

PROBABILIDAD Y ESTADÍSTICA (AVANZADAS)

PARA BACHILLERATO

Ignacio Vallés Oriola



Índice general

0.1.	¿Qué es la Estadística?	12
0.2.	Estructura de este libro	13
0.3.	Guía de lectura	14
0.4.	Acerca de este libro	16
I	Estadística Descriptiva	17
1.	Estadística Descriptiva unidimensional	19
1.1.	Terminología	19
1.2.	Frecuencias	20
1.2.1.	Frecuencias para datos agrupados	21
1.3.	Representaciones gráficas	23
1.3.1.	Diagrama de barras	24
1.3.2.	Histogramas	24
1.3.3.	Polígonos de frecuencias	25
1.3.4.	Diagrama de sectores	26
1.3.5.	Otros tipos de representaciones gráficas en estadística	28
1.3.6.	Diagramas de tallos y hojas	29
1.3.7.	Diagrama de cajas y bigotes	29
1.4.	Parámetros estadísticos	29
1.4.1.	Parámetros de centralización	30
1.4.2.	Parámetros de posición	38
1.4.3.	Parámetros de dispersión	45
1.4.4.	Parámetros de forma	48
1.4.5.	Interpretación conjunta de la media y la desviación típica	50
1.5.	Tipificación	50

1.6. Ejercicios	52
1.7. Curiosidades	58
2. Distribuciones bidimensionales: Correlación y Regresión lineal	63
2.1. Distribuciones estadísticas bidimensionales	63
2.1.1. Distribuciones de frecuencias	64
2.1.2. Representación gráfica de variables bidimensionales	66
2.1.3. Parámetros de las distribuciones bidimensionales	67
2.2. Dependencia o Correlación	69
2.3. Correlación lineal	70
2.4. Regresión lineal	71
2.5. Coeficiente de determinación	74
2.5.1. Valoración de las predicciones. Interpolación y extrapolación	75
2.6. Recta de Tukey	80
2.7. Ejercicios	84
2.7.1. Problemas propuestos (con solución)	87
2.8. Curiosidades	89
II Probabilidad	95
3. Cálculo de probabilidades	97
3.1. Introducción	97
3.2. Sucesos	98
3.2.1. Tipos de sucesos	99
3.2.2. Operaciones con sucesos	100
3.3. Probabilidad	106
3.4. Probabilidad condicionada	116
3.5. Teorema de la probabilidad total y teorema de Bayes	119
3.6. Ejercicios	127
3.6.1. Problemas propuestos (con solución)	157
3.7. Curiosidades	165
4. Distribuciones de Probabilidad	173

4.1.	Variable aleatoria	174
4.2.	V.A. Discreta	174
4.2.1.	Función de probabilidad	174
4.2.2.	Función de distribución	175
4.2.3.	Esperanza matemática, varianza y desviación típica	176
4.3.	Distribución binomial	178
4.4.	Variable Aleatoria Continua	187
4.4.1.	Función de distribución y función densidad	188
4.4.2.	Esperanza matemática y desviación típica de una v.a. continua	189
4.5.	Distribución Normal	190
4.5.1.	Normal standard (típica o tipificada)	196
4.5.2.	Cálculo de probabilidades en una $N(\mu, \sigma)$. Tipificación	203
4.6.	Aproximación de la Binomial por una Normal	206
4.7.	Ejercicios	210
4.7.1.	Problemas propuestos	216
4.8.	Curiosidades	223
III	Estadística Inferencial	227
5.	Distribuciones muestrales. Estimación	229
5.1.	Introducción	229
5.2.	Intervalos característicos	231
5.2.1.	Intervalos característicos en $N(0,1)$	231
5.2.2.	Intervalos característicos en una $N(\mu, \sigma)$	233
5.3.	Estimación de las medias muestrales. Teorema central del límite	234
5.3.1.	Estimación por intervalos: intervalos de confianza para la muestra	238
5.3.2.	Relación entre nivel de confianza, error admisible y tamaño de la muestra	242
5.3.3.	Ejercicios de estimación de las medias muestrales	243
5.4.	Estimación de la proporción	248
5.4.1.	Intervalo de confianza para una proporción	250
5.4.2.	Ejercicios de estimación de las proporciones	251
5.5.	Problemas tipo de distribuciones muestrales	257

5.6. Ejercicios	258
5.6.1. Problemas propuestos	258
5.7. Curiosidades	268
6. Contraste de hipótesis	273
6.1. Introducción	273
6.2. Elementos de un contraste de hipótesis	274
6.2.1. Hipótesis	274
6.2.2. Errores	276
6.2.3. Nivel de significación y potencia	279
6.2.4. Región de aceptación y región crítica	280
6.2.5. Tipos de contraste: bilateral o de colas	280
6.3. Metodología general de un test de hipótesis	281
6.4. Contraste de hipótesis para la media poblacional	282
6.5. Contraste de hipótesis para la proporción poblacional	284
6.6. Contraste de hipótesis para la diferencia de las medias de dos poblaciones	286
6.7. Ejercicios	288
6.7.1. Problemas propuestos	297
6.8. Curiosidades	303
IV Apéndices	307
A. Ideas básicas de la teoría de conjuntos	309
A.1. Conjuntos	309
A.1.1. Subconjuntos	310
A.1.2. Diagramas de Venn	312
A.2. Operaciones con conjuntos	313
A.2.1. Unión e Intersección de conjuntos	313
A.2.2. Diferencia de conjuntos	315
A.2.3. Producto cartesiano de dos conjuntos	316
A.3. Propiedades combinadas de las operaciones con conjuntos	316
B. Combinatoria. Técnicas de recuento	319

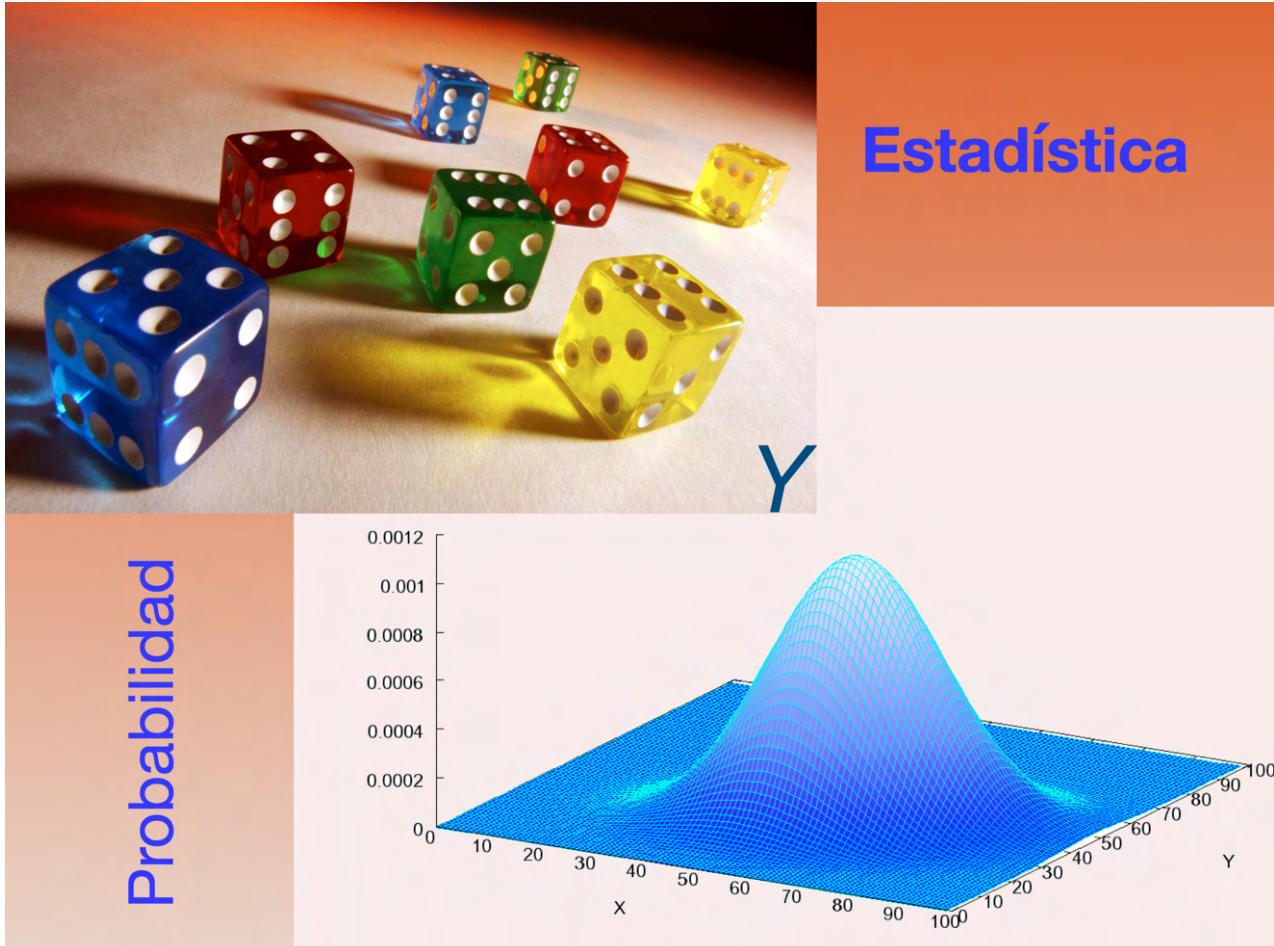
B.1. Principio de multiplicación	319
B.2. Permutaciones	321
B.2.1. Permutaciones con repetición	322
B.2.2. Permutaciones circulares	323
B.3. Variaciones	323
B.3.1. Variaciones con repetición	324
B.4. Combinaciones	325
B.4.1. Combinaciones con repetición	326
B.5. Resumen	327
B.6. Números combinatorios	328
 C. Tablas distribución Binomial y Normal	 331
 D. El problema del borracho *	 339
D.1. Esquema general de la resolución	339
D.2. Resolución del problema	339
D.2.1. Distribución binomial	340
D.2.2. Distribución en función de la posición final	340
D.2.3. Probabilidad de volver al farol: caso particular $m = 0$	341
D.2.4. Análisis gráfico de las ecuaciones D.4 y D.5	342

Probabilidad Y Estadística

Hic sunt dracones



Ignacio Vallés Oriola



- I - Estadística descriptiva
 - 1 - Estadística descriptiva
 - 2 - Distribuciones bidimensionales: Correlación y Regresión lineal
- II - Probabilidad
 - 3 - Cálculo de Probabilidades
 - 4 - Distribuciones de probabilidad
- III - Estadística inferencial
 - 5 - Distribuciones muestrales. Estimación
 - 6 - Contraste de Hipótesis.

0.1. ¿Qué es la Estadística?

Podemos decir que la Estadística es la *herramienta* que se utiliza cuando se quiere estudiar un hecho, el que sea, y no se conocen las leyes que lo rigen.

En estos casos lo único que se puede hacer es *observar el fenómeno* o suceso de interés y *tomar datos*, y luego analizar esos datos y ver si tienen relación con otras variables, si se comportan de alguna forma especial...

La Estadística consiste en todo eso, recoger los datos, estudiarlos, analizarlos y sacar conclusiones. Y basándonos en las observaciones, a veces seremos capaces de proponer una ley que explique o que al menos describa el comportamiento del suceso.

Como herramienta al servicio de otras ciencias la Estadística aparece en todas partes, Física, Química, Biología... y también ciencias sociales, Economía, Sociología, Psicología... Hasta en Lingüística se usa la estadística para determinar, por ejemplo, las letras más frecuentes en los textos de una lengua.

Un poco de historia...

Las antiguas civilizaciones, como la egipcia, la china y la azteca, ya hacían estadísticas sobre el número de personas que vivían en las ciudades, normalmente para organizar el pago de impuestos y el ejército. En general a lo largo de toda la historia los gobiernos y dirigentes de las distintas naciones han procurado disponer de datos sobre la población con fines organizativos.

La Estadística como ciencia experimentó un gran avance gracias al desarrollo de la Matemáticas, en especial de la *Teoría de la Probabilidad*, cuyas bases no fueron establecidas hasta el siglo XVII por los matemáticos franceses Pierre de Fermat y Blaise Pascal. ¿Y qué tiene que ver la Probabilidad con la Estadística? ¡Mucho! Hemos dicho que la Estadística es un instrumento para el estudio de un fenómeno cuando no se conocen que leyes lo rigen, y si hay fenómenos que no están regidos por leyes eso son los fenómenos aleatorios. Y la base matemática para el estudio estadístico de los fenómenos aleatorios la proporciona la Teoría de la Probabilidad.

INSTITUTOS NACIONAL DE ESTADÍSTICA. INE
<https://www.ine.es/explica/explica.htm>



La estadística se divide en dos partes, la *Estadística Descriptiva* que se encarga de la recolección de datos de un proceso aleatorio, clasificarlos, representarlos gráficamente y reducirlos a números estadísticos y la *Estadística Inferencial* que se encarga de deducir consecuencias a partir de los datos proporcionados por la estadística descriptiva y hacer predicciones. Para ello se basa en la *Teoría de las probabilidades*.

0.2. Estructura de este libro

La estructura del libro se presenta del siguiente modo:

Definición 0.1:

De esta manera aparecerán las definiciones.

Teorema 0.1:

En estos recuadros aparecerá la teoría: teoremas, propiedades, ...

Ejemplo 0.1:

Estos recuadros están reservados a los ejemplos que ilustran los distintos apartados.

Ampliación

Aquí aparecerán las ampliaciones de la teoría.

Curiosidades

Reservamos este cuadro para las curiosidades relacionadas con el tema que se esté tratando.

Resumenes

Al final de cada tema aparece un resumen del mismo.

Ejercicio resuelto 0.1. *Así pondremos los ejercicios del tema.*

Las soluciones a los ejercicios del tema aparecerán fuera del recuadro (dentro de ellos en los que he llamado ‘ejercicios resueltos’).

Los ejercicios propuestos con solución que acompañan a todos los temas aparecen sin ningún tipo de resalte.

Párrafo destacado: reservamos esta forma de resaltar para hacer incapié sobre determinados aspectos importantes del tema.

0.3. Guía de lectura

Tema 1. Estadística descriptiva unidimensional

En este capítulo se explican los conceptos básicos de la estadística descriptiva: tablas, gráficos y parámetros estadísticos: de centralización, de posición, de dispersión y de forma.

Para finalizar, se presenta el coeficiente de variación de Pearson para la comparación entre distribuciones estadísticas distintas. Se introduce el concepto de “tipificación de la variable”.

Tema 2. Distribuciones bidimensionales. Correlación y regresión lineal

En el tema se estudia la correlación lineal entre dos variables estadísticas y, en su caso (que así lo indique el diagrama de dispersión y que el coeficiente de correlación sea, en valor absoluto, próximo a la unidad), encontrar la recta que mejor se ajusta a la nube de puntos, la recta de regresión.

Se hacen predicciones con las dos rectas de regresión haciendo hincapié en que la mayor fiabilidad de las interpolaciones.

Se menciona la recta de Tukey como alternativa a la de regresión ante la presencia de *outliers* y, como ampliación, se habla de las correlaciones exponencial y potencial, ambas no lineales.

Tema 3. Probabilidad

Después de introducida el álgebra de los sucesos de experimentos aleatorio se dan varias definiciones de probabilidad: a posteriori o frecuencialista y a priori o regla de Laplace. A continuación se enuncia la definición axiomática de Kolmogorov.

Seguimos con la definición de probabilidad condicionada y los teoremas de la probabilidad total y de Bayes. Este tema, por su interés y dificultad, se ve con más detenimiento y se acompaña de gran cantidad de ejemplos y ejercicios resueltos y propuestos con solución.

Tema 4. Distribuciones de probabilidad

Las distribuciones de probabilidad son idealizaciones matemáticas de las distribuciones estadísticas. En el presente tema se analizan detalladamente las distribuciones de probabilidad más importantes: la binomial (para variable aleatoria discreta) y la normal (para variable aleatoria continua).

Como distribuciones de variable aleatoria discreta se analizan también la distribución uniforme, la de Bernoulli y la de Poisson. Para variable continua se estudian, someramente, la distribución uniforme continua y la exponencial.

En la distribución normal se estudió el uso de tablas para la normal típica o standard $N(0,1)$ y la “tipificación de la variable” para cualquier distribución normal $N(\mu, \sigma)$.

El tema termina con el estudio de la aproximación de la binomial por la normal y la correspondiente “corrección por continuidad”.

Este tema se acompaña de gran cantidad de ejemplos y ejercicios resueltos y propuestos con solución.

Tema 5. Distribuciones muestrales. Estimación

Comenzamos con la definición y cálculo de los intervalos característicos en una distribución normal típica y en una normal cualquiera, que usaremos más tarde en la estimación de parámetros por intervalos.

Estudiamos la distribución de las medias muestrales y analizamos el teorema central del límite. Se aprende a hacer estimaciones puntuales y por intervalos de confianza de la media de la población (conocida una muestra) y de la media de una muestra (conocida la población). Se hace notar la relación entre el nivel de confianza, el error máximo admisible y el tamaño de la muestra.

Como ampliación, vemos la distribución de la diferencia de medias de dos muestras.

El tema acaba analizando la distribución de la proporción y haciendo estimaciones para la proporción de una población y de una muestra.

Este tema se acompaña de gran cantidad de ejemplos y ejercicios resueltos y propuestos con solución.

Tema 6. Contraste de hipótesis

En el estudio del contraste de hipótesis se estudian los elementos que la componen así como la metodología general de su aplicación, teniendo en cuenta los dos tipos de posibles errores en que se puede caer.

Estudiamos el contraste de hipótesis para la media de un población y para la proporción de una población. Como ampliación, vemos el contraste de hipótesis para la diferencia de medias de dos poblaciones.

Este tema se acompaña de gran cantidad de ejemplos y ejercicios resueltos y propuestos con solución.

Apéndices

A - “Ideas básicas de la teoría de conjuntos”: analizamos el álgebra de Boole como sistema lógico de amplia aplicación en probabilidad.

B - “Combinatoria. Técnicas de conteo”: principio de multiplicación, principio del palomar, permutaciones, variaciones y combinaciones (con y sin repetición).

Estrategia de los diagramas de árbol. Números combinatorios, triángulo de Tartaglia y binomio de Newton.

C - “Tablas de las distribuciones Binomial y Normal”.

D - “El problema del borracho”. Analizamos en este apéndice el famoso problema cuyo enunciado dice: *“Un borracho parte de un farol dando pasos de igual longitud hacia ambos lados. ¿Cuál es la probabilidad de que después de N pasos vuelva al farol?”*

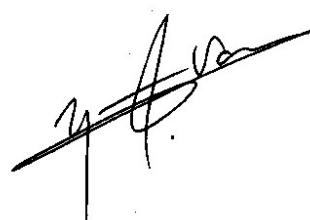
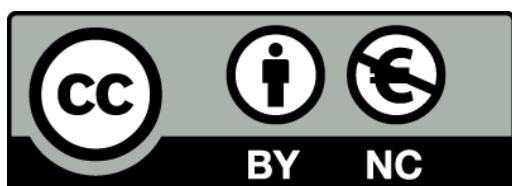
0.4. Acerca de este libro

El presente libro, colección de apuntes, contempla la parte de estadística de los temarios de bachillerato. Junto a los otros dos, “Cálculo infinitesimal (avanzado) para bachillerato” y “Álgebra lineal y geometría (avanzadas) para bachillerato”, este “Probabilidad y estadística (avanzadas) para bachillerato” contemplan todo el temario actual de matemáticas de este nivel de enseñanza.

