

Guía Quartz: Limpieza de datos



CC by SA NC Thomas Hawk

Esta guía fue [originalmente escrita](#) por Christopher Groskopf para la revista Quartz. La traducción es de [Gibrán Mena](#) para Escuela de Datos.

Contents

[hide]

1 Índice

- 1.1 Problemas que debería resolver tu fuente
- 1.2 Cuestiones que deberías resolver tú mismx
- 1.3 Problemas que un tercero experto debería ayudarte a solucionar
- 1.4 Problemas que un programador debería ayudarte a resolver

2 Lista detallada de problemas

- 2.1 Problemas que tu fuente debería resolver
 - 2.1.1 Valores faltantes
 - 2.1.2 Hay datos faltantes que fueron reemplazados con ceros
 - 2.1.3 Faltan datos que sabes deberían estar ahí
 - 2.1.4 Filas o valores que están duplicados
 - 2.1.5 La ortografía es inconsistente
 - 2.1.6 El orden de las palabras es inconsistente
 - 2.1.7 Formatos de fecha inconsistentes
 - 2.1.8 Las unidades no están especificadas.
 - 2.1.9 Las categorías fueron mal elegidas
 - 2.1.10 Los nombres de los campos son ambiguos
 - 2.1.11 El origen de los datos no fue documentado
 - 2.1.12 Hay valores sospechosos
 - 2.1.13 Los datos son muy burdos
 - 2.1.14 Los totales difieren de los agregados publicados
 - 2.1.15 La hoja de cálculo tiene 65536 filas
 - 2.1.16 Hojas de cálculo que tienen fechas en 1900, 1904, 1969 o 1970

- 2.1.17 Texto que fue convertido a números
- 2.1.18 Números que fueron guardados como texto

2.2 Problemas que deberías resolver tú mismx

- 2.2.1 El texto es confuso
- 2.2.2 Los espacios al final de la línea están mal codificados
- 2.2.3 Los datos están en PDF
- 2.2.4 Los datos son demasiados granulares
- 2.2.5 Los datos fueron capturados por humanos
- 2.2.6 Los datos están mezclados con formateo y anotaciones
- 2.2.7 Los agregados fueron calculados con valores que faltan
- 2.2.8 La muestra no es aleatoria
- 2.2.9 El margen de error es demasiado amplio
- 2.2.10 El margen de error es desconocido.
- 2.2.11 La muestra está sesgada
- 2.2.12 Los datos fueron editados manualmente
- 2.2.13 La inflación distorsiona los datos
- 2.2.14 Variaciones naturales/de temporada distorsionan los datos
- 2.2.15 La escala de tiempo fue manipulada
- 2.2.16 El marco temporal fue manipulado

2.3 Problemas que deberían ayudar a resolver terceros

- 2.3.1 El autor no es confiable
- 2.3.2 El proceso de recolección es opaco
- 2.3.3 Los datos son de una precisión irreal
- 2.3.4 Hay valores atípicos inexplicables
- 2.3.5 Un índice enmascara variaciones subyacentes
- 2.3.6 Hay p-hacking en los resultados
- 2.3.7 La ley de Benford falla
- 2.3.8 Demasiado bueno para ser verdad

2.4 Problemas que un programador debería ayudarte a resolver

- 2.4.1 Los datos están agregados en categorías o geografías incorrectas
- 2.4.2 Los datos están en documentos escaneados

Índice

Problemas que debería resolver tu fuente

- Valores faltantes
- Valores faltantes reemplazados por ceros
- Datos que faltan pero sabes que deberían estar ahí
- Columnas o datos están duplicados
- La ortografía es inconsistente

- Orden en los nombres es inconsistente
- Formatos en las fechas son inconsistentes
- No se especifican unidades
- Las categorías fueron mal elegidas
- Nombres de los campos son ambiguos
- El origen no está documentado
- Hay valores sospechosos
- La data es muy burda
- Los totales difieren de los agregados publicados
- La hoja de cálculo tiene 65536 filas
- La hoja de cálculo tiene fechas en 1900, 1904, 1969 o 1970
- El texto fue convertido a números
- Números que fueron guardados como texto

Cuestiones que deberías resolver tú mismx

- El texto es confuso
- Las líneas finales son confusas
- Los datos están en PDF
- Los datos son demasiado granulares
- Los datos entraron fueron capturados por humanos
- Los datos están mezclados con formatos y anotaciones
- Se hicieron sumatorias tomando en cuenta datos faltantes
- La muestra no es aleatoria
- El margen de error es demasiado alto
- El margen de error es desconocido
- La muestra está sesgada
- Los datos fueron editados manualmente
- Se distorsionan los datos por la inflación
- Variaciones naturales/de temporada distorsionan los datos
- La escala de tiempo ha sido manipulada
- El contexto o las referencias han sido manipuladas

Problemas que un tercero experto debería ayudarte a solucionar

- El autor no es confiable
- El procesos de recolección es opaco
- Los datos suponen una precisión irreal
- Hay valores atípicos inexplicables
- Un índice oculta variaciones subyacentes
- Hay P-hacking en los resultados
- La ley de Benford falla
- Demasiado bueno para ser verdad

Problemas que un programador debería ayudarte a resolver

- Los datos están agregados por la categoría o geografía errónea
- Los datos están en documentos escaneados

Lista detallada de problemas

Problemas que tu fuente debería resolver

Valores faltantes

Cuidado con los valores en blanco o “*null*” en cualquier *dataset*, a menos que estés seguro de lo que significan. Si los datos son anuales, ¿el dato para ese año no fue levantado? ¿Si es una encuesta, algún encuestado se rehusó a contestar la pregunta?

