Diseño y lógica de los estudios cuantitativos

Grupo de estadística para el estudio del lenguaje | GESEL
Reunión n.º 2 – 6 de septiembre de 2018

Santiago Gualchi
santiagogualchi@gmail.com

> Statistical methods were adopted, not because they were easy to master, but because people realized that they were worth the effort. The new methods provided solutions to problems that had frustrated the field for years.
>
> Abney, S. (2011). "Data-Intensive Experimental Linguistics".

1. Introducción

En nuestra primera reunión, 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), la oferta académica excluye 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 presentación. 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 pueden ayudarnos a entender y explicar mejor los fenómenos estudiados. Entre los campos señalados incluimos (sin ser exhaustivos) la psicolingüística y la neurolingüística, la lingüística de corpus, la lingüística computacional y NLP, 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.

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 variables que determinará 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

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

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

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.

3. Hipótesis y operacionalización

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

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*

3.1. Hipótesis científicas en forma de texto

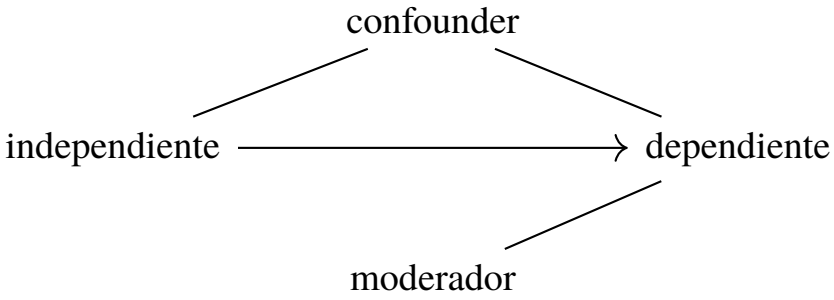
Características de las hipótesis de tipo 1:

- Es un enunciado general ocupado de más de un evento singular.
- Es un enunciado con una estructura condicional (*si..., entonces...*), o, al menos, puede ser parafraseado como tal.
- Es potencialmente falsable (i.e., se pueden pensar eventos o situaciones que contradigan al enunciado) y testeable (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 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.

Figura 1: Tipos de variables según su influencia.



Tipos de variables 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 (véase Figura 1).

Existe otro tipo de hipótesis (de tipo 2) que contiene 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 de muestra [*sampling context*], 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.

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 from the table*);

- ANIMACIDAD: *animado* o *inanimado*;
- CONCRECIÓN: *concreto* o *abstracto*; y
- TIPO del OD: *pronominal* (e.g., *He picked him up this morning*), *semi-pronominal* (e.g., *He picked something up from the floor*), *léxico* (e.g., *He picked people up this morning*) o *nombre propio* (e.g., *He picked 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.

Tipos de variable según sus niveles de medida:

- 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).
- variable cuantitativa [*ratio variable*]: además de distinguir categorías y *rankear* objetos, también permiten comparar las diferencias y los ratios entre valores (e.g., LARGO EN SÍLABAS) de forma significativa.

3.3. Hipótesis científicas 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 significación 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:

$$
\begin{aligned}
H_{1 direccional}: &\quad n_{SSPPdir.enVPO} < n_{SSPPdir.enVOP} \\
H_{1 no\ direccional}: &\quad n_{SSPPdir.enVPO} \neq n_{SSPPdir.enVOP} \\
H_0: &\quad n_{SSPPdir.enVPO} = n_{SSPPdir.enVOP}
\end{aligned}
$$

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 más representativa y balanceada es a partir de la aleatorización [*randomization*]. Este es uno de los principios más importantes de la recolección de datos.

5. Almacenamiento

Una vez que recolectamos los datos, 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.

Formato *case-by-variable*:

- 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 "NA" (véase Cuadro 3), y este símbolo *solo* debe usarse con dicho significado.

6. La decisión

Cuando ya almacenamos los datos, procedemos a evaluarlos con alguna prueba estadística. Sin embargo, la forma de proceder que se acostumbra en ciencias

Cuadro 3: 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
1	vpo	2
2	vpo	2
3	vop	2
4	vop	2
5	vop	2
6	vop	2
7	vpo	3
8	vpo	3
9	vop	3
...	...	...

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.

