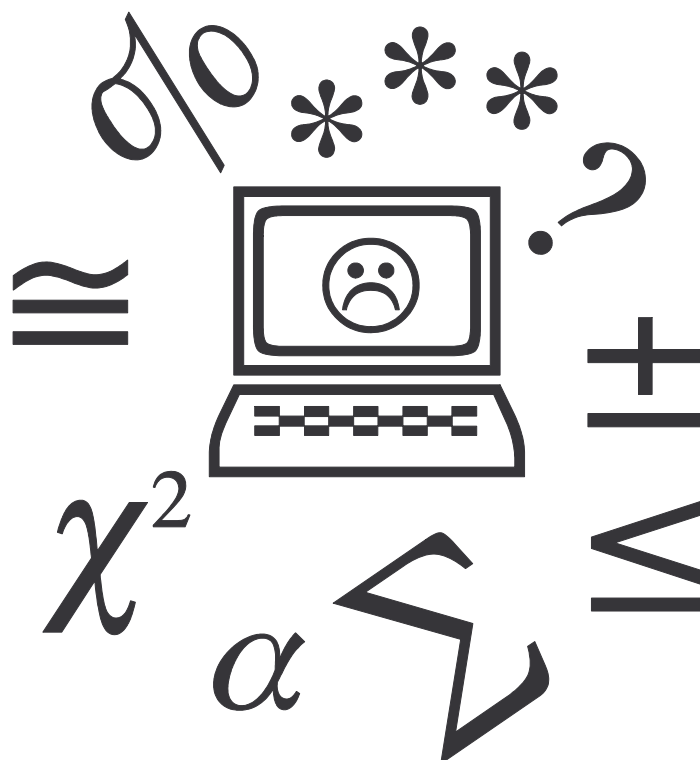

INTRODUCCIÓN A LA ESTADÍSTICA PARA ORNITÓLOGOS QUE ODIAN EL *ARDEOLA*

VERSIÓN 2.7

Enero 2009

por

Marcos Méndez Iglesias, Catedrático en Burrología y Letras



ÍNDICE

1. Introducción	3
2. Conceptos básicos	
El cambio de mentalidad	4
Las tres utilidades de la Estadística	5
Poblaciones y muestras, variables y datos	6
Tipos de variables	7
3. Estadística descriptiva	
Por qué es útil	9
Distribuciones de frecuencias	10
Medidas de tendencia central	11
Medidas de dispersión	12
Tablas	14
Representación gráfica	15
4. Contraste de hipótesis	
Por qué es necesario	17
Hipótesis nula y tipos de error	18
Cómo funciona un test	19
Elección del test adecuado	20
Tests paramétricos y no paramétricos	21
Comparando dos grupos	23
Comparando más de dos grupos	24
5. Asociación entre variables	
Por qué es útil	25
Correlación y regresión	27
Tests paramétricos y no paramétricos	28
Tablas de contingencia	29
Guía de utilización	30
6. Estadística y ordenadores	31
7. Más allá de este manual: autoaprendizaje de la Estadística	32
6. Bibliografía recomendada	33

1. INTRODUCCIÓN

Para la mayor parte de la gente corriente, la Estadística es sólo algo que tiene que ver con el fútbol o con las elecciones generales. Pero muchos aficionados a la Ornitología saben que es algo más que eso. Principalmente, su contacto con esta rama de las Matemáticas se produce cuando tratan de leer artículos en revistas como *Ardeola*. Normalmente, tras un título atrayente se oculta una jerga incomprensible de letras griegas, valores de probabilidad y nombres de "tests" en inglés que acaban con el ánimo del aficionado más interesado. Como mucho, se logra entender la introducción y el resumen del artículo, pero es imposible hacerse una idea clara de qué demonios ha averiguado el autor. Tras varias malas experiencias, esto produce un odio profundo por la Estadística y por revistas como *Ardeola* y, como efecto secundario, los aficionados limitan sus lecturas a las revistas más digeribles y con menos farrago matemático.

Yo también odiaba la Estadística. Cuando estudié Biología en la universidad tuve que soportar una horrible asignatura de Bioestadística, que es lo más cerca que he estado en mi vida del lenguaje marciano. Empollé como un burro y aprobé, pero mis conocimientos de la materia tras ese año se reducían a... ¡menudo rollo sin sentido! y el nombre gracioso que recibía alguno de los "tests" (por ejemplo, el "test" de Kolmogorov-Smirnov... ¿Serviría para comprobar el contenido alcohólico del vodka?). Todo esto cambió cuando, tiempo después, empecé a hacer mis propios estudios de campo sobre Ecología. En cuanto tuve mis datos en la mano y tuve que ponerme a interpretarlos, la Estadística se transformó de algo incomprensible en algo que tenía su lógica; de ser el enemigo pasó a ser un aliado; y de ser algo muy difícil y abstruso, a ser lo más fácil del Universo (bueno, sin pasarse...).

Lo primero que averigüé fue que la Estadística es **una herramienta**. Resulta útil cuando trabajas con algún tipo de información en forma de números recogida durante observaciones de campo o cualquier otro tipo de estudio. Y como pasa con toda herramienta, hay que saber: (1) por qué se utiliza, y (2) cómo utilizarla. Este manual proporciona una primera base para comprender ambas cosas. Espero que tras leerlo: 1) entenderás por qué en los artículos de *Ardeola* se utiliza la Estadística, 2) la lectura de *Ardeola* no sea un infierno y todas las letras griegas, los valores de probabilidad y los nombres estrafalarios cobren algún sentido y 3) si te planteas alguna vez hacer tus propios estudios de campo, sepas que tienes una herramienta que puede sacarte de varios apuros.

Pero si sigues sin entender una sola palabra en los artículos de *Ardeola* y sin ver ningún sentido a la Estadística... ¡Que te devuelvan tu dinero!

2. CONCEPTOS BÁSICOS: EL CAMBIO DE MENTALIDAD

Comenzar a entender la Estadística requiere un cambio de mentalidad, no difícil, pero sí fundamental. Has de pasar de una mentalidad **determinista** a una mentalidad **probabilista**. Normalmente, tendrás la idea (bastante natural, por cierto), de que existe una relación de causa a efecto entre las cosas. Si tiras al suelo agua (causa), el suelo se moja (efecto). Esto es algo determinista; si lo haces un millón de veces, observarás una y otra vez el mismo resultado. No se trata de algo probable, que unas veces ocurre y otras no.

La Estadística, en cambio, se basa precisamente en lo contrario; analiza cosas que unas veces ocurren y otras no, es decir, fenómenos que implican cierta incertidumbre, cierta probabilidad. En principio, parece que tales hechos deben ser raros o, al menos, que en la vida cotidiana no te los vas a encontrar. Pero resulta que la incertidumbre es algo tan común que casi pasa inadvertida. Por ejemplo, si coges una regla y mandas medir la longitud de una mesa a 10 amigos, probablemente se obtengan varias medidas diferentes. ¿Cómo es posible, si se trata de la misma mesa y de la misma regla? Simplemente las diferencias entre personas en la forma de coger la regla, en cómo consideran el ajuste entre el borde de la mesa y la escala de la regla, en su tendencia a redondear si la medida de la mesa no se ajusta exactamente a una marca en la regla, etc., introducen variación. Esto se conoce por **error de medida**.

Pero hay otras fuentes de variación que no dependen de la meticulosidad de una medida ni de lo sofisticado del aparato utilizado. Imagínate que no se trata de medir una mesa sino de determinar el peso de una especie cualquiera de ave, recogida durante una sesión de anillamiento. Probablemente, si tus 10 amigos realizan ese ejercicio se obtendrán 10 pesos diferentes. A las diferencias entre personas, en este caso, hay que añadir que, probablemente, cada uno ha capturado y pesado individuos diferentes, que difieren en edad, sexo y condición física, por no hablar de la constitución genética, etc. La **variación** es algo omnipresente en la naturaleza, y esa es la fuente de incertidumbre, o sea, de probabilidades. Uno puede tratar de describir esa variación con palabras (muchos, pocos, grandes, pequeños, azules, amarillos, ...). En algunos casos, eso es suficiente. Pero no cuando uno pretende acercarse a los detalles de un fenómeno natural. La Estadística permite describir esa variación en forma de **números**, lo cual resulta mucho más conveniente.

Por tanto, el primer paso a dar para comprender la utilidad de la Estadística es admitir que existe variación en la naturaleza e incertidumbre en cualquier conjunto de observaciones; en resumen, adquirir una mentalidad probabilista. No es casualidad que en Estadística se use el término "variable". Las tres aplicaciones de la Estadística que se discuten en este manual tienen como objetivo, en el fondo, tratar con esa variación.

2. CONCEPTOS BÁSICOS: LAS TRES UTILIDADES DE LA ESTADÍSTICA

¿Para qué sirve la Estadística? *Toda su complejidad aparente se reduce a su utilidad en el manejo de la variación, en forma de números.* En concreto, permite realizar tres tareas, relacionadas con la variabilidad en las observaciones. El primer paso, pues, es hacerte una idea clara de cuál de esas tres tareas quieres realizar.

1. ¿Resumir y describir la información? Estadística descriptiva

Te permite organizar, resumir y describir la información recogida de modo que sea fácilmente comprensible para tí y para los demás, mediante el uso de:

- a) medidas de tendencia central: Media, Mediana, Moda,
- b) medidas de dispersión: Desviación Típica, Rango, Coeficiente de Variación,
- c) tablas,
- d) representación gráfica.

2. ¿Comparar valores medios entre grupos de datos? Contraste de hipótesis

Te permite decidir si la variación entre grupos de datos es sistemática (debida a un fenómeno biológico) o meramente “ruido” debido a la variación natural existente en todo grupo de organismos, como se comentó al hablar de la mentalidad probabilista. Esto se logra mediante el uso de tests estadísticos. La elección del test adecuado depende de:

- a) el tipo de variables,
- b) el número de muestras o tratamientos que se desea comparar,
- c) el cumplimiento de los requisitos necesarios para cada test (tests paramétricos y tests no paramétricos).

3. ¿Descubrir si hay relación entre dos variables? Medidas de asociación

Te permite descubrir la existencia, dirección y fuerza de la relación entre dos variables, mediante:

- a) correlación,
- b) regresión,
- c) tablas de contingencia.

Reconozco que ahora mismo todo esto parece misterioso, pero si sigues leyendo verás que tiene bastante sentido. Aquí simplemente he introducido el tipo de preguntas que la Estadística ayuda a resolver. El resto del manual trata con más detalle cada una de estas tres utilidades e intenta dejar claro por qué son interesantes e incluso necesarias cuando tienes que analizar tus datos. Pero antes es conveniente dar una serie de definiciones. Voy a ello.