Es cierto que falta la parte de “Programación lineal” que haré en una nueva pequeña entrega. La confección de estos textos es fruto de una larga experiencia como profesor de matemáticas de secundaria y para ello me he basado en mis más de treinta años de docencia y en la de tantos autores que han contribuido a la explicación de estos conceptos a multitud de alumnos. He usado también apuntes y problemas de libros de texto de segundo de bachillerato así como apuntes y ejercicios encontrados en la web y pruebas de acceso a la universidad de distintas comunidades autónomas. Gracias a todos sus autores por su inestimable ayuda para la confección de estos textos que espero que sirva a alguien y que escribo libre de todo tipo de derechos.

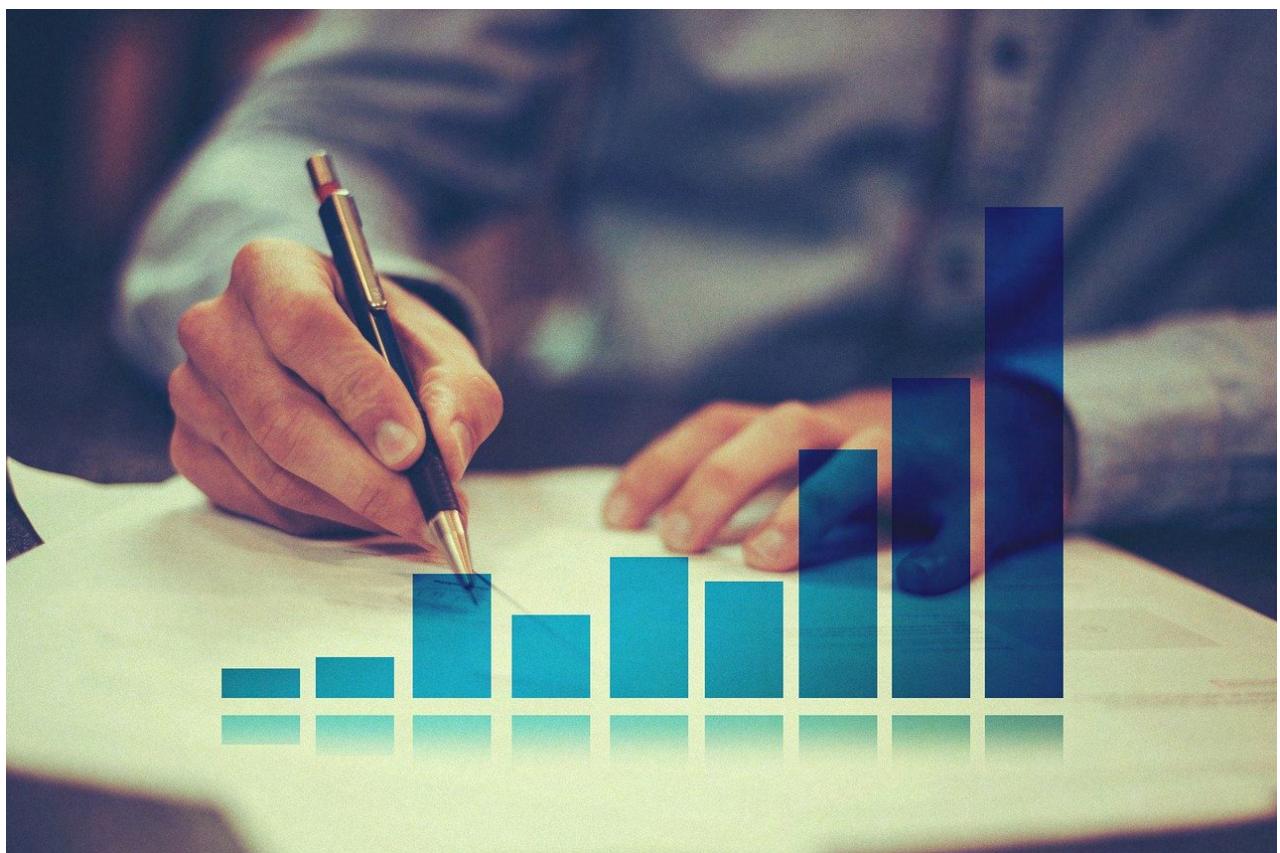
*Este material es un conjunto de apuntes personales que comparto gratuitamente en la red.
Se agradecería la comunicación de la detección de cualquier error.*

Este documento se comparte bajo licencia ‘Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)’



Parte I

Estadística Descriptiva



Capítulo 1

Estadística Descriptiva unidimensional

1.1. Terminología

Definición 1.1:

- **Población**

Conjunto sobre el que se va a realizar es estudio estadístico.

- **Muestra**

Cualquier subconjunto de la población.

- **Individuo**

Cualquier elemento de la población.

- **Serie o distribución estadística**

Conjunto de datos (cuantitativos o cualitativos) que se obtienen al estudiar un *carácter* de los individuos de una población o muestra.

- **Carácter estadístico**

Propiedad o característica de los individuos que se desea someter a estudio.

Tipos de caracteres estadísticos:

- **Cualitativos o *Atributos*.** (Sexo, Profesión, ...)

No se pueden medir, describen *modalidades*.

- **Cuantitativos o *Variables*.** (Edad, Altura, ...)

Se pueden medir, toman valores numéricos.

Tipos de variables estadísticos:

- **Discretas.** (Número de hijos, Edad –en años–,...)
Toman valores aislados.
- **Continuas.** (Altura de personas, Peso, ...)
Pueden tomar cualquier valor en un intervalo.

1.2. Frecuencias

Definición 1.2:

Dada una serie estadística que contenga N datos, llamaremos:

- **Frecuencia absoluta** de un dato, al número de veces que aparece ese dato en la serie. n_i
- **Frecuencia relativa** de una dato, al cociente entre la frecuencia absoluta del mismo y el número total de datos de la serie. $f_i = n_i/N$

*“Supuesta la variable estadística **ordenada de mayor a menor**”, se definen:*

- **Frecuencia absoluta acumulada** del dato que ocupa el lugar k -ésimo en la serie, x_k , es el valor que se obtiene sumando todas las frecuencias absolutas de los datos anteriores hasta llegar al x_k :

$$N_i = \sum_{i=1}^k n_i = n_1 + n_2 + \cdots + n_k$$

- **Frecuencia absoluta relativa** del dato x_k es el valor que se obtiene sumando todas las frecuencias relativas hasta llegar al valor x_k :

$$\begin{aligned} F_i &= \sum_{i=1}^k f_i = f_1 + f_2 + \cdots + f_k = \\ &= \frac{n_1 + n_2 + \cdots + n_k}{N} = \frac{n_k}{N} \end{aligned}$$

Teorema 1.1:

$$\sum_{i=1}^N n_i = N; \quad \sum_{i=1}^N f_i = 1$$

$$N_N = N; \quad F_N = 1; \quad 0 \leq n_1 \leq N; \quad 0 \leq f_i \leq 1; \quad N_i = N_{i-1} + n_i$$

Para facilitar la lectura de los datos estadísticos, éstos se suelen agrupar y presentar en forma de **Tablas de Frecuencias**:

Valores de la Variable	Frecuencias			
	Ordinarias		Acumuladas	
	n_i	f_i	N_i	F_i
x_i	(absoluta)	(relativa)	(absoluta)	(relativa)

Ejemplo 1.1:

Una empresa se propone reestructurar las remuneraciones de sus trabajadores, se estudia los años de éstos determinándose los siguientes resultados:

4 - 5 - 4 - 6 - 7 - 9 - 7 - 7 - 5 - 8 - 8 - 7 - 6 - 7 - 7
 4 - 6 - 8 - 8 - 9 - 6 - 8 - 9 - 5 - 6 - 5 - 4 - 7 - 9 - 6
 7 - 6 - 5 - 4 - 4 - 4 - 6 - 8 - 8 - 7 - 8 - 9 - 5 - 5 - 4
 6 - 7 - 9 - 5 - 4

Agrupamos y ordenamos decrecientemente los 50 valores.

Años de servicio	n_i	f_i	N_i	F_i
4	9	0.18	9	0.18
5	8	0.16	17	0.34
6	9	0.18	26	0.52
7	10	0.20	36	0.72
8	8	0.16	44	0.88
9	6	0.12	50	1.00
TOTALES	50	1.00		

1.2.1. Frecuencias para datos agrupados

Si la variable es continua o el número de datos es muy grande, estos se *agrupan en intervalos* llamados '**clases**'. En general se toman intervalos de la misma amplitud y cerrados por la izquierda y abiertos por la derecha:

$$[a, b_1[, [b_1, b_2[, \dots, [b_{k-1}, b_k[, \dots, [b_N, b[$$

Definición 1.3:

Se llama **marca de clase** al punto medio de cada intervalo:

$$\text{Para el intervalo } [b_{k-1}, b_k[, \text{ su marca de clase es } m_k = \frac{b_{k-1} + b_k}{2}.$$

Este valor representará a todos los datos que estén en su interior.

Es evidente que al agrupar datos en intervalos *se pierde información* ya que se supone que los datos se distribuyen homogéneamente alrededor de la marca de clase y esto no siempre es cierto.

Para el número de intervalos y forma de agrupamiento de los datos hay muchos criterios diferentes, entre ellos mostramos los que aparecen en la siguiente figura:

- **Distribución de frecuencias:**
 - **Número de clases k** (en general entre 10 y 15)
 - Criterio de Norcliffe $= \sqrt{N}$
 - Criterio de Sturgess $= 1 + 3.322(\log N)$
 - Criterio de Huntsberger $= 1 + 3,3 \log N$
 - Criterio de Brooks and Carruthers $= < 5 \log N$
 - Otros $= 1 + \log_2 N$

- **Distribución de frecuencias**
- **Rango de los datos**

$$R = X_{\max} - X_{\min}$$
- **Amplitud de cada clase** número entero un poco mayor que sea divisible por el número de intervalos queremos establecer

$$a = \frac{R}{k} \approx \text{mayor}$$
- **Marca de clase**

$$m = \frac{L_{\sup} + L_{\inf}}{2}$$

Veamos un ejemplo:

Ejemplo 1.2:

Las edades de un conjunto de niños son: 3, 7, 10, 10, 10, 6, 5, 4.5, 12, 11, 10, 15, 10.5, 6, 2, 10, 9, 10, 4, 15, 13, 14, 12, 7, 10, 6, 8. Agrupando los datos, construir la tabla de frecuencias.

En los 26 datos hay 15 valores distintos (desde el 2 hasta el 15). Aplicando, p.e., el criterio de Norcliffe, tendremos que construir $\sqrt{15} \approx 4$ intervalos.

La amplitud de cada intervalo debe ser $\frac{15 - 2}{4} = \frac{13}{4} = 3.25 \rightarrow 3.5$ ó 4. Tomaremos 4 como amplitud. Así, los intervalos en que agrupar los datos serán [2, 6[, [6, 10[, [10, 14[, [14, 18[

Distribuyendo, ahora, los datos en sus clases construimos la tabla de frecuencias.

Edades x_i	Marcas de clase	Frecuencias		F. acumuladas	
		n_i	f_i	N_i	F_i
[2, 6[4	5	5/26	5	5/26
[6, 10[8	7	7/26	12	12/26
[10, 14[12	11	11/26	23	23/26
[14, 18[16	3	3/26	26	26/26
		N=26			

Intervalos no solapados: si los datos ya aparecen agrupados pero lo están en intervalos no solapados se tomarán intervalos que contengan a estos pero sin modificar las frecuencias.

Lo vemos con un ejemplo:

Ejemplo 1.3:

Internavo	n_i	Intervalo	n_i
10 a 19	10	[9.5,19.5[10
20 a 29	15	[19.5,29.5[15
30 a 39	17	[29.5,39.5[17
39 a 49	11	[39.5,49.5[17

1.3. Representaciones gráficas

En los análisis estadísticos, es frecuente utilizar representaciones visuales complementarias de las tablas que resumen los datos de estudio. Con estas representaciones, adaptadas en cada caso a la finalidad informativa que se persigue, se transmiten los resultados de los análisis de forma rápida, directa y comprensible.

Cuando se muestran los datos estadísticos a través de representaciones gráficas, se ha de adaptar el contenido a la información visual que se pretende transmitir. Para ello, se barajan múltiples formas de representación. Los tipos más frecuentes de representaciones estadísticas son:

- **Diagramas de barras:** muestran los valores de las frecuencias absolutas sobre un sistema de ejes cartesianos, cuando el carácter estadístico en estudio es un atributos o una variable discreta.
- **Histogramas:** formas especiales de diagramas de barras para distribuciones de variable continua.
- **Polígonos de frecuencias:** formados por líneas poligonales abiertas sobre un sistema de ejes cartesianos.
- **Gráficos de sectores:** circulares o de tarta, dividen un círculo en porciones proporcionales según el valor de las frecuencias relativas. Este tipo de diagramas no es nada recomendado.¹
- **Pictogramas:** o representaciones visuales figurativas. En realidad son diagramas de barras en los que las barras se sustituyen con dibujos alusivos a la variable.
- **Cartogramas:** expresiones gráficas a modo de mapa.
- **Pirámides de población:** para clasificaciones de grupos de población por sexo y edad.

¹Lo veremos más adelante, en su apartado correspondiente.

1.3.1. Diagrama de barras

Definición 1.4:

Un **diagrama de barras** es una representación gráfica en un eje cartesiano de las frecuencias de una variable cualitativa (atributo) o cuantitativa discreta.

Se representan, en el eje de abcisas, los valores de la serie estadística, y en el de ordenadas, los valores de las frecuencias absolutas.

Se suelen usar para comparar magnitudes de varias categorías, ver la evolución en el tiempo de una magnitud concreta, etc.

Ejemplo 1.4:

Producción Agrícola 2007	
Cereales	Miles de toneladas
Cebada	11.945
Trigo	6.436
Avena	4.310
Centeno	261

Fuente: Ministerio de Agricultura,
Alimentación y Medio Ambiente

Producción de cereales en España. 2007

Millones de toneladas

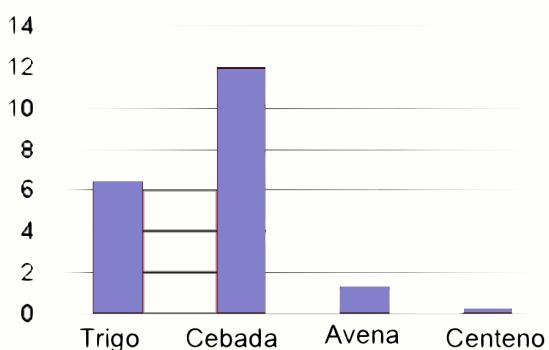


Diagrama de barras. (INE)

1.3.2. Histogramas

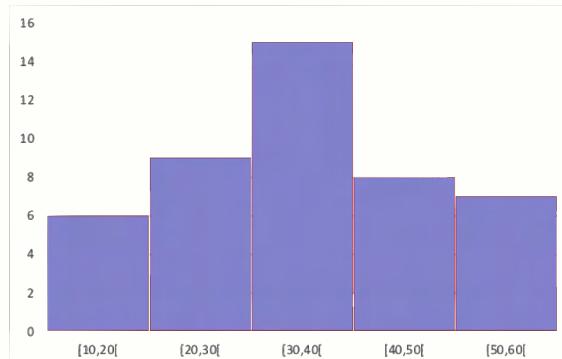
Definición 1.5:

Los **histogramas** se usan para representar las frecuencias de una variable cuantitativa continua.

En uno de los ejes se posicionan las clases de la variable continua (los intervalos) y en el otro eje las frecuencias absolutas. No existe separación entre las barras.

Ejemplo 1.5:

Clases	frecuencia
[10,20[6
[20,30[9
[30,40[15
[40,50[8
[50,60[7

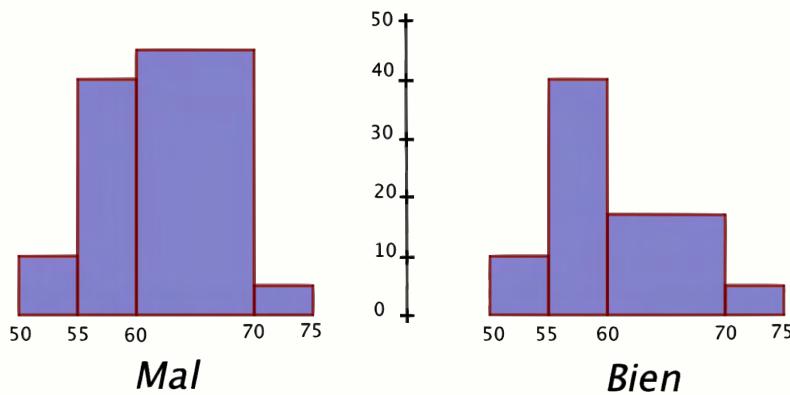


Histograma.

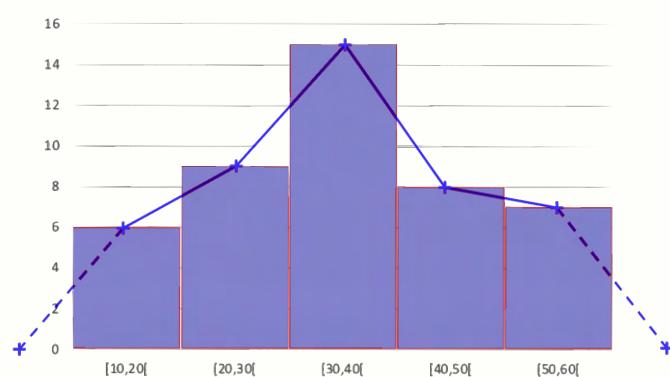
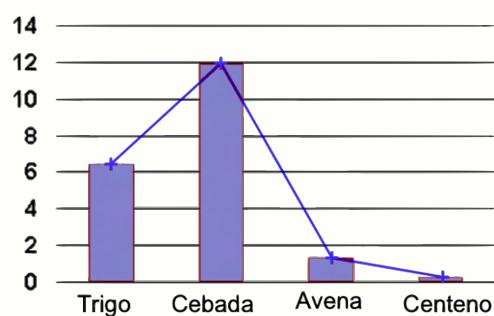
Ejemplo 1.6:

Si las clases no son de la misma amplitud, no es la altura de los rectángulos sino su área el que debe ser proporcional a la frecuencia absoluta de cada una de ellas.

Clase	frecuencia	amplitud	corrección: f/a
[50,55[10	5	2 → 10
[55,60[40	5	8 → 40
[60,70[45	10	4.5 → 18
[70,75[5	5	1 → 5

**1.3.3. Polígonos de frecuencias****Definición 1.6:**

Un **polígono de frecuencias** es una linea poligonal que une o bien el extremos de las barras en un diagrama de barras o bien los puntos medios de las partes superiores de los rectángulos que forman un histograma .

Ejemplo 1.7:**Producción de cereales en España. 2007**
Millones de toneladas

Polígonos de frecuencias (absolutas) en diagrama de barras y en histograma.

Si en lugar de usar en el eje de ordenadas las frecuencias absolutas se usan las frecuencias absolutas acumuladas, se obtienen los *polígonos de frecuencias absolutas acumuladas*. También se habla de polígonos de frecuencias relativas y de frecuencias relativas acumuladas.

