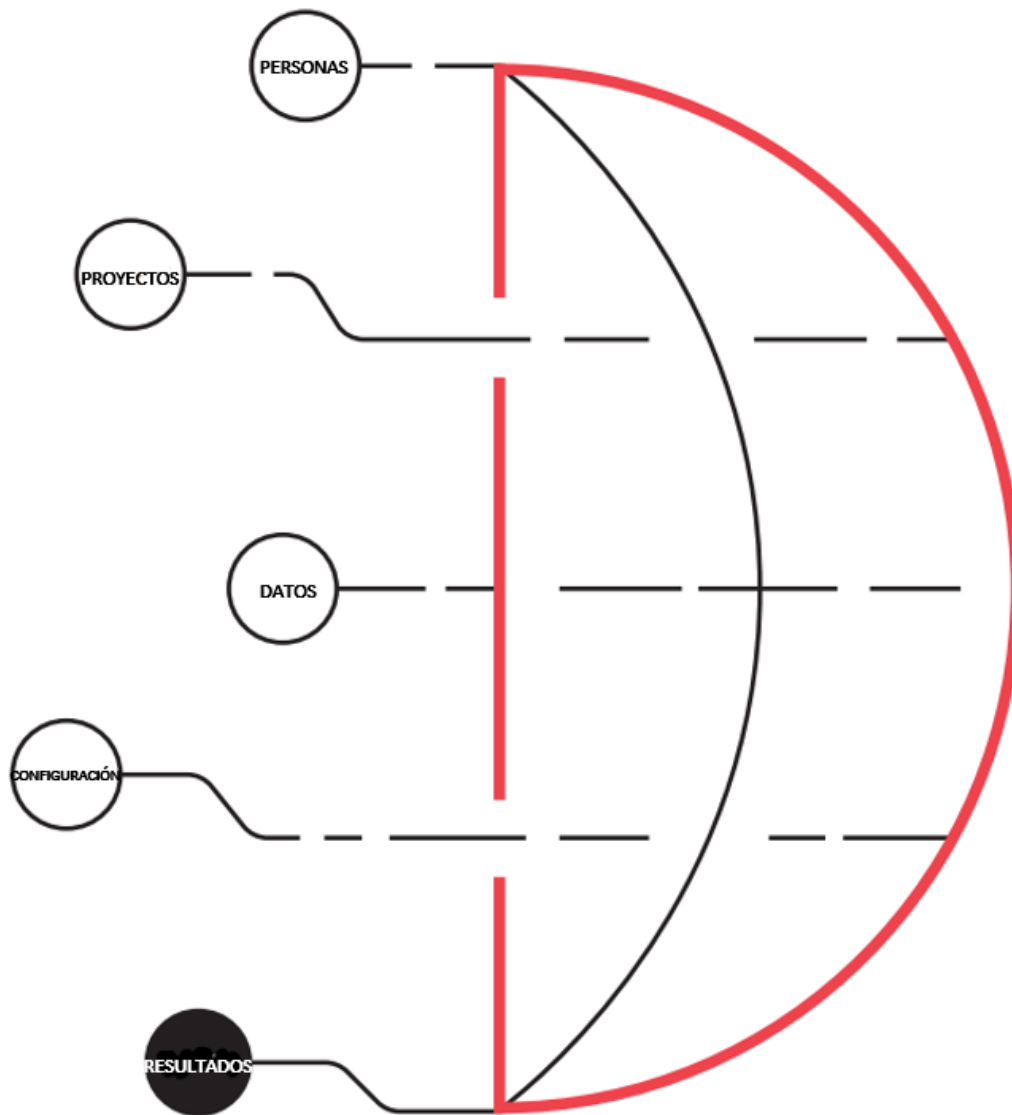


Manual de Control de Divulgación Estadística de Resultados

Emily Griffiths (University of Manchester)
Carlotta Greci (The Health Foundation)
Yannis Kotrotsios (Cancer Research UK)
Simon Parker (Cancer Research UK)
James Scott (UK Data Archive, University of Essex)
Richard Welpton (The Health Foundation)
Arne Wolters (The Health Foundation)
Christine Woods (UK Data Archive, University of Essex)



Traducido al español por Natalia Volkow, June 2020

Reconocimientos

Los autores desean agradecer a las siguientes personas por su guía y apoyo; y por revisar este documento.

Anthea Springbett, Felix Ritchie (Universidad del Oeste de Inglaterra), Mark Elliot (Universidad de Manchester), Matthew Woollard (Archivo de Datos del Reino Unido, Universidad de Essex), y participantes del 15 de Agosto de 2018 en los talleres del Grupo de Profesionales de Acceso Seguro a Datos (GPASD) en CDE (incluyendo personal del Laboratorio de datos de la oficina de la Hacienda Pública y Aduanas del Reino Unido, el Servicio de Datos Seguros de la Oficina Nacional de Estadística del Reino Unido, Servicio de Datos del Reino Unido, Investigación de Cáncer del Reino Unido, Servicio Médico Social del Escocia y la Fundación para la Salud).

Los autores agradecen el apoyo generoso de la Fundación para la Salud para la publicación de este Manual.

Sobre esta versión

Bienvenidos a la versión 1.0 de nuestro Manual CDE. Después de la publicación de la versión beta en marzo de 2019, estamos contentos de escuchar que el personal que lleva a cabo revisiones de resultados de servicios seguros de acceso a datos, lo están usando y han encontrado que es útil para poder realizar su trabajo.

Esto refleja dos hechos:

- En años recientes, se han establecido muchos servicios de acceso a datos en enclaves seguros tanto en el gobierno, la academia y en otras instancias, que ha contratado personal cuya responsabilidad es asegurar que los resultados producidos a partir de datos confidenciales conllevan un riesgo mínimo de revelar la identidad e información personal.
- El personal contratado consideraba que había muy pocos recursos de los que pudieran echar mano, cuando revisaban las primeras versiones de este Manual, empezaron a usarlo para poder realizar su trabajo.

Para producir este Manual, los autores han compartido su conocimiento colectivo de CDE y explicaciones de cómo llevan a cabo la revisión de CDE. Sin embargo, queremos dejar sentado que hay más trabajo que se debe llevar a cabo, las versiones subsecuentes serán actualizadas. Hay otras personas que llevan a cabo este trabajo y que no han podido compartir sus experiencias y podrían contribuir con sus propios ejemplos. Hay investigadores en el tema de confidencialidad y privacidad que quisiéramos escuchar, en particular para asegurar la exactitud técnica de los contenidos. Quisiéramos escuchar a cualquiera que tuviera ideas de cómo se puede mejorar este Manual. Por favor contáctenos visitando [https:// securedatagroup.org/sdc-handbook](https://securedatagroup.org/sdc-handbook).

Sobre el Grupo de Profesionales de Servicios Seguros de Acceso a Datos

El Grupo de Profesionales de Acceso Seguro a Datos (GPASD) se creó en 2011 para juntar personal de los servicios de acceso segura para compartir experiencias y desarrollar mejores prácticas.

.....

Esta red está integrada por miembros que trabajan para poder brindar acceso seguro a datos confidenciales de negocios, salud y socioeconómicos dentro de gobierno, la academia y las organizaciones no gubernamentales. Este Manual es parte de un trabajo de mayor envergadura que desarrolla el grupo GPASD para compartir mejores prácticas y experiencias en administrar el acceso a datos sensibles.

Si desea tener más información sobre GPASD puede encontrarla en securedatagroup.org

La responsabilidad de la información y puntos de vista que se presentan en este Manual es por completo de los autores y no necesariamente representa los puntos de vista de la organización donde trabajan.

Tabla de Contenidos

Introducción	6	Diagramas de simetría	48
Audiencia	9	Árboles de decisiones y criterio de exclusión	51
Desarrollo de este Manual	10	Análisis de supervivencia: curva de Kaplan-Meier	53
Alcance	12	Análisis espacial (mapas)	55
Objetivos	12	Coeficientes de Gini	57
Estructura	13	Tasas de concentración	59
Conceptos principales	14	Regresiones	61
SECCIÓN A		Residuales	64
Sobre el Control de Divulgación Estadística	15		
El riesgo estadístico: ¿de qué se trata?	16	Gráficos de margen	66
¿Qué es Control de Divulgación Estadística (CDE)?	19	Pruebas estadísticas	68
Evaluación de riesgo	22	Sección C	
		Requerimientos organizaciones y de administración	69
Los riesgos estadísticos: principio y reglas	26	Introducción	70
SECCIÓN B		¿Cómo implementar CDE en una organización?	71
¿Cómo evaluar resultados estadísticos ?	28		
Introducción: CDE	29	Administración de los analistas	77
Usando la guía	30	¿Qué es un buen conjunto de resultados estadísticos?	79
		Los buenos	82
Estadísticas descriptivas	31	Los malos	83
Percentiles	33	Administración de expectativas	84
Histogramas	35	Otros recursos	87
Diagrama de caja	38	Glosario	90
Coeficientes de correlación	40	Derecho de autor	94
Análisis de factores	42		
Índices	44		
Gráficos de dispersión	46		

Introducción

Los analistas están demandando acceso a más datos sobre individuos y organizaciones, con más detalle como antes nunca. Al usar datos más específicos es posible llevar a cabo análisis más innovadores y robustos, explorar nuevos temas y vertientes, y generar resultados que fundamentan de mejor manera la definición de política pública y las decisiones de los negocios.

.....

Estas fuentes de datos ya están disponibles. Sin embargo, el nivel del detalle es tal que esos datos se consideran “personales, sensibles y confidenciales” y están al amparo de las leyes de Protección de Datos. Por ello, el acceso a estos datos se brinda en general in “Configuraciones seguras” (servicio de acceso seguro a datos sensibles) para asegurar que la confidencialidad de los sujetos de los datos se salvaguarda.

SOBRE LOS CINCO SEGUROS

Los Cinco Seguros es un marco de referencia sobre cómo dar acceso a datos. Fue conceptualizado en su origen por el Profesor Felix Ritchie quien en 2006 trabajaba en la Oficina Nacional de Estadística del Reino Unido (en la actualidad se encuentra en la Universidad del Oeste de Inglaterra).

Este marco proporciona un proceso de toma de decisión que permite el “uso seguro” de datos personales confidenciales. “Objetivo estadístico” era el propósito que se tenía en mente cuando se definió este marco de referencia.

De manera específica, este marco de referencia considera que hay cinco aspectos que hay que considerar para tener un acceso seguro a datos:

- Datos Seguros (¿cuáles son las características de los datos? ¿Son tan detallados que pueden contener información confidencial que puede atribuirse a individuos?)
- Personas Seguras (¿Quién va a tener acceso a los datos? ¿Tienen las credenciales debidas, experiencia y motivación para tener acceso a los datos?)
- Proyectos Seguros (¿cuál es el objetivo que buscan obtener con los datos? ¿Es con fines de investigación, o la persona que solicita tener acceso a los datos busca

identificar a alguien? ¿o hay alguna otra razón por la que quiera tener acceso los datos? ¿Hay un beneficio público?

- Configuración segura (¿considera el ambiente en el que se brindará acceso a los datos? ¿Existen las medidas de seguridad para proteger la confidencialidad de los datos?)
- Resultados seguros (¿Qué información se entregará? ¿Estadísticas? ¿Esta información se puede usar para identificar a alguien? ¿se podría entregar información confidencial?)

Este es un marco de referencia muy usado, no solo por la Oficina Nacional de Estadística del Reino Unido, otras agencias estadísticas en todo el mundo (incluyendo Australia y Nepal); así como otras organizaciones como el Servicio de Datos del Reino Unido, Centro de Investigación de Cáncer del Reino Unido y la Fundación para la Salud. Más información de cómo aplicar el Marco de Referencia de Cinco Seguros se puede encontrar en Desai, Ritchie and Welpton (2016): ver Otros Recursos.

Este Manual ha sido diseñado de manera específica para personas que quieren asegurarse, que se producen Resultados Seguros de fuentes de datos personales y confidenciales (el quinto punto de los Cinco Seguros que se enlistó antes). Ver Audiencia para más información.

Configuración Segura y Resultados Seguros

En este marco de referencia, los analistas pueden usar las Cinco Seguros para tener acceso y analizar datos sensibles. Los resultados estadísticos que generen en la Configuración Segura les serán entregados después de que hayan sido revisados para evaluar riesgo de divulgación de confidencialidad (por ejemplo, Control de Divulgación Estadística, CDE) para asegurar que los resultados que se publiquen no revelan la identidad ni contienen información confidencial sobre el sujeto de los datos (una observación en los datos). Solo se liberan de la Configuración Segura Resultados Seguros.

En el Reino Unido operan varios servicios con Configuración Segura. Esto incluye agencias de gobierno como la Oficina Nacional de Estadística que tiene el Servicio Seguro de Investigación y el Datalab de la Oficina de Ingresos y Aduanas; infraestructuras académicas con apoyos financieros como el Laboratorio del Servicio Seguro de Datos del Reino Unido y organizaciones no gubernamentales como la Fundación para la Salud y el Centro de Investigación de Cáncer del Reino Unido. En fechas recientes, Conectando Salud de Ciudades, que es una colaboración entre la academia y el sector salud, estableció una red de Ambiente de Investigación Confiables para analizar datos de pacientes. Estas organizaciones han establecido sus Configuraciones Seguras para adquirir y brindar acceso a datos confidenciales para uso de sus analistas y proveer acceso a analistas de terceras partes. Dentro de este marco de referencia, es fundamental asegurar que todo resultado estadístico que se publica no presenta riesgo de violentar la privacidad de los datos de los

sujetos a los que refieren los datos. CDE tiene un papel crítico en mitigar el riesgo y representa un componente clave del sistema de seguridad de la información de la Configuración Segura.

Este Manual busca alcanzar dos objetivos:

- Introducir los principios del CDE con la presentación de los retos más importantes y las mejores prácticas en el Reino Unido;
- Apoyar a organizaciones y su personal que aplican revisión de CDE a resultados estadísticos generados a partir de datos confidenciales.

Audiencia

Este Manual se escribió para dos audiencias:

- Personal que trabaja en Configuraciones Seguras que son responsables de aplicar CDE a resultados estadísticos derivados de datos confidenciales:
- Analistas que están planeando publicar sus hallazgos derivados de datos que se mantienen en Configuraciones Seguras.

.....

Para cualquiera de estas dos audiencias, el Manual asume que el lector está familiarizado con los conceptos estadísticos básicos y experiencia en el manejo de datos. El Marco de Referencia de Competencias del GPASD (securedatagroup.org/guides-a-and-resources) proporciona guías de como personal nuevo puede desarrollar sus habilidades, que incluye CDE.

Es la responsabilidad de la organización que opera la Configuración Segura el capacitar a sus analistas en cómo revisar resultados estadísticos antes de entregarlos. Sin embargo, buscamos que este Manual sea un recurso de referencia útil y, que ayude a generar un enfoque consistente de CDE en la evaluación de resultados estadísticos generados con datos confidenciales. Por ejemplo, los resultados estadísticos que se presentan en este Manual con un conjunto de ‘sería útil saber’ que puedan usarlos en cualquier sistema de CDE y cualquier tipo de usuario.

Desarrollo de este Manual

En abril de 2017, los autores, junto con personal de la Oficina Nacional de Estadística del Reino Unido, se reunieron para discutir y comparar requerimientos y lineamientos para el personal que revisa resultados estadísticos para asegurar que son seguros cuando se entregan. Una característica del grupo es que los miembros del GPASD son personal que llevan a cabo de alguna manera CDE. La revisión de CDE es en general un componente clave en el manejo de la Configuración Segura.

.....

Los miembros del GPASD en general han seguido la publicación ESSnet/DwB, “Lineamientos para la Revisión de Resultados”; y en salud, la “ISB1523 Anonymisation Standard” para llevar a cabo el CDE. La Oficina del Comisionado de Información (OCI) publicó un “Código de Prácticas de Anonimización”. Sin embargo, el grupo considera que estos lineamientos se beneficiarían de mejoras, debido a los avances en los métodos estadísticos, y el mayor uso de datos confidenciales en Configuraciones Seguras.

Los lineamientos de ESSnet son una fuente, de un número reducido de referencias de información del que el GPASD tiene identificado, que solo provee asesoría específica sobre la evaluación de resultados estadísticos para control de divulgación (aun cuando hay una literatura vasta en privacidad estadística, esta se orienta a los responsables de la publicación de información estadística nacional). Hoy en día los analistas producen una amplia variedad de resultados estadísticos, no solo tablas, sino mapas, estadística descriptiva, resultados de modelos. Las Configuraciones Seguras debe estar capacitados para entregar cualquier tipo de resultado que los analistas producen, y asegurar que estos resultados no quebrantan la confidencialidad de los datos, cuando se entreguen. Esta flexibilidad hace necesario que el personal responsable de revisar y entregar los resultados estadísticos tenga la suficiente experiencia en aplicar los principios de CDE a una variedad de tipos de análisis.

Mucha literatura académica se ha publicado en el tema de control de divulgación estadística. Por ejemplo, uno puede revisar artículos en la Journal of Privacy and Confidentiality or Transactions in Data Privacy para asesoría técnica en el control de

divulgación estadística. GPASD considera que esta literatura proporciona conceptos útiles; sin embargo, es vital que las lecciones puedan adoptarse en la práctica y evaluarlas.

Con esto en mente, el GPASD consideró que sería útil desarrollar un Manual de CDE para usarse:

Como guía práctica sobre cómo evaluar los requerimientos más comunes de requerimientos de control de divulgación estadística y proveer una base para evaluar casos más complejos;

Para asesoría de cómo promover la producción de resultados estadísticos buenos y aceptables, de los cuales se facilite el proceso de CDE.

Esto ayudará a asegurar:

- Que los resultados estadísticos son evaluados de manera precisa y comprensiva.
- Que los resultados estadísticos se revisen de manera consistente y, en la medida de lo posible, estén alineadas con las prácticas de otras Configuraciones Seguras.

Los autores y el GPASD tienen la intención que el Manual se revise de manera continua, para asegurar que es útil, esté actualizado, y atienda las necesidades de los individuos que evalúan estadísticas en las Configuraciones Seguras; proveedores de datos, que requieren asegurar que la confidencialidad de los datos a los cuales ofrecen acceso no se compromete, y a los que desarrollan e imparten cursos de capacitación a personal y analistas externos.

ACTUALIZACIONES FUTURAS

Como este es un trabajo en proceso, el Manual se actualizará de manera regular. securdatagroup.org/sdc-handbook

ALCANCE

Este Manual es sobre la evaluación, administración y entrega de resultados estadísticos, producidos de fuentes de datos confidenciales que se mantienen en Configuraciones Seguras. Por “resultados estadísticos”, entendemos estadísticas derivadas de datos y no otro resultado que se pueda producir, como

- Conjunto de datos (por ejemplo, subconjuntos de datos, o conjunto de datos agregados;
- Versiones sintéticas de fuentes de datos confidenciales;
- Ponderadores derivados de fuentes de datos confidenciales.

OBJETIVOS

Para resumir, los objetivos de este Manual son:

- Traducir los conceptos de control de divulgación en asesoría práctica, medidas y pasos para evaluar riesgo de divulgación de resultados estadísticos;
- Asegurar a los dueños de los datos, que el acceso a los datos que ellos proporcionaron es administrado de manera segura, y que la confidencialidad de los datos no será comprometida;
- Facilitar el proceso de solicitud de entrega de resultados estadísticos;
- Para ser usado en la capacitación del personal que trabaja en Configuraciones Seguras con ejemplos que ellos puedan usar de manera directa en sus objetivos de capacitación.

Pensamos que este Manual provee una fuente de referencia robusta para el personal que trabaja en Configuraciones Seguras para obtener la acreditación en gobernanza de información, a la luz del desarrollo de nuevas técnicas estadísticas, el que se compartan y combinen fuentes de datos de diferentes disciplinas (como es salud y ciencias sociales), y cambios del marco regulatorio.

