

## ÍNDICE

|                                                                     |    |
|---------------------------------------------------------------------|----|
| PREFACIO .....                                                      | 9  |
| PRÓLOGO .....                                                       | 19 |
| <b>TEMA 1</b>                                                       |    |
| <b>INTRODUCCIÓN A LA ESTADÍSTICA</b> .....                          | 23 |
| 1.1. LA ESTADÍSTICA: DEFINICIÓN Y TERMINOLOGÍA .....                | 27 |
| 1.2. APLICACIONES DE LA ESTADÍSTICA EN CONTEXTOS LINGÜÍSTICOS ..... | 31 |
| 1.2.1. La estadística en el análisis del discurso .....             | 31 |
| 1.2.2. La estadística en la sociolingüística .....                  | 32 |
| 1.2.3. La estadística en la psicolingüística .....                  | 34 |
| 1.2.4. La estadística en la adquisición de lenguas .....            | 35 |
| Actividad 1.1. La estadística en la adquisición de lenguas .....    | 35 |
| 1.3. TIPOS DE ESTADÍSTICA: DESCRIPTIVA E INFERENCIAL .....          | 36 |
| 1.3.1. La estadística descriptiva .....                             | 37 |
| Actividad 1.2. La estadística descriptiva .....                     | 42 |
| 1.3.2. La estadística inferencial .....                             | 43 |
| 1.4. HERRAMIENTAS DIGITALES PARA LA INVESTIGACIÓN LINGÜÍSTICA ..... | 45 |
| 1.5. CONCLUSIONES .....                                             | 48 |
| EJERCICIOS .....                                                    | 49 |
| REFERENCIAS BIBLIOGRÁFICAS .....                                    | 52 |
| <b>TEMA 2</b>                                                       |    |
| <b>PREPROCESAMIENTO DE DATOS Y SELECCIÓN DE VARIABLES</b> .....     | 55 |
| 2.1. PREPROCESAMIENTO DE DATOS EN ESTUDIOS LINGÜÍSTICOS .....       | 59 |

|                                                               |    |
|---------------------------------------------------------------|----|
| 2.1.1. Limpieza de datos .....                                | 60 |
| 2.1.2. Integración de datos .....                             | 62 |
| Actividad 2.1. Integración de datos.....                      | 63 |
| 2.1.3. Transformación de datos .....                          | 65 |
| Actividad 2.2. Transformación de datos.....                   | 66 |
| 2.1.4. Reducción de datos .....                               | 68 |
| 2.2. SELECCIÓN DE VARIABLES PARA EL ANÁLISIS ESTADÍSTICO..... | 70 |
| 2.3. CONCLUSIONES.....                                        | 73 |
| EJERCICIOS .....                                              | 74 |
| REFERENCIAS BIBLIOGRÁFICAS .....                              | 77 |

### TEMA 3

#### ESTADÍSTICA DESCRIPTIVA PARA LA ENSEÑANZA Y EL TRATAMIENTO DE LENGUAS .....

|                                                                                   |     |
|-----------------------------------------------------------------------------------|-----|
| 3.1. Población y muestra.....                                                     | 85  |
| 3.1.1. Una aproximación a los diferentes tipos de muestreo.....                   | 89  |
| Actividad 3.1. Muestro para estudios lingüísticos.....                            | 91  |
| Actividad 3.2. Muestro para estudios lingüísticos.....                            | 91  |
| 3.2. Estadísticos básicos.....                                                    | 91  |
| Actividad 3.3. Estudio de aprendizaje del léxico con estadística descriptiva..... | 95  |
| 3.3. VISUALIZACIÓN DE DATOS DESCRIPTIVOS Y CÁLCULOS EN JASP.....                  | 96  |
| Actividad 3.4. Análisis de datos descriptivos con gráficos.....                   | 100 |
| EJERCICIOS .....                                                                  | 100 |
| REFERENCIAS BIBLIOGRÁFICAS .....                                                  | 104 |

### TEMA 4

#### INTRODUCCIÓN A LA ESTADÍSTICA INFERENCIAL .....

|                                                                        |     |
|------------------------------------------------------------------------|-----|
| 4.1. INTRODUCCIÓN A LA ESTADÍSTICA INFERENCIAL.....                    | 109 |
| Actividad resuelta 4.1. Introducción a la estadística inferencial..... | 110 |