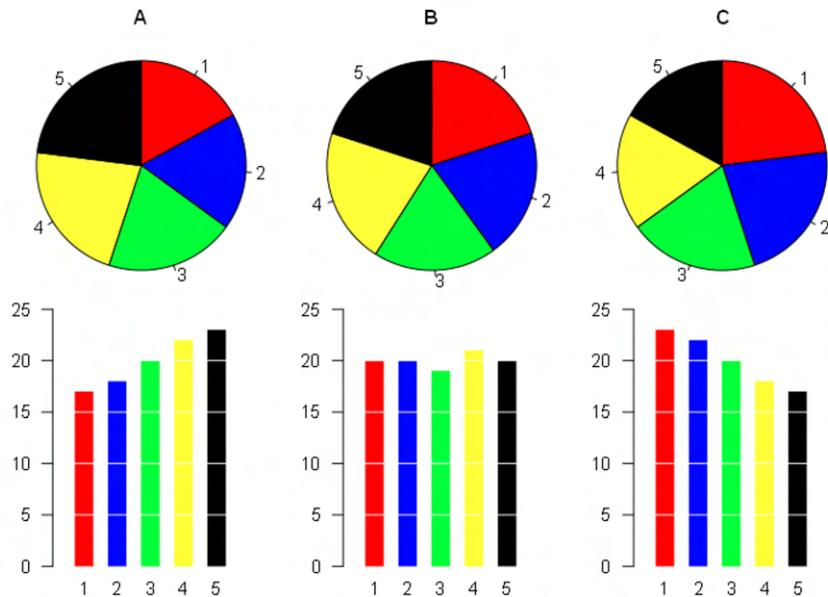
El uso de este tipo de representaciones gráficas es muy útil para el cálculo aproximado de parámetros de posición, como veremos en el próximo apartado.

1.3.4. Diagrama de sectores**Definición 1.7:**

En un **diagrama de sectores**, la serie estadística se representa como sectores circulares de ángulo (área) proporcional a la frecuencia de cada dato.

Se suele usar para datos cualitativos y con pocos valores.

“Tartas no, gracias”.



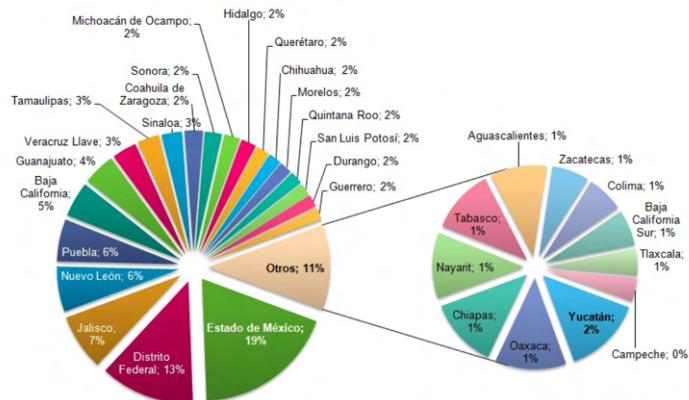
* su uso es casi anecdótico en publicaciones científicas

* [...] es más difícil comparar el tamaño de los objetos de una gráfica cuando éstos, en lugar de en longitud, varían en área o forma. De acuerdo con la ley de las potencias de Stevens, el exponente asociado al área es 0,7, mientras que el de la longitud es 1. Esto sugiere que la longitud es mejor escala: las diferencias percibidas se corresponden linealmente con las verdaderas.

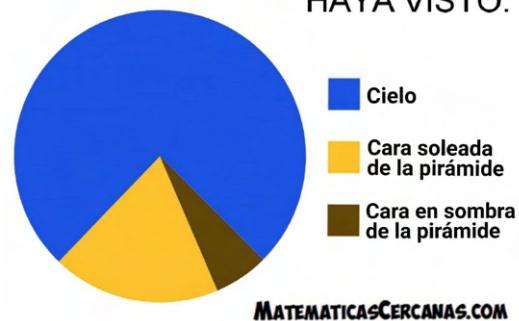
Por todo ello **¿Tartas? No, gracias.**

Los diagramas de sectores no son aconsejables.

Usuarios de Facebook en México



PROBABLEMENTE SEA EL GRÁFICO DE SECTORES MÁS EXPLÍCITO QUE JAMÁS HAYA VISTO.

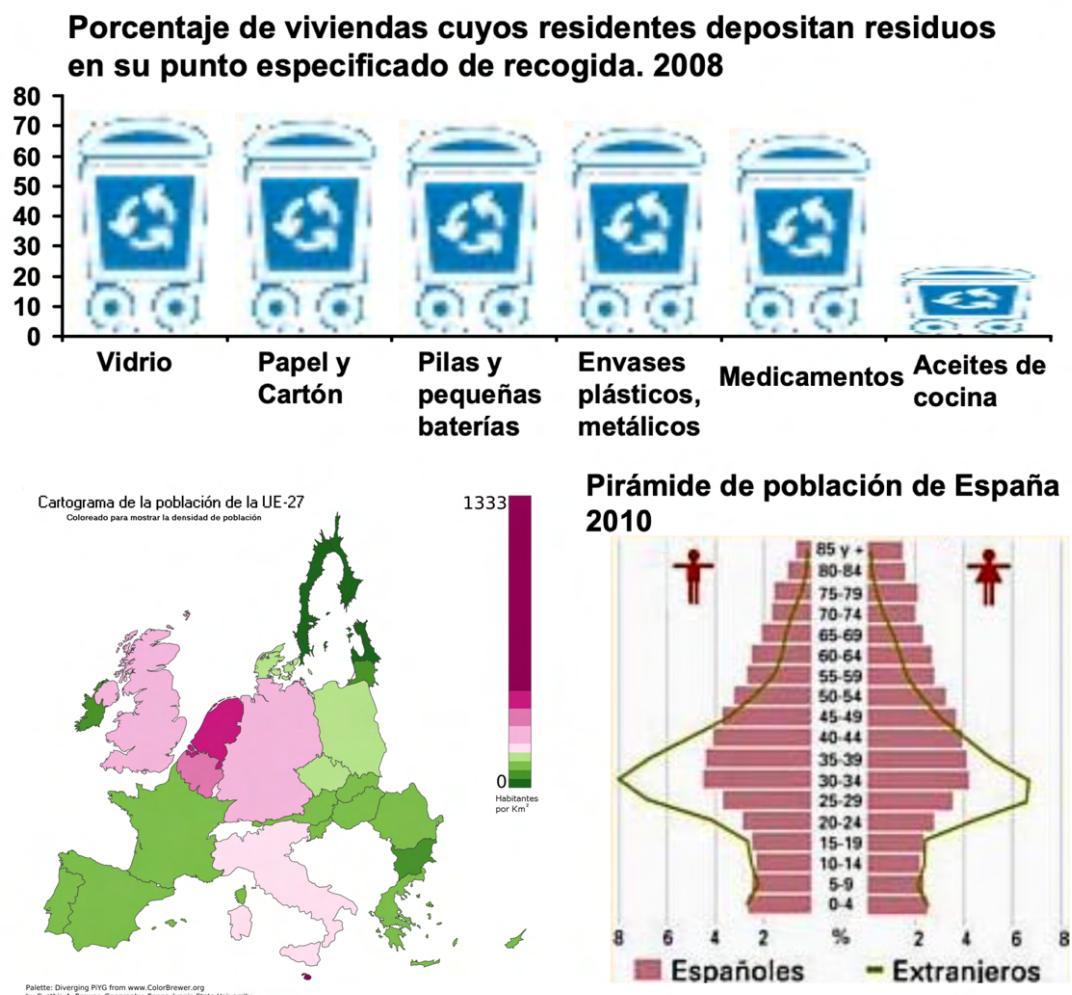


MATEMATICASCERCANAS.COM

Mucha información → nula información

--- Chiste XDD ---

1.3.5. Otros tipos de representaciones gráficas en estadística



Pictogramas, cartogramas y pirámides de población. (INE)

Definición 1.8:

Pictogramas: se usan para datos no agrupados. Se representan con imágenes alusivas a la distribución y de tamaños proporcionales (o repitiendo éstas un determinado número de veces) proporcional a la frecuencia. Tiene escasa precisión.

Cartogramas: se usan para representar sobre un mapa datos relacionados con un área geográfica.

Pirámides de población: son un par de histogramas verticales con un eje común, normalmente la edad y separando a izquierda y derecha por sexos.

1.3.6. Diagramas de tallos y hojas

Definición 1.9:

Un tipo de representación estadística que además es útil para el recuento de datos es el **diagrama de tallos y hojas** (Stem-and-Leaf Diagram) es un semigráfico que permite presentar la distribución de una variable cuantitativa. Consiste en separar cada dato en el último dígito (que se denomina hoja) y las cifras delanteras restantes (que forman el tallo).

Ejemplo 1.8:

Los datos siguientes corresponden a los tiempos de reacción de una muestra de 33 sujetos, medidos en centésimas de segundo, son: 55, 51, 60, 56, 64, 56, 63, 63, 61, 57, 62, 50, 49, 70, 72, 54, 48, 53, 58, 66, 68, 45, 74, 65, 58, 61, 62, 59, 64, 57, 63, 52, 67.

Construir el diagrama de tallos y hojas.

Tallos	Hojas
4	5 8 9
5	0 1 2 3 4 5 6 6 7 7 8 8 9
6	0 1 1 2 2 2 3 3 3 4 4 5 6 7 8
7	0 2 4

1.3.7. Diagrama de cajas y bigotes

Veremos este tipo de diagramas al final de la sección parámetros de posición.

1.4. Parámetros estadísticos

En los datos recogidos en tablas se pierde información si se agrupan en intervalos, aunque ello resulte necesario; aún se pierde más información al presentarlos en forma de gráficos pero se gana rapidez en la transmisión visual de la información. Pues, si lo que hacemos es asignar a toda la distribución estadística un limitado número de parámetros que la represente, la pérdida de la información es aún más grande pero inevitable para poder comparar rápidamente distintas distribuciones de datos, es lo que ocurre con los parámetros estadísticos.

Los parámetros estadísticos se dividen en:

- Parámetros de centralización.
 - Media aritmética (ponderada, geométrica, armónica).
 - Moda.

- Mediana.
- Parámetros de Posición.
 - Cuartiles, Deciles, Percentiles.
- Parámetros de dispersión.
 - Recorrido.
 - Desviación media.
 - Varianza. Desviación típica.
 - Coeficiente de variación de Pearson.

1.4.1. Parámetros de centralización

En matemáticas y estadística, una media o promedio es una medida de tendencia central. Resulta al efectuar una serie determinada de operaciones con un conjunto de números y que, en determinadas condiciones, puede representar por sí solo a todo el conjunto. Existen distintos tipos de medias, tales como la media geométrica, la media ponderada y la media armónica aunque en el lenguaje común, tanto en estadística como en matemáticas la elemental de todas ellas es el término que se refiere a la media aritmética.

Definición 1.10:

La **media aritmética** o simplemente media de una serie de valores aislados $\{x_1, x_2, \dots, x_n\}$, agrupados con sus frecuencias absolutas $\{n_1, n_2, \dots, n_n\}$, que se designa por \bar{x} , se define como:

$$\bar{x} = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + \dots + x_n \cdot n_n}{N} = \frac{\sum_{i=1}^n x_i \cdot n_i}{N}$$

donde $n_1 + n_2 + \dots + n_n = \sum_{i=1}^n n_i = N$ es el número total de datos.

Si la distribución está agrupada en clases, x_i representa a la ‘marca de clase’ o punto medio del intervalo.

Teorema 1.2:

1. La suma de las desviaciones respecto de la media es cero.

$$(x_1 - \bar{x}) \cdot n_1 + (x_2 - \bar{x}) \cdot n_2 + \dots + (x_n - \bar{x}) \cdot n_n = 0$$

2. Si a cada valor de los datos x_i se le añade una cantidad constante b , la media aumenta en b .

$$\overline{\{x_1 + b, n_1; x_2 + b, n_2; \dots; x_n + b, n_n\}} = \bar{x} + b$$

3. Si a cada valor de los datos x_i se le multiplica por una cantidad constante a , la media queda multiplicada por a .

$$\overline{\{ax_1, n_1; ax_2, n_2; \dots; ax_n, n_n\}} = a\bar{x}$$

4. La media es '*invariante*' frente a transformaciones lineales, cambio de origen y escala, de las variables; es decir si X es una variable aleatoria e Y es otra variable aleatoria que depende linealmente de X , $Y = a \cdot X + b$ (donde a representa la magnitud del cambio de escala y b la del cambio de origen). De otro modo $\{x_i\} \rightarrow \{y_i = a \cdot x_i + b\}$ se tiene que:

$$\bar{Y} = a \cdot \bar{X} + b \quad \text{ó} \quad \overline{\{y_i\}} = a \cdot \overline{\{x_i\}} + b$$

Observaciones sobre la media aritmética:

- La media se puede hallar sólo para variables cuantitativas.
- La media es independiente de las amplitudes de los intervalos.
- La media es muy sensible a las puntuaciones extremas, también conocidos como valores atípicos.
- La media no se puede calcular si hay un intervalo con una amplitud indeterminada (clases abiertas, p.e. [10, 20[; [20, 50[; más de 50) , en este caso no es posible hallar la media porque no podemos calcular la marca de clase de último intervalo.

Definición 1.11:

Otras medias:

$$\sum_{i=1}^N x_i \cdot p_i$$

Media ponderada: $\bar{x}_p = \frac{\sum_{i=1}^{i_1} x_i \cdot p_i}{\sum_{i=1}^N p_i}; \quad p_i \rightarrow \text{pesos}$

A veces puede ser útil otorgar pesos o valores a los datos dependiendo de su relevancia para determinado estudio. Si a los valores $\{x_1, x_2, \dots, x_n\}$ se le conceden los pesos $\{p_1, p_2, \dots, p_n\}$, la expresión anterior representa su 'media ponderada'. Sirve para dar más peso a determinados valores.

Media Geométrica: $\bar{x}_g = G = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdots x_n^{n_n}}$

La media geométrica es un promedio muy útil en conjuntos de números que son interpretados en orden de su producto, no de su suma (tal y como ocurre con la media aritmética). Por ejemplo, las velocidades de crecimiento.

Media Armónica: $\bar{x}_a = H = \frac{N}{\sum_{i=1}^n \frac{n_i}{x_i}}$

La media armónica es un promedio muy útil en conjuntos de números que se definen en relación con alguna unidad, por ejemplo la velocidad (distancia por unidad de tiempo).

Ejemplo 1.9:

Cálculo de la media del ejemplo 1.2.1:

edades	x_i	n_i	$x_i \cdot n_i$
[2,6[4	5	20
[6,10[8	7	56
[10,14[12	11	132
[14,18[16	3	48
	26	256	

$$\bar{x} = \frac{\sum x_i \cdot n_i}{\sum n_i} = \frac{256}{26} = 9.85$$

Ejemplo 1.10:

Una empresa ha generado un 20 % de rentabilidad el primer año, un 15 % el segundo año, un 33 % el tercer año y un 25 % el cuarto año. Si sumamos las cantidades y dividir entre cuatro, este valor no representa nada en correcto. Para calcular la media de varios porcentajes debemos hacer uso de la media geométrica. Aplicado al caso anterior, tendríamos lo siguiente:

$\bar{x}_g = \sqrt[4]{1.20 \cdot 1.15 \cdot 1.33 \cdot 1.25} = 1.23 \rightarrow 23\% \text{ de beneficio medio en los cuatro años, es decir, si cada año hubiese ganado un } 23\%, \text{ hubiera ganado lo mismo que ganando un } 20\% \text{ el primer año, un } 15\% \text{ el segundo, un } 33\% \text{ el tercero y un } 25\% \text{ el último año.}$

Ejemplo 1.11:

Un coche recorre 100 km a una velocidad de 60 Km/h, otros 100 km a 70 Km/h y otros 100 km a 80 Km/h. La velocidad media viene determinada por la *media armónica*:

$$\bar{x}_h = \frac{3}{\frac{1}{60} + \frac{1}{70} + \frac{1}{80}} = 69.04 \text{ km/h} \quad \left(\text{velocidad media} = \frac{\text{espacio total}}{\text{tiempo empleado}} \right)$$

Definición 1.12:

Dada una serie estadística, se llama **moda**, **mo**, al valor de la serie que tiene mayor frecuencia absoluta.

Una serie estadística puede ser *plurimodal* (tener más de una moda) y *amodal*, sin moda (si todas las frecuencias son iguales).

Teorema 1.3:

La **moda para datos agrupados** se obtiene aplicando la fórmula:

$$mo = L_k + \frac{n_k - n_{k-1}}{(n_k - n_{k-1}) + (n_k - n_{k+1})} \cdot c_k , \quad \text{donde:}$$

- L_k : extremo inferior de la **clase modal** (*clase con mayor frecuencia absoluta*) .
- n_k : frecuencia absoluta de la clase modal.
- n_{k-1} y n_{k+1} : frecuencias absolutas de las clases anterior y posterior a la clase modal.
- c_k : amplitud de la clase modal.

Ejemplo 1.12:

Cálculo de la moda del ejemplo 1.2.1:

edades	x_i	n_i
[2,6[4	5
[6,10[8	7
[10,14[12	11
[14,18[16	3
		26

Intervalo modal: [10, 14[

$$n_i = n_3 = 11; \quad n_{i-1} = n_2 = 7; \quad n_{i+1} = n_4 = 3$$

$$L_i = L_3 = 10; \quad c_i = 4$$

$$mo = 10 + \frac{(11 - 7)}{(11 - 7) + (11 - 3)} \cdot 4$$

$$mo = 11.33$$

En muchas ocasiones es suficiente con determinar el intervalo modal (el de mayor frecuencia absoluta).

Definición 1.13:

Dada una serie estadística de un *número impar de valores*, se llama **mediana** , **me**, al valor central de la serie, es decir, aquel valor tal que la mitad de los valores son mayores o iguales a él y la otra mitad, menores o iguales a él. Suponiendo la serie estadística *ordenada en sentido creciente*.

Si la serie tiene un *número par de valores*, se define la media como la *semisuma de los dos valores centrales*.

Para un determinado número de individuos N , el valor central viene determinado por $\frac{N+1}{2}$. Si el resultado es decimal, hay dos individuos centrales, el $\frac{N}{2}$ y el $\frac{N}{2} + 1$.

