Estadística Descriptiva

Carlos Ignacio Patiño (cpatinof@gmail.com)

Julio 25, 2015

1 Objetivo

El objetivo del presente taller en clase (grupal) es permitir a los estudiantes practicar la aplicación de los conceptos de la estadística descriptiva, muy empleados en el análisis exploratorio de información.

2 ENUT

Los datos que serán empleados en el presente taller corresponden a la información de la Encuesta Nacional sobre Uso del Tiempo, llevada a cabo por el DANE durante el periodo 2012-2013. El archivo `enutVivienda.csv` contiene información relacionada con los estudiantes (de todos los niveles) encuestados que contestan de manera afirmativa a la pregunta de si durante la jornada de referencia (usualmente el día anterior a realizada la encuesta) dedicó algún tiempo, fuera de la jornada normal escolar, a estudiar o hacer tareas (relacionadas con sus estudios actuales) desde la casa.

En el archivo, las columnas corresponden a la siguiente información:

- `DIA_REFERENCIA_2`: corresponde a las 24 horas (de las 00:00 horas a las 23:59 horas) del día anterior al día de visita asignado Ej.: si se le ha asignado como día de visita al hogar el martes 14 de agosto el día de referencia será el lunes 13 de agosto de 2012 desde las 00:00 (o 12 de la noche) hasta las 23:59 (u 11:59 de la noche). Lunes (1) a Domingo (7), y (8) para día festivo.
- `P6040`: Edad. Si es menor de 1 año, el valor es 0.
- `P6020`: Género. (1: Masculio, 2: Femenino).
- `P6175`: Establecimiento (1: Público, 2: Privado).
- `P1158S1`: Nivel educativo: se refiere al nivel más alto de instrucción alcanzado por la persona, dentro del sistema formal de enseñanza, sea éste, educación preescolar, básica primaria, educación básica secundaria, superior o universitaria y postgrado.
- `P1161S1A1`: Horas de estudio (fuera de jornada escolar) en la VIVIENDA.
- `P1161S1A2`: Minutos de estudio (fuera de jornada escolar) en la VIVIENDA.

En todos los casos, las columnas han sido re-codificadas con el fin de tener en sus campos los valores descriptivos adecuados. Es decir, para el caso de la columna referente al género, en lugar de mantener los códigos 1 o 2, éstos se han recodificado a "Masculino" y "Femenino", con el objetivo de hacer más fácil la interpretación de cualquier resultado. Sin embargo, note que este tipo de estrategias puede ser poco adecuada para bases de datos muy grandes, en las que por eficiencia (no sólo por espacio sino también por velocidad en los procesos), se deben mantener estos códigos que ocupan menos espacio.

Igualmente, note que la información correspondiente al tiempo dedicado a estudiar en casa se encuentra separada en dos columnas diferentes. Esto no es una práctica muy recomendada en la gestión y análisis de bases de datos, por lo que el primer punto del presente taller le pedirá al estudiante que realice los pasos necesarios para obtener una base de datos apta para el análisis. En el proceso de análisis de información, los pasos necesarios para convertir una base de datos "cruda" en una base de datos analítica (o lista para el análisis) se denominan "Pasos de Procesamiento" o "Código de Procesamiento". En este caso, ya el instructor ha llevado a cabo este componente del proceso analítico y los estudiantes interesados en conocer dicha documentación, la podrán encontrar en el siguiente link.

3 Taller

Parte I

1. (Limpieza de datos) Genere una nueva columna denominada `tiempoCasa` que combine las dos columnas correspondientes al tiempo dedicado al estudio desde la Vivienda. Esta columna puede ser expresada en horas o minutos.
2. (Estadística descriptiva) Caracterice el tiempo dedicado al estudio en la vivienda en Colombia. ¿Qué forma tiene su distribución? ¿Qué puede usted decir al respecto?
3. (Estadística descriptiva) ¿Existen diferencias en el comportamiento (tiempo de estudio en la Vivienda) entre hombres y mujeres?
4. (Estadística descriptiva) ¿Cuáles medidas de tiempo presentan mayor variabilidad? ¿Las asociadas a los estudiantes de Primaria, Secundaria y Media? O las asociadas a los estudiantes de Pregrado, programas Técnicos y Tecnológicos?
5. (Regla Empírica) Un estudiante de Pregrado reporta que durante el día de referencia estudió en su vivienda durante 7 horas. ¿Qué tan probable es dicho registro? ¿Se trata esto de un dato "normal", o de un posible dato atípico?

Parte II

6. (Probabilidad y variables aleatorias) Grafique la distribución de probabilidad de la asistencia a una institución pública o privada.
7. (Probabilidad) ¿Cuál es $P(P/E)$?, donde P es igual a asistir a una institución pública y E es igual a ser un estudiante que asegura haber estudiado desde su casa durante el día de referencia?
8. (Regla de Bayes -Opcional-) Empleando la respuesta anterior, calcule $P(E/P)$. (Ayuda: 18.884 estudiantes reportan haber estudiado desde la casa durante el día de referencia. El total de estudiantes encuestados es 40.753. De los estudiantes que reportan no haber estudiando en la vivienda, 16.381 asisten a una IE Pública).
9. (Variables aleatorias) De las siguientes, ¿cuáles son variables aleatorias discretas? ¿Cuáles son continuas? a) El número de periodicos vendidos por El País en un mes; b) La cantidad de tinta empleada por El País para imprimir la edición dominical; c) El número de personas que están en fila en McDonald's a medio día el último viernes de cada mes; d) El tiempo en que una farmaceútica obtiene la aprobación por parte de las autoridades para lanzar un nuevo medicamento.
10. (Distribución Normal Estandarizada) ¿Cuál es la probabilidad de que un estudiante que atiende a una IE en el nivel de primaria, dedique más de 2 horas al día estudiando y haciendo tareas en su vivienda? (Chequee primero la normalidad de la distribución)