BLOQUE 7:

ESTADÍSTICA Y PROBABILIDAD

- Combinatoria
- Probabilidad
- Medidas de centralización y dispersión
- E. descriptiva bidimensional
- Regresión lineal. Estimación

7. ESTADÍSTICA Y PROBABILIDAD

7.1 COMBINATORIA

Las permutaciones, variaciones y combinaciones son sencillas técnicas de recuento que permiten abordar y resolver problemas que serían demasiado complicados usando los métodos habituales. Su estudio constituye el llamado **análisis combinatorio**.

- **Factorial de un número**

Si n es un número natural, se llama factorial de n y se escribe n! al producto:

$$n! = n(n-1)(n-2)\cdot \ldots \cdot 2 \cdot 1$$

1!=1
Por convenio 0!=1

- **Permutaciones ordinarias**

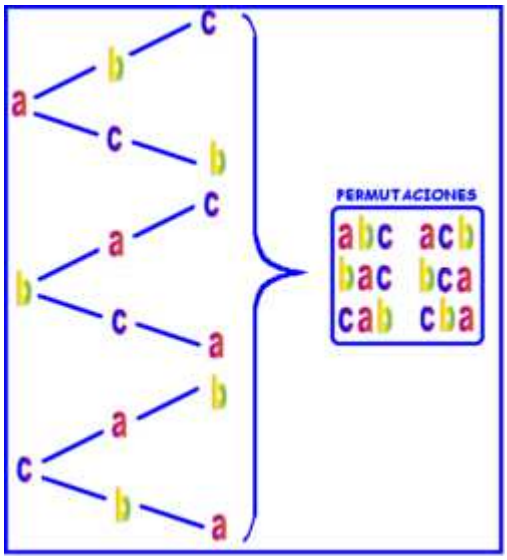
Se llama permutación ordinaria de m elementos, a cada una de las posibles ordenaciones de esos m elementos.

Dados m elementos hay m opciones para colocar el primero. Una vez colocado, quedan m-1 opciones para el segundo, para el tercero habrá m-2 y así sucesivamente. Por lo tanto:

$$P_m = m(m-1)(m-2)\cdot \ldots \cdot 2 \cdot 1 = m!$$

Dos permutaciones sólo son iguales si todos sus elementos están exactamente en el mismo orden.

Ejemplo:



$$P_3 = 3! = 3 \cdot 2 \cdot 1 = 6$$

- **Variaciones ordinarias**

Variaciones ordinarias, o simplemente variaciones, de m elementos tomados de n en n $(n \leq m)$ son las posibles agrupaciones que se pueden formar con n elementos distintos tomados entre los m dados, de modo que dos agrupaciones son distintas si difieren en algún elemento o en el orden de colocación de los mismos.

Si se dispone de los elementos $\{a,b,c,d,e\}$ abc y acb son dos variaciones distintas de 3 elementos tomados de los 5 dados.

El número de variaciones ordinarias de m elementos tomados de n en n se escribe $V_{m,n}$.

Ejemplo:

Con los dígitos 1, 2, 3 y 4 las variaciones de uno en uno son: 1, 2, 3, 4, tenemos, por tanto, $V_{4,1} = 4$

Tomadas de dos en dos: 12 13 14
21 23 24
31 32 34
41 42 43

Por lo tanto: $V_{4,2} = 4 \cdot 3 = 12$

En general: $$V_{m,n} = m(m-1)(m-2)\cdot \ldots \cdot (m-n+1) = \frac{m!}{(m-n)!}$$

- **Variaciones con repetición**

Variaciones con repetición de m elementos tomados de n en n son las posibles agrupaciones que se pueden formar con n elementos tomados de entre los m dados, de modo que dos agrupaciones son distintas si difieren en algún elemento o en el orden de colocación de los mismos.

Las variaciones con repetición de m elementos tomados de n en n se escriben: $VR_{m,n}$.

Ejemplo:

Formaremos las variaciones con repetición de los dígitos 3, 4 y 5 tomados de dos en dos. Puesto que pueden repetirse, bastará con añadir todos y cada uno de ellos a los dígitos dados.

33 34 35 43 44 45 53 54 55

Si tenemos m elementos, hay m posibilidades para la primera posición, como se pueden repetir, seguimos teniendo m posibilidades para la segunda y así sucesivamente hasta la posición n.

$$VR_{m,n} = m^n$$

En este caso puede ser n > m.

- **Permutaciones con repetición**

Se llama permutación con repetición de m elementos a las muestras ordenadas que se pueden obtener con todos los elementos de un conjunto de los cuales hay n_1 repetidos, n_2 repetidos,…

Para determinar su número consideramos que si fuesen m elementos diferentes, habría m! permutaciones pero como un elemento está repetido n_1 veces, habrá que dividir entre n_1! Y así sucesivamente con cada uno de los elementos que se repitan.

Por tanto: $$PR_m^{n_1,n_2,\ldots} = \frac{m!}{n_1!\cdot n_2!\cdot\ldots}$$

- **Combinaciones ordinarias**

Combinación de m elementos tomados de n en n $(n \le m)$ es cada uno de los conjuntos que se pueden formar eligiendo n elementos distintos de entre los m dados.

Se escribe $C_{m,n}$ o C_m^n .
Dos combinaciones son iguales si tienen los mismos elementos aunque estén colocados en distinto orden.

Ejemplo:
Con los dígitos 1, 2, 3 y 4 formamos las combinaciones tomados de dos en dos:

12 13 14
23 24
34

Tomados de tres en tres: 123 124 132 134
Observamos que las 6 permutaciones de 3 elementos dados forman una única combinación.

Para determinar el número de combinaciones $C_4^3 = \frac{V_{4,3}}{P_3} = \frac{4\cdot 3\cdot 2}{3\cdot 2\cdot 1} = 4$

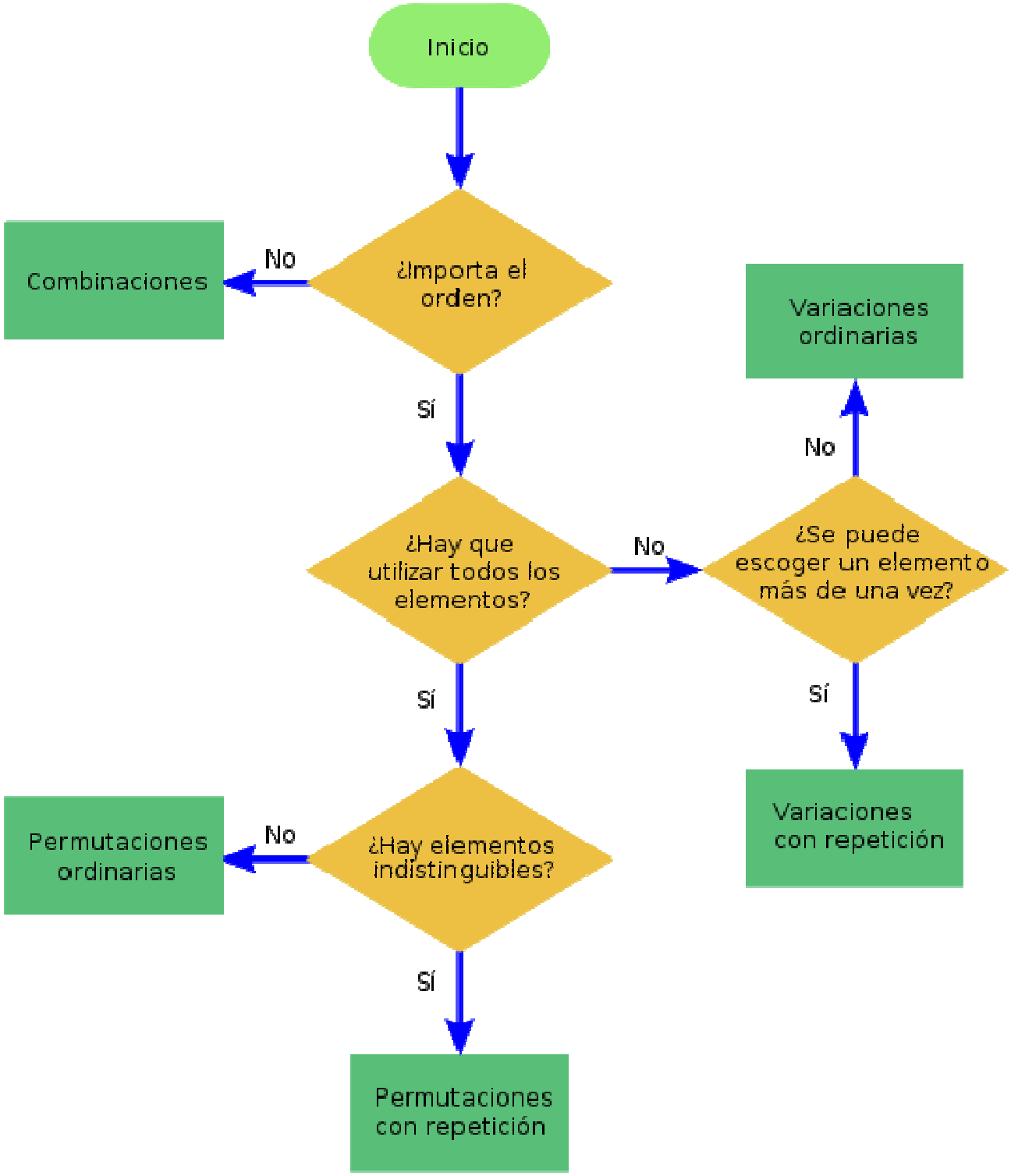
En general:

$$C_m^n = \frac{V_{m,n}}{P_m} = \frac{m!}{n!(m-n)!}$$

- **Combinaciones con repetición**

Llamaremos combinaciones con repetición de orden n definidas en un conjunto con m elementos, a los diferentes grupos de n elementos, iguales o distintos, que pueden formarse con los m elementos dados, de modo que dos grupos sean distintos cuando difieran, al menos, en un elemento.
En este caso n puede ser mayor que m.

Su número viene dado por: $CR_m^n = \binom{n+m-1}{n} = \frac{(n+m-1)!}{n!(m-1)!} = C_{n+m-1}^n$


Inicio
¿Importa el orden?
No
Combinaciones
Sí
¿Hay que utilizar todos los elementos?
No
¿Se puede escoger un elemento más de una vez?
No
Variaciones ordinarias
Sí
Variaciones con repetición
Sí
¿Hay elementos indistinguibles?
No
Permutaciones ordinarias
Sí
Permutaciones con repetición

7.2 NÚMEROS COMBINATORIOS

Se llama número combinatorio "m sobre n" y se escribe $\binom{m}{n}$ al número de combinaciones de m elementos tomados de n en n:

$$\binom{m}{n} = C_m^n = \frac{m!}{n!(m-n)!}$$

Por convenio $\binom{0}{0} = 1$

- **Propiedades**
  1. Un número combinatorio siempre es un número natural.
  2. $\binom{m}{0} = 1 \quad y \quad \binom{m}{m} = 1$
  3. $\binom{m}{m-n} = \binom{m}{n}$

     $$\binom{m}{m-n} = \frac{m!}{(m-n)![m-(m-n)]!} = \frac{m!}{(m-n)!n!} = \binom{m}{n}$$
  4. $\binom{m}{n} + \binom{m}{n+1} = \binom{m+1}{n+1}$

     $$\binom{m}{n} + \binom{m}{n+1} = \frac{m!}{(m-n)!n!} + \frac{m!}{(m-n-1)!(n+1)!} = \frac{m!(n+1)}{(m-n)!(n+1)!} + \frac{m!}{(m-n-1)!(n+1)!} =$$
     $$\frac{m!(n+1)}{(m-n)!(n+1)!} + \frac{m!(m-n)}{(m-n)!(n+1)!} = \frac{m!(n+1+m-n)}{(m-n)!(n+1)!} = \frac{(m+1)!}{(m-n)!(n+1)!} = \binom{m+1}{n+1}$$

7.3 BINOMIO DE NEWTON

El binomio de Newton es una fórmula que se utiliza para hacer el desarrollo de la potencia de un binomio elevado a una potencia cualquiera de exponente natural. Es decir, se trata de una fórmula para desarrollar la expresión:

$$(a+b)^n \qquad n \in \mathbb{N}$$

Las sucesivas potencias son:

$(a+b)^0 = 1$

$(a+b)^1 = a+b$

$(a+b)^2 = a^2 + 2ab + b^2$

$(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$

$(a+b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4.$

Los coeficientes de cada polinomio resultante siguen la secuencia:

$$\begin{array}{ccccccccc} & & & & 1 & & & & \\ & & & 1 & & 1 & & & \\ & & 1 & & 2 & & 1 & & \\ & 1 & & 3 & & 3 & & 1 & \\ 1 & & 4 & & 6 & & 4 & & 1 \end{array}$$

Además las potencias del primer sumando del binomio, a, comienzan por n y en cada sumando van disminuyendo de uno en uno hasta llegar a 0. Por el contrario, las potencias del segundo sumando del binomio, b, empiezan en 0 y van aumentando de uno en uno hasta llegar a n.

La estructura en triángulo anterior recibe el nombre de **Triángulo de Pascal o Triángulo de Tartaglia.**
El vértice superior es un 1 y en la segunda fila son siempre dos "unos". A partir de la tercera fila, el método de construcción es el siguiente:
- Primer número: 1.
- Números siguientes: la suma de los dos que se encuentran inmediatamente por encima.
- Último número: 1.

