ISSN 2796-8952

CUADERNOS DEL GRUPO DE ESTADÍSTICA PARA EL ESTUDIO DEL LENGUAJE

Volumen I

Julia Carden (ed.)
Ezequiel Koile (ed.)
Analí Taboh (ed.)
Federico Alvarez
Santiago Gualchi
Mercedes Güemes
Laura Tallon
Charo Zacchigna

GESEL

Corrección y edición
Julia Carden
Ezequiel Koile
Analí Taboh

Cuadernos del Grupo de Estadística para el Estudio del Lenguaje, I volumen / Federico Alvarez; Santiago Gualchi; Mercedes Güemes; Laura Tallon y Charo Zacchigna.

1era edición.
25 de mayo, 217; primer piso.
Ciudad Autónoma de Buenos Aires, Argentina.
Publicación digital, diciembre 2021.
ISSN: 2796-8952

CUADERNOS DEL GRUPO DE ESTADÍSTICA PARA EL ESTUDIO DEL LENGUAJE

Volumen I

Federico Alvarez, Santiago Gualchi , Mercedes Güemes, Laura Tallon y Charo Zacchigna.

Editado por Julia Carden, Ezequiel Koile y Analí Taboh.

ISSN 2796-8952

Prólogo

Sobre este cuaderno

El Cuaderno del Grupo de Estadística para el Estudio del Lenguaje - Volumen I es una obra colectiva. Los distintos capítulos que lo integran son el resultado de una recopilación bibliográfica llevada a cabo dentro del marco del subsidio FiloCyT "Métodos cuantitativos en el estudio del lenguaje". Cada capítulo fue elaborado para acompañar una reunión del equipo en la que se trataron los distintos temas con mayor profundidad, a modo de presentación oral. Este volumen no pretende ser una obra acabada sobre los temas de estadística que se abordan, sino una guía para la lectura de bibliografía más compleja y completa sobre el tema. Por este motivo, estos capítulos deben ser acompañados de la lecturas que se sugieren en las referencias.

Estos cuadernos fueron concebidos para aportar material de consulta sobre estadística aplicada al estudio del lenguaje que, al menos en español, sigue siendo un área de vacancia bibliográfica. El objetivo es que proporcionen información, acceso a nuevos materiales y una explicación breve y simple sobre los puntos esenciales que deben ser conocidos para quienes se introduzcan en el estudio cuantitativo del lenguaje y no posean una base firme en matemática.

Capítulo II

Diseño y lógica de los estudios cuantitativos

Santiago Gualchi

En este capítulo se propone un recorrido por las distintas etapas de los estudios cuantitativos. Las características específicas de cada una de ellas dependerán de cada caso particular, pero la información ofrecida puede servir de guía general para entender cómo se planifican y ejecutan los estudios cuantitativos y por qué se hace de ese modo. Contrario a la creencia de que los métodos estadísticos arrojan conclusiones objetivas acerca de los fenómenos bajo estudio, las investigaciones cuantitativas son el resultado de una serie de decisiones que el investigador debe tomar: qué hipótesis considerar, cómo operacionalizar las variables, cómo conformar la muestra, entre otras. Aprender a conducir estudios cuantitativos exitosos es, en buena medida, aprender a tomar buenas decisiones y basarse en buenos criterios.

1. Introducción

Los métodos estadísticos requieren de tiempo y práctica para su dominio. No obstante, el esfuerzo invertido es ampliamente compensado por sus diversas ventajas. Los estudios cuantitativos permiten simplificar procesos analíticos, testear la reproducibilidad de resultados, reutilizar herramientas existentes y arribar a soluciones para problemas irresueltos, entre otras ventajas.

La dificultad para adquirir los conocimientos necesarios para la realización de estudios cuantitativos se relaciona, en parte, con la amplia variedad de herramientas que existen y la especificidad de los criterios para su uso adecuado. Las características de un estudio en particular dependerán de diversos factores, como la naturaleza del fenómeno bajo estudio, la estructura de las hipótesis, la calidad y la cantidad de los datos disponibles, entre otros. Todo esto determinará la necesidad de usar unos u otros métodos cuantitativos. En este punto, es importante remarcar que el investigador debe decidir acerca de numerosos aspectos de su estudio que influirán en las conclusiones a las que se arribará finalmente. Aprender a conducir estudios cuantitativos exitosos es, en buena medida, aprender a tomar buenas decisiones y adquirir buenos criterios.

A pesar de la amplia variación que pueden presentar los estudios cuantitativos, existen etapas compartidas por casi todos ellos (si no por todos). En este capítulo, se revisarán estas etapas en común.

2. Entender el fenómeno

El primer paso a la hora de comenzar un estudio cuantitativo consiste en estudiar y caracterizar el fenómeno de interés tanto como sea posible. Las tareas específicas dependerán de cada caso particular, pero en general implica estudiar la bibliografía relevante, observar el fenómeno en escenarios naturales para permitir una primera generalización inductiva, solicitar información adicional, que puede venir de colegas u otras fuentes (Gries, 2021). La omisión o subestimación de este paso muy probablemente conduzca a frustración y tiempo perdido más adelante.

