CO3321: Estadística

PROYECTO (30 puntos)

Profesor: Pedro Ovalles *Enero - Marzo 2020 (en junio-julio)*

1. Datos

1.1. Notas

En el archivo ***notas.txt***, hay observaciones de 10 variables. Las observaciones o unidades muestrales corresponden a estudiantes de la USB que presentan un examen estandarizado de conocimientos científicos. Los datos que se registran para cada estudiante son:

`MA1` = Nota acumulada la primera vez que cursan matemáticas 1.
`MA2`= Nota acumulada la primera vez que cursan matemáticas 2.
`MA3` = Nota acumulada la primera vez que cursan matemáticas 3.
`CS1` = Nota acumulada la primera vez que cursan sociales 1.
`CS2`= Nota acumulada la primera vez que cursan sociales 2.
`LL1` = Nota acumulada la primera vez que cursan lenguaje 1.
`LL2` = Nota acumulada la primera vez que cursan lenguaje 2.
`Y` = Nota en el examen estandarizado.
`Gen` = Género del estudiante, H para hombres y M para mujeres.
`Car` = Carrera del estudiante, en este caso solo se tienen estudiantes de cinco carreras, ingeniería en computación (IC), ingeniería electrónica (IE), ingeniería mécanica (IM), ingeniería en producción (IP) e ingeniería química (IQ).

1.2. Puntajes

El archivo de datos ***puntajes.txt*** contiene datos acerca de varias películas. Estos datos tienen relación con los puntajes que le colocan 7 expertos de la industria cinematográfica. Cada observación contiene información sobre 11 variables. Estas son:
`P1` = Puntaje de un productor de cine.
`P2` = Puntaje de una actriz de larga trayectoria.
`P3` = Puntaje de un conocido director.
`P4` = Puntaje de una directora artística.
`P5` = Puntaje de una seleccionadora de elencos.
`P6` = Puntaje de un director de fotografía.
`P7` = Puntaje de un creador de efectos visuales.
`IMDb` = Puntaje en la página web Internet Movie Database (IMDb.com).
`Genre` = Género principal de la película.
`Year` = Año registrado de estreno de la película.
`Name` = Nombre de la película.

2. Trabajo asignado

En cada proyecto deberán hacer entregas semanales basadas en lo que se les pide a continuación. Luego todo esto se debe consolidar en el informe final que se entregará al final del periodo.

2.1. Notas

Para este grupo de variables se solicita el siguiente trabajo:

1. Realizar un análisis descriptivo de los datos.
2. Realice un intervalo de confianza del 97 % para la media de cada variable en estudio. Analice lo obtenido.
3. Pruebe, a un nível de 0.05, que el promedio de la prueba estandarizada es mayor a 60 puntos.
4. Estudie si las notas promedio entre matemáticas 1 y sociales 1 son iguales.
5. Realizar una prueba de bondad de ajuste para determinar si la variable "Y" tiene distribución normal.
6. Con un nível de significancia de 0.02, pruebe si las proporciones de estudiantes por carrera son iguales.
7. Realizar un gráfico de dispersión y una matriz de correlación de las variables.
8. Halle un modelo lineal que explique mejor la variable "Y". Incluya todas las pruebas necesarias para llegar a este modelo, así como un análisis de residuos del modelo final.
9. Con los datos *notas_pre.txt* haga una predicción de la variable "Y" (con el mejor modelo) y haga un histograma, diagrama de cajas y resumen estadístico de los residuos de predicción (valor observado vs. predicción del modelo) para concluir con relación al poder predictivo del modelo.
10. Realice un análisis de varianza para decidir si las medias por carrera y género son iguales en cada una de las mátematicas y la variable "Y".
11. En el caso de obtener en el análisis de varianza que existe una diferencia significativa, por medio de pruebas de hipótesis, decidir cuales son los factores con diferencias.