|                                                                                                        |            |
|--------------------------------------------------------------------------------------------------------|------------|
| 4.2. UNA APROXIMACIÓN PRÁCTICA A LAS DIVERSAS PRUEBAS ESTADÍSTICAS PARA LA COMPARACIÓN DE MEDIAS ..... | 114        |
| 4.2.1. Contraste de hipótesis sobre dos medias en muestras independientes .....                        | 115        |
| 4.2.2. Contraste de hipótesis en muestras relacionadas.....                                            | 125        |
| Estudio de caso: una intervención de AD didáctica...                                                   | 126        |
| 4.2.3. Análisis de la varianza .....                                                                   | 134        |
| Estudio de caso: comparando la efectividad de métodos y técnicas de idiomas .....                      | 138        |
| 4.3. APLICACIONES A LA ENSEÑANZA Y EL TRATAMIENTO DE LENGUAS .....                                     | 144        |
| EJERCICIOS .....                                                                                       | 145        |
| REFERENCIAS BIBLIOGRÁFICAS .....                                                                       | 148        |
| <br><b>TEMA 5</b>                                                                                      |            |
| <b>LA REGRESIÓN EN LA ENSEÑANZA Y EL TRATAMIENTO DE LENGUAS .....</b>                                  | <b>151</b> |
| 5.1. LA REGRESIÓN LINEAL .....                                                                         | 155        |
| EJERCICIOS .....                                                                                       | 163        |
| REFERENCIAS BIBLIOGRÁFICAS .....                                                                       | 167        |
| <br><b>TEMA 6</b>                                                                                      |            |
| <b>CORPUS PARA LA ENSEÑANZA Y EL TRATAMIENTO DE LENGUAS .....</b>                                      | <b>169</b> |
| 6.1. LA LINGÜÍSTICA DE CORPUS: DEFINICIÓN Y ENFOQUE.....                                               | 173        |
| 6.2. LA COMPILACIÓN DE CORPUS LINGÜÍSTICOS.....                                                        | 174        |
| 6.3. LA EXPLORACIÓN DE CORPUS LINGÜÍSTICOS.....                                                        | 177        |
| 6.3.1. Sketch Engine como herramienta para la exploración de corpus.....                               | 178        |
| 6.3.2. UAM Corpus Tool como herramienta para el estudio de corpus de aprendices.....                   | 189        |
| 6.4. CONCLUSIONES.....                                                                                 | 195        |
| EJERCICIOS .....                                                                                       | 196        |
| REFERENCIAS BIBLIOGRÁFICAS .....                                                                       | 200        |

## **TEMA 7**

|                                                                                                        |     |
|--------------------------------------------------------------------------------------------------------|-----|
| <b>INTRODUCCIÓN A LA PSICOMETRÍA EN LA ENSEÑANZA DE LENGUAS</b> .....                                  | 205 |
| 7.1. UNA BREVE INTRODUCCIÓN A LA PSICOMETRÍA.....                                                      | 209 |
| 7.2. CONCEPTOS BÁSICOS: TEST, ESCALA Y CUESTIONARIO .....                                              | 210 |
| 7.3. UNA APROXIMACIÓN A LA EVALUACIÓN EN ENSEÑANZA DE LENGUAS DESDE UNA PERSPECTIVA PSICOMÉTRICA ..... | 212 |
| 7.4. PREGUNTAS DE RESPUESTA MÚLTIPLE: REDACCIÓN Y ANÁLISIS PSICOMÉTRICO BÁSICO .....                   | 216 |
| 7.3.1. Análisis psicométrico básico.....                                                               | 218 |
| EJERCICIOS .....                                                                                       | 224 |
| REFERENCIAS BIBLIOGRÁFICAS .....                                                                       | 226 |
| <b>WEBGRAFÍA</b> .....                                                                                 | 229 |
| <b>EPÍLOGO</b> .....                                                                                   | 231 |
| <b>SOLUCIONARIO</b> .....                                                                              | 235 |

# TEMA 1

## INTRODUCCIÓN A LA ESTADÍSTICA

### Introducción

En un contexto académico cada vez más orientado al análisis empírico y al rigor metodológico, la estadística se presenta como una aliada indispensable para el estudio científico del lenguaje. Lejos de ser una disciplina abstracta o exclusivamente numérica, esta herramienta permite transformar datos lingüísticos en conocimiento sistemático, verificable y útil. A lo largo de esta unidad, el alumnado iniciará un recorrido progresivo por los fundamentos estadísticos, descubriendo su valor para observar, medir y comprender fenómenos lingüísticos diversos. Así, se sientan las bases para una investigación responsable, crítica y basada en evidencias, que conecte teoría, práctica y tecnología en el ámbito de la lingüística aplicada.

### Índice de contenidos

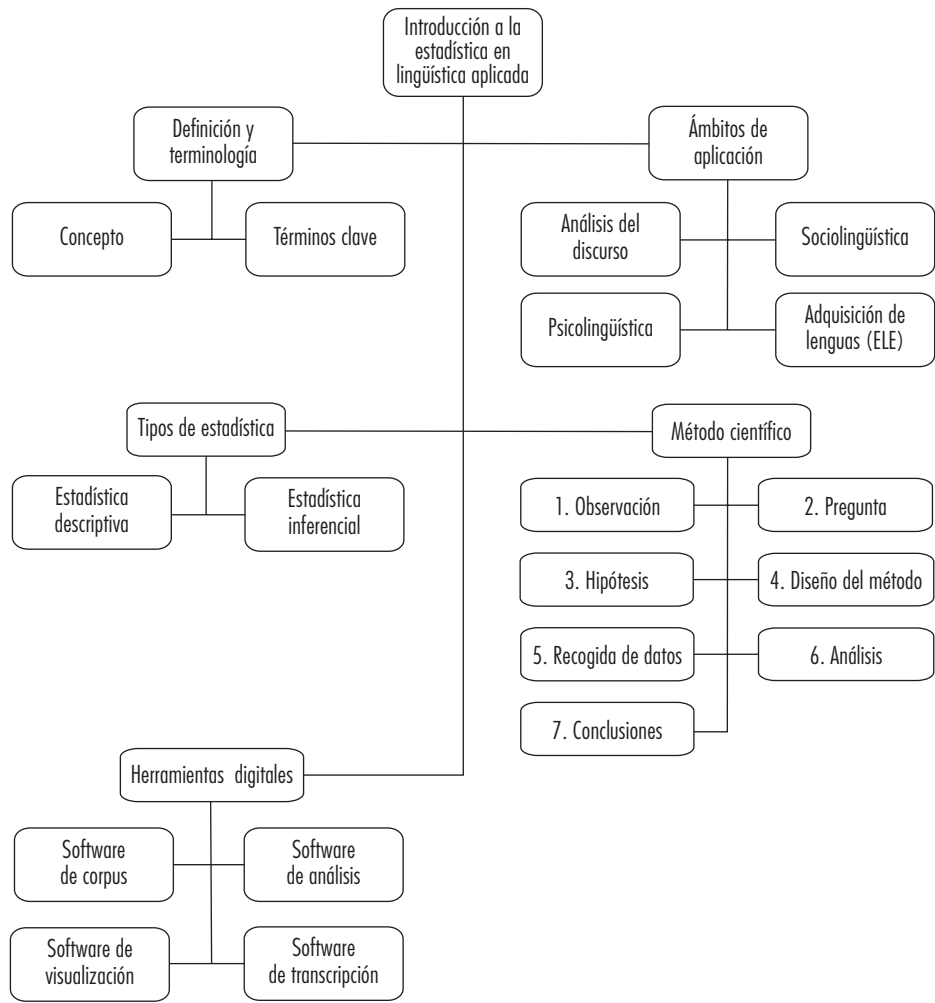
- 1.1. La estadística: definición y terminología
- 1.2. Aplicaciones de la estadística en contextos lingüísticos
  - 1.2.1. La estadística en el análisis del discurso
  - 1.2.2. La estadística en la sociolingüística
  - 1.2.3. La estadística en la psicolingüística
  - 1.2.4. La estadística en la adquisición de lenguas
- 1.3. Tipos de estadística: descriptiva e inferencial
  - 1.3.1. La estadística descriptiva
  - 1.3.2. La estadística inferencial
- 1.4. Herramientas digitales para la investigación lingüística
- 1.5. Conclusiones

