Apuntes para

Estadística [Descriptiva, Probabilidad e Inferencia Estadística]

(asignatura de los Grados de Ingeniería Química e Ingeniería Informática del Software de la Universidá d'Uviéu en el año 2022)[1]

Última actualización: 14 de abril de 2022

[1]D. N. Barba (Universidá d'Uviéu, Departamento de Estadística e Investigación Operativa y Didáctica de la Matemática). E-mail: nietodavid@uniovi.es.

Índice

I. Un vistazo general

- ☞ La Estadística Descriptiva aborda las diversas herramientas habituales en la manipulación de datos.
- ☞ Trabaja con "colectivos", "grupos" o "conjuntos" de datos recogidos experimentalmente con anterioridad.

Tales colectivos tendrán, en su forma más básica, una notación del tipo $X = \{x_1, \ldots, x_n\}$, la cual llamaremos "notación conjuntista". "X" sería un colectivo de datos concreto; "x_i", donde i es un índice que recorre los números naturales desde 1 hasta n — lo cual abreviaremos por $i = 1, \ldots, n$—, sería cada dato que forma parte de aquel, y "n", el número de datos que hay en el colectivo.

- ☞ La Probabilidad dispone de modelos a priori y procedimientos de cálculo.
- ☞ Desde la Estadística Descriptiva se analizan a posteriori colectivos de datos.
- ☞ La Inferencia Estadística utiliza herramientas probabilísticas para extraer conclusiones a partir de información estadística de un colectivo parcial (muestra), buscando deducir aspectos de un colectivo más amplio en el que el primero está incluido (población).

En el proceder experimental, una vez contextualizado adecuadamente, planificado y diseñado un experimento concreto, la organización, presentación y análisis de los datos que de él se obtengan caen en el ámbito de la Estadística Descriptiva.

II. Estadística Descriptiva

i. Variables estadísticas y representaciones tabulares y gráficas de colectivos.

- ☞ Llamaremos **variable estadística** al rasgo o rasgos sometidos a estudio en un experimento concreto.
- ☞ Las variables estadísticas se concretan en **colectivos de datos** tras llevar a cabo el experimento asociado.

Las variables estadísticas se clasifican en **cuantitativas** o **cualitativas** en función de si los valores que los datos puedan tomar, por la naturaleza del experimento planteado, son numéricos o no, respectivamente.

A su vez, se dirá que una variable estadística es cualitativa nominal u ordinal dependiendo de si, por la naturaleza del experimento planteado, no se sobreentiende que haya una ordenación entre los valores que aquella pueda tomar o sí, respectivamente. Se dirá que una variable es cuantitativa discreta o continua atendiendo a si, por la naturaleza del experimento planteado, los datos podrían en principio tomar una cantidad finita o numerable de valores distintos, en el primer caso, o si podrían tomar una cantidad infinita no numerable de valores, en el segundo caso. En síntesis:

- ☞ Variables estadísticas **cualitativas**: **nominales** (no se contempla una ordenación de los valores que los datos pueden tomar) u **ordinales** (se contempla una ordenación de los valores que los datos pueden tomar).

☞ Variables estadísticas **cuantitativas**: **discretas** (se entiende que, en principio, los datos del correspondiente experimento podrían tomar una cantidad finita o infinita numerable de valores distintos) o **continuas** (se entiende que, en principio, los datos del correspondiente experimento podrían tomar una cantidad infinita no numerable de valores distintos).

Presentar un colectivo de datos en la forma $X = \{x_1, \ldots, x_n\}$ puede hacer que mucha información del experimento se pierda por el camino. Aportar una tabla con tantas entradas como se consideren necesarias, añadiendo eventual información extra, es un modo más adecuado de presentar los datos en muchos casos. Por ejemplo, trabajar con variables cuantitativas suele requerir especificar las unidades en las que los datos numéricos se expresan. Para ello, se debe dar información adicional sobre las medidas realizadas.

Ejemplo (contaminación tanque 1). Considérese el experimento consistente en medir las emisiones de CO_2 de un tanque concreto en el desfile militar del 12 de octubre de 2021, en Madrid, en intervalos de 10 segundos consecutivos durante 1 minuto y 40 segundos a partir de las 11:30 am (GTM). Imagínese que los datos recogidos por un grupo de expertas, expresados en gramos de CO_2 (gCO_2) y ordenados cronológicamente, son:

$$\begin{aligned} Y &= \{122, 118, 119, 117, 110, 121, 118, 119, 122, 121\} = \\ &= \{\ y_1\ ,\ y_2\ ,\ y_3\ ,\ y_4,\ y_5\ ,\ y_6\ ,\ y_7\ ,\ y_8,\ y_9\ ,\ y_{10}\} \ . \end{aligned}$$

En este experimento la ordenación temporal de los datos es muy importante, así como el tiempo de medición asociado a los mismos y las unidades en las que se expresan. Esta información se puede aportar en una representación tabular del colectivo como es la siguiente:

Medidas (gCO_2/tiempo)	122	118	119	117	110	121	118	119	122	121
Nombres (y_i)	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
Etiqueta cronológica (i)	1	2	3	4	5	6	7	8	9	10
Tiempo (segundos)	10	10	10	10	10	10	10	10	10	10

Cuadro 1: Tabla que recoge datos ficticios de emisión de CO_2 de un tanque durante un minuto y cuarenta segundos, en intervalos de diez segundos consecutivos, junto con el nombre, etiqueta cronológica y tiempo asociado a cada medición.

☞ Un colectivo de datos puede presentarse en formato de tabla, lo cual permite además adjuntar información extra considerada relevante.

En muchas ocasiones la cantidad de datos con las que se trabaja es enorme. Por ello, en lugar de presentar cada dato uno por uno, se suele recurrir a representaciones tabulares y gráficas que sinteticen gran parte de la información más relevante y vistosa.

La representación tabular por excelencia es la tabla de frecuencias. Para un colectivo de datos dado, esta consiste en una tabla con las entradas explicadas a continuación. En las primeras, se colocan los distintos **valores** que los datos toman, $v_1, \ldots, v_k$, ordenados[2] de menor a mayor (k sería el número de valores distintos que el colectivo toma). En la siguiente, se indica el número de veces que aparece cada valor, $n_1, \ldots, n_k$ (**frecuencia**

[2]En caso de que en el experimento asociado se haya considerado un orden. De no ser así, se colocan como se quiera.

absoluta). Después, se anota el número de veces que aparecen valores menores o iguales a cada valor, $N_1, \ldots, N_k$ (**frecuencia absoluta acumulada**). Tras ello, se indica la proporción de ocasiones que aparece cada valor, $f_1, \ldots, f_k$ (**frecuencia relativa**). Y, finalmente, se indica la proporción de ocasiones que aparecen valores menores o iguales a cada valor, $F_1, \ldots, F_k$ (**frecuencia relativa acumulada**). Naturalmente, las frecuencias absolutas y relativas acumuladas solo tienen sentido para datos que admiten un orden. De no ser ese el caso, esas entradas se omitirían.

Ejemplo (contaminación tanque 2). En el colectivo de datos del ejemplo anterior, Y, se tienen seis valores distintos. Llámense estos v_j, con $j = 1, \ldots, 6$, y $v_1 < \cdots < v_6$. La tabla de frecuencias asociada a ese colectivo es la siguiente:

Medidas (gCO_2/10secs)	110	117	118	119	121	122
Nombres (v_j)	v_1	v_2	v_3	v_4	v_5	v_6
Frecuencia absoluta (n_j)	1	1	2	2	2	2
Frecuencia absoluta acumulada (N_j)	1	2	4	6	8	10
Frecuencia relativa (f_j)	0.1	0.1	0.2	0.2	0.2	0.2
Frecuencia relativa acumulada (F_j)	0.1	0.2	0.4	0.6	0.8	1

Cuadro 2: Tabla de frecuencias asociada al colectivo de datos de la Tabla 1.

☞ La **tabla de frecuencias** sintetiza la información de un colectivo de datos a través del número y la proporción de apariciones de cada valor distinto.

Otra forma de resumir o, simplemente, mostrar la información de un colectivo de datos es acudir a representaciones gráficas. Hay una enorme variedad de ellas. A continuación se muestra un diagrama de barras, uno de sectores y un histograma con conteo absoluto, en ese orden, asociados al colectivo de datos del ejemplo anterior, Y.

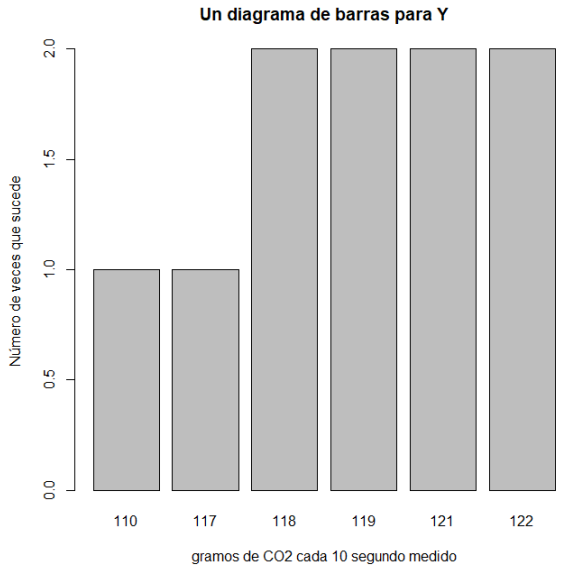


Figura 1: Diagrama de barras asociado al conjunto de medidas Y, el cual recoge 10 medidas ficticias de la emisión de CO_2 (en gramos) estimada cada 10 segundos consecutivos durante 100 segundos.

☞ En los **diagramas de barras** se dibuja tantas barras como valores distintos haya, siendo la altura de cada barra igual a la frecuencia absoluta de cada uno de ellos. El ancho de la barra no tiene significado alguno y tampoco importa el orden.

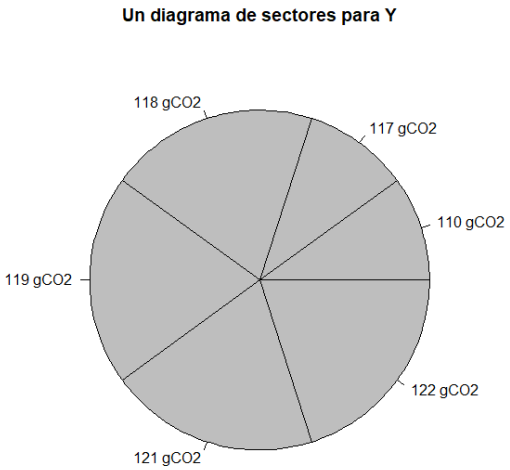


Figura 2: Diagrama de sectores asociado al conjunto de medidas Y, el cual recoge 10 medidas ficticias de la emisión de CO_2 (en gramos) estimada cada 10 segundos consecutivos durante 100 segundos.

☞ El **diagrama de sectores** de un colectivo de datos dado se basa en asociar a cada valor distinto del mismo un sector circular cuyo ángulo sea la frecuencia relativa asociada al correspondiente valor, multiplicada por $360°$ o 2π radianes.

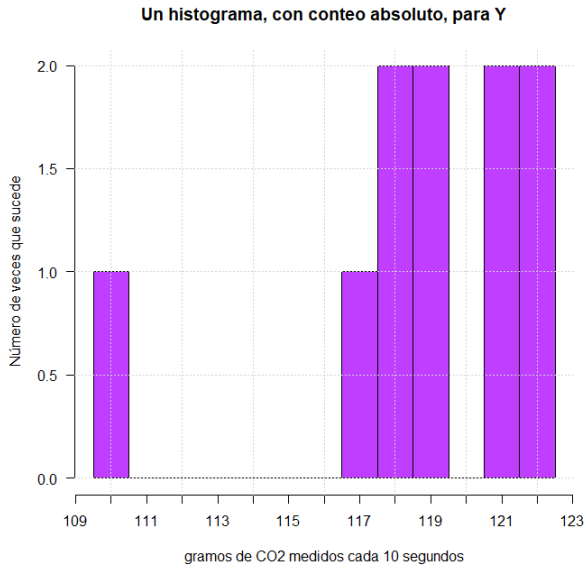
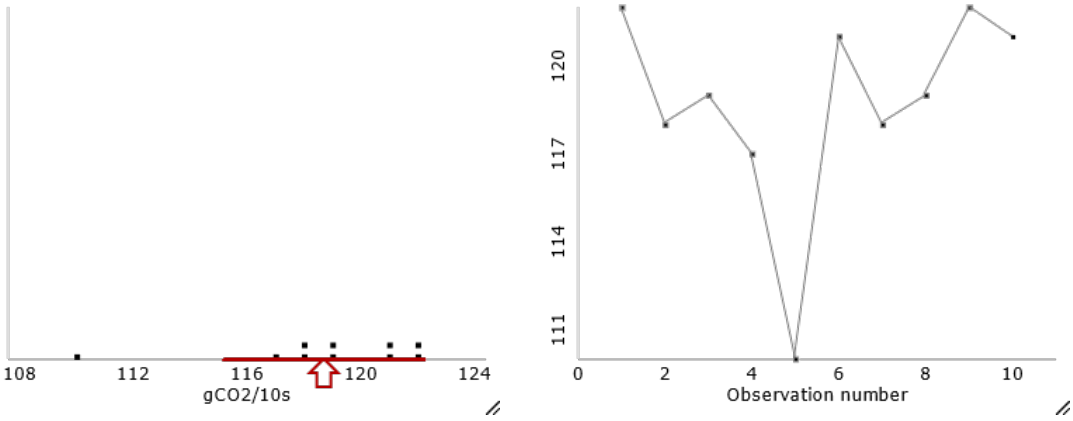


Figura 3: Un histograma, con conteo absoluto, asociado al conjunto de medidas Y, el cual recoge 10 medidas ficticias de la emisión de CO_2 (en gramos) estimada cada 10 segundos consecutivos durante 100 segundos. En este caso, para la base de cada rectángulo se han tomado segmentos centrados en cada dato y con ancho igual a 1.

☞ Para representar un **histograma con conteo absoluto** (respectivamente, **con conteo relativo**) de un colectivo de datos dado, el primer paso es agrupar los valores en los intervalos preestablecidos y no superpuestos que se quieran. Tras ello, se levanta sobre cada uno de los intervalos, dispuestos en el eje horizontal de unos ejes de coordenadas cartesianas, una barra cuya altura haga que el área de la misma sea igual a la suma de las frecuencias absolutas (respectivamente, relativas) de los datos que caen en dicho intervalo.

Le daremos especial importancia a las representaciones anteriores, además del diagrama de caja, que se introducirá más adelante. En cualquier caso, existen otras muchas representaciones gráficas, como pueden ser el diagrama de puntos y la representación gráfica como serie temporal, mostradas a continuación.



(a) Diagrama de puntos, con indicaciones, para Y.

(b) Serie temporal asociada a Y.

Figura 4: A la izquierda: diagrama de puntos asociado al conjunto de datos Y, que recoge 10 medidas ficticias de la emisión de CO_2 (en gramos) estimada cada 10 segundos consecutivos durante 100 segundos. La flecha roja añadida indica el valor de la media de Y, $\overline{Y}$, y la barra roja delimita el intervalo centrado en la media y de longitud igual a la desviación típica, s_Y. A la derecha: representación de esos mismos datos como serie temporal. Las representaciones gráficas se obtuvieron mediante la aplicación de este enlace.

☞ Para hacer un **diagrama de puntos** de un colectivo de datos dado, se toma un eje horizontal en el que representar los datos numéricos. Después, sobre cada valor distinto que aquel tome, se va dibujando un punto sobre otro por cada vez que dicho valor se repite en el conjunto de datos con el que se está trabajando.

☞ Para hacer una representación como **serie temporal** de un grupo de datos dado que cuente con un orden cronológico específico, lo primero es tomar ejes de coordenadas cartesianas. En el eje horizontal se representa el orden temporal en el que tuvieron lugar los datos y en el vertical, los valores de los datos. Para cada instante temporal, se sitúa una marca en el punto del plano caracterizado por el par que forman cada "instante" de tiempo y el valor que la variable tomó en ese instante de tiempo. Es habitual, además, unir las marcas consecutivas por líneas rectas.

Algo fundamental a comprender es que no todas las representaciones gráficas son adecuadas para una variable estadística concreta sometida a estudio. Dependiendo de los valores distintos que los datos puedan tomar (lo cual recibe el nombre de **modalidades** de la correspondiente variable estadística), en ocasiones ya se sabría de antemano que

algunas representaciones gráficas no cumplirán con el objetivo de sintetizar o hacer visible o intuitiva la manera en que se pueden distribuir los datos de un colectivo.

Ejemplo (contaminación tanque 3). Retomemos el colectivo de los ejemplos anteriores, en los que se tenía las mediciones $Y = \{122, 118, 119, 117, 110, 121, 118, 119, 122, 121\} = \{y_1, \ldots, y_{10}\}$ (gramos de CO_2 cada 10 segundos). Para este grupo de datos concreto, puede apreciarse en las figuras anteriores que todas las representaciones gráficas parecen tener sentido. No obstante, la adecuación de cada representación gráfica depende de varias cuestiones, como puede ser el contexto experimental, y no solo de un colectivo de datos concreto que nos hayan dado.

Para empezar, habría que tener en cuenta la precisión que en el proceder experimental se tiene para medir o estimar esos gramos de CO_2 y la que se podría llegar a tener. Es decir, en este ejemplo, cabe plantearse si podrían haberse obtenido mediciones más precisas (esto es, haber recogido del experimento datos con decimales y no solo hasta las unidades) con unos instrumentos o técnicas experimentales más avanzadas.

La importancia de esto radica en que, de ser así, el colectivo Y podría en realidad haber tomado, por poner un ejemplo, la forma siguiente:

$$Y_{\text{un decimal}} = \{122.1, 118.3, 119.2, 117.1, 110.9, 121.5, 118.4, 119.7, 122.1, 121.8\} \ .$$

De este modo, habríamos pasado de trabajar con $m = 10$ datos y $k = 6$ valores distintos en Y, a tener $m = 10$ datos y $k_{\text{un decimal}} = 9$ valores distintos en $Y_{\text{un decimal}}$. Así, al haber tenido en cuenta mayor precisión, se ha pasado de tener 6 valores distintos a tener 9 valores distintos. Es decir, se tendrían casi tantos valores distintos como datos hay en el colectivo $Y_{\text{un decimal}}$.

En efecto, en $Y_{\text{un decimal}}$ solo se repite el primer y el penúltimo dato, que son 122.1 gCO_2. El resto son todos distintos, a pesar de que en Y algunos eran iguales. Más aún, si se tuviera más y más precisión, es decir, si se pudieran conseguir más y más decimales en los datos, entonces idealmente se hubiera podido recoger colectivos de datos con dos, tres decimales ($Y_{\text{dos decimales}}$, $Y_{\text{tres decimales}}$), etc. **La cuestión es que sería de esperar que en algún momento todos los datos sean distintos entre sí al ir precisando cada vez más decimales**.

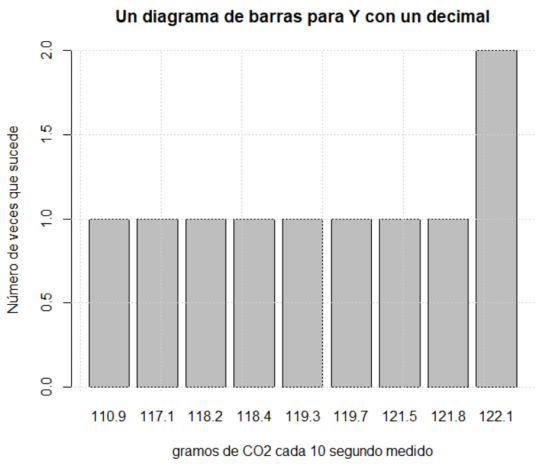
De ser así, ¿qué ocurriría con las representaciones gráficas del diagrama de barras y el diagrama de sectores? Al ser prácticamente todos los datos distintos unos de otros, casi todos los valores distintos recogidos aparecerían una sola vez. En la Figura 5 puede apreciarse que la información que aporta el diagrama de barras y el de sectores es prácticamente la misma que aporta la lista completa de datos, porque apenas hay repeticiones.

Es toda esta cuestión lo que realmente caracteriza a las variables estadísticas cuantitativas llamadas continuas. Los diagramas de barras y de sectores resultan potencialmente inadecuados o no apropiados.

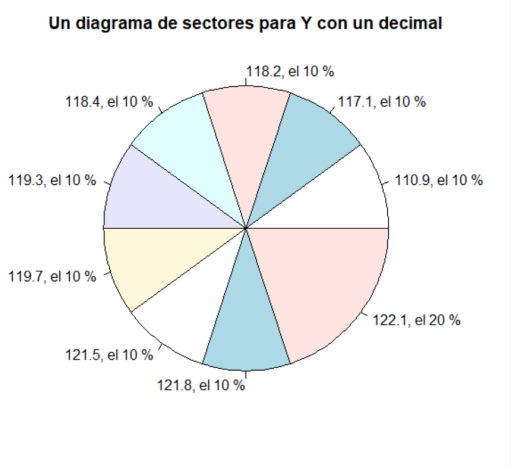
☞ Algo característico de las variables estadísticas aleatorias cuantitativas (o numéricas) continuas es que, por la naturaleza del experimento planteado, al considerar cada vez más precisión sobre las mediciones, cabe esperar que todos los datos de cualquier colectivo que se recoja tiendan a ser distintos entre sí.

De hecho, para este tipo de variables el proceder habitual es agrupar los posibles valores en unos intervalos (llamados **clases**) preestablecidos y no superpuestos — justo lo que se hace en un histograma, por eso un histograma es desde luego una representación adecuada para variables de naturaleza continua—. Es de destacar que la tabla de frecuencias podría entonces hacerse para esas clases, entendiéndolas como modalidades de una nueva variable estadística.

Por poner un ejemplo, podrían escogerse los intervalos siguientes: $[109, 111]$, $[116, 117.5)$, $[117.5, 118.5]$, $(118.5, 119.5]$, $[120.5, 121.5)$ y $[121.5, 122.5]$. De una parte, viendo cuántos datos o qué proporción de ellos cae en cada intervalo, se podría hacer el histograma con conteo absoluto o con conteo relativo, asociado a esa clasificación, para los datos iniciales. Sin embargo, también se podría hacer el diagrama de barras y de sectores para las clases, entendidas como valores de otra variable estadística. Ello, no obstante, no es muy adecuado. Además de los motivos anteriores, con tales diagramas se pierde la intuición de la posición relativa de los intervalos.



(a) Diagrama de barras para $Y_{\text{un decimal}}$.



(b) Diagrama de sectores para $Y_{\text{un decimal}}$.

Figura 5: Diagrama de barras (izquierda) y de sectores (derecha) asociados al conjunto de medidas $Y_{\text{un decimal}}$, el cual recoge 10 medidas ficticias de la emisión de CO_2 (en gramos) estimada cada 10 segundos consecutivos durante 100 segundos, con un decimal de precisión.

A parte de los histogramas, acudir a lo que se conoce como diagrama de caja es otro modo habitual de representar colectivos de datos provenientes de una variable estadística de naturaleza continua, siendo aquella también adecuada para variables numéricas discretas. Para introducir esta última representación, se requiere de ciertas herramientas previas.

ii. Algunas operaciones habituales y sus propiedades.

Llamaremos "operación" al resultado de hacer cualquier cuenta que involucre a un colectivo de datos concreto. Coger el primer dato y multiplicarlo por dos, quedarse con un solo dato o tomar el último y restarle el primero son ejemplos de operaciones. En el contexto de la Inferencia Estadística, cuando el conjunto de datos se trata de una muestra de la población sometida a estudio, las operaciones reciben el nombre de **estadísticos**.

Las operaciones se emplean para tratar de sintetizar información sobre la distribución de los datos o para señalar cualidades suyas. Además, aquellas resultan de especial interés para obtener rangos de los que los datos (numéricos) no "suelan" escaparse o, más bien, para dar cuenta nuevamente de cómo están distribuidos los datos de un colectivo dado. En ello, intervienen medidas u operaciones llamadas de posición (las de tendencia central son un caso particular de estas) y las llamadas de dispersión.

Un ejemplo de operación muy útil es el **máximo** o el **mínimo** de un colectivo de datos (claro está, siempre que sean datos provenientes de una variable estadística cuantitativa). El **recorrido** se define como la diferencia entre el máximo y el mínimo y es un ejemplo de medida de dispersión.

En el marco de la Estadística Descriptiva, las operaciones de posición y de dispersión se utilizan conjuntamente para dar resumida cuenta de la distribución de los datos. Por ejemplo, la media y la desviación típica se utilizan habitualmente como las indicaciones en rojo de la Figura 4a. Una representación más completa de la distribución de los datos la proporciona el diagrama de caja, el cual permite además identificar a simple vista lo que se denomina como "valores atípicos".

En el contexto probabilístico, la media y la varianza guardan relación con parámetros que dan cuenta de cierta tendencia central y dispersión de poblaciones. En Inferencia Estadística, cuando el colectivo de datos se corresponde con una muestra (aleatoria simple), la media y la varianza muestral se emplean como estimadores de esos parámetros poblacionales que, en muchos casos, son innacesibles experimentalmente.

Operaciones de tendencia central y de posición.

Es adelante supondremos que nos han dado un colectivo de datos numéricos denotado por $X = \{x_1, \ldots, x_n\}$, cuyos valores distintos son $v_1 < \ldots < v_k$, ordenados de menor a mayor. Es importante tener siempre presente el número de datos (n) y el número de valores distintos (k) que tiene el colectivo con el que se trabaja. Las medidas de tendencia central más utilizadas son las siguientes:

- ☞ La **media**, definida como la suma de todos los datos entre el número total de datos. Esta se denota por $\overline{X}$ y se puede escribir de las siguientes maneras:

$$\overline{X} = \frac{x_1 + \cdots + x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n} = \sum_{j=1}^{k} \frac{n_j v_j}{n} = \sum_{j=1}^{k} f_j v_j \ .$$

- ☞ La **moda**, definida como el valor o valores más repetidos en el colectivo y denotada por $\mathrm{Mo}(X)$. Nótese que esta puede no ser un único valor.

- ☞ La **mediana**, por la cual entenderemos el primer valor del colectivo que deja por debajo a la mitad de los datos, incluyéndose el propio valor. Se denotaría $\mathrm{Me}(X)$.

Supón que tenemos un colectivo con n datos, siendo todos ellos distintos. Si n es impar, entonces la mediana se corresponde con el dato o valor que cae justo en la mitad, $v_{\frac{n+1}{2}}$. Si n es par, no habría un valor que caiga justo en la mitad y, de acuerdo con la definición que hemos dado, la mediana sería $v_{\frac{n}{2}}$, ya que su frecuencia relativa acumulada sería $F_{\frac{n}{2}} = f_1 + \ldots + f_{\frac{n}{2}} = 0.5$ (porque estamos sumando $n/2$ veces $1/n$, que es la frecuencia

relativa de todos los valores de este caso).

Lo habitual cuando hay un valor cuya frecuencia relativa acumulada coincide con el 0.5, en cambio, es considerar la mediana como el valor intermedio a ese y el siguiente. Esto es, en el caso anterior (n par) se tomaría $(v_{\frac{n}{2}} + v_{\frac{n+1}{2}})/2$. Ello se debe a la pretensión de que la mediana sea una medida de tendencia central.

Por otro lado, recuérdese que, al trabajar con variables numéricas de naturaleza continua, se podría tratar con las clases directamente (los intervalos preestablecidos no superpuestos), y, si se quiere, incluso entender esas clases como modalidades de una nueva variable estadística. La tabla de frecuencias podría hacerse para estas clases y tendría sentido hablar de las frecuencias acumuladas de estas. En particular, tendría sentido hablar de la mediana, que sería uno de los intervalos o clases, de acuerdo con la definición que acabamos de dar. En la práctica, en cambio, se suele escoger un valor concreto de tal intervalo (interpolación) y no entraremos en ello. Con el procedimiento aquí dado, simplemente calcularíamos la mediana del colectivo concreto con el que se trabaje y no de sus clases aunque la variable estadística de partida sea entendida continua.

La mediana es un caso particular de una familia más amplia de medidas de posición, los percentiles. También son casos particulares de los percentiles los deciles y los cuartiles, cuyo uso es frecuente.

- ☞ Si α es un número natural entre 1 y 99 ($\alpha = 1, \ldots, 99$), el **percentil** α de un colectivo X es el primer valor del mismo que deja al $\alpha\,\%$ de los datos por debajo, incluyendo al propio valor. Se denota $\mathrm{Pe}_\alpha(X)$.
- ☞ Si $\beta = 1, \ldots, 3$, se define el primer, segundo o tercer **cuartil** como $Q_1(X) = \mathrm{Pe}_{25}(X)$, $Q_2(X) = \mathrm{Pe}_{50}(X) = \mathrm{Me}(X)$ y $Q_3(X) = \mathrm{Pe}_{75}(X)$. De manera similar, se definen los **deciles** (desde el primero hasta el noveno) como $D_\gamma(X) = P_{\gamma\cdot 10}(X)$ para cada $\gamma = 1, \ldots, 9$.

