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

Índice

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 I

Estadística para el estudio del lenguaje: una introducción

Federico Alvarez

A lo largo de su historia, la tradición lingüística ha mantenido un enfoque preponderantemente cualitativo en el abordaje de su objeto de estudio. Por distintas razones (algunas de las cuales pueden rastrearse en la bibliografía, otras pueden tener más que ver con la apreciación personal de cada investigadore), las teorías sobre el lenguaje no han aplicado de manera extensiva herramientas cuantitativas para articular sus hipótesis, cosa que sí ha ocurrido en la mayoría de las disciplinas científicas. Esta tendencia se ha ido modificando durante las últimas décadas, y al día de hoy podemos encontrar una gran cantidad de estudios que hacen uso de modelos estadísticos para formular y verificar sus predicciones. Sin embargo, este viraje en la metodología no ha tenido aún un impacto fuerte en la formación de les lingüistas, y tampoco hay demasiado material disponible para iniciarse en el estudio de la estadística que haga foco en su aplicación en las ciencias del lenguaje. Uno de los problemas que se derivan de esta situación es que aquelles lingüistas interesades en adoptar una metodología con una base formal[1] más sólida no necesariamente tienen un andamiaje institucional que facilite el abordaje de la temática. Este trabajo, entonces, se ofrece como un repaso sobre las bases y los requisitos de un enfoque cuantitativo y las áreas donde este tipo de enfoque puede ser de utilidad en el campo disciplinar de la lingüística, con el propósito de servir como guía en el aprendizaje teórico de la estadística.

1. ¿Qué es la estadística?

Es fácil dar por sentado que sabemos qué es la estadística, al punto que varios textos introductorios al tema no proveen ninguna definición del término. Más aún, al contrastar las definiciones que brindan distintes autores nos encontramos con tantas discrepancias como coincidencias.

Freedman et al. (1998) ofrecen la siguiente definición (trad. mía):

> La estadística es el arte de proponer conjeturas numéricas sobre preguntas desconcertantes.

Esta definición, aunque escueta, nos indica que vamos a trabajar con números con el objetivo de responder preguntas. Una cosa importante que falta es una idea sobre cómo se vinculan las conjeturas con las preguntas. La siguiente definición, de Levshina (2015), dice (trad. mía):

> [...] la estadística es un conjunto de herramientas para describir y analizar datos.

Nuevamente una definición bastante escueta. Aquí nos presenta un concepto central para la disciplina que es el de **dato**. Al hacer estadística necesariamente estamos trabajando en función de datos, ya sea para

[1] El término "formal" no refiere a una oposición entre forma y significado sino al empleo de abstracciones con el fin de explicitar la relación entre los conceptos teóricos y datos empíricos de manera tal que cada observación se pueda interpretar de manera unívoca.

describirlos o para analizarlos. La próxima definición, de Downey (2011), aúna un poco algunos elementos de las dos anteriores (trad. mía):

> La estadística es la disciplina dedicada a usar muestras de datos para respaldar afirmaciones sobre poblaciones. La mayoría del análisis estadístico está basado en la probabilidad, debido a lo cual estos temas suelen presentarse conjuntamente.

Aquí podemos ver un vínculo explícito entre la noción de dato y el objetivo de responder preguntas. Al mismo tiempo, se introducen varios conceptos importantes que también tendremos que definir para orientar nuestro aprendizaje: **muestra**, **población** y **probabilidad**. Finalmente, la siguiente definición, de la página en inglés de Wikipedia (2018), es la más abarcativa (trad. mía):

> La estadística es una rama de la matemática que abarca la recolección, el análisis, la interpretación, la presentación y la organización de datos. Al emplearla en, por ejemplo, un problema científico, industrial o social, es convencional comenzar con una población estadística o un modelo estadístico a estudiar. Las poblaciones pueden ser tan diversas como "todas las personas que viven en un país" y "cada átomo que compone un cristal". La estadística abarca todos los aspectos de los datos, incluyendo el planeamiento de la recolección de datos en términos de diseño de encuestas y experimentos.

Esta definición apunta a una serie de cuestiones fundamentales que es necesario plantearse para abordar una investigación de manera cuantitativa. Hace mucho énfasis en el concepto de dato, alude al concepto de población e introduce el concepto de **modelo**.

Provisionalmente, vamos a definir los conceptos que aparecieron en estas definiciones para construir una definición operativa de estadística que los tome en cuenta.

- **Dato**: Unidad de observación sobre un fenómeno empírico determinado.
- **Muestra**: Conjunto de observaciones.
- **Población**: Grupo que representa los objetos bajo estudio. Dependiendo del estudio, la población puede ser cosas muy diferentes como "los hablantes de español rioplatense", "los conectores adversativos" o "las oraciones que usan verbos intransitivos". La población es aquello sobre lo cual nos estamos haciendo una pregunta que queremos responder a través de la recolección y el análisis de observaciones.
- **Modelo**: Constructo diseñado para dar cuenta de algún aspecto de la realidad de manera esquemática. Es un conjunto de supuestos explícitos sobre aspectos puntuales de fenómenos de interés.
- **Probabilidad**: Estudio de **eventos aleatorios**. Es una forma de representar numéricamente nuestras intuiciones y nuestros supuestos sobre las chances de que ocurran ciertos eventos y no otros. Normalmente pensamos informalmente en la verosimilitud de ciertos sucesos en relación a otros, más o menos verosímiles. La teoría de la probabilidad nos permite hacer afirmaciones cuantitativas sobre estas ideas.

Con este conjunto de definiciones podemos pasar a proponer una definición operativa de la estadística. Para nuestros propósitos, vamos a considerar que:

> La estadística es la disciplina que abarca todos los aspectos del estudio de observaciones empíricas en función de nuestras creencias sobre fenómenos particulares. Supone la

ISSN: 2796-8952

> recolección de datos de una población de interés, su descripción y análisis en función de modelos probabilísticos, y su subsecuente interpretación.

Esta definición implica que desde el momento en el que decidamos usar observaciones empíricas para responder una pregunta sobre un fenómeno de interés vamos a estar dentro del marco de la estadística, y los procedimientos que elijamos para obtener observaciones, las decisiones sobre cómo vamos a observar el fenómeno, la construcción de modelos que lo expliquen, la formulación de preguntas sobre dichos modelos y la interpretación de los datos con el fin de obtener una respuesta son todos aspectos importantes en la práctica. Esto no quiere decir que no haya lugar para las consideraciones cualitativas sino todo lo contrario: el análisis cualitativo precede al diseño de los estudios, ya que determina la selección misma del objeto de estudio, así como de las variables de interés para realizar un análisis, y también determina la interpretación final de los resultados obtenidos.

2. ¿Cómo hacemos estadística?

De manera muy general, un estudio estadístico supone la caracterización de un aspecto o conjunto de aspectos (**variables**) de una población a partir del análisis de la distribución de los mismos en una muestra. Los distintos objetivos de cada estudio resultan en dos modos complementarios de analizar los datos, **exploratorio** y **confirmatorio**.

Une puede explorar las observaciones a su disposición con el objetivo general de encontrar patrones, pautas que se repitan y den lugar a preguntas sobre el fenómeno observado. Según Behrens (1997), el análisis exploratorio supone el trabajo básico de familiarización con los datos que permite responder la pregunta general "¿qué está pasando acá?" e implica un énfasis en la representación gráfica de los datos, un foco en un proceso iterativo de generación de hipótesis y construcción de modelos tentativos, y una postura flexible respecto a los métodos a adoptar. Para el análisis confirmatorio, en cambio, es necesario especificar a priori las hipótesis y los métodos a emplear. No se trata de una instancia de generación de preguntas sino de respuestas.

Tanto la exploración como la confirmación recurren al mismo conjunto de herramientas; si bien pueden poner énfasis en elementos diferentes, la diferencia crucial entre ambos modos de análisis no pasa por las herramientas que usan, sino que radica en que en el análisis confirmatorio los datos se miran una única vez. En cuanto volvemos a consultarlos, ya estamos realizando una exploración y no podemos interpretar los patrones que encontremos como la confirmación de una hipótesis. Esto implica, necesariamente, que los datos que se emplean para proponer una hipótesis y los que se usan para confirmar jamás pueden ser los mismos.

La exploración y la confirmación se suceden iterativamente, donde los datos que se recolectaron para confirmar una hipótesis son luego explorados para proponer hipótesis diferentes, que luego se ponen a prueba con nuevas observaciones. Es importante considerar que tanto la exploración como la confirmación son importantes, que la exploración va primero y que cualquier estudio puede, y usualmente *debe,* incluir ambas.

Reflexionemos sobre dos ejemplos:

- Queremos estudiar las diferencias de realización del fonema /r/ en hablantes de Argentina. Nuestro conocimiento previo nos sugiere por lo menos dos: una vibrante múltiple dental sonora y una fricativa palatal sonora.

ISSN: 2796-8952

- Tras una búsqueda bibliográfica nos quedamos con estas dos opciones y llegamos a la suposición de que el uso de una u otra tiene relación con la región de origen de le hablante, donde distinguimos entre Cuyo, el Litoral, el Norte, el Centro, la Pampa, la Patagonia y el Río de la Plata.
- También suponemos que hay una influencia del registro, que distinguimos entre formal y coloquial (¿son las únicas posibilidades? ¿cómo determinamos qué registro corresponde a cada caso?).
- Asimismo, consideramos probable que haya una variabilidad importante entre individuos.
- Suponemos que la realización fricativa ocurre con una frecuencia superior en personas del Norte y que es más común en el registro coloquial que en el formal.
- En función de lo anterior, decidimos realizar grabaciones de personas de las seis regiones en un ambiente laboral de oficina, que identificamos como situación formal, y en un ambiente doméstico, que caracterizamos como informal. De las grabaciones extraemos, para cada hablante, la cantidad de veces que produjo cada realización. Luego, realizamos un test para verificar nuestra hipótesis. Al analizar nuestros resultados, nos encontramos con que efectivamente hay una diferencia en la cantidad de realizaciones consistente con nuestra hipótesis anterior, pero nos encontramos con que hay mucha variación entre sujetos que no se explica a través de las variables que consideramos. A partir de eso, agregamos a nuestras observaciones información sobre la edad y el género de cada hablante, y llevamos a cabo un nuevo test para verificar si explican parte de la variación.

- Queremos estudiar las oraciones en las que el objeto directo es un nominal coordinado, del tipo de "los vi a vos y a tu hermano juntos el viernes".
  - Tras hacer un relevamiento bibliográfico llegamos a la propuesta teórica de que, en una tarea de juicios de aceptabilidad, la concordancia del clítico acusativo con el objeto directo –que puede concordar en singular con el primer coordinado o con el segundo, o en plural con ambos– y el modo en el que se interprete el evento descrito por el verbo –un único evento que afecta a ambos OD o dos eventos disjuntos– tienen un efecto en la aceptabilidad percibida. Medimos este efecto en una escala del 0 al 10, donde la concordancia con el primer coordinado tendrá valores altos de aceptabilidad, la concordancia con el último coordinado tendrá bajos valores de aceptabilidad (¿qué serían valores altos y bajos? ¿por qué?) y la concordancia con ambos coordinados tendrá baja aceptabilidad para la lectura de eventos disjuntos, pero alta para un mismo evento.
  - Consideramos que hay variabilidad propia de cada hablante y que el nivel de dominio de otras lenguas y el lugar de origen también son una fuente de variación.
  - Para verificar la hipótesis, diseñamos una encuesta para administrar en línea, con múltiples oraciones que presenten las seis combinaciones posibles de concordancia del clítico y la lectura del evento. En el experimento consultamos a les participantes por su dominio de otras lenguas y su lugar de origen, además de otras preguntas respecto de edad, profesión y carrera. Para testear la relación entre variables, eliminamos de la muestra a aquellas personas que reportaran un dominio avanzado de cualquier lengua además del castellano y a aquellas personas que vinieran de cualquier lugar que no fuera Capital Federal o provincia de Buenos Aires. Tras testear la relación entre las respuestas de los juicios y los valores de lectura y concordancia, encontramos que las diferencias eran coherentes con nuestras hipótesis iniciales, excepto por el caso de concordancia total con lectura de eventos disjuntos, que predijimos con baja aceptabilidad pero mostró valores altos en la muestra. Aparte, encontramos que la lectura por sí sola no explicaba una cantidad significativa de variación, sino sólo en interacción con la concordancia. Cuando exploramos los valores de las demás

ISSN: 2796-8952

preguntas encontramos que cerca de la mitad de les participantes eran estudiantes o graduades de lingüística, por lo que realizamos un test para ver si había una relación significativa entre la profesión y las respuestas de los juicios que no arrojó diferencias significativas entre lingüistas y no lingüistas.

En ambos ejemplos puede verse una estructura común: primero se define un fenómeno de interés que se releva y a partir del cual se identifican variables de interés. Utilizando esas variables, se proponen hipótesis y se diseña un protocolo para recolección de datos que luego se lleva adelante. Esos datos se testean para comprobar las hipótesis iniciales y luego se proponen nuevas hipótesis en función de los hallazgos. En el segundo caso, también se condujo un nuevo test para evaluar la pertinencia de una nueva hipótesis. Considerando esto, puede decirse que entre exploración y confirmación, el análisis cuantitativo supone los siguientes pasos:

1. Se identifica un fenómeno de interés.
2. A partir de datos ya disponibles e investigación previa, se hace un relevamiento sobre el fenómeno. A partir de este relevamiento se identifican las variables que intervienen en el mismo. Es importante la exhaustividad en la identificación de las variables, dado que vamos a condensar el objeto de estudio en el modo en el que estas variables se distribuyen en las observaciones.
3. Se operacionalizan las variables. Esto quiere decir que se definen explícitamente el tipo de variable y el procedimiento para medirla.
4. Se proponen hipótesis sobre el comportamiento de las variables. Es necesario formularlas de manera tal que pueda expresarse una relación cuantitativa entre ellas.
5. Se define el procedimiento de recolección de observaciones, esto es, se define si se toman muestras en un ambiente natural, si se realiza una encuesta o si se conduce un experimento controlado y el protocolo con el que se hace la toma.
6. Si el propósito es conducir un análisis confirmatorio, se determinan también los modelos estadísticos a emplear para testear las hipótesis.
7. Se realiza la toma de datos.
8. Se llevan a cabo los análisis confirmatorios.
9. Los datos obtenidos se exploran para generar nuevas hipótesis.

