Análisis de las propinas de un restaurante

Estadística

El objetivo de esta práctica es realizar un análisis pormenorizado de un dataset o conjunto de datos, denominado "propinas", que contiene información sobre las propinas dadas en un restaurante en función de algunos parámetros, que explicaremos más abajo. La idea es utilizar toda la potencia de las técnicas de Estadística Descriptiva que nos proporciona R para extraer la mayor cantidad de información posible.

Introducción

El dataset "propinas" fue descrito por primera vez por P.G. Bryant y M.A. Smith en su libro "Practical Data Analysis: Case Studies in Business Statistics" (1995). Consta de un total de 244 observaciones de 8 variables, a saber:

- OBS: Identificador entero del número de observación.
- TOTAL: Valor total de la cuenta de un determinado servicio del restaurante, en $.
- PROPINA: Valor de la propina entregada, en $.
- SEXO: Indica el sexo de la persona que pagó la cuenta.
- FUMADOR: Indica si los comensales pidieron sentarse en la zona de fumadores o de no fumadores del restaurante.
- DIA: Representa el día de la semana de la visita al restaurante.
- TURNO: Representa el turno de comidas del restaurante.
- TAMANO: Indica el número de personas que formaban el grupo de comensales.

Importar a R el fichero CSV

El dataset está en formato de valores separados por comas (CSV). Podemos leerlo directamente desde R con la función read.csv.

```
> datos <- read.csv("ruta/dataset-propinas.csv")
```

Donde *ruta* hace referencia a la carpeta que contiene el fichero CSV. Puede ser la ruta absoluta o la relativa. Las barras para indicar las carpetas son las inclinadas hacia la derecha (tecla 7).

Preguntas generales a resolver

- ¿Hay alguna relación entre el TAMAÑO de los grupos de comensales y la PROPINA?
- ¿En las comidas más caras se dejan más o menos PROPINAS que en las comidas más modestas?
- Información vital para los camareros: ¿en qué TURNO y qué DÍA de la semana son más suculentas las PROPINAS?
- ¿Cuál es el cliente medio del restaurante?
- En definitiva, ¿qué variables afectan más a la PROPINA?

Tareas concretas

1. Inspeccionar de qué **tipo** son cada una de las 8 variables, según los dos criterios de clasificación que hemos estudiado en clase:

- Según si los **valores** se pueden representar por números enteros o reales (discreta, continua).
- Según la **escala de medida** (cualitativa nominal u ordinal, cuantitativa de intervalo o proporción).

Responder:

- ¿La variable OBS es cuantitativa o cualitativa? ¿Qué dice R? ¿Estás de acuerdo?
- ¿Están bien ordenadas las variables cualitativas ordinales según el orden que establece R por defecto? Si crees que no, indícale a R explícitamente la ordenación adecuada en cada caso.

2. Calcular **medidas de sumarización** de cada variable.

- Para las variables continuas, calcular el recorrido, los cuartiles, la media aritmética, desviación estándar, o cualquier otra medida de centralidad y dispersión que creas conveniente. Dibujar también un diagrama de cajas.
- Para las variables discretas, indicar los valores diferentes que toman y calcular la moda. Dibujar un diagrama de barras.

Responder:

- ¿Cuál es el valor TOTAL medio?
- ¿Cuál es la PROPINA media?
- ¿Cuál es el valor medio de la proporción PROPINA/TOTAL?
- ¿Tienen las variables continuas algún valor anómalo? ¿Cuáles?
- ¿Cómo varía la PROPINA media según el DÍA de la semana?
- ¿Cómo varía la PROPINA según el SEXO del pagador de la cuenta?
- ¿Cuál es el TAMAÑO medio de los grupos de comensales?
- ¿Qué DÍA de la semana hay más clientes?
- ¿Qué DÍA de la semana se hace más caja? ¿En qué TURNO?

3. Considerar la variable TOTAL de la cuenta y **agruparla** en intervalos de anchura 10 $. Asimismo considerar la variable PROPINA y **agruparla** en intervalos de anchura 1 $. Calcular la **tabla de contingencia** y las **frecuencias marginales** de estas dos variables, tanto absolutas como relativas.

Responder:

- ¿En qué intervalo cae el valor TOTAL de la cuenta más frecuente?
- ¿En qué intervalo cae la PROPINA más habitual?

4. Considerando el SEXO del pagador y si el grupo era FUMADOR o no, calcular la **tabla de contingencia** y de **frecuencias marginales**, tanto absolutas como relativas. Dibujar también un **diagrama de mosaico**. Se recomienda representar en color AZUL los datos correspondientes a los hombres y en color ROSA los correspondientes a las mujeres.

