UNIVERSIDAD AUTÓNOMA DE MADRID

33677 - SEMINARIO: ANÁLISIS DE VALORES PERDIDOS E IMPUTACIÓN DE RESPUESTAS

Información de la asignatura

Código - Nombre: 33677 - SEMINARIO: ANÁLISIS DE VALORES PERDIDOS E IMPUTACIÓN DE RESPUESTAS
Titulación: 787 - Máster en Metodología de las Ciencias del Comportamiento y de la Salud (2023)
Centro: 105 - Facultad de Psicología
Curso Académico: 2023/24

1. Detalles de la asignatura

1.1. Materia

-

1.2. Carácter

Optativa

1.3. Nivel

Máster (MECES 3)

1.4. Curso

2 y 1

1.5. Semestre

Segundo semestre

1.6. Número de créditos ECTS

3.0

1.7. Idioma

Castellano

1.8. Requisitos previos

-

1.9. Recomendaciones

Se recomienda que el estudiante tenga algo de manejo con R para seguir más fácilmente el ritmo de las clases (aunque no es imprescindible y el código de R se proporciona al alumno para evitarle tener que crearlo él mismo).

1.10. Requisitos mínimos de asistencia

No es obligatoria la asistencia, pero sí es altamente recomendable para trabajar sobre los ejemplos prácticos que se exponen en el seminario

1.11. Coordinador/a de la asignatura

-

https://autoservicio.uam.es/paginas-blancas/

1.12. Competencias y resultados del aprendizaje

1.12.1. Competencias

CG1 - Tomar conciencia de la importancia de la metodología en la adquisición del conocimiento científico, así como de la diversidad metodológica existente para abordar distintos problemas de conocimiento.

CG2 - Desarrollar el razonamiento crítico y la capacidad para realizar análisis y síntesis de la información disponible.

CG3 - Saber identificar las necesidades y demandas de los contextos en los que se exige la aplicación de herramientas metodológicas y aprender a proponer las soluciones apropiadas.

CG5 - Obtener información de forma efectiva a partir de libros, revistas especializadas y otras fuentes.

CG6 - Desarrollar y mantener actualizadas competencias, destrezas y conocimientos según los estándares propios de la profesión.

CB10 - Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida auto-dirigido o autónomo.

CE4 - Analizar datos identificando diferencias y relaciones. Esto implica conocer las diferentes herramientas de análisis así como su utilidad y aplicabilidad en cada contexto.

1.12.2. Resultados de aprendizaje

-

1.12.3. Objetivos de la asignatura

El objetivo general de este seminario es el de proporcionar a los estudiantes la formación teórica y práctica fundamental para enfrentarse de forma crítica y razonada a los valores perdidos que pueden generarse en cualquier investigación empírica. Los datos o valores perdidos se producen cuando las observaciones que se pretenden medir no consiguen recogerse por la razón que sea. Este seminario pretende demostrar las implicaciones asociadas a la pérdida de datos, en particular el peligro de que los resultados estén sesgados debido. El estudiante aplicará en todo momento los conceptos que se van viendo con tareas y ejercicios que le permitan ir poniendo en práctica los contenidos del seminario. Se pondrá especial hincapié en:

a) comprender las implicaciones de los datos perdidos

b) cómo diagnosticar el impacto probable de los datos perdidos en los análisis estadísticos

c) los métodos estadísticos para evitar los sesgos.

Se tratará, en la medida de lo posible, que los conceptos y las técnicas teóricas básicas se alternen con ejercicios prácticos mediante el soporte informático necesario (v.g,. R, SPSS, etc.).

1.13. Contenidos del programa

Unidad temática 1.- Introducción a los valores perdidos

1.1.- Problema omnipresente en investigación

1.2.- Patrones de pérdida de datos

1.3.- Mecanismos de pérdida de datos (MCAR, MAR y MNAR)

Unidad temática 2.- Métodos clásicos de imputación

2.1.- Métodos convencionales

2.1.2.- Eliminación por lista

2.1.2.- Análisis de casos disponibles

2.1.3.- Reemplazamiento por la media

2.1.4.- Método hot-deck

2.1.5.- Imputación por regresión

2.1.6.- Imputación por regresión estocástica

Unidad temática 3.- Métodos modernos de imputación: Máxima verosimilitud

3.1.- Introducción a máxima verosimilitud

3.2.- Aplicación del método de estimación por máxima verosimilitud con librerías de R (v.g., lavaan) a modelos estadísticos concretos

Unidad temática 4.- Métodos modernos de imputación: Imputación múltiple

4.1.- Introducción a la imputación múltiple

4.2.- Aplicación del método de imputación múltiple en R (v.g., mice, mdmb, jomo, etc.) a modelos estadísticos concretos

Unidad temática 5.- Cuestiones avanzadas de pérdida de datos

5.1.- Compatibilidad entre el modelo sustantivo y el modelo de imputación

5.2.- El problema de los modelos con efectos no lineales (interacción, polinómicos, etc.)