2. CONCEPTOS BÁSICOS: POBLACIONES Y MUESTRAS, VARIABLES Y DATOS

Has salido al campo, has hecho un montón de observaciones y quieres ver si tienen algún sentido. ¡Cuidado! Esos números pueden referirse a cosas muy diversas. Para no liarse en lo sucesivo, es conveniente dar unas definiciones previas.

Unidad de estudio.- Es cada *elemento* que va a ser estudiado. Lo más normal es que se trate de *individuos*, pero no tiene por qué ser así. Si se estudian, por ejemplo, las longitudes de los picos de las gaviotas, las unidades de estudio son *partes de un individuo*. El estudio también puede centrarse en bandos, colonias, nidadas, etc, que son *grupos de individuos*. También se denomina unidad de muestreo.

Población.- En Estadística, este término tiene un significado diferente del habitual en la vida cotidiana. Se refiere al *grupo total* de unidades de estudio sobre las que quiere llegarse a alguna conclusión. Así que en Estadística puede hablarse de una "población de picos de gaviota" o una "población de bandos" sin estar loco. El tamaño de la población varía según la generalidad que quiera darse a las conclusiones (p. ej., las gaviotas del Cabo Peñas, las gaviotas de Asturias, las gaviotas en general).

Muestra.- Es un *subconjunto de la población*, que se elige mediante algún procedimiento específico¹. El tamaño muestral se representa por *n*, y es igual a la cantidad de unidades de estudio tomadas de la población.

Variable.- Una variable es cualquier *característica* que ha sido medida, registrada o cuantificada durante el estudio realizado, para cada una de las unidades de estudio incluidas en la muestra.
Ejemplos: temperatura, sexo de las aves, especie, hábitat utilizado, número de huevos.

Dato.- Es cada uno de los registros o *valores individuales* que toma la variable que se ha medido. Puede recibir otros nombres (*caso, observación, medida*).
Ejemplos: 5°C (temperatura), macho (sexo de las aves), *Hirundo rustica* (especie), bosque (hábitat), 3 (número de huevos).

En resumen, en tu estudio de campo estás interesado en extraer conclusiones sobre una *población* de determinadas *unidades de estudio*. Para ello tomas *muestras* de dicha población y en ellas mides u observas *datos* para cada una de las *variables* que te interesan.

¹ Esos procedimientos corresponden a la planificación del estudio, tratada en el primer manual de esta serie.

2. CONCEPTOS BÁSICOS: TIPOS DE VARIABLES

Un aspecto crucial de la Estadística son las variables, o sea, aquello que uno mide. Dado que constituyen la materia prima fundamental en cualquier análisis estadístico, es importantísimo saber distinguir las distintas formas que pueden tomar. He aquí los principales tipos de variables.

Variable mensurable.- Es aquella que puede ser expresada en forma de números.

Variable continúa.- La que puede tomar infinitos valores entre dos puntos fijos cualesquiera.

Ejemplos: longitudes, áreas, pesos, volúmenes, temperatura, tiempo, porcentajes.

Variable discreta.- La que únicamente puede tomar un número fijo de valores y además no puede tomar valores intermedios entre esos valores.

Ejemplos: números (de huevos en una puesta, de plumas, de nidos, etc.).

Variable ordinal.- Aquella que no puede ser expresada en forma de números, pero que puede ser ordenada o clasificada según su magnitud.

Ejemplos: escalas de abundancia, probabilidades de cría.

Atributo.- Cualquier variable no expresable en forma de números, pero que puede expresarse de un modo cualitativo en forma de categorías, aunque sin establecer ninguna relación de orden entre ellas.

Ejemplos: sexo, hábitat, color.

La diferencia entre estos tipos no es del todo radical; a veces depende simplemente de cómo han sido medidas las variables. Las variables continuas se miden normalmente con una exactitud limitada y eso puede hacerlas discretas en la práctica. Por otra parte, algunas variables discretas tienen tantos valores posibles (p. ej. número de aves) que pueden considerarse continuas a todos los efectos. Las variables mensurables pueden expresarse a veces como variables ordinales. Por ejemplo, la longitud del pico en varias especies de aves, puede expresarse en milímetros (variable continua) o como un rango desde la especie con el pico más corto a la especie con el pico más largo (variable ordinal). La operación contraria no siempre es posible: algunas variables ordinales no pueden ser expresadas como variables mensurables.

Como resumen, la Fig. 1 muestra la aplicación en un caso concreto de todos esos conceptos.

Villaviciosa						
Sexo:	Machos			Hembras		
Longitud del pico (mm)	20 datos			20 datos		
Nivel de marea:	Bajo	Medio	Alto	Bajo	Medio	Alto
Nº sondeos por minuto	20 datos	20 datos	20 datos	20 datos	20 datos	20 datos
Nº presas por minuto	20 datos	20 datos	20 datos	20 datos	20 datos	20 datos

Zeluán						
Sexo:	Machos			Hembras		
Longitud del pico (mm)	20 datos			20 datos		
Nivel de marea:	Bajo	Medio	Alto	Bajo	Medio	Alto
Nº sondeos por minuto	20 datos	20 datos	20 datos	20 datos	20 datos	20 datos
Nº presas por minuto	20 datos	20 datos	20 datos	20 datos	20 datos	20 datos

Fig. 1. Esquema de un estudio sobre la alimentación del Combatiente en Asturias. En dos localidades (Zeluán y Villaviciosa) se midió la longitud del pico, el número de sondeos realizados en el fango por minuto y el número de presas capturadas en dicho tiempo para 20 machos y 20 hembras por localidad, en función del nivel de marea donde los individuos se estaban alimentando (bajo, medio, alto).

Población: los Combatientes en Asturias.

Variables: 5 variables (longitud del pico -continua-, número de sondeos -discreta-, número de presas -discreta-, sexo -atributo-, nivel de marea -ordinal-).

Muestras: dos muestras (Zeluán y Villaviciosa).

Datos: 20 datos por variable, localidad, y sexo.

3. ESTADÍSTICA DESCRIPTIVA: POR QUÉ ES ÚTIL

Después de realizar un estudio sobre cualquier tema, te encontrarás con un buen montón de muestras, variables, datos, etc., etc. En ese momento te enfrentas a dos problemas: 1) cómo extraer de todo ese montón una idea general de lo que pasa, ese "mensaje oculto en los datos", y 2) cómo contárselo a los demás de forma que lo entiendan.

El examen de los datos uno a uno no es práctico. Necesitarás algún modo de resumir u ordenar esos datos para que los patrones que muestran resulten evidentes. Por ejemplo, puedes tomar miles de observaciones sobre las fechas de aparición de la Golondrina Común en Asturias. Pero sería absurdo que cuando alguien te pregunte cuándo puede observar la especie tú le contestases dándole, una a una, todas las fechas en que la has visto. Lo más razonable es resumir esas fechas en forma de un estatus para la especie (invernante, estival, etc.). Algo similar ocurre con la información de tipo numérico. La Estadística ofrece varias posibilidades para resumir la información disponible. Las formas más habituales son el cálculo de valores medios y la realización de gráficas o tablas, donde se muestran frecuencias o medias de los datos obtenidos. Si alguna vez has hecho una gráfica, una tabla o una media, has estado utilizando la Estadística, aunque te fastidie reconocerlo.

En esta sección se introducen algunos conceptos que conviene saber cuando resumes tus datos y se dan algunas normas para la presentación clara de los mismos.

3. ESTADÍSTICA DESCRIPTIVA: DISTRIBUCIONES DE FRECUENCIAS

La forma más elemental de resumir un grupo de datos referentes a una variable es hacer una representación gráfica de los mismos en forma de *distribución de frecuencias*: en el eje horizontal de la gráfica se representan los diferentes valores que puede tomar la variable, y en el eje vertical se representan el número de veces (o sea, la frecuencia) que se ha obtenido cada valor de la variable (Fig. 2). Esta representación gráfica se aplica a variables continuas, tras agrupar en clases los infinitos valores posibles de la variable (Fig. 2). Para representar las variables discretas, ordinales o atributos se utiliza un diagrama de barras (descrito más adelante en este manual).

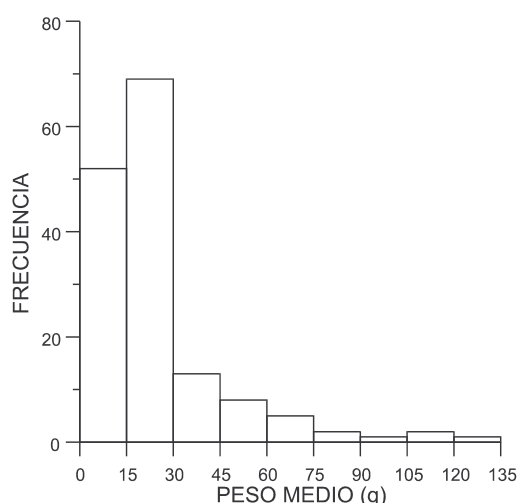


Fig. 2. Distribución de frecuencias del peso medio de las aves paseriformes de Europa.

La distribución de frecuencias te permite conocer de un vistazo la "forma" de los datos: 1) ver si existe una *tendencia a agruparse* en torno a un valor dado, y 2) qué grado de variabilidad o *dispersión* existe en los datos. Las posibilidades son, en principio, infinitas pero hay varios tipos de distribución de frecuencias que aparecen comúnmente y merecen un nombre.

Distribución unimodal.- Los datos se agrupan en torno a un valor y la frecuencia de datos disminuye a medida que los valores se alejan del mismo (Fig. 3 A y B). Un tipo especial de distribución unimodal es la *distribución normal*, en la que se cumplen una serie de requisitos sobre el valor central y la dispersión en torno al mismo.

Distribución uniforme.- Los datos se distribuyen con una frecuencia aproximadamente similar a lo largo de todos los valores medidos (Fig. 3 C).

Distribución bimodal.- Los datos se agrupan en torno a dos valores, con una disminución en la frecuencia de valores extremos e intermedios entre esos dos valores (Fig. 3 D).

3. ESTADÍSTICA DESCRIPTIVA: MEDIDAS DE TENDENCIA CENTRAL

Las distribuciones de frecuencias no son el único método, ni a veces el más conveniente, de resumir o presentar la información recogida. Las *medidas de tendencia central* permiten expresar todos los datos correspondientes a una variable mediante un solo número, que sirve como representante de los mismos. Cumplen, por tanto, la primera función de las distribuciones de frecuencias. Las principales medidas de tendencia central son tres: la media, la mediana y la moda.

1. Media.- Se representa por \bar{X} y se calcula mediante la siguiente fórmula:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{\text{suma de todos los datos}}{\text{n}^\circ \text{ de datos}}$$

2. Mediana.- Es un valor que divide la distribución de frecuencia de los datos en dos partes, tal que la mitad de los datos son menores que la mediana y la otra mitad son mayores. Es decir, es el punto medio de la distribución de frecuencias.

3. Moda.- Es el valor, o clase, que más se repite dentro del conjunto de datos, es decir, el "punto más alto" de la distribución de frecuencias.

