Tema 7: Estadística y probabilidad

* En este tema revisaremos:
  1. Representación de datos e interpretación de gráficas.
  2. Estadística descriptiva.
  3. Probabilidad elemental.

* Dos ideas para empezar:
  * Vivimos rodeados de estadísticas, no siempre bien usadas y/o entendidas.

    "Hay mentiras, grandes mentiras y estadísticas"
    Mark Twain

    Una TED talk (informal): https://goo.gl/vo1zQV

  * La probabilidad puede ser, a veces, muy poco intuitiva.

Representaciones de datos

* Cuatro observaciones sobre el tema:

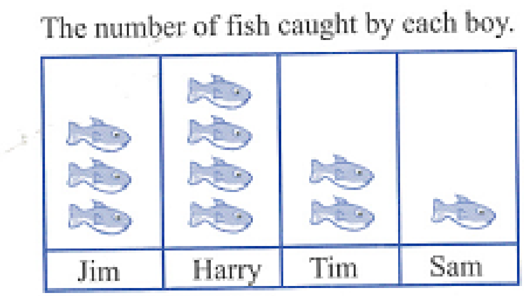
(a) la dificultad conceptual (del nivel en que se trata el tema en primaria-secundaria) no es mucha, hay que medir bien el tiempo que se le dedica en clase.

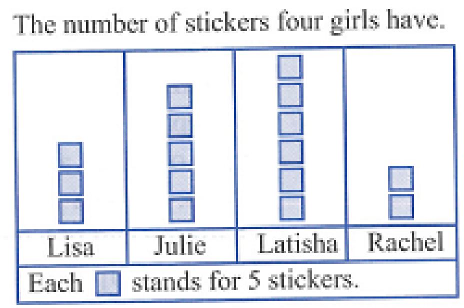
(b) con una presentación adecuada, los diagramas de barras son la mejor preparación para las gráficas de funciones.

(c) primero, interpretación; después, construcción.

(d) es importante elegir bien las actividades de recogida de datos.

Tipos de representaciones: secuencia curricular





Primer ciclo

Pictograma

Pictograma abstracto

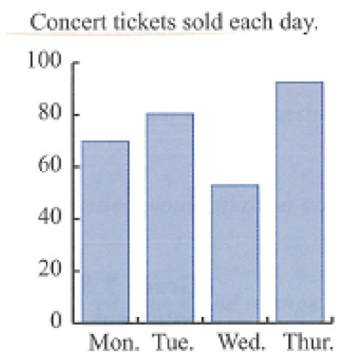
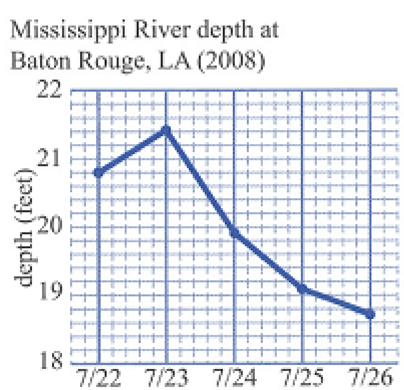


Diagrama de barras

Segundo ciclo

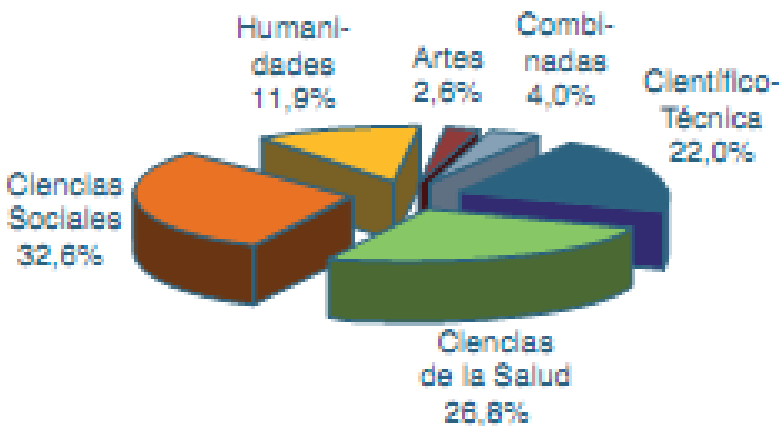


Gráfica

Tercer ciclo

Diagrama de sectores

* El área de cada sector circular representa los datos.

