Estadística básica*

Matías Raja

Universidad de Murcia

1. Introducción

Para una gran mayoría de la gente la Estadística es un arte de hacer pronósticos, casi siempre menos acertados que la predicción meteorológica cuando se trata de Murcia, sobre asuntos que involucran la voluntad de un gran número de personas, como pueden ser unas elecciones, la opinión sobre un tema de actualidad o un estudio de mercado para saber si un determinado producto se vendería bien. Lo cierto es que la Estadística tiene mucho más alcance y se emplea para obtener información de conjuntos de datos, validar o rechazar hipótesis a partir de ellos. Más detalladamente daremos tres razones inherentes a la naturaleza y alcance de los métodos estadísticos y dos razones adicionales más relacionas con el uso actual que se hace de ellos.

1. **Describir**. La Estadística proporciona el soporte teórico para la toma sistemática de datos, asegurando precisión y objetividad. Los análisis posteriores podrían recomendar una selección optima de variables a medir, mejorando el valor descriptivo de los datos.

2. **Informar**. Los métodos estadísticos permiten un análisis inicial de la estructura de los datos, la elaboración de resúmenes numéricos y gráficos para condensar los aspectos esenciales con poca pérdida de información.

3. **Inferir**. Métodos de Estadística más avanzados permiten descubrir relaciones no triviales entre las variables o justificar la ausencia de ellas, poner a prueba las hipótesis formuladas dotando de objetividad a las conclusiones, cuantificando la verosimilitud de las afirmaciones.

*Notas de clase del **Máster de Universitario en Historia y Patrimonio Histórico**

4. **Tendencia**. Hay un cierto consenso en que el método científico, tal como se entiende en ciencias experimentales, debe aplicarse, en la medida que sea posible, en cualquier disciplina que implique generar conocimiento a partir de un conjunto de datos. A veces un grupo de investigación pionero aplica una determinada técnica con éxito (*seminal paper*) y con el tiempo se convierte en estándar.

5. **Factibilidad**. Ciertos análisis involucran un importante volumen de cálculo o software especializado. Cada año los ordenadores son más potentes y hay más disponibilidad de software libre para análisis de grandes volúmenes de datos.

En las próximas secciones trataremos de cubrir los primeros tres puntos de manera muy general. Para la selección de las técnicas entre las muchas disponibles prestaremos especialmente a las que son tendencia en Arqueología y para ilustrar la parte numérica usaremos el software libre **R**[1].

Las técnicas estadísticas deben formar parte de la "caja de herramientas" del profesional que maneja datos. Es importante saber lo que se puede hacer con ellas, pero es casi más importante conocer sus limitaciones. Para ello hay que ir un poco más lejos del nivel de usuario y mirar dentro de la *caja negra* de la Estadística. En esta breve discusión alrededor de algunos métodos estadísticos intentaremos dar ese punto de vista esquivando los detalles matemáticos más difíciles. Sabiendo los fundamentos que hay detrás de los algoritmos del software estadístico estaremos más preparados para saber que quieren decir realmente los resultados de los análisis, y también lo que no quieren decir.

2. Variables

La Estadística toma parte de su terminología de su objeto de estudio original, las poblaciones humanas, tal como su etimología revela. Por una *población* podemos entender un conjunto relativamente homogéneo del que observaremos una serie de características, algunas de ellas cuantificables. Una población puede estar compuesta de artefactos, restos humanos, o poblados de una determinada cultura o periodo. En general se una población se presenta como un ente no totalmente conocido. La acción de observar una determinada característica sobre los individuos de la población se expresa matemáticamente

[1]Las prácticas se desarrollan en un documento aparte

mediante el concepto de función. Denotemos por Ω la población, que es un conjunto compuesto de elementos

$$\Omega = \{\omega_1, \omega_2, \omega_3, \dots\}.$$

Una variable es una función X definida sobre Ω que toma valores en un cierto conjunto que podemos pensar numérico, pero no necesariamente lo es. Los posibles tipos de variables están jerarquizados en tipos que vamos a describir a continuación, basados en la clasificación de Shennan [1].

1. **binaria**. Expresa la presencia o ausencia de una determinada característica o atributo, por lo que sólo puede tener dos valores, que de hecho, pueden codificarse con 0 y 1.

2. **nominal**. La variable puede tomar un número limitado de valores correspondientes a una clasificación del individuo en tipos que no guardan un orden entre si. Los posibles valores se denotan con etiquetas.

3. **ordinal**. Es como la nominal, pero los valores si que presentan un orden. Una variable ordinal puede aparecer al agrupar una numérica, ya sea por conveniencia o desconocimiento.

4. **de intervalo**. Es una variable numérica que no corresponde a la medición de una magnitud.

5. **proporcional**. Una variable numérica que corresponde a la medida de una magnitud.

6. **vectorial**. Un conjunto de variables numéricas proporcionales que deben ser consideradas simultáneamente para representar una información compleja sobre el individuo.

La escala nominal no es más sencilla por no requerir de medición, al contrario, la clasificación dentro de una determinada tipología no es un problema trivial en el plano abstracto[2]. La forma de proceder más objetiva consiste en

[2]Los atributos de los artefactos son casi infinitos y sería absurdo pensar que el arqueólogo debería analizarlos todos por igual. En la práctica, nadie lo intenta. Nuestras observaciones arqueológicas se limitan a la búsqueda y anotación de atributos que el observador cree que han sido hechos por el hombre o seleccionados por y para el hombre. Este juicio último es otro juicio arbitrario y depende del observador y sus ideas o modelos sobre la mentalidad del hombre antiguo. Podemos suponer que el canto bifacial no fue seleccionado por su ra-

una toma de numerosos datos numéricos para interpretarlos como variables vectoriales en un espacio multimensional. En ese contexto los diferentes los individuos más parecidos entre si aparecerán más agrupados que los que no lo son. Es posible que para ello haya que transformar los datos (e.g. aplicando logaritmos). Si la separación entre grupos es suficientemente nítida tendremos así definidos tipos y los individuos así clasificados, en su mayoría.

3. Muestreo y probabilidad

Uno de los principales objetivos de la Estadística es conocer la población Ω. Concretamente, como una determinada característica representada por una variable X se distribuye en Ω, en qué proporciones. La observación de una variable X sobre un individuo $\omega \in \Omega$ se expresa como $X(\omega)$. A lo largo de esta discusión asumiremos que la variable X toma valores numéricos. En la medida que el conocimiento perfecto de la población es imposible, la única forma de obtener información es el *muestreo* que consiste en la selección de un cierto conjunto de elementos de la población $\{\omega_1, \ldots, \omega_n\}$ sobre los que observaremos X lo que dará una serie de valores $\{X(\omega_1), \ldots, X(\omega_n)\}$, que para simplificar renombraremos como $\{x_1, \ldots, x_n\}$ (muestra de tamaño n). Para que el muestreo arroje información fiable sobre la distribución de X sobre Ω deben satisfacerse dos requisitos: el muestreo debe ser *aleatorio* y el tamaño de la muestra debe ser suficientemente grande. Ambos requisitos no pueden satisfacerse en general en ciertas disciplinas por cuestiones metodológicas y presupuestarias. Por ejemplo, los artefactos procedentes de una excavación no son seleccionados por un procedimiento aleatorio sino que aparecen contiguamente unos con otros como consecuencia de una extracción ordenada, si embargo el desconocimiento de lo que hay bajo el suelo permite considerar la extracción de artefactos como un muestreo aleatorio a falta de algo mejor. En cuanto al tamaño necesario de la muestra para que ésta revele características de Ω dependerá de lo que queramos saber y el nivel de *confianza*, que desarrollaremos después.