### Objetivos de aprendizaje

- Identificar y definir conceptos clave de estadística, como media, mediana, desviación estándar o hipótesis nula, así como reconocer su terminología en español e inglés.
- Formular hipótesis y aplicar medidas estadísticas básicas a conjuntos de datos lingüísticos, utilizando herramientas digitales.

- Diseñar un estudio de investigación lingüística aplicando el método científico, desde la formulación de hipótesis hasta la recogida, análisis y presentación de datos estadísticos.

Mapa conceptual



## **Orientaciones para el aprendizaje**

- Leer de forma comprensiva los conceptos básicos de estadística y su terminología.
- Relacionar los contenidos con ejemplos de la lingüística aplicada.
- Identificar la diferencia entre estadística descriptiva e inferencial.
- Practicar el cálculo de medidas como media, mediana y desviación estándar.
- Formular hipótesis a partir de datos reales o simulados.
- Utilizar herramientas digitales para analizar información.
- Interpretar gráficos y tablas correctamente.
- Revisar los supuestos de cada prueba estadística.
- Aplicar el método científico en pequeños proyectos.
- Fomentar una actitud crítica ante los resultados.

## **Resumen**

La unidad introduce la estadística como herramienta clave en la lingüística aplicada, abordando conceptos básicos, terminología, tipos de estadística, método científico y análisis de datos. Destaca su valor para formular hipótesis, interpretar resultados y apoyar investigaciones rigurosas mediante herramientas digitales y ejemplos contextualizados.

### 1.1. LA ESTADÍSTICA: DEFINICIÓN Y TERMINOLOGÍA

La estadística, término popularizado por el filósofo y matemático alemán Gottfried Achenwall (1719-1772) y que deriva de la palabra alemán *Statistik* inspirándose en el término latín *statisticus* según el *Diccionario Etimológico Castellano en Línea*, es definida como el “Estudio de los datos cuantitativos de la población, de los recursos naturales e industriales, del tráfico o de cualquier otra manifestación de las sociedades humanas” por el *Diccionario de la Lengua Española*, y se considera una disciplina matemática con raíces antiguas que ha evolucionado constantemente a lo largo del tiempo. En sus inicios (Coronado Padilla, 2009), especialmente hasta el siglo XVII, se entendía como la recopilación de datos con fines administrativos y gubernamentales; ya en la época del Imperio Romano (Casanova Leal, 2021), se realizaban censos y recuentos de recursos como soldados, medios de transporte y riquezas. Se sabe, por ejemplo, que durante el mandato de César Augusto se llevó a cabo un censo poblacional. De hecho, el término *estadística* parece derivar del latín *status* (González Murias, Ribóo Almanzor y Tato Rodríguez, 1997), que alude al “estado de las cosas”, lo que da origen a lo que hoy conocemos como estadística descriptiva.

En la actualidad, vivimos rodeados de datos cuantitativos que reflejan múltiples aspectos de la sociedad y nos permiten entenderlos mejor: tasas de desempleo, fluctuaciones de precios como el IPC, gastos familiares, hábitos de consumo, evolución de los mercados financieros, siniestralidad vial, resultados electorales, uso de tecnologías, prevalencia de enfermedades, predicciones meteorológicas, y muchos otros indicadores. Incluso en disciplinas como la genética de poblaciones, se analizan frecuencias genotípicas y fenotípicas, así como diversas correlaciones entre factores biológicos y demográficos. Así, los datos numéricos y los porcentajes son herramientas fundamentales para comunicar y analizar la realidad, que luego se cuenta en medios de comunicación, textos académicos o publicaciones especializadas. Sin embargo, para poder interpretar la información recibida y aprobarla o reprobarla, el ciudadano debe adoptar una actitud crítica que solo puede conseguir mediante el desarrollo de ciertas habilidades. En otras palabras, conviene potenciar la “alfabetización estadística” (Pino y Estrella, 2012) para fortalecer el buen funcionamiento de la sociedad y favorecer la contribución democrática de cada persona.