Supongamos que una enfermedad bacteriana mata a la mitad de los pacientes que afecta mientras que la otra mitad se logra recuperar. Ahora supongamos que probamos un medicamento en 10 pacientes y encontramos que la proporción de pacientes que se curaron fue del 70 %. ¿Podemos atribuir esta diferencia al medicamento administrado? Ahora supongamos que un médico hacia finales del siglo xviii procede de igual forma y obtiene los mismos resultados, pero en vez de recetar una droga aplica sanguijuelas sobre el cuerpo de los pacientes. Una vez más, ¿podemos atribuir la diferencia de pacientes recuperados respecto de la media a la aplicación de este tipo de terapia? El testeo de hipótesis nos permite decidir si un efecto observado se debe a relaciones reales entre las variables o al azar[1]. Tanto el estadístico como las editoriales, y los demás actores sociales quieren evitar perder tiempo y plata analizando/interpretando/considerando datos irreales. Para sortear esto existen distintas técnicas.

Uno de los procedimientos que se utilizan (especialmente en psicología) es la Prueba de Significación de la Hipótesis Nula [*Null Hypothesis Significance Testing*] (NHST)[2]:

[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

Cuadro 4: Significación de los valores de p.

Valor	Significación	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. definición del *nivel de significación* $p_{crítico}$, 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 la estadística en las hipótesis estadísticas;

3. computación de la *probabilidad de error 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 $p_{crítico}$ 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 $p_{crítico}$ 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 p*. El Cuadro 4 recoge la semántica de dicho valor. Al analizar la significación 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 5). Sin embargo, existen dos combinaciones lógicas más. Un error de tipo I ocurre cuando la hipótesis nula es verdadera y se rechaza. Los errores de este tipo 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 son menos graves que los de tipo I, pero, de todos modos, debemos reducir la posibilidad de que ocurran.

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

Supongamos que estudiamos morfos cero en la derivación. Para ello encuestamos 3 sujetos acerca de si la palabra *camino* es un verbo o un nombre,

Perezgonzalez (2015).

Cuadro 5: 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

asumiendo que los nombres son más prototípicos y, 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 6 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, 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.

[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.

Cuadro 6: 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	$p_{\text{resultado}}$
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 de la Figura 2. Como se observa en dicha figura, a medida que aumenta el número de sujetos la distribución 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 3).

6.2. 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 significación que discutimos es una *prueba de una cola* [*one-tailed test*], 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

Figura 2: Distribución de probabilidad para resultados de 3, 6, 12, 25, 50 y 100 intentos binarios igualmente probables.

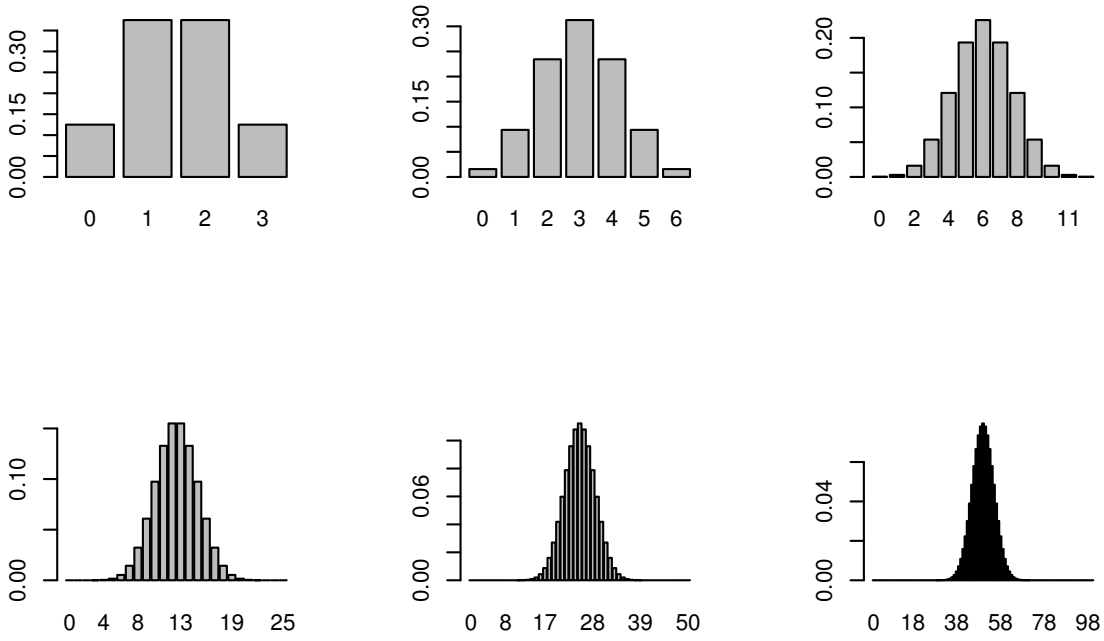
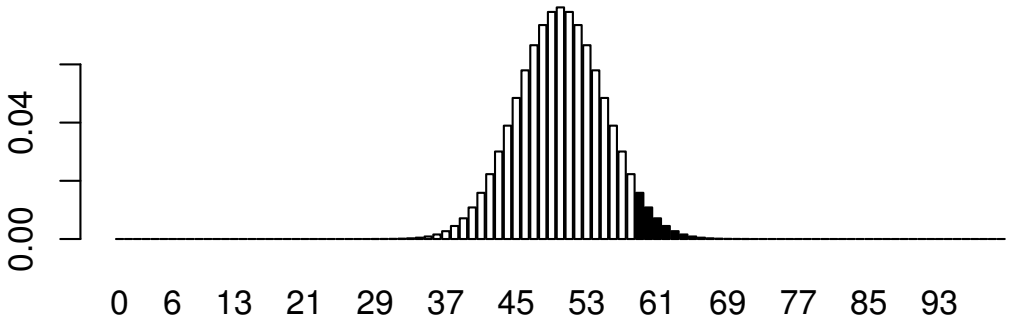