El cálculo de estas medidas de tendencia central sólo es posible para variables mensurables y ordinales. No tiene sentido calcular la media de un atributo como el sexo, pues no se puede sumar "macho + hembra" y dividirlo entre dos. Para los atributos siempre hay que trabajar con *frecuencias* (número de datos en cada categoría del atributo), que pueden representarse gráficamente o expresarse como porcentajes del total.

¿Cómo elegir entre ellas?- En general, la media es la medida de tendencia central más útil y más empleada. El uso de la **media** es el más apropiado cuando las variables son mensurables y la distribución de los datos es unimodal y aproximadamente simétrica (Fig. 3 A). Cuando valores extremos distorsionan la distribución de los datos (Fig. 3 B), el uso de la **mediana** es más apropiado pues se ve menos afectada, pero en la práctica esta medida de tendencia central no se utiliza demasiado. Si se trata de una variable ordinal, o sólo necesitas una descripción rápida y aproximada de la tendencia central, puedes utilizar la **moda**, que también es útil cuando la distribución está distorsionada por valores extremos o la distribución es bimodal (Fig. 3 D).

3. ESTADÍSTICA DESCRIPTIVA: MEDIDAS DE DISPERSIÓN

No todos los individuos son iguales. Si mides una variable en varios individuos obtendrás valores por lo general diferentes. Una manera de resumir esa variación es mediante las *medidas de dispersión*. Como las medidas de tendencia central, resumen en un solo número todo un grupo de datos, que expresa la variación presente en los mismos. Cumplen la segunda función de las distribuciones de frecuencia. Las medidas de dispersión más comunes son las tres siguientes.

1. Desviación típica.- Se representa por S o d.t. y se calcula mediante la fórmula:

$$S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n}} = \text{Raiz cuadrada de } \left\{ \frac{\text{Suma del cuadrado de (cada valor menos la media)}}{\text{N}^\circ \text{ de datos}} \right\}$$

En la práctica se usa la desviación típica muestral, en que se divide por $n - 1$, no por n .

2. Rango.- Es la diferencia entre el valor más alto y el valor más bajo de la variable. También se expresa a veces indicando ambos valores, separados por un guión.

3. Coeficiente de variación.- Es igual a la desviación típica dividida por la media, y expresado como un porcentaje:

$$CV = \frac{S}{\bar{X}} \times 100$$

¿Cómo elegir entre ellas?- La medida de dispersión más útil es la **desviación típica**. Sólo debes usar el **rango** cuando dispones de pocas medidas y otra medida de dispersión sería muy grosera, o cuando todo lo que necesitas conocer es la dispersión general de las medidas. Utiliza el **coeficiente de variación** cuando quieras tener una idea de la variabilidad relativa de dos o más variables cuyas medias son muy diferentes en magnitud (por ejemplo, longitudes de picos y longitudes de alas). Esto se ve facilitado por su carácter adimensional, es decir, no depende de las unidades en que se mida la media (p. ej., milímetros para los picos y centímetros para las alas).

La forma más breve de resumir los valores que toma una variable es mediante una medida de tendencia central junto con una medida de dispersión. La forma habitual es utilizar la media junto con la desviación típica y el tamaño de muestra ($\bar{X} \pm \text{d.t.} (n)$). Todo ello es equivalente a una distribución de frecuencias, pero más sencillo de calcular y en mucho menos espacio. Además se obtienen fácilmente usando una calculadora normal y corriente.

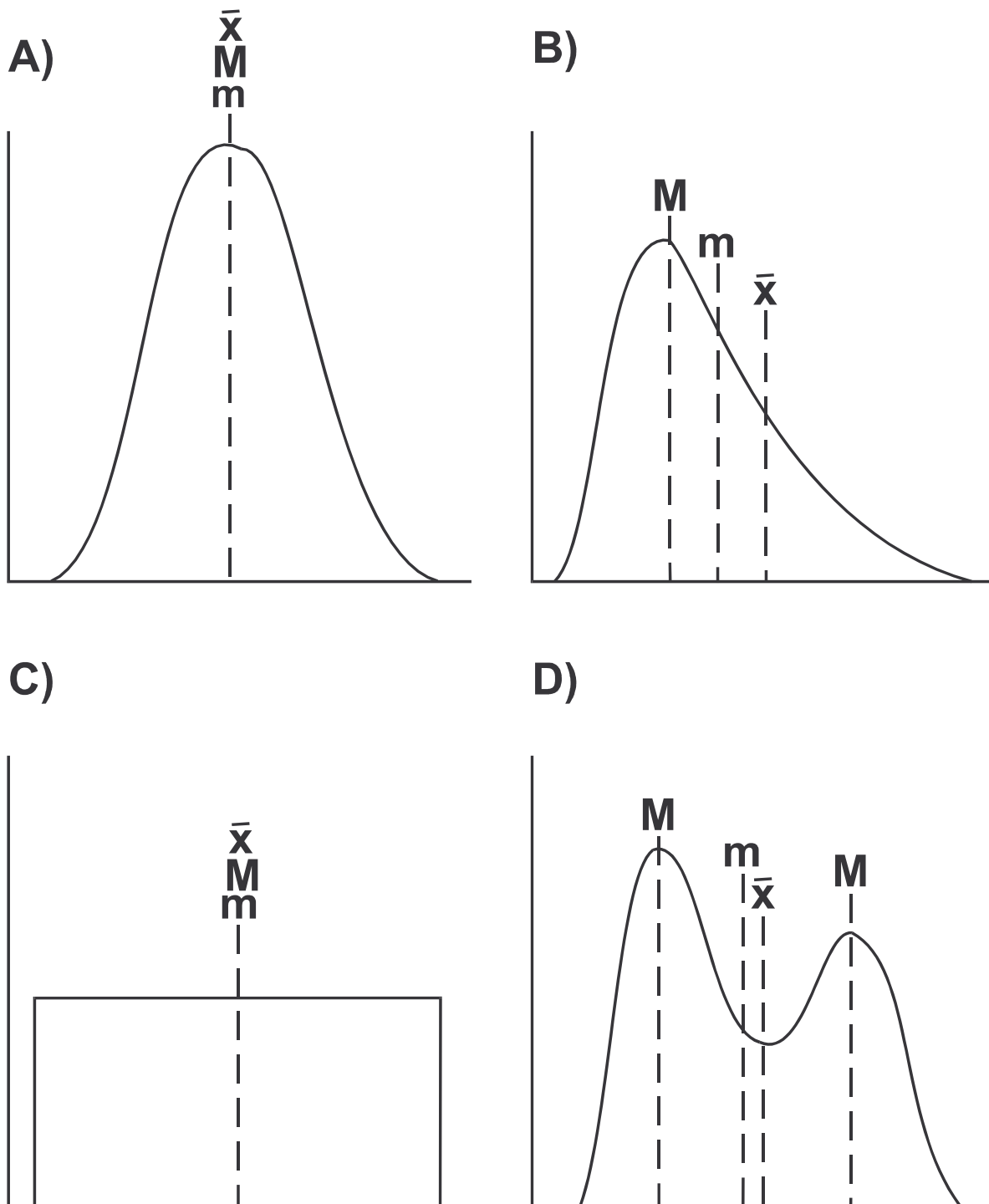


Fig. 3. Varios tipos de distribuciones de frecuencia, con la posición de diversas medidas de tendencia central: A) distribución unimodal simétrica, B) distribución unimodal asimétrica o sesgada, C) distribución uniforme, y D) distribución bimodal. M = moda; m = mediana, \bar{x} = media.

3. ESTADÍSTICA DESCRIPTIVA: TABLAS

Las tablas son una forma habitual de presentar los datos recogidos, las frecuencias, las medidas de tendencia central y de dispersión, etc., etc. Para conseguir una mayor claridad en su elaboración conviene seguir algunas reglas sencillas.

1. Cada tabla debe tener una leyenda autoexplicativa, que permita entenderla sin tener que recurrir al texto. Por ejemplo:

- el uso de abreviaturas debe evitarse o, en caso de usarse, deben ir explicadas,
- cuando se proporcionen medidas, deben indicarse las unidades,
- debe darse el tamaño de muestra para las medias tabuladas.

2. Dentro de la tabla, las columnas deben ordenarse de un modo que resulte lógico y comprensible, y que permita una comparación fácil de los datos.

3. Debe procurarse no dejar espacios vacíos en las tablas. En caso de que no sea posible, debes distinguir claramente cuando se trata de un valor cero (0) y cuando se trata de falta de datos (-).

4. Su uso debe ser racional. No presentes datos innecesarios. Tampoco hagas tablas de modo abusivo. Muchas veces varias tablas pueden "fusionarse" en una sola en la que figura toda la información relevante.

¿Cuántos decimales?- Cuando resumes información y calculas medias, te enfrentarás a la aparición de números con decimales. Esto no es extraño en las variables continuas, pero en algunas variables discretas (p. ej. número de huevos de una puesta) está claro que los decimales son algo "ilusorio" (no existe el "medio huevo"). Para decidir cuántos decimales utilizar en las tablas presentadas sigue los siguientes consejos:

1. para las variables continuas, no des más decimales de la precisión con que mediste esas variables,
2. para las variables discretas, no des más de un decimal.

Cómo redondear decimales.- Para redondear esos decimales de más que aparecen al usar la calculadora o el ordenador, sólo hay que seguir unas reglas muy sencillas:

(1) si el último número después de la última cifra a considerar es igual o mayor que 6, se suma 1 a la última cifra; por ejemplo 6,32654 se redondea a dos cifras decimales como 6,33;

(2) si el último número después de la última cifra a considerar es menor que 5, se deja la última cifra como está; por ejemplo 6,32654 se redondea a una cifra decimal como 6,3;

(3) si el último número después de la última cifra a considerar es igual a 5, se redondea la última cifra al número par más próximo; por ejemplo, 6,32654 se redondea a tres cifras decimales como 6,326.

3. ESTADÍSTICA DESCRIPTIVA: REPRESENTACIÓN GRÁFICA

Aunque una tabla con medias y desviaciones típicas es un buen modo de resumir los datos, una imagen vale más que mil palabras. Por eso es útil hacer representaciones gráficas.

A continuación se describen los principales tipos de representación gráfica.

1) Histograma o distribución de frecuencias.- Se utiliza para representar frecuencias en variables continuas (Fig. 2). Las barras están pegadas unas a otras, para indicar el carácter continuo de la variable.

2) Diagrama de barras.- Se utiliza para representar frecuencias en variables discretas, ordinales o atributos. Se diferencia de un histograma en que las barras están separadas entre sí, para indicar el carácter discreto de las variables (Fig. 4 A).

3) Gráficos con barras de error.- Se utilizan para representar valores medios, indicados por medio de puntos o de barras, a los que se añade un segmento o un semisegmento que indica una medida de dispersión (Fig. 4 B). Se aplican a variables mensurables y a veces a las ordinales, nunca a atributos.