5.3.- Modelos MNAR: Modelo de Heckman para el sesgo de selección de la muestra

1.14. Referencias de consulta

Hay una documentación elaborada por el profesor muy extendida para el seminario que se puede utilizar para seguir perfectamente el seminario. Además, las referencias que pueden ayudar a comprender el seminario se listan a continuación. Precedidas de un asterisco están las referencias básicas y recomendadas para una primera aproximación al problema de los valores perdidos.

* Allison, P. D. (2001). Missing Data. Sage University Papers Series on Quantitative Applications in the Social Siences , 07-136. Thousand Oaks, CA: Sage.

Asparouhov T. & Muthén B. (2010). Multiple Imputation with Mplus . Technical Report. www.statmodel.com

Enders, C. K. (2001). The Impact of Nonnormality on Full Information Maximum-Likelihood Estimation for Structural Equation Models With Missing Data. Psychological Methods, 6, 4, 352 – 370.

Enders, C. K. (2003). Using the EM algorithm to estimate coefficient alpha for scales with item-level missing data. Psychological Methods, 8, 322-337.

Enders, C. K. (2004). The Impact of Missing Data on Sample Reliability Estimates: Implications for Reliability Reporting Practices. Educational and Psychological Measurement, 64, 419–436. DOI: 10.1177/0013164403261050

* Enders, C. K. (2010). Applied missing data analysis. New York, NY: Guilford Press.

Enders, C. K. (2011). Missing Not at Random Models for Latent Growth Curve Analysis. Psychological methods, 16 (1), 1 – 16.

* Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle (Ed.), Statistical strategies for small sample research (pp. 1–29). Thousand Oaks, CA: Sage.

Graham, J. W., Taylor, B. J., Olchowski , A. E., & Cumsille , P. E. (2006). Planned missing data designs in psychological research. Psychological Methods, 11, 323–343.

Holman, R. & Glas , C. A. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. The British Psychological Society, 58, 1–17.

Heckman, J. T. (1979). Sample selection bias as a specification error. *Econometrika , 47,* 153–161. Horton N.J., Lipsitz S.R., & Parzen , M. (2003) A potential for bias when rounding in multiple imputation. American Statistician, 57, 229-232.

Kadengye , D.T., Cools, W., Ceulemans , E., & van den Noortgate , W. (2012). Simple imputation methods versus direct likelihood analysis for missing item scores in multilevel educational data. Behavior Research Methods, 44 (2), 516 – 531.

Korobko , O. B., Glas , C. A., Bosker , R. J., & Luyten , J. W. (2008). Comparing the Difficulty of Examination Subjects with Item Response Theory. Journal of Educational Measurement, 45, 2, 139 – 157.

Little, R. J. A., & Rubin, D. B. (2002). Statistical analysis with missing data (2nd ed .). Hoboken, NJ: Wiley.

Muthen , B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. Psychometrika , 52, 431-462.

Muthén , L. K., & Muthén , B. O. (1998–2011). Mplus user's guide (Sixth Edition). Los Angeles, CA: Muthén & Muthén .

Pimentel, J. L. (2005). Item Response Theory modeling with nonignorable missing data. Ph.D. thesis, University of Twente, The Netherlands.

Robitzsch , A. & Rupp, A. A. (2008). Impact of Missing Data on the Detection of Differential Item Functioning: The Case of Mantel- Haenszel and Logistic Regression Analysis. Educational and Psychological Measurement, 69, 1, 18 – 34. DOI: 10.1177/0013164408318756

Rose, N., von Davier , M., & Xu, X. (2010). Modeling Nonignorable Missing Data With Item Response Theory. Research Report.

ETS, Princeton, New Jersey

Schafer, J.L. (1997) Analysis of Incomplete Multivariate Data. Chapman & Hall, London.

* Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. Psychological Methods, 7, 147–177. doi:10.1037/1082-989X.7.2.147

Sijtsma , K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika , 74, 1, 107–110.

