Honoris Causa de Ingrid Van Keilegom

Laudatio pronunciada polo Profesor Ricardo Cao Abad

Excelentísimo Reitor Magnífico da Universidade da Coruña, autoridades civís, militares e académicas, profesoras, profesores, amigas e amigos, *beste Vlaamse collega's en vrienden, beste Ingrid*:

Hoxe comparto con todos os que, coma min, tiveron a sorte de traballar coa Profesora Ingrid Van Keilegom, o orgullo que supón expoñer os seus logros científicos, parte deles en colaboración cunha chea de investigadoras e investigadores galegos e doutras CCAA do Estado Español, moitos deles aquí presentes hoxe.

En primeiro lugar, gustaríame amosar o meu agradecemento ao Departamento de Matemáticas da Universidade da Coruña, por formular a proposta desta distinción á Profesora Van Keilegom, tras a iniciativa do grupo de investigación de Modelización, Optimización e Inferencia Estatística (MODES) da Universidade da Coruña, que foi secundada polo Departamento de Economía da nosa universidade. Agradecemento que me gustaría extender aos departamentos de Estatística, Análise Matemática e Optimización da Universidade de Santiago de Compostela, de Estatística e Investigación Operativa da Universidade de Vigo, de Estadística de la Universidad Carlos III de Madrid, de Estadística e Investigación Operativa de la Universidad de Granada, de Estadística e Investigación Operativa y Didáctica de la Matemática de la Universidad de Oviedo, de Estadística e Investigación Operativa de la Universidad de Valladolid, de Métodos Estadísticos de la Universidad de Zaragoza, así como ao Centro de Investigación en Tecnoloxías da Información e as Comunicacións (CITIC) da Universidade da Coruña e á Sociedade Galega para a Promoción da Estatística e da Investigación de Operacións (SGAPEIO), que deron o seu apoio á proposta.

Ingrid naceu en Wilrijk (Antwerpen) [Amberes], Bélxica, o día de Noiteboa de 1971. Rematou os seus estudos da Licenciatura de Matemáticas na Universitaire Instelling Antwerpen no ano 1993. Eu tiven a sorte de coñecela en Hasselt (Bélxica) en 1996, na miña primeira visita a esa universidade (entón chamada Limburgs Universitair Centrum) para colaborar cos profesores Noël Veraverbeke e Paul Janssen, hoxe aquí presentes. Neses momentos Ingrid tiña moi avanzada a súa Tese de Doutoramento, dirixida polo Profesor Veraverbeke, sobre "Nonparametric estimation of the conditional distribution in regression with censored data", que defendeu brillantemente no ano 1998. Tras ocupar prazas como profesora axudante doutora na Pennsylvania State University (EEUU) na Eindhoven University of Technology (Países Baixos) e na Université Catholique-de-Louvain (Louvain-la-Neuve, Bélxica), Ingrid Van Keilegom obtén unha praza de Profesora Titular (Associate Professor) en 2005 e de Catedrática (Full Professor) en 2008, nesa última universidade. No ano 2016 a profesora Van Keilegom trasladouse como Full Professor á KU Leuven -antes chamada Katholieke Universiteit Leuven- (Leuven, Bélxica), aínda que mantén unha praza simultánea de Full Professor a tempo parcial na Université Catholique de Louvain.

Ao longo da súa carreira académica foi profesora visitante en mais de vinte ocasións en diversas universidades e centros de investigación de Australia, Bélxica, España, Francia e Países Baixos. Ademais, a profesora Van Keilegom desenvolveu mais de 130 visitas curtas de investigación en universidades e centros de investigación de todo o mundo, incluíndo Alemaña, Australia, China, Chipre, Dinamarca, EEUU, España, Francia, Noruega e Sudáfrica. Gustaríame salientar que 29 das súas visitas de investigación tiveron lugar nas tres universidades galegas.

As liñas de investigación de Ingrid Van Keilegom comprenden a análise de supervivencia (datos censurados, modelos de curación, censura con dependencia), a inferencia estatística baixo erros de medida, a regresión cuantílica, a regresión non paramétrica e semiparamétrica, a estatística matemática e as variables instrumentais, a endoxeneidade e os modelos fronteira en Econometría. Os seus traballos conteñen avances fundamentais cunha grande repercusión na Inferencia Estatística: nas extensións da verosimilitude empírica (empirical likelihood) a campos como a estimación non paramétrica e semiparamétrica de curvas, os datos censurados, os problemas de igualdade entre dúas poboacións e o caso de funcións criterio non suaves; o método bootstrap e a súa aplicación en modelos de previsión de mortalidade, na construción de bandas de confianza para a regresión e as súas derivadas, e na análise de supervivencia; así como o estudio de modelos estatísticos non paramétricos e semiparamétricos para a análise de datos funcionais, os contrastes de especificación ou as curvas ROC, entre moitas outras. Os seus resultados metodolóxicos recentes nos modelos de curación están a ter unha importante repercusión na Estatística moderna e na súa aplicación a outras ciencias, como ten sido recoñecido nun dos premios acadados por ela recentemente.