Estructura

EL DOCUMENTO TIENE TRES SECCIONES.

Sección A, para aquellos que buscan una introducción sobre cómo hacer resultados estadísticos ‘seguros’. Se pueden incluir a los analistas que revisan sus resultados, y personal que lleva a cabo evaluación de CDE.

Sección B proporciona lineamientos prácticos paso por paso para llevar a cabo la evaluación de control de divulgación de confidencialidad.

Sección C es para el personal que desarrolla y administra un “servicio” de datos en el cual el Control de Divulgación Estadística es un elemento importante en su trabajo.

.....

Las secciones A y B proveen un panorama del Control de Divulgación Estadística y definen los conceptos de riesgo de divulgación. Se describe una lista extensiva de los resultados estadísticos más comunes, con ejemplos y guías prácticas de cómo llevar a cabo el CDE. Éstos incluyen tablas de frecuencias, coeficientes de regresión, histogramas, gráficos de sedimentación, y técnicas más complejas como análisis espacial y de sobrevivencia. Para cada tipo de resultado, se brinda asesoría de qué es lo que se debe buscar, y sugerimos preguntas que se le pueden hacer a la persona que solicitó la entrega de resultados y a quien evalúa los resultados que se van a entregar. Reconocemos que hay otros tipos de resultados estadísticos que se requieran evaluar el riesgo de divulgación de confidencialidad, proporcionar una guía para saber cómo manejar la divulgación estadística en nuevos tipos de resultados estadísticos que nunca antes, se han evaluado.

Sección C proporciona asesoría y guía para organizaciones de cómo administrar el proceso de CDE y las cargas de trabajo. También se consideran otros métodos para reducir el riesgo de entregar resultados que lleven el riesgo de violentar la confidencialidad de los datos, que también incluyen como incentivar el requerimiento de buenos resultados.

Conceptos principales

A lo largo de este libro, nos referimos a cierto número de términos. Los hemos enlistado aquí para a manera de guía, también se pueden encontrar en el Glosario.

.....

Revisores de Resultados

Son las personas responsables de revisar los riesgos de divulgación de los resultados estadísticos (alias productos estadísticos) creados en la Configuración Segura.

CDE

El proceso que se aplica a los resultados estadísticos (productos estadísticos) para mitigar el riesgo de divulgación potencial de los resultados que se extraen de la Configuración Segura.

Configuración Segura

Los medios técnico -ya sea ubicados de manera física o virtual- mediante la cual los analistas trabajan con los datos.

Sujeto de los Datos

La unidad de observación en un grupo de datos. De manera usual son individuos (personas físicas) o negocios (personas morales) depende de la fuente de los datos.

SECCIÓN A

Sobre el Control de Divulgación Estadística

El riesgo estadístico:

¿de qué se trata?

EN ESTA SECCIÓN APRENDERÁ SOBRE:

1. El concepto detrás de cómo alguien puede ser identificado en las estadísticas publicadas.
2. Cómo las estadísticas publicadas pueden quebrantar la confidencialidad.

EL CONCEPTO DETRÁS DEL CONTROL DE DIVULGACIÓN ESTADÍSTICA

La idea básica es que, usando cierta información estadística, sería posible inferir información confidencial, e incluso identificar a alguien, a partir de ciertos resultados que se liberen.

Supongamos que tenemos ciertos datos de un grupo de empresas que operan en diferentes sectores. Los datos que se incluyeron son volumen de ventas y empleo para cada empresa. ¿Cómo podría ocurrir la divulgación estadística y la identificación?

Empecemos por los datos de agregación de las ventas por sectores. Nosotros conocemos el número de empresas en cada sector.

Podríamos encontrar que, para un sector, hay dos empresas. Podríamos publicar las ventas totales para el sector, que sería la suma de las ventas de las dos empresas.

Por ejemplo, si las ventas del sector ascendieron a £1,000,000, y la Empresa A sabría que sus ventas fueron de £400,000, si solo hay dos empresas en el sector, los restantes £600,000 serían las ventas de la Empresa B, que es su competidor.

Ahora supongamos que sumamos una tercera empresa, Empresa C. Ahora se publica el resultado agregado para las tres empresas del sector asciende a £1,200,000. Si bien la empresa A conoce su parte que es £400,000, y puede saber por cierto que los £800,000 restantes lo comparten la Empresa B y Empresa C. Sin embargo, no puede saber más, no puede saber qué proporción de las ventas corresponde a la Empresa B y cuál a la Empresa C, por ello, no hay divulgación estadística.

Esto explica porque, desde la perspectiva estadística, siempre debe haber un “límite” de al menos tres observaciones, es decir la frecuencia mínima en la que una estadística se basa. Sin embargo, las Configuraciones Seguras en general usan un límite mayor que éste, para hacer frente a la “divulgación secundaria”. Esto se explica en las siguientes dos secciones.

QUEBRANTO DE CONFIDENCIALIDAD ESTADÍSTICA

En el ejemplo anterior, ilustramos cómo puede ocurrir la divulgación de confidencialidad usando un ejemplo de datos de negocios (dado que las empresas, de un cierto tamaño; están obligadas a publicar su contabilidad, quizás se pregunte porqué usamos este ejemplo: la Oficina Nacional de Estadística del Reino Unido (ONE) proporciona acceso seguro a muchas fuentes de datos de negocios, que se colectan con encuestas bajo el compromiso de confidencialidad.

Aquí hay otro ejemplo.

Estadísticas de cáncer

Un número de agencias el sector salud publican de manera regular estadísticas sobre la incidencia de cáncer. Desgraciadamente, el cáncer es algo común, por lo que uno esperaría que los resultados de incidencia de cáncer desagregados por autoridad local serían números grandes. Este no siempre es el caso. Por ejemplo, la City de Londres, si bien tiene una población importante de instituciones financieras con muchos trabajadores, si tiene algunos residentes. Por ello, el publicar incidencia de un cáncer común, como el de pulmón, probablemente desplegaría una frecuencia pequeña. Para cánceres más raros, el número sería de un solo dígito, y es probable que los resultados agregados arrojarían cifras cerca de 1 o 2. Si usted fuera la persona con el diagnóstico de esa enfermedad, usted podría reconocerse en las estadísticas. Otros también podrían reconocerlo. Aquí hay dos aspectos:

1. Usted ha sido identificado
2. Se ha rebelado información médica sobre usted.

(y tercera, información médica confidencial se ha asociado con usted).

DISPUTAS

Departamento de Salud, estadísticas de aborto, 2011

En ocasiones puede resultar difícil la decisión de publicar estadísticas. El caso de 2011 que involucró al Departamento de Salud, que decidió no publicar estadísticas sobre aborto por un periodo de 24 semanas, es un caso en cuestión.

Cuando se fragmentaba en ciertas categorías, las frecuencias terminaban como cifras de un sólo dígito. En algún caso, era posible derivar que un aborto se debió a una probable malformación del bebé. Los números era muy pequeños, el Departamento de Salud argumento en contra de su publicación sobre la base de que se podría identificar a las

mujeres. Un requerimiento de la Libertad de Información demandó que las estadísticas se entregaran, y fueron negadas, se sobrevino un caso legal.

Eventualmente la Suprema Corte dictaminó que las estadísticas se debían publicar, porque siendo estadísticas, los datos personales se presentaban anonimizados con los datos no personales. Esto hace destacar el tema de cuando los datos estadísticos publicados se pueden revertir en datos personales.

Incluimos este ejemplo para explicar que el CDE no es un tema de corte raso. Hemos escrito esta guía con una perspectiva de evaluación de riesgo. El hecho es que puede ser por completo seguro publicar estadística de frecuencias pequeñas en las que nadie pueda ser identificado o dañado. Por el otro lado, puede resultar beneficioso para la sociedad que se publiquen estadísticas aun cuando puede haber divulgación de la confidencialidad. Sin embargo, los revisores de resultados no siempre son los responsables de tomar estas decisiones, los dueños de los datos son lo que tendrán que decidir esto.

LECTURAS SUGERIDAS

1. ICO Anonymization Code of Practice, pp 14-15
<https://ico.org.uk/media/1061/anonymisation-code.pdf>

¿Qué es Control de Divulgación Estadística (CDE)?

EN ESTA SECCIÓN APRENDERÁ SOBRE:

1. Control de Divulgación Estadística
2. Cómo este Manual aborda el CDE
3. Porque la fuente de los datos es importante
4. La importancia de la unidad de observación

El Control de Divulgación Estadística (CDE) busca:

- Prevenir que se revele la identidad del sujeto o unidad de observación al que refieren los datos;
- o/y publicar información asociada o que pertenece al sujeto o unidad de observación

Tradicionalmente el CDE se ha aplicado:

- A tabulados estadísticos (a menudo generados por los Institutos Nacionales de Estadística), antes de que se publiquen;
- A los microdatos, con el propósito de crear versiones anonimizadas de los datos originales.

Sin embargo, debido al crecimiento en el número de Configuraciones Seguras y la posibilidad de llevar a cabo análisis complejos se han desarrollado técnicas para evaluar el riesgo de divulgación de los resultados estadísticos producidos en estos servicios. Después de todo, es poco lo que se logra si al establecer Configuraciones Seguras con acceso seguro a datos confidenciales, los resultados estadísticos que se entregan quebrantan la confidencialidad de los datos.

El riesgo de re-identificar el sujeto o unidad de observación (a lo que refieren los datos, por ejemplo: una persona, organización) de este tipo de fuentes de datos es real y se necesita administrar de una manera muy cuidadosa y eficiente. En este Manual, el CDE se relaciona a la evaluación de resultados estadísticos generados de datos confidenciales, más que de tabulados y microdatos que se publiquen para el público.

Cabe aclarar que, para los propósitos de este Manual, no hacemos distinción si los individuos o unidades de observación a los que refieren los datos están vivos o muertos (para el caso, si una empresa a la que refieren los datos es solvente o está en bancarota o cerrada). Si bien la Ley de Protección de Datos cubre a personas vivas, la Ley del Censo contiene una cláusula que limita a 100 la publicación de detalle de los individuos; y la Ley de Comercio Estadística no distingue entre empresas vivas o muertas. Para hacer las cosas más sencillas, asumimos que debemos proteger la confidencialidad de los datos y la identidad de todos los sujetos o unidades de observación a los que refieren los datos.

Si bien el CDE busca prevenir la re-identificación del sujeto o unidad de observación, es una estrategia de “minimización de riesgo” más que una estrategia de “eliminación de riesgo”. Muchos podrán sostener con gran entusiasmo que eliminaron por completo el riesgo de divulgación, en particular cuando cambia el contexto mediante el cual los datos personales se mantienen confidenciales. Por ejemplo, las personas pueden publicar información personal en sus redes sociales cuando antes hubieran esperado que esa información debía mantenerse confidencial. El tratar de eliminar el riesgo de divulgación estadística, ignoraría por completo los cambios de como pensamos sobre los datos y nuestro apetito por divulgar información confidencial sobre nosotros; y como resultado al final fracasaremos en el CDE. En cambio, la minimización de riesgo toma en cuenta que afuera hay condiciones que no podemos controlar.

El enfoque de este Manual toma en cuenta estas consideraciones; no se debe ignorar el riesgo de divulgar confidencialidad, pero a cambio busca administrar y mitigarlo lo más posible. La única manera de eliminar el riesgo de divulgación es no publicar ningún resultado estadístico de los servicios de Configuración Segura; pero adoptar esta posición repercutiría de manera negativa para los análisis, investigación y el bien público al cual sirven. Se establece un balance cuando se pueden entregar resultados estadísticos al mismo tiempo que se minimiza el riesgo de quebrantar confidencialidad lo más posible, y que se cumple la debida diligencia.

ENFOQUES

La Configuración Segura facilita el acceso a datos confidenciales para análisis. Como un analista puede elaborar cualquier resultado de su análisis, las Configuraciones Seguras necesitan estar preparados para evaluar cualquier tipo de resultados estadístico. Todo resultado estadístico debe ser evaluado para saber si presenta riesgo de que los sujetos o unidad de observación a los que refieren los datos pueden ser re-identificados, o si pueden revelar información confidencial.

En algunos casos, el beneficio público de que se publique cierta estadística puede pesar más que el riesgo. Por ejemplo, en una prueba clínica podría ser muy beneficioso publicar el hecho de que una persona tuvo una reacción muy severa al nuevo

tratamiento, si la conclusión de estos resultados es que el medicamento se retira para desarrollos y pruebas posteriores. Una persona puede destacarse y quizás de manera indirecta ser identificada, pero el beneficio a los otros puede considerarse que sobrepasa el riesgo.

¿LA FUENTE DEL DATO IMPORTA?

La fuente de los datos es relevante cuando se hace una evaluación del CDE. Por ejemplo, se puede argumentar que los datos de una encuesta conllevan un riesgo menor de identificar a alguno de los sujetos o unidades de observación por la naturaleza de la muestra (y el marco muestral), mientras que los datos de registros administrativos incluyen a todos los que son relevantes, por lo que es más probable que una posible identificación sea más precisa. Cuando se toma una decisión de CDE, la fuente de donde provienen los datos -si es de una encuesta o de registros administrativos- puede ser un factor, a considerar, en la evaluación del riesgo.

UNIDAD DE OBSERVACIÓN

Se debe tener cuidado en entender la unidad de observación o sujeto de los datos cuya confidencialidad estamos tratando de proteger. El CDE puede aplicarse a datos de individuos, negocios, viviendas, o subgrupos particulares de población como pacientes, contribuyentes o personas vulnerables. Si bien la aplicación del CDE sigue el mismo proceso, el contexto y significado de la divulgación asociada a una unidad de observación puede variar.

Por ejemplo, un analista de un conjunto de datos de negocios puede producir estadística referida a lugares (incluyendo baja desagregación geográfica como códigos postales). Si los lugares (por ejemplo, sucursales de supermercados) pertenecen a la misma organización, entonces el riesgo de re-identificación de la organización puede prevalecer, aun cuando, por ejemplo, todas las estadísticas sean agregadas y se cumplan los requerimientos de frecuencias para los individuos o unidades de observación asociados. En una primera mirada, podría parecer que los resultados estadísticos se pueden entregar; pero si la unidad de observación no es lo que parece en un principio, entonces se debe tener mayor cuidado en considerar el contexto del análisis y las implicaciones de divulgación.

LECTURAS SUGERIDAS

1. Duncan, D., Elliot, M., Salazar-Gonzalez, J., (2011) "Statistical Confidentiality: Principles and Practice", published by Springer

Evaluación de riesgo

EN ESTA SECCIÓN USTED APRENDERÁ SOBRE:

2. Identificación, atribución y divulgación secundaria
 3. La importancia de la información del contexto
 4. Si se debe revisar todos los resultados que genera un usuario o no
 5. Tipos de riesgo que preocupan a los dueños de los datos
 6. Concepto de límites y dominancia.
-

Conceptos Básicos

IDENTIFICACIÓN

Esta ocurre cuando el sujeto o unidad de observación del dato es identificado a partir de resultados estadísticos. El CDE busca minimizar el riesgo de que esto suceda.

La identificación pudo ocurrir cuando un número pequeño de observaciones están aisladas y quedan presentes en los resultados. Por ejemplo, calcular el ingreso máximo de un conjunto de datos probablemente revelaría el ingreso exacto de una persona. Cuando se combine con otros resultados estadísticos como edad o ubicación geográfica, la persona con el ingreso más alto se podría identificar y con ello se revelaría sus datos confidenciales.

ATRIBUCIÓN

Cuando se pueden poner juntas un número de características y asociarlas con la misma observación (incluso sin tener identificadores directos como nombre y dirección), un sujeto de los datos o unidad de observación puede ser identificado. Divulgación por atribución puede ocurrir cuando características, que parecen individualmente anónimas, se pueden poner juntas. Como con la identificación, esto puede ocurrir cuando un número pequeño de observaciones se presentan en un resultado estadístico.

Por ejemplo, un gráfico disperso que despliegue el valor de dos variables para cada sujeto de datos o unidad de observación probablemente permita que se asocie la información a cada uno de los sujetos de datos o unidad de observación. Además, hay otros escenarios en los que la atribución puede darse. Por ejemplo, si un grupo de sujeto de datos o unidades de observación comparten una característica, que se presenta como un resultado estadístico, la característica se puede atribuir a cada uno de los sujetos de datos o unidad de observación.

DIVULGACIÓN SECUNDARIA

La divulgación secundaria ocurre cuando se combina un resultado publicado con otra información que se producen con nuevas estadísticas que son reveladoras.

Por ejemplo, dos tablas de estadísticas descriptivas producidas de una fuente de datos pueden presentarse en varios cortes en diferentes formas. Una observación puede aislarse, por ejemplo, si una tabla se deduce de la otra.

De manera alternativa, las estadísticas pueden combinar datos disponibles fuera de Configuraciones Seguras donde los datos son analizados. Siempre hay un riesgo potencia que aún con control de divulgación los resultados estadísticos pueden combinarse con otros datos u otros resultados para revelar la identidad del sujeto de datos o unidad de observación.

INFORMACIÓN CONTEXTUAL

Evaluar los riesgos de resultados estadísticos puede facilitarse de manera significativa cuando está disponible información relevante del contexto.

Esto incluye, pero de manera cierta, no es limitativo a:

- Cuando se estudia una industria u ocupación específica.
- Cuando los datos y resultados han sido reunidos y promediados para un periodo de tiempo y área geográfica específicos;
- Cuando se tiene cierto conocimiento de la población subyacente en los datos
- Las muestras de diferentes tamaños y sin ponderadores subyacentes para distintos subgrupos de población en más de un tabulado o gráfica;
- Otros, como, información que está disponible en el dominio público o la tiene quienes recibirán los resultados.

Sin contexto o sin información sobre los resultados estadísticos hay un grado mayor de incertidumbre con relación al riesgo de divulgación de confidencialidad de los resultados estadísticos. En medio de esta incertidumbre los resultados estadísticos no deben ser liberados de la Configuración Segura a menos que se provea información adecuada.

¿ES NECESARIO REVISAR CADA RESULTADO?

Los analistas tienden a producir más resultados estadísticos que los que requieren se les revisen para publicar. Como los métodos son revisados y los analistas siguen su camino iterativo, el usuario decidirá qué resultados estadísticos valen la pena solicitar para entrega.

En el marco de referencia de la Configuración Segura, cada resultado que solicite el usuario se le entregue se debe revisar para CDE. Algunos resultados son más complejos que otros e implicarán mayor tiempo y demanda técnica para su evaluación. Por ejemplo, en la evaluación de los resultados de análisis de regresiones se tomará menos tiempo que un tabulado de descriptivos estadísticos. Estas diferencias se explican con más detalle más adelante en el Manual. Sin embargo, todos los resultados que solicitan los analistas deben ser revisados para CDE. Imagine las consecuencias que tendría si un resultado no se revisa antes de que se entregue y revela la identidad del sujeto de datos o unidad de observación, causando daño y molestia al individuo o unidad de observación, y ocasionando penalización por el quebranto de la ley.

LÍMITES

La mayoría de las evaluaciones de riesgo de divulgación juzgan los resultados estadísticos con base en una serie de límites, definidas como el mínimo de observaciones que contienen los estadísticos.

En una tabla de frecuencias esto es algo obvio: si una celda tiene un valor de frecuencia menor que el que se establece en el límite, entonces el tabulado se considera “inseguro” para entregarse, a menos que se pruebe después de un análisis más minucioso con información del contexto, que no permite la identificación de ningún sujeto de los datos o unidad de observación.

Dicho esto, es importante hacer notar que los tabulados no pueden ser evaluados de manera aislada. Como se explicó antes, puede ocurrir divulgación secundaria cuando dos tabulados se diferencian, aun cuando ambos cumplan con los límites, de manera individual en cada tabulado.