El soporte matemático de la aleatoriedad, y por lo tanto el muestreo, así

diactividad o por su índice de refracción; sino que se seleccionó por el tamaño y el peso, y probablemente, el material también ¿Qué del color y todos los otros atributos? Obviamente, algunos atributos fueron regular y cuidadosamente seleccionados por algún motivo y éstos nos aportan una información y datos arqueológicos. Otros no fueron ni normal ni intencionadamente seleccionados y producen una "interferencia parásita" o una "no-información" (texto tomado de Clarke [3]).

como en general los métodos estadísticos, lo proporciona la Teoría de la Probabilidad. Supongamos que una una cierta proporción $0 \leq \alpha \leq 1$ tiene una determinada característica. La probabilidad de obtener un individuo con esa característica en un muestreo aleatorio es exactamente α, lo que significa que si extrajéramos n individuos al azar de Ω aproximadamente $n\alpha$ deberían tener esa característica, que es lo que intuitivamente esperamos. Sin embargo, el uso positivista de la probabilidad es algo más complicado: Supongamos que arrojamos tres veces una moneda y salen dos veces "cruz" y una "cara" ¿Cabe esperar como resultado "cara" si se tira una cuarta vez porque es el reparto más adecuado para una moneda en la que caras y cruces son equiprobables? La respuesta es que incluso para la cuarta tirada cara y cruz son igualmente probables porque la moneda no tiene memoria, y gracias a eso cada tirada es un suceso *independiente*. Sin embargo, para una moneda que se tira cuatro veces la combinación 2 caras - 2 cruces es más probable que 3 cruces - 1 cara, pero no mucho más. Concretamente la primera puede suceder de 6 formas distintas mientras que la segunda sólo de 4 formas, de un total de 16 posibles resultados.

Estas diferencias son aún más dramáticas si en vez de cuatro lanzamientos se hacen muchos más. Pero, si bien es altamente improbable un reparto equitativo de caras y cruces, sí que es posible delimitar lo que es más probable de lo que no lo es: un reparto desequilibrado hasta la anomalía entre caras y cruces es altamente improbable para una moneda simétrica, por lo tanto debería ser más probable que los resultados se mantengan dentro de ciertos límites que se pueden calcular: en una tirada de la moneda 8 veces, que salgan sólo caras o cruces es tiene una probabilidad menor que 0.01, por lo tanto más del 99 % de las veces habrá caras y cruces en la tirada. Como veremos más adelante, esto tiene relación con los *tests de hipótesis estadísticas*. Más aún, el teorema conocido como "ley de los grandes números" [5] afirma que la proporción entre caras y cruces tiende a igualarse cuando el número de lanzamientos tiende a infinito, casi siempre... la Teoría de la Probabilidad se cura en salud.

4. Estimación y confianza

Sigamos examinando la variable numérica X. Sus valores sobre los individuos de Ω se reparten en un cierto intervalo y es posible que se agrupen alrededor de un valor en particular, que es típicamente la *media* de la población, denotada μ, pero no siempre, lo que discutiremos en la siguiente sección. La idea de μ que debemos tener en mente es que los valores de X, visualizando

los números sobre una recta, se reparten a derecha e izquierda de μ. Lo agrupados, o esparcidos, que estén los valores respecto a μ es también susceptible de medida, lo que puede hacerse de varias maneras. Habitualmente se toma la *desviación* σ, sobre la que ampliaremos información después. Sin entrar en el significado preciso de σ, la idea es que cuanto más pequeña es σ mas concentrados están los valores de X alrededor de μ. En este sentido, el par de números (μ, σ) permite resumir de manera muy eficiente multitud de valores de X sobre Ω.

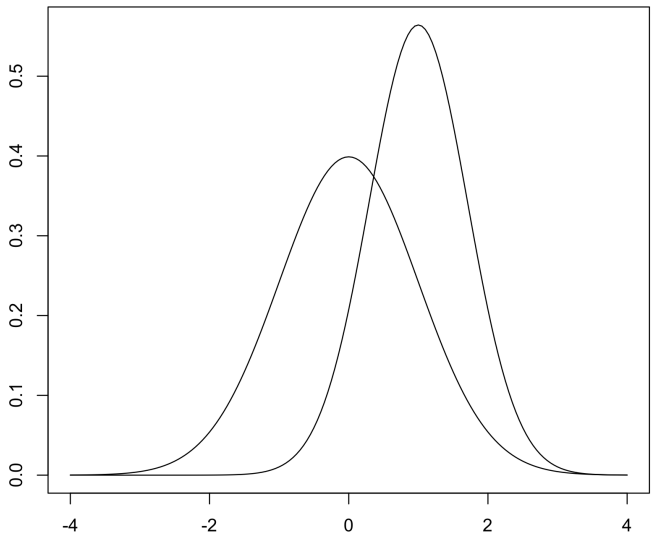
Para tener una interpretación más precisa del significado de μ y σ para una variable X numérica definida sobre una población tenemos que ir a un ejemplo muy particular, pero importante. Ocurre con frecuencia en las ciencias empíricas o sociales que las variables tienen una distribución *normal*, descubierta por Gauss, lo que da una información totalmente precisa de como están repartidos los valores de X conociendo únicamente μ y σ.[3] Las variables con distribución aproximadamente normal aparecen como consecuencia de procesos en los que se superponen incontables efectos independientes gracias al llamado *teorema central del límite* [5], que sólo mencionamos para espolear la curiosidad del lector. Si se quiere pensar en un ejemplo concreto de variable con distribución normal, pongamos que X representa la estatura y Ω es un conjunto homogéneo de individuos: adultos, mismo sexo, grupo étnico (si no se filtran estas características la aproximación no es tan buena) ¿Qué significa exactamente que X sigue una distribución normal de media μ y desviación σ? Simplemente, que la proporción de individuos de Ω para los cuales la observación de X cae dentro del intervalo $[a, b]$, que denotaremos abreviadamente como $\mathbb{P}(X \in [a, b])$ haciendo uso del lenguaje de la Teoría de la Probabilidad, es

$$\mathbb{P}(X \in [a, b]) = \frac{1}{\sqrt{2\pi}\sigma} \int_a^b e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx.$$

La función dentro de la integral recibe el nombre de *campana de Gauss*, *gaussiana* o simplemente *curva normal*. Sobre la media μ se sitúa el máximo de la función y σ da una medida de lo afilada o chata que es la gráfica. No hay que preocuparse por el calculo de la integral que, si bien no es posible resolverla por cálculo de primitivas, sus valores están tabulados (e implementados en el software estadístico) con precisión más que suficiente. Lo que sí es interesante

[3]Por ello es asombroso que siga habiendo malinterpretaciones, o peor, iterpretaciones interesadas, de fenómenos totalmente naturales basados en la distribución normal. Véase el capítulo "Empresa acusada de prejuicios étnicos en las contrataciones" de Paulos [9].

visualizar la integral como un *área* bajo la curva (la extraña constante que antecede a la integral tiene como finalidad garantizar que el área total bajo la curva es 1).



