salida, o de desempeño. Los datos de entrada son cantidades inciertas sobre las cuales no se tiene control y que no se alterarán por los cambios que se hagan para mejorar el sistema. Los datos de salida por su parte, representan el desempeño del sistema ante las entradas. Los datos de desempeño son útiles para validación pero el objetivo

2. Identificación de la distribución de los datos

2.1 Carga de archivo de datos. Usaremos el archivo babyboom.txt.

DESCRIPCIÓN RESUMIDA

El archivo babyboom.txt contiene un registro del la hora de nacimiento, sexo y peso al nacer de 44 bebés nacidos en un periodo de 24 horas en un hospital de Brisbane, Australia. También se incluye el número de minutos transcurridos desde la media noche para cada nacimiento.

NAME: Time of Birth, Sex, and Birth Weight of 44 Babies

TYPE: Observational

SIZE: 44 observations, 4 variables

SOURCE:

The data appeared in the Brisbane newspaper _The Sunday Mail_ on

December 21, 1997.

VARIABLE DESCRIPTIONS:

Columns

 1 - 8 Time of birth recorded on the 24-hour clock

 9 - 16 Sex of the child (1 = girl, 2 = boy)

17 - 24 Birth weight in grams

25 - 32 Number of minutes after midnight of each birth

Values are aligned and delimited by blanks. There are no missing values.

STORY BEHIND THE DATA:

Forty-four babies -- a new record -- were born in one 24-hour period at the Mater Mothers' Hospital in Brisbane, Queensland, Australia, on December 18, 1997. For each of the 44 babies, The Sunday Mail recorded the time of birth, the sex of the child, and the birth weight in grams.

Additional information about these data can be found in the "Datasets and Stories" article "A Simple Dataset for Demonstrating Common Distributions" in the _Journal of Statistics Education_ (Dunn 1999).

#Ejercicicio con datos babyboom.txt

datos de "Datasets and Stories" article "A Simple Dataset for Demonstrating Common Distributions" in the _Journal of Statistics Education_ (Dunn 1999).

```
# cargar datos del archivo
babies=read.table("babyboom.txt");
# verificar datos
# V1 Hora nacimiento 24:00
# V2 Sexo 1:F, 2:M
# V3 Peso, gramos
# V4 Número de minutos entre la medianoche y el nacimiento
babies;
```

```
> babies;
     V1 V2   V3   V4
1     5  1 3837    5
2   104  1 3334   64
3   118  2 3554   78
4   155  2 3838  115
5   257  2 3625  177
6   405  1 2208  245
```

```
7   407  1 1745  247
8   422  2 2846  262
9   431  2 3166  271
10  708  2 3520  428
```

Las siguientes secciones están basadas en https://cran.r-project.org/doc/contrib/Ricci-distributions-en.pdf

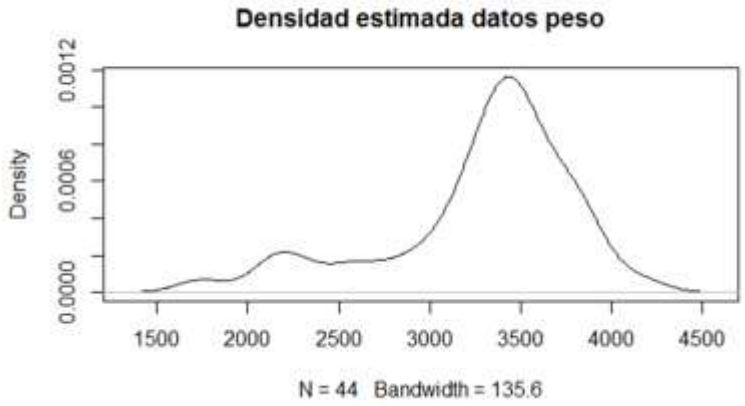
2.1 Identificar la forma de la distribución

Un histograma de frecuencias es útil para identificar la forma de una distribución. La instruccion hist() sirve para hacer el histograma.

```
# histograma variable peso al nacer, P
hist(babies$P,main="Histograma peso al nacer");
```