Teorema 1.4:

Para datos agrupados, si la media está en la clase $[L_k, L_{k+1}[$, de frecuencia absoluta n_k , frecuencia absoluta acumulada de la clase anterior a la mediana N_{k-1} y de c_i amplitud de la clase mediana, la **mediana** la determina la fórmula:

$$me = L_k + \frac{\frac{N}{2} - N_{k-1}}{n_k} \cdot c_k$$

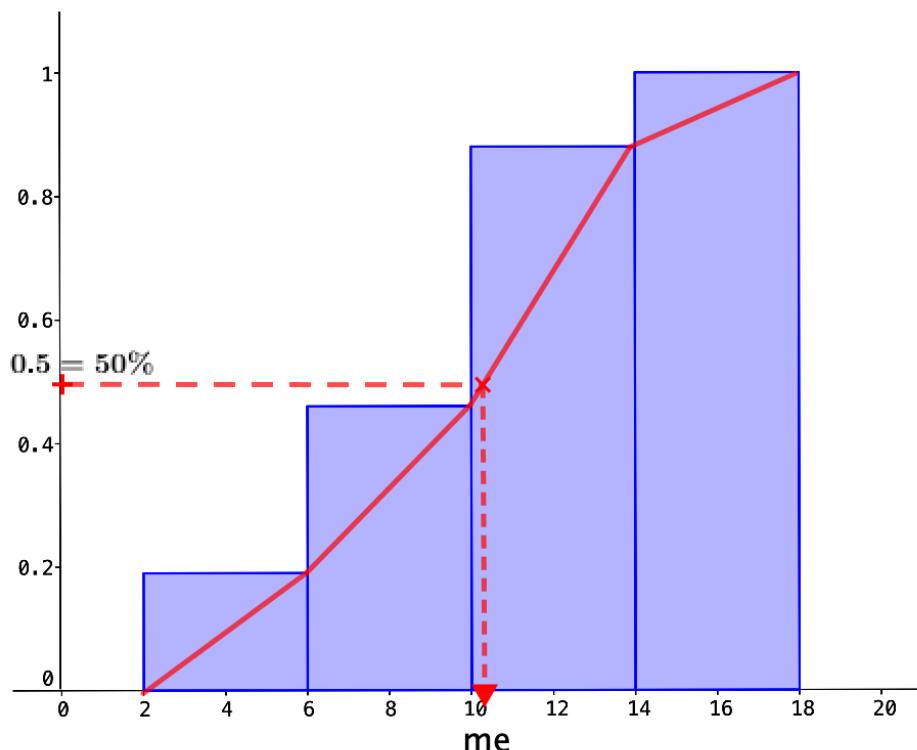
Ejemplo 1.13:

- En la serie: 2, 4, 6, **8**, 9, 10, 12, la mediana es $me = 8$
- En la serie: 2, 2, 3, **5**, **8**, 10, 10, 12, la mediana es $me = \frac{5+8}{2} = 6.5$
- En la serie del ejemplo 1.2.1, la mediana será:

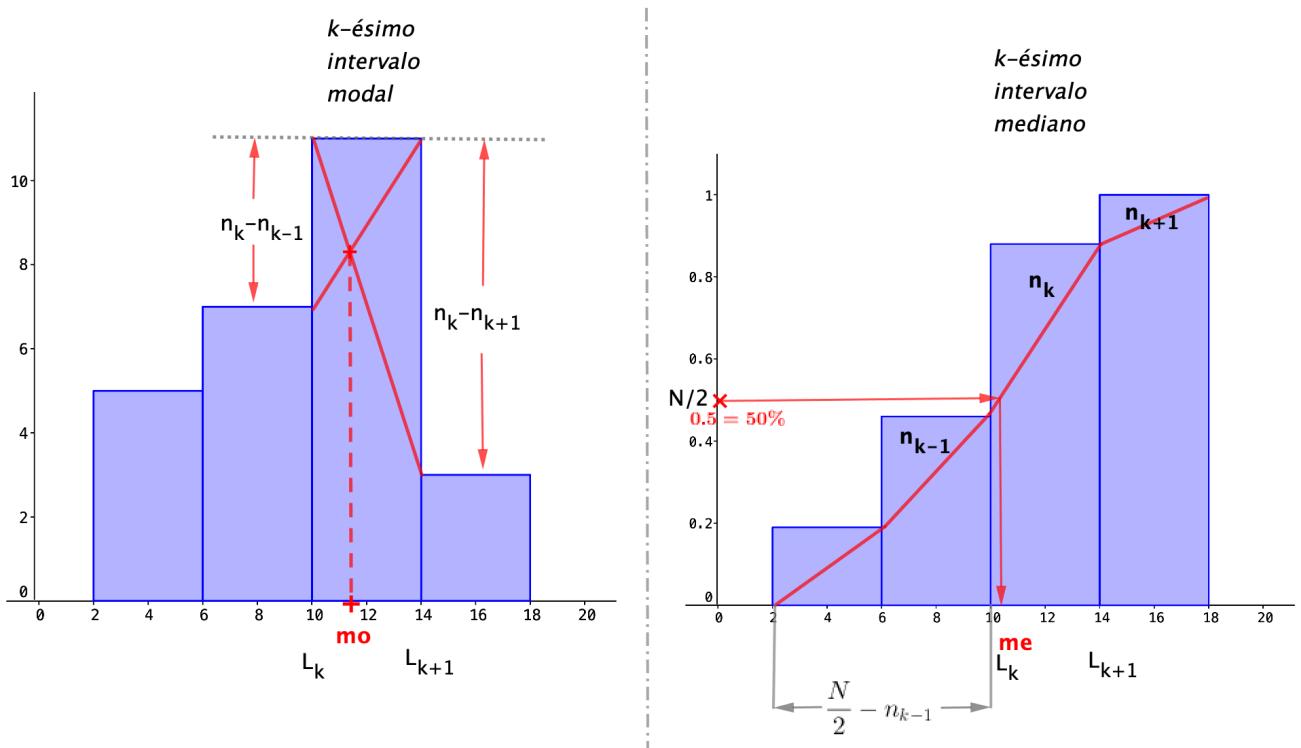
edades	x_i	n_i	N_i	H_i	$\frac{N}{2} = 13$; $L_k = 10$,
[2,6[4	5	5	0.19	$n_k = 11$; $N_{k-1} = 12$
[6,10[8	7	12	0.46	$k = 3$, clase mediana [10, 14[
[10,14[12	11	23	0.88	
[14,18[12	3	26	1.00	$me = 10 + \frac{13-12}{10} \cdot 4 = 10.36$
		26			

En ocasiones es suficiente con determinar la clase mediana.

Se puede obtener un valor aproximado de la mediana a través del histograma de frecuencias relativas acumuladas (ahora, la línea poligonal no une los puntos medios de cada clase sino los extremos superiores de cada clase).



Tanto en el caso de la mediana como en el de la moda para datos agrupados, las fórmulas se deducen la una *simple interpolación lineal*.



Ejercicio resuelto 1.1. Calcular Media, moda y mediana para la siguiente serie estadística.

Clases	x_i	n_i	N_i	$x_i \cdot n_i$
[20,30[25	20	20	500
[30,40[35	35	25	1225
[40,50[45	50	105	2250
[50,60[55	49	154	2695
[60,70[65	25	179	1625
[70,80[75	15	194	1125
[80,90[85	6	200	510
		200	9930	

$$\bar{x} = \frac{9930}{200}$$

$$\bar{x} = 49.65$$

$$mo = 40 + \frac{50 - 35}{(50 - 35) + (50 - 49)} \cdot 10$$

$$mo = 49.38$$

$$me = 40 + \frac{\frac{200}{2} - 55}{50} \cdot 10$$

$$me = 49.00$$

Ventajas y desventajas de la medidas de tendencia central.

■ MEDIA

- Ventajas
 - Es la medida de tendencia central más usada.
 - Emplea en su cálculo toda la información disponible.
 - Es un valor único.
 - Se emplea a menudo en cálculos estadísticos posteriores.
 - Tiene un sentido claro como valor de tendencia del agrupamiento de los datos.

- Es sensible a cualquier cambio en los datos (puede ser usado como un detector de variaciones en los datos).
- Es útil para llevar a cabo procedimientos estadísticos como la comparación de medias de varios conjuntos de datos.
- Presenta rigor matemático.
- En la gráfica de frecuencia representa el centro de gravedad.

- Desventajas

- Es sensible a los valores extremos.
- No es recomendable emplearla en distribuciones muy asimétricas.
- Si se emplean variables discretas, la media aritmética puede no pertenecer al conjunto de valores de la variable.
- No se puede calcular para datos cualitativos.
- No se puede calcular para datos que tengan clases de extremo abierto, tanto superior como inferior.

- MEDIANA

- Ventajas:

- No se ve influenciada por valores extremos, ya que solo influyen los valores centrales.
- Fácil de entender.
- Se puede calcular para cualquier tipo de datos cuantitativos, incluso los datos con clase de extremo abierto.
- Es la medida de tendencia central más representativa en el caso de variables que solo admiten la escala ordinal.

- Desventajas

- No utiliza en su cálculo toda la información disponible.
- No pondera cada valor por el número de veces que se ha repetido.
- Hay que ordenar los datos antes de determinarla.

- MODA

- Ventajas

- No requiere cálculos.
- Puede usarse para datos tanto cuantitativos como cualitativos.
- Fácil de interpretar.
- No se ve influenciada por valores extremos.
- Se puede calcular en clases de extremos abiertos.

- Desventajas

- Solo tiene significado en el caso de una gran cantidad de datos.
- No utiliza toda la información disponible.
- Difícil de interpretar si los datos tiene 3 o más modas.

¿Cuál de los parámetros de posición es más representativo?

Cuando estudiamos las medidas de tendencia central, como la media, la moda y la mediana, nos preguntamos, tantas medidas y al final, ¿cuál es la más representativa?, ¿es suficiente con las medidas de tendencia central para caracterizar a un conjunto de datos?^a

Aclaremos esto con un ejemplo (tomado de “Probabilidades y Estadística. Su Enseñanza” de J. Foncuberta, Red Federal de Formación Docente Continua, MECyT, 1996).

El Sr. J., gobernante de un remoto país, se jactaba de que el salario medio en su país era de 3000 €. El señor Modulador, de visita por esos lugares observó las extrañas costumbres de sus habitantes: vestían muy mal, comían peor, se guarecían donde podían y padecían graves enfermedades. Como el señor Modulador es sumamente ingenuo creyó que en ese país el ahorro más que una virtud era una obsesión pero por más que preguntó nadie supo sacarlo de su perplejidad. Hasta que cierto día encontró a un colega Modulador de aquellas regiones que le aclaró el misterio: sucedía que entre 1 millón de habitantes, la renta se distribuía así:

	Renta	perceptores	promedio per cápita
Sr. J	1.800.000.000	1	1.800.000.000
Allegados Sr. J	1.000.120.000	999	1.0001.121
Resto de habitantes	199.880.000	999.000	200
	3.000.000.000	1.000.000	3.000

(cualquier parecido con la realidad es mera coincidencia)

Efectivamente el salario promedio era de 3.000 €!, ahora bien esto no es para nada representativo de la realidad del salario en dicho país!!! o si?

Desde entonces el Sr. Modulador desconfía mucho del valor medio.



No es lo mismo un país donde el salario medio es de 3.000 € con valores máximo y mínimo 3.500 y 2.000 que otro como el del ejemplo con igual valor medio pero con valores extremos 1.800.000.000 € y 200 €.

Sin duda en el primer caso los datos están menos dispersos que en el segundo, de hecho en el ejemplo podemos hablar de valores extremos (lo que gana Don J!). Una de las desventajas del valor medio es justamente la sensibilidad a valores extremos y en este ejemplo vemos claramente como la media es ‘tironeada’ hacia arriba por estos valores extremos.

Y que pasaría si para el mismo ejemplo, simplificando la situación y pensando que los salarios son exactamente los que figuran en la columna promedio per cápita, calculáramos la Mediana? Como la hallamos? ordenando todos los datos de menor a mayor y encontrando el valor central. Tenemos un millón de datos, donde una vez ordenados el valor central sería el promedio del

dato en la posición 500.000 y la 500.001. Si ordenamos y los primeros 999.000 ganan 200 €, entonces $\text{Mediana} = (200+200)/2 = 200 \text{ €}$.

En este caso y en este ejemplo cuál de las dos medidas es más representativa? la media o la mediana? sin duda que la MEDIANA! Lo mismo podríamos decir de la MODA (también 200 €).

¿Cómo podemos cuantificar esta variabilidad del conjunto de datos? Necesitamos una medida para la desviación o dispersión de los datos, y es allí donde aparecen las medidas de variabilidad. Las vemos en el siguiente apartado.

^aDel blog “<https://estadisticaestasahi.blogspot.com>”

Media, mediana y moda coinciden en las distribuciones “normales”, aprenderemos que son estas distribuciones en el tema 4 de ‘distribuciones de probabilidad’.

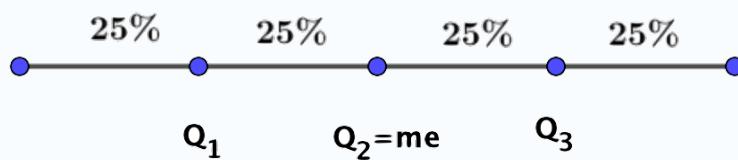
1.4.2. Parámetros de posición

Aunque hemos definido a la **mediana** como un **parámetro de centralización**, en realidad es un **parámetro de posición**, aunque representa la *posición central* de la distribución (o serie estadística) y por eso se puede considerar en los dos tipos de parámetros.

Los **cuantiles** son medidas de posición y proporcionan los valores en que se encuentra la distribución. Reciben distintos nombres según el número de partes en que si divida la distribución o serie estadística, así, se habla de **Cuartiles, Deciles y Percentiles**.

Definición 1.14:

Cuartiles: Q_1 , $Q_2 = me$, Q_3 , dividen a la población (N) en cuatro partes iguales ($N/4$, $N/2$, $3N/4$). Entre dos cuartiles sucesivos se encuentra el 25 % de la población.



Evidentemente, $Q_2 = me$.

Teorema 1.5:

Para datos agrupados, como en la mediana,

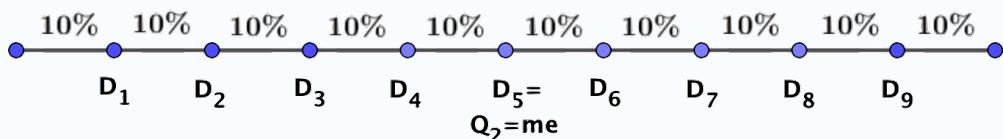
$$Q_1 = L_k + \frac{\frac{N}{4} - N_{k-1}}{n_k} \cdot c_k$$

$$Q_2 = L_k + \frac{\frac{N}{2} - N_{k-1}}{n_k} \cdot c_k = me$$

$$Q_3 = L_k + \frac{\frac{3N}{4} - N_{k-1}}{n_k} \cdot c_k$$

Definición 1.15:

Los **Deciles** son 9 valores, $D_1, D_2, D_3, \dots, D_9$, que dividen a la población en 10 partes iguales. Entre dos deciles consecutivos se encuentra el 10 % de la población.



Evidentemente, $D_5 = Q_2 = me$.

Teorema 1.6:

Para datos agrupados, como en la mediana,

$$D_m = L_k + \frac{\frac{mN}{10} - N_{k-1}}{n_k} \cdot c_k \quad m = 1, 2, 3, \dots, 9$$

Definición 1.16:

Los **Percentiles** son 99 valores, $P_1, P_2, P_3, \dots, P_{67}, \dots, P_{99}$, que dividen a la población en 100 partes iguales. Entre dos percentiles consecutivos se encuentra el 1 % de la población.

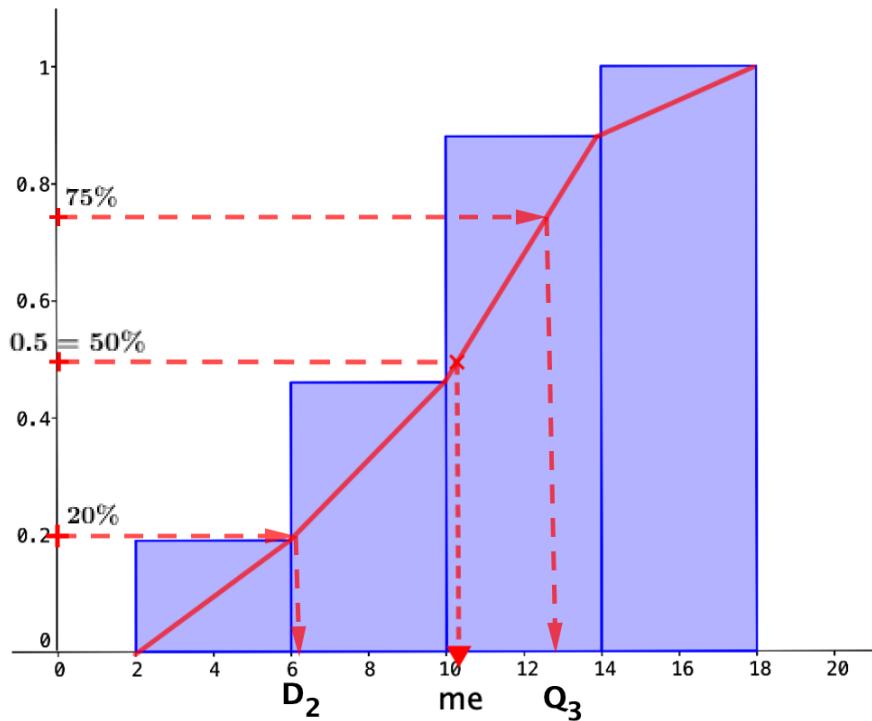
Teorema 1.7:

Para datos agrupados, como en la mediana,

$$P_m = L_k + \frac{\frac{mN}{100} - N_{k-1}}{n_k} \cdot c_k \quad m = 1, 2, 3, \dots, 99$$

Evidentemente, $P_{50} = D_5 = Q_2 = me; P_{25} = Q_1; P_{75} = Q_3$

Como ocurría con la mediana, se puede obtener un valor aproximado de los cuartiles a través del histograma de frecuencias relativas acumuladas (ahora, la línea poligonal no une los puntos medios de cada clase sino los extremos superiores de cada clase y el extremo inferior de la primera clase).



Ejemplo 1.14:

Los datos siguientes corresponden a los tiempos de reacción de una muestra de 33 sujetos, medidos en centésimas de segundo, son: 55, 51, 60, 56, 64, 56, 63, 63, 61, 57, 62, 50, 49, 70, 72, 54, 48, 53, 58, 66, 68, 45, 74, 65, 58, 61, 62, 59, 64, 57, 63, 52, 67.

Calcúlese la moda, la mediana, el primer y el tercer cuartil, directamente a partir de los datos. Calcúlese, así mismo, los deciles segundo y quinto y los percentiles 33 y 58.

Construimos un diagrama de tallos y hojas, que nos permitirá contar los datos y tenerlos ordenados (una forma alternativa a las tablas de frecuencias absolutas).

Tallos	Hojas
4	5 8 9
5	0 1 2 3 4 5 6 6 7 7 8 8 9
6	0 1 1 2 2 3 3 3 3 4 4 5 6 7 8
7	0 2 4

$$mo = 63;$$

$$(33 + 1)/2 = 17 \rightarrow me = x_{17} = 60$$

$$\frac{1}{4}33 = 8.25 \rightarrow Q_1 = x_9 = 55$$

$$\frac{3}{4}33 = 24.75 \rightarrow Q_3 = x_{25} = 34$$

$$D_2 = \frac{2}{10}33 = 6.6 \rightarrow D_2 = x_7 = 53$$

$$D_5 = me = 60$$

$$\frac{33}{100}33 = 10.89 \rightarrow P_{33} = x_{11} = 66$$

$$\frac{58}{100}33 = 19.14 \rightarrow P_{58} = x_{20} = 68$$

Ejemplo 1.15:

Con los datos del ejemplo anterior, construya una tabla estadística agrupados los datos en 5 intervalos de igual amplitud, calcule la mediana, el primer y el tercer cuartil. Calcúlese, así mismo, los deciles segundo y quinto y los percentiles 33 y 58.

Para llegar a la anterior tabla se ha calculado en primer lugar el rango de la distribución que es el mayor valor 74 menos el menor 45, lo que nos da 29. Como 29 no es divisible entre 5 redondeamos hasta el valor más próximo por exceso que es 30, dividiendo este rango entre el número de intervalos que deseamos, cinco, obtenemos la amplitud que deben tener los intervalos, seis. A partir del primer valor, 45 se han calculado los restantes extremos sumando 6, sucesivas veces. Posteriormente se ha contado el número de observaciones comprendidas dentro de cada intervalo, recuérdese que los intervalos se toman abiertos a la derecha, y de esta forma se han obtenido las frecuencias que aparecen en la tabla, en la que se han añadido las frecuencias acumuladas.