Los dueños de los datos tienen requerimientos diferentes para el valor de los límites. Los rangos pueden ser de un mínimo de tres a un máximo de 30 sujetos de datos o unidades de observación. Esto refleja el apetito del riesgo de los dueños de los datos y la naturaleza de la información.

A lo largo de este Manual, se asume que el límite se fijará en 10 (es decir $N=10$). Este valor lo usa la Oficina Nacional de Estadística de Inglaterra y ha sido adoptado de manera subsecuente por otras Configuraciones Seguras y departamento de gobierno.

Algunas veces es posible que el resultado estadístico no tiene problemas de confidencialidad aún sin llegar a cumplir el límite. Por el contrario, la regla del límite se puede alcanzar, pero el resultado estadístico puede ser inseguro para entregar. Esto

dependerá de la naturaleza del resultado estadístico: hacer una evaluación del riesgo de cada resultado, es un ejercicio muy valioso. Más que usar un enfoque de rígido de una regla con base en un parámetro (es decir que los resultados no se pueden liberar a menos que se cumpla con el límite), estamos a favor de un enfoque con base en principios, en el cual cada resultado se evalúa por el riesgo potencial de divulgación; y si consideramos que ese riesgo es insignificante, entonces los resultados se entregan.

DOMINANCIA

Es cuando una observación aporta gran parte del valor de la medida estadística, y por ello es identificable. Esto puede aplicar algunas veces a individuos, pero es de mayor preocupación en las estadísticas de negocios donde empresas puede dominar el mercado o sector, o que esté llevando a cabo grandes inversiones en un año particular.

Véase “Proporciones de concentración” para la presentación de un ejemplo.

APETITO DE RIESGO

Las Configuraciones Seguras debe asegurar a los dueños de la información que se mantiene la confidencialidad de sus datos. Esto implica proporcionar la seguridad de que las prácticas de CDE mitigan varios riesgos que se exponen abajo. Los diferentes dueños de información tienen apetito distinto al riesgo, lo cual se refleja, por ejemplo, en cómo evalúan el riesgo de divulgación de los resultados y en el valor de los límites. Este Manual busca definir una serie de lineamientos para poder llevar a cabo la evaluación de riesgo de resultados estadísticos para mantener la confidencialidad de los datos como sea razonablemente posible. Estos lineamientos son flexibles, y los parámetros dentro de los lineamientos se pueden ajustar para adaptarse al apetito de riesgo de cualquier dueño de información.

PERCEPCIÓN PÚBLICA Y RIESGO REPUTACIONAL

Si se publican datos confidenciales, se dañan la confianza del público y la reputación del dueño de la información. También puede conllevar sanciones penales y se puede inhibir la operación del dueño de la información que puede ser una institución pública que provea un servicio. Una frecuencia pequeña, validada contra riesgo, puede resultar riesgosa para el público, que puede cuestionar porque esos resultados se publican, aún si el riesgo es muy pequeño o no lo hay de quebrantar la confidencialidad de los datos.

RIESGO DE PRIVACIDAD

Existe el riesgo que se entreguen resultados de Configuraciones Seguras que lleven a la identificación del sujeto de datos o unidad de observación. También se puede entregar información confidencial del sujeto de datos o unidad de observación. Esto puede conllevar a daño y una experiencia angustiosa para el sujeto de los datos o de gente asociada de manera cercana a él o ellos. También puede haber consecuencias legales.

Los riesgos estadísticos: principio y reglas

EN ESTA SECCIÓN APRENDEREMOS SOBRE:

1. CDE basado en reglas versus basado en principios
2. Cómo este Manual aborda el CDE

.....

El “riesgo estadístico” trata sobre el mínimo de información estadísticas que se requiere para revelar la identidad de los datos individuales de una unidad de observación y/o información confidencial sobre esta. CDE trata de prevenir que esto ocurra, lo más que sea razonablemente posible, mientras que balancea riesgo de privacidad y el beneficio público de publicar cierta información estadística generada de datos confidenciales.

De manera general, hay dos escuelas de pensamiento de cómo llevar a cabo el CDE:

- CDE basado en reglas;
- CDE basado en principios.

Muchos servicios adoptan un enfoque basado en reglas, que establece que, dado un conjunto de reglas fijas que establecen que se puede y que no se puede entregar, el resultado estadístico que presenta el usuario cubre el criterio, (si es el caso, la decisión es que se entrega), si no (no se entrega). Por ejemplo, si un límite de frecuencias que se presenta en un tabulado es fijo, como ninguna celda en el tabulado puede contener frecuencias de menos de 10, entonces el régimen basado en reglas permite que el resultado se entregue si todas las frecuencias cumplen con este límite; no se entregará si alguna de las frecuencias del tabulado no cumple con el límite.

El método basado en principios implica preguntar si dada una evaluación de riesgo, ¿el resultado estadístico presentado es seguro para poder liberarlo o no? (De acuerdo con el elemento “Resultado Seguro” del modelo de los Cinco Seguros). Podría ser que en un tabulado de frecuencias que presenta el usuario, ciertas celdas no cubren el mínimo del límite establecido; pero si se toman en cuenta cierto número de factor, se decide en la evaluación de riesgo que es adecuado entregar el tabulado.

Un ejemplo puede incluir frecuencias de incidencias de cáncer por tipo de cáncer para todo el Reino Unido. Las celdas con valores de 2 o 3 en general se entrega porque si no se tienen

información con más desglose, no tampoco otra información, es casi imposible poder saber a quiénes refieren esos 2 o 3 casos.

Por el contrario, podría haber un ejemplo donde, aun cuando las frecuencias cumplan con los límites, podría ser problemático publicar los resultados por el riesgo de divulgación.

Para más información, Ritchie and Elliot (2015), han escrito sobre este tema (véase Lecturas sugeridas)

¿Cuáles son las ventajas de los dos enfoques? El basado en reglas es fácil de seguir, incluso sin tener que pensar. Por el otro lado, algunos resultados estadísticos no se van a entregar aun cuando fuera seguro hacerlo, lo que impacta de manera negativa a los esfuerzos de investigación. Además, como se sugiere arriba, puede proveer de un sentido de seguridad falso, que puede hacer que se entreguen estadísticas que no son seguras, solo porque cumplieron con el límite.

El enfoque basado en principios requiere mayor compromiso y requiere más tiempo y habilidades para poderlo aplicar. El revisor de resultados tiene que tomar en cuenta un número de factores (contexto de la investigación, por ejemplo, el sector industrial, ocupación, etc.) así como considerar la información que ya está en el dominio público que podría combinarse y causar divulgación de confidencialidad. La justificación por razones estadísticas para entregar puede ser más fuerte que simplemente que “los resultados cumplieron con la regla”. Pero posiblemente, este enfoque lleve a mejores resultados de investigación, ya que potencialmente se pueden entregar de manera segura más resultados estadísticos.

En este Manual no hemos hecho una elección a favor de ninguno de los dos enfoques, sino que ofrecemos asesoría de cómo hacer una evaluación de riesgo. Potencialmente esto está más cercano al enfoque basado en principios, dado que ayuda a los revisores de resultados a realizar un juicio ponderado, usando un número de factores para evaluar si se entrega o no el resultado estadístico. En contraste, podríamos haber simplemente estipulado reglas a seguir, pero creemos que no se debe hacer esto, en el interés de promover investigaciones robustas que usan fuentes de datos con información confidencial.

LECTURAS SUGERIDAS

1. Ritchie and Elliot (2015) “Principles-versus rules-based output Statistical Disclosure Control in remote access environments”, IASSIST Quarterly v39 pp5-13
2. Desai, T., Ritchie, F., Welpton, R.
2. “Five Safes: designing data access for research”, University of West of England Working Paper Series, <http://eprints.uwe.ac.uk/28124/1/1601.pdf>

SECCIÓN B

¿Cómo evaluar resultados estadísticos?

Introducción: CDE

EN ESTA SECCIÓN APRENDERÁ SOBRE:

1. ¿Cómo llevar a cabo una evaluación de riesgo de Control de Divulgación Estadística para cierto tipo de resultados estadísticos?
2. ¿Qué se les debe preguntar a los analistas cuando solicitan que se les entreguen resultados?

.....

Esta sección presenta los tipos más comunes de resultados. Para cada uno, se describe un ejemplo numérico, con asesoría de cómo evaluar el riesgo de divulgación estadística. En un recuadro se resume la información mínima que se requiere para evaluar el resultado. Otras consideraciones adicionales sobre divulgación secundaria pueden identificarse con base en las experiencias de los autores.

Es imposible cubrir todos los tipos de resultados estadísticos que pueden producir los analistas: El Control de Divulgación Estadística no se puede limitar solo a la siguiente sección.

Usando la guía

Cada una de las siguientes páginas se dedican a un resultado estadístico particular. Se presenta una explicación del resultado estadístico.

.....

A esto le sigue una descripción de los aspectos a considerar en el CDE, una “regla de oro” para revisar los resultados y técnicas que pueden aplicarse para proteger la confidencialidad de los datos usados para crear los resultados estadísticos. Para los casos donde no sea necesario o posible alterar los resultados, no hemos incluido una sección sobre Reducción de Riesgo de Divulgación. En cambio, recomendamos seguir los lineamientos que contiene esta sección – por ejemplo, asegurarse que las estadísticas se basan en un número suficiente de observaciones.

Estas guías no son recetas, y no deben seguirse de manera ciega por los revisores de resultados. En algunos casos, puede ser perfectamente aceptable entregar los resultados estadísticos aún que no se cumpla la “regla de oro”, porque el revisor de resultados llega a la conclusión de que es poco probable el riesgo de divulgación. En cambio, estos lineamientos están diseñados para promover la identificación del riesgo probable, y de hacer las preguntas que un revisor de confidencialidad podría hacerle al analista.

También tratamos de implementar un enfoque de semáforos:

No entregar a menos que se este absolutamente cierto que es seguro, sino no entregarlo

Puede ser seguro entregarlo, si se aplican algunas técnicas de CDE

En general se puede entregar, sin gran evaluación

Estadísticas descriptivas

REQUERIMIENTOS MÍNIMOS:

- Especificación de la cohorte
- Información contextual (por ejemplo: datos de negocios, sector, empresas, periodo de tiempo)

Tabla 3:
Tabla de frecuencias

ANTECEDENTES ÉTNICO/EDAD	16	17	18	19	20
Africano Asiático	0	0	0	0	0
Bangladesh	3	4	6	5	4
Africano Negro	3	5	7	6	8
Caribe Indio Occidental	0	4	2	4	3
Chino	0	2	1	1	0
Lejano Oriente	1	0	1	2	0
Indio	5	9	4	4	4
Medio Oriente	1	1	0	0	1
Mezcla Caribe Hindú Occidental	0	0	1	0	2
Mezcla Hindú	0	0	0	0	0
África del Norte	1	0	1	0	1
Pakistán	6	10	5	4	3
Sri Lanka	0	0	2	0	0
Turcos	1	1	1	0	1
Blancas	54	135	141	146	130

Considere agrupar
columnas o filas

Títulos
comprensibles

Frecuencias con
celdas pequeñas

Las estadísticas descriptivas se pueden presentar en maneras diferentes (por ejemplo: tablas, histogramas, gráficas de pie, gráfica de barras). En general incluyen una medida para el centro de la distribución (media, mediana o moda). Una medida de la forma de la distribución (varianza, desviación estándar, simetría y curtosis), y una medida de que tantas veces se observan una combinación de atributos (frecuencias, frecuencias relativas y sumas).

Tabla 2:
Otros estadísticos descriptivo

	MÍNIMO	1 ^{er} CUARTIL	MEDIANA	MEDIA	3 ^{er} CUARTIL	MÁXIMO
Ingreso	0	2894	6046	7178	10350	41113
Edad	16	30	41	43	54	111
Ventas	140	3345993	7100730	8327156	12034046	38333022

Considere redondear o promediar valores por ejemplo de 10 observaciones

Probablemente información precisa de una observación

CONSIDERACIÓN DE CDE

Considere el grado de detalle en el resultado. Por ejemplo, si un tabulado simplemente presenta el desplegado por edad y genero (tabulado de un solo cruce), o si el resultado presenta un tabulado de varios cruces de variables por rango – por ejemplo, edad y sexo presentado por rango de nivel de ingreso y área geográfica (tabulado de cuatro cruces) – entonces el nivel de detalle se incrementa y puede ocurrir la atribución (cuando un individuo es asociado a valores (datos confidenciales) y puede ser identificado).

REGLA DE ORO

Todas las frecuencias, ya sea que se presenten como sumas o frecuencias relativas, debe alcanzar o superar el valor del límite “N”. De manera similar, la media, mediana y moda, desviación estándar y varianza de la distribución/variable debe presentarse con base en al menos “N” observaciones.

REDUCIENDO EL RIEGO DE DIVULGACIÓN

El agregar es la forma más simple de mitigar el riesgo asociado a un resultado; otras técnicas incluyen:

Juntar/combinar columnas y filas:

Promediar valores por ejemplo de 10 observaciones en un grupo;

Redondear valores;

Suprimir celdas (considere que en general se tienen que suprimir dos celdas en una fila/columna, sino es posible deducir los valores originales).

Percentiles

REQUERIMIENTO MÍNIMO:

- Frecuencia de cada percentil

Un percentil indica el valor al cual o por debajo del por ciento en el cual la observación cae. El percentil 50º representa la mediana (centro) de la distribución. Los analistas también pueden presentar deciles (que dividen la distribución en 10 partes iguales), quintiles (cinco partes iguales) y/o cuartiles (cuatro partes iguales).

La Tabla 3 muestra que el 95 por ciento de los hombres que viven en el Reino Unido tiene un ingreso neto anual de £17,734 o menos, y los cinco por ciento superiores tienen un ingreso mayor de £17,734 (sin especificar periodo de tiempo).

CONSIDERACIONES DE CDE

Los resultados pueden ser reveladores porque los valores de los ingresos para cada percentil se presentan de manera precisa (es decir no están redondeados), y por ello es probable que presente el ingreso de alguna persona en específico, o el promedio de un grupo reducido de individuos en el grupo del percentil en el cual, combinado con otras características, una observación puede ser identificada.

Dependiendo del contexto del análisis (por ejemplo: población, método de muestreo) los sujetos de los datos o unidades de observación pueden ser fácilmente identificables y atribuirles el dato del ingreso a partir del tabulado.

REGLA DE ORO

Ninguna celda debe contener menos de “N” observaciones. Revise que el número de observaciones que está detrás de cada valor de ingreso cumple con el límite de la regla de oro.

REDUCIENDO EL RIESGO DE DIVULGACIÓN

Redondee los valores (como en la Tabla 4, por ejemplo, al centenar o millar más cercano). Esto aplica a todos los valores de los percentiles, incluyendo la mediana, y es consistente con el enfoque recomendado para valores mínimos y máximos.

Si el analista desea mostrar la distribución del ingreso, hay otras opciones que se incluyen:

- Agregación/agrupamiento de los valores en categorías (por ejemplo: menos de £12000) para asegurar que se cumple el límite de la regla de oro. Los resultados

pueden desplegarse en un tabulado o en forma de gráfica (tome en cuenta que a los resultados producidos de manera similar se les tendrá que aplicar la técnica de redondeo, para evitar la diferenciación);

- Presentar los rangos de inter-cuartiles.

Tabla 3:
Percentiles sin CDC

Percentil	Ingreso Neto Semanal %
1	99
5	554
10	1129
20	2346
30	3464
40	453
50	6046
60	7555
70	9287
80	11473
90	14750
95	17734
99	22997

Sin frecuencias

La mediana puede referir a una sola observación por lo que vale la pena revisar

Títulos comprensibles

Tabla 4:
Percentiles con CDC

Percentil	Ingreso Neto Semanal (%)	N
1	100	10
5	600	15
10	1100	17
20	2300	11
30	3500	10
40	5000	10
50	6000	15
60	7600	19
70	9300	12
80	11500	10
90	15000	11
95	17800	16
99	23000	12

Datos redondeados, es menos probable que se pueda atribuir a un individuo específico

Proporciona frecuencias

Histogramas

REQUERIMIENTOS MÍNIMOS:

- Tabla de frecuencias subyacente
- Etiquetas para los ejes y las variables y título
- Detalle de los sujetos de los datos y sumas totales

.....

Un histograma muestra la distribución de la frecuencia de una variable. El ancho de las barras o bien representa los intervalos de la clase o un solo valor, y la altura representa la densidad de la frecuencia. Una gráfica de densidad muestra la distribución de los datos sobre el continuo o una variable discreta y suele usar para mostrar la forma de la distribución de una variable específica.

CONSIDERACIONES DE CDE

Las gráficas deben ser sujetas a la misma metodología de CDE que un tabulado estadístico. Idealmente la manera más sencilla de revisar una gráfica es aplicar CDE a la tabla subyacente.

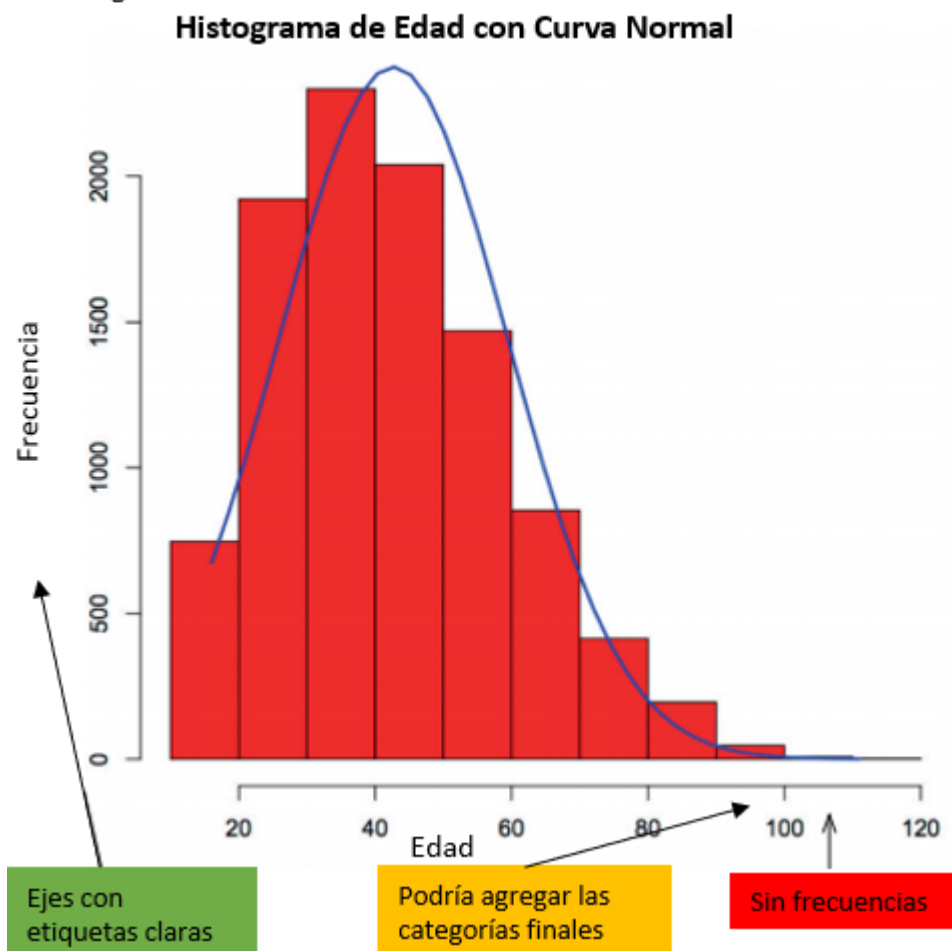
Los aspectos comunes de CDE de histogramas y gráficos de densidad derivan de celdas con sumas pequeñas (menores al “N” número de sujetos de datos o unidades de observación) y los valores máximos y mínimos. Como los histogramas y las gráficas de dispersión se usan para mostrar la distribución del valor, los valores pequeños son siempre un asunto, especialmente en las colas. Mínimos y máximos generalmente quedan enmascarados en la escala del eje de las X en los puntos inicial y final. Muchos programas estadísticos (por ejemplo, Stata) presenta por default al valor máximo como último punto del valor del eje.