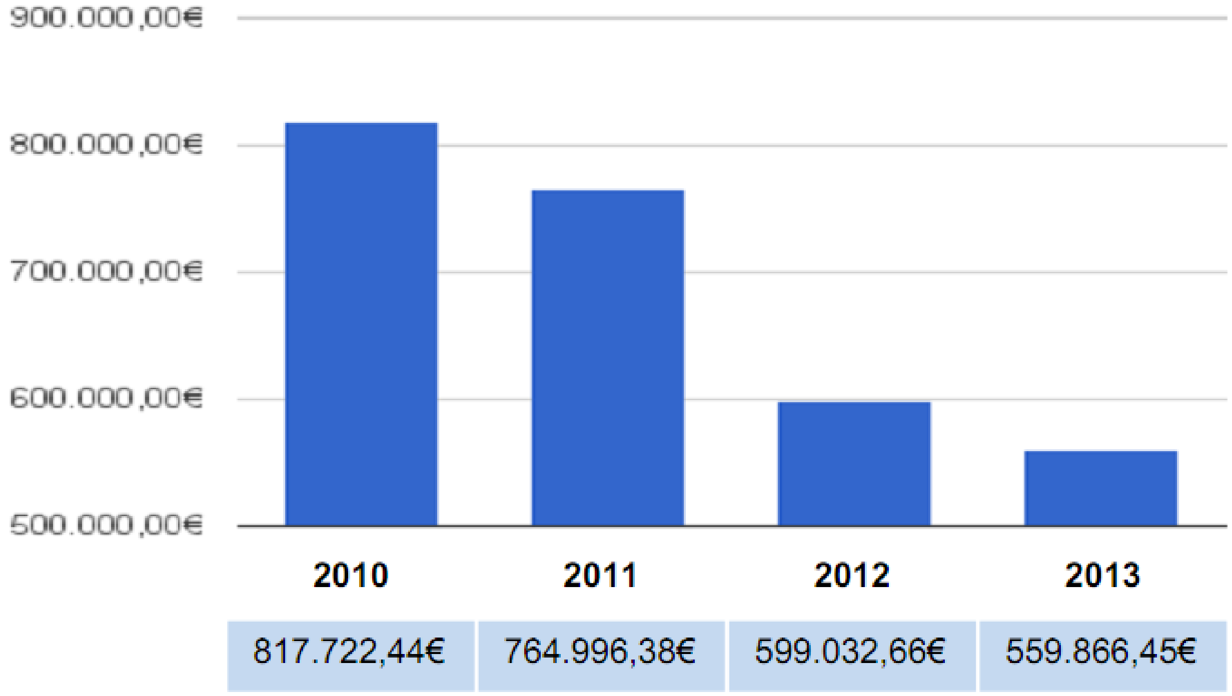


Aprobados en selectividad (año 2009)

* Observaciones sobre las actividades de recogida de datos:
  (a) mejor si son cortas, y hay que evitar que sean tediosas.
  (b) lo ideal es que los datos aporten algo.

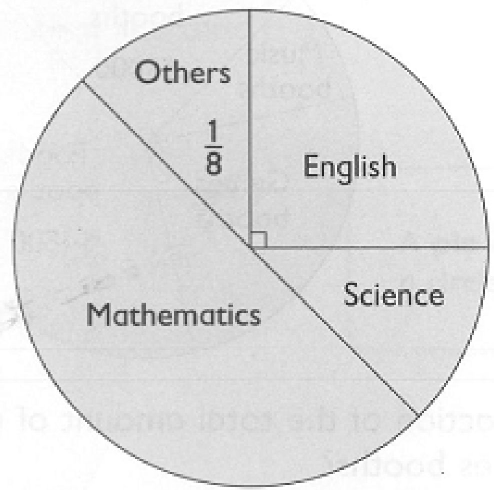
Presentación de datos

* A veces puede ser engañosa.



∗ Un ejemplo de 6º de Singapur.
La representación de datos puede ser una buena ocasión para repasar otros conceptos.

4. The students in a school were asked to name their favorite subject. The pie chart represents their choices.



(a) What fraction of the students liked Mathematics?
(b) What percentage of the students liked English?
(c) What fraction of the students liked Science?
(d) If 1200 students liked Mathematics, how many students liked English?

Workbook Exercise 18

Estadística descriptiva

* Tenemos un conjunto de datos numéricos. Buscamos medidas que los describan.

* En ocasiones, los valores son demasiados, y se agrupan en intervalos.

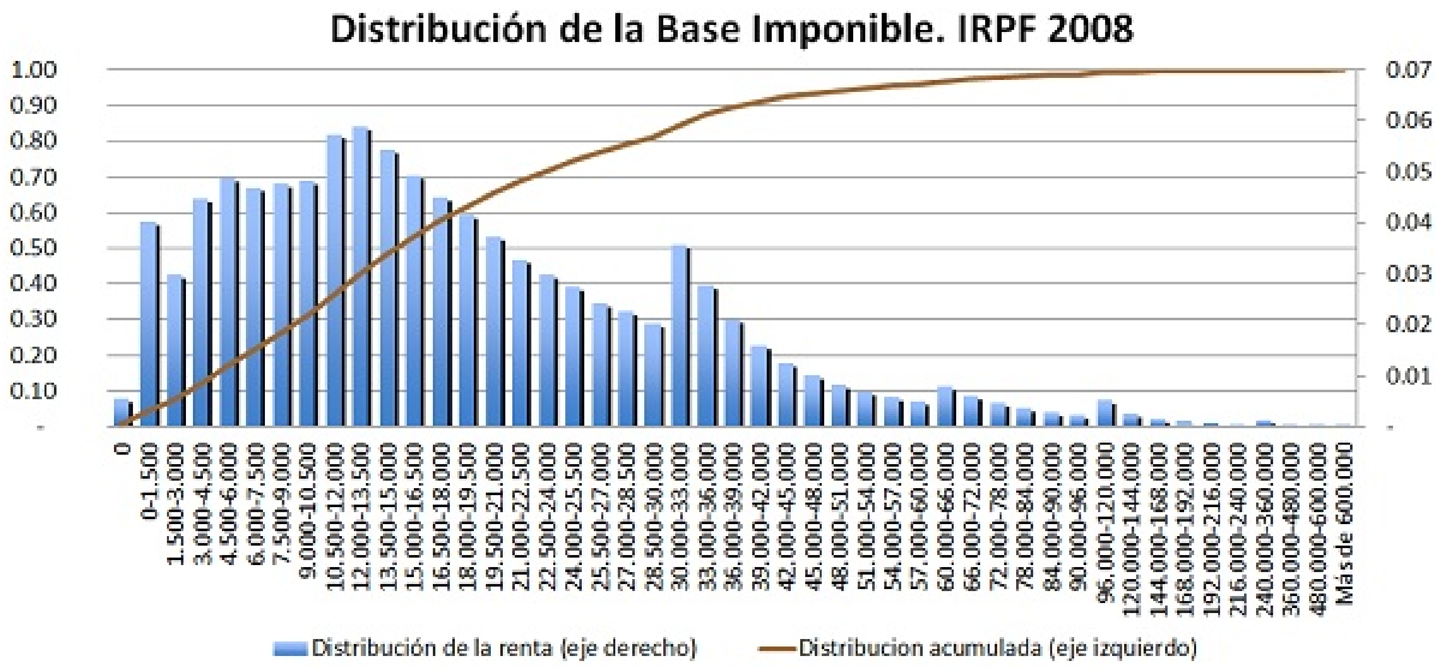
* Por ejemplo, los datos de salario no van a decir cuánta gente gana 15677 euros, sino para cuánta gente su salario esta entre 15000 y 16000 euros (u otro intervalo parecido).