Nestes campos citados, Ingrid publicou mais de 160 artigos de investigación en revistas de prestixio internacional, destacando os seus mais de trinta artigos en The Annals of Statistics, Biometrika, Journal of the Royal Statistical Society e Journal of the American Statistical Association, catro das revistas de estatística matemática mais recoñecidas do mundo; así como as publicacións en revistas de primeira liña en campos afíns, entre as que destaca a revista Econometrica. Destes más de 160 artigos, 24 deles son con coautores dalgunha das tres universidades galegas. Un deles, conxunto con tres coautores da UDC, foi recentemente galardoado co premio como mellor contribución metodolóxica en Estatística pola

Sociedad de Estadística e Investigación Operativa (SEIO) de España, no ano 2021.

Un dos aspectos mais salientables da profesora Van Keilegom é o de o seu liderado en proxectos e traballos de investigación. Dentro dos proxectos internacionais dirixidos por ela son de destacar os seus dous proxectos financiados polo prestixioso European Research Council. Tamén dirixiu 13 teses de doutoramento (6 mais baixo dirección actualmente). Así mesmo, participou en 49 tribunais de teses de doutoramento, 8 delas no programa de doutoramento en Estatística e Investigación de Operacións coorganizado pola Universidades da Coruña, Santiago de Compostela e Vigo.

Unha boa mostra da actividade de Ingrid Van Keilegom relacionada coas universidades galegas en materia de formación doutoral queda reflectida na fotografía que poden ver na pantalla. Corresponde aos instantes posteriores ás defensas das Teses de Doutoramento de Juan Carlos Pardo, hoxe profesor na Universidade de Vigo, dirixida por Ingrid Van Keilegom e Wenceslao González Manteiga, profesor da Universidade de Santiago, e de Luís Filipe Meira Machado, hoxe profesor na Universidade do Minho, dirixida por Jacobo de Uña Álvarez, profesor da Universidade de Vigo, e Carmen Cadarso, profesora da Universidade de Santiago de Compostela. Todos eles están hoxe presentes neste acto de investidura (Luís por videoconferencia), agás a nosa colega e entrañable amiga Carmen Cadarso, que faleceu prematuramente o día 3 deste mes. Carmen, tamén coautora dunha publicación con Ingrid, foi un xenuino exemplo da colaboración entre as comunidades estatísticas de Galicia e Flandes. Botarémola moito de menos, mais sempre a teremos no noso corazón.

Como queda patente no exposto anteriormente, as relacións científicas e académicas da profesora Ingrid Van Keilegom coa Universidade da Coruña e, en xeral, coas tres universidades galegas, é moi intensa. Como pequena mostra adicional cabe citar que ela acolleu 4 estadías de estudantes de doutoramento da UDC e 3 de investigadores senior da UDC na Université Catholique de Louvain e na KU Leuven ao longo dos últimos 15 anos.

A profesora Ingrid Van Keilegom ten desenvolvido unha chea de actividade editorial en revistas do máximo prestixio internacional. Foi editora conxunta do Journal of the Royal Statistical Society - Series B e editora asociada de The Annals of Statistics, Biometrika, Electronic Journal of Statistics, Annual Review of Statistics and its Application, Econometrics and Statistics, International Journal of Biostatistics, Scandinavian Journal of Statistics, Annals of the Institute of Statistical Mathematics, Statistics and Probability Letters e no antes mencionado Journal of the Royal Statistical Society - Series B. En moitas delas está a desempeñar actividade editorial na actualidade.

As súas labores en organización de actividades de investigación e sociedades científicas abranguen a presidencia de comités científicos dunha decena de congresos internacionais (por so citar un exemplo cómpre mencionar o International Symposium on Nonparatric Statistics que se celebrará en Chipre a vindeira semana, despois de ter que ser adiado desde o 2020 por mor da pandemia) e de workshops internacionais; a súa participación en comités da American Statistical Association, o Institute of Mathematical Statistics, a Bernoulli Society, a International Biometric Society e a Belgian Statistical Society e as de avaliación científica para o European Research Council e o Research Council of Flanders, entre moitas outras institucións.

