Diseño y lógica de los estudios cuantitativos

Santiago Gualchi - Federico Alvarez

21 de septiembre de 2019

1. Introducción

En la clase pasada, realizamos una aproximación a la estadística y discutimos su relevancia a la hora de estudiar el lenguaje. Establecimos que la estadística es la disciplina que se ocupa de todas las etapas que involucran a los datos, incluyendo el diseño previo a su recolección, su análisis e interpretación, y su comunicación. En esta línea, introdujimos también una serie de conceptos, entre ellos: población, muestra, variable, hipótesis, espacio muestral y probabilidad.

Asimismo, establecimos la diferencia entre estadística descriptiva y estadística inferencial. La primera se refiere al conjunto de técnicas matemáticas que se limitan a describir las propiedades de la muestra estudiada. La segunda alude a las pruebas que permiten generalizar las observaciones sobre la muestra a la población relevante.

En esta reunión, vamos a avanzar sobre las líneas propuestas en el encuentro anterior. Nos vamos a concentrar en las etapas de diseño de una investigación para lo cual vamos a profundizar algunos de los conceptos que ya introdujimos. Vamos a centrarnos en las hipótesis y variables, y a estudiar sus propiedades y las repercusiones que traen las distintas formas de operacionalizarlas. Vamos a explicar cómo recolectar los datos de forma rigurosa y ver algunos buenos hábitos para su almacenamiento.

1.1. Ejercitación

1. Antes de avanzar, escribí definiciones para los siguientes conceptos: hipótesis, variable y operacionalización. No te preocupes si algunas de estas nociones te resultan muy nuevas. Está bien si las definís como te salga.

2. ¿Cómo caracterizarías los procesos de recolección y almacenamiento de datos? ¿Qué cuidados tendrías a la hora de llevarlos a cabo?

3. ¿Cómo pensás que podemos hacer para saber si nuestra hipótesis es correcta o incorrecta?

2. *Scouting*

Al principio de una investigación se suelen llevar a cabo las siguientes tareas:

- una primera caracterización del fenómeno;
- estudio de la bibliografía relevante;
- observación del fenómeno en escenarios naturales para posibilitar una primera generalización inductiva;
- recolección de información adicional (e.g., de colegas, estudiantes, etc.);
- razonamiento deductivo.

Si estudiamos el orden de palabras de los verbos frasales del inglés, encontramos la siguiente alternancia:

(1) a. He picked up [$_{SN}$ the book].
Orden: *VPO* (verbo - partícula - objeto)
b. He picked [$_{SN}$ the book] up.
Orden: *VOP* (verbo - objeto- partícula)

Al observar este fenómeno podemos encontrar un gran número de posibles variables que podrían influir en la elección de una u otra forma. Las **variables** son símbolos que pueden tomar, por lo menos, dos estados o niveles diferentes (e.g., la edad de un grupo de estudiantes de secundaria). En este sentido, se oponen a las **constantes**, que siempre presentan un mismo valor sin experimentar variación (e.g., la edad de un grupo de jóvenes de 12 años). Entre las variables que pueden afectar al posicionamiento de la partícula en los verbos frasales del inglés, las siguientes han sido propuestas en la bibliografía:

- Complejidad del OD (Fraser, 1966);
- Largo del OD (Chen, 1986; Hawkins, 1994);
- Presencia de un SP direccional (Chen, 1986);
- Animacidad (Gries, 2003);

- Concreción (Gries, 2003); y
- Tipo del OD (Van Dongen, 1919), entre otras.

Esta información puede ser más fácilmente visualizada en formato tabular, que permite reconocer qué variables han sido consideradas en los distintos estudios y cuántas variables consideró cada estudio (véase Cuadro 1).

Cuadro 1: Resumen de la bibliografía sobre posicionamiento de partículas en inglés I.