* En estadística, estos intervalos se llaman clases. No vamos a entrar en cómo definirlos. (Depende del problema, y puede no ser sencillo).

* Si se trabaja con clases, para los cálculos se trabaja con un número (normalmente el punto medio), que se llama marca de clase.

Ejemplo de un error

* Una mala elección de los intervalos puede dar resultados engañosos.



http://tinyurl.com/7gy8ucx

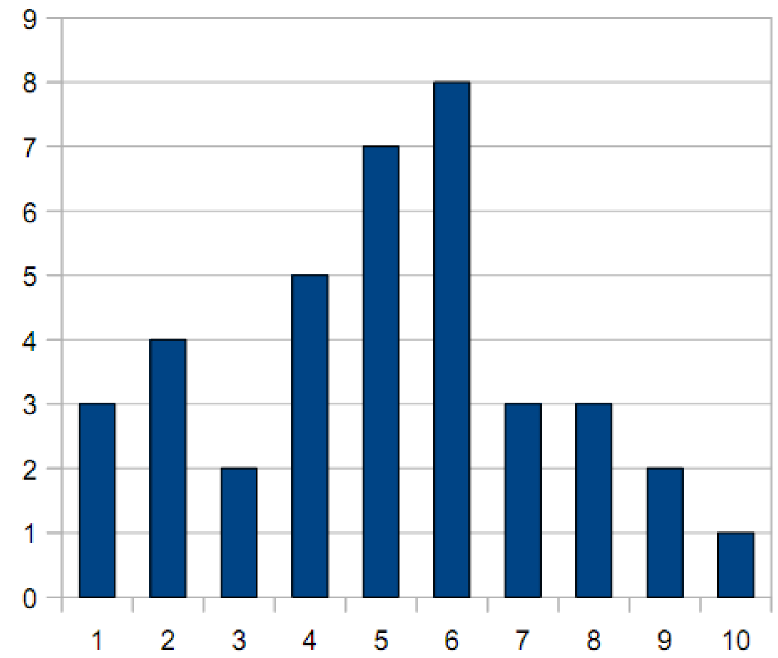
Datos. Frecuencia absoluta, frecuencia relativa.

∗ Supongamos que tenemos una serie de 38 datos con este aspecto (números del 1 al 10): 1, 5, 2, 7, 4,..., 3, 9, 2.

∗ Podemos organizar los datos en una tabla como esta:

dato (o intervalo)	frecuencia absoluta	frecuencia relativa
x_i	h_i	f_i
1	3	0,08
2	4	0,11
3	2	0,05
4	5	0,13
5	7	0,18
6	8	0,21
7	3	0,08
8	3	0,08
9	2	0,05
10	1	0,03
	38	1
	n	

∗ Y representarlos en un diagrama de barras.



Medidas de posición: la media aritmética

* La media aritmética:

(a) Si tenemos la lista completa, $\{x_1, x_2, \ldots, x_n\}$,

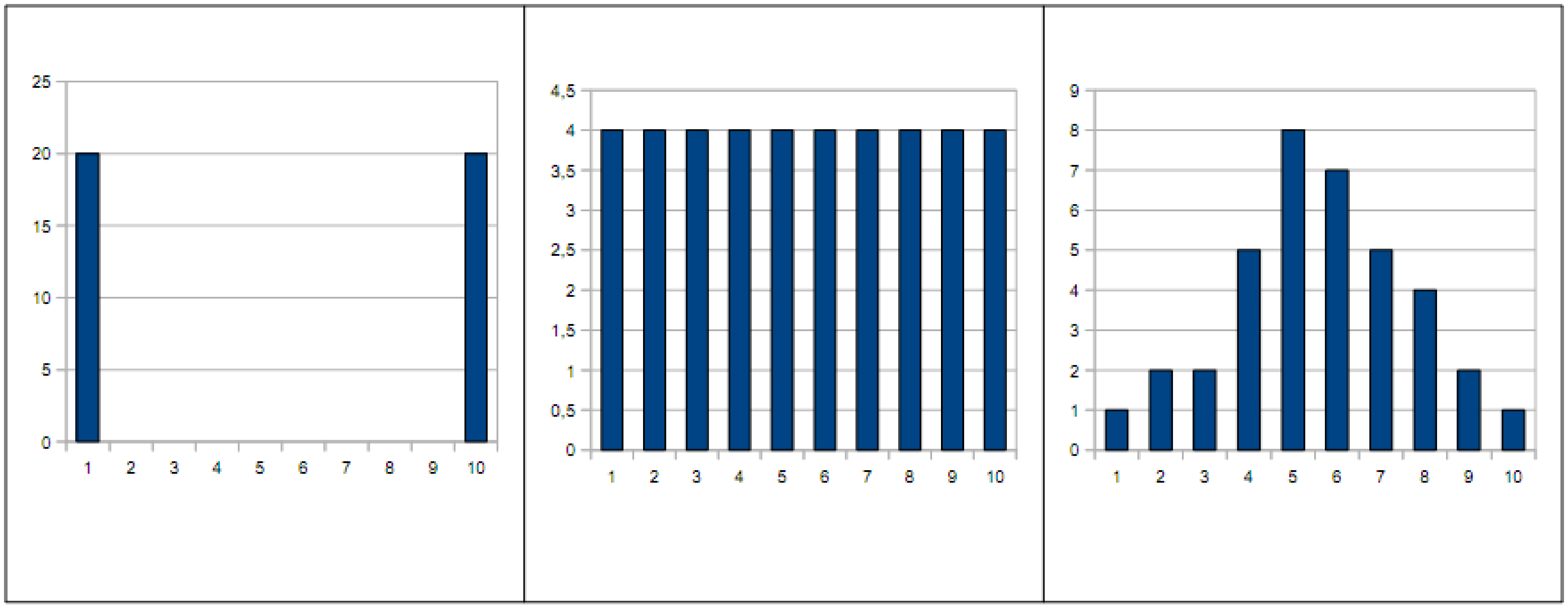
$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

(b) En la práctica, lo normal es tener cada dato x_i con su frecuencia absoluta h_i (como en la transparencia anterior). Entonces

$$\bar{x} = \frac{h_1 x_1 + h_2 x_2 + \cdots + h_k x_k}{h_1 + h_2 + \cdots + h_k}$$

Medidas de posición: la media aritmética

* La media aritmética da poca información sobre los datos.



