Diseño y lógica de los estudios cuantitativos (versión Python)

Santiago Gualchi

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. Señalamos la conocida problemática de que, si bien los estudios cuantitativos están cada vez más presentes en la investigación lingüística, en nuestro país (y, en particular, en nuestra universidad), los planes de estudio excluyen la estadística de la formación del lingüista. Esto supone un grave problema para quienes nos formamos en el área en tanto que el desconocimiento de los métodos usados actualmente en investigación amenaza con dejarnos fuera del debate científico (no poder publicar, no poder leer).

A su vez, mencionamos 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.

Finalmente, discutimos distintas áreas donde el uso de modelos estadísticos puede ayudarnos a entender y a explicar mejor los fenómenos que estudiamos. Entre los campos señalados incluimos (sin ser exhaustivos) la psicolingüística y la neurolingüística, la lingüística de corpus, la lingüística computacional, la tipología y la lingüística histórica, y la teoría lingüística.

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 cuáles son los buenos hábitos para su almacenamiento. Por último, vamos a introducir la forma de proceder para la aceptación (o no) de nuestras hipótesis y por qué se realiza de este modo.

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. Preparación del directorio de trabajo

Cuando llevamos a cabo una investigación cuantitativa, trabajamos con datos. Casi siempre estos datos son muy numerosos y necesitan ser analizados con tests estadísticos que pueden involucrar un también muy alto número de operaciones matemáticas. Por esta razón, hoy resulta prácticamente imprescindible hacer uso de un software de análisis estadístico (e.g., SPSS, SAS, Stata) o, mejor todavía, un lenguaje de programación con buen soporte estadístico (e.g., R, Python, Julia). Para mantener el espacio de trabajo organizado, facilitar la comprensión y reproducibilidad y evitar errores prevenibles, es importante mantener una buena estructuración de nuestro directorio. Existen distintas prácticas sugeridas. A continuación voy a resumir las recomendaciones desarrolladas por DrivenData para la estructuración de proyectos de data science en Python:

El archivo README Todos nos hemos encontrado con archivos que se llaman "README", "README.md", "README.txt", "LÉAME", "LEEME.md", "LEER", "LEEMEEEE!!.txt" y otras variantes cuando descargamos algunos contenidos de internet. El archivo README es uno de los más importantes de cualquier proyecto. Como su nombre sugiere es el primer lugar por el que se accede al contenido. Un buen archivo README resume la información esencial que el desarrollador considera que el usuario debe conocer al hacer uso de su solución, y es el primer archivo que el usuario debe consultar cuando accede a un proyecto. Los archivos README suelen estar escritos en texto plano (sin ningún tipo de formato) o usando algún lenguaje de marcado (por lo general, MarkDown). Es fundamental que todo proyecto que encaremos cuente con un buen archivo README! Pueden ver un ejemplo de una buena plantilla de README para proyectos en Data Science en el sitio de Code for San Francisco.

El archivo LICENSE Este es otro documento muy importante. Por un lado, resguarda tus derechos de autor y te permite definir qué usos pueden hacer otros de tu código. Pero también, permite a otros reutilizar los recursos que desarrollaste. No incluir una licencia en tu proyecto puede suponer un impedimento legal para que otros usen, modifiquen o reciclen tu código. Con el fin de hacer una ciencia más abierta se anima cada vez más el uso de licencias libres, que permiten al usuario el uso, el estudio, la redistribución y la mejora del trabajo. Algunos ejemplos conocidos de este tipos de licencias son GNU GPL, BSD y MIT.

El directorio data Todos los datos que vayamos a usar van a estar albergados en esta carpeta. A su vez, dentro de esta carpeta, vamos a incluir una carpeta **raw** en la que vamos a almacenar la versión original de nuestros datos. Estos no deben ser *nunca* modificados, sino que se deben crear nuevas versiones y almacenarlas en otros directorios dentro nuestra carpeta data.

El directorio models Acá almacenamos nuestros modelos entrenados y serializados, sus predicciones y sus resúmenes.