	Van Dongen (1919)	Fraser (1966)	Chen (1986)	Hawkins (1994)	Gries (2003)
Complejidad		×			
Largo			×	×	
SP			×		
Direccional					
Animacidad					×
Concreción					×
Tipo	×				

2.1. Ejercitación

1. Planteá un problema de investigación de tu interés.
2. ¿Cómo lo caracterizarías?

3. Hipótesis y operacionalización

Una vez que tenemos una visión general del fenómeno que queremos estudiar, es el momento de formular **hipótesis**.

3.1. Hipótesis científicas en forma de texto

Las hipótesis son:

- enunciados generales ocupados de más de un evento singular;
- enunciados con una estructura condicional (*si. . . , entonces. . .*), o que, al menos, puede ser parafraseados como tal; y

- potencialmente **falsables** (i.e., se pueden pensar eventos o situaciones que contradigan al enunciado) y **testeables** (i.e., se pueden realizar pruebas que determinen la verdad o falsedad del enunciado).

Para el estudio del posicionamiento de partículas en inglés, podemos pensar, por ejemplo, las siguientes hipótesis:

- Si el objeto directo de un verbo frasal transitivo es sintácticamente complejo, entonces los hablantes nativos producirán el orden de constituyentes *VPO* más frecuentemente que cuando el objeto directo es sintácticamente simple;
- Si el objeto directo de un verbo frasal transitivo es largo, entonces los hablantes nativos producirán el orden de constituyentes *VPO* más seguido que cuando el objeto directo es corto; o
- Si una construcción de verbo-partícula es seguida por un SP direccional, entonces los hablantes nativos producirán el orden de constituyentes *VOP* más seguido que cuando el SP direccional no está presente.

A su vez, las variables que consideremos pueden clasificarse según su influencia:

Variable independiente: Es la variable presente en la prótasis, y suele referirse a la causa de los cambios/efectos. La variable independiente representa tratamientos o condiciones que el investigador controla (directa o indirectamente) para entender sus efectos sobre la variable dependiente.

Variable dependiente: Es la variable presente en la apódosis, cuyos valores, variación o distribución se quieren explicar. La variable dependiente es la salida que depende del tratamiento experimental o de lo que el investigador cambia o manipula.

Variables de confusión o *confounders*: Son variables que interactúan tanto con la variable independiente como con la variable dependiente. Es importante identificar los confounders para realizar mejores diseños experimentales y obtener resultados con menos ruido.

Una vez que formulamos nuestra hipótesis, a la que vamos a llamar **hipótesis alternativa** (H_1), y antes de recolectar datos, tenemos que definir las condiciones que van a falsar nuestra hipótesis. De este modo, definimos la **hipótesis nula** (H_0) como el opuesto lógico de H_1 (predice la ausencia del efecto que enuncia H_1). La llamamos hipótesis nula porque se postula para ser anulada

con los datos de la investigación. Esto es importante, porque la idea es que ambas hipótesis cubran todo el espacio de resultados o **espacio muestral**, i.e., el conjunto de todos los resultados teóricamente posibles. Por ejemplo:

- Si el objeto directo de un verbo frasal transitivo es sintácticamente complejo, entonces los hablantes nativos *no* producirán el orden de constituyentes VPO más seguido que cuando el objeto directo es sintácticamente simple (H_0 correspondiente a la primera hipótesis de tipo 1); o
- Los dos niveles de Orden (VPO y VOP) ~~*no* no~~ son igualmente frecuentes (H_0 correspondiente a la hipótesis de tipo 2).

Ahora bien, en algunas investigaciones es posible suponer que los efectos o relaciones entre variables ocurran en una dirección determinada (se desvíen de la H_0 hacia *un* lado). En estos casos, se dice que se establece una **hipótesis direccional**. Por el contrario, las **hipótesis no direccionales** solo predicen que existe un efecto o relación sin especificar la dirección del efecto.