Entre as distincións da profesora Van Keilegom cómpre sinalar os seus nomeamentos como *Fellow* da American Statistical Association, *Fellow* do Institute of Mathematical Statistics, *Elected member* do International Statistical Institute e a Cátedra da Janssen Research Foundation sobre Análise de Supervivencia.

Podería alongarme moito falando hoxe das achegas da Profesora Ingrid Van Keilegom á Estatística. Mais non tería tempo dabondo no resto do día para simplemente describir sucintamente a maioría do seus profundos resultados de investigación na inferencia estatística moderna. Se non estou mal informado, ela falará hoxe dun tema apaixonante e de rabiosa actualidade co auxe dos Big Data: a análise estatística dos datos imperfectos. Porén, gustaríame dedicar uns minutos a falar da súa gran calidade humana. Mais isto fareino en inglés.

A colleague (and friend of mine) uses to say that "human beings don't have defects and virtues; we just have properties". As you can imagine, my friend is a mathematician. I do not agree 100 % with his statement, but I subscribe the message it reflects. So I would like to briefly mention what, in my humble opinion, are the most outstanding "properties" of Ingrid as a human being. First, Ingrid has a brilliant mind. She is very curious, and she likes to look at the problems from different viewpoints before deciding the one to focus on. Second, Ingrid is an extremely hard worker. She is an example for all of us, for young researchers and for her PhD students. And third, and more important, Ingrid is a great person, friendly, open-minded, generous and always willing to cooperate with others. Please, Ingrid, keep being as you are! Finally, I would like to say a few words in Flemish, Ingrid's native language.

Beste Ingrid, de gemeenschap van Galicische statistici heeft veel geluk. Langs het pad van ons onderzoek hebben we het geluk gehad om met u samen te werken en met vele andere Belgische collega's, met name Vlamingen, zoals Paul en Noël die vandaag bij ons zijn. Hartelijk dank voor uw vrijgevigheid, talent en goede werk. Bedankt dat je bent zoals je bent!

Ten sido para min unha honra poder glosar hoxe algúns dos méritos mais salientables da Profesora Ingrid Van Keilegom que a fan merecedora da distinción de Doutora *Honoris Causa* pola nosa benquerida Universidade da Coruña.

Con isto remato o meu discurso. Moitas grazas pola súa atención.

Discurso pronunciado pola Profesora Ingrid Van Keilegom

Reitor, autoridades aquí presentes hoxe, señoras e señores,

En primeiro lugar quero dicir o honrada que me sinto por recibir esta distinción. Tamén quero manifestar un gran aprecio a todos aqueles que estiveron involucrados no nomeamento para este grao de Doutora *Honoris Causa*.

As miñas colaboracións coa Universidade da Coruña e, en xeral, coas tres universidades galegas, incluídas tamén as de Santiago de Compotela e Vigo, remóntanse a hai máis de 25 anos. Daquela eu aínda era estudante de doutoramento na Universidade de Hasselt en Bélxica, e Ricardo Cao e Wenceslao González Manteiga visitaban regularmente ao meu director de tese Noël Veraverbeke e Paul Janssen, ambos presentes hoxe aquí. Despois do doutoramento comecei a traballar tanto con Ricardo como con Wenceslao e despois con moitos outros investigadores galegos, e dende entón visitei A Coruña, Santiago de Compostela e Vigo moitas veces, para traballar en proxectos conxuntos, para asistir a congresos ou para formar parte de tribunais de teses. Vindo a A Coruña e a Galicia síntome case como volvendo a casa, aínda que debo recoñecer que nunca conseguín acostumarme aos xantares e ceas tardíos. A familia galega é incriblemente xenerosa, hospitalaria e simpática, e sempre tiven a sensación de formar parte dela en canto cheguei aquí. Compartimos moitos momentos moi bonitos pero tamén tristes xuntos, o que fixo que os nosos lazos fosen aínda máis fortes. Estes fortes lazos coa familia galega non só son comigo. De feito, moitos estatísticos belgas teñen fortes conexións coa familia galega, tanto a nivel profesional como persoal. Polo tanto, en nome de toda a comunidade estatística flamenca, quero trasladar o noso pésame á estatística galega polo falecemento da nosa querida compañeira e amiga Carmen Cadarso Suárez o pasado 3 de xuño. A comunidade flamenca, e en particular os grupos de estatística de Lovaina e Hasselt, tiñan moi bos contactos con ela. Botarémola moito de menos, pero queda no noso corazón para sempre.