En cualquier momento en que estés trabajando con datos faltantes deberías preguntarte: “¿conozco el significado de la ausencia de este valor?” Si la respuesta es negativa, deberías preguntarle a tu fuente.

Hay datos faltantes que fueron reemplazados con ceros

Peor que un dato faltante es el uso de un valor arbitrario en su lugar. Esto puede ser el resultado de un humano que no esté pensando en las consecuencias de ese uso o puede suceder como resultado de un proceso automatizado que simplemente no sabe cómo manejar valores nulos. En cualquier caso, si ves ceros en una serie de números deberías preguntarte si esos valores corresponden realmente al número 0 o más bien, corresponden al significado “nada”. (-1 también se usa a veces así). Si no estás seguro, pregúntale a tu fuente.

La misma precaución debería valer para otros valores no-numéricos donde un 0 pueda ser representado de otra manera. Por ejemplo, un falso 0 para una fecha suele ser representado como 1970-01-01T00:00:00Z o 1969-12-31T24:59:59Z, que es el [comienzo del registro de tiempo en Unix](#). Un falso 0 para una ubicación puede ser representado como 0°00'00.0"N+0°00'00.0"E o simplemente 0°N 0°E, que es un punto en el Océano Atlántico justo al sur de Ghana, frecuentemente llamado [Null Island](#).

Ver también:

- Hay valores sospechosos
- La hoja de cálculo tiene fechas en 1900, 1904, 1969 o 1970

Faltan datos que sabes deberían estar ahí

A veces hacen falta datos y no lo puedes saber simplemente con el *dataset* mismo. Pero aún así puedes saberlo porque sabes de qué se supone que son los datos. Si tienes un *dataset* que cubre Estados Unidos de América entonces puedes asegurarte de que los 50 estados están representados. (Y no olvides los territorios, 50 no es el número correcto si no incluyes Puerto

Rico). Si estás trabajando con un *dataset* de jugadores de béisbol asegúrate de que tiene el número de equipos que tú esperas. Verifica que unos pocos jugadores que tú conozcas estén incluidos. Confía en tu intuición si algo parece faltar y vuelve a verificar con tu fuente. El universo de tus datos puede ser más pequeño de lo que tú crees.

Filas o valores que están duplicados

Si la misma fila aparece en tu *dataset* más de una vez, deberías averiguar por qué. A veces no necesita ser una fila entera. Algunos datos de financiamiento de campañas incluyen “correcciones” que usan los mismos identificadores únicos que la transacción original. Si no sabías eso, entonces cualquier cálculo que hayas hecho con los datos sería incorrecto. Si hay algo que parezca debe ser único, verifica que lo sea. Si descubres que no lo es, pregunta a tu fuente por qué.

La ortografía es inconsistente

Los dedazos son una de las maneras más obvias de saber si los datos se compilaron a mano. No sólo te fijas en los nombres de la gente, ese es uno de los sitios donde más difícil es hallar dedazos. En lugar de esto, busca lugares donde los nombres de estados o ciudades no sean consistentes. (Los Angeles es un error muy común).

Si encuentras errores de este tipo, puedes estar seguro de que los datos fueron compilados o editados a mano y esa es razón suficiente para guardar cierto escepticismo. Los datos editados a mano son los más proclives a fallas. Esto no significa que no deberías usarlos pero entonces deberías corregirlos manualmente o publicarlos como errores en tu reporte.

La herramienta de [Open Refine](#) para [agrupar texto](#) puede ayudarte a hacer ese proceso sencillo y eficiente al sugerir coincidencias cercanas entre valores inconsistentes en una columna (por ejemplo, igualando Los Angeles con Los Angeles). Cérciorate, no obstante, de [documentar los cambios que haces](#), de modo que garantices un [buen origen de los datos](#).

Ver también:

- Datos fueron introducidos por humanos

El orden de las palabras es inconsistente

¿Entres tus datos están nombres de Oriente Medio o Asia del Este? ¿Estás seguro de que los apellidos están en los mismos lugares? ¿Existe cualquier posibilidad de que alguien en tu set de datos [use un monónimo](#)?

Estas son la clase de cosas en que se equivocan habitualmente quienes hacen datos. Si estás trabajando con una lista de nombres éticamente diversos —que es prácticamente cualquier lista de nombres— entonces deberías hacer al menos una revisión somera antes de asumir que tomar en cuenta las primeras columnas nombre y apellido te dará algo que es apropiado publicar.

Formatos de fecha inconsistentes

¿Qué fecha es en septiembre?:

- 10/9/15
- 9/10/15

Si la primera fue escrita por un latinoamericano o europeo y la segunda por un estadounidense, entonces [ambas lo son](#). Si no conoces la historia de los datos no puedes estar seguro. Averigua de dónde provinieron tus datos y cerciérate de que fue creada por personas del mismo continente.