3.2. Operacionalización de variables

Una vez que formulamos nuestra hipótesis, es importante encontrar un modo de **operacionalizar** las variables. Esto supone decidir qué será observado, contado, medido, etc. cuando investiguemos nuestras hipótesis. Por ejemplo, si volvemos a las variables consideradas en la bibliografía sobre el orden de palabras en los verbos frasales del inglés, podemos operacionalizarlas como sigue:

- Complejidad: OD *simple* (e.g., *the book*), *OD modificado sintagmáticamente* (e.g., *the book on the table*) u *OD modificado por cláusula* (e.g., *the book I had bought in Europe*);
- Largo: el largo del OD medido en sílabas;
- SP direccional: *presencia* o *ausencia* de un SP direccional (e.g., *He picked the book up [$_{SP}$ from the table]*);
- Animacidad: *animado* o *inanimado*;
- Concreción: *concreto* o *abstracto*; y
- Tipo del OD: *pronominal* (e.g., *He picked [$_{pron}$ him] up this morning*), *semipronominal* (e.g., *He picked [$_{semi}$ something] up from the floor*), *léxico* (e.g., *He picked [$_{léx}$ people] up this morning*) o *nombre propio* (e.g., *He picked [$_{prop}$ Peter] up this morning*).

Otro ejemplo, si queremos operacionalizar el conocimiento de una lengua extranjera de una persona, podemos tomar en consideración:

- La complejidad de las oraciones que una persona puede formar en la lengua en cuestión;
- El tiempo en segundos entre dos errores en la conversación;
- El número de errores cada 100 palabras en un texto que la persona escriba en 90 minutos.

La operacionalización de variables involucra el uso de **niveles** numéricos para representar estados de variables. Un número puede ser una medida (e.g., 402 ms de tiempo de reacción), pero estados discretos no numéricos también pueden, teóricamente, ser codificados usando números. Según los niveles de medida las variables pueden clasificarse en:

Variable nominal (o binaria): Sólo pueden tomar dos niveles diferentes y sus valores sólo revelan que los objetos con estos valores exhiben características diferentes (e.g., animacidad);

Variable categórica: Son una generalización del caso anterior a tres niveles o más. (e.g., aspecto);

Variable ordinal: Distinguen categorías que presentan alguna clase de secuencia o progresión.

Variable cuantitativa: Consisten en valores numéricos que pueden ser continuos o discretos, para los cuales la diferencia absoluta entre los valores (variables de intervalo) o la proporción entre los valores (variable de razón) son significativas (e.g., largo en sílabas).

3.3. Hipótesis estadísticas en formato estadístico/matemático

Después de formular las hipótesis (H_0 y H_1) en forma de texto y definir cómo operacionalizar las variables, es necesario formular dos **versiones estadísticas** de las hipótesis. Esto significa expresar los resultados numéricos esperados sobre la base de las hipótesis textuales. Dichos resultados suelen involucrar una de las siguientes formas matemáticas:

- Frecuencias
- Promedios

- Dispersiones
- Correlaciones
- Distribuciones

Este va a ser el formato que vamos a usar para evaluar la **significancia** de nuestras hipótesis, y su definición va a depender directamente de cómo operacionalizamos las variables. Por ejemplo, si nuestra hipótesis involucra la variable largo del OD, su forma estadística no va a ser la misma si la operacionalizamos cuantitativamente como largo medido en número de sílabas o de forma discreta como una variable categórica con niveles *corto*, *mediano* y *largo*. En el primer caso, nuestras hipótesis estadísticas van a poder referirse a la media del largo, mientras que esto no es posible en el segundo caso. Tomando largo como una variable categórica podríamos operacionalizar nuestras hipótesis, por ejemplo, basándonos en conteos o frecuencias.

Retomemos la H_1 respecto de la presencia/ausencia de un SP direccional: si una construcción de verbo-partícula es seguida por un SP direccional, entonces los hablantes nativos producirán el orden de constituyentes VOP más seguido que cuando el SP direccional no está presente. Si formulamos nuestras hipótesis matemáticamente, obtenemos los siguientes resultados:

$$H_1 \text{ direccional} : {}^{n}\text{SSPP dir. en VPO} < {}^{n}\text{SSPP dir. en VOP}$$

$$H_1 \text{ no direccional} : {}^{n}\text{SSPP dir. en VPO} \neq {}^{n}\text{SSPP dir. en VOP}$$