Medidas de posición: moda, mediana

* La moda es el valor que más se da (el valor con mayor frecuencia absoluta).

* La mediana es un valor que es mayor o igual que la mitad de los valores y menor o igual que la mitad de los valores.

* Para calcular la mediana es útil ordenar los valores

$$x_1 < x_2 < \cdots < x_k$$

y considerar la frecuencia absoluta acumulada:

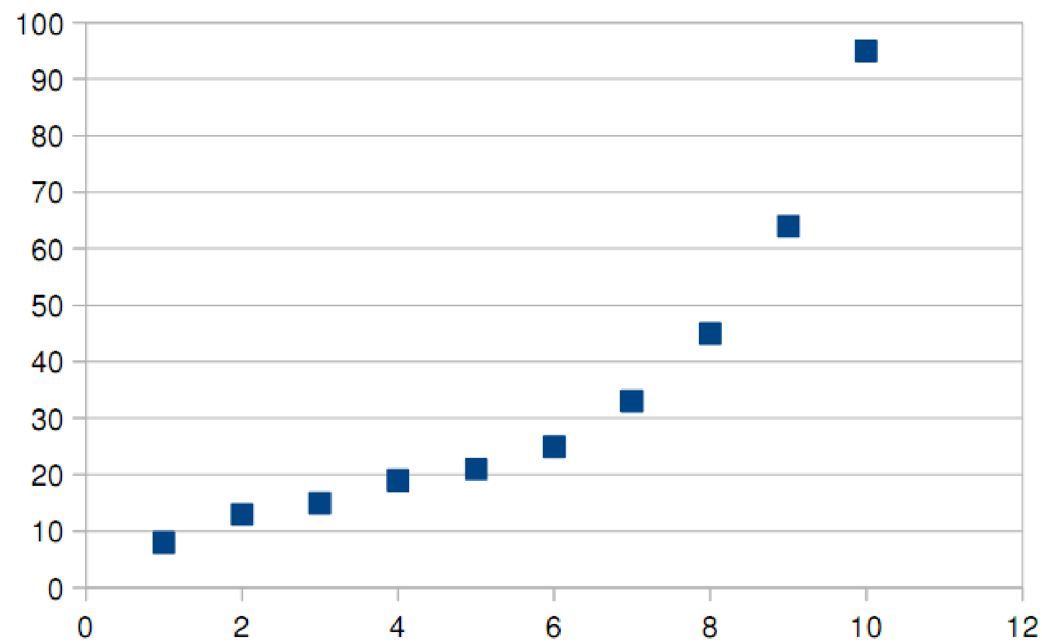
$$H_k = h_1 + h_2 + \cdots + h_k$$

* Ejemplo: calcular la mediana de la distribución de la tabla de la transparencia 6.

Comparación media-mediana

Ejemplo

Datos	
	8
	13
	15
	19
	21
	25
	33
	45
	64
	95
Media	**51,5**



Medidas de posición: cuartiles, percentiles

* Consideramos un número α entre 0 y 1.
  El cuantil de orden α, C_α, es el menor valor que es mayor o igual que una proporción α de los valores.

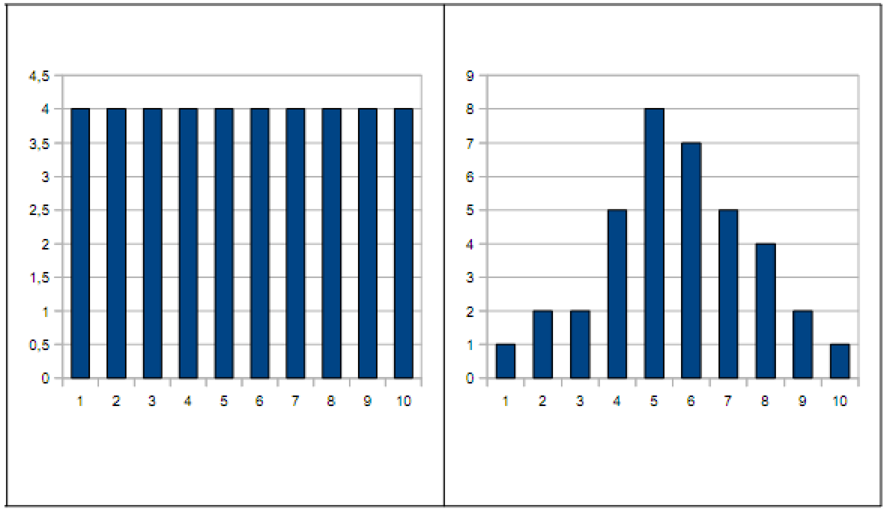
* Ejemplo: $\alpha = \frac{1}{4}$, distribución de las tablas de las transparencias 6 y 11.

$\alpha = \frac{1}{4}$, $\alpha = \frac{3}{4}$ cuartiles

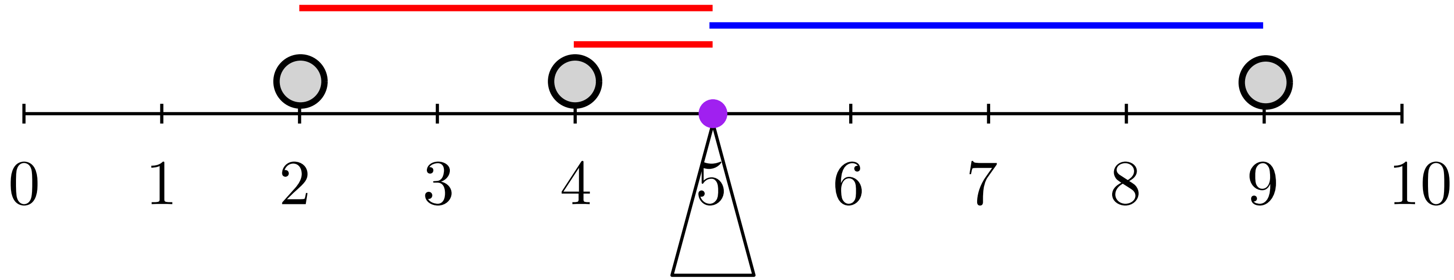
$\alpha = \frac{i}{100}$ percentiles $(i = 1, 2, \ldots, 99, 100)$

Medidas de dispersión

* ¿Están los datos concentrados cerca de la media?



