Lección 1 Introducción a los datos

Hans Sigrist

2026-03-04

Tabla de contenidos

Figura 1: Quarto

1. Google Colaboratory (Colab)

colab es un servicio alojado de Jupyter Notebook que no requiere configuración y ofrece acceso gratuito a recursos informáticos, como GPU y TPU. Colab es especialmente adecuado para el aprendizaje automático, la ciencia de datos y la educación.

En esta asignatura, lo utilizaremos para representar, graficar y resumir datos que no es posible implementar en otras aplicaciones como Excel u Hojas de Cálculo de Google (al menos sin la fricción propia de éstos).

Pero **colab** solo es el espacio de trabajo, necesitamos un motor, un lenguaje que compile nuestros datos. Actualmente, existen dos principalmente que permiten desarrollar análisis matemático y/o estadístico: python y R.

Usualmente, un usuario interesado en ejecutar este tipo de programas necesita:

1. Un **editor**
2. Un **kernel**
3. Conocer los **comandos del lenguaje** correspondiente

Con **colab** contamos con los dos primeros recursos: (1) **colab** en sí es un editor y (2) Google ofrece por medio de sus servicios distintos kernel, entre ellos python y R. Es cosa de "instalarlos" (en realidad nada se instala en tu computador solo se instalan en el servidor) y luego (3) usar los comandos respectivos para manipular datos -por ejemplo.

2. Tu cuenta institucional

Solo necesitas eso, tu cuenta institucional @colegioportaliano.cl:

1. Ingresa en colab
2. Presiona el botón "New Notebook", te advertirá que se requiere el acceso a Google y presionas el botón Acceder.
3. Ingresas tus credenciales, posiblemente te solicite que ingreses un segundo nivel de seguridad (mensaje de texto en tu celular, presionar "si" en un celular, etc.)
4. Listo ya estás en **colab**.

3. R

R es un entorno de software libre para computación estadística y gráficos. Se compila y ejecuta en una amplia variedad de plataformas UNIX, Windows y macOS. Este será nuestro kernel por defecto. Y por sus características, versatilidad y amplios usos en ciencias, economía, salud, ingeniería, etc. es alimentado por una gran comunidad que provee paquetes o librerías de dedicación especial. ¿Qué significa esto? Bueno, que usuarios alrededor del mundo han facilitado mucho las cosas para que otros usuarios -como ustedes y yo- podamos utilizar bases de datos originales y actualizadas que junto con la potencia del kernel R nos permiten realizar simulaciones que no podríamos hacer con otros software convencionales (mucha fricción).

Y eso es justamente lo que haremos ahora. Prepárate.

4. Estudio de caso: uso de stent para prevenir ACV (accidente cerebrovascular)

En esta sección, consideraremos un experimento que estudia la eficacia de los stents (también conocido como *endoprótesis vascular*: una malla o férula diseñada para abrir arterias, venas u otros conductos estrechados o bloqueados) en el tratamiento de pacientes con riesgo de ACV. Los stents son dispositivos que se colocan dentro de los vasos sanguíneos y que facilitan la recuperación del paciente tras un evento cardíaco y reducen el riesgo de sufrir otro ACV o incluso la muerte. Muchos médicos esperaban que se obtuvieran beneficios similares para los pacientes con riesgo de ACV.

Comenzamos por escribir la *pregunta principal* que los investigadores esperan responder:

¿Reduce el uso de stents el riesgo de un ACV?

4.1. El experimento

Los investigadores que formularon esta pregunta realizaron un experimento con 451 pacientes en riesgo.

Cada paciente voluntario fue asignado aleatoriamente a uno de dos grupos:

- **Grupo de tratamiento.** Los pacientes del grupo de tratamiento recibieron un stent y tratamiento médico. El tratamiento médico incluyó medicamentos, control de los factores de riesgo y ayuda para modificar el estilo de vida.
- **Grupo de control.** Los pacientes del grupo de control recibieron el mismo tratamiento médico que el grupo de tratamiento, pero no recibieron stents.

Los investigadores asignaron aleatoriamente a 224 pacientes al grupo de tratamiento y a 227 al grupo de control.

En este estudio, el grupo de control proporciona un punto de referencia para medir el impacto médico de los stents en el grupo de tratamiento.

Los investigadores estudiaron el efecto de los stents en dos momentos: 30 días después de la inscripción y 365 días después de la inscripción.

4.2. El resumen

Los resultados de 5 pacientes se resumen en la tabla Tabla 1. Los resultados de los pacientes se registran como “stroke” o “no event”, lo que indica si el paciente sufrió o no un ACV al final de un período.

Tabla 1: Tabla de resultados por paciente en el estudio de stents

Patient	Group	0–30 days	0–365 days
1	treatment	no event	no event
2	treatment	stroke	stroke
3	treatment	no event	no event
450	control	no event	no event
451	control	no event	no event

4.3. Primer problema

Considerar los datos de cada paciente individualmente sería un proceso largo y engorroso para responder a la pregunta de investigación original.

En cambio, realizar un análisis estadístico de datos nos permite considerar todos los datos a la vez.

Aquí entra R.

Existe una librería llamada "openintro" que a su vez es la creadora de muchas bases de datos, entre ellas stent30 y stent365, para usarlas debemos primero, instalar la librería:

Presiona Shift+Enter dentro de esta celda (usualmente al final del texto):

```
install.packages("openintro")
```

Podrás ver cómo colab instala la librería y sus componentes e incluso algunas otras cosas que necesita para su funcionamiento (dependencias). Puede durar desde algunos pocos segundos hasta un minuto aproximadamente. No hagas nada, solo espera que termine.

Ya instalada la librería, debemos invocarla:

```
library(openintro)
```

Listo, librería activa. No se observa nada.