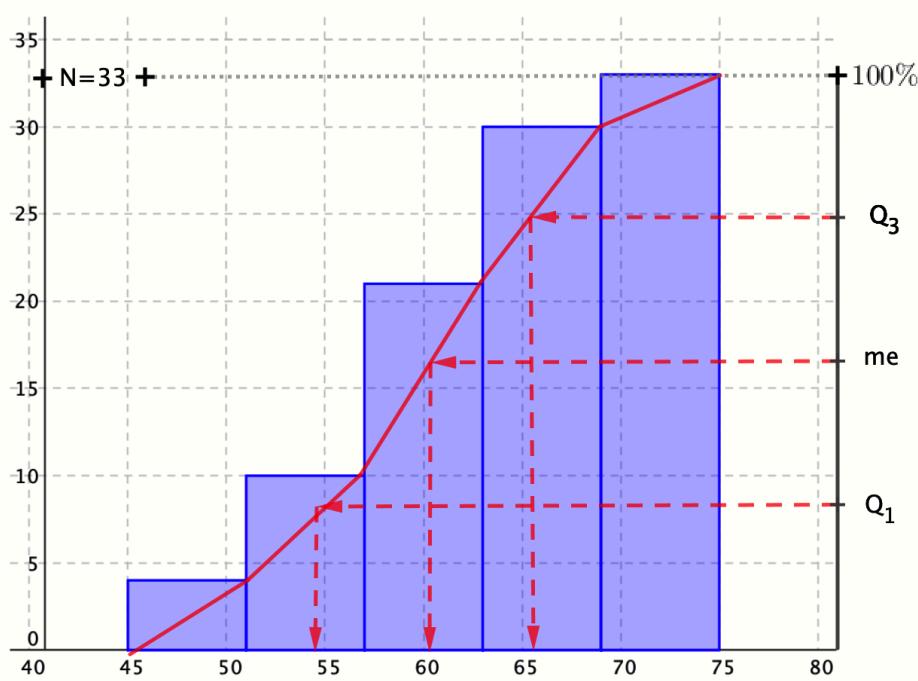
Clases	n_i	N_i
[45,51[4	4
[51,57[6	10
[57,63[11	21
[63,69[9	30
[69,75[3	33

$$\begin{aligned} & \text{— Intervalo modal: } [57, 63[\\ & \text{— } mo = 57 + \frac{11 - 6}{(11 - 6) + (11 - 9)} 6 = 61.3 \\ & \text{— } 33/2 = 16.5 \rightarrow \text{Int. mediano: } [57, 63[\\ & \text{— } me = 57 + \frac{16.5 - 10}{11} 6 = 60.5 \end{aligned}$$

$$\begin{aligned} & \text{— } 25\%33 = 8,25 \rightarrow [51, 57[\rightarrow Q_1 = 57 + \frac{8.25 - 4}{6} 6 = 55.3 \\ & \text{— } 75\%33 = 24,75 \rightarrow [63, 69[\rightarrow Q_3 = 63 + \frac{24.75 - 21}{9} 6 = 65.5 \\ & \text{— } 20\%33 = 6,6 \rightarrow [51, 57[\rightarrow D_2 = 51 + \frac{6.6 - 4}{6} 6 = 53.6 \\ & \text{— } D_5 = me = 60.5 \\ & \text{— } 33\%33 = 10.9 \rightarrow [57, 63[\rightarrow P_{33} = 57 + \frac{10.9 - 10}{11} 6 = 57.1 \\ & \text{— } 58\%33 = 19,1 \rightarrow [57, 63[\rightarrow P_{58} = 57 + \frac{19.1 - 10}{11} 6 = 62.0 \end{aligned}$$

De manera aproximada, aparece en la figura siguiente.

Ejercicio resuelto 1.2. Para la siguiente distribución estadística, responder a las preguntas que se formulan.



Histograma de frecuencias relativas acumuladas. del ejemplo 1.15

x_i	n_i	N_i	
[0,10[8	8	a) ¿Entre qué valores se encuentra el 50 % central de los individuos?
[10,20[22	30	b) Calcule el percentil 27.
[20,30[32	62	c) ¿A partir de qué puntuación se encuentra el 12 % de los sujetos con puntuación más alta?
[30,40[44	106	d) Si descontamos el 15 % de los individuos con menor puntuación y al 15 % de los de mayor puntuación, ¿en qué intervalo de puntuación se encuentran los restantes?
[40,50[28	134	
[50,60[20	154	
[60,70[6	160	

— a) El 50 % de la población estará entre el Q_1 (que deja tras sí al 25 % de la población) y el Q_3 (que deja tras sí al 75 %, luego tiene por delante al 25 % de la población).

$$160/4 = 40 \rightarrow [20, 30[\rightarrow Q_1 = 20 + \frac{40 - 30}{32} \cdot 10 = 23.13$$

$$3 \cdot 160/4 = 120 \rightarrow [40, 50[\rightarrow Q_3 = 40 + \frac{120 - 106}{28} \cdot 10 = 45.00$$

El 50 % de la población está en [23.13, 45.00[

— b) $27 \% \cdot 160 = 43.2 \rightarrow [20, 30[\rightarrow P_{27} = 20 + \frac{43.2 - 30}{32} \cdot 10 = 20.13$

— c) El valor que deja por encima el 12 % de los sujetos con mayor puntuación es el mismo que deja por debajo el 88 % con menor puntuación, por tanto debemos calcular el percentil 88.

$$88 \% \cdot 160 = 140.8 \rightarrow [50, 60[\rightarrow P_{88} = 50 + \frac{140.8 - 134}{20} \cdot 10 = 53.40$$

— d) Se trata de calcular el percentil 15 y el percentil 85. El 15

$$15\% \text{ de } 160 = 24 \rightarrow [10, 20] \rightarrow P_{15} = 10 + \frac{24 - 8}{22} \cdot 10 = 17.27$$

$$85\% \text{ de } 160 = 136 \rightarrow [50, 60] \rightarrow P_{85} = 50 + \frac{136 - 134}{20} \cdot 10 = 51.00$$

El intervalo solicitado es [17.27, 51.00], es el llamado **intervalo intercuartílico**.

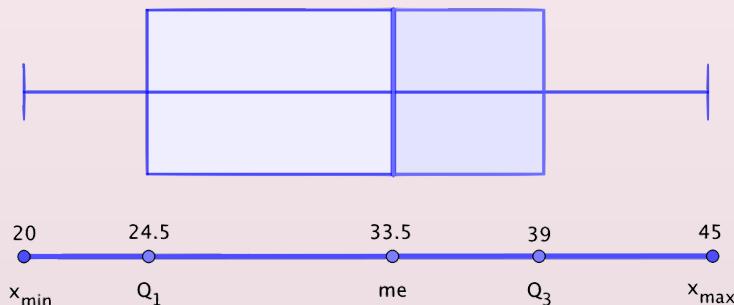
Diagrama de cajas y bigotes

Los **diagramas de cajas y bigotes** (*boxplots* o *box and whiskers*) son una representación visual que describe varias características de la serie estadística al mismo tiempo (dispersión, simetría).

Para su construcción hay que calcular Q_1, me, Q_3 que se representarán sobre un rectángulo, **caja**, y los valores x_{max}, x_{min} que se representarán por segmentos, **bigotes**.

Veamos un ejemplo. Para la serie: 36, 25, 37, 24, 39, 20, 36, 45, 31, 31, 39, 24, 29, 23, 41, 40, 33, 24, 34, 40; obtenemos:

$$Q_1 = 25; me = 33.5; Q_3 = 39; x_{min} = 20; x_{max} = 45$$



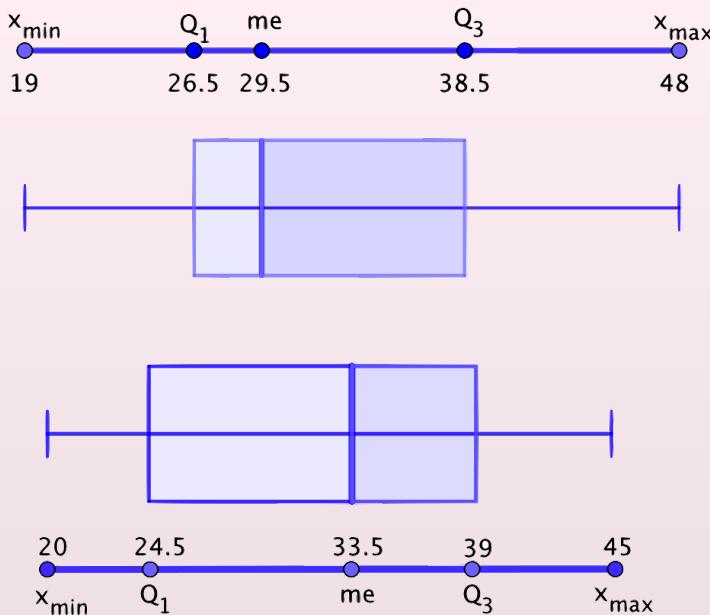
Algunas de las informaciones que se desprenden de la observación de este gráfico son:

- La parte izquierda de la caja es mayor que la derecha. Los más jóvenes, entre el 25 % y el 50 % de la población tienen edades más dispersas que los mayores de entre el 50 % y 75 % de la población.
- El bigote de la izquierda es menor que el de la derecha. El 25 % de los más jóvenes están más concentrados que el 25 % de los más mayores.
- El *rango intercuartílico*, $Q_3 - Q_1 = 14.5$. El 50 % de la población tiene edades que se diferencian en menos de 14.5 años.
- La $me = 33.5$. El 50 % de la población son menores de 33.5 años y el otro 50 % tiene edades mayores a 33.5 años.

Además, los diagramas de cajas y bigotes son muy útiles para comparar dos distribuciones: supongamos otra distribución por edades de 20 personas cuyas edades son, 35, 38, 32, 28, 30, 29, 27, 19, 48, 40, 39, 24, 24, 34, 26, 41, 29, 48, 28, 22.

Los valores que ahora se obtienen son: $Q_1 = 26.5, Q_3 = 38.5, me = 29.5, x_{min} = 19, x_{max} = 48$. Comparando ambos diagramas:

Ahora, puede obtenerse información respecto a las dos distribuciones.

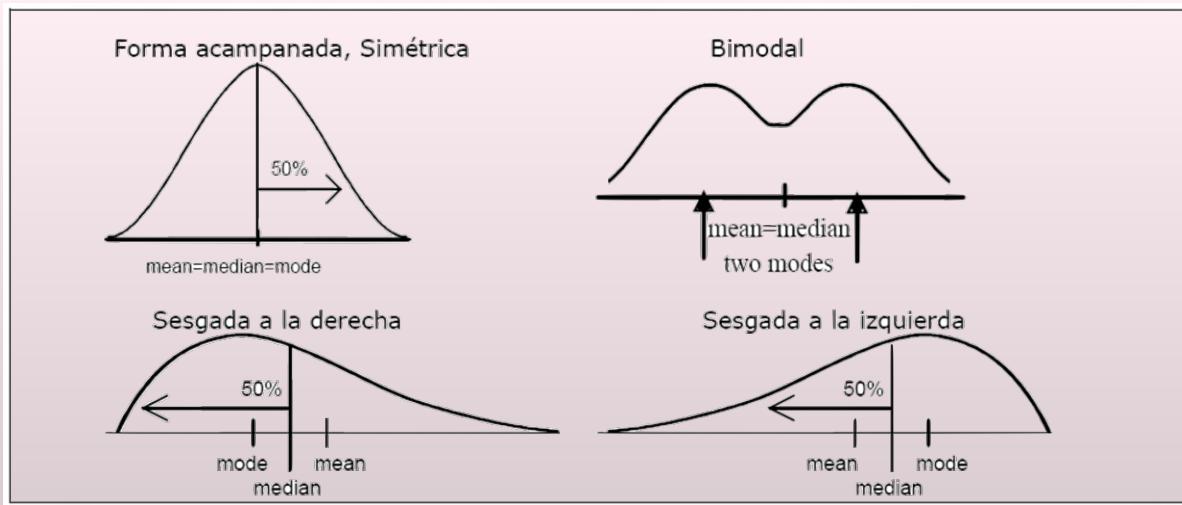


- Rápidamente se observa que la segunda distribución, representada por el diagrama de cajas y bigotes de arriba, está más concentrada en edades que la representada abajo que es más dispersa (distribución primera).
- En la distribución de abajo (primera distribución), los datos eran más dispersos entre el 25 % y 50 % de la población. En la distribución de arriba (segunda distribución), ocurre al revés, los datos son más dispersos entre el 50 % y el 75 % de la población.
- En cuanto a los bigotes, para la distribución de arriba se observa que son más largos, sobre todo la rama derecha lo cual indica que el 25 % de la población de esta distribución es más dispersa que la otra.

También se suelen representar en los diagramas de cajas y bigotes los llamados *Valores atípicos* son aquellos que muestran una gran distancia a la mediana del resto de puntuaciones en la variable, es decir, o son demasiado bajas o son demasiado altas. Numéricamente, se consideran *valores atípicos* aquellos que sean menores que $Lim_{inf} = Q_1 - 1.5 \cdot (Q_3 - Q_1)$ o mayores que $Lim_{sup} = Q_3 + 1.5 \cdot (Q_3 - Q_1)$, de otro modo, los que están fuera del intervalo (Lim_{inf}, Lim_{sup}) , es decir, los que se separan de los cuartiles más de 1.5 veces el rango intercuartílico ($Q_3 - Q_1$),

Relación entre media, mediana y moda.

- si $\bar{x} = me = mo$, la distribución es *simétrica*.
- si $me > \bar{x}$, la distribución es *asimétrica*, con cola a la derecha (*sesgada a la derecha*).
- si $me < \bar{x}$, la distribución es *asimétrica*, con cola a la izquierda (*sesgada a la izquierda*).



1.4.3. Parámetros de dispersión

Las medidas de dispersión, o de variabilidad (o propagación), expresan cómo se distribuyen los datos en torno a alguna de las medidas de centralización definidas antes (nos informan sobre cuánto se alejan del centro los valores de la distribución) y son un complemento a estas últimas para describir más fielmente un conjunto de datos.

Cuanto más pequeña sea la medida de dispersión, mayor grado de agrupamiento de los datos respecto a la medida de centralización considerada (generalmente la media) y más representativa será este parámetro de centralización para definir el comportamiento de la distribución.

Son de medidas de dispersión estadística son la varianza, la desviación típica, desviación media, el rango y rango intercuartílico.

Definición 1.17:

Se define el **recorrido** o **rango** de la distribución como la diferencia entre los valores mayor y menor de la variable.

$$R = x_{\max} - x_{\min}$$

Se define el **rango intercuartílico** como la diferencia entre el tercer y el primer cuartil. Proporciona la longitud del intervalo en el que se encuentra el 50 % de las observaciones centrales.

$$R_I = Q_3 - Q_1$$

Definición 1.18:

Se llama **desviación media** respecto al valor medio a la media de los valores absolutos de las diferencias de cada valor respecto del valor medio:

$$DM = \frac{\sum_{i=1}^n |x_i - \bar{x}| \cdot n_i}{N}$$

Definición 1.19:

Varianza, s^2 y desviación típica, s . La desviación típica es la medida de dispersión más utilizada, representa la dispersión de los datos de la distribución respecto del valor medio y se expresa en sus mismas unidades.

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot n_i}{N}; \quad s = \sqrt{s^2}$$

Teorema 1.8:

- La varianza será siempre un valor positivo o cero (si todos los datos son iguales).
- Si a todos los valores de la variable se les suma un número la varianza no varía.
 $\{x_i\} \rightarrow s^2 \Rightarrow \{x_i + d\} \rightarrow s^2$
- Si todos los valores de la variable se multiplican por un número la varianza queda multiplicada por el cuadrado de dicho número.
 $\{x_i\} \rightarrow s^2 \Rightarrow \{ax_i\} \rightarrow a^2 s^2$

- Si tenemos varias distribuciones con la misma media y conocemos sus respectivas varianzas se puede calcular la varianza total.

Para n muestras de varianza s_i^2 formadas cada una de ellas por k_i datos, la varianza total de la nueva distribución de $k_1 + \dots + k_n$ datos es: $s^2 = \frac{k_1 s_1^2 + \dots + k_n s_n^2}{k_1 + \dots + k_n}$

- Las siguientes definiciones de la varianza son equivalentes:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot n_i}{N} \leftrightarrow s^2 = \frac{\sum_{i=1}^n (x_i)^2 \cdot n_i}{N} - \bar{x}^2$$

Demostración. .

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot n_i}{N} = \frac{\sum_{i=1}^n (x_i^2 - 2x_i \bar{x} + \bar{x}^2) \cdot n_i}{N} = \\ &= \frac{\sum_{i=1}^n x_i^2 \cdot n_i}{N} - 2\bar{x} \frac{\sum_{i=1}^n x_i \cdot n_i}{N} + \bar{x}^2 = \frac{\sum_{i=1}^n x_i^2 \cdot n_i}{N} - 2\bar{x}^2 + \bar{x}^2 = \frac{\sum_{i=1}^n x_i^2 \cdot n_i}{N} - \bar{x}^2 \end{aligned} \quad \square$$

Observaciones respecto a la varianza:

- La varianza, al igual que la media, es un índice muy sensible a las puntuaciones extremas.
- En los casos que no se pueda hallar la media tampoco será posible hallar la varianza.

Definición 1.20:

El **coeficiente de variación**, también denominado como *coeficiente de variación de Pearson*, es una medida estadística adimensional que nos informa acerca de la dispersión relativa de varias distribuciones.

Si los valores sean positivos y su media dé, por tanto, un valor positivo, *a mayor valor del coeficiente de variación mayor heterogeneidad (dispersión) de los valores de la variable*; y a menor C.V., mayor homogeneidad (menor dispersión) en los valores de la variable. Por ejemplo, si el C.V es menor o igual al 80 %, significa que la media aritmética es representativa del conjunto de datos, por ende el conjunto de datos es “*Homogéneo*”. Por el contrario, si el C.V supera al 80 %, el promedio no será representativo del conjunto de datos (por lo que resultará “*Heterogéneo*”).

El coeficiente de variación se define como la razón entre la desviación típica y el valor medio:

$$CV = \frac{s}{\bar{x}} \quad (\bar{x} \neq 0)$$

Ejercicio resuelto 1.3. Calcula las medidas de dispersión para la siguiente distribución estadística.

x_i	n_i	N_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$	$ x_i - \bar{x} $	$ x_i - \bar{x} \cdot n_i$
1	7	7	7	7	1.6	11.2
2	14	21	18	56	0.6	8.4
3	9	30	27	81	0.4	3.6
4	8	38	32	128	1.4	11.2
5	2	40	10	50	2.4	4.8
N=40		Total: 104	Total: 322		Total: 39.2	

— Rango: $R = x_{max} - x_{min} = 5.1 - 1 = 4$

— Rango: $R = x_{max} - x_{min} = 5.1 - 1 = 4$

$$25\%(40) = 10 \rightarrow Q_1 = 2; \quad 75\%(40) = 30 \rightarrow Q_3 = \frac{3+4}{5} = 3.5$$

— Recorrido intercuartílico: $R_I = Q_3 - Q_1 = 3.5 - 2 = 1.5$

$$\bar{x} = \frac{\sum x_i \cdot n_i}{N} = \frac{104}{40} = 2.6$$

- Desviación media: $DM = \frac{\sum_{i=1}^n |x_i - \bar{x}| \cdot n_i}{N} = \frac{39.2}{40} = 0.98$
- Varianza: $s^2 = \frac{\sum_{i=1}^n (x_i)^2 \cdot n_i}{N} - \bar{x}^2 = \frac{332}{40} - 2.6^2 = 1.29$
- Desviación típica: $s = \sqrt{s^2} = \sqrt{1.29} = 1.14$
- Coeficiente de variación (de Pearson): $CV = \frac{s}{\bar{x}} = \frac{1.14}{2.6} = 0.44 = 44\%$

Si los datos estuviesen agrupados en intervalos, procederíamos de la misma forma, considerando como x_i las marcas de clase.