A modo de ejemplo, se puede pensar en el fenómeno de los órdenes de palabras en términos translingüísticos. En este sentido, es conocido que las distintas lenguas presentan distintos ordenamientos de

los constituyentes de sus oraciones. Por ejemplo, si se considera el orden de genitivo y nombre, en español se observa que el orden dominante es nombre-genitivo (NGen):

1. la oreja del conejo,

pero en chino mandarín el orden dominante es genitivo-nombre (GenN) (Dryer, 2013a):

2. rùzi de ěrduō

   conejo PART oreja

   "La oreja del conejo".

Un breve repaso de la bibliografía del tema, permite conocer que los autores que abordaron esta temática estudiaron y propusieron distintas variables para dar cuenta del fenómeno de la variación en los órdenes de palabras. Greenberg (1966) postuló que (i) el orden de sujeto, verbo y objeto, (ii) el orden de adjetivo y nombre, y (iii) la ocurrencia de preposiciones o posposiciones son los principales rasgos sintácticos que permiten explicar un amplio número de otras propiedades formales. Para el caso específico del orden de genitivo y nombre, Greenberg reportó una asociación con el tipo de adposiciones en la lengua. Posteriormente, Dryer (1992) sostuvo que la variable que mejor explica el comportamiento de los órdenes de constituyentes es el orden de objeto y verbo. Por su parte, Dunn et al. (2011) consideraron que los órdenes de palabras varían libremente respecto de otras características formales, pero que se encuentran fuertemente condicionados por el devenir filogenético de la lengua. Esto no quiere decir que Greenberg (1966) o Dryer (1989, 1992) desconocieran la influencia de factores extralingüísticos en las propiedades formales de las lenguas, pero sus estudios buscaron descontar el efecto de estos factores y concentrarse en sus propiedades sintácticas.

Posteriormente, el estudio de Dunn et al. (2011), que arribaba a conclusiones disruptivas, fue fuertemente criticado (véase Baker 2011, Croft et al. 2011, Dryer 2011, Longobardi y Robert 2011, entre otros). La mayoría de estas críticas estuvieron asociadas a la etapa de los estudios cuantitativos que da nombre a esta sección: entender el fenómeno. Siguiendo a estos autores, el estudio de Dunn et al. (2011) presentaba problemas metodológicos originados en una mala comprensión del fenómeno de interés y de sus antecedentes en la literatura. Entre otras dificultades, Dunn et al. (2011) malinterpretaron los modelos que se habían propuesto en la bibliografía (y que ellos criticaron) (Dryer, 2011) y emplearon métodos cuantitativos tomados de la biología que no contemplan el efecto del contacto lingüístico sobre las propiedades sintácticas de las lenguas (Croft et al., 2011).

3. Hipótesis y operacionalización

Una vez que se tiene una visión general del fenómeno que queremos estudiar, el siguiente paso consiste en formular las hipótesis.

3.1. Hipótesis científicas en forma de texto

Las hipótesis son enunciados que refieren a más de un evento singular y que pueden ser falseadas (i.e., se pueden pensar eventos o situaciones que contradigan al enunciado) y testeadas (i.e., están dados los medios para llevar a cabo las acciones requeridas para conocer el valor de verdad del enunciado). Por ejemplo, siguiendo esta definición, un enunciado como "Si no le hablo a mi hijo durante su primer año, desarrollará más rápidamente el lenguaje" no es una hipótesis porque se ocupa de un evento singular (el desarrollo del

ISSN: 2796-8952

lenguaje en mi hijo). Tampoco son hipótesis “Si no se le habla a los niños durante su primer año, podrían desarrollar más rápidamente el lenguaje” ni “Si no se le habla a los niños durante su primer año, desarrollarán más rápidamente el lenguaje”. La primera no es falsable. El uso de “podría” implica que la hipótesis será verdadera independientemente de si la falta de estímulo en el primer año acelera el desarrollo de la lengua. La segunda no es testeable, ya que por razones éticas resultaría difícil obtener sujetos experimentales que puedan ser sometidos a la pruebas necesarias para conocer su valor de verdad.

Más allá de estas características generales, las hipótesis pueden presentar distintas estructuras lógicas. Gries (2021) propuso clasificarlas en:

- Hipótesis de diferencia/independencia
- Hipótesis de bondad de ajuste

Las hipótesis del primer tipo presentan una estructura condicional (si X, entonces Y), o, al menos, pueden ser parafraseadas como tales. Por ejemplo, para el estudio de la variación en el orden de genitivo y nombre, se puede pensar la siguiente hipótesis:

(1) Si una lengua presenta orden objeto-verbo (OV), entonces también presentará orden GenN, y si presenta orden verbo-objeto (VO), entonces también presentará orden NGen.