$$H_0 : {}^{n}\text{SSPP dir. en VPO} = {}^{n}\text{SSPP dir. en VOP}$$

3.4. Ejercitación

1. ¿Cuáles de los siguientes enunciados podrían ser hipótesis científicas?
   a) ¿La frecuencia fundamental aumenta con la edad?
   b) El sujeto X aprenderá la palabra *casa* antes que la palabra *examen*.
   c) La presencia de tonos en una lengua está influida por la humedad de la zona en que se habla.
   d) Las lenguas con menos hablantes posiblemente tienden a cambiar más rápidamente que las lenguas con muchos hablantes.
   e) La relación entre forma y significado es arbitraria.
   f) Los tweets de Donald Trump hacen caer el valor de la industria china.
   g)

2. Donde sea posible, reformulá los enunciados en forma de hipótesis.

3. Operacionalizá las variables involucradas y determiná de qué tipo es cada una.

4. ¿Qué variables podrían estar influyendo y no están siendo tomadas en cuenta?

5. Operacionalizá matemáticamente las hipótesis.

4. Recolección de datos

La **recolección** de datos comienza solo después de haber operacionalizado las variables y formulado las hipótesis. Por lo general, no se estudia la población entera sino una muestra. Si queremos que nuestros datos puedan generalizarse a la población, esta muestra debe ser **representativa** (i.e., las distintas partes de la población deben estar reflejadas en la muestra) y **balanceada** (i.e., los tamaños de las partes de la muestra deben corresponderse con las proporciones que presentan en la población). Esto muchas veces es un ideal teórico porque con frecuencia no conocemos todas las partes y las proporciones de la población. Una forma de obtener una muestra representativa y balanceada es a partir de la randomización. Este es uno de los principios más importantes de la recolección de datos.

4.1. Tipos de estudios

De acuerdo al grado en el que se pueden manipular las condiciones del estudio pueden distinguirse dos tipos:

Experimento: Es un estudio en el cual las condiciones son asignadas de manera deliberada (y usualmente aleatoria) a individuos/sujetos/instancias temporales con el objetivo de ver si el efecto de estas condiciones en alguna característica particular. Las condiciones suelen llamarse *tratamientos* y se corresponden con los niveles de la(s) variable(s) independiente(s). La característica observada, por su parte, se corresponde con la variable dependiente. Los individuos/objetos/instancias temporales son las *unidades experimentales*.

Estudio observacional: Es un estudio donde las condiciones no son asignadas ni controladas por la persona que investiga, sino simplemente observadas. Las condiciones son características inherentes de cada sujeto/objeto/instancia temporal. El interés sigue puesto en comparar los valores de la variable dependiente en función de los niveles de la varia-

ble independiente. Aquí podemos distinguir entre estudios prospectivos, donde van recolectándose observaciones a medida que van teniendo lugar sucesos de interés, y estudios retrospectivos, donde se recolectan los datos luego de ocurridos los sucesos.

En un estudio experimental es muy importante la manera en la que se seleccionen las observaciones que integran la muestra, dado que para poder generalizar los resultados a la población de interés es importante que las observaciones sean independientes entre sí y que los distintos grupos dentro de la población estén debidamente representados. En términos de la selección de las unidades experimentales, se pueden identificar por lo menos cuatro formas distintas de muestreo:

Muestreo aleatorio: Supone la posibilidad de elegir al azar las unidades experimentales sobre la totalidad de la población.

Muestreo estratificado: Supone la división de la población en estratos, subgrupos que comparten semejanzas respecto de un determinado conjunto de variables. La muestra del experimento se realiza en base a la proporción de la población total que representa cada estrato.

Muestreo por *clusters*: Consiste en dividir a la población en grupos de manera aleatoria, seleccionar **n** grupos, y dentro de cada grupo muestrear aleatoriamente.

Muestreo por conveniencia: Consiste en una selección basada en la accesibilidad relativa de distintos miembros de la población. Puede generar una cantidad importante de sesgo en los resultados, dado que puede estar asociado a la recolección de muestras no independientes entre sí, o a la subrrepresentación de grupos.