2.1. Hipótesis

Una hipótesis estadística supone una relación entre dos conjuntos de variables, donde las variables del primer conjunto permiten explicar la variación en las variables del segundo conjunto. Las del primer conjunto se llaman **variables independientes o explicativas**, mientras que las del segundo conjunto son **variables dependientes o de respuesta**. Un estudio empírico impone restricciones sobre la formulación teórica (Gries, 2013). Para ajustarse a ellas, una hipótesis necesita cumplir con los siguientes criterios:

- Debe ser una **afirmación general** que abarque más que un **evento singular**.
- Debe tener la estructura de una **oración condicional** o ser parafraseable como una.
- Debe ser **falsable**, o sea, es necesario que exista la posibilidad de concebir eventos o situaciones que contradigan la afirmación. Esto a su vez requiere que **exista la posibilidad de comprobarla**. Sin embargo, estas dos características no son idénticas: hay afirmaciones que son falsables pero no testeables.

Asimismo, siempre hay que tener en cuenta que el modo en el que las variables de interés se definan y operacionalicen determina en gran medida la formulación de las hipótesis.

ISSN: 2796-8952

Al definir hipótesis debe definirse también la condición –o conjunto de condiciones– que falsaría cada una. Hay distintos marcos teóricos para la comprobación estadística de una hipótesis, pero la práctica más convencional, NHST (*null hypothesis significance testing*), estipula que para cada hipótesis propia –**hipótesis alternativa H_1**– es necesario formular una hipótesis opuesta –**hipótesis nula H_0**– tal que si H_1 predice la existencia de un fenómeno determinado, H_0 consiste en la inexistencia de ese fenómeno. Aun si no se adhiere al marco teórico de NHST, es importante tener en cuenta que las hipótesis sólo son relevantes en tanto compiten unas con otras. No hay valor en contrastar empíricamente dos hipótesis que realizan las mismas predicciones (ni es realmente posible, estrictamente hablando).

2.2. Diseño del estudio, probabilidades y espacio muestral

Los aspectos formales y materiales de la recolección de datos son también extremadamente importantes. Por ejemplo, para el caso de la realización de la /r/, múltiples maneras de extraer información son posibles: podríamos tomar una única muestra por sujeto, o seleccionar una cantidad fija de participantes y grabarles en condiciones de uso de registro formal e informal. En el caso de las oraciones con OD coordinado, ¿todes les participantes ven los mismos estímulos o hay diferentes condiciones de presentación? De haber diferentes condiciones, ¿todas contienen la misma cantidad de estímulos? ¿Son todos del mismo tipo? También debemos tener en cuenta constantemente cuestiones como la cantidad de variables independientes y dependientes en consideración, y si hay alguna interacción observable entre ellas. Distintas respuestas a estas preguntas van a suponer el uso de diferentes pruebas. Además, distintos tipos de estudios suponen una diferencia no sólo en la metodología de prueba de hipótesis, sino también en el modo en el que se interpretan las pruebas. Únicamente en el caso de experimentos controlados pueden inferirse relaciones causales entre variables; en todos los demás, pueden establecerse asociaciones significativas, pero no inferirse causalidad.

Estas consideraciones sobre los datos no son sólo relevantes a la hora de pensar en cómo poner a prueba una hipótesis; son constitutivas respecto de cómo se concibe el objeto de estudio. Sin abogar por ninguna metodología en particular, es importante analizar los métodos de recolección de datos sobre los cuales descansan los supuestos previos sobre el lenguaje empleados para construir las teorías con las que se trabaja (Abney, 2011). Aun si se hallara que los procesos son defectuosos, eso no implica que las hipótesis sean falsas ni que las teorías sean insuficientes. Sin embargo, sí significa que hay valor en intentar reproducir las predicciones de una teoría empleando una metodología más rigurosa que la que se utilizó para construirla. Además, no hay que perder de vista que el protocolo de recolección de datos y su ejecución son los eslabones más débiles en la realización de un estudio. Si los datos están mal recolectados, no hay test de hipótesis que valga, y los errores de toma y medición son extremadamente frecuentes y fáciles de cometer.

El conjunto de todos los posibles resultados de un estudio constituye lo que se llama **espacio muestral** (Casella y Berger, 2001). Las hipótesis, en términos cuantitativos, representan la asignación de un valor numérico entre 0 y 1 para cada uno de los elementos del espacio muestral, de forma tal que la suma (o integración) del valor asignado a todos los eventos sume 1. El número asignado a cada elemento del espacio muestral es la probabilidad de ese elemento. Diferentes marcos teóricos estadísticos van a hacer un trabajo distinto sobre el espacio muestral. El marco de NHST supone evaluar la probabilidad de las observaciones de las que disponemos según la hipótesis nula H_0 y, si tienen una probabilidad muy baja, va a sugerir que aceptemos la hipótesis alternativa H_1. Un marco teórico distinto, como el bayesianismo, supone calcular la probabilidad que cada hipótesis le asigna a cada observación en el espacio muestral, y permite así cuantificar y comparar la credibilidad de varias hipótesis en competencia.

ISSN: 2796-8952

2.3. Estadística descriptiva y exploración visual

Cuando se dispone de una muestra sobre la cual proponer o verificar hipótesis, es muy útil agregar la información presente en cada una de las observaciones de nuestra muestra en una métrica que dé cuenta de las propiedades de la muestra en su conjunto (Johnson, 2014). La **estadística descriptiva** refiere al conjunto de valores que dan cuenta de las propiedades de una muestra sin, por sí solas, responder preguntas acerca de las propiedades de la población de la que proviene. Principalmente, los valores medidos en una estadística descriptiva dan cuenta de la **tendencia central** de la muestra (un único valor que representa al conjunto de las observaciones), la variabilidad observada entre distintos valores de una misma variable (**dispersión**) y la **asociación** entre diferentes variables, o sea, cuánta información da una variable sobre otra. La elección de diferentes estadísticos descriptivos, ya sea de tendencia central, dispersión o asociación, depende del modo en el que las variables se hayan operacionalizado.

La representación visual de la muestra es también de gran importancia, ya que permite observar patrones que no necesariamente se evidencian en las medidas descriptivas e interpretar de manera rápida e intuitiva dichas medidas. Las visualizaciones a emplear están fuertemente asociadas a los tipos de variables de interés, así como a su cantidad.

2.4. Estadística inferencial

La **estadística inferencial** constituye el paso en el cual se proyecta la información obtenida en la muestra sobre la población de interés. La batería de pruebas estadísticas disponibles representa, en su forma más básica, la asignación de una probabilidad a las observaciones de una muestra tomando en cuenta una hipótesis determinada. En el marco NHST, como ya se dijo, se considera la probabilidad de las observaciones bajo el supuesto de que la hipótesis nula es correcta. De acuerdo con el resultado de la prueba que apliquemos, se acepta H_1 o se la rechaza en favor de H_0. La aceptación o rechazo de una hipótesis depende de la probabilidad de obtener las observaciones de la muestra si la hipótesis nula fuera cierta. Dos tipos de errores son posibles: el de rechazar la hipótesis nula siendo la misma verdadera en la población, conocido como **error de tipo 1 (falso positivo)**, y el de no rechazar la hipótesis nula siendo esta falsa en la población, conocido como **error de tipo 2 (falso negativo)** (Arunachalam, 2013).

Las diferentes pruebas a realizar dependen del cumplimiento de distintos supuestos sobre las variables, su distribución y su asociación. Las familias de modelos estadísticos se pueden dividir en dos grandes grupos: paramétricos y no paramétricos. Los **modelos paramétricos** son aquellos cuya estructura está determinada por un conjunto finito de valores, por lo cual su complejidad está delimitada ante un aumento en el tamaño de la muestra. Los **modelos no paramétricos**, en cambio, no poseen una estructura predefinida. Sin entrar en detalles, las pruebas paramétricas suelen ser más sencillas en términos de procesamiento, pero tienen una mayor cantidad de supuestos sobre la muestra, mientras que las pruebas no paramétricas suelen ser más simples en términos de lo que presuponen, pero son menos sensibles a tendencias en los datos.

Una cuestión crucial sobre la inferencia es que, en su forma más abstracta, supone la capacidad de "predecir el futuro". Al aceptar una hipótesis une se compromete con la idea de que, ante una eventual nueva muestra de la misma población, pueda predecirse la tendencia de las observaciones. Esto quiere decir que no basta sólo con la adecuación a los datos ya observados: entre dos teorías en competencia, una que explica perfectamente observaciones disponibles pero resulta insuficiente ante la aparición de nuevos datos es menos **robusta** que una teoría que, sin explicar la totalidad de la información disponible, realiza predicciones de manera consistente. La inferencia estadística, entonces, supone una formalización sobre la probabilidad de replicar observaciones.

ISSN: 2796-8952

3. ¿En qué áreas de la lingüística puede proponerse un abordaje estadístico?

Hay estudios sobre el lenguaje que parecen prestarse naturalmente a una óptica cuantitativa. Sin embargo, las áreas donde el uso de modelos estadísticos puede enriquecer nuestro conocimiento sobre el lenguaje son más de las que podríamos suponer si nos guiáramos por las teorías más difundidas sobre la gramática. En esta sección se pasa revista a algunas de estas áreas.

3.1. Psicolingüística y neurolingüística – Abordajes experimentales

Tanto en estudios de adquisición como de comprensión y producción de lenguaje, los experimentos realizados suelen estar sujetos a un análisis cuantitativo. De acuerdo al aspecto del lenguaje bajo estudio y a la metodología utilizada las pruebas requeridas van a ser diferentes, pero en general las variables independientes son estructuras de diferentes niveles lingüísticos y/o diferentes grupos dentro de la población y las variables dependientes son comportamentales (p. ej., tiempo de reacción, movimientos sacádicos) o fisiológicas (p. ej., ERP, flujo sanguíneo). En el caso de los estudios *offline*, pueden ser diferentes cosas como la cantidad de palabras diferentes empleadas para completar un espacio en blanco, la cantidad de errores cometidos en una tarea, etcétera.

3.2. Lingüística de corpus

La lingüística de corpus tiene la particularidad de que cuando el corpus es la población de interés y ya está disponible en su totalidad, anulando la necesidad de tomar muestras, todo el análisis cuantitativo posible va a ser una caracterización del mismo, sin necesidad de realizar inferencia alguna (siempre y cuando el alcance de un estudio se limite al corpus y no busque generalizar a otros textos). Las variables de análisis van a ser básicamente frecuencias, ya se trate de palabras, tipos, lexemas, locuciones, morfemas, estructuras sintácticas, o algún otro constructo teórico. Las medidas más complejas son asociaciones y proporciones entre estas frecuencias (Gries, 2009). Cabe aclarar que nada impide, a priori, el uso de un corpus como herramienta para realizar inferencias sobre una población diferente que lo abarque.

3.3. Lingüística computacional y procesamiento de lenguaje natural

Estos dos campos, que al utilizar herramientas similares suelen presentarse de manera asociada, pueden distinguirse por diferencias en su propósito: mientras que la lingüística computacional supone el uso de medios computacionales para la comprensión del lenguaje a diferentes niveles de análisis, el procesamiento de lenguaje natural busca optimizar las herramientas disponibles para resolver diferentes problemas utilizando un *input* lingüístico, aun si las soluciones propuestas no responden al procesamiento que hace un hablante (Cohen, 2016). Hay una tradición importante de modelos computacionales basados en teorías sobre todos los niveles de la gramática, que emplean sistemas determinísticos de reglas. La disponibilidad de grandes volúmenes de datos y el incremento en la capacidad de cómputo han favorecido el desarrollo de modelos basados en la probabilidad de hallazgo de patrones en corpus de entrenamiento (Klavans y Resnik, 1996; Manning y Schütze, 1999). Los abordajes computacionales estadísticos abarcan varios niveles de análisis, desde gramáticas probabilísticas que asignan una probabilidad a diferentes estructuras sintácticas y modelos de lenguaje basados en las probabilidades de transiciones entre secuencias de morfemas o palabras hasta modelos de las temáticas abordadas en un corpus (Blei, 2012) o modelos de procesamiento acústico de habla.

ISSN: 2796-8952

3.4. Tipología y lingüística histórica

Dunn et al. (2005) muestran el empleo de métodos basados en árboles filogenéticos binarios para la identificación de familias lingüísticas y del grado de semejanza que poseen las distintas lenguas entre sí, basándose en la codificación binaria –presencia o no presencia– de diferentes rasgos gramaticales. Varios métodos de inferencia son empleados para la selección entre los múltiples árboles posibles. Los métodos no se limitan al uso de rasgos gramaticales: es posible emplear variables extralingüísticas como la distancia geográfica, u otras variables relacionadas con el lenguaje, como la similitud entre palabras a través de cognados. De hecho, esta última opción es la prevalente en los métodos filogenéticos actuales.

3.5. Teoría lingüística

Al incorporar herramientas computacionales como parte de las opciones disponibles a le lingüista para desarrollar sus teorías, es posible generar hipótesis cuantitativas en todos los niveles de estudio del lenguaje, ya se trate del estudio de la sintaxis a través de juicios de aceptabilidad (Schütze y Sprouse, 2014), semántica (Goodman y Lassiter, 2015), o básicamente cualquier otro nivel (Bod et al., 2003). De hecho, incluso abordajes teóricos ya consolidados pueden incorporar modelado estadístico a su aparato teórico, como la gramática generativa para modelar adquisición y variación (Yang, 2004) o la lingüística cognitiva para modelar fenómenos semánticos, morfológicos y construccionales (Kuznetsova, 2013; Milin et al., 2016). Dado que la lógica de los análisis estadísticos se desprende del modo en que la teoría misma defina sus variables de interés, en tanto estas definiciones sean formalmente explícitas y tengan procedimientos de medición que permitan la recolección de muestras, todo constructo lingüístico es susceptible de un análisis estadístico.

4. Conclusión – ¿Por qué hay que usar métodos cuantitativos para estudiar el lenguaje?

Sin importar la filiación teórica o el nivel de análisis en el que une se especialice, es posible utilizar el conocimiento previo como asidero para introducirse en un paradigma más riguroso formal y empíricamente. Las razones para usar un abordaje cuantitativo tienen que ver, fundamentalmente, con la posibilidad de replicar hallazgos y de comparar teorías entre sí, no con una perspectiva teórica en particular. Uno de los puntos más importantes de esto es que la adopción de métodos cuantitativos permite la integración de diferentes puntos de vista de manera sistemática y la consideración de información de múltiples fuentes distintas bajo un mismo marco metodológico, independientemente de la filiación teórica de las personas que recolectan la información.