Aquí podedes ver unha serie de imaxes, tomadas en diversas ocasións no pasado na Coruña, Santiago de Compostela e Vigo. Foron tomadas durante conferencias, visitas de traballo ou algunha das moitas actividades de fin de semana nas que Ricardo e outros compañeiros mostráronme distintos lugares da fermosa Galicia. Son testemuñas da especial relación que teño con Galicia e coa Coruña en particular, construída nos últimos 25 anos.

Comezarei esta presentación falando do crecente papel que a estatística está a xogar na nosa vida diaria, e cales son ao meu ver as posibles explicacións disto. Despois, presentareivos algunhas das investigacións que estiven facendo durante os últimos anos. Escollín tres exemplos de áreas nas que traballei, concretamente, os erros de medición, a censura dependente e a inferencia causal. Os tres caracterízanse polo feito de que os datos son imperfectos dun xeito ou doutro, e que se pode corrixir esta imperfección se se ten a man un conxunto de datos suficientemente grande. Rematarei con algunhas conclusións e mensaxes para levar a casa.

Sendo unha persoa dedicada á investigación en estatística dende hai case 30 anos, é moi interesante ver a evolución que está a atravesar esta área. Está claro que nos últimos 10 anos aproximadamente, o papel que xoga a estatística na nosa vida diaria é cada vez máis relevante, e que tamén está reclamando un rol máis destacado nos debates e discusións sociais sobre problemas importantes como o cambio climático, climas extremos, pandemias, cuestións económicas sobre benestar ou desemprego e, por último, pero non menos importante, tamén xoga un papel importante no auxe da intelixencia artificial, a aprendizaxe automática, a internet das cousas e áreas relacionadas. Esta evolución tamén se aprecia polo que as persoas influentes din sobre a estatística nas entrevistas, ou por informes de importantes organizacións nos que se fala do papel da estatística na sociedade. Cítanse aquí dous exemplos: o empresario Mark Cuban mencionou nunha entrevista en 2017 que "os datos son o novo ouro" e unha nota do Parlamento Europeo en 2020 titulábase "Son os datos o novo petróleo?"

Outra forma de ver este fenómeno é analizando a evolución do número de estudantes que deciden cursar estatística e ciencia de datos a nivel de mestrado e a evolución do posto de traballo de estatístico/estatística respecto doutros postos de traballo. Na miña universidade, o número de estudantes de mestrado en estatística e ciencia de datos é actualmente de máis de 400 (para os dous anos dos mestrados combinados), mentres que hai 10-12 anos era só uns 50, e as condicións de acceso eran menos severas nese momento do que son agora. Escoito que as cifras tamén están aumentando noutras universidades. Observando a evolución do número de postos de traballo, esta gráfica amosa que a oferta laboral dos estatísticos está a crecer moito máis rápido que a media e, ao mesmo tempo, o traballo de estatístico ou estatística é valorado como un dos mellores segundo criterios como contidos laborais interesantes, salario, taxa de paro, presenza de estrés e posibilidade de medrar.

Entón, unha pregunta natural que un pode facerse é cales son as razóns desta evolución? Na miña opinión hai varios aspectos que se reforzan mutuamente. En primeiro lugar, existe a crecente conciencia de que ter datos de boa calidade é importante en todas as ramas da sociedade, e as decisións baséanse cada vez máis en datos sólidos en lugar de intuición ou xuízos cualitativos. Un exemplo é a evolución das competicións deportivas profesionais, onde as estratexias se determinan cada vez máis a partir das análises estatísticas dos puntos fortes e débiles do competidor. Outra razón do éxito da estatística é o papel destacado que desempeñou e segue xogando en importantes problemas sociais como a crise da COVID-19 e as consecuencias do quecemento global, como os cambios climáticos e as condicións meteorolóxicas extremas. Os modelos, estimacións e predicións dos estatísticos e bio-

estatísticos axudan considerablemente aos gobernos na definición de políticas adecuadas para controlar a evolución destas crises. Unha terceira razón é que áreas que están estreitamente relacionadas coa estatística, como a aprendizaxe automática, a intelixencia artificial, a bioinformática ou a epidemioloxía, están a utilizar moitos coñecementos estatísticos, polo que existe un efecto de reforzo mutuo entre a estatística e estas disciplinas. Finalmente, non se pode subestimar a importancia que teñen os desenvolvementos recentes na investigación estatística na evolución exitosa deste campo. No que resta desta presentación explicarei algunhas das achegas nas que estiveron traballando o meu grupo de investigadores e investigadoras predoutorais e posdoutorais xunto con colaboradores externos, así como os resultados que conseguimos.