Normales con diferentes medias y desviaciones

Naturalmente, para usar la fórmula anterior hay que conocer los *parámetros* μ y σ, además de tener la certeza de que la variable X sigue una distribución normal. Vayamos por partes. La estimación de los parámetros, para la normal u otra distribución, se puede hacer a través de ciertas funciones construidas a partir de los valores muestrales que reciben en nombre de *estadígrafos* (o *estadísticos*, pero tal nombre induce a confusión). La media muestral m es el mejor estimador para μ

$$m = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{k=1}^{n} x_k.$$

Como la media muestral depende de la muestra escogida es ella misma una *variable aleatoria* cuya media es también μ y cuya desviación es proporcional a $1/\sqrt{n}$, lo que nos da una idea de como se aproxima m a μ. Gracias al teorema central del límite (que volvemos a invocar), se sabe además que m sigue una distribución normal, sin embargo el desconocimiento exacto de su desviación

nos impide ser más precisos a priori. Se puede usar la *desviación muestral* s como estimación y sustituto provisional de σ

$$s = \sqrt{\frac{(x_1 - m)^2 + \cdots + (x_n - m)^2}{n-1}} = \sqrt{\frac{\sum_{k=1}^{n}(x_k - m)^2}{n-1}}.$$

Resulta extraño dividir por $n-1$ en lugar de n, pero esto es lo que nos dicen los cálculos. De esta manera, con una muestra inicial relativamente "pequeña" se puede estimar σ y con ese conocimiento proponer un nuevo tamaño muestral n para estimar μ con un margen de confianza ajustado a nuestros propósitos. Más concretamente, se tiene para la diferencia entre μ (desconocida) y m la siguiente fórmula debida a Chebyshev

$$\mathbb{P}\left(|m - \mu| \leq \frac{k\sigma}{\sqrt{n}}\right) \geq 1 - \frac{1}{k^2}$$

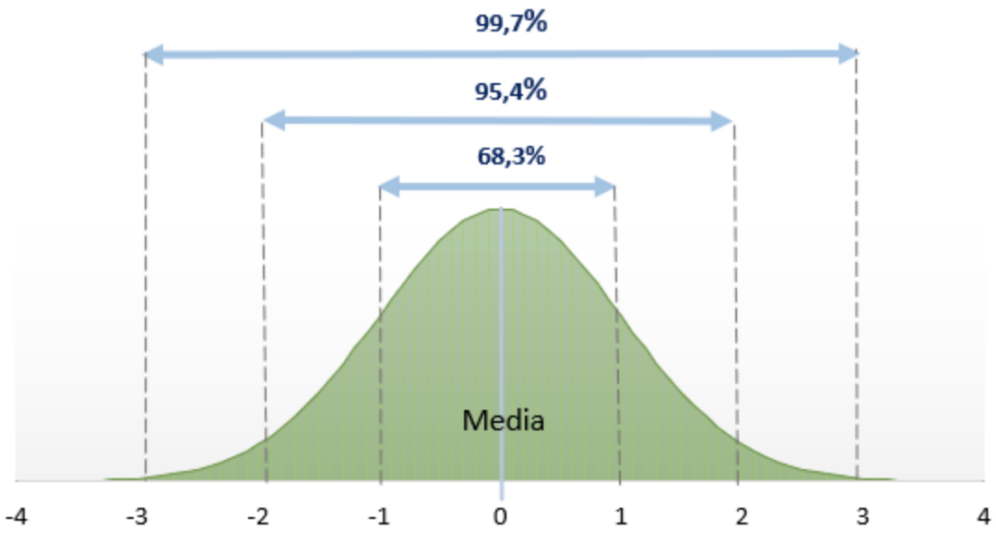
donde k es un parámetro para ajustar la confianza, por ejemplo $k = 10$ da una confianza de 0,99 (si quisiéramos 0,95 tendríamos que poner $k = 4{,}47$). Esta fórmula es válida para cualquier distribución, no sólo la normal, y permite construir *intervalos de confianza* para μ centrados en m. La presencia de $\sqrt{n}$ nos dice que cada vez que queramos reducir el tamaño del intervalo de confianza a la mitad hay que multiplicar n por 4. También sirve para reducir el tamaño muestral el conocimiento previo de la distribución. Por ejemplo, si X fuese normal, el tamaño del intervalo se reduciría a menos de la mitad del dado por la fórmula anterior. Explícitamente el intervalo para μ al 0,95 de confianza es

$$m \pm \frac{1{,}96\sigma}{\sqrt{n}}.$$

Señalemos también que el "malabarismo" descrito anteriormente de estimar en primer lugar σ por medio de s para después estimar μ, con un n mayor, se puede evitar usando una distribución especial llamada t de Student que sustituye a la normal cuando se desconoce σ.

Dependiendo de las aplicaciones pueden ser convenientes distintos niveles de confianza, pero también hay que tener presente que fijarlo muy próximo a 1 dispara el tamaño de las muestras. Más adelante, en relación con la comprobación estadística de las hipótesis veremos que para detectar relaciones débiles un alto nivel de confianza es contraproducente. Por este motivo existe un cierto

consenso en adoptar el 0,95 en ciencias sociales.[4] En ciencias empíricas puede llegarse al 0,99 e incluso más ya que la relación causal entre las variables se traduce fuertemente en los datos, o bien porque se pueden repetir los experimentos un suficiente número de veces. A falta de más datos, existe una relación inversa entre la cantidad de información y la su certeza. En efecto, si queremos pasar de una certeza del 0,95 al 0,99 será a costa de ampliar el intervalo de confianza para el dato que se estima con la subsecuente pérdida de información.



Interpretación de la distribución normal en términos porcentuales

5. Otros estadígrafos descriptivos

En la sección anterior nos hemos centrado en la media y la desviación por ser los estadígrafos de uso más extendido y por ser los que más se benefician de la Teoría de la Probabilidad. No obstante, si lo único que queremos es condensar en unos pocos números la distribución de los valores observados de una cierta variable X hay otros estadígrafos que pueden cumplir ese propósito incluso mejor. La media no siempre es el valor más representativo: pensemos en si la renta media de un país da una buena descripción de los ingresos de su "clase media", por ejemplo. Llamamos *mediana* de X un valor m_d para el cual los conjuntos $\{\omega : X(\omega) \leq m_d\}$ y $\{\omega : X(\omega) \geq m_d\}$ tienen el mismo tamaño.

[4]The Sacredness of .05: A Note concerning the Uses of Statistical Levels of Significance in Social Science, J.K. Skipper, A.L. Guenther y G. Nass, The American Sociologist 1967.

Para los valores observados en un muestreo $\{x_1, \ldots, x_n\}$ la mediana se obtiene reordenando los valores de manera creciente

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

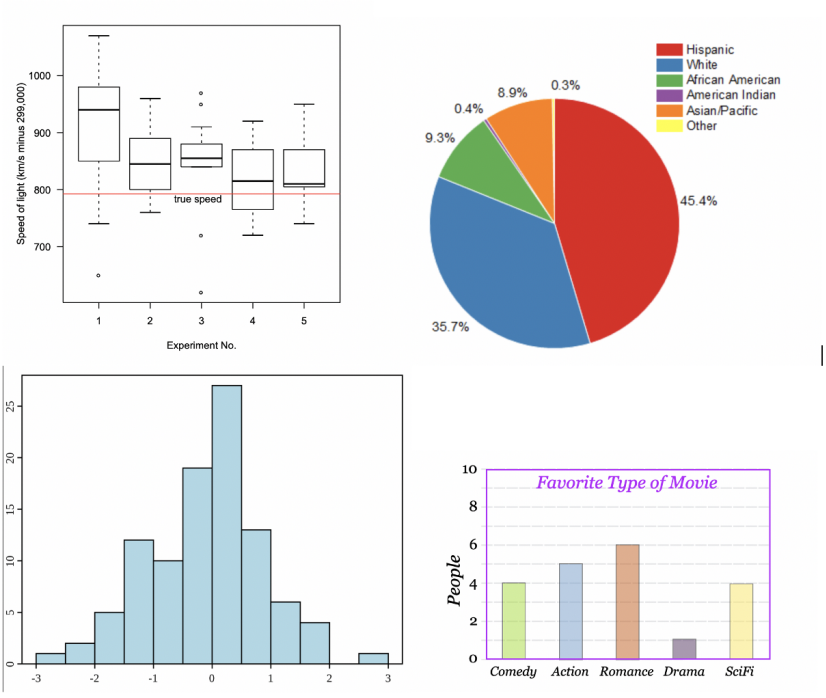
y la media será el valor que ocupe el lugar central (n impar) o la semisuma de los dos valores centrales (n par). Es evidente que la mediana es más informativa que a media en aquellas situaciones en las que la presencia y distribución de valores extremos desplaza a esta última hasta posiciones que son menos representativas de la muestra.

