FACULTAD DE CIENCIAS SOCIALES

Q-LAB

# SÍLABO 2025

## I. INFORMACIÓN GENERAL

| | |
|---|---|
| Nombre del curso | : Big Data y Analytics para la Gestión Pública Peruana |
| Código del curso | : -- |
| Carácter | : Curso de Formación Continua |
| Créditos | : 2 |
| Número de horas de teoría | : 18 |
| Número de horas de práctica | : 6 |
| Profesor del curso | : Víctor Fuentes (vfuentes@pucp.edu.pe) |
| Horario clases | : Sábados y domingos, 8:30am-11:30am<br>Miércoles, 7:00pm-10:00pm |

## II. SUMILLA

La explosión del Big Data puso en duda la utilidad de otras herramientas como los datos de encuestas. Sin embargo, estas se han transformado y explotado nuevas tecnologías mediante encuestas web, telefónicas, etc. En este contexto, existe un debate sobre qué es lo que más útil: big data o good data. Este curso busca explorar ambos lados del debate, por lo que se divide en dos bloques: técnicas de muestreo y computación en paralelo. En ambos casos se combinan clases teóricas con prácticas y se utilizarán programas estadísticos como Stata, R (primer bloque) y Python (segundo bloque).

En el primer bloque, se estudiarán los conceptos básicos sobre muestreo, el enfoque basado en el diseño vs en el modelo y el cálculo del poder estadístico. Luego, se verán las técnicas de muestreo: muestreo aleatorio simple, muestreo aleatorio estratificado, muestreo por conglomerados y diseños complejos en los cuales se aplican las técnicas previamente vistas en simultáneo. Asimismo, se estudiarán conceptos vinculados al muestreo como los factores de expansión, efecto de diseño y la correlación intracluster, que serán calculados de forma empírica. Finalmente, se explorarán otras formas de muestreo, como el muestreo proporcional al tamaño o con probabilidades desiguales.

En el segundo bloque, se estudiarán los conceptos básicos sobre computación en paralelo, los teoremas de la computación en escala y los paradigmas de los tipos de procesamiento serial y en paralelo. Además, se abordará la medición de aceleración tras aplicar métodos en paralelo y se diferenciará el procesamiento en CPU y GPU (tarjetas gráficas). Luego, se presentará la computación en paralelo a través del estándar Message Passing Interface (MPI) y su aplicación en Python. Finalmente, se presentará la biblioteca de computación en paralelo y distribuida Dask; la cual se utilizará para realizar ejercicios de ETL (Extract, Transform and Load) y ML (Machine Learning).

## III. RESULTADOS DE APRENDIZAJE

Este curso proporcionará herramientas conceptuales, herramientas de implementación, pero también aplicaciones en temas de ciencias sociales. Así, los estudiantes podrán conocer las ventajas y desventajas de las principales técnicas de muestreo, y cómo el diseño muestral de una encuesta debe ser considerado al utilizarla. Por otro lado, las herramientas proporcionadas sobre la computación en paralelo permitirán agilizar ciertos procesos, cuando estos son repetitivos o cuando se trabaja con bases de datos de gran tamaño. Esto brindará nuevas oportunidades de procesamiento al estudiante que complete satisfactoriamente el curso.

## IV. CONTENIDO DEL CURSO

### TÉCNICAS DE MUESTREO (TM)

**1. Conceptos generales**

En esta sección se presentan de manera general los conceptos básicos sobre el muestreo (población, muestra, muestreos con y sin reemplazamiento, entre otros). El trade-off sesgo-varianza, el paradigma Total Survey Error. Se contrastará el enfoque basado en el modelo con el enfoque basado en el diseño. Se estudiará bajo el enfoque basado en el modelo el cálculo de poder y tamaños de muestra.