En este enunciado aparecen tres variables. Las **variables** son símbolos que pueden tomar, por lo menos, dos estados o niveles diferentes, y que se oponen a las **constantes**, que siempre presentan un mismo valor sin experimentar variación. En el ejemplo, las variables involucradas son (i) el orden de genitivo y nombre, que es la variable que se quiere explicar, (ii) el orden de objeto y verbo, que es la variable mediante la cual se quiere predecir el comportamiento de la anterior, y (iii) la lengua, que sirve como identificador de cada observación (dado que solo podemos tener una observación de estas características por lengua). Una constante de los datos que competen a cualquier estudio que aborde esta hipótesis puede ser el entorno al que pertenecen las lenguas bajo estudio. En todos los casos será la Tierra. Esta consideración puede resultar ridícula a primera vista, pero dado que las propiedades formales de una lengua dependen en buena medida de su filogenia y su entorno, la posibilidad de que las muestras de lenguas que podamos obtener estén afectadas por características accidentales de la historia de la humanidad debe ser tenida en consideración (para una discusión más elaborada del tema, véase Dryer, 1989).

Volviendo a las variables, es posible clasificarlas de acuerdo a cómo se relacionan entre sí (Gries, 2021):

- **variable independiente (o predictor)**: es la variable presente en la prótasis, y suele referirse a la causa de los cambios/efectos (aunque no necesariamente). La variable independiente representa tratamientos o condiciones que el investigador controla (directa o indirectamente) para entender sus efectos sobre la variable dependiente.
- **variable dependiente (o de respuesta)**: es la variable presente en la apódosis, cuyos valores, variación o distribución se quieren explicar. La variable dependiente suele ser 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.

ISSN: 2796-8952

Hasta ahora se refirieron las características de las hipótesis de diferencia/independencia. En las hipótesis de bondad de ajuste, se tiene 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:

(2) Si una lengua presenta orden OV, entonces también presentará orden GenN.

A simple vista, podría parecer que en esta hipótesis encontramos dos variables, pero, dado que la hipótesis no expresa nada acerca del orden VO, el orden de objeto y verbo es en verdad una constante que restringe la población. Una versión más débil de esta hipótesis, pero que sirve para resaltar estas diferencias, es la siguiente:

(3) En las lenguas con orden OV, el orden GenN es más frecuente que el orden NGen.

Si bien este no es el único modo de proceder, en lingüística (y otras ciencias) el paso que suele suceder a la formulación de la hipótesis que se presenta como novedosa –la hipótesis alternativa (H_1)– consiste en definir las situaciones y estados de cosas que falsarán dicha hipótesis –la hipótesis nula (H_0). Es importante señalar que la hipótesis nula no necesariamente enuncia la ausencia de un efecto, sino que es la hipótesis que se asume en un contexto de incertidumbre. "Nula" no hace referencia a una hipótesis que enuncia la falta de una diferencia, sino a la hipótesis que se pretende anular (Gigerenzer, 2004). Por ejemplo, si se careciera de evidencia acerca de si los cocodrilos sienten o no dolor, la hipótesis nula (la que aceptaríamos por defecto a falta de mayor evidencia) podría ser que sí sienten dolor (ya sea por razones éticas o por conocimiento del dolor en otras especies). Es importante remarcar que, si bien la hipótesis nula no necesariamente debe ser el opuesto lógico estricto de la hipótesis alternativa, ambas hipótesis deben cubrir todo el espacio de resultados o espacio muestral, esto es, el conjunto de todos los resultados teóricamente posibles de nuestro experimento. Por ejemplo, se pueden plantear las siguientes hipótesis nulas para los ejemplos de hipótesis alternativas dados anteriormente:

(4) Hipótesis nula de (1)
El orden de genitivo y nombre es independiente del orden de objeto y verbo[1].

(5) Hipótesis nula de (2) y (3)
En las lenguas con orden OV, no existen diferencias en la frecuencia de los niveles de orden de genitivo y nombre.

3.2. Operacionalización de variables

Una vez formuladas las hipótesis, es importante encontrar un modo de operacionalizar las variables. Esto supone decidir qué será observado, contado, medido, etc. cuando se investiguen las hipótesis (Gries, 2021). La operacionalización de variables involucra el uso de niveles numéricos para representar sus estados. Un número puede ser una medida (p. ej., 402 milisegundos de tiempo de reacción), pero los niveles, esto es, estados discretos no numéricos, también pueden ser codificados usando números. Por ejemplo, volviendo a la variable de orden de genitivo y nombre, es posible operacionalizarla como una variable dicotómica: GenN vs NGen. En este caso, si se quieren codificar los niveles numéricamente, podríamos asignar 0 al orden GenN y 1 al orden NGen (o viceversa). También podría asignarse -½ y ½, respectivamente, o se podría querer incluir un tercer nivel para lenguas que no presentan un orden dominante de genitivo y nombre. Otra

[1] En probabilidad, un evento A es independiente de un evento B, si la probabilidad de observar A no varía según B haya sido observado o no (Casella y Berger, 2002). Para entenderlo más fácilmente, la interpretación es la misma que se hace con el adverbio "independientemente" cuando se dice, por ejemplo, "Los padres harán el reclamo independientemente de que el director presente una disculpa", donde la probabilidad del evento A (que los padres hagan el reclamo) no varía según ocurra o no el evento B (que el director presente una disculpa).