En estatística pódese considerar normalmente tres pedras angulares ou bloques de construción que xogan un papel crucial en calquera problema estatístico: temos que ter datos adecuados, necesitamos ter un modelo axeitado que represente o mundo real do que proveñen estes datos, e precisamos ter procedementos ou métodos potentes para estimar, predicir ou contrastar a cantidade ou a hipótese que nos interesa. Nesta presentación, centrareime nos datos e, en particular, explicarei o que se pode facer cando os datos son imperfectos. Responderanse preguntas sobre como podemos modelar a natureza imperfecta dos datos, como podemos identificar as cantidades de interese ou como podemos corrixir estes datos imperfectos ao construír procedementos de estimación ou contraste. Esta será a primeira liña vermella da miña presentación. A segunda liña vermella preocúpase polo feito de que a cantidade de datos é cada vez máis grande na actualidade, debido ao aumento das capacidades de almacenamento, á concienciación da importancia de ter datos, á crecente cultura de código aberto de compartir datos, etc. Isto é certo tanto para o número de observacións (o tamaño da mostra) como para o número de características que se almacenan para cada observación (a dimensión dos datos). Por ese motivo, adoita ocorrer que parámetros difíciles de estimar con mostras pequenas, se fan identificables ou estimables en mostras máis grandes, e isto é algo que queremos aproveitar á hora de desenvolver novas metodoloxías.

A análise estatística dos datos imperfectos ilustrarase centrándose en tres tipos deles, concretamente, os datos con erros de medición, os datos censurados e os datos recollidos para responder a preguntas sobre relacións causais, que requiren unha forma de pensar bastante diferente á dos datos nos que a causalidade non é a cuestión de interese. Hai moitos outros tipos de datos imperfectos dos que podería falar, pero non terei tempo para facelo. Estes inclúen, por exemplo, o caso de datos faltantes, o caso no que os datos son substituídos por datos subrogados que son máis baratos, menos invasivos ou máis rápidos de recoller, e o caso no que se sabe que os datos só se atopan nun intervalo determinado, como os datos de enquisas que están representados en categorías.

Daquela, imos comezar cos datos que conteñen erros de medición, que ás veces tamén se denominan datos borrosos ou datos ruidosos. Os erros de medición poden ser causados por un dispositivo de medición inexacto (como unha báscula ou un termómetro), pero tamén por rexistros imprecisos (por exemplo, se se lle pregunta a unha persoa cantos cafés bebeu o día anterior, probablemente non o lembrará exactamente), ou por variación temporal (por exemplo, a presión arterial pode flutuar arredor dun determinado valor nunha xanela de tempo determinada). Nalgunhas áreas os erros de medición poden considerarse moi pequenos e insignificantes, pero noutras áreas poden ter un gran impacto na análise se non se teñen en conta. Por exemplo, en estudos de enquisas ou en estudos de nutrición (como o exemplo sobre a inxesta de café), os erros de medición moitas veces non se poden ignorar. De non ser así, poden provocar un nesgo grave, unha perda de potencia e o enmascaramento de determinadas características presentes nos datos.

Isto pódese ver nesta figura, onde o panel superior representa datos simulados que non conteñen ningún erro de medición, e o panel inferior representa os datos nos que se engaden erros de medición no eixo horizontal. En lugar de ver unha relación sinusoidal clara, vemos unha imaxe moi borrosa, na que a parte superior e a parte inferior da función do seno apenas son visibles. Se queremos estimar esta relación sinusoidal, necesitamos impoñer un modelo para os erros de medida. Moitas veces asúmese que os erros de medida son aditivos e teñen unha distribución normal con media cero e certa varianza sigma ao cadrado.

Na literatura adóitase asumir que esta varianza se coñece cando non hai información adicional dispoñible, como datos de validación, medicións repetidas ou instrumentos. Demostramos, xunto con algúns estudantes de doutoramento, que a suposición de varianza coñecida non é necesaria sempre que a verdadeira variable non observada teña un soporte finito. Este é un gran paso adiante, xa que a varianza do erro adoita ser descoñecida na práctica. Especificamente, demostramos que se pode identificar a varianza, propuxemos un procedemento de estimación e os resultados de simulación mostraron que dito procedemento se comporta de forma axeitada en mostras grandes. Polo contrario, en mostras pequenas, pode que non haxa información dabondo nos datos para capturar a varianza do erro de forma adecuada.

