OPSO79-1-UCSH2021

Tipos de datos cuantitativos e introducción a dplyr en R. Bloque teórico.

27/08/2021

Introducción

Introducción

Cuando analizamos datos cuantitativos, representaciones numéricas del mundo social, nos encontraremos con cosas muy diferentes.

Antes de comprender cómo se **producen**, **distribuyen** y **analizan** los datos, revisaremos los distintos tipos de datos cuantitativos que nos encontraremos.

Definiremos cuestiones básicas como que es una variable, que es una observación, cuáles son los conjuntos de datos y tipos de variables a los que comúnmente nos enfrentaremos.

Esto es relevante dado que según el tipo de datos que tengamos dependerá el tipo de análisis a realizar.

Introducción

Los datos rectangulares, con variables listadas en columnas y con filas que refieren a diferentes observaciones, es una producción que demanda tiempo, trabajo y dinero.

Cuando aplicamos encuestas los datos son un producto de la interacción social entre investigador, encuestador y encuestado, pero también pasan por un proceso de **validación** y **ordenamiento**.

Las hojas marcadas por el encuestador deben pasar a una tabla, las respuestas abiertas codificadas, los valores no deben ser contradictorios entre sí ni imposibles, etcétera.

Este proceso de ordenamiento es aún más evidente cuando codificamos documentos como cotizaciones, contratos colectivos, noticias, películas, páginas webs, post en redes sociales, entre otros.

Producción de un orden

Cerca de 1.300 trabajadores de la División Andina de Codelco podrían llegar a estar en huelga si es que se confirma la decisión del Sindicato de Unión Plantas (Suplant) de rechazar las propuestas de la empresa.

Hasta el cierre de esta edición, se esperaba que la empresa presentara una nueva oferta -cerca de las 20:30 horas- para intentar desactivar el conflicto en el marco de la negociación colectiva de la minera, después de que la organización rechazara una propuesta previa, lo que hasta el mediodía de este lunes hacía pensar que la huelga se haría efectiva a partir de las 8 horas de este martes.

Un punto que mantenía trabadas las conversaciones, era el traspaso de beneficios para trabajadores nuevos de la estatal, según explicó el presidente del sindicato.

Esta negociación colectiva se da casi en paralelo a la que llevan el Sindicato Industrial de Integración Laboral (SIIL) y Sindicato Unificado de Trabajadores (SUT) de esa misma división de la estatal, los que ya habían rechazado las últimas ofertas de la compañía y habían comenzado sus paralizaciones el jueves de la semana pasada.

De confirmarse la decisión de Suplant, los trabajadores paralizados en esa faena llegarían hasta las 1.300 personas movilizadas.

Además, se está a la espera de la negociación del Sindicato de Supervisores, el que recibiría este martes la última oferta por parte de la estatal.

Presión en la industria

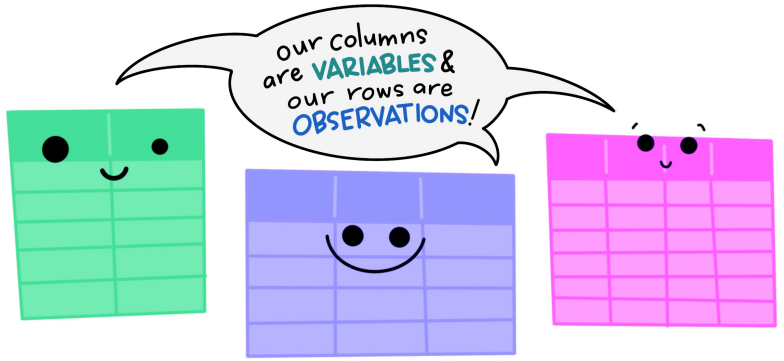
Consultado por los motivos que los alejan de la empresa, el presidente de Suplant, Clodomiro Vásquez, señaló que se pretende eliminar los beneficios en salud de los trabajadores nuevos, entre otros aspectos como la indemnización por año de servicio.