ISSN: 2796-8952

posibilidad sería considerar al orden de genitivo y nombre como una variable continua y asignar como valor la proporción de construcciones que presentan el orden GenN del total de construcciones de genitivo y nombre en un corpus de la lengua. En este sentido, es crucial remarcar que la forma en que se operacionalice es una decisión que toma el investigador y que dependerá de aspectos teóricos, pero también de aspectos prácticos como los datos de los que se dispone o los métodos estadísticos a ser usados. Según los niveles de medida que resulten de dicha operacionalización, podemos clasificar las variables en (Johnson, 2013):

- **variable categórica:** si bien estas variables pueden codificarse numéricamente, no expresan una cantidad, sino la pertenencia a un grupo. En el caso en que la variable pueda tomar solo dos niveles, se dice que es **dicotómica** (p. ej., orden de genitivo y nombre con niveles GenN y NGen). Y, en el caso en que pueda tomar tres o más niveles y estos presentan un orden natural, se dice que es una variable **ordinal** (p. ej., orden de genitivo y nombre con niveles "sin construcciones con orden GenN", "solo construcciones pronominales con orden GenN", "siempre orden GenN").
- **variable numérica** además de distinguir categorías y rankear objetos, también permiten comparar las diferencias y las razones entre valores de forma significativa (p. ej., orden de genitivo y nombre como la proporción de construcciones con orden GenN en un corpus).

3.3. Hipótesis científicas en formato estadístico/matemático

Después de formular las hipótesis en forma de texto y definir cómo operacionalizar las variables, es necesario formular la versión estadística de las hipótesis (Gries, 2021). Esto significa expresar los resultados cuantitativos esperados sobre la base de las hipótesis textuales. Dichos resultados suelen involucrar formas matemáticas como frecuencias, promedios o distribuciones.

Este va a ser el formato que se usará para evaluar la significancia de las hipótesis (véase más abajo), y su definición va a depender directamente de cómo se hayan operacionalizado las variables. Retomando la hipótesis de que si una lengua presenta orden OV, entonces también presentará orden GenN, su forma matemática y la de la hipótesis nula serán:

(6) Hipótesis alternativa
En las lenguas con orden OV, el número de observaciones con orden GenN es mayor que el número de observaciones con orden NGen.

(7) Hipótesis nula
En las lenguas con orden OV, el número de observaciones con orden GenN no es distinto al número de observaciones con orden NGen.

La hipótesis alternativa que planteamos es direccional, se espera encontrar más observaciones para un determinado nivel de la variable orden de genitivo y nombre (GenN). También se podría haber propuesto una variable no direccional:

(8) Hipótesis alternativa no direccional textual
En las lenguas con orden OV, el orden GenN y el orden NGen no son igualmente frecuentes.

(9) Hipótesis alternativa no direccional matemática
En las lenguas con orden OV, el orden GenN y el orden NGen son igualmente frecuentes .

ISSN: 2796-8952

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 estudian todos los individuos u objetos de interés (esto es, la población) sino una muestra (Gries, 2021). Si se quiere que los datos puedan generalizarse a la población, esta muestra debe ser (Gries, 2021):

- **representativa**: las distintas partes de la población deben estar reflejadas en la muestra, y
- **balanceada**: 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 se conocen todas las partes y las proporciones de la población. Una forma de obtener una muestra más representativa y balanceada es mediante randomización, es decir, seleccionando los elementos que conforman la muestra en forma aleatoria.

A modo de ejemplo, si el estudio trata acerca de las propiedades de las lenguas humanas habladas en la actualidad, para que la muestra sea representativa debería incluir lenguas de todas las regiones del mundo y de todas las familias lingüísticas, y para que sea balanceada debería incluirlas en forma proporcional a la porción de la población que representan. Qué lenguas incluir podría ser determinado mediante randomización, o en función de los datos a los que es posible acceder. En el segundo caso, se dice que empleamos una muestra de conveniencia.

En cambio, si nuestra población no son las lenguas habladas en la actualidad, sino las lenguas humanas posibles, el diseño de la muestra será más complejo y requerirá de supuestos teóricos más fuertes (como se señaló en "Entender el fenómeno"). En este caso, las lenguas habladas en la actualidad, si bien son un subconjunto, no son una muestra representativa y balanceada de las lenguas humanas posibles. Esto es así dado que la distribución de lenguas en el mundo no es solo el resultado de las características del soporte cognitivo que habilita el lenguaje humano, sino también de factores culturales e históricos accidentales que podrían introducir sesgos en la muestra y conducir a generalizaciones incorrectas acerca de la población (véase Dryer, 1989).

5. Almacenamiento

Una vez recolectados los datos, es necesario almacenarlos en un formato que nos permita anotarlos, manipularlos y evaluarlos fácilmente. Siempre que sea posible, se deben usar formatos libres y estándar. Además, se debe procurar que los datos puedan ser fácilmente interpretados por humanos y computadoras (clases y conversaciones con Johann-Mattis List). Si bien el formato específico dependerá de las características de los datos y de los métodos mediante los cuales se analizarán, usualmente la configuración más cómoda para el análisis es el formato *case-by-variable* (Gries, 2021):

- la primera fila contiene los nombres de las variables;
- las otras filas representan cada una un ***data point*** (esto es, 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);

ISSN: 2796-8952

- las otras columnas representan una sola variable o característica correspondiente a un determinado *data point*; y
- siempre debe usarse el mismo símbolo para la información faltante y solo debe ser usado con ese fin. Si se elige, por caso, usar "NA" para los datos faltantes, ninguna variable podrá usar "NA" como valor para ninguno de sus niveles. Si se tiene una variable "Región", por ejemplo, se tendrá que evitar usar "NA" para referir a Norteamérica.