Por lo tanto, antes de explorar de manera más exhaustiva este concepto en el contexto de la lingüística aplicada, conviene adoptar una definición que pueda encajar y englobar las distintas aplicaciones de la estadística (Martínez Bencardino, 2019) en nuestro ámbito de estudio. De esta manera, proponemos que la **estadística** se puede entender como una metodología que emplea análisis matemático para obtener información sobre diferentes realidades con las que convivimos, como pueden ser la adquisición de segundas lenguas, la enseñanza de lenguas, la lingüística forense, la lingüística computacional o el análisis del discurso, por mencionar solo algunos. Además, es necesario poder conocer los términos más significativos en el campo de la estadística antes de proceder con su puesta en marcha. Por esta razón, la *Tabla 1.1.* muestra un listado de terminología relevante (en español e inglés) usada en este ámbito junto con sus definiciones (Brezina, 2018; Holosko y Thyer, 2011; INEI, 2006; OECD, 2008; Posada Hernández, 2016).

**Tabla 1.1. Terminología relevante en el ámbito de la estadística en español e inglés junto con sus definiciones**

| Término en español  | Término en inglés | Definición                                                                                                                                                 |
|---------------------|-------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Caso                | Case              | Una unidad de análisis en un conjunto de datos. Puede ser una persona, un texto, una oración, entre muchos otros, dependiendo del tipo de estudio.         |
| Conjunto de datos   | Dataset           | Una colección organizada de datos que se usa para análisis. Puede ser una hoja de cálculo con columnas (variables) y filas (casos).                        |
| Desviación estándar | Standard          | Mide cuánto se alejan los datos, en promedio, de la media. Cuanto mayor sea, más dispersos están esos datos.                                               |
| Dispersión          | Dispersion        | Mide cuánto varían los datos entre sí. Si todos los datos son parecidos, la dispersión es baja; si son muy diferentes, es alta.                            |
| Distribución        | Distribution      | La forma en que se reparten los valores de una variable. Por ejemplo, si la mayoría de los textos tienen entre 100 y 200 palabras, esa es su distribución. |
| Distribución normal | Normal            | Una forma de distribución donde la mayoría de los valores están cerca de la media. Es muy común en fenómenos naturales y lingüísticos.                     |
| Hipótesis           | Hypothesis        | Una afirmación o suposición que se hace sobre un caso concreto o sobre la relación entre variables, y que se pone a prueba mediante el análisis de datos.  |

| <b>Término en español</b> | <b>Término en inglés</b> | <b>Definición</b>                                                                                                                                                                                                                                                                                                                  |
|---------------------------|--------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Hipótesis alternativa     | Alternative              | Es la suposición de que sí hay efecto o diferencia, por ejemplo, “Sí hay diferencia en el uso de conectores discursivos entre textos académicos y periodísticos”.                                                                                                                                                                  |
| Hipótesis nula            | Null                     | Es la suposición de que no hay efecto o diferencia entre dos variables, por ejemplo, “No hay diferencia en                                                                                                                                                                                                                         |
| Intervalo de confianza    | Confidence               | Un rango de valores que probablemente contiene el valor real que estamos buscando (p. ej., la media de una población), con un cierto nivel de seguridad (en torno al 95%).                                                                                                                                                         |
| Media                     | Mean                     | Una medida que indica el valor promedio de un conjunto de datos. Es una forma de resumir todos los valores en un solo número que representa el “centro” de los datos.                                                                                                                                                              |
| Mediana                   | Median                   | En un conjunto de datos ordenados de menor a mayor (o de mayor a menor), la mediana es el valor que los divide en dos grupos iguales, donde la mitad de los valores son menores y la otra mitad son mayores. Es decir, se trata de una medida de posición central que permite ver la distribución de los datos a un lado y a otro. |
| Moda                      | Mode                     | El valor que más se repite en un conjunto de datos.                                                                                                                                                                                                                                                                                |
| Medida estadística        | Statistical              | Cualquier número que resume o describe un conjunto de datos, como la media, la mediana o la desviación estándar.                                                                                                                                                                                                                   |
| Prueba estadística        | Statistical              | Un procedimiento para decidir si los datos apoyan o no una hipótesis, como t de                                                                                                                                                                                                                                                    |
| Robusto / Resistente      | Robust                   | Se dice de un método o prueba que sigue funcionando bien en diversas condiciones, incluso cuando los datos no cumplen todos los supuestos ideales.                                                                                                                                                                                 |
| Supuesto / Premisa        | Assumption               | Una condición que se da por cierta para poder aplicar una técnica estadística. Tras la aplicación de esta técnica estadística, se puede corroborar o refutar ese supuesto.                                                                                                                                                         |
| Tamaño del efecto         | Effect                   | Una medida cuantitativa que refleja la magnitud de una diferencia o relación entre variables. Es útil para saber si un resultado es importante, más allá de si es “estadísticamente significativo”.                                                                                                                                |

| Término en español            | Término en inglés | Definición                                                                                                                                                                                   |
|-------------------------------|-------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Valor atípico / Valor extremo | Outlier           | Un dato que es muy diferente al resto (un valor anormal). Puede tratarse de un error o de un dato interesante que merece atención, por lo que su identificación es extremadamente relevante. |
| Valor                         | p-                | Un valor comprendido entre 0 y 1 que se usa en el contraste de hipótesis para rechazar o aceptar la hipótesis nula. Concretamente, la hipótesis nula se rechaza si el valor                  |
| Variable                      | Variable          | Una característica que se mide y que puede cambiar entre casos, por ejemplo, la longitud de una oración o el número de sustantivos en un texto.                                              |

