Probabilidades y Estadística **10 de junio de 2022**

Tarea 2: Inferencia Estadística e Intervalos de Confianza

Profesores: *Elisa Irarrázaval, Ignacio Montegu*

1. Introducción

En la presente tarea, Ud. deberá aplicar sus conocimientos del curso para responder algunas preguntas prácticas. Para ello, deberá modelar un conjunto de problemas, ayudarse de librerías de `Python` que hagan parte del trabajo por usted, y elaborar un informe con análisis de los resultados obtenidos. El trabajo debe efectuarse en grupos de 3 personas, los mismos de la Tarea 1.

"Un tema es tener los datos, otro tema es saber qué hacer con ellos". El objetivo de esta tarea es que aprendan a trabajar con unas bases de datos *DataFrames* en `Python`, para que así puedan interactuar con variables y sus muestras, aplicando los conocimientos vistos en clases e investigando. También especificamente aprender a calcular intervalos de confianza e interpretar los resultados obtenidos con una mirada práctica.

2. Contexto

A pesar de que las criptomonedas existen hace más de 10 años, pero los últimos 3 años han tenido un crecimiento exponencial en su compra/venta y uso como instrumento de inversión. Desde la primera criptomoneda, el Bitcoin, han nacido nuevas divisas digitales, haciendo necesario diferenciarlas y clasificarlas. Las 3 categorías más comunes son:

- **Altcoins:** Originalmente aquellas con una capitalización bursátil más baja, hoy simplemente obedecen a ser las sucesoras del Bitcoin (Ej: Ethereum)
- **Stablecoins:** Este nuevo tipo de criptomonedas son 'tokens' que están asociados al valor de una moneda 'fiat' (como el dólar o el euro), a bienes materiales como el oro o los inmuebles, o a otra criptomoneda, reduciendo su volatilidad. (Ej: USDCoin)
- **Shitcoins:** Son criptomonedas que no aportan nada nuevo ni mejoran en lo absoluto la tecnología que usan para su funcionamiento, o incluso a veces, estafas. (Ej: Dogecoin)

3. Desarrollo

"Kaggle" es una página web especializada en temas de *Data Science*, que ofrece una serie de bases de datos y estudios estadísticos de interés. Para esta tarea ustedes deben descargar múltiples bases de datos del siguiente link, y utilizar los archivos correspondientes a Bitcoin, Ethereum, USDCoin y Dogecoin. Dentro del mismo enlace podrán encontrar la descripción de las variables. Deben hacerse un usuario para poder acceder al contenido, es recomendable entrar con su correo institucional. Las bases de datos a utilizar, junto con la descripción de las variables, está en la sección "Data". **Hint:** En la sección "Code"pueden encontrar una serie de estudios hechos con la misma base de datos, por lo que se pueden ahorrar muchos pasos de aprendizaje y al mismo tiempo indagar en temas estadísticos más complejos (recuerde usar referencias si es que utiliza algo de ahí, para que no sea plagio).

En base a sus conocimientos, y asumiendo poblaciones normales en las muestras de las criptomonedas, responda las siguientes preguntas ejecutando los comandos necesarios en `Python`:

(1) Calcule los intervalos de confianza para la media para Bitcoin para el total de datos (2013-2022) con una confianza del 95 % y con tamaños de muestra $n = 25, 100$ y 1,000 (`pandas` tiene métodos específicos para obtener muestras). ¿Qué es lo que sucede con el intervalo? Explique la relación entre tamaño de muestra y los intervalos obtenidos.

(2) Calcule los intervalos de confianza para la media para Bitcoin para el total de datos (2013-2022) con confianzas del 90 %, 95 % y 100 %, con tamaño de muestra $n = 100$. ¿Qué es lo que sucede con el intervalo? Explique la relación entre la confianza y los intervalos obtenidos.

(3) Conforme a lo explicado en la sección Contexto existe la teoría de que la variabilidad de las Altcoins es mayor a la de Bitcoin, la variabilidad de las Shitcoins es mayor a la de las Altcoins y finalmente, que las Stablecoins prácticamente no varían. ¿Qué puede decir al respecto? Construya intervalos de confianza que sustenten su argumentación.

(4) En el mundo de las inversiones se habla de *Bull-market* cuando el mercado es alcista, mientras que un *Bear-market* es un mercado a la baja. Un analista de criptomonedas dice que el 2021 fue un *Bull-market* y el 2022 hasta ahora está siendo un *Bear-market*. ¿Está de acuerdo con su afirmación? Utilice intervalos de confianza para parámetros de interés para respaldar su postura.

(5) Elija muestras adecuadas (representativas) para cada una de las 4 criptomonedas (Bitcoin, Ethereum, USD-Coin y Dogecoin) tratadas en esta tarea y caracterícelas en términos de media y varianza, para analizar su evolución histórica (año a año), represente esta información a través de tablas o gráficos ad-hoc.