Operaciones de dispersión.

Entre las medidas de dispersión más importantes y utilizadas se encuentra la varianza.

- ☞ Se define la **varianza** de un colectivo X, denotada por $\mathrm{Var}(X)$ o σ_X^2, como el promedio de la distancia al cuadrado de los datos a la media. Desarrollando el cuadrado, es fácil llegar a la siguiente fórmula:

$$\mathrm{Var}(X) = \frac{(x_1-\overline{X})^2 + \cdots + (x_n-\overline{X})^2}{n} = \sum_{i=1}^{n} \frac{(x_i-\overline{X})^2}{n} =$$
$$= \sum_{i=1}^{n} \frac{x_i^2}{n} - 2\overline{X}\sum_{i=1}^{n} \frac{x_i}{n} + \sum_{i=1}^{n} \frac{\overline{X}^2}{n} = \overline{X^2} - 2\overline{X}^2 + \overline{X}^2 = \overline{X^2} - \overline{X}^2 \,,$$

  donde se ha denotado por X^2 al colectivo de datos formado por cada dato de X, pero al cuadrado. Es decir, $X^2 = \{x_1^2, \ldots, x_n^2\}$.

Asimismo, cabe resaltar que si los datos numéricos tuvieran unidades, entonces la media se mide en las mismas unidades que los datos (ya que simplemente se suman los datos y se dividen por algo adimensional). Sin embargo, para calcular la varianza se suman

números que resultan de elevar al cuadrado. Así, las unidades de la varianza son las mismas que los datos, pero al cuadrado. Es por ello que resulta natural y necesario introducir la desviación típica, que sí se mide en las mismas unidades que los datos.

☞ La **desviación típica**, denotada por σ_X, es la raíz cuadrada de la varianza, es decir, $\sigma_X = \sqrt{\sigma_X^2}$. Se mide en las mismas unidades que los datos.

La desviación típica suele utilizarse junto a la media para sintetizar un colectivo de datos en un intervalo: aquel centrado en la media y de longitud igual a la desvación típica. Es decir,

$$\left[\overline{X} - \frac{\sigma_X}{2}, \overline{X} + \frac{\sigma_X}{2}\right] ,$$

siendo también habitual resaltar en ese intervalo la media cuando se representa gráficamente. Un ejemplo de ello son justamente las indicaciones en rojo en el diagrama de puntos en la Figura 4a.

En realidad, lo que en esa figura se está representado es el intervalo anterior, pero usando lo que se conoce como desviación típica muestral, que denotaremos por s_X, en lugar de σ_X. Es muy habitual que programas estadísticos utilicen únicamente la varianza muestral (que denotaremos por s_X^2) y la desviación típica muestral.

☞ La **varianza muestral**, s_X^2, se define al igual que la varianza, pero diviendo por $n-1$ en lugar de entre n. Es decir:

$$s_X^2 = \frac{n}{n-1}\sigma_X^2 .$$

☞ Se llama **desviación típica muestral**, s_X, a la raíz cuadrada de la varianza muestral, esto es: $s_X = \sqrt{s_X^2}$.

La importancia de la varianza y desviación típica muestrales radica en que juegan un papel especial en el contexto de la Inferencia Estadística. Cuando el colectivo de datos se corresponde con una muestra (aleatoria simple) y con ella se pretende estimar la varianza de una población, es la varianza muestral la que lo hace de manera óptima.

Aparte de la varianza y la desviación típica (muestral o a secas), otra medida de dispersión frecuente es el coeficiente de variación, pensada para colectivos de datos que no cambian de signo (o bien positivos o bien negativos).

☞ El **coeficiente de variación** (de Pearson), que denotaremos $\mathrm{CV}(X)$, se define como el resultado de dividir la desviación típica por el valor absoluto de la media. También se suele usar la desviación típica muestral en lugar de la desviación típica.

Nótese que el coeficiente de variación es siempre positivo y que es adimensional. Cuando toma valores altos, indica que la desviación típica es mucho mayor que la media; valores bajos, indican que la desviación típica es mucho menor que la media.

Como medida de dispersión que no involucra a la media, es destacable el rango intercuartílico.

☞ Se define el **rango intercuartílico**, denotado por $\mathrm{IQR}(X)$, como la diferencia entre el tercer cuartil y el primer cuartil. Es decir, $\mathrm{IQR}(X) = Q_3(X) - Q_1(X)$.

La última representación gráfica que veremos hace vistosa la distribución de los datos a través de los cuartiles e involucra al rango intercuartílico. Este último se usa para identificar lo que se conoce como valores atípicos (outliers).

☞ Se dice que un **valor es atípico** si cae fuera del intervalo siguiente:

$$[Q_1(X) - 1.5\cdot\mathrm{IQR}(X), Q_3(X) + 1.5\cdot\mathrm{IQR}(X)] \ .$$

Diagrama de caja.

☞ Para hacer un **diagrama de caja** se debe calcular primero los tres cuartiles, así como el rango intercuartílico. Tras ello, se toma un eje vertical de referencia en el que indicar los datos y se dibuja una caja cuyo lado inferior esté a la altura del primer cuartil y cuyo lado superior esté a la altura del tercer cuartil. El ancho de la caja no tiene significado alguno. También se marca la mediana, es decir, el segundo cuartil, como una línea horizontal dentro de la caja. Finalmente, por los extremos de la caja impuestos por el primer y tercer cuartil salen dos segmentos, uno por cada lado. El inferior irá hasta $Q_1(X) - 1.5\mathrm{IQR}(X)$ en caso de que hubiera valores atípicos por debajo de este valor. De no ser así, tal segmento irá tan solo hasta el mínimo del colectivo, mín(X). De manera similar, el segmento superior terminará en $Q_3(X) + 1.5\mathrm{IQR}(X)$, siempre que haya valor atípicos por encima de ese valor. De no ser ese el caso, el segmento terminará en máx(X). Finalmente, se indican los valores atípicos con una marca.

Ejemplo (contaminación tanque 4). Para $Y = \{122, 118, 119, 117, 110, 121, 118, 119, 122, 121\} = \{y_1, \ldots, y_{10}\}$ (gramos de CO_2 cada 10 segundos), el colectivo de los ejemplos anteriores, las operaciones tienen los siguientes resultados e interpretaciones. Recuérdese que se tienen $m = 10$ datos y $k = 6$ valores distintos.

Los diez segundos más contaminantes fueron el primero y el penúltimo, pues alcanzan el valor máximo, máx$(Y) = 122$ gCO_2. El periodo menos contaminante fue el quinto, donde se tomó el mínimo: mín$(Y) = 110$ gCO_2. El recorrido del colectivo es la distancia entre el máximo y el mínimo, es decir, 12 gCO_2.

La media, que puede ser calculada a mano a partir de la lista de datos o bien a partir de la tabla de frecuencias, resulta ser $\overline{Y} = 118.7$ gCO_2. En efecto:

$$\overline{Y} = \frac{y_1 + \cdots + y_{10}}{10} = \frac{110\cdot 1 + 117\cdot 1 + 118\cdot 2 + 119\cdot 2 + 121\cdot 2 + 122\cdot 2}{10}\ \mathrm{gCO_2} = \\ = 118.7\ \mathrm{gCO_2}\ ,$$

En cuanto a la varianza, esta se puede calcular por la definición:

$$\sigma_Y^2 = \frac{(y_1-\overline{Y})^2 + \cdots + (y_{10}-\overline{Y})^2}{10} = \frac{(110-118.7)^2+(117-118.7)^2+2(118-118.7)^2}{10} + \\ + \frac{2(119-118.7)^2 + 2(121-118.7)^2 + 2(122-118.7)^2}{10}\ (\mathrm{gCO_2})^2 = \\ = \frac{1}{10}\left(8.7^2 + 1.7^2 + 2\cdot 0.7^2 + 2\cdot 0.3^2 + 2\cdot 2.3^2 + 2\cdot 3.3^2\right)\ (\mathrm{gCO_2})^2 = \\ = \frac{112.1}{10}(\mathrm{gCO_2})^2 \simeq 11.2\ (\mathrm{gCO_2})^2\ .$$

Algo más eficiente, quizá, hubiera sido calcular $\overline{Y^2}$, recordando que por Y^2 entendemos $Y^2 = \{y_1^2, \ldots, y_m^2\}$. La misma tabla de frecuencias para Y, pero con los valores al cuadrado, sirve para calcular la media de Y^2. Así, también se podría haber hecho:

$$\sigma_Y^2 = \overline{Y^2} - \left(\overline{Y}\right)^2 = \sum_{j=1}^{6} \frac{n_j v_j^2}{10} - \left(\sum_{j=1}^{6} \frac{n_j v_j}{10}\right)^2 ,$$

llegando al mismo resultado. La desviación típica es $\sigma_Y \simeq 3.35$ gCO$_2$ y la varianza muestral se puede calcular a partir de la varianza como:

$$s_Y^2 = \frac{m}{m-1}\sigma_Y^2 = \frac{10}{9}\sigma_Y^2 \simeq 12.45 \text{ (gCO}_2)^2 .$$

De ser así, la desviación típica muestral quedaría $s_Y \simeq 3.53$ gCO$_2$ (puedes comprobar que este es el valor que, por ejemplo, el software libre R entiende por desviación típica).

El coeficiente de variación resulta de dividir la desviación típica o la desviación típica muestral (la que se quiera, pero diciéndolo) por el valor absoluto de la media. En este caso, se tendría:

$$\text{CV}(Y) = \frac{s_Y}{\overline{Y}} \simeq 0.0297 \quad \text{o bien} \quad \text{CV}(Y) = \frac{\sigma_Y}{\overline{Y}} \simeq 0.0282 .$$

Volviendo a las medidas de posición, la moda en este caso son cuatro valores $\text{Mo}(Y) = = \{118, 119, 121, 122\}$ (en gCO$_2$). En cuanto a los percentiles, se pueden calcular todos a partir de la tabla de frecuencias (Tabla 2). Para cada $\alpha = 1, \ldots, 10$, $\text{Pe}_\alpha(Y) = 110$ gCO$_2$; para cada $\alpha = 11, \ldots, 20$, $\text{Pe}_\alpha(Y) = 117$ gCO$_2$; para cada $\alpha = 21, \ldots, 40$, $\text{Pe}_\alpha(Y) = = 118$ gCO$_2$; para cada $\alpha = 41, \ldots, 60$, $\text{Pe}_\alpha(Y) = 119$ gCO$_2$; para cada $\alpha = 61, \ldots, 80$, $\text{Pe}_\alpha(Y) = 121$ gCO$_2$ y, para cada $\alpha = 81, \ldots, 99$, $\text{Pe}_\alpha(Y) = 122$ gCO$_2$.

Especial mención merecen los percentiles 10, 20, 40, 60 y 80. Pues 0.1, 0.2, 0.4, 0.6 y 0.8 son justamente las frecuencias relativas acumuladas de 110, 117, 118, 119, 121 y 122, respectivamente. De acuerdo con nuestra definición, no habría lugar a dudas y los percentiles serían en cada caso esos valores. Recuérdese, no obstante, que cuando ello ocurre se suele entender que es más apropiado tomar como percentil el valor intermedio al elegido anteriormente y el siguiente. A nosotros nos valdrá cualquier valor entre medias, incluyendo a los propios valores. Es decir, consideraremos correcto lo anterior.

Los deciles, los cuartiles y la mediana se obtienen como caso particular de los percentiles. La mediana sería $\text{Me}(Y) = 119$ gCO$_2$, el tercer cuartil $Q_3(Y) = 121$ gCO$_2$ y el primero, $Q_1(Y) = 118$ gCO$_2$. Con estos dos últimos se calcula la única medida de dispersión que no está referida a la media, de las que hemos visto: el rango intercuartílico. Es claro que $\text{IQR}(Y) = Q_3(Y) - Q_1(Y) = 3$ gCO$_2$.

Con estos últimos cálculos, estamos en disposición de representar el diagrama de caja para el colectivo Y. Lo primero, es identificar si hay valores atípicos superiores o inferiores. Recuérdese que un valor de Y se dice atípico si queda fuera del intervalo:

$$[Q_1(Y) - 1.5{\cdot}\text{IQR}(Y), Q_3(Y) + 1.5{\cdot}\text{IQR}(Y)] = [118 - 1.5{\cdot}3, 121 + 1.5{\cdot}3] = [113.5, 125.5] .$$

Así, como el máximo de nuestro colectivo es más pequeño que el extremo superior del intervalo ($\text{máx}(Y) = 122 < 125.5 = Q_3(Y) + 1.5{\cdot}\text{IQR}(Y)$), entonces no hay valores atípicos superiores y la rama superior que sale de la caja delimitada por el primer y el tercer

cuartil solamente llega hasta máx$(Y) = 122$ (ver Figura 6).

Por el lado inferior sucede, en cambio, que $Q_1(Y) - 1.5 \cdot \text{IQR}(Y) > \text{mín}(Y)$. En consecuencia, hay valores atípicos y la rama inferior llegará hasta $Q_1(Y) - 1.5 \cdot \text{IQR}(Y)$. Recordando que dentro de la caja hay que señalar también la mediana y que hay que marcar los valores atípicos (en este caso, el único valor atípico es el mínimo), ya se tiene toda la información para hacer el diagrama de caja de Y, que está en la siguiente figura.

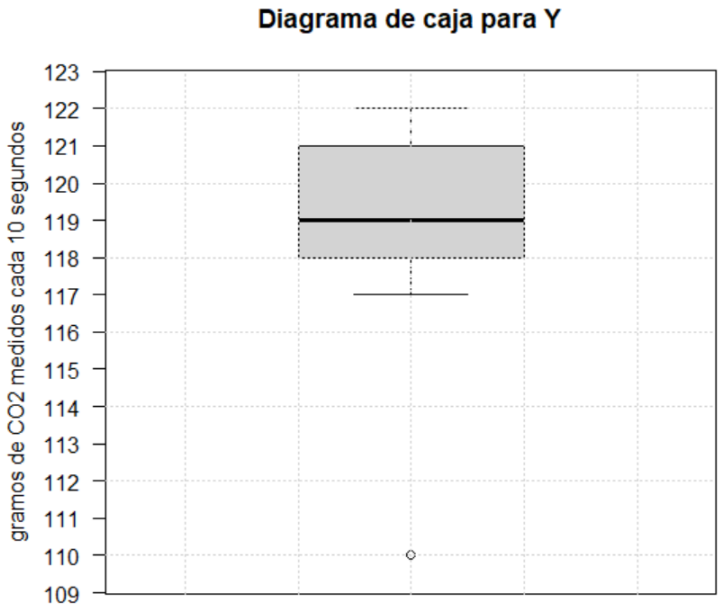


Figura 6: Diagrama de caja asociado al conjunto de medidas Y, el cual recoge 10 medidas ficticias de la emisión de CO_2 (en gramos) estimada cada 10 segundos consecutivos durante 100 segundos.

iii. Algunas cuestiones fundamentales.

Sobre la adecuación de cada operación según el tipo de variable estadística.

En definitiva, con el recurrente ejemplo de los gramos de CO_2 emitidos por el vehículo militar, hemos podido calcular todas las operaciones (o estadísticos) introducidas. Hay dos cuestiones a tener muy claras: ¿por qué hemos podido hacerlo? ¿Se pueden calcular todas las operaciones para cualquier tipo de variable?

Por ejemplo, para obtener la media solo necesitamos poder sumar los datos y dividirlos por un número natural (el número total de datos[3] [4]). ¿Podríamos hacer eso si hemos considerado la variable como cualitativa? Lógicamente, **NO**, ya que si hemos considerado la variable cualitativa es porque estamos asumiendo que los datos no son números, sino

[3] El número total de un grupo de datos dado es siempre finito, entero y positivo. Es decir, es siempre un número natural. Si llamamos n al número de datos de un conjunto, en notación matemática esto se puede escribir como $n \in \mathbb{N} = \{1, 2, 3, 4 \ldots, 731, \ldots, 1312, \ldots, 4 \cdot 10^{30}, \ldots\}$ y se lee "el número n pertenece al conjunto de los números naturales, $\mathbb{N}$". Date cuenta de que n en principio puede ser tan grande como quieras, pero $+\infty$ no está en el conjunto $\mathbb{N}$.

[4] Conviene que tengas también muy claro qué son los números enteros, $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. Es como el conjunto de los naturales, pero contando también con los números sin decimales negativos, además del cero.

etiquetas, nombres o factores que expresan cualidades o atributos. Habríamos eliminado cualquier carácter numérico de los datos al hacerlo.

En cambio, **SÍ** tiene sentido hacer la media para variables numéricas que se puedan sumar y dividir por un número natural, como puede ser cualquier número real, sea racional o irracional[5]. Si, además, es posible elevar los números al cuadrado (como es el caso de los números reales), entonces **también** podemos calcular la varianza y la desviación típica. En consecuencia, se puede calcular el coeficiente de variación porque se puede hacer la media y la desviación típica. De otra parte, como es posible ordenar los valores distintos ($v_1 < \ldots < v_k$) de menor a mayor (porque los números de la recta real se pueden ordenar), entonces tiene sentido hablar de frecuencias absolutas y relativas acumuladas. Por lo tanto, al poder hablar de frecuencias acumuladas, **también** tiene sentido calcular los percentiles (y quien dice percentil, dice también mediana, cuartiles, deciles y rango intercuartílico).

La única operación que queda por discutir es la moda. La moda solo depende del número de veces que aparece cada valor distinto de un conjunto de datos dado. Por lo tanto, **puede** calcularse para cualquier tipo de variable. Date cuenta de que la moda es la única operación que se puede hacer para variables estadísticas de tipo cualitativo nominal. Si tuviésemos una variable cualitativa ordinal, en cambio, **además** de la moda sería discutible si podría tener sentido calcular las frecuencias acumuladas. En tal caso, se podrían calcular los percentiles y el resto de operaciones que se deducen de los percentiles.

Sobre el carácter continuo o discreto de las variables estadísticas numéricas.

A todo esto, ¿cuál era la variable estadística asociada al experimento del tanque? El colectivo de datos propuesto es una concreción de la variable estadística "los gramos de CO_2".

Pues bien, ¿de qué tipo es esa variable? La mejor forma de identificar el tipo de variable con la que se esté tratando es tener completamente claro cuáles son sus modalidades. Esto es: los posibles valores que los datos del colectivo pueden tomar, más allá de los que hallan tomado en un ejemplo concreto. Para los gramos de CO_2, es evidente que los datos tienen que ser, para empezar, numéricos. Por lo tanto, un buen inicio es decir que la variable estadística "los gramos de CO_2" es cuantitativa.

Así, "solamente" nos hace falta responder a la pregunta: ¿es "los gramos de CO_2 emitidos" continua o discreta? De acuerdo con la definición que hemos dado, tenemos que preguntarnos si los posibles valores (las modalidades) forman un conjunto discreto (es decir, un conjunto con un número finito o infinito numerable[6] de elementos) o continuo (infinito no numerable[7]).

[5] Los números reales o los números de la recta real, con los que todes trabajamos constantemente, pueden dividirse en dos grupos: los racionales ($\mathbb{Q}$) y los irracionales ($\mathbb{I}$). Los números racionales son todas las fracciones (es decir, aquellos números que pueden expresarse como división entre números enteros, siendo el denominador distinto de cero) y los irracionales son aquellos números reales que no pueden expresarse como una fracción. En notación matemática, $\mathbb{Q} = \{\frac{a}{b}$, siendo a un número entero y siendo b un número entero distinto de cero$\}$.

[6] Un conjunto cualquiera se dice infinito numerable si puede contarse como los números naturales, es decir, si existe una correspondencia uno a uno con los números naturales. Por ejemplo, los número enteros y los racionales ($\mathbb{Z}$ y $\mathbb{Q}$) son infinitos numerables. Esto es, son discretos.

[7] Un conjunto cualquiera se dice no numerable si no puede ponerse en una lista como los naturales, es decir, si no existe una correspondencia uno a uno con los números naturales. Los números reales y lo

Pero... ¿cuántos gramos pesa una molécula de CO_2? Si pudiéramos cuantificar precisamente cuánto pesa una sola molécula de CO_2 y pudiéramos cuantificar precisamente cuántas moléculas de CO_2 expulsa el tanque cada diez segundos, entonces los posibles valores de "los gramos de CO_2" en ese experimento concreto serían múltiplos de la masa de una sola molécula de CO_2. Por ejemplo, si se emitieran un millón (10^6) de partículas de CO_2 en diez segundos y una molécula de CO_2 pesase $7.30654 \cdot 10^{-23}$ gramos, el tanque habría emitido $10^6 \cdot 7.30654 \cdot 10^{-23} = 7.3 \cdot 10^{-17}$ gramos de CO_2. Suponiendo que esto fuera así, las modalidades de los gramos de CO_2 sería:

$$N \cdot 7.30654 \cdot 10^{-23} \ , \quad \text{donde} \quad N \text{ es natural} \ .$$

En notación matemática, esto puede escribirse como $\{N \cdot 7.30654 \cdot 10^{-23}$ tales que N pertenece a $\mathbb{N}\}$ y se diría que las modalidades son "el conjunto de los números que se escriben como $N \cdot 7.30654 \cdot 10^{-23}$, siendo N un número natural".

De acuerdo con el párrafo anterior, la variable estadística "los gramos de CO_2" **sería discreta**. Los posibles valores que los datos pueden tomar forman una cantidad infinita, pero numerable: ocurre exactamente lo mismo que con los números naturales. Sin embargo, la discusión en torno a la Figura 5 sobre la **naturaleza continua** de esta variable se mantiene válida aquí[8]. Hay tantísimas moléculas de CO_2 y cada una de ellas pesa tan poco que hay una inevitable precisión finita experimental. En otras palabras, es prácticamente imposible conocer con absoluta precisión un dato. Además, al ir aumentando la precisión, cabe esperar que cada vez se vayan obteniendo más datos distintos entre sí, como se discutió anteriormente.

De hecho, cualquier magnitud física (longitud, área, volumen, tiempo, masa, temperatura, presión, etc) suele entenderse como una variable estadística continua. A pesar de que, como acabamos de ver y **con la definición que hemos dado** de variable estadística numérica "discreta" o "continua", la masa podría llegar a entenderse discreta.

Nosotras vamos a eludir todas estas discusiones complejas y trascendentes, para ser prácticas. Ello lo vamos a hacer **dando otras "definiciones" de variable estadística continua y discreta**.

☞ Nosotras entenderemos que una **variable estadística** es **continua** si se ha asumido que es apropiado o conveniente agrupar los datos en intervalos, por el motivo que sea. De otro lado, diremos que una **variable estadística** es **discreta** si se entiende oportuno operar con los datos resultantes sin agruparlos en intervalos.

Con esta nueva definición, si los datos de una variable estadística pudieran tomar **cualquier valor racional** en el intervalo $[0, 1]$, **por ejemplo**, aquella se debería entender continua, ¡a pesar de que los números racionales en el intervalo $[0, 1]$, $\mathbb{Q} \cap [0, 1]$ es infinito numerable! Al igual que podría aparecer el valor $1/3$, podrían aparecer los valores 0.3, 0.33, 0.333, 0.3333, etc. Entonces, ¡sería **muy apropiado** agrupar los datos en intervalos! Ello se debe a que podemos llegar a obtener datos tan cerca como queramos y que,

números irracionales son un ejemplo de ello ($\mathbb{R}$ y $\mathbb{I}$). Lo mismo pasa con cualquier intervalo de número reales, como puede ser $[0, 1] \subset \mathbb{R}$ y los irracionales en cualquier intervalo, como $\mathbb{I} \cap [0, 1]$. El símbolo "$\subset$" se lee "contenido en" y el símbolo "$\cap$" es una intersección (es lo mismo que la conjunción como operador lógico, pero para conjuntos).

[8] ¡Date cuenta de que he supuesto que la masa de una molécula de CO_2 es $7.30654 \cdot 10^{-23}$!... ¿No podría haber dado más decimales?

siendo distintos al fin y al cabo, en realidad no lo son tanto. Sobre todo si la precisión del experimento es finita, que siempre lo es. **Tienes que tener muy presente que, por ejemplo, toda magnitud física se entiende continua por este razonamiento**.

Ejemplo (el número de personas en paro en la península ibérica). Si contamos el número de personas en el paro de la península cada mes, ¿cuál es la variable estadística? La variable estadística sería "número de personas en el paro" y sus modalidades es cualquier número natural, $\mathbb{N} = \{1, 2, 3, 4, \ldots\}$. O, si se prefiere, desde 1 hasta n, siendo n el número máximo de personas que ha habido, hay y habrá en la península ibérica, es decir: $\{1, 2, 3, 4, 5 \ldots, n\}$. Se trata de una variable numérica, desde luego. Con la nueva definición, ¿esta variable sería continua o sería discreta? Pues depende de si entiendes apropiado trabajar con intervalos, pero **con los número naturales** ($\mathbb{N}$) no ocurre lo mismo que con los números racionales: **no puedes obtener datos tan cerca como quieras**. Así, si las modalidades son los números naturales, **es habitual considerar la variable como discreta**. Pero **si vas a clasificar esos valores en intervalos o entiendes que ello es apropiado o necesario, deberías decir que es continua**.

Ejemplo (la proporción como variable estadística, caso 1). Si nos preguntamos ahora, en lugar de por el número de personas que hay en paro, por la proporción de personas que hay en paro en la península ibérica... ¿cuál sería la variable estadística? La variable estadística sería ahora "proporción de personas en paro". Si los datos están expresados en tanto por uno (en vez de tanto por ciento, %), entonces los posibles valores que toman los datos están entre 0 y 1, incluidos. El valor cero significaría que ninguna persona de la península ibérica está en paro, mientras que el valor 1 significaría que todas las personas de la península ibérica están en paro. Para calcular la proporción de personas en paro, simplemente tengo que dividir el número de personas que está en paro entre el número total de personas de la península ibérica. Es decir... en principio los posibles datos son todas las fracciones en el intervalo $[0, 1]$. En otras palabras, las modalidades de esta variable estadística es $\mathbb{Q} \cap [0, 1]$. De nuevo, con la definición anterior debería entenderse discreta... pero debe estar claro que sería más apropiado clasificar los valores por intervalos o clases y que, de acuerdo a la nueva definición de variable continua, la proporción puede entenderse continua. Como podemos tener datos tan cerca como queramos (variando las personas en paro y el número de personas totales en la península), sería más apropiado clasificar los datos, esto es, las proporciones de personas en paro, en intervalos.

Ejemplo (la proporción como variable estadística, caso 2). Supón que estamos tomando muestras de agua de mar en las costas de Perú, donde recientemente ha habido vertidos masivos de petróleo de los que es responsable la empresa Repsol. Supón que estamos interesadas en observar, sobre cada muestra de "agua", la variable estadística "proporción de volumen de la muestra que es petróleo". A diferencia del caso anterior, aquí en principio se puede tomar cualquier valor real entre el 0 y el 1. Es decir, las modalidades son $[0, 1]$, incluyendo racionales e irracionales. Si la muestra de agua hubiera sido tomada en un cilindro de altura $h = 1$ cm y radio $r = 1$ cm, entonces el volumen de la muestra sería $\pi r^2 h = \pi$ cm^3 (cm=centímetros). Si el petróleo estuviera concentrado en un cubo que esté contenido en ese cilindro y de lados de longitud $l = 1/5$ cm, entonces el volumen del cubo sería $1/125 = 1/5^3$ cm^3. ¿Cuál sería en ese caso la proporción de volumen de la muestra que es petróleo? Sería la división entre el volumen que es petróleo entre el volumen de la muestra, es decir, $1/125$ cm^3 entre π cm^3. Este es un número en el intervalo $[0, 1]$, pero que no es racional (porque el número π no es racional, esto es: no

puede escribirse como una fracción). Por lo tanto, en este ejemplo se pueden tomar todos los valores en el intervalo $[0, 1]$, racionales e irracionales, y, tanto con la definición anterior como con la nueva, la variable estadística "proporción de volumen" podría ser considerada continua.

Sobre la adecuación de cada gráfica según el tipo de variable estadística.

Ahora nos planteamos otra cuestión esencial en nuestro propósito, dentro de la Estadística Descriptiva, de hacer accesible la información recogida en un grupo de datos y de hacerlo de la manera más adecuada posible: ¿son todas las representaciones gráficas igualmente apropiadas para cualquier tipo de variable? Evidentemente, **NO**. A continuación se revisará los distintos tipos de variables, una a una, y se discutirá qué representaciones gráficas cabría esperar que, en principio, fueran adecuadas y cuáles no. Y, sobre todo, se discutirá brevemente por qué. Nos limitaremos a los diagramas de sectores, de barras, de puntos, de caja y a los histogramas.

