Universidad Simón Bolívar
Departamento de Cómputo Científico y Estadística
Trimestre Enero-Marzo 2018
Estadística para ingenieros (CO3321)
Estadística para matemáticos (CO3322)

Proyecto Final

1. Datos

1.1. Estudiantes. Los datos **caracteristicasfisicas.txt**, contiene 300 observaciones de 9 variables. Las observaciones o unidades muestrales corresponden a estudiantes españoles y las variables a sus características físicas, éstas son:

```
Estatura (centímetros)
Peso (kilogramos)
Longitud de pie (centímetros)
Longitud del brazo (centímetros)
Anchura de espalda (centímetros)
Diámetro del cráneo (centímetros)
Longitud entre la rodilla y el tobillo (centímetros)
IMC índice de masa corporal
Zona del aspirante
```

1.2. Estudio de algas. El archivo de datos **algas.txt** contiene 300 muestras recogidas en ciertos puntos de investigación destinados al crecimiento de un determinado tipo de alga de agua dulce. Cada observación contiene información sobre 10 variables. Estas son:

```
*''Long'' = Longitud promedio de las algas en centímetros.
*''Temp'' = Temperatura del agua en grados centigrados.
*''mxPH'' = Máximo valor de pH del agua en unidades de pH.
*''mnO2'' = Valor mínimo de oxígeno del agua en partes por millón.
*''NA'' = Valor medio de sodio del agua en partes por millón.
*''mFe'' = Máximo valor del hierro del agua en partes por millón.
*''Mg'' = Valor medio de magnesio del agua en partes por millón.
*''Chla'' = Media de la clorofila en partes por millón.
*''zona'' = Zona donde se toma la medición.
*''contaminante'' = Tipo de contaminante presente en el agua.
```

1.3. Capacitación. En el archivo **Capacitacion.txt** se encuentran los datos de 300 aspirantes que forman parte de un programa de inserción laboral, para personas que hayan desertado del sistema de educación preparatoria formal. Cada observación contiene información sobre 9 variables. Estas son:

```
*''JTS'' = puntaje en la prueba de capacitación aplicada al final del programa.
*''ING'' = Ingreso en el hogar del aspirante como porcentaje por encima del nivel de pobreza.
*''EDU'' = Nivel educativo efectivo del responsable del hogar del aspirante, medido en años de educación.
*''DIS'' = Número de faltas de disciplina del aspirante durante sus dos últimos años en la escuela.
*''SEM'' = Número de semanas transcurridas entre la deserción de la escuela y el inicio del programa de capacitación.
*''EMP'' = Número de empleos ocupados desde la deserción de la preparatoria.
```

*''PAS'' = Porcentaje de asistencia durante los últimos dos años en la escuela.
*''IPR'' = Ingreso promedio mensual del aspirante en sus anteriores empleos.
*''zona'' = Zona donde reside el aspirante.

2. Trabajo asignado

2.1. Estudiantes.

1. Realizar un análisis descriptivo de los datos.
2. Realizar un gráfico de dispersión y una matriz de correlación de las variables.
3. Realizar una prueba de bondad de ajuste para determinar si la variable "estatura" se distribuye en forma normal.
4. Halle un modelo lineal que explique mejor la variable "estatura". Incluya todas las pruebas necesarias para llegar a este modelo, así como un análisis de residuos del modelo final.
5. Con los datos *Caracteristicasfisicas_prediccion.txt* haga una predicción de la variable "estatura" (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.
6. El Índice de Masa Corporal sirve como una medida para clasificar a las personas de la siguiente manera:

IMC (kg/m^2)	Categoría	Masa de una persona de 1,80 m con este IMC
Menos de 16,00	Infrapeso severo *(criterio de ingreso)*	menos de 51,8 kg
De 16,00 a 16,99	Infrapeso moderado	51,8 - 55,1 kg
De 17,00 a 18,49	Bajo peso	55,1 - 59,9 kg
De 18,50 a 24,99	Peso normal	59,9 - 81,0 kg
De 25,00 a 29,99	Sobrepeso	81,0 - 97,2 kg
De 30 a 34,99	Sobrepeso crónico *(obesidad de grado I)*	97,2 - 113,4 kg
De 35,00 a 39,99	Obesidad premórbida *(obesidad de grado II)*	113,4 - 129,6 kg
De 40,00 a 44,99	Obesidad mórbida *(obesidad de grado III)*	129,6 - 145,8 kg
A partir de 45,00	Obesidad hipermórbida *(obesidad de grado IV)*	más de 145,8 kg

   Clasifique a cada estudiante según esta tabla.
7. Reporte la estatura media de los estudiantes según su clasificación por IMC y zona.
8. Realice un análisis de varianza para estudiar si existen diferencias entre las estaturas medias de los estudiantes según su clasificación por IMC y zona.
9. En el caso de conseguir diferencias significativas, realice pruebas de medias para determinar cuales son las medias que difieren.