Teniendo en cuenta la definición de número combinatorio es fácil comprender que el Triángulo de Pascal o Triángulo de Tartaglia es, de hecho, el siguiente:

$$\begin{array}{ccccccccc} & & & & \binom{0}{0} & & & & \\ & & & \binom{1}{0} & & \binom{1}{1} & & & \\ & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & \\ & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & \\ \binom{4}{0} & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & \binom{4}{4} \end{array}$$

Así podemos generalizar el desarrollo de la potencia de un binomio:

$$(a+b)^n = \binom{n}{0}a^n + \binom{n}{1}a^{n-1}b + \binom{n}{2}a^{n-2}b^2 + \ldots + \binom{n}{n-2}a^2b^{n-2} + \binom{n}{n-1}ab^{n-1} + \binom{n}{n}b^n$$

que se conoce como **binomio de Newton.**

El término que ocupa el lugar k en el desarrollo $(a+b)^n$ tiene la forma:

$$T_k = \binom{n}{k-1}a^{n-(k-1)}b^{k-1}$$

EJERCICIOS Y PROBLEMAS

1. Un número telefónico consta de siete cifras enteras. Supongamos que la primera cifra debe ser un número entre 2 y 9, ambos inclusive. La segunda y la tercera cifra deben ser números entre 1 y 9, ambos inclusive. Cada una de las restantes cifras es un número entre 0 y 9, ambos inclusive. ¿Cuántos números de teléfono distintos pueden formarse con estas condiciones?

2. Una empresa produce cerraduras de combinación. Cada combinación consta de tres números enteros del 0 al 99, ambos inclusive. Por el proceso de construcción de las cerraduras cada número no puede aparecer más de una sola vez en la combinación de la cerradura. ¿Cuántas cerraduras diferentes pueden construirse?

3. ¿De cuántas formas pueden repartirse 3 entradas para un concierto de rock entre 6 amigos y amigas sin que ninguno pueda llevarse más de una?

4. Para formar un equipo de baloncesto hacen falta 5 jugadores y el entrenador dispone de 10.
   a) ¿Cuántos equipos distintos puede formar?
   b) Si elige a dos jugadores y los mantiene fijos, ¿cuántos equipos distintos podrá hacer con los ocho que le quedan?

5. Se van a celebrar elecciones en la ANPA y hay que elegir al presidente, al secretario y al tesorero. ¿De cuántas maneras se pueden elegir estos tres cargos, si se presentan ocho candidatos?

6. El lenguaje de un ordenador se traduce a secuencias de dígitos formados por ceros y unos. Un byte es una de estas secuencias y está formado por 8 dígitos. Por ejemplo: 00100011.
   ¿Cuántos bytes diferentes se pueden formar?

7. El profesor de Matemáticas nos ha propuesto diez problemas de los que tenemos que resolver cinco.
   a) ¿Cuántas formas hay de seleccionarlos?
   b) De los 10 problemas propuestos hay 2 de los que no tienes “ni idea”. ¿Se reducen mucho las posibilidades de selección?

8. Para acceder a una caja fuerte se tiene que introducir un número de 10 cifras. Se sabe que dicho número está formado por 5 doses, 3 cincos y 2 seises. ¿Cuántas claves diferentes se pueden formar?

9. Para decidir los ganadores de un concurso de poesía, un profesor debe elegir de jurado a 3 de sus 22 alumnos. ¿De cuántas formas diferentes puede realizar su elección?

10. En un restaurante de comida rápida se puede elegir entre hamburguesa con queso, sándwich vegetal, sándwich mixto, ensalada César y perrito caliente. ¿Cuántos pedidos diferentes puede hacer un grupo de 6 amigos?

11. Para un nuevo club deportivo se quiere hacer una bandera tricolor (tres colores distintos) que conste de tres franjas verticales. Si para crearla se dispone de 10 colores distintos, ¿cuántas banderas diferentes se pueden realizar?

12. Los números de los décimos de la Lotería Nacional tienen cinco cifras que se pueden repetir. Si por error un día se les olvida introducir en los cinco bombos el número 0, ¿cuántos posibles números habrá como candidatos al premio?

13. a) Con las letras de la palabra AMIGO, ¿cuántas ordenaciones distintas se pueden hacer?
    b) ¿Cuántas empiezan por A?
    c) ¿Cuántas empiezan por AMI?

14. Catorce montañeros deciden acampar, para lo cual disponen de tres tiendas de campaña de diferentes capacidades. En una pueden dormir ocho personas; en otra, cuatro, y en otra, dos. ¿De cuántas formas diferentes se pueden organizar para dormir en las tres tiendas?

15. A una reunión acudieron 20 personas. Para saludarse, dos personas se daban la mano. Si todo el mundo se saludó, ¿cuántos estrechamientos de manos hubo?

16. Si se lanzan simultáneamente 12 monedas de 1 €, ¿cuántos resultados distintos se pueden obtener?

17. Simplifica la expresión: $\dfrac{\left[\binom{29}{3}+\binom{29}{25}\right]\cdot 4!}{630}$

18. Desarrolla el binomio $(2x+y)^5$

19. Calcula el coeficiente de x^4 en el desarrollo de $(x-2)^6$

20. Calcula el término en el que aparece x^{26} en el desarrollo de $(x^2+\frac{y}{2})^{18}$

21. Resuelve las siguientes ecuaciones:

a) $\binom{1000}{750-2x}=\binom{1000}{3x}$ b) $V_{x,2}=190+\binom{x}{2}$ c) $\binom{x-3}{9}+\binom{x-3}{10}=\binom{18}{10}$

AUTOEVALUACIÓN 1

1. Ocho ciclistas van por el carril bici en fila. ¿De cuántas formas pueden ir ordenados? (40320 formas)
2. A una familia de 6 personas les ha tocado un viaje para dos personas. ¿De cuántas formas se pueden repartir el viaje? (15 formas)
3. En un concurso de radio participan 7 personas, de las cuales, 2 pueden conseguir los premios, que son: una enciclopedia y una radio. Sabiendo que una persona no puede conseguir los dos premios, ¿cuántas posibles distribuciones hay? (42)
4. Para hacer una transferencia bancaria, Marta tiene que teclear una clave de acceso que consta de 8 cifras con los dígitos 0 y 1. ¿Cuántas claves distintas puede formar? (256)
5. Calcula el número de palabras que pueden formarse con las cinco letras de la palabra SOSAS. (20)
6. Se lanzan simultáneamente 4 dados. ¿Cuántos resultados diferentes se pueden obtener? (126)
7. Calcula: $(x+1)^5-(x-1)^5$ $\left(10x^4+20x^2+2\right)$
8. Resuelve: $2\binom{n}{3}=V_{n,2}$ (n=5)
9. Averigua el término de grado 7 del desarrollo de $\left(\frac{x}{2}-\frac{x^2}{3}\right)^5$ $\left(\frac{5x^7}{36}\right)$
10. Una persona tiene 6 chaquetas y 10 pantalones. ¿De cuántas formas distintas puede combinar estas prendas? (60)

7.4 PROBABILIDAD

En la vida cotidiana, y en la práctica científica, existen experimentos en los que el resultado se puede predecir con antelación. Por ejemplo, cuando dejamos caer una piedra, podemos determinar cuánto tiempo tardará en llegar al suelo. Estos experimentos se llaman **deterministas** o causales, ya que la causa determina unívocamente el resultado.

Existen, por el contrario, experimentos en los que es imposible asegurar cuál será el resultado. Por ejemplo, al lanzar un dado o al extraer una carta de la baraja. Estos experimentos reciben el nombre de **aleatorios**.

Un experimento se llama aleatorio cuando se conocen todos sus posibles resultados, pero no puede predecirse cuál de ellos se producirá en una experiencia concreta.

Suceso elemental es cada uno de los posibles resultados de un experimento aleatorio, siempre que no se pueda descomponer en otros más simples.

Espacio muestral, E, de un experimento aleatorio es el conjunto de todos los sucesos elementales.

Suceso es cualquier subconjunto del espacio muestral E o, lo que es lo mismo, la unión de cualquier número de sucesos elementales.

En todo espacio muestral tenemos:

Suceso seguro, que es el propio espacio muestral E.

Suceso imposible, que es el que no se verifica nunca y se representa por $\varnothing$.

Ejemplo:

En el lanzamiento de dos monedas.
Un suceso seguro sería "obtener cualquier resultado"=$\{cc, cx, xc, xx\}$
Un suceso imposible sería "obtener 3 caras".
Un suceso elemental sería "obtener 2 caras" =$\{cc\}$.

Sea E el espacio muestral de un experimento y S uno de los sucesos. Si al realizar el experimento resulta un suceso elemental S, se dice que S ha sido un **éxito**, de lo contrario se dice que ha sido un **fracaso**.

Suceso **contrario o complementario** de un suceso A es el formado por los sucesos elementales que no pertenecen a A. Se representa por $\overline{A}$.

Ejemplo:
En el experimento de lanzar un dado, si A es el suceso "salir múltiplo de 3"=$\{3, 6\}$, el suceso contrario será "no salir múltiplo de 3" = $\{1, 2, 4, 5\}$.

Sucesos **incompatibles** son aquellos que no se pueden verificar simultáneamente, en caso contrario, son **compatibles**.

La **diferencia de dos sucesos A y B** es el suceso A-B formado por los sucesos elementales de A que no están en B.

Frecuencias

Si realizamos N veces un experimento aleatorio y el resultado x_i se ha presentado f_i veces, f_i es la frecuencia absoluta del resultado x_i y $h_i = \frac{f_i}{N}$ la frecuencia relativa.

$$0 < f_i < N \qquad 0 < h_i < 1$$

Idea intuitiva de la probabilidad

En el experimento aleatorio de lanzar una moneda, parece razonable que la frecuencia relativa de obtener cara debe ser $\frac{1}{2}$.

Tal y como se ve en la figura del margen, a medida que aumenta el número de tiradas, la frecuencia del suceso tiende a estabilizarse en torno al 0,5.

Este hecho característico de los experimentos aleatorios se llama **estabilidad de las frecuencias relativas** o **ley de los grandes números**, y nos va a permitir calcular la probabilidad de un suceso.

La probabilidad asigna a cada resultado la frecuencia relativa obtenida tras numerosos experimentos.

$$\mathrm{p(A)} = \lim_{N \to \infty} f_r\left(\mathrm{A}\right)$$

Este número variará entre 0 y 1, como las frecuencias relativas. Valdrá 0 para el suceso imposible, 1 para el seguro y valores comprendidos entre ambos para los demás.
Dos o más sucesos con la misma probabilidad se dice que son **equiprobables.**

Ley de Laplace

Este método de cálculo de probabilidad requiere que los sucesos sean equiprobables.

Dice así:
La probabilidad de un suceso A es igual al cociente entre el número de casos favorables al suceso A y el número de casos posibles:

$$p(\mathrm{A}) = \frac{\mathrm{n^o\ de\ casos\ favorables\ al\ suceso\ A}}{\mathrm{n^o\ de\ casos\ posibles}}$$

Ejemplo:

En una bolsa hay 5 bolas negras, 3 amarillas y 2 naranjas. ¿Cuál es la probabilidad de extraer una bola naranja?

$$p(\mathrm{A})=\frac{\text{casos favorables}}{\text{casos posibles}}=\frac{2}{10}=0,2$$

Unión e intersección de sucesos

Dados los sucesos A y B de un mismo experimento:

- Se llama **suceso intersección**, $\mathrm{A}\cap\mathrm{B}$, al formado por los sucesos elementales que pertenecen a A y B a la vez.
- Se llama **suceso unión**, $\mathrm{A}\cup\mathrm{B}$, al formado por los sucesos que pertenecen a A o a B o a ambos.

Diremos que dos sucesos A y B son incompatibles si $\mathrm{A}\cap\mathrm{B}=\varnothing$

Propiedades

1.–Asocativa

$$(\mathrm{A}\cup\mathrm{B})\cup\mathrm{C}=\mathrm{A}\cup(\mathrm{B}\cup\mathrm{C})$$
$$(\mathrm{A}\cap\mathrm{B})\cap\mathrm{C}=\mathrm{A}\cap(\mathrm{B}\cap\mathrm{C})$$

2.–Conmutativa

$$\mathrm{A}\cup\mathrm{B}=\mathrm{B}\cup\mathrm{A}$$
$$\mathrm{A}\cap\mathrm{B}=\mathrm{B}\cap\mathrm{A}$$

3.–Idempotente $\quad \mathrm{A}\cup\mathrm{A}=\mathrm{A} \quad \mathrm{A}\cap\mathrm{A}=\mathrm{A}$

4.–Simplificativa

$$\mathrm{A}\cup(\mathrm{B}\cap\mathrm{A})=\mathrm{A}$$
$$\mathrm{A}\cap(\mathrm{B}\cup\mathrm{A})=\mathrm{A}$$

5.–Distributiva

$$\mathrm{A}\cup(\mathrm{B}\cap\mathrm{C})=(\mathrm{A}\cup\mathrm{B})\cap(\mathrm{A}\cup\mathrm{C})$$
$$\mathrm{A}\cap(\mathrm{B}\cup\mathrm{C})=(\mathrm{A}\cap\mathrm{B})\cup(\mathrm{A}\cap\mathrm{C})$$

6.–$\mathrm{A}\cup\overline{\mathrm{A}}=\mathrm{E}$ y $\quad \mathrm{A}\cap\overline{\mathrm{A}}=\varnothing$

Leyes de Morgan:

$$(\mathrm{A}\cup\mathrm{B})=\overline{\mathrm{A}}\cap\overline{\mathrm{B}}$$
$$(\mathrm{A}\cap\mathrm{B})=\overline{\mathrm{A}}\cup\overline{\mathrm{B}}$$

Probabilidad de la unión de dos sucesos

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

En el caso de que sean incompatibles: $p(A \cup B) = p(A) + p(B)$

Ejemplo:

Disponemos de una baraja de 40 cartas. Sea el suceso A = $\{\text{sacar un oro}\}$ y el suceso B ={sacar una figura} Calcular la probabilidad de obtener un oro o una figura, al extraer una carta de la baraja.