Como ilustración mostramos aquí unha situación na que a estimación da función de regresión está moi nesgada cando se ignoran os erros de medición. A verdadeira recta de regresión é 1+X pero, debido ao erro de medición, a estimación naif está lonxe da verdade. O noso procedemento consiste en estimar primeiro a varianza do erro de medida, e despois utilizar ese estimador nun procedemento de estimación corrixido que teña en conta dito erro de medida.

Chegamos agora ao segundo tema, que trata dos datos

censurados e máis precisamente da censura dependente. A censura ocorre con moita frecuencia nos estudos nos que interesa o tempo que transcorre ata que acontece un determinado suceso. Este evento pode ser a morte dun paciente que está tratado por unha determinada enfermidade, pero tamén pode ser o tempo ata que un demandante de emprego atopa un novo traballo ou o tempo ata que falla unha máquina. Neste tipo de estudos adoita ocorrer que o suceso de interese non se poda observar, ben porque o suceso aínda non se produciu ao final do período de estudo, ou ben porque o suxeito abandonou o estudo antes do remate. Este fenómeno, denominado censura pola dereita, aparece graficamente nesta figura, onde unha cruz indica que o suceso ocorreu e un círculo indica que o suxeito está censurado. O tempo do evento de interese chámase comunmente tempo de supervivencia, denotado por T, e o tempo ata que o suxeito é censurado denomínase tempo de censura e é denotado por C. Observamos o menor de T e C, e a outra variable de tempo (o máis grande destes dous valores) normalmente non se observa. Polo tanto, este é outro exemplo dunha situación na que os datos son imperfectos e temos que tratar esta imperfección dunha forma adecuada. A área da estatística que trata este tipo de datos denomínase análise de supervivencia, e é moi relevante nun gran número campos como a medicina e a economía do traballo, pero tamén nos estudos industriais ou calquera outro campo de aplicación onde o tempo ata que ocorre algo sexa de interese. As principais contribucións foron realizadas por Sir David Cox, pioneiro na análise de supervivencia e en moitas outras áreas da estatística, que morreu hai uns meses aos 97 anos.

O feito de que só se poida observar o máis pequeno de T e C é un gran obstáculo na análise de supervivencia. Implica que a distribución do tempo de supervivencia, T, non se pode identificar a non ser que se faga unha suposición sobre a relación entre T e C. Na literatura adóitase asumir que T e C son independentes, o que é en certo sentido a suposición máis natural que se pode facer. Esta hipótese ten sentido en moitas situacións prácticas, por exemplo cando un suxeito non experimentou o evento de interese ao final do estudo, ou cando a censura ocorre por outros motivos alleos ao evento de interese. Porén tamén son numerosas as situacións nas que a independencia do tempo de supervivencia e do tempo de censura é dubidosa. Por exemplo, cando un paciente decide abandonar un estudo médico por motivos relacionados coa súa saúde, podemos ter un problema de censura dependente. Outras situacións nas que o mecanismo de censura depende do tempo de supervivencia pódense atopar nos estudos de desemprego, en estudos de transplantes e en estudos onde o custo dun tratamento médico é de interese.

Estes gráficos ilustran o que ocorre cando se ignora a dependencia entre T e C cando se estima a función de supervivencia de T. Cando a correlación entre T e C é cero, que é a imaxe da esquina superior esquerda, a curva estimada, representada pola función escalonada, está preto da verdadeira curva suave. Polo contrario, cando a correlación aumenta, as dúas curvas se separan cada vez máis unha da outra, polo que se mostra claramente que ignorar a dependencia crea un nesgo que pode ser substancial no caso de que esta sexa forte.

Xunto con algúns estudantes de doutoramento e posdoutorais do meu equipo demostramos que esta dependencia entre T e C pode identificarse nun modelo de cópula baixo certas condicións. Comezamos co caso totalmente paramétrico, no que tanto as distribucións marxinais como a cópula son paramétricas. Actualmente estamos traballando na extensión a modelos máis flexibles, como modelos semiparamétricos de Cox, onde resulta que a presenza de covariables axuda a identificar o modelo. Os resultados son sorprendentes xa que só observamos T ou C pero nunca ambos, e aínda así demostramos que podemos identificar a relación entre estas dúas variables. Aínda que as simulacións mostran que hai que ter unha mostra suficientemente grande para obter un estimador preciso, cremos que este resultado é moi útil na práctica, xa que permite corrixir a censura dependente sen ter que asumir que a relación entre T e C sexa coñecida ou precisar facer unha análise de sensibilidade.