Para empezar, si tenemos una variable cualitativa, entonces no estamos entendiendo los datos como números, sino como nombres. Dado que los datos pierden todo posible carácter numérico por esta elección, no es apropiado hacer ninguna representación gráfica en la que se utilice un eje para representar los datos. Por lo tanto...

- ☞ Para una variable estadística entendida como **cualitativa**, no son apropiados un diagrama de puntos, ni un diagrama de caja, ni un histograma.

De las representaciones gráficas planteadas, solo quedaría discutir la adecuación del diagrama de barras y el de sectores. Ello depende de si la variable es nominal u ordinal. Si fuera ordinal, en un diagrama de sectores se pierde la noción del orden... por ello, esa representación no sería muy adecuada. Sin embargo, si se utiliza un diagrama de barras y las barras de cada valor distinto se ordenan de acuerdo al orden de esa variable cualitativa ordinal, en la representación gráfica no solo no se pierde esa ordenación de los datos, sino que se hace muy visible.

- ☞ Para variables **cualitativas ordinales**, el **diagrama de sectores** no es del todo apropiado, porque se pierde la intuición visual del orden de los datos. En cambio, el **diagrama de barras** puede respetar esa ordenación de los datos si en la representación se colocan los valores distintos y sus respectivas barras conforme al orden de la variable.

De otro lado, para variables cualitativas nominales pasa lo contrario: el diagrama de sectores es apropiado porque no da lugar a que se piense, al verlo, que hay un orden de los datos. Sin embargo, el diagrama de barras puede no ser del todo adecuado, pues el orden de los datos en la gráfica podría cambiarse a placer, dando lugar a distintas gráficas según el orden escogido.

- ☞ Para variables **cualitativas nominales**, el **diagrama de barras** no es muy adecuado, puesto que dependiendo de como se coloquen los datos se pueden obtener gráficas distintas. Además, al ver una gráfica concreta podría pensarse que hay una ordenación de los datos. De otro lado, el **diagrama de sectores** suele considerarse más apropiado. Ello se debe, justamente, a que no da pie a que se piense que hay una ordenación de los datos.

Ya solo nos quedar discutir la adecuación de las gráficas para variables numéricas. Lo primero a notar es que los números (reales, que son con los que trabajaremos) están

ordenados. Por lo tanto, un diagrama de sectores no es adecuado porque dificulta la visualización de ese orden (ver, por ejemplo, la Figura 4b). A pesar de que un diagrama de barras pudiera llegar a respetar ese orden, como sucede en la Figura 4a... el diagrama de barras no permite visualizar la distancia entre los distintos datos. Dicho más fino, se pierde la intuición de la posición relativa de los datos.

☞ Para variables **cuantitativas** (es decir, numéricas), ni un **diagrama de sectores** ni uno de **barras** son adecuados. El segundo, porque pierde la intuición de la distancia entre los datos; el primero, además de eso, pierde también el orden de los números.

Por otro lado, en el resto de representaciones gráficas se utiliza un eje (la recta real) para representar los datos. Por lo tanto, no se siguen teniendo los problemas anteriores. No obstante, esto no quiere decir que todas sean adecuadas siempre. Ello puede depender de si la variable numérica es considerada continua o discreta[9].

☞ Para variables cuantitativas **continuas**, tanto un **histograma** como un **diagrama de caja** son representaciones adecuadas. El histograma, precisamente, porque es resultado de agrupar los datos en intervalos. El diagrama de caja, porque básicamente utiliza los cuartiles para delimitar también los valores por intervalos. Sin embargo, un **diagrama de puntos** es o bien inadecuado o bien potencialmente inadecuado, porque cabe esperar que todos los valores sean distintos entre sí al considerar más precisión y el diagrama de puntos aportaría casi la misma información que la lista literal de datos.

☞ Para variables cuantitativas **discretas**, un **histograma** no es adecuado ya que no se entiende que haya necesidad de agrupar los datos por intervalos. El **diagrama de puntos**, por su parte, cumple justamente con la función del histograma, pero para variables discretas. Por ello, esta suele entenderse adecuada. Finalmente, el mismo argumento para el histograma se aplica al **diagrama de caja**.

iv. Estadística bivariante.

Cuando se lleva a cabo un experimento, siempre podemos fijarnos en más de un rasgo o característica. Es decir, siempre puede observarse más de una variable estadística. Cuando se observan varias variables, los procedimientos estadísticos se enmarcan en lo que se conoce como Estadística Descriptiva multidimensional. Nosotres nos limitaremos a la teoría para pares de variables estadísticas y desarrollaremos algunas nociones básicas de lo que recibe el nombre de Estadística bivariante o bidimensional.

Colectivos de datos bidimensionales: colectivos marginales y condicionados y representaciones tabulares.

Supón que consideramos un experimento en el que estamos fijándonos en dos variables estadísticas a la vez. Entonces, por cada observación o realización del experimento, se obtiene un dato por cada variable, llamémoslos x_ℓ e y_ℓ, para $\ell = 1, \ldots, n$, si se hicieron n observaciones. Tendríamos entonces dos colectivos de datos con el mismo número de datos (digamos, n, y llamemos a cada colectivo como $X = \{x_1, \ldots, x_n\}$ e $Y = \{y_1, \ldots, y_n\}$).

Pero estos datos se han observado conjuntamente, de manera que los datos de un colectivo están ligados a los del otro. Por ello, es natural considerar los datos como pares,

[9] Continua o discreta, digamos, según la nueva definición. Entenderemos que una variable numérica es continua cuando se crea conveniente agrupar los valores en intervalos. Discreta, en caso contrario.

$(x_1, y_1), \ldots, (x_n, y_n)$, y no solo hablar de los colectivos X e Y por separado. Al colectivo de datos formado por esos pares de observaciones lo denotaremos por (X, Y) y escribiremos:

$$(X, Y) = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{n-1}, y_{n-1}), (x_n, y_n)\} \ .$$

Si en este último ignorásemos una de las dos variables, entonces recuperaríamos el colectivo de datos individual de la otra variable. Por ejemplo, si ignoramos los valores de X (respectivamente, Y) en (X, Y), entonces obtendríamos el colectivo Y (respectivamente, X). El colectivo Y (respectivamente, X) recibe el nombre en este caso de colectivo de datos marginales de Y (respectivamente, de X).

- ☞ Un **colectivo** de datos **bidimensional** (X, Y) está formado por datos apareados de dos colectivos de datos unidimensionales.
- ☞ Cada uno de esos colectivos de datos unidimensionales (X e Y) recibe el nombre de **colectivo marginal** del colectivo bidimensional, al considerarlos individualmente e ignorar la dependencia del otro colectivo.
- ☞ De manera similar, se habla de **variable estadística bidimensional** y de **variables estadísticas marginales**.

Una vez tenemos claro lo que es un colectivo de datos bidimensional y sus colectivos marginales, la primera cuestión es cómo se podrían presentar los datos. Al igual que en Estadística Descriptiva unidimensional, es muy útil acudir a una tabla: cada fila (o columna, como hace R) puede corresponderse con los colectivos marginales de cada variable estadística y cada columna (respectivamente, fila) sería el par de datos recogido en cada observación hecha del experimento. A continuación concretamos las definiciones y este formato de tabla (**tabla de datos apareados**) continuando el ejemplo del tanque.

Ejemplo (contaminación tanque 5). De la tabla Tabla 1 debe resultar claro que cada medida de CO_2 puede entenderse ligada al orden cronológico. El orden cronológico puede también entenderse como una variable estadística. Recuérdese que Y era el nombre del conjunto de datos de las emisiones ficticias, $Y = \{122, 118, 119, 117, 110, 121, 118, 119, 122, 121\}$ (en gramos), y llámese $T = \{t_1, \ldots, t_{10}\} = \{1, \ldots, 10\}$ al colectivo de datos de las correspondientes etiquetas cronológicas. De acuerdo con la notación recién introducida, nosotres llamaríamos y escribiríamos el colectivo de datos conjunto como:

$$(Y, T) = \{(y_1, t_1), (y_2, t_2), (y_3, t_3), (y_4, t_4), (y_5, t_5), (y_6, t_6), (y_7, t_7), (y_8, t_8), (y_9, t_9), (y_{10}, t_{10})\} =$$
$$= \{(122, 1), (118, 2), (119, 3), (117, 4), (110, 5), (121, 6), (118, 7), (119, 8), (122, 9), (121, 10)\} \ .$$

El colectivo de datos $Y = \{y_1, \ldots, y_{10}\}$ recibiría el nombre de colectivo marginal de Y asociado a la variable estadística bidimensional, (Y, T). El colectivo de datos T, sería el colectivo marginal de T asociado a (Y, T). En la tabla siguiente, cada fila se corresponde con los colectivos marginales y las columnas se corresponden con las distintas observaciones llevadas a cabo.

Y (en gCO_2)	122	118	119	117	110	121	118	119	122	121
T (orden cron.)	1	2	3	4	5	6	7	8	9	10

Cuadro 3: Tabla de datos apareados de la emisión de CO_2 de un tanque y la etiqueta cronológica de cada medición.

Al igual que se hizo en Estadística Descriptiva unidimensional, es de gran utilidad clasificar las variables bidimensionales por los valores distintos que los datos recogidos toman. Si estamos trabajando con un colectivo genérico, (X, Y) en la notación anterior, con n observaciones realizadas, lo primero es tener claro cuántos y qué valores distintos toman los colectivos marginales.

Llamemos $v_1 < \cdots < v_k$ a los valores distintos del colectivo marginal de X y, de otro lado, $\tilde{v}_1 < \cdots < \tilde{v}_{\tilde{k}}$ a los valores distintos del colectivo marginal de Y (ordenados de menor a mayor, de nuevo, en caso de que se cuente con un orden en cada variable). Con esta notación, k sería el número de valores distintos en el colectivo marginal de X y $\tilde{k}$, el número de valores distintos en el colectivo marginal de Y. Para cada par de valores distintos, $(v_i, \tilde{v}_j)$, vamos a definir la frecuencia absoluta y relativa como el número y la proporción, respectivamente, de ocasiones en las que aparece el par $(v_i, \tilde{v}_j)$, con $i = 1, \ldots, k$ y $j = 1, \ldots, \tilde{k}$.

Denotaremos por n_{ij} (siendo $i = 1, \ldots, k$ y $j = 1, \ldots, \tilde{k}$ dos índices que recorren los valores distintos de X e Y, respectivamente) a la frecuencia absoluta de $(v_i, \tilde{v}_j)$. Esto es, al número de veces que aparece cada par de valores distintos. De manera similar, denotaremos por f_{ij} a la proporción de ocasiones en las que aparece el par $(v_i, \tilde{v}_j)$ en el colectivo (X, Y).

- ☞ La **frecuencia absoluta** de cada par de valores distintos de un colectivo bidimensional, $(v_i, \tilde{v}_j)$, se define como el número de veces que aparece el valor $(v_i, \tilde{v}_j)$ en el colectivo de datos dado, (X, Y), y se denota por n_{ij}.
- ☞ La **frecuencia relativa** de cada par de valores distintos de un colectivo bidimensional, $(v_i, \tilde{v}_j)$, se define como la proporción de veces que aparece el valor $(v_i, \tilde{v}_j)$ en el colectivo de datos dado, (X, Y), y es $f_{ij} = n_{ij}/n$.

Con esta información se puede hacer lo que se conoce como tabla de frecuencias (absolutas o relativas) bidimensional, que podrían escribir como sigue:

$X \backslash Y$	$\tilde{v}_1$	$\tilde{v}_2$	$\cdots$	$\tilde{v}_j$	$\cdots$	$\tilde{v}_{\tilde{k}-1}$	$\tilde{v}_{\tilde{k}}$
v_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	$n_{1,\tilde{k}-1}$	$n_{1\tilde{k}}$
v_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	$n_{2,\tilde{k}-1}$	$n_{2,\tilde{k}}$
$\vdots$	$\vdots$	$\ddots$	$\ddots$	$\vdots$	$\iddots$	$\iddots$	$\vdots$
v_i	n_{i1}	$\cdots$	$\cdots$	n_{ij}	$\cdots$	$\cdots$	$n_{i\tilde{k}}$
$\vdots$	$\vdots$	$\iddots$	$\iddots$	$\vdots$	$\ddots$	$\ddots$	$\vdots$
v_{k-1}	$n_{k-1,1}$	$n_{k-1,2}$	$\cdots$	$\cdots$	$\cdots$	$n_{k-1,\tilde{k}-1}$	$n_{k-1,\tilde{k}}$
v_k	n_{k1}	n_{k2}	$\cdots$	n_{kj}	$\cdots$	$n_{k,\tilde{k}-1}$	$n_{k\tilde{k}}$

Cuadro 4: Tabla de frecuencias absolutas bidimensional asociada al colectivo de datos genérico (X, Y). Los valores distintos de los colectivos marginales de X e Y son $v_1, \ldots, v_k$ y $\tilde{v}_1, \ldots, \tilde{v}_{\tilde{k}}$, respectivamente. En caso de que las variables estadísticas marginales cuenten con un orden, los valores distintos se ordenan de menor a mayor. La frecuencia absoluta de cada par de valores es n_{ij}, donde i y j son índices con posibles valores $i = 1, \ldots, k$ y $j = 1, \ldots, \tilde{k}$.

- ☞ La **tabla de frecuencias absolutas bidimensional** (llamada **tabla de contingencia**) de un colectivo de datos consiste en representar en forma de tabla los valores distintos de los colectivos marginales, enfrentados, y asociar a cada par su correspondiente frecuencia absoluta.

$X\backslash Y$	$\tilde{v}_1$	$\tilde{v}_2$	$\cdots$	$\tilde{v}_j$	$\cdots$	$\tilde{v}_{\tilde{k}-1}$	$\tilde{v}_{\tilde{k}}$
v_1	f_{11}	f_{12}	$\cdots$	f_{1j}	$\cdots$	$f_{1,\tilde{k}-1}$	$f_{1\tilde{k}}$
v_2	f_{21}	f_{22}	$\cdots$	f_{2j}	$\cdots$	$f_{2,\tilde{k}-1}$	$f_{2,\tilde{k}}$
$\vdots$	$\vdots$	$\ddots$	$\ddots$	$\vdots$	$\iddots$	$\iddots$	$\vdots$
v_i	f_{i1}	$\cdots$	$\cdots$	f_{ij}	$\cdots$	$\cdots$	$f_{i\tilde{k}}$
$\vdots$	$\vdots$	$\iddots$	$\iddots$	$\vdots$	$\ddots$	$\ddots$	$\vdots$
v_{k-1}	$f_{k-1,1}$	$f_{k-1,2}$	$\cdots$	$\cdots$	$\cdots$	$f_{k-1,\tilde{k}-1}$	$f_{k-1,\tilde{k}}$
v_k	f_{k1}	f_{k2}	$\cdots$	$f_{\tilde{k}j}$	$\cdots$	$f_{k,\tilde{k}-1}$	$f_{k\tilde{k}}$

Cuadro 5: Tabla de frecuencias relativas bidimensional asociada al colectivo de datos genérico (X, Y). Los valores distintos de los colectivos marginales de X e Y son $v_1, \ldots, v_k$ y $\tilde{v}_1, \ldots, \tilde{v}_{\tilde{k}}$, respectivamente. En caso de que las variables estadísticas marginales cuenten con un orden, los valores distintos se ordenan de menor a mayor. La frecuencia relativa de cada par de valores es f_{ij}, donde i y j son índices con posibles valores $i = 1, \ldots, k$ y $j = 1, \ldots, \tilde{k}$.

☞ La **tabla de frecuencias relativas bidimensional** de un colectivo de datos consiste en representar en forma de tabla los valores distintos de los colectivos marginales, enfrentados, y asociar a cada par su correspondiente frecuencia relativa.

Nótese que las tablas de frecuencias bidimensionales recuerdan a matrices y que, a diferencia de las tablas de frecuencias unidimensionales, se debe dar una tabla entera para la frecuencia escogida. Es decir, solo se representa un tipo de frecuencia: la absoluta o relativa.

Ejemplo (contaminación tanque 6). De la Tabla 3, se puede deducir la tabla de frecuencias absolutas bidimensional asociada al colectivo (T, Y), que es la siguiente:

$T\backslash Y$	110	117	118	119	121	122
1	0	0	0	0	0	1
2	0	0	1	0	0	0
3	0	0	0	1	0	0
4	0	1	0	0	0	0
5	1	0	0	0	0	0
6	0	0	0	0	1	0
7	0	0	1	0	0	0
8	0	0	0	1	0	0
9	0	0	0	0	0	1
10	0	0	0	0	1	0

Cuadro 6: Tabla de frecuencias absolutas bidimensional asociada al colectivo de datos (T, Y).

En este caso, en cada nivel temporal (en cada fila de la tabla anterior) aparece una única frecuencia absoluta distinta de cero, ya que solamente se hizo una medición en cada nivel temporal. Si se hubieran tomado más datos en cada tiempo (con distintos aparatos, por ejemplo), entonces habría más valores distintos de cero en cada fila.

Sea como fuere, para el ejemplo que nos atañe, las únicas frecuencias absolutas distintas de cero son n_{16}, n_{23}, n_{34}, n_{42}, n_{51}, n_{65}, n_{73}, n_{84}, n_{96} y $n_{10,5}$, que son las frecuencias absolutas de $(v_1, \tilde{v}_6) = (t_1, y_1)$, $(v_2, \tilde{v}_3) = (t_2, y_2)$, $(v_3, \tilde{v}_4) = (t_3, y_3)$, $(v_4, \tilde{v}_2) = (t_4, y_4)$,

$(v_5, \tilde{v}_1) = (t_5, y_5)$, $(v_6, \tilde{v}_5) = (t_6, y_6)$, $(v_7, \tilde{v}_3) = (t_7, y_7)$, $(v_8, \tilde{v}_4) = (t_8, y_8)$, $(v_9, \tilde{v}_6) = (t_9, y_9)$ y $(v_{10}, \tilde{v}_5) = (t_{10}, y_{10})$, respectivamente. De otro lado, dividiendo la tabla anterior por el número total de datos ($n = 10$), se obtendría la tabla de frecuencias relativas. Esta sería la misma tabla que la Tabla 6, pero con 1/10 donde hay escrito 1: las frecuencias relativas son en este caso $f_{ij} = n_{ij}/10$, porque $n = 10$ es el número de datos. Nótese que los índices recorren en este caso los valores $i = 1, \ldots, 10$ y $j = 1, \ldots, 6$, ya que el número de valores distintos en T es $k = 10$ y en Y, es $\tilde{k} = 6$.

Además de los colectivos de datos marginales, es importante saber cómo obtener lo que llamaremos "colectivos condicionados" a partir de una tabla de frecuencias de un colectivo bidimensional dado. Volviendo a nuestro colectivo genérico, que habíamos llamado (X, Y), si queremos saber cómo se comporta uno de los dos colectivos marginales cuando nos restringimos a un valor o unos valores de la otra, podemos hacerlo eficientemente a partir de las tablas de frecuencias bidimensionales. Tan solo debemos fijarnos, dentro de las tablas, en las columnas o filas asociadas a los valores a los que nos restrinjamos, e ignorar el resto. El colectivo marginal de una de las dos variables al restringirse a algunos valores de la otra, es lo que llamaremos colectivos condicionados.

- ☞ Si tenemos una variable estadística (X, Y), llamaremos **colectivo condicionado de X a que en Y se dé el valor $\tilde{\mathbf{v}}_\mathbf{j}$**, para un $j = 1, \ldots, \tilde{k}$ concreto, al resultado de filtrar los datos de (X, Y) por la condición $Y = \tilde{v}_j$ (es decir, de imponer que el segundo valor del par sea $\tilde{v}_j$) y, tras ello, quedarse con el colectivo de datos marginales de X. Al colectivo resultante lo denotaremos por $X|_{Y=\tilde{v}_j}$. De manera similar, llamaremos **colectivo condicionado de Y a que en X se dé el valor $\mathbf{v}_\mathbf{i}$**, para un $i = 1, \ldots, k$ concreto, al colectivo de datos marginales de Y, entre aquellos datos en (X, Y) que verificaban que el primer valor del par fuera v_i. Lo denotaremos por $Y|_{X=v_i}$.
- ☞ Del mismo modo, podemos establecer condiciones más complejas sobre una u otra variable, como: $X|_{\{Y=\tilde{v}_j \text{ ó } Y=\tilde{v}_{j'}\}}$ o $Y|_{\{X=v_i \text{ ó } X=v_{i'}\}}$. El primer colectivo sería el colectivo condicionado de X a que en Y se tome el valor $\tilde{v}_j$ o $\tilde{v}_{j'}$, donde $j, j' = 1, \ldots, \tilde{k}$. El segundo, el colectivo condicionado de Y a que en X se tome el valor v_i o $v_{i'}$, siendo $i, i' = 1, \ldots, k$. Otros ejemplos de condiciones que se pueden imponer (si la variable por la que se condiciona cuenta con un orden) son $Y|_{X \leq v_i}$ (colectivo condicionado de Y a que en X los datos sean menores o iguales que v_i) o $X|_{\{Y>\tilde{v}_j \text{ ó } Y=\tilde{v}_{j'}\}}$ (colectivo condicionado de X a que en Y los datos sean mayores que $\tilde{v}_j$ o sean iguales a $\tilde{v}_{j'}$), etc.

Ejemplo (contaminación tanque 7). De la Tabla 3 o la Tabla 6 puede apreciarse a simple vista que el colectivo condicionado $T|_{Y=119}$ (es decir, el colectivo de los datos en T que cumplen la condición $Y = 119$) es $T|_{Y=119} = \{3, 8\}$. Otros ejemplos son $T|_{Y=122} = = \{1, 9\}$, $T|_{Y=117} = \{4\}$ e $Y|_{T=5} = \{110\}$. También: $Y|_{T\geq 5 \text{ ó } T\leq 8} = \{110, 121, 118, 119\}$, $T|_{Y\geq 117} = \{1, 2, 3, 4, 6, 7, 8, 9, 10\}$, etc. Cuando se condiciona por Y, en este ejemplo, debemos mirar solo las columnas (en la tabla de frecuencias) que se corresponden con los valores que cumplen la condición impuesta. Al condicionar por valores de T, en cambio, basta mirar únicamente las filas (en la tabla de frecuencias) que cumplan tal condición.

Algunas relaciones e independencia estadística.

Retomemos el grupo de datos bidimensional genérico, $(X, Y) = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, en el que se recogen n medidas. Como hicimos antes, llamemos $v_1 < \cdots < v_k$ a los valores

distintos del colectivo marginal de X y $\tilde{v}_1 < \cdots < \tilde{v}_{\tilde{k}}$ a aquellos del colectivo marginal de Y (ordenados de menor a mayor en caso de que se pueda). Es decir, en el colectivo marginal de X se tienen k valores distintos y en el de Y, $\tilde{k}$.

Si n_{ij} es el número de veces que aparece el par $(v_i, \tilde{v}_j)$[10] y f_{ij} es la proporción de veces que aparece ese par en el colectivo (X, Y), deben estar claras las siguientes relaciones:

$$f_{ij} = \frac{n_{ij}}{n} \ , \quad \sum_{i,j} n_{ij} = n \ , \quad \sum_{i,j} f_{ij} = 1 \ .$$

Es decir, la frecuencia relativa cada valor $(v_i, \tilde{v}_j)$ es su frecuencia absoluta entre el número total de datos. Además, la suma de todas las frecuencias absolutas es el número total de datos y, naturalmente, la suma de todas las frecuencias relativas es igual a uno.

En lo que a los colectivos marginales respecta, escribiremos en adelante "$n_{i\cdot}$" para referirnos a la frecuencia absoluta de cada valor distinto del colectivo marginal de X y llamaremos "$n_{\cdot j}$" a la frecuencia absoluta de cada valor distintos del colectivo marginal de Y. De manera similar, denotaremos por "$f_{i\cdot}$" y "$f_{\cdot j}$" a las respectivas frecuencias relativas. Entonces, se satisfacen las siguientes relaciones:

$$n_{i\cdot} = \sum_{j=1}^{\tilde{k}} n_{ij} \ , \quad \sum_{i=1}^{k} n_{i\cdot} = n \ , \quad f_{i\cdot} = \frac{n_{i\cdot}}{n} \quad \text{y} \quad \sum_{i=1}^{k} f_{i\cdot} = 1 \ , \quad \text{respecto al colectivo marginal de } X \ ,$$

y

$$n_{\cdot j} = \sum_{i=1}^{k} n_{ij} \ , \quad \sum_{j=1}^{\tilde{k}} n_{\cdot j} = n \ , \quad f_{\cdot j} = \frac{n_{\cdot j}}{n} \quad \text{y} \quad \sum_{j=1}^{\tilde{k}} f_{\cdot j} = 1 \ , \quad \text{respecto al colectivo marginal de } Y \ .$$

Respecto al colectivo condicionado de X a cierto valor $\tilde{v}_j$ de Y, $X|_{Y=\tilde{v}_j}$, escribiremos "$n_{i|Y=\tilde{v}_j}$" para referirnos a la frecuencia absoluta de cada valor del colectivo condicionado. En otras palabras, cada $n_{1|Y=\tilde{v}_j}, \ldots, n_{k|Y=\tilde{v}_j}$, para j fijo, es el número de veces que se repite cada valor $(v_1, \tilde{v}_j), \ldots, (v_n, \tilde{v}_j)$ en el colectivo $X|_{Y=\tilde{v}_j}$. Para cada $i = 1, \ldots, k$, se puede leer como: "$n_{i|Y=\tilde{v}_j}$ es la frecuencia absoluta del valor v_i, sabiendo que ocurrió $\tilde{v}_j$". Análogamente, en $Y|_{X=v_i}$ escribiremos "$n_{j|X=v_i}$" para hablar del número de veces que se recogió el valor $\tilde{v}_j$, entre los pares de datos cuya primera componente es v_i. Con estos nombres, se verifican las siguientes igualdades:

$$n_{i|Y=\tilde{v}_j} = n_{ij} \ , \quad \sum_{i=1}^{k} n_{i|Y=\tilde{v}_j} = n_{\cdot j} \ , \quad f_{i|Y=\tilde{v}_j} = \frac{n_{ij}}{n_{\cdot j}} = \frac{f_{ij}}{f_{\cdot j}} \quad \text{y} \quad \sum_{i=1}^{k} f_{i|Y=\tilde{v}_j} = 1 \ ,$$

para el colectivo $X|_{Y=\tilde{v}_j}$, y

$$n_{j|X=v_i} = n_{ij} \ , \quad \sum_{j=1}^{\tilde{k}} n_{j|X=v_i} = n_{i\cdot} \ , \quad f_{j|X=v_i} = \frac{n_{ij}}{n_{i\cdot}} = \frac{f_{ij}}{f_{i\cdot}} \quad \text{y} \quad \sum_{j=1}^{\tilde{k}} f_{j|X=v_i} = 1 \ ,$$

para el colectivo $Y|_{X=v_i}$.

[10] Recuerda que i y j son dos índices que recorren los valores distintos de cada colectivo ($i = 1, \ldots, k$ y $j = 1, \ldots, \tilde{k}$).

Nótese que esta forma de escribir las frecuencias de un colectivo marginal, condicionado a un valor del otro, se puede adaptar para imponer otras condiciones más generales.

La mayor parte de las relaciones anteriores se pueden avistar a partir de una tabla de frecuencias en la que se añada una columna y fila al final que exprese la suma de las frecuencias de cada fila y columna, respectivamente. Es decir, algo como:

$X\backslash Y$	$\tilde{v}_1$	$\tilde{v}_2$	$\cdots$	$\tilde{v}_{\tilde{k}}$	Total
v_1	n_{11}	n_{12}	$\cdots$	$n_{1\tilde{k}}$	$n_{1\cdot}$
v_2	n_{21}	n_{22}	$\cdots$	$n_{2\tilde{k}}$	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
v_k	n_{k1}	n_{k2}	$\cdots$	$n_{k\tilde{k}}$	$n_{k\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot\tilde{k}}$	n

(a) Tabla frecuencias absolutas.

$X\backslash Y$	$\tilde{v}_1$	$\tilde{v}_2$	$\cdots$	$\tilde{v}_{\tilde{k}}$	Total
v_1	f_{11}	f_{12}	$\cdots$	$f_{1\tilde{k}}$	$f_{1\cdot}$
v_2	f_{21}	f_{22}	$\cdots$	$f_{2\tilde{k}}$	$f_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
v_k	f_{k1}	f_{k2}	$\cdots$	$f_{k\tilde{k}}$	$f_{k\cdot}$
Total	$f_{\cdot 1}$	$f_{\cdot 2}$	$\cdots$	$f_{\cdot\tilde{k}}$	1

(b) Tabla frecuencias relativas.