Figura 3: En negro, probabilidad de obtener 59 respuestas o más que clasifican *camino* como un nombre.



es no direccional y queremos 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 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 4). Dado que este resultado es mayor al $p_{\text{crítico}}$ que definimos, no podemos rechazar H_0.

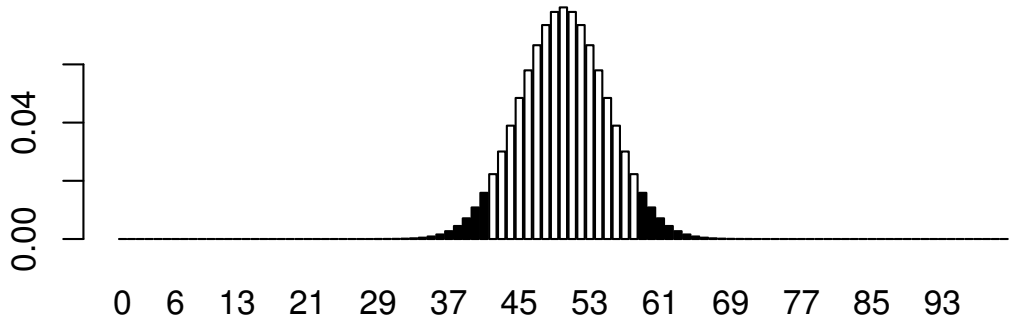
Cuando tenemos conocimiento previo sobre un fenómeno, podemos formular una hipótesis direccional. Esto nos habilita a que el resultado necesario para una conclusión significativa sea menos extremo que en el caso de una hipótesis no direccional. En la mayoría de los casos, el valor p que obtenemos para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

7. El informe

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 informe de una investigación cuantitativa consiste, por lo general, de cuatro

Figura 4: En negro, probabilidad de obtener 41 respuestas o menos y 59 respuestas o más que clasifican *camino* como un nombre.



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.

8. Conclusión

En esta reunión, discutimos cómo llevar a cabo una investigación cuantitativa. Para ello, analizamos las distintas etapas que involucra llevar a cabo un estudio siguiendo esta esta metodología, desde el planeamiento hasta la redacción del informe. Es crucial para obtener resultados rigurosos atravesar cada una de las etapas de forma responsable y detallada.

Bibliografía consultada

Arunachalam, S. (2013). Experimental methods for linguists. *Language and Linguistics Compass*, *7*(4), 221–232.

Casella, G., y Berger, R. L. (2002). Probability theory. En *Statistical inference* (pp. 1–45). Pacific Grove: Duxbury.

Chen, P. (1986). Discourse and particle movement in English. *Studies in Language*, *10*(1), 79–95.

Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003.

Fraser, B. (1966). Some remarks on the VPC in English. En F. P. Dinneen (Ed.), *Problems in semantics, history of linguistics, linguistics and English* (pp. 45–61). Washington DC: Georgetown University Press.

Gries, S. T. (2003). *Multifactorial analysis in corpus linguistics: A study of particle placement*. London, New York: Continuum.

Gries, S. T. (2013a). Basic significance testing. En R. J. Podesva y D. Sharma (Eds.), *Research methods in linguistics* (pp. 316–336). Cambridge: Cambridge University Press.

Gries, S. T. (2013b). Some fundamentals of empirical research. En *Statistics for linguistics with R* (pp. 1–55). Berlin: De Gruyter Mouton.

Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology*, *6*, 1–11.

Salkind, N. J., y Rasmussen, K. (2007). *Encyclopedia of measurement and statistics*. Thousand Oaks: SAGE Publications.

Van Dongen, W. A. S. (1919). He puts on his hat & he puts his hat on. *Neophilologus*, *4*, 322-353.