Ejercicio resuelto 1.4. *El cóndor de los Andes tiene una envergadura media (alas extendidas) de 285 cm con una desviación estándar de 30 cm, mientras que una especie de murciélagos tiene una envergadura media de 10 cm y su población presenta una desviación estándar de 3 cm.*

¿Cuál de las dos poblaciones presenta una mayor dispersión en lo que se refiere a su envergadura?

$$\text{Condor: } CV = \frac{s}{\bar{x}} = \frac{30}{285} = 0.11 = 11\%$$

$$\text{Murciélagos: } CV = \frac{s}{\bar{x}} = \frac{3}{10} = 0.30 = 30\%$$

Aunque la desviación típica de la envergadura del cóndor de los Andes es muy superior a la de esa especie de murciélagos, su dispersión es menor.

1.4.4. Parámetros de forma

Se usan para describir numéricamente la forma de la distribución. Miden la *simetría* o sesgo y la *curtosis* o apuntamiento y son parámetros adimensionales.

Definición 1.21:

Índices de asimetría:

- Coeficiente de asimetría de Fisher: $As = \frac{\sum_{i=1}^n (x_i - \bar{x})^3 \cdot n_i}{Ns^3}$

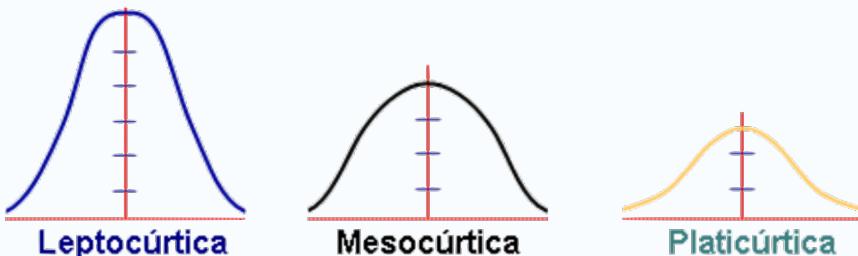
Interpretación de los coeficientes: Si $As > 0$ la distribución presenta una asimetría positiva, si $As < 0$ la asimetría es negativa y si $As \approx 0$ la distribución es simétrica.

**Definición 1.22:**

El **apuntamiento o curtosis** proviene de una comparación de la distribución con la *distribución normal o campana de Gauss* por lo que solo tendrá sentido para las distribuciones que se asemejen a la curva normal (unimodales y prácticamente simétricas). Para su medida se usa el

$$\text{— Coeficiente de apuntamiento de Fisher: } K = \frac{\sum_{i=1}^n (x_i - \bar{x})^4 \cdot n_i}{Ns^4} - 3$$

Si $K < 0$ tenemos una distribución *platicúrtica*, en las colas de la distribución hay más casos que en la curva normal. Si $K > 0$ la distribución es *leptocúrtica*, ocurre lo contrario que en el caso anterior. Para $K \approx 0$ la distribución es como la normal, se llama *mesocúrtica*.

**Ejemplo 1.16:**

Calcula el coeficiente de asimetría y la curtosis para la siguiente distribución.

Clase	x_i	n_i	$x_i \cdot n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2 \cdot n_i$	$(x_i - \bar{x})^3 \cdot n_i$	$(x_i - \bar{x})^4 \cdot n_i$
[45,55[50	6	300	-19,4	2258,16	-43808,304	849881,0976
[55,65[60	10	600	-9,4	883,6	-8305,84	78074,896
[65,75[70	19	1330	0,6	6,84	4,104	2,4624
[75,85[80	11	880	10,6	1235,96	13101,176	138872,4656
[85,95[90	4	360	20,6	1697,44	34967,264	720325,6384
		50	3470		6082	-4041,6	1787156,56

$$\bar{x} = \frac{3470}{50} = 69.4 ; \quad \sigma = \sqrt{\frac{6082}{50}} = 11.0$$

$$As = \frac{-4041.6}{50 \cdot 11.0^3} = -0.06 \simeq -0.1$$

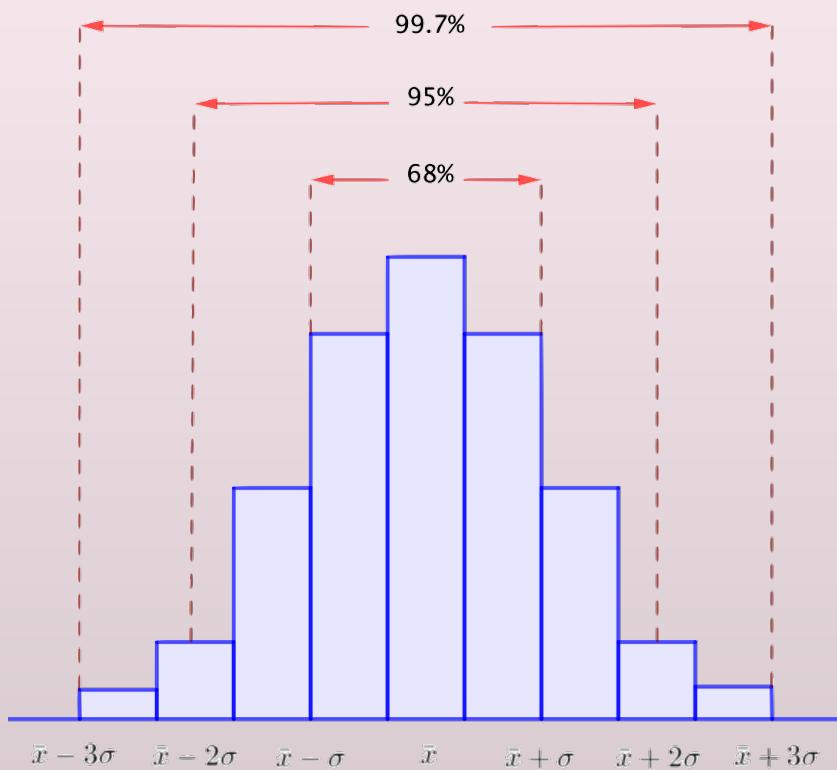
$$K = \frac{1787156.56}{50 \cdot 11.0^4} - 3 = 2.4 - 3 = -0.6$$

La distribución es prácticamente simétrica (ligero sesgo o asimetría negativa o hacia la izquierda) y platicúrtica.

1.4.5. Interpretación conjunta de la media y la desviación típica

\bar{x} y σ .

Cuando la distribución en estudio es “*normal*” (de momento no sabemos que significa esto de ‘normal’ – lo veremos en distribuciones de probabilidad – pero supongamos que son aquellas que no presentan valores ‘extraños’) resulta que datos de la distribución que se alejen una desviación típica del valor medio hay alrededor de los 2/3 del total de datos, es decir, individuos con $x_i \in [\bar{x} - \sigma, \bar{x} + \sigma]$ son aproximadamente el 68 % del total. Los que se alejan dos desviaciones típicas son el 95 % del total y tres desviaciones típicas el 99.7 %.



1.5. Tipificación

Definición 1.23:

Haciendo uso de las propiedades de las medidas estadísticas, podremos facilitar y simplificar los cálculos de parámetros estadísticos mediante un cambio de variable.

Así, si todos los valores son muy altos, podremos restarles una cantidad (normalmente la Moda) y, si poseen cifras decimales o son múltiplos de un mismo número, podremos multiplicarlos o dividirlos por el valor adecuado.

Una vez calculados los parámetros estadísticos, en virtud de las propiedades descritas, obtendremos el valor final real de tales parámetros.

Mención especial merecen dos cambio de variables particular :

1. **Diferenciales:** partiendo de la variable inicial x_i (puntuaciones directas), si a todos los valores les restamos la media, obtenemos una nueva variable $y_i = x_i - \bar{x}$ (puntuaciones diferenciales) cuya media es cero (la desviación típica no se modifica).
2. **Tipificadas:** Si a todos los valores de la variable inicial x_i les restamos la media y el resultado lo dividimos por la desviación típica, obtenemos una nueva variable $z_i = \frac{x_i - \bar{x}}{s_x}$, *puntuaciones típicas o tipificadas* cuya media es cero y cuya desviación típica es siempre la unidad.

Este último cambio de variable recibe el nombre de **TIPIFICACIÓN** y es muy útil a la hora de comparar dos valores de individuos pertenecientes a distribuciones distintas.

Ejemplo 1.17:

Un alumno obtiene en un examen de matemáticas una calificación de 8 en un grupo cuya media es 5 y desviación típica 1. Su amigo, en otro grupo de media 6 y desviación típica 2 obtiene un 10. ¿Cuál de las dos notas, comparativamente a su grupo, es mejor calificación?.

Nos encontramos con dos distribuciones de calificaciones medidas en distintas escalas. Para poder comparar tendremos que referir ambas series de valores a otras equivalentes entre sí (igual media y desviación típica).

El proceso de *tipificación* nos proporciona lo que deseamos, siempre obtendremos una distribución con media 0 y desviación típica 1, donde los datos ya son comparables.

Tipificando ambas calificaciones se obtiene :

$$\text{Nota del primer amigo: } z_1 = \frac{8 - 5}{1} = 3$$

$$\text{Nota del segundo amigo: } z_2 = \frac{10 - 6}{2} = 2$$

La nota obtenida del primer amigo es, comparativamente, superior a la del segundo.

1.6. Ejercicios

Ejercicio 1.1. Considera las siguientes distribuciones:

Distribución 1	
clase	n_i
[10,20[3
[20,30[17
[30,40[13
[40,50[9
[50,60[8

Distribución 2	
clase	n_i
[10,20[5
[20,30[9
[30,40[23
[40,50[9
[50,60[4

- a) Dibuja el histograma de cada distribución y deduce, de su observación, cuál de las dos distribuciones es más dispersa.
- b) Calcula mo , me , Q_1 , D_1 , P_{66} para la primera distribución.
- c) Calcula \bar{x} , \bar{x}_g , \bar{x}_a para la primera distribución.
- d) Calcula el rango, la desviación media y la desviación típica, para la primera distribución.
- e) Calcula los coeficientes de variación de Pearson para ambas distribuciones y deduce, a partir de ellos, cuál es la distribución más dispersa.
- f) Da, para la primera distribución, una interpretación conjunta de media y desviación típica. ¿A partir de qué valor se está entre los 10% valores más altos? ¿Cuántos valores hay menores que 47?
- g) Encuentra los coeficientes de asimetría y de curtosis de ambas distribuciones. ¿Qué puedes decir de la forma de estas distribuciones?
- h) Dibuja el diagrama de cajas y bigotes.
- i) ¿Entre qué valores se encuentra el 20% de los valores centrales?

Cálculos:

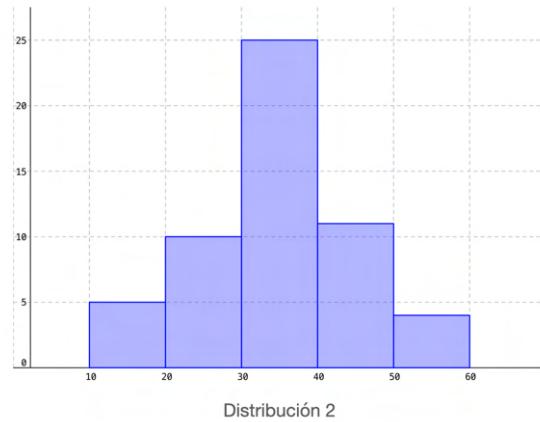
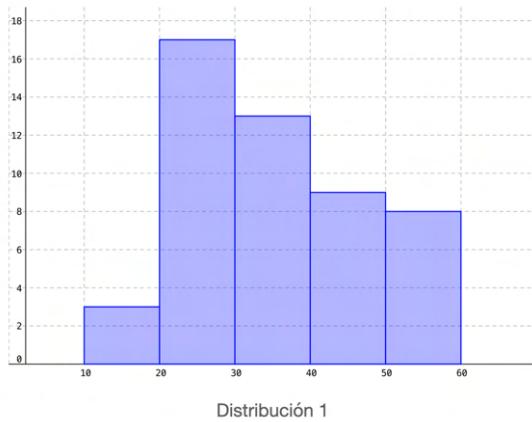
Distribución 1									
x_i	n_i	N_i	$x_i \cdot n_i$	$x_i - \bar{x}$	$ x_i - \bar{x} \cdot n_i$	$(x_i - \bar{x})^2 \cdot n_i$	$(x_i - \bar{x})^3 \cdot n_i$	$(x_i - \bar{x})^4 \cdot n_i$	
15	3	3	45	-20.4	61.2	1248.48	-25468.99	519567.44	
25	17	20	425	-10.4	176.8	1838.72	-19122.69	198875.96	
35	13	33	455	-0.4	5.2	2.08	-0.83	0.33	
45	9	42	405	9.6	86.4	829.44	7962.62	76441.19	
55	8	50	440	19.6	156.8	3073.28	60236.29	1180631.24	
N=50		1770			486.4	6992.00	23606.40	1975516.16	

Distribución 2									
x_i	n_i	N_i	$x_i \cdot n_i$	$x_i - \bar{x}$	$ x_i - \bar{x} \cdot n_i$	$(x_i - \bar{x})^2 \cdot n_i$	$(x_i - \bar{x})^3 \cdot n_i$	$(x_i - \bar{x})^4 \cdot n_i$	
15	5	5	75	-19.6	98	1920.8	-37647.68	737894.53	
25	9	14	225	-9.6	86.4	829.44	-7962.62	76441.19	
35	23	37	805	0.4	9.2	3.68	1.47	0.59	
45	9	46	405	10.4	93.6	973.44	10123.78	105287.27	
55	4	50	220	20.4	81.6	1664.64	33958.66	692756.58	
N=50		1730			368.8	5392.00	-1526.40	1612380.16	

Para el cálculo de las desviaciones es necesario, previamente, calcular la medias:

$$\bar{x}_1 = \frac{1770}{50} = 34,5; \quad \bar{x}_2 = \frac{1730}{50} = 34,6$$

— a) Histogramas:



A simple vista, la distribución 1 parece más dispersa que la distribución 2. En el apartado e) lo veremos con más precisión al calcular los respectivos coeficientes de variación (Pearson).

— b) mo, me, Q_1, D_1, P_{66} para la primera distribución:

El intervalo modal, de mayor frecuencia absoluta, es el $[20, 30]$, su frecuencia absoluta es $n_k = 17$ y las frecuencias absolutas de los intervalos anterior y posterior son, respectivamente, $m_{k-1} = 3$ y $n_{k+1} = 13$. La amplitud del intervalo modal es $c_k = 10$, como la de todos ellos.

$$\Rightarrow mo = 20 + \frac{17 - 3}{(17 - 3) + (17 - 13)} 10 = 27.8$$

Calculamos la moda sin usar la fórmula:

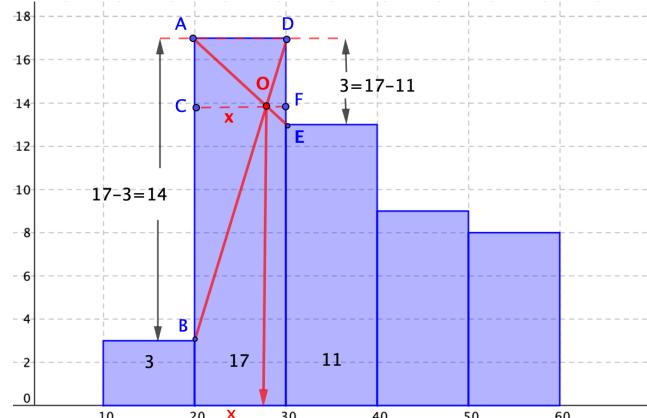
Los triángulos $AOB \sim DOE$ son semejantes.

$$\frac{AB}{OC} = \frac{DE}{OF}$$

$$\frac{14}{x} = \frac{4}{10-x}$$

$$x = 7.8 \Rightarrow$$

$$mo = 20 + 7.8 = 27.8$$



El intervalo mediano es el tercero, $[30, 40]$ de frecuencia absoluta $n_k = 13$ y cuya frecuencia acumulada del intervalo anterior es $N_{k-1} = 20$. La amplitud de éste y todos los intervalos es $c_k = 10$.

$$\Rightarrow me = 30 + \frac{\frac{50}{2} - 20}{13} 10 = 33.8$$

Calculamos la mediana sin usar la fórmula:

Los triángulos $AOB \sim ADC$

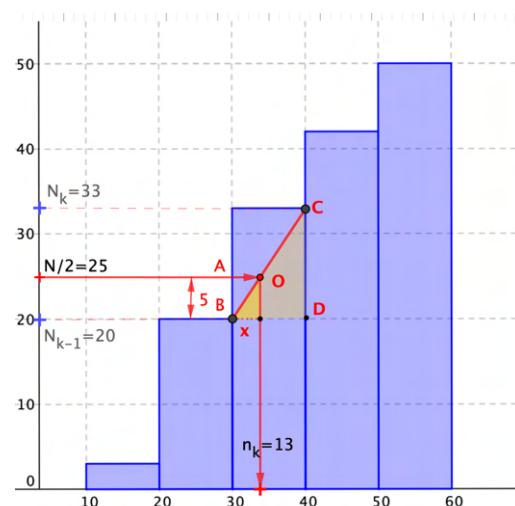
son *semejantes*.

$$\frac{OB}{AB} = \frac{DC}{AC}$$

$$\frac{5}{x} = \frac{13}{10}$$

$$x = 3.8 \Rightarrow$$

$$mo = 30 + 3.8 = 33.8$$



$$\Rightarrow 25\% (50) = 12.5 \rightarrow Q_1 = 20 + \frac{12.5 - 3}{17} 10 = 25.6$$

$$\Rightarrow 10\% (50) = 5 \rightarrow D_1 = 20 + \frac{5 - 3}{17} 10 = 21.2$$

$$\Rightarrow 65\% (50) = 32.5 \rightarrow P_{66} = 30 + \frac{32.5 - 20}{13} 10 = 39.6$$

— c) $\bar{x}, \bar{x}_g, \bar{x}_a$ de distribución-1.