El directorio notebooks En este directorio vamos a guardar las **Jupyter Notebooks** que escribamos. Estas son una muy buena solución para combinar código ejecutable

y documentación. De este modo, podemos escribir reportes en forma sencilla, pero también podemos aprovechar su versatilidad para correr análisis exploratorios.

El directorio src En esta carpeta va a estar contenido todo el código que escribamos para llevar a cabo nuestra análisis, como el que se ocupa de descargar, generar o limpiar los datos, definir y entrenar los modelos, o crear visualizaciones. La única excepción es que vamos a incluir código que usa el que está contenido en este directorio en las notebooks.

Esta es una simplificación de la propuesta, la estructura completa de directorio incluye archivos y carpetas específicos para Python, además de otros subdirectorios para usos especializados. Además, cabe señalar que este diseño puede resultar útil para un gran número de proyectos, pero no quiere decir que no existan casos en los que apartarse un poco de estas recomendaciones pueda resultar productivo.

2.1. Ejercitación

1. Armá un directorio para una investigación cuantitativa siguiendo los lineamientos propuestos por el equipo de DrivenData.
2. Escribí un README siguiendo las sugerencias de Code for San Francisco.
3. Investigá en internet qué lenguajes de marcado existen, cuál es la sintaxis básica de MarkDown y cuáles son sus ventajas y desventajas. Formateá el "README" que escribiste usando MarkDown y renombrá el archivo "README.md".
4. Incluí una licencia GNU LGPL. ¿En qué se diferencia con la licencia GNU GPL?
5. Contestá verdadero o falso y justificá:
   a. No es necesario guardar una copia de nuestros datos originales.
   b. Las notebooks son especialmente útiles para la manipulación de datos y su modelado.
   c. Un proyecto sin una licencia es por defecto de dominio público.
   d. La mayor parte del código fuente debe ser almacenada en "src".

3. Control de versiones

El control de versiones se refiere a la gestión de los cambios en un archivo o un grupo de archivos. Es lo que hacemos cuando en una carpeta tenemos "trabajo.odt", "trabajo2.odt", "trabajo final.odt", "trabajo FINAL.odt", "trabajo FINALISIMO.odt" y así *ad infinitum*. No obstante, existen mejores formas para hacer esto. Una de las soluciones más usadas es Git, que nos permite restaurar versiones pero también trabajar colaborativamente, entre otras muchas ventajas. No vamos a introducir su funcionamiento, pero pueden acceder a la guía en español escrita por Macarena Fernández.

3.1. Ejercitación

1. Comprobá si Git está instalado en tu computadora. Si no sabés cómo hacerlo, buscá en internet.
2. Si no tenés Git instalado, instalalo.
3. Investigá qué es un repositorio.
4. Creá un repositorio en tu computadora desde la terminal.

5. Abrí una cuenta en una plataforma de desarrollo colaborativo y creá un repositorio desde ahí.
6. Buscá en internet qué otros sistemas de control de versiones existen y qué servicios que usamos frecuentemente los usan.

4. *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). Otra tabla útil es la que sintetiza los niveles de las variables y sus preferencias para uno u otro orden. Como se ve en el Cuadro 2, el orden *VPO* sería usado con OODD cognitivamente más complejos (SSNN complejos y largos con sustantivos léxicos que refieren a entidades abstractas). *VOP*, en cambio, es usado en los casos opuestos.

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	×				

Cuadro 2: Resumen de la bibliografía sobre posicionamiento de partículas en inglés II.

	Nivel de variable para *VPO*	Nivel de variable para *VOP*
Complejidad	modificado sintagmáticamente modificado por cláusula	
Largo	largo	
SP Direccional	ausente	presente
Animacidad	inanimado	animado
Concreción	abstracto	concreto
Tipo		pronominal

4.1. Ejercitación

1. Planteá un problema de investigación de tu interés.
2. ¿Cómo lo caracterizarías?
3. Completá la siguiente tabla con bibliografía que sea relevante para el problema que planteaste. En la columna observación, escribí por qué ese trabajo es particularmente útil para tu propuesta de investigación.

