Probabilidad y Estadística Aplicada

Tarea 3

12 de Julio de 2025

Estudiantes:

Franco Filardi (5.507.925-6)

Stefano Francolino (5.482.263-4)

Mateo Hernández (5.565.681-8)

Agustín Pose (5.386.051-8)

1. Introducción

El presente informe tiene como objetivo aplicar conceptos fundamentales de Probabilidad y Estadística para analizar fenómenos sociales y técnicos en el contexto uruguayo. A través de dos estudios de caso, se busca poner en práctica herramientas como la estadística descriptiva, el cálculo de percentiles y los test de hipótesis, interpretando los resultados en relación con preguntas relevantes desde el punto de vista económico y tecnológico.

En el **primer problema**, se analiza la distribución territorial de los hogares con mayores ingresos per cápita en Uruguay. A partir de datos de la Encuesta Continua de Hogares (ECH) correspondientes al período 2009–2019, se investiga si estos hogares están uniformemente distribuidos entre los 19 departamentos del país. La hipótesis nula supone una distribución equitativa, y se utiliza la prueba de bondad de ajuste chi-cuadrado para contrastar. Este análisis busca aportar evidencia empírica sobre la equidad territorial en la distribución del ingreso.

El **segundo problema** se enmarca en el ámbito universitario y aborda la calidad del servicio de internet en distintos edificios de la Universidad Católica del Uruguay. En particular, se compara la velocidad promedio de conexión entre el edificio Central y el edificio Semprún. Para ello, se toma una muestra de mediciones de velocidad realizadas en diversos horarios y se aplica una prueba t de Student para muestras independientes, con el objetivo de determinar si existen diferencias significativas entre ambas locaciones.

Ambos casos permiten ilustrar la utilidad de las herramientas estadísticas para fundamentar decisiones basadas en datos. Asimismo, se enfatiza el trabajo con muestras reales, la implementación de métodos con Python y la interpretación contextualizada de los resultados. El desarrollo de esta tarea también apunta a fortalecer el pensamiento crítico y la capacidad de análisis cuantitativo de los estudiantes.

2. Marco Teórico

El presente informe se apoya en conceptos fundamentales de la estadística descriptiva, la teoría de probabilidades y la inferencia estadística, aplicados a dos problemas concretos: la distribución territorial del ingreso per cápita **(Problema 1)** y la comparación de la calidad de la conexión a internet en dos edificios del campus universitario **(Problema 2)**. A continuación se presentan los fundamentos teóricos involucrados en ambos análisis.

2.1 Ingreso per cápita y quintiles

El ingreso per cápita se calcula dividiendo el ingreso total del hogar entre la cantidad de personas que lo integran. Este indicador es más representativo del bienestar individual que el ingreso total, ya que corrige por el tamaño del hogar y permite comparaciones más equitativas entre unidades domésticas.

A partir del ingreso per cápita, se pueden construir grupos ordenados según nivel de ingreso. Una clasificación habitual es en quintiles, que divide a la población en cinco grupos de igual tamaño. El uso de esta segmentación permite focalizar el análisis en los hogares con mayores recursos (quintil superior), y observar su distribución geográfica como indicador de equidad territorial.

2.2 Distribución empírica y frecuencias observadas

La distribución empírica de una variable categórica se construye a partir de la frecuencia relativa de ocurrencia de sus valores en una muestra. En este caso, se analizan las frecuencias observadas de hogares pertenecientes al quintil más alto de ingreso per cápita en cada uno de los 19 departamentos del país. Estas frecuencias se utilizan como base para contrastar si la riqueza está equitativamente distribuida entre las regiones o si se concentra en algunas de ellas.

2.3 Hipótesis de uniformidad y prueba chi-cuadrado

Para evaluar la equidad territorial se plantea como hipótesis nula (H_0) que los hogares de altos ingresos están distribuidos de forma uniforme entre los departamentos. La hipótesis alternativa (H_1) plantea que la distribución no es uniforme.

La herramienta estadística utilizada para este contraste es la prueba de bondad de ajuste chi-cuadrado (χ^2), que compara las frecuencias observadas (Oi) con las frecuencias esperadas (Ei) bajo H_0. El estadístico se calcula como:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

donde k representa la cantidad de categorías (departamentos). Bajo la hipótesis nula, este estadístico sigue una distribución chi-cuadrado con $k - 1$ grados de libertad. Si el valor calculado excede el valor crítico de referencia o si el p-valor es menor que el nivel de significancia α, se rechaza la hipótesis de uniformidad **(Turney, 2023)**.