Asimismo, no hay que entender las opciones disponibles como una limitación que vuelve la perspectiva cuantitativa inabarcable, sino como la presencia de una gran cantidad de puntos de entrada al tema, de forma tal que es posible seleccionar el que sea más afín al bagaje previo de cada une. La ganancia en precisión teórica que se obtiene con el modelado estadístico del lenguaje representa un avance importante en las capacidades de descripción, explicación y predicción de la teoría lingüística.

Finalmente, es importante no perder de vista que el uso de una metodología cuantitativa no elimina la subjetividad en la investigación, pero sí facilita la integración de distintas subjetividades con un objetivo común.

Abney (2011) dice respecto de la lingüística computacional que (trad. mía):

ISSN: 2796-8952

> El lenguaje es un sistema computacional y hay una profundidad de entendimiento que es inalcanzable sin un conocimiento profundo de la computación. Pero incluso por sobre eso, el nuevo enfoque [la lingüística computacional] refleja un entendimiento más profundo del método científico y sitúa a la investigación lingüística firmemente dentro del paradigma de investigación intensiva de datos que ha dado en caracterizar a la ciencia moderna.

El abordaje estadístico está íntimamente ligado al empleo de herramientas computacionales, con lo cual es razonable considerar que la afirmación de arriba es apropiada también para referirse al abordaje cuantitativo. El giro conceptual asociado al uso de métodos cuantitativos tiene que ver con las posibilidades de refinamiento teórico que proporciona una metodología cuantitativa, en consonancia con el resto de la ciencia contemporánea.

5. Referencias

Abney, S. (2011). Data-intensive experimental linguistics. *Linguistic Issues in Language Technology*. https://doi.org/10.33011/lilt.v6i.1235

Arunachalam, S. (2013). Experimental methods for linguists. *Language and Linguistics Compass*, 7(4), 221–232. https://doi.org/10.1111/lnc3.12021

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods, 2*(2), 131–160. https://doi.org/10.1037/1082-989X.2.2.131

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84. https://doi.org/10.1145/2133806.2133826

Bod, R., Hay, J., & Jannedy, S. (2003). *Probabilistic linguistics*. MIT Press.

Casella, G., & Berger, R. L. (2001). *Statistical Inference* (2da ed.). Duxbury Press.

Cohen, S. (2016). *Bayesian Analysis in Natural Language Processing*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00719ED1V01Y201605HLT035

Downey, A. B. (2011). *Think Stats: Probability and Statistics for Programmers* (2da ed.). O'Reilly Media.

Dunn, M., Terrill, A., Reesink, G., Foley, R. A., & Levinson, S. C. (2005). Structural phylogenetics and the reconstruction of ancient language history. *Science*, *309*(5743), 2072–2075. https://doi.org/10.1126/science.1114615

Freedman, D., Pisani, R., & Purves, R. (1998). *Statistics*. W.W. Norton.

Goodman, N. D., & Lassiter, D. (2015). Probabilistic semantics and pragmatics: Uncertainty in language and thought. *The handbook of contemporary semantic theory* (2da ed.). Wiley-Blackwell.

Gries, S. T. (2009). *Quantitative Corpus Linguistics with R: A Practical Introduction*. Routledge.

Gries, S. T. (2013). *Statistics for Linguistics with R: A Practical Introduction* (2da ed.). De Gruyter Mouton.

Johnson, D. E. (2014). Descriptive statistics. En R. J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 288–315). Cambridge University Press.

Klavans, J., & Resnik, P. (1996). *The Balancing Act: Combining Symbolic and Statistical Approaches to Language* (1era ed.). The MIT Press.

Kuznetsova, J. (2013). *Linguistic profiles: Correlations between form and meaning*. [Doctoral dissertation, Universitetet i Tromsø].

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. John Benjamins Publishing Company.

Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Milin, P., Divjak, D., Dimitrijević, S., & Baayen, R. H. (2016). Towards cognitively plausible data science in language research. *Cognitive Linguistics*, 27(4), 507–526. https://doi.org/10.1515/cog-2016-0055

Schütze, C. T., & Sprouse, J. (2014). Judgment data. En R. J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 27–50). Cambridge University Press.

Wikipedia (2018). Statistics — wikipedia, the free encyclopedia. Recuperado el 26 de enero de 2018, de https://en.wikipedia.org/wiki/Statistics

Yang, C. D. (2004). Universal grammar, statistics or both? *Trends in cognitive sciences*, *8*(10), 451–456. https://psycnet.apa.org/doi/10.1016/j.tics.2004.08.006

ISSN: 2796-8952

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[2].

(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

[2] 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);

ISSN: 2796-8952

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

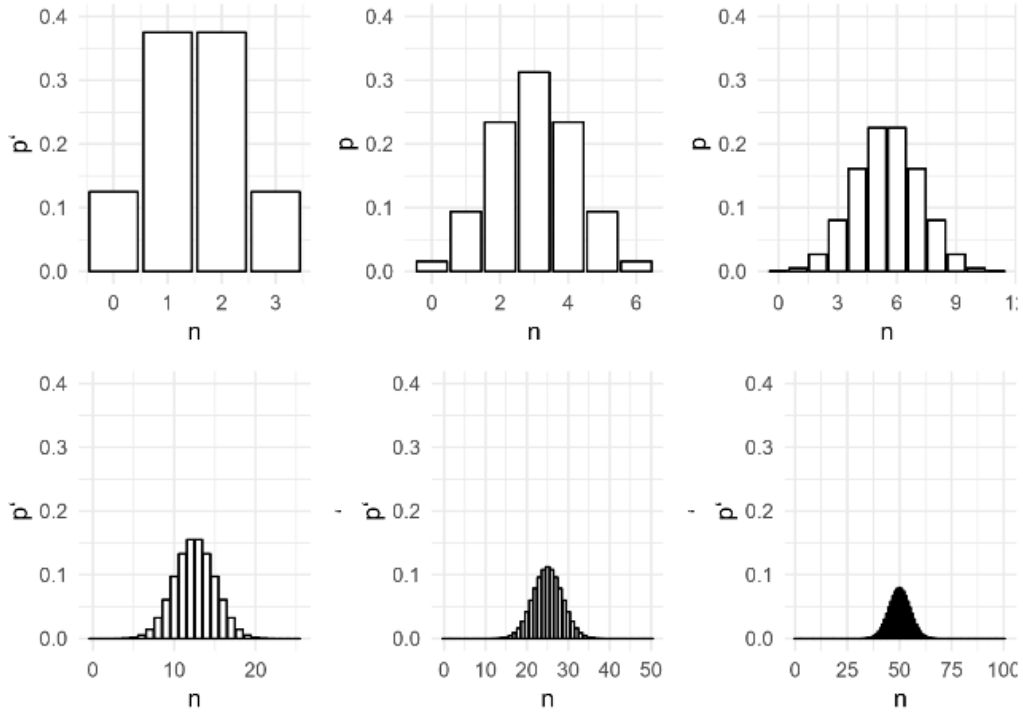

Figura 1. Distribución de probabilidad para resultados de 3, 6, 12, 25, 50 y 100 intentos binarios igualmente probables.

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).

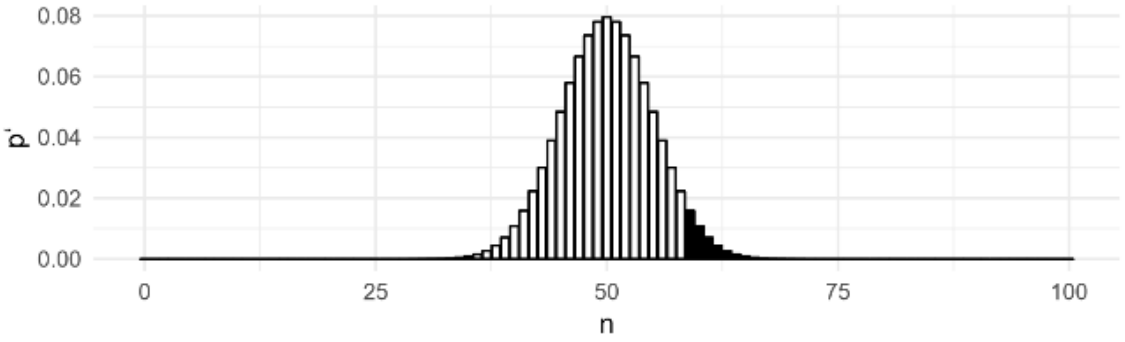

Figura 2. En negro, probabilidad de observar 59 lenguas o más con orden GenN en una muestra de 100 lenguas.

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.

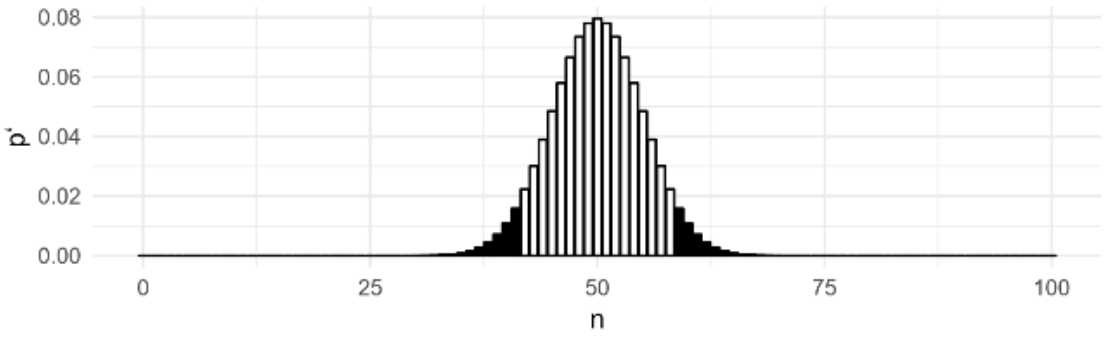

Figura 3. En negro, probabilidad de observar 41 lenguas o menos con orden GenN o 59 lenguas o más con dicho orden en una muestra de 100 lenguas.

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

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

Capítulo III

Estadística descriptiva

Laura Tallon y Charo Zacchigna

Antes de realizar cualquier análisis cuantitativo es imprescindible conocer la naturaleza del conjunto de datos y ciertas características básicas. Como su nombre lo indica, la estadística descriptiva es la rama de la estadística que se centra en organizarlos y describirlos, de manera de tener un pantallazo general de los datos a analizar. En el presente capítulo se trabajan nociones básicas por donde comenzar el estudio cuantitativo, tales como: los tipos de variables, la distribución, la importancia de visualizar los datos, diversas medidas de tendencia y dispersión, valores de normalización y correlación entre variables.

1. Introducción

La estadística descriptiva generalmente constituye el primer paso del análisis cuantitativo de un conjunto de datos. El objetivo es transformar la información a un número menor y más manejable de datos, dejando de lado los detalles y el ruido, para describir las propiedades básicas y generales de los datos. No se busca hacer inferencias sobre una población más grande que la muestra estudiada, sino que se utiliza como material exploratorio del que pueden surgir preguntas que conduzcan a una hipótesis sobre una población más grande. La estadística descriptiva, entonces, se circunscribe solo a los datos de la muestra estudiada.

1.1. Variables

El primer paso en el análisis es identificar los tipos de variables en estudio, ya que esto definirá los métodos y tipos de medidas estadísticas a utilizar.

-
    - **Variable de razón**: es una variable numérica que toma valores de una escala, cuyos intervalos pueden ser medidos (se puede decir que la diferencia entre 2 sílabas y 4 sílabas es igual a la diferencia entre 5 y 7 sílabas, y podemos afirmar que una palabra de 6 sílabas tiene el doble que una palabra de 3). Además la escala tiene un cero definido que marca la inexistencia (no existe una palabra de 0 sílabas) y por supuesto no cuenta con valores negativos. Otros ejemplos son los tiempos de reacción y la duración de emisiones, pausas, fijaciones, etc.
    - **Variable de intervalo**: a diferencia de la anterior, la escala de esta variable tiene un cero arbitrario, cuyo valor no significa la inexistencia de dicha variable, por lo tanto esta escala también puede tomar valores negativos. Un caso que ejemplifica este tipo de variables es el año de nacimiento, o los datos de una encefalografía en estudios de potenciales relacionados con eventos.
- **Variables ordinales**: son aquellas que tienen tres o más categorías y pueden ser listadas en un orden natural (primero, segundo, tercero o bajo, medio, alto). Con estas variables no se puede saber si los intervalos entre los valores son iguales. Es decir, la diferencia que hay entre bajo y

medio no necesariamente es la misma diferencia que hay entre medio y alto. Un ejemplo puede ser el curso escolar cuando estudiamos grupos de niños con diferentes niveles educativos, los valores de un juicio de aceptabilidad con escala de likert, o el nivel de conocimiento del idioma (alto/medio/bajo).

- **Variables nominales**: son aquellas que tienen tres o más categorías sin un orden natural. Al no estar cuantificadas no pueden realizarse operaciones matemáticas con ellas. Un ejemplo de variable nominal puede ser la clase de palabras y los puntos o modos de articulación.

- **Variables dicotómicas**: son un caso particular de variables nominales que solo tiene dos categorías opuestas. Pueden ser la presencia/ausencia de un rasgo y suelen cuantificarse como 0 y 1. Esta cuantificación puede usarse para aplicar algunos procedimientos de las variables continuas. Un ejemplo puede ser el rasgo sonoro/sordo de los fonemas consonánticos, la presencia/ausencia de pausas, verbos regulares/irregulares o finitos/no finitos.

1.2. Distribución

En segundo lugar, es importante observar la manera en que se distribuyen los datos recolectados. La distribución hace referencia a la probabilidad de cada valor de una variable. En el caso de las variables continuas, una distribución típica que se pueden tomar las variables es la llamada **distribución normal** (Figura 1), también conocida como "campana de Gauss" por su forma, que está definida por dos parámetros: la media y el desvío estándar.