* Primer intento: medir las distancias $x_i - \bar{x}$ y sumarlas (con signo)



* La media es el punto que "equilibra" los datos.

Medidas de dispersión: varianza

* Varianza: es la media aritmética de la distribución $(x_i - \bar{x})^2$. Notación: $\mathrm{Var}(x)$

* Ejemplo: calcula la varianza de la distribución de datos de esta tabla.

	4
	5
	3
	7
	6
	4
	8
Media	**5,29**

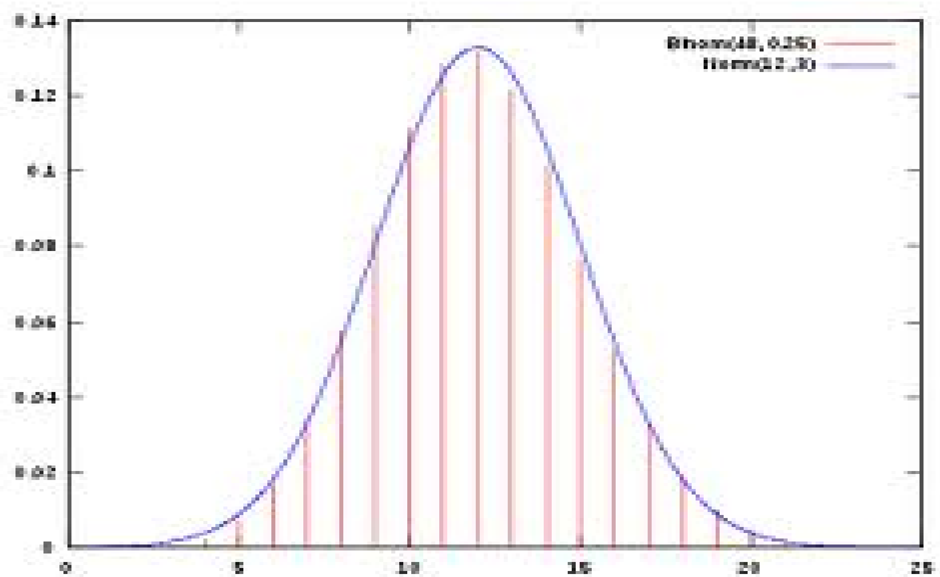
* Pregunta: ¿qué conjuntos de datos tienen varianza cero?

Medidas de dispersión: desviación típica

* En la mayoría de las ocasiones, lo que se usa es

$$\sigma(x) = \sqrt{\mathrm{Var}(x)} \qquad \text{(Desviación típica)}$$

* Una distribución muy importante:

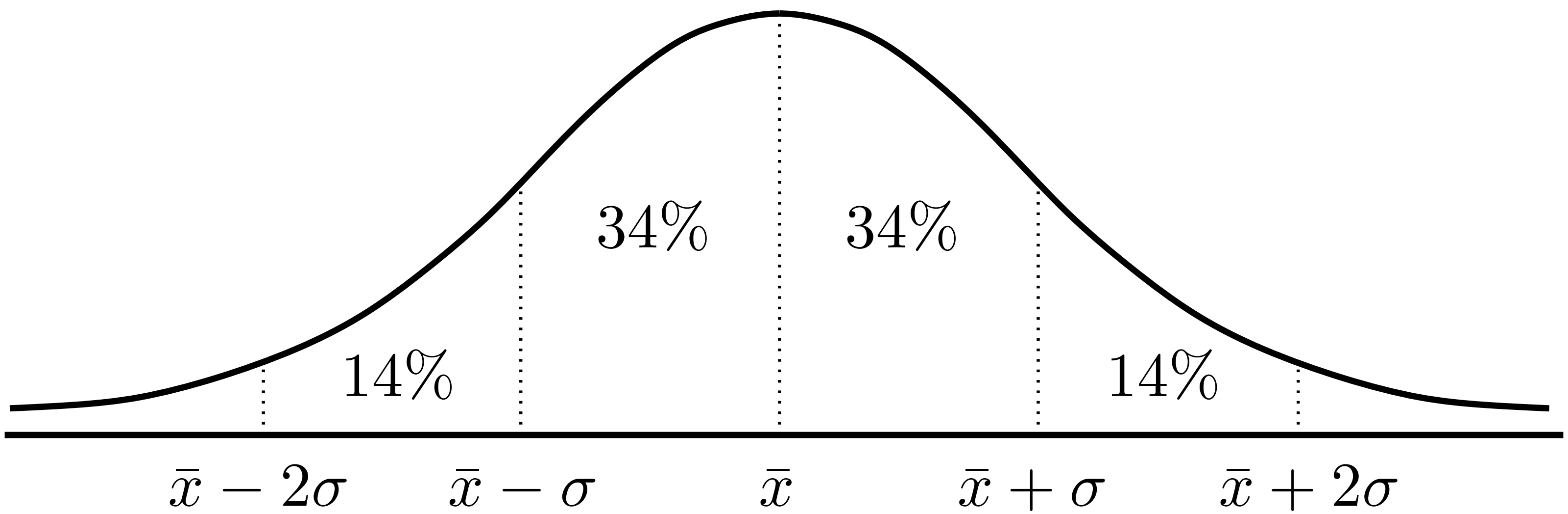


La distribución normal o campana de Gauss.

* Aparece en muchas ocasiones.

La distribución normal

* En el modelo teórico (los fenómenos se pueden parecer, pero no serán nunca iguales), una distribución normal con media $\bar{x}$ y desviación típica σ tiene este aspecto:



* Un ejemplo: Los tests de inteligencia están diseñados para que el coeficiente de inteligencia siga (se parezca a) una distribución normal de media 100 y desviación típica 15.