La relevancia de la estadística en el campo de la lingüística aplicada se debe a varias razones. En primer lugar, los datos cuantitativos permiten describir fenómenos lingüísticos de forma objetiva y sistemática, anulando las intuiciones de la persona investigadora. Segundo, ayudan a identificar patrones de uso, variación o adquisición del lenguaje, de manera que permite proponer nuevos métodos de enseñanza. Tercero, son esenciales para validar hipótesis de estudio y teorías lingüísticas con cierta confianza y certeza mediante la comparación de grupos o sujetos y la obtención de resultados. En cuarto lugar, favorecen la replicabilidad y la transparencia en la investigación, contribuyendo así a una ciencia abierta. En resumen, los datos cuantitativos son necesarios para que la lingüística aplicada pueda avanzar como ciencia empírica. Sin embargo, ¿qué rol juega la estadística en la validación de resultados? La estadística es la herramienta que permite interpretar datos lingüísticos de forma rigurosa y fundamentada conduciendo a conclusiones válidas basadas en esos datos y no en la casualidad o en la intuición. Por ejemplo, esta herramienta permite, además de generalizar resultados, detectar diferencias significativas, evaluar relaciones y estimar la fiabilidad de los resultados mediante medidas estadísticas.

Para conocer mejor qué papel pueden adquirir los análisis basados en estadística en sus áreas de interés, la siguiente sección (Sección 1.2.) recoge varios ejemplos de aplicaciones de esta herramienta en el ámbito de la lingüística aplicada para abordar, después, los conceptos de la estadística descriptiva e inferencial (Sección 1.3.) y, por último, qué herramientas digitales pueden ser útiles en el ámbito de la investigación lingüística (Sección 1.4.).

## 1.2. APLICACIONES DE LA ESTADÍSTICA EN CONTEXTOS LINGÜÍSTICOS

Dado que los análisis de corte matemático y numérico suelen resultar poco amigables a estudiosos de las Humanidades, este apartado ofrece ideas para facilitarle su puesta en práctica. Por esta razón, mostramos ejemplos accesibles en contextos realistas y comprensibles en las siguientes áreas: (1) el análisis del discurso, (2) la sociolingüística, (3) la psicolingüística, (4) la adquisición de lenguas. En esta última será usted quien determine si ha entendido la terminología presentada anteriormente.

### 1.2.1. La estadística en el análisis del discurso

En el ámbito del análisis del discurso, tomemos como ejemplo el análisis del uso de conectores discursivos en textos académicos y periodísticos donde nuestro objetivo es comparar el uso de este tipo de elementos lingüísticos (p. ej., *por lo tanto*, *sin embargo*, *además*) en estos dos géneros discursivos para ver si hay diferencias significativas.

- Cada texto que es analizado constituye un **caso** (p. ej., un artículo académico, una noticia) y el conjunto de todos ellos conforma el **conjunto de datos**, si se consideran uno solo, o dos conjuntos de datos si tenemos en cuenta cada género discursivo independiente. Por ejemplo, podría tratarse de 20 textos académicos y 20 textos periodísticos cuyos títulos aparecen en una tabla junto con el número de conectores discursivos usados en cada uno. Este número de conectores por texto sería la **variable** y se podría medir usando distintas **medidas estadísticas**, por ejemplo, la media, la media, entre otras.
- Así, partimos de un **intervalo de confianza** de 10-12 conectores por texto. Podríamos obtener una **media** (valor promedio) de conectores encontrados en los textos académicos de 12 (dentro de nuestro intervalo de confianza), mientras que en los textos periodísticos podríamos hallar 7 (fuera de ese intervalo). El número de conectores encontrados en el primer grupo organizado de menor a mayor sería 3-3-3-5-5-7-7-8-8-8-8-8-9-9-10-10-10-12-12, por lo tanto, la **mediana** tendría un valor de 8. En los textos periodísticos, el conector más empleado sería *no obstante*, con una recu-

rrencia de 7 veces, conformando la **moda**. La **desviación estándar** mediría cuánto varían los textos respecto a la media. Un **valor atípico** sería encontrar 50 conectores en un texto.

- La **distribución** del número de conectores por texto en cada conjunto de datos permitiría observar cómo se reparten estos valores. Por ejemplo, si la mayoría de los textos contienen entre 5 y 10 conectores, eso sugiere una concentración de datos en ese rango. En contextos ideales, podríamos asumir que esta variable sigue una **distribución normal**, caracterizada por una simetría alrededor de la media. Sin embargo, dado que estamos trabajando con conteos (número de conectores por texto), es más probable que los datos sigan una distribución de Poisson o una distribución asimétrica, especialmente si hay muchos textos con pocos conectores y pocos textos que contienen muchos conectores. La **dispersión** nos indica cuán variable es el uso de conectores entre textos: una mayor dispersión implica que algunos textos usan muchos más conectores que otros. Si los datos no cumplen con los **supuestos** de normalidad o homocedasticidad, es recomendable aplicar métodos **robustos**, como pruebas no paramétricas o modelos de regresión adecuados para conteos, en lugar de técnicas paramétricas clásicas.
- En nuestro análisis estadístico partiremos de una **hipótesis nula** que establece que no hay diferencia en el uso de conectores entre textos académicos y periodísticos. O de una **hipótesis alternativa** que dice que sí hay diferencia en el uso de conectores entre ambos tipos de texto. Una posible **prueba estadística** para descubrirlo podría ser chi-cuadrado de independencia porque permite determinar si hay una asociación significativa entre tipo de texto y frecuencia de uso de conectores (por categorías). Si el **valor  $p$**  es menor de 0,05, se rechaza la hipótesis nula, pues significaría que la diferencia observada es estadísticamente significativa. Por otro lado, se podría medir el **tamaño del efecto** para saber la magnitud de la diferencia entre los dos grupos de textos.