2.4 Comparación de medias y prueba t de Student

En el segundo problema se desea evaluar si la velocidad promedio de internet en el edificio Central es significativamente menor que en el edificio Semprún. Para ello, se utiliza una prueba de hipótesis para dos medias poblacionales a partir de muestras independientes.

La prueba t de Student para muestras independientes con varianzas desiguales (conocida como test de Welch) permite evaluar la diferencia entre medias cuando no se puede asumir igualdad de varianzas **(Wasserman, 2004)**. El estadístico t se calcula a partir de las medias muestrales, los desvíos estándar y los tamaños de muestra de ambos grupos.

2.5 p-valor y decisión estadística

El p-valor representa la probabilidad, bajo la hipótesis nula, de obtener un resultado tan extremo o más que el observado. En pruebas unilaterales como la del Problema 2, si el p-valor es menor que el nivel de significancia α (por ejemplo, 0,05), se rechaza H_0, concluyendo que existe una diferencia significativa entre las medias.

2.6 Aplicaciones prácticas de los métodos

Ambos métodos aplicados —prueba χ^2 y prueba t— permiten responder preguntas relevantes desde el punto de vista social y organizacional. La evaluación de la distribución del ingreso aporta evidencia sobre desigualdades regionales, mientras que el análisis de la velocidad de internet permite diagnosticar problemas de infraestructura tecnológica y justificar mejoras en zonas críticas.

3. Metodología de Trabajo

La presente tarea fue abordada combinando análisis de datos reales simulados, implementación computacional en Python, cálculo manual de estadísticos y contrastes de hipótesis. Se adoptó una metodología empírico-computacional orientada a la exploración, validación e interpretación de los resultados. Las principales etapas fueron las siguientes:

3.1 Procesamiento de datos y limpieza

Para el Problema 1, se trabajó con una muestra de 5.000 hogares extraída de la Encuesta Continua de Hogares (ECH), abarcando el período 2009–2019. Se calculó el ingreso per cápita dividiendo el ingreso total del hogar entre la cantidad de personas, y se clasificaron los hogares en quintiles con la función qcut de Pandas. Posteriormente, se filtraron los hogares del quintil 5 (mayores ingresos) y se agruparon por departamento.

Para el Problema 2, se procesó una base de datos con mediciones de velocidad de internet en distintos edificios del campus universitario. Se filtraron los registros correspondientes a los edificios Central y Semprún, y se eliminaron observaciones faltantes.

3.2 Análisis descriptivo y estadísticos resumen

Se calcularon medidas de tendencia central y dispersión como media, desvío estándar y tamaño muestral para los subconjuntos definidos. En el caso del ingreso, se construyó una tabla de frecuencias de hogares del quintil superior por departamento. En el caso de la velocidad de internet, se resumieron los valores por edificio para evaluar diferencias preliminares.

3.3 Implementación del test chi-cuadrado manual

Con el objetivo de contrastar la hipótesis de distribución uniforme de hogares ricos por departamento, se calcularon las frecuencias esperadas bajo H_0 (asumiendo distribución equiprobable), y se implementó manualmente el estadístico χ^2 siguiendo la fórmula clásica. Se utilizó la función chi2.sf de scipy.stats para obtener el p-valor asociado.

3.4 Implementación del test t de Student

Para comparar las velocidades promedio entre Central y Semprún, se implementó el test t para muestras independientes con varianzas no iguales. Se calculó el estadístico t, los grados de libertad mediante la fórmula de Welch-Satterthwaite, y el p-valor correspondiente con la función t.cdf. La hipótesis alternativa planteada fue unilateral: H_1: μ_Central < μ_Semprún.

3.5 Visualización y representación gráfica

Se generaron gráficos de barras y diagramas de caja (boxplots) para visualizar la distribución de hogares ricos por departamento y las velocidades de conexión por edificio. Estas representaciones complementaron el análisis estadístico y facilitaron la comunicación de los hallazgos.

3.6 Interpretación y conclusiones

Se interpretaron los resultados obtenidos desde una perspectiva crítica, destacando las implicancias en términos de equidad territorial y calidad del servicio. El análisis permitió identificar zonas con mayor concentración de hogares de altos ingresos y edificios con deficiencias significativas en conectividad.