Autor	Año	Título	Observación

4. Escribí una lista de las personas en tu lugar de estudio o trabajo que podrían sugerirte bibliografía o líneas de análisis para este problema en particular.
5. ¿Qué motores de búsqueda podés usar para encontrar investigaciones científicas?

Usalos para encontrar bibliografía de utilidad para la investigación que propusiste.

5. Hipótesis y operacionalización

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

5.1. Hipótesis científicas en forma de texto

Las hipótesis pueden clasificarse en **hipótesis de tipo 1** y **de tipo 2**. Las primeras se caracterizan por ser:

- 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 seguido 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.

confounder es una variable que interactúa 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.

variable moderadora es una variable independiente secundaria que se selecciona para determinar si afecta la relación entre la variable independiente y la dependiente.

Por su parte, las hipótesis de tipo 2 contienen solo una variable dependiente y ninguna variable independiente. En estos casos, la hipótesis es un enunciado sobre los valores, variación o distribución de la variable dependiente. Por ejemplo, “los dos niveles de Orden no son igualmente frecuentes”. Así, podemos definir una hipótesis como un enunciado

acerca de la relación entre dos o más variables, o acerca de una variable en un **contexto muestral**, que se espera que aplique en contextos similares y/o para objetos similares de la población.

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 situaciones y estados de cosas 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.

5.2. 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 los niveles, i.e., 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) solo pueden tomar dos niveles diferentes y sus valores solo revelan que los objetos con estos valores exhiben características diferentes (e.g., animacidad);

variable categórica pueden tomar tres niveles diferentes o más y sus valores solo revelan que los objetos con estos valores exhiben características diferentes (e.g., aspecto);

variable ordinal permiten distinguir categorías, pero también permiten rankear los objetos de forma significativa (e.g., complejidad del OD); y

variable cuantitativa (o de razón) además de distinguir categorías y rankear objetos, también permiten comparar las diferencias y los ratios entre valores de forma significativa (e.g., largo en sílabas).

5.3. 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; o
- distribuciones.

Este va a ser el formato que vamos a usar para evaluar la **significancia** de nuestras hipótesis (véase más abajo), 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 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:

$$H_{1 \text{ direccional}} : n_{\text{SSPP dir. en VPO}} < n_{\text{SSPP dir. en VOP}}$$

$$H_{1 \text{ no direccional}} : n_{\text{SSPP dir. en VPO}} \neq n_{\text{SSPP dir. en VOP}}$$

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

5.4. Ejercitación

1. ¿Cuáles de los siguientes enunciados podrían ser hipótesis?
   a. La variación lingüística está dada por diferencias en las propiedades de las categorías funcionales.
   b. ¿La frecuencia fundamental aumenta con la edad?
   c. El sujeto LC aprenderá la palabra *casa* antes que la palabra *examen*.
   d. La presencia de tonos en una lengua está influida por las condiciones climáticas de la zona en que se habla.
   e. La presencia de tonos en una lengua está influida por la humedad de la zona en que se habla.
   f. Las lenguas con menos hablantes posiblemente tienden a cambiar más rápidamente que las lenguas con muchos hablantes.
   g. La relación entre forma y significado es arbitraria.
   h. Oumuamua era una nave interestelar que llevaba criaturas con un sistema simbólico doblemente articulado.
2. Reformulá los enunciados que no tienen las características de las hipótesis para que las tengan.
3. ¿De qué tipo es cada hipótesis?
4. Operacionalizalas matemáticamente.
5. Operacionalizá las variables involucradas y determiná de qué tipo es cada una.
6. ¿Qué *confounders* y variables moderadoras podrían estar influyendo y no están siendo tomadas en cuenta?

6. 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 más representativa y balanceada es a partir de la randomización. Este es uno de los principios más importantes de la recolección de datos.

6.1. Ejercitación

1. Querés caracterizar los distintos usos del prefijo *in-* en una población de 150.000 hablantes,
   a. ¿cuántos participantes necesitarías para la recolección de datos?
   b. ¿qué subgrupos de tu población deberías representar?

   c. ¿creés que esos subgrupos presentarán un comportamiento diferenciado?
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. ¿Cómo conformarías la muestra?