- Datos ingresados por humanos
- Origen de los datos no está documentado

Las unidades no están especificadas.

Ni el peso ni el costo transmiten ningún tipo de información sobre la unidad de medida. No te apresures a asumir que los datos producidos en Estados Unidos están en libras y dólares. Los datos científicos a menudo están en sistema métrico decimal. Algunos precios extranjeros pueden estar en su propia moneda local. Si los datos no explicitan sus unidades, regresa a tu fuente y hálalos. Incluso si explicita sus unidades, manténte prevenido sobre significados que puedan haber cambiado con el tiempo. Un peso en 2010 no es un peso de hoy. Y una tonelada corta no es una tonelada imperial ni una tonelada, a secas.

Ver también:

- Los nombres de los campos son ambiguos
- La inflación distorsiona los datos

Las categorías fueron mal elegidas

Ten cuidado con valores que se supone validen como verdadero o falso, pero que no lo hagan. Este es usualmente el caso con encuestas donde el que la gente se rehúse a contestar o no dio respuesta se incluyan como valores válidos y con significado.

Otro problema común es el uso de cualquier tipo de categoría. Si las categorías en un set de datos son un montón de países y “otros” ¿eso qué significa? ¿Significa que la persona que compiló los datos no sabía la respuesta correcta? ¿Estaban en aguas internacionales? ¿Expatriados, refugiados?

Las malas categorías también pueden excluir datos artificialmente. Este es frecuentemente el caso con estadísticas de crimen. El FBI ha definido el crimen de violación en una variedad de maneras a lo largo del tiempo. De hecho, han hecho tan mal trabajo categorizando la violación que muchos criminólogos arguyen que esas estadísticas ni siquiera deberían ser usadas. Una mala definición puede significar que un crimen sea contabilizado en una categoría distinta a la que esperas o que no sea contabilizada del todo. Manténte excepcionalmente prevenido sobre este problema cuando trabajes con temas donde las definiciones tienden a ser arbitrarias, tal como ocurre con la raza o etnicidad.

Los nombres de los campos son ambiguos

¿Qué es una residencia? ¿Es el lugar donde vive alguien o el lugar donde paga sus impuestos? ¿Es una ciudad o un condado? Los nombres de los campos en las bases de datos nunca son tan

específicos como nos gustaría, pero es necesario especial cuidado con aquellos que obviamente significan dos o más cosas. Incluso si infieres válidamente lo que esos datos se supone signifiquen, esa ambigüedad pudo haber causado que la persona que compiló los datos haya ingresado el valor incorrecto.

El origen de los datos no fue documentado

Los datos son creados por una variedad de individuos y organizaciones que incluyen empresas, gobiernos, OSCs y gente loca con teorías insostenibles. Los datos son reunidos en muchas maneras diferentes, las cuales incluyen encuestas, sensores y satélites. Puede ser escrita en máquina o garabateada a mano. Saber de dónde provienen tus datos puede darte abrirte enormemente la percepción sobre sus límites.

Los datos de encuestas, por ejemplo, rara vez son exhaustivos. Los sensores tienen diferencias en precisión. Los gobiernos usualmente no están inclinados a dar información sin sesgos. Los datos de una zona de guerra pueden tener un sesgo geográfico importante debido al peligro de cruzar líneas de combate. Para empeorar esta situación, estas fuentes distintas entre sí están habitualmente encadenadas. Los analíticas de políticas públicas frecuentemente redistribuyen los datos que obtienen de gobiernos. Los datos que fueron escritos por una doctora pueden haber sido tecleados por un enfermero. Cada paso de esa cadena es una oportunidad para el error. Sabe siempre de dónde vienen tus datos.

Ver también:

- Unidades no especificadas

Hay valores sospechosos

Si ves cualquiera de estos valores en tus datos, trátalos con abundancia de cuidado:

Números:

- 65,535
- 2,147,483,647
- 4,294,967,295
- 555-3485
- 99999 (o cualquier otra secuencia de nueves)
- 00000 (o cualquier otra secuencia de ceros)

Fechas:

- 1970-01-01T00:00:00Z
- 1969-12-31T23:59:59Z
- 1ero de enero de 1900
- 1ero de enero de 1904

Ubicaciones:

- 0°00'00.0"N+0°00'00.0"E o, simplemente, 0°N 0°E
- Código postal estadounidense 12345 (Schenectady, New York)

- Código postal 90210 (Beverly Hills, California)

Cada uno de estos números indica un error en particular, cometido ya sea por un humano o una computadora. Si los ves, asegúrate de que signifiquen realmente lo que crees que significan.

Ver también:

- La hoja de cálculo tiene 65536 filas
- La hoja de cálculo tiene fechas en 1900, 1904, 1969 o 1970

Los datos son muy burdos

Tienes estados y necesitas países. Tienes empleadores y necesitas empleados. Te dieron años, pero quieres los meses. En muchos casos, obtienes datos que han sido agregados demasiado para nuestros propósitos.

Los datos usualmente no pueden ser desagregados una vez que fueron mezclados. Si te dieron datos que son muy burdos, necesitarás pedirle a tu fuente algo más específico. Puede que no lo tengan. Si lo tienen, es posible que se declaren incapacitados a ofrecértelo o que simplemente no quieran. Hay muchos sets de datos federales a los que no hay acceso a nivel local para proteger la privacidad de los individuos que podrían estar identificados de manera única por esos mismos datos. (Por ejemplo, la única persona somalí que vive en Texas occidental). Lo único que puedes hacer, es preguntar.