Cuadro 7: Tabla de frecuencias absolutas (izquierda) y relativas (derecha) bidimensional asociada al colectivo de datos genérico (X, Y), a la que se han añadido una fila y una columna expresando las frecuencias de los valores de los colectivos marginales. Los valores distintos de los colectivos marginales de X e Y son $v_1, \ldots, v_k$ y $\tilde{v}_1, \ldots, \tilde{v}_{\tilde{k}}$, respectivamente. Para cada $i = 1, \ldots, k$, $n_{i\cdot}$ es la frecuencia absoluta de cada valor, v_i, del colectivo marginal de X. Para cada $j = 1, \ldots, \tilde{k}$, $n_{\cdot j}$ es la frecuencia absoluta de cada valor, $\tilde{v}_j$, del colectivo marginal de Y.

Antes de pasar al siguiente apartado, introduciremos brevemente la idea de independencia estadística. Diremos que dos colectivos de datos son estadísticamente independientes cuando al condicionar un colectivo a cierto valor o valores del otro se encuentra que las frecuencias relativas del colectivo condicionado son las mismas que las del colectivo marginal. Es decir, cuando el comportamiento de los colectivos condicionados es la misma que el comportamiento del colectivo marginal. Recordando la relación introducida anteriormente:

$$f_{i|Y=\tilde{v}_j} = \frac{f_{ij}}{f_{\cdot j}} \ ,$$

aquello viene a decir que, por ejemplo, debe ocurrir $f_{i|Y=\tilde{v}_j} = f_{i\cdot}$, en cuyo caso de la fórmula anterior se deduce que $f_{ij} = f_{i\cdot} \cdot f_{\cdot j}$. De manera análoga, también debe ocurrir que $f_{j|X=v_i} = f_{\cdot j}$. De ello se puede deducir que la independencia estadística tiene lugar cuando las filas de la tabla de frecuencias son proporcionales entre sí.

☞ Se dice que dos colectivos de datos son **estadísticamente independientes** cuando para cada $i = 1, \ldots, k$ y para todo $j = 1, \ldots, \tilde{k}$ se cumple que $f_{ij} = f_{i\cdot} \cdot f_{\cdot j}$. Ello es equivalente a que las filas de la tabla de frecuencias sean proporcionales entre sí. También es equivalente a que las columnas de una tabla de frecuencias sean proporcionales entre sí.

Representaciones gráficas para colectivos bidimensionales.

Los diagramas de barras y los histogramas se pueden llevar al caso bidimensional y su uso es frecuente. Le daremos especial importancia a otra representación gráfica de colectivos bidimensionales que se conoce como **diagrama de dispersión** o **nube de puntos**.

El diagrama de dispersión se utiliza para variables numéricas y, dado un colectivo de datos (X, Y), consiste en tomar dos ejes de coordenadas (uno para cada colectivo marginal,

X e Y) y representar cada dato del colectivo (X, Y) como un punto del plano, con sus coordenadas.

Ejemplo (contaminación tanque 8). Una curiosidad es que la nube de puntos del recurrente ejemplo, (T, Y), se concreta prácticamente en su serie temporal (Figura 4b, quitando las líneas). Ello se debe a que en (T, Y) se observa simultáneamente el orden cronológico y los gramos de CO_2, además de solo haber recogido una medición en cada nivel temporal.

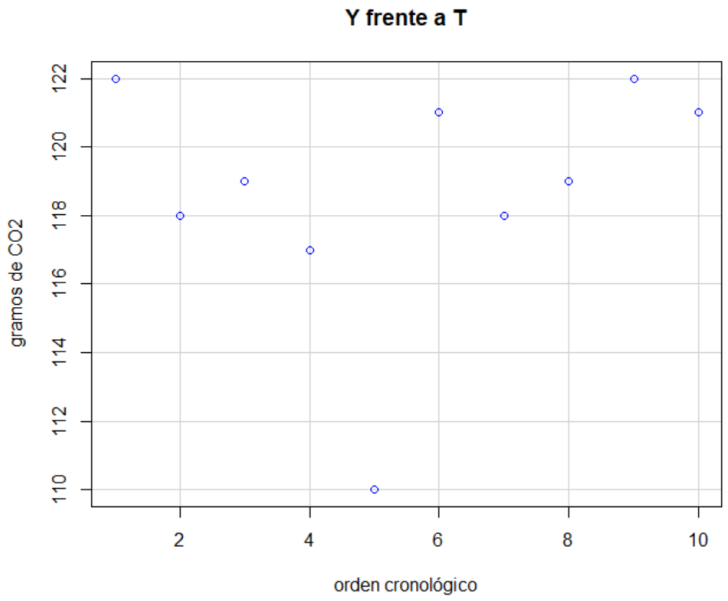


Figura 7: Diagrama de dispersión del colectivo bidimensional (Y, T) del colectivo de datos ficticios sobre los gramos medidos de emisión de un tanque en cada nivel temporal.

Covarianza y covarianza muestral. Correlación lineal de Pearson.

Dado un colectivo de n datos bidimensionales, $(X, Y) = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, hay una operación de especial importancia: la covarianza. Al igual que la varianza de un colectivo unidimensional, la covarianza da cuenta de la dispersión, pero para un colectivo bidimensional. Esta operación es un modo de cuantificar la dispersión conjunta de los datos, se denota por σ_{XY} y es la siguiente:

$$\sigma_{XY} = \sum_{i=1}^{n} \frac{\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{n} \ .$$

Nótese que la relación de la covarianza y la varianza es muy similar a la relación que hay entre la norma de un vector al cuadrado (que es el producto escalar consigo mismo) y el producto escalar de un vector con otro distinto. De hecho, multiplicando en la expresión anterior, se encuentra:

$$\sigma_{XY} = \overline{X \cdot Y} - \overline{X}\,\overline{Y} \quad \left(\text{ de modo que } \sigma_{XX} = \overline{X^2} - \overline{X}^2 = \sigma_X^2 \right) \ .$$

Es decir, la covarianza de un colectivo de datos consigo mismo es justamente su varianza.

Cuando se trabaja en contextos de Inferencia Estadística, en cambio, es más conveniente, como se verá, utilizar la **covarianza muestral**, denotada por S_{XY} y definida como:

$$S_{XY} = \sum_{i=1}^{n} \frac{(x_i - \overline{X})(y_i - \overline{Y})}{n-1} \ .$$

La covarianza muestral se puede calcular a partir de la covarianza, al igual que la varianza muestral se calculaba a partir de la varianza, usando la siguiente ecuación:

$$S_{XY} = \frac{n}{n-1}\sigma_{XY} = \frac{n}{n-1}\left[\overline{X \cdot Y} - \overline{XY}\right] \quad \left(\text{ entonces, } S_{XX} = S_X^2 \ \right) \ .$$

☞ Las covarianzas permiten cuantificar la dispersión conjunta de los colectivos marginales de un colectivo bidimensional. Sus unidades son las mismas que las del producto de las unidades de cada colectivo.

Si dividimos la covarianza entre el producto de las desviaciones típicas de los colectivos marginales, se obtiene una medida adimensional: el coeficiente de correlación lineal de Pearson. Esta cantidad se define como sigue, y está siempre entre -1 y 1:

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad \text{y} \quad -1 \leq r_{XY} \leq 1 \ .$$

El coeficiente de correlación es una medida de la relación lineal entre los colectivos marginales. Cuando $r_{XY} = 1$, el diagrama de dispersión de (X, Y) es una perfecta recta creciente; cuando es $r_{XY} = -1$, se tiene una perfecta recta decreciente (en estos dos casos, se habla de colectivos correlacionados) y cuando $r_{XY} = 0$ (colectivos incorrelados), no existe relación lineal, aunque puede que sí exista una relación no lineal.

De otro lado, es interesante e importante saber que cuando dos colectivos de datos son estadísticamente independientes, entonces son incorrelados (**independencia implica incorrelación**). Sin embargo, al revés no es cierto: como acabamos de comentar, dos colectivos pueden ser incorrelados pero estar relacionados no linealmente (**incorrelación no implica independencia**).

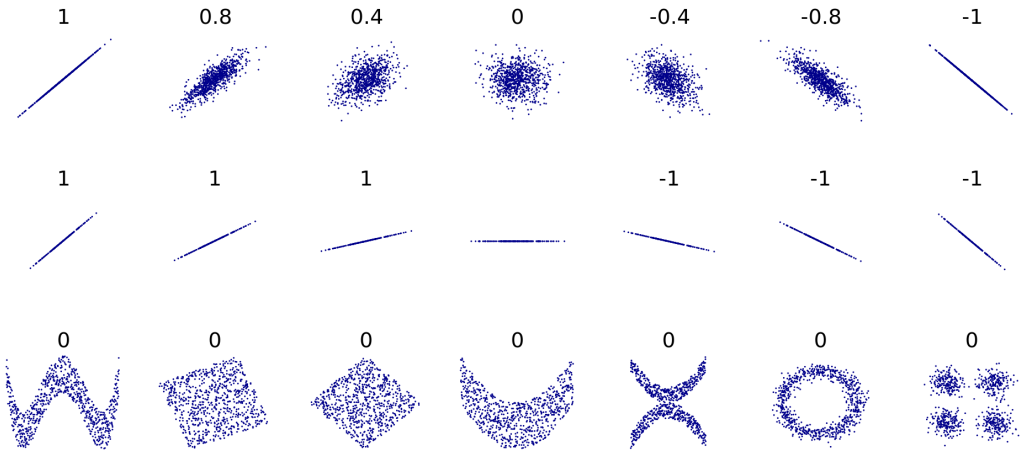


Figura 8: Distintos diagramas de dispersión junto al coeficiente de correlación lineal de Pearson de cada uno. Fuente: en este enlace.

☞ El **coeficiente de correlación lineal** (de Pearson), r_{XY}, es el resultado de dividir la covarianza por el producto de las desviaciones típicas. El resultado de hacer esta operación con las desviaciones típicas y la covarianza es el mismo que hacerlo con la covarianza muestral y las desviaciones típicas muestrales. Asimismo, es un valor entre -1 y 1 adimensional que permite graduar la existencia de relación lineal entre las variables, además de si esa relación es creciente o decreciente (ver Figura 8).

v. Regresión lineal: método de mínimos cuadrados.

Recta de mínimos cuadrados.

Cuando se observan varias variables numéricas simultáneamente, es de especial interés estudiar cómo influye cada una de ellas en la otra y tratar de obtener un modelo que se ajuste "bien" a los datos. El objetivo del **análisis de regresión** es relacionar, en un colectivo de datos multidimensional, una variable estadística marginal (llamada **variable explicada**, dependiente o de respuesta) con el resto (llamadas **variable explicativas**, independientes o predictoras), a través de una función.

Al hacerlo, es habitual que los valores recogidos experimentalmente y los valores predichos por el modelo sean distintos. A esa diferencia se la suele llamar **error** y denotaremos por $\varepsilon = \{\varepsilon_1, \ldots, \varepsilon_n\}$ al colectivo de errores en cada dato.

Nuestro objetivo será explicar Y en función de X, siendo X e Y colectivos marginales de un colectivo bidimensional dado, $(X, Y) = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. Esto es, vamos a tratar a Y como variable explicada y a X como variable explicativa. Para cada función f que escojamos, se cometerá un error sobre cada dato, lo cual escribiremos como:

$$Y = f(X) + \varepsilon \ , \quad \text{queriendo decir que} \quad y_i = f(x_i) + \varepsilon_i \ , \quad \text{para } i = 1, \ldots, n \ .$$

Cada $f(x_i)$ sería una **estimación** de y_i y lo escribiremos como $\hat{y}_i$. Si concretamente **buscamos una relación lineal** (es decir, $f(X) = a + bX$), entonces:

$$f(x_i) = a + bx_i = \hat{y}_i \ , \quad \text{para cada } i = 1, \ldots, n \ .$$

El error cometido, por lo tanto, sería:

$$\varepsilon_i = y_i - \hat{y}_i = y_i - (a + bx_i) \ .$$

Entre todas las rectas posibles (es decir, entre todos los valores de a y b posibles), el método de mínimos cuadrados da un criterio para elegir una. El **método de mínimos cuadrados** consiste en tomar la recta que minimice la media de los cuadrados de los errores cometidos. Es decir, se busca encontrar los valores de a y b que hagan lo más pequeño posible lo siguiente:

$$\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 = \frac{1}{n}\sum_{i=1}^{n} \left(y_i - (a + bx_i)\right)^2 \ .$$

El resultado de minimizar lo anterior es una recta que pasa por el punto $(\overline{X}, \overline{Y})$ en el diagrama de dispersión y que tiene pendiente S_{XY}/S_X^2. Recordando la ecuación punto-pendiente de una recta, el ajuste lineal viene entonces dado por la relación:

$$\hat{Y} - \overline{Y} = \frac{S_{XY}}{S_X^2}(X - \overline{X}) \ , \quad \text{es decir,} \quad \hat{Y} = \overline{Y} - \overline{X}\frac{S_{XY}}{S_X^2} + \frac{S_{XY}}{S_X^2}X \ .$$

Es decir, para cada valor de la variable explicativa, $X = x$, obtenemos una estimación o predicción, $\hat{Y}(X = x)$, de la variable explicada, Y. Los valores de los coeficientes de la recta de regresión, por lo tanto, son:

$$\hat{b} = \frac{S_{XY}}{S_X^2} \quad \text{y} \quad \hat{a} = \overline{Y} - \hat{b}\overline{X} \ ,$$

óptimos en el sentido de que hacen mínimo el error cuadrático. $\hat{b}$ es la pendiente de la recta de mínimos cuadrados y, de otro, $\hat{a}$ es la ordenada en el origen.

- ☞ El signo de la covarianza (S_{XY}) determina el signo de la pendiente de la recta. La recta será creciente si la covarianza es positiva y decreciente, si es negativa. La covarianza también determina el signo del coeficiente de correlación lineal.
- ☞ El coeficiente de correlación lineal (r_{XY}) permite comparar el grado de relación lineal que existe entre distintos colectivos bidimensionales, al ser adimensional. Su signo también indica si la recta de mínimos cuadrados es creciente o decreciente (ver Figura 8).

Los modelos de regresión se utilizan para obtener valores estimados de la variable explicada ($\hat{y}$ para estimar y) en función de valores de la variable explicativa (x), más allá de los datos concretos recogidos en el colectivo de partida. Pero esto solo es apropiado hacerlo cuando x está en el rango del colectivo X, es decir, cuando está en el intervalo $[\text{mín}(X), \text{máx}(X)] = \text{Rango}(X)$. En tal caso, se habla **interpolación** y la calidad de la estimación dependerá de cuán bueno sea el ajuste lineal. Sin embargo, es inapropiado aplicar este modelo para valores de x fuera del rango de X, en cuyo caso se habla de **extrapolación**.

Ejemplo (contaminación tanque 9). El ajuste de mínimos cuadrados de Y sobre T (Y como colectivo explicado y T como colectivo explicativo) es el siguiente:

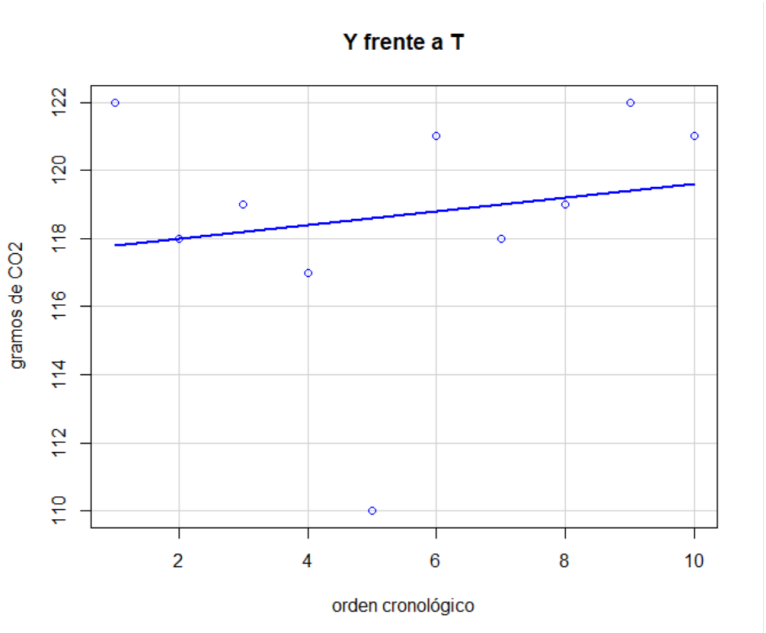


Figura 9: Diagrama de dispersión del colectivo bidimensional (Y, T) del colectivo de datos ficticios sobre los gramos medidos de emisión de un tanque en cada nivel temporal y la recta de regresión de Y explicada por T: $\hat{Y} = a + bT$, con $a = 117.6$ y $b = 0.2$. El coeficiente de correlación es $r_{YT} = 0.17157$.

La recta de regresión es:

$$\hat{Y} = 117.6 + 0.2T \ , \quad \text{es decir,} \quad a = 117.6 \quad \text{y} \quad b = 0.2 \ .$$

La recta se representa en el diagrama de dispersión en la Figura 9. La idea es que esa relación lineal podría servir para estimar los gramos de CO_2 en distintos tiempos entre el mínimo y el máximo (entre 1 y 10). No obstante, aquí no tendría demasiado sentido porque nuestra variable 'orden cronológico' es discreta y no tiene mucho sentido considerar valores intermedios.

Coeficiente de determinación.

La varianza de Y (σ_Y^2 o SST) se puede descomponer en este caso en la varianza explicada por la regresión ($\sigma_{\hat{Y}}^2$ o SSR) y la varianza residual (que es la suma de los errores cuadráticos dividida por el número de datos, σ_r^2 o SSE) como sigue:

$$\frac{\sum_{i=1}^n (y_i - \overline{Y})^2}{n} = \frac{\sum_{i=1}^n (\hat{y}_i - \overline{Y})^2}{n} + \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n} \ , \quad \text{es decir} \quad \sigma_Y^2 = \sigma_{\hat{Y}}^2 + \sigma_r^2 \ .$$

Hay una medida basada en esta división que es de gran utilidad para graduar cuán bueno es un ajuste realizado, sea lineal o no. Aquella se conoce como **bondad del ajuste** o **coeficiente de determinación** y se denota por R^2.

El coeficiente de determinación es un valor entre cero y uno que se define como la división entre la varianza explicada por el modelo y la varianza del colectivo, es decir:

$$R^2 = \frac{\sigma_{\hat{Y}}^2}{\sigma_Y^2} = \frac{\sigma_Y^2 - \sigma_r^2}{\sigma_Y^2} = 1 - \frac{\sigma_r^2}{\sigma_Y^2} \in [0, 1] \ .$$

Este coeficiente se interpreta como la proporción de la variabilidad de Y que es explicada por el modelo de regresión. Valores próximos a cero indican que el ajuste es poco adecuado, mientras que, a medida que los valores se aproximan a uno, más adecuado se considera el ajuste. **En el modelo de regresión lineal, el coeficiente de determinación es justamente el coeficiente de correlación lineal al cuadrado**, es decir:

$$R^2 = r_{XY}^2 \ , \quad \text{para el modelo de regresión lineal} \ .$$

Recta de regresión de X sobre Y.

Cuando tratamos de explicar el colectivo X a partir de Y, la recta no es la misma que la descrita anteriormente. Siguiendo el procedimiento anterior, querríamos relacionar X con Y a través de una función g, cometiendo un error ε. Esto es:

$$X = g(Y) + \varepsilon \ , \quad \text{queriendo decir que} \quad x_i = g(y_i) + \varepsilon_i \ , \quad \text{para } i = 1, \ldots, n \ .$$

Estaríamos estimando x_i como $g(y_i) = c + dy_i$, para ciertos coeficientes de la recta, c y d. Los coeficientes de la recta de regresión de X sobre Y minimizan

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n \left(x_i - (c + dy_i)\right)^2 \ ,$$

y son los siguientes:

$$\hat{d} = \frac{S_{XY}}{S_X^2} \quad \text{y} \quad \hat{c} = \overline{X} - \hat{d} \cdot \overline{Y} \ .$$

Es decir, la recta de mínimos cuadrados de X sobre Y sería

$$\hat{X} - \overline{X} = \frac{S_{XY}}{S_Y^2}(Y - \overline{Y}) \quad \text{o, si se prefiere,} \quad \hat{X} = \overline{X} - \overline{Y}\frac{S_{XY}}{S_Y^2} + \frac{S_{XY}}{S_Y^2}X \ .$$

Por otra parte, la bondad del ajuste de X sobre Y es el mismo que de Y sobre X: $R^2 = r_{XY}^2$.

Otros tipos de regresión: modelo hiperbólico, potencial y exponencial[11].

[11] Los comentarios sobre modelos de regresión no lineales son ampliación.

III. Probabilidad

i. La noción de probabilidad.

En el marco de la Estadística Descriptiva, hemos visto operaciones, representaciones gráficas y tabulares y modelos de regresión. Todo ello son herramientas para analizar un grupo de mediciones experimentales dado — es decir, un colectivo de datos— recogido con anterioridad a ese análisis estadístico. Toda la información que se pueda obtener desde la Estadística Descriptiva es información de fenómenos o experimentos concluidos, es decir, de fenómenos que forman parte del pasado.

De otro lado, la noción de 'probabilidad' se utiliza para cuantificar afirmaciones sobre los posibles resultados de un experimento — llamados **sucesos**—, a partir de la frecuencia con la que se espera — de antemano— que estos ocurran. La idea básica es que la **probabilidad de un suceso** sea la proporción de ocasiones que se espera que ese suceso ocurra, al repetir un experimento concreto una y otra vez exactamente en las mismas condiciones. En la naturaleza de la idea de probabilidad está considerar los experimentos o fenómenos como '**aleatorios**'. Considerar un experimento como aleatorio quiere decir que, aunque el experimento se produzca exactamente bajo las mismas condiciones, su resultado podría no ser el mismo.

Bajo esta lógica, la probabilidad de un suceso sería la proporción de veces que este tendría lugar al repetir "muchísimas" veces el mismo experimento bajo, exactamente, las mismas circunstancias. Este modo de entender la probabilidad de un suceso se conoce como **interpretación frecuentista**.

Ejemplo (composición del aire, interpretación frecuentista de la probabilidad). Si alguien nos dice que el aire está formado principalmente por nitrógeno, en un 78 %; oxígeno, en un 21 % y argón, en un 1 %, esos porcentajes vendrían a decir que 78 de cada 100 moléculas que hay en el aire, son nitrógeno; 21 de cada 100, son oxígeno; 1 de cada 100, argón... Naturalmente, no se han contado todas las moléculas que hay y, sin embargo, parece que se pueden dar estimaciones de la proporción de moléculas de cada tipo que hay en el aire. Bajo la interpretación frecuentista de la probabilidad, si esas estimaciones fuesen correctas, esos porcentajes querrían decir que si repetimos una y otra vez, en las mismas condiciones, el experimento consistente en coger una molécula del aire (al azar y con la misma probabilidad todas ellas[12]) y ver qué molécula es, entonces las proporciones o **frecuencias relativas** que obtendríamos de cada tipo de molécula, al recoger "muchísimas", deberían aproximarse a aquellos porcentajes.

Otra cuestión es cómo podríamos llegar a estimar adecuadamente esos porcentajes (parámetros de la población total de moléculas en el aire) y cuánto podemos confiar en nuestras estimaciones, partiendo de que desconocemos la proporción real de la población. De ello se encarga la Inferencia Estadística, que utiliza herramientas de la Probabilidad aplicadas a estadísticos (es decir, operaciones sobre colectivos de datos que son muestras).

El enfoque frecuentista, además de una interpretación, es un modo de asignar probabilidades a un suceso. Repetimos un experimento "muchísimas" veces y entendemos que la

[12]La idea de repetir un experimento aleatorio, una y otra vez, en las mismas condiciones, de manera independiente, y recoger los resultados en un colectivo de datos se corresponde con la noción de *muestra aleatoria simple*, central en Inferencia Estadística.

probabilidad de un suceso es el límite de la proporción de veces (de la frecuencia relativa) con la que este ocurre al coger más y más datos cada vez.

Otro modo más sencillo, práctico y en ocasiones más adecuado, es asignar probabilidades a sucesos con lo que se conoce como **regla de Laplace**. Si llamamos 'A' a un suceso y $P(A)$ a su probabilidad, la regla de Laplace consiste en asignar la siguiente probabilidad:

$$P(A) = \frac{\text{número de casos favorables a que suceda } A}{\text{número de casos posibles}} \ .$$

La regla de Laplace se aplica siempre bajo la asunción de que todos los resultados posibles son posibles con la misma probabilidad. Cuando todos los posibles resultados de un experimento son igualmente probables, se dice que son **equiprobables**.

Ejemplo (composición del aire, regla de Laplace). Si fuera cierto que el aire está compuesto en un 78 % de moléculas de nitrógeno, en un 21 % de oxígeno y en un 1 % de argón, los posibles resultados del experimento que consiste en observar una partícula y ver de qué tipo es serían los elementos del conjunto siguiente: $\{\text{nitrógeno}, \text{oxígeno}, \text{argón}\}$. La regla de Laplace consistiría en asignar a cada uno de los posibles resultados la siguiente probabilidad:

$$P(A) = \frac{1}{100} = 0.01 \ , \quad P(O) = \frac{21}{100} = 0.21 \quad \text{y} \quad P(N) = \frac{78}{100} = 0.78 \ ,$$

donde A es el suceso 'la molécula detectada es argón', O es el suceso 'la molécula detectada es oxígeno' y N es el suceso 'la molécula detectada es nitrógeno'.

Los sucesos anteriores reciben el nombre de **sucesos elementales**. También podemos calcular la probabilidad de **sucesos compuestos**, como los siguientes. Llamemos $A \cup N$ al suceso 'la molécula detectada es argón o nitrógeno', $O \cup N$ al suceso 'la molécula detectada es oxígeno o nitrógeno' y $A \cup O$ al suceso 'la molécula detectada es argón u oxígeno'. Sus probabilidades, en este ejemplo, serían:

$$P(A \cup N) = \frac{79}{100} = 0.79 \ , \quad P(O \cup N) = \frac{99}{100} = 0.99 \quad \text{y} \quad P(A \cup O) = \frac{22}{100} = 0.22 \ .$$

En este caso, en el que asumimos que solo hay partículas de esos tres tipos, la probabilidad del suceso 'la molécula detectada es nitrógeno, argón u oxígeno' sería:

$$P(A \cup N \cup O) = \frac{100}{100} = 1 \ .$$

Lo que en Estadística Descriptiva recibía el nombre de modalidades, en probabilidad se conoce como **espacio muestral** o **suceso seguro** y lo escribiremos como S[13]. El espacio muestral es el suceso compuesto formado por todos los posibles resultados del experimento con el que se trabaja, de modo que su probabilidad es siempre 1 (naturalmente, el 100 % de las veces que repitamos el experimento en las mismas condiciones el resultado será uno de los posibles resultados).

Si llamamos $O \cap N$ al suceso 'la molécula detectada es argón y oxígeno' — es decir,

[13] El espacio muestral de un experimento suele denotarse por la letra griega omega mayúscula, Ω. También suele denotarse por E, de 'espacio', y, como nosotres haremos, por S, de 'space'.

los dos tipos al mismo tiempo—, como esto no puede ocurrir, le asignaremos una probabilidad cero. Un suceso que no puede ocurrir se denota por $\emptyset$ (conjunto vacío) y se llama **suceso imposible**. Escribiríamos:

$$P(O \cap N) = P(\emptyset) = 0 \ .$$

Cuando dos sucesos concretos no pueden ocurrir a la vez, se dice que son **sucesos incompatibles**.

Además de mediante la interpretación frecuentista y la regla de Laplace, también se pueden asignar probabilidades a través de modelos teóricos que involucren longitudes, áreas o volúmenes y de otros modos como en el enfoque bayesiano.

Para empezar, hay que tener en mente los siguientes conceptos fundamentales:

- ☞ Se dice que un experimento es aleatorio cuando, aun repitiéndolo en las mismas condiciones exactamente, el resultado no es necesariamente el mismo.
- ☞ Un suceso es un posible resultado o una afirmación sobre los posibles resultados de un experimento aleatorio.
- ☞ La probabilidad de un suceso es la proporción de ocasiones que se espera que ese suceso ocurra de antemano.

ii. Cálculo de probabilidades

Operaciones entre sucesos.

Escribiremos los sucesos que no sean el suceso seguro y el imposible con letras mayúsculas (A, B, C, D...). Hay una relación muy clara entre las operaciones entre sucesos y algunos operadores en lógica proposicional.

- ☞ $A \cup B$: unión de A y B (disyunción no excluyente en lógica proposicional). $A \cup B$ es el suceso 'ocurre A u ocurre B'.
- ☞ $A \cap B$: intersección de A y B (conjunción en lógica proposicional). Expresa el suceso 'sucede A y sucede B', a la vez.
- ☞ $\overline{A}$ o A^c: complementario de A (negación en lógica). Se trata del suceso 'no ocurre A'.
- ☞ $B - A$: diferencia de sucesos. Se corresponde con el suceso 'ocurre B, pero no ocurre A'. Se verifica la igualdad $B - A = B \cap \overline{A}$.