Una web para explorar

* National Library of Virtual Manipulatives: http://tinyurl.com/pbetcs

Introducción a la probabilidad

* A veces, la probabilidad es poco intuitiva.

(1) El problema de Monty Hall

http://tinyurl.com/bls8yvp

(2) El problema del cumpleaños.

Hay n personas en una habitación (las personas están elegidas al azar). Ahora apostamos a que hay al menos dos personas que cumplen años el mismo día.

¿Para qué valores de n apostarías a favor o en contra de este hecho?

Bastante bien explicado aquí: http://tinyurl.com/bqhqpc8

Introducción a la probabilidad

* La probabilidad se ocupa del estudio de experimentos aleatorios.
  No se puede predecir el resultado del experimento, pero sí se pueden decir cosas sobre conjuntos de experimentos.

* Ejemplos:
  a) lanzar una moneda al aire.
  b) lanzar un dado al aire.
  c) lanzar dos dados y sumar los puntos.

* El conjunto de posibles resultados se llama espacio muestral.

Un ejemplo

* Tenemos tres bolas en una urna, de colores blanco, negro y rojo. Sacamos una bola al azar, la devolvemos a la urna, y sacamos otra.
  ¿Cuál es la probabilidad de haber sacado una bola negra al menos una de las veces?

* Se pueden considerar espacios muestrales distintos.

Modelos de probabilidad

* Un modelo de probabilidad es una asignación de probabilidades a cada uno de los posibles resultados del experimento, de manera que la suma total sea 1.

  (Si damos la probabilidad en %, la suma debe ser 100).

* Una forma intuitiva de representar modelos de probabilidad es el modelo de áreas.

  El espacio muestral se representa como un rectángulo, dividido en los posibles resultados. El área de cada posible resultado es proporcional a su probabilidad.

Moneda

cara	cruz

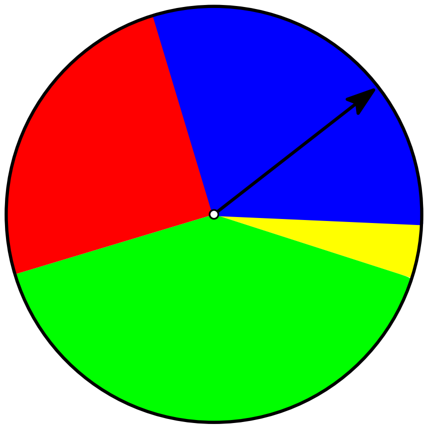
Dado

1	2	3	4	5	6

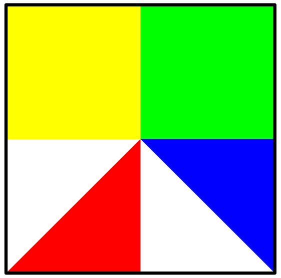
Modelo de áreas

* Es la vía para generalizar la probabilidad a experimentos con infinitos posibles resultados.

* ¿De qué depende la probabilidad de que la aguja se pare en cada una de las regiones en la ruleta de la figura?



* Elegimos un punto al azar en el cuadrado de la figura. ¿Cuál es la probabilidad de que sea rojo?



* Se trata de un modelo matemático. Cómo se podría hacer esto en la práctica, es otra cuestión ...

Sucesos. Operaciones con sucesos

* Un suceso es un conjunto de posibles resultados.
  Suceso elemental: formado por un único resultado.

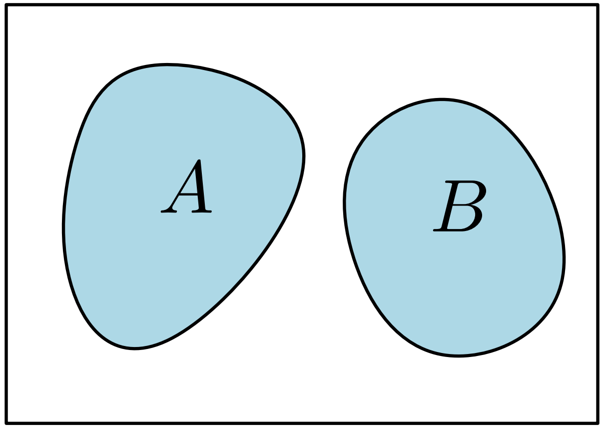
* Lanzamos un dado. Calcula las probabilidades de los sucesos:

  $A \equiv$ "sale un 6" $B \equiv$ "sale un múltiplo de 3"

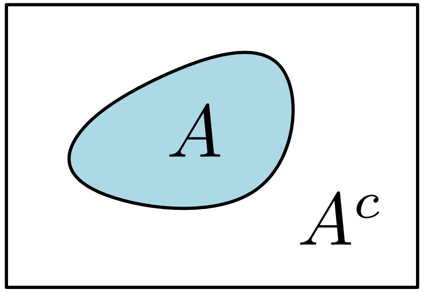
* Operaciones con sucesos:
  * A^c (complementario de A). Es el que ocurre cuando no ocurre A.
  * $A \cup B$ (A unión B). Es el que ocurre cuando ocurre alguno de los dos (pueden suceder los dos).
  * $A \cap B$ (A intersección B). Es el que ocurre cuando ocurren los dos.