Una cosa que nunca deberías hacer es dividir el valor anual entre 12 y llamarlo el “promedio por mes”. Si no conoces la distribución de los valores, ese número no tendrá significado.

(Quizá todas las instancias ocurrieron en un mes o una estación. Quizá los datos siguen una tendencia exponencial en lugar de una lineal). Es equivocado. No lo hagas.

Ver también:

- Los datos son demasiado granulares
- Los datos se han agregado a las categorías o geografías equivocadas

Los totales difieren de los agregados publicados

Imagina que después de una larga pelea con solicitudes de acceso a la información recibes una lista “completa” de incidentes en el uso de la fuerza policiaca. La abres y descubres que tiene 2 mil 467 filas. Genial, hora de publicarlo. No tan rápido. Antes de que publiques cualquier cosa sobre el set de datos averigua la última vez que el jefe de policía fue entrevistado sobre el uso público de la fuerza.

Podrías hallar que en una entrevista que dio hace seis semanas dijo que se había usado la fuerza “menos de 2 mil veces” o que dijo un número específico que no cuadra con tus datos.

Este tipo de discrepancias entre las estadísticas publicadas y la *raw data* puede ser una gran fuente de pistas. Con frecuencia, la respuesta puede ser simple. Por ejemplo, los datos que te fueron dados pueden no cubrir el mismo periodo de tiempo del que el jefe de policía está

hablando. Pero algunas veces simplemente los descubrirás mintiendo. De cualquier modo, debes asegurarte que los números publicados empatan con los totales de los datos que te dieron.

La hoja de cálculo tiene 65536 filas

El número máximo de filas que una vieja hoja de cálculo de Excel permitía era de 65 mil 536. Si recibes un dataset con ese número de filas es casi seguro que recibiste datos truncos. Llama de vuelta y pide el resto. Versiones de Excel más recientes permitían 1 millón 48 mil 576 filas, así que es menos probable que trabajes con datos que le peguen a ese límite.

Hojas de cálculo que tienen fechas en 1900, 1904, 1969 o 1970

Por razones que van más allá de ser oscuras, la fecha predeterminada de Excel, desde donde cuenta el resto de ellas, es el 1 de enero de 1900, a menos que estés usando Excel en una Mac, en cuyo caso es el 1 de enero de 1904.

Hay una diversidad de formas en que los datos en Excel puedan ser ingresados o calculados de manera incorrecta y terminen por dar una de estas dos fechas. Si las ves entre tus datos, probablemente se trate de un problema.

Muchas bases de datos y aplicaciones tendrán por lo general la fecha 1970-01-01T00:00:00Z o 1969-12-31T24:59:59Z, que corresponde al [comienzo del registro del tiempo Unix](#). En otras palabras, esto es lo que sucede cuando un sistema trata de mostrar un valor nulo o un valor de 0 como una fecha.

Texto que fue convertido a números

No todos los numerales son números. Por ejemplo, la oficina de censos estadounidense usa códigos numéricos para identificar cada sitio en el país. Estos códigos numéricos son de distintas longitudes. No obstante, *no son números*. 037 es el código para el condado de Los Angeles, no el número 37. Los numerales 37 son, sin embargo, un código de censo válido: para Carolina del Norte. Excel y otras hojas de cálculo con frecuencia cometerán el error de asumir que los numerales son números y les quitarán el cero que les antecede. Esto puede causar toda clase de problemas si tratas de convertir el archivo a otra extensión o cruzarlo con otro set de datos. Cuidado con los datos en los que esto ha ocurrido antes de dárteles a ti.

Números que fueron guardados como texto


Cuando trabajas con hojas de cálculo, los números pueden ser almacenados como texto por un formato no deseado.

Lo anterior ocurre con frecuencia cuando una hoja de cálculo está optimizada para presentar datos más que para ser reutilizada. Por ejemplo, en vez de representar un millón de dólares con el número «1000000», una celda puede contener el *string* (secuencia que no es un número) “1,000,000” o “1 000 000” o “USD 1,000,000” con el formato de comas, unidades y espacios ingresados como caracteres. Excel se puede hacer cargo de ciertos casos simples con sus funciones integradas, pero usualmente necesitarás usar fórmulas para sacar dejar puros caracteres hasta que las celdas estén lo suficientemente limpias para ser reconocidas como

números. Una buena práctica es almacenar números sin formato e incluir información contextual en los nombres de columnas o metadatos.

Problemas que deberías resolver tú mismx

El texto es confuso

Todas las letras son representadas por las computadoras como números. Los problemas de codificación surgen cuando el texto es representado por un específico grupo de números (llamado “*encoding*” o “codificación”) el cual desconoces. Esto origina un fenómeno llamado [mojibake](#), donde el texto en tus datos parece basura, o se ve así: .

En la gran mayoría de los casos, tu editor de textos u hoja de cálculo averiguará automáticamente el cifrado correcto, sin embargo, si no funciona podrías publicar el nombre de alguien más con un caracter raro en medio. Tu fuente debe ser capaz de decirte la codificación de tus datos, en caso de que no pueda, hay maneras bastante confiables de de adivinarla. Pregunta a un programador.