El código completo desarrollado para ambos problemas puede consultarse en el repositorio oficial del equipo: **https://github.com/AgustinPose/PyE_tarea3**

4. Problema 1: Distribución del Ingreso por Departamento

Preparación y transformación de los datos

Se trabajó con una muestra representativa de 5.000 hogares extraída de la Encuesta Continua de Hogares (ECH) para el período 2009–2019. El objetivo principal fue analizar la distribución territorial del 20% de los hogares con mayores ingresos per cápita.

4.1 Cálculo del ingreso per cápita

El primer paso fue calcular el ingreso per cápita de cada hogar, dividiendo el ingreso total del hogar entre la cantidad de personas que lo componen. Este indicador es más representativo del bienestar individual que el ingreso total, ya que corrige por el tamaño del hogar:

Ingreso per cápita= Cantidad de habitantes del hogar / Ingreso del hogar

4.2 Clasificación en quintiles

A partir del ingreso per cápita, se clasificaron los hogares en quintiles utilizando la función *pd.qcut* de Python, que divide la muestra en cinco grupos con la misma cantidad de observaciones. Se asignaron etiquetas del 1 (más bajos ingresos) al 5 (más altos ingresos).

4.3 Selección del quintil superior

Para focalizar el análisis, se filtraron los hogares que pertenecen al **quintil 5**, es decir, al 20% de mayor ingreso per cápita. Este subconjunto representa los hogares más acomodados de la muestra, objeto del análisis de equidad territorial.

4.4 Frecuencias observadas por departamento

Finalmente, se construyó una tabla con las **frecuencias observadas de hogares del quintil superior en cada uno de los 19 departamentos** del país. Se utilizó el método *value_counts()* para contar cuántos hogares ricos aparecen en cada categoría departamental.

La distribución resultante no fue homogénea: algunos departamentos como Montevideo concentran una proporción considerablemente mayor de hogares de altos ingresos, mientras que otros presentan cantidades muy reducidas. Esta información se representa en la Figura 1.

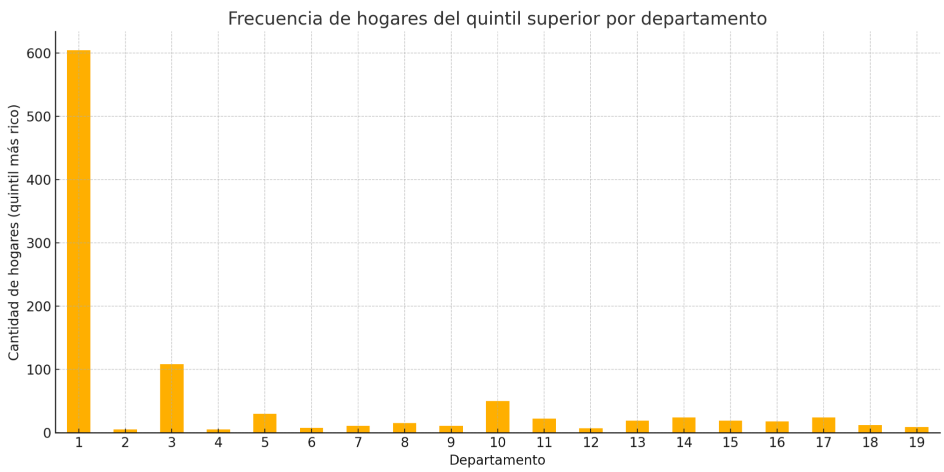

Figura 1 - **Frecuencia de hogares con alto ingreso per cápita por departamento.**

4.5 Frecuencias esperadas bajo la hipótesis de distribución uniforme

Bajo la hipótesis nula, se supone que los hogares del quintil superior se encuentran distribuidos de forma equitativa entre los 19 departamentos del país. En función del total de hogares en el quintil 5 (1001 hogares), la frecuencia esperada por departamento es constante y se calcula como:

$$Ei = \frac{1001}{19} = 52,68$$

Es decir, si la distribución fuera verdaderamente uniforme, se esperaría encontrar aproximadamente 52,68 hogares del quintil más alto en cada departamento.