2.2. Estudio de algas. Para este grupo de variables se solicita el siguiente trabajo:

1. Realizar un análisis descriptivo de los datos.
2. Realizar un gráfico de dispersión y una matriz de correlación de las variables.
3. Realizar una prueba de bondad de ajuste para determinar si la variable "Long" se distribuye en forma normal.

4. Encuentre el modelo de regresión simple que mejor se ajuste a los datos, donde "Long" es la variable dependiente.
5. Halle un modelo lineal que explique mejor la variable "Long". Incluya todas las pruebas necesarias para llegar a este modelo, así como un análisis de residuos del modelo final.
6. Con los datos *algas_prediccion.txt* haga una predicción de la variable "Long" (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.
7. Reporte el crecimiento medio de las algas según contaminante y zona.
8. Realice un análisis de varianza para estudiar si existen diferencias entre los crecimientos medio de las algas según contaminante y zona.
9. En el caso de conseguir diferencias significativas, realice pruebas de medias para determinar cuales son las medias que difieren.

2.3. Capacitación.

1. Realizar un análisis descriptivo de los datos.
2. Realizar un gráfico de dispersión y una matriz de correlación de las variables.
3. Realizar una prueba de bondad de ajuste para determinar si la variable "JTS" se distribuye en forma normal.
4. Halle un modelo lineal que explique mejor la variable "JTS". Incluya todas las pruebas necesarias para llegar a este modelo, así como un análisis de residuos del modelo final.
5. Con los datos *Capacitacion_prediccion.txt* haga una predicción de la variable "JTS" (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.
6. Cree una nueva variable cualitativa que clasifique los aspirantes según el ingreso promedio mensual de sus anteriores empleos. Se deben dividir en cuatro categorías, menos de 840, entre 840 y 1200, entre 1200 y 1550, y mayor de 1550.
7. Reporte el puntaje medio de los aspirantes en la prueba según la clasificación anterior y zona.
8. Realice un análisis de varianza para estudiar si existen diferencias entre los puntajes medio de los aspirantes según la clasificación del ingreso y zona.
9. En el caso de conseguir diferencias significativas, realice pruebas de medias para determinar cuales son las medias que difieren.

Estudiantes

Grupo 1		Grupo 2		Grupo 3		Grupo 4	
14-10708	Irune Morales	13-10439	David Ferere	11-10093	María Bello	14-10607	Stefano Manelli
14-10843	Oriana Petitjean	11-11431	Aimee Salazar	11-10278	Albert Díaz	14-10853	Luis Pocaterra
14-11050	Maria Solorzano	14-11117	Horelyz Vásquez	12-10578	Fabio Suárez	14-11059	Silvio Strefezza
Grupo 5		Grupo 6		Grupo 7		Grupo 8	
		11-10041	Anna Antonini	12-11469	Juan Casilla	12-10613	Lelezka Duque
12-11018	Luis Millán	14-11052	María Sotillo	13-10805	Irina Marcano	13-10708	Genesis Kuffaty
10-10705	Edwin Sosa	14-11091	Doriana Troconiz	14-10630	Karen Martínez	13-11264	Rubmary Rojas
						13-11347	Carlos Serrada
Grupo 9		Grupo 10		Grupo 11		Grupo 12	
13-10131	Fabiana Becker	14-10063	Grileudy Avilan	14-10205	Aurivan Castro	11-10156	José Carmona
12-10040	Miriam Cedeño	07-41797	Wuidiana Ramos	14-11130	Sandra Vera	13-10248	Paola Castro
13-10228	Carmen Moncada			14-11127	Miguel Vélez	12-11499	Orlando Chaparro

Estudio de algas

Grupo 1		Grupo 2		Grupo 3		Grupo 4	
14-10406	Ian Goldberg	14-10005	Angélica Acosta	13-10886	Valentina Merola	12-10199	Nairelys Hernández
13-11223	Manuel Rodríguez	14-10515	Ana Ibarra	14-11140	Ángel Villamizar	10-10488	Edwin Murillo
13-11303	Abelardo Salazar	14-10869	Elvin Quero	12-11543	Mónica Zárate	13-11199	Jawil Ricauter
Grupo 5		Grupo 6		Grupo 7		Grupo 8	
14-10440	Angely Granados	13-10094	Laura Avila	03-35752	Gabriel Casique	14-10551	Guillermo Lapelosa
12-10882	Edymar Mijares	14-10820	Julio Pérez	12-11119	Carlos Leal	14-10650	Guillermo Matos
14-11069	Sara Teppa	13-11389	Guilianne Tavano	11-11509	Ariela Pardo	14-11167	Edgardo Zerpa
Grupo 9		Grupo 10		Grupo 11		Grupo 12	
12-10965	Luis Pacheco	14-10213	Annerys Chacín			14-10381	Isaac García
13-11238	Yaremi Rodríguez	14-10271	Steven Da Silva			14-10481	Ana Henriquez
		14-10529	Javier Herrera			14-10610	Arístides Mantoupulos