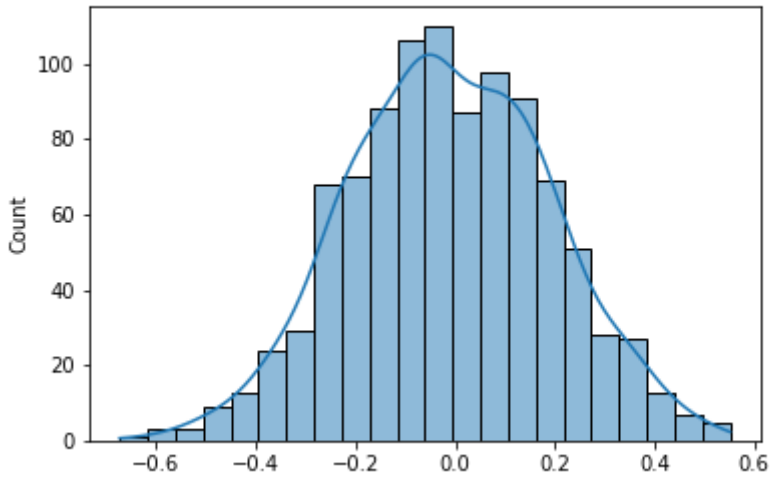

Figura 1. Se grafica un histograma con datos superpuestos con la curva de su distribución normal.

Por supuesto, los datos reales no tienen porqué estar distribuidos como se espera, pero ciertos métodos estadísticos usan esta distribución como base para el análisis. Un método que asume una distribución específica para la población estudiada, se denomina **método paramétrico**. Existen también métodos que no asumen una distribución específica, como los **métodos no-paramétricos** (los más apropiados para datos con variables ordinales o nominales).

Por esto es necesario conocer la distribución antes de aplicar cualquier tipo de método estadístico. Esto puede hacerse graficando la información en un eje cartesiano o representándola en una tabla de frecuencias.

2. Medidas de tendencia central

Estas medidas hacen referencia al valor más representativo de una variable. Hay distintas concepciones de qué es lo más representativo y esto se refleja en distintas formas de medir la tendencia central. En una distribución normal los valores de estas medidas se asemejan (excepto la media geométrica).

- **Media** ($\overline{X}$/μ): también conocida como promedio, se obtiene sumando todos los valores de la variable y dividiendo por la cantidad total de observaciones. Una variable solo tiene una media, que se ve muy afectada por la presencia de ***outliers*** (un valor cuya magnitud es muy diferente a la del resto de los valores de una muestra, ver Figura 3).[3]

$$\overline{X} = \frac{\Sigma X_i}{n}$$

- **Media geométrica**: es la raíz *N*-ésima del producto de todos los valores, siendo *N* la cantidad total de observaciones. Al igual que la anterior, una variable solo tiene una media geométrica, y se ve muy afectada por la presencia de *outliers*.[4]

$$\sqrt[n]{\Pi x_i}$$

- **Mediana**: si se ordenan todos los valores de una variable, desde el más pequeño hasta el más grande, la mediana será el valor que demarca la mitad de los datos. Al tomar por ejemplo la Figura 2, la mediana es 50,37; si empezamos a contar desde el valor más pequeño, al llegar a la mediana habremos contado la mitad de los casos. Si tenemos un número par de valores, la mediana se calcula sacando el promedio de los dos valores centrales. Se utiliza para variables ordinales, continuas y de razón. Una variable solo tiene una mediana. Es menos sensible a los *outliers* que la media.

- **Moda**: Es el valor más común en una distribución, el que más se repite. Una variable puede tener más de una moda. No se ve afectada por la presencia de *outliers*.

[3] **Σ** es una notación matemática que permite representar sumas de varios sumandos. Se conoce como suma o sumatoria. Por ejemplo, $\mathbf{\Sigma}X_i$ significa que se sumarán todos los valores de X; $\mathbf{\Sigma}(X_i/2)$ significa que a cada valor de X se lo dividirá por dos y luego se sumarán los resultados de cada división: $((X_i/2) + (X_i+2/2) + (....) + (X_n/2))$

[4] Π es una notación matemática que permite representar la multiplicación de una cantidad arbitraria. Se conoce como productoria o multiplicatoria. Por ejemplo, ΠX_i significa que se multiplicaran todos los valores de X; $\Pi(X_i/2)$ significa que a cada valor de X se lo dividirá por dos y luego se multiplicarán los resultados de cada división: $((X_i/2) . (X_{i+2}/2) . (....) . (X_n/2))$

ISSN: 2796-8952

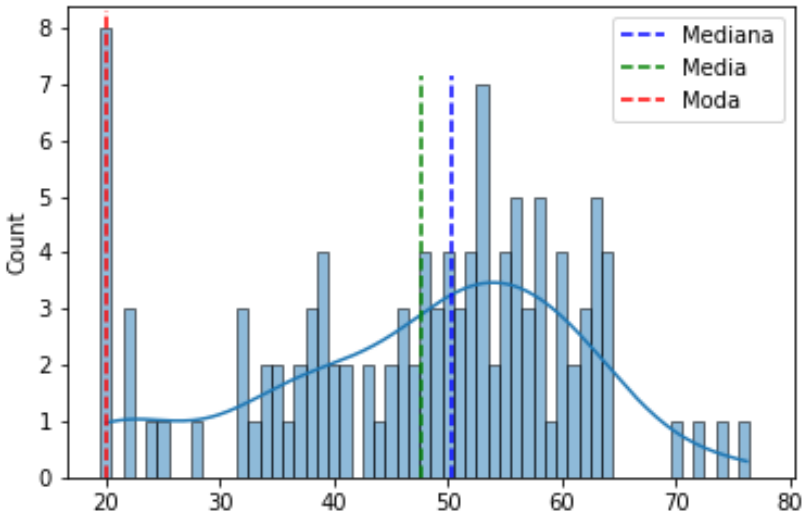

Figura 2. Moda, mediana y media. El gráfico muestra que las medidas de tendencia central no siempre coinciden. En este caso, la moda toma el valor 20, la mediana el valor 50,37 y la media o promedio el valor de 47,57.

Si el conjunto de datos presenta muchos *outliers* es asimétrico, es preferible reportar la mediana. También se utiliza esta medida para variables ordinales, mientras que para variables nominales solo puede utilizarse la moda.

ISSN: 2796-8952

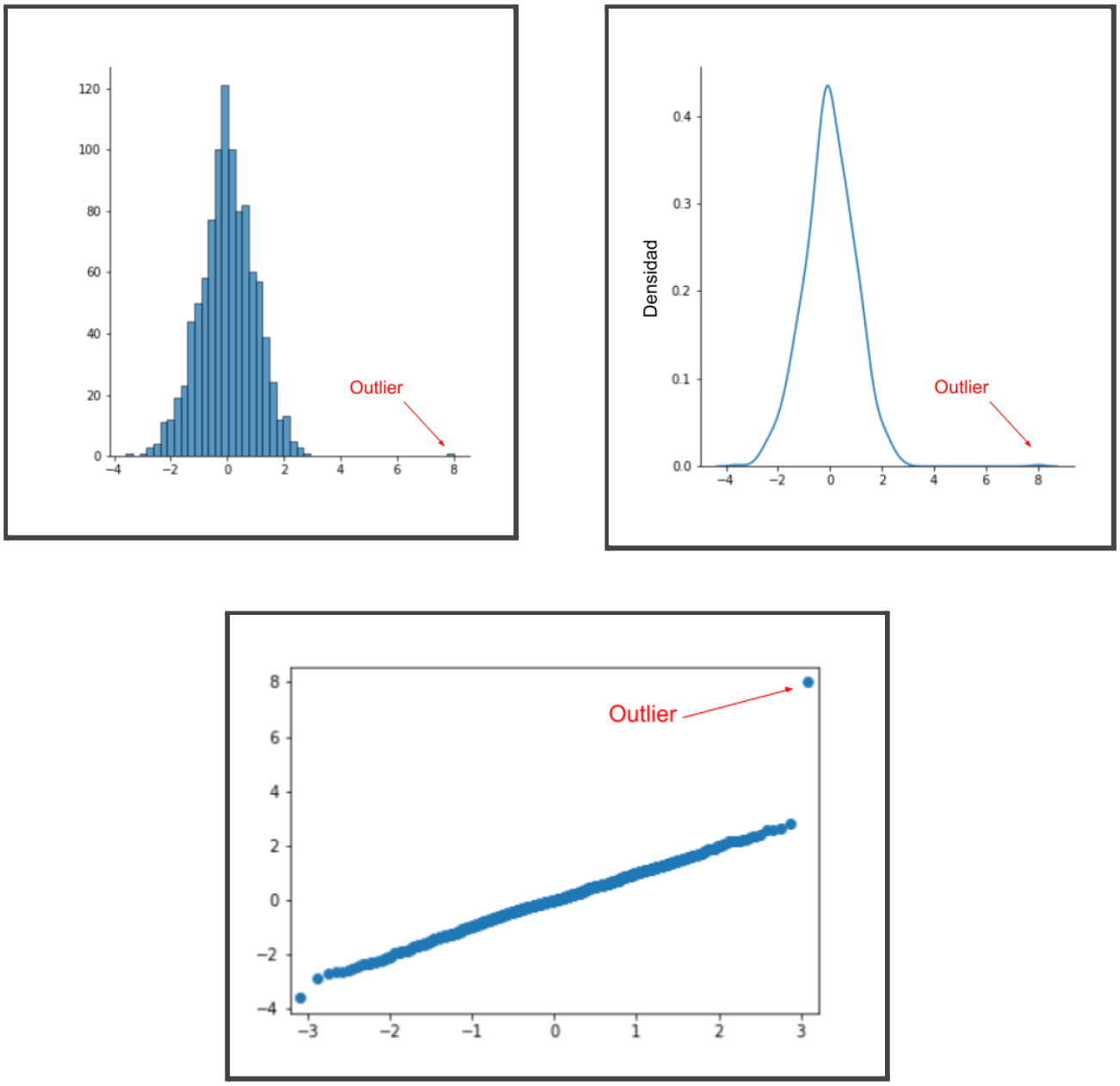

Figura 3. Outlier o "valor atípico": valor cuya magnitud es significativamente diferente a la del resto de los valores de una muestra. Usualmente se utilizan 2 o 3 desvíos estándar por debajo o por encima de la media para delimitar cuáles son los outliers.
La Figura muestra cómo se ven los outliers en un histograma, un gráfico de densidad y un gráfico cuantil-cuantil (qqplot) hechos ad hoc.

ISSN: 2796-8952

3. Medidas de dispersión

3.1. Medidas de dispersión dimensionales

Antes de comenzar con los tipos de medidas es importante conocer los conceptos de percentil y cuartil. El **percentil** es una medida de posición que indica, una vez ordenados los datos de menor a mayor en un grupo de observaciones, el valor por debajo del cual se encuentra un porcentaje dado de observaciones. Por ejemplo, si el 50% de los datos se encuentran por debajo de 34, el percentil 50 será 34. La forma de calcularlo es una regla de tres simple:

$$N = 100\%$$

$$I = P\%$$

Siendo N el total de los datos, P el percentil buscado e I el valor del percentil P_i. Si el valor de I es decimal, se redondea para arriba. A los percentiles 25, 50 (la mediana), 75 y 100 se los denominan **cuartiles**.

Al hacer estadística descriptiva resulta necesario medir cómo se comporta la variable estudiada. La medida de tendencia central no es suficiente por sí misma para describir su distribución (ver §7). Las medidas de dispersión representan qué tanto varían los valores con respecto a la tendencia central. Existen varias medidas, que toman las mismas unidades que la variable:

- **Rango**: es la diferencia entre el valor máximo y el valor mínimo de una variable. Es muy sensible a *outliers*.
- **Rango intercuartil** (**IQR**): si se calcula la diferencia entre el tercer (75) y el primer cuartil (25), se obtiene lo que se llama "rango intercuartil" (IQR por sus siglas en inglés). La diferencia entre los cuartiles muestra la heterogeneidad de los datos (mientras más difieran los cuartiles, más heterogénea es la muestra). Utilizando el IQR se disminuye la influencia de los *outliers*.
- **Desviación estándar** (**SD o σ**): es la medida más utilizada y se obtiene sacando la raíz cuadrada de la varianza. La **varianza** (**σ^2**) se calcula, primero, obteniendo para cada valor de la variable el cuadrado de su distancia con la media, luego realizando la sumatoria de estos cuadrados, y finalmente dividiendo todo por N. Entonces:

$$varianza = \frac{\sum_{i=1} (x_i - \overline{x})^2}{n}$$

$$\sigma = \sqrt{varianza}$$

  Cuando la muestra es grande hay que cambiar el divisor por n-1. A esto se le llama "corrección de Bessel" y se usa para no sobreestimar la muestra con respecto a la población. Dos distribuciones pueden tener medias similares, pero desvíos muy diferentes (ver §7).
- **Error estándar** (**SE** por sus siglas en inglés): se define como la desviación estándar de las medias de muestras igualmente grandes de una misma población. La media de una muestra no es lo mismo que la media de toda la población, a menos que la muestra sea perfectamente representativa. Una forma de saber qué tan representativa es la media de la muestra en relación con toda la población es computar la media de varias muestras semejantes (tomadas al azar, pero del mismo tamaño) y

ISSN: 2796-8952

calcular la desviación estándar de todos esos promedios. A este cálculo se lo conoce como error estándar:

$$SE = \sqrt{\frac{var}{n}} = \frac{\sigma}{\sqrt{n}}$$

Mientras más grande es el error, menos representativa es la estimación para el resto de la población. Y mientras más grande sea *N* (el tamaño de la muestra), más pequeño será el error estándar. Si al comparar las medias de dos muestras aproximadamente iguales los errores estándar se superponen, las medias no son significativamente diferentes. Ahora bien, que los errores no se superpongan no significa que las medias sean diferentes significativamente. La forma de calcular la diferencia entre medias es la siguiente:

$$SE_{(\text{diferencia entre medias})} = \sqrt{SE_{m1}^{\ 2} + SE_{m2}^{\ 2}}$$

Esta medida sólo es útil cuando se aplica a una distribución normal o cuando la muestra es igual o mayor a 30.

Al analizar una variable usualmente se elige la media y el desvío estándar, o bien la mediana y el IQR. Sin embargo, siempre es mejor graficar la muestra y luego utilizar todas o varias de las medidas de tendencia central y dispersión.