Los espacios al final de la línea están mal codificados

Los textos y archivos de datos en texto, como los .csv, usan caracteres invisibles para representar una nueva línea. Windows, Mac y Linux han tenido un desacuerdo histórico sobre cuáles deberían ser estos caracteres. Intentar abrir un programa guardado en un sistema operativo con otro a veces provoca que Excel u otras aplicaciones fallen al identificar estas nuevas líneas.

Normalmente, esto se resuelve fácilmente abriendo un archivo con un editor de texto general y volviéndolo a guardar. Si el archivo es excepcionalmente grande deberías considerar usar una herramienta con interfaz de línea de comandos o buscar la ayuda de un programador. Puedes leer más acerca de este problema [aquí](#).

Los datos están en PDF

Una gran cantidad de data -especialmente la data de gobierno- son sólo accesibles en formato PDF. Si tienes datos en un PDF, existen varias opciones para extraerlos. (Sin embargo si tienes los archivos escaneados ese es un problema diferente). Una excelente herramienta es [Tabula](#). De cualquier forma, si tienes Adobe Creative Cloud entonces tienes acceso a Acrobat Pro, el cual tiene una excelente herramienta para exportar tablas de PDF a Excel.

Vea también:

- Los datos están en documentos escaneados

Los datos son demasiados granulares

Esto es lo opuesto a [Los datos son demasiado burdos](#). Esta vez tienes condados, pero quieres estados, o tienes meses, pero quieres años. Por fortuna, normalmente esto es muy sencillo.

Los datos pueden ser agregados usando tablas dinámicas de Excel o Google Docs, usando una base de datos SQL o con código hecho para la ocasión. Las tablas dinámicas son una fabulosa herramienta que cualquier reportero debería aprender, pero tienen sus límites. Para sets de datos

excepcionalmente largos o para agregar datos de grupos inusuales, pregúntale a un programador, ellos pueden crear una solución que sea fácil de verificar y reutilizar.

Ver también:

- Los datos son muy toscos (poco granulares)
- Los datos fueron agregados a las categorías o geografías equivocadas

Los datos fueron capturados por humanos

Los datos capturados por humanos son un problema tan común que sus síntomas son mencionados en al menos 10 otros problemas que son descritos aquí. No hay peor forma de arruinar los datos que dejar que una sola persona los capture, sin validación alguna. Por ejemplo, una vez adquirí la base completa de licencias para perros en el condado de Cook, Illinois. En vez de pedir a la personas que eligieran una raza de la lista al registrar a sus perros, los creadores de la lista les dieron un espacio de texto para escribirla. El resultado de esa base de datos obtuvo al menos 250 maneras diferentes de la palabra Chihuahua. Incluso con la mejores herramientas, el desastre en el que estaban esos datos no se podía arreglar. Esto no es tan importante con datos de perros, pero no quieres que esto suceda con soldados heridos o tableros de cotizaciones. Toma tus precauciones con datos capturados por humanos.

Los datos están mezclados con formateo y anotaciones

Representaciones complejas de datos como HTML o XML permiten una clara separación entre datos y formato, pero no es ese el caso por representaciones de datos tabulares comunes, como los de las hojas de cálculo. Aún así, algunos tratan de hacerlo. Un problema común con los datos así provistos es que las primeras filas de datos serán descripciones o notas de los datos más que encabezados de columna o datos. Una llave o diccionario de datos puede también estar a media hoja. Algunos encabezados de filas pueden repetirse, o la hoja de cálculo puede incluir muchas tablas (que pueden tener diferente longitud en sus encabezados de columnas) una tras la otra en la misma página en lugar de estar separadas en distintas páginas.

En todos estos casos la solución centra es simplemente identificar el problema. Obviamente cualquier análisis en una hoja con este tipo de problemas fallará, algunas veces por razones que no son obvias. Al mirar nuevos datos por primera vez siempre es una buena idea asegurarse de que no hay filas extra de encabezados u otros caracteres de formato insertos entre los datos.

Los agregados fueron calculados con valores que faltan

Imagina un dataset con 100 filas y una columna llamada costo. En 50 de las columnas el costo está en blanco. ¿Cuál es el promedio de esa columna? ¿Es una suma del costo entre 50 o una suma del costo entre 100? No hay respuesta definitiva. En general, si vas a calcular agregados en columnas con datos que faltan, puedes hacerlo con confianza dejando fuera las filas faltantes, ¡pero ten cuidado de no comparar agregados de dos columnas en las que dos filas distintas fueron valores faltantes! En algunos casos los valores que faltan pueden también ser interpretados legítimamente como 0. Si no estás segura, pregúntale a una experta, o simplemente no lo hagas.

Este es un error que puedes cometer en tu análisis, pero es también un error que otros cometen y te lo pasan a ti, así que observa con cuidado si en tus datos ya vienen agregados.

