UNIVERSIDAD AUTÓNOMA DE MADRID

30962 - SEMINARIO: ANÁLISIS DE VALORES PERDIDOS E IMPUTACIÓN DE RESPUESTAS

Información de la asignatura

Código - Nombre: 30962 - SEMINARIO: ANÁLISIS DE VALORES PERDIDOS E IMPUTACIÓN DE RESPUESTAS
Titulación: 385 - Máster en Metodología de las Ciencias del Comportamiento y de la Salud
Centro: 105 - Facultad de Psicología
Curso Académico: 2021/22

1. Detalles de la asignatura

1.1. Materia

-

1.2. Carácter

Optativa

1.3. Nivel

Máster (MECES 3)

1.4. Curso

1

1.5. Semestre

Segundo semestre

1.6. Número de créditos ECTS

2.0

1.7. Idioma

Español

1.8. Requisitos previos

Para este seminario no hay requisitos específicos imprescindibles , si bien unos conocimientos mínimos de inferencia estadística (v.g., conceptos como el contraste de hipótesis, la estimación de parámetros, nivel de significación, nivel crítico o error típico) facilitará n la comprensión de los contenidos que se exponen . Asimismo, saber desenvolverse con algún *software* estadístico como SPSS también facilitará el seguimiento del seminario.

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).

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

Ricardo Olmos Albacete

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

1.12. Competencias y resultados del aprendizaje

1.12.1. Competencias

Las competencias básicas y generales del seminario son las siguientes:

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 .
Desarrollar el razonamiento crítico y la capacidad para realizar análisis y síntesis de la información disponible.
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.
Planificar una investigación identificando problemas y necesidades, y ejecutar cada uno de sus pasos (diseño, medida, proceso de datos, análisis de datos, modelado, informe).
Obtener información de forma efectiva a partir de libros, revistas especializadas y otras fuentes.
Desarrollar y mantener actualizadas competencias, destrezas y conocimientos según los estándares propios de la profesión.

Todas estas competencias no se ven en “vacío”, sino que se estudian dentro de múltiples modelos estadísticos que el estudiante ve en otras asignaturas del máster: T para muestras relacionadas, T para muestras independientes, ANOVAS MR y ANOVAS CA, Análisis Factorial Exploratorio, Análisis Factorial Confirmatorio o Modelos de regresión lineal. Por lo tanto, se espera que el estudiante sepa cómo abordar un tratamiento adecuado de los valores perdidos en muchos de los modelos estadísticos que se ven en el máster. Además, también se espera que el estudiante sepa cómo abordar un estudio psicométrico cuando hay omisiones en una encuesta o en un test de rendimiento óptimo.

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.

CG4 - Planificar una investigación identificando problemas y necesidades, y ejecutar cada uno de sus pasos (diseño, medida, proceso de datos, análisis de datos, modelado, informe).

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.

1.12.2. Resultados de aprendizaje

Los resultados de aprendizaje serán valorados a través de la combinación de las metodologías y sistemas de evaluación específicos indicados para esta materia/asignatura

1.12.3. Objetivos de la asignatura

Los objetivos del seminario son varios:

En primer lugar, se espera que el estudiante tome conciencia de que tratar los valores perdidos con los métodos clásicos provoca, muy probablemente, que los resultados estadísticos tengan deficiencias (por ejemplo, en términos de consistencia, eficiencia y potencia estadística).
En segundo lugar, el estudiante debe saber que existen tres tipos de mecanismos qu e generan los valores perdidos: MCAR (Missing Completely At Random), MAR (Missing At Random) y MNAR (Missing Not At Random) y que algunos son más “benignos” que otros.
En tercer lugar, el estudiante debe saber identificar y conocer los métodos clásicos de imputación, algunos de los cuales representan las opciones por defecto en los paquetes estadísticos más difundidos: la eliminació n por lista (listwise deletion) o la eliminación por pares (pairwise deletion) . Otros métodos clásicos son la sustitución por la media (mean substitution), imputación hot deck , imputación por la media condicional (conditional mean imputation) o imputación por regresión estocástica.
En cuarto lugar, el estudiante debe conocer y saber aplicar los métodos modernos de imputación: Máxima Verosimilitud (FML – Full Maximum L ikelihood) e Imputación Múltiple (MI – Multiple Imputation). Y la aplicación se hará en mútiples modelos estadísticos que el estudiante conoce de otras asignaturas del máster para que sepa aplicar lo que ha aprendido en relación con los valores perdidos.