D	E	F	G	H	I	J	K	L
nnot	prensa	yr	mes	inicio	fin	duracion	ddpp	leg
0		2015	11	2015-11-27	2015-12-04	8,0	6,0	1
0		2015	11	2015-11-30	2015-12-11	12,0	10,0	1
0		2015	11	2015-11-30	2015-12-18	19,0	15,0	1
0		2015	11	2015-11-30	2015-12-04	5,0	5,0	1
6	1	2015	11	2015-11-30	2015-12-17	17,0	13,0	2
0		2015	12	2015-12-03	2015-12-14	12,0	10,0	1
0		2015	12	2015-12-03	2015-12-04	2,0	2,0	1
0		2015	12	2015-12-04	2015-12-10	7,0	5,0	1
0		2015	12	2015-12-04	2015-12-04	1,0	1,0	1
0		2015	12	2015-12-04	2015-12-18	15,0	11,0	1
0		2015	12	2015-12-04	2015-12-07	4,0	4,0	1
0		2015	12	2015-12-07	2015-12-16	10,0	8,0	1
0		2015	12	2015-12-07				1
3	4	2015	1	2015-12-09	2016-01-15	38,0	28,0	1
0		2015	12	2015-12-09	2016-01-20	42,0	30,0	1
0		2015	12	2015-12-10	2015-12-14	5,0	5,0	1
2	2	2015	12	2015-12-10	2015-12-11	0,1	0,1	2
2	2	2015	12	2015-12-11	2015-12-13	2,0	2,0	2
1	1	2015	12	2015-12-12	2015-12-12	1,0	1,0	2
1	2	2015	12	2015-12-12				2
0		2015	12	2015-12-14	2015-12-22	9,0	7,0	1
0		2015	12	2015-12-15	2015-12-17	3,0	3,0	1
1	4	2015	1	2015-12-16	2016-01-07	23,0	21,0	1
0		2015	12	2015-12-16	2016-01-06	20,0	16,0	1
1	2	2015	12	2015-12-16	2015-12-16	1,0	1,0	2
184	3	2015	12	2015-12-16	2015-12-20	4,0	4,0	2
0		2015	12	2015-12-17	2015-12-30	14,0	10,0	1
23	3	2015	12	2015-12-17	2016-01-17	27,0	17,0	2
1	1	2015	12	2015-12-17	2015-12-17	1,0	1,0	2
1	1	2015	12	2015-12-18	2015-12-22	4,0	4,0	2
0		2015	12	2015-12-21	2016-01-12	23,0	17,0	1
0		2015	12	2015-12-21	2016-01-07	17,0	13,0	1

Producción de un orden

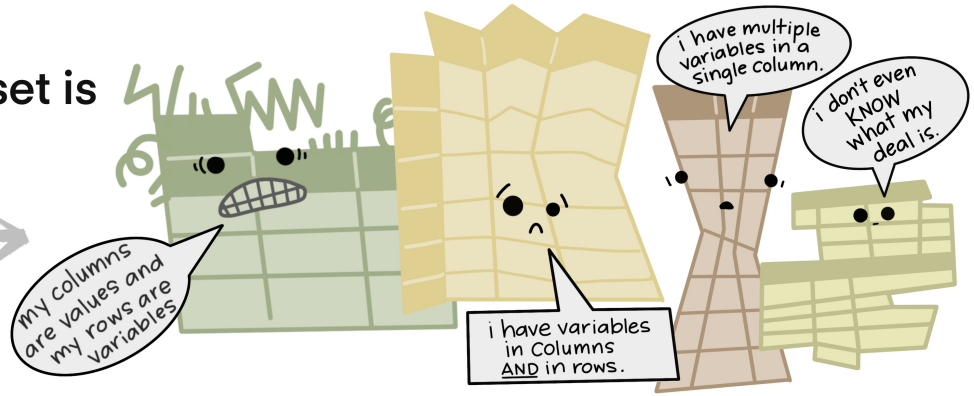
Producción de un orden

The standard structure of tidy data means that "tidy datasets are all alike..."



"...but every messy dataset is messy in its own way."

–HADLEY WICKHAM



El proceso estadísticos

GSBPM

Datos ordenados

Combinación de observaciones y variables.

Cada fila es una observación.

Cada variable es una propiedad de las observaciones.

El término variable hace referencia a que el valor que asume **varía** entre las distintas observaciones.

Los valores pueden ser milímetros de precipitaciones, o respuestas "sí" o "no" a la pregunta de sí llovió.

las variables se pueden clasificar por su **nivel de medida**.

La variable será **categórica** si cada observación pertenece a un set de diferentes categorías (2 o más, pueden tener un orden o no).

La variable será **cuantitativa** si la observación toma valores numéricos que representan diferentes magnitudes de una variable.

Datos ordenados

Aunque ocupemos números para representar categorías como un "sí" (1) o un "no" (2), la variable sigue siendo cuantitativa.

Características claves de las **variables cuantitativas** serán su **centro** y **variabilidad**,

¿cuántas personas iban al día en las protestas de 2019?, ¿Cómo varió la participación desde octubre hasta fin de año?

Para las categóricas será clave saber el número de observaciones en cada categoría.

Existiendo protestas violentas y pacíficas, ¿Cuál es el porcentaje de protestas violentas en 2019?

Variables cuantitativas

Pueden ser **discretas** o **continuas**.

La variable es discreta si sus valores forman un set de número separados (1, 2, 3, 0, 10).

Si tiene un número finito de valores es discreta (hijos, protestas, terremotos, etc.).