Distribución 1

x_i	n_i	$x_i \cdot n_i$	$x_i^{n_i}$	n_i/x_i
15	3	45	3.30E+03	0.20
25	17	425	5.82E+23	0.68
35	13	455	1.18E+20	0.37
45	9	405	7.57E+14	0.20
55	8	440	8.37E+13	0.15
N=50		1770	$\Pi = 1.47E+76$	1.60

$$\bar{x} = \frac{1770}{50} = 35.4; \quad \bar{x}_g = \sqrt[50]{1.47 \times 10^{76}} = 33.4; \quad \bar{x}_a = \frac{50}{1.60} = 31.3$$

— d) R, DM, s , de distribución-1.

$$\Rightarrow R = 60 - 10 = 50$$

$$\Rightarrow DM = \frac{486.4}{50} = 9.7$$

$$\Rightarrow s = \sqrt{\frac{6992.00}{50}} = 11.83$$

— e) CV para ambas distribuciones:

$$\bar{x}_1 = 35.5; \quad s_1 = 11.83; \quad x_2 = 34.6; \quad s_2 = \sqrt{\frac{5392}{50}} = 10.38$$

$$\Rightarrow CV_1 = \frac{\bar{s}_1}{x_1} = 0.343 = 34.3\%$$

$$\Rightarrow CV_2 = \frac{\bar{s}_2}{x_2} = 0.300 = 30.0\%$$

Como habíamos predicho en el apartado a) la primera distribución tiene los datos más dispersos.

— f) $n_i \in [\bar{x} - \sigma, \bar{x} + \sigma]$; $10\% \uparrow \rightarrow P_{90}$; ¿cuantos individuos tiene $x_i < 47$?

En el intervalo $[\bar{x} - \sigma, \bar{x} + \sigma] = [22.67, 46, 33]$ están todos los individuos del intervalo $[30, 40[$, esto es, 13 individuos más la parte proporcional de individuos de cada uno de los intervalos adyacentes.

En $[20, 30[$ hay 17 individuos, para una amplitud de 10. En $[22.67, 30[$, de amplitud 7.33 habrán: $17 \frac{7.33}{10} = \underline{12.46}$

\Rightarrow En $[40, 50[$ hay 9 individuos, para una amplitud de 10. En $[40, 46.33[$, de amplitud 6.33 habrán: $9 \frac{6.33}{10} = \underline{5.70}$

Por lo que, en $[\bar{x} - \sigma, \bar{x} + \sigma] = [22.67, 46, 33]$ hay $12.46 + 13 + 5.70 = 31.16$, lo que significa una proporción del $\frac{31.16}{50} = 62.32\%$ de la población. (Esta es la interpretación conjunta de media y desviación típica).

Estar entre el 10% de los individuos de mayor puntuación es lo mismo que dejar atrás al 90% de ellos, nos piden en percentil 90.

$90\%(50) = 45 \rightarrow P_{90}$ estamos en el último intervalo:

$$\Rightarrow P_{90} = 50 + \frac{45 - 42}{8} 10 = 53.7$$

$47 \in [40, 50]$ luego con $x_i < 47$ estarán todos los individuos de los intervalos anteriores, $3 + 17 + 13$, y la parte proporcional de los 9 individuos del intervalo de amplitud 10, $[40, 50[$:

$$\Rightarrow 3 + 17 + 13 + \frac{47-40}{9} 10 = 40.78, \text{ hay } 40 \text{ individuos con puntuación menor que } 47.$$

— g) $As_1, K_1; As_2, K_2$

$$As_1 = \frac{23606.40}{5011.83^3} = 3.51$$

$$As_2 = \frac{-1526.40}{5010.38^3} = -0.03$$

$$K_1 = \frac{1975516.16}{50 \cdot 11.83^4} - 3 = -0.98$$

$$K_2 = \frac{1612380.16}{50 \cdot 10.38^4} - 3 = -0.22$$

La primera distribución está sesgada a la derecha (positivamente) y más platicúrtica que la segunda.

— g) Diagrama de cajas y bigotes.

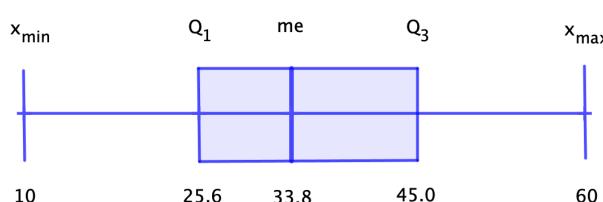
En apartados anteriores obtuvimos: $me = 33.8$; $Q_1 = 25.6$; $x_{min} = 10$; $x_{max} = 60$

$$75\%(50) = 37.5 \rightarrow Q_3 = 40 + \frac{37.5 - 33}{9} 10 = 45.0$$

$$\text{Valores atípicos: } Lim_{inf} = 25.6 - 3 \frac{45 - 25.6}{2} = -17.0; Lim_{sup} = 45.0 - 3 \frac{45 - 25.6}{2} = 87.6$$

Fuera del intervalo $[-17.0, 87.6[$ no hay ningún valor, por lo que no existen valores atípicos para esta distribución.

Diagrama de cajas y bigotes:



Se observa un cierto sesgo a la derecha.

— i) $]P_{40}, P_{60}[$

El valor central es la mediana, el P_{50} ; el 20 % de los valores a su alrededor estarán comprendidos entre el P_{40} y el P_{60}

$$40\% (40) = 20 \rightarrow P_{40} = 30$$

$$60\% (40) = 30 \rightarrow P_{60} = 30 + \frac{30 - 20}{13} 10 = 37.69$$

El intervalo pedido es $[30, 37.69]$

Ejercicio 1.2. Para los datos 7, 3, 4, 6, 4, 3, 3; calcula la media aritmética, geométrica y armónica.

x_i	n_i	$x_i \cdot n_i$	$x_i^{n_i}$	n_1/x_i
3	3	9	27	1.00
4	2	8	16	0.50
6	1	6	6	0.17
7	1	7	7	0.14
$N = 8$		$\Sigma \rightarrow 30$	$\Pi \rightarrow 18144$	$\sigma \rightarrow 1.81$

$$\bar{x} = \frac{\sum_{i=1}^4 x_i \cdot n_i}{N} = 3.75; \quad \bar{x}_g = \sqrt[8]{\prod_{i=1}^4 x_i^{n_i}} = 3.41; \quad \bar{x}_a = \frac{N}{\sum_{i=1}^4 \frac{n_i}{x_i}} = 4.42$$

Diferencia entre variable discreta y continua a la hora del cálculo de parámetros de posición.

En variable discreta para encontrar la posición del individuo adecuado usaremos una aproximación por defecto.

Ejercicio 1.3. Para las dos siguientes distribuciones, encontrar el P_{33}

Distribución 1		
x_i	n_i	N_i
3	3	3
5	10	10
7	15	25
9	7	35
$N=35$		

Distribución 2			
clase	x_i	n_i	N_i
[2,4[3	3	3
[4,6[5	10	10
[6,8[7	15	25
[8,10[9	7	35
$N=25$			

Para la distribución 1: $33\%(35) = 11.55 \rightarrow 12^{\text{o}}$ individuo deja tras de sí al (más) 33 % de los individuos de la población.

El valor de la variable para este 12^{o} individuo es $P_{33} = 7$

Para la distribución 2: $33\%(35) = 11.55 \rightarrow P_{33} = 6 + \frac{11.55 - 10}{15} 2 = 6.21$

Ejercicio 1.4. Una variable estadística X tiene una media $\bar{x} = 20$ y una desviación típica $s_x = 5$. Si elevamos al cuadrado todos los datos, ¿cuál será ahora el valor de la media?

$$x_i \rightarrow y_i = x_i^2 \Rightarrow s_y^2 = \frac{\sum n_i \cdot x_i^2}{N} - \bar{x}^2 = \bar{y} - \bar{x}^2$$

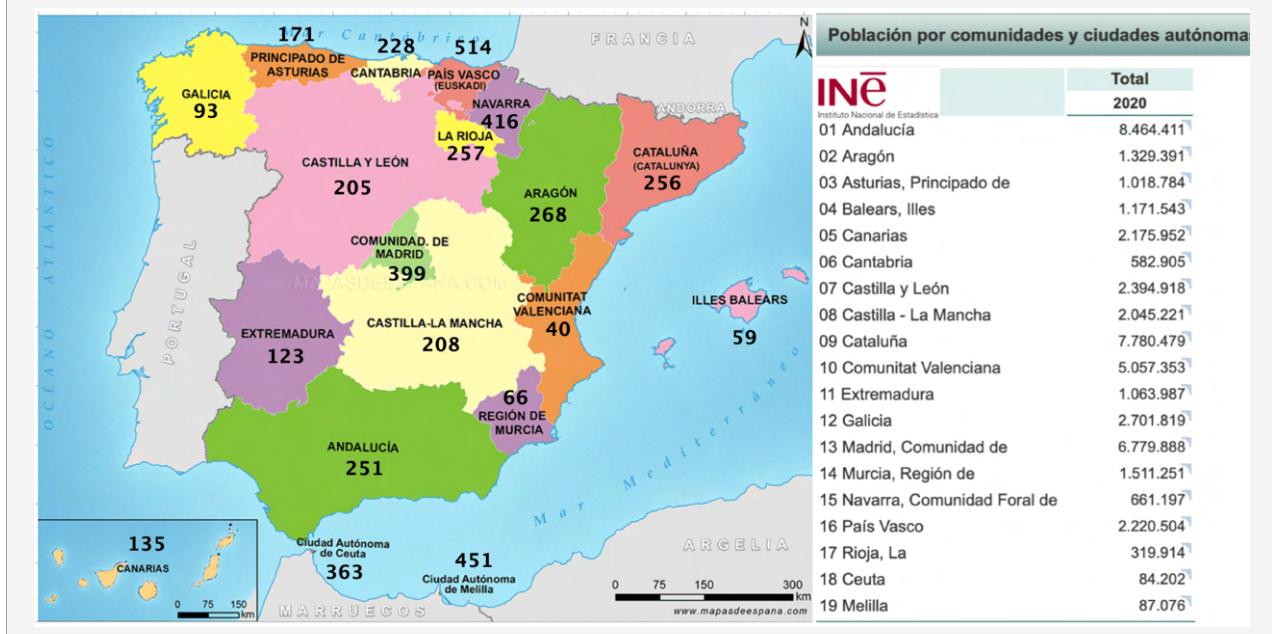
$$\bar{y} = \bar{x}^2 + s^2 x \rightarrow \bar{y} = 20^2 + 5^2 = 425$$

Ejercicio 1.5. Una variable estadística tiene $\bar{x} = 8$ y $s_x = 2$. ¿Qué transformación hay que hacer a la variable para que su nueva media sea $\bar{y} = 42$ y la nueva desviación típica $s_y = 10$

El cambio lineal general es: $y_i = a + bx_i \vee Y = a + bX$

$$\begin{cases} \bar{y} = a + bx \\ s_y = bs_x \end{cases} \rightarrow \begin{cases} 42 = a + 8b \\ 10 = 2b \end{cases} \rightarrow b = 5; a = 2 \Rightarrow Y = 2 + 5X \text{ or } y_i = 2 + 5x_i$$

Ejercicio 1.6. Calcula la tasa covid 14 días para todo el territorio español → “media ponderada”.



En el mapa (cartograma) aparece la incidencia acumulada de contagios en 14-días por cada 100.000 habitantes, a la derecha se muestra la población de cada comunidad autónoma.

Tomaremos como variable x_i la incidencia acumulada por 100K habitantes de cada una de las 19 comunidades y ciudades autónomas de España y como frecuencia absoluta de cada una, n_i , la población. Así,

$$\bar{x}_p = \sum_{i=1}^{19} x_i \cdot n_i / \sum_{i=1}^{19} n_i = 232.2$$

Todo lo visto en este tema se puede hacer, más rápidamente, con el múltiple sw existente, pero se ha optado por presentar todos los cálculos a mano por motivos didácticos.

1.7. Curiosidades

Cómo mentir con estadísticas

“Hay tres tipos de mentiras: mentiras, malditas mentiras y estadísticas”^a; popularizada en los Estados Unidos por Mark Twain, esta frase describe el abuso de estadísticas para reforzar argumentos falaces.

Un buen gráfico permite sintetizar información numérica, es difícil no convencerse de cualquier cosa cuando se acompaña la idea con gráficos.

La misma información se puede utilizar para crear dos gráficos con mensajes totalmente opuestos. Detrás de una estadística hay, sin duda, mucho más allá que figuras representativas de números, y es debido a esto que hoy existen expertos en el “entendimiento de riesgo”, estadísticos que trabajan con matemáticos, psicólogos y otros profesionales para transmitir al público de forma precisa sus hallazgos.

“Los gráficos pueden ser tan manipuladores como las palabras”, por eso, ser capaces de leer, entender y evaluar estadísticas es una habilidad fundamental y estar atentos a los trucos que pueden hacernos caer en un error o engaño es una necesidad.

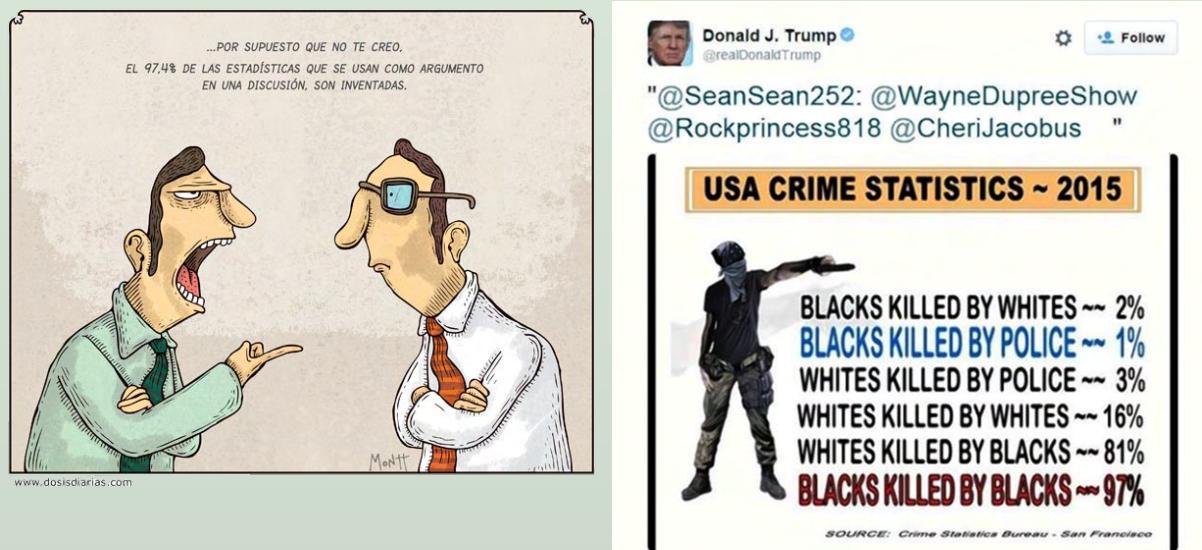
Algunos casos de mentiras estadísticas, que rescatan los matemáticos Andrew Gelman y Deborah Nolan en su libro *Teaching Statistics: A Bag of Tricks*, son:

1. Inventar cifras

Según un estudio de la Universidad de Cambridge, un 53,4% de las estadísticas son inventadas. Suena inteligente, ¿no?. Inventar estadísticas es la forma más directa de engaño de la que hablaremos y es, lamentablemente, más común de lo que se pensaría, gracias al aura de verdad o seriedad que un porcentaje o un grupo de barras puede darle al mensaje que lo acompaña.

Uno de los ejemplos más recientes y notables por su impacto, y que afortunadamente fue desbaratado con rapidez, fue un tuit publicado por el candidato republicano Donald Trump.

En diciembre del año pasado, Trump retuiteó estadísticas que señalaban, entre otras cosas, que un 81% de los asesinatos de gente blanca era cometidos por gente negra.



Una rápida consulta de las cifras oficiales del FBI arrojó que el porcentaje, en realidad, asciende a un 15% y que la supuesta fuente, la Oficina de Estadísticas Criminales de San Francisco, jamás ha existido.

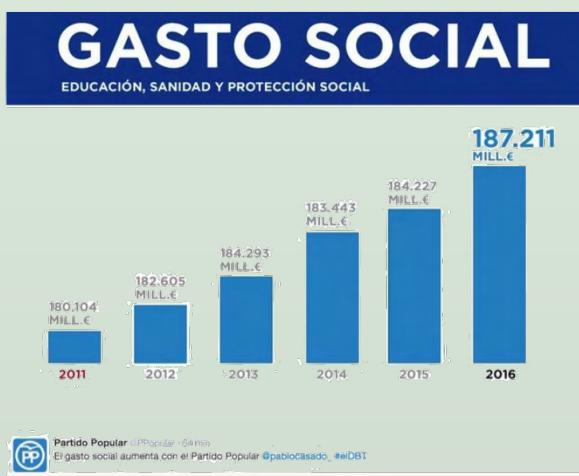
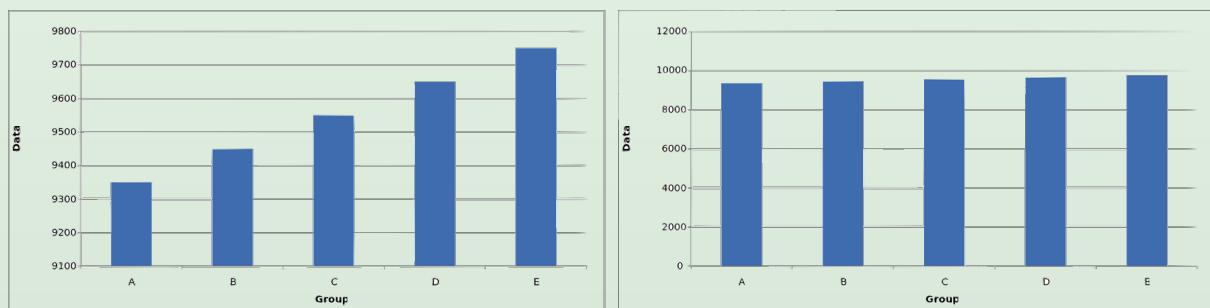
Si bien Twitter, y en especial Trump, no dejan la vara muy alta en cuanto a veracidad de información se trata, no deja de ser impactante el peso que una estadística totalmente inventada de una oficina que jamás existió pudo tener en un mensaje.

La lección es clara y se aplica a todo tipo de información sensible: *hay que revisar la fuente de cualquier estadística que nos presenten.*

2. Cortar ejes

En el colegio nos enseñan que los ejes, tanto el “X” (horizontal) como el “Y” (vertical), parten de 0, pero esto no siempre ocurre en los gráficos. No es algo prohibido ni malo necesariamente, pero se presta para exagerar o minimizar deliberadamente comparaciones de cifras.

Existen muchas formas de alterar ejes, pero la más común es la que se aprecia en las siguientes imágenes. Nótese que se trata de exactamente los mismos datos, solo que en el primero buena parte del eje Y (el vertical) fue truncado para hacer, suponemos, leña (;-). Así, la diferencia entre cada barra se ve enorme, cuando al verla en el contexto total (segundo gráfico), la diferencia es mínima.



Lo mismo ocurre con las secuencias de tiempo (habitualmente situadas en el eje X). El crecimiento o disminución de la delincuencia, economía, desempleo, IPC, etc. puede verse impresionante o irrelevante según qué período se tome. Habitualmente los políticos seleccionan los marcos de tiempo que mejor hacen ver su gestión.