Para representar los sucesos, es muy útil recurrir a **diagramas de Venn**[14]. En ellos, se representa el espacio muestral, S, como una región en el plano y los sucesos como regiones dentro de S. En esos diagramas, la probabilidad de cada suceso está representada por la proporción de área asociada a cada suceso, respecto al área del espacio muestral, S. Es decir, en un diagrama de Venn:

$$P(A) = \frac{\text{área de } A}{\text{área de } S} \ , \quad P(B) = \frac{\text{área de } B}{\text{área de } S} \ , \quad \text{etc.}$$

Recordemos que:

[14] Mira el segundo ejercicio de la tercera tanda de ejercicios o este enlace.

- ☞ El espacio muestral (S) es el suceso compuesto de todos los posibles resultados de un experimento. También lo llamaremos suceso seguro y siempre cumple la igualdad $P(S) = 1$.
- ☞ Un suceso imposible ($\emptyset$) será todo aquel que no pueda ocurrir. Siempre satisface la igualdad $P(\emptyset) = 0$.
- ☞ Se dice que dos sucesos, A y B, son incompatibles o mutuamente excluyentes si $A \cap B = \emptyset$. Es decir, si no pueden ocurrir a la vez. Si dos sucesos incompatibles se dibujan en un diagrama de Venn, las respectivas regiones no solapan.

Es muy importante, tener en mente las siguientes propiedades de la unión y la intersección de sucesos:

- ☞ $A \cup A = A$ y $A \cap A = A$ (idempotencia).
- ☞ $A \cup B = B \cup A$ y $A \cap B = B \cap A$ (conmutatividad).
- ☞ $(A \cup B) \cup C = A \cup (B \cup C) = A \cup B \cup C$ y $(A \cap B) \cap C = A \cap (B \cap C) = A \cap B \cap C$ (asociatividad).
- ☞ $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ y $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$ (distributividad).

A continuación se definirá una probabilidad de forma axiomática (es decir, a partir de unos pocos principios o postulados) y se verán algunas consecuencias de tomar esa definición.

Definición axiomática de probabilidad y propiedades.

Llamaremos función de probabilidad a toda función, P, que a cada suceso de un espacio muestral, S, le asigne un número y verifique además las siguientes tres propiedades:

1. (No negatividad) Cualquier suceso, A, cumple $P(A) \geq 0$.
2. El espacio muestral cumple $P(S) = 1$.
3. (Aditividad) Si dos sucesos, A y B, son incompatibles (es decir, $A \cap B = \emptyset$), entonces $P(A \cup B) = P(A) + P(B)$.

Tan solo con estos tres principios (o sea, axiomas) son ciertas las siguientes relaciones: (i) $P(\overline{A}) = 1 - P(A)$; (ii) $0 \leq P(A) \leq 1$; (iii) $P(A - B) = P(A) - P(A \cap B)$; (iv) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Probabilidad condicionada.

En probabilidad, al igual que ocurría en estadística bidimensional[15], hay algunas relaciones que permiten calcular la probabilidad de un suceso condicionado a que otro ocurra. Cuando se condiciona por un suceso se habla de **probabilidad condicionada** y se cumple la siguiente relación análoga a aquella de las frecuencias relativas:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \ ,$$

[15] Recuerda las relaciones que se daban como, por ejemplo, $f_{j|X=v_i} = f_{ij}/f_{i\cdot}$.

para cualesquiera sucesos A y B. $P(A|B)$ es la probabilidad de A, condicionada a que ocurra B y $P(\cdot|B)$ es una probabilidad. Es decir, la probabilidad condicionada a cualquier suceso verifica también la definición axiomática de probabilidad y, por lo tanto, sus propiedades. Para cualquier suceso A: (i) $P(\overline{A}|B) = 1 - P(A|B)$; (ii) $0 \leq P(A|B) \leq 1$; (iii) $P(A \cup C|B) = P(A|B) + P(C|B) - P(A \cap C|B)$, para cualquier suceso C.

La probabilidad de la intersección de conjuntos se puede entonces calcular también como:

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) \ ,$$

al despejar de la intersección de $P(A|B)$ o $P(B|A)$, respectivamente. Estas dos formar de tratar la intersección son muy útiles.

Independencia de sucesos.

De nuevo, al igual que en estadística bivariante[16]:

☞ Se dice que dos **sucesos**, A y B, son **independientes** si se cumple que:

$$P(A \cap B) = P(A)P(B) \ .$$

Es decir, si la probabilidad de la intersección es el producto de las probabilidades de cada suceso.

Hay que tener en cuenta que, sustituyendo en las expresiones anteriores de la probabilidad condicionada, se tiene que si dos sucesos son independientes, entonces cumplen también:

$$P(A|B) = P(A) \ \ (\text{si } P(B) \neq 0) \quad \text{y} \quad P(B|A) = P(B) \ \ (\text{si } P(A) \neq 0) \ .$$

Teorema de la probabilidad total y Teorema de Bayes.

Hay dos teoremas (es decir, dos consecuencias especialmente importantes de los axiomas) que son de gran utilidad y podemos deducir de manera sencilla a partir de las propiedades introducidas.

El primero es el Teorema de la probabilidad total y sirve para calcular la probabilidad de un suceso, A, cuando esta no se conoce directamente, sino que se conoce la probabilidad de A condicionada a varios sucesos $B_1, \ldots, B_m$ que forman un 'sistema completo de sucesos', además de conocer la probabilidad de estos últimos.

☞ Se dice que unos sucesos, $B_1, \ldots, B_m$, forman un **sistema completo de sucesos** si la unión de todos ellos es el suceso seguro y, además, esos sucesos son incompatibles dos a dos. Esto es, si no intersecan dos a dos ($B_j \cap B_{j'} = \emptyset$ para cada $j, j' = 1, \ldots, m$, con $j \neq j'$) y juntos 'cubren' todas las posibilidades ($B_1 \cup \cdots \cup B_n = S$).

El **Teorema de la probabilidad total** viene a decir lo siguiente:

$$P(A) = \sum_{j=1}^{m} P(A \cap B_j) = \sum_{j=1}^{m} P(A|B_j)P(B_j) = P(A|B_1)P(B_1) + \cdots + P(A|B_m)P(B_m) \ .$$

[16] La independencia estadística de dos colectivos de datos se traducía en que se verifique la relación $f_{ij} = f_{i\cdot} f_{\cdot j}$

De otro lado, en las mismas condiciones, el Teorema de Bayes permite calcular las probabilidades condicionadas $P(B_j|A)$ conociendo las probabilidades $P(A|B_j)$ y $P(B_j)$ (es decir, en los mismos supuestos que en el Teorema de la probabilidad condicionada). El **Teorema de Bayes** viene a decir que:

$$P(B_j|A) = \frac{P(B_j \cap A)}{P(A)} = \frac{P(A|B_j) \cdot P(B_j)}{P(A|B_1)P(B_1) + \cdots + P(A|B_m)P(B_m)} \ .$$

iii. La noción de variable aleatoria (v.a.).

En Estadística Descriptiva hablábamos de variables estadísticas (que eran rasgos o características bajo observación en un experimento) y estas se concretaban en colectivos de datos tras llevar a cabo el experimento.

En cambio, en Probabilidad pasaremos de hablar de variables estadísticas a hablar de variables aleatorias. Con cada variable aleatoria, llevaremos cada uno de los posibles resultados del experimento a los números reales y estas permitirán trasladar la probabilidad del espacio muestral (S) a los números reales ($\mathbb{R}$), donde tendremos funciones con las que calcular probabilidades.

- ☞ Llamaremos **variable aleatoria** (**v.a.**), X, a una función que a cada posible resultado de un experimento aleatorio (con su espacio muestral, S) le asigna un número real, lo cual abreviaremos por $X : S \longrightarrow \mathbb{R}$.
- ☞ Todo suceso de S, A, se llevará a los números reales a través de X. Por lo tanto, todo suceso puede ser también visto como un conjunto de números (trabajaremos solo con intervalos y con conjuntos de números puntuales).
- ☞ La probabilidad de un conjunto de números reales, a través de una variable aleatoria específica, será justamente la probabilidad de los sucesos que caen en ese conjunto de números a través de la variable aleatoria.

Ejemplo (baraja de cartas). Considerar el experimento que consiste en tomar una carta de una baraja española y ver las naranjas que gana (o pierde) Rigoberta, quien apostó dos naranjas a que salía una carta de bastos. Una forma de transformar este enunciado a una v.a. es tomar el espacio muestral $S = \{\text{sale bastos, no sale bastos}\}$ y llamar 'Y' a la v.a. $Y =$'variación del número de naranjas que Rigoberta tendrá'. Si sale una carta de bastos, Rigoberta tendrá dos naranjas más; mientras que si no sale bastos, tendrá dos naranjas menos. Es decir:

$$\begin{array}{rccc} Y: & S & \longrightarrow & \mathbb{R} \\ & \{\text{bastos}\} & \longmapsto & 2 \\ & \{\text{no bastos}\} & \longmapsto & -2 \ , \end{array}$$

la v.a. llamada Y asigna un 2 o un -2 en función de si sale o no bastos, respectivamente. ¿Cuál es la probabilidad de que Y sea igual a -2?

$$P(Y = -2) = P(\text{Rigoberta pierda dos naranjas}) = P(\{\text{no bastos}\}) = \frac{3}{4} \ .$$

¿Cuál es la probabilidad de que gane dos naranjas?

$$P(\text{Rigoberta gane dos naranjas}) = P(Y = 2) = P(\{\text{bastos}\}) = \frac{1}{4} \ .$$

Como hemos llevado los sucesos a los números reales, podemos preguntarnos también, por ejemplo, por: ¿cuál es la probabilidad de que Y sea menor estricto que menos 1? ¿Y de que sea mayor o igual que cero? ¿Y de que sea mayor estricto que -5 y menor o igual que 3.14?

$$P(Y<-1)=\frac{3}{4}=P(Y=-2)\ ,\quad P(Y\geq 0)=\frac{1}{4}=P(Y=2)\ ,$$
$$P(-5<Y\leq 3.14)=P((Y=-2)\cup(Y=2))=P(Y=-2)+P(Y=2)=1\ ,\ \text{etc.}$$

iv. Función de masa o densidad y función de distribución de variables aleatorias. Modelos e interpretaciones.

Al igual que en Estadística Descriptiva, clasificaremos las v.a. en discretas y continuas en función de si los valores numéricos que estas pueden tomar forman un conjunto discreto o de si se trabaja por intervalos, respectivamente. Una v.a. puede tomar cualquier valor en, al menos, un intervalo.

Variables aleatorias discretas.

☞ Diremos que una v.a., X, es discreta si los posibles valores distintos que la v.a. toma son números separados bien identificados $\{v_1,\ldots,v_k,\ldots\}$ (posiblemente infinitos, pero en todo caso numerables). Cada uno de esos valores tiene una probabilidad de ocurrir: $p_1,\ldots,p_k,\ldots$.

Las v.a. discretas están caracterizadas por su función de masa y por su función de distribución.

☞ La **función de masa** o función de probabilidad, $P(X=x)$, de una v.a. discreta es una función que da la probabilidad de cada número real. Es decir, para cada número x en $\mathbb{R}$:

$$P(X=x)=\begin{cases} p_j\ , & \text{si } x=v_j \text{ para algún } j=1,\ldots,k,\ldots \\ \\ 0\ , & \text{en otro caso .}\end{cases}$$

☞ La **función de distribución**, $F(x)$, de una v.a. discreta es una función que, para cada $x\in\mathbb{R}$:

$$F(x)=P(X\leq x)=\sum_{v_j\leq x}p_j$$

Nótese que la función de distribución recuerda a la idea de frecuencia relativa acumulada; la de probabilidad de cada valor, a la de frecuencia relativa y la de los posibles valores de una v.a., a la de modalidades.

Dada una v.a. discreta, X, llamaremos **esperanza** o **valor esperado** de X ($\mathbb{E}[X]$) y llamaremos **varianza** de X ($\mathrm{Var}[X]$) a:

$$\mathbb{E}[X]=\sum_{j=1,\ldots,k,\ldots}v_jp_j\quad\text{y}\quad \mathrm{Var}[X]=\mathbb{E}[(X-\mathbb{E}[X])^2]=\mathbb{E}[X^2]-(\mathbb{E}[X])^2\ .$$

Es decir, el valor esperado de una v.a. recuerda a la media de un colectivo de datos y su varianza, a la varianza de un colectivo de datos.

Para cualquier par de números reales a y b, y cualquier par de v.a., X e Y, la esperanza cumple las siguientes propiedades:

$$\mathbb{E}[aX+b] = a\mathbb{E}[X] + b \quad \text{y} \quad \mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y] \ .$$

De otro lado, la varianza verifica:

$$\text{Var}[X] \geq 0 \ , \quad \text{Var}[aX] = a^2 X \quad \text{y} \quad \text{Var}[X+b] = \text{Var}[X] \ .$$

Finalmente, cabe mencionar la idea y propiedades que cumplen dos variables estadísticas discretas cuando son independientes. Se dice que dos variables estadísticas discretas, X e Y, son **independientes** si:

$$P((X=x) \cap (Y=y)) = P(X=x) \cdot P(Y=y) \ .$$

Cuando dos v.a. discretas son independientes, además de las propiedades anteriores se cumple lo siguiente:

$$\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y] \quad \text{y} \quad \text{Var}[X \pm Y] = \text{Var}[X] + \text{Var}[Y] \ .$$

La noción de independencia de v.a. es bastante similar a las ya introducidas anteriormente[17].

Algunos modelos discretos habituales.

Entre los modelos de variables aleatorias discretas, son especialmente frecuentes los siguientes:

Distribución de Bernoulli ($\mathcal{B}(p)$).

Supón que realizamos un experimento y estamos interesados en ver si ocurre algo (suceso de interés) o no ocurre y supón que la probabilidad de que ocurra ese suceso de interés es p (y de que no ocurra, $1-p$). La variable aleatoria $X =$'número de veces que ocurre el suceso de interés al hacer una vez el experimento' tiene dos posibles valores: 1, si ocurrió; 0, si no ocurrió. Entonces, se dice que X sigue una distribución Bernoulli de parámetro p y escribiremos $X \equiv \mathcal{B}(p)$.

La función de masa de una distribución de Bernoulli es:

$$P(X=x) = \begin{cases} p \ , & \text{si } x=1 \\ 1-p \ , & \text{si } x=0 \\ 0 \ , & \text{en otro caso} \ . \end{cases}$$

Es decir, para una distribución Bernoulli con parámetro p:

$$P(\text{ocurre el suceso de interés}) = P(X=1) = p \quad \text{y} \quad P(\text{no ocurre}) = P(X=0) = 1-p \ .$$

La esperanza y la varianza de esta v.a. son las siguientes:

$$\mathbb{E}[X] = p \quad \text{y} \quad \text{Var}[X] = p \cdot (1-p) \ .$$

[17] Es reseñable que en Estadística Descriptiva bidimensional también se cumple la relación $\overline{X \cdot Y} = \overline{X} \cdot \overline{Y}$ cuando dos colectivos de datos, X e Y, son estadísticamente independientes. Es decir, bajo la independencia estadística se tiene que la media del producto es el producto de las medias. Una consecuencia es que tanto la recta de regresión de Y sobre X como la de X sobre Y son constantes, porque la covarianza es justamente: $\sigma_{XY} = \overline{X \cdot Y} - \overline{X} \cdot \overline{Y} = 0$. (Independencia implica incorrelación)

Distribución Binomial ($B(n,p)$)**.**

Si repetimos n veces, de manera independiente, un experimento de Bernoulli de parámetro p (p probabilidad de que ocurra el suceso de interés), entonces la v.a. discreta X ='número de veces que ocurre el suceso de interés al hacer n veces el experimento' puede tomar los valores $0, 1, \ldots, n$. Es decir, puede ocurrir ese suceso de interés cero veces, una vez... o hasta un máximo de n veces. En estas condiciones, la v.a. sigue una distribución binomial de parámetros n y p. Escribiremos $X \equiv B(n,p)$.

La función de masa de una distribución de Binomial de parámetros n y p es:

$$P(X=x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} \ , & \text{si } x = 0, 1, \ldots, n \\ 0 \ , & \text{en otro caso} \ , \end{cases}$$

donde

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \quad \text{y} \quad n! = n \cdot (n-1) \cdot \cdots \cdot 2 \cdot 1 \ .$$

Es decir, para una distribución binomial con n repeticiones y probabilidad p de que en cada repetición ocurra el suceso de interés:

$$P(\text{el suceso de interés ocurre } k \text{ veces en } n \text{ repeticiones}) = P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \ .$$

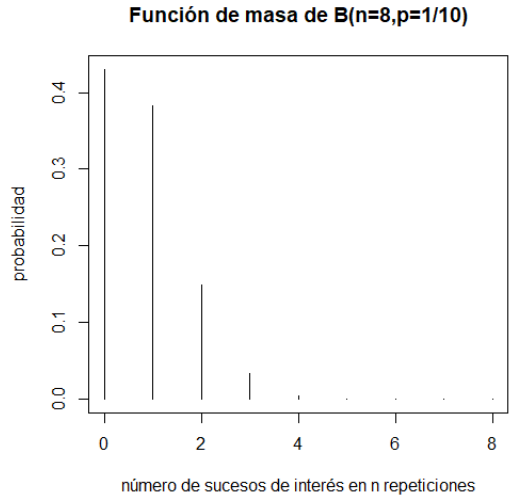
Su esperanza y su varianza son las siguientes:

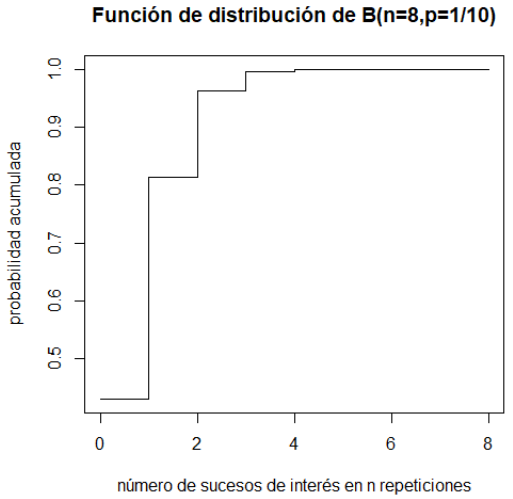
$$\mathbb{E}[X] = np \quad \text{y} \quad \text{Var}[X] = np \cdot (1-p) \ .$$

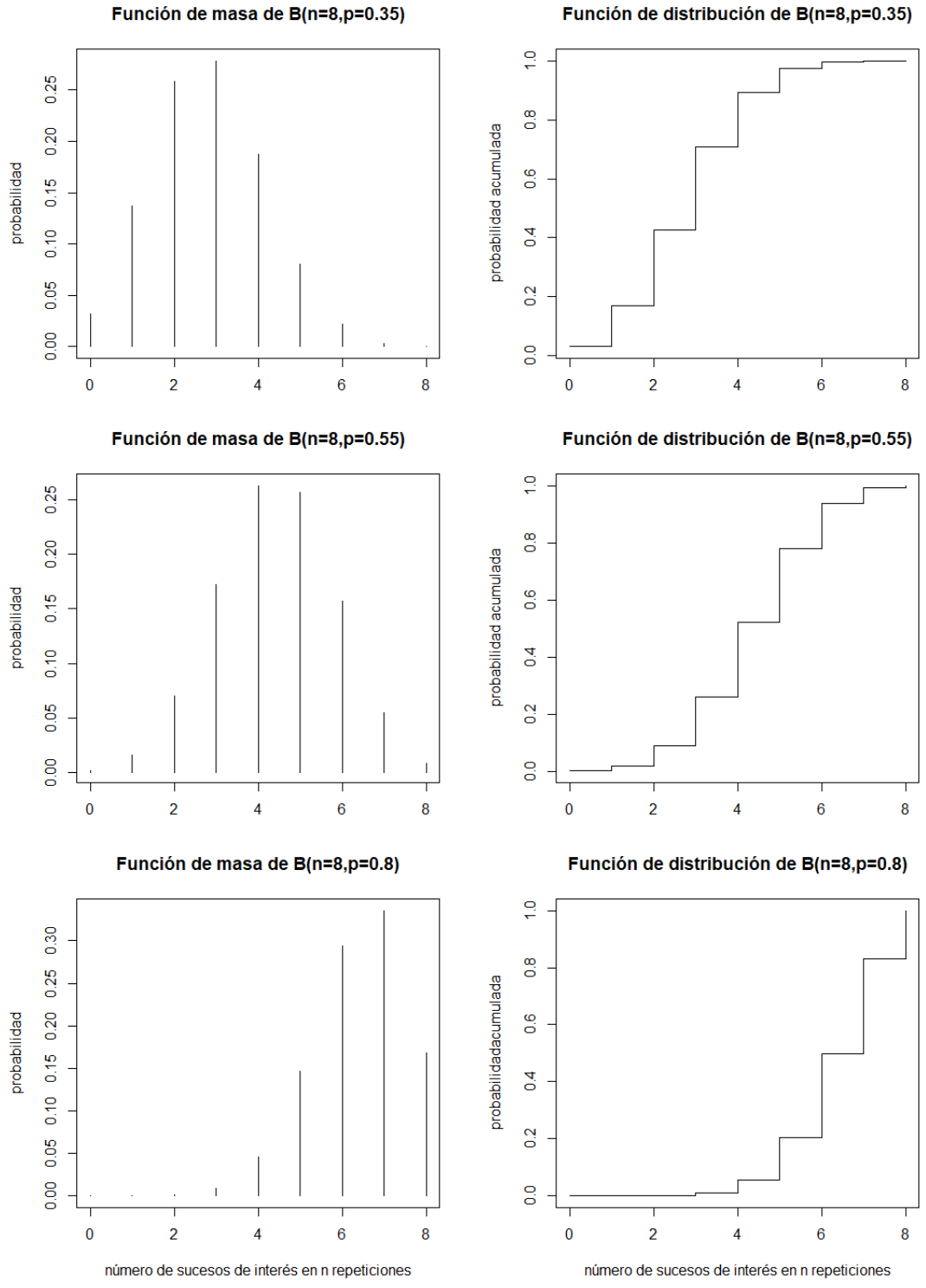
Es destacable que si $X \equiv B(n,p)$, entonces X puede escribirse como la suma de n v.a. independientes que siguen una distribución de Bernoulli:

$$X = X_1 + \cdots + X_n \ , \quad \text{con} \quad X_1 \equiv \mathcal{B}(p) \ , \ \ldots \ , \ X_n \equiv \mathcal{B}(p) \ .$$

Además, una distribución de Bernoulli es exactamente lo mismo que una binomial con una repetición. Es decir: $\mathcal{B}(p) \equiv B(1,p)$. El resultado anterior es entonces un caso particular de la siguiente **propiedad reproductiva**: si X e Y son dos v.a. independientes, donde $X \equiv B(n_1,p)$ e $Y \equiv B(n_2,p)$ (con la misma probabilidad, p, del suceso de interés), entonces: $X + Y \equiv B(n_1+n_2,p)$.







<u>Distribución de Poisson</u> ($\mathcal{P}(\lambda)$).

La distribución de Poisson se utiliza cuando se está observando el número de veces que ocurre un suceso de interés a lo largo de un intervalo concreto (de, por ejemplo, tiempo, área, longitud, etc) y siempre se tienen las mismas condiciones experimentales a lo largo de la observación. Además, los resultados del experimento deben ser independientes si se observan en intervalos que no se solapan y no puede ocurrir más de una vez al mismo

tiempo el suceso de interés. También, el promedio de veces que ocurre el suceso de interés por unidad de tiempo, longitud, área... debe ser constante e igual a λ.

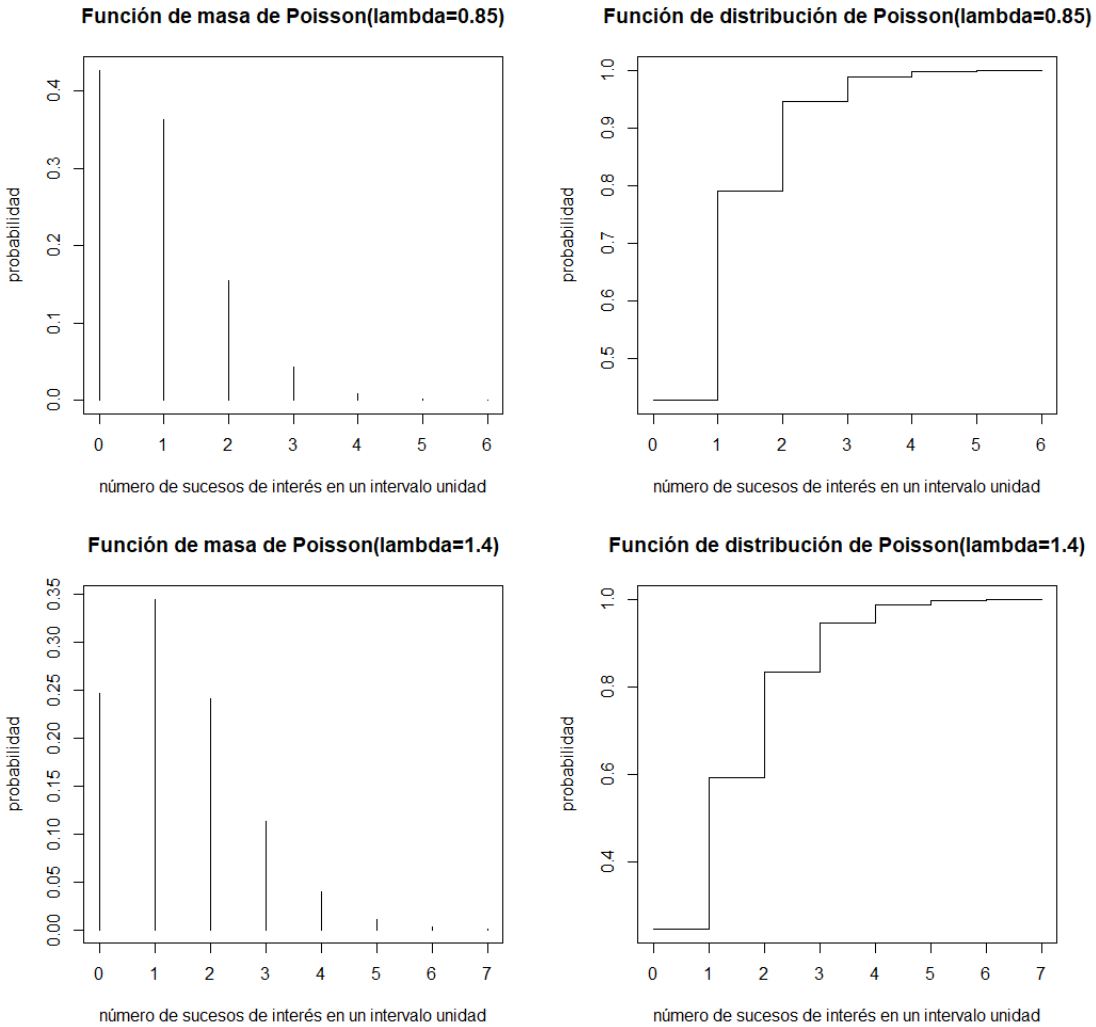
En estas circunstancias, si llamamos X a la v.a. $X =$'número de veces que ocurre el suceso de interés en un intervalo unidad', entonces $X \equiv \mathcal{P}(\lambda)$. Los posibles valores que esta variable puede tomar, son todos los números naturales, incluyendo al 0: $\mathbb{N} \cup \{0\} = \{0, 1, 2, \ldots\}$.