7. 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 4: 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 broked the project down into three separate parts.
7	vpo	3	I'm looking for a red dress.
...	...	...	...

7.1. Ejercitación

1. El dataset en https://tinyurl.com/y3vd3x3g contiene información sobre el orden de sujeto, verbo y objeto de distintas lenguas y de la familia lingüística a la que pertenecen. 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*.
2. Investigá sobre la importancia de usar formatos estándar para el almacenamiento

de datos.
3. Cross-Linguistic Data Formats es una iniciativa que busca proponer estándares para la representación de datos translingüísticos para la investigación tipológica e histórica. ¿Qué principios guían su diseño? ¿Qué tipos de datos soporta actualmente?
4. ¿Qué bases de datos disponibles abiertamente en internet se ajustan a estándares definidos por esta iniciativa?

8. Cómo decidir

Cuando ya almacenamos los datos, procedemos a evaluarlos con algún **test estadístico**. Sin embargo, la forma de proceder que se acostumbra en ciencias biológicas, psicología, ciencias sociales y humanidades consiste no en probar que H_1 es correcta, sino que la versión estadística de H_0 es improbable y, por lo tanto, pueda ser rechazada. Ya que H_0 es la contracara lógica de H_1, esto apoya H_1.

El **testeo de hipótesis** nos permite decidir si un **efecto** observado se debe a relaciones reales entre las variables o al azar[1]. Dicho de otra forma, nos permite justificar la preferencia por explicar un fenómeno por medio de una interacción entre variables contra considerar que dicha interacción no tiene influencia sobre el efecto observado. Tanto el estadista como las editoriales, y los demás actores sociales quieren evitar perder tiempo y plata analizando/interpretando/considerando conclusiones incorrectas. Para sortear esto existen distintas técnicas. Uno de los procedimientos que se utilizan (especialmente en psicología) es la **Prueba de Significancia de la Hipótesis Nula** (NHST, por sus siglas en inglés)[2]. En líneas muy generales, podemos definirla como sigue:

1. definición del **nivel de significancia** $\boldsymbol{p_{\text{crítico}}}$, que por lo general es 0,05;
2. análisis de los datos computando la probabilidad de un efecto e (e.g., una distribución, una diferencia de medias, una correlación) usando las hipótesis estadísticas;
3. computación de la **probabilidad de error** $\boldsymbol{p}$ (qué tan probable es encontrar e o algo que se desvía aún más de H_0 cuando H_0 es verdadera); y
4. comparación de $p_{\text{crítico}}$ y p y decidir si $p < p_{\text{crítico}}$; entonces podemos rechazar H_0 y aceptar H_1.

Si la probabilidad de error p de un fenómeno es menor a $p_{\text{crítico}}$ podemos rechazar H_0 y aceptar H_1. Esto no significa que hayamos probado H_1, sino que la probabilidad de error p es lo suficientemente baja como para aceptar H_1. La probabilidad de error p es conocida como **valor** $\boldsymbol{p}$. El Cuadro 5 recoge la semántica estándar de dicho valor.

Cuadro 5: Semántica de los valores de p.

Valor	Significancia	Indicación
$p < 0{,}001$	altamente significativo	***
$0{,}001 \leq p < 0{,}01$	muy significativo	**
$0{,}01 \leq p < 0{,}05$	significativo	*
$0{,}05 \leq p < 0{,}1$	marginalmente significativo	*ms* o .

[1] Realmete el testeo de hipótesis no siempre nos permite decidir. Algunos procedimientos (como vamos a ver a continuación) solo nos permiten calcular la probabilidad de que una observación se dé por azar.

[2] La NHST ha sido fuertemente discutida. Para un análisis crítico de este procedimiento, véase Cohen (1994) y Perezgonzalez (2015).