Ver también:

- Valores que faltan
- Ceros que reemplazan valores faltantes

La muestra no es aleatoria

Un error de muestreo no aleatorio ocurre cuando una encuesta u otro dataset de muestra intencional o accidentalmente no cubre la población entera. Esto puede ocurrir por una diversidad de razones que van de la hora del día al lenguaje que usa la persona que responde y es una fuente común de error en la investigación sociológica. También puede pasar por razones menos obvias, como cuando el investigador piensa que tiene una base de datos completa y escoge sólo trabajar con una parte de ella. Si la base de datos original estuviera incompleta por cualquier razón la muestra también sería incorrecta. Lo único que puede hacer para escoger una muestra no aleatoria es evitar el uso de estos datos.

Ver también:

- Los datos están sesgados

El margen de error es demasiado amplio

No conozco ningún otro problema que cause más errores de información que el uso irreflexivo de números con márgenes de error demasiado amplios (MOE, por sus siglas en inglés). MOE está asociado normalmente con datos que provienen de encuestas. El lugar más habitual en que un reportero lo encuentra es al usar datos de encuestas o de la Encuesta sobre la Comunidad Estadounidense de la Oficina del Censo de los Estados Unidos. El MOE es una medida de rangos de posibles valores verdaderos. Se puede expresar como un número (400 +/- 80) o como un porcentaje de la totalidad (400 +/- 20%). Cuanto menor es la población relevante, mayor será el MOE. Por ejemplo, de acuerdo con los estimados de la ACS 2014 a 5 años, el número de asiáticos que viven en Nueva York es 1.106.989 +/- 3,526 (0,3%). El número de filipinos es 71.969 +/- 3.088 (4,3%). El número de samoanos es de 203 +/- 144 (71%)

Los dos primeros números pueden reportarse con seguridad. El tercer número nunca se debe utilizar para publicaciones. No hay una regla sobre cuándo un número no es lo suficientemente preciso para utilizar, pero como regla general, se debe tener cuidado al usar cualquier número con un MOE mayor al 10%.

Vea también:

El margen de error es desconocido

El margen de error es desconocido.

A veces el problema no es que el margen de error es demasiado grande, es que nadie se molestó en averiguar cuál era en primer lugar. Este es un problema con las encuestas no científicas. Sin

computarizar un MOE, es imposible saber la exactitud de los resultados. Como regla general, cada vez que tenga los datos que son de una encuesta, pregunte por el MOE. Si la fuente no se puede decir, no vale la pena utilizar utilizar dichos datos para ningún análisis que sea serio.

Ver también:

- El margen de error es demasiado amplio

La muestra está sesgada

Como una muestra que no es aleatoria, una muestra sesgada resulta de la falta de cuidado sobre cómo se ejecuta una muestra. O, es intencionalmente engañosa. Una muestra puede ser sesgada porque se llevó a cabo en Internet y las personas más pobres no utilizan internet con tanta frecuencia como los ricos. Las encuestas deben ser ponderadas cuidadosamente para asegurarse que cubren segmentos proporcionales de cualquier población que pudiera sesgar los resultados. Es casi imposible hacer esto perfectamente, por lo que se hace a menudo mal.

See a:

- La muestra no es aleatoria

Los datos fueron editados manualmente

La edición manual es casi el mismo problema el que los datos sean capturados por humanos, excepto que ocurre *a posteriori*. De hecho, los datos son editados manualmente en un intento de arreglar datos que fueron originalmente capturados por humanos. Se empiezan a filtrar problemas cuando la persona que está editando no tiene total conocimiento de los datos originales. Alguna vez atestigüé a alguien “corregir” espontáneamente un nombre en una base de datos de Smit a Smith. ¿Era realmente el nombre de esa persona Smith? No lo sé, pero sé que ahora ese valor es un problema. Sin un registro de ese cambio, es imposible verificar cuál de los dos debería ser.

Estos problemas con la edición manual son una de la razones por las que deberías asegurarte de que tus datos tengan un origen bien documentado. La falta de éste puede ser un buen indicador de que alguien haya estado jugando con ella. Los académicos y analistas de políticas públicas obtienen datos del gobierno con frecuencia, los manosean y luego se los redistribuyen a periodistas. Sin ningún registro de los cambios que hacen es imposible saber si fueron justificados. Siempre que sea posible trata de acceder a la *fuentes primaria* o al menos a la versión más antigua disponible y haz tu propio análisis a partir de ella. A lack

Ver también:

- El origen no está documentado
- Los datos fueron capturados por humanos

La inflación distorsiona los datos

La inflación monetaria implica que con el tiempo el valor del dinero cambia. No hay manera de saber si los números fueron ajustados a la inflación sólo con mirarlo. Si obtienes datos y no estás seguro de que hayan sido ajustados, verifícalo con tu fuente. Si no se ha ajustado entonces

probablemente quieras hacerlo tú mismx. Este [ajustador de la inflación](#) es un buen sitio por donde comenzar.

Ver también:

- Variaciones naturales o de temporada distorsionan los datos

Variaciones naturales/de temporada distorsionan los datos

Muchos tipos de datos fluctúan naturalmente debido a algunas fuerzas subyacentes. El ejemplo más conocido de lo anterior son las variaciones que fluctúan por temporada. Los economistas han desarrollado una varios métodos para compensar esta variación. Los detalles de esos métodos no son particularmente importantes, pero sí lo es que sepas que si los datos que estás usando fueron “ajustados a la temporada”. Si no fue así y quieres comparar empleos de mes a mes seguramente querrás que tu fuente te de datos ajustados. (Ajustarlos tú mismo es mucho más complejo que ajustar la inflación).