Sijtsma , K., & Van der Ark. L.A. (2003). Investigation and Treatment of Missing Item Scores in Test and Questionnaire Data. Multivariate Behavioral Research, 38:4, 505-528, DOI: 10.1207/s15327906mbr3804_4

Van Ginkel , J.R., Van der Ark, L.A., & Sijtsma , K. (2007). Multiple Imputation of Item Scores in Test and Questionnaire Data, and Influence on Psychometric Results. Multivariate Behavioral Research, 42:2, 387 – 414.

Van Ginkel , J. R., Sijtsma , K., Van der Ark, L. A., & Vermunt , J. K. (2010). Incidence of missing item scores in personality measurement, and simple item-score imputation. Methodology, 6, 17–30.

Zhang, B. & Walker, C. M. (2008). Impact of Missing Data on Person--Model Fit and Person Trait Estimation. Applied Psychological Measurement , 32, 466 – 479. DOI: 10.1177/0146621607307691

2. Metodologías docentes y tiempo de trabajo del estudiante

2.1. Presencialidad

* No se requiere un porcentaje mínimo de presencialidad (como se ha señalado más arriba, se recomienda mucho ir a las clases para seguir el ritmo de la asignatura)

2.2. Relación de actividades formativas

Para conseguir que el estudiante desarrolle las competencias descritas en el apartado 1.12 el seminario se desarrolla siguiendo la siguiente dinámica:

Clases teórico-prácticas:

Se darán un total de cuatro sesiones teórico-prácticas (opcionalmente, y si el calendario lo permite, se da una quinta sesión práctica para enseñar simulación de pérdida de datos en Mplus o R) . Se exponen los conceptos teóricos de cada uno de los temas del seminario durante la primera mitad de la clase y luego, los estudiantes, individualmente o por parejas, utilizando el ordenador , ponen en práctica los conocimientos explicados previamente.

El trabajo práctico del estudiante en el seminario es el siguiente:

En las dos primeras sesiones el estudiante aprende a explorar e inspeccionar los valores perdidos en bases de datos proporcionadas por el profesor. Esto se hace con R y SPSS. Además, el estudiante aprende a simular con el programa R o SPSS los mecanismos de valores perdidos explicados durante la teoría y así como a utilizar los métodos de imputación clásicos. El objetivo es que el estudiante compruebe por sí mismo los problemas de eficiencia, consistencia y potencia que acarrea el tratamiento de los valores perdidos con los métodos clásicos y, especialmente, cuando la pérdida de datos es MNAR.

En las dos siguientes sesiones el estudiante trabaja con los métodos de imputación modernos, máxima verosimilitud (ML) e imputación múltiple (MI) con el software R (librerías mice y lavaan) y con Mplus. El estudiante trabajará los valores perdidos en modelos estadísticos básicos habituales (T de Student , ANOVAs …) y en modelos de regresión lineal múltiple, análisis factorial exploratorio y análisis factorial confirmatorio. Es importante terminar el seminario aprendiendo a manejar los procedimientos modernos de tratamientos de valores perdidos con modelos estadísticos concretos.

Trabajo del estudiante:

Se reserva una sesión del seminario para que al final de la misma el estudiante presente el trabajo de la asignatura. Los trabajos serán de carácter empírico y/ o de simulación (ver apartado 3. Sistemas de evaluación y porcentaje en la calificación final)

Tutorías:

El estudiante podrá solicitar las tutorías que considere necesarias al profesor, ya sean individuales o en grupos. En estas tutorías habitualmente se establecen las líneas de los trabajos finales de la

asignatura y se ayuda al estudiante a llevar a cabo un trabajo de simulación bajo los distintos mecanismos de pérdida de datos explicados.

TIEMPO DE TRABAJO ESTIMADO PARA EL ESTUDIANTE

Tipo de actividad	Lugar	Horas
Clases teórico/prácticas. Teoría	Aula de clase	11/12 horas de exposición

Clases teórico/prácticas Prácticas	*Aula de clase	11/12 horas
Exposiciones trabajos	Aula de clase	2 horas
Trabajo del estudiante + tutorías		50 horas

* Normalmente como se necesita un portátil para cada dos personas no es necesario desplazarse al aula de informática para realizar las prácticas

Aproximadamente, 75 horas de trabajo (6 ECTs) deben ser suficientes para que el estudiante desarrolle las competencias descritas.