La Tabla 1 muestra un ejemplo de este formato.

ID	Lengua	Orden de objeto y verbo	Orden de genitivo y nombre
1	Euskera	OV	GenN
2	Japonés	OV	GenN
3	Urdu	OV	GenN
4	Farsi	OV	NGen
5	Selknam	OV	NA

Tabla 1. Ejemplo de formato case-by-variable. Datos tomados de Dryer (2013a, 2013b)

Aunque este formato es conveniente en la mayoría de las situaciones, existen excepciones (Wickham, 2017). Para estructurar familias lingüísticas, por ejemplo, puede ser más efectivo utilizar el formato Newick (véase Felsenstein, 2021), que permite representar diagramas de árboles con un número irrestricto de niveles y de nodos por nivel (como ejercicio, imagínese cómo estructurar la información filogenética de las lenguas indoeuropeas y de lenguas aisladas como el euskera en el formato *case-by-variable*). Este formato puede encontrarse en los datos de familias lingüísticas puestos a disposición en Glottolog (Hammarström et al. 2021).

6. El testeo

Cuando ya almacenamos los datos, se continúa 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 la hipótesis alternativa es correcta, sino en probar que la versión estadística de la hipótesis nula es improbable y, por lo tanto, puede ser rechazada. Considerando que las hipótesis nula y alternativa agotan el universo de eventos posible, rechazar la hipótesis nula resulta en un sustento a la hipótesis alternativa. El testeo de hipótesis nos ayuda a decidir si un efecto observado se debe a relaciones reales entre las variables o al azar. Dicho de otra forma, nos permite justificar la preferencia por explicar un fenómeno por medio de una relación entre variables contra considerar que dicha relación no tiene influencia sobre el efecto observado.

A nadie le sirve perder tiempo y dinero analizando/interpretando/considerando conclusiones incorrectas. Para sortear esto existen distintas técnicas. Uno de los procedimientos que se utilizan es la **Prueba de Significancia de la Hipótesis Nula** (**NHST**, por sus siglas en inglés). En líneas muy generales, es posible definirla como sigue (Gries, 2021):

1. definición del nivel de significancia, *a* (que por lo general es 0,05);

2. análisis de los datos computando la probabilidad de un efecto *e* (p. ej., una distribución, una diferencia de medias, una correlación) usando las hipótesis estadísticas;
3. computación de la probabilidad de error, o valor *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 *a* y *p*, y decisión: si $p < a$, entonces rechazo de H_0 y aceptación de H_1.

Si el valor *p* de un fenómeno es menor a *a* 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 Tabla 2 recoge la semántica estándar de dicho valor:

Valor	Significancia
$p < 0{,}001$	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
$0{,}1 \leq p$	no significativo

Tabla 2. Semántica estándar del valor p.

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. Sin embargo, existen dos combinaciones lógicas más (véase Tabla 3). Un error de tipo I (*falso positivo)* ocurre cuando la hipótesis nula es verdadera y se la rechaza. Por lo general, estos errores deben ser evitados tanto como sea posible, y es lo que se busca hacer cuando llevamos a cabo pruebas de testeo de hipótesis. Asimismo, un error de tipo II (*falso negativo)* ocurre cuando la hipótesis alternativa es verdadera y se la rechaza. Estos errores suelen ser menos graves que los de tipo I, pero de todos modos debemos reducir la probabilidad de que ocurran.

	H0 es verdadera	H1 es verdadera
Rechaza H0	error de tipo I	correcto
No rechaza H0	correcto	error de tipo II

Tabla 3. Tipos de error.

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

Supóngase que se quiere estudiar la asociación entre el orden de objeto y verbo y el orden de genitivo y nombre en términos translingüísticos a partir de la siguiente hipótesis alternativa:

(10) Si una lengua presenta orden OV, entonces también presentará orden GenN.
(11) "En las lenguas con orden OV, el número de observaciones con orden GenN es mayor que el número de observaciones con orden NGen".

La hipótesis nula correspondiente será la siguiente:

(12) En las lenguas con orden OV, no existen diferencias en la frecuencia de los niveles de orden de genitivo y nombre.

(13) Hipótesis nula
En las lenguas con orden OV, el número de observaciones con orden GenN no es distinto al número de observaciones con orden NGen.

Para testear las hipótesis, se necesitará una muestra balanceada y representativa de la población: las lenguas humanas posibles. Para simplificar el ejemplo, se asumirá que la muestra cumple con estos requisitos. Si se obtienen datos de tres lenguas con orden OV y en los tres casos tienen orden genitivo-nombre, ¿es posible rechazar H_0 y aceptar H_1 asumiendo un $a = 0{,}05$? La Tabla 4 sintetiza la probabilidad de cada posible resultado bajo el supuesto de que H_0 es verdadera. Las columnas GenN y NGen 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). La columna $p_{resultados}$ representa la probabilidad de que ocurra la combinación de valores para las tres lenguas. Dado que bajo la H_0 asumimos que GenN y NGen son valores igualmente probables, la probabilidad de que una lengua sea GenN o NGen es la misma (P (GenN) + P (NGen) = 1; P(GenN) = P(NGen) = 0,5). La probabilidad de cada combinación puede calcularse de dos formas. La primera es, 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 ÷ 8 = 0,125. Otra posibilidad consiste en calcular el producto de las probabilidades de las 3 respuestas correspondientes a la combinación: 0,5 × 0,5 × 0,5 = 0,125. Dado que bajo H_0 la probabilidad de que las tres lenguas sean GenN es de $p = 0{,}125$ y que $p > a$, no es posible rechazar H_0.