1.13. Contenidos del programa

Los contenidos del seminario están organizados en cuatro temas:

TEMA 1: INTRODUCCIÓN A LOS VALORES PERDIDOS. Problema omnipresente dentro del análisis de datos. Cómo se ha abordado el problema. Patrones diferentes de valores perdidos. Tres mecanismos que generan la pérdida de datos: MCAR (Missing Completely At Random), MAR (Missing At Random) y MNAR (Missing Not At Random).

TEMA 2: MÉTODOS CLÁSICOS DE IMPUTACIÓN DE VALORES PERDIDOS. Eliminación por lista. Eliminación por pares. Imputación por la media. Métodos Hot Deck . Imputación por regresión. Imputación por regresión estocástica.

TEMA 3: MÉTODOS MODERNOS DE IMPUTACIÓN DE VALORES PERDIDOS (I). Máxima verosimilitud (ML). Introducción a estimación ML. Estimación ML con valores perdidos. Utilización de la librería lavaan de R.

TEMA 4: MÉTODOS MODERNOS DE IMPUTACIÓN DE VALORES PERDIDOS (II). Imputación Múltiple (MI). Introducción a la MI. Estimación MI con valores perdidos (modelización conjunta y ecuaciones encadenadas). Utilización de la librería mice en R.

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, l as referencias que pueden ayudar a completar la formación dada en 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: 10.1177/014662160730769

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 (opcionalmente se puede ver con SPSS, pero se prioriza utilizar R en todas las sesiones actualmente). Además, el estudiante aprende a simular con el programa R los mecanismos de valores perdidos explicados durante la teoría, 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 lavaan y mice, respectivamente) . 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:
Los trabajos serán de carácter empírico y/ o de simulación (ver apartado 3. Sistemas de evaluación y porcentaje en la calificación final) . Se proporciona en Moodle una guía a modo orientativo con la rúbrica de los puntos que tiene el trabajo para facilitarle la tarea al estudiante. A lo largo de las sesiones el estudiante puede decidir sobre qué tema quiere hacer el trabajo.

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	6/7 horas de exposición
Clases teórico/prácticas Prácticas	*Aula de clase	7/8 horas
Trabajo del estudiante + tutorías		25 horas

* Normalmente como se necesita un portátil para cada dos personas no es necesario desplazarse al aula de informática para realizar las prácticas
Aproximadamente, 40 horas de trabajo 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 cinco sesiones se vayan pidiendo como trabajo de clase. El porcentaje que constituye de la nota final es del 7 0%.

El restante 3 0% se valora con la presentación del trabajo final. Como se describía en el apartado 2.2, el estudiante contará con un documento para orientarle a hacer el trabajo optativo.
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 o con el mismo programa R) 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 perdidos con 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

-

4. Cronograma orientativo

Las dos 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 (con el programa R).

La tercera y cuarta sesión (y normalmente un poco de la segunda) 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 (R), primero, aprendiendo el manejo de las librerías que se utilizan para gestionar desde la perspectiva moderna los valores perdidos (mice para imputación múltiple y lavaan para máxima verosimilitud) y, posteriormente, abordando con estas librerías modelos estadísticos ya conocidos (pruebas Ts, ANOVAs, A. Factorial, etc.), pero esta vez bajo la presencia y, por lo tanto, ante el problema, de valores perdidos en sus variables. Estas sesiones están orientadas a que el estudiante sepa tratar los valores perdidos dentro de esos modelos estadísticos desde la perspectiva moderna con el objetivo de que sepa dar respuesta a este tipo de escenarios tan comunes.

La quinta sesión se destina a terminar de dar, si es que no ha dado tiempo, los últimos modelos estadísticos con lavaan y mice (las dos librerías de R previamente comentadas). El resto de esta sesión está pensada para dar una alfabetización básica para simular algunos modelos estadísticos y generar en ellos pérdida de datos. Eso le permite al estudiante poder estudiar cómo se recuperan los parámetros simulados bajo las distintas perspectivas que el estudiante ha visto durante el seminario.