En una baraja española de 40 cartas hay 10 de cada palo y 12 figuras en total, 3 figuras por palo.

$$p(A) = \frac{10}{40} = \frac{1}{4} \qquad p(B) = \frac{12}{40} = \frac{3}{10}$$

El suceso A $\cap$ B lo forman las tres figuras del palo de oros, $p(A \cap B) = \frac{3}{40}$

Por lo tanto: $p(A \cup B) = p(A) + P(B) - p(A \cap B) = \frac{1}{4} + \frac{3}{10} - \frac{3}{40} = \frac{19}{40}$

Probabilidad de sucesos contrarios

Si A y $\overline{A}$ son sucesos contrarios, entonces $p(\overline{A}) = 1 - p(A)$

Ejemplo:

En el experimento consistente en lanzar tres monedas. Hallar la probabilidad de S={obtener por lo menos una cara}

E = {ccc, ccx, cxc, xcc, cxx, xcx, xxc, xxx}

Es más fácil calcular la probabilidad del suceso contrario $\overline{S}$ = {obtener ninguna cara} = {obtener tres cruces}

$$p(\overline{S}) = \frac{1}{8}$$

Por lo tanto $p(S) = 1 - p(\overline{S}) = 1 - \frac{1}{8} = \frac{7}{8}$

Probabilidad condicionada

Normalmente el suceso objeto de estudio es, en realidad, un suceso compuesto por dos o más experimentos y en tales casos el cálculo de probabilidades puede simplificarse mediante el concepto de **probabilidad condicionada.**

Ejemplo:

Para tratar de curar una enfermedad, se ha aplicado un nuevo tratamiento a una serie de individuos.
La siguiente tabla refleja los resultados obtenidos.

	Curados	No curados	
Tratamiento nuevo	60	21	81
Tratamiento anterior	43	36	79
	103	57	160

- Probabilidad de que se haya curado:

$$p(\text{curado}) = \frac{103}{160}$$

- Probabilidad de que haya recibido el nuevo tratamiento:

$$p(\text{nuevo tratamiento})) = \frac{81}{160}$$

- Probabilidad de que se haya curado con el tratamiento nuevo:

$$p(\text{curado con el tratamiento nuevo}) = \frac{60}{103}$$

- Probabilidad de que se haya curado sabiendo que ha recibido el tratamiento nuevo:

$$p(\text{curado/tratamiento nuevo}) = \frac{60}{81}$$

- Probabilidad de que haya recibido el tratamiento nuevo sabiendo que se ha curado:

$$p(\text{tratamiento nuevo/curado}) = \frac{60}{103} = \frac{p(\text{tratamiento nuevo} \cap \text{curado})}{p(\text{curado})} = \frac{\frac{60}{160}}{\frac{103}{160}} = \frac{60}{103}$$

Si A y B son dos sucesos de un mismo espacio muestral, se llama **probabilidad del suceso A condicionado al de B**, y se escribe $p(A/B)$, a la probabilidad del suceso A sabiendo que el suceso B se ha verificado.

$$p(A/B) = \frac{p(A \cap B)}{p(B)} \qquad p(B/A) = \frac{p(A \cap B)}{p(A)}$$

Regla del producto

$$\left.\begin{array}{l} p(A\cap B)=p(A)\cdot p(B/A) \\ p(A\cap B)=p(B)\cdot p(A/B) \end{array}\right\} \Rightarrow p(A\cap B)=p(B)\cdot p(A/B)=p(A)\cdot p(B/A)$$

Diremos que dos sucesos son **independientes** si la realización de uno no modifica la probabilidad de realización del otro. Por tanto A y B son independientes si:

$$p(A/B)=p(A)$$
$$p(B/A)=p(B)$$

Y entonces, por la regla del producto: $p(A\cap B)=p(A)\,p(B)$

En caso contrario se dirá que los sucesos son **dependientes**.

Tablas de contingencia y diagramas de árbol

Una **tabla de contingencia** es una forma de presentar los datos que permite abordar de forma sencilla la resolución de problemas de cálculo de probabilidades.

En general una tabla de contingencia refleja todas las posibilidades que pueden presentar dos sucesos y es de la forma:

	A	$\overline{A}$	Total
B	$p(A\cap B)$	$p(\overline{A}\cap B)$	P(B)
$\overline{B}$	$p(A\cap \overline{B})$	$p(\overline{A}\cap \overline{B})$	$p(\overline{B})$
Total	P(A)	$p(\overline{A})$	1

En el ejemplo anterior se utilizó una tabla de contingencia para resolver el problema.

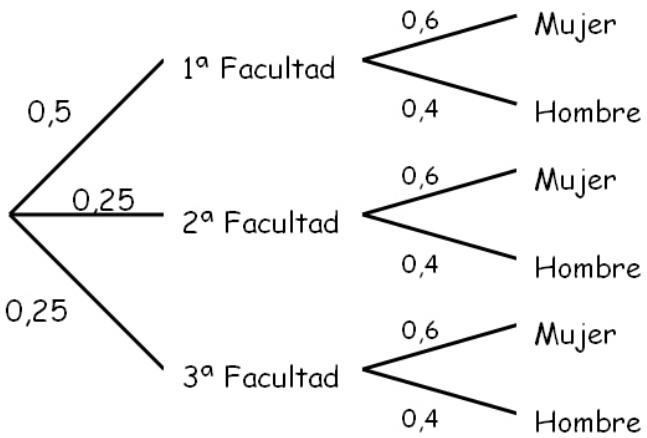
Un **diagrama de árbol** es otra forma de representar determinadas situaciones.

<u>Ejemplo</u>:

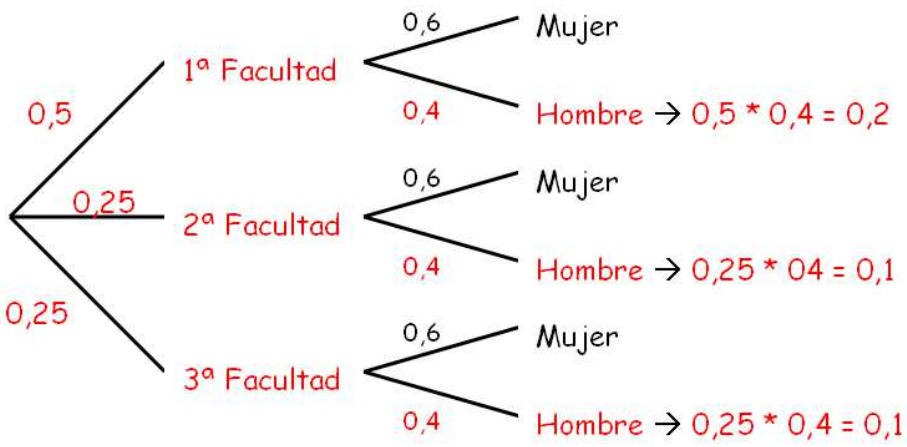
Una universidad está formada por tres facultades:

- La 1ª con el 50% de estudiantes.
- La 2ª con el 25% de estudiantes.
- La 3ª con el 25% de estudiantes.

Las mujeres están repartidas uniformemente, siendo un 60% del total en cada facultad.



¿Cuál es la probabilidad de encontrar un alumno varón?



p(varón) = 0.2 + 0.1 + 0.1 =0.4

Probabilidad total

Si $A_1, A_2, ..., A_n$ son sucesos incompatibles dos a dos tales que $A_1 \cup A_2 \cup ... \cup A_n = E$, la probabilidad de cualquier suceso B de E es:

$$p(B)=p(B/A_1)\cdot p(A_1)+p(B/A_2)\cdot p(A_2)+...+p(B/A_n)\cdot p(A_n)$$

Esta fórmula se conoce como de la probabilidad total.

Teorema de Bayes

La probabilidad condicionada de A_i dado B es:

$$p(A_i/B)=\frac{p(A_i \cap B)}{p(B)}$$

Aplicando la probabilidad total y la probabilidad de la intersección:

$$p(A_i/B)=\frac{p(A_i \cap B)}{p(B)}=\frac{p(A_i)\cdot p(B/A_i)}{p(B)}=\frac{p(A_i)\cdot p(B/A_i)}{p(B/A_1)\cdot p(A_1)+p(B/A_2)\cdot p(A_2)+...+p(B/A_n)\cdot p(A_n)}$$

Por tanto la fórmula de Bayes es:

$$p(A_i/B)=\frac{p(A_i)\cdot p(B/A_i)}{p(B/A_1)\cdot p(A_1)+p(B/A_2)\cdot p(A_2)+...+p(B/A_n)\cdot p(A_n)}$$

Las probabilidades $p(A_i)$ se denominan probabilidades **a priori.**

Las probabilidades $p(B/A_i)$ se denominan probabilidades **a posteriori.**

EJERCICIOS Y PROBLEMAS

1. En la lotería primitiva, ¿cuál es la probabilidad de que los seis números extraídos sean pares?

2. Calcular la probabilidad de obtener una copa o una espada al sacar una carta de la baraja.

3. En un instituto se ha realizado una encuesta y los resultados se muestran en la tabla siguiente:

	SI Juegan al baloncesto	NO Juegan al baloncesto	TOTAL
1º CURSO	45	15	60
2º CURSO	32	58	90
TOTAL	77	73	150

   Se selecciona un alumno al azar, si sabemos que ese alumno es de 1º, ¿cuál es la probabilidad de que juegue al baloncesto?

4. Halla la probabilidad de que al extraer sucesivamente dos cartas de una baraja de 40, resulten ser dos ases:
   a. Sin devolver al mazo la primera carta extraída.
   b. Devolviéndola antes de la segunda extracción.

5. En el experimento de lanzar 3 monedas, sean los sucesos A={obtener una cara como máximo} y B = {obtener cara y cruz} . ¿Son A y B dependientes o independientes?

6. En cierta ciudad el 40% de la población tiene cabello castaño, el 25% tiene los ojos castaños y el 15% tiene cabello y ojos castaños. Se escoge una persona al azar.
   Calcular:
   a. Si tiene cabello castaño, ¿cuál es la probabilidad de que también tenga los ojos castaños?
   b. Si tiene los ojos castaños, ¿cuál es la probabilidad de que no tenga cabello castaño?
   c. ¿Cuál es la probabilidad de que no tenga ni cabello ni ojos castaños?

7. En cierto país, donde la enfermedad X es endémica, sabemos que un 12% de la población padece esta enfermedad. Se dispone de una prueba para detectarla pero no es muy fiable ya que da positiva en el 90% de los casos de personas realmente enfermas pero también da positivo en el 5% de las personas sanas. ¿Cuál es la probabilidad de que esté sana una persona a quien la prueba ha dado positiva?

8. Tenemos dos urnas, la urna A con 3 bolas rojas y 5 azules y la urna B con 6 rojas y 4 azules. Sacamos una bola al azar. ¿Cuál es la probabilidad de que sea azul?

9. Un dado está cargado de forma que la probabilidad de obtención de cada cara es proporcional al número de dicha cara. Si se lanza el dado una vez, halla:
    a. La probabilidad de obtener cada cara.
    b. La probabilidad de obtener un número par.

10. En unos grandes almacenes, para celebrar su aniversario, por cada 50€ de compra tienes una oportunidad de sacar premio de 4 bolsas de colores amarillo, rojo, verde y azul.
    En la bolsa amarilla hay 10 boletos, 4 de ellos premiados, en la roja hay 15 boletos, 8 premiados, en la verde 8 boletos, 1 premiado y por último en la azul 10 boletos, 2 premiados. Seleccionamos una bolsa y sacamos un boleto. ¿Cuál es la probabilidad de que esté premiado?

11. Se tienen los sucesos A y B tales que p(A) = 0,7; p(B) = 0,6 y $p\left(\overline{A} \cup \overline{B}\right)$ =0,58. ¿Son independientes A y B?

12. Un estudiante cuenta, para un examen con la ayuda de un despertador, el cual consigue despertarlo en un 80% de los casos. Si oye el despertador, la probabilidad de que realice el examen es 0.9 y, en caso contrario, de 0.5.
    a. Si va a realizar el examen, ¿cuál es la probabilidad de que haya oído el despertador?
    b. Si no realiza el examen, ¿cuál es la probabilidad de que no haya oído el despertador?

13. Tenemos para enviar tres cartas con sus tres sobres correspondientes. Si metemos al azar cada carta en uno de los sobres, ¿cuál es la probabilidad de que al menos una de las cartas vaya en el sobre que le corresponde?

14. Se hace una encuesta en un grupo de 120 personas, preguntando si les gusta leer y ver la televisión. Los resultados son:
    A 32 personas les gusta leer y ver la tele.
    A 92 personas les gusta leer.
    A 47 personas les gusta ver la tele.
    Si elegimos al azar una de esas personas:
    a. ¿Cuál es la probabilidad de que no le guste ver la tele?
    b. ¿Cuál es la probabilidad de que le guste leer, sabiendo que le gusta ver la tele?
    c. ¿Cuál es la probabilidad de que le guste leer?

15. Tenemos dos bolsas, A y B. En la bolsa A hay 3 bolas blancas y 7 rojas. En la bolsa B hay 6 bolas blancas y 2 rojas. Sacamos una bola de A y la pasamos a B. Después extraemos una bola de B.
    a. ¿Cuál es la probabilidad de que la bola extraída de B sea blanca?

b. ¿Cuál es la probabilidad de que las dos bolas sean blancas?