Las gráficas deben entregarse como imágenes fijas (por ejemplo, formato PNG), como muchos softwares estadísticos (por ejemplo, Stata) pueden guardar los datos de una gráfica. Si los analistas requieren recrear la gráfica en un diseño o formato, deben usar las frecuencias subyacentes una vez que han sido revisadas para riesgo de divulgación.

REGLA DE ORO

Aplica la misma regla que la de los tabulados de frecuencias, y mínimos y máximos.

Figura 1:
Histograma



REDUCIENDO EL RIESGO DE DIVULGACIÓN

Es importante que el analista especifique el propósito de su gráfica, ya que la mejor opción dependerá del significado detrás del resultado. Si tanto la gráfica como la tabla de frecuencias se entregan, las acciones de mitigación deben aplicarse a ambas.

Si la gráfica tiene la intención de mostrar que la distribución tiene una cola muy larga (por ejemplo, que hay muchos valores atípicos) entonces el analista debe agrupar todos estos valores en una sola clase. Este enfoque enmascara los valores máximos y mínimos. Esto no es el caso cuando los mínimos y máximos son valores estructurales o están definidos por el analista. En el histograma que se presenta, el mínimo estructural es cero, como no puede haber un valor menor para la edad, no es necesario encubrirlo. Lo mismo se aplicaría si los individuos fueran seleccionados de acuerdo con un rango de edades (es decir, solo individuos menores de 65 años de edad), con ello el límite superior coincidiría con el valor máximo de 65.

Si las sumas se distribuyen fuera de la cola, podría ser una buena solución agregar la variable en el eje de las X en un grupo de clases, en la cual cada clase tuviera un número suficiente de sujetos de datos o unidades de observación para evitar la re-identificación.

Si el objetivo es mostrar la forma de la distribución probabilística, es posible mantener toda la gráfica si se omiten los valores en el eje de las X. Con esta solución, es posible relajar las reglas de los límites para las sumas pequeñas, y los mínimos y máximos, porque no sería posible asociar un valor a una barra específica o a un punto en la gráfica.

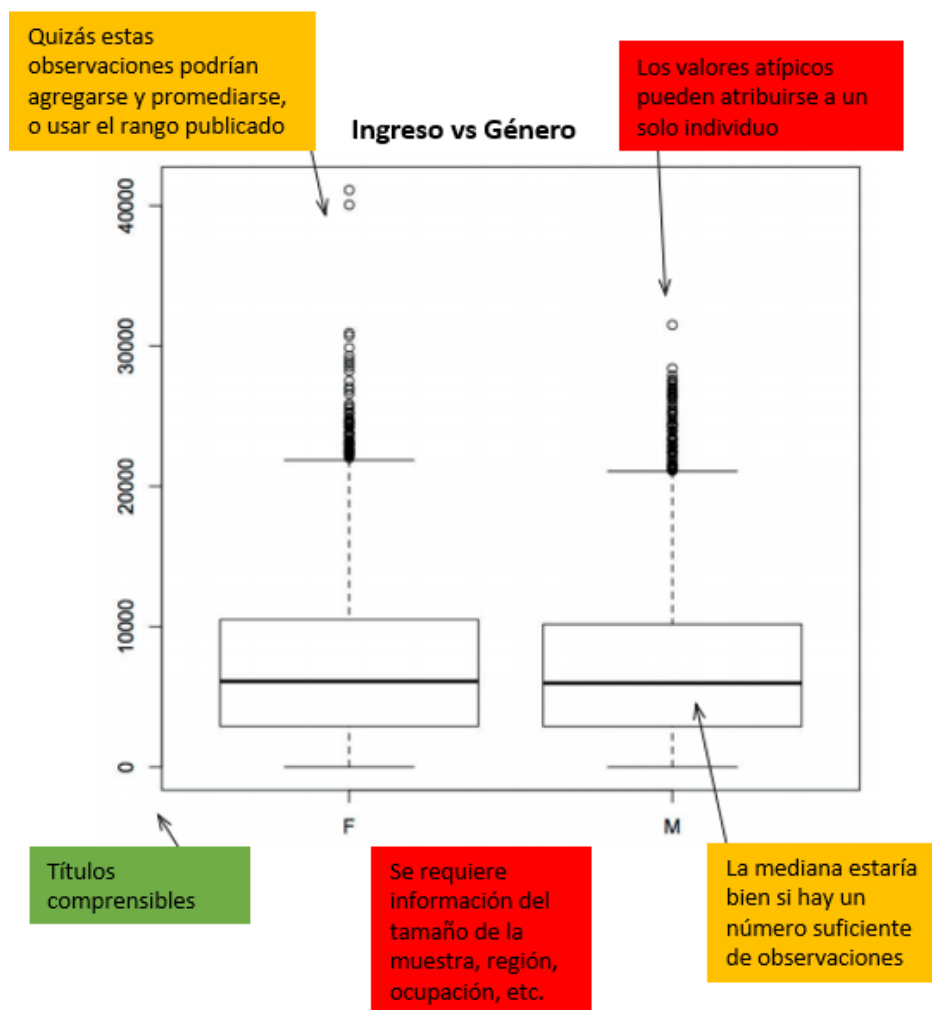
Con lo antes dicho, nótese que se puede hacer reingeniería inversa con la escala si solo un par de estadísticos descriptivos (por ejemplo, la mediana y rangos inter-cuartiles son publicados en otro lugar.

Diagrama de caja

REQUERIMIENTOS MÍNIMOS:

- Información de la fuente de datos y cohorte
- Etiquetas de los ejes y títulos
- Breve descripción del diagrama de caja.

Figura 2:
Diagrama de caja



Un diagrama de caja (algunas veces conocida como “caja y bigote”) es una gráfica que representa como los datos están distribuidos. Típicamente, un diagrama de caja tiene:

- “Colas” en alguno de los dos extremos del gráfico, representando los valores máximos y mínimos de la distribución;
- El rango de inter-cuartiles (del 75º al 25º percentil), representado por la “caja”;
- El valor de la mediana.

El diagrama de caja de la Gráfica 2 muestra la distribución del ingreso de hombres y de mujeres de la misma fuente de datos. El mínimo y el máximo del salario por hora están representados por las colas de cada gráfica; con los rangos de inter-cuartiles representados en los rangos en las cajas. La gráfica muestra que la distribución de salarios de los hombres es más “amplia”, es decir el salario mínimo es menor, pero el salario máximo es mayor, para las mujeres que para los hombres.

CONSIDERACIONES DE CDE

Las colas son los valores mínimo y máximo. Si estos valores refieren a una sola observación, la entrega del diagrama de caja podría divulgar información confidencial sobre los individuos.

REGLA DE ORO

Entregue si el analista puede demostrar que los valores de las colas no se pueden atribuir a la observación de un solo individuo del conjunto de datos.

REDUCIENDO EL RIESGO DE DIVULGACIÓN

Los valores mínimos y máximos pueden ser agrupados en un promedio. Por ejemplo, en lugar de desplegar la cola como un valor mínimo de salario por hora de £7.40 por hora, agrúpelo en un rango de £7 -8 por hora (siempre y cuando en el rango se cumpla con el límite de número de observaciones, por ejemplo 10).

Coeficientes de correlación

REQUERIMIENTOS MÍNIMOS:

- Etiquetas de variables
- Número de observaciones

.....

Un coeficiente de correlación indica la fuerza y la dirección de la relación lineal entre dos variables.

Los coeficientes de correlación tienen un rango de -1 a 1, el -1 indica una relación lineal negativa perfecta, 0 no hay relación lineal y 1 indica una relación lineal positiva perfecta.

Los coeficientes de correlación normalmente se despliegan en una tabla (por ejemplo, Tabla 5).

CONSIDERACIONES DE CDE

En general los coeficientes de correlación se consideran seguros siempre y cuando se cumpla con el límite de “N” observaciones.

EL riesgo puede incrementarse en casos en los que el coeficiente de correlación es exactamente igual 1(-/+) y también se presentan en los resultados estadísticas descriptivas como la mediana, percentiles y, mínimos y máximos. Por ejemplo, si hay una relación lineal positiva perfecta (=1) entre ventas y costos de empleo para una muestra de empresas, y también se presenta el valor de la mediana para cada variable, esos valores de las variables se relacionan a una empresa, dado que la relación entre las dos variables tiene una correlación perfecta. Esto incrementaría el riesgo de que una empresa se pueda identificar y se pueda asociar información confidencial es ésta.

REGLA DE ORO

Un coeficiente de correlación se tiene que derivar de al menos “N” observaciones.

REDUCIENDO EL RIESGO DE DIVULGACIÓN

Raramente los coeficientes de correlación son iguales +/-1, ya que las variables tienden a no estar perfectamente correlacionadas. Cuando la correlación es igual a +/-1, y se presentan un resumen de estadísticas, considere el suprimir o cambiar el coeficiente de correlación (por ejemplo, =>0.80).

Tabla 5:
Coeficientes de correlación

Títulos claros
de etiquetas

	Ventas	Salarios	Incremento total de inventarios	Importaciones
Ventas	1.00	0.54	0.33	0.03
Salarios	0.54	1.00	0.31	0.09
Incremento total de inventarios	0.33	0.31	1.00	0.19
Importaciones	0.03	0.09	0.19	1.00

El número de observaciones subyacente a cada coeficiente = 10, 789

Se proporciona el número
de observaciones

Análisis de factores

REQUERIMIENTOS MÍNIMOS:

- Etiquetas de variables
- Número de observaciones
- Breve descripción del análisis

El análisis de factores es una técnica para identificar factores que explique la interrelación entre los datos. Se lleva a cabo mediante la creación de un nuevo conjunto de datos agregados: estos se basan en varios elementos de la base de datos original. El número de nuevos datos es menor que el número de datos originales.

El análisis de factores se basa en el supuesto de que hay un número de factores que explican la correlación entre los elementos de los datos originales. Si los factores se mantienen constantes, la correlación parcial entre los elementos de los datos observados es cero, y por ello, los factores explican los valores de los elementos de datos observados.

Supongamos que tenemos una base de datos de desempeño escolar de los alumnos. Llevaremos a cabo un análisis de factores de la base de datos, mirando las correlaciones entre los elementos observados de los datos y creando dos factores:

Tabla 6:
Tabla de carga de factores

VARIABLE	FACTOR 1	FACTOR 2
Calificación en Aritmética	0.40	0.89
Calificación en Álgebra	0.56	0.81
Calificación en Lógica	0.46	0.62
Calificación en Vocabulario	0.44	0.09
Calificación en Lectura	0.50	0.12

Estos valores no son correlaciones, son carga de factores.

Comery and Lee (1992) sugieren que la carga de factores cuando es mayor a 0.71 indica que más del 50% de la varianza se explica por el factor. Podemos interpretar de estos resultados que el Factor 2 se relaciona con habilidades matemáticas.

CONSIDERACIONES DE CDE

Por la metodología que involucra el llevar a cabo análisis de factores, es poco probable que resulte problemático la divulgación de datos confidenciales o la identificación de individuos. Considere que, en el ejemplo, el Factor 2, relacionado con competencia matemática, se relacionara con un alumno de la escuela. Esto no podría pasar a menos que el dato refiera a un número muy pequeño de observaciones, porque de lo contrario no se podría establecer una correlación. No se puede llevar a cabo un análisis de factores sin una muestra grande, con certeza por encima del límite que haya establecido el proveedor de los datos.

REGLA DE ORO

La guía de Datos sin Fronteras sugiere que, como regla de oro, el factor debe integrarse con al menos dos variables. Esto puede evitar cualquier correlación directa entre un factor y un individuo, como es poco probable que esto sea así con un número suficiente de observaciones en los datos.

Índices

REQUERIMIENTOS MÍNIMOS:

- Número de observaciones
- Descripción de su construcción
- Interpretación

Los índices proveen estadísticas agregadas útiles para describir los datos y varían mucho en cómo se construyen. Los economistas pueden producir un índice de precios, para describir la fluctuación de los precios de los bienes o servicios en un periodo de tiempo. Un investigador de ciencias sociales o un epidemiólogo pueden producir índices para mediar la equidad en el acceso a los servicios de salud en las distintas áreas del país.

CONSIDERACIONES DE CDE

Un rango se puede considera que es un índice, pero es una simple deducción del valor mínimo de un dato del máximo valor. Si se basa en un número pequeño de sujeto de datos o unidades de observación, el riesgo de divulgación sería más alto a que si se usan muchas observaciones.

Rango = Max – Min

Los lineamientos para evaluar la divulgación estadística serían los mismos que para evaluar una tabla de frecuencias, es decir los valores deben estar basados en un número suficientes de observaciones a menos de que se pueda demostrar de una manera razonable que es poco probable la divulgación de información confidencial y o riesgo de identificación.

En contraste, un índice más complejo se puede crear. Por ejemplo, tomadas de las guías 2010 de ESSnet, reproducimos el Índice de Precios Fisher.

Figura 3:

Índice de Precios Fisher $PF = \sqrt{PL \cdot Pp} = \sqrt{\frac{\sum_{j=1}^m p_{1,j} \cdot q_{0,j}}{\sum_{j=1}^m p_{0,j} \cdot q_{0,j}} \cdot \frac{\sum_{j=1}^m p_{1,j} \cdot q_{1,j}}{\sum_{j=1}^m p_{0,j} \cdot q_{1,j}}}$

Como se puede observar, la construcción de este es un índice es más compleja, y los datos que usan para producir la estadística se va a haber transformado un número de veces. Derivado de esta transformación, la divulgación estadística es poco probable que sea problemática.

REGLA DE ORO

Cuando se haga una evaluación de CDE, el revisor de resultados debe considerar el grado de la complejidad. Los índices simples pueden ser más problemáticos que los de tipo complejo.

Es esencial contar con información de cómo se produjo el índice para poder hacer la evaluación.

Gráficos de dispersión

REQUERIMIENTOS MÍNIMOS:

- Etiquetas en los ejes y las variables
- Título y descripción de la cohorte

.....

Típicamente despliega los valores de dos variables y cada punto de la gráfica representa un sujeto de dato o unidad de observación, como en la Gráfica 4.

CONSIDERACIONES DE CDE

En general, se considera que los gráficos de dispersión revelan confidencialidad. En el ejemplo anterior, cada punto representa un sujeto de datos o unidad de observación, lo que significa que los ingresos y edad de cada observación de la muestra se pueden ver de manera fácil en la gráfica. El presentar datos a nivel individual hace que sea relativamente sencillo identificar el sujeto del dato o unidad de observación y atribuirle el dato. Por ejemplo, en la Gráfica 4 se aísla al individuo de 35 años con un ingreso superior a £42,000.

Las variables que se presenta aquí (ingreso y edad) están en su formato original, como fueron captadas en la encuesta. No han sido transformadas en ninguna manera y por ello se pueden atribuir a los sujetos de los datos relativamente fácil. Si se transforman los datos lo puede hacer más seguros.

REGLA DE ORO

La regla de oro es que ninguna celda debe contener menos de N observaciones. En este ejemplo, esto no se cumple ya que cada punto de la gráfica representa un sujeto del dato o unidad de observación.

REDUCIENDO EL RIESGO DE DIVULGACIÓN

Una opción sería agrupar los sujetos de los datos, para asegurarse que las estadísticas que se presentan se basan en un número suficiente de observaciones. En la Tabla 7 se muestra el promedio de ingreso por hora para tres grupos de edad, el número de observaciones de cada grupo de edad cumple con el límite de la regla de oro.

Figura 4:
Gráfico de dispersión

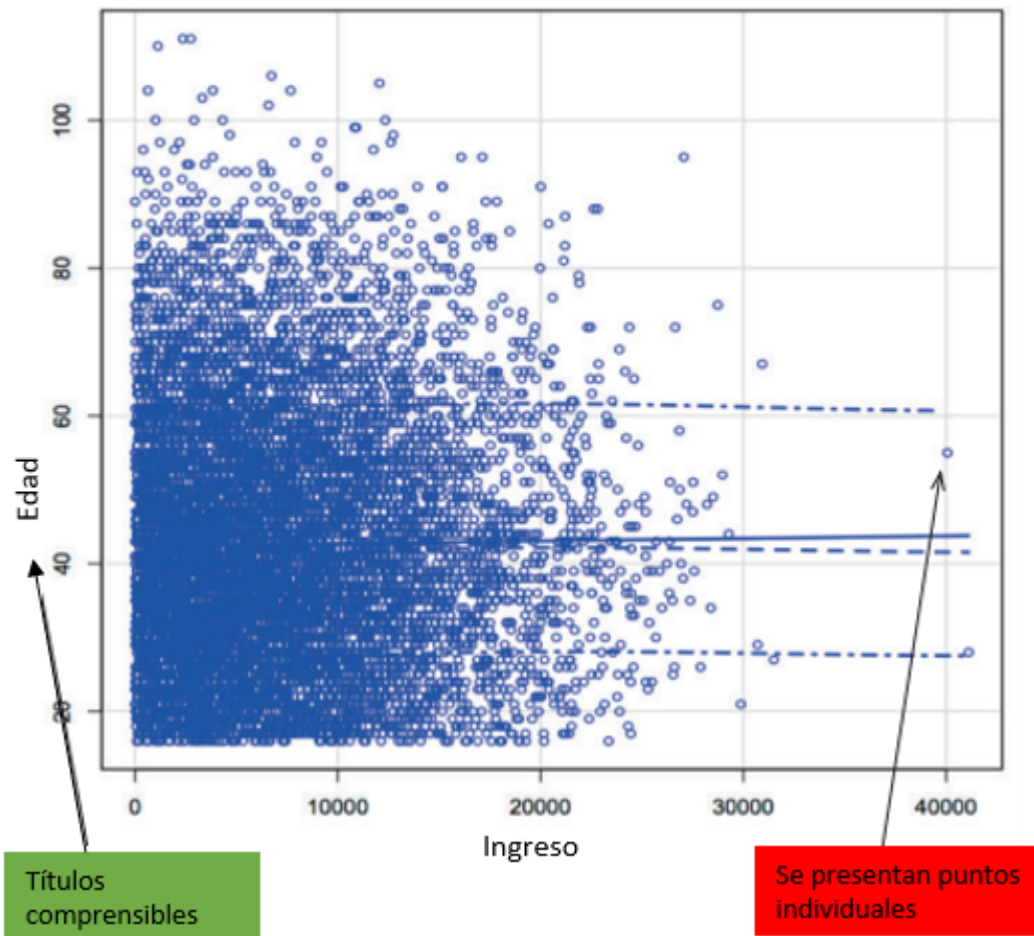


Tabla 7:
Datos modificados para nueva gráfica de dispersión

Se podría modificar los datos originales y crear un gráfico de dispersión

GRUPO DE EDAD	INGRESO POR HORA (£)	N
16-24	6.10	24
25-34	8.20	20
35-44	8.90	18
45-55	8.80	22
55 o más	10.20	14
Total		98

Diagramas de simetría

REQUERIMIENTOS MÍNIMOS:

- Sumas y una explicación clara de lo que cada punto representa
-

Un diagrama de simetría es una técnica gráfica para evaluar si una variable presenta una distribución simétrica. Los valores de la muestra se ordenan de menor a mayor. El diagrama grafica la distancia entre el valor más grande y la mediana contra la distancia del valor más pequeño y la mediana. Esto se repite para el segundo valor más grande y el segundo valor más pequeño, luego el tercer valor más grande y el tercer valor más pequeño, y así sucesivamente hasta que todos los pares son diagramados. Si la variable tiene una distribución simétrica todos los puntos se mantendrán dentro de la línea de referencia (definida como $y = x$).