Valor	Significancia	Indicación
$0{,}01 \leq p$	no significativo	

Al analizar la significancia del efecto, es correcto rechazar la hipótesis nula cuando la hipótesis alternativa es verdadera, y aceptar la hipótesis nula cuando esta es verdadera (véase Cuadro 6). Sin embargo, existen dos combinaciones lógicas más. Un **error de tipo I** ocurre cuando la hipótesis nula es verdadera y se rechaza. Por lo general, estos errores deben ser evitados tanto como sea posible en la carrera del investigador, y es lo que buscamos hacer cuando llevamos a cabo pruebas de testeo de hipótesis. Asimismo, un **error de tipo II** ocurre cuando la hipótesis alternativa es verdadera y se rechaza. Estos errores suelen ser menos graves que los de tipo I, pero, de todos modos, debemos reducir la posibilidad de que ocurran.

Cuadro 6: Errores de tipo I y II.

	H_0 es verdadera	H_1 es verdadera
Rechaza H_0	error de tipo I	correcto
Acepta H_0	correcto	error de tipo II

8.1. Valores *p* de una cola en distribuciones de probabilidad discretas

Supongamos que queremos estudiar ambigüedad categorial. Para ello encuestamos 3 sujetos acerca de si la palabra *camino* es un verbo o un nombre, asumiendo que los nombres son cognitivamente más prominentes y que, por lo tanto, las respuestas serán mayores para esta categoría:

H_0 textual ambas respuestas son igualmente frecuentes.
H_1 textual *nombre* es una respuesta más frecuente que *verbo*.
H_0 estadística los sujetos van a responder *nombre* tantas veces como *verbo*: $n_{\text{nombre}} = n_{\text{verbo}}$.
H_1 estadística los sujetos van a responder *nombre* más veces que *verbo*: $n_{\text{nombre}} > n_{\text{verbo}}$.

¿Si los 3 sujetos responden que *camino* es un nombre podemos rechazar H_0 y aceptar H_1 asumiendo un $p_{\text{crítico}} = 0{,}05$? El Cuadro 7 sintetiza la probabilidad de cada respuesta bajo el supuesto de que H_0 es verdadera. Las columnas N y V representan **variables aleatorias**. Cada una de ellas contabiliza la frecuencia de ocurrencia de un nivel para una variable. De esta forma creamos dos nuevos espacios muestrales con una cantidad reducida (y, por lo tanto, más manejable) de elementos (i.e., posibles resultados)[3]. La columna $p_{\text{resultado}}$ representa la probabilidad de que ocurra la combinación de respuestas de los sujetos correspondiente. Dado que, bajo la H_0 asumimos que *nombre* y *verbo* son respuestas igualmente probables, la probabilidad de que un sujeto dado responda *nombre* o responda *verbo* es la misma ($P(nombre) + P(verbo) = 1$; $P(nombre) = P(verbo) = 0{,}5$). La probabilidad de cada combinación puede calcularse de dos formas. La primera es,

[3]En este caso, solo es necesario definir una variable aleatoria, ya que ambas son complementarias : $N = 3 - V$. Definimos dos variables aleatorias a modo ilustrativo.

sabiendo que la suma de las probabilidades de las combinaciones debe sumar 1 y asumiendo que cada combinación es igualmente probable, dividir 1 por el total de elementos en el espacio muestral: $1 \div 8 = 0{,}125$. Otra posibilidad consiste en calcular el producto de las probabilidades de las 3 respuestas correspondientes a la combinación: $0{,}5 \times 0{,}5 \times 0{,}5 = 0{,}125$. Dado que, bajo H_0, la probabilidad de que los 3 sujetos respondan nombre es de $p = 0{,}125$ y que $p > p_{\text{crítico}}$ no podemos rechazar H_0.

Cuadro 7: Todos los resultados posibles de pedir a 3 sujetos que clasifiquen *camino* como un nombre o un verbo.

Sujeto 1	Sujeto 2	Sujeto 3	N	V	$p_{\text{resultados}}$
nombre	nombre	nombre	3	0	0,125
nombre	nombre	verbo	2	1	0,125
nombre	verbo	nombre	2	1	0,125
nombre	verbo	verbo	1	2	0,125
verbo	nombre	nombre	2	1	0,125
verbo	nombre	verbo	1	2	0,125
verbo	verbo	nombre	1	2	0,125
verbo	verbo	verbo	0	3	0,125

La distribución de probabilidad de cada resultado posible para 3 sujetos queda recogida en el primer histograma Figura 1. Como se observa en dicha figura, a medida que aumenta el número de sujetos la distribución se asemeja cada vez más a la de la distribución gaussiana o normal.