4) Diagramas de puntos, o de dispersión.- A diferencia de los tipos anteriores, no representan frecuencias o valores medios en el eje vertical. En su lugar, representan dos variables, mensurables u ordinales, una en el eje horizontal y la otra en el eje vertical (Fig. 6).

Recuerda que las representaciones gráficas:

- no añaden nada a los datos que ya no estuviese allí; su función es simplemente mostrarlos de modo más claro,
- son una necesidad, no un lujo; no emplees más gráficos de los necesarios,
- deben resaltar los patrones de interés sin comprometer la integridad de los datos,
- deben carecer de "chatarra gráfica" como colores, sombreados, líneas y volumen innecesarios.

Nunca:

- repitas la misma información en un gráfico y una tabla, o con dos representaciones gráficas distintas,
- utilices gráficos de "tarta" (Fig. 4 C) ni gráficos de barras apiladas (Fig. 4 D); ¡más que ayudar a ver los patrones, los oscurecen!

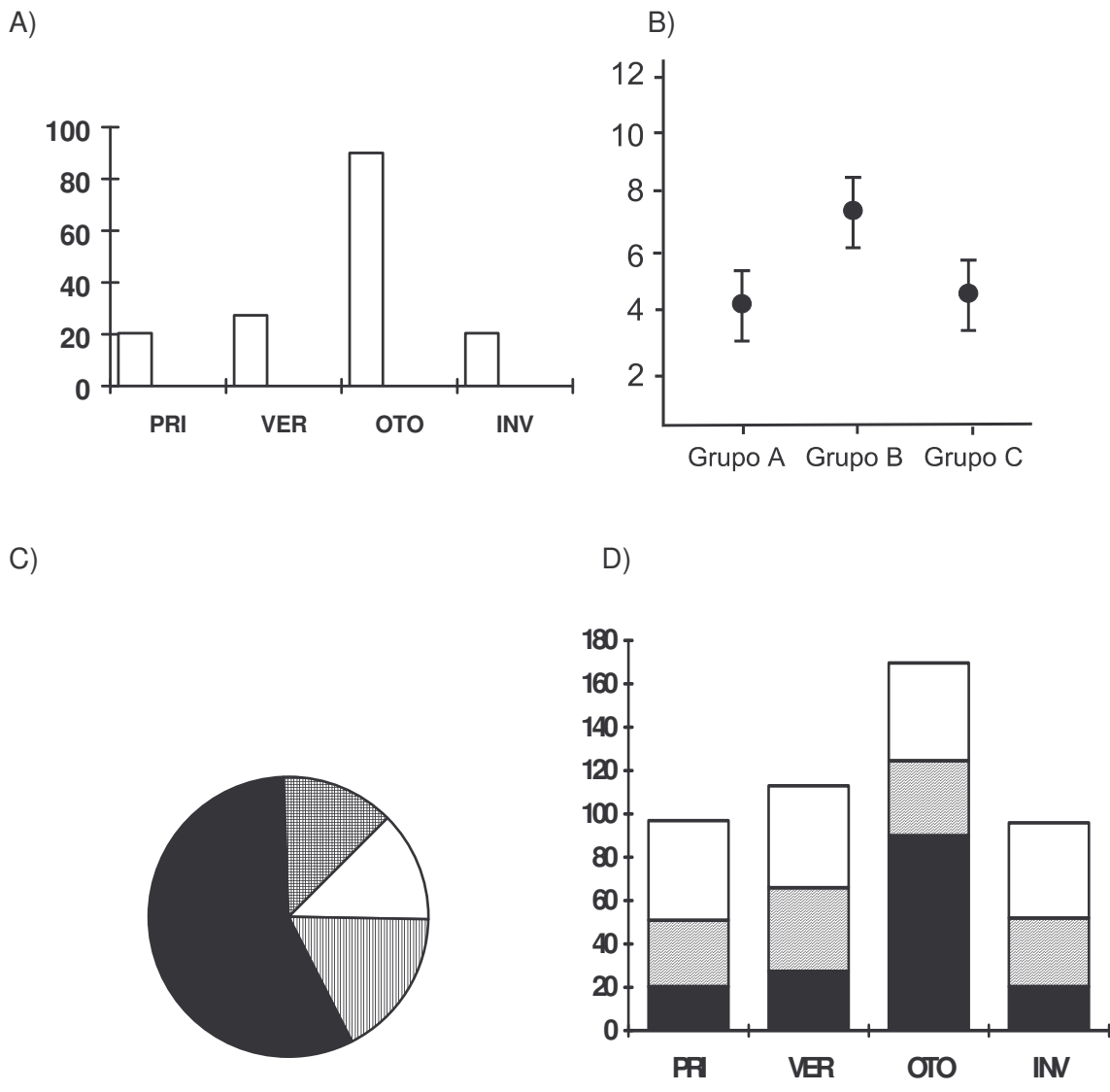


Fig. 4. Varios tipos de representaciones gráficas: A) diagrama de barras, B) gráfico con barras de error, C) gráfico de "tarta", D) diagrama de barras apiladas.

4. CONTRASTE DE HIPÓTESIS: POR QUÉ ES NECESARIO

En Ornitología, muchas veces te interesa comparar valores medios de una variable entre distintos grupos de datos. Imagina que te has planteado el estudio descrito en la Fig. 1 sobre los Combatientes. Has ido a Villaviciosa y has tomado tus datos. Cuando acabas de tomar los datos en Zeluán te encuentras con un amigo y tras charlar un rato, resulta que él ha tenido exactamente la misma idea que tú y también ha tomado datos en Villaviciosa y Zeluán. Así que decidís compararlos pero, para vuestra sorpresa, las medias son distintas en todas las variables ¿Alguno de los dos ha tomado mal los datos? La respuesta es que no y que esa diferencia era esperable, por lo que comparar a vista los valores medios no es suficiente.

Por un lado, dos muestras de una misma población serán casi con toda seguridad diferentes, aunque las tome la misma persona. Está claro que si fuese posible medir todos los individuos de la población, obtendrías la auténtica media. Al tomar una muestra, simplemente dispones de una aproximación a dicha media real. Esa aproximación estará tanto más cerca de la media auténtica cuanto mayor sea el tamaño de muestra. Pero como en la mayor parte de los casos no es posible medir la variable que interesa en todos los individuos de la población (cuando la población es demasiado grande o cuando la medida es de carácter destructivo [p. ej., peso del hígado]) tienes que asumir que la media que has obtenido tiene asociada una incertidumbre. Dicha incertidumbre tiene varias causas: (1) distintas personas pueden cometer distinto error al medir una variable, (2) cada Combatiente es diferente, y al medir distintos animales habeis encontrado distintas medias, (3) el tamaño de muestra fue diferente.

En vista de ello, es de esperar que tus muestras de Villaviciosa y Zeluán también sean diferentes. La cuestión es ¿difieren por esos efectos asociados al muestreo de toda población, o existe una componente añadida de diferencia real, debida a causas biológicas? ¿Qué hacer ante esas discrepancias en los datos, que siempre van a aparecer? Hay varias alternativas. La primera es no medir nada que varíe... ¡pero son las cosas que varían las que tienen interés! Contar patas, picos o alas no tiene ningún misterio. Así que esa opción es poco útil. Como alternativa, ya que hay que medir cosas variables, lo mejor es hacerlo sobre toda la población, para que no haya discrepancias debidas a que una persona mide unos animales y otra persona mide otros diferentes. Pero resulta que hay más de 1000 combatientes en Villaviciosa. Medir todo eso llevaría siglos y no os queda más remedio que usar una muestra. La tercera opción es usar la Estadística. Los tests de hipótesis se utilizan para contrastar si los valores medios obtenidos para una serie de grupos de datos son realmente diferentes", es decir, existe una causa subyacente que tiene algún sentido más allá de esos errores asociados con el uso de muestras (sexo, edad, poblaciones diferentes) o si dichas diferencias se deben simplemente a cuestiones del muestreo (error de medida, medida de individuos diferentes).

4. CONTRASTE DE HIPÓTESIS: HIPÓTESIS NULA Y TIPOS DE ERROR

Siguiendo con el ejemplo de la Fig. 1, has medido las longitudes de los picos de los Combatientes macho y hembra en Villaviciosa, pues sospechas que los machos pueden tener picos mayores y conseguir alimento en el fango a mayor profundidad que las hembras. Tienes dos medias pero a estas alturas ya sabes que, dado que provienen de muestras, están afectadas por un error. La Estadística te permite comparar esas medias, y una analogía con lo que ocurre en un juicio servirá para aclarar la manera de hacerlo.

El procedimiento a seguir en este caso es el siguiente.

- 1) Especificar una **hipótesis nula** y una **hipótesis alternativa**, excluyentes, basadas en los motivos, ideas, etc. que han llevado a recoger los datos. En este caso, sospechas que existe una *diferencia* entre machos y hembras, en concreto, los machos pueden tener picos más largos que las hembras. La hipótesis nula siempre afirma que **NO** existen diferencias. Es una hipótesis de trabajo que, obviamente, se intenta rechazar. La hipótesis alternativa (que es la que te interesa) afirma que sí existen esas diferencias y se acepta siempre que se rechaza la hipótesis nula. Estas dos hipótesis pueden considerarse veredictos posibles en un juicio (hipótesis nula = no culpable; hipótesis alternativa = culpable).
- 2) La Estadística hace de juez que dicta una sentencia. La sentencia puede ser: (a) no rechazar la hipótesis nula, es decir, no rechazar la posibilidad de que los machos de Combatiente tienen picos de tamaño similar al de las hembras, o (b) rechazar la hipótesis nula (y aceptar la hipótesis alternativa) de que existen diferencias. Es decir, debes decidir entre la "no culpabilidad" o la "culpabilidad". Recuerda que la hipótesis nula nunca se acepta, sólo se rechaza o no se rechaza. Siempre cabe la posibilidad de que la hipótesis nula sea falsa (y el acusado sea culpable) aunque no pueda rechazarse, como se verá en el punto siguiente.
- 3) Al emitir el juicio puedes acertar o equivocarte. Las equivocaciones pueden ser de dos tipos: **error de tipo I**, consiste en rechazar la hipótesis nula cuando era cierta (has condenado a un inocente); y **error de tipo II**, consiste en aceptar la hipótesis nula cuando es falsa (has dejado libre a un culpable). Obviamente te interesa minimizar ambos errores, aunque en Estadística se da prioridad al error de tipo II (es mejor dejar libre a un culpable que condenar a un inocente). Reducir ambos errores a cero es imposible y, además, al disminuir uno se aumenta el otro, por lo que hay que elegir bien el test empleado para llegar al mejor balance posible entre el error de tipo I y el error de tipo II. La única manera de disminuir ambos es aumentar el tamaño de muestra. En los siguientes apartados verás cómo los intentos de minimizar cada tipo de error tienen su papel en el contraste de hipótesis.