Responder:

- ¿Cuál es la combinación más habitual de estas dos variables?

5. Considerando el DIA de la semana y el TURNO de comida, calcular también la **tabla de contingencia** y de **frecuencias marginales**, tanto absolutas como relativas. Dibujar también un **diagrama de mosaico**.

Responder:

- ¿Cuál es la combinación más habitual de estas dos variables?
- ¿El patrón de trabajo del restaurante es el mismo los días de entre semana que los fines de semana?

6. Realizar los **histogramas** siguientes:

- En concreto para la variable PROPINA, donde las anchuras de los intervalos sean de 2, 1, 0,50, 0,25 y 0,05 $.
- A continuación dibujar el **diagrama de tallo y hojas** de la variable PROPINA.

Responder:

- A la vista de estos diagramas, ¿se aprecia algún patrón en la cantidad concreta que la gente deja de PROPINA?

7. Realizar el **histograma** del TOTAL con intervalos de anchura 5 $.

Responder:

- A la vista del histograma, indicar si los datos presentan una cola positiva o negativa.
- Calcular la asimetría de esta variable para comprobar que el grado de asimetría es el esperado por el apartado anterior.
- ¿Cuánto vale la curtosis? ¿Se aprecia en el histograma?

8. Realizar el **histograma** de la proporción PROPINA/TOTAL, con intervalos de 0.05 (5%).

Responder:

- ¿En qué intervalo cae el máximo?
- A la vista del histograma, ¿hay datos anómalos? ¿Por qué? Comprobarlo con un diagrama de cajas.

9. Para todas las variables cualitativas, dibujar los **diagramas de barras** o **de mosaico** (o ambos) de cada una de ellas, utilizando colores según el SEXO del pagado.

Responder:

- ¿Es un restaurante al que la gente vaya a comer sola o en grandes grupos? ¿Predominan las parejas?
- En el caso de las parejas, ¿quién suele pagar la cuenta la mayor parte de las veces?

10. Para las variables cuantitativas, representar los **diagramas de dispersión** correspondientes y calcular los **coeficientes de correlación** entre las siguientes variables:

- El valor de la PROPINA frente al valor TOTAL de la cuenta.
- La proporción que supone la PROPINA respecto del valor TOTAL de la cuenta, frente al valor TOTAL de la cuenta.
- La proporción que supone la PROPINA por persona respecto del valor TOTAL de la cuenta, frente al valor TOTAL de la cuenta.

Responder:

- ¿Es en general mayor la PROPINA para comidas más caras, es decir, para mayor valor TOTAL?
- ¿Y la proporción de PROPINA/TOTAL? ¿Cómo varía en % esta proporción por cada dólar de incremento en la factura total?
- ¿Y la proporción de PROPINA por persona respecto del valor TOTAL?
- ¿Quién da mejores propinas según el SEXO del pagador de la cuenta y de si el grupo era FUMADOR o no?

11. Con las gráficas anteriores del apartado anterior, calcular y dibujar las **rectas de regresión** para todos los datos, y para los datos separados en función de la variable SEXO.

Responder:

- ¿Son iguales las pendientes de las rectas para cada SEXO? ¿Qué signo tienen?

12. Analizar la influencia del TAMAÑO del grupo en la proporción PROPINA/TOTAL. Para ello, calcular la **correlación** entre ambas variables y dibujar un **diagrama de dispersión**. Asimismo calcular un ajuste lineal a estos datos.

Responder:

- ¿Cómo varía en % la proporción PROPINA/TOTAL por cada persona añadida al grupo de comensales?

Sugerencias

Además de las preguntas que aquí se plantean, se deja total libertad para utilizar todo tipo de representaciones gráficas, como diagramas de dispersión, histogramas, diagramas de tartas, diagramas de barras apiladas, diagramas de mosaico, etc. También puedes calcular cualquier otra medida estadística que pueda resultar de interés.

Se recomienda, aunque no será imprescindible, que los gráficos generados tengan un tamaño de fuente suficientemente grande para ser legible, y que se utilicen colores vistosos.

Pero lo importante será interpretar tanto los resultados numéricos como las gráficas. Será fundamental explicar con palabras las conclusiones extraídas. No se evaluará una gráfica o cálculo que no vaya acompañado de su explicación justificada y razonada.

Evaluación

Este trabajo se contabiliza dentro del 30% de la nota final de la asignatura. Se penalizará entregarlo fuera de plazo.