Ahora bien, si encuestamos a 100 personas, ¿podemos rechazar H_0 si 59 responden *nombre*? La respuesta es sí: asumiendo H_0, la probabilidad de que los sujetos respondan *nombre* 59 veces o más es de $p = 0{,}044$ (véase Figura 2).

8.2. Valores *p* de dos colas en distribuciones de probabilidad discretas

En la sección anterior, nuestra H_1 era direccional: "Si una palabra puede ser analizada como nombre o como verbo, los sujetos responderán *nombre* más frecuentemente". La prueba de significancia que discutimos es una prueba de una cola, porque solo nos interesaba una dirección en la que el resultado observado se desviaba del resultado esperado. Si, en cambio, asumimos una H_1 no direccional (por ejemplo, "Si una palabra puede ser analizada como nombre o como verbo, los hablantes responderán verbo y nombre con distinta frecuencia"), tenemos que mirar hacia ambos lados del desvío:

H_0 estadística los sujetos van a responder *nombre* tantas veces como *verbo*: $n_{\text{nombre}} = n_{\text{verbo}}$.

H_1 estadística los sujetos van a responder *nombre* en un número distinto de veces que *verbo*: $n_{\text{nombre}} \neq n_{\text{verbo}}$.

Ahora imaginemos que, una vez más, de 100 sujetos que responden si *camino* es un nombre o un verbo, 59 deciden que es un nombre. Ya que nuestra hipótesis es no direccional y queremos calcular la probabilidad de que ocurra dicho resultado u otro que se desvíe