2.2. Puntaje

Para este grupo de variables se solicita el siguiente trabajo:

1. Realizar un análisis descriptivo de los datos.
2. Realice un intervalo de confianza del 97 % para la media de cada variable en estudio.
3. Pruebe, a un nível de 0.05, que el promedio de puntajes de IMDb es menor a 7 puntos.
4. Estudie si los puntajes promedios del director y el productor son iguales.
5. Realizar una prueba de bondad de ajuste para determinar si la variable "IMDb" tiene distribución normal.
6. Con un nível de significancia de 0.02, pruebe las proporciones de películas por género son iguales.
7. Realizar un gráfico de dispersión y una matriz de correlación de las variables.
8. Halle un modelo lineal que explique mejor la variable "IMDb". Incluya todas las pruebas necesarias para llegar a este modelo, así como un análisis de residuos del modelo final.
9. Con los datos *puntajes_pre.txt* haga una predicción de la variable "IMDb" (con el mejor modelo) y haga un histograma, diagrama de cajas y resumen estadístico de los residuos de predicción (valor observado vs. predicción del modelo) para concluir con relación al poder predictivo del modelo.
10. Realice un análisis de varianza para decidir si las medias por género y año, son iguales, para las variables "P1","P5","P6" y "IMDb". En este punto, se deben excluir los años con menos de 10 películas.
11. En el caso de obtener en el análisis de varianza que existe una diferencia significativa, por medio de pruebas de hipótesis, decidir cuales son los factores con diferencias.

3. Criterios de corrección para el informe

La estructura que debe tener el informe es:

- Portada con resumen (en la misma hoja).
- Planteamiento del problema (incluyendo los objetivos del trabajo), descripción de la base de datos y la metodología a emplear.
- Desarrollo (donde se realizan las asignaciones).
- Conclusiones y recomendaciones.
- Bibliografía.
- Anexos (+ códigos en R).

En la portada se debe encontrar el título del proyecto, el resumen y la identificación de los autores.
Una de las partes más importantes del informe es el resumen; en este se deben plantear los objetivos del proyecto y una breve descripción de la base de datos y de la metodología empleada. También se deben encontrar los resultados del proyecto (o por lo menos, los más substanciales), y se debe aclarar las implicaciones de estos resultados, las conclusiones y recomendaciones (simplificadas) que hace el analista.

El cuerpo principal del informe, debe comenzar con el planteamiento del problema, y luego describir la base de datos y la metodología que se empleará durante la resolución del mismo. Se deben usar tablas y gráficos para facilitar la lectura del informe y obtener la atención del cliente; las tablas y gráficas deben estar comentadas, no se permiten tablas o gráficas a las que no se hacen referencia. Debido a que el informe no debe tener más de diez (10) páginas (desde la portada a la bibliografía), se debe resumir la información en tablas o diagramas y se deben seleccionar los gráficos más relevantes.

En las conclusiones se presentan los resultados obtenidos conjuntamente con las implicaciones que tienen esos resultados (sin profundizar en terrenos del área en el que se desenvuelve el cliente, a menos de que se esté seguro del impacto de las implicaciones). Recuerde que este es un trabajo parecido al de asesoría y que el cliente es el que toma las decisiones, el analista sólo plantea alternativas y puede sugerir alguna de las soluciones al problema.
Presentación de resultados

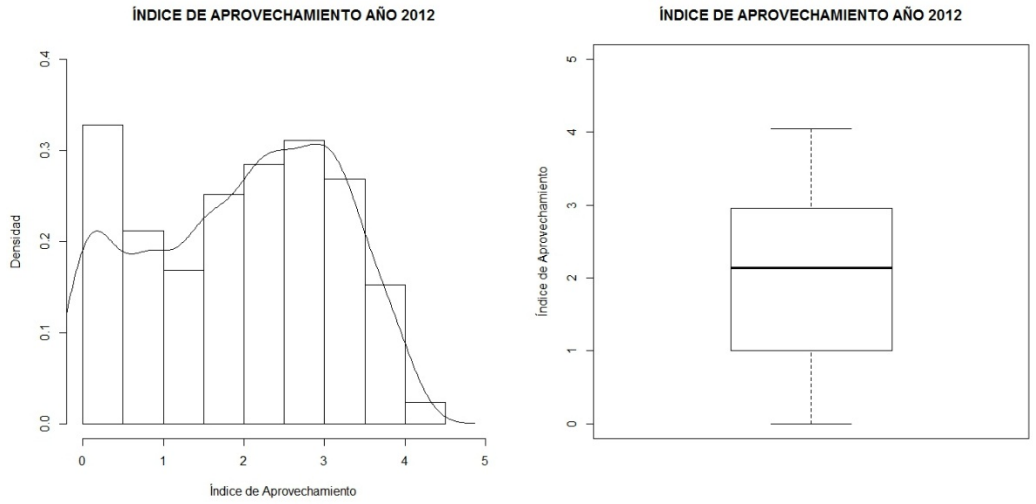
- Presente sus resultados en tablas ordenadas e interprete.
- Identifique en los diagramas de caja si hay datos atípicos, cómo es la distribución de los datos, si es sesgada a la derecha, etc.
- Los gráficos tiene que tener su título y los nombres de los ejes (todo en español).