3.2. Medidas de dispersión adimensionales

Las medidas nombradas hasta el momento se expresan en la misma unidad de la variable, es decir, si la variable toma valores de F0 (Hz) o segundos, su desvío estándar también será un valor de F0 o segundos. También existen medidas adimensionales que son útiles para comparar datos con diferentes unidades de medida:

- **Coeficiente de dispersión cuartil**: medida no-paramétrica adimensional que se obtiene al dividir el IQR por la suma del primer y tercer cuartil:

$$[(\text{cuartil}_3 - \text{cuartil}_1) / (\text{cuartil}_1 + \text{cuartil}_3)]$$

- **Coeficiente de variación**: un problema con la desviación estándar es que su tamaño depende de la media. Cuando los valores y, por lo tanto, también la media se incrementan, también aumenta la desviación estándar. Es por ello que no se pueden comparar desviaciones estándar de distribuciones con diferentes medias si no se normalizan primero. En cambio, al dividir la desviación estándar por su media se consigue el coeficiente de variación, una medida paramétrica que no se ve afectada por el aumento de los valores de la distribución.

$$Cv = \frac{\sigma}{\bar{X}}$$

4. Otras medidas

- **Asimetría** (***skewness***): una distribución puede tener mayor cantidad de valores por encima o por debajo de la media, lo que en un gráfico se percibe como una cola considerablemente más larga que

la otra (Figura 4). Para conseguir esta medida se debe calcular el "coeficiente de asimetría de Fisher", a partir de la sumatoria de la diferencia de cada valor con la media elevada al cubo y esto dividirlo por el cubo de la desviación estándar. Si la distribución tiene muchos valores por encima de la media, cuando estos se potencien al cubo van a resultar en grandes números positivos (asimetría derecha). Si la distribución tiene muchos valores por debajo de la media, el resultado va a ser negativo (asimetría izquierda). Esta es una medida adimensional, no tiene unidades. Una distribución simétrica posee una asimetría con valor 0 (por ende, una distribución normal posee una asimetría con valor 0, pero es importante resaltar que una asimetría con valor 0 no implica que la distribución sea normal).

$$\text{Asimetría} = \frac{\sum_{i=1} (x_i - \bar{x})^3}{\sigma^3}$$

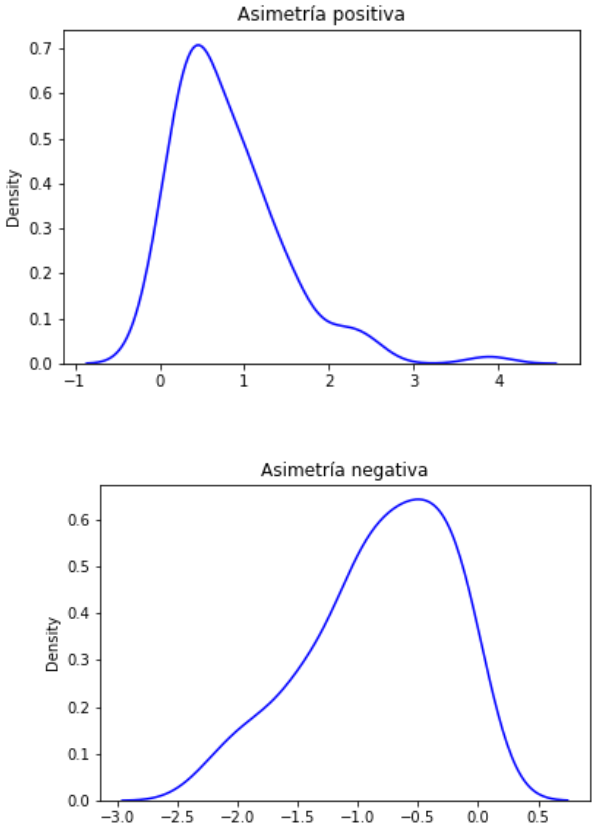

Figura 4. Representación de distribuciones asimétricas.

- **Curtosis**: se utiliza para saber si la distribución tiene un pico puntiagudo (leptocúrtica), redondeado (platicúrtica) o similar a una distribución normal (mesocúrtica) (Figura 5). La fórmula es igual que para la asimetría, pero se reemplaza el cubo por la cuarta potencia en numerador y denominador. Hay que restarle 3 para hacer una corrección, llamada "exceso de curtosis", y así se obtienen

valores positivos para una distribución leptocúrtica y negativos para una distribución platicúrtica. No tiene unidades. En una distribución mesocúrtica tendría un valor de 0.

$$\text{Curtosis} = \frac{\sum_{i=1} (x_i - \bar{x})^4}{\sigma^4} - 3$$

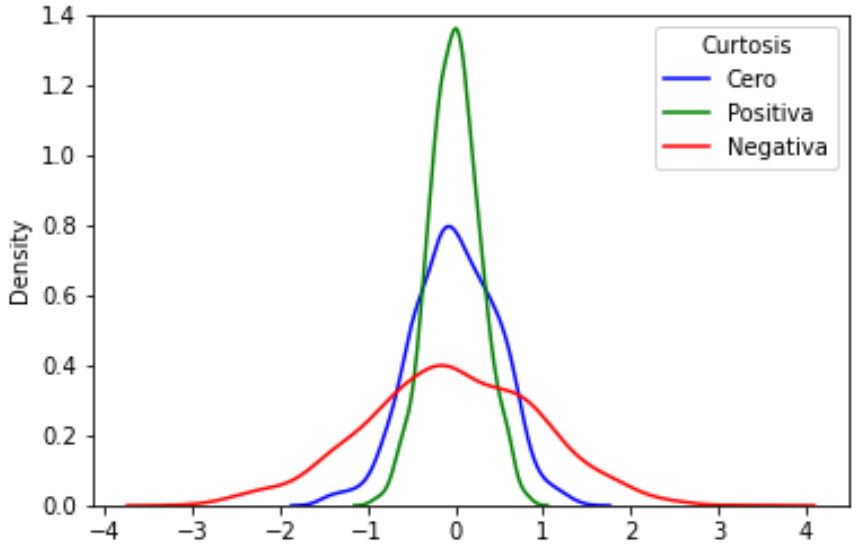

Figura 5. Distribuciones según sus valores de Curtosis. En el gráfico puede verse una curva leptocúrtica (la superior, en color verde), una curva mesocúrtica (la intermedia, en color azul) y una curva platicúrtica (la inferior, en color rojo).

- **Normalización**: no se pueden comparar valores o medidas que surgen de diferentes escalas, para esto es necesario normalizar o relativizar estos valores. Dos formas de hacerlo son:
  - **Centrado**: de cada valor individual se resta la media de la escala. Gráficamente, consistiría en centrar la media en 0.

$$C = X_i - \bar{x}$$

  - **Estandarización**: es la conversión de valores individuales en una forma estándar llamada "valores-*z*" (*z-scores*) o "valores-*t*" (*t-scores*).
- ***z-scores***: esto indica a cuántos desvíos estándar se encuentra un valor en relación a la media, y se utiliza cuando el tamaño de la muestra es igual o mayor a 30. El valor-*z* de *X* es la diferencia de *X* y la media, dividido por la desviación estándar de la población.

$$Z_{xi} = \frac{x_i - \bar{x}}{\sigma}$$

ISSN: 2796-8952

- ***t-scores***: esta forma estandarizada se utiliza cuando no se conoce la desviación estándar de la población y el tamaño de la muestra es menor a 30. En este caso, se calcula la diferencia entre el valor y la media, y se divide por el error estándar.

$$T_{xi} = \frac{x_i - \bar{x}}{\frac{\sigma}{\sqrt{n}}}$$

- **Intervalo de confianza**: permite saber qué tan bien se puede generalizar el resultado de una muestra a su población. Las siguientes fórmulas se basan en el error estándar, por lo que si los datos no se distribuyen normalmente o la muestra es demasiado pequeña, conviene usar otros métodos para computar el I.C.

  - **I.C. de la media**: provee un intervalo de valores aproximados, alrededor de los cuales se asume que no hay diferencia significativa entre este y la media poblacional. Para calcularlo debemos tener en cuenta el tamaño de la muestra:
    Si la muestra es mayor a 30, los límites del intervalo se calculan sumando y restando, respectivamente, a la media muestral el error estándar multiplicado cierta cantidad de veces. Esta última cantidad, representada en la fórmula como $z_{1\text{-}p/2}$, varía de acuerdo al porcentaje definido para el intervalo de confianza. Generalmente se utiliza un intervalo del 95%, en cuyo caso se multiplicará el error estándar por 1,96[5].

$$CI = (SE * Z_{\frac{1-p}{2}}) \pm \bar{X}$$

    El intervalo de confianza de 95% nos indica que si recolectáramos datos de infinitas muestras diferentes, el valor real de la media poblacional estaría dentro del intervalo de confianza en el 95% de esas repeticiones.
    Si se comparan los intervalos de confianza de dos muestras y estos no se superponen, las medias de las muestras son significativamente diferentes.

5. Estadística con dos variables: medidas de correlación

5.1. Medidas de correlación paramétricas

Cuando existe una dependencia entre dos variables se dice que están asociadas o que existe una correlación. Por ejemplo, en una correlación lineal, si *X* aumenta una cierta cantidad, *Y* aumentará o disminuirá una cierta cantidad proporcionalmente. Siempre que una variable nos ayude a predecir otra se puede hacer el camino inverso. Si *X* nos ayuda a predecir *Y, Y* nos ayudará a predecir *X*. Esto, sin embargo, no implica causalidad.

- **Correlación de Pearson** (***r***): antes de calcularla, tenemos que graficar las variables a fin de asegurarnos de que la relación entre estas sea lineal. Esta correlación se define como la covarianza

[5]El valor 1,96 es un valor constante para los intervalos de confianza del 95% que surge de partir del producto de restarle a 1 la probabilidad (*p*) que se quiere para el intervalo de confianza (en este caso 0,95 para un intervalo de 95%) y dividirlo por dos. La *z* refiere a *z-score*, debido a que se asume una distribución-z, es decir, una distribución normal cuya media es 0 y su desvío estándar 1.

de las dos variables, dividida por el producto de sus desviaciones estándar, y siempre toma un valor entre -1 y 1. En tanto que es un método paramétrico, la correlación de Pearson funciona mejor cuando las dos variables tienen una distribución normal. No tiene unidades y es sensible a los *outliers*.

- **Covarianza**: para cada dato se mide la diferencia entre su valor de *X* y la media de *X*; esto se multiplica por la diferencia entre su valor de *Y* y la media de *Y*. La covarianza es el promedio del producto de estas dos diferencias[6]:

$$\mathrm{Cov}_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$$

Entonces:

$$r = \frac{Cov_{xy}}{\sigma_x * \sigma_y}$$

El signo (+/-) de *r* define la dirección de la correlación, mientras que su valor absoluto define la fuerza de la correlación. Cuando *r* es igual a 0 no hay correlación entre las variables. La Tabla 1 muestra la interpretación de distintos valores de *r*.

Coeficiente de correlación *(r)*	**Etiqueta de correlación**	**Tipo de correlación**
0,7 < r ≤ 1	Muy alta	Un valor positivo de r representa una correlación positiva*: "cuando* X *aumenta…* Y *aumenta…"* o *"cuando* X *disminuye…* Y *disminuye…".*
0,5 < r ≤ 0,7	*Alta*	
0,2 < r ≤ 0,5	*Intermedia*	
0 < r ≤ 0,2	*Baja*	
r ≈ 0	*No hay correlación estadística (H_0)*	
0 >r ≥ -0,2	*Baja*	Un valor negativo manifiesta una correlación negativa*: "cuando* X *aumenta…* Y *disminuye…"* o *"cuando* X *disminuye…* Y *aumenta…".*
-0,2 >r ≥ -0,5	*Intermedia*	
-0,5 >r ≥ -0,7	*Alta*	
-0,7 >r ≥ -1	*Muy alta*	

Tabla 1. Interpretación de los valores de un coeficiente de correlación. Tomado de Gries (2013), traducción nuestra.

[6]En la fórmula el denominador es *n*-1 porque se hace una corrección, así que estrictamente no sería el promedio que se suele utilizar cotidianamente.

ISSN: 2796-8952

- **Coeficiente de determinación** (***r-squared***): sirve para resumir cómo se ajusta un modelo a los datos, es decir, qué tanto de la varianza de la variable dependiente es explicada por la variable independiente.
- **Regresión lineal**: otra forma de investigar la correlación es tratando de predecir valores de la variable dependiente con base en la independiente.[7]

5.2. Medidas de correlación no paramétricas

Cuando la relación entre variables no es lineal, es mejor utilizar alguna de las siguientes medidas no paramétricas (que también son más apropiadas para variables ordinales).

- **Rho (ρ) de Spearman**: mide la "monotonía" de una asociación. En una relación perfectamente monótona, si una variable aumenta la otra va a aumentar o disminuir consistentemente (pero no ambas). Si ambas se mueven en la misma dirección, rho = 1; si se mueven en direcciones opuestas, rho = -1. Se calcula utilizando el mismo método que *r* (covarianza dividida por el producto del desvío estándar), pero los datos primero son transformados en rangos. En este caso, el concepto de **rango** se refiere al orden de los valores de la escala (primero, segundo, etc). Los métodos no paramétricos usualmente involucran rangos, convirtiendo variables continuas a una escala ordinal. Esto hace más débil al método (se necesitan más observaciones), pero es más fuerte frente a *outliers*, asimetrías o distribuciones multimodales.
- **Tau (τ) de Kendall**: si unimos dos puntos cualesquiera de los datos con una línea recta, la *tau* de Kendall es la probabilidad de que esta línea tenga una pendiente positiva menos la probabilidad de que tenga una pendiente negativa. Este resultado va a ser entre -1 y 1; la *Tau* de Kendall tiende a ser menor que la *Rho* de Spearman, pero bastante similar.

6. Análisis de variables categóricas

Hasta ahora, muchas de las medidas descritas solo pueden aplicarse a variables continuas, pero no a variables categoriales, dado que precisan de valores numéricos.

- Variables nominales: no es posible calcular la media o la mediana de una variable nominal. Lo mismo sucede con el desvío estándar. La moda es el valor más frecuentemente utilizado. Para dar la dispersión de una variable nominal se puede usar el **índice de dispersión**, que es cercano a cero si la mayoría de los datos pertenece a una sola categoría y es igual a 1 si se distribuye en todas las categorías equitativamente. Si n es el total de observaciones, K el número de categorías y f describe la cantidad de datos en cada categoría, entonces el índice de dispersión es:

$$\frac{K*(n^2-\sum(f^2))}{n^2*(k-1)}$$

[7]La regresión lineal será abordada con mayor detalle en el siguiente volumen de estos cuadernos.