16. En un control de tráfico fueron multados 10 conductores: siete, por no llevar puesto el cinturón de seguridad, y los otros tres, por circular a mayor velocidad de la permitida. Elegidos al azar dos de los conductores sancionados, calcula la probabilidad de que ambos hayan sido multados por exceso de velocidad.

17. Una bolsa contiene 5 bolas rojas, 10 negras y 12 azules. Se extraen 2 bolas al azar. Calcula la probabilidad de que ambas sean del mismo color.

18. Completa la siguiente tabla de contingencia que muestra la distribución de las tres clases de 1º de bachillerato de un centro escolar.

	Alumnos	Alumnas	
A	30		
B		60	100
C			78
	100		232

Se escoge un estudiante al azar. Calcula la probabilidad de que:

a. Pertenezca a la clase A.
b. Sea una alumna.
c. Sea una alumna y esté en la clase B.
d. Pertenezca a la clase C sabiendo que es alumna.
e. Sea un alumno sabiendo que pertenece a la clase A.

19. María y Paula juegan un partido de tenis de mesa. La vencedora será la primera que gane dos de los tres sets de que consta el encuentro.

a. Dibuja un diagrama de árbol con todos los posibles resultados.
b. Calcula la probabilidad de que Paula gane el partido si la probabilidad de que María logre un set es de 0,4.

20. En un aula con 24 estudiantes de 1º de ESO, los profesores de Matemáticas, Lengua e Inglés piden cada día al azar los cuadernos a algunos alumnos para revisarlos. El de Matemáticas se lo reclama a 4 alumnos; el de Lengua, a 6, y el de Inglés, a 8.
Halla la probabilidad de que a un alumno concreto, en un día:

a. Le pidan 2 cuadernos.
b. No le reclamen ninguno.
c. Le soliciten los 3 cuadernos.

AUTOEVALUACIÓN 2

1. En un aula hay 18 chicos y 20 chicas, de los que $\frac{1}{3}$ de los chicos y la mitad de las chicas tienen el pelo negro.
    a. ¿Cuál es la probabilidad de que al elegir un alumno al azar sea chico o tenga el pelo negro?
    b. Si el alumno elegido tiene el pelo negro, ¿cuál es la probabilidad de que no sea chico?

    $$\left(a.\frac{14}{19} \quad b.\frac{5}{8} \right)$$

2. En un centro escolar los alumnos pueden optar por cursar como lengua extranjera inglés o francés. En un determinado curso, el 90% de los alumnos estudia inglés y el resto francés. El 30% de los que estudian inglés son chicos y de los que estudian francés son chicos el 40%. El elegido un alumno al azar, ¿cuál es la probabilidad de que sea chica?

    (0.69)

3. Un taller sabe que por término medio acuden: por la mañana tres automóviles con problemas eléctricos, ocho con problemas mecánicos y tres con problemas de chapa, y por la tarde dos con problemas eléctricos, tres con problemas mecánicos y uno con problemas de chapa.
    a. Hacer una tabla ordenando los datos anteriores.
    b. Calcular el porcentaje de los que acuden por la tarde.
    c. Calcular el porcentaje de los que acuden por problemas mecánicos.
    d. Calcular la probabilidad de que un automóvil con problemas eléctricos acuda por la mañana.

    (b. 30% c.55% d.0.6)

4. Si A y B son dos sucesos tales que: p (A)= 0,4 p(B/A)=0,25 y $p(\overline{B})$= 0,75
    a. ¿Son A y B independientes?
    b. Calcula $p(A \cup B)$ y $p(A \cap B)$

    (a. Son independientes b. 0.1 y 0.55)

5. En unas oposiciones, el temario consta de 85 temas. Se eligen tres temas al azar de entre los 85. Si un opositor sabe 35 de los 85 temas, ¿cuál es la probabilidad de que sepa al menos uno de los tres temas?

    (0.802)

6. De un estuche que contiene 5 bolígrafos azules y 6 negros, se sacan sin mirar dos de ellos. Halla la probabilidad de que ambos sean de distinto color.

    $$\left(\frac{6}{11} \right)$$

7. Completa la siguiente tabla de contingencia, que muestra el tipo de medio de transporte que utilizan para llegar hasta su puesto de trabajo los 200 empleados de una empresa situada en la periferia de una gran ciudad.

	Hombres	Mujeres	
Público		50	85
Privado			
	120		

Se escoge un trabajador al azar. Calcula la probabilidad de que:

a. Sea un hombre y utilice el transporte público.
b. Utilice el transporte público sabiendo que es un hombre.
c. Sea una mujer sabiendo que usa el transporte privado.
d. ¿Los sucesos "ser hombre" y "utilizar el transporte público" son dependientes o independientes?

Razona tu respuesta.

(a.0.175 b.0.29 c.0.26 d. Dependientes)

8. Se pide a dos chicas que escriban, por separado, una de las cinco vocales.
   a. ¿Cuál es la probabilidad de que las dos escriban la a?
   b. ¿Cuál es la probabilidad de que las dos escriban la misma?

$$\left(a.\frac{1}{25} \quad b.\frac{1}{5}\right)$$

9. El 20% de los empleados de una empresa son ingenieros y otro 20% son economistas. El 75% de los ingenieros ocupan un puesto directivo y el 50% de los economistas también, mientras que los no ingenieros y los no economistas solamente el 20% ocupa un puesto directivo. ¿Cuál es la probabilidad de que un empleado directivo elegido al azar sea ingeniero? (0.405)

10. En la caja A hay un dado de cuatro caras, en la caja B uno de seis caras y en la caja C otro de ocho. Elegimos una caja al azar y lanzamos el dado que contiene.
    Calcula las probabilidades de que:
    a. Salga un 3.
    b. Salga un 3, si resultó elegida la caja A.
    c. Hayamos tirado el dado de la caja A, si sabemos que ha salido un 3.
    d. Salga un 6.
    e. Hayamos sacado el dado de la caja A, si sabemos que salió un 6.
    f. Salga un 6, si la caja seleccionada fue la caja C.

$$\left(a.\frac{13}{72} \quad b.\frac{1}{4} \quad c.\frac{6}{13} \quad d.\frac{7}{72} \quad e.0 \quad f.\frac{1}{8}\right)$$

7.5. *MEDIDAS DE CENTRALIZACIÓN Y DISPERSIÓN*

DEFINICIONES:

La **Estadística** es la rama de las Matemáticas que utiliza conjuntos de datos numéricos para obtener inferencias basadas en el cálculo de probabilidades. Es una ciencia relativamente reciente, pues sus orígenes se remontan al siglo XVIII. Pero su implantación hoy en día es muy acusada:

- Se diseñan encuestas para recopilar información previa al día de elecciones y así predecir el resultado de las mismas.
- Se seleccionan al azar consumidores para obtener información con el fin de predecir la preferencia con respecto a ciertos productos y/o servicios.
- Los economistas consideran varios índices de la situación económica durante cierto periodo y utilizan la información para predecir la situación económica futura.
- Su utilidad es evidente también para los asesores financieros que han de evaluar las oportunidades de inversión a través de las bolsas de valores.
- Los portales de apuestas deportivas online recurren a la Estadística para, de acuerdo con todos los datos hasta la fecha, determinar el nivel de confianza de cada una de los posibles resultados.

La Estadística se divide en dos grandes ramas:

- La **Estadística Descriptiva** se dedica a recolectar, ordenar, analizar y representar un conjunto de datos, con el fin de describir apropiadamente las características de éstos. Utiliza para ello las llamadas «medidas de centralización».

- La **Estadística Inductiva** o **Inferencial** tiene como objeto obtener conocimientos sobre un colectivo, utilizando para ello las observaciones de una muestra, para sí poder inferir resultados. En este proceso se utiliza el cálculo de probabilidades.

Para todo lo anterior, la Estadística trabaja con una serie de aspectos, cualidades o propiedades de los individuos de la población, llamados caracteres; los valores que recorre un determinado carácter se llaman variables estadísticas. Pueden ser de varios tipos:

- **cuantitativas:** son medibles, es decir, se describen mediante números
  - **discretas:** sólo toman valores puntuales (p. ej. nº de hijos …)
  - **continuas**: puede tomar cualquier valor entre dos cualesquiera (p.ej estatura…)
- **cualitativas:** no son medibles, se describen mediante modalidades (p. ej. color del pelo…)

Población es el conjunto de elementos que se investigan, **muestra** es una parte representativa de la población, e **individuo** es cada uno de los elementos que forman la población.

FRECUENCIAS Y TABLAS:

Ejemplo 1

En un instituto hay una clase de 1º de Bachillerato cuyos 20 alumnos presentan las siguientes edades: 16, 16, 16, 17, 19, 16, 17, 18, 16, 17, 16 , 18, 16, 16, 16, 16, 18, 16, 16 y 16.

La variable que vamos a estudiar en esta distribución es la edad (variable cuantitativa), que llamaremos x_i, para ello vamos a construir la siguiente tabla:

Edad(años)	f_i	F_i	h_i	H_i
16	13	13	0,65	0,65
17	3	16	0,15	0,80
18	3	19	0,15	0,95
19	1	20	0,05	1
	$\sum = 20$		$\sum = 1$	

La 2ª columna recoge la **frecuencia absoluta**, f_i, que es el número de veces que aparece cada valor de la variable.

La 3ª columna refleja la **frecuencia absoluta acumulada**, F_i, que se obtiene sumando la frecuencia absoluta de cada fila con las anteriores.

En la 4ª columna tenemos la **frecuencia relativa**, h_i, que es la frecuencia absoluta dividida por el nº de datos, N:

$$h_i = \frac{f_i}{N}$$

Obviamente, la suma de las frecuencias absolutas es N, y la de las relativas es 1.

Finalmente, la última columna recoge la frecuencia relativa acumulada, H_i, que se obtiene sumando la frecuencia relativa de cada fila con las anteriores. Por ejemplo, el dato 0,95 de la 3ª fila significa que el 95% de los alumnos tienen 18 años o menos.

Ejemplo 2:

En la misma clase del ejemplo anterior los 20 alumnos presentan las siguientes estaturas (en cm):164, 175, 165, 170, 168, 157, 167, 172, 177, 160, 168, 160, 164, 174, 170, 182, 161, 171, 173 y 194.

La variable que vamos a estudiar ahora es la estatura (variable cuantitativa). Para confeccionar la tabla utilizaremos intervalos de amplitud 10 llamados - **intervalos de clase** –, comenzando por 155:

Estatura(cm)	f_i	F_i	h_i	H_i
[155-165)	6	6	0,30	0,30
[165-175)	10	16	0,50	0,80
[175-185)	3	19	0,15	0,95
[185-195)	1	20	0,05	1
	$\sum = 20$		$\sum = 1$	

En cada intervalo de clase se incluye el extremo inferior pero no el superior, salvo en el último. El punto medio de cada intervalo se llama **marca de clase**, y lo denotaremos por x_i.

¿Cuándo utilizar un tipo de tabla u otro?:

1º) Tablas con los valores de la variable individualizados (como en el ejemplo 1): cuando, ya sean pocos o muchos datos, **la variable toma pocos valores diferentes** (es decir, los valores se repiten mucho).

2º) Tablas con los valores de la variable agrupados en intervalos de clases (como en el ejemplo 2): **cuando el número de datos y de valores diferentes que toma la variable son grandes**. Con ello perderemos algo de información pero ganaremos en claridad.

En este último caso, ¿cuántos intervalos de clase utilizar? Existe un criterio orientativo según el cual el nº de clases debe ser aproximadamente igual a la $\sqrt{\ }$ del número de datos:

$$\text{nº de clases}=\sqrt{N}$$

En el ejemplo anterior sería $\sqrt{20} \simeq 4,47$, es decir, 4 intervalos vendrían bien. A continuación, se determina la amplitud de los intervalos teniendo en cuenta los valores mínimo (157 cm) y máximo (194 cm) de la distribución: $\dfrac{194\text{cm}-157\text{cm}}{4 \text{ intervalos}} \simeq 9,25\text{cm/intervalo}$

de modo que 10 cm por intervalo es lo apropiado. A la hora de decidir dónde comienza el primer intervalo se recomienda que, finalmente, los extremos de los intervalos no coincidan con ninguno de los datos.

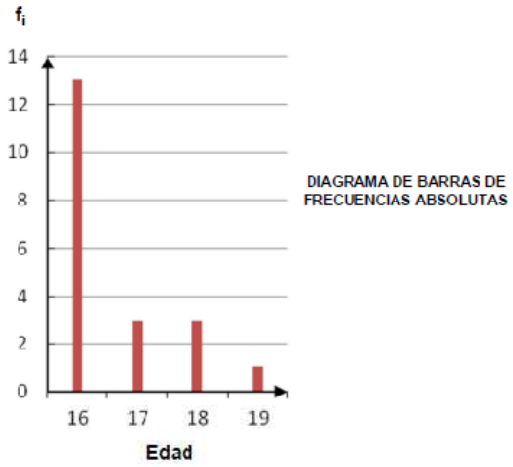
Nótese que en la práctica el elegir un tipo de tabla u otro puede ser relativo: ¿qué se entiende por "pocos datos"? Veremos que una misma distribución se puede estudiar con dos tablas no necesariamente iguales, y las dos pueden ser perfectamente válidas.