El agrupamiento alrededor de la mediana para los valores observados puede ser también cuantificado con la misma filosofía. Podemos fijar un número $\alpha \in [0, 1]$ y buscar un número Q_α con la propiedad de que el conjunto $\{n : x_n \leq Q_\alpha\}$ tenga exactamente αn elementos y $\{n : x_n > Q_\alpha\}$ tenga $(1 - \alpha)n$ elementos. Naturalmente, esto tendrá que interpretarse de manera aproximada si αn no es entero y el número Q_α se determinará de manera unívoca con reglas similares a las empleadas con m_d. La mediana se corresponde con $m_d = Q_{,5}$ (hemos suprimido el 0 a la izquierda del punto), que junto con $Q_{,25}$ y $Q_{,75}$ reciben en nombre de *cuartiles*. La agrupación viene mediada por el número $RQ = Q_{,75} - Q_{,25}$ que se denomina *recorrido intercuartílico*. A partir de m_d y RQ se establecen reglas para etiquetar los *valores atípicos* en una observación.

Cuando se tiene una variable que toma un número finito de valores (e.g. nominal, número entero) ocurre a veces que uno se repite con más frecuencia. En ese caso recibe el nombre de *moda*. Para una variable continua X cuya distribución puede ser representada con una curva continua, como es el caso de la normal, la moda se corresponde con el valor donde esta función alcanza su máximo. Puede ocurrir que este máximo absoluto esté acompañado por otro máximo relativo dando a la curva el aspecto de las jorobas de un camello. En este caso se dice que la distribución es *bimodal*, o *multimodal* si hubiera más de dos. Esta situación suele poner de manifiesto que la población es en realidad la superposición de dos o más poblaciones y en cada una de ellas X se manifiesta de manera más regular. La definición de moda para valores observados numéricos de una muestra presenta algunos problemas, ya que los números no enteros difícilmente se repiten. En este caso, los valores pueden ser agrupados en intervalos entre el mínimo y el máximo, ni pocos ni demasiados (aproximadamente $\sqrt{n}$ suele ser una buena elección). Si hay un intervalo que contenga más que los demás, podemos elegir como moda el punto medio de

dicho intervalo (visualmente, esto se corresponde con la columna más alta del histograma construido sobre esos intervalos). Debe entenderse que esto únicamente pretende señalar lo que sería el valor más repetido del muestreo si se admitiera un cierto "redondeo".

Decimos que media, mediana y moda son medidas de *centralización*, mientras que la desviación y recorrido intercuartílico son medidas de *dispersión*. Sin embargo, una distribución puede seguir patrones más irregulares y mostrar, por ejemplo, una fuerte asimetría alrededor de su centro. Por este motivo se consideran también medidas de asimetría o *kurtosis*. No dejemos de mencionar aquí que la Estadística Descriptiva ofrece numerosos recursos gráficos para observar y resumir distribuciones: *box-plot*, que resume gráficamente el intervalo de distribución de valores y cuartiles; *gráfico de barras*, para representar cantidades sobre variables nominales (no confundir con *histograma*); *gráficos circulares*, para lo mismo que el gráfico de barras pero con proporciones o porcentajes; *diagramas triángulares*, para representar simultáneamente varias proporciones cuando sólo hay tres variables; etc.



De izquierda a derecha y de arriba a abajo: box-plot, gráfico circular, histograma, gráfico de barras.

6. Los datos toman forma

Las medidas de centralización y dispersión descritas son adecuadas para elaborar resúmenes numéricos, pero se pierde mucha información. Para resolver ciertas cuestiones, como saber por ejemplo, saber si una variable X sigue una distribución normal, o si los valores de X se reparten igual entre dos subgrupos de Ω, necesitamos trabajar simultáneamente con todos los valores observados en el muestreo. Al igual que en las dos secciones anteriores tenemos que dar dos nociones paralelas, una para la población y otra para la muestra. Consideremos un variable X sobre la población Ω. Su *función de distribución* es la siguiente

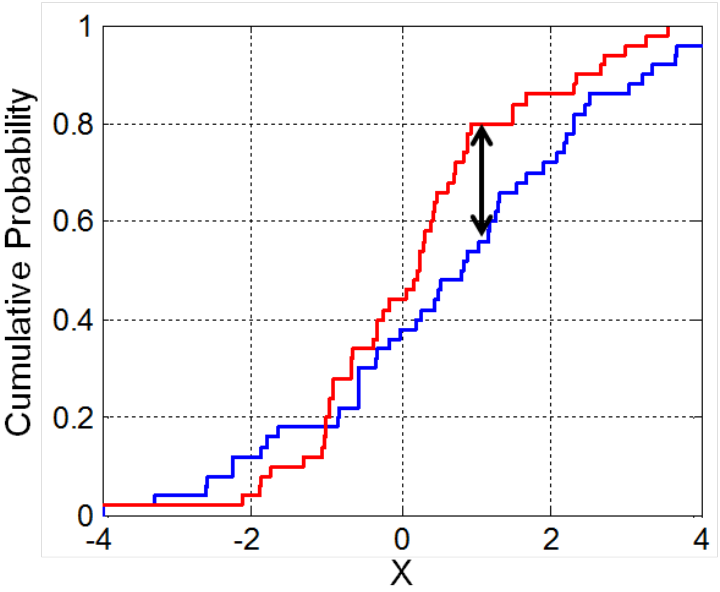
$$F_X(t) = \mathbb{P}(X \leq t)$$

es decir, la proporción de individuos $\omega \in \Omega$ para los cuales $X(\omega) \leq t$. Notemos que la función está definida para todo $t \in \mathbb{R}$, es creciente y toma valores entre 0 y 1. Para un conjunto de valores observados $\{x_1, \ldots, x_n\}$ la función de distribución *muestral* es una *función escalonada* construida con las *frecuencias acumuladas* del siguiente modo

$$F_x(t) = n^{-1}\#\{k : x_k \leq t\}$$

donde # indica el *cardinal* o número de elementos del conjunto al que antecede. Es interesante notar que con las funciones de distribución, sea F_X o F_x, ya no es necesario volver usar explícitamente la población Ω ya que toda la información que nos interesa sobre la variable X está contenida en ellas.

Un teorema fundamental de Kolmogorov dice que F_x se aproxima *uniformemente* a F_X cuando n es grande y la muestra se ha obtenido aleatoriamente. En realidad, la diferencia $\max |F_X - F_x|$ se puede mirar como un estadígrafo muestral cuya distribución es conocida y a partir de ahí formular el parecido entre F_X y F_x en términos de confianza. Por lo tanto, si X sigue una distribución normal debería poder notarse en la gráfica de F_x. Así mismo, si dos variables X e Y tienen la misma distribución debería manifestarse en la aproximación entre sí de las funciones de distribución muestrales asociadas F_x y F_y. Este es el fundamento teórico del test de Kolmogorv-Smirnov del que hablaremos en otra sección.



Comparación entre las gráficas de frecuencias acumuladas para dos muestras

Para las variables numéricas X a veces es posible representar su distribución mediante una *función de densidad* f_X que se relaciona con la de distribución por la fórmula

$$F_X(t) = \int_{-\infty}^{t} f_X(s)\, ds.$$

La función de densidad, entre otras ventajas, permite una visualización más sencilla de como se distribuyen los valores de la variable ya que se corresponden con áreas en lugar de diferencias

$$\mathbb{P}(X \in [a, b]) = \int_{a}^{b} f_X(t)\, dt.$$

La campana de Gauss es el principal ejemplo de función de densidad y cualquier otra función de densidad se interpreta siguiendo las pautas similares vistas anteriormente.