### 1.2.2. La estadística en la sociolingüística

En el ámbito de la sociolingüística, tomemos como ejemplo el análisis del uso del voseo (uso de *vos* en lugar de *tú*) en hablantes jóvenes de dos re-

giones, Buenos Aires (Argentina) y Montevideo (Uruguay), para ver si hay diferencias significativas en su frecuencia de uso en contextos informales.

- Cada hablante entrevistado constituye un **caso** (p. ej., un joven de 18 a 25 años), y el conjunto de todos ellos conforma el **conjunto de datos**. Si se analizan por separado, tendríamos dos conjuntos de datos: uno para Buenos Aires y otro para Montevideo. Por ejemplo, podríamos tener 30 entrevistas por ciudad, y en cada una se cuenta cuántas veces se usa el voseo en una conversación informal grabada. El número de usos del voseo por hablante sería la **variable** principal. Esta variable se puede analizar con distintas **medidas estadísticas** como la media, la mediana o la moda.
- Supongamos que el **intervalo de confianza** para el grupo de Buenos Aires es de 18 a 22 usos por entrevista. Si la **media** de usos en Buenos Aires es 20 (dentro del intervalo) y en Montevideo es 12 (fuera del intervalo), podríamos estar ante una diferencia relevante. Si los valores de Buenos Aires son: 10, 12, 14, 15, 16, 17, 18, 18, 19, 20, 20, 20, 20, 21, 21, 21, 22, 22, 23, 24, 24, 25, 25, 26, 27, 28, 28, 29, 30, 32, la **mediana** sería 21. Si el uso más frecuente es 20, entonces esa sería la **moda**. La **desviación estándar** nos diría cuánto varían los hablantes entre sí respecto a la media. Un hablante que usara el voseo 50 veces podría considerarse un **valor atípico**.
- La **distribución** de los datos mostraría cómo se reparten los valores, por ejemplo, la mayoría de los hablantes usan el voseo entre 15 y 25 veces. Si los datos se agrupan de forma simétrica alrededor de la media, estaríamos ante una **distribución normal**. La **dispersión** nos indicaría la homogeneidad o diversidad en el uso del voseo entre los hablantes. En este análisis, asumimos el **supuesto** de que los datos son independientes y siguen una distribución normal. Si no se cumple, podríamos usar un método **robusto** como una prueba no paramétrica.
- En nuestro análisis estadístico partiremos de una **hipótesis nula** que establece que no hay diferencia en la frecuencia de uso del voseo entre jóvenes de Buenos Aires y Montevideo. La **hipótesis alternativa** sería que sí hay diferencia. Una **prueba estadística** adecuada podría ser la prueba t de Student para comparar las medias de ambos grupos. Si el **valor p** es menor de 0,05, se rechaza la hipótesis nula, pues indicaría que la diferencia observada es esta-

dísticamente significativa. Finalmente, el **tamaño del efecto** nos ayudaría a saber la magnitud de la diferencia entre los dos grupos, más allá de que sea significativa.

### 1.2.3. La estadística en la psicolingüística

En el ámbito de la psicolingüística, tomemos como ejemplo el análisis del tiempo de reacción en tareas de decisión léxica en hablantes bilingües de español e inglés (hablantes bilingües con la misma competencia lingüística en ambas lenguas vs. hablantes con dominio dominante del español), a quienes se pide que indiquen lo más rápido posible si una palabra presentada en pantalla es una palabra real o no (p. ej., *mesa*, *florp*), para ver si hay diferencias significativas en esos tiempos.

- Cada participante constituye un **caso** (p. ej., un hablante bilingüe), y el conjunto de todos ellos conforma el **conjunto de datos**. Si se analizan por separado, tendríamos dos conjuntos de datos: uno para bilingües con la misma competencia en ambas lenguas y otro para bilingües con un mayor dominio del español. Por ejemplo, podríamos tener 25 participantes por grupo y en cada uno se registra el tiempo medio de reacción (en milisegundos) ante palabras reales, siendo este tiempo la **variable** principal. Esta variable se puede analizar con distintas **medidas estadísticas** como la media, la mediana o la desviación estándar.
- Supongamos que el **intervalo de confianza** para los bilingües con una competencia equilibrada es de 580–620 ms. Si la **media** de tiempos en ese grupo es 600 ms (dentro del intervalo) y en el grupo dominante en español es 670 ms (fuera del intervalo), podríamos estar ante una diferencia relevante. Si los valores del grupo dominante son: 580, 600, 600, 610, 620, 630, 640, 650, 660, 670, 690, 690, 690, 690, 700, 730, 730, 730, 730, 730, 770, 780, 790, 800, 810, la **mediana** sería 690 ms. Si el valor más frecuente es 730, esa sería la **moda**. La **desviación estándar** nos diría cuánto varían los tiempos de reacción entre los participantes. Un valor de 950 ms podría considerarse un **valor atípico**.
- La **distribución** de los datos mostraría cómo se reparten los tiempos, por ejemplo, la mayoría de los participantes responden entre 600

- y 700 ms. Si los datos se agrupan de forma simétrica alrededor de la media, estaríamos ante una **distribución normal**. La **dispersión** nos indicaría cuán homogéneos o variables son los tiempos de reacción. En este análisis, asumimos el **supuesto** de que los datos son independientes y siguen una distribución normal. Si no se cumple, podríamos usar un método **robusto**, como una prueba no paramétrica.
- La **hipótesis nula** de nuestro análisis estadístico sería que no hay diferencia en los tiempos de reacción entre los dos grupos de bilingües. La **hipótesis alternativa** sería que sí hay una diferencia. Para comprobar cualquiera de estas hipótesis, se usaría una **prueba estadística** apropiada. Si el **valor  $p$**  es menor de 0,05, se rechaza la hipótesis nula, lo que indicaría que la diferencia observada es estadísticamente significativa. Finalmente, el **tamaño del efecto** nos ayudaría a saber la magnitud de esa diferencia entre los dos grupos, más allá de que sea significativa.