4.6 Cálculo del estadístico χ^2

Para determinar si las frecuencias observadas se desvían significativamente de las esperadas, se utilizó la prueba de bondad de ajuste chi-cuadrado. El estadístico se define como:

$$X^2 = \sum_{i=1}^{19} \frac{(O_i - E_i)^2}{E_i}$$

donde O_i representa la frecuencia observada y E_i la frecuencia esperada para cada departamento. Al aplicar la fórmula con los datos obtenidos, se obtuvo el siguiente valor:

$$X^2 = 6276,65$$

4.7 Contraste de hipótesis

Con un nivel de significancia de $\alpha = 0,05$ y k = 19 categorías (departamentos), los grados de libertad son df = 18. El valor crítico correspondiente de la distribución χ^2 es:

$$X^2 0,95; 18 = 28,87$$

Dado que el estadístico calculado es ampliamente superior al valor crítico (6276,65 > 28,87), se concluye que se debe rechazar la hipótesis nula. Por lo tanto, existe una diferencia estadísticamente significativa entre la distribución observada y la distribución uniforme teórica.

4.8 Interpretación de los resultados

Los resultados obtenidos indican que los hogares con mayores ingresos per cápita no se distribuyen de forma homogénea entre los departamentos del país. La fuerte discrepancia entre las frecuencias observadas y las esperadas evidencia la concentración del ingreso en ciertas zonas, especialmente en el departamento de Montevideo, que presenta una cantidad de hogares ricos muy por encima del promedio.

Desde el punto de vista estadístico, el rechazo de la hipótesis nula confirma la existencia de desigualdades territoriales marcadas. Este hallazgo tiene implicaciones relevantes en el diseño de políticas públicas orientadas a la equidad regional, ya que refuerza la necesidad de implementar estrategias diferenciadas de desarrollo que atiendan a las disparidades económicas existentes entre departamentos.

5. Problema 2: Velocidad del internet en una Universidad

5.1 Filtrado de los datos

Se trabajó con un archivo que contiene mediciones de velocidad de conexión a internet (en Mb/s) realizadas en distintos edificios del campus universitario. Para este análisis comparativo se seleccionaron exclusivamente las observaciones correspondientes a los edificios “Central” y “Semprún”, con el objetivo de evaluar diferencias en el rendimiento del servicio entre ambas instalaciones.

El filtrado se realizó mediante la variable categórica "Edificio", utilizando funciones de la biblioteca Pandas en Python. Se generaron dos subconjuntos de datos: uno con las velocidades medidas en el edificio Central y otro en Semprún, conservando para cada caso las observaciones válidas sin valores perdidos.

5.2 Estadísticos descriptivos

Para cada grupo de datos (Central y Semprún), se calcularon los siguientes estadísticos:

- La media aritmética, que representa la velocidad promedio observada.

- El desvío estándar muestral, que indica la dispersión de las mediciones respecto al promedio.

- El tamaño muestral, es decir, la cantidad de mediciones disponibles por edificio.

Estos valores se utilizarán luego para la prueba de hipótesis. Un desvío estándar elevado en relación a la media, como se observó especialmente en el edificio Central, sugiere alta variabilidad en la calidad del servicio.

5.3 Cálculo del estadístico t

El objetivo es evaluar si la velocidad promedio en el edificio Central es significativamente menor que en Semprún. Para ello, se plantea la siguiente hipótesis:

- H_0: μ_Central = μ_Semprún (no hay diferencia)

- H_1: μ_Central < μ_Semprún (la media en Central es menor)

Se aplica un *test t* para dos muestras independientes con varianzas posiblemente diferentes. El estadístico t se calcula según la fórmula:

$$t = \frac{X_1 - X_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Donde:

- X1 y X2 son las medias de Central y Semprún
- s1, s2 son los desvíos estándar muestrales
- n1, n2 son los tamaños muestrales

El valor obtenido del *estadístico t* servirá luego para calcular el p-valor o comparar con un valor crítico, a fin de decidir si se rechaza o no la hipótesis nula.

5.4 Cálculo del p-valor

Una vez obtenido el estadístico $t = -8,6605$, se procedió a calcular el **p-valor** asociado a la prueba de hipótesis unilateral inferior. Este valor representa la probabilidad, bajo la hipótesis nula, de obtener un resultado tan extremo o más en la dirección planteada por la hipótesis alternativa.