ISSN: 2796-8952

- Variables ordinales: se puede utilizar la mediana y algunas medidas relacionadas, como el IQR. Sin embargo, a menos que la variable tenga muchas categorías, esto no es muy útil.
- Variables binarias o dicotómicas: se puede representar las variables con 1 y 0, y calcular un tipo de media usando esos números (promedio de la cantidad de 1 o 0). A esto se le llama **media de una proporción** o "*p*". Para medir la dispersión se puede calcular el desvío estándar, pero el resultado no es independiente de la media, por esto el desvío estándar de una proporción no es muy útil.

Otra medida de correlación es el **Coeficiente *phi*** que determina que una variable independiente y una dependiente son intercambiables. Se utiliza para tablas de contingencia[8] si al menos una de las variables es nominal, o ambas son dicotómicas. El resultado es un número entre -1 y 1 cuya interpretación es similar a la de la *r* de Pearson.

$$phi = \sqrt{p * (1 - p)}$$

En relación a la correlación de dos variables, Kendall y Spearman son apropiadas para variables ordinales, o en el caso de una ordinal y una continua.

7. La importancia de graficar la distribución

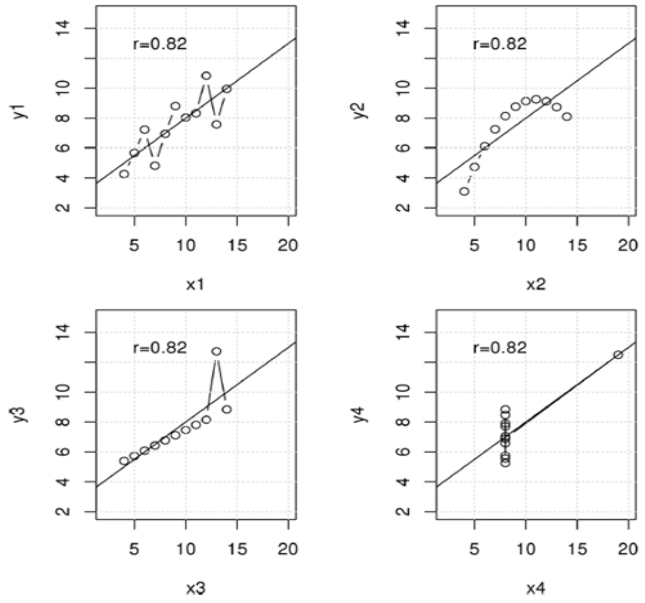

Figura 6. Diferencias de distribución, tomado de Gries (2013, p. 155).

Finalmente, la media, la varianza y la línea de regresión pueden ser las mismas para dos conjuntos de datos, pero esto no significa que su distribución sea la misma. En los gráficos superiores de la Figura 6

[8] Tabla de doble entrada donde se detallan las frecuencias, es decir, la cantidad de apariciones de una variable (en relación con otra). Se emplean para registrar y analizar la asociación entre dos o más variables. Por ejemplo:

	Respuestas correctas	Respuestas incorrectas	Total
Grupo A	35	15	50
Grupo B	22	28	50
Total	57	43	100

ISSN: 2796-8952

tenemos una situación donde *X* e *Y* están relacionados en una forma curvilínea, por lo que utilizar el método de correlación lineal no tiene mucho sentido. En los dos gráficos inferiores, los *outliers* tienen una gran influencia en el valor de *r* y la regresión lineal. A simple vista podemos notar que estas cuatro distribuciones son diferentes, a pesar de presentar el mismo valor de *r* (0,82). Estos cuatro gráficos ejemplifican lo indispensable que es inspeccionar visualmente los datos.

8. Referencias

Gries, S. T. (2013). Descriptive statistics. En *Statistics for linguistics with R* (2da ed., pp. 102-156). Walter de Gruyter GmbH.

9. Bibliografía sugerida

Gries, S. T. (2013). Descriptive statistics. En *Statistics for linguistics with R* (2da ed., pp. 102-156). Walter de Gruyter GmbH.

Johnson, D. E. (2013). Descriptive statistics. En R. J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 288–315). Cambridge University Press.

Levshina, N. (2015). “Chapter 3. Descriptive statistics for quantitative variables” en Levshina, N. *How to do Linguistics with R. Data exploration and statistical analysis* (pp. 41-68). John Benjamins Publishing Company.

ISSN: 2796-8952

Capítulo IV

Estadística inferencial

María Mercedes Güemes

La estadística inferencial permite generalizar desde una serie de observaciones representativas a un universo más grande de posibles observaciones, usando pruebas de hipótesis (como los *t*-test y ANOVA). En el presente capítulo, se abordan las nociones básicas de la estadística inferencial, como la obtención de la muestra, la postulación y el testeo de hipótesis, los errores y su relación con la significancia estadística. Posteriormente, se presentan tres teorías en las que se basan las pruebas estadísticas actuales, sus ventajas y sus limitaciones (Fisher y Neyman-Pearson, NHST). Por último, se exponen las interpretaciones incorrectas más habituales que se hacen sobre los test estadísticos, valores *p*, intervalos de confianza y poder según Greenland (2016).

1. Introducción

Realizar un análisis descriptivo es un paso crucial para conocer las características de un conjunto de datos. Sin embargo, los resultados se limitan a ese set particular de datos (muestra) y no sirven para deducir propiedades de una población mayor.

A diferencia de la estadística descriptiva, la estadística inferencial posibilita derivar las conclusiones obtenidas de una muestra a un conjunto de datos más amplio. Gracias a la estadística inferencial, es posible estudiar un fenómeno sin necesidad de tener datos sobre toda la población. A modo de ejemplo, si el objetivo de una investigación es determinar si los hablantes del español del Río de la Plata presentan mayor velocidad de habla que los hablantes del español peninsular, no es necesario (ni es posible) tener información sobre todos los hablantes de cada variedad lingüística. A partir de una pequeña parte de cada población (muestra) se pueden aplicar distintos métodos y procedimientos para generar inferencias sobre toda la población.

1.1. Muestras

El punto de partida de la estadística inferencial es la obtención de una *muestra*. La muestra debe ser lo suficientemente representativa para garantizar que todo el análisis estadístico sea correcto. Es importante que la recolección de datos, es decir, el diseño de muestreo, presente un protocolo de base científica (para más información sobre el tema, ver Capítulo II).

Según Johnson (2011) siempre se debería poder balancear las muestras con datos aleatorios de la población. No obstante, en lingüística, es muy habitual que se utilicen muestras no aleatorias por varios motivos. Por ejemplo, cuando se requiere estudiar un fenómeno que sucede en contextos específicos, como en el bilingüismo, o cuando se intenta controlar la variabilidad lingüística de los socio- y cronolectos. Si bien es frecuente que en lingüística experimental se trabaje con muestras no aleatorias, desde el punto de vista estadístico esto no es adecuado. La capacidad de generalizar los resultados a la población depende de que se obtenga una buena muestra. Se debe considerar que un buen método estadístico no compensa una mala muestra (Johnson, 2011).

Una vez que se obtuvo la muestra sobre la que se va a trabajar, el objetivo es la estimación de parámetros de la población y el testeo de hipótesis. Dado que no se cuenta con información sobre los valores reales de la población, lo que se suele hacer es estimar los parámetros a través de los estadísticos que presenta la muestra. Para ello, se van a distinguir dos tipos de medidas. Las medidas de la población se representan con letras griegas y las de la muestra con letras romanas (para más información sobre cómo se calculan estas medidas, ver Capítulo III).

Parámetros (población)

μ = media poblacional

σ = desvío estándar poblacional

σ^2 = varianza poblacional

Estadísticos (muestra)

$\bar{x}$ = media muestral

s = desvío estándar muestral

s^2 = varianza muestral

1.2. Testeo de hipótesis

Frente a las medias de dos muestras ($\bar{x}$ de la muestra x e $\bar{y}$ de la muestra y), ¿cómo se puede establecer si esas medias son diferentes entre sí? Lo más importante es, en realidad, saber si las diferencias estimadas para las medias $\bar{x}$ e $\bar{y}$ pueden representar las diferencias entre μ_1 y μ_2. De esta forma, se puede establecer el valor de confianza de $\bar{x}$ sobre μ_1.

Por ejemplo, se puede estudiar un fenómeno fonético vinculado a dos tipos de poblaciones: determinar si la emisión de la conjunción *pero* varía según el género del grupo que la emite (datos tomados de Ferrari et al., 2021[9]). La investigación toma como observación la duración de la conjunción *pero* en milisegundos en dos contextos: voces femeninas y voces masculinas. De esta forma, cuenta con dos muestras y sus medidas. La media de emisión de *pero* para las voces masculinas es de 200 ms (media $\bar{x}$ de la muestra x) con un desvío estándar de 8 ms, mientras que la media para las emisiones del grupo de voces femeninas es 172 ms (media $\bar{y}$ de la muestra y) con 32 ms de desvío estándar. A partir de estas medidas, es posible definir, luego de generar una serie de cálculos estadísticos, si esos dos tipos de emisiones (muestras) difieren en la duración de la emisión *pero,* o si esas diferencias numéricas en la duración de *pero* se deben a una variación posible de la toma de dos muestras al azar que no se relaciona con el tipo de voces que emiten la conjunción.

2. Teorema del Límite Central

Una de las bases teóricas que permiten generar inferencias sobre las medias de la muestra, sin tener información sobre la población total, es el **Teorema del Límite Central** (**TLC**). Dado que es difícil conocer la distribución de toda la población, se toma como dato la distribución de las muestras. Esto significa que si se toman muchas muestras sobre la población, se pueden tener muchas medias muestrales y con eso se puede

[9] Los datos que se toman en este capítulo corresponden a material no publicado que se obtuvo de la investigación llevada a cabo en ese trabajo.

ISSN: 2796-8952

elaborar una distribución muestral. En ese caso, si se toman muchas muestras (*sampling*), ¿cómo se vería la distribución de todas las medias de esas muestras (la distribución de las muestras o *sampling distribution*)?

El TLC explica que, sin importar la distribución de la población (sea normal, asimétrica u otras), es una propiedad de las medias tener una distribución cercana a la normal. Además, cuando el número de observaciones de cada muestra (N) aumenta, la distribución de las medias se acerca más a la normalidad (Figura 1). En esto se basa el Teorema del Límite Central y sirve para hacer inferencias sobre la media, aunque no se conozca la distribución de la población total. Así, se puede usar la distribución normal o cercana a la normal para generar conclusiones de probabilidad sobre la media (como se hace con los *z-scores*) aunque la distribución de la población no sea normal.

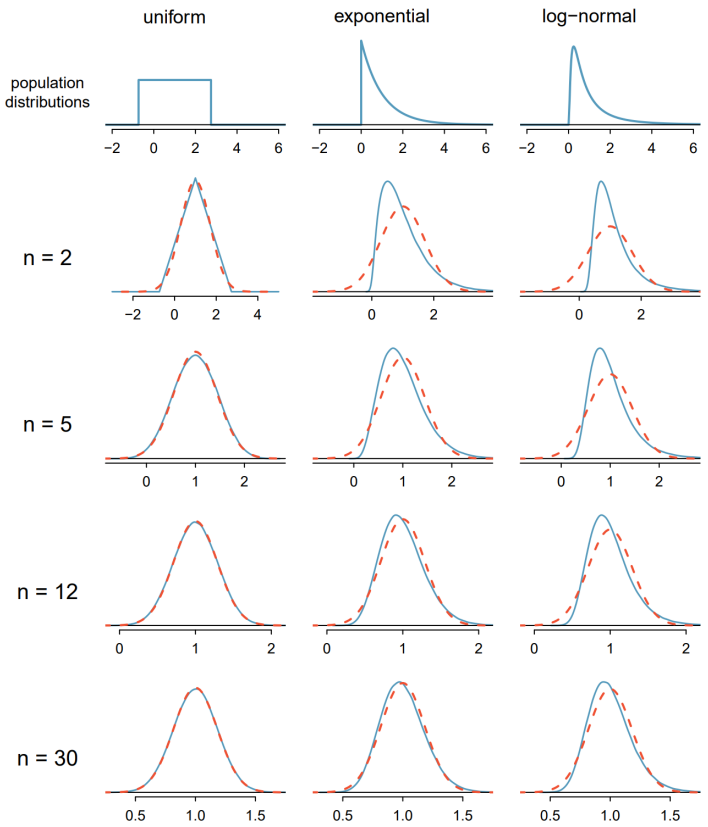

Figura 1. Distribución de las medias en relación al número de casos. Se grafica la distribución de las medias de n *muestras de variables con distribución uniforme, exponencial y log-normal. En línea punteada se superpone una distribución normal con los mismos parámetros (Gráfico obtenido de OpenIntro Statistics, Diez et al., 2012, p. 186)*

Sobre las medidas de dispersión relacionadas con el TLC, se observa que el error estándar (SE, el desvío estándar de todas las medias, ver Capítulo III) disminuye cuando el N de las muestras aumenta. El principio general que rige esto es el siguiente: las estimaciones de la población son más acertadas si se tiene una muestra más grande. Si el SE es muy alto, hay que preguntarse si la muestra es adecuada. El SE también depende de σ (desviación estándar de la población). Si la población tiene menor σ, el SE es menor. No es necesario contar con una distribución de medias para calcular el SE. Si no se conoce sigma, se usa s (desviación estándar de la muestra).

ISSN: 2796-8952

$$SE = \frac{s}{\sqrt{n}}$$

Gracias a estas propiedades de las medias y los cálculos de dispersión, se pueden hacer inferencias sobre la media de una muestra. A pesar de que se puede utilizar este principio para el testeo de hipótesis, el TLC requiere que se cumplan dos condiciones: la muestra debe ser independiente (es decir, las observaciones deben ser aleatorias) y el número de observaciones no debe ser menor a 30 (N > 30). Si bien es importante saber que la distribución muestral posibilita hacer inferencias a partir de una sola muestra, en la práctica la distribución real de las muestras es desconocida. El TLC es un concepto abstracto que permite generar deducciones sobre la muestra con la que se va a trabajar.