REPRESENTACIONES GRÁFICAS:

Diagrama de barras/Histograma

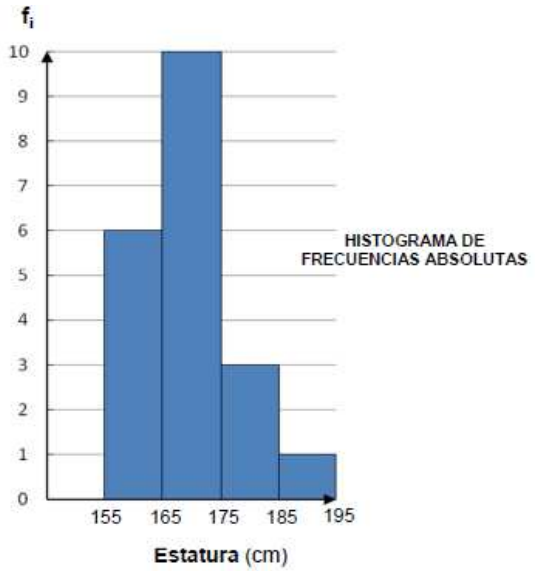
Consideremos la distribución del ejemplo 1. En unos ejes cartesianos situamos en el eje horizontal las edades y en el vertical la frecuencia absoluta fi. Levantamos, a continuación, barras cuya altura es la frecuencia. Obtendremos así un **diagrama de barras**:

Edad (años)	f_i
16	13
17	3
18	3
19	1
	Σ=20



Si hacemos lo mismo con el ejemplo 2, pero, esta vez, dando a las barras el ancho de los intervalos, obtendremos un **histograma**:

Estatura (cm)	f_i
[155–165)	6
[165–175)	10
[175–185)	3
[185–195]	1
	Σ=20



En el histograma anterior, y puesto que los intervalos tenían la misma amplitud, los rectángulos tienen la altura correspondiente a la f_i. Ahora bien, si son de distinta amplitud, entonces habrá que ajustar la altura a_i de cada rectángulo mediante la siguiente fórmula:

$$\text{área} = f_i = \text{amplitud} \cdot a_i \Rightarrow a_i = \frac{f_i}{\text{amplitud}}$$

Veámoslo en el siguiente ejemplo:

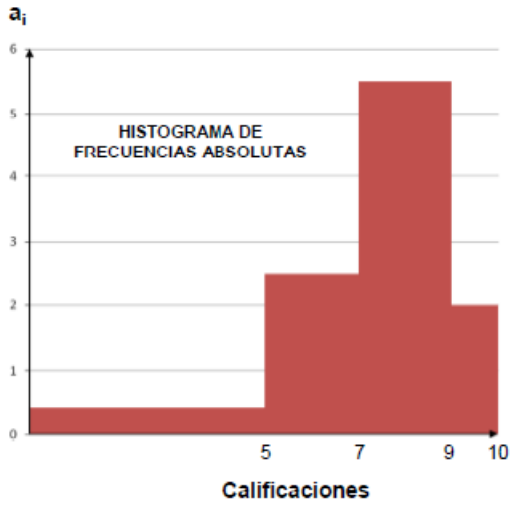
Ejemplo 3:

Las calificaciones de un examen de 20 alumnos son: 10, 8, 7, 5, 5, 7, 7, 7, 5, 7, 10, 7, 4, 6, 6, 8, 7, 8, 7 y 4.

Confeccionar la correspondiente tabla agrupando los datos en intervalos de suspensos (0 a 5), aprobados (5 a 7), notables (7 a 9) y sobresalientes (9 a 10). Construir el histograma de frecuencias absolutas.

Calificaciones	f_i	F_i	$a_i = f_i/\text{amplitud}$	h_i	H_i
[0-5)	2	2	2/5=0,4	0,10	0,30
[5-7)	5	7	5/2=2,5	0,25	0,35
[7-9)	11	18	11/2=5,5	0,55	0,90
[9-10]	2	20	2/1=2	0,10	1
	$\sum = 20$			$\sum = 1$	

El histograma quedaría:



¿Cuál es la diferencia entre el diagrama de barras y el histograma?

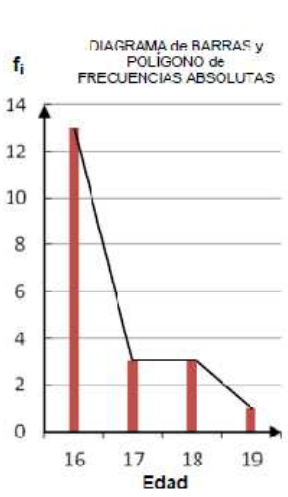
- El diagrama de barras visualiza las frecuencias como alturas, y se utiliza para variables discretas (y también para cualitativas).
- En el histograma el área de cada rectángulo representa la frecuencia correspondiente, y se utiliza para datos agrupados en intervalos.

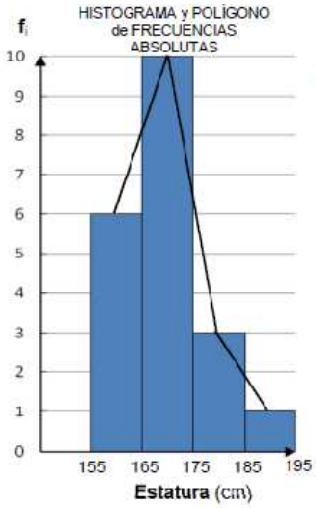
En el histograma anterior, y puesto que los intervalos tenían la misma amplitud, los rectángulos tienen la altura correspondiente a la fi.

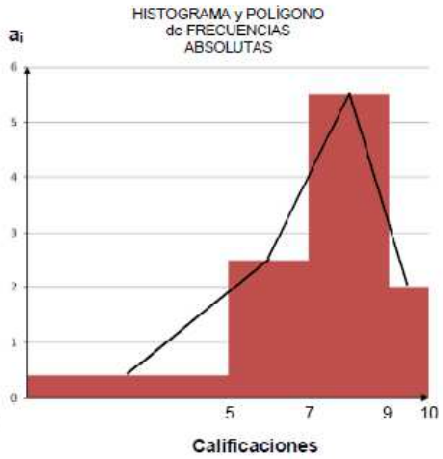
Polígono de frecuencias

Uniendo los extremos superiores de las barras de un diagrama de barras, o los puntos medios del lado superior de cada rectángulo de un histograma, obtenemos el llamado **polígono de frecuencias.**

Veámoslo para los ejemplos anteriores:



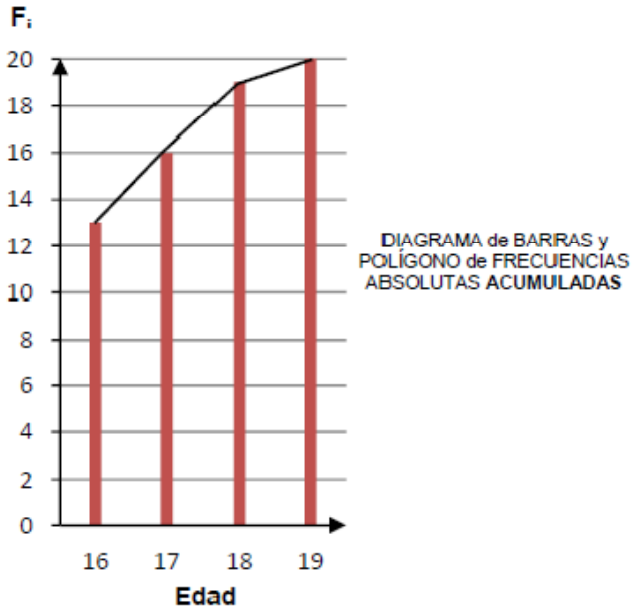




Observaciones:

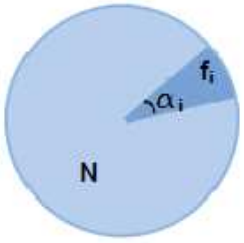
En los ejemplos anteriores lo que se representaba era la frecuencia absoluta. Pero, evidentemente, también existen diagramas de barras o histogramas de frecuencias relativas, de frecuencias acumuladas, polígonos de frecuencias absolutas o relativas acumuladas, etc.

Por ejemplo, en el caso del ejemplo 1:



Obviamente, si la variable es cualitativa, no tienen sentido los gráficos acumulativos.

Gráfico de sectores:

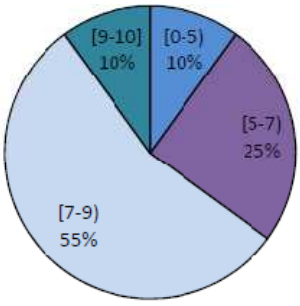


Si dividimos un círculo en **sectores circulares** de amplitud angular proporcional a su frecuencia absoluta, obtendremos un **gráfico de sectores**. Vamos a obtener, mediante regla de tres, la fórmula que nos indique cuántos grados le corresponden a cada sector.

$$\left.\begin{array}{l} 360^o \to N \\ \alpha_i \to f_i \end{array}\right\} \Rightarrow \boxed{\alpha_i = \frac{f_i}{N} \cdot 360^o = h_i \cdot 360^o}$$

Vamos a aplicarla al caso concreto del ejemplo 3:

Calificaciones	f_i	h_i	$\alpha_i = h_i \cdot 360^o$
[0-5)	2	0,10	36º
[5-7)	5	0,25	90º
[7-9)	11	0,55	198º
[9-10]	2	0,10	36º
	$\sum = 20$	$\sum = 1$	



Los ángulos suman 360^o. Además, se suele indicar el % de cada sector.

Por último, indicar que, en el caso de variables cualitativas, existen otros tipos de representaciones gráficas como pictogramas, cartogramas, diagramas de columnas apiladas o pirámides de población.

CUADRO-RESUMEN

- variable
  - cuantitativa
    - discreta: diagrama de barras ↔ polígono de frecuencias; diagrama de sectores
    - continua: histograma ↔ polígono de frecuencias; diagrama de sectores
  - cualitativa: diagrama de barras (nunca frecuencias acumuladas) ↔ polígono de frecuencias; diagrama de sectores; pictogramas, cartogramas, pirámides de población...

Medidas de centralización

Media aritmética: Se define como la suma de todos los valores x_i dividida por el número de valores, N. Se designa como $\overline{x}$:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_n}{N}$$

Ahora bien, en general, cada valor x_i se repetirá con una frecuencia f_i. En ese caso, la fórmula sería:

$$\boxed{\overline{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_n x_n}{N}}$$

Observaciones:

- En el caso de valores agrupados, x_i indica la marca de clase, es decir, el punto intermedio de cada intervalo.
- Obviamente, no existe la media si los datos son cualitativos. Ni tampoco si los datos están agrupados y alguna clase está abierta (p.ej. si en una encuesta el último grupo fueran "mayores de 60").

Vamos a ver cómo se calcula la media en los ejemplos 1 y 2.

Edad(años)	f_i	$f_i x_i$
16	13	208
17	3	51
18	3	54
19	1	19
	$\sum = 20$	$\sum = 332$

$$\overline{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N} = \frac{332}{20} = 16,6 \text{ años;}$$

Estatura(cm)	f_i	$f_i x_i$
[155-165)	6	960
[165-175)	10	1700
[175-185)	3	540
[185-195)	1	190
	$\sum = 20$	$\sum = 3390$

$$\overline{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N} = \frac{3390}{20} = 169,5 \text{ cm}$$

Moda: Es el valor de la variable que tiene mayor frecuencia. Representa el valor dominante de la distribución; por ejemplo, en unas elecciones sería el partido más votado.

Ejemplo 1:

Edad(años)	f_i
16	13
17	3
18	3
19	1
	$\sum = 20$

Moda=16

Ejemplo 2:

Estatura(cm)	f_i
[155-165)	6
[165-175)	10
[175-185)	3
[185-195)	1
	$\sum = 20$

Intervalo modal: [165-175)

Observaciones:

- La moda existe siempre, en cualquier tipo de distribuciones (cualitativas o cuantitativas).
- Vemos que en las que presentan datos agrupados en intervalos se habla de intervalo modal o clase modal.

Mediana: Es el valor central de los datos, una vez ordenados de menor a mayor. Si el número de datos es par, se toma el valor medio de los dos centrales.

Ejemplo 4:

Los sueldos mensuales (€) de los 7 trabajadores de una empresa son:

3000 915 650 825 700 775 1580

Si calculamos la media (1206,43€) observamos que no es muy representativa de los datos de esta distribución, ya que éstos se encuentran muy dispersos. Obtengamos la mediana:

650 700 775 875 915 1580 3000

El valor obtenido 875€, es más representativo de la mayoría de los datos.

En una variable discreta, cuando tenemos más datos, construimos la tabla de frecuencias acumuladas, F_i , calculamos N/2 y buscamos los valores $F_i$que verifiquen

$$F_{i-1} < \frac{N}{2} < F_i$$

La mediana será el valor x_i de la variable correspondiente a F_i, es decir, el primer valor de la variable que excede a N/2.

Nota: Si coincide $F_{i-1} = \frac{N}{2} < F_i$, se toma $M_e = \frac{x_{i-1} + x_i}{2}$

En el **ejemplo 1**:

Edad(años)	f_i	F_i
16	13	13
17	3	16
18	3	19
19	1	20
	$\sum = 20$	

Mediana M_e=16 años, ya que $\frac{N}{2} = 10$

Si la variable está agrupada en intervalos calcularemos el intervalo mediano, en el ejemplo 2 será:

Intervalo mediano [165-175), ya que $\frac{N}{2}=10$ →

Estatura(cm)	f_i	F_i
[155-165)	6	6
[165-175)	10	16
[175-185)	3	19
[185-195)	1	20
	$\sum = 20$	

Observaciones:

- Curiosamente, la mediana depende del orden de los datos y no de su valor.
- La mediana se puede calcular también en distribuciones de tipo cualitativo cuyas modalidades se pueden ordenar.
- En una distribución sólo hay una media y una mediana, pero puede haber varias modas.