Capacitación

Grupo 1		Grupo 2		Grupo 3		Grupo 4	
14-10880	Yuni Quintero	13-10055	Ramsés Antolines	13-11148	Mario Quintero	14-10334	Jaqueline Farrach
14-10924	Germán Robayo	14-11377	David Piza	14-10950	Javier Rodríguez	14-10930	David Rodríguez
14-10944	Cristopher Rodríguez	11-10886	Ninfa Rodríguez	14-11088	Miguel Trejo	14-11082	Andrés Torres
Grupo 5		Grupo 6		Grupo 7		Grupo 8	
13-10428	Adriana Estrada	11-11252	Marisabel Cedeño	09-10066	Gabriel Austin	12-10403	Joseph Corona
14-11268	Luis Martín	14-10290	Juan De La Calle	11-10959	Alejandro Segovia	13-11520	Ángel Zambrano
14-10913	Andrés Rivas	14-10685	Simón Milano	07-41475	Daniel Rodríguez		
Grupo 9		Grupo 10		Grupo 11		Grupo 12	
12-11285	Raúl Bander	12-10774	Miguel Parra	11-10421	Andy Guevara	13-10191	David Cabeza
10-11005	Yezabel Rincón	10-11133	Anakaren Sosa	12-11002	Nathalie Vivas	13-10575	Pablo González
11-11020	Sergio Téran	11-11508	Vanessa Villalba			13-10838	Fabiola Martínez

4. Criterios de corrección para el proyecto

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.

4.1. Acerca de la 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

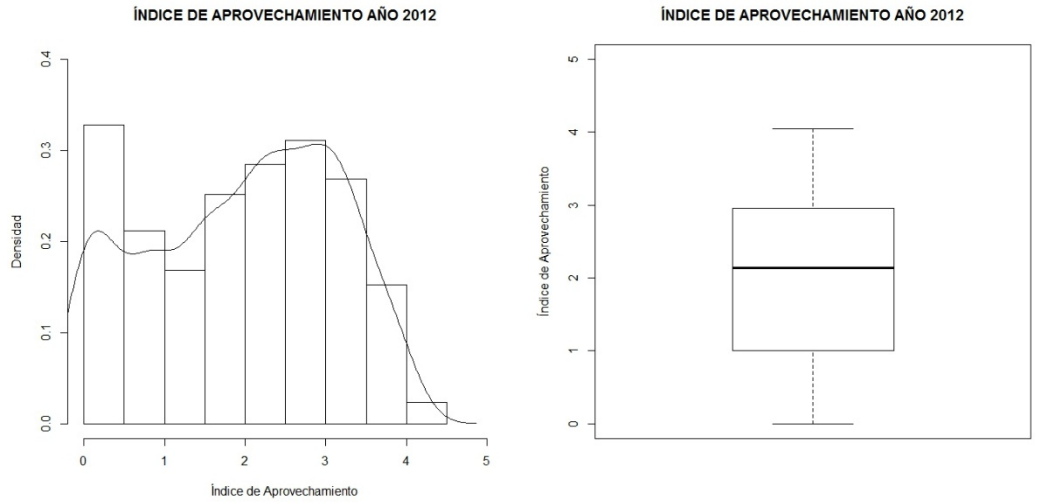
- No se aceptará presentación de resultados con manuscritos escaneados.
- Se anulará la evaluación de aquellos que compartan fotografías tomadas desde la pantalla de la computadora.
- No se aceptará un copy y paste de los resultados.
- No se aceptarán títulos de las gráfidcas generados por degecto en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

4.2. 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.

Tabla 1. Resumen estadístico para la variable Índice de Aprovechamiento

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

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.

5. Condiciones de entrega

a El informe debe ser entregado en forma electrónica y en formato ".pdf".

b La entrega se realizará al correo electrónico povallesgarcia@usb.ve a más tardar el viernes 23 de marzo de 2018. El asunto del correo DEBE ser: "*Proyecto. CO3321*".

c No se corregirá informes entregados fuera del tiempo establecido para la entrega.