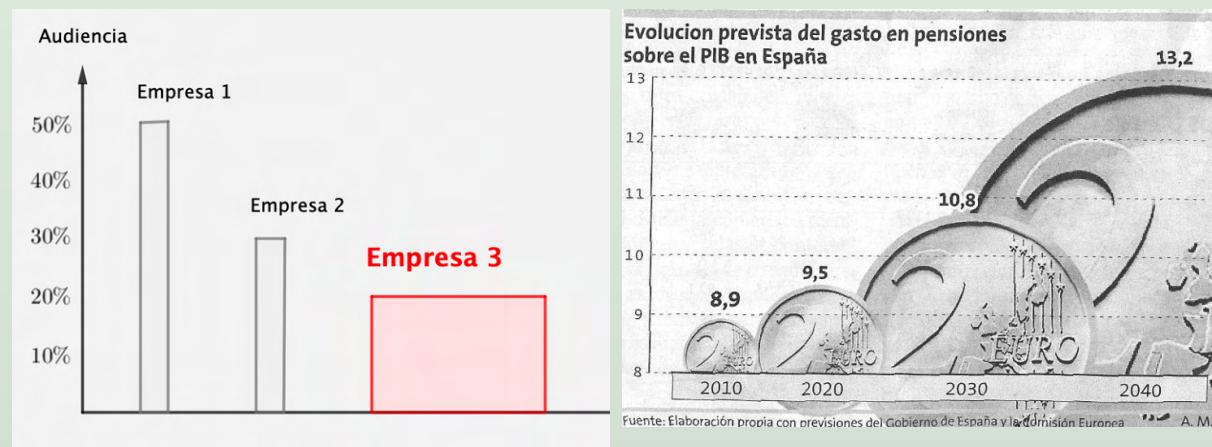
La próxima vez que veas uno de estos gráficos *¡Asegúrate que partan de 0! O al menos, que consideren un marco de referencia suficientemente amplio para tener claro cuán relevante es la estadística.*

3. Otras representaciones erróneas

En el siguiente gráfico, a igualdad de frecuencia relativa para los sectores rojo y morado, el efecto 3D hace que al aparecer el sector morado más próximo al observador aumente su proporción respecto al sector rojo. También aumenta el efecto aparente de tamaño el hecho de extraer un sector de la tarta (el verde).



Uno de los más frecuentes errores de diseño en gráficos, y que ocurre hasta en los diarios y revistas más prestigioso, es usar dibujos con área o volumen, para representar un dato que debería tener un solo eje. Aunque se hace con un objetivo decorativo, el lector no sabe si debe comparar sólo la altura, el área o el volumen del objeto.



Aunque parezcan muchas cosas de las que estar preocupado, en realidad se trata de tener la misma actitud que frente a un estudio: *ser escéptico*. En otras palabras, verificar que la fuente sea fiable (y que exista), que sus elementos tengan las proporciones correctas, y que la información que transmite (los números) sea la misma que nos dicen que transmite con los gráficos que la acompañan.

^aBasado en el artículo de Francisco J. Lastra

RESUMEN Estadística descriptiva unidimensional

x_i	n_i	f_i	N_i	F_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
	N	1			$\sum \rightarrow \bar{x}$	$\sum \rightarrow s$

▷ Media(s):

$$\bar{x} = \frac{\sum_{i=1}^n x_i \cdot n_i}{N} \quad \bar{x}_p = \frac{\sum x_i \cdot p_i}{\sum p_i} \quad \bar{x}_g = \sqrt[N]{\prod x_i^{n_i}} \quad \bar{x}_a = \frac{N}{\sum \frac{x_i}{n_i}}$$

▷ Moda:

$$mo = L_k + \frac{n_k - n_{k-1}}{(n_k - n_{k-1}) + (n_k - n_{k+1})} \cdot c_k$$

▷ Mediana, Cuartiles, Deciles y Percentiles:

$$me = L_k + \frac{\frac{N}{2} + N_{k-1}}{n_k} \cdot c_k ; \quad Q_i = L_k + \frac{i \cdot \frac{N}{4} + N_{k-1}}{n_k} \cdot c_k; \quad i = 1, 2, 3$$

$$D_i = L_k + \frac{i \cdot \frac{N}{10} + N_{k-1}}{n_k} \cdot c_k; \quad i = 1, 2, \dots, 9; \quad P_i = L_k + \frac{i \cdot \frac{N}{100} + N_{k-1}}{n_k} \cdot c_k; \quad i = 1, 2, \dots, 99$$

▷ Varianza y desviación típica:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot n_i}{N} = \frac{\sum x_i^2 \cdot n_i}{N} - \bar{x}^2 \quad s = \sqrt{s^2}$$

▷ Coeficiente de variación (Pearson) y tipificación:

$$CV = \frac{s}{\bar{x}} \quad z_i = \frac{x_i - \bar{x}}{s}$$

Capítulo 2

Distribuciones bidimensionales: Correlación y Regresión lineal

2.1. Distribuciones estadísticas bidimensionales

Definición 2.1:

Si de una determinada muestra o población se estudian dos caracteres simultneamente, se obtienen dos series de datos de la forma:

x_i	x_1	x_2	x_3	\dots	\dots
y_i	y_1	y_2	y_3	\dots	\dots
n_i	n_1	n_2	n_3	\dots	\dots

$\rightarrow \Sigma N$

Donde x_i e y_i son las variables estadísticas correspondientes a las características observadas y n_i es la frecuencia absoluta de cada pareja de datos (x_i, y_i) , es decir, n_i es el número de veces que se repite el par (x_i, y_i) .

A la lista de estos pares ordenados de datos con sus respectivas frecuencias se le llama **variable estadística bidimensional**, (x_i , y_i ; n_i).

La frecuencia relativa del par (x_i, y_j) es $f_{ij} = n_{ij}/N$, donde N denota el número total de pares observados.

Las **variables marginales**^a X e Y toman los valores $\{x_1, x_2, \dots, x_m\}$, $\{y_1, y_2, \dots, y_r\}$ para ciertos m , r , respectivamente, y la variable bidimensional (X, Y) tomará los valores $\{(x_i, y_j)\} \quad 1 \leq i \leq m, \quad 1 \leq j \leq r$.

^aLas distribuciones marginales son las distribuciones unidimensionales que nos informan del número de observaciones para cada valor de una de las variables, prescindiendo de la información sobre los valores de las demás variables.

Ejemplo 2.1:

Se muestran algunos ejemplos de variables bidimensionales:

- (X, Y) (sexo, color del ojos); X cualitativa; Y cualitativa.
- (X, Y) (nivel de estudios, lustros en la empresa); X cualitativa; Y cuantitativa (discreto -años-).
- (X, Y) (número de hermanos, número de hijos); X cuantitativa (v.e. discreta); Y cuantitativa (v.e. discreta).
- (X, Y) (peso, estatura); X cuantitativa (v.e. continua); Y cuantitativa (v.e. continua).

En este tema nos interesarán las variables estadísticas bidimensionales en que ambas sean cuantitativas.

Teorema 2.1:

Nótese que:

$$\sum_{i=1}^m \sum_{j=1}^r n_{ij} = N; \quad \sum_{i=1}^m \sum_{j=1}^r f_{ij} = 1$$

2.1.1. Distribuciones de frecuencias**Definición 2.2:**

Se define la **frecuencia (absoluta) marginal** del valor x_i como la suma de las frecuencias correspondientes a los pares (x_i, y_j) , para $1 \leq j \leq r$:

Análogamente se define la frecuencia (absoluta) marginal del valor y_j :

$$n_{x_i} = \sum_{j=1}^r n_{ij}; \quad n_{y_j} = \sum_{i=1}^m n_{ij}$$

Teorema 2.2:

Nótese que n_{x_i} (respectivamente, n_{y_j}) representa el número de veces que aparece el valor x_i (respectivamente, y_j) en el total de pares obtenidos. Se verifica:

$$\sum_{i=1}^m n_{x_i} = \sum_{i=1}^m \sum_{j=1}^r n_{ij} = N; \quad \sum_{j=1}^r n_{y_j} = \sum_{j=1}^r \sum_{i=1}^m n_{ij} = N$$

A partir de las frecuencias absolutas marginales se obtienen las frecuencias relativas marginales.

Definición 2.3:

Se definen las *frecuencias relativas marginales* como:

$$f_{x_i} = \frac{n_{x_i}}{N}; \quad f_{y_j} = \frac{n_{y_j}}{N}$$

Teorema 2.3:

Las frecuencias relativas marginales cumplen:

$$\sum_{i=1}^n f_{x_i} = \frac{1}{N} \sum_{i=1}^m n_{x_i} = 1 \quad \sum_{i=1}^n f_{y_j} = \frac{1}{N} \sum_{j=1}^r n_{y_j} = 1$$

Todos estos datos se pueden representar en una **“tabla de doble entrada”** (así se entenderá mejor):

Y / X	$\mathbf{y_1}$	$\mathbf{y_2}$	\cdots	$\mathbf{y_r}$	marg X	rel X
$\mathbf{x_1}$	n_{11}	n_{12}	\cdots	n_{1r}	n_{x_1}	f_{x_1}
$\mathbf{x_2}$	n_{21}	n_{22}	\cdots	n_{2r}	n_{x_2}	f_{x_2}
\cdots	\cdots	\cdots	\cdots	\cdots	\cdots	\cdots
$\mathbf{x_m}$	n_{m1}	n_{m2}	\cdots	n_{mr}	n_{x_m}	f_{x_m}
marg Y	n_{y_1}	n_{y_2}	\cdots	n_{y_r}	\mathbf{N}	
rel. Y	f_{y_1}	f_{y_2}	\cdots	f_{y_r}		$\mathbf{1}$

En la práctica, alguna de las n_{ij} puede ser cero (la podemos dejar en blanco).

Ejemplo 2.2:

Construir una tabla simple y una de doble entrada para los siguientes datos de alturas, x_i (m) y pesos y_i (m).

(1.55,65), (1.62,71), (1.73,75), (1.77,72), (1.72,81), (1.58,79), (1.68,72), (1.85,85), (1.77,77), (1.85,83), (1.67,81), (1.68,78), (1.79,78), (1.62,69)

Agrupamos los datos en intervalos: $X \rightarrow [150, 160[, [160, 170[, [170, 180[$ e $Y \rightarrow [60, 70[, [70, 80[, [80, 90[$. Vamos poniendo cada dato en la casilla correspondiente. (Puede ponerse la X/Y en vertical u horizontal en la tabla de doble entrada, como se desee). Para cálculos posteriores se trabajará con las *marcas de clase*.

$\downarrow X / Y \rightarrow$	[60,70[→ 65	[70,80[→ 75	[80,90[→ 85	n_{x_i}
[150,160[→ 155	1	1	0	2
[160,170[→ 165	1	3	1	5
[170,180[→ 175	0	4	1	5
[180,190[→ 185	0	0	2	2
n_{y_j}	2	8	4	14

Para la tabla de simple entrada, o bien leyendo los datos originales, o bien los de la tabla de doble entrada, los escribimos ahora en horizontal y con las variables representadas por sus marcas de clase (leemos por filas en la tabla de doble entrada).

x_i	155	155	165	165	165	175	175	185	
y_i	65	75	65	75	85	75	85	85	
n_i	1	1	1	3	1	4	1	2	→ 14

Para trabajar numéricamente con la tabla es mejor la forma sencilla que la de doble entrada.

En las tablas de doble entrada, cuando algún valor de los n_{ij} es cero, puede dejarse la celda en blanco.

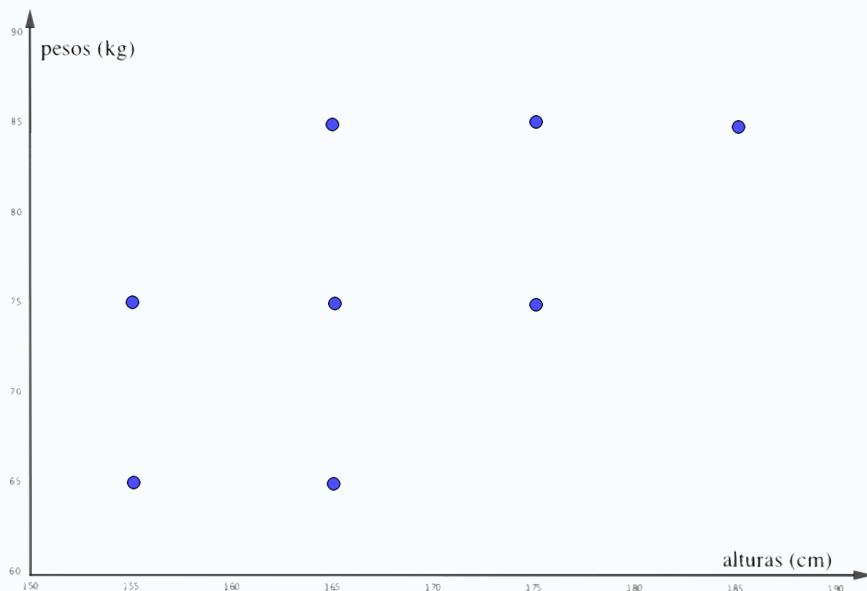
$\downarrow X / Y \rightarrow$	[60,70] → 65	[70,80] → 75	[80,90] → 85	n_{x_i}
[150,160] → 155	1	1		2
[160,170] → 165	1	3	1	5
[170,180] → 175		4	1	5
[180,190] → 185			2	2
n_{y_j}	2	8	4	14

2.1.2. Representación gráfica de variables bidimensionales

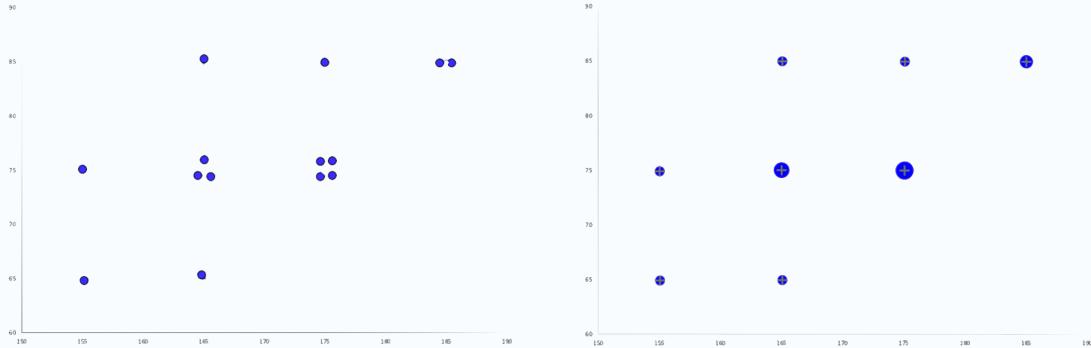
Definición 2.4:

Si en la variable estadística bidimensional todas las frecuencias absolutas de los pares de valores son 1, basta con representar los pares de valores en un diagrama X-Y. Es el llamado **diagrama de dispersión** o **nube de puntos**.

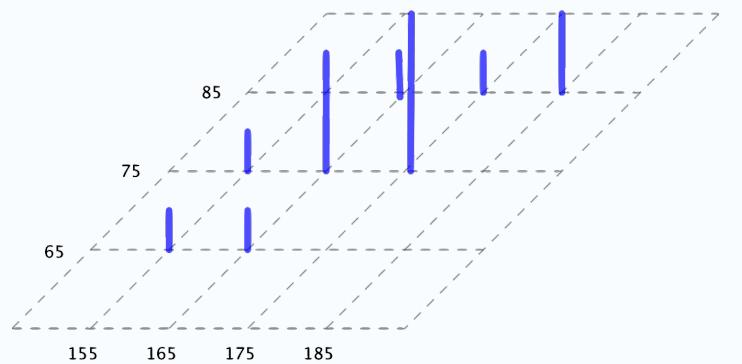
Siguiendo con la distribución del ejemplo anterior,



Si cada par de variables tiene su propia frecuencia tenemos dos soluciones: o bien dibujamos tantos puntos próximos entre sí como indique la frecuencia absoluta, o bien dibujamos para cada pareja de valores círculos de área proporcional a la frecuencia absoluta de los mismos ($A = \pi r^2 \propto n_{ij} \rightarrow r \propto \sqrt{n_{ij}}$; — \propto es el símbolo de proporcionalidad).



Otra posibilidad es usar el *prismograma*, es una representación 3D, en la base se representa la variable X-Y, (x_i, y_i) , y como altura la frecuencia absoluta de cada valor de la variable bidimensional, n_{ij} , formando una barra o un prisma (todos de la misma base).



2.1.3. Parámetros de las distribuciones bidimensionales

Definición 2.5:

Considerando las variables X e Y por separado, como variables unidimensionales, tenemos las **variables marginales**.

En la tabla de doble entrada, la última fila y la última columna son las frecuencias absolutas de las variables marginales. En las tablas de simple entrada, para trabajar con las variables marginales, basta con obviar cualquiera de ellas dos para considerar la otra.

Para cada una de las variables **marginales** podemos calcular su **valor medio**, su **varianza** y su **desviación típica**. Así, tendremos:

$$X \rightarrow \bar{x}; \quad s_x^2; \quad s_x; \quad Y \rightarrow \bar{y}; \quad s_y^2; \quad s_y$$

Definición 2.6:

Covarianza. Se define la *covarianza* de una distribución bidimensional, s_{xy} , a la media aritmética de los productos de las desviaciones da cada uno de los valores de la variable bidimensional respecto a sus respectivos valores medios:

$$s_{xy} = \frac{\sum_{i=1}^m \sum_{j=1}^r (x_i - \bar{x}) \cdot (y_i - \bar{y}) \cdot n_{ij}}{N} = \frac{\sum_{i=1}^m \sum_{j=1}^r x_i y_i n_{ij}}{N} - \bar{x} \cdot \bar{y}$$

La última igualdad es fácilmente demostrable sin más que desarrollar el producto que aparece en la definición de covarianza.

La covarianza nos proporciona información acerca del grado de dependencia existente entre las variables. El signo de la covarianza nos proporciona información sobre el sentido de esa dependencia:

- Si la covarianza es positiva las dos variables varían en el mismo sentido; si aumenta una, aumentará la otra ($X \uparrow \Rightarrow Y \uparrow$).
- Si la covarianza es negativa las dos variables varían en sentido contrario; si aumenta una, disminuirá la otra ($X \uparrow \Rightarrow Y \downarrow$).

Ejemplo 2.3:

Vamos a realizar, a mano, todos los cálculos para la distribución de los ejemplos anteriores. Como dijimos en el tema de estadística, también para estadística bidimensional, con el sw. apropiado o calculadora adecuada, todos estos cálculos se obtienen muy rápidamente. Usamos el método manual para reforzar el significado de los parámetros.

$\downarrow X / Y \rightarrow$	[60,70[→ 65	[70,80[→ 75	[80,90[→ 85	n_{xi}
[150,160[→ 155	1	1	0	2
[160,170[→ 165	1	3	1	5
[170,180[→ 175	0	4	1	5
[180,190[→ 185	0	0	2	2
n_{yj}	2	8	4	14

x_i	y_i	n_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$	$y_i \cdot n_i$	$y_i^2 \cdot n_i$	$x_i \cdot y_i \cdot n_i$
155	65	1	155	24025	65	4225	10075
155	75	1	155	24025	75	5625	11625
165	65	1	165	27225	65	4225	10725
165	75	3	495	81675	225	16875	37125
165	85	1	165	27225	85	7225	14025
175	75	4	700	122500	300	22500	52500
175	85	1	175	30625	85	7225	14875
185	85	2	370	68450	170	14450	31450
		14	2380	405750	1070	82350	182400

$$\bar{x} = \frac{2380}{14} = 170$$

$$\bar{y} = \frac{1070}{14} = 76.43$$

$$s_x = \sqrt{\frac{405750}{14} - 170^2} = 9.06$$

$$s_y = \sqrt{\frac{82350}{14} - 76.43^2} = 6.37