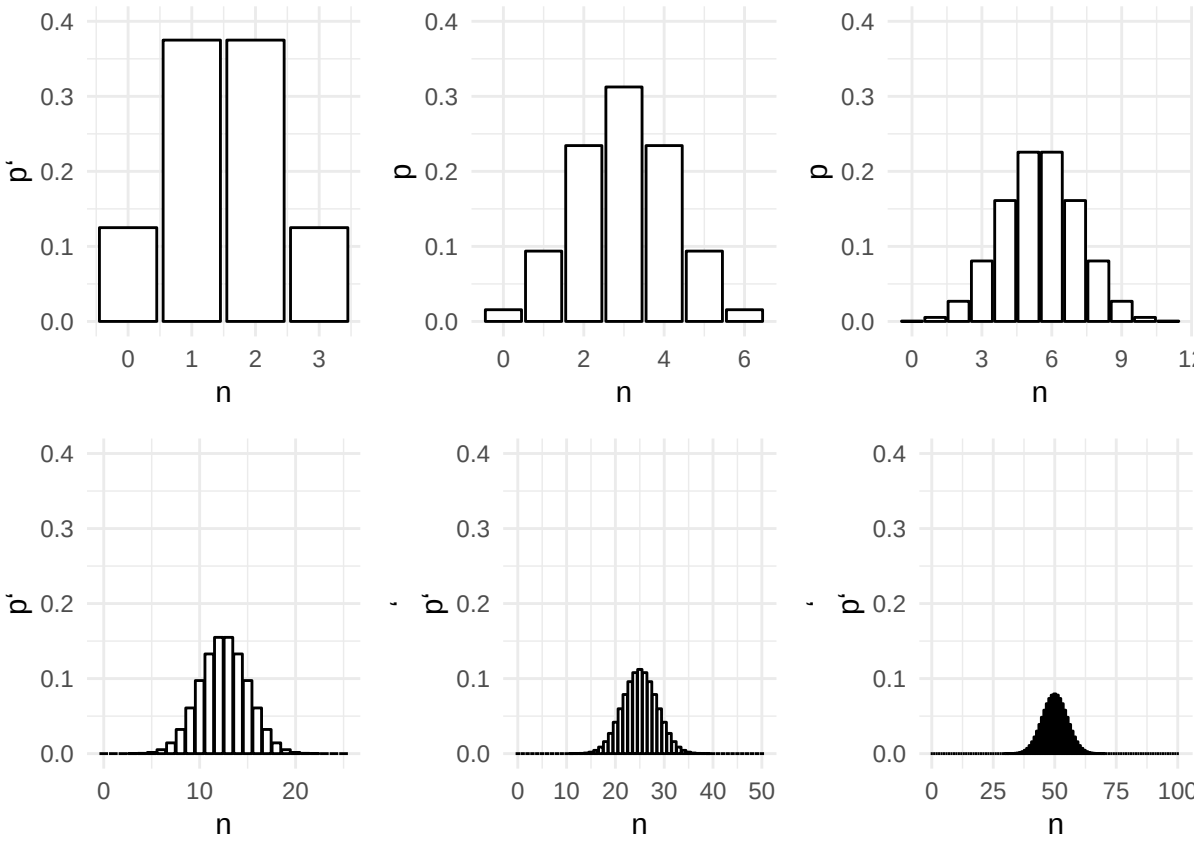

Figura 1: Distribución de probabilidad para resultados de 3, 6, 12, 25, 50 y 100 intentos binarios igualmente probables.

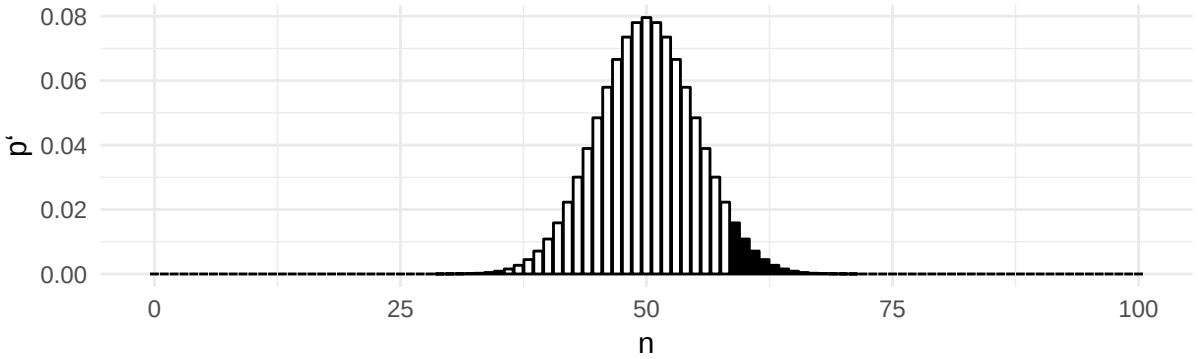

Figura 2: En negro, probabilidad de obtener 59 respuestas o más que clasifican camino como un nombre.

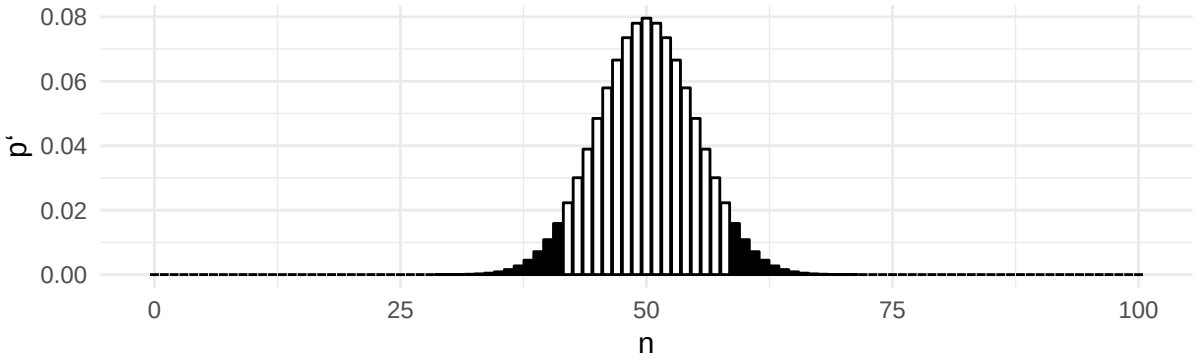

Figura 3: En En negro, probabilidad de obtener 41 respuestas o menos y 59 respuestas o más que clasifican camino como un nombre.

aún más de H_0 cuando H_0 es verdadera, tenemos que mirar hacia ambos lados de la distribución (que los sujetos respondan que es nombre 59 o más veces y que los sujetos respondan que es nombre 41 o menos veces). Así, obtenemos una frecuencia acumulada de $p = 0{,}089$ (véase Figura 3). Dado que este resultado es mayor al $p_{\text{crítico}}$ que definimos, no podemos rechazar H_0.