Propiedades de la probabilidad

* Si $A \cap B = \emptyset$ (sucesos disjuntos),
  $P(A \cup B) = P(A) + P(B)$



* $P(A) + P(A^c) = 1$

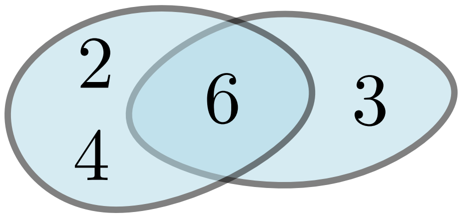


Ejercicio

- Lanzamos un dado. Considera los sucesos:

  $A \equiv$ "sale un número par"

  $B \equiv$ "sale un múltiplo de 3"

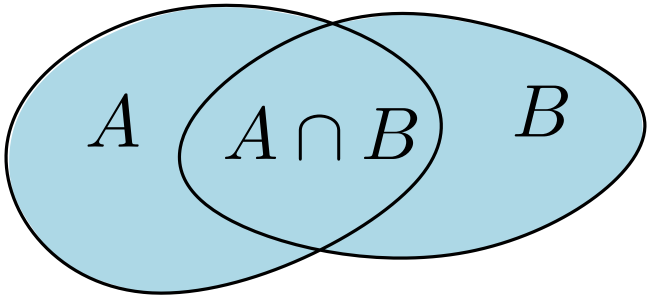


  - Determina los sucesos $A \cup B$ y $A \cap B$.
  - Calcula $P(A \cup B)$ y $P(A \cap B)$.
  - Comprueba que $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$\downarrow$

Se cumple siempre y se conoce como principio de inclusión-exclusión.

$A \cup B$



Fórmula de Laplace

* Fórmula de Laplace: Nos dice las probabilidades de los sucesos de experimentos con n posibles resultados, todos igualmente probables.

$$P(A) = \frac{\text{casos favorables}}{\text{casos posibles}}$$

* Ejemplo: tenemos una bolsa con 2 bolas blancas y 3 bolas negras. Extraemos una bola al azar. ¿Cuál es la probabilidad de que sea blanca?

* De la misma bolsa ahora extraemos dos bolas (sin reemplazamiento).
¿Cuál es la probabilidad de que las dos sean negras?

* Dos opciones: diagrama de árbol y “saber contar”.

Ejercicios

* Tiramos dos dados y sumamos los resultados. ¿Cuál es la probabilidad de que la suma sea 7?

* Tiramos una moneda al aire tres veces. ¿Cuál es la probabilidad de sacar exactamente dos veces cara?

* Hay k personas en una habitacion. ¿Cuál es la probabilidad de que todas tengan cumpleaños distintos?

  Ya podríamos resolver el problema del cumpleaños, si "supiéramos contar".

$\ast$ Formas de obtener k caras al tirar n monedas: el triángulo de Pascal.

1 moneda: 1 (1 C) $\quad$ 1 (0 C)

2 monedas: 1 (2 C) $\quad$ 2 (1 C) $\quad$ 1 (0 C)

3 monedas: 1 (3 C) $\quad$ 3 (2 C) $\quad$ 3 (1 C) $\quad$ 1 (0 C)

Triángulo de Pascal

```
                  1
                1   1
              1   2   1
            1   3   3   1
          1   4   6   4   1
        1   5  10  10   5   1
      1   6  15  20  15   6   1
    1   7  21  35  35  21   7   1
  1   8  28  56  70  56  28   8   1
1   9  36  84 126 126  84  36   9   1
```

Sucesos independientes

* Se dice que los sucesos A y B son independientes si conocer información sobre uno no cambia la probabilidad del otro.

* Ejemplo: dos tiradas de un dado.

* En ocasiones la independencia puede no estar tan clara: Tiramos un dado rojo y otro azul y consideramos los sucesos

  $A \equiv$ "en el dado rojo sale un número par"

  $B \equiv$ "la suma es impar"

  ¿Son A y B sucesos independientes?

Sucesos independientes: Principio del producto

* Nos centraremos en sucesos independientes en que la independencia está clara, y la usaremos para calcular probabilidades utilizando el

  Principio del producto: Si los sucesos A y B son independientes, entonces

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

* Ejercicio: Hacemos tiradas sucesivas de un dado.
  1. ¿Cuál es la probabilidad de no sacar ningún 6 si lo tiramos 2 veces? ¿Y si lo tiramos k veces?
  2. ¿Cuál es la probabilidad de sacar al menos un 6 si lo tiramos k veces?

Ejercicios

* Ojo: Un error muy frecuente es confundir sucesos disjuntos con sucesos independientes.

  Ejemplo: Tiramos un dado rojo y otro azul y definimos los sucesos

  $A \equiv$ "sale el mismo número en los dos dados"

  $B \equiv$ "sale un 4 en el dado rojo"

  $C \equiv$ "la suma de los resultados es impar".

  Considera los sucesos por parejas y estudia si son disjuntos o independientes.

Ejercicios

* Te proponen la siguiente apuesta:
  De una baraja española de 40 cartas extraemos una carta al azar. Te dicen que la carta es una figura o es del palo de oros. ¿Apostarías a favor, o en contra?

* Hacemos dos veces el experimento de tirar dos dados y sumar los resultados.
  ¿Cuál es la probabilidad de sacar 7 al menos una vez?

Probabilidad condicionada

* El objetivo del resto del tema es intentar contestar a la siguiente pregunta:
* La prueba de una enfermedad ofrece una tasa de falsos positivos del 5% (y no hay falsos negativos). La enfermedad afecta a una de cada mil personas. Se elige una persona de forma aleatoria, se le hace la prueba y da positivo. ¿Cuál es la probabilidad de que tenga la enfermedad?

Probabilidad condicionada

* Si dos sucesos A y B no son independientes, saber algo sobre el suceso B hace cambiar las probabilidades del suceso A.
* Ejemplo: Tiramos dos dados y consideramos los sucesos
  $A \equiv$ "la suma es 8"
  $B \equiv$ "en el primer dado ha salido un 5"
* Sin información adicional, sabemos que $P(A) = \dfrac{5}{36}$.

  Si sabemos que ha ocurrido B, ¿cuál será ahora la probabilidad de que haya ocurrido A?
* La probabilidad de A, condicionada por B, que escribiremos $P(A|B)$, es la probabilidad de que ocurra A si sabemos que ha ocurrido B.

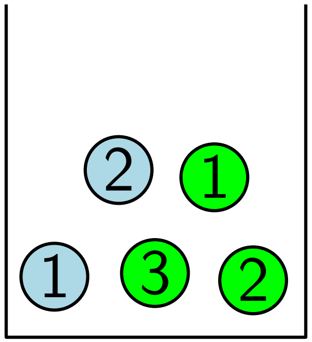
Ejemplo

* En la caja de la figura elegimos una bola al azar y consideramos los sucesos

  $A \equiv$ "la bola es un 2"

  $B \equiv$ "la bola es azul"

  Calcula $P(A)$, $P(A|B)$ y $P(A|B^c)$.