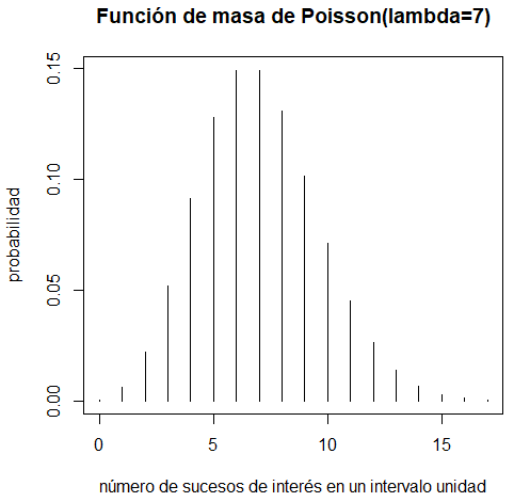
La función de masa de una distribución de Poisson de parámetro λ es:

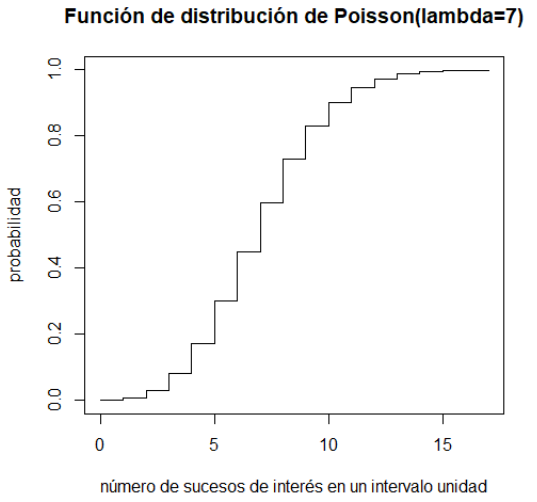
$$P(X = x) = \begin{cases} e^{-\lambda}\dfrac{\lambda^x}{x!}\ , & \text{si } x \in \mathbb{N} \cup \{0\} \\ \\ 0\ , & \text{en otro caso}\ , \end{cases}$$

Es decir, para una distribución de Poisson con parámetro λ:

$$P(\text{el suceso de interés ocurre } k \text{ veces en un intervalo unidad}) = P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}\ .$$







Tanto la esperanza como la varianza de una distribución de Poisson son iguales a λ:

$$\mathbb{E}[X] = \lambda = \text{Var}[X] \ .$$

Es también importante conocer la propiedad reproductiva de las distribuciones de Poisson: si X e Y son dos v.a. independientes, con $X \equiv \mathcal{P}(\lambda_1)$ e $Y \equiv \mathcal{P}(\lambda_2)$, entonces se tiene que: $X + Y \equiv \mathcal{P}(\lambda_1 + \lambda_2)$.

De hecho, si llamamos X_t a la v.a. X_t =‘número de veces que ocurre el suceso de interés en un intervalo de longitud t’, entonces $X_t \equiv \mathcal{P}(\lambda \cdot t)$.

Variables aleatorias continuas.

☞ Diremos que una v.a., X, es continua si los posibles valores de la v.a. cubren, al menos, todo un intervalo.

Las v.a. continuas están caracterizadas por su función de densidad y por su función de distribución.

☞ La **función de densidad**, $f(x)$, de una v.a. continua es una función que da la probabilidad de un conjunto de números a través de su integral. Para un intervalo $[a, b]$,

$$P(a \leq X \leq b) = \int_a^b f(x)dx \ .$$

En otras palabras, la probabilidad de un conjunto de números viene dada por el área bajo la función de densidad. En el caso discreto, la probabilidad venía dada por las longitudes bajo la función de masa. A diferencia del caso discreto, la probabilidad de que una v.a. continua tome un valor concreto es siempre nula:

$$P(X = x) = \int_x^x f(x)dx = 0 \ .$$

Naturalmente, una función de densidad siempre es positiva y su integral en todo $\mathbb{R}$ es 1:

$$0 \leq f(x) \quad \text{y} \quad \int_{-\infty}^{+\infty} f(x)dx = 1 \ .$$

☞ La **función de distribución**, $F(x)$, de una v.a. discreta es una función que, para cada $x \in \mathbb{R}$:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

Por lo tanto, como se cumple lo siguiente:

$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)$, se tiene que , $\boxed{P(a \leq X \leq b) = F(b) - F(a)}$.

Dado que la probabilidad de un solo número es cero en el caso continuo, las siguientes probabilidades son idénticas:

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b) \ .$$

De otro lado, el valor esperado o esperanza de una v.a. continua y la varianza se definen como sigue:

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} xf(x) \quad \text{y} \quad \text{Var}[X] = \int_{-\infty}^{+\infty} (x - \mathbb{E}[X])^2 f(x)dx = \mathbb{E}[(X - \mathbb{E}[X])^2] \ .$$

Además, bajo la independencia de v.a. continuas se cumple las mismas relaciones que en el caso discreto.

Algunos modelos continuos habituales.

Distribución exponencial (Exp(λ)).

Supón que trabajamos con la v.a. Y =‘número de veces que ocurre un suceso de interés en un intervalo unidad’ en un experimento de Poisson, es decir, $Y \equiv \mathcal{P}(\lambda)$. Recordemos que si llamamos Y_t =‘número de veces que ocurre el suceso de interés en un intervalo de longitud t’, entonces $Y_t \equiv \mathcal{P}(\lambda \cdot t)$.

Pues sucede que, si llamamos X a la v.a. X =‘tamaño del intervalo entre dos ocurrencias del suceso de interés’ (el mismo suceso de interés que en Y), esta v.a. sigue una distribución Exponencial de parámetro λ (el mismo λ que en Y) y se escribe $X \equiv \text{Exp}(\lambda)$.

La función de densidad de una distribución Exponencial de parámetro λ es:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} \ , & \text{si } x \geq 0 \\ 0 \ , & \text{en otro caso} \ . \end{cases}$$

La función de distribución de una exponencial es:

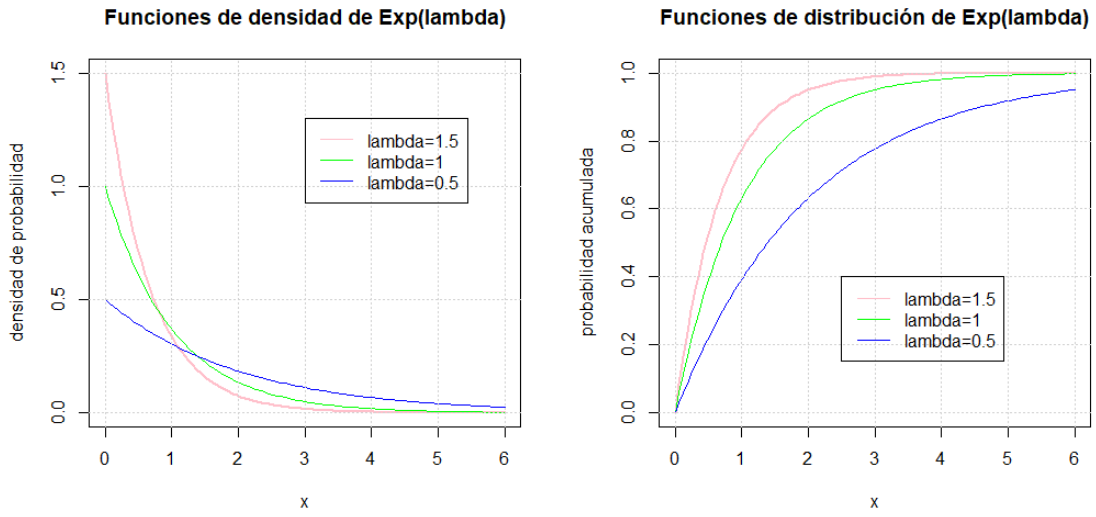
$$F(x) = \begin{cases} 1 - e^{-\lambda x} \ , & \text{si } x \geq 0 \\ \\ 0 \ , & \text{en otro caso} \ , \end{cases}$$

La relación entre $X \equiv \text{Exp}(\lambda)$ e $Y_t \equiv \mathcal{P}(\lambda \cdot t)$ puede verse por la función de distribución, a partir de las igualdades siguientes:

$$P(X \leq t) = 1 - P(X > t) = 1 - P(Y_t = 0) = 1 - e^{-\lambda t} \ .$$

En cuanto al valor esperado y la varianza de esta distribución, se tiene que:

$$\mathbb{E}[X] = \frac{1}{\lambda} \quad \text{y} \quad \text{Var}[X] = \frac{1}{\lambda^2} \ .$$



Por último, es destacable que la distribución exponencial verifica la siguiente propiedad de pérdida de memoria:

$$P(X > t + x | X > x) = P(X > t) \ , \quad \text{para cualesquiera } x, t > 0 \ .$$

Distribución de Weibull[18] **($W(k, \lambda)$).**

Se dice que una v.a., X, sigue una distribución de Weibull de parámetros $k > 0$ y $\lambda > 0$ si su función de densidad es:

$$f(x) = \begin{cases} \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k} \ , & \text{si } x \geq 0 \\ 0 \ , & \text{en otro caso} \ . \end{cases}$$

La función de distribución asociada es:

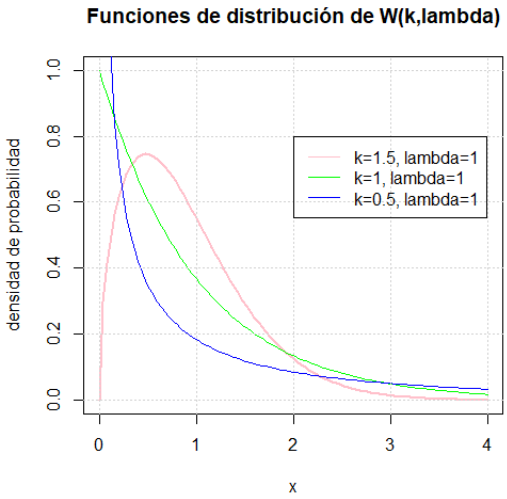
$$F(x) = \begin{cases} 1 - e^{-\left(\frac{x}{\lambda}\right)^k} \ , & \text{si } x \geq 0 \\ \\ 0 \ , & \text{en otro caso} \ , \end{cases}$$

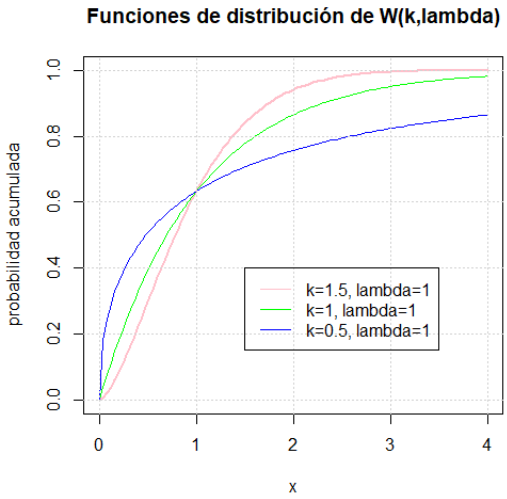
La distribución de Weibull, con parámetros $k > 0$ y $\lambda > 0$, es útil para modelar (entre otras cosas) el tiempo que transcurre hasta que ocurre un suceso de interés, cuando el promedio de veces que ocurre ese suceso de interés varía con el tiempo. En concreto, cuando la tasa de ocurrencias del suceso en cada tiempo t es justamente el siguiente término de la función de densidad:

$$\frac{k}{\lambda} \left(\frac{t}{\lambda}\right)^{k-1} \ ,$$

de modo que si $k > 1$, el promedio crece con el tiempo; si $k < 1$, el promedio decrece y si $k = 1$, el promedio es constante en el tiempo.

[18] Esta distribución es ampliación.





<u>Distribución normal</u> ($\mathcal{N}(\mu, \sigma^2)$).

La distribución normal es una de las distribuciones más utilizadas, ya que es muy habitual que multitud de fenómenos estén razonablemente bien descritos por ella. Ello encuentra explicación en el Teorema del Límite Central.

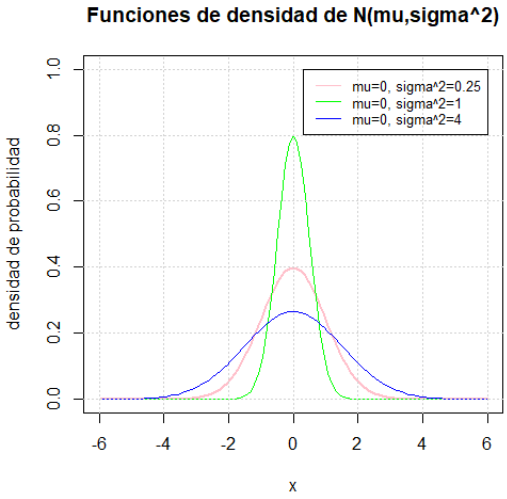
Una v.a., X, sigue una distribución normal de parámetros μ y σ^2 ($X \equiv \mathcal{N}(\mu, \sigma^2)$) si su función de densidad es la siguiente:

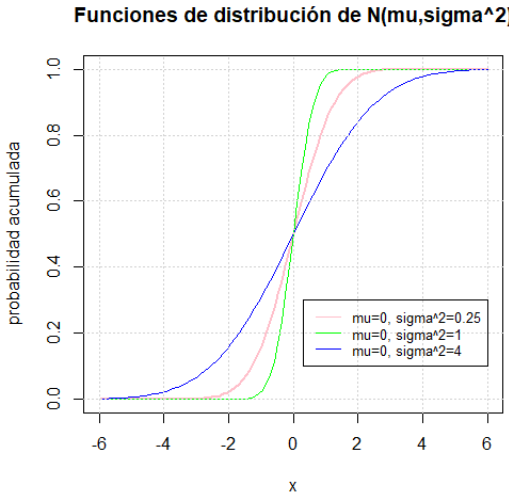
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad , \quad \text{para } x \in \mathbb{R} \ .$$

Los parámetros coinciden el valor esperado y la varianza:

$$\mathbb{E}[X] = \mu \quad \text{y} \quad \text{Var}[X] = \sigma^2 \ .$$

La representación gráfica de la densidad normal alcanza su máximo en $x = \mu$ y es simétrica respecto a ese valor.





Para cualesquiera constantes a y b, con $b \neq 0$, una distribución normal, $X \equiv \mathcal{N}(\mu, \sigma^2)$, verifica la siguiente propiedad:

$$a + bX \equiv \mathcal{N}\left(a + \mu, b^2\sigma^2\right)$$

De este modo, multiplicar por una constante simplemente modifica la varianza de la distribución, mientras que sumar una constante traslada la función de densidad.

También se tiene la siguiente propiedad reproductiva. Si $X \equiv \mathcal{N}(\mu_1, \sigma_1^2)$ e $Y \equiv \mathcal{N}(\mu_2, \sigma_2^2)$ son independientes, entonces:

$$X + Y \equiv \mathcal{N}\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right) \ .$$

De la primera propiedad se deduce que si $X \equiv \mathcal{N}(\mu, \sigma^2)$[19], entonces:

$$\boxed{\frac{X - \mu}{\sigma} \equiv \mathcal{N}(0, 1)}$$

Denotaremos por Z a la distribución normal de media nula y varianza 1, $Z \equiv \mathcal{N}(0, 1)$. A ella acudiremos para calcular las probabilidades acumuladas ($P(X \leq x)$) de cualquier distribución normal, tras realizar el proceso anterior (conocido como **tipificación**). Es decir, para calcular $P(X \leq x)$, haríamos:

$$\boxed{P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = P\left(Z \leq z\right)} \ ,$$

donde se ha llamado z a $z = (x - \mu)/\sigma$. Las probabilidades de $P\left(Z \leq z\right)$ se obtendrán a partir de una tabla. No obstante, en las tablas solo aparecen las colas izquierdas para valores de z positivos. Si queremos hallar la cola izquierda de un valor de z negativo, debemos proceder como sigue, utilizando la simetría respecto al cero de Z:

$$\boxed{\text{si } \ -z < 0 \ , \quad P(Z \leq -z) = P(Z \geq z) = 1 - P(Z \leq z)} \ .$$

Teorema del Límite Central.

El Teorema del Límite Central establece un resultado de gran importancia: si tenemos una cantidad muy grande, $n \gg 1$, de v.a. independientes, $X_1, \ldots, X_n$, que tienen la misma distribución, con valor esperado μ y varianza σ^2, entonces su suma sigue aproximadamente una distribución normal. Concretamente:

$$X_1 + \cdots + X_n \overset{\text{aprox.}}{\longrightarrow} \mathcal{N}(n\mu, n\sigma^2) \ .$$

De manera equivalente:

$$\frac{X_1 + \cdots + X_n}{n} \overset{\text{aprox.}}{\longrightarrow} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \ .$$

Es decir, sea cual sea la distribución que un gran grupo de v.a. siga (aunque no se conozca siquiera cuál es), la v.a. que es la suma de todas ellas se asemeja a una distribución normal.

[19] ¡Cuidado! También es muy habitual usar la notación $\mathcal{N}(\mu, \sigma)$, es decir, escribir la desviación típica como parámetro de la normal. Por ejemplo, en R el segundo parámetro denota la desviación típica.

Teorema del Límite Central para ($B(n,p)$).
Si $X \equiv B(n,p)$ y n es muy grande; entonces, como X puede escribirse como la suma de n v.a. de Bernoulli independientes de parámetro p, $X_1, \ldots, X_n \equiv \mathcal{B}(p)$:

$$X \equiv B(n,p) = X_1 + \cdots + X_n \overset{\text{aprox.}}{\longrightarrow} \mathcal{N}(np, np(1-p)) \ ,$$

ya que $\mathbb{E}[X_1] = \cdots = \mathbb{E}[X_n] = p$ y $\text{Var}[X_1] = \cdots = \text{Var}[X_n] = p(1-p)$.

Teorema del Límite Central para ($\mathcal{P}(\lambda)$).
Si $X \equiv \mathcal{P}(\lambda)$ y n y λ son muy grandes; entonces, como X puede escribirse como la suma de n v.a. de Poisson independientes de parámetro λ/n, $X_1, \ldots, X_n \equiv \mathcal{P}(\lambda/n)$:

$$X \equiv \mathcal{P}(\lambda) = X_1 + \cdots + X_n \overset{\text{aprox.}}{\longrightarrow} \mathcal{N}(\lambda, \lambda) \ ,$$

ya que $\mathbb{E}[X_1] = \cdots = \mathbb{E}[X_n] = \lambda/n = \text{Var}[X_1] = \cdots = \text{Var}[X_n]$.

IV. Inferencia Estadística

Como primera intuición, diremos que la Inferencia Estadística trata los métodos matemáticos que permiten extraer conclusiones acerca de un colectivo de datos global (llamado población, de tamaño N) a partir de un colectivo de datos parcial (llamado muestra, de tamaño n), siempre que la muestra sea escogida "apropiadamente". En ello media la probabilidad, pues con ella se cuantifica el riesgo de error en las conclusiones.

☞ Llamaremos **población**, de un experimento concreto con su contexto específico, al colectivo de datos formado por todos los datos existentes o posibles para ese experimento. Siempre representaremos la población mediante una variable aleatoria, con una **distribución de probabilidad concreta**, llevando cada elemento de la población (llamado individuo) a los números reales.

☞ Llamaremos **muestra observada** a cada selección concreta de datos de la población.

De un lado, en los estudios de Inferencia Estadística se parte del desconocimiento total o parcial de la población. Cada muestra observada, que debe ser "bien obtenida" si se quiere que sea representativa de la población, es la información experimental disponible acerca de esa población desconocida. Por otra parte, son herramientas de probabilidad las que permiten, a partir de cada posible muestra "adecuadamente extraída", sacar conclusiones con una confianza (probabilidad) controlada.

Debe estar claro que el método de obtención de una muestra es peliagudo y puede conllevar sesgos. Para establecer una noción de muestra observada "bien", "apropiada" o "adecuadamente" obtenida de antemano se recurre a técnicas de muestreo: nosotres trabajaremos con la idea de *muestra aleatoria simple observada*. Esto es, cada dato que forma parte de la muestra observada debe interpretarse como el resultado o realización de una variable aleatoria con la misma distribución de probabilidad que la población, siendo además cada una de las variables aleatorias —de las que proviene cada dato recogido en la muestra observada— independientes entre sí.

En otras palabras, el experimento consistente en extraer cada uno de los datos que conformarán la muestra se debe considerar aleatorio. Además, la aleatoriedad en la elección de cada dato se rige por la distribución de probabilidad de la población. También, la elección de cada dato debe ser independiente de la elección de otro dato (es decir, cada dato de la muestra es obtenido en las mismas condiciones y sin influirse entre sí).

☞ Sea X una v.a. representando una población, se llama **muestra aleatoria simple** de tamaño n a n variables aleatorias independientes e idénticamente distribuidas (v.a.i.i.d.), $X_1, X_2, \ldots, X_n$, cuya distribución de probabilidad es la misma que la de la población, X.

☞ Tras realizar el experimento consistente en recoger una muestra aleatoria simple (m.a.s.) de tamaño n, se obtienen n datos concretos, $x_1, x_2, \ldots, x_n$. Llamaremos a cada posible grupo de datos así obtenido **muestra aleatoria simple observada** (m.a.s. observada) de tamaño n.

Para poblaciones finitas, la noción de muestra aleatoria simple se basa en la siguiente idea: asignar a todos los posibles datos que conforman la población la misma probabilidad de aparecer como resultado de cada una de las v.a. de la m.a.s., $X_1, \ldots, X_n$, sin influir el resultado de alguna de estas en otra.

Ejemplo (personas de la Tierra que verifican una propiedad). Considera una propiedad o cualidad, "A", que pueda verificar o tener una persona de la Tierra. Supón que quieres conocer la proporción de personas de la Tierra que verifican la propiedad A. Por muchos motivos, no siempre es factible o de interés obtener todos los datos de una población: como en este caso, el tiempo o el coste podría ser inasumible o, directamente, podría entenderse imposible llegar a toda la población.

Como colectivo de datos, la población consistiría en las $N \simeq 8000$ millones de personas en la Tierra, cada una de las cuales puede verificar la propiedad A o no verificarla. Una m.a.s. observada de tamaño $n = 50$, por poner un caso, sería una selección de 50 personas de la población, $x_1, \ldots, x_{50}$, que han sido elegidos de forma aleatoria entre las 8000 millones de personas de la Tierra, teniendo todas ellas la misma probabilidad de aparecer como cada dato y sin influir ninguna elección en otra.

Partiendo de una m.a.s. observada, se procederá a extraer conclusiones acerca de la población. Cabe distinguir dos casos, según el objetivo del método de inferencia:

- ☞ Cuando se asume conocida la distribución de probabilidad que sigue la población, pero se desconocen uno o varios parámetros de aquella, se habla de **inferencia paramétrica**.
- ☞ En otro caso, se habla de **inferencia no paramétrica**.

Por ejemplo, si se asumiera que la población sigue una distribución Normal ($\mathcal{N}(\mu, \sigma^2)$) o Bernoulli ($\mathcal{B}(p)$), pero se desconoce algún parámetro (μ o σ^2, para la normal, o p, para la Bernoulli) y el objetivo es obtener conclusiones sobre esos parámetros, entonces se está haciendo inferencia paramétrica. Si, en cambio, se desconoce y se está interesada en conocer la distribución de probabilidad de una población o se quiere saber si dos variables aleatorias son independientes, se estaría frente a un problema de inferencia no paramétrica.

Ejemplo (composición molecular del aire). Supón que queremos conocer la proporción de moléculas de nitrógeno que hay en una región cerrada de aire y que contamos para ello con un detector capaz de identificar cualquier tipo de molécula. La población como colección de datos sería cada una de las moléculas del aire que hay en esa región, sea nitrógeno, oxígeno, argón, agua, etc. La primera cuestión a tratar es qué distribución de probabilidad se debería asociar a la población.

Si llamamos $p_{\text{nitrógeno}}$ a la proporción de moléculas del aire que son de nitrógeno, la variable aleatoria representando a la población debe tomarse como una Bernoulli de parámetro $p_{\text{nitrógeno}}$, $X \equiv \mathcal{B}(p_{\text{nitrógeno}})$:

$$\begin{array}{rccc} X: & S & \longrightarrow & \mathbb{R} \\ & \{\text{nitrógeno}\} & \longmapsto & 1 \\ & \{\text{no nitrógeno}\} & \longmapsto & 0\,. \end{array}$$

La proporción de moléculas que son nitrógeno, en la que se está interesada y que es desconocida, queda guardada como parámetro de la Bernoulli, $\mathcal{B}(p_{\text{nitrógeno}})$. Se tendría que $P(X=1) = P(\text{la partícula detectada es nitrógeno}) = p_{\text{nitrógeno}}$, mientras que $P(X=0) = $ $= P(\text{la partícula detectada no es nitrógeno}) = 1 - p_{\text{nitrógeno}}$.

Una muestra aleatoria simple de tamaño n consiste en n v.a. con la misma distibución que

la población, X, e independientes. En este caso, $X_1 \equiv \mathcal{B}(p_{\text{nitrógeno}}), \ldots, X_n \equiv \mathcal{B}(p_{\text{nitrógeno}})$. Si se obtiene una m.a.s. observada de tamaño n, ello quiere decir que se ha realizado el experimento consistente en recoger n datos, $x_1, \ldots, x_n$, de manera independiente y en las mismas condiciones: cada uno de ellos es un posible resultado de cada una de las v.a. de la m.a.s., $X_1, \ldots, X_n$. De esta manera, los datos recogidos en la muestra al repetir n veces el experimento (consistente en detectar una partícula de la población de forma independiente, en las mismas condiciones, y ver si cada una de ellas es nitrógeno o no lo es) han sido llevados a los números reales y se tratan como tales.

En este ejemplo, cada dato puede ser un cero o un uno y la media de los datos es justamente la proporción de moléculas que son nitrógeno en la muestra observada de tamaño n (lo que se llama *proporción muestral*).

$$\left.\begin{array}{rcl} x_1 & = & 0 \text{ ó } 1 \\ x_2 & = & 0 \text{ ó } 1 \\ \vdots & & \vdots \\ x_n & = & 0 \text{ ó } 1 \end{array}\right\} \Longrightarrow \quad \frac{x_1 + \cdots + x_n}{n} = \text{proporción muestral de Nitrógeno .}$$

En ello, como incidiremos más adelante, radica la importancia, en este ejemplo, de entender la población como una variable aleatoria $\mathcal{B}(p_{\text{nitrógeno}})$. Si quisiéramos hacer lo mismo para la proporción de moléculas de otro tipo, habría que cambiar la asignación de la v.a. poblacional por una Bernoulli con distinto parámetro: la proporción poblacional de moléculas que son de ese tipo.

Ejemplo (longitud de árboles). Supón que quieres conocer la altura de los árboles que hay en cierta región. Debido a la naturaleza continua de esa característica, es apropiado agrupar las longitudes por intervalos. La distribución de probabilidad de la población (continua, representando las posibles longitudes de los árboles y las frecuencias de cada intervalo de longitudes) estaría caracterizada por una función de densidad. Esa función de densidad debería ser el límite de un histograma con conteo relativo de una m.a.s. de tamaño n cuando $n \to +\infty$ (interpretación frecuentista).

Con poblaciones continuas, como es la longitud de los árboles, se podría estar interesada en saber si la población, X, sigue una distribución normal ($\mathcal{N}(\mu, \sigma^2)$, con ciertos parámetros μ y σ desconocidos), lo cual sería un problema de inferencia no paramétrica. Si se supiera que la población sigue una distribución normal, se podría estar interesada en conocer la media o la varianza de las longitudes de los árboles, es decir, los parámetros μ y σ^2 de aquella distribución normal. Esto último sería un problema de inferencia paramétrica.

Nosotres nos centramos sobre todo en aspectos teóricos y prácticos de inferencia paramétrica, pasando por estimación puntual, estimación por intervalos y test de hipótesis. Es decir, se estará centrado en parámetros que caracterizan la distribución de probabilidad de la población. En estimación puntual, el objetivo es dar un valor concreto a partir de cada m.a.s. observada, que sea "la mejor" aproximación a priori al verdadero valor de ese parámetro poblacional de interés. En cambio, en estimación por intervalos se busca

construir un intervalo para cada m.a.s. observada de tal manera que se tenga una confianza (probabilidad) controlada de que el parámetro poblacional de interés esté en esos intervalos.

La principal herramienta en nuestra ocupación inferencial son lo que conoce como *estadísticos*:

☞ Si X es una v.a. representando una población y $X_1, \ldots, X_n$ es una m.a.s. de esa población (recordemos, v.a.i.i.d. con la misma distribución de probabilidad que X), entonces cualquier función de esa m.a.s., $T(X_1, \ldots, X_n)$, recibe el nombre de **estadístico**. Por lo tanto, un estadístico **es una variable aleatoria**.

i. Estimación puntual.

Recordemos que el objetivo principal de la inferencia paramétrica es obtener información sobre parámetros poblacionales desconocidos (como p, si $X \equiv \mathcal{B}(p)$, o como μ ó σ^2, si $X \equiv \mathcal{N}(\mu, \sigma^2)$) a partir de una muestra.

Llamemos θ al parámetro poblacional que queremos conocer ($\theta = p$ ó μ ó σ^2...). En estimación puntual se busca dar, a partir de cada muestra observada, con un único valor, $\hat{\theta}$, que sirva como estimación del parámetro poblacional de interés, θ. Para ello, se recurre a *estimadores*:

☞ Un **estimador** para un parámetro θ es un estadístico cuyos posibles resultados son los posibles valores del parámetro θ.