Escenario	Lengua 1	Lengua 2	Lengua 3	GenN	NGen	$p_{resultados}$
A	GenN	GenN	GenN	3	0	0,125
B	GenN	GenN	NGen	2	1	0,125
C	GenN	NGen	GenN	2	1	0,125
D	GenN	NGen	NGen	1	2	0,125
E	NGen	GenN	GenN	2	1	0,125
F	NGen	GenN	NGen	1	2	0,125
G	NGen	NGen	GenN	1	2	0,125
H	NGen	NGen	NGen	0	3	0,125

Tabla 4. Probabilidad de cada escenario bajo H_0.

La distribución de probabilidad de cada resultado posible para tres lenguas queda recogida en el primer diagrama de barras de la Figura 1. Como se observa en los sucesivos gráficos de dicha figura, a medida que aumenta el número de lenguas la distribución se asemeja cada vez más a la de la distribución gaussiana o normal.

ISSN: 2796-8952

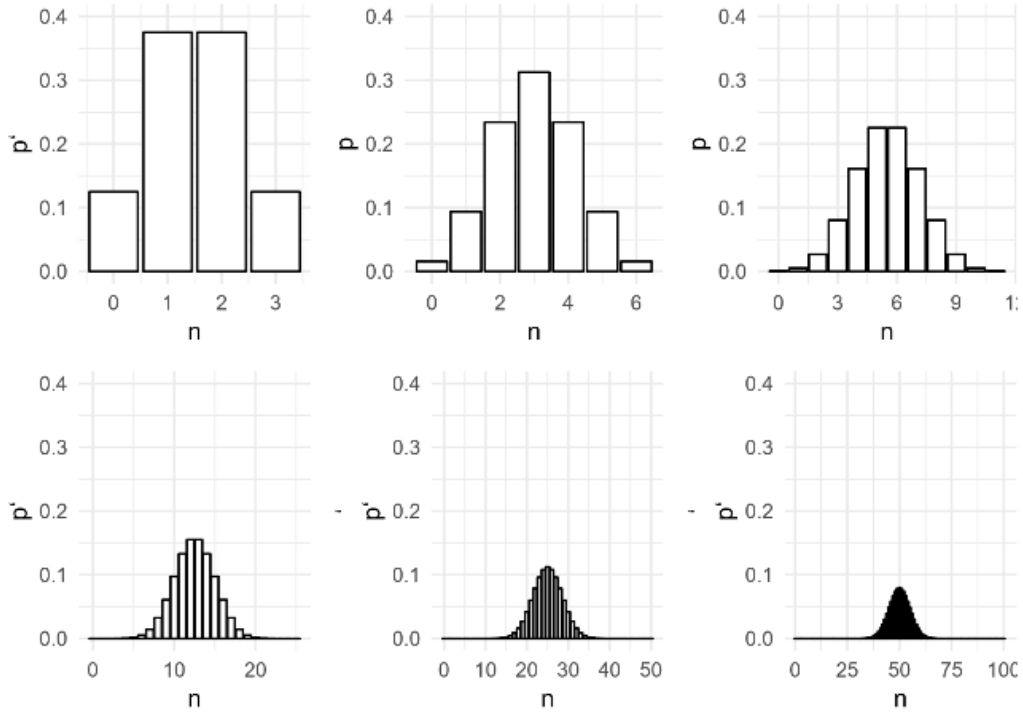


Figura 1. Distribución de probabilidad para resultados de 3, 6, 12, 25, 50 y 100 intentos binarios igualmente probables.

Ahora bien, si recolectamos datos de 100 lenguas, ¿podemos rechazar H_0 si 59 tienen orden GenN? La respuesta es sí: asumiendo H_0, la probabilidad de que 59 lenguas o más tengan orden GenN es de p = 0,044 (véase Figura 2).

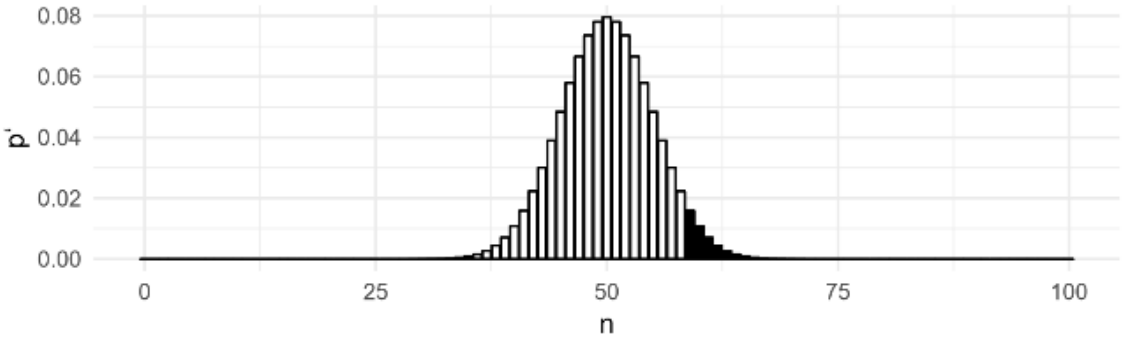


Figura 2. En negro, probabilidad de observar 59 lenguas o más con orden GenN en una muestra de 100 lenguas.