(6) Elija otra criptomoneda dentro del pool posible del repositorio de "Kaggle", analice su comportamiento histórico (mediante intervalos de confianza) y comparando con su análisis del ítem 5 identifique a que categoría de criptomoneda dentro de las ya mencionadas se ubicaría ésta.

4. Sobre la entrega

En un archivo comprimido en `zip` deberá entregar:

- Un informe en `pdf` con portada y hasta diez páginas de contenido (incluyendo figuras). En la portada debe explicitarse el nombre completo y RUT de cada uno de los integrantes del grupo. En el informe, deben agregar todos los "outputs"del código (gráficos, tablas, etc.), respuestas a las preguntas y comentarios requeridos. No agreguén los códigos ejecutados.
- El código en `Python` utilizado en cada pregunta. El nombre de archivo deberá tener de la forma: `P` + *nro. de pregunta* + `.py` o `.ipynb` (p.ej., `P2.py`). Recuerden que cualquier código extraído de "Kaggle" o de otra fuente debe ser especificado, caso contrario será considerado como plagio. También recuerde que otro grupo no se considera como fuente válida de referenciar.

El archivo comprimido deberá tener por nombre `Tarea2_` + *Primer apellido de cada integrante del grupo* + `.zip`. P.ej., `Tarea2_SanchezVidalBravo.zip`. El incumplimiento de estas restricciones de nombres será sancionado con un descuento de un punto en la nota final de la tarea. El archivo comprimido deberá entregarse via buzón de Canvas. El plazo para entregar vence impostergablemente el **Viernes 24 de Junio a las 23:59 hrs.**

Aquél equipo que no entregue alguno de los documentos anteriormente específicados, o que entregue en plazo posterior al determinado, o que sus archivos no compilen con el dataset **original**, será evaluado con nota mínima 1,0. Cualquier caso de copia será calificado de la misma forma y será repasado a las autoridades correspondientes sujeto a sanciones adicionales.

5. Librerías de Python

En esta sección se entrega una ayuda introductoria para el uso de dos librerías que necesitará para esta tarea: `pandas` y `matplotlib` Finalmente se explica cómo instalarlas, si no están disponibles en el computador en que Ud. va a trabajar. Se asume que `Python` sí está instalado, junto con algún entorno de programación.

Si necesita más ayuda, recuerde que en internet puede encontrar tutoriales de ambas librerías y de `Python` en general para todos los niveles de dificultad, y que también puede preguntar a los profesores y ayudantes del curso.

5.1. Pandas para trabajar bases de datos

Necesitaran de esta libreria para trabajar de manera más ágil y fácil con las bases de datos. Se importa de la siguiente manera:

```
import pandas as pd
```

Si este paso arroja un mensaje de error, vea la sección 5.4.

Como norma general se le asigna la abreviación "pd". La recomendación es que mantengan esa norma porque se les hará más fácil encontrar bibliografía de esta manera. Para llamar a la librería solo deben usar `pd.` seguido de la función de pandas que quieran ocupar.

Dentro de las ventajas de usar esta librería es que permite crear tablones a partir de archivos y luego crear nuevas variables al aplicar funciones comunes de `python` en sus columnas manteniendo el orden original de la base de datos. Además, encontrarán de utilidad que la librería permite realizar filtros tanto a nivel de variables como de observaciones (filas), dándoles la oportunidad de crear nuevos DataFrames más pequeños y fáciles de manejar.

En este enlace encontrarán un resumen con los comandos más conocidos de `pandas`

5.2. Para cargar la base de datos

Una vez que se importe `pandas`, podrán importar desde un csv (archivo de valores separado por comas) la base de datos y crear un objeto DataFrame para poder seguir usandola.

Este comando es uno de los que cumple esa función:

```
bdd = pd.read_csv("Mi_Archivo.csv")
```

Para que no haya errores, es importante que `Mi_Archivo.csv` se encuentre en la misma carpeta que el archivo `python` (`.py` o `.ipynb` si utilizan Jupyter Notebook)

5.3. Estadística

La librería `scipy.stats` es muy útil para análisis estadístico, incluyendo cálculo de intervalos de confianza. Se importa de la siguiente manera:

```
import scipy.stats as st
```

Si este paso arroja un mensaje de error, vea la sección 5.4.

En particular la librería cuenta con métodos asociados a las distribuciones vistas en clases (Normal, t-Student y Chi-Cuadrado) para calcular intervalos de confianza, y probabilidades puntuales, de ser necesario.

5.4. Instalación de librerías

Si una librería no está instalada junto a `Python` en un computador dado, la consola entregará un mensaje de error de la forma `ImportError: No module named 'pepito'`. Para instalarla, escriba las siguientes dos líneas de código, reemplazando "pepito" por el nombre de la librería, entre comillas:

```
import pip
pip.main(["install", "pepito"])
```