Los estimadores con los que trabajaremos son la **media muestral**, $\overline{X_n}$, y la **varianza muestral** o cuasivarianza, S_n^2, definidas como:

$$\overline{X_n} = \frac{X_1 + \cdots + X_n}{n} \quad \text{y} \quad S_n^2 = \frac{(X_1 - \overline{X_n})^2 + \cdots + (X_1 - \overline{X_n})^2}{n-1} = \sum_{i=1}^{n} \frac{(X_i - \overline{X_n})^2}{n-1} \ .$$

La media muestral y la varianza muestral son funciones que dependen de la m.a.s., de modo que son estadísticos. Un estadístico es también una variable aleatoria y, como tal, puede calcularse su esperanza y su varianza.

Un buen estimador debe ser *insesgado*, *consistente* y *eficiente*. Nosotres desarrollaremos más la noción de estimadores insesgados:

☞ Se dice que un **estimador** es **insesgado** para estimar un parámetro θ cuando la esperanza de ese estimador es justamente el parámetro que se quiere estimar, θ.

Caso $\mathcal{N}(\mu, \sigma^2)$.

Si se tiene una población continua que sigue una distribución normal, $X \equiv \mathcal{N}(\mu, \sigma)$, entonces la media muestral y la varianza muestral son estimadores insesgados para μ y σ^2, respectivamente, ya que:

$$\mathbb{E}\left[\ \overline{X_n}\ \right] = \mathbb{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \frac{1}{n}\left(\mathbb{E}\left[X_1\right] + \cdots + \mathbb{E}\left[X_n\right]\right) = \frac{n\mu}{n} = \mu \ ,$$

y

$$\mathbb{E}\left[S_n^2\right] = \sigma^2 \ ,$$

para una m.a.s., $X_1, \ldots, X_n$[20]. Por lo tanto, $\overline{X_n}$ es un estimador insesgado para μ y S_n^2 lo es para σ^2. Es posible calcular la varianza de $\overline{X_n}$ de manera sencilla, con propiedades vistas de la Normal, gracias a la independencia de las v.a. de la m.a.s.:

$$\mathrm{Var}\left[\ \overline{X_n}\ \right] = \frac{1}{n^2}\left(\mathrm{Var}\left[X_1\right] + \cdots + \mathrm{Var}\left[X_n\right]\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \ .$$

La media muestral y la varianza muestral no solo son estimadores insesgados, sino también eficientes y consistentes, motivo por el cual se utilizan y los utilizaremos como estimadores puntuales de μ y σ^2, respectivamente.

Caso $\mathcal{B}(p)$.

Supón que se tiene una población discreta que sigue una distribución Bernoulli de parámetro p, $X \equiv \mathcal{B}(p)$, donde p representa la proporción de individuos de la población que satisfacen una propiedad. Entonces, la proporción muestral (que denotaremos por $\hat{p}$ y que coincide con la media muestral), es un estimador insesgado. Con los mismos razonamientos que en el caso anterior, se deduce fácilmente que su esperanza y su varianza son:

$$\mathbb{E}\left[\hat{p}\right] = p \ \ \text{y} \ \ \mathrm{Var}\left[\hat{p}\right] = \frac{p(1-p)}{n} \ .$$

La proporción muestral no solo es un estimador insesgado, sino que también es consistente y eficiente. Por ello, se utiliza y lo utilizaremos para estimar puntualmente el parámetro p de $\mathcal{B}(p)$.

ii. Estimación por intervalos.

El objetivo en estimación por intervalos es asignar a cada m.a.s. observada un intervalo, de manera que para el $(1-\alpha)\cdot 100\,\%$ de las m.a.s. observadas el parámetro de interés esté en su correspondiente intervalo. α es un número preestablecido y se llama **nivel de significación**, mientras que $1-\alpha$ es lo que se conoce como **confianza**.

Supón que queremos conocer un parámetro, θ, de una población. Un **intervalo de confianza** $1-\alpha$ para ese parámetro será un intervalo (aleatorio) delimitado por dos estadísticos, $[T_1(X_1, \ldots, X_n), T_2(X_1, \ldots, X_n)]$, verificando:

$$P(T_1(X_1, \ldots, X_n) \leq \theta \leq T_2(X_1, \ldots, X_n)) = 1 - \alpha \ .$$

Es decir, con una probabilidad $1-\alpha$ el parámetro poblacional de interés, θ, estará en el intervalo $[T_1(X_1, \ldots, X_n), T_2(X_1, \ldots, X_n)]$. La confianza es justamente esa probabilidad $1-\alpha$, mientras que el nivel de significación, α, es la probabilidad de que el parámetro no caiga dentro del intervalo de confianza (esto es, el riesgo de errar).

[20] La deducción de que la esperanza de la varianza muestral es justamente la varianza poblacional, σ^2, no es (muy) inmediata y por ello se omite. Lo sorprendente es que el mejor estimador para la varianza poblacional es lo que en Estadística Descriptiva llamábamos varianza muestral (diviendo por $n-1$) y no la propia varianza (dividiendo por n).

Para cada m.a.s. observada, $x_1, \ldots, x_n$, el intervalo aleatorio se concreta en un intervalo numérico: $[T_1(x_1, \ldots, x_n), T_2(x_1, \ldots, x_n)]$. Tener un intervalo de confianza $1-\alpha$ viene a decir que para el $(1-\alpha)\cdot 100\,\%$ de las m.a.s. observadas posibles, el verdadero valor del parámetro θ estará en el correspondiente intervalo numérico. En un $\alpha\cdot 100\,\%$ de las ocasiones, ello no ocurre.

La construcción de intervalos de confianza está basada en encontrar un *estadístico pivote* adecuado para ello:

☞ Un **estadístico pivote** para la construcción de intervalos de confianza para un parámetro θ se trata de un estadístico que, aun dependiendo del parámetro θ (que en general se desconoce), sí se conoce su distribución de probabilidad y ello permite construir intervalos de confianza.

Caso $\mathcal{B}(p)$.

Para construir un intervalo de confianza $1-\alpha$ para la proporción poblacional, p, con un α concreto y dada una población $X \equiv \mathcal{B}(p)$, lo primero es considerar una m.a.s., $X_1, \ldots, X_n$. La proporción muestral, que llamaremos $\hat{p}$, se trata del siguiente estimador:

$$\hat{p} = \frac{X_1 + \cdots + X_n}{n} \ .$$

Un posible intervalo de confianza para el parámetro p, siempre que el tamaño muestral sea suficientemente grande ($n \gg 1$), es el siguiente:

$$\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \ ,$$

donde $z_{\alpha/2}$ es un número real que deja a la derecha, en una distribución $\mathcal{N}(0,1)$, $\alpha/2$ de probabilidad. Como la $\mathcal{N}(0,1)$ es simétrica respecto al origen, $-z_{\alpha/2}$ deja a la izquierda una probabilidad igual también a $\alpha/2$. Es decir:

$$P(\mathcal{N}(0,1) \geq z_{\alpha/2}) = \alpha/2 \ \text{ y } \ P(\mathcal{N}(0,1) \leq -z_{\alpha/2}) = \alpha/2 \ .$$

De esta manera, si calculamos la probabilidad de que el verdadero parámetro poblacional esté en tal intervalo y hacemos unas pequeñas manipulaciones, se puede deducir que, como se buscaba:

$$\begin{aligned} &P\left(T_1(X_1, \ldots, X_n) \leq p \leq T_2(X_1, \ldots, X_n)\right) = \\ =&P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = \\ =&P\left(-z_{\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \leq z_{\alpha/2}\right) = 1 - 2P(\mathcal{N}(0,1) \geq z_{\alpha/2}) = 1-\alpha \ . \end{aligned}$$

El penúltimo paso se debe a una aproximación al Teorema del Límite Central. Como $\mathbb{E}[\hat{p}] = p$ y $\mathrm{Var}[\hat{p}] = \dfrac{p(1-p)}{n}$, si n es suficientemente grande:

$$\frac{\hat{p}-p}{\sqrt{\dfrac{p(1-p)}{n}}} \overset{\text{aprox.}}{\longrightarrow} \mathcal{N}(0,1) \ ,$$

aplicando el Teorema del Límite Central directamente. Ahora bien, con el estadístico de la ecuación anterior no sería posible construir un intervalo de confianza, de modo que es necesario hacer una aproximación más allá: aproximar $p(1-p)$ del denominador por $\hat{p}(1-\hat{p})$. En efecto, esto es una aproximación al Teorema del Límite Central, de modo que:

$$\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \overset{\text{aprox.}}{\longrightarrow} \mathcal{N}(0,1) \ .$$

Este último estadístico sería un **estadístico pivote**. Su distribución de probabilidad es conocida, aun involucrando al parámetro poblacional desconocido (p), y aquella permite construir intervalos de confianza $1-\alpha$ para este.

Es de destacar que para la construcción del intervalo de confianza, en este caso es necesario que la muestra sea suficientemente grande (pongamos, $n \geq 30$). También, que para hallar el intervalo de confianza $1-\alpha$ a partir de una m.a.s. observada, $x_1, \ldots, x_n$, simplemente se debe calcular la proporción muestral y sustituirla, junto al tamaño de la muestra, en el intervalo de partida. Además, hay que hallar la constante $z_{\alpha/2}$, para lo cual recurriremos a la tabla de la $\mathcal{N}(0,1)$.

Caso $\mathcal{N}(\mu, \sigma^2)$, parámetro de interés: μ (varianza conocida, $\sigma^2 = \sigma_0^2$).

Asume que la población sigue una distribución normal, $X \equiv \mathcal{N}(\mu, \sigma_0^2)$, donde σ_0^2 es conocido. Si tomamos una m.a.s. de esta población, la media muestral también sigue una distribución normal que tiene como esperanza a μ y como varianza, σ_0^2/n. Por lo tanto, por propiedades básicas de la normal,

$$\frac{\overline{X_n}-\mu}{\sigma_0/\sqrt{n}} \equiv \mathcal{N}(0,1) \ .$$

Este es el estadístico pivote que utilizaremos para construir, en este caso, intervalos de confianza $1-\alpha$ para μ. Tomando $z_{\alpha/2}$ como en el caso anterior, se tendría que:

$$1-\alpha = P\left(-z_{\alpha/2} \leq \frac{\overline{X_n}-\mu}{\sigma_0/\sqrt{n}} \leq z_{\alpha/2}\right) = P\left(\overline{X_n} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \overline{X_n} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) \ .$$

De esta manera, el intervalo

$$\left[\overline{X_n} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}, \overline{X_n} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right] \ ,$$

es un intervalo de confianza $1-\alpha$ para μ. Para calcular un intervalo concreto a partir de una m.a.s., simplemente habría que cacular la media de la muestra, el valor $z_{\alpha/2}$ y sustituirlos en la expresión anterior junto al tamaño de la muestra, n, y a la varianza poblacional, $\sigma^2 = \sigma_0^2$, que se conocería en este caso.

Caso $\mathcal{N}(\mu, \sigma^2)$, parámetro de interés: σ^2.

Para una población normal, en la que estamos interesadas en la varianza poblacional, utilizaremos el estadístico de contraste:

$$\frac{(n-1)S_n^2}{\sigma^2} \equiv \chi^2_{n-1} \ .$$

Este estadístico sigue una distribución **chi-cuadrado** con $n-1$ grados de libertad (que denotaremos por χ^2_{n-1}). Para cada grado de libertad posible se tiene una distribución de probabilidad distinta. Su función de densidad es, en cualquier caso, distinta de cero solo para valores positivos.

Considerando los valores $\chi^2_{n-1,\alpha/2}$ y $\chi^2_{n-1,1-\alpha/2}$ que dejan, respectivamente, $\alpha/2$ de probabilidad a la derecha y $\alpha/2$ de probabilidad a la izquierda en una χ^2_{n-1}, se encuentra que:

$$1-\alpha = P\left(\chi^2_{n-1,1-\alpha/2} \leq \frac{(n-1)S_n^2}{\sigma^2} \leq \chi^2_{n-1,\alpha/2}\right) = P\left(\frac{(n-1)S_n^2}{\chi^2_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S_n^2}{\chi^2_{n-1,1-\alpha/2}}\right).$$

Siendo así, un intervalo de confianza $1-\alpha$ resulta ser:

$$\left[\frac{(n-1)S_n^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S_n^2}{\chi^2_{n-1,1-\alpha/2}}\right] .$$

Para cada m.a.s. observada, el intervalo anterior se concretará en un intervalo numérico al sustituir la varianza muestral, el tamaño de la muestra, y las constantes $\chi^2_{n-1,\alpha/2}$ y $\chi^2_{n-1,1-\alpha/2}$, para lo cual acudiremos a tablas.

Cabe destacar que la distribución χ^2_{n-1} es exactamente la misma que la distribución de la suma de n v.a. independientes $(\mathcal{N}(0,1))^2$, de ahí que el estadístico pivote siga una χ^2_{n-1}:

$$\frac{(n-1)S_n^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X_n}}{\sigma}\right)^2 \equiv \chi^2_{n-1} \ .$$

Caso $\mathcal{N}(\mu,\sigma^2)$, parámetro de interés: μ (varianza desconocida).

Supón que nos enfrentamos a una población normal en la que desconocemos el valor de la varianza poblacional, σ^2, y estamos interesadas en el parámetro μ. Entonces, el estadístico pivote del penúltimo caso no se podría utilizar, ya que en él aparecía la varianza poblacional, que ahora es desconocida. Una opción fructífera es sustituir la varianza poblacional por la varianza muestral. Es decir, se empleará como estadístico pivote el siguiente:

$$\frac{\overline{X_n} - \mu}{S_n/\sqrt{n}} \equiv t_{n-1} \ ,$$

donde S_n es la desviación típica muestral. Este estadístico pivote sigue una distribución **t de Student** con $n-1$ grados de libertad, aquí denotada por t_{n-1}. Se trata de una distribución continua determinada por los grados de libertad (el tamaño de la muestra menos uno) y cuya función de densidad, al igual que la $\mathcal{N}(0,1)$, es simétrica respecto al cero.

Llamemos $t_{n-1,\alpha/2}$ al valor que deja a la derecha, en la distribución t_{n-1}, una probabilidad $\alpha/2$. Como la función de desnidad de t_{n-1} es simétrica, entonces $-t_{n-1,\alpha/2}$ deja a la izquierda una probabilidad también igual a $\alpha/2$. Por lo tanto,

$$\begin{aligned} 1-\alpha &= P\left(-t_{n-1,\alpha/2} \leq \frac{\overline{X_n} - \mu}{S_n/\sqrt{n}} \leq t_{n-1,\alpha/2}\right) = \\ &= P\left(\overline{X_n} - t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}} \leq \mu \leq \overline{X_n} + t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}\right) \ , \end{aligned}$$

y el siguiente intervalo es un intervalo de confianza $1-\alpha$ para μ:

$$\left[\overline{X_n} - t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}, \overline{X_n} + t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}\right] \ .$$

Para hallar el intervalo de confianza correspondiente a una m.a.s. observada, $x_1, \ldots, x_n$, habría que calcular la media y la varianza muestral de esta, y sustituirlas en la expresión anterior junto al tamaño muestral, n, y el valor de la constante $t_{n-1,\alpha/2}$ (que obtendremos mediante tablas).

Es destacable que la distribución t_{n-1} es la que sigue la división de una $\mathcal{N}(0,1)$ entre la raíz cuadrada de una χ^2_{n-1} que ha sido dividida por su grados de libertad, siendo además las dos variables independientes. Así, en efecto, se puede comprobar que:

$$\frac{\overline{X_n}-\mu}{S_n/\sqrt{n}} = \frac{\dfrac{\overline{X_n}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S_n^2}{(n-1)\sigma^2}}} \equiv t_{n-1} \ , \quad \text{ya que} \quad \frac{\overline{X_n}-\mu}{\sigma/\sqrt{n}} \equiv \mathcal{N}(0,1) \quad \text{y} \quad \frac{(n-1)S_n^2}{\sigma^2} \equiv \chi^2_{n-1} \ ,$$

y las dos últimas distribuciones son independientes.

Ejercicio: construye un intervalo de confianza al 95 % para la media de una población $\mathcal{N}(\mu, 9)$, a partir de una muestra observada de tamaño $n = 20$, cuya media muestral es $\overline{x_n} = 10$ y desviación típica muestral, $s_n = 5$.

Resolución. La confianza que se quiere es $1-\alpha = 0.95$, de modo que el nivel de significación es $\alpha = 0.05$ y $\alpha/2 = 0.025$. Suponiendo que la muestra proporcionada venga de una m.a.s., un intervalo de confianza 0.95 se concretaría en:

$$\left[\overline{x_n} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}, \overline{x_n} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right] \ ,$$

donde $\overline{x_n} = 10$, $z_{\alpha/2} = z_{0.025} = 1.96$ (tabla $\mathcal{N}(0,1)$), $\sigma_0 = 3 = \sqrt{9}$ y $n = 20$. Sustituyendo, el intervalo de confianza 0.95 para la muestra observada es, aproximadamente, $[8.685, 11.315]$.

Ejercicio: construye un intervalo de confianza con nivel de significación 0.04 para la varianza poblacional de una distribución normal, a partir de una muestra de tamaño 41 y desviación típica muestral igual a 5.

Resolución. Se quiere un nivel de significación $\alpha = 0.04$, de modo que $\alpha/2 = 0.02$ y la confianza, $1-\alpha = 0.96$. Suponiendo que la muestra observada sea un m.a.s. observada, un intervalo de confianza 0.96 para esa muestra sería:

$$\left[\frac{(n-1)s_n^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)s_n^2}{\chi^2_{n-1,1-\alpha/2}}\right] \ .$$

Dado que $n = 41$, $s_n^2 = 5^2 = 25$, $\chi^2_{n-1,\alpha/2} = \chi^2_{40,0.01} = 63.691$ y $\chi^2_{n-1,1-\alpha/2} = \chi^2_{40,0.99} = = 22.164$ (tabla χ^2), entonces el intervalo de confianza 0.96 para la varianza poblacional, con la muestra observada sería, aproximadamente, $[15.70, 45.12]$.

Ejercicio: construye un intervalo de confianza con confianza 0.998 para la proporción de una población de Bernoulli a partir de una muestra observada de tamaño 123 cuya proporción muestral es 0.43.

Resolución. Se quiere una confianza de $\alpha = 0.998$ o, lo que es lo mismo, un nivel de significación $1 - \alpha = 0.002$, de modo que $\alpha/2 = 0.001$. Suponiendo que la muestra observada provenga de una m.a.s., un intervalo de confianza sería:

$$\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] ,$$

donde $\hat{p} = 0.46$ y $n = 123$, para la muestra observada. Además, $z_{\alpha/2} = z_{0.001} = 3.09$. Es decir, el intervalo para esa muestra observada es $[0.3211, 0.5989]$, aproximadamente.

Finalizaremos la estimación por intervalos con las siguiente observaciones:

- ☞ Al aumentar la confianza, $1 - \alpha$, la longitud de los intervalos de confianza aumenta (es decir, disminuye la precisión de la estimación).
- ☞ Al aumentar el tamaño de la muestra, n, la longitud de los intervalos de confianza se reduce (es decir, la precisión de la estimación aumenta).

iii. Contrastes de hipótesis.

El objetivo en contrastes o test de hipótesis es construir una regla de decisión que, a partir de una m.a.s. observada, permita decidir si rechazar o no rechazar una hipótesis acerca de la población.

Cabe distinguir entre test paramétricos y test no paramétricos, en función de si las hipótesis son afirmaciones sobre los parámetros de una población o no, respectivamente. Nos centraremos en test paramétricos cuya construcción está muy relacionada con los intervalos de confianza. Las nociones aprendidas nos permitirán también llevar a cabo test no paramétricos con R Commander.

Supón que tenemos una población, X, que sigue una distribución de probabilidad concreta, pero que no conocemos uno de sus parámetros, θ (como puede ser el parámetro p de una $\mathcal{B}(p)$ o el parámetro μ ó σ^2 de una $\mathcal{N}(\mu, \sigma^2)$). Para un número concreto, θ_0, podríamos plantearnos las siguientes cuestiones: ¿es el parámetro θ igual a θ_0? ¿Es $\theta \leq \theta_0$? ¿Es $\theta \geq \theta_0$? ¿Es θ distinto de θ_0?

Para construir un test de hipótesis, plantearemos dos afirmaciones complementarias sobre el parámetro de interés. Es decir, dos suposiciones sobre el parámetro de interés que sean la una la negación de la otra. A una de esas hipótesis la llamaremos hipótesis nula (H_0) y, a la otra, hipótesis alternativa (H_1).

- ☞ La **hipótesis nula** (H_0) se rechaza si se observa una gran evidencia experimental en contra. En otro caso, se diría que no se rechaza H_0, es decir, que sería factible que la hipótesis nula fuese cierta.
- ☞ La **hipótesis alternativa** (H_1) solo se acepta si se rechaza la hipótesis nula.

Concretamente, nos centraremos en tres tipos de test: **test bilaterales**, **test unilaterales izquierda** y **test unilaterales derecha**.

Test bilateral	Test unilateral izquierda	Test unilateral derecha
$\begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta \neq \theta_0 \end{cases}$	$\begin{cases} H_0: & \theta \geq \theta_0 \\ H_1: & \theta < \theta_0 \end{cases}$	$\begin{cases} H_0: & \theta \leq \theta_0 \\ H_1: & \theta > \theta_0 \end{cases}$

Hay que tener muy presente que a la hora de tomar una decisión en un test de hipótesis (rechazar o no rechazar H_0) hay dos tipos de errores que pueden cometerse: rechazar H_0 cuando esta hipótesis era cierta en realidad o no rechazar H_0 cuando la misma era falsa. Al primer tipo de error se le llama **error de tipo I** y al segundo, **error de tipo II**.

Decisión \ Realidad	H_0 es cierta	H_0 es falsa
se aceptó H_0	ok	error tipo II
se rechazó H_0	error tipo I	ok

Lo ideal sería construir test de hipótesis pudiendo controlar la probabilidad de cometer los dos tipos de errores: el de tipo I y el de tipo II. No obstante, esto no es posible. Ello se debe a que, al minimizar el error de tipo I aumenta el de tipo II y viceversa.

Lo habitual es controlar la probabilidad de cometer un error de tipo I (lo cual se llama nivel de significación de un test).

☞ El **nivel de significación**, α, de un test es la probabilidad de rechazar la hipótesis H_0 bajo la suposición de que H_0 es cierta.

A partir de un nivel de significación preestablecido, α, se utilizará un *estadístico de contraste* para establecer con qué m.a.s. observadas se rechazaría H_0 y con cuáles no. El procedimiento es el siguiente: elegir un nivel de significación (por ejemplo, $\alpha = 0.01$, 0.02, 0.03...) y utilizar un estadístico de contraste que permita, a ese nivel de significación, decidir para qué m.a.s. observadas se rechazaría la hipótesis nula y para cuáles no.

☞ Diremos que una m.a.s. observada está en la **región de rechazo** para un test de hipótesis si aquella conduce a rechazar la hipótesis nula, H_0.

☞ Un **estadístico de contraste** es un estadístico, de distribución conocida, que permite cuantificar la discrepancia de una m.a.s. observada con la suposición de que H_0 fuera cierta en un test concreto. Dado un nivel de significación, α, el estadístico de contraste determina la región de rechazo de ese test.

Según la situación (población y parámetro de interés), trabajaremos con distintos estadísticos de contraste que pueden ser calculados explícitamente con una m.a.s. observada. Junto al nivel de significación, el estadístico de contraste determina la región de rechazo: si al evaluar una m.a.s. observada en el estadístico de contraste esa evaluación cae en la región de rechazo, entonces el resultado del test será rechazar la hipótesis nula (al nivel de significación dado). Si, por el contrario, aquella evaluación no cae en la región de rechazo, el resultado del test será no rechazar H_0.

La región de rechazo de un test bilateral serán las colas izquierda y derecha con probabilidad $\alpha/2$ del estadístico de contraste. La región de rechazo de un test unilateral izquierda será la cola izquierda con probabilidad α del estadístico de contraste. La región de rechazo de un test unilateral derecha será la cola derecha con probabilidad α del estadístico de contraste.

☞ Considérese un estadístico de contraste, $W(X_1,\ldots,X_n)$, un nivel de significación α, y una m.a.s. observada, $x_1,\ldots,x_n$. Llámese w_{obs} a $W(x_1,\ldots,x_n)$. Para un test bilateral, se rechazará la hipótesis nula si $w_{\text{obs}} > w_{\alpha/2}$ o si $w_{\text{obs}} < w_{1-\alpha/2}$, donde $w_{\alpha/2}$ y $w_{1-\alpha/2}$ son tales que:

$$P(W(X_1,\ldots,X_n) > w_{\alpha/2}) = \frac{\alpha}{2} \text{ y } P(W(X_1,\ldots,X_n) < w_{1-\alpha/2}) = \frac{\alpha}{2} .$$

Para un test unilateral izquierda, se rechazará la hipótesis nula si $w_{\text{obs}} < w_{1-\alpha}$, donde $w_{1-\alpha}$ es tal que:

$$P(W(X_1,\ldots,X_n) < w_{1-\alpha}) = \alpha .$$

Para un test unilateral derecha, se rechazará la hipótesis nula si $w_{\text{obs}} > w_{\alpha}$, donde w_{α} es tal que

$$P(W(X_1,\ldots,X_n) > w_{\alpha}) = \alpha .$$

En esencia, los estadísticos de contraste que utilizaremos para hacer test de hipótesis a partir de una muestra serán los estadísticos pivote vistos en la construcción de intervalos de confianza. La única diferencia es que se cambiará el parámetro desconocido que aparece en aquellos, θ, por el valor que delimite las hipótesis, θ_0 (para controlar el error de tipo I, se asume cierta H_0). De esta manera, el estadístico de contraste puede, efectivamente, calcularse para cada m.a.s. observada.

Test de hipótesis con una muestra.

Caso $\mathcal{B}(p)$.

El estadístico de contraste que utilizaremos para hacer un test de hipótesis sobre el parámetro $\theta = p$ de una Bernoulli es:

$$Z(X_1,\ldots,X_n) = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} ,$$

donde $p_0 = \theta_0$ es el valor concreto que aparece en el correspondiente test de hipótesis. Este estadístico seguirá aproximadamente una distribución $\mathcal{N}(0,1)$, bajo el supuesto de que H_0 sea cierta, siempre que $n \geq 30$.

Si se escoge un nivel de significación, α, y se plantea un test bilateral ($H_0 : p = p_0$), entonces se rechazaría la hipótesis nula para toda m.a.s. observada para la cual se tenga que $z_{\text{obs}} > z_{\alpha/2}$ ó $z_{\text{obs}} < -z_{\alpha/2}$, donde $z_{\text{obs}} = Z(x_1,\ldots,x_n)$. Para un test unilateral izquierda, se rechazaría la hipótesis nula ($H_0 : p \geq p_0$) si $z_{\text{obs}} < -z_{\alpha}$. Para un test unilateral derecha, se rechazaría la hipótesis nula ($H_0 : p \leq p_0$) si $z_{\text{obs}} > z_{\alpha}$.

Caso $\mathcal{N}(\mu,\sigma^2)$, parámetro de interés: μ (varianza conocida, $\sigma^2 = \sigma_0^2$).

El estadístico de contraste que utilizaremos para hacer un test de hipótesis sobre el parámetro $\theta = \mu$ de una población normal de varianza conocida, $\mathcal{N}(\mu,\sigma_0^2)$, es:

$$Z(X_1,\ldots,X_n) = \frac{\overline{X_n} - \mu_0}{\sigma_0/\sqrt{n}} ,$$

donde μ_0 es el valor concreto que aparece en el correspondiente test de hipótesis. Este estadístico seguirá $\mathcal{N}(0,1)$, bajo la asunción de que H_0 sea cierta.

Si se escoge un nivel de significación, α, y se plantea un test bilateral ($H_0 : \mu = \mu_0$), entonces se rechazaría la hipótesis nula para toda m.a.s. observada para la cual se tenga que $z_{\text{obs}} > z_{\alpha/2}$ ó $z_{\text{obs}} < -z_{\alpha/2}$, donde $z_{\text{obs}} = Z(x_1, \ldots, x_n)$. Para un test unilateral izquierda, se rechazaría la hipótesis nula ($H_0 : \mu \geq \mu_0$) si $z_{\text{obs}} < -z_\alpha$. Para un test unilateral derecha, se rechazaría la hipótesis nula ($H_0 : \mu \leq \mu_0$) si $z_{\text{obs}} > z_\alpha$[21]

Caso $\mathcal{N}(\mu, \sigma^2)$, parámetro de interés: σ^2.

Para hacer test de hipótesis sobre la varianza, $\theta = \sigma^2$ de una distribución normal, el estadístico de contraste que utilizaremos es:

$$T(X_1, \ldots, X_n) \equiv \frac{(n-1)S_n^2}{\sigma_0^2} \ ,$$

donde σ_0^2 es el valor concreto que aparece en el correspondiente test de hipótesis. Bajo la asunción de que H_0 sea cierta, este estadístico seguirá una distribución χ^2_{n-1}.