Ver también:

- La inflación distorsiona los datos

La escala de tiempo fue manipulada

Un fuente puede distorsionar el mundo accidental o intencionalmente al darte datos que comienzan o se detienen en determinado lapso de tiempo. Para ver un ejemplo potente nótese la “ola de crimen” ampliamente reportada en 2015. No hubo ninguna “ola de crimen”. Lo que hubo fue una serie de repuntes in ciudades en particular al ser comparadas con el último par de años. Si los periodistas hubiesen examinado un marco temporal más amplio habrían descubierto que los crímenes violentos fueron mayores virtualmente en cualquier sitio de EUA diez años atrás. Y veinte años antes sumaban casi el doble.

Si tienes datos que cubren un marco temporal limitado trata de evitar cálculos con el primer periodo de tiempo que tienes. Si empiezas unos años (o meses o días) más adelante puedes tener certeza de que no estarás haciendo una comparación que sería invalidada con la adición de un solo punto de datos.

Ver también:

- El marco de referencia fue manipulado

El marco temporal fue manipulado

Las estadísticas de crimen son manipuladas frecuentemente por razones políticas al comparar contra un año en el que el crimen era alto. Lo anterior puede expresarse o bien como un cambio (disminuyó 60% desde 2004) o como un índice (40, donde 2004=100). En cualquiera de estos casos, 2004 puede ser o no un año apropiado para la comparación. Pudo haber sido un año inusualmente alto en tasa de crímenes. Lo mismo ocurre cuando se comparan lugares. Si quiero hacer ver mal a algún país, simplemente expreso los datos de ese país contra los cuales cualquier otra nación tiene mejores datos.

Este problema tiende a brotar de la nada en temas donde el sujeto tiene un fuerte prejuicio de confirmación (“¡Tal como lo creí, el crimen ha aumentado!”) Cuando sea posible trata de comparar las tasas a partir de diversos puntos para ver cómo cambian los números. Y, por lo que más quieras, *no uses esta técnica tú mismx* para hacer un argumento que creas que es importante. No tiene excusa.

Ver también:

- El marco temporal de referencia fue manipulado

Problemas que deberían ayudar a resolver terceros

El autor no es confiable

En ocasiones los únicos datos que tienes son de una fuente de la que en realidad desconfías. En algunos casos, está bien. Los únicos que saben cuántas armas se manufactura son quienes hacen armas. No obstante, si tienes datos cuestionables siempre verifícalos con algún otro experto. Mejor aún, verifícalo con dos o tres . No publiques datos de una fuente sesgada a menos que tengas evidencia sustancial que la corrobore.

El proceso de recolección es opaco

Es muy sencillo que se introduzcan falsas suposiciones, errores o incluso crasas falsedades en el proceso de recolección de datos. Por esta razón es importante que los métodos usados sean transparentes. Será raro que sepas exactamente cómo fueron levantados los datos de un dataset, pero hay indicadores rojos: que las cifras incluidas sean de una precisión irreal o que los datos sean demasiado buenos para ser verdad.

A veces el origen de los datos puede simplemente ser sospechoso: ¿realmente los académicos x y y entrevistaron 50 miembros activos de pandillas en el sur de Chicago? Si la manera en que los datos fueron reunidos parece cuestionable y tu fuente no puede ofrecerte un origen a prueba de balas, entonces deberías verificar siempre con otro experto que los datos hayan podido ser levantados razonablemente en la forma en que está descrito (en la documentación).

Ver también:

- El origen no está documentado
- Los datos son de una precisión irreal
- Los datos son demasiado buenos para ser reales

Los datos son de una precisión irreal

Fuera del mundo de la ciencia dura, pocas cosas son medidas rutinariamente con mayor precisión que la de dos puntos decimales. Si te llega un dataset que pretende mostrar las emisiones de una fábrica con siete decimales, eso de alta que éstas fueron estimadas a través de otros valores. Por sí mismo, eso puede no ser un problema, pero es importante ser transparentes acerca de estimados. Usualmente están equivocados.

Hay valores atípicos inexplicables

Recientemente cree un dataset del tiempo que toman diferentes mensajes para llegar a diferentes destinos a través de internet. Todos los tiempos estaban en el rango de los 0.05 a los 0.8 segundos, excepto tres de ellos. Esos estaban por encima de los 5 mil segundos. Esta es un signo significativo de que algo estaba mal con la producción de los datos. En este caso en particular, un error en el código que escribí provocó algunas fallas para seguir contando mientras todos los otros mensajes eran enviados y recibidos.

Valores atípicos como estos pueden fastidiar tu estadística dramáticamente —especialmente si estás usando promedios—. (Probablemente deberías estar usando medianas). Cuando tienes un nuevo dataset es una buena idea echar un vistazo a los valores mínimos y mayores y asegurarte de que están en un rango razonable. Si los datos lo justifican puedes también hacer un análisis estadístico más riguroso usando [desviaciones estándar](#) o [desviaciones medias](#).

Como un beneficios adicional de hacer esto, los valores atípicos son generalmente una buena manera para encontrar buenos encabezados. Si realmente hubiera un sólo país donde enviar un mensaje por internet tomase 5 mil veces el tiempo que en el resto, esa sería una gran historia.