4. CONTRASTE DE HIPÓTESIS: CÓMO FUNCIONA UN TEST

Para emitir sentencias sobre diferencias entre medias procedentes de muestras se utilizan diversos distintos procedimientos de cálculo, o "tests" estadísticos. Puedes considerar un test como un juez que decide. Pero ¿cómo funciona en la práctica?

Los tests se basan en el cálculo de una **distribución de muestras** o de diferencias entre muestras. Éstas son distribuciones teóricas compuestas de todas las diferencias posibles entre medias del mismo tamaño que la muestra que tú has tomado, bajo la hipótesis nula de que no existen diferencias "reales". También pueden interpretarse como distribuciones de las probabilidades de obtener el valor de la media que efectivamente has encontrado, suponiendo que la hipótesis nula es cierta. La manera en que los tests generan esas distribuciones teóricas varía en cada caso y a veces tienen fórmulas complicadas. Un ejemplo sencillo corresponde a las probabilidades de obtener un seis al lanzar un dado. Si tiras el dado un número alto de veces obtendrás varios resultados que, en promedio (si el dado no está trucado), deberían indicar que esa probabilidad es $\frac{1}{6}$. Al final puedes hacer una tabla donde figuran cada resultado y la probabilidad correspondiente al mismo. Los libros de Estadística incluyen tablas para numerosos tests.

Sabes que has tomado unas muestras y que, incluso si pertenecen a la misma población (estadística) existirá una diferencia entre ellas. Esa diferencia puede ser mayor o menor. Los tests estadísticos comparan esas diferencias y evalúan si es posible que ocurran procediendo de la misma población. El **nivel de significación (α)** es la probabilidad límite que se elige para rechazar la hipótesis nula. En otras palabras, establece el grado de diferencia entre muestras que estás dispuesto a aceptar antes de decidir que esa diferencia es demasiado extrema y no puede obtenerse si ambas muestras proceden de la misma población. La obtención de una diferencia mayor sería algo tan improbable que, si ocurre, lo más probable es que eso indique que ambas medias proceden de poblaciones (estadísticas) diferentes. Normalmente se fija $\alpha = 0.05$, pero a veces conviene tomar un valor mayor (0,1) o menor (0,01). Esto simplemente significa que si la diferencia observada entre muestras tiene una probabilidad de ocurrir del 5% (o el 10% o el 1%) o menos no puede deberse a meros errores de muestreo sino que indica diferencias reales entre muestras.

La **región de rechazo** es la parte de la distribución de muestreo del test cuya probabilidad es menor o igual que α , es decir, el nivel de significación. Al realizar un test estadístico lo que haces es comparar la diferencia que has obtenido entre muestras con la distribución de muestreo teórica generada por el test, utilizando las tablas disponibles. Si tu diferencia cae dentro de la región de rechazo, concluyes que la hipótesis nula debe ser rechazada, la hipótesis alternativa aceptada y las diferencias son significativas.

4. CONTRASTE DE HIPÓTESIS: ELECCIÓN DEL TEST ADECUADO

El valor de α equivale al error de tipo I que cometes en tu juicio. Obtener diferencias mayores que α es muy improbable, pero no imposible. Así que en ocasiones eso ocurrirá. ¿Con qué frecuencia? puedes preguntarte angustiado. Eso dependerá del valor de α utilizado. Cuanto menor sea α , menor será la probabilidad de equivocarte, pero más difícil será rechazar la hipótesis nula; cuanto mayor sea α , mayor cantidad de resultados posibles se condicionarán "improbables" y la hipótesis nula se rechazará más a menudo, pero la probabilidad de hacerlo de modo erróneo aumentará. En otras palabras, existe un balance entre los errores de tipo I y de tipo II.

Existen diferentes medios para decidir sobre la aceptación o rechazo de la hipótesis nula. Esos medios son los distintos tests, que difieren tanto en su posibilidad de error de tipo I como de error de tipo II. Por tanto, es importante elegir el test adecuado para cada caso concreto. En primer lugar te interesa tener en cuenta (a) el tipo de datos (variables continuas, discretas, rangos o atributos) y (b) el número de muestras que se desean comparar (dos muestras o más de dos muestras).

Después de ello aún hay diversas opciones. En este momento es cuando interesa tener en cuenta la probabilidad de error de tipo II. Los diferentes tests difieren en esa probabilidad de error de tipo II. La **potencia** de un test es su capacidad de rechazar la hipótesis nula cuando es falsa (de condenar culpables). Esto equivale a la inversa del error de tipo II. Obviamente te conviene elegir el test más potente dentro de los disponibles en la situación concreta que estés estudiando. Cual será ese test depende del tipo de datos tomados y del cumplimiento de una serie de requisitos en los mismos, como se explica en el apartado siguiente.

4. CONTRASTE DE HIPÓTESIS: TESTS PARAMÉTRICOS Y NO PARAMÉTRICOS

Para calcular la distribución de muestreo de algunos tests hay que hacer ciertos supuestos sobre los datos:

1. los datos corresponden a una variable continua,
2. la variable sigue una distribución normal,
3. las variancias de los subgrupos son similares u homogéneas,
4. los datos son independientes.

Esto es lo que se denominan **tests paramétricos**. Cuando se cumplen esos supuestos, son muy potentes. Sin embargo, si esos supuestos no se cumplen, los resultados que producen no serán fiables. Por eso, antes de realizar un test, deben comprobarse que esas suposiciones son ciertas. Esto se realiza mediante otro grupo de tests (tranquilo, no hay que comprobar requisitos para estos tests mediante otros tests y así hasta el infinito).

Otros tests no requieren los supuestos 1 a 3. Estos son los **tests no paramétricos** y son casi o igual de potentes que los tests paramétricos en esas situaciones en que no los datos no cumplen los supuestos 1 a 3. Además, son más **robustos**, es decir, cuando muestran diferencias significativas, están mejor fundadas. También requieren tamaños de muestra menores.

Cómo elegir entre tests paramétricos y tests no paramétricos

1. Cuando el tamaño de muestra es muy bajo (< 10), son preferibles los tests no paramétricos. Esto no es una panacea; por debajo de cierto número de datos, ningún test estadístico es posible.
2. Comprueba el cumplimiento de los requisitos para los tests paramétricos:
 - (a) normalidad, mediante un **test de Kolmogorov-Smirnov**,
 - (b) homogeneidad de variancias, mediante el **test de Bartlet** u otro,
3. Si se cumplen los requisitos, usa un test paramétrico; si no se cumplen, usa un test no paramétrico.

Como resumen de todo el proceso necesario para el contraste de hipótesis echa un vistazo a la Fig. 5.

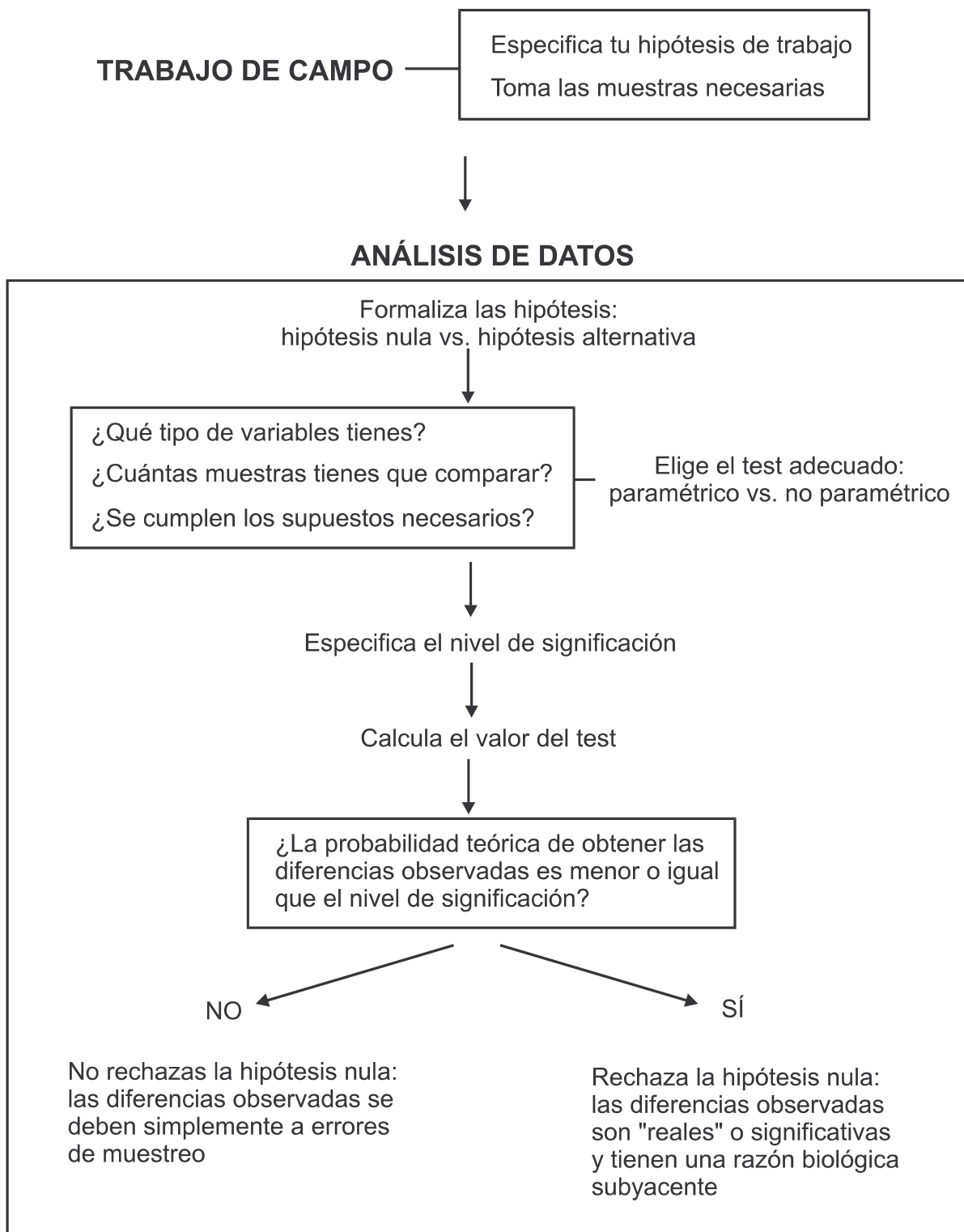


Fig. 5. Pasos a seguir en el contraste de hipótesis. Para más detalles, no te quedará más remedio que releer el texto.

4. CONTRASTE DE HIPÓTESIS: COMPARANDO DOS GRUPOS

Supongamos que quieres comparar dos grupos de datos, por ejemplo el valor medio de la longitud del pico en machos y hembras de Combatiente. En este caso los datos son independientes, pues los grupos de datos a comparar se han obtenido por el muestreo de individuos distintos. De modo indicativo, el procedimiento a seguir es el siguiente.