Para inspeccionar qué contiene esta librería usamos el comando `data()`:

```
data(package = "openintro")
```

Se desplegará un menú vertical a la derecha con todos los conjuntos de datos del paquete "openintro" y su descripción. Si navegas podrás ver a stent30 y a stent365.

Ahora carguemos nuestras bases de datos de estudio, recuerda que son de la librería openintro. Usamos `data()` nuevamente:

```
data(stent30,stent365)
```

Un comando útil en estos momentos es `ls()`, éste lista el contenido cargado:

```
ls()
```

Podrás ver la salida: `'stent30'` y `'stent365'`

Ahora queremos saber qué estructura tiene el archivo, pues aún no lo vemos. Para ello usamos `str()`:

```
print(str(stent365))
```

Interpretación: se trata de una tabla de dos factores, cada uno con dos niveles.

¿Podrías decir cuáles son cada uno?

También podemos usar el comando `head()` que nos muestra el encabezado de la tabla:

```
head(stent365)
```

¿Cómo interpretas lo anterior?

Justo en este momento sería bueno que ejecutaras el siguiente comando que te permitirá conocer finalmente la base de datos `stent365`:

```
write.csv(stent30,  "stent30.csv",  row.names = FALSE)
write.csv(stent365, "stent365.csv", row.names = FALSE)
list.files(pattern = "^stent(30|365)\\.csv$")
```

Ahora ve al panel de la izquierda y abre el penúltimo ícono Archivos, actualiza y verás dos nuevos archivos para descargar. ¿Te diste cuenta? Es simplemente un archivo `csv` (comma separated values) de valores separados por coma, que leídos por esta hoja de cálculo corresponden simplemente a dos columnas de datos.

Por las características de los datos (estructura, encabezados, extensión), a veces es imposible usar hojas de cálculo para su lectura. Eso es justamente lo que estamos haciendo aquí. Como ambas variables son categóricas, el resumen natural es una tabla de contingencia (doble entrada).

Para una tabla de contingencia usamos el comando `table()` de esta forma:

```r
tab_stent365 <- table(stent365$group, stent365$outcome)
tab_stent365
```

Con la primera linea estamos pidiendo a R que use el comando `table()` para agrupar tanto por group como por outcome. Con la segunda linea, le pedimos que muestre dicha tabla.

Intenta hacer algunos simples cálculos aritméticos y te sorprenderás…

Ahora haremos algo más matemático y desde el punto de vista estadístico muy importante. Introduciremos la probabilidad, en este contexto: la proporción de pacientes según el grupo y el nivel.

Para ello usamos el comando `prop.table()` de esta forma:

```r
prop.table(tab_stent365, margin = 1)
```

Esta vez no copiaré el resultado porque ya te has dado cuenta de que son exactamente las salidas de los comandos que he ido proponiéndote.

A partir de tabla, ¿Qué estás observando? ¿Cómo lo interpretas? ¿Podrías hacer una conclusión? ¿Podrías responder a la pregunta inicial?

Para ayudar a tus conclusiones, siempre es bueno una perspectiva visual de la problemática. Para ello usamos el comando `barplot()` de esta forma:

```r
barplot(prop.table(tab_stent365, 1),
        beside = FALSE,
        legend = TRUE
)
```

¿Te atreves ahora a responder a la pregunta inicial? Escribe tu respuesta en tu cuaderno y compártela cuando se te solicite.

5. Cuestionario

5.1. Diseño del estudio

1. ¿Cuál es la pregunta de investigación del estudio?
2. ¿Cuál es la variable explicativa?
3. ¿Cuál es la variable respuesta?
4. ¿Por qué es importante que la asignación a los grupos haya sido aleatoria?
5. ¿Qué función cumple el grupo de control?

5.2. Naturaleza de las variables

6. Clasifica las variables `group` y `outcome`. ¿Son:

   - cuantitativas discretas?
   - cuantitativas continuas?
   - categóricas nominales?
   - categóricas ordinales?

   Justifica tu respuesta.

7. ¿Por qué el resumen natural de estos datos es una tabla de contingencia?

5.3. Cálculo manual de proporciones

Sabemos que:

- En el grupo tratamiento hubo 224 pacientes, de los cuales 45 tuvieron ACV en 365 días.
- En el grupo control hubo 227 pacientes, de los cuales 28 tuvieron ACV en 365 días.

8. Calcula la proporción de pacientes con ACV en el grupo tratamiento.
9. Calcula la proporción de pacientes con ACV en el grupo control.
10. Calcula la diferencia absoluta entre ambas proporciones.
11. Expresa esa diferencia en porcentaje.

5.4. Interpretación

12. Según tus cálculos, ¿el uso de stents reduce el riesgo de ACV?
13. ¿El resultado coincide con lo que esperaban los médicos?
14. ¿Puede una diferencia observada deberse al azar? Explica usando el ejemplo de lanzar una moneda muchas veces.

5.5. Pensamiento crítico

15. ¿Podemos generalizar estos resultados a todos los pacientes con riesgo de ACV? Explica.
16. ¿Qué implica que los pacientes hayan sido voluntarios?
17. Menciona al menos dos posibles limitaciones del estudio.
18. ¿Qué crees que significa que el estudio encontró “evidencia convincente de daño”?

5.6. Profundización matemática

19. Calcula la razón de proporciones:

$$\frac{0.20}{0.12}$$

Interpreta el resultado.

20. Si el tratamiento fuera realmente beneficioso, ¿qué esperarías observar en las proporciones?
21. Si el tamaño de la muestra fuera mucho menor, ¿sería más o menos confiable la diferencia observada? Explica.

5.7. Reflexión final

22. ¿Qué enseñanza deja este estudio respecto a la intuición médica y la necesidad de evidencia estadística?