Medidas de dispersión

Tienen por objeto dar una idea de la mayor o menor concentración de los valores de una distribución alrededor de los valores centrales.

Recorrido:

Es la diferencia entre el mayor y el menor valor de una distribución.

Ejemplo 5: Las notas de dos alumnos son: Carlos: 5, 7, 7, 7, 9 y Ana 2, 6, 8, 10, 9
Comprueba que ambos tienen de media 7. Pero Ana tiene sus notas mucho más dispersas que Carlos.
Recorrido de Carlos: 9-7=2 Recorrido de Ana: 10-2=8

Varianza y Desviación típica:

Antes de definir la varianza, conviene considerar en una serie de datos su desviación respecto a la media, que sería la diferencia entre cada dato y la media, en valor absoluto (para que siempre sea >0): $d_i = |x_i - \overline{x}|$

La varianza, σ^2, se define como «la media aritmética de los cuadrados de las desviaciones respecto a la media»:

$$\sigma^2 = \frac{\sum_{i=1}^{n} f_i \left(x_i - \overline{x}\right)^2}{N}$$

Observaciones:

- No es necesario el valor absoluto de las desviaciones, puesto que están al cuadrado.
- La fórmula es equivalente a la siguiente que es más utilizada:

$$\boxed{\sigma^2 = \frac{\sum_{i=1}^{n} f_i x_i^{\,2}}{N} - \overline{x}^2}$$

- Las unidades de la varianza van al cuadrado.

Como la varianza, por estar expresada en unidades cuadradas, no se puede comparar con la media, se utiliza la **desviación típica**, σ, que se define como la raíz cuadrada de la varianza:

$$\boxed{\sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i x_i^{\,2}}{N} - \overline{x}^2}}$$

- La σ tiene las mismas unidades que la variable x_i.
- La σ nos dice cómo de alejados de la media, es decir, cómo de dispersos se encuentran los datos: cuanto más agrupados estén los datos en torno a la media, menor será σ. De hecho, para poder comparar varias distribuciones y ver cuál está más dispersa respecto a la media se utiliza el llamado **<u>coeficiente de variación o dispersión</u>**:

$$C.V. = \frac{\sigma}{\overline{x}} \text{ que habitualmente se expresa en \%.}$$

<u>Ejemplo 5</u>: Las notas de dos alumnos son: Carlos: 5, 7, 7, 7, 9 y Ana 2, 6, 8, 10, 9.
En ambos casos la media es 7.

$$\sigma^2_{\text{Carlos}} = 1,6 \Rightarrow \sigma_{\text{Carlos}} = \sqrt{1,6} \cong 1,26$$
$$\sigma^2_{\text{Ana}} = 8 \Rightarrow \sigma_{\text{Ana}} = \sqrt{8} \simeq 2,83$$
$$\sigma_{\text{Carlos}} < \sigma_{\text{Ana}}$$

<u>Ejemplo 1:</u>

$$\overline{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N} = \frac{332}{20} = 16,6 \text{ años}$$

$$\sigma^2 = \frac{\sum_{i=1}^{n} f_i x_i^{\,2}}{N} - \overline{x}^2 = \frac{5528}{20} - 16,6^2 = 0,84 \text{ años}^2$$

$$\sigma = \sqrt{0,84} \cong 0,92 \text{ años}$$

Edad(años)	f_i	$f_i x_i$	$f_i x_i$
16	13	208	3328
17	3	51	867
18	3	54	972
19	1	19	361
	$\sum = 20$	$\sum = 332$	$\sum = 5528$

<u>Ejemplo 3:</u>

Estatura(cm)	x_i	f_i	$f_i x_i$	$f_i x_i^2$
2[155-165)	160	6	960	153600
[165-175)	170	10	1700	289000
[175-185)	180	3	540	97200
[185-195)	190	1	190	36100
		$\sum = 20$	$\sum = 3390$	$\sum = 575900$

$$\overline{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N} = \frac{3390}{20} = 169,5 \text{ cm}$$

$$\sigma^2 = \frac{\sum_{i=1}^{n} f_i x_i^{\,2}}{N} - \overline{x}^2 = \frac{575900}{20} - 169,5^2 = 64,75 \text{ cm}^2$$

$$\sigma = \sqrt{64,75} \cong 8,05 \text{ cm}$$

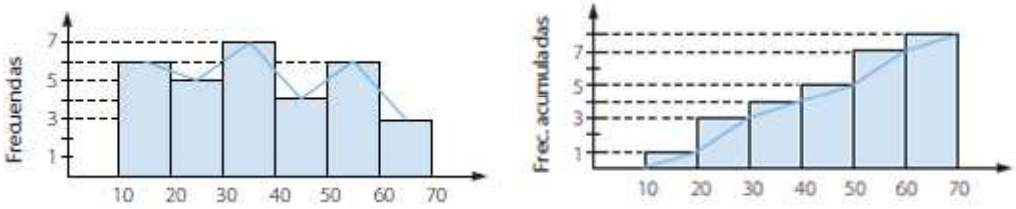
EJERCICIOS Y PROBLEMAS

1. El número de hermanos de 40 alumnos es:

3	4	2	3	4	3	4	4	4	2
3	4	4	3	4	1	2	3	5	4
2	2	2	5	3	4	4	6	2	6
4	3	2	1	2	3	2	4	3	1

   a) ¿De qué tipo de variable se trata? Construir una tabla estadística en la que figuren todas las frecuencias.
   b) ¿Cuántos alumnos tienen 5 o más hermanos? ¿Cuántos 3 o menos?
   c) Construir el diagrama de barras y polígono de frecuencias absolutas, y el diagrama de sectores.

2. Construye las tablas de frecuencias que corresponden a los siguientes gráficos estadísticos, indicando de qué tipo es cada uno.



3. Se aplica un test de inteligencia para averiguar el cociente intelectual de 40 alumnos de 1º de Bachillerato, obteniéndose los siguientes resultados:

106	136	81	110	95	92	99	106	81	95
110	103	88	81	81	99	110	114	128	103
103	74	95	136	95	88	106	121	106	114
117	92	85	125	95	110	132	95	103	81

   a) Razonar qué tipo de variable es. Construir una tabla estadística en la que figuren todas las frecuencias.
   b) ¿Cuántos alumnos tienen un CI por debajo de 100?
   c) Si se consideran superdotados a los que tienen CI>130, ¿hay alguno en clase?
   d) ¿Qué porcentaje de alumnos tiene de CI 110 o más?
   e) Construir el histograma y polígono de frecuencias relativas y el diagrama de sectores.

4. Organiza, en una tabla de frecuencias, estos datos relativos al peso, en kg, de 20 personas. Calcula sus medidas de centralización.

42	51	56	66	75
69	59	50	70	59
47	51	45	63	79
62	54	60	63	58

5. En la figura adjunta aparecen los resultados de la primera jornada de la Liga de fútbol 2014-2015. Se pide:

Local	Resultado	Visitante
Real Madrid	2 - 0	Córdoba
Rayo	0 - 0	Atlético
Málaga	1 - 0	Athletic
Sevilla	1 - 1	Valencia
Granada	2 - 1	Deportivo
Almería	1 - 1	Espanyol
Éibar	1 - 0	R. Sociedad
Barcelona	3 - 0	Elche
Celta	3 - 1	Getafe
Levante	0 - 2	Villarreal

   a) Formar con los goles/partido una tabla estadística apropiada para responder a los apartados siguientes.
   b) Dibujar el diagrama de barras, polígono y diagrama de sectores, todos ellos de frecuencias absolutas
   c) Calcular la media de goles/partido, moda y mediana.

6. El número de horas de sol registradas en un determinado mes en 50 estaciones meteorológicas es:

83	82	78	72	107	107	93	72	85
98	71	76	75	83	72	126	102	76
112	99	155	118	150	129	119	148	181
151	167	156	180	173	149	80	131	121
110	200	162	214	176	186	187	186	141
212	186	199	198	219				

   a) Razonar de qué clase de variable se trata. Confeccionar una tabla estadística de cara a los siguientes apartados.
   b) Dibujar el histograma de frecuencias relativas y el polígono de frecuencias absolutas acumuladas.
   c) Calcular la media de horas de sol, la moda y la mediana.

7. En la tabla figuran los datos de las pulsaciones de un equipo de atletismo después de una carrera:

Pulsaciones	70–74	75–79	80–84	85–89	90–94	95–99
Nº de atletas	3	3	7	10	12	8

   a) ¿Qué tipo de variable es? Construir una tabla apropiada para lo que se pide a continuación.
   b) Hallar la media.
   c) ¿Cuál es el intervalo mediano? ¿y el modal?
   d) Construir el histograma y el polígono de frecuencias absolutas.

8. A un conjunto de cinco notas cuya media es 7,31 se le añaden las calificaciones siguientes: 4,47 y 9, 15 ¿Cuál es la nueva media?

9. Los datos que siguen corresponden a las medidas del tórax, en cm, de cien hombres adultos. Construir la tabla estadística necesaria para responder a las siguientes cuestiones:
   a) Hallar la media.
   b) Obtener la desviación típica.
   c) Dibujar el histograma y el polígono de frecuencias absolutas acumuladas.

Medida del tórax (cm)	Nº de hombres
[80,85)	4
[85,90)	10
[90,95)	24
[95,100)	32
[100,105)	22
[105,110]	8

10. Estos son los pesos de los últimos 20 pacientes de una consulta médica.

42	51	56	66	75	47	51	45	63	79
69	59	50	70	59	62	54	60	63	58

Organiza los datos en una tabla de frecuencias y calcula sus medidas de centralización.

11. De un estudio sobre el peso de los elefantes y el peso de los ratones se tiene esta información.

Peso de los elefantes: $\overline{x} = 2000$ kg; σ=100 kg

Peso de los ratones: $\overline{x} = 0,05$ kg; σ=0,02 kg

Compara la dispersión en las variables.

12. De los 30 alumnos de una clase, el 10 % aprobó todo, el 20 % suspendió una asignatura, el 50 % suspendió dos asignaturas y el resto suspendió más de dos asignaturas.
Realiza una tabla de frecuencias con estos datos. ¿Hay algún tipo de frecuencia que responda a la pregunta de cuántos alumnos suspendieron menos de dos asignaturas?

13. Se va a valorar la eficiencia de dos baterías para cámaras fotográficas. Se repite el siguiente proceso 50 veces:
Se recarga totalmente la batería. Se coloca en la cámara y se hace una fotografía cada tres segundos. Se cuenta el número de fotografías que ha sido posible hacer.
Los resultados han sido:

BATERÍA *A*	
N.º de fotos	f_i
[300, 350)	3
[350, 400)	12
[400, 450)	20
[450, 500)	13
[500, 550)	1
[550, 600)	1

BATERÍA *B*	
N.º de fotos	f_i
[320, 360)	5
[360, 400)	9
[400, 440)	19
[440, 480)	15
[480, 520)	2

   a) Valora cuál es la media aritmética de fotografías que se puede hacer con una recarga de cada tipo de batería y su desviación típica.
   b) ¿En cuál de los dos casos hay menor dispersión?
   c) ¿Qué batería recomendarías comprar, sin considerar el precio? ¿Por qué?

AUTOEVALUACIÓN 3

1. En un centro militar se ha tomado una muestra de 16 jóvenes, obteniéndose las siguientes estaturas (en cm):

   160 172,4 168 167 175 179 180 198
   164 166 174 177 182,5 185 191 173,5

   i. Construir una tabla estadística apropiada.
   ii. Obtener el histograma y el polígono de frecuencias absolutas acumuladas.
   iii. Calcular la media, moda, mediana y desviación típica.

   { $\bar{x} \cong 176{,}25$ cm; Me=175 cm; $\sigma \cong 9{,}9$ cm}

2. Se han medido los pesos y las alturas de 6 personas, obteniéndose los datos siguientes:

   a) ¿Qué medidas están más dispersas, los pesos o las alturas? Utilizar el coeficiente de variación.

Peso (kg)	65	60	65	63	68	68
Altura (cm)	170	150	168	170	175	180

   {*Las alturas, puesto que CVa=5%>CVp=4%*}

   b) Comprobar lo anterior gráficamente.

3. La tabla siguiente nos da las puntuaciones obtenidas por un grupo de 20 alumnos en un test:

Puntuaciones	0–20	20–40	40–50	50–60	60–80	80–100
Nº alumnos	3	6	5	3	–	3

   Al mismo grupo de alumnos se le hace otra prueba y las puntuaciones obtenidas son:

   10 11 20 5 10 8 11 12 5 9
   14 11 3 9 11 12 11 8 9 11

   ¿Qué datos se hallan más dispersos? Utilizar el coeficiente de variación.

   {Las puntuaciones, ya que CV1≅54,7%>CV2≅34,6%}

4. Las calificaciones de los 25 alumnos de 1º de Bach. A son:

   6 6 7 6 7 5 5 6 7 5 4 5 4
   9 3 3 5 5 5 9 5 4 5 4 8

   mientras que las de los 20 alumnos de 2º Bach. B son:

   6 6 7 3 10 3 5 5 2 5 4 3 9
   4 9 5 6 6 6 7

   Calcular en qué grupo las notas están más dispersas.

   {En el B, puesto que CVa≅28,6%<CVp≅38,9%}

5. En un examen, en el que la puntuación varía entre 0 y 10, la media aritmética de los 12 primeros datos de la lista, en un grupo de 20 alumnos, fue 6,5.
   ¿Cuáles son los valores mínimo y máximo que puede tomar la media del grupo?