3. Sistemas de evaluación y porcentaje en la calificación final

3.1. Convocatoria ordinaria

La evaluación del seminario se basa en las prácticas/ejercicios que en las sesiones se vayan pidiendo como trabajo de clase y casa. El porcentaje que constituye de la nota final es del 70%.

El restante 30% se valora con la presentación del trabajo final.

El trabajo final . El estudiante tiene libertad para escoger un trabajo que considere relevante para él. Habitualmente las opciones de trabajos son las siguientes:

Trabajo de simulación. El estudiante quiere estudiar el sesgo, *coverage*y/o la potencia estadística que se producen en la estimación de parámetros (v.g., los pesos de un modelo de regresión, los pesos factoriales, índices de dificultad o de discriminación en un test, etc.). Para ello simula diferentes mecanismos de valores perdidos en una base de datos simulada (pérdida bajo MCAR, bajo MAR y/o bajo MNAR). Además, habitualmente compara distintos métodos de imputación (por ejemplo, uno moderno como ML y otro inadecuado como eliminación por lista). Todo ello se lleva a cabo en un modelo estadístico en el que tenga interés el estudiante. En las tutorías se enseña a cómo simular datos (v.g., con el software Mplus) y se acuerdan las condiciones del trabajo de simulación (v.g.,tamaño de la muestra,porcentaje de valores perdidos, tipo de pérdida de datos, métodos de imputación, etc.).

Trabajo empírico. El estudiante cuenta con una base de datos de su interés y decide estudiar los modelos estadísticos que ya conoce utilizando los métodos modernos de imputación que ha aprendido en el curso. En este tipo de trabajos, habitualmente, el interés se centra en comprobar las diferencias en los resultados de sus modelos estadísticos al aplicar tratamientos modernos de valores perdidoscon métodos clásicos. También el estudiante aprende a identificar las posibles causas de pérdida de datos que tiene en sus datos.

Trabajo de otra asignatura. En ocasiones el estudiante puede solicitar como trabajo final introducir los métodos modernos de imputación a otro trabajo anteriormente realizado en el máster en el que no se haya tenido en cuenta el problema de los valores perdidos. Este tipo de trabajos sirve para que el estudiante compruebe qué le aporta (en términos, por ejemplo, de consistencia, eficiencia o potencia) un uso adecuado de tratamiento de valores perdidos en relación a los resultados que ya ha obtenido.

3.1.1. Relación actividades de evaluación

Actividad de evaluación	**%**
Trabajo final	30
Evaluación continua (prácticas-ejercicios)	70

3.2. Convocatoria extraordinaria

Igual que en la convocatoria ordinaria.

3.2.1. Relación actividades de evaluación

Las primeras sesiones se utilizan para que el estudiante conozca los problemas generales que ocasionan los valores perdidos, los patrones de los mismos, los mecanismos (o causas) que los generan y los métodos clásicos de tratamiento de valores perdidos que existen. El estudiante aprenderá a simular algunas de estas causas con el ordenador (SPSS o R).

Las siguientes sesiones se emplean para que el estudiante conozca los métodos modernos de imputación: máxima verosimilitud e imputación múltiple. En estas dos sesiones el estudiante trabaja con el ordenador (Mplus o R) abordando modelos estadísticos ya conocidos (pruebas Ts, ANOVAs, A. Factorial, etc.) con presencia de valores perdidos. Estas sesiones

están orientadas a que el estudiante sepa tratar los valores perdidos dentro de esos modelos estadísticos desde la perspectiva moderna para que sepa dar respuesta a este tipo de escenarios tan comunes.

Se reserva alguna sesión en la que el estudiante aprende a simular con estructuras básicas de programación algunos escenarios más interesantes y complejos para que "vea" cómo los valores perdidos pueden afectar a los resultados estadísticos y como estrategia para poder valorar el impacto de la pérdida de datos (no analiticamente, sino por medio de simulaciones). En estas simulaciones se utiliza normalmente R.

4. Cronograma orientativo

Si las sesiones son de tres horas (a veces se organizan cada dos horas), tentativamente el estudiante puede saber que las primeras dos sesiones se dedican a la introducción de valores perdidos, el tipo de problema que representan para la inferencia estadística, los patrones y mecanismos implicados (MCAR, MAR y MNAR). Las siguientes dos sesiones están pensadas para trabajar con Imputación Múltiple (librerias mice, mdmb de R) y las siguientes dos con máxima verosimilitud completa (lavaan). Hay otra sesión reservada para trabajar con modelos de simulación y otra para exponer los trabajos de los estudiantes (30% de la evaluación).