#### 1.2.4. La estadística en la adquisición de lenguas

A continuación, exponemos un ejemplo contextualizado en el ámbito de la enseñanza de español como lengua extranjera (ELE), donde usted podrá encontrar la información necesaria para responder a las preguntas que aparecen más abajo. Una vez encuentre las respuestas a esas preguntas, escriba un breve mensaje en el foro indicándolas. Recuerde que el conjunto de datos es ficticio y sus fines son didácticos, por lo que los resultados no tienen por qué coincidir necesariamente con la realidad.

##### *Actividad 1.1. La estadística en la adquisición de lenguas*

En el marco de la lingüística aplicada, uno de los objetivos clave es comprender cómo se manifiesta el desarrollo lingüístico en aprendientes de lenguas extranjeras. Una forma de observar este desarrollo es a través de la producción escrita, analizando aspectos cuantificables que reflejan el nivel de competencia lingüística. En este ejercicio, nos centramos en la longitud media de las oraciones como indicador del grado de complejidad sintáctica en textos escritos por estudiantes de ELE. Se han recopilado 80 textos escritos por estudiantes de dos niveles distintos



según el MCER: 40 de nivel A2 (básico) y 40 de nivel B2 (intermedio alto). Cada texto ha sido segmentado en oraciones y se ha calculado la media de palabras por oración. El objetivo del análisis es determinar si existe una diferencia significativa en la longitud media de las oraciones entre los dos niveles, lo que podría reflejar un mayor dominio de estructuras complejas en los estudiantes de nivel B2.

1. ¿Qué representa un *caso* en este estudio?
2. ¿Cuál es la *variable* que se está analizando?
3. ¿Qué diferencia hay entre *media*, *mediana* y *moda* en este contexto?
4. ¿Qué significa que haya una *alta dispersión* en los textos de nivel B2?
5. ¿Qué es un *valor atípico* en este estudio? Aporte un ejemplo.
6. Redacte con sus palabras la *hipótesis nula* y la *hipótesis alternativa* de este análisis.
7. Si el *valor p* es 0,01, ¿qué decisión tomaría respecto a la hipótesis nula? ¿Por qué?
8. Si la *media* de palabras por oración en A2 es 9 y en B2 es 14, ¿qué nos indica esto sobre la competencia lingüística de los sujetos?

### 1.3. TIPOS DE ESTADÍSTICA: DESCRIPTIVA E INFERENCIAL

Tras haber comentado varios ejemplos relacionados con la lingüística aplicada donde podemos llevar a cabo un análisis de corte estadístico, le planteamos ahora lo siguiente dentro también del área de la adquisición de lenguas. Imagine que está analizando un corpus de producciones escritas de estudiantes de ELE y que ha recogido 100 redacciones para estudiar el uso del conector *sin embargo*. ¿Qué tipo de análisis sería más adecuado si quisiera diseñar una intervención didáctica basada en el uso de conectores contrastivos? ¿Por qué?

Por un lado, podría calcular cuántas veces aparece *sin embargo*, en qué posición del texto suele usarse (inicio de párrafo, entre oraciones, etc.), y qué otros conectores se usan con mayor o menor frecuencia. Este análisis le permitiría describir patrones observables en su muestra. Por otra parte, podría preguntarse: ¿Usan los estudiantes de nivel B2 significativamente más conectores contrastivos que los de nivel B1? Para responder, aplicaría una prueba estadística (por ejemplo, t de Student o chi-cuadrado) que le permitiría inferir si las diferencias observadas son

generalizables a toda la población de estudiantes de ELE. Así pues, nos encontramos ante dos tipos de análisis diferentes, uno basado en estadística descriptiva y otro basado en estadística inferencial, respectivamente, que representan dos ramas diferentes en el ámbito de la estadística.

La toma de decisiones sobre qué tipo de análisis queremos llevar a cabo es un paso necesario dentro del llamado **método científico**, el cual se impuso durante todo el siglo XVIII en toda rama científica para llegar también a los estudios del lenguaje (Bernal Leongómez, 2021). De tal forma, en cualquier investigación, independientemente de su ámbito o área de estudio, se seguirán una serie de pautas para su desarrollo, las cuales se organizarán de la siguiente manera (Doquin de Saint-Preux, 2024; Jiménez Ruiz, 2006; Martínez-Hincapié, 2013):

1. Efectuar una tarea de observación para definir claramente el objeto de estudio.
2. Formular una pregunta que nos permita conocer más allá del fenómeno observado.
3. Elaborar una hipótesis (posible explicación de los hechos) que pueda ser validada o refutada.
4. Diseñar el método para probar esa hipótesis (p. ej., qué pruebas o experimentos se llevarán a cabo, qué datos necesito, con qué herramientas debo contar, etc.).
5. Idear un plan que permita la recogida de datos medibles (bien en términos cualitativos, cuantitativos, o ambos).
6. Realizar un análisis del conjunto de datos que poseo aplicando el método elegido.
7. Anotar los descubrimientos que se han hallado o conclusiones sacadas de ese análisis, de manera que se pueda corroborar o anular la hipótesis establecida.
8. Pasamos a explicar con más detalle el papel de las dos ramas de la estadística mencionadas (descriptiva e inferencial) teniendo en cuenta los pasos destacados del método científico.