3. Hipótesis

Como se observó en el Capítulo III, para sacar conclusiones de probabilidad de las observaciones se las convierte en *z-scores*; de la misma manera, para hacer inferencias sobre la media de la población, se calcula el valor de *t*.

$$t = \frac{\bar{x} - \mu}{s}$$

Dado que no se conoce σ y se utiliza s en su lugar, no es del todo correcto asumir una distribución normal. Por eso, se utiliza la distribución *t*, que es muy similar y da cuenta de cuán certeras son las inferencias sobre σ. La distribución de *t* también es unimodal y simétrica (Figura 2), pero tiene una mayor área en los extremos y menor altura en el centro, lo que permite que más observaciones se sitúen a más de 2 desviaciones estándar de la media (esto compensa que σ se estime mal). Tiene un solo parámetro, los grados de libertad (gl), que son los que determinan el grosor de las colas (a mayor gl, más cercana a la normal). Con la fórmula para el valor de *t*, se puede asignar un valor a μ.

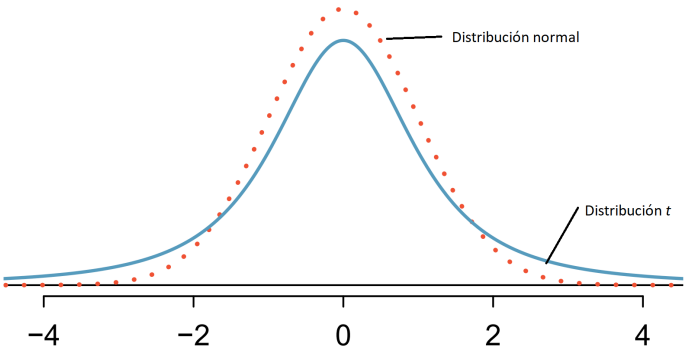

Figura 2. Distribución normal y distribución t (Gráfico obtenido de Open Intro Statistics de Diez, 2012, página 223)

En este punto, antes de realizar los cálculos, es necesario postular las hipótesis que se van a testear. Se distinguen entonces por un lado, la H_0 o hipótesis nula que es aquella que postula la "no diferencia" entre

ambas poblaciones x e y. Representa la mirada escéptica sobre un fenómeno. Si en una investigación se quiere probar un nuevo fenómeno, la H_0 postula que este no ocurre. Por otra parte, la H_1 o hipótesis alternativa presenta una postura nueva que se sostiene en la investigación y va en contra de la H_0.

En el ejemplo del caso de la duración de *pero*, los datos necesarios para comparar ambos grupos son la media ($\bar{x}$), el desvío estándar (s) y el número de observaciones (n) para cada grupo.

Duración de *pero*	Media ($\bar{x}$)	Desvío (s)	Observaciones (n)
Voces femeninas	172 ms	32 ms	18
Voces masculinas	200 ms	8 ms	20

Tabla 1. Media y desvío estándar de la duración de la emisión de "pero" por grupos en milisegundos.

Comparando estos datos en bruto, no se puede sacar una conclusión fiable sobre la duración de *pero* en estas dos muestras. No se sabe si la diferencia de esos dos números obedece a una variabilidad esperable en la toma de datos o si esa diferencia revela que hay un fenómeno que la produce. Esto conduce a la postulación de las hipótesis de investigación:

H_0 = no hay diferencias entre esos dos grupos, ambas poblaciones tienen la misma duración en la emisión de la conjunción, que la media sea 172 o 200 no es relevante. La relación entre las variables duración de *pero* y género (f-m) es independiente.

H_0: $\mu_f - \mu_m = 0$

H_1= sí hay diferencias entre los grupos. Las medias revelan que algo está ocurriendo. Una población presenta mayor duración en la emisión de la conjunción que la otra, no se debe al azar la diferencia numérica entre ambas muestras. La relación entre las variables duración de *pero* y género es dependiente.

H_1: $\mu_f - \mu_m \neq 0$

Si bien hay otros tipos de testeo de hipótesis, uno de los más habituales es el de comparar dos medias. Podría compararse la media de una muestra con un valor determinado, o postular que la media de la población es mayor o menor a un número (en ese caso se puede usar la primera fórmula de *t*). En este caso, al comparar dos medias, se utiliza una versión adaptada para el cálculo de *t*. Antes de esto, se debe calcular el error estándar para la diferencia de dos medias. A continuación, se observa que la forma de calcular el SE para dos muestras es obtener la raíz de la suma del desvío de las dos muestras. Esta es la fórmula (a la derecha está aplicada con el ejemplo de los datos de la duración de *pero*):

$$SE = \sqrt{\frac{s1^2}{n1} + \frac{s2^2}{n2}} = \sqrt{\frac{32^2}{18} + \frac{8^2}{20}} = 7{,}75$$

En el cálculo de t, se puede ver que el valor estimado es la diferencia de las medias de las muestras y se le resta el valor de la hipótesis nula, es decir 0, y esto se divide por el valor del error.

ISSN: 2796-8952

$$t = \frac{valor\ estimado - valor\ nulo}{error\ estándar} = \frac{(200 - 172) - 0}{7{,}75} = 1{,}03$$

Si se aplica la fórmula, se obtiene el valor de t = 1,03. El valor de t indica en qué parte de la distribución se encuentra el valor del estimado (en este caso, diferencias entre las medias). Por este motivo, al valor de t se le puede asignar una probabilidad o valor de p que permite interpretar el resultado en relación a las hipótesis de investigación.

3.1. El valor de *p*

El valor de p o el p-valor se define como la probabilidad de obtener el valor observado o uno más extremo asumiendo que la hipótesis nula es verdadera. Esto significa que, para hacer un testeo de hipótesis, se parte de la veracidad de la hipótesis nula. Así, la pregunta que conduce el testeo de las hipótesis es: dado que la H_0 es verdadera, ¿cuál es la probabilidad de que una muestra tenga el valor observado?

Si bien el valor de p asociado a un estadístico (z, t, F, entre otros) se puede obtener mediante ecuaciones, la forma más habitual es utilizar un método que lo compute automáticamente. Se pueden usar plataformas como R o SPSS para obtener estos cálculos estadísticos. En el próximo volumen de estos cuadernos se abordará el análisis estadístico de datos lingüísticos con R.

En el ejemplo tomado anteriormente, se contaba con un valor de t de 1,03 para la diferencia de las dos muestras. El valor de p asociado a este valor es 0.31. Esto significa que, si se asume que la diferencia de las medias es 0, la probabilidad de que la diferencia entre dos muestras aleatorias sea 28 ms es de 31 veces en 100. Si bien parece poco que las probabilidades de obtener esta diferencia de medias sea de aproximadamente el 31%, hay una manera de vincular este número con la verificación de las hipótesis de investigación.

3.2. Errores

Para rechazar o aceptar una hipótesis, se cuenta con el valor de p que manifiesta una probabilidad asociada a un estadístico, calculado a partir de las características de una muestra. Junto con esa probabilidad, se encuentra el porcentaje de error asumido de que la interpretación sobre las hipótesis sea incorrecta. Existen dos tipos de errores asociados al valor de p:

	H_0 es verdadera	H_0 es falsa
Aceptar H_0	Correcto	**Error de Tipo II**
Rechazar H_0	**Error de Tipo I**	Correcto

Tabla 2. Tipos de errores asociados al valor de p.

¿Cómo se cuantifica esta diferencia? Hay que elegir una probabilidad del Error de Tipo I (falso positivo) que estemos dispuestos a tolerar. El criterio para establecer esta probabilidad de Error de Tipo I se llama **nivel de significancia (α)**. Un nivel de α aceptable y que se usa habitualmente en lingüística y otras disciplinas similares es 0,05.

En el caso de la duración de *pero*, el valor de p es 0,096. Si se establece como límite aceptable α menor a 0,05, el valor p para las muestras no tiene significancia estadística. Es decir, ese valor de p no es

suficiente para rechazar la H_0. Dicho de otra manera, a partir de las muestras obtenidas no hay evidencias suficientes para afirmar que H_0 no es verdadera. Como no se puede rechazar H_0, se asume que no hay diferencias en la emisión de *pero* en relación al género.

También se puede definir el margen de probabilidad del Error de Tipo II (falso negativo). Esto se conoce como **beta (β)**. Un nivel de beta aceptable es 0,2. El **poder o potencia estadística** se basa en beta y se define como 1-β. El poder estadístico se aumenta con un incremento de N, ya que hace que la prueba sea más sensible a pequeñas diferencias. Sin embargo, el poder estadístico tiene un nivel de máxima que se alcanza después de un valor determinado de N.

4. Teorías de testeo de hipótesis

4.1. La propuesta de Fisher

A partir de 1925, Fisher se encargó de desarrollar y promover tests de significancia (Perezgonzalez, 2015). La perspectiva de Fisher sobre el análisis de datos se puede resumir en cinco pasos:

1. Elegir el test correcto de acuerdo con la naturaleza de las variables con las que se trabaja.

El tipo de test lo determina la distribución elegida, otras características vienen con la muestra.

2. Establecer la hipótesis nula.

Puede ser direccional (si se espera un resultado determinado) o no direccional (no hay predicciones sobre los datos). La H_0 no tiene que ser siempre nula (= 0). Puede ser que una diferencia no supere un valor específico.

3. Calcular la probabilidad teórica de los resultados observados considerando que H_0 es cierta (valor de *p*).

Cuando los datos de la muestra se acercan a la media de la distribución nula, la probabilidad aumenta (el valor de *p* es más alto); cuanto más se aleja el valor del centro de esa distribución, menos probable resulta ese valor (el valor de *p* es más bajo). El valor de *p* no es una probabilidad de un punto exacto, es una probabilidad acumulada del área que va desde el punto observado hasta la cola de la distribución. Es importante recordar que la H_0 siempre es verdadera, no se puede falsear, porque toda la distribución del test está basada en esta hipótesis.

4. Evaluar la significancia estadística α de los resultados.

Para determinar si el valor de *p* es lo suficientemente bajo para rechazar la H_0, se debe establecer el nivel de significancia. Los niveles de confianza más elegidos son 5% ($\alpha \approx 0,05$) o 1% ($\alpha \approx 0,01$). Los test estadísticos pueden ser de una cola de dos colas. En este último caso, el nivel de significancia se divide en áreas de ambos extremos (positivo y negativo). Así, el 5%, por ejemplo, cubriría el 2,5% de cada lado. Si se realizan múltiples tests, se incrementa la probabilidad de encontrar significancia estadística en los resultados. Para ello, se utilizan las correcciones que nivelan hacia abajo el valor de *p* (p. ej., la corrección de Bonferroni, que es el que más se usa a pesar de ser muy conservadora).

5. Interpretar el nivel de significancia de los resultados.

Un resultado significativo se explica de una de las siguientes dos formas. O este resultado ocurre raramente con probabilidad *p*, o la H_0 no explica los resultados satisfactoriamente. En general, los resultados no significativos no se reportan. Sin embargo, según Fisher aportan información y pueden resultar un medio para confirmar o reforzar la H_0. [Nota: rechazar o dudar de la H_0 bajo la perspectiva de Fisher, estadísticamente, no convierte a lo opuesto en verdadero, el valor de *p* no sirve para apoyar la H_1].

Para destacar los puntos positivos de la perspectiva de Fisher, su propuesta es flexible porque puede ser utilizada para investigación exploratoria. Además, es inferencial, ya que permite ir de la muestra a la población. Tiene también puntos negativos. Fisher nunca explicitó el poder estadístico de las pruebas. Sí habló de mayor "sensibilidad" cuando se aumentaba la muestra, pero nunca hizo un cálculo matemático para controlar este poder. Otra de las críticas es que en el marco teórico de Fisher no hay declaración explícita de la H_1, esta es la negación implícita de la H_0.

4.2. La propuesta de Neyman y Pearson

Al intentar mejorar la propuesta de Fisher, los autores terminaron por elaborar una forma alternativa para el testeo de datos, más matemática que la anterior. La propuesta de Neyman y Pearson (Neyman y Pearson, 1928, 1933; Neyman, 1956) se puede resumir en los siguientes pasos:

1. Establecer el tamaño del efecto esperado para una población.

Una de las innovaciones de esta perspectiva es explicitar la H_1 cuando se están explorando los datos. La H_1 representaría una segunda población que se sitúa junto a la población principal con el mismo continuo de valores. Estas dos poblaciones difieren en el tamaño del efecto. El tamaño del efecto es el grado de posibilidad de visualizar diferencias entre poblaciones. Cuanto menor es el tamaño del efecto, menos posibilidad de identificar pequeñas diferencias entre ellas. Como, en general, se desconocen los parámetros de la población, se recurre al tamaño del efecto de la población de muestras (*sampling distribution*). Las muestras no se superponen, están separadas. El porcentaje de distribución conocido es llamado beta (β) o MES (*Minimum Effect Size*) y representa la parte de la hipótesis principal HM (H_0)que no va a ser rechazada por el test.

2. Seleccionar un test óptimo.

Se debe elegir un test estadístico según su poder (los paramétricos tienen más poder que los no paramétricos) y en condiciones que aumenten el poder (por ejemplo., ampliar el N).

3. Definir la hipótesis principal (HM)

De acuerdo con Neyman y Pearson (1928), siempre se deben definir al menos dos hipótesis. La principal es la que debe ser testeada y debe incorporar el MES [Nota: H_0 y HM son muy similares y se postulan igual. Neyman y Pearson la llaman hipótesis nula también. Sin embargo, HM se postula explícitamente, incorpora un MES y compite con otra hipótesis, Figura 4].

ISSN: 2796-8952

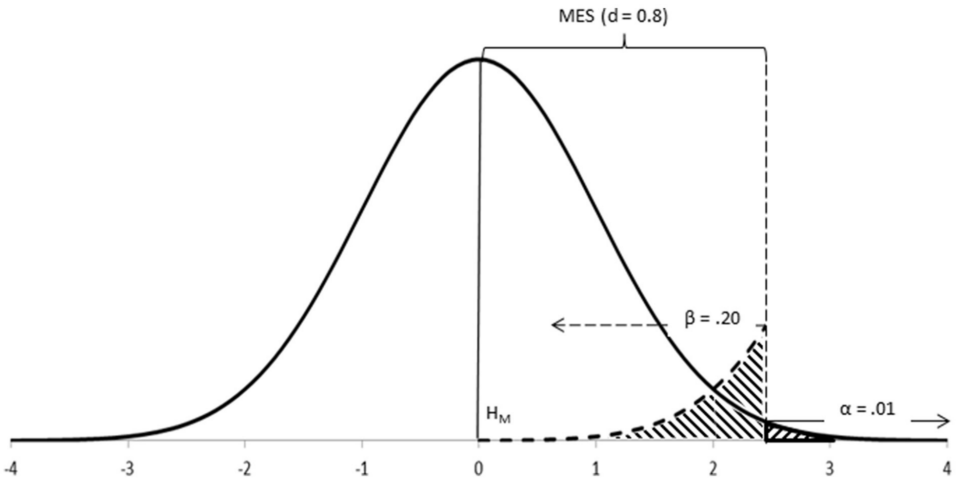

Figura 4. MES para calcular el valor de beta (Gráfico obtenido de Perezgonzalez, 2015, página 5)

Error de Tipo I: rechazar incorrectamente la HM (y aceptar H_1 incorrectamente). Para Neyman-Pearson este error tiene relevancia a largo plazo (no es identificable en un solo ensayo) y debe controlarse durante el diseño de un proyecto, no se puede controlar *a posteriori*.

