Teaching Guide				
Identifying Data				2023/24
Subject (*)	Statistical Analysis of Complex Data		Code	614G02031
Study programme	Grao en Ciencia e Enxeñaría de Datos			
Descriptors				
Cycle	Period	Year	Type	Credits
Graduate	1st four-month period	Fourth	Optional	6
Language	SpanishGalician			
Teaching method	Face-to-face			
Prerequisites				
Department	Matemáticas			
Coordinador	Oviedo de la Fuente, Manuel	E-mail	manuel.oviedo@udc.es	
Lecturers	López Cheda, Ana Oviedo de la Fuente, Manuel	E-mail	ana.lopez.cheda@udc.es manuel.oviedo@udc.es	
Web	https://dm.udc.es/modes/			
General description	In this course, the main mechanisms which lead to missing data will be studied. Specifically, this course provides a first contact with the main statistical techniques to analyse problems with missing data, functional data, censored data or biased data. The limitations of the different approaches will be examined. The presented methodologies will be applied to simulated or real datasets. The conclusions will be interpreted accordingly.			

Study programme competences / results	
Code	Study programme competences / results
A3	CE3 - Capacidade para a análise de datos e a comprensión, modelado e resolución de problemas en contextos de aleatoriedade.
A17	CE17 - Capacidade para a construción, validación e aplicación dun modelo estocástico dun sistema real a partir dos datos observados e a análise crítica dos resultados obtidos.
A20	CE20 - Coñecemento das ferramentas informáticas no campo da análise dos datos e modelización estatística, e capacidade para seleccionar as máis adecuadas para a resolución de problemas.
B2	CB2 - Que os estudantes saiban aplicar os seus coñecementos ao seu traballo ou vocación dunha forma profesional e posúan as competencias que adoitan demostrarse por medio da elaboración e defensa de argumentos e a resolución de problemas dentro da súa área de estudo
B3	CB3 - Que os estudantes teñan a capacidade de reunir e interpretar datos relevantes (normalmente dentro da súa área de estudo) para emitir xuízos que inclúan unha reflexión sobre temas relevantes de índole social, científica ou ética
B4	CB4 - Que os estudantes poidan transmitir información, ideas, problemas e solucións a un público tanto especializado como non especializado
B6	CG1 - Ser capaz de buscar e seleccionar a información útil necesaria para resolver problemas complexos, manexando con soltura as fontes bibliográficas do campo.
B7	CG2 - Elaborar adecuadamente e con certa orixinalidade composicións escritas ou argumentos motivados, redactar plans, proxectos de traballo, artigos científicos e formular hipóteses razoables.
B8	CG3 - Ser capaz de manter e estender formulacións teóricas fundadas para permitir a introdución e explotación de tecnoloxías novas e avanzadas no campo.
B9	CG4 - Capacidade para abordar con éxito todas as etapas dun proxecto de datos: exploración previa dos datos, preprocesado, análise, visualización e comunicación de resultados.
B10	CG5 - Ser capaz de traballar en equipo, especialmente de carácter multidisciplinar, e ser hábiles na xestión do tempo, persoas e toma de decisións.
C1	CT1 - Utilizar as ferramentas básicas das tecnoloxías da información e as comunicacións (TIC) necesarias para o exercicio da súa profesión e para a aprendizaxe ao longo da súa vida.
C4	CT4 - Valorar a importancia que ten a investigación, a innovación e o desenvolvemento tecnolóxico no avance socioeconómico e cultural da sociedade.

Learning outcomes

Learning outcomes	Study programme competences / results		
Know and understand the basics of missing data	A3 A20	B6	C1 C4
To know the main techniques to analyse problems with missing data	A3 A17 A20	B3 B4 B9	C1
To know the main techniques to analyse functional data	A3 A17 A20	B3 B4 B9	C1
To know the main techniques to analyse censored data	A3 A17 A20	B3 B4 B9	C1
To know the main techniques to analyse problems with biased data	A3 A17 A20	B3 B4 B9	C1
To be able to apply different techniques for missing data, functional data, censored data and biased data to a real or a simulated dataset	A20	B2 B3 B4 B9	C1
To be able to interpret the results and to know the limitations of the different methods	A3	B6 B7 B8 B10	C1 C4

Contents	
Topic	Sub-topic
Introduction to functional data	Motivation and examples Functional data registration and smoothing Metrics and semimetrics for functional data Representing functional data: basis expansions
Functional data analysis	Estimation of mean and covariance operator On the concept of depth for functional data: functional anomaly detection Functional principal component analysis Functional linear models
Introduction to missing data	Challenges and problems with missing data Missing data mechanisms: missing at random (MAR) and missing completely at random (MCAR) The consequences of discarding missing data
Imputation methods	Mean imputation Single imputation methods Maximum likelihood multiple imputation under MAR Expectation{Maximization (EM) algorithm Multiple imputation methods under MAR
Biased data	Selection bias: length, time and size The consequences of disregarding bias Mean and variance estimation for biased data Likelihood principle for biased data Situations with unspecified bias function

Censored data	Missing data and censoring The consequences of discarding censored data Parametric estimation for censored data Nonparametric estimation for censored data: the Kaplan-Meier estimator Cox model: conditional survival

Planning				
Methodologies / tests	Competencies / Results	Teaching hours (in-person & virtual)	Student?s personal work hours	Total hours
Oral presentation	A3 B2 B3 B4 C4	21	31.5	52.5
ICT practicals	A17 A20 A3 B2 B3 B4 B6 B7 B8 B9 B10 C1	7	24.5	31.5
Supervised projects	A17 A20 A3 B2 B3 B4 B6 B7 B9 B10 C1	3.5	15.75	19.25
Problem solving	A17 B2 B7 B8 B10	7	28	35
Mixed objective/subjective test	A20 A3 B2 B3 B4 B8 C1	1.5	3	4.5
Mixed objective/subjective test	A20 A3 B2 B3 B4 B8 C1	1.5	3.75	5.25
Personalized attention		2	0	2
(*)The information in the planning table is for guidance only and does not take into account the heterogeneity of the students.				