En la gráfica 5, se diagraman los datos del ingreso de una muestra de 50 mujeres. La que gana más, tiene un ingreso de £15, 780 más alto que la mediana, contra la que gana menos con un ingreso de £10,230 menor que la mediana.

CONSIDERACIONES DE CDE

En este ejemplo, cada punto, grafica la distancia de dos sujetos de datos o unidades de observación a la mediana. Si se conoce el valor de ingreso de la mediana entonces los valores de los ingresos de cada sujeto de datos o unidad de observación se pueden calcular de cada grupo de pares.

Si hay valores atípicos en la muestra, puede ser relativamente fácil identificar el sujeto del dato o unidad de observación y atribuirle el dato (si se conoce la mediana). La variable del ingreso que se despliega en este ejemplo está en su formato original, como se captó en la encuesta. No ha sido transformada de ninguna manera y como resultado de ello cada valor se relacionan con un sujeto de datos o unidad de observación.

REGLA DE ORO

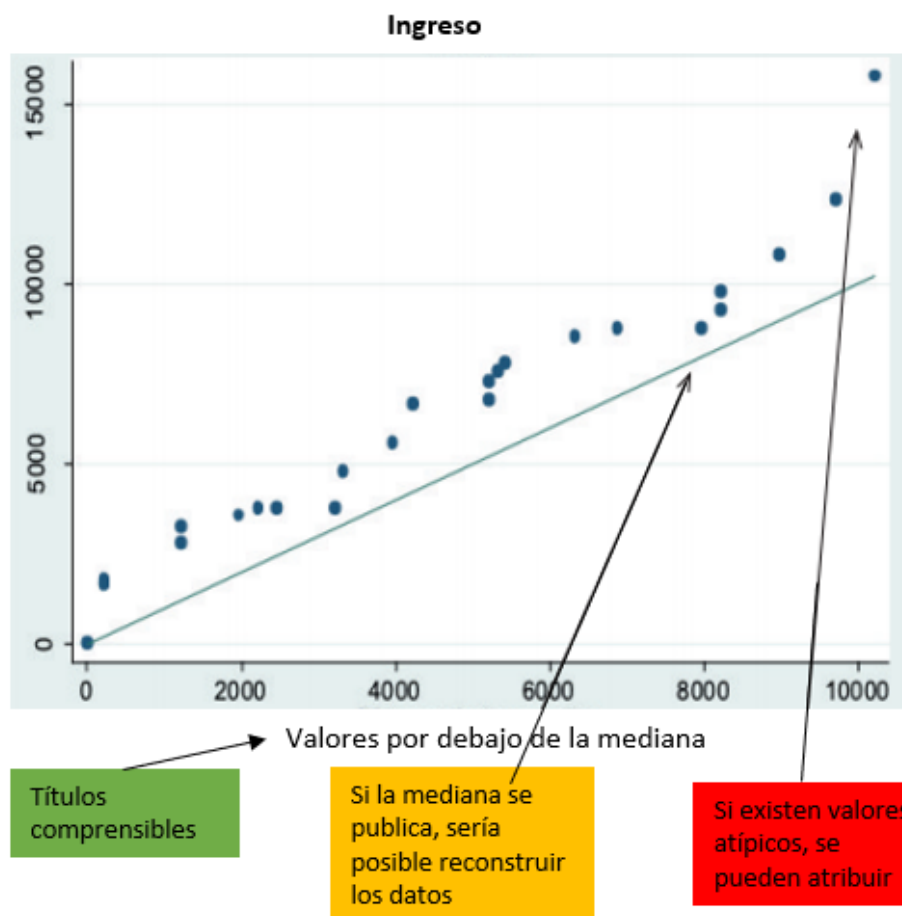
Ninguna celda debe contener menos de 'N' observaciones. En este ejemplo, esta regla de oro no se cumple ya que cada punto de la gráfica representa dos sujetos de datos y los valores de cada sujeto de datos o unidad de observación pueden ser calculados (si se conoce la mediana).

REDUCIENDO EL RIESGO DE DIVULGACIÓN

Se pueden redondear los valores o agregarles ruido. Stata proporciona una función “jitter” que agrega ruido aleatorio a los datos antes de diagramarlos, con la opción de especificar la magnitud del ruido como porcentaje del área de la gráfica. La Gráfica 6 muestra diagramas de dispersión en el cual los datos de ingreso han sido redondeados y se les ha agregado ruido aleatorio.

La estandarización es un método que transforma los valores para que tenga una media de cero y una desviación estándar de uno. Cuando la variable estandarizada se presenta en un diagrama de simetría, cada valor de la observación de la variable indica su diferencia de la media del valor original en un número de las desviaciones estándar. Sin embargo, esta opción de mitigación no es suficiente: si se presentan en los resultados de la media, la desviación estándar y la mediana; entonces estos valores pueden usarse para revertir la transformación y calcular el valor de la variable para cada sujeto de datos o unidad de observación.

Figura 5:
Diagrama simetría sin valores alterados



Si hay un sujeto de dato o unidad de observación con un valor extremo (por ejemplo, en particular alguien del alto ingreso en la Figura 5), entonces la gráfica de simetría puede hacerlo ver en los resultados. Si bien no podemos calcular el valor del monto del ingreso para el sujeto del dato o unidad de observación, el diagrama nos permite identificar a un sujeto de dato o unidad de observación que es problemático. Se tienen que hacer una evaluación del riesgo al diagrama de simetría o a una gráfica similar (por ejemplo, el histograma) suprimiendo las etiquetas de los ejes.

Figura 6:
Diagrama de simetría (después de aplicar CDC)

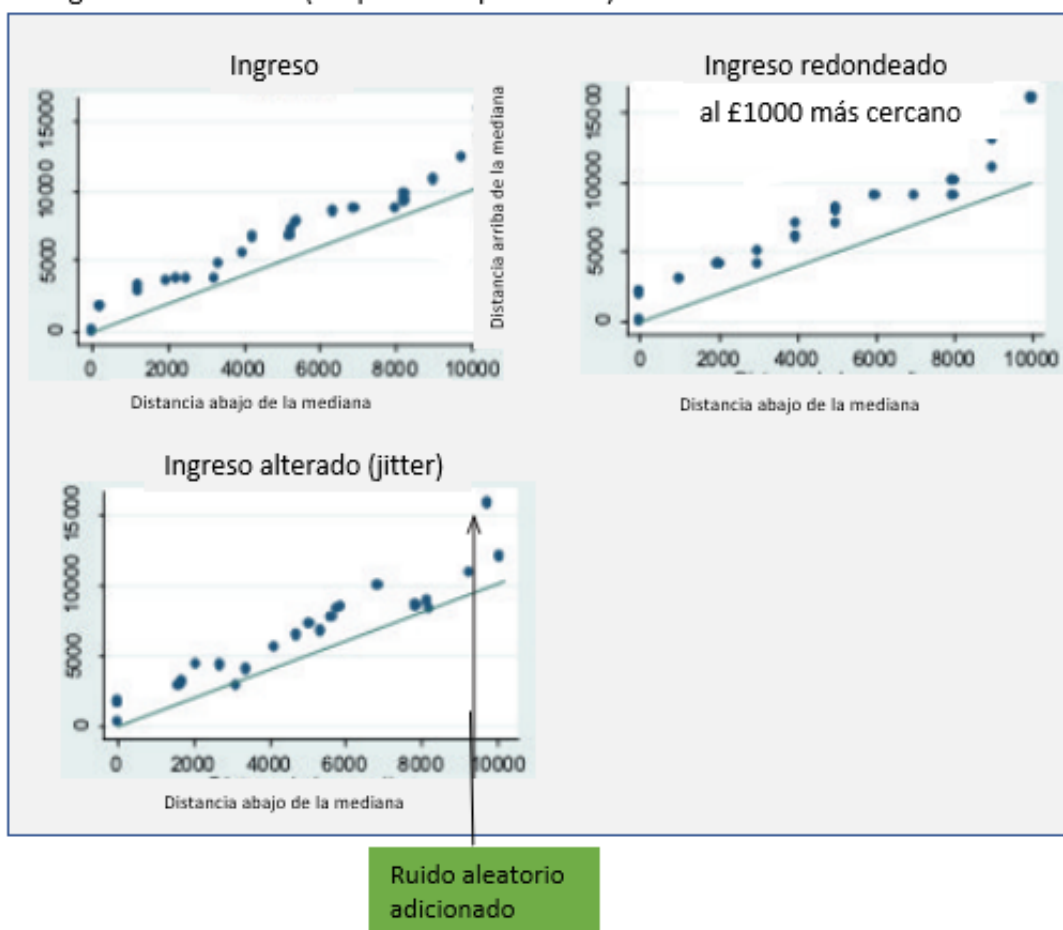


Diagrama de simetría de ingreso de mujeres. El eje Y (distancia por arriba de la mediana) = $Y(n - i + 1)$ mediana y eje X (distancia por debajo de la mediana) = $-Y(i)$ en donde la mediana es la mediana de la muestra, Y es la variable de la muestra, y Y (i) indica el orden i^o – orden estadístico de Y.

Árboles de decisiones y criterio de exclusión

REQUERIMIENTOS MÍNIMOS:

- Número de sujeto de datos o unidades de observación y remanente en cada etapa
- Número de sujeto de datos o unidades de observación descartados en cada etapa
- Especificaciones de cohortes

.....

Un árbol de decisión es como un diagrama de flujo, en el cual se incluyen pruebas estadísticas en cada etapa para evaluar la significancia estadística de las diferencias entre los flujos.

El modelaje con árboles de decisiones en muchas ocasiones se basa en un proceso automático para seleccionar el orden en que se incluyen las variables en el diagrama de flujo de datos. En general, las variables se eligen de tal manera, que las variables que explican el monto mayor de la varianza se incluyen primero en el modelo. Sin embargo, en algunos casos, el orden de las variables puede determinarse de manera a priori.

CONSIDERACIONES DE CDE

Puede ser potencialmente riesgoso de divulgar confidencialidad si los criterios de exclusión se aplican a grupos pequeños de sujetos de datos o unidades de observación, por ejemplo, si se usan criterios muy específicos, o la población es pequeña (ver Figura 7 en el que 4 sujetos de datos se excluyen en el paso 3).

Se puede presentar potencialmente identificación espontánea de un sujeto de datos o unidad de observación, o incluso que, la atribución de características no conocida si la población a elegir está bien definida.

Por ejemplo, en la Figura 7 nosotros sabemos de antemano que los sujetos de datos o unidades de observación son pacientes de la tercera edad que están en casas de descanso en un área específica, y en el paso 3, también podemos notar que una casa de descanso ha

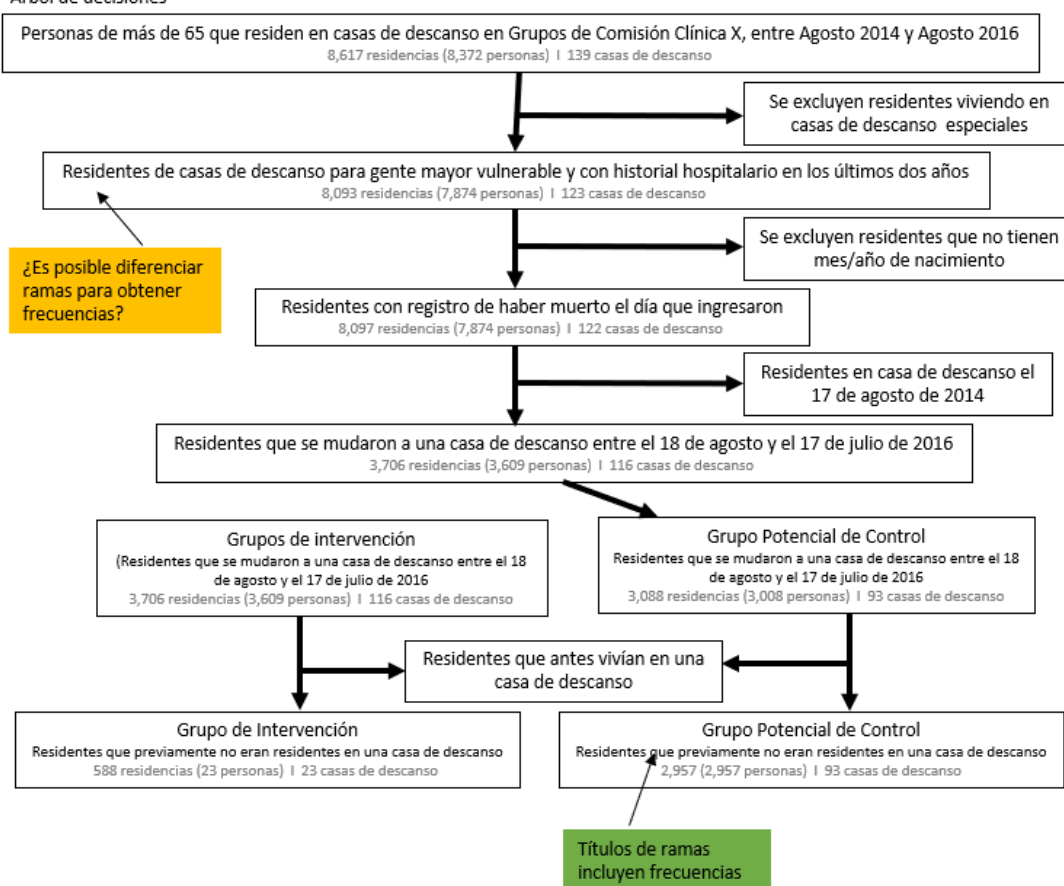
sido excluida, lo que sugiere que los cuatro pacientes se encontraban en la misma casa de descanso (atribución).

Se puede presentar divulgación secundaria cuando se generan las mismas estadísticas más de una vez en subgrupos similares de la misma población. Por ejemplo, si el resultado en la Figura 7 se replica en un subgrupo de la población original mediante la exclusión de paciente con más de 85 años de edad, el nuevo resultado puede divulgar confidencialidad, aún si se cumple con la regla de oro de todos los criterios de exclusión. Si la diferencia marginal que se obtiene de comparar los totales de los subgrupos de cada criterio es muy pequeña, se podría saber por ejemplo que un paciente que murió en una de las casas de descanso el día que lo ingresaron y tenía más de 85 años. Junto con otra información de ambos resultados, podría llevar a la re-identificación del paciente.

REGLA DE ORO

El número de sujeto de datos o unidades de observación que se descarta en cada criterio de exclusión debe ser mayor a N.

Figura 7:
Árbol de decisiones



Análisis de supervivencia: curva de Kaplan-Meier

REQUERIMIENTO MÍNIMO:

- Frecuencia en cada etapa

.....

La curva de Kaplan-Meier es una representación gráfica de supervivencia. La supervivencia no necesariamente tiene que relacionarse con mortalidad, puede ser del tiempo de un evento (por ejemplo, el número de días entre que se le dio el alta del hospital y su readmisión).

La gráfica en la figura 8 representa el número de sujetos de datos o que “sobrevivieron” (eje vertical), en el tiempo (eje horizontal). Cada etapa en la gráfica representa un sujeto de datos, o grupo de sujetos de datos, que no sobrevivieron en ese punto del tiempo.

El modelo de riesgo proporcional Cox intenta explicar la tasa aleatoria (tasa inversa de supervivencia) usando variables explicativas a lo largo de la gráfica (vea estadística descriptiva).

CONSIDERACIONES DE CDE

Cada etapa de cambio en la curva de supervivencia debe representar al menos ‘N’ sujetos de datos.

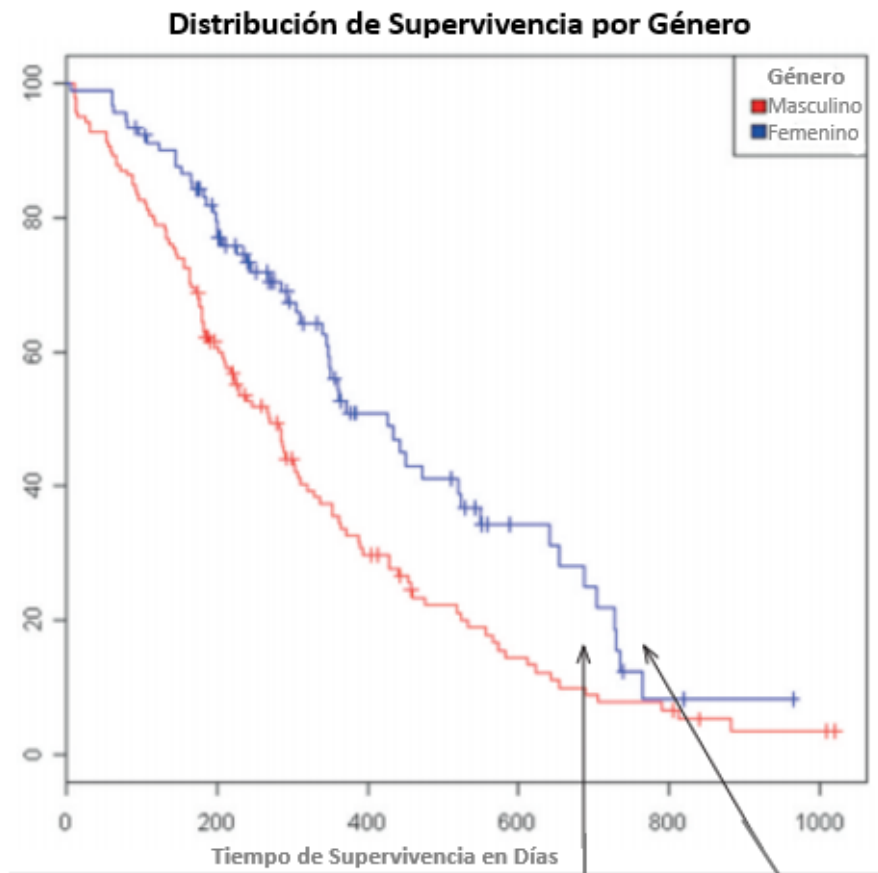
REGLA DE ORO

Cada etapa de cambio en la curva de supervivencia debe representar al menos ‘N’ sujetos de datos.

REDUCIENDO EL RIESGO DE DIVULGACIÓN

Considere agrupar las etapas de cambio, de tal manera que cada etapa en la gráfica represente al menos ‘N’ observaciones, o considere publicar la curva de supervivencia sin incluir escalas en los ejes. Esto es especialmente relevante si se busca ilustrar la diferencia entre dos grupos.

Figura 8:
Curva Kaplan-Meier



Títulos
comprensibles

La descripción y
tamaño total de la
muestra puede
ayudar en la
evaluación

¿Se podrían identificar
un número pequeños
de individuos en el
cambo de etapa?

¿Hay suficiente
número de
observaciones
entre cada etapa
de cambio?

Análisis espacial (mapas)

REQUERIMIENTOS MÍNIMOS:

- Sumas y una explicación clara de lo que cada punto representa

Un mapa de puntos es una forma geográfica de desplegar datos. Por ejemplo, puede presentar la ubicación de negocios o servicios públicos.

Figura 9:
Mapa de puntos



CONSIDERACIONES DE CDE

- Cada punto puede representar una sola observación.
- ¿Tiene las observaciones “características no comunes” (por ejemplo, la condiciones rara X)?

- Los puntos pueden estar ubicados de manera precisa. Podrían ser vinculados con otra información como género, edad, o etnicidad, lo cual incrementaría el riesgo de re-identificación.