Lamentablemente, no hay una noción que represente el mismo papel que la función de densidad para una muestra. Lo más adecuado consiste en realizar una división en intervalos iguales del rango de valores y definir una función escalonada que valga k/n sobre un intervalo que contiene k valores del conjunto $\{x_1, \ldots, x_n\}$. Este tipo de gráfico se llama *histograma* y será representativo

de la distribución en la medida en que se parezca a f_X. Por este motivo la elección y número de intervalos es fundamental: con muy pocos difícilmente se ajustará a una curva y con demasiados tendrá intervalos vacíos dejando el histograma fraccionado. Para un número moderado de valores se obtiene un reparto proporcionado si el número de intervalos es aproximadamente $\sqrt{n}$, lo que intuitivamente se corresponde con que los valores deben rellenar un área en lugar de un segmento. Para valores grandes de n se prefiere que el número de intervalos crezca más despacio, logarítmicamente por ejemplo (regla de Sturges). Con un buen histograma será evidente la normalidad, o ausencia de ella, de los valores observados, por ejemplo, o nos dará una pauta para proponer hipótesis sobre X.

7. Causalidad e independencia

Cuando dos variables X e Y (sigamos pensando en variables numéricas para fijar ideas) mantienen entre sí una fuerte relación causal, por ejemplo temperatura y dilatación para una barra metálica, los pares de valores observados (x_k, y_k) se dispondrán sobre una curva que es en realidad la gráfica de la función f que las liga $Y = f(X)$ (en el caso de la dilatación de barra f seguirá una ley lineal para un intervalo moderado de temperaturas). La situación totalmente opuesta es que X e Y sean independientes, como lo son dos lanzamientos de la misma moneda, esto se traduce en que cualquier conjunto de valores definido por X, la variable Y se reparte en la misma proporción que en todo Ω. Cuantitativamente eso se traduce en

$$\mathbb{P}((X, Y) \in [a, b] \times [c, d]) = \mathbb{P}(X \in [a, b]) \cdot \mathbb{P}(Y \in [c, d])$$

y gráficamente en un reparto caótico de los valores muestrales en los que no se percibe la cercanía a ninguna curva. Muchas situaciones para dos variables oscilan entre los dos extremos dados: o bien la relación causa efecto es más débil, por ejemplo, edad y estatura, o bien la falta de un ajuste más exacto entre los datos se puede achacar a la falta de precisión de las mediciones u otro *ruido*. En tal caso, aunque los valores observados (x_k, y_k) no se dispondrán sobre una curva pero se aproximarán a ella.

Ahora bien, dadas dos variables que no son independientes necesitamos una forma de proponer una función cuya gráfica mejor se ajuste a los valores observados. El problema es que sin un conocimiento previo de la posible relación causal no se puede proponer la familia de funciones más adecuada para la

búsqueda de la que mejor describe los datos. Para lo que sigue supondremos que queremos explicar Y como consecuencia de X, lo que rompe la simetría de la discusión. Dado que la dependencia a través de una función derivable es lineal a pequeña escala lo normal es explorar las relaciones lineales entre los datos. Esto también es más sencillo porque las líneas en el plano dependen únicamente de dos parámetros, que suelen ser la pendiente a y el valor de $b = y$ cuando $x = 0$, esto es, la recta de ecuación $y = ax + b$. La recta que mejor se ajusta, en el sentido que minimiza la suma de los cuadrados de las distancias verticales a los puntos, se llama *recta de regresión*, pero como es ampliamente utilizada en otras disciplinas donde toma el nombre de la técnica *mínimos cuadrados*. Pondremos la fórmula explícita de los coeficientes

$$a = \frac{n\sum_{k=1}^{n} x_k y_k - (\sum_{k=1}^{n} x_k)(\sum_{k=1}^{n} y_k)}{n\sum_{k=1}^{n} x_k^2 - (\sum_{k=1}^{n} x_k)^2}$$

$$b = \frac{(\sum_{k=1}^{n} x_k^2)(\sum_{k=1}^{n} y_k) - (\sum_{k=1}^{n} x_k)(\sum_{k=1}^{n} x_k y_k)}{n\sum_{k=1}^{n} x_k^2 - (\sum_{k=1}^{n} x_k)^2}$$

que afortunadamente están implementados en el software estadístico como **R**. El *modelo lineal* consiste tratar las diferencias entre los valores muestrales y la recta de regresión como el efecto de una variable aleatoria desconocida que sigue una distribución normal. El modelo lineal tiene interés *predictivo*.

Hay que decir que las fórmulas anteriores aparecen más complicadas de lo que podrían ser porque deliberadamente hemos evitado que aparezcan las medias muestrales. Si llamamos $\overline{x}$ a la media muestral de X e $\overline{y}$ a la de X, podemos escribir el *coeficiente de correlación*

$$r = \frac{\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y})}{\sqrt{\sum_{k=1}^{n}(x_k - \overline{x})^2(y_k - \overline{y})^2}}.$$

El coeficiente r está entre -1 y 1 y mide el grado de dependencia lineal y el sentido en el que esta ocurre, aunque se prefiere r^2 porque da una medida más precisa de la relación (lineal) entre las variables. Al estar r^2 comprendido entre 0 y 1, al igual que la probabilidad, se suele interpretar como un grado de certidumbre. Es importante notar que r o r^2, al contrario que los coeficientes a y b, no depende del orden de las variables.

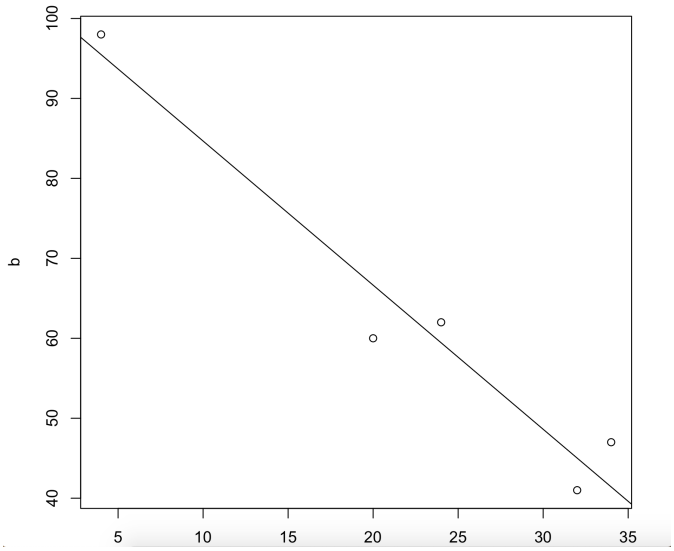


Gráfico con cinco pares de datos y recta de regresión

El grado de dependencia entre variables en la escala nominal también es susceptible de medida. Cuando se observan simultáneamente dos variables nominales X e Y que pueden tomar n y m valores respectivamente hay $n \cdot m$ pares de valores posibles. La anotación ordenada de las frecuencias de cada par da lugar una matriz con números enteros llamada *tabla de contingencia*. Una de las cuestiones habituales es preguntarse si las variables son independientes, lo que se puede evaluar con el test de la χ^2 que veremos en la sección siguiente. En el caso de que no lo sean se puede cuantificar la medida de dependencia entre ellas de la siguiente forma. Nos limitaremos únicamente al caso de un par de variables binarias. En tal caso la tabla de contingencia será 2×2 y tendrá este aspecto