Por último, como aplicación, gustaríame mostrar os resultados dun estudo sobre o cálculo de primas nos seguros de vida de persoas casadas. O feito de que as parellas compartan hábitos alimentarios similares e teñan un estilo de vida semellante leva a un problema de censura dependente interesante e orixinal neste contexto. Mostramos nestas dúas figuras o que acontece se se ignora esta dependencia entre o tempo de vida dos dous cónxuxes, á esquerda para o caso da renda vitalicia conxunta e á dereita no caso da renda vitalicia do último superviviente da parella, o que demostra que ignorar a estrutura de dependencia leva a nesgo, especialmente na figura da dereita.

Chegamos agora ao terceiro e último tema do que me gustaría falar, que é o da causalidade. Na estatística clásica adoitamos interesarnos en identificar ou estimar relacións entre variables, pero na inferencia causal imos un paso máis alá e gústanos inferir, por exemplo, o efecto causal dun tratamento médico na evolución dunha enfermidade, ou o efecto causal das políticas de cambio climático sobre as emisións de gases. É fundamental observar que neste caso nos interesa unha relación causal e non só unha asociación ou correlación. Polo tanto, queremos saber se unha variable ten un impacto directo sobre outra, ou se quizais hai unha terceira, que se chama confusora, e que é a responsable da relación entre estas dúas variables.

Que a causalidade é un tema importante pódese ver, por exemplo, no feito de que o último premio Nobel de economía recaeu en tres economistas, Angrist, Imbens e Card, polo seu traballo innovador sobre a causalidade nas Ciencias Sociais. Eles demostraron que os modelos causais son cruciais se un está interesado en como afectan os salarios mínimos aos postos de traballo e ás empresas,

e tamén en cuestións relacionadas co impacto económico da inmigración. Sen un modelo causal só poderían atopar correlacións, pero ningún efecto causal.

Aquí temos un bo exemplo para ilustrar o concepto de causalidade. Se representamos o consumo de chocolate fronte ao número de gañadores do premio Nobel de varios países do mundo, vemos unha aparente tendencia crecente. Por exemplo, se observamos o caso de Suíza e Bélxica, vemos que os suízos consumen máis chocolate que os Belgas (o que é sorprendente), pero tamén teñen (falando relativamente) moitos máis premios Nobel. Entón, parece suxerir que comer moito chocolate fai que sexas máis intelixente ou, se es un xenio, adoitas comer máis chocolate. Pero, é certo isto?

O estudo chamou moita atención e no sitio web ConfectionaryNews.com, os xornalistas deduciron que "Comer chocolate produce gañadores do premio Nobel".

A empresa de medios de comunicación Forbes concluíu na súa páxina web que os chocolates e os premios Nobel están ligados (o que non deixa de ser unha interpretación neutral), pero despois continúa o artigo dicindo que os xenios son máis propensos a comer moito chocolate, o que parece suxerir que hai unha relación causal. Entón, que está a pasar aquí? Quen di a verdade? E somos capaces de reconstruír a verdade absoluta a partir deste estudo?

Se representamos o consumo de chocolate mediante a variable X, e o feito de obter un premio Nobel por Y, entón o primeiro artigo concluíu que X implica Y (polo que comer chocolate faiche intelixente), e o segundo artigo concluíu que Y implica X (polo que ser un xenio fai que comas chocolate). Pero quizais haxa outra explicación, é dicir, pode ser que haxa un factor externo ou un factor de confusión oculto, que teña un impacto tanto en X como en Y. No caso do noso exemplo de chocolate, este factor de confusión oculto podería ser, por exemplo, a riqueza dun país. Os países máis ricos teñen un maior consumo de chocolate e tamén gastan máis diñeiro en investigación científica, polo que hai máis premios Nobel. Polo tanto, ben podería ser que non haxa ningunha relación causal entre X e Y, pero que a aparente tendencia crecente sexa causada polo factor de confusión oculto.