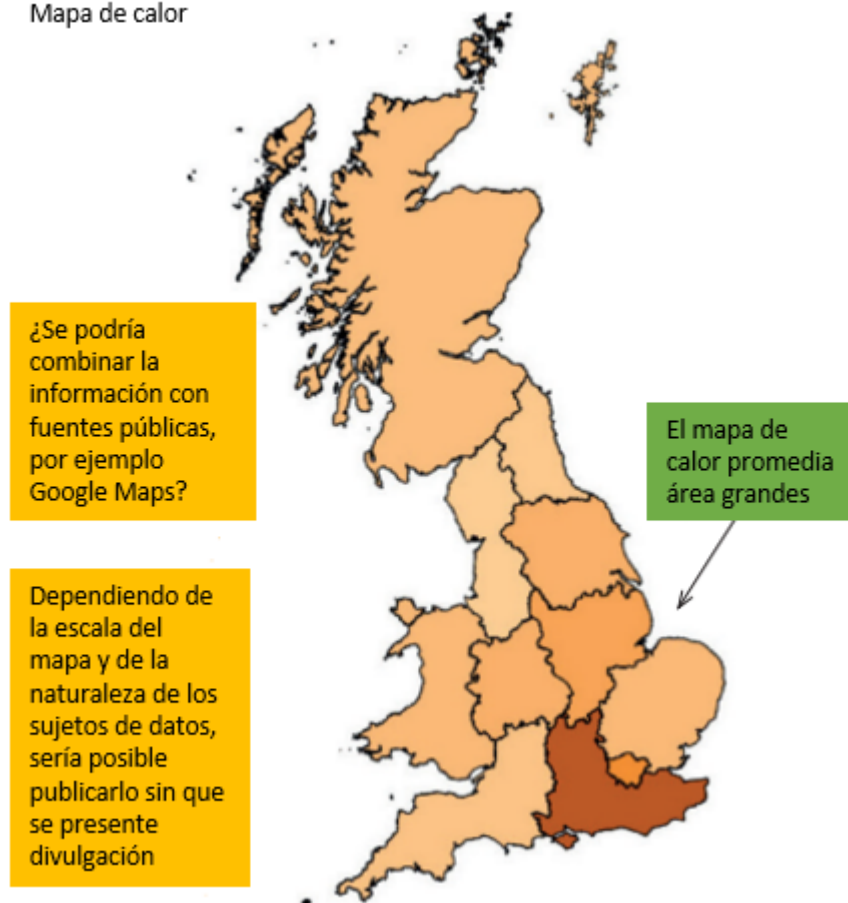
REGLA DE ORO

Ninguna celda debe contener menos de 'N' observaciones. En este caso cada punto representa una sola observación ($N=1$). Si esto se presentara en forma de tabulado se vería de manera muy obvia que no cumple con el límite de 'N' observaciones.

REDUCIENDO EL RIESGO DE DIVULGACIÓN

Un mapa de calor para indicar niveles de actividad, intensidad, concentración, etc. Por ejemplo, los colores oscuros se pueden usar para indicar una actividad más alta mientras que los colores claros para indicar menor actividad. Al convertir la Figura 9 en un mapa de calor (Figura 10) se reduce la especificidad de la ubicación y se quitan el número específico de pacientes en cada área, con lo cual se reduce de manera significativa el riesgo.

Figura 10:
Mapa de calor



Coeficientes de Gini

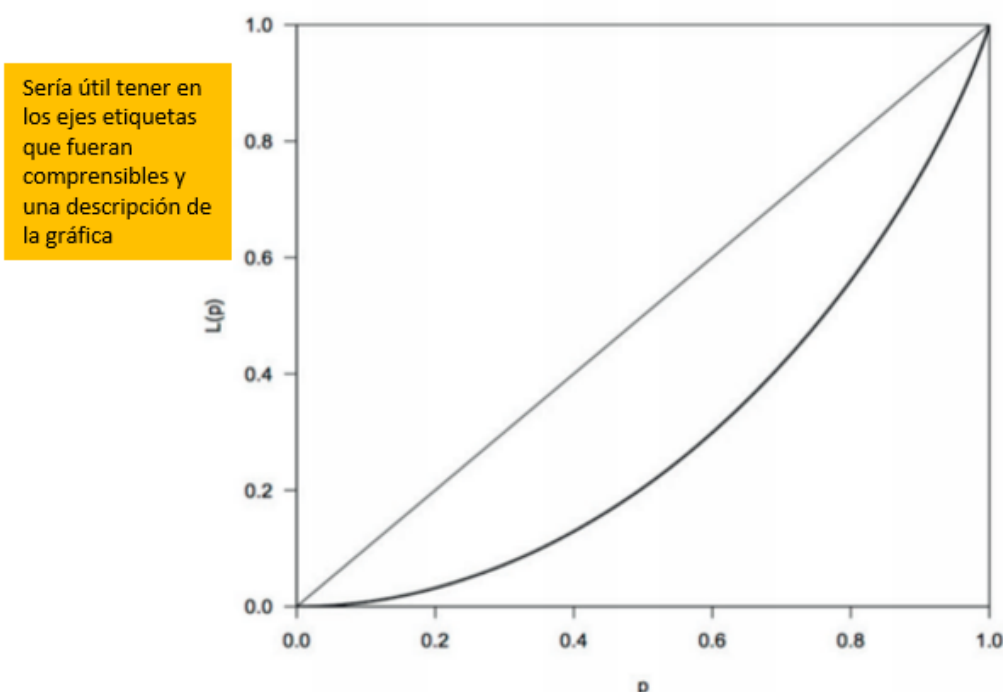
REQUERIMIENTOS MÍNIMOS:

- Frecuencias

Una curva de Lorenz mide el grado de distribución de un valor de una población o una muestra. Se usa de manera frecuente para evaluar la distribución del ingreso. Por ejemplo, un coeficiente de Gini puede demostrar que el 90% de la riqueza de un país la posee el 10% de su población.

Un analista puede producir una tabla de frecuencia que ilustra la distribución del ingreso en dos años, y luego calcular el coeficiente de Gini para cada año. El cambio en el coeficiente entre los años determinará si la desigualdad en el ingreso ha cambiado. En la Tabla 8 la información el Coeficiente de Gini muestra que la desigualdad se ha incrementado entre 2000 y 2010. Una curva de Lorenz (Gráfica 11) visualiza el grado de la desigualdad, donde la igualdad se presenta por la línea a 45 grados.

Figura 11:
Curva de Lorenz



CONSIDERACIONES DE CDE

Los coeficientes de Gini son medidas agregadas. Es poco probable que un analista produzca un coeficiente de Gini a menos que tenga suficientes observaciones para hacer significativo el coeficiente de Gini. En general son Resultados Seguros.

Tabla 8:
Desigualdad del ingreso

Rangos de Ingreso	% Hogares (2000)	% Hogares (2000)
0 – 20,000	20	15
20,001– 50,000	45	40
51,000– 75,000	25	20
75,001 +	10	25
Coeficiente de Gini	0.46	0.52

Se requiere información del tamaño de la muestra, región, ocupación, etc.

Generalmente estos son estadísticas agregadas basadas en una muestra grande

REGLA DE ORO

Asegúrese que el coeficiente de Gini se ha calculado con al menos N observaciones. Por ejemplo, en la Tabla 8, cada rango de ingreso debe incluir una cierta frecuencia de hogares. Estas deben cumplir con el requerimiento mínimo.

Para las curvas de Lorenz, asegúrese que ninguna de las partes de la distribución se refiere a menos de N observaciones.

Tasas de concentración

REQUERIMIENTOS MÍNIMOS:

- Explicación de cohorte
- Información contextual (por ejemplo, datos de negocios, sector, empresas, periodo de tiempo)

Tabla 9:
Las 3 empresas más grandes por industria (sin modificar)

Industria	Número de empresas	Aportación de los 3 más grandes en 2010
Manufactura	1000	0.05
Comercio	9000	0.01
Servicios digitales	150	0.2
Seguros	58	0.3
Refinación de petróleo	13	0.7

¿Si sugiere que las 3 empresas más grandes controlan el 70% del mercado, hay un problema de dominancia?

Las tasas de concentración (en particular, una derivación conocida como Índice de Herfindahl), se producen con frecuencia en economía de negocios para medir que tanto de la “medida” se puede atribuir a un número pequeño de observaciones. La Tabla 9 ilustra Las Tasas de Concentración de las Tres Empresas más Grandes dentro de un número de industrias. Las tres empresas más grandes por participación de mercado (ventas) en el sector de manufacturas aportaron el 5 por ciento en 2010. En la industria de refinación de petróleo, las tres empresas más grandes por participación del mercado (ventas) aportaron el 70% del mercado. En este último caso, es probable que la industria está “dominada” por una sola observación.

CONSIDERACIONES DE CDE

Si una observación es dominante: es decir que aporta la mayor parte de la medida, por ejemplo, ventas, ingreso, etc., entonces el dato se puede atribuir a una sola observación. Es algo que debe observarse en industrias poco competitivas.

REGLA DE ORO

Asegúrese que la ‘regla de dominancia’ se cumpla (ninguna observación en lo individual puede ‘dominar’ la medida por X por ciento. En general, es el 40 por ciento).

REDUCIENDO EL RIESGO DE DIVULGACIÓN

Los resultados pueden alterarse: adicionándole ruido, o redondeo, o presentarlos en ‘rango’, o usar una medida más grande de concentración (por ejemplo, considerar las 10 empresas más grandes en lugar de las 3 más grandes). Esto dependerá de los efectos en los cambios de los resultados totales.

Cuando se compara un número de industrias, se pueden “ordenar” por nivel de competitividad en lugar de presentar las tasas de competitividad.

En la Tabla 10, en lugar de producir una Tasa de Concentración para cada industria, se ordenaron las industrias de la ‘más competitiva’ a la ‘menos competitiva’.

Tabla 10:
Industrias ordenadas por tasa de competitividad (sin publicar tasa de competitividad)

Industria	Número de empresas	Orden
Comercio	9000	1
Manufactura	1000	2
Servicios digitales	150	3
Seguros	58	4
Refinación de petróleo	13	5

Muestra que es un mercado poco competitivo, pero evita el problema de dominancia de una sola empresa.

Puede ser que se tengan que suprimir todos los resultados – en cambio el analista presenta una descripción del grado de competencia en el mercado sin referirse a los resultados. Sin embargo, este sería un último recurso, dada la necesidad de tener políticas fundamentadas en evidencia.

Regresiones

REQUERIMIENTOS MÍNIMOS:

- Número de observaciones
- Grado de libertad
- Etiquetas de variables
- Parámetros omitidos
- Especificación de cohorte

Una regresión refiere a un grupo de modelos para estimar la relación estadística entre dos o más variables.

Hay muchos tipos. Principalmente hay regresiones lineales y no lineales:

- Modelos de regresión lineal estiman el “cambio de paso” entre dos variables, por ejemplo, el cambio de productividad en respuesta a cambio en inversión.
- Los modelos no lineales en general usan un máximo de probabilidad estimada para calcular la probabilidad, por ejemplo, de que mamás solteras en X por ciento es menos probable que entren a un trabajo de tiempo completo después del nacimiento de su primer hijo.

CONSIDERACIONES DE CDE

Cantidad requerida: muchas veces los analistas solicitaran tener los resultados de todas las regresiones (como en la Tabla 11). En otras ocasiones, solo solicitarán un número de resultados destacados que le son relevantes. Por ejemplo, es común que los analistas del mercado laboral en el modelo de regresión incluyan parámetros para ocupación, educación e industria; pero en las conferencias los analistas no se referirán de manera específica a estos resultados, sino que se enfocarán en los principales resultados de su interés.

La estimación de los parámetros de regresiones son un proceso iterativo. Se puede llegar a requerir un número de iteraciones para el modelo de estimación (ajustes, añadir y eliminar parámetros). Los analistas pueden solicitar resultados de regresiones que no necesariamente van a terminar en su publicación.

¿Por qué puede divulgar confidencialidad?

- Si la regresión se lleva a cabo en una sola unidad.
- Si se llevan a cabo una secuencia de regresiones, pero el número de observaciones en la cohorte cambia cada vez por un número pequeño (y las diferencias en los resultados son significativas) entonces esto se puede asociar con las observaciones adiciones que incluye la cohorte.
- Si en las regresiones solo se usan variables categóricas (estas son variable que pueden tomar valores 0/1, Falso/Verdadero, Si/No, etc.).

Tabla 11:

Tabla de estimación de regresiones

TERM	ESTIMACIÓN	ERROR ESTÁNDAR	ESTADÍSTICO T	VALOR-P
Intercepción	842.32	32.13	26.21	0.00
Femenino	-26.71	5.23	-5.11	0.00
Age20-24	170.06	14.89	11.42	0.00
Age 25-29	111.23	14.64	7.50	0.17
Age 30-34	20.39	14.76	1.38	0.39
Age 35-39	-13.01	15.23	-0.85	0.06
Age 40-44	-28.19	15.05	-1.47	0.16
Age 45-49	-21.26	14.98	-1.42	0.04
Age 50-54	-30.52	15.09	-2.02	0.00
Age 55-59	-72.56	15.38	-4.72	0.00
Age 60-64	-111.01	15.77	-7.04	0.00
Age 65-69	-122.34	15.68	-7.80	0.00
Age 70+	87.95	31.77	2.77	0.00
Estado civil (Casada/ Cohabitando/Pareja Civil)	34.81	7.50	-4.64	0.00
Estado civil (Divorciada/Viuda/Previa Pareja Civil)	57.83	10.30	5.61	0.00
Más Alta Calificación Sin Respuesta	168.35	55.01	-3.06	0.00
Grado de más Alta Calificación	121.51	30.04	-4.00	0.00
Grado de más Alta CGE nivel A	143.01	31.19	-5.15	0.00

Títulos claros y fáciles de entender

No se debe llevar a cabo en una sola observación, (series de tiempo)

REGLA DE ORO

En general, los resultados de regresiones sí se pueden entregar; una revisión de grados de libertad (deben ser al menos N), y las secuencias de regresiones que se llevan a cabo no debe diferir en número de sumas de observaciones de menos de N).

REDUCIENDO EL RIESGO DE DIVULGACIÓN

Asegúrese que hay un número suficiente de observaciones, en el modelo y entre las estimaciones. En la práctica, es raro que esto sea un aspecto de cuidado.

Residuales

REQUERIMIENTOS MÍNIMOS:

- Etiquetas en los ejes

El diagrama de residuales es una gráfica que representa los residuales (o errores) después de la estimación de un modelo de regresión. Los residuales en general se grafican contra las covariables usadas en el modelo (como en la Figura 12).

CONSIDERACIONES DE CDE

Los residuales individuales nunca se deben reportar. Sin embargo, la figura o la distribución de los residuales (observados en los errores) después de una regresión puede ayudar a describir la pertinencia de un modelo particular o de la heteroscedasticidad de los datos.

Los valores atípicos merecen consideración importante cuando se evalúa un diagrama de residuales, y qué variable se usan en el eje de las x.

Para poder atribuir un valor individual residual, uno tiene que poder ordenar de manera exitosa todas las observaciones a lo largo de eje de las x que se presente. Esto en general es difícil de lograr con un número grande de observaciones y un diagrama. Cuando se presentan varios diagramas de residuales después de la estimación de un mismo modelo, esto hace que se incremente el riesgo de divulgación ya que los valores atípicos son fáciles de identificar, si los residuales se presentan a lo largo del eje de las x.

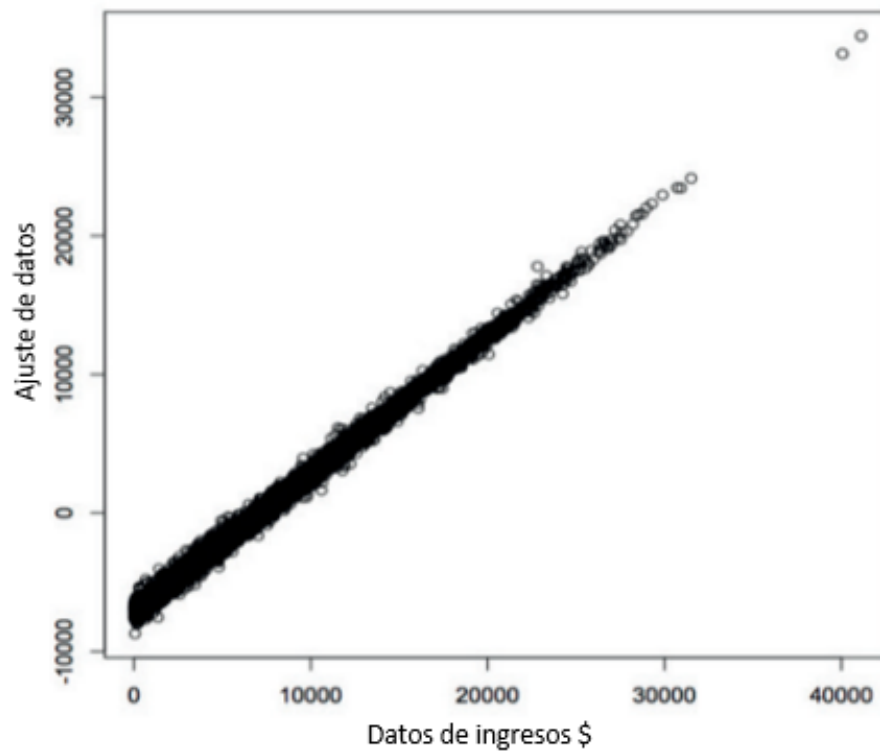
REGLA DE ORO

Como cada residual representa una observación, se debe evitar presentar diagramas de residuales.

REDUCIENDO EL RIESGO DE DIVULGACIÓN

Considere describir la forma del diagrama de residuales, o la conclusión que deriva de analizarla, más que presentar el diagrama de residuales. Si se requiere un diagrama de residuales, el riesgo se puede mitigar eliminando la escala de los ejes, y en lo posible use una covariable en el eje de las x que es difícil de observar fuera de conjunto de datos.

Figura 12:
Diagrama individual de residuales



La forma de la distribución es adecuada para poder entregar

Una solución para poder entregar es que se modifiquen los datos

Puntos individuales

Gráficos de margen

REQUERIMIENTOS MÍNIMOS:

- Número de observaciones
- Grados de libertad
- Descripción de las variables dependiente e independiente

Un gráfico de margen grafica los márgenes predecibles de una modelo de regresión. Los márgenes se generan de los modelos de regresión y muestran el valor de la media predecible del valor dependiente para las categorías o valores de una(s) variable(s) independientes. Pueden ser desplegadas en tablas (por ejemplo, Tabla 12 o en gráficas (por ejemplo, Figura 13)

CONSIDERACIONES DE CDE

En general se consideran seguras ya que los gráficos de márgenes son predicciones/estimaciones.

REGLA DE ORO

El modelo resultante debe tener al menos el límite de 'N' del número de grados de libertad y al menos usar 'N' unidades para producir el modelo. El modelo no se debe de basar en una sola unidad (por ejemplo, series de tiempo de una sola empresa).

Figura 13:
Gráfico de márgenes predictivos

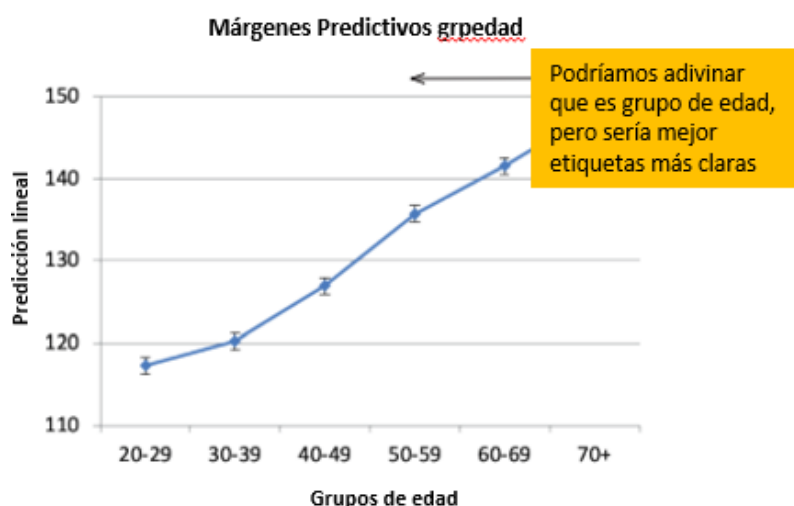


Tabla 12:
Tabla de márgenes predictivos

	Método Delta					
Grupo de edad	Margen	Error estándar	t	P>[t]	[Intervalo de confianza 95%]	
20-29	117.2684	0.419845	279.31	0.000	116.4454	118.0914
30-39	120.2383	0.502813	239.48	0.000	119.2541	121.2225
40-49	126.9255	0.56699	223.86	0.000	125.8141	128.0369
50-59	135.682	0.562859	241.06	0.000	134.5787	136.7853
60-69	141.5285	0.37812	374.30	0.000	140.7873	142.2696
70+	148.1096	0.644507	229.80	0.000	146.8463	149.373

¿Se basa el modelo en un número suficiente de observaciones?