4.2. Ejercitación

1. De los siguientes ejemplos, ¿cuáles son experimentos y cuáles estudios observacionales?

   a) Se intentó identificar los efectos de distractores externos (ninguno, constante, variables) y de tipos de palabras (frutas, sustantivos, cualquier tipo) en la habilidad para memorizar palabras. Hay nueve (3 x 3) combinaciones de distracción y tipo de palabra. 36 sujetos fueron asignados al azar a las nueve combinaciones, con 4 sujetos por combinación. Para cada tipo de palabra, se preparó una lista con 30 términos. Cada sujeto estudió la lista que le fue asignada por 5 minutos. Luego de una espera de 2 minutos, se le pidió a los sujetos

que reciten la totalidad de palabras que recordaban y se midió la cantidad de palabras correctamente reportadas.

b) Se intentó comparar la dificultad de lectura en dos revistas, **Personas** y **Personas Jóvenes**. Se seleccionaron cien oraciones al azar de la última edición de cada revista y, para cada una, se comparó el largo promedio de las oraciones en cantidad de letras.

c) Se intentó determinar si el CI de les niñes está relacionado con haber sido amamantades o no. Investigadores midieron el CI de una gran cantidad de estudiantes de primer grado en una ciudad grande. Les investigadores además consultaron a las madres si les habían amamantado o no.

d) Se intentó determinar si una reducción en el número de pop-ups mejoraba la experiencia de uso de un sitio web. Un grupo de 1000 suscriptores fueron seleccionados al azar. La mitad de ellos vieron aproximadamente la mitad de pop-ups al visitar el sitio, mientras que la otra mitad vio la cantidad usual. Tras dos semanas, se le solicitó a les suscriptores que completaran una encuesta de satisfacción.

2. En una población compuesta por hablantes de hasta 25 años en un 43 %, hablantes de entre 26 y 50 años en un 29 %, hablantes de entre 51 y 75 años en un 19 % y mayores de 76 en un 9 %, querés estudiar la influencia del trap en la lengua. Diseñá el estudio.

5. Almacenamiento de datos

Una vez que recolectamos los datos (o mientras lo hacemos), es necesario **almacenarlos** en un formato que nos permita anotarlos, manipularlos y evaluarlos fácilmente. Para esto es recomendable el uso de hojas de cálculo (e.g., LibreOffice Calc), bases de datos o R.

Un formato recomendado para el almacenamiento es el *case-by-variable* (véase Cuadro 4):

- la primera fila contiene los nombres de las variables;
- las otras filas representan cada una un *data point* (i.e., una observación determinada de la variable dependiente);
- la primera columna numera todos los n casos de 1 a n (esto permite identificar cada fila y restaurar el orden original);
- las otras columnas representan una sola variable o característica correspondiente a un determinado *data point*; y

- la información faltante se anota usando *un* símbolo (por ejemplo, “NA”) y el mismo solo debe usarse para representar dicho significado.

Cuadro 2: Una tabla que usa el formato *case-by-variable* para codificar información sobre el posicionamiento de partículas en inglés en función del largo del OD medido en sílabas.

Caso	Orden	Largo	Oración
1	vpo	2	He turned on the lights.
2	vpo	2	The police broke into the house.
3	vop	2	Mary asked Susan out.
4	vop	2	I had to hold my dog back because there was a cat in the park.
5	vop	2	You can warm your feet up in front of the fireplace.
6	vop	3	Our teacher finally broke the project down into three separate parts.
7	vpo	3	I’m looking for a red dress.
...	...	...	...

5.1. Ejercitación

1. El dataset en el archivo datos_lenguas.csv contiene información sobre el orden de sujeto, verbo y objeto de distintas lenguas y de la familia lingüística a la que pertenecen. Supongamos que querés investigar cómo este orden se ve influido por la familia lingüística. Creá un dataset que contenga esta información y se ajuste al formato *case-by-variable*.