Biemer, P. (2018)
GFCLST, Cap 1-2.2
Valdivieso (2020). Cap. 1

**2. Muestreo Aleatorio Simple (SRS)**

En esta sección se presenta la justificación para usar muestreos probabilísticos y el primer diseño, el Muestreo Aleatorio Simple. Esta es la técnica de muestreo básica que servirá de benchmark para las más sofisticadas. Se presentan las propiedades de la distribución muestral: insesgadez, variabilidad entre muestras diferentes y que el diseño sea medible. Se estudia como el Teorema Central del Límite y la Ley de los Grandes Números se utilizan en el muestreo. Se discuten las ventajas y desventajas de este método, su versión con y sin reemplazamiento y por qué funciona como un punto de comparación para los demás métodos. También se discuten cuestiones sobre el tamaño de la muestra, la corrección por finitud y los pesos probabilísticos.

Lumley (2010). Cap. 2.1
GFCLST, Cap 3-4.3
Valdivieso (2020). Cap. 2

**3. Muestreo aleatorio estratificado**

En esta sección se presenta el muestreo aleatorio estratificado como una mejora de la eficiencia del SRS. Justificación para estratificar. Dominios vs estratos. Sobremuestreo, estimadores, factores de expansión y tamaños de muestra. Se discuten las ventajas de este esquema respecto al MAS y se presenta el efecto del diseño (DEFF).

Lumley (2010). Cap. 2.2-2.6
GFCLST, Cap 3 y 4.5
Valdivieso (2020). Cap. 3

**4. Muestreo por conglomerados**

En esta sección se presenta el muestreo por conglomerados como una forma de ahorrar costos de la encuesta. Justificación para usar clústeres pese a la reducción de la precisión. Clusterización a uno y múltiples niveles, estimadores y tamaños de muestra. Se presenta el concepto de correlación intra-cluster, su impacto en el muestreo y su relación con el DEFF. Enfoques alternativos como el muestreo con probabilidades desiguales y probabilidad proporcional al tamaño (PPS sampling). Finalmente, se presentan los estimadores más usados en la literatura: Horvitz-Thompson y Hájek

Lumley (2010). Cap. 3
GFCLST, Cap 4.4-4.10
Valdivieso (2020). Cap. 4

**5. Diseños complejos**

En esta sección se discuten los diseños complejos de muestreo en los que se aplican en simultáneo las técnicas estudiadas. Como parte de esto, se estudia los pesos de muestreo en mayor detalle y su importancia. Se derivan por ingeniería inversa los tamaños de muestra. Se desarrollará un ejemplo con encuestas peruanas para resolver dudas prácticas.

Lumley (2010). Cap. 3, 8
Couper, P. (2013)
Valdivieso (2020). Cap. 5

## COMPUTACIÓN EN PARALELO (CP)

**6. Introducción a la CP y la paralelización a mano**

En esta sección se presentan los conceptos introductorios sobre computación en paralelo. Computación serial vs CP. Justificación de uso de CP. Memoria compartida vs distribuida. Tareas embarrassingly parallelizable. Escalabilidad, la ley de Amdahl y la ley de Gustafson. El método PCAM de Foster. Además, se estudia la interfaz MPI y, en específico, la distribución para Python mpi4py. Este permitirá dividir tareas y utilizar los procesadores de una computadora en simultáneo bajo el método PCAM de Foster para entender cómo funciona la paralelización.

Pacheco, P. (2011) Cap. 1-3
Robey & Zamora (2020) Cap. 1, 8
Whitaker, S. D. (2018)
Dalcín, L., Paz, R., Storti, M., & D'Elía, J. (2008).

**7. Unidades de Procesamiento Gráfico (GPU) y OpenCL**

En esta sección se discuten las diferencias entre los CPUs y GPUs, la medición de poder computacional, eficiencia y costos de comunicación. Se utiliza el paquete OpenCL para realizar trabajos en paralelo utilizando la tarjeta gráfica GPU. Se realizan ejemplos utilizando computación en la nube serverless mediante COLAB.

Robey & Zamora (2020) Cap. 9

### 8. DASK

En esta sección se presenta Dask, una distribución nativa a Python que permite escalar aplicaciones de Pandas, Scikit-Learn, Numpy fácilmente. Dask permitirá disparar un cluster de manera local y en la nube (mediante COLAB) para trabajar con múltiples núcleos y luego con clúster distribuido. Se realizarán aplicaciones de Dask para ejercicios de ETL (Extract, Transform and Load) y ML (Machine Learning).