Si se escoge un nivel de significación, α, y se plantea un test bilateral ($H_0 : \sigma^2 = \sigma_0^2$), entonces se rechazaría la hipótesis nula para toda m.a.s. observada para la cual se tenga que $\chi^2_{\text{obs}} > \chi^2_{n-1,\alpha/2}$ ó $\chi^2_{\text{obs}} < \chi^2_{n-1,1-\alpha/2}$, donde $\chi^2_{\text{obs}} = T(x_1, \ldots, x_n)$. Para un test unilateral izquierda, se rechazaría la hipótesis nula ($H_0 : \sigma^2 \geq \sigma_0^2$) si $\chi^2_{\text{obs}} < \chi^2_{n-1,1-\alpha}$. Para un test unilateral derecha, se rechazaría la hipótesis nula ($H_0 : \sigma^2 \leq \sigma_0^2$) si $\chi^2_{\text{obs}} > \chi^2_{n-1,\alpha}$[22].

Caso $\mathcal{N}(\mu, \sigma^2)$, parámetro de interés: μ (varianza desconocida).

Para hacer contrastes de hipótesis sobre la media de una población normal de la que se desconoce la varianza, emplearemos el estadístico de contraste siguiente:

$$\frac{\overline{X_n} - \mu_0}{S_n/\sqrt{n}} \equiv t_{n-1} \ ,$$

donde μ_0 es el valor concreto que aparece en el correspondiente test de hipótesis. Asumiendo que H_0 sea cierta, este estadístico seguirá una distribución t_{n-1}.

Si se escoge un nivel de significación, α, y se plantea un test bilateral ($H_0 : \mu = \mu_0$), entonces se rechazaría la hipótesis nula para toda m.a.s. observada para la cual se tenga que $t_{\text{obs}} > t_{n-1,\alpha/2}$ ó $t_{\text{obs}} < -t_{n-1,\alpha/2}$, donde $t_{\text{obs}} = T(x_1, \ldots, x_n)$. Para un test unilateral izquierda, se rechazaría la hipótesis nula ($H_0 : \mu \geq \mu_0$) si $t_{\text{obs}} < -t_{n-1,\alpha}$. Para un test unilateral derecha, se rechazaría la hipótesis nula ($H_0 : \mu \leq \mu_0$) si $t_{\text{obs}} > t_{n-1,\alpha}$[23].

[21] Recuerda que llamamos z_α al valor que deja a su derecha una probabilidad α en una $\mathcal{N}(0,1)$. Al ser la normal simétrica respecto al cero, $-z_\alpha$ deja a su derecha una probabilidad $1-\alpha$, es decir, deja a su izquierda una probabilidad α.

[22] Recuerda que definimos $\chi^2_{n-1,\alpha}$ como el valor que deja a la derecha de una χ^2_{n-1} una probabilidad α. El valor $\chi^2_{n-1,1-\alpha}$, por su parte, es el que deja una probabilidad $1-\alpha$ a su derecha, es decir, deja una probabilidad α a su izquierda.

[23] Recuerda que llamamos $t_{n-1,\alpha}$ al valor que deja a su derecha una probabilidad α en una distribución t_{n-1}. Al ser la función de densidad de la t de Student simétrica respecto al cero, entonces $-t_{n-1,\alpha}$ deja a su derecha una probabilidad $1-\alpha$, es decir, deja a su izquierda una probabilidad α.

Ejercicio: supón que tienes una población normal de la que se recogió una muestra observada de tamaño 17, con desviación típica muestral igual a 1.3 y media muestral, 15. Lleva a cabo un test de hipótesis con la muestra anterior para cada cuestión que sigue: (a) A un nivel de significación 0.001, ¿se rechazaría la hipótesis de que la varianza es mayor que 1.8? ¿Y a un nivel de significación 0.025?; (b) ¿Se podría aceptar la hipótesis de que la media es mayor que 13 al nivel de significación 0.075? ¿Hay evidencias suficientes para rechazar la hipótesis de que la media es mayor que 20 al mismo nivel de significación, 0.075?; (c) Si supiéramos que la varianza poblacional es 4, ¿a qué niveles de significación se rechazaría la hipótesis de que la media poblacional es menor que 14?

Resolución. Hay que suponer que la muestra recogida es una m.a.s. observada de una población $X \equiv \mathcal{N}(\mu, \sigma^2)$. La información que nos dan de la muestra observada es: $n = 17$, $s_n = 1.3$ y $\overline{x_n} = 15$

(a) Para $\alpha = 0.001$, un test de hipótesis que, potencialmente, pueda permitir rechazar la afirmación "la varianza (poblacional) es mayor que 1.8" sería el siguiente:

$$\begin{cases} H_0: & \sigma^2 \geq 1.8 \\ H_1: & \sigma^2 < 1.8 \end{cases}$$

Evaluando la muestra observada en el correspondiente estadístico de contraste, se tiene que $\chi^2_{\text{obs}} = (n-1)s_n^2/\sigma_0^2 = 16 \cdot 1.3^2/1.8 = 15.022$, donde $\sigma_0^2 = 1.8$ para este test de hipótesis. Para ver si se rechaza o no se rechaza la hipótesis nula, hay que comprobar si χ^2_{obs} ha caído o no dentro de la región de rechazo que marca el nivel de significación junto al estadístico de contraste.

La región de rechazo de este test, con $\alpha = 0.001$, está delimitada por el valor $\chi^2_{n-1,1-\alpha} = \chi^2_{16,0.999} = 3.942$ (recurriendo a la tabla de la χ^2, es el valor que deja a su derecha una probabilidad 0.999 en una distribución χ^2_{16}). Simplemente hay que comparar $\chi^2_{\text{obs}} = 15.022$ con $\chi^2_{16,0.999} = 3.942$. Al tratarse de un test unilateral izquierda, se rechazará la hipótesis nula si $\chi^2_{\text{obs}} < \chi^2_{n-1,1-\alpha}$. Como no ocurre eso, sino que $15.022 > 3.942$, el resultado del test planteado es no rechazar H_0. Es decir, en la muestra observada no hay suficientes evidencias, al nivel de significación $\alpha = 0.001$, para rechazar la hipótesis de que la varianza poblacional sea mayor que 1.8.

Para responder al mismo test con un nivel de significación $\alpha = 0.025$, lo único que cambia respecto a lo que acabamos de hacer es la región de rechazo: estaría delimitada por $\chi^2_{n-1,1-\alpha} = \chi^2_{16,0.975} = 6.908$. Cualquier χ^2_{obs} más pequeño que $\chi^2_{16,0.975}$ nos llevaría a rechazar la hipótesis nula a un nivel de significación $\alpha = 0.025$. De nuevo, como $\chi^2_{\text{obs}} = 15.022$ y este valor es más grande que 6.908, entonces el resultado del test de hipótesis es no rechazar H_0.

(b) Para poder, potencialmente, aceptar la hipótesis "la media poblacional es mayor que 13" al nivel de significación $\alpha = 0.075$, debemos establecer un test en el que esa hipótesis sea la alternativa, H_1 (la única que se puede aceptar, por rechazar H_0). Es decir, nos planteamos el test:

$$\begin{cases} H_0: & \mu \leq 13 \\ H_1: & \mu > 13 \end{cases}$$

Evaluando el correspondiente estadístico de contraste en la muestra observada, se tiene que $t_{\text{obs}} = (\overline{x_n} - \mu_0)/(s_n/\sqrt{n}) = (15-13)/(1.3/\sqrt{17}) = 6.34$. Como se trata de un test

unilateral derecha, se rechazará la hipótesis nula a nivel de significación $\alpha = 0.075$ si $t_{\text{obs}} > t_{n-1,\alpha}$, donde $t_{n-1,\alpha} = t_{16,0.075} = 1.512$ (mirando la tabla de la t de Student). Dado que $6.34 > 1.512$, entonces el resultado del test es rechazar la hipótesis nula (rechazar que $\mu < 13$) al nivel de significación $\alpha = 0.075$. Al rechazar la hipótesis nula, se acepta la alternativa. Respondiendo a la pregunta inicial: sí, con la muestra dada se puede aceptar la hipótesis (al nivel de significación $\alpha = 0.075$) de que la media poblacional es mayor que 13.

La segunda cuestión nos pregunta sobre si se puede rechazar la hipótesis "la media es mayor que 20", al mismo nivel de significación. En este caso, el test a plantearse debe llevar esa hipótesis como hipótesis nula, ya que es la que se busca potencialmente rechazar.

$$\begin{cases} H_0: & \mu \geq 20 \\ H_1: & \mu < 20 \end{cases}$$

El estadístico de contraste y su evaluación en la muestra son iguales, cambiando μ_0 por 20. Se obtiene que $z_{\text{obs}} = -15.86$. En esta ocasión, se trata de un test unilateral izquierda, de modo que se rechazará la hipótesis nula siempre que $t_{\text{obs}} < -t_{n-1,\alpha}$. Dado que $-t_{n-1,\alpha} = -t_{16,0.075} = -1.512$, se tiene que $-15.86 < -1.512$ y, por lo tanto, al nivel de significación $\alpha = 0.02$, hay evidencias suficientes en la muestra para rechazar que la media poblacional sea mayor que 20.

(c) Si la varianza poblacional fuese $4 = \sigma_0^2$, el estadístico de constraste que hay que emplear es distinto al anterior. Su evaluación en la muestra observada es: $z_{\text{obs}} = (\overline{x_n} - \mu_0)/(\sigma_0^2/\sqrt{n}) = (15-14)/(2/\sqrt{17}) = 2.06$, donde $\mu_0 = 14$. El test a plantear para decidir si rechazar la hipótesis "la media poblacional es menor que 14" es:

$$\begin{cases} H_0: & \mu \leq 14 \\ H_1: & \mu > 14 \end{cases}$$

ya que la hipótesis rechazable debe situarse como hipótesis nula.

Al tratarse de un test unilateral derecha, para cada nivel de significación preestablecido, α, se rechazará H_0 si $z_{\text{obs}} > z_\alpha$. Es decir, se rechazará H_0 para los niveles de significación para los que se verifique $2.06 > z_\alpha$[24]. Buscando en la tabla de la normal, se encuentra que el valor 2.06 deja a la derecha una probabilidad $1 - 0.9803 = 0.0197$. Por lo tanto, para todo nivel de significación $\alpha > 0.0197$, se cumplirá que $2.06 > z_\alpha$, lo cual conducirá a rechazar la hipótesis nula. Por ejemplo, para niveles de significación $\alpha = 0.01, 0.009, 0.005...$, no se rechazaría la hipótesis nula a ese nivel de significación. En cambio, para niveles de significación como $\alpha = 0.02, 0.05, 0.06...$, se rechazaría la hipótesis nula. Respondiendo a la pregunta: para cualquier nivel de significación, α, más grande que 0.0197, se rechazaría la hipótesis de que la media poblacional es menor que 14.

El último apartado del ejercicio anterior permite intuir el último concepto que introduciremos: el p-valor. El p-valor de un test de hipótesis es un valor numérico, entre 0 y 1, concreto para cada m.a.s. observada.

Volviendo al planteamiento genérico de los contrastes de hipótesis, supón que tenemos una población X y estamos interesadas en un parámetro, θ, de la correspondiente distribución

[24] Recuerda que la región de rechazo cambia con el nivel de significación.

de probabilidad. Considera un estadístico de contraste, $W(X_1, \ldots, X_n)$, que depende de la m.a.s., $X_1, \ldots, X_n$. Para cada posible m.a.s. observada, $x_1, \ldots, x_n$, su p-valor da cuenta de la fuerza con que se rechazaría o no se rechazaría la hipótesis nula para esa m.a.s. observada. Llamando $w_{\text{obs}} = W(x_1, \ldots, x_n)$:

- En un test unilateral izquierda para θ ($H_0 : \theta \geq \theta_0$), el p-valor de una m.a.s. observada se define como p-valor$= P(W(X_1, \ldots, X_n) < w_{\text{obs}})$.
- En un test unilateral derecha para θ ($H_0 : \theta \leq \theta_0$), el p-valor de una m.a.s. observada se define como p-valor$= P(W(X_1, \ldots, X_n) > w_{\text{obs}})$.
- En un test bilateral para θ ($H_0 : \theta = \theta_0$), el p-valor se define como

  $$p\text{-valor} = 2 \cdot P(W(X_1, \ldots, X_n) > w_{\text{obs}}) \quad \text{ó} \quad p\text{-valor} = 2 \cdot P(W(X_1, \ldots, X_n) < w_{\text{obs}}) \ ,$$

  el que sea menor que uno.

El p-valor permite reducir la regla de decisión en un test de hipótesis (rechazar o no rechazar H_0) para un nivel de significación dado, α, a comprobar si el p-valor es mayor o menor que el nivel de significación.

- Si el **p-valor** de una m.a.s. observada es mayor que el nivel de significación, entonces no se rechaza la hipótesis nula. Si, en cambio, el p-valor de una m.a.s. observada es menor que el nivel de significación, entonces se rechaza la hipótesis nula.

Ejercicio: Supón que se toma una m.a.s. observada proveniente de una población normal, de tamaño $n = 28$; media muestral, $\overline{x_n} = 3$ y desviación típica muestral, $s_n = 1.71$. Decide, a partir del p-valor, si se rechazaría o no la hipótesis de que la media poblacional es igual a 2.2, con un nivel de significación $\alpha = 0.01$.

Resolución. Se está en el caso de una población normal con varianza desconocida en el que estamos interesadas en la media poblacional. El estadístico de contraste que utilizamos en este caso sigue una distribución t de Student con $n - 1 = 27$ grados de libertad y el resultado de evaluarlo en la m.a.s. observada es: $t_{\text{obs}} = (\overline{x_n} - \mu_0)/(s_n/\sqrt{n}) = 2.475$. Como se trata de un test bilateral, su p-valor es[25]:

$$p\text{-valor} = 2 \cdot P(t_{n-1} > 2.475) = 2 \cdot 0.01 = 0.02 \ .$$

Por lo tanto, al nivel de significación $\alpha = 0.01$, como el p-valor es mayor que el nivel de significación, con la m.a.s. observada no se puede rechazar la hipótesis de que la media poblacional sea igual a 2.2.

Ejercicio: responde a los apartados (a), (b) y (c) del penúltimo ejercicio en términos del p-valor.

Contrastes de normalidad.

Un contraste o test de normalidad es un test de hipótesis no paramétrico en el que la hipótesis nula es "la población sigue una distribución normal". Es decir:

$$\begin{cases} H_0: & X \text{ sigue una distribución normal} \\ H_1: & X \text{ no sigue una distribución normal} \end{cases}$$

[25] Si calculásemos $2 \cdot P(t_{n-1} < 2.475)$, saldría mayor que uno.

Existen diferentes tipos de test que utilizan distintos estadísticos de contraste, uno de los más habituales es el test Shapiro-Wilk. Con R Commander podremos llevar a cabo test de normalidad para una m.a.s. observada a partir del p-valor. Para ello, habrá que decidir, a partir del p-valor que proporciona el software, si hay evidencias suficientes para rechazar o no rechazar la hipótesis de que la población sigue una distribución normal, comparándolo como veníamos haciendo con un nivel de significación dado o con los niveles de significación habituales.

Test de hipótesis con dos muestras independientes.

En este apartado consideraremos dos poblaciones, X e Y, cada una de las cuales tendrá asociada una distribución de probabilidad. Nos planteamos como objetivo establecer contrastes de hipótesis sobre la relación entre dos parámetros (cada uno de una población). Para ello, tomaremos dos m.a.s. (cada una de una población) y asumiremos que las dos muestras aleatorias son independientes entre sí:

$$\begin{array}{l} \text{m.a.s. de } X \rightsquigarrow X_1, \ldots, X_{n_X} \ , \ \text{ de tamaño } n_X \ , \\ \text{m.a.s. de } Y \rightsquigarrow Y_1, \ldots, Y_{n_Y} \ , \ \text{ de tamaño } n_Y \ . \end{array}$$

Trabajaremos con casos en los que las poblaciones sean distribuciones normales ($X \equiv \mathcal{N}(\mu_X, \sigma_X^2)$ e $Y \equiv \mathcal{N}(\mu_Y, \sigma_Y^2)$, cada una con su media y su varianza) y en los que las poblaciones sean distribuciones de Bernoulli ($X \equiv \mathcal{B}(p_X)$ e $Y \equiv \mathcal{B}(p_Y)$, cada una con su proporción poblacional). Los test de hipótesis que formularemos será sobre la diferencia de las medias poblacionales ($\mu_X - \mu_Y$ ó $p_X - p_Y$) y sobre la división de las varianzas poblacionales (σ_X^2/σ_Y^2).

Para estimar la diferencia entre las medias de una población normal, emplearemos distintos estadísticos, en función de si: (i) conocemos las varianzas de las dos poblaciones, (ii) desconocemos las varianzas poblacionales, pero estas son iguales o bien (iii) desconocemos las varianzas poblacionales. Por ello, a nivel operativo, tras "confirmar" con un test de normalidad que las poblaciones siguen una distribución normal, se debe realizar un test sobre las división de las varianzas poblacionales: en caso de que sean desconocidas, no se utiliza el mismo estadístico de contraste si se puede asumir que las varianzas poblacionales son iguales, que si hay evidencias experimentales suficientes para afirmar que las varianzas son distintas.

Caso normal, interés: σ_X^2/σ_Y^2.

Tenemos dos poblaciones, $X \equiv \mathcal{N}(\mu_X, \sigma_X^2)$ e $Y \equiv \mathcal{N}(\mu_Y, \sigma_Y^2)$. Queremos saber si las varianzas son iguales (o cuál de las varianzas poblacionales es mayor). Los test que nos plantearemos serán de la siguiente forma:

$$\left\{ \begin{array}{ll} H_0: & \sigma_X^2/\sigma_Y^2 = 1 \\ H_1: & \sigma_X^2/\sigma_Y^2 \neq 1 \end{array} \right. \qquad \left\{ \begin{array}{ll} H_0: & \sigma_X^2/\sigma_Y^2 \geq 1 \\ H_1: & \sigma_X^2/\sigma_Y^2 < 1 \end{array} \right. \qquad \left\{ \begin{array}{ll} H_0: & \sigma_X^2/\sigma_Y^2 \leq 1 \\ H_1: & \sigma_X^2/\sigma_Y^2 > 1 \end{array} \right.$$

Respectivamente, test bilateral, unilateral izquierda y unilateral derecha para σ_X^2/σ_Y^2.

El estadístico de contraste que utilizaremos sigue una nueva distribución de probabilidad: la distribución F de Snedecor con $n_X - 1$ y $n_Y - 1$ grados de libertad (F_{n_X, n_Y}),

donde n_X y n_Y son los tamaños de la m.a.s. de cada población. El estadístico de contraste que emplearemos es:

$$F(X_1, \ldots, X_n) = \frac{S^2_{n_X}}{S^2_{n_Y}} \ . \tag{1}$$

Bajo la asunción de que H_0 es cierta, la división de las varianzas muestrales, $S^2_{n_X}/S^2_{n_Y}$ sigue una distribución F_{n_X-1,n_Y-1}. Hay que tener presente que la división inversa, $S^2_{n_Y}/S^2_{n_X}$, no sigue la misma distribución, sino F_{n_Y-1,n_X-1} (y, en general, no es igual poner un grado de libertad primero que poner otro).

Cabe destacar que la distribución F de Snedecor surge de dividir dos distribuciones χ^2 que están divididas a su vez por sus respectivos grados de libertad, es decir:

$$\frac{\chi^2_{n_X-1}/(n_X-1)}{\chi^2_{n_Y-1}/(n_Y-1)} = F_{n_X-1,n_Y-1} \ .$$

Recordando de apartados anteriores que $(n_X-1)S_{n_X}/\sigma^2_X \equiv \chi^2_{n_X-1}$ y que $(n_Y-1)S_{n_Y}/\sigma^2_Y \equiv \chi^2_{n_Y-1}$, es claro que:

$$\frac{S^2_{n_X}/\sigma^2_X}{S^2_{n_Y}/\sigma^2_Y} \equiv F_{n_X-1,n_Y-1} \ . \tag{2}$$

Bajo la hipótesis nula y en el caso menos favorable a H_0, se tiene que $\sigma^2_X = \sigma^2_Y$ ($\sigma^2_X/\sigma^2_Y = 1$), lo cual justificaría utilizar el estadístico de contraste presentado anteriormente.

Para llevar a cabo el test, se podría proceder del mismo modo que se venía haciendo: se escoge un nivel de significación y, utilizando la distribución F_{n_X-1,n_Y-1}, se construye la región de rechazo. Nos centraremos en calcular el p-valor, procediendo como en el siguiente ejemplo.

Ejercicio: se tienen dos poblaciones normales, de cada una de las cuales se extrajo una m.a.s. observada, de tamaños 21 y 16, respectivamente. Si para la primera muestra se observó una media y varianza muestrales de 3.6 y 1.36, respectivamente, y para la segunda muestra se observó una media y varianza muestral de 4.2 y 3.75, ¿se rechazaría la hipótesis de que las varianzas poblacionales son iguales al nivel de significación 0.01? ¿Y al nivel de significación 0.06?

Resolución. Llamemos $X \equiv \mathcal{N}(\mu_X, \sigma^2_X)$ a la población con mayor varianza muestral e $Y \equiv \mathcal{N}(\mu_Y, \sigma^2_Y)$ a la población con menor varianza muestral (ello simplifica los razonamientos a hacer). De este modo, $n_X = 16$ y $n_Y = 21$, luego $\overline{x_{n_X}} = 4.2$, $s^2_{n_X} = 3.75$, $\overline{y_{n_Y}} = 3.6$, $s^2_{n_Y} = 1.36$. Nos estamos planteando un test de hipótesis bilateral en el que la hipótesis nula es $H_0 : \sigma^2_X/\sigma^2_Y = 1$. El resultado de evaluar el estadístico de contraste (una distribución $F_{20,15}$) en la muestra dada es $F_{\text{obs}} = s^2_{n_X}/s^2_{n_Y} = 2.757$. El p-valor sería entonces, aproximadamente:

$$p\text{-valor} = 2 \cdot P(F_{20,15} \geq F_{\text{obs}}) = 2 \cdot 0.025 = 0.05 \ ,$$

mirando las tablas.

Por lo tanto, respondiendo a la pregunta: al nivel de significación 0.01, el p-valor calculado es mayor que el nivel de significación. Por lo tanto, no se rechazaría la hipótesis de

que las varianzas sean iguales (a ese nivel de significación). En cambio, al nivel de significación 0.06, el p-valor resulta ser menor que el nivel de significación y, en consecuencia, se rechazaría la hipótesis de que las varianzas poblacionales sean iguales (a ese nivel de significación).

Caso normal, interés: $\mu_X - \mu_Y$ (varianzas conocidas, σ_X^2 y σ_Y^2).

Tenemos dos poblaciones, $X \equiv \mathcal{N}(\mu_X, \sigma_X^2)$ e $Y \equiv \mathcal{N}(\mu_Y, \sigma_Y^2)$. Queremos hacer test de hipótesis sobre la diferencia entre las medias poblacionales, y conocemos las varianzas de cada población, σ_X^2 y σ_Y^2. Los test que nos plantearemos serán de la siguiente forma:

$$\begin{cases} H_0: & \mu_X - \mu_Y = \mu_0 \\ H_1: & \mu_X - \mu_Y \neq \mu_0 \end{cases} \qquad \begin{cases} H_0: & \mu_X - \mu_Y \geq \mu_0 \\ H_1: & \mu_X - \mu_Y < \mu_0 \end{cases} \qquad \begin{cases} H_0: & \mu_X - \mu_Y \leq \mu_0 \\ H_1: & \mu_X - \mu_Y > \mu_0 \end{cases}$$

Respectivamente, test bilateral, unilateral izquierda y unilateral derecha para $\mu_X - \mu_Y$.

El estadístico de contraste que utilizaremos es

$$\frac{\overline{X_{n_X}} - \overline{Y_{n_Y}} - \mu_0}{\sqrt{\dfrac{\sigma_X^2}{n_X} + \dfrac{\sigma_Y^2}{n_Y}}} \equiv \mathcal{N}(0,1) \ ,$$

el cual sigue una distribución $\mathcal{N}(0,1)$ en caso de que la hipótesis nula sea cierta. Por lo tanto, según el tipo de test, la región de rechazo se construiría (para un nivel de significación dado) de la misma forma que veníamos haciendo para una muestra.

Caso normal, interés: $\mu_X - \mu_Y$ (varianzas desconocidas, pero iguales $\sigma_X^2 = \sigma_Y^2$).

Tenemos dos poblaciones, $X \equiv \mathcal{N}(\mu_X, \sigma_X^2)$ e $Y \equiv \mathcal{N}(\mu_Y, \sigma_Y^2)$. Queremos hacer test de hipótesis sobre la diferencia entre las medias poblacionales. No conocemos las varianzas de cada población, ni σ_X^2 ni σ_Y^2, pero se puede suponer que son iguales: $\sigma_X^2 = \sigma_Y^2$[26]. Los test que nos plantearemos serán de la misma forma que en el apartado anterior.

El estadístico de contraste que utilizaremos es:

$$\frac{\overline{X_{n_X}} - \overline{Y_{n_Y}} - \mu_0}{S_p\sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}} \equiv t_{n_X+n_Y-2} \ , \quad \text{donde} \quad S_p = \sqrt{\frac{(n_X-1)S_{n_X}^2 + (n_Y-1)S_{n_Y}^2}{n_X+n_Y-2}} \ .$$

Este estadístico de contraste sigue una distribución $t_{n_X+n_Y-2}$ en caso de que la hipótesis nula sea cierta. Por lo tanto, según el tipo de test, la región de rechazo se construiría (para un nivel de significación dado) de la misma forma que veníamos haciendo para una muestra.

Caso normal, interés: $\mu_X - \mu_Y$ (varianzas desconocidas).

Tenemos dos poblaciones, $X \equiv \mathcal{N}(\mu_X, \sigma_X^2)$ e $Y \equiv \mathcal{N}(\mu_Y, \sigma_Y^2)$. Queremos hacer test

[26] Por ejemplo, si se ha realizado un test bilateral para la división de las varianzas poblacionales y ha salido un p-valor "muy" alto.

de hipótesis sobre la diferencia entre las medias poblacionales. No conocemos las varianzas de cada población, ni podemos suponer que son iguales. Los test que nos plantearemos serán de la misma forma que en el apartado anterior.

El estadístico de contraste que utilizaremos es:

$$\frac{\overline{X_{n_X}} - \overline{Y_{n_Y}} - \mu_0}{\sqrt{\frac{S^2_{n_X}}{n_X} + \frac{S^2_{n_Y}}{n_Y}}} \equiv t_f \ , \quad \text{donde} \quad f \simeq \frac{\left(\frac{S^2_{n_X}}{n_X} + \frac{S^2_{n_Y}}{n_Y}\right)^2}{\frac{\left(\frac{S^2_{n_X}}{n_X}\right)^2}{n_X - 1} + \frac{\left(\frac{S^2_{n_Y}}{n_Y}\right)^2}{n_Y - 1}} \ .$$

Este estadístico de contraste sigue aproximadamente (aproximación de Welch) una distribución t_f, en caso de que la hipótesis nula sea cierta, donde f se toma como el número entero más cercano al cálculo anterior. Por lo tanto, según el tipo de test, la región de rechazo se construiría (para un nivel de significación dado) de la misma forma que veníamos haciendo para una muestra.

Caso $\mathcal{B}(p)$.

Tenemos dos poblaciones, $X \equiv \mathcal{B}(p_X)$ e $Y \equiv \mathcal{B}(p)$. Queremos hacer test de hipótesis sobre la diferencia entre las medias poblacionales, y conocemos las varianzas de cada población, σ^2_X y σ^2_Y. Los test que nos plantearemos serán de la siguiente forma:

$$\begin{cases} H_0: & p_X - p_Y = p_0 \\ H_1: & p_X - p_Y \neq p_0 \end{cases} \qquad \begin{cases} H_0: & p_X - p_Y \geq p_0 \\ H_1: & p_X - p_Y < p_0 \end{cases} \qquad \begin{cases} H_0: & p_X - p_Y \leq p_0 \\ H_1: & p_X - p_Y > p_0 \end{cases}$$

Respectivamente, test bilateral, unilateral izquierda y unilateral derecha para $p_X - p_Y$.

El estadístico de contraste que utilizaremos es

$$\frac{\hat{p}_X - \hat{p}_Y - p_0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \equiv \mathcal{N}(0,1) \ , \quad \text{donde} \quad \hat{p} = \frac{n_X \hat{p}_X + n_Y \hat{p}_Y}{n_X + n_Y}$$

y $\hat{p}_X$ y $\hat{p}_Y$ son las proporciones muestrales de cada m.a.s. De nuevo, según el tipo de test, la región de rechazo se construiría (para un nivel de significación dado) de la misma forma que veníamos haciendo para una muestra. Distinto estadístico de contraste, pero mismo procedimiento.