a	b
c	d

El grado de relación entre las variables enfrentadas se calcula con la fórmula

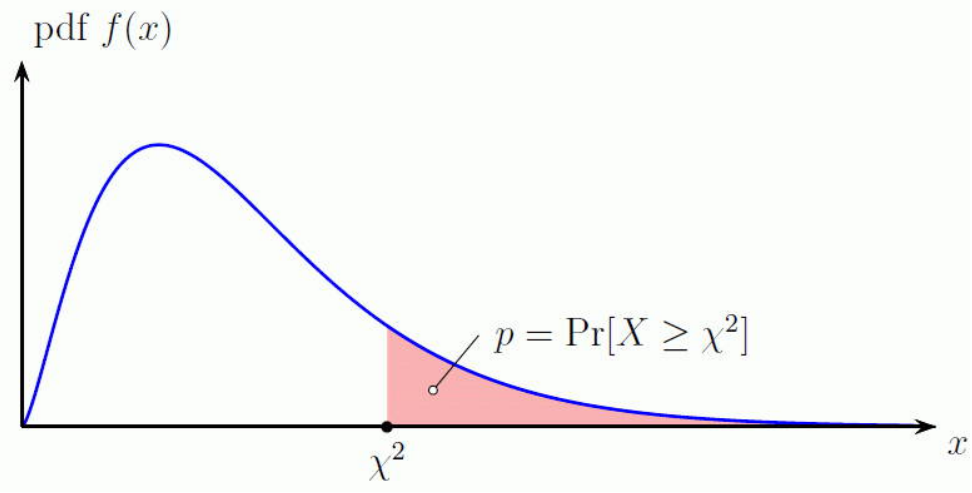
$$q = \left| \frac{ad - bc}{ad + bc} \right|.$$

Notemos que si las variables fueran independientes el numerador estaría próximo a 0 mientras que la dependencia carga los coeficientes en una de las diagonales.

8. Validación estadística de las hipótesis

Hasta ahora han aparecido varias situaciones en las que se afirma una característica de la población o las variables, pero no hemos dado una herramienta efectiva para comprobar si son ciertas, hasta cierto grado de confianza, su complementario el nivel de *significación*. Ejemplos de esto son saber si una cierta variable X sigue una distribución normal, sabe si un par de variables X e Y tienen la misma distribución o si esas mismas variables son independientes. Para resolver estas cuestiones y otras parecidas se forma una función o estadígrafo a partir de los valores muestrales que tiene una cierta distribución conocida si la hipótesis que se desea verificar (llamada normalmente *hipótesis nula* o H_0) es cierta, por lo que con una cierta probabilidad caerá en un determinado intervalo. Si es así, esto es todo cuadra, aceptamos la hipótesis H_0, y se rechaza en caso contrario, que se puede ver como aceptar la hipótesis alternativa H_1.

Por ejemplo, queremos comprobar si una moneda es simétrica. Sabemos que en caso de que lo sea, si se lanza 8 veces la probabilidad de que salgan todo caras o todo cruces es inferior a 0,01. El test que vamos a realizar consistirá en lanzar la moneda 8 veces: si salen caras y cruces aceptamos que la moneda es simétrica; en caso contrario se rechaza la hipótesis. Hagamos un poco de crítica del método. Sabemos que una moneda simétrica puede tener rachas de caras o cruces arbitrariamente largas, por lo que eventualmente podemos rechazar la hipótesis siendo verdadera. Esto es lo que se denomina *error de tipo I*. Por otra parte, si la moneda estuviera desequilibrada, tendría que estarlo mucho para repetirse 8 veces el mismo resultado, de manera que se aceptaría fácilmente la hipótesis de simetría siendo falsa. Esto se denomina *error de tipo II*. Bien, el error de tipo I es el único sobre el que se tiene una estimación precisa de la probabilidad de cometerlo. El error de tipo II es más esquivo y difícil de estimar, sólo se puede reducir a costa de la calidad del test empleado (su *potencia*) y del tamaño de la muestra. Tras la epidemia Covid-19 cierta terminología se ha incorporado al lenguaje cotidiano: podemos referirnos al error de tipo I como *falso negativo* y al de tipo II como *falso positivo*. Cuando se realiza un test implementado en software estadístico normalmente el resultado es arrojado en forma de número llamado *p-valor*. Explicaremos lo que significa. El test, como estadígrafo, sigue una cierta distribución T, y cuando se evalúa sobre la muestra arroja un valor τ. El p-valor es la probabilidad de que T sea mayor que τ, de manera que para aceptar la hipótesis nula el p-valor debe ser mayor que 0,05 si este es el nivel de significación escogido.



Gráfica de la χ^2 mostrando el p-valor (área en color rosa)

Una advertencia importante: el hecho de pasar el test no quiere decir que la hipótesis sea cierta, ni que sea descartable si no lo pasara. Ha quedado claro en la discusión que son posibles errores, en los dos sentidos, y su ocurrencia es totalmente aleatoria por fino que sea el umbral fijado. Si realmente hubiera un hecho escondido en los datos éste terminaría apareciendo eventualmente de alguna manera, independientemente del azar subyacente a la metodología. Por otra parte, un test podría ahorrarnos tiempo descartando una apreciación subjetiva errónea sobre los datos.

El test de Kolmogorov-Smirnov se basa en la aproximación de la función de distribución muestral a la función de distribución subyacente. Aunque normalmente se exige que las variables sean numéricas puede también funcionar con variables ordinales asociando a los valores de éstas números en el mismo orden. Con variables numéricas se puede contrastar la hipótesis de que ésta siga una distribución dada, por ejemplo la normal. Esto se denomina uso como test de *bondad de ajuste*. Con esta finalidad funciona a partir de un tamaño muestral de $n \geq 20$, lo que es verdaderamente notable. Se puede usar para comprobar que dos muestras siguen la misma distribución (*test de homogeneidad*), lo que incidentalmente sirve también para comprobar la independencia de dos variables (*test de independencia*) de las cuales una debe ser binaria y la otra ordinal o numérica. En efecto, clasificada la muestra en dos grupos de acuerdo a los valores de la variable binaria se reduce a comparar que las dos muestras siguen la misma distribución. El test de Kolmogorov-Smirnov da buenos resultados si ambas muestras tienen un tamaño $n \geq 40$.

El test de la χ^2 de Pearson se puede usar también para bondad de ajuste, pero tiene el inconveniente de que los datos deben ser agrupados por intervalos. En este sentido, funciona mejor cuando los datos ya vienen dados en esta forma, particularmente una tabla de contingencia de dos variables nominales. Utiliza una distribución que depende de un parámetro entero, los *grados de libertad*. Sirve para comprobar independencia donde el test de Kolmogorov-Smirnov no llega porque las variables no están ordenadas o toman 3 o más valores. Funciona bien a partir de $n \geq 50$ siempre que no queden categorías vacías o con números muy bajos. La corrección para ese tipo de situaciones está implementada por lo general en el software estadístico.

Ambos test pertenecen al tipo llamado *no paramétrico*, reservándose el término *paramétrico* que sirve para comprobar hipótesis bajo el supuesto que la variable que interviene es normal. Esta es la situación frecuente en muchas disciplinas y se ha desarrollado una panoplia de técnicas que no nos serán de mucha ayuda por la naturaleza de los datos arqueológicos.

9. Perspectivas

Hay técnicas estadísticas que no hemos mencionado aquí que pueden ser de interés para el profesional de la Arqueología. Mencionemos algunas de ellas.

Se tiene la *regresión lineal multivariante* en la que se trabaja con más variables, en lugar de una, como hemos hecho anteriormente. En ellos se puede además cuantificar la influencia de cada variable, o mejor aún proponer nuevas variables que son combinación lineal de las primitivas y resumen mejor la variación de los datos. Esto es el llamado *análisis de componentes principales*. Para ampliar información sobre técnicas de análisis multivariante véase [8].