### 1.3.1. La estadística descriptiva

Como se puede conocer si partimos de su propio nombre, la estadística descriptiva implica un análisis de datos basado en la descripción de

los mismos, aportando un resumen que resulte de medidas estadísticas, tablas o gráficos. Es decir, la estadística descriptiva no busca sacar conclusiones o inferencias, sino describir un conjunto de datos para poder entenderlos, y es el tipo de estadística que se ha venido haciendo desde la antigüedad (Madsen, 2016).

Siguiendo en el contexto de los aprendices de ELE, tome como ejemplo el análisis de conectores causales usados en las producciones escritas de estos sujetos, donde podríamos preguntarnos cuáles son más frecuentes y responder a esta cuestión mediante el conteo de palabras o el cálculo de diversidad léxica (índice de tipo-*token*). Así pues, se decide aplicar la estadística descriptiva, materializándose los pasos del método científico de la siguiente manera:

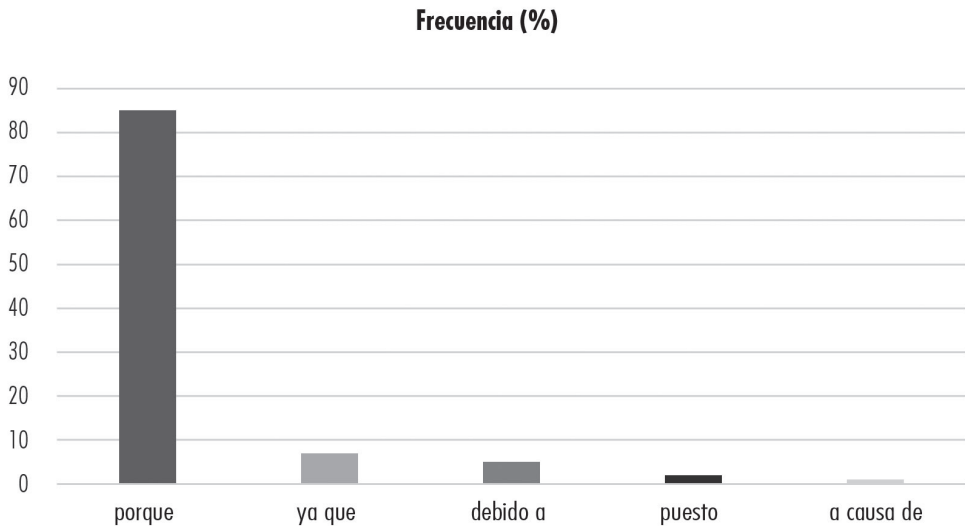
1. **Observación:** Durante la corrección de redacciones de estudiantes de nivel B1, se observa que un gran número de textos carece de conectores causales como *porque*, *ya que*, *debido a*, lo que afecta a la coherencia argumentativa.
2. **Pregunta:** ¿Cuáles son los conectores causales más utilizados por los estudiantes de nivel B1 en sus redacciones? ¿Con qué frecuencia los emplean?
3. **Hipótesis:** Los estudiantes de nivel B1 tienden a usar mayoritariamente el conector *porque*, mientras que otros conectores causales más formales o variados son poco frecuentes o inexistentes.
4. **Método:** Se recopilarán 30 redacciones escritas por estudiantes de nivel B1. A continuación, se identificarán y clasificarán todos los conectores causales utilizados. Finalmente, se utilizará una hoja de cálculo para registrar la frecuencia de cada conector.
5. **Plan de recogida de datos:** Se hará acorde al criterio de longitud, es decir, se recogerán redacciones argumentativas de entre 150 y 200 palabras. Se tendrán en cuenta dos tipos de variables, esto es, el tipo de conector causal y la frecuencia de aparición. Por último, se usará Excel para el conteo y análisis (ver *Tabla 1.2.*).
6. **Análisis de datos:** Se aplicará la estadística descriptiva para ofrecer el conteo total de conectores causales, el porcentaje de uso de cada tipo y una representación gráfica (p. ej., gráfico de barras o diagrama circular; ver *Figura 1.1.* y *Figura 1.2.*, respectivamente).

7. **Conclusiones:** Se confirma la hipótesis, ya que el 85 % de los conectores causales usados son *porque*, mientras que *ya que* y *debido a* aparecen en menos del 10 % de los casos, lo que sugiere una escasa variedad en el uso de conectores.

Este tipo de análisis en el contexto de la enseñanza de ELE puede ayudarle a conocer las debilidades del estudiantado y a diseñar actividades específicas, por ejemplo, de ampliación léxica y cohesión textual en este caso.

**Tabla 1.2. Ejemplo de recogida de datos en la estadística descriptiva**

| Conector causal   | Frecuencia absoluta | Frecuencia relativa |
|-------------------|---------------------|---------------------|
| <i>porque</i>     | 51                  | 85 %                |
| <i>ya que</i>     | 4                   | 6,67 %              |
| <i>debido a</i>   | 2                   | 3,33 %              |
| <i>puesto que</i> | 1                   | 1,67 %              |
| <i>a causa de</i> | 2                   | 3,33 %              |



*Figura 1.1. Ejemplo de representación de datos en la estadística descriptiva; gráfico de barras*