O problema de descoñecer ?que provoca que? é típico nos estudos de observación, onde non temos un control total sobre o experimento, xa que non foi deseñado previamente. O único xeito de distinguir a correlación da causalidade é impoñer un modelo ou gráfico causal, que describa a dinámica e os efectos subxacentes dunha forma adecuada. Un xeito de facelo é empregando o chamado marco de resultados potenciais, tamén chamado modelo causal de Neyman-Rubin. Pártese da idea de que para cada xenio do exemplo do chocolate poderiamos facer un experimento hipotético no que medimos a produción científica se ao xenio non se lle dan bombóns, e tamén medimos os resultados se ao xenio se lle dá moito chocolate todos os días ao longo da súa carreira científica. Ao final, queremos saber se este xenio obtivo un premio Nobel, si ou non. Obviamente, hai dous problemas con este experimento hipotético: primeiro, ninguén aceptaría participar nun experimento deste tipo xa que obrigar a alguén a comer ou non comer chocolates nun experimento de por vida simplemente non é un experimento realista. Por outra parte, aínda que fósemos capaces de montar un experimento deste tipo, só poderíamos observar un dos dous resultados para un xenio dado, pero non os dous. Polo tanto, temos, en certo sentido, un problema de datos faltantes, onde non está dispoñible a metade dos datos. Con isto, necesitamos suposicións sobre as dinámicas subxacentes que describan de forma adecuada o que está a suceder neste estudo.

Xunto a un contratado posdoutoral, algúns estudantes de doutoramento e colaboradores externos, estudamos o problema da inferencia causal nun contexto onde o resultado de interese é un tempo de supervivencia, que está suxeito a censura. Queremos saber cal é o efecto causal dun tratamento sobre o resultado, cando o tratamento non se dá de forma aleatoria. Os pacientes poden, por exemplo, rexeitar un medicamento experimental por razóns relacionadas co seu resultado. Se ignoramos isto, o efecto do tratamento que se estima dun xeito naif a partir dos datos podería estar seriamente nesgado. Polo que necesitamos un modelo causal axeitado que teña en conta a natureza observacional dos datos. Facemos isto mediante as chamadas variables instrumentais, relacionadas coa variable de tratamento (canto máis forte mellor) e que non están directamente relacionadas coa variable de resultado. Estas variables permiten corrixir a natureza non aleatoria do experimento. Neste contexto, mostramos unha serie de resultados importantes, como por exemplo, como se pode identificar o efecto do tratamento sobre toda a poboación. Ademais, cabe destacar que en traballos anteriores nos centramos no efecto do tratamento sobre os que cumpren co tratamento ou sobre os subgrupos tratados. Así mesmo, construímos un modelo flexible que permite variar o tratamento ao longo do tempo, modelos nos que o resultado está suxeito a censura dependente ou modelos nos que os cuantís dos efectos do tratamento son de interese.

Como ilustración do noso traballo metodolóxico, analizamos datos nos que se examina o efecto do cribado mamario. Queremos saber se o exame regular de mama axuda a reducir a mortalidade por cancro de mama. Neste estudo, as mulleres que son asignadas ao chou ao cribado poden rexeitar o tratamento. Dado que o motivo da negativa pode estar relacionado coas súas posibilidades de desenvolver cancro de mama, non temos un experimento aleatorio e necesitamos corrixir o incumprimento. Calculamos os efectos do tratamento por cuantís e descubrimos que non son significativos, agás nalgúns cuantís baixos, o que suxire que o cribado non é útil para reducir a mortalidade por cancro de mama. Os nosos descubrimentos contrastaban cos dos artigos anteriores

que ignoraron o asunto do factor de confusión.

Gustaríame rematar esta presentación cunhas conclusións xerais. Ao amosar algúns exemplos de traballos recentes, espero tervos convencido de que a investigación científica en estatística é un mundo fascinante, con moitos problemas aínda abertos á espera dunha solución adecuada. A estatística ás veces ten mala reputación no público xeral, quen cre que pódese demostrar calquera cousa con estatísticas. Dicíndoo cunha famosa cita de Mark Twain, "hai mentiras, malditas mentiras e estatísticas!" Pero os métodos estatísticos, se se usan dun xeito correcto, son realmente moi poderosos e poden contribuír a numerosos campos de aplicacións, como a medicina, a economía, a psicoloxía, etc. A lista é realmente infinita e a nube de palabras que se mostra aquí dá algúns dos máis destacados ámbitos nos que a estatística xoga un papel moi relevante. Por iso, estou moi orgullosa de recibir este grao de *Honoris Causa* en estatística.

Moitas grazas!