1. Para cada grupo por separado, comprueba si los datos siguen una distribución normal, por medio de un **test de Kolmogorov-Smirnov**.
2. Comprueba la igualdad de variancias entre ambos grupos, por medio de un **test de Bartlett**.
3. Si se cumplen los requisitos de normalidad e igualdad de variancias, puedes comparar las medias mediante uno de estos dos tests:
 - (a) un **test de la t**,
 - (b) un análisis de variancia o **ANOVA**, de idéntico resultado pero más potente.
4. Si los datos son normales, pero las variancias son heterogéneas,
 - (a) utiliza una **transformación de los datos** (logarítmica, raíz cuadrada o inversa), comprueba si las variancias se vuelven homogéneas y en caso afirmativo utiliza un ANOVA,
 - (b) utiliza el **test de la t aproximado de Welch**, que no asume igualdad de variancias,
 - (c) utiliza un **test de la U de Mann-Whitney**, que no es paramétrico.
5. Si los datos no son normales, pero las variancias son homogéneas,
 - (a) utiliza una transformación de los datos (logarítmica, raíz cuadrada o inversa), comprueba que si datos se vuelven normales y en caso afirmativo utiliza un ANOVA. El ANOVA es robusto ante la falta de normalidad, pero si los datos no son normales conviene usar otro test,
 - (b) utiliza un test de la U de Mann-Whitney, que no es paramétrico.
6. Si los datos no son normales ni las variancias son homogéneas,
 - (a) intenta una transformación de los datos, como ya se ha dicho en los puntos previos,
 - (b) utiliza un test de la U de Mann-Whitney, que no es paramétrico.

Si los datos no son independientes, es decir, los grupos a comparar provienen del muestreo de los mismos individuos (p. ej. el tamaño de puesta entre dos puestas sucesivas de una misma hembra), se procede del siguiente modo.

1. Si los datos son normales y las variancias homogéneas, puedes utilizar un **ANOVA de medidas repetidas**.
2. Si la diferencia entre cada par de valores sigue una distribución normal, puedes utilizar un **test de la t de medidas repetidas**. Este test requiere comprobar menos requisitos que el anterior.
3. Si no se cumplen los requisitos paramétricos, usa un **test de Wilcoxon de rangos con signo**. Hay otros tests no paramétricos para comparar dos grupos de datos no independientes, pero son menos potentes.

4. CONTRASTE DE HIPÓTESIS: COMPARANDO MÁS DE DOS GRUPOS

Cuando se trata de comparar más dos grupos de datos, como el crecimiento de pollos a tres tasas de cebs diferentes, la cantidad de posibilidades de análisis aumenta. Aquí trataré sólo las más simples, para datos independientes y una clasificación sencilla de los grupos (es decir, los grupos se constituyen en base a un sólo factor; en el ejemplo, la tasa de cebs). En los restantes casos, y como con el uso de la Estadística en general, es necesario consultar los libros más especializados. El procedimiento indicativo a seguir es el siguiente.

1. Para cada grupo por separado, comprueba si los datos siguen una distribución normal, por medio de un **test de Kolmogorov-Smirnov**.
2. Comprueba la igualdad de variancias entre grupos, por medio de un **test de Bartlett**.
3. Si se cumplen los requisitos de normalidad e igualdad de variancias compara las medias mediante un **ANOVA** (el test de la t no es aplicable a más de dos grupos).
 - (a) Si las diferencias entre grupos son significativas, utiliza un test a posteriori (hay muchos) para descubrir qué grupo o grupos son los responsables de dichas diferencias.
4. Si los datos son normales, pero las variancias son heterogéneas,
 - (a) utiliza una **transformación de los datos** (logarítmica, raíz cuadrada o inversa), comprueba si las variancias se vuelven homogéneas y en caso afirmativo utiliza un ANOVA,
 - (b) transforma los datos en forma de **rangos** y realiza el ANOVA sobre los mismos.
5. Si los datos no son normales, pero las variancias son homogéneas,
 - (a) utiliza una transformación de los datos (logarítmica, raíz cuadrada o inversa), comprueba si los datos son normales y en caso afirmativo utiliza un ANOVA. El ANOVA es robusto ante la falta de normalidad, pero si los datos no son normales conviene usar otro test,
 - (b) utiliza un **test de Kruskal-Wallis**, que no es paramétrico. Este test requiere igualdad de variancias, por lo que no es aplicable al caso (4) ni al (6) (aunque al parecer es bastante robusto ante la heterogeneidad de variancias).
 - (i) Si las diferencias entre grupos son significativas, utiliza un **test a posteriori de Dunn-Sidák** para descubrir qué grupo o grupos son los responsables de dichas diferencias.
6. Si los datos no son normales ni las variancias son homogéneas,
 - (a) intenta una transformación de los datos, como ya se ha dicho en los puntos previos,
 - (b) transforma los datos en forma de **rangos** y realiza el ANOVA sobre los mismos.

5. ASOCIACIÓN ENTRE VARIABLES: POR QUÉ ES ÚTIL

Muchas cuestiones que se plantean en Ornitología requieren establecer la existencia de relaciones entre dos o más variables. ¿Aumenta la densidad de aves de una especie al aumentar la cantidad de alimento disponible? ¿Varía el éxito de las polladas con la altitud a la que se encuentra el nido? ¿Aumenta la mortalidad de una especie al disminuir la temperatura invernal? Una forma rápida de examinar esa **asociación** entre variables consiste en representar una frente a otra mediante un diagrama de puntos y examinar visualmente si existe alguna tendencia de las variables a variar conjuntamente. Esa relación puede ser **positiva** (al aumentar los valores de una variable, aumentan también los de la otra, Fig. 6 A) o **negativa** (al aumentar los valores de una variable, los de la otra disminuyen, Fig. 6 B).

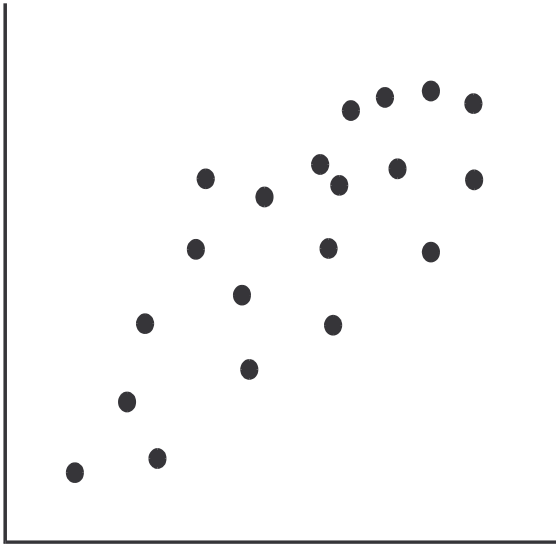
En una asociación positiva, por ejemplo, una mentalidad determinista esperaría que si aumenta el valor de la variable en el eje horizontal, o eje X, también debería aumentar el valor de la variable en el eje vertical, o eje Y (Fig. 6 D). Pero como a estas alturas ya habrás podido imaginar, la cosa no es tan fácil. De nuevo, la variación presente en los datos puede hacer difícil descubrir o interpretar esas asociaciones. La representación gráfica puede producir simplemente una "nube de puntos" (Fig. 6 C) o revelar tendencias aproximadas (Fig. 6 A y B). En las situaciones ilustradas en las figuras 6 A, B y C, una mentalidad determinista no saldría de su asombro, pues valores cada vez mayores de la variable en el eje X van ligados a valores a veces mayores, pero a veces menores, de la variable en el eje Y. Es necesario tener una mentalidad probabilista para poder ver las tendencias más o menos ocultas por la variación.

En el estudio de la asociación entre variables, la Estadística nos permite dos cosas:

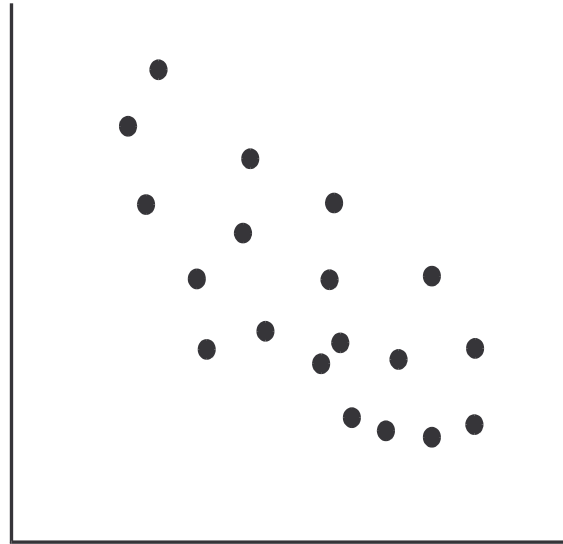
1) decidir si esa asociación entre variables es o no "real", es decir, si se debe a causas tales como errores de medida, variaciones derivadas de la medida de individuos diferentes de una misma población, uso de muestras, etc., o se debe a una relación biológicamente relevante entre las variables. Esto se consigue mediante la asignación de un valor de probabilidad para la asociación encontrada,

2) asignar una "fuerza" o un "grado" a la asociación encontrada. Esto es interesante porque, además de saber si la asociación existe, conviene saber en qué punto del continuo entre las situaciones representadas en la Fig. 6 C y D se encuentra. El grado de asociación se expresa mediante un **coeficiente**, que varía entre -1 (relación negativa perfecta) y 1 (relación positiva perfecta). Un valor del coeficiente de 0 indica que no existe ningún tipo de asociación.

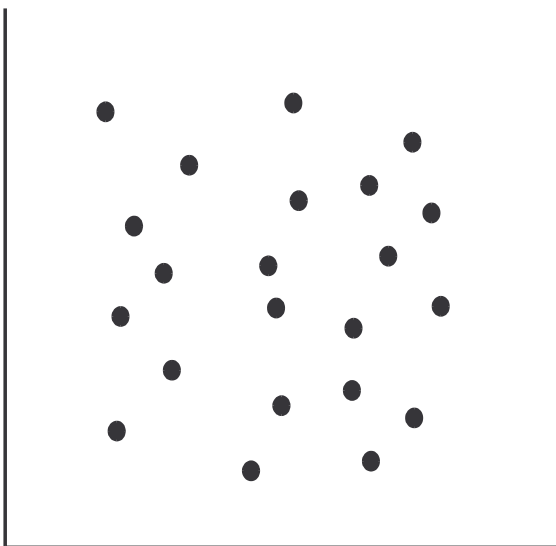
A)



B)



C)



D)

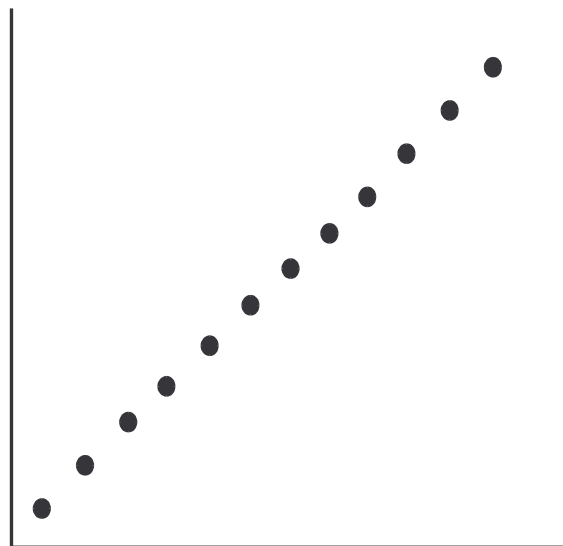


Fig. 6. Cuatro ejemplos de asociación entre variables: A) positiva, B) negativa, C) ausente ("nube de puntos") y D) perfecta (positiva).