Para definir tipos se usan técnicas de *análisis de conglomerados* o *clusters* [8]. El principio subyacente es examinar vectores de datos en un espacio multidimensional y generar grupos a partir de la distancia entre ellos. Los métodos también proporcionan jerarquías o taxonomías con las que se pueden adscribir nuevos individuos en los grupos ya definidos. Debido al auge de la *ciencia de los datos* hay muchos más procedimientos para descubrir estructuras y agrupaciones en un conjunto de datos multivariante (*data mining*). Estas técnicas, como las *redes neuronales*, trascienden la Estadística. Un conjunto de datos

sirve para *entrenar* un sistema (*Machine Learning*) que después realizará clasificaciones automáticamente. Estos sistemas pueden implementarse de manera que aprenden de sus errores, con lo que el funcionamiento mejora con el tiempo y el uso.

La práctica arqueológica implica la excavación de yacimientos y prospección del terreno para encontrar nuevos lugares. El *análisis espacial* es una herramienta para tener en cuenta. El estudio de la variabilidad de los artefactos a lo largo de las cuadrículas excavadas podría servir para planificar las siguientes campañas de la excavación. En Geografía se han propuesto distintos modelos para evaluar la influencia de un conjunto de núcleos de población en un territorio, lo cual es perfectamente aplicable a sociedades pasadas [4]. Mencionemos también que las técnicas de *Geoestadística* desarrolladas para la estimación de recursos mineros podrían ser interesantes en un contexto arqueológico [10].

Finalmente, hay una corriente en la Estadística que ha ganado una tremenda presencia en las últimas décadas llamada *Estadística Bayesiana*, cuyo principio básico es el reemplazar la *interpretación frecuentista*, basada en la teoría matemática de la probabilidad de Kolmogorov, por una interpretación filosófica del mundo. Para mí como científico, con esto queda todo dicho.

Referencias

[1] S. Shennan, *Arqueología Cuantitativa*, Crítica 1992.

[2] D. L. Carlson, *Quantitative Methods in Archaeology Using R*, Cambridge University Press, 2017.

[3] D. L. Clarke, *Arqueología Analítica*, Bellaterra 1984.

[4] K. W. Butzer, *Arqueología, una ecología del hombre*, Bellaterra, 2007.

[5] S. Ríos, *Métodos Estadísticos*, Ediciones del Castillo 1983.

[6] M. R. Spiegel, *Estadística*, Serie Schaum, McGraw-Hill 1975.

[7] G. I. Ivichenko, Y.I. Medvedev, *Mathematical Statistics*, URSS 1990.

[8] D. Peña, *Análisis de Datos Multivariantes*, Mc Graw Hill 2002.

[9] J. A. Paulos, *Un matemático lee el periódico*, Metatemas 44, Tusquets.

[10] E. Pardo, *Geomatemáticas*, IGME - Catarata, 2012

Prácticas de R

R es un leguaje de programación orientado a la Estadística. Como cualquier otro lenguaje de programación, utiliza *comandos* para realizar las operaciones deseadas. Una vez que está instalado **R**, los comandos se ejecutan a través de una *terminal.* Para que la "experiencia de usuario" sea más agradable se han desarrollado entornos en los que usar **R**, es decir programas que hacen de intermediarios y ayudan en la utilización de **R**.

Nosotros usaremos **RStudio**, que está instalado en todas las ALAs de la UMU. Quien desee tenerlo en su ordenador personal, puede descargarlo e instalarlo desde https://posit.co/download/rstudio-desktop/ Se requiere instalar previamente **R**, pero esto ya está explicado convenientemente en las instrucciones que se encontrarán en dicha página.

Todo lo que se haga se guardará en el directorio en el que se está trabajando, lo que se puede ajustar al comienzo de la sesión (se especifica el path)

```
> getwd()
> setwd("Documents/R")        (Sirve para especificar otro directorio)
```

Puede ser interesante distinguir entre directorios cuando se usan paquetes sospechosos de ser "incompatibles" (e.g. se usa el mismo nombre para un comando definido en dos paquetes, no necesariamente con la misma finalidad).

Cuando se sale con q() da la opción de guardar todo lo que se ha escrito como .txt
Si algún ejemplo o comando resultara particularmente interesante lo mejor es copiarlo en un archivo de texto plano, ya que la sesión de **R** completa contendrá mucha "morralla" que no interesa.

- R como Calculadora

El objeto de esta sección es comenzar a familiarizarse con **R** a través de las operaciones aritméticas más básicas.

```
> 2+2
> 2*2/3
> 2^3
> 2/3
> 50 %/% 8                    ¿Qué efecto tienen estas operaciones?
> 50 %% 8
> sqrt(7)
> 1.2e3                       (Cuidado, los decimales van con puntos)
```

Si queremos más (o menos) dígitos, la precisión se puede ajustar con
```
> options(digits=17)
> 5/3                          (observamos redondeo hacia arriba)
```

```
> options(digits=22)
> pi                          (R conoce algunos números)
> 1/0
>1/Inf                       (aritmética del infinito)
> 1/0 – 1/0                   (esto ya es demasiado… not a number)
> options(digits=4)           (para no cansarnos de ver decimales)
> floor(pi)                   (parte entera de un número)

> 2==2                        (lógica)
> 2!=2
> (2!=3)==(2!=4)              ¿Qué significa esto?
> a <- F
> a == FALSE

> a=2                         (asignación de valor a una constante)
> a
> 2=a                         (al revés no funciona con el signo = )
> a <- 7                      (asignación con flecha)
> 7 -> a                      (no hay problema de inversión)
```

- Sucesiones y vectores

Aparte de números, **R** trabaja con distintas estructuras. La siguiente en complejidad son los vectores (sucesiones de números). **R** dispone de comandos para generar vectores con patrones regulares.

```
> 1:10
> seq(0,20, by=3)
> rep(7,10)

> a <- 1:10                   (los vectores pueden ser asignados a constantes)
> b <- rep(3,10)
> a + b                       (operaciones con vectores)
> c <- rep(1,20)
> a+c                         ¿Cómo es esto posible?
> a*b
> a^b
> log10(1000)
> a+2

> c(a,b)                      (pegado de vectores)

> a <- 1:12
> a[7:12]                     (seleccionar coordenadas del vector)
> a[-3]                       (suprimir coordenadas del vector)
```

Escriba un vector largo y desordenado x

```
> min(x)
> max(x)
```

- Matrices

Tras los vectores van las matrices. Una manera de producir una matriz es tener previamente los coeficientes escritos en un vector.

```
> a <- 1:16
> b <- a^2
> matrix(b,4,4)                    ¿Cómo se rellena la matriz?
> matrix(b,2,8)

> A <- matrix(c(1,4,3,1),2,2)      (el comando se puede aplicar sobre vectores)
> A*A
> A^2
```

Para multiplicar matrices en la manera ordinaria hay que usar %*%

```
> B <- solve(A)                    (matriz inversa, por si alguna vez hiciera falta)
> A%*%B
```

Hasta ahora hemos visto distintas estructuras en orden de complejidad creciente: números, vectores y matrices. Exploremos más relaciones entre ellas.

```
> matrix(1:100,10,10) -> D
> t(D)
> D[2,3]
> D[1, ]
> D[ ,2]
> D[1:5,6:10]
> D[c(3,9,10),c(1,2,6)]            (posiciones múltiples tienen estructura de vector)

> matrix(c(1,2,3),c(11,12,13),c(11,21,31))
```

Por defecto una matriz recibe como etiquetas las posiciones de fila y columna. Esto puede cambiarse

```
> colnames(D) <- c("A", "B", "C", "D", "E", "F", "G", "H", "I", "J")
> rownames(D) <- c("M", "N", "O", "P", "Q", "R", "S", "T", "U", "V")
> D
> D["P","D"]                       (R distingue entre D y "D")
```