   {mín=3,9;máx=7,9}

7.6. *ESTADÍSTICA DESCRIPTIVA BIDIMENSIONAL*

En numerosas ocasiones interesa estudiar simultáneamente dos (o más) caracteres de una población. En la práctica unas variables pueden estar relacionadas con otras. Por ejemplo, el peso de una persona y su talla; por el contrario, entre la altura y el cociente intelectual no parece que exista relación alguna.

En el caso de dos (o más) variables estudiadas conjuntamente se habla de **variable bidimensional.** La lista de pares de datos correspondientes a cada individuo de la población (repetidos o no), es lo que llamamos **variable estadística bidimensional.**

Ejemplos

1. A cada uno de los reclutas de un reemplazo se les talla y pesa. Se trata de dos variables cuantitativas.

2. Entre los empleados de una empresa se ha realizado una encuesta sobre el consumo del tabaco, que ha arrojado los siguientes resultados:

Hábito Sexo	Fumadores	No fumadores	Totales de filas
Varones Mujeres	**49** **43**	**64** **37**	**113** **80**
Totales de columnas	**92**	**101**	**Total general** **193**

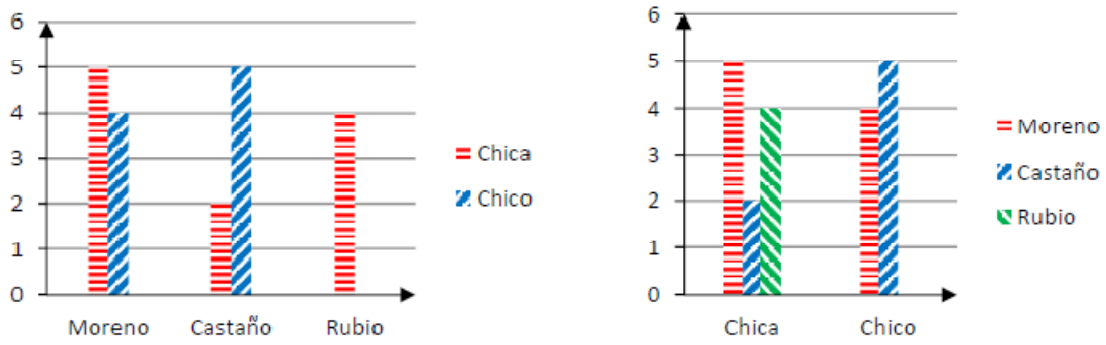
TABLAS DE CONTINGENCIA

Se disponen las frecuencias en una tabla de doble entrada. Recibe el nombre de tabla de frecuencias o **tabla de contingencia**.

Ejemplo: Los 20 alumnos de 1º de Bachillerato A se clasifican por sexo y color de pelo de acuerdo con los datos de la siguiente tabla de doble entrada, en la que también hemos reflejado los porcentajes:

		Pelo				
		Moreno	Castaño	Rubio	f_i	%
Sexo	Chica	5	2	4	11	55
	Chico	4	5		9	45
	f_j	9	7	4	Σ=20	
	%	45	35	20		Σ=100

Se pueden construir diagramas de barras de frecuencias absolutas bidimensionales agrupadas por una u otra variable:



Las frecuencias absolutas marginales divididas por el total de observaciones, N, nos dan las frecuencias relativas marginales, hi y hj»:

		Pelo			
		Moreno	Castaño	Rubio	h_i
Sexo	Chica	0,25	0,1	0,2	0,55
	Chico	0,2	0,25		0,45
	h_j	0,45	0,35	0,2	Σ=1

Vemos que, por ejemplo, el 20% de la clase son chicos morenos.

Frecuencias condicionadas:

Supongamos que nos planteamos preguntas de este estilo: ¿Qué % de alumnos de pelo moreno son chicas? ¿Qué % de chicas tienen el pelo moreno?, etc.

En primer lugar, si para cada color de pelo (TOTALES) calculamos las frecuencias relativas, se obtienen las frecuencias relativas condicionadas al color del pelo:

	Moreno	Castaño	Rubio
Chica	5/9=0,56	2/7=0,29	4/4=1
Chico	4/9=0,44	5/7=0,71	
	Σ=1	Σ=1	Σ=1

TOTALES

Por ejemplo, del total de alumnos castaños, el 29% son chicas y el 71% chicos.

Pero también podemos hallar, para cada sexo, las frecuencias relativas, obteniéndose así las frecuencias relativas condicionadas al sexo:

TOTALES

	Moreno	Castaño	Rubio	
Chica	5/11=0,46	2/11=0,18	4/11=0,36	Σ=1
Chico	4/9=0,44	5/9=0,56		Σ=1

Del total de chicos el 44% son morenos y el 56% castaños. No hay rubios

Las frecuencias relativas condicionadas nos permiten conocer el % de los valores de una variable condicionada a cada valor de la otra variable.

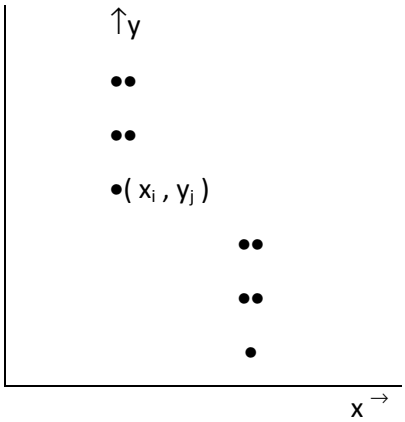
Parámetros de la v. e. bidimensional

Considerando las distribuciones marginales, como son unidimensionales es posible calcular los siguientes parámetros:

	Variable X	Variable Y
Media marginal	$\overline{x} = \frac{\sum_{i=1}^{n} x_i f_i}{N}$	$\overline{y} = \frac{\sum_{i=1}^{n} y_i f_i}{N}$
Varianza marginal	$s_x^{\,2} = \frac{\sum_{i=1}^{n} f_i x_i^{\,2}}{N} - \overline{x}^2$	$s_y^{\,2} = \frac{\sum_{i=1}^{n} f_i y_i^{\,2}}{N} - \overline{y}^2$

NUBE DE PUNTOS:

Los pares de valores observados (x_i , y_j) se pueden representar en unos ejes coordenados:

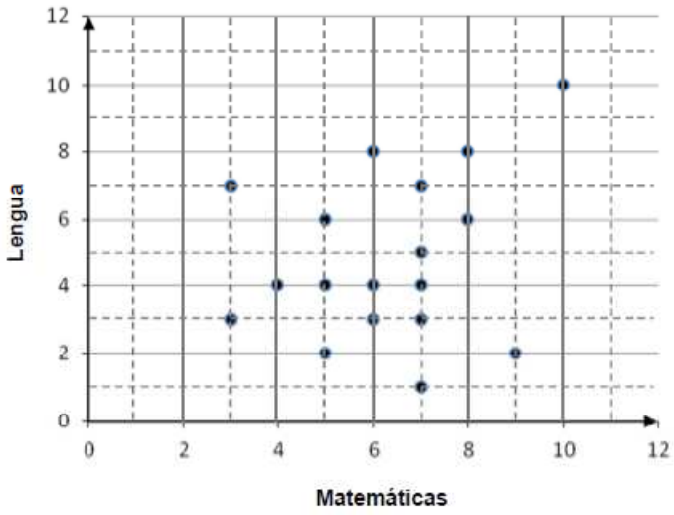


El conjunto de puntos que resulta se llama diagrama de dispersión o **nube de puntos** de la distribución.

La nube de puntos nos permite apreciar si hay una mayor o menor relación o dependencia entre las variables.

El proceso inverso también es posible, es decir, a partir de una nube de puntos podemos obtener la distribución:

Ejemplo: Las calificaciones en Matemáticas y Lengua de 18 alumnos son las representadas por la siguiente nube de puntos:



A partir de esta nube es inmediato deducir la distribución:

Matemáticas	3	3	4	5	5	5	6	6	6	7	7	7	7	7	8	8	9	10
Lengua	3	7	4	2	4	6	3	4	8	1	3	4	5	7	6	8	2	10

CORRELACIÓN:

Nuestro objetivo es determinar si existe relación entre dos variables y medir el sentido y la intensidad de esa relación. Cuando exista relación estadística diremos que existe **correlación** entre ellas; esto permitirá estimar una variable a partir de la otra.

La **correlación es lineal o curvilínea** según que el diagrama de puntos se condense en torno a una línea recta o a una curva.

La **correlación es positiva o directa** cuando a medida que crece una variable, la otra también crece.

La **correlación es negativa o inversa** cuando a medida que crece una variable, la otra decrece.

La **correlación es nula** cuando no existe ninguna relación entre ambas variables; en este caso los puntos están esparcidos al azar, sin formar ninguna línea.

La **correlación es de tipo funcional** si existe una función que satisface todos los valores de la distribución.

LA COVARIANZA: Para analizar la correlación con precisión necesitamos un nuevo parámetro estadístico: la covarianza es la media aritmética de los productos de las desviaciones de cada variable respecto de sus medias respectivas:

$$s_{xy} = \frac{\sum_{i=1}^{n} f_{ij}\,(x_i - \overline{x})(y_i - \overline{y})}{N} = \frac{\sum_{i=1}^{n} f_{ij} x_i y_i}{N} - \overline{x}\,\overline{y}$$

Ejemplo: Dada la distribución bidimensional:

X	1	2	1	2	3	2	2	2	3	1
Y	3	5	2	3	5	4	3	5	5	3

la tabla correspondiente es:

X / Y	1	2	3	Frec. absolutas marginales de Y
2	1			1
3	2	2		4
4		1		1
5		2	2	4
Frec. absolutas marginales de X	3	5	2	N=10

Comprueba que las medias marginales, las varianzas y la covarianza son:

$$\overline{x} = 1,9;\ \overline{y} = 3,8;\ s_x{}^2 = 0,49;\ s_y{}^2 = 1,16$$

$$s_{xy} = 0,58$$

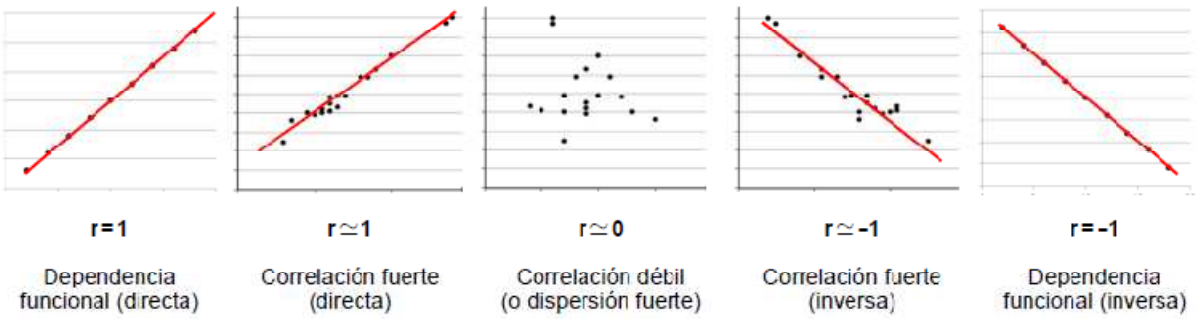
La covarianza no es un parámetro adecuado para mirar la intensidad de la correlación, pues se ve afectado por las escalas de medida de X e Y.

Para subsanar esas deficiencias, se define el coeficiente de correlación lineal o de Pearson:

$$r = \frac{s_{xy}}{s_x s_y}$$

Propiedades de *r* coeficiente de regresión

- r varía entre -1 y 1: $-1 \le r \le 1$
- Si ***r*** es próximo a -1 o a +1 significa que hay correlación lineal fuerte.
- Si ***r*** es próximo a 0 significa que la correlación lineal, si la hay, es débil.
- Si ***r*** > 0: correlación directa. Al aumentar una variable cabe esperar aumento de la otra
- Si ***r*** < 0: correlación inversa. Al aumentar una variable disminuye la otra.
- Si ***r*** =-1 o ***r*** = 1 quiere decir que la correlación lineal es perfecta para los datos analizados. En este caso, los puntos de la nube caen exactamente en la recta.



REGRESIÓN LINEAL:

Si entre dos variables existe una fuerte correlación, el diagrama de puntos se condensa en torno a una recta. Sea X la variable independiente e Y la variable dependiente de X, entonces el problema consiste en encontrar la ecuación de la recta que mejor se ajuste a la nube de puntos.

La ecuación buscada se obtiene mediante el método de los mínimos cuadrados y es:

$$y - \overline{y} = \frac{s_{xy}}{s_x^{\,2}}(x - \overline{x})$$

La ecuación de la recta de Y sobre X es:

$$x - \overline{x} = \frac{s_{xy}}{s_y^{\,2}}(y - \overline{y})$$

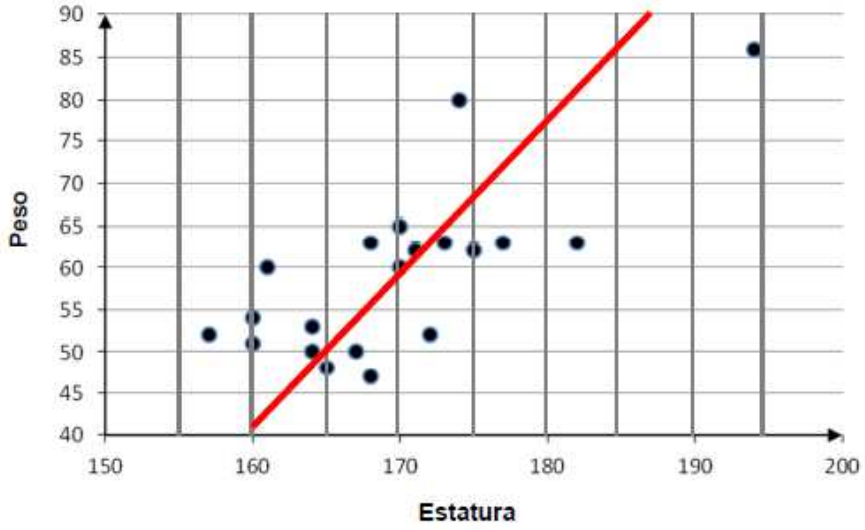
La fiabilidad qué podemos darle a calcular una variable a partir de la otra será tanto mejor cuanto mejor sea el coeficiente de correlación lineal en valor absoluto.