Methodologies	
Methodologies	Description
Oral presentation	Presentation using the computer
ICT practicals	Statistical data analysis using R
Supervised projects	Statistical analyses of some databases applying the studied methodologies
Problem solving	Deciding statistical tools and strategies for problem solving with missing data, functional data, censored data or biased data
Mixed objective/subjective test	Test related to concepts and/or practical exercises with R (in the middle of the semester)
Mixed objective/subjective test	Test related to concepts and/or practical exercises with R (official exam in January)

Personalized attention	
Methodologies	Description
Problem solving ICT practicals Supervised projects	Attendance and participation in the theoretical lessons Practical cases of study using R Problem solving using R Test related to theoretical and/or practical questions

Assessment			
Methodologies	Competencies / Results	Description	**Qualification**
Mixed objective/subjective test	A20 A3 B2 B3 B4 B8 C1	Test related to concepts and/or practical exercises with R (official exam in January)	40

Supervised projects	A17 A20 A3 B2 B3 B4 B6 B7 B9 B10 C1	Practical and research projects	40
Mixed objective/subjective test	A20 A3 B2 B3 B4 B8 C1	Test related to concepts and/or practical exercises with R (in the middle of the semester)	20

Assessment comments
The assessment scoring will be the following: ? Two or three practical and research works: 4 points. ? Concept/practical test related to Topics 1-2: 2 points. It will take place in the middle of the quadrimester. Students will avoid this test in the official exam if they obtain, at least, a score of 3.5 out of 10 in this first exam, unless they want to get a higher score (in this case, the score will be the one obtained in the official exam). Moreover, if students do not attend the first exam or if they obtain a score lower than 3.5 out of 10, then they will be evaluated (the day of the official exam) of this. ? Concept/practical test related to Topics 3 - 6: 4 points. It will take place the date of the official exam. In order to pass the subject it is necessary to obtain a score of at least 3.5 out of 10 in this test. To pass the subject it is necessary to obtain a score of at least 5 out of 10 overall.On the second opportunity (July), students must attend the exams in which they obtained a lower score than 3.5 out of 10 in January tests. If they want to get a higher score, then the final score will be the one obtained in July.Only students that didn't take any test will be qualified as NON ATTENDANT in the first opportunity (January-February). In July (2nd opportunity) only students that didn't take the final exam will be qualified as NON ATTENDANT. If a student wants to take a test in a specific official language (Spanish or Galician), he/she must inform the professor at least 1 week in advance. Fraud in tests or evaluation activities will directly imply the failure grade "0" in the subject in the corresponding call, thus invalidating any grade obtained in all the evaluation activities for the extraordinary call.

Sources of information	
Basic	- Little R. J., Rubin D. B. (2019). Statistical analysis with missing data (Vol. 793). John Wiley & Sons - Ramsay J. O., Silverman B. W. (2005). Functional Data Analysis. 2nd Edition. Springer - Ferraty F., Vieu P. (2006). Nonparametric functional data analysis : theory and practice. Springer - Hosmer D. W., Lemeshow S., May S. (2008). Applied survival analysis: regression modeling of time-to-event data. Wiley-Interscience - Lee E. T., Wang J. W. (2013). Statistical Methods for Survival Data Analysis. 4th Edition. Wiley - Qin J. (2017). Biased sampling, over-identified parameter problems and beyond (Vol. 5). Springer - Cox D. R. (2005). Some sampling problems in technology. . Selected Statistical Papers of Sir David Cox
Complementary	- Van Buuren, S. (2018). Flexible imputation of missing data. CRC Press - Febrero-Bande M, Oviedo de la Fuente M. (2012). Statistical Computing in Functional Data Analysis: The R Package fda.usc. Journal of Statistical Software, 51(4), 1?28 - Therneau T. M., Grambsch P. M. (2000). Modeling Survival Data: Extending the Cox Model. Springer - Therneau T. (2021). A Package for Survival Analysis in R. CRAN

Recommendations
Subjects that it is recommended to have taken before
Statistical Analysis of Dependent Data/614G02022 Regression Models/614G02012 Statistical Modeling of High Dimensional Data/614G02013 Statistical Inference/614G02007 Probability and Basic Statistics/614G02003
Subjects that are recommended to be taken simultaneously
Spatiotemporal Data Representation and Management/614G02035 Simulation and Resampling Techniques/614G02036
Subjects that continue the syllabus

Omics Data Management and Modeling /614G02042
Other comments
As stated in the different applicable regulations for university teaching, the gender perspective must be incorporated in this course (non-sexist language will be used, bibliography of authors of both genders will be used, intervention in class of both male and female students will be encouraged, etc.)Work will be done to identify and modify prejudices and sexist attitudes. The environment will be influenced to modify these prejudices and attitudes, to promote values of respect and equality.Situations of discrimination based on gender must be detected. Actions and measures to correct them will be proposed.