- Funciones

R contiene algunas funciones matemáticas.

```
> sin(pi/3)
> log(10)                    (logaritmo neperiano)
> log10(100)                 ¿Recordamos que son los logaritmos?
> exp(1)                     (el famoso número “e”)
```

También se pueden definir funciones para realizar operaciones aritméticas repetitivas. La sintaxis es un poco más complicada pues hay que precisar las variables antes de la fórmula. Por ejemplo:

```
> f <- function(x,y){x^2+y^2}
> f(2,3)
> f(c(1,2),c(3,2))           (si se puede calcular sobre vectores, también lo hace)
```

- Números aleatorios

Los números aleatorios se pueden generar respecto a distintas distribuciones de probabilidad. Usaremos la uniforme, la normal y la Poisson.

```
> runif(n=10,min=0,max=10)
> round(runif( 100, 0, 100))    (sabiendo la posición de cada parámetro es suficiente)

> N <- rnorm(100,0,1)

> hist(N)                    (aumentar n hasta que el histograma se vea bonito)

> p <- -10+ (1:10)*2
> hist(N,p)
```

Simular el lanzamiento de un dado y hacer histograma (floor)

```
> rpois(20,1)                (llenado aleatorio de 20 cajas, media = 1)
```

- IRIS

Un “data frame” es la estructura en la que se suelen guardar las observaciones. Iris es un data frame que con datos de 150 flores que está contenido en **R** como ejemplo para ensayar las funciones y comandos.

```
> iris                       (cuidado, R distingue entre minúsculas y mayúsculas)
```

```
> str(iris)                          (números/factores)
> summary(iris)

> iris[1:50, 2]                      ¿Qué hace?

> iris[1:50,2] -> x
> mean(x)
> median(x)
> quantile(x)

> sort(x)
> stem(x)                            (una forma sencilla de ver la agrupación de datos)
> hist(x)

> summary(x)
> sd(x)

> iris[iris$Species %in% "versicolor", ]          (otra manera de filtrar)
> which(iris$Sepal.Length>6)
> iris[which(iris$Sepal.Length>6), ]
> split(iris$Sepal.Length,iris$Species)          (al revés, a ver qué pasa)
```

- Dibujos con R

Con **R** se pueden dibujar gráficas de funciones usando el comando plot.

```
> x <- seq(-pi,pi,by=0.1)
> plot(x,sin(x))                                 (como hemos dado puntos, salen puntos)

> plot(x,sin(x), type="l", col="blue")           (mejor así ¿qué es cada cosa?)
> plot(x, exp(-x^2), type="l", col="red")        ¿A qué se parece?
```

- Regresión lineal y correlación

Dados dos vectores x, y de la misma longitud (usar runif y rnorm para generarlos)

```
> plot(x,y)
> cor(x,y)
```

Ejemplo tomado del libro de Shennan

```
> a <- c(4,20,32,34,24)
> b <- c(98,60,41,47,62)
> D <- data.frame(a,b)
```

```
> D
> plot(D)
> cor(a,b)
> cor(a,b)^2                    (grado de relación entre a y b)
> lm(b~a)
> M <- lm(b~a)
> M
> summary(M)                    (M contiene más información de lo que parece)
> abline(M)
> str(M)

> attach(iris)                  (R se limita a “iris” y permite abreviar nombres)

> plot(iris)                    (observe las correlaciones entre variables)
> plot(iris$Sepal.L, iris$Petal.L)   (dos variables selecionadas, por ejemplo)
> plot(iris$Sepal.L,iris$Petal.L,col=iris$Species)
> cor(iris$Sepal.L, iris$Petal.L)
```

Separe una de las especies, tal como se ha hecho antes, y se calcule de nuevo la correlación.

- Test de hipótesis

Haremos un primer experimento con datos aleatorios generados por **R**.

```
> x <- rnorm(50)
> y <- rnorm(30)

> ks.test(x,y)

> z <- runif(30)
> ks.test(x,z)                  (comparar el resultado)

> a <- c(6,8,11,29,19,3)
> b <- c(23,21,25,36,27,4)
> T <- matrix(c(a,b),6,2)
> colnames(T) <- c("ricos","pobres")
> rownames(T) <- c("infantil I", "infantil II", "juvenil", "adulto","maduro","senil")
```

Lamentablemente el test ks en **R** no funciona sobre tablas de contigencia ni en la escala nominal. Hay que “engañar” a **R** creando un conjunto de datos individuales que produzca la misma tabla de contingencia.

```
> A <- c(rep(1,6),rep(2,8),rep(3,11),rep(4,29),rep(5,19),rep(6,3))
> B <- c(rep(1,23),rep(2,21),rep(3,25),rep(4,36),rep(5,27),rep(6,4))
> ks.test(A,B)                  (interpretar el resultado)
```

- Paquete “archdata”

El paquete “archdata” fue creado por D.L. Carlson para ilustrar las tecnicas explicadas en su libro *Quantitative Methods in Archaelogy Using R*, Cambridge 2017. Para poder usarlo, hay que instalarlo previamente.

```
> install.packages(“archdata”)
> library(archdata)                (una vez instalado sólo se necesita invocarlo)

> ?archdata                        (para conocer el contenido)

> data(Fibulae)                    (carga el data frame con el que se va a trabajar)
> Fibulae
> table(Fibulae$Coils) -> N        (genera tabla para una variable nominal)
> pie(N)
> barplot(N)

> M <- Fibulae[,"BW"]
> sort(M)
> stem(M)
> sort(Fibulae$Length)
> stem(Fibulae$Length, scale=2)    (experimentar con la escala 1,3)


> data(DartPoints)

> DartPoints
> boxplot(Length~Name, DartPoints)

> T <- c(DartPoints$Length, DartPoints$Width)
> plot(T)

> data(PitHouses)

> table(PitHouses$D,PitHouses$S)        (nótese la forma abreviada)
> xtabs(~PitHouses$D+PitHouses$S)
> xtabs(~Size+Depth, PitHouses)         (escribe los nombres como se los dan)

> T <- table(PitHouses$D,PitHouses$S)
> chisq.test(T)                    (df, P-value, warning)
```

- Intercambio de datos con Excel

El programa Excel de Microsoft es muy utilizado para recoger datos. Por defecto guarda los ficheros en su propio formato (.xls, .xlsx). Si queremos que el fichero pueda ser utilizado por **R** en las opciones de guardar se puede elegir la extensión .csv (comma separated values) que es un fichero de texto plano que se puede abrir y modificar con un editor de texto sencillo como Notepad o editores de TeX (no usar Word o similares porque introducen caracteres extraños). A veces se pueden evitar mensajes de error haciendo pequeños retoques “a mano” en un .csv antes de empezar a trabajar con él.

Para transformar un .csv en data frame se usa esta sintaxis (el .csv debe estar previamente en el working directory)

```
> read.csv("MiFile.csv", header=TRUE, sep=";") -> miDataFrame
```

Excel suele producir ficheros .csv separados por “;” no obstante, se puede comprobar antes viendo el fichero en un editor.

La transformación de un data frame en .csv se utiliza

```
> write.csv(MiDataFrame, file="MiFile.csv")
```

La separación en el .csv será por defecto con “,”. Si queremos después abrir el fichero con Excel vamos a tener problemas porque se requiere “;”. Hay dos maneras de resolverlo… si aún estamos a tiempo de escribirlo, así

```
> write.csv2(MiDataFrame, file="MiFile.csv")
```

En caso de que ya esté escrito, hay que abrir una hoja nueva en Excel y seguir esta línea en el menú: data -> Get Data -> From Text y especificar el .csv deseado. En el cuadro de diálogo se puede especificar que la separación se hace con “,” entre otras cosas.