Un decimal no tiene sentido (no pueden haber 3,5 terremotos en un año)

La variable es continua si sus valores posibles forman un intervalo (1.2, 3.5, 100.1).

Los valores posibles que puede tomar forman un continuo infinito (peso, ingresos,)

Distribución de una variable

Un paso clave a la hora de entender un conjunto de datos, es conocer la distribución de sus variables.

La distribución describe como las observaciones "caen" a lo largo de un rango posible de valores.

En las variables categóricas los valores posibles son las diferentes categorías. Cada observación cae en una categoría.

Se puede reportar el número de observaciones:

```
table(guaguas::guaguas$sexo)
```

```
##
##      F      M
## 523623 321777
```

Esto es una **tabla de frecuencias** (absolutas). Es un listado de los posibles valores de una variable junto al número de observaciones de cada una.

Distribución de una variable

También podemos observas una tabla de **frecuencias relativas**.

La proporción de observaciones en cada categoría corresponde al número de observaciones en dicha categoría dividido por el total del número de observaciones.

Porcentaje de mujeres en guaguas:

```
523623/(523623+321777)
```

```
## [1] 0.619379
```

Con función:

```
prop.table(table(guaguas::guaguas$sexo))
```

```
##
##        F        M
## 0.619379 0.380621
```

El porcentaje es cuando la proporción se multiplica por 100.

Distribución de una variable

La categoría que concentra la mayor frecuencia se llama la **categoría modal**.

En las cuantitativas también es importante visualizar la distribución, pero cuando el número de valores posibles es muy alto y cada valor lo asume un reducido número de observaciones **no es pertinente**.

Por ejemplo:

```
table(guaguas::guaguas$n)[1:100]
```

```
##
##      1      2      3      4      5      6      7      8      9     10     1
## 524247  91398  42081  25451  17365  12677   9888   7930   6694   5756   487
##     12     13     14     15     16     17     18     19     20     21     2
##   4299   3811   3469   2965   2827   2510   2284   2156   1929   1888   178
##     23     24     25     26     27     28     29     30     31     32     3
##   1606   1450   1445   1338   1181   1147   1079   1022   1020    957    87
##     34     35     36     37     38     39     40     41     42     43     4
##    888    841    825    784    762    729    656    652    703    605    61
##     45     46     47     48     49     50     51     52     53     54     5
##    579    585    555    533    528    521    510    466    428    492    45
```

Distribución de una variable

Para una variable cuantitativa resulta más pertinente observar la **forma** de la distribución, el **centro** y su **variabilidad**.

```
summary(guaguas::guaguas$n)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     1.00     1.00     1.00    25.96     3.00 21448.00
```

En sesión de estadística descriptiva revisaremos en detalle estos conceptos y sus respectivos estadísticos.

Variables categóricas

Las variables categóricas se pueden usar tanto para **ordenar** como para **clasificar**

Asún.

Otras distinciones relevantes

Datos válidos y no válidos

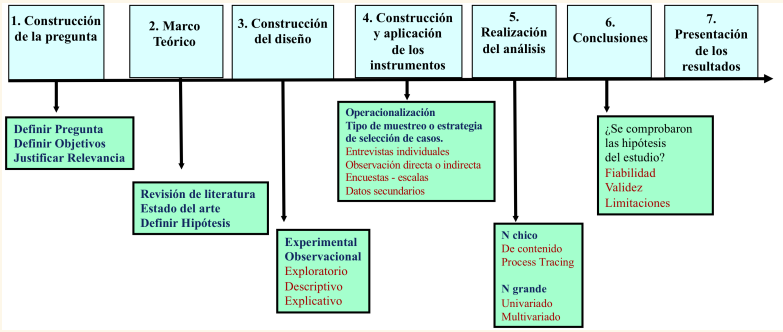
Tabulados v/s microdatos.

Transversal v/s longitudinal

Diseños de investigación

Solo si alcanzamos

El proceso de investigación



El diseño

Plan estructurado de acción y ordenamiento de la situación de investigación, que está orientado a responder empíricamente (evidencia observable) una pregunta de investigación.

- ¿Qué aspecto de la teoría va a ser testeado?
- ¿Qué observaciones se harán para responder a la pregunta de investigación?
- ¿Cómo se levantarán o producirán los datos (observaciones)?
- ¿Cómo se analizará la información recolectada?

Todo buen diseño de investigación busca el mismo objetivo: Sacar conclusiones fundamentadas y relevantes a partir de un correcto tratamiento de la evidencia empírica.

Bibliografía utilizada

Agresti, A. and C. Franklin (2018). *Statistics the Art and Science of Learning from Data*. Pearson Education Limited.

Wickham, H. (2021). *R Para Ciencia de Datos*. URL: https://es.r4ds.hadley.nz/.

Un poco más de R base e introducción a paquete dplyr

10:30