Alfa (α): probabilidad de cometer el Error de Tipo I. Los niveles de alpha más utilizados son 0,05 y 0,01. A diferencia de la significancia estadística, alpha se debe incorporar en la postulación de la HM, no admite gradación y no se basa en rechazar la HM sino en aceptar una H_1.

Región crítica: alfa permite generar una región crítica, a partir de la cual se define la probabilidad de ese valor según las hipótesis de investigación. Si el valor del test cae fuera de la región, es probable dentro de la HM; en cambio, si entra en la región crítica, es un valor probable de la H_1.

Valor crítico: es el valor que delimita la región crítica y se define a priori, en la postulación de la HM.

HM : $Media_1–Media_2 = 0 \pm MES$, $\alpha = 0.05$, $CVt = 2.38$

4. Definir la hipótesis alternativa (H_1)

No es necesario establecer valores definidos para la H_1 (incluso los autores las definen vagamente), solo es una oposición de la HM que debe incluir el MES.

Error de Tipo II: rechazar incorrectamente la H_1. Es menos grave que el Error de Tipo I.

Beta (β): probabilidad de cometer el Error de Tipo II. No puede ser más pequeña que alpha (porque lo que testeamos es la HM), pero debe reducirse lo más posible. Neyman-Pearson proponen fijar β en el máximo 0,20 y el mínimo en α. Debe incorporarse este valor en la H_1.

H_1: $Media_1–Media_2 \neq 0 \pm MES$, $\beta = 0,20$

5. Calcular el tamaño de la muestra (N) para un buen poder estadístico (1–β).

El poder o potencia estadística es la probabilidad de rechazar correctamente HM en favor de la H_1. Es matemáticamente opuesta al error del Tipo II, es decir que es 1–β.

6. Calcular el valor crítico del test.

A partir de N y alpha podemos definir el valor crítico para delimitar los rangos desde donde evaluar nuestras hipótesis.

7. Calcular el valor del test de la investigación.

Cuando este valor está más cerca de cero, los datos están más cerca de la HM, mientras que cuanto más se alejan de cero, más se alejan de la HM. [Nota: desde la perspectiva de Neyman-Pearson se pueden usar los *p*-valores, solo que hay que recordar que los valores van en dirección opuesta].

8. Decidir a favor de la HM o la H1

Neyman (1955) dice:

- Si el resultado observado cae en la región crítica, rechazar HM y aceptar H_1.
- Si el resultado cae fuera de la región crítica y el test tiene mucha potencia, aceptar HM y rechazar H_1.
- Si el resultado cae fuera de la región crítica y el test tiene baja potencia, no se puede sacar conclusiones.

Como puntos a favor, la perspectiva de Neyman-Pearson incorpora la potencia o poder estadístico para evitar errores de Tipo II y, además, es una perspectiva mejor para investigaciones que toman diversas muestras de la misma población. Tiene en contra una menor flexibilidad y puede confundirse con la teoría de Fisher si no se tienen en cuenta el MES y beta.

4.3. NHST (Null Hypothesis Significance Testing)

Es la perspectiva que más se usa actualmente. Es una amalgama, sin criterio, de las propuestas de Fisher y Neyman-Pearson. Según el autor que la usa, puede tender más hacia una teoría o la otra. En general, se usa la perspectiva práctica de Neyman-Pearson y la filosofía de Fisher. Según Perezgonzalez (2015) se utiliza en todos los ámbitos (editores, investigadores, revisores), pero es una pseudociencia. El autor propone solucionar esto acercando la NHST a las dos teorías. Recomienda el uso de G*Power para implementar las recomendaciones para ambos casos. La teoría de Fisher es más cercana a la NHST. Muchos paquetes estadísticos la tienen como base (por ejemplo, SPSS). Se puede mejorar esto introduciendo los conceptos de poder estadístico y tamaño del efecto. No habla de H_1 ni nada por el estilo. La NHST es muy dañina para la teoría de Neyman-Pearson. Un error grave es la utilización de los valores *p* como evidencia de errores de Tipo I. Una solución es ajustar alfa y beta.

5. Malas interpretaciones sobre los test estadísticos, valores *p*, intervalos de confianza y poder (Greenland et al., 2016)

Debemos tener cuidado con la forma en que interpretamos los conceptos estadísticos en la práctica, pues estos no siempre son entendidos completamente por el investigador. Listamos aquí algunos ejemplos típicos de cómo se los suele malinterpretar.

ISSN: 2796-8952

5.1. Formas habituales de malinterpretar un único valor de *p*

1. El valor de *p* es la probabilidad de que la hipótesis principal sea cierta (p. ej., *p* = 0,01 es 1% de probabilidad de que H_0 sea verdad).

No. El test asume que H_0 es verdadera. El valor de *p* indica el grado en el cual los datos de la muestra se ajustan al modelo que predice el test. *p* = 0,01 indica que los datos no se acercan al modelo y a la hipótesis sostenida en el test (si todos los supuestos se cumplen).

2. El valor de *p* indica la probabilidad de que la asociación se deba al azar (chance).

No. Esta es una falacia que deriva de la anterior. El valor de *p* es la probabilidad de un estadístico asumiendo que los supuestos son verdaderos (un supuesto sería que la asociación se deba al azar).

3. Un resultado significativo ($p \leq 0,05$) significa que la H_0 es falsa o debe rechazarse.

No. Un valor bajo de *p* indica que los datos analizados son inusuales dados los supuestos estadísticos del test (entre ellos la H_0). Puede ser un valor causado por un gran error aleatorio o por una violación de algún otro supuesto.

4. Un resultado no significativo ($p > 0,05$) indica que la H_0 es verdadera o debe ser aceptada.

No. Un valor alto de *p* indica que el valor no es inusual dados los supuestos asumidos en el test. Puede ser un valor causado por un gran error aleatorio o por una violación de algún otro supuesto. Puede ser que los datos sean usuales para otra hipótesis también o bajo otros supuestos.

5. Un valor alto de *p* es evidencia a favor de la H_0.

No, salvo que *p* sea 1. Cualquier *p* menor a 1 es indicio de que la hipótesis del test no es compatible con los datos.

6. Un valor de *p* mayor a 0,05 significa que ningún efecto fue observado.

No. Significa que la H_0 sigue siendo una entre otras hipótesis sobre los datos. Salvo que el punto observado coincida exactamente con el punto nulo, es un error definir los resultados como "sin evidencia de un efecto".

7. La significancia estadística indica que se encontró una relación importante a nivel científico.

No. Puede deberse a una violación de supuestos o a errores de una muestra grande. El hecho de que sea inusual no significa que sea científicamente relevante.

8. La falta de significancia estadística indica que el tamaño del efecto es pequeño.

No. Especialmente en un estudio chico, grandes efectos pueden estar tapados por ruido.

9. El valor de *p* indica el porcentaje de chances de que ocurra nuestra observación si la H_0 es real.

No. Muestra la distribución de la frecuencia de la data si los supuestos son ciertos.

10. Si se rechaza H_0 porque $p < 0,05$ significa que la probabilidad del error de "falso positivo" es 5%.

No. Si rechazo H_0 y es real, el error es del 100%.

ISSN: 2796-8952

11. Un valor de $p = 0,05$ y $p \leq 0,05$ significan lo mismo.

No. Uno es un punto determinado y el otro incluye los valores más extremos.

12. Una forma correcta de reportar *p* es a partir de un valor que no incluya el = (p. ej., reportar $p < 0,02$ cuando $p = 0,015$)

No. No se pueden interpretar bien los resultados. Únicamente cuando los valores *p* son muy bajos pierden precisión (0,001).

13. La significancia estadística es una propiedad del fenómeno, entonces una prueba estadística detecta esta significancia.

Este error se manifiesta en la frase "se encontró evidencia de significancia estadística…". La significancia estadística es una descripción del valor de *p*, no una propiedad de la población.

14. Siempre se debe usar un valor de *p* bilateral.

No. Es lo más habitual y se usa si la H_0 incluye un valor específico (p. ej., 0). Si la pregunta de investigación es unilateral, es válido usar un valor de *p* unilateral.

5.2. Formas habituales de malinterpretar varios valores *p* y comparaciones

15. Cuando se testea una hipótesis y todos, o casi todos los resultados muestran $p > 0,05$, se asume que es una forma de verificar la H_0.

No. Que un grupo de muestras individuales no den significancia estadística no implica que sea una evidencia total de "falta de efecto".

16. Cuando los valores *p* obtenidos de dos muestras se encuentran en lados diferentes de 0,05, los resultados se interpretan como opuestos.

No. Los test estadísticos son sensibles a las variaciones de las muestras. Eso se puede reflejar en *p*, pero pueden presentar las mismas asociaciones. Las diferencias de resultados pueden evaluarse con pruebas (llamadas análisis de la heterogeneidad, interacción o modificación).

17. Cuando la misma hipótesis es testeada en diferentes poblaciones y se obtienen valores *p* similares, los resultados concuerdan.

No. Dos estudios pueden mostrar valores *p* similares pero no mostrar las mismas asociaciones, debido a diferencias en las poblaciones.

18. Si se observa un valor de *p* muy bajo en un estudio, es altamente probable que en el próximo estudio se obtenga un resultado similar.

No. El valor de *p* indica justamente la probabilidad de obtener un valor similar o menor. Es decir, si $p = 0,03$, la probabilidad de que sea igual o menor es 3%. La probabilidad de la réplica es exactamente el valor de *p*. Esto siempre que sean muestras independientes y que los supuestos sean ciertos en ambos estudios. Si no es así, el valor de *p* es muy sensible al tamaño de la muestra y a las violaciones de los supuestos.

ISSN: 2796-8952

5.3. Formas habituales de malinterpretar los intervalos de confianza

El problema empieza cuando se interpreta que $p > 0,05$ significa que la H_0 tiene 5% de chances de ser falsa y surgen las siguientes falacias:

19. El intervalo de confianza específico presentado en un estudio tiene 95% de probabilidad de contener el verdadero tamaño del efecto.

No. La frecuencia con la cual un intervalo contiene el efecto real es 100% si lo contiene y 0%, si no lo contiene El 95% refiere a la cantidad de intervalos de confianza que contienen el valor real. Esto significa que el 95% de los intervalos de confianza creados por muestras del mismo tamaño de la misma población van a contener el parámetro real de la población, si todos los supuestos son reales.

20. Un resultado fuera del intervalo de confianza del 95% ha sido refutado o excluido por los datos.

No. Lo que significa es que la combinación de los datos con los supuestos asumidos, en relación al criterio del 95%, es incompatible con los datos observados por algún motivo.

21. Si dos intervalos de confianza se superponen significa que las diferencias entre los valores estimados o de los estudios no es significativa.

No. Los intervalos pueden coincidir, pero sus diferencias pueden producir $p < 0,05$ por diferencias entre estudios.

22. Un intervalo de confianza de 95% observado predice que el 95% de los puntos estimados de los futuros estudios van a estar en dentro de ese intervalo.

No. El 95% se refiere a la cantidad de otros intervalos no observados que van a contener el verdadero efecto, no cuál es la frecuencia de futuros puntos estimados sobre ese intervalo.

23. Si un intervalo incluye el valor nulo y otro lo excluye, el primero es más preciso.

No. La precisión depende de la anchura del intervalo.

24. Si se rechaza la hipótesis nula porque $p > 0,05$ y el poder estadístico es de 90%, las chances de cometer un falso negativo son del 10%.

No. El 10% se refiere a la cantidad de veces que se cometería este error en este test estadístico a través de muchos estudios si la hipótesis alternativa es cierta, no se refiere a un uso singular del test.

25. Si *p* supera el 0,05 y el poder es del 90%, los resultados apoyan la hipótesis nula y no la alternativa.

Si bien es un razonamiento intuitivo, hay contraejemplos en los cuales el valor de *p* para la hipótesis nula está entre 0,05 y 0,10 y, sin embargo, las alternativas tienen un valor de *p* que excede el 0,10 y el poder es del 90%. Este tipo de cálculos pueden explicar la crisis de replicación, ya que definir los resultados de un estudio único de forma dicotómica anticipa un problema.

ISSN: 2796-8952

6. Referencias

Cetinkaya-Rundel, M. (2019). Data Analysis. Duke University. Coursera.

Diez, D. M., Barr, C. D., & Cetinkaya-Rundel, M. (2012). *OpenIntro statistics*. CreateSpace Independent Publishing Platform.

Ferrari L., Güemes M. M., Tallon L., Torres H., & Urcelay M. B. (2021). Correlatos prosódicos de los distintos valores de la conjunción pero. *Verba: Anuario Galego de Filoloxía, 48*. https://doi.org/10.15304/verba.48.6477

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society: Series B (Methodological), 17*, 69–78. https://doi.org/10.1111/j.2517-6161.1955.tb00180.x

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology, 31*(4), 337-350. https://doi.org/10.1007/s10654-016-0149-3

Johnson, K. (2011). *Quantitative methods in linguistics*. John Wiley & Sons.

Neyman, J. (1967). R. A. Fisher (1890-1962): an appreciation. *Science, 156*, 1456–1460. https://doi.org/10.1126/science.156.3781.1456

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: part I. *Biometrika, 20A*(1/2), 175–240. https://doi.org/10.2307/2331945

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, 231*, 289–337. https://doi.org/10.1098/rsta.1933.0009

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology, 6*, 223. https://doi.org/10.3389/fpsyg.2015.00223

ISSN: 2796-8952