Pruebas estadísticas

REQUERIMIENTOS MÍNIMOS:

- Número de observaciones

Las pruebas estadísticas se usan de manera frecuente en los análisis estadísticos. La naturaleza y derivación de una prueba estadística dependerá de qué hipótesis se está probando. Algunos ejemplos de pruebas estadísticas incluyen: t-prueba, R², F-prueba, etc. En general, una prueba estadística es una representación compleja de un gran número de observaciones, se usan para probar la diferencia entre dos o más grupos de observaciones, se prueba si un parámetro particular es diferente a cero o se prueba si las distribuciones observadas se parecen a la distribución teórica.

CONSIDERACIONES DE CDE

Las pruebas estadísticas en general se calculan tomando como base un número grande de observaciones. El riesgo asociado a la prueba estadística se mitiga por la complejidad misma de la prueba estadística.

Nota: las pruebas estadísticas de dos muestras marginalmente diferentes pueden resultar en divulgación, al observar las diferencias entre las dos estadísticas.

REGLA DE ORO

La prueba estadística se debe derivar con base en al menos 'N' observaciones.

Sección C

Requerimientos organizacionales y de administración

Introducción

Este Manual ha sido diseñado tanto por personal trabajando en servicios que proporcionan acceso a datos confidenciales, como por usuarios de esos servicios; y personal de organizaciones que proporcionan un ambiente de datos seguros para el uso interno de ese personal, que es probable que tengan ellos mismo que evaluar el riesgo de divulgación estadísticas de los resultados.

La Sección C está destinada sobre todo a los primeros: personal que trabaja en un servicio que provee acceso seguro a datos confidenciales, y para los cuales los usuarios probablemente serán externos (sin embargo, una solución interna bien fundamentada también puede trabajar bajo estos mismos principios: por ejemplo, el Servicio Seguro de Datos de la ONE provee acceso seguro a personal interno de la ONE de la misma manera que lo ofrece a analistas externos).

Hay una razón por la que es importante hacer esta distinción. La forma en que el Control de Divulgación Estadística está organizado puede ser un soporte en la entrega de resultados seguros. Si se administra de manera ineficiente, se presenta un riesgo muy alto de que algo falte, o se cometa un error, que resulte en la publicación de estadística que pudieran revelar la identidad del sujeto de datos o unidad de observación o/y de información confidencial.

Sin embargo, la forma en que las organizaciones manejen los requerimientos de resultados estadísticos y el proceso de Control de Divulgación Estadística puede variar de acuerdo con sus propias circunstancias. Por ejemplo, una “solución hecha en casa” puede hacer que las evaluaciones de Control de Divulgación Estadística se lleven a cabo una o dos veces por semana, dependiendo del número de proyectos que se estén desarrollando. Esto puede contrastar con un servicio grande, como el Servicio de Datos del Reino Unido, que recibe en promedio entre 5 y 10 requerimientos de resultados estadísticos cada día laboral.

Cuales quiera que sean las circunstancias, los consejos y lineamientos que se presentan a continuación pueden aplicarse a ambas situaciones. Estos reflejan las mejores prácticas compartidas y la sabiduría del personal que ha laborado para varios años en ambos tipos de servicios.

¿Cómo implementar CDE en una organización?

EN ESTE CAPÍTULO USTED APRENDERÁ SOBRE:

1. ¿Cómo alentar a los analistas para que produzcan resultados que puedan ser fáciles de evaluar para divulgación de confidencialidad?
2. ¿Por qué es importante que dos personas revisen los resultados?
3. ¿Por qué es importante que el personal que lleva a cabo la revisión sea independiente del análisis que se desarrolla?
4. ¿Cómo administrar cargas de trabajo y presión?
5. La importancia de mantener registros y auditorías
6. El contexto más amplio de llevar a cabo Control de Divulgación Estadística

.....

El Control de Divulgación Estadística representa una herramienta muy valiosa para cualquier organización que produzca de manera rutinaria resultados estadísticos de datos sensibles. El tener un enfoque robusto de la revisión de resultados estadísticos proporciona un marco de referencia que asegura que la información que se publica en el dominio público no quebranta la confidencialidad de los datos de los sujetos o unidades de observación y al mismo tiempo satisface el apetito del riesgo de los proveedores de los datos.

El diseño de cualquier proceso de CDE debe reflejar los objetivos y propósitos analíticos de la organización, y debe ser consistente con el enfoque global de seguridad de la información. No existe una receta de cómo implementar este proceso y en gran parte debe estar hecho a la medida del tipo del acceso a los datos que se proporciona. Esta sección presenta un panorama de algunos aspectos específicos que se tienen que tomar en consideración cuando se pone en operación un servicio de CDE, los retos asociados, y ejemplos prácticos basados en mejores prácticas que han adoptado cierto número de centros de datos en el Reino Unido.

Incentivando Buenos Resultados

Un proceso exitoso de CDE se basa en gran parte en la calidad de los resultados que generan los analistas. Resultados claros y fáciles de entender permite a los revisores proporcionar una mejor evaluación y reducir el tiempo y esfuerzo en la revisión de resultados estadísticos.

Una organización debe alentar formas para incentivar a los analistas a producir buenos resultados. Una capacitación básica en Control de Divulgación Estadística que cubra los tipos de resultados más comunes se puede proporcionar a todos los analistas que tienen acceso a datos sensibles. Como otra medida más, se les puede pedir a los analistas que apliquen estos principios a cada uno de los resultados que produzcan. Nosotros recomendamos que se den con anticipación lineamientos de la información que requieren los revisores para poder evaluar cada uno de los tipos de resultado estadísticos.

PRINCIPIO DE LOS CUATRO- OJOS

Es una mejor práctica general que los resultados estadísticos los revisen dos revisores distintos. Éstos pueden ser miembros de una organización o analistas internos, pero en general, cada resultado estadístico requiere ser escudriñado por al menos una persona que no está involucrada con los resultados estadísticos originales. Cada revisión se debe llevar a cabo de manera independiente de la otra revisión, y la decisión de entregar o no ciertos resultados debe ser tomada de manera conjunta para asegurar consistencia.

Las organizaciones grandes o servicios de datos pueden recibir requerimientos de una gran variedad de resultados, la evaluación de algunos puede ser desafiante para los revisores. El tener un segundo par de ojos que revisen los resultados estadísticos, puede ayudar a crear confianza en la decisión de entregar o no, y mitigar el riesgo de un error al pedir una segunda opinión.

Para otras organizaciones, por ejemplo, el proporcionar ‘en-casa’ ambientes seguros de datos para el personal, puede no ser necesario (si, por ejemplo, solo se divulgan uno dos grupos de resultados estadísticos muy de vez en cuando). Esto es porque el personal que tiene que evaluar el CDE de los resultados estadísticos tiene más tiempo para hacerlo. En un servicio con mucha demanda como el Servicio de Datos el Reino Unido, el principio de los cuatro-ojos es más apropiado.

INDEPENDENCIA DE LOS REVISORES

Todo proceso de CDE consistente depende de la habilidad de la persona responsable de revisar los resultados estadísticos para proporcionar una evaluación imparcial. Como organización, es importante asegurarse que hay un nivel adecuado de segregación entre el analista que produce los resultados estadísticos y la persona responsable de revisarlos.

Organizaciones pequeñas y servicios “de la casa” tienden a confiar en los propios analistas para hacer la revisión de los resultados. En estas circunstancias, es importante que el rol de

revisor se mantenga separado del trabajo producido por el analista. Una regla básica es que no se permita a revisores la liberación de sus propios resultados estadísticos, se debe requerir una segunda evaluación de otro analista, posiblemente de otro proyecto. En equipos pequeños, esto puede resultar todo un reto para los analistas que puede tener un conflicto de interés potencial si están trabajando en el mismo proyecto. En este caso, un proceso de auditoría robusto y revisión aleatoria de entregas hechas podría proporcionar una buena solución para monitorear la calidad de los resultados estadísticos que han sido liberados.

ADMINISTRACIÓN DE CARGAS DE TRABAJO Y PRESIÓN

El CDE es la salvaguardia final antes de que se libere en el dominio público información potencialmente sensible, una organización debe asignar recursos suficientes para asegurar que los revisores están en buena posición para hacer sus juicios. Estrés, presión y grandes cargas de trabajo pueden llevar a que se libere resultados que divulguen confidencialidad. Como organización es importante proveer un ambiente en el cual los revisores se sienten confiados de rechazar malos resultados y no sientan presión por parte de los analistas para hacer excepciones.

Poner en operación un Acuerdo del Nivel del Servicio (ANS) es una forma de liberar presión del sistema y dar a los revisores suficiente tiempo para evaluar cada requerimiento de manera adecuada. Los analistas necesitan estar conscientes que los requerimientos de revisión de resultados los deben enviar con el debido tiempo y su plan de trabajo debe prever el tiempo necesario para la revisión de sus resultados. En algunos casos, el que planeen un periodo de tiempo para la revisión de sus resultados puede alentar a que los analistas se concentren en solicitar todos los resultados que necesitan en un solo requerimiento, lo que evita la fragmentación y facilita el trabajo del revisor.

Un sistema de rotación de revisores también puede mejorar el desempeño del proceso, y puede evitar favoritismo o desbalance en las cargas de trabajo hacia los revisores “más amigables”. Esto también puede alentar a que los revisores trabajen diferentes variedades de resultados estadísticos y ultimadamente puede incrementar la consistencia en la aplicación del CDE.

El asegurarse que los revisores de resultados trabajen en otras tareas durante la semana laboral puede ayudar a evitar “fatiga de resultados” y agotamiento, que podría llevar a una reducción en la calidad del servicio.

De manera ideal, el personal debe revisar los resultados con diferentes miembros del equipo; si un par de revisores tienen ambos un error consistente, esto no se podría percibir.

Algunos servicios pueden considerar que no todos los resultados estadísticos se necesitan revisar. Por ejemplo, resultados de regresiones que contienen un riesgo muy bajo de divulgación; por el contrario, es más común el riesgo inherente en las tablas de frecuencias

o en las estadísticas descriptivas. Podría ser que los revisores solo lleven a cabo CDE para éstos últimos, y de manera automática liberen, por resultados de regresiones sin cuestionarlas. Podría ser una posición comprensible de adoptar: sin embargo, el GPASD recomienda que todo resultado estadístico se revise, aun cuando solo sea una mirada superficial la que se aplique a resultados como las regresiones. Un enfoque de salvaguardia sería el alentar a los analistas prolíferos que solicitan muchos requerimientos que vean si realmente necesitan la entrega de tantos resultados estadísticos, sobre todo si saben que es poco probable que usen o necesiten referir la mayoría de los resultados que están solicitando.

MANTENIMIENTO DE REGISTROS Y RENDICIÓN DE CUENTAS

Como una herramienta para mitigar riesgo, es importante mantener un registro de todos los resultados que se solicitan para poder tener evidencias para evaluar si el proceso opera de acuerdo con la necesidad de la organización. En particular, hay un número de razones por las cuales mantener un registro exhaustivo del proceso de CDE puede ser útil para una organización.

Primero, esto ayuda a monitorear el volumen y tipo de resultados estadísticos que se solicitan y permite generar información de la calidad de los resultados estadísticos y las decisiones de los revisores, con margen para mejora continua en el proceso. También es una herramienta para identificar quien es responsable de cada requerimiento y para asegurar que el proceso de CDE se lleva de una manera correcta sin confusiones.

Segundo, un registro histórico de la actividad de CDE ofrece una base útil de conocimiento para los revisores, especialmente cuando les toca revisar resultados de fuentes de datos con los cuales están menos familiarizados. También proporciona seguridad en el caso de decisiones pasadas de entregas en el caso de auditorías de los dueños de los datos. Finalmente, es una forma fácil de generar registros de la administración de riesgo para auditorías internas y externas (por ejemplo, ISO 27001, Kit de Herramientas de Seguridad y Protección de Datos) y para permitir controles puntuales. Al mismo tiempo, proporciona un enfoque consistente con las condiciones que establece la regulación de protección de datos (por ejemplo, privacidad por diseño).

No existe una regla general de cómo hacer el registro del proceso de CDE. Se pueden diseñar registros auditables a la medida de las necesidades de una organización, del volumen de servicio, los recursos disponibles y los requerimientos de la gobernanza de la seguridad de la información (por ejemplo, ISO 2700 Herramienta de Gobernanza de Información. Es importante que el proceso no se vuelve muy pesado de seguir, ya que ello puede llevar a problemas en el registro tanto del analista como del revisor.

En general es una buena práctica mantener un registro de todos los resultados estadísticos que se han solicitado;

- La decisión del revisor (esto es, resultados estadísticos liberados, rechazados y retirados);
- ¿Quién solicitó los resultados estadísticos?;
- ¿Quién revisó los resultados estadísticos?;
- Cualquier aspecto o cambios aplicados;
- El razonamiento detrás de cada decisión;
- Fecha cuando el requerimiento fue hecho, aprobado y los resultados entregados.

Además, podría ser útil mantener una copia de todos los resultados estadísticos entregados en su forma original. Esto es para permitir revisiones puntuales y minimizar divulgación secundaria que podría presentarse por copias múltiples que se pudieran entregar de los mismos resultados estadísticos.

AUDITORÍA

Un aspecto clave de la administración del riesgo de la revisión de resultados estadísticos se basa en la auditoría continua del proceso de CDE.

Se aconseja tener una auditoría regular interna de la actividad de CDE, para poder monitorear la calidad de los resultados estadísticos que se han entregado e identificar cualquier aspecto que pueda causar problemas de seguridad. Por ejemplo, la tasa de rechazo de resultados estadísticos puede representar un buen indicador del problema. Una tasa inusualmente alta se puede relacionar con una mala calidad de los resultados estadísticos de un analista particular o bien de dar mayor capacitación a los revisores de resultados, o en última instancia puede indicar que el CDE se está aplicando de manera incorrecta. Por el contrario, una tasa baja o nula de rechazos puede ser una señal de que no se cumplen los procedimientos de CDE, lo cual puede llevar a liberar resultados del sistema que potencialmente pueden divulgar confidencialidad.

Un proceso interno de revisión permite a la organización identificar el espectro de mejora continua y facilita un debate constructivo del sistema adoptado. En la misma línea, las auditorías externas pueden proporcionar una oportunidad valiosa para tener una opinión fresca del proceso de CDE. Es importante designar auditores con el propósito de mejorar el proceso y para alentar a los revisores y analistas a informar sobre aspectos relacionados con fallas que así se cometieron. Puede resultar beneficioso llevar a cabo reuniones regulares con revisores para discutir resultados que particularmente presentan grandes retos, para identificar la necesidad de capacitación y compartir experiencias en la forma en que aplican el CDE. Los controles puntuales de los resultados estadísticos que han sido liberados también pueden ser usados como una herramienta adicional para monitorear el desempeño del proceso, y para retirar del dominio público resultados que fueron entregados y que se consideran pueden divulgar confidencialidad.

El concepto de divulgación estadística es dependiente del contexto en el cual se liberan los resultados, de la interpretación del revisor y del apetito al riesgo en ese momento. Estos factores pueden cambiar en el tiempo, especialmente cuando más resultados estadísticos del mismo tipo han sido liberados lo cual conlleva a un incremento en el riesgo de divulgación secundaria. Las revisiones puntuales aleatorias y las auditorías proporcionan una forma de revisión del riesgo y aseguran que el enfoque es consistente en el tiempo y está alineado a los objetivos de la organización.

ENTENDIENDO DEL PANORAMA MÁS AMPLIO

El CDE es una herramienta ampliamente usada, que atiende un marco normativo cada vez más complejo en materia de confidencialidad de los datos y gobernanza de la información. En este ambiente de rápidos cambios, una organización puede tener problemas para identificar cuáles son las habilidades correctas para este puesto. Al mismo tiempo, los revisores pueden sentirse desafiados por las nuevas metodologías estadísticas y las nuevas fuentes de datos, las cuales pueden conllevar potencialmente formas no conocidas de divulgación de confidencialidad.

En este contexto, es importante identificar oportunidades de capacitación y mantener a los revisores actualizados de mejores prácticas, cambios en la legislación y nuevas técnicas de CDE. Además, una organización debe alentar a los revisores para que compartan y discutan de manera informal resultados estadísticos especialmente desafiantes y los retos que enfrentan. En última instancia, el Control de Divulgación Estadística es una herramienta que refleja el enfoque de una organización hacia la administración de riesgo. Por esta razón, es importante que los revisores se sientan confiados cuando revisan resultados estadísticos y tengan una ruta para compartir cualquier aspecto o puedan solicitar asesoría de un experto.

Administración de los analistas

EN ESTE CAPÍTULO USTED APRENDERÁ SOBRE:

1. La importancia de mantener un enfoque consistente del CDE
2. Como organizaciones que proveen un servicio similar, o proporcionan acceso a datos similares, pueden trabajar de manera conjunta para tener un enfoque consistente.

.....

Puede ser desafiante el recibir un gran número de requerimientos frecuentes de analistas para revisión de resultados estadísticos. Por ejemplo, ¿se tiene que priorizar los requerimientos? ¿Qué tan consistente se puede lograr la aplicación del CDE? Algo muy importante, como se puede cultivar la relación con los analistas para lograr los resultados más eficientes. En esta sección se consideran cierto número de aspectos que afectan tanto a los analistas como al personal del servicio de acceso a datos.

ENFOQUE CONSISTENTE DE CDE EN TODA LA ORGANIZACIÓN Y TODOS LOS REVISORES

Cuando hay distinto personal en una organización que lleva a cabo el CDE, es importante que cada revisor tenga un enfoque consistente. Esto incluye:

- Evaluar estadísticas para Control de Divulgación;
- Llevar a cabo CDE en la misma manera;
- Asegurar que el tipo de resultados que presentan los analistas cumple con los requerimientos de la organización.

Las organizaciones pueden diferir en el enfoque para evaluar el CDE de las estadísticas en relación con el tipo de resultados estadísticos que requieren que los analistas envíen (véase “¿Qué son buenos resultados?” adelante para recomendaciones). Sin embargo, es crucial que los revisores apliquen el enfoque de manera consistente. Si esto no sucede, entonces los analistas pueden buscar en particular ciertos revisores (por ejemplo, porque son más indulgentes), lo cual incrementa el riesgo de que se publiquen estadísticas no seguras.

Los revisores deben tratar a los analistas de una manera consistente deben asegurar que los mensajes de comunicación son consistentes. Por ejemplo, cuando se requiere información adicional de un analista para que el revisor puede llevar a cabo la evaluación para saber si una estadística es segura o no, el revisor debe explicar de una manera clara la razón de por qué lo requiere. Igualmente, si un resultado estadístico se considera no seguro y se requieren revisiones, el revisor debe asegurarse de que el analista entiende la(s) razón(es) de por qué y con ello puede hacer los cambios requeridos para producir un conjunto de resultados estadísticos seguros.

Finalmente, es importante que el enfoque adoptado por la organización para proteger la confidencialidad de los sujetos de los datos o unidades de observación es consistente a través del tiempo. Recomendamos que lo antes expuesto se monitoree mediante auditorías regulares, cuando se identifiquen ciertas cuestiones, se implementen los cambios necesarios (por ejemplo, mediante capacitación del personal).