Ejemplo:

La estatura y el peso de 20 alumnos de 1º de Bachillerato son:

(164,53)	(175,62)	(165,48)	(170,60)	(168,47)	(157,52)	(167,50)
(172,52)	(177,63)	(160,54)	(168,63)	(160,51)	(164,50)	(174,80)
(170,65)	(182,63)	(161,60)	(171,62)	(173,63)	(194,86)	

Dibujamos la nube de puntos:



Nuestro objetivo es obtener la recta de la figura, es decir, la que mejor se ajusta a la nube de puntos.

Estatura (cm) x_i	Peso (kg) y_i	$x_i y_i$	x_i^2	y_i^2
164	53	8692	26896	2809
175	62	10850	30625	3844
165	48	7920	27225	2304
170	60	10200	28900	3600
168	47	7896	28224	2209
157	52	8164	24649	2704
167	50	8350	27889	2500
172	52	8944	29584	2704
177	63	11151	31329	3969
160	54	8640	25600	2916
168	63	10584	28224	3969
160	51	8160	25600	2601
164	50	8200	26896	2500
174	80	13920	30276	6400
170	65	11050	28900	4225
182	63	11466	33124	3969
161	60	9660	25921	3600
171	62	10602	29241	3844
173	63	10899	29929	3969
194	86	16684	37636	7396
Σx_i=3392 cm	Σy_i=1184 kg	$\Sigma x_i y_i$=202032 cm·kg	Σx_i^2=576668 cm^2	Σy_i^2=72032 kg^2

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{N} = \frac{3392}{20} = 169,6 \text{ cm}$$

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{N} = \frac{1184}{20} = 59,2 \text{ kg}$$

$$s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{N} - \overline{x}\,\overline{y} = \frac{202032}{20} - 169,6 \cdot 59,2 = 61,28 \text{ cm} \cdot \text{kg}$$

$$s_x{}^2 = \frac{\sum_{i=1}^{n} x_i{}^2}{N} - \overline{x}^2 = \frac{576668}{20} - 169,6^2 = 69,24 \text{ cm}^2$$

Con estos parámetros, la recta de regresión (de y sobre x) será:

$$y - 59,2 = \frac{61,28}{69,22}(x - 169,6) \Rightarrow \boxed{y = 0,89x - 91,74}$$ que es la recta del gráfico.

Utilidad de la recta de regresión: Con ella podemos hacer estimaciones. Por ejemplo, un alumno que mida 180 cm pesará:

y = 0,89·180 − 91,74 = 68,5kg

La validez de la estimación estará en función del grado de correlación de ambas variables:

$$s_y{}^2 = \frac{\sum_{i=1}^{n} y_i{}^2}{N} - \overline{y}^2 = \frac{72032}{20} - 59,2^2 = 96,96 \text{ kg}^2 \Rightarrow s_y \cong 9,85 \text{ kg}$$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{61,28}{8,32 \cdot 9,85} \cong 0,75$$

es decir, las estimaciones a partir de la recta de regresión tienen una fiabilidad del 75%.

En resumen, cuanto más se aproxime *r* a 1 o -1 mayor validez tendrá toda estimación obtenida de la recta de regresión.

EJERCICIOS Y PROBLEMAS

1. Completa la siguiente tabla:

Y \ X	A	B	C	Total
a	2		1	4
b		2	2	5
c			0	7
Total	7	6		

a) ¿Qué porcentaje de datos presentan la característica B en la variable unidimensional X?
b) ¿Qué porcentaje de datos presentan la característica c en la variable unidimensional Y?
c) Porcentaje de datos que presenta la característica (B, c) en la variable bidimensional (X, Y).

2. Los ingresos mensuales, en €, en cuatro sucursales de loterías y apuestas, han sido los siguientes:

	Lotería Nacional	Primitiva	Apuestas por Internet	Bonoloto	Euromillones	Gordo de la primitiva
Centro	10529	5139	1288	1053	568	311
Aeropuerto	3179	1259	314	218	200	97
Estación de tren	2115	1495	376	229	135	106
Barrio del Pilar	7386	4875	1015	950	417	309

a) ¿Qué % supone la Lotería Nacional por sucursal?
b) ¿Qué % se ha jugado en la sucursal del aeropuerto en cada tipo de apuestas?

3. Observa las siguientes variables bidimensionales:

A)

N.º de cigarrillos consumidos al día	3	6	8	20	25
Índice de mortalidad	0,2	0,4	0,5	1,2	1,7

B)

N.º de horas de estudio	1	2	3	4	5
N.º de horas de televisión	5	4	3	3	1

En cada caso:

a) Representa la nube de puntos.
b) Indica el tipo de correlación.

4. Se pregunta a los 20 alumnos de una clase sobre la talla de su calzado y el número de hermanos (contándose él), obteniéndose los siguientes resultados:

(38,4) (41,2) (38,2) (37,2) (37,3) (36,2) (37,3) (38,4) (42,5) (38,3)(40,2) (38,3) (42,3) (46,4) (38,2) (39,3) (43,1) (48,1) (40,2) (36,2)

a) Obtener la nube de puntos.

b) ¿Si comparáramos talla de calzado y estatura la nube sería tan dispersa?

5. Para cada una de las distribuciones bidimensionales que siguen dibujar la nube de puntos y construir una tabla apropiada para hallar el coeficiente de correlación y, si procede, la recta de regresión de y sobre x:

x	2	3	4	5	6
y	4	2	5	4	6

Gastos en publicidad (miles €)	1	2	3	4	5	6	7	8
Ventas (miles €)	15	16	14	17	20	18	18	19

Clasificación	1	2	3	4	5	6	7	8	9	10
Ganados (G)	27	24	20	18	19	16	18	15	15	15
Empatados (E)	12	15	11	11	9	13	9	14	12	12
Perdidos (P)	5	5	13	15	16	15	17	15	17	17
Puntos (PT)	66	63	51	47	47	45	45	44	42	42

para las variables (G,P), (G,PT) y (G,E).

Nº horas estudio (E)	**Nº horas TV** (T)	**Nº suspensos** (S)
4	2	1
5	1,5	0
4	2,5	3
2,5	4	2
6	0,5	0
0,5	5,5	6
1	5	2
2	4	5
3	2,5	3
4,5	1,5	2
3	3,5	4
1,5	5	3
3,5	2,5	4
5,5	3,5	1
2,5	3,5	3

para las variables (E,T), (E,S) y (T,S)

6. La evolución de los asuntos que entraron por la vía civil y penal en una determinada localidad se expresa en la siguiente tabla:

¿Cuántos asuntos cabe esperar en el juzgado penal un año que se hayan recibido 300 en el civil?

Años	Juzgado civil	Juzgado penal
2010	134	87
2011	171	107
2012	196	135
2013	199	143
2014	216	168

7. Se determina la pérdida de efectividad de un determinado preparado farmacéutico con el tiempo y se obtiene el siguiente resultado:

Tiempo (meses)	1	2	3	4	5
Actividad (en %)	90	75	42	30	21

   a) ¿Qué % de actividad quedará a los 6 meses?

   b) ¿En cuánto tiempo la actividad se reduce al 50 %?

8. La evolución del IPC y la tasa de inflación en los primeros nueve meses de un año viene reflejada en la siguiente tabla:

	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP
IPC	0,7	1,1	1,7	2	1,9	1,9	2,9	2,9	3,8
Inflación	6	6	6,3	6,2	5,8	4,9	4,9	4,5	4,4

   ¿Qué tasa de inflación es razonable esperar en un mes en el que el IPC sea 4,5?

9. Las tallas y los pesos de 10 personas vienen recogidos en la siguiente tabla:

talla (cm**)**	160	165	170	180	185	190	192	175	182	172
pesos (kg**)**	58	61	65	73	80	85	83	68	74	67

   Estimar el peso medio de una persona que mida 168cm.

10. El número de licencias de caza, en miles, y el número de votantes a un determinado partido en 6 comunidades autónomas, en decenas de miles, está expresado en la siguiente tabla:

Nº de licencias (**X**)	103	26	3	7	26	5
Nº de votantes (**Y**)	206	26	27	14	24	12

   Determinar:

   a) Media y varianza de las variables X e Y.

   b) Coeficiente de correlación, interpretando su valor.

   c) En el caso de que exista correlación: si en una determinada comunidad existen 50 decenas de millar de votantes, ¿cuántas licencias de caza, en miles, se puede estimar que existen.

11. Se ha preguntado a los alumnos de un centro el número de horas de estudio diario, X, y el número de asignaturas aprobadas al final del curso, Y. A la nube de puntos resultado de la encuesta se ha ajustado la recta de regresión $y = 3,8x + 0,2$.

   a) Para aprobar 4 asignaturas, ¿cuánto tiempo de estudio deberían emplear?

   b) Y para superar las 11 asignaturas, es decir, todas, ¿cuál sería la recomendación de horas de estudio?

12. Una variable bidimensional (X, Y) tiene de coeficiente de correlación r = 0,78, y las medias de las distribuciones marginales son $\overline{x} = 2;\ \overline{y} = 9$. Razona cuál de las siguientes rectas se ajusta más a dicha variable.

   $y = -3x + 12;\quad y = 1,5x + 6;\quad y = -2,5x + 14$

AUTOEVALUACIÓN 4

1. En la variable bidimensional del ejercicio 3A.

   Calcula las medias y las desviaciones típicas de las variables X e Y. Calcula la covarianza de la variable (X, Y).

   $$\{\overline{x}=12,40;\overline{y}=0,8;s_x=8,5463;s_y=0,5621;s_{xy}=4,78\}$$

2. Calcula el coeficiente de correlación correspondiente a la recta de regresión calculada para las temperaturas máxima y mínima (en ºC) en una determinada localidad a lo largo del año. Representa la nube de puntos.

	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
Máxima	16	17	19	19	21	25	28	27	23	21	18	17
Mínima	7	10	12	13	15	19	21	22	18	15	12	9

   $$\{r=0,98\}$$

3. La evolución del SIDA en España desde su aparición viene expresada en la siguiente tabla:

Año	Nº de casos	Nº de muertos
1981	1	1
1982	6	6
1983	17	17
1984	46	43
1985	164	128
1986	437	274
1987	624	322

   En un año que se vayan a presentar 700 casos, ¿cuántos muertos se prevén? ¿Tendría sentido hacer una estimación para, por ejemplo, 1000 casos?

   {y=0,53x+14,21}

4. Una asociación dedicada a la protección de la infancia decide estudiar la relación entre la mortalidad infantil en cada país y el número de camas de hospitales por cada mil habitantes.

   Datos

x	50	100	70	60	120	180	200	250	30	90
y	5	2	2,5	3,75	4	1	1,25	0,75	7	3

   Donde **x** es el nº de camas por mil habitantes e **y** el tanto por ciento de mortalidad.

   Se pide calcular las rectas de regresión y el coeficiente de correlación lineal.

   ¿Si se dispusiese de 175 camas por mil habitantes que tanto por ciento de mortalidad cabría esperar? ¿La estimación es fiable? Razona la respuesta. Las rectas de regresión serán por tanto:

   $$\left\{\begin{array}{l} y-3,025=-0,022(x-115);\quad x-115=-30,205(y-3,025) \\ r=-0,824\rightarrow \text{inversa alta};\quad 1,678 \text{ fiable} \end{array}\right\}$$

5. El número de separaciones matrimoniales y divorcios en una determinada provincia en el período 2010-2014 se distribuye según la siguiente tabla:

Año	Separaciones	Divorcios
2010	2357	4000
2011	2586	3428
2012	2689	2903
2013	2821	2711
2014	3073	2910

¿Cuántas separaciones se prevé que se produzcan en un determinado año sabiendo que hubo 3600 divorcios?

{ y=-1,63x+7605,97; 2457 separaciones}

6. En una variable bidimensional (X, Y), su coeficiente de correlación lineal es 0,48 y la pendiente de su recta de regresión es 1,34.
Sabiendo que la suma de las desviaciones típicas de X e Y es 7,33, calcula cada una de ellas y la covarianza de la variable bidimensional.

$$\{s_x = 1,93;\ s_y = 5,4;\ s_{xy} = 4,99\}$$

7. Una variable bidimensional viene dada por la siguiente tabla:

X	2	3	5	a
Y	1	25	b	3

Sabiendo que $s_{xy} = 1$ y $s_x^2 = 3$ y que a es el valor máximo de la variable X, calcula a y b.

$$\{a = 6,46; b = 30,46\}$$

8. Se ha medido experimentalmente el área de distintos triángulos equiláteros de lados 1, 2, 3 decímetros, sucesivamente, y se han obtenido los siguientes resultados.

$(Lado)^2$	2	4	9	16	25
Área	0,42	1,65	3,7	6,5	10,2

a) Calcula el coeficiente de correlación lineal entre el cuadrado del lado y el área del triángulo. ¿Qué tipo de correlación existe?

b) ¿Debería haber una relación funcional? ¿A qué se debe que la relación no llegue a ser funcional?

$$\{r = 0,99 \rightarrow \text{positiva y fuerte; Sí, pero el coeficiente no es 1 debido al redondeo}\}$$