5. ASOCIACIÓN ENTRE VARIABLES: CORRELACIÓN Y REGRESIÓN

La asociación entre dos variables mensurables u ordinales puede estudiarse por dos procedimientos: correlación o regresión. Es muy importante tener clara la diferencia entre ambos, porque aunque comparten algunos de los métodos estadísticos de cálculo y tienen representaciones gráficas casi idénticas, su significado biológico es completamente diferente.

La **correlación** te dice si existe una asociación real (significativa) entre dos variables y mide el grado de asociación utilizando el **coeficiente de correlación (r)** que, como ya sabes, varía entre -1 y +1. La correlación **NO** implica que la variación en una variable sea la causa de la variación en la otra. Por ejemplo, si mides la longitud del ala y la longitud del tarso en una especie cualquiera de ave, encontrarás que ambas están asociadas positivamente: individuos con mayor longitud del tarso tendrán generalmente alas de mayor longitud. Pero eso no significa que la variación en la longitud del tarso provoque la variación en la longitud del ala ni viceversa. Eso se comprueba fácilmente si piensas que cortar el ala de un ave no causará ninguna variación en la longitud de su tarso. En realidad, ambas variables son expresión de algo más general, el tamaño del ave, que causa las variaciones conjuntas en ambas variables. En la correlación, el diagrama de puntos puede hacerse indistintamente con cualquiera de las variables en el eje X.

En cambio, la **regresión** se utiliza cuando te interesa predecir los valores de una variable en función de los valores de otra variable relacionada con ella. Se considera la variable cuyo valor se quiere predecir como **dependiente** y la variable que se quiere usar para predecirla se denomina **independiente**. La representación gráfica siempre sitúa la variable independiente en el eje X y la variable dependiente en el eje Y, y ese orden no puede invertirse. La asociación entre variable dependiente e independiente puede ser auténticamente *causal* (p. ej., entre tasa de ceba y crecimiento de los pollos) o simplemente *funcional* (p. ej. entre envergadura alar y peso). En este caso también se obtiene un coeficiente r, que mide la fuerza de la asociación. Además se obtienen otros dos índices que indican la relación funcional entre las dos variables. El primero es el **coeficiente de determinación (R^2)** y mide la cantidad de variación en la variable dependiente que es explicada por la variación en la variable independiente. Está comprendido entre 0 (no explica ninguna variación) y 1 (la explica toda). El segundo es la **recta de regresión**, que es una ecuación gracias a la cuál puede saberse, dado un valor de la variable independiente, el valor correspondiente de la variable dependiente. La representación gráfica de una regresión suele incluir esa recta de regresión, que **nunca** debes emplear cuando haces una correlación.

Tanto en la correlación como en la regresión, es posible relacionar simplemente dos variables, o muchas más simultáneamente. En este último caso se habla de **correlación o regresión múltiple**. En la regresión múltiple, existe una sola variable dependiente y lo que se examina es la variación presente en la misma que explican las diversas variables independientes.

5. ASOCIACIÓN ENTRE VARIABLES: TESTS PARAMÉTRICOS Y NO PARAMÉTRICOS

Como en los tests de hipótesis, el uso de la correlación y la regresión depende del cumplimiento de una serie de requisitos, cuyas violaciones tienen consecuencias diversas. Al ser un aspecto más complejo del estudio de la asociación entre variables, se tratará aquí de modo muy preliminar y sin dar todas las explicaciones necesarias, que serían demasiado largas y técnicas. No obstante, debe ser mencionado para no dar una falsa impresión de que los tests pueden usarse en cualquier circunstancia. Al contrario, además del tipo de variables y de datos, el ajuste a los requisitos es una parte fundamental en la elección del test adecuado.

Los requisitos necesarios para poder realizar un análisis de regresión paramétrico son: a) las medidas son independientes, b) la relación entre las variables es lineal, no curvilínea, c) los valores de la variable independiente son fijados por el observador, d) la variable independiente se mide sin error y e) la variación en la variable dependiente es similar en todo el rango de valores de la variable independiente. Los requisitos (a) y (c) se cumplirán si la toma de datos en el campo ha sido planificada de modo correcto. Las condiciones (b) y (e) pueden contrastarse examinando los residuos de la regresión, es decir, la diferencia entre los valores observados de la variable dependiente y los predichos de acuerdo con la recta de regresión. Si la condición (b) no se cumple, es posible transformar los datos para hacer que la relación se convierta en lineal. Si las condiciones (c), (d) y (e) no se cumplen, puede utilizarse un modelo II de regresión. Violaciones del resto de los supuestos hacen imposible el análisis. Existe un método no paramétrico de regresión, el método robusto de Kendall de ajuste de rectas, pero no es muy utilizado.

Los requisitos necesarios para poder realizar un análisis de correlación paramétrico son: a) la relación entre las variables es lineal, y b) ambas variables siguen una distribución normal. La normalidad de las variables se comprueba mediante un test de Kolmogorov-Smirnov, como se explicó para los tests de hipótesis. Puede recurrirse a la transformación de las variables, si se viola el supuesto (a). En el resto de los casos, lo mejor es recurrir a un test no paramétrico.

El método de correlación no paramétrico más habitual es el **coeficiente de correlación de Spearman (r_s)**. Este método simplemente evalúa si los valores de la variable dependiente aumentan o disminuyen con cambios en la variable independiente. Puede usarse como sustituto del análisis de regresión paramétrico, pero no proporciona ninguna ecuación para predecir los valores de la variable dependiente. Puede usarse no sólo para variables mensurables sino también para variables ordinales.

La **correlación de Kendall** es también un método no paramétrico de correlación. Puede utilizarse para examinar la asociación entre dos variables, mensurables u ordinales. Pero, a diferencia de la correlación de Spearman, permite el examen no paramétrico de la correlación entre más de dos variables.

5. ASOCIACIÓN ENTRE VARIABLES: TABLAS DE CONTINGENCIA

Las **tablas de contingencia** se utilizan para medir la asociación entre dos atributos. Cumplen un papel análogo a las otras medidas de asociación entre variables (correlación y regresión) pero también sirven como contraste de hipótesis, dado que en las variables medidas como atributos ambos tipos de análisis no se diferencian claramente.

El nombre de "tablas" deriva del tipo de representación que suele hacerse para el análisis. Se construye una tabla con tantas filas como niveles tiene una de las variables y tantas columnas como niveles toma la otra variable. En el ejemplo de la Fig. 7 cada variable tiene sólo dos niveles, pero eso no es obligatorio; se habla de tablas 2 X 2 si cada variable tiene dos niveles, de tablas 2 X 4 si una tiene dos niveles y la otra cuatro o, en general, de tablas n X m. En cada casilla de la tabla se escribe la frecuencia observada de individuos dentro de cada combinación de niveles. El análisis consiste en averiguar si el reparto de los individuos en las diferentes casillas sigue algún patrón o es simplemente debido al azar (Fig. 7 A y B). Habrá una asociación entre variables si determinadas combinaciones de las mismas aparecen con una frecuencia estadísticamente mayor que otras. Se obtiene un valor de significación para la asociación por medio de un **test Chi-cuadrado** (χ^2) o un **test de la G**. También es posible calcular la "fuerza" de la misma, mediante el **coeficiente de Cramér (C)** o el **coeficiente Phi de asociación (r_ϕ)** (sólo en tablas 2 X 2). Estos coeficientes varían entre 0 (ninguna asociación) y 1 (asociación perfecta).

Como en la correlación y la regresión, es posible examinar la asociación entre más de dos variables simultáneamente, mediante el uso de tablas multidimensionales y el uso de **modelos loglineares**.

		variable 1	
		1	2
variable 2	a	0	10
	b	10	0

		variable 1	
		1	2
variable 2	a	5	5
	b	5	5

Fig. 7. Ejemplo de asociación entre dos atributos para los que se han medido 20 valores: A) asociación perfecta (todos los a son 2 y todos los b son 1), B) asociación inexistente (los a son 1 ó 2 con idéntica frecuencia, y lo mismo sucede con los b).

5. ASOCIACIÓN ENTRE VARIABLES: GUÍA DE UTILIZACIÓN

A continuación se resumen los pasos a seguir para el estudio de la asociación entre variables.

1. Asegúrate de que lo que te interesa es la asociación entre variables y no un test de hipótesis.
2. Si las variables son mensurables, ordinales o una combinación de estos tipos:
 - (a) decide si te interesa la asociación (correlación) o la relación funcional (regresión),
 - (b) comprueba los supuestos necesarios (independencia, normalidad: test de Kolmogorov-Smirnov, linealidad de la relación y variación homogénea de la variable Y en todos los valores de la variable X: examen de los residuos),
 - (c) decide el test a utilizar en función del número de variables y del cumplimiento de los supuestos.

Supuestos

paramétricos	Correlación	Regresión
	<i>Dos variables</i>	<i>Dos variables</i>
Se cumplen	Coeficiente de correlación de Pearson	Coeficiente de determinación
No se cumplen	Coeficiente de correlación de Spearman Correlación de Kendall	Regresión modelo II Método robusto de Kendall de ajuste de rectas Utiliza correlación no paramétrica
	<i>Más de dos variables</i>	<i>Más de dos variables</i>
Se cumplen	Correlación múltiple o parcial	Regresión múltiple
No se cumplen	Correlación parcial de Kendall	No existe ningún test

3. Si las variables son atributos o una combinación de atributos y variables ordinales:
 - (a) no deben existir casillas vacías (si existen, agrupa niveles hasta evitarlo, o emplea un test exacto de Fisher para tablas 2 x 2),
 - (b) no deben existir frecuencias esperadas menores que 5 en más del 20% de las casillas,
 - (c) si tienes dos variables y es una tabla 2 X 2,
 - (i) tienes al menos 25 datos: utiliza un test χ^2 o, preferentemente, un test de la G,
 - ◆ entre 25 y 200 datos, emplea la corrección para continuidad,
 - ◇ más de 200 datos, no es necesaria la corrección para continuidad,
 - (ii) tienes menos de 25 datos: utiliza un test exacto de Fisher,
 - (b) tienes dos variables y es una tabla 2 X m o n X m: test χ^2 o, preferentemente, test de la G,
 - (c) utiliza modelos loglineares para el caso de más de dos variables.