Podría haber ventajas, si se identifican servicios que ofrecen acceso a tipos de datos similares, o incluso que proporcionen acceso a los mismos analistas, que se apliquen enfoques consistentes de CDE. Sería más eficiente porque los analistas no tendrían que aprender “dos o más” enfoques y recordar el aplicarlos dependiendo de qué servicio con el cual están teniendo acceso a los datos. Con diferentes servicios de acceso el objetivo a alcanzar sería la convergencia.

¿Qué es un buen conjunto de resultados estadísticos?

EN ESTE CAPÍTULO USTED APRENDERÁ SOBRE:

1. ¿En qué consiste un ‘buen’ resultado estadístico?
2. ¿Qué información se le debe solicitar que proporcione el analista cuando va a solicitar revisión de resultados?

Las organizaciones tendrán diferentes requerimientos en relación con el tipo de resultados estadísticos que tiene que revisar. Se presentan algunas sugerencias de los que constituyen un conjunto de buenos resultados estadísticos:

- Bien explicados
- Descripción del proyecto
- Conjunto de datos que se usó
- Criterio de selección de la muestra
- Método
- Descripción de las variables
- Descripción de los resultados
- Presentado de manera pulcra
- Tablas y gráficas numeradas
- Variables con etiquetas claras
- Fácil de leer, etc.

INCLUYE LA INFORMACIÓN REQUERIDA

Esta es la información que necesita proporcionar el analista para que él (los) revisor(es) puedan evaluar si las estadísticas son seguras. Esto incluye información como el número de observaciones, así como una descripción clara de las variables. Véase ‘Requerimientos Mínimos’ (por ejemplo, la información que necesita proporcionar el analista) en cada una de las técnicas de CDE en la sección B de este Manual.

MÍNIMA CANTIDAD DE INFORMACIÓN A LIBERARSE

Recomendamos que se le solicite al analista que reflexione y seleccione las estadísticas que necesita para presentar sus hallazgos, más que solicite todas las estadísticas que haya producido durante su análisis.

NO LOG FILES

Recomendamos que no se entreguen log files o resultados estadísticos que se hayan copiado de un log file. Esto es porque este tipo de resultados no cumplen con los requisitos antes descritos (por ejemplo, explicados, presentados con pulcritud, etc.) y es probable que incluyan estadísticas que no se requieren en la publicación.

RAZÓN PARA LIBERAR

Los analistas deben exponer la razón que justifique la liberación (por ejemplo, revista donde busca publicar, presentación).

Las organizaciones quizás quieran utilizar un formato para tener un registro de los requerimientos del analista en el cual se capture parte de la información que se presenta arriba, cuando ello sea apropiado. Esto podría estar disponible en línea o dentro de la Configuración Segura.

¿PORQUE SON IMPORTANTES LOS BUENOS RESULTADOS ESTADÍSTICOS?

Cuando no se producen buenos resultados estadísticos es más probable que se presente un mayor riesgo de divulgación.

Por ejemplo, el asegurarse que los resultados estadísticos estén bien explicados y presentados de manera pulcra significa que los revisores entienden las estadísticas y pueden hacer de manera rápida y fácil una evaluación para saber si son seguras. Cuando se presenta un conjunto de resultados malos (por ejemplo, con descripciones pobres y mala presentación), los revisores encontrarán difícil poder entender los resultados y aun cuando puedan solicitar información adicional (por ejemplo, requerimiento de más información al usuario), los resultados estadísticos serán más difíciles de revisar y por ello mayor el riesgo de divulgación.

El requerir al analista que proporcione un mínimo de información que necesita para su publicación ayuda a reducir el riesgo de divulgación secundaria. Cuando se solicita un gran número de estadísticas para revisión (por ejemplo, muchas estadísticas descriptivas), el riesgo de divulgación secundaria (por ejemplo, divulgación por diferenciación) será mayor.

Para asegurar que los analistas producen buenos resultados estadísticos, se recomienda que la organización tenga un sistema en operación que incentive el buen comportamiento y evite la producción y publicación de malas estadísticas (por ejemplo, estadísticas riesgosas). Podría ser que se priorice la liberación de ‘buenos resultados’ sobre los ‘malos resultado’ y se explique a los analistas que presentaron los conjuntos de resultados, porque

la revisión de los resultados se priorizó. Véase Ritchie & Welpton (2013) para enfoque similares a este.

LA REVISIÓN DE BUENOS Y MALOS RESULTADOS

En agosto de 2018 el GPASD organizó un taller de trabajo de CDE, que reunió a un grupo de personas dedicadas a revisión de CDE. Cada participante escribió un ejemplo de un ‘buen resultado’ y también un ejemplo de un ‘mal resultado’.

Los resultados se muestran en la siguiente página

- Fácil de entender (etiquetas claras en gráficas, etc.);
- Bien explicado (metodología e interpretación de resultados.)

En contraste, malos resultados:

- Contienen poca o ninguna explicación de lo que los resultados muestran;
- Muchas veces era un poco más que un log file creado por el analista como parte de su trabajo diario.

Los buenos

Campo de información requerida con explicación de porque era segura la liberación de los valores atípicos

Pidió asesoría antes de solicitar la revisión

Los resultados incluían por separado cálculo para mostrar que se cumplía la regla de dominancia

Proporcionó borrador del artículo que facilita el entendimiento del contexto de los resultados

Solo solicitó lo que realmente requería

Explicación clara de los resultados

Resultados presentados de manera clara, listo para publicarse

Los resultados se acompañaron del plan del analista, para que fuera fácil entender los resultados

Mando su solicitud de revisión con mucho tiempo de anticipación a su presentación

Proporcionó cita de los datos

Explicación clara del significado de los resultados

Se proporcionó suma de frecuencias de las gráficas

Voluntad para explicar metodología

Fácil de entender los nombres de las variables

Los malos



Administración de expectativas

EN ESTE CAPÍTULO USTED APRENDERÁ SOBRE:

1. ¿Cómo construir una buena relación con los analistas que producen estadística?
 2. ¿Por qué es importante colaborar con los analistas?
 3. ¿Cómo se puede apoyar a los nuevos analistas?
-

ACUERDO DE NIVEL DE SERVICIO (ANS)

El establecer un Acuerdo del nivel de servicio (ANS) de revisión de CDE que ofrece la organización y mantener el estándar en la provisión del servicio, ayudará a manejar las ‘expectativas’ de los analistas. Al mismo tiempo el manejar las ‘expectativas’ de los analistas ayuda a administrar la presión y las cargas de trabajo de los revisores, ambos aspectos reducen el riesgo de que se liberen estadísticas que no sean seguras. Por ejemplo, si una organización no tiene estándares en términos del tiempo que le lleva la revisión de los resultados y un analista solicita la revisión de resultados al final de día, solicitando que se lo entreguen ese mismo día, es probable que esto le provoque presión al revisor.

La revisión de resultados estadísticos bajo presión, con poco tiempo, es riesgoso y pueden ocurrir errores, incrementando el riesgo de divulgación de información confidencial. Teniendo un estándar establecido que determine el tiempo requerido para la revisión de resultados y que sea adecuado para los revisores, y al mismo tiempo atienda las necesidades de los analistas, asegura que el personal no revisará resultados estadísticos bajo presión indebida.

SER CLARO SOBRE LOS SERVICIOS Y RESPONSABILIDADES

Los servicios de revisión de CDE varían entre organizaciones, así como las responsabilidades de los analistas dentro del proceso. Sin embargo, es importante que las organizaciones comuniquen de manera clara lo que incluye el servicio de revisión del CDE, así como lo que excluye (por ejemplo, una organización puede proporcionar asesoría y guía de expertos de cómo producir resultados estadísticos buenos y seguros, pero no va a modificar ni alterar los resultados estadísticos de ninguna manera para hacerlos seguros). De igual manera, el ser claro sobre la responsabilidad de los analistas (por ejemplo, producir resultados buenos

y seguros y estar disponible para responder a dudas, etc.) asegura que los analistas están conscientes de lo que se les puede pedir a ellos. Ambas cosas ayudan a manejar las expectativas de los analistas y asegura que la revisión del CDE se lleva a cabo en un ambiente sin presiones.

ADMINISTRACIÓN DE RELACIONES

Administrar las relaciones con analistas de manera efectiva es un elemento clave para manejar el riesgo de divulgación.

El establecimiento de la relación se inicia con la sesión de capacitación que los analistas tienen que tomar antes de tener acceso a la Configuración Segura. Recomendamos que en la capacitación se proporcione a los analistas el entendimiento del CDE y de las habilidades requeridas para producir resultados estadísticos buenos y seguros. También es importante proporcionar a los analistas una visión clara de sus responsabilidades y cómo están involucrados en el proceso de CDE, y de la relación laboral conjunta entre ellos y los revisores.

Después de la sesión de capacitación, el proceso de CDE debe ser un proceso colaborativo, basado en entendimiento y respeto mutuo. Los revisores deben estar abiertos a revisar cualquier estadística que se les presente, y cuando tenga conocimiento limitado de la estadística, trabajar con el analista para entenderlo. Los revisores también deben proporcionar ayuda y asesoría a los analistas de cómo hacer seguros, resultados estadísticos ‘inseguros’. De igual manera, los analistas deben asumir la responsabilidad de sus resultados estadísticos, y deben tener cuidado de producir resultados estadísticos de buena calidad y seguros, y tener la disponibilidad para discutir sobre éstas con los revisores y hacer los cambios que se necesiten. Ambas partes deben trabajar de manera conjunta para identificar formas en que los resultados estadísticos puedan liberarse, este trabajo no debe recaer solo en el (los) revisor(es).

Como Desai and Ritchie (2009, p.8) demuestran en su artículo “Administración Efectiva de Investigación”, la capacitación y el involucrar a los analistas en el CDE promueve una cultura de entendimiento de la seguridad de los datos, en la cual los analistas sienten que son responsables de la seguridad de los datos. Esto disminuye el riesgo de divulgación para los sujetos de los datos (véase este artículo para una discusión completa de los beneficios del sistema).

APOYO A ‘NUEVOS’ ANALISTAS

Nuevos analistas pueden requerir apoyo extra para producir resultados estadísticos seguros y se les debe proporcionar. De manera ideal esas conversaciones se deben hacer mediante teléfono ya que le da al analista la oportunidad de hacer preguntas y el revisor se puede asegurar que sus puntos se han entendido. Se recomienda posteriormente el envío de un correo electrónico para confirmar la conversación y los acuerdos.

Se pueden requerir recursos adicionales para nuevos analistas, sin embargo, esto resulta un uso eficiente de recursos en el largo plazo. Al proporcionar apoyo adicional en las primeras etapas asegurará que los analistas entendieron sus responsabilidades y cuenten con las habilidades para producir resultados estadísticos de calidad y seguros – lo que ahorrará tiempo y esfuerzo a los revisores en el largo plazo.

Otros recursos

No existen muchos recursos que proporcionen guía sobre cómo poner en operación un proceso de CDE, o de cómo administrar a los usuarios del servicio con respecto al CDE. Esto porque no existen muchas Configuraciones Seguras, y porque son una invención relativamente reciente. Sin embargo, en esta sección se incluyen referencias que pueden resultar útiles para el personal que trabaje en Configuraciones Seguras.

AMINISTRACION EFECTIVA DE INVESTIGACIÓN

Professor Felix Ritchie (University of the West of England) y Tanvi Desai (former Assistant Director of the Administrative Data Service and Data Manager at the London School of Economics) publicaron un artículo sobre cómo administrar usuarios de manera efectiva. Si bien este artículo no se enfoca de manera específica al CDE, si promueve que el personal de Configuraciones Seguras piense cómo administrar sus usuarios de una manera positiva y proactiva, y cómo crear incentivos.

<https://pdfs.semanticscholar.org/5ad1/c7b30b8310f448cabe61386c77889048a44b.pdf>

RECURSOS DE CONTROL DE DIVULGACION ESTADISTICAS DE LA OFICINA NACIONAL DE ESTADÍSTICA (ONE) DEL REINO UNIDO

La ONE produjo una guía de CDE para la liberación de microdatos, incluye escenario con intrusos y un lineamiento específico para liberación de estadísticas de salud (tablas agregadas). Si bien estos lineamientos no fueron diseñados para productos de investigación, los conceptos que explican son útiles.

<https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol>

SERVICIO ESTADISTICO DEL GOBIERNO

El Servicio Estadístico del Gobierno apoya a los estadísticos que trabajan en departamentos y agencias de gobierno. Han producido unos lineamientos para llevar a cabo CDE para tablas agregadas creadas de fuentes de registros administrativos.

<https://gss.civilservice.gov.uk/guidances/methodology/statisticaldisclosure-control/#tables-produced-from-administrativesources>

LINEAMIENTOS DIGITALES DEL SERVICIO NACIONAL DE SALUD

Los investigadores que tienen acceso a los datos del Servicio Nacional de Salud Pública Digital de Inglaterra tienen que cumplir con los lineamientos de publicación de tablas de datos de salud, que se establecen en este documento:

<https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/isb1523-anonymisation-standard-for-publishing-health-and-social-care-data>

LINEAMIENTOS DE ESSNET

Se refirieron antes en el Manual, los lineamientos originales de ESSNet producidos en 2010, se pueden encontrar aquí:

http://neon.vb.cbs.nl/casc/ESSnet/guidelines_on_outputchecking.pdf

Después se actualizaron como parte del proyecto Datos sin Frontera. Se pueden encontrar como parte de los Entregables en la sección 11.

<http://dwbproject.org>

https://ec.europa.eu/eurostat/cros/content/guidelines-outputchecking_en

ARTÍCULO DE INVESTIGACIÓN SOBRE CDE EN PRODUCTOS (RESULTADOS)

El Professor Felix Ritchie (University of West of England) escribió este artículo técnico sobre CDE para productos, u porque es necesario considerar el CDE para investigación de manera separada.

https://wiserd.ac.uk/sites/default/files/documents//WISERD_WDR_006.pdf

GUÍA DE CDE PARA PRODUCTOS (RESULTADOS)

Producido por la Red de Datos Administrativos para Investigación, este artículo del Profesor Felix Ritchie y Philip Lowthian de la ONE describe CDE para productos (resultados) de investigación.

https://adrn.ac.uk/media/174254/sdc_guide_final.pdf

RED DEL REINO UNIDO DE ANONIMIZACIÓN

La Red del Reino Unido de Anonimización proporciona asesoría para anonimización de datos confidenciales. El sitio proporciona un número de recursos que pueden ser de interés.

<https://ukanon.net>

CODIGO DE ANONIMIZACION DE LA OFICINA DEL COMISIONADO DE INFORMACIÓN (OCI)

La Oficina del Comisionado de Información (OCI) publicó un Código de Práctica en 2012. En poco tiempo lo actualizará para tomar en cuenta los cambios que han ocurrido desde la introducción de la Regulación General de Protección de Datos (RGPD).

<https://ico.org.uk/media/1061/anonymisation-code.pdf>

LAS CINCO SEGURAS

Para una explicación del marco de referencia de las Cinco Seguras para facilitar el uso seguro de datos, véase: Desai, Ritchie and Welpton (2016).

<https://uwe-repository.worktribe.com/output/914745>

Glosario

Analista

Personas que analizan microdatos con objetivos de investigación y producen resultados estadísticos. En algunos campos de estudio, se les conoce como ‘investigadores’; para claridad usamos ‘analistas’ a lo largo de todo el Manual.

Apetito de riesgo

El nivel de riesgo que un dueño de datos está dispuesto a asumir en relación con el uso de sus datos.

Atribución

Es cuando un sujeto de dato o unidad de observación puede ser identificado por características, que parecen anónimas de manera individual, pero se unen para crear una imagen más clara que potencialmente puede divulgar confidencialidad.

Cinco Seguras

Es un marco de referencia de gobernanza adoptado por muchas Configuraciones Seguras para describir el enforque de administración del acceso a los datos.

Configuraciones Seguras

Un ambiente tecnológico (y a veces físico) en el cual los analistas tienen acceso a datos para llevar a cabo sus análisis, y sus resultados estadísticos se les entregan después de una revisión que hace el personal de Control de Divulgación Estadística. También se les denomina Centro de Análisis de Datos, Enclave Seguro, Ambiente Seguro de datos, Ambiente Confiable para Investigación, Puerto Seguro de Datos, Refugio de Datos Seguros. Puede incluir ‘acceso en sitio’ o acceso ‘remoto seguro’.

Control de Divulgación Estadística

El proceso que se aplica a los productos estadísticos (resultados estadísticos) para mitigar el riesgo potencial de los resultados de divulgar confidencialidad una vez que salen de la Configuración Segura.

Controlador de datos

Este es un término definido en la legislación de protección de datos, y refiere a la organización o el individuo que determinan el propósito para el cual los datos se procesan. En general este será un dueño de los datos, pero no siempre.

Divulgación secundaria

Cuando dos, o más, al parecer estadísticas ‘seguras’ (por ejemplo, tablas, gráficas) que se presentan como parte de un producto -o incluso de varios productos- que pueden producir nuevas estadísticas potencialmente riesgosas de divulgar confidencialidad cuando se combina.

Dueño de los datos

Es la organización responsable de los datos. Puede ser un controlador bajo los términos de la Ley de Protección de Datos, pero no siempre. Por ejemplo, la Oficina Nacional de Estadística es el dueño de los datos que capta mediante encuestas.

Identificación/re-identificación

Es la identificación de un sujeto de datos o unidad de observación, es lo que el proceso de CDE busca evitar. Cuando los analistas usan microdatos con mucho detalle, aun cuando los datos no tengan identificadores directos, existe el riesgo de re-identificación, mediante la combinación de variables (identificable de manera indirecta).

Información confidencial (sensible)

Refiere a datos que contienen detalles y han sido captados en términos confidenciales, puede atribuirse a un individuo, y se le puede haber quitado identificadores directos como nombre o dirección.

Información del contexto

Para poder hacer que el proceso de CDE se dé sin contratiempos, muchos proveedores de datos requieren que los resultados estadísticos sean tablas o gráficas, estén acompañados de materiales como: Ns sin ponderados, información de la muestra, método, etc.; aun cuando la extensión de los requerimientos puede variar entre los distintos proveedores de datos.

Límite

Es el número de observaciones subyacentes a la derivación de la estadística que debe cumplirse para que se considere 'seguro'. A lo largo de este Manual se ha usado el límite común de 10.

Principio de los cuatro-ojos

Es el mejor principio de práctica, que cuando es posible, los productos estadísticos (también denominados resultados estadísticos) los deben revisar dos personas y no solo una.

Producto estadístico

El resultado del análisis que el analista desea se libere de la Configuración Segura y al cual se le aplica el proceso de CDE. También se denominan resultados estadísticos.

Producto seguro

Un conjunto de estadísticas que se puede considerar que no revelan información confidencial y /o revelan la identidad de un sujeto de datos.

Proveedor de datos

La organización que ofrece los datos para la Configuración Segura. El proveedor de datos puede operar como servicio una Configuración Segura

Regla de oro

Un enfoque práctico a ciertos aspectos del CDE, basados en la experiencia de aquello es probable que mitigue el riesgo de divulgación en casi todas las situaciones.

Revisor de resultados

Son las personas responsables de revisar los productos estadísticos (también denominados resultados estadísticos) generados en una Configuración Segura para aspectos que potencialmente divulguen confidencialidad.

Sujeto de datos

La unidad de observación en un conjunto de datos. En general son individuos o empresas, dependiendo de la fuente de datos.

Derecho de autor

¿Si quieres usar nuestro trabajo, por favor hazlo, pero por favor danos reconocimiento!
Grupo de Profesionales de Acceso Seguro a Datos.

Este trabajo tiene licencia bajo la Licencia Internacional de Atribución Creativa Común- No comercial- No derivativa4.0