Documentación oficial Dask: Getting started, fundamentals. https://docs.dask.org/en/latest/
Documentación oficial Dask-ML: Use, integration, develop. https://ml.dask.org

## V. METODOLOGÍA

El contenido del curso se desarrollará mediante sesiones sincrónicas. El alumno aprenderá el contenido del curso, también, mediante el desarrollo de ejercicios. Finalmente, el alumno plasmará los conocimientos aprendidos en dos trabajos.

## VI. EVALUACIÓN

| TIPO DE EVALUACIÓN | PONDERACIÓN |
|---|---|
| Una tarea breve (TB) | 20% |
| Trabajo 1: Técnicas de Muestreo (T1) | 45% |
| Trabajo 2: Computación en paralelo (T2) | 35% |

Fórmula de calificación: Nota Final = 0.25*TB+0.40*T1+0.35*T2

## VII. BIBLIOGRAFÍA[1]


**TÉCNICAS DE MUESTREO (TM)**

Amaya, A., Biemer, P., Kinyon, D., (2020) Total Error in a Big Data World: Adapting the TSE Framework to Big Data. Journal of Survey Statistics and Methodology, Volume 8, Issue 1, Pages 89–119, https://doi.org/10.1093/jssam/smz056

Biemer, P. (2018) Big Data Can't Replace Surveys, But They Can Work Together. RTI International. https://www.rti.org/insights/big-data-can%E2%80%99t-replace-surveys-they-can-work-together

Couper, P. (2013). Is the sky falling? New technology, changing media, and the future of surveys. In Survey Research Methods (Vol. 7, No. 3, pp. 145-156).

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). Survey methodology. John Wiley & Sons. [GFCLST]

Lumley, T. (2011). Complex surveys: a guide to analysis using R (Vol. 565). John Wiley & Sons.

Valdivieso, L. (2020) Notas de técnicas de muestreo. Departamento Académico de Ciencias. PUCP

**COMPUTACIÓN EN PARALELO (CP)**

Dalcín, L., Paz, R., Storti, M., & D'Elía, J. (2008). MPI for Python: Performance improvements and MPI-2 extensions. Journal of Parallel and Distributed Computing, 68(5), 655-662.


[1] La bibliografía obligatoria debe estar disponible en el PAIDEIA del curso.

Pacheco, P. (2011). An introduction to parallel programming. Elsevier.

Robey, R., & Zamora, Y. (2020). Parallel and High Performance Computing. Simon and Schuster.

Whitaker, S. D. (2018). Big data versus a survey. The Quarterly Review of Economics and Finance, 67, 285-296.

## VIII. CRONOGRAMA

Cada clase comprenderá horas de clases teóricas como prácticas dirigidas. En total, el curso comprende 24 horas de clases.

| FECHA | TEMA/CONTENIDOS | HORAS |
|---|---|---|
| 4 enero | Muestreo, conceptos, paradigma modelo vs diseño, cálculos de poder | 3 |
| 5 enero | Muestreo aleatorio simple | 3 |
| 8 enero | Muestreo aleatorio estratificado **[Tarea Breve]** | 3 |
| 11 enero | Muestro por conglomerados | 3 |
| 12 enero | Diseños complejos y enfoques alternativos **[Trabajo 1]** | 3 |
| 15 enero | Introducción a la CP y Paralelización “a mano” | 3 |
| 18 enero | Unidades de Procesamiento Gráfico (GPU) y OpenCL | 3 |
| 19 enero | DASK (ETL + ML) **[Trabajo 2]** | 3 |
| | **Total** | **24** |

La evaluación de todos los trabajos contemplará el respeto de los derechos de autor. En este marco, cualquier indicio de plagio tendrá como consecuencia la nota cero. Esta medida es independiente del proceso disciplinario que la Secretaría Académica de la facultad estime iniciar según cada caso. Para obtener más información sobre el citado visitar el siguiente sitio web: www.pucp.edu.pe/documento/pucp/plagio.pdf