6. ESTADÍSTICA Y ORDENADORES

Aunque algunos de los métodos estadísticos pueden hacerse fácilmente a mano, en muchos casos la Estadística requiere el uso de ordenadores y de programas especiales. Esos programas son de dos tipos, básicamente. En primer lugar se necesita una "hoja de cálculo" en la que introducir los datos. Las hojas de cálculo son un análogo informático de una libreta de campo; permiten almacenar los datos en un forma de filas y columnas, donde las columnas indican las variables y las filas indican las observaciones o unidades de estudio. Las hojas de cálculo más utilizadas son *Excel* de Microsoft y *Lotus 1-2-3* de Lotus, para ordenadores que trabajan en entorno Windows. Existen programas similares para ordenadores de entorno MacIntosh. Con estas hojas de cálculo puede realizarse estadística descriptiva (tablas, gráficas, medias, desviaciones típicas, etc.), algunos test de hipótesis sencillos y medidas de asociación entre variables.

El segundo tipo de programas que se necesita son programas estadísticos especializados. En muchos casos las hojas de cálculo se quedan cortas en sus posibilidades de análisis estadístico. Los programas estadísticos especializados incluyen muchas más opciones y tipos de tests y permiten resolver prácticamente cualquier problema estadístico que uno pueda plantearse. Los programas estadísticos suelen tener una opción para introducir los datos, a modo de hoja de cálculo. También ofrecen posibilidades de representación gráfica. Y además cuentan con toda una gama de tests estadísticos para contraste de hipótesis y medidas de asociación entre variables. Existen numerosos programas; quizá los más populares son: *SPSS-PC*, *SAS*, *Minitab*, *Statistica* y *Statgraphics*. Como opción libre, descargable desde Internet, está *R*, que es el más completo y recomendable, pero que tiene el inconveniente de que hay que teclear las instrucciones "a mano" pues no usa el sistema de "ventanitas" clásico de Windows.

7. MÁS ALLÁ DE ESTE MANUAL: AUTOAPRENDIZAJE DE LA ESTADÍSTICA

Este manual no es una panacea, es simplemente una introducción. Para comprender y aprender realmente la Estadística se requiere un esfuerzo adicional; *hay que estudiarla*. Y eso supone irse a los libros y leer atentamente la base y el procedimiento para realizar cada test. En la sección siguiente viene una selección de libros sobre Estadística. Dentro del apartado *Manuales generales* incluyo una puntuación de los libros en función de su sencillez (* sencillo, ** nivel medio, *** complicado o especializado). No he puntuado los libros de los que tengo poco conocimiento. Personalmente, recomiendo pasar rápidamente de los libros de nivel * a los de nivel ** y consultar sólo para casos concretos los de nivel ***.

Mi recomendación personal es utilizar Fowler & Cohen (1995) o su versión en castellano (1999) para hacerse una idea inicial sobre las posibilidades de la Estadística, y complementar dicho libro con Dytham (1999), que proporciona una guía muy completa y sencilla (aunque en inglés) sobre cómo elegir los tests estadísticos, y con McKillup (2005), muy didáctico para los fundamentos de los tests estadísticos. Además, Dytham (1999) proporciona instrucciones sobre cómo realizar los tests estadísticos utilizando algunas de las hojas de cálculo y programas estadísticos más comunes. Una vez que uno tiene una base estadística media, lo mejor es utilizar manuales como Sokal & Rohlf (1994), Zar (1999) o Quinn & Keough (2002). Estos tres manuales permiten afrontar el 99% de los análisis estadísticos que uno necesita. Recomiendo especialmente Quinn & Keough (2002); parecía imposible escribir un libro más sobre estadística que no repitiese simplemente lo ya dicho, pero Quinn & Keough no solo lo logran sino que además alcanzan un nivel pedagógico altísimo. Gotelli & Ellison (2004) es un buen complemento. Sólo en casos puntuales uno necesitará consultar bibliografía especializada o artículos concretos.

Subrayo la necesidad de *estudiar* la Estadística. Cuando uno se enfrenta a la situación de tener que aprender un nuevo test, debe leer cuidadosamente el capítulo correspondiente **con lápiz y papel** para tomar las notas necesarias. Es la única manera de dominar la base de dicho test. Eso requiere cierto esfuerzo, pero no está fuera del alcance de nadie.

¡Cursos de Estadística en Internet!

Las ciencias adelantan que es una barbaridad. Ahora puedes encontrar en Internet cursillos introductorios sobre Estadística. Recomiendo visitar la siguiente dirección, que incluye una serie de cursillos:

<http://colargol.ibg.uit.no/biologi/botanikk/lennart/bot-frame.htm>

Finalmente, una nota de advertencia... aquellos que os escudais en la falta de bibliografía en castellano para no ir a los libros de estadística cada vez lo teneis más chungo; ya he visto hasta una "Estadística en cómic", en castellano (aunque la verdad es que no parecía muy buena).

8. BIBLIOGRAFÍA RECOMENDADA

No tengo todos los libros, pero sí los artículos que menciono en la bibliografía. Si no puedes conseguirlos, mándame un mensaje por correo electrónico:

marcos.mendez@urjc.es

También puedes escribirme para señalarme errores o sugerirme mejoras a este manual.

Libros de iniciación

- Dytham, C. (1999). *Choosing and using statistics. A biologist's guide*. Blackwell Science, Oxford. 218 pp. (*)
- Fowler, J.; Cohen, L. (1991). *Practical statistics for field biology*. Wiley, Nueva York. 224 pp. (*)
- Fowler, J.; Cohen, L. (1999). *Estadística básica en ornitología*. SEO, Madrid. 144 pp. (*)
- McKillup, S. (2005). *Statistics explained: an introductory guide for life scientists*. Cambridge University Press, Cambridge. 267 pp. (*)

Manuales generales

- Gotelli, N. J.; Ellison, A. M. (2004). *A primer of ecological statistics*. Sinauer Associates, Sunderland, MA. 510 pp. (**)
- Mead, R.; Curnow, R. N. (1983). *Statistical methods in agriculture and experimental biology*. Chapman & Hall, Londres. 335 pp.
- Quinn, G. P.; Keough, M. J. (2002). *Experimental design and data analysis for biologists*. Cambridge University Press, Cambridge. 537 pp. (**)
- Sokal, R. R.; Rohlf, F. J. (1994). *Biometry: The principles and practice of statistics in biological research*. 3ª ed. W. H. Freeman & co., Nueva York. 887 pp. (Traducción española de la 1ª ed.: [1979] *Biometría. Principios y métodos estadísticos en la investigación biológica*. H. Blume, Madrid. 832 pp.) (**)
- Underwood, A. J. (1997). *Experiments in ecology. Their logical design and interpretation using analysis of variance*. Cambridge University Press, Cambridge. 504 pp. (**)
- Zar, J. H. (1999). *Biostatistical analysis*. 4ª ed. Prentice-Hall, Englewood Cliffs. 929 pp. (**)

Manuales avanzados

- Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. (1993). *Classification and regression trees*. Chapman & Hall/CRC, Boca Ratón, FL. 358 pp.
- Draper, N. R.; Smith, H. (1998). *Applied regression analysis*, 3ª ed. Wiley and Sons, Nueva York. 706 pp.
- Hosmer, D. W.; Lemeshow, S. (2000). *Applied logistic regression*, 2ª ed. Wiley and Sons, Nueva York. 375 pp.
- Legendre, P.; Legendre, L. (1998). *Numerical ecology*, 2ª ed. Elsevier, Amsterdam. 853 pp.
- Lepš, J.; Šmilauer, P. (2003). *Multivariate analysis of ecological data using CANOCO*. Cambridge University Press, Cambridge. 269 pp. (**)
- McCullagh, P.; Nelder, J. A. (1989). *Generalized linear models*, 2ª ed. Chapman & Hall/CRC, Boca Ratón, FL. 511 pp.
- McCune, B.; Grace, J. B. (2002). *Analysis of ecological communities*. MjM Software Design, Gleneden Beach, OR. 300 pp. (*)

- Ruiz-Maya Pérez, L. (coord.) (1990). *Metodología estadística para el análisis de datos cualitativos*. CIS, Madrid. 382 pp. (**)
- Scheiner, S. M.; Gurevitch, J. (2001). *Design and analysis of ecological experiments*, 2^a ed. Oxford University Press, Oxford. 415 pp. (**)
- Shipley, B. (2000). *Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference*. Cambridge University Press, Cambridge. 317 pp.
- Siegel, S.; Castellan, N. J. Jr (1988). *Nonparametric statistics for the behavioral sciences*. 2^a ed. McGraw-Hill, Nueva York. 399 pp. (**)

Artículos avanzados

- Day, R. W.; Quinn, G. P. (1989). Comparisons of treatments after an analysis of variance in ecology. *Ecological Monographs* 59: 433-463.
- Fienberg, S. E. (1970). The analysis of multidimensional contingency tables. *Ecology* 51: 419-433.
- Jackson, D. A.; Somers, K. M. (1991). The spectre of "spurious" correlations. *Oecologia* 86: 147-151.
- Matson, P.; Potvin, C.; Travis, J. (eds.) (1993). Statistical methods: an upgrade for ecologists. *Ecology* 74: 1615-1676.
- McArdle, B. H. (1988). The structural relationship: regression in biology. *Canadian Journal of Zoology* 66: 2329-2339.
- Mitchell, R. J. (1992). Testing evolutionary and ecological hypotheses using path analysis and structural equation modelling. *Functional Ecology* 6: 123-129.
- Parkhurst, D. F. (2002). Statistical significance tests: equivalence and reverse tests should reduce misinterpretation. *BioScience* 51: 1051-1057.
- Rice, W. R. (1989). Analyzing tables of statistical tests. *Evolution* 43: 223-225.
- Rice, W. R.; Gaines, S. D. (1994). 'Heads I win, tails you lose': testing directional alternative hypotheses in ecological and evolutionary research. *Trends in Ecology and Evolution*. 9: 235-237.
- Seaman, J. W. Jr; Walls, S. C.; Wise, S. E.; Jaeger, R. G. (1994). Caveat emptor: rank transform methods and interaction. *Trends in Ecology and Evolution* 9: 261-263.
- Scheiner, S. M.; Gurevitch, J. (eds.) (1993). *Design and analysis of ecological experiments*. Chapman & Hall, Nueva York. 445 pp. (**)
- Tufte, E. R. (1990). Data-ink maximization and graphical design. *Oikos* 58: 130-144.
- Vepsäläinen, K.; Savolainen, R.; Penttinen, A. (1988). Causal reasoning in modelling multiway contingency tables. *Oikos* 53: 281-285.