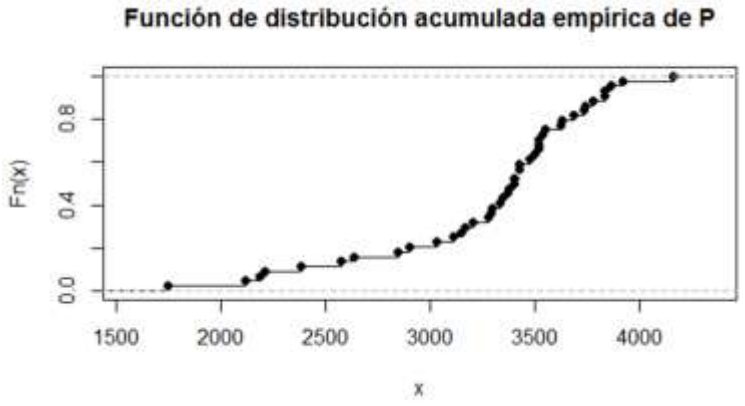
La densidad de las frecuencias se puede estimar usando density() y plot() para graficar

```
plot(density(babies$P),main="Densidad estimada datos peso")
```



R encuentra la función de densidad acumulada empírica con ecdf():

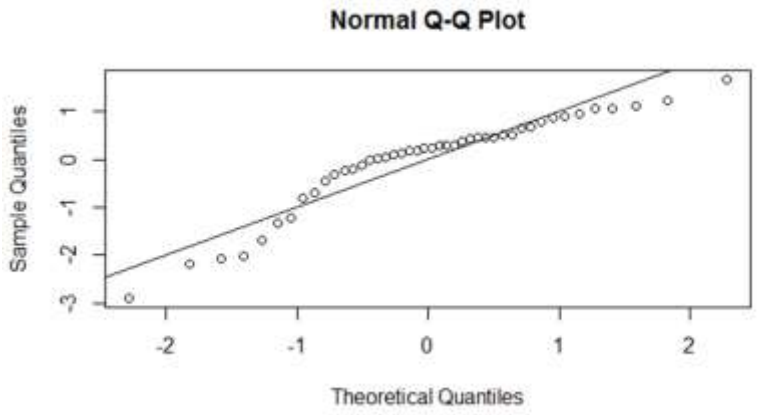
```
plot(ecdf(babies$P),main="Función de distribución acumulada empírica de P")
```



Un gráfico Q-Q es un gráfico de dispersión que compara las distribuciones ajustadas y empíricas en términos de los valores dimensionales de la variable (cuantiles empíricos). Esta es una técnica gráfica para determinar si un conjunto de datos viene de una población conocida. En la ordenada están los cuantiles empíricos y en la abscisa los del modelo teórico. El ajuste a una distribución gaussiana se puede probar con qqnorm() y el ajuste de cualquier otra distribución con qqplot()

Prueba normalidad

```
# datos de peso estandarizados
> znorm<-(babies$P-mean(babies$P))/sd(babies$P)
# gráfico qq de los datos de peso estandarizados
> qqnorm(znorm)
# línea m=1
```



Ejercicio 2.1

¿Qué concluye de la normalidad del peso?

Repita el procedimiento separando los datos por sexo.

```
babyF=read.table("babyboomF.txt", col.names=c("HNF","SF","PF","MF"));
hist(babyF$PF,main="Histograma peso al nacer F");
plot(density(babyF$PF),main="Densidad estimada datos peso F");
zpesoF<-(babyF$PF-mean(babyF$PF))/sd(babyF$PF);
qqnorm(zpesoF);
abline(0,1);

babyM=read.table("babyboomM.txt", col.names=c("HNM","SM","PM","MM"));
hist(babyM$PM,main="Histograma peso al nacer M");
plot(density(babyF$PM),main="Densidad estimada datos peso M");
zpesoM<-(babyM$PM-mean(babyM$PM))/sd(babyM$PM);
qqnorm(zpesoM);
abline(0,1);
```

2.2 Ajuste de funciones de distribución de probabilidad

1) mle() included in package stats4

2) fitdistr()included in package MASS