Es INACEPTABLE

- No se aceptará presentación de resultados con manuscritos escaneados.
- Se anulará la evaluación de aquellos que compartan fotografías tomadas desde la pantalla de la computadora.
- No se aceptará un copy y paste de los resultados.
- No se aceptarán títulos de las gráficas generados por degecto en el programa.

Tabla 1. Resumen estadístico para la variable Índice de Aprovechamiento

Reusmen Estadístico							
Variable	**Mínimo**	**Primer**	**Mediana**	**Media**	**Tercer**	**Máximo**	**Desviación**
IAP	0	1	2.14	1.97	2.95	4.05	1.16



3.1. Ejemplos

Por último se exponen unos ejemplos para la presentación de los resultados (Gráficas y Tablas), para mayor información se puede consultar las normas de la Universidad Simón Bolívar para la elaboración de trabajos.

Gráfico 1. Histograma y gráfico de caja para la variable Índice de aprovechamiento.

NOTA: recuerde que existen normas para la elaboración de trabajos propias de la USB, es recomendable revisar las mismas para la escritura del proyecto. Por ejemplo, es muy común cometer errores en la bibliografía. Recuerde que el autor debe ser mencionado en el texto, y posteriormente señalar la referencia en la bibliografía.

Ejemplo:
"Para Gelman y otros (2014), el muestreador de Gibbs es un método de gran utilidad en problemas donde el espacio de parámetros es multidimensional."

"En este trabajo se aplicó el programa R Development Core Team (2015)."

"Según Gil, J. (s/f), los métodos..."

En la bibliografía

Gleman, A., Carlin, J., Stern, H. y Rubien, D. (2004). Bayesian data analysis. Second Edition. Chapman & Hall/ CRC.

Gil, J. (s/f). Modelos de medición: desarrollos actuales, supuestos, ventajas e inconvenientes. Universidad de Sevilla. [Revista en Línea]. Disponible: http://innoevalua.us.es/files/irt.pdf [Consulta: 2015, Diciembre, 09].

R Development Core Team (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0, Disponible: http://www.R-project.org.

4. Condiciones de entrega

a Tanto las entregas semanales como el informe final deben ser entregados en forma electrónica y en formato ".pdf".

b Las entregas semanales serán de la siguiente forma:

Punto 1, a más tardar el sábado 13-06 (semana 1) Punto 2, a más tardar el sábado 27-06 (semana 3) Puntos 3, 4, 5 y 6, a más tardar el sábado 04-07 (semana 4) Puntos 7, 8 y 9, a más tardar el sábado 11-07 (semana 5) Puntos 10 y 11, a más tardar el jueves 16-07 (semana 6)

al correo electrónico povallesgarcia@usb.ve. El asunto del correo debe ser "*Entrega (número de la semana). CO3321*". Por ejemplo, la entrega del primer punto debe ir en un correo cuyo asunto debe ser *Entrega 1. CO3321.*

c La entrega del informe final se realizará al correo electrónico **povallesgarcia@usb.ve** a más tardar el sábado 18 de julio de 2020 a las 8:00 a.m. El asunto del correo DEBE ser: "*Proyecto. CO3321*".

d No se corregirán informes entregados fuera del tiempo establecido para la entrega.