Se utilizó una **prueba t de Student para muestras independientes**, sin asumir igualdad de varianzas, tal como lo indica el enunciado del problema. Los grados de libertad fueron calculados utilizando la fórmula general para este tipo de pruebas, obteniéndose un valor de $df \approx 132,47$.

El p-valor se obtuvo con la función `t.cdf()` de `scipy.stats` y resultó ser prácticamente nulo:

$$\text{p-valor} \approx 0{,}00000$$

Esto indica una altísima evidencia estadística en contra de la hipótesis nula.

5.5 Decisión estadística

El valor del p-valor obtenido fue claramente inferior al nivel de significancia prefijado $\alpha = 0{,}05$. Dado que $p < \alpha$, se **rechaza la hipótesis nula**. Esto significa que existe evidencia estadística suficiente para afirmar que la velocidad promedio de internet en el edificio Central es significativamente menor que en el edificio Semprún.

5.6 Interpretación del resultado

Los resultados de la prueba t sugieren una diferencia estadísticamente significativa entre las velocidades promedio de internet de los dos edificios. El edificio Central presenta un rendimiento significativamente menor, lo que podría impactar negativamente en la experiencia académica de estudiantes y docentes.

Estas diferencias podrían justificar intervenciones específicas en la infraestructura tecnológica del edificio Central, con el fin de equiparar el acceso y la calidad del servicio de internet dentro del campus universitario.

6. **Conclusión**

En el presente trabajo se abordan dos problemas aplicados de estadística inferencial, con el objetivo de evaluar situaciones reales mediante herramientas como intervalos de confianza y pruebas de hipótesis.

En el primer problema, se analizó la distribución territorial del 20% de los hogares con mayores ingresos per cápita a partir de una muestra proveniente de la Encuesta Continua de Hogares (ECH). A través del cálculo del ingreso per cápita, la clasificación en quintiles y el análisis de frecuencias observadas por departamento, se evidenció una alta concentración de hogares ricos en determinadas zonas del país. La hipótesis de una distribución uniforme fue contrastada mediante una prueba de bondad de ajuste chi-cuadrado, cuyo resultado ($\chi^2 = 6276{,}65$) permitió rechazar claramente la hipótesis nula. Este hallazgo confirma la existencia de desigualdades territoriales marcadas, con implicancias relevantes para el diseño de políticas públicas orientadas a una mayor equidad regional.

En el segundo problema, se evaluó si la velocidad promedio de internet en el edificio Central de la universidad era menor que en el edificio Semprún. Para ello, se aplicó una prueba t de Student para muestras independientes sin asumir igualdad de varianzas. El cálculo del p-valor permitió determinar si existía evidencia suficiente para rechazar la hipótesis nula. El análisis contribuye a detectar posibles desigualdades en el acceso a servicios digitales entre distintos edificios, lo cual puede tener impacto en la experiencia académica de estudiantes y docentes.

Ambos estudios muestran cómo el análisis estadístico puede ser una herramienta útil para fundamentar decisiones en contextos sociales, económicos y educativos **(Montgomery & Runger, 2014)**, permitiendo ir más allá de observaciones descriptivas hacia conclusiones respaldadas por evidencia cuantitativa.

7. Bibliografía

- Decuadro, S. (2025, 8 de junio). Test de hipótesis [Presentación de clase]. Facultad de Ingeniería, Universidad Católica del Uruguay.
- Decuadro, S. (2025, 8 de junio). Intervalos de confianza [Presentación de clase]. Facultad de Ingeniería, Universidad Católica del Uruguay.
- Devore, J. L. (2011). Probabilidad y estadística para ingeniería y ciencias (8.ª ed.). Cengage Learning.
- Montgomery, D. C., & Runger, G. C. (2014). Applied statistics and probability for engineers (6th ed.). Wiley.
- Ross, S. M. (2014). Introduction to probability and statistics for engineers and scientists (5th ed.). Academic Press.
- Statology. (s. f.). How to Report Chi-Square Results in APA Format. Recuperado de https://www.statology.org/how-to-report-chi-square-results/
- Statology. (2023, junio 22). Turney, S. Chi-Square Goodness of Fit Test | Formula, Guide & Examples. Recuperado de https://www.scribbr.com/statistics/chi-square-goodness-of-fit/
- Wasserman, L. (2004). All of statistics: A concise course in statistical inference. Springer.