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

En la sección anterior, la H_1 era direccional: "Si una lengua presenta orden OV, entonces también presentará orden GenN.". 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 por H_0. Si, en

ISSN: 2796-8952

cambio, asumimos una H_1 no direccional (por ejemplo, "En las lenguas con orden OV, el orden GenN y el orden NGen no son igualmente frecuentes"), tenemos que mirar ambas colas de la distribución:

(14) H_0 matemática:
En las lenguas con orden OV, el orden GenN y el orden NGen no son igualmente frecuentes.

(15) H_1 matemática:
En las lenguas con orden OV, el orden GenN y el orden NGen son igualmente frecuentes.

Ahora imaginemos que, una vez más, en una muestra de 100 lenguas con orden OV, 59 tienen orden GenN. Ya que la hipótesis es no direccional y se quiere calcular la probabilidad de que ocurra dicho resultado u otro que se desvíe aún más de H_0 cuando H_0 es verdadera, tenemos que mirar hacia ambos lados de la distribución (que el orden GenN ocurra 59 veces o más, o que ocurra 41 veces o menos). Así, obtenemos una probabilidad de p = 0,088 (véase Figura 3). Dado que este resultado es mayor al a que definimos, no podemos rechazar H_0.

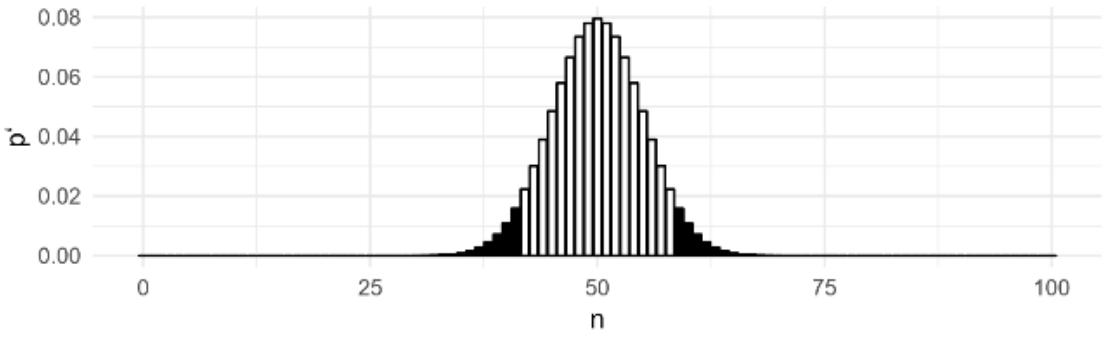


Figura 3. En negro, probabilidad de observar 41 lenguas o menos con orden GenN o 59 lenguas o más con dicho orden en una muestra de 100 lenguas.

Cuando se tiene conocimiento previo sobre un fenómeno, es posible formular una hipótesis direccional. Esto habilita que el resultado necesario para una conclusión significativa sea menos extremo que en el caso de una hipótesis no direccional. Siempre que la distribución sea simétrica, el valor p que se obtiene para un resultado con una hipótesis direccional es la mitad del valor p obtenido para una hipótesis no direccional.

7. El reporte

Uno de los puntos fundamentales sobre los que se basa la ciencia es la replicabilidad de los resultados de las investigaciones. Para asegurar que un estudio presente esta característica, es necesario ser tan detallados como sea posible al comunicarlo. El reporte de una investigación cuantitativa consiste, por lo general, en cuatro secciones: 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

ISSN: 2796-8952

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.

8. Referencias

Baker, M. C. (2011). The interplay between Universal Grammar, universals, and lineage specificity. *Linguistic Typology*, *15*(2), 473-482. https://doi.org/10.1515/LITY.2011.031

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Duxbury.

Croft, W., Bhattacharya, T., Kleinschmidt, D., Smith, D. E., & Jaeger, T. F. (2011). Greenbergian universals, diachrony, and statistical analyses. *Linguistic Typology*, *15*(2), 433-453. https://doi.org/10.1515/lity.2011.029

Dryer, M. S. (1989). Large linguistic areas and language sampling. *Studies in language*, *13*(2), 257-292. https://doi.org/10.1075/sl.13.2.03dry

Dryer, M. S. (1992). The Greenbergian word order correlations. *Language*, *68*, 81-138. https://doi.org/10.2307/416370

Dryer, M. S. (2011). The evidence for word order correlations. *Linguistic Typology*, *15*(2), 335-380. https://doi.org/10.1515/lity.2011.024

Dryer, M. S. (2013a). Order of Genitive and Noun. En M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/86

Dryer, M. S. (2013b). Order of Object and Verb. En M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/83

Dunn, M., Greenhill, S. J., Levinson, S. C., & Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, *473*(7345), 79-82. https://doi.org/10.1038/nature09923

Felsenstein, J. (2021, diciembre 2). The Newick tree format. *PHYLIP*. Recuperado el 2 de diciembre de 2021 de https://evolution.genetics.washington.edu/phylip/newicktree.html

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*, 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Greenberg, J. H. (1966). Some universals of grammar with particular reference to the order of meaningful elements. En J. H. Greenberg (Ed.), *Universals of language* (2da ed., pp. 73-113). The MIT Press.