Un índice enmascara variaciones subyacentes

Los analistas que quieren seguir la tendencia de algún tema crean con frecuencia índices de varios valores para seguir su progreso. No hay nada intrínsecamente incorrecto al usar un índice. Pueden tener gran poder de explicación. No obstante, es importantes ser cauteloso con índices que combinan medidas dispares.

Por ejemplo, el [Índice de Desigualdad de Género de las Naciones Unidas](#) combina varias mediciones relacionadas con el progreso de las mujeres hacia la igualdad. Una de las medidas usadas en el Índice es la “representación de mujeres en el parlamento”. Dos países que tienen leyes vinculantes para representación de género en los parlamentos son Pakistán y China. Como resultado, estos dos países tienen un desempeño mucho mejor que países que tienen condiciones similares en el resto de los parámetros. ¿Es esto justo? No importa, realmente, porque es confuso para cualquiera que no sepa de este factor. El GII e índices similares deberían siempre usarse con un análisis cuidadoso para cerciorarse de que sus variables subyacentes no inclinen el índice en maneras inesperadas.

Hay p-hacking en los resultados

P-hacking es la maniobra de alterar datos, cambiar estadísticas o reportar selectivamente los resultados de un análisis, de manera intencional, con el objetivo de mostrar hallazgos estadísticamente significativos. Algunos ejemplos: dejar de recolectar datos una vez que tienes un resultado estadísticamente significativo, borrar observaciones para obtenerlo o implementar numerosos análisis pero sólo publicar aquellos que te den resulten significativos. Ha habido bastante [buen reporte](#) sobre este problema.

Si vas a publicar los resultados de tu estudio necesitas comprender qué son los «valores p», qué significan y entonces hacer una decisión informada sobre si los resultados que estás usando valen

la pena. Una gran cantidad de estudios basura son publicados en grandes revistas científicas porque los periodistas no entienden lo que son los «valores p».

Ver también:

- El margen de error es demasiado amplio

La ley de Benford falla

[La ley de Benford](#) es una teoría que propone que los dígitos pequeños (1, 2, 3) aparecen al comienzo de un número con mucha mayor frecuencia que números más grandes (7,8,9).

Teóricamente, la Ley de Benford puede usarse para detectar anomalías en conteos o resultados de elecciones, aunque en la práctica se aplica de manera equivocada con frecuencia. Si sospechas que un dataset fue creado o modificado para engañar a la audiencia, la Ley de Benford puede ser un primer test excelente, pero deberías siempre verificar tus resultados con un experto antes de concluir que tus datos fueron manipulados.

Demasiado bueno para ser verdad

No hay ningún dataset global de opinión pública. Nadie sabe el número exacto de personas que viven en Siberia. Las estadísticas de crimen no son comparables más allá de las fronteras. El gobierno de Estados Unidos de América There is no global dataset of public opinion. Nobody knows the exact number of people living in Siberia. Crime statistics aren't comparable across borders. The US government is not going to tell you how much fissile material it keeps on hand.

Beware any data that purport to represent something that you could not possibly know. It's not data. It's somebody's estimate and it's probably wrong. Then again... it could be a story, so ask an expert to check it out.

Problemas que un programador debería ayudarte a resolver

Los datos están agregados en categorías o geografías incorrectas

A veces tus datos tienen el suficiente nivel de detalle (ni demasiado burdos ni demasiado granulares), pero fueron agregados en diferentes categorías de las que tú necesitabas. Un ejemplo clásico son los datos que son agregados por códigos postales pero que tú preferirías que fueran agregados por colonias y ciudad. En muchos casos esta problema es imposible de resolver sin tener datos más granulares de tu fuente, pero a veces los datos pueden ser mapeados proporcionalmente de un grupo al otro. Esta tarea se debe emprender siempre en el entendido de que un margen de error puede ser introducido en el proceso. Si tienes datos que fueron agregados en grupos equivocados, pregunta a un programador si es posible reagregarlos.

Ver también:

- Los datos son muy burdos
- Los datos son demasiado granulares
- El margen de error es demasiado amplio

Los datos están en documentos escaneados

Gracias a las leyes de transparencia y acceso a la información, es frecuente que los gobiernos sean emplazados a proporcionarte datos –incluso cuando en realidad no quieren hacerlo–. Una táctica muy común en estos casos es que te den hojas escaneadas o fotografías de las páginas. Se puede tratar de hecho de archivos de imagen, o, mucho más frecuentemente, de un PDF.

Es posible extraer textos de imágenes y transformarlos en datos. Esto se hace a través de un procedimiento que se conoce como OCR (reconocimiento óptico de caracteres, por sus siglas en inglés). El OCR moderno puede tener un porcentaje de precisión cercano al 100%, pero muchas veces esto depende de la naturaleza del documento. En cualquier ocasión en que uses un OCR para extraer datos querrás tener a la mano un proceso de validación de los resultados, que deben acercarse al original.

Hay muchos sitios a los que puedes subir un documento para un proceso OCR, pero también hay herramientas gratuitas que un programador puede ajustar para tus documentos específicos. Pregúntales cuál es la mejor estrategia para los documentos que tú tienes en particular.

Ver también:

- Los datos están en un PDF

Los comentarios están cerrados.