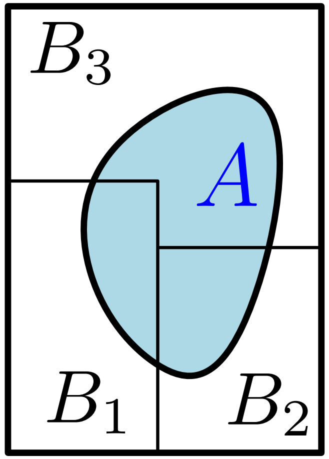
* Una fórmula para la probabilidad condicionada. Para todos los sucesos A y B se verifica que

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

* Compruébalo en el ejemplo anterior y comprueba qué dice esta fórmula sobre $P(A|B)$ cuando A y B son independientes.

Probabilidad total

* En muchas ocasiones, no sabemos calcular la probabilidad de A, pero si sabemos cómo calcularla “a trozos”.



* La fórmula de la probabilidad total dice que, en la situación de la figura,

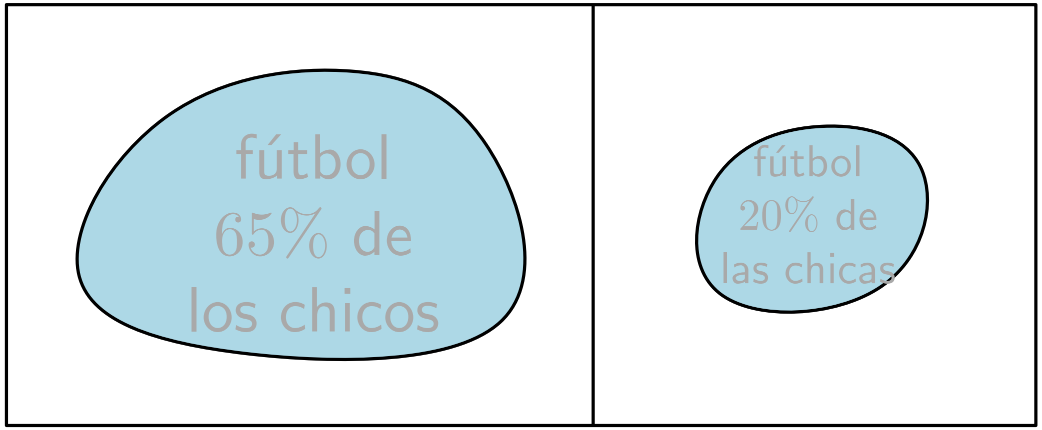
$$P(A) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3)$$

* Ejemplo: Tenemos dos cajas, una con 2 bolas blancas y 3 bolas negras, y otra con 4 bolas blancas y 1 bola negra. Tiramos un dado y elegimos la caja según el resultado: si sale un número menor que 5, extraemos una bola al azar de la primera caja. En caso contrario, extraemos la bola de la segunda caja. Calcula la probabilidad de que la bola extraída sea blanca.

Problema

* En un grupo hay un 60% de chicos y un 40% de chicas. Sabemos que a un 65% de los chicos les gusta el fútbol, y que a un 20% de las chicas les gusta el fútbol.

1. Si elegimos una persona al azar, ¿cuál es la probabilidad de que le guste el fútbol?
2. Si elegimos una persona al azar, y resulta que no le gusta el fútbol, ¿cuál es la probabilidad de que sea un chico?



chicos: 60% chicas: 40%

Problemas

* Tenemos una caja con 7 bolas blancas y 3 bolas negras. Extraemos dos bolas al azar (sin reemplazamiento).
  Te piden apostar a favor o en contra del suceso $A \equiv$ "al menos una de las bolas es negra". ¿Qué harías?

* Una fábrica de enlatados produce 5000 envases diarios. La máquina A produce 3000 envases y la máquina B el resto. Sabemos que el 2% de los envases que produce la máquina A son defectuosos, y que el 4% de los que produce la máquina B son defectuosos.
  1. Calcula la probabilidad de que un envase elegido al azar sea defectuoso.
  2. Si el envase seleccionado es defectuoso, ¿cuál es la probabilidad de que fuera fabricado por la máquina B?

Problema

* La prueba de una enfermedad ofrece una tasa de falsos positivos del 5% (y no hay falsos negativos). La enfermedad afecta a una de cada mil personas. Se elige una persona de forma aleatoria, se le hace la prueba y da positivo. ¿Cuál es la probabilidad de que tenga la enfermedad?

* Interpretación:

Un ejemplo con cifras reales

* Ejemplo sacado de http://nadaesgratis.es/?p=40761

  La prevalencia[a] del cáncer de mama es del 1%.
  Las sensibilidad del test es del 90%, es decir, si una mujer está enferma, el test lo detecta con probabilidad de 0.9.
  El porcentaje de falsos positivos[b] es del 9%.
  Si a una mujer el test le da positivo, ¿cuál es la probabilidad de que esté afectada por la enfermedad?

---

[a] Porcentaje de mujeres afectadas en la población de riesgo
[b] Que el test sea positivo cuando la mujer no está afectada

Un comentario sobre aplicaciones de la estadística

* La estadística tiene innumerables aplicaciones.

* Un ejemplo muy sencillo: estimación del número de peces en un estanque.

  1. Pescamos 20 peces, les hacemos una marca y los devolvemos al estanque.
  2. Al día siguiente, volvemos al estanque y pescamos 50 peces. Si hay 4 que tienen la marca, ¿cuántos peces dirías que hay en el estanque?

* Es la técnica mas sencilla de estimación de poblaciones. Se conoce como método de captura y recaptura.

* Por supuesto, se trata de una estimación. Estudiar cómo de fiable es esa estimación bajo diversos supuestos es uno de los temas de la estadística avanzada.