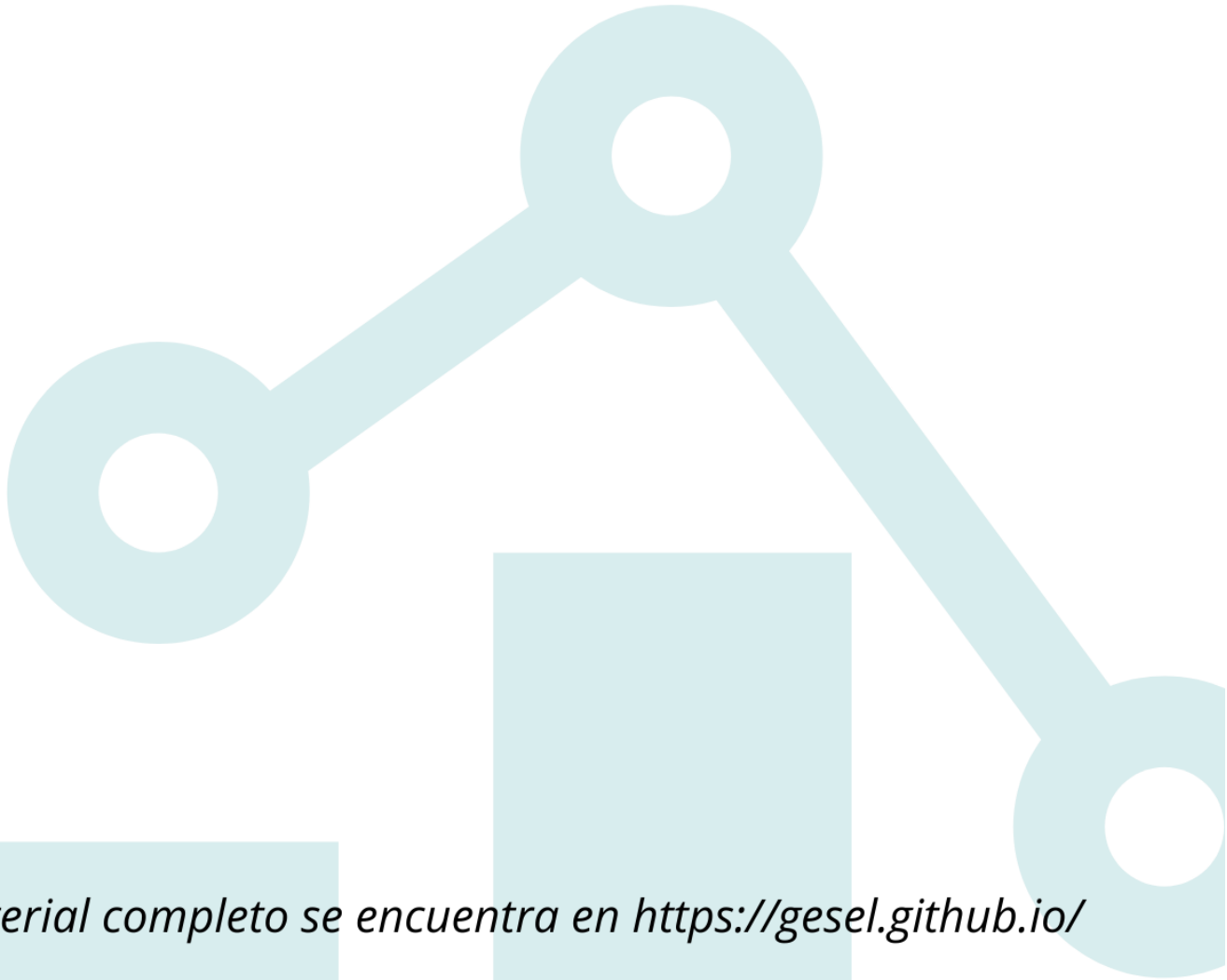
Gries, S. T. (2021). Some fundamentals of empirical research. En *Statistics for linguistics with R* (3a ed., pp. 1-50). de Gruyter Mouton. https://doi.org/10.1515/9783110718256-001

Hammarström, H., Forkel,R., Haspelmath, M., & Bank, S. (2021). *Glottolog 4.5*. Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.5772642

Longobardi, G., & Roberts, I. (2011). Non-arguments about non-universals. *Linguistic Typology*, *15*(2), 483-495. https://doi.org/10.1515/lity.2011.032

Wickham, H., & Grolemund, G. (2017). *R for data science*. O'Reilly.

ISSN: 2796-8952



El material completo se encuentra en https://gesel.github.io/