Cuando tenemos conocimiento previo sobre un fenómeno, podemos formular una hipótesis direccional. Esto nos habilita a que el resultado necesario para una conclusión significativa sea menos extremo que en el caso de una hipótesis no direccional. En la mayoría de los casos, el valor p que obtenemos para un resultado con una hipótesis direccional es la mitad del valor p obtenido para una hipótesis no direccional.

8.3. Ejercitación

1. Completá el siguiente cuadro:

Valor	Significancia	Indicación
$p < 0{,}001$	_____	***
$0{,}001 \leq p < 0{,}01$	muy significativo	_____
$0{,}01 \leq p <$_____	significativo	*
0,05_____ p _____ 0,1	marginalmente significativo	*ms* o .
_____	_____	_____

2. Un investigador acepta su hipótesis alternativa (que la sonorización de oclusivas sordas en español está influida por el signo zodiacal y el ascendente) y rechaza la nula. No obstante, el corpus de datos presentaba un sesgo importante y, en realidad, tal efecto no existe. ¿Qué tipo de error cometió este investigador?

3. Se quiere estudiar, el uso de tiempos compuestos en español rioplatense a partir de respuestas a preguntas. Se obtuvieron los siguientes resultados:

Caso	Sujeto	Respuesta	Tiempo
1	1	1	S
2	1	2	S
3	2	3	S
4	2	4	S
5	3	5	C
6	3	6	C
7	4	7	S
8	4	8	C

¿Cuál es la probabilidad de obtener estos datos bajo la hipótesis nula de que no hay diferencias de frecuencia entre el uso de tiempos simples y tiempos compuestos u otros que se desvíen todavía más de esta hipótesis? ¿La probabilidad obtenida es significativa?

4. Graficá la distribución de probabilidades de la variable aleatoria tiempo simple e indicá en el gráfico dónde se ubican los datos obtenidos.

9. El reporte

Uno de los pilares sobre los que se basa la ciencia es la replicabilidad de los resultados de una investigación. Para asegurar que nuestro estudio presente esta característica, tenemos que ser tan detallados como sea posible.

El **reporte** de una investigación cuantitativa consiste, por lo general, de cuatro partes: **introducción**, **métodos**, **resultados**, y **discusión**. Si se discute más de un caso de estudio, en el informe, cada caso suele requerir sus propias secciones de métodos, resultados y discusión, seguido de una discusión general. Entre la información a presentar, tenemos que incluir la población estudiada, las hipótesis consideradas y las variables (sin dejar de lado su operacionalización), la confección de la muestra (y las consideraciones que tuvimos en cuenta para que la misma sea representativa y balanceada), la forma en que se almacenaron los datos y los distintos pasos del testeo de hipótesis. Por último, es importante no olvidar que el objetivo final de toda investigación es la comunicación. En este sentido, tenemos que tener en cuenta el poder ilustrativo que tienen los gráficos y las tablas para transmitir información y, por supuesto, tratar de ser tan claros en las explicaciones como sea posible!

9.1. Ejercitación

1. Buscá un artículo sobre una investigación experimental que organice su comunicación siguiendo este formato. ¿Qué información incluye cada sección?

10. Conclusión

En esta clase, discutimos cómo llevar a cabo una investigación cuantitativa. Para ello, analizamos las distintas etapas que involucra llevar a cabo un estudio siguiendo esta

metodología, desde el planeamiento hasta la redacción del reporte. Es crucial para obtener resultados rigurosos atravesar cada una de las etapas de forma responsable y detallada.