Laboratorio 4
Estadística Computacional

Universidad Técnica Federico Santa María
Departamento de Informática

Sebastián Bórquez
<sborquez@alumnos.inf.utfsm.cl>

José García
<jigarcia@alumnos.inf.utfsm.cl>

Héctor Allende
<hallende@inf.utfsm.cl>

Rodrigo Naranjo
<rodrigo.naranjo@alumnos.usm.cl>

24 de julio de 2019

1. Método de Remuestreo: Bootstrapping

Contexto

Bootstrap es método para la inferencia estadística. Es utilizado ampliamente para estimar un estadístico (por ejemplo, la media) y su distribución. Esta distribución de muestreo se utiliza para calcular el error estándar de los parámetros estadísticos o del modelo y sus intervalos de confianza. A diferencia del enfoque estadístico clásico del cálculo de errores estándar e intervalos de confianza con fórmulas, este método tiene la ventaja de no necesitar ningún supuesto de los datos.

Su potencial se esconde en la capacidad de estimar estos estadísticos a partir de una muestra, cuando obtener toda la población o una muestra más grande es demasiado costoso.

La idea detrás del bootstrap es sencilla, esta consiste en realizar la inferencia sobre un estadístico (o estimador) para una población sobre datos de una muestra conocida. Para esto, el método toma **varias muestras con remplazo de la muestra conocida**. Luego, calcula el estadístico utilizando estas nuevas muestras.

Actividad - Students Performance in Exams (100 pts.)

En esta sesión, deberá utilizar y corroborar el método de bootstrapping. Para esto se les entrega dos archivos con las calificación de estudiantes en tres asignaturas. Estos corresponden a *population.csv* con todos los 450 estudiantes, y *sample.csv* una pequeña muestra de 45 estudiantes.

Debe contestar las siguientes preguntar utilizando para el desarrollo el software estadístico **R** o **Python**. **Ustedes deben implementar el algoritmo de Bootstrap**, ya que se evaluará su comprensión y aplicación de este.

Recuerde argumentar sus respuesta con sus resultados y agregar títulos, etiquetas y nombres a los ejes de sus gráficos.

1.- Escriba el **algoritmo** de boostrap para estimar la distribución de la media de una asignatura utilizando R remuestreos generados a partir de una muestra X.**(7 pts.)**

2.- Describa a la población (*population.csv*) para comparar los resultados. ¿Cuántos individuos hay? ¿Cuánto es la media de notas para cada asignatura? Grafique un histograma de notas para cada asignatura. **(9 pts.)**

3.- Describa su muestra (*sample.csv*). ¿Cuántos individuos hay? ¿Cuánto es la media de notas para cada asignatura? Grafique un histograma de notas para cada asignatura. **(9 pts.)**

4.- Determine un valor de R lo suficientemente grande. Para esto usted debe realizar los siguientes pasos para $R \in (1, 5, 25, 100, 1000)$ y usando la **asignatura de matemáticas**: **(45 pts.)**

- Obtenga la distribución de la media utilizando su algoritmo del ítem 1.
- Grafique la distribución de la media obtenida utilizando un histograma.
- Obtenga el intervalo de confianza de la media para un valor de confianza de 90 % (los valores de los percentiles 5 y 95).
- Compare con el valor de la media de la población, ¿se encuentra dentro del intervalo?

¿Cómo afecta el valor de R al resultado? ¿Qué valor es conveniente para utilizar bootstrapping?

5.- Para cada uno de las asignaturas, obtenga los intervalos de confianza para la media utilizando bootstrap para cada uno de los valores de confianza: 70 %, 80 %, 90 %, 95 % y 99 %. Comente sus resultados. ¿Concuerdan con los valores de la población? ¿Cómo afecta el valor de confianza? **(30 pts.)**

2. Sobre el desarrollo

Las sesiones y material usados serán hechas en R y Python. El desarrollo puede ser realizado con R o Python utilizando las herramientas presentadas en las sesiones. Las herramientas para el desarrollo son R Markdown y Jupyter Notebooks, respectivamente. Para usar R se recomienda trabajar en RStudio, y para Python usar Jupyter Notebooks junto con Spyder, recomendado trabajar con Anaconda.

3. Sobre la Entrega

El informe puede realizarse en parejas o tríos. El informe **debe incluir el código** que usó en la ejecución, por lo que es necesario que use notebooks en el trabajo. Se aplicarán **descuentos** por código desordenado, ilegible o no modularizado. Se recomienda leer las siguientes convenciones de código: https://github.com/google/styleguide. La fecha de entrega es **POR DEFINIR**. El archivo a subir **debe ser el notebook** con el que trabajaron con los scripts ejecutados en formato HTML (o .ipynb en caso de usar Jupyter Notebooks) con nombre "Nombre1Apellido1-Nombre2Apellido2" a la sección de entregas de Moodle. En caso de atrasos, si el atraso es de 1 día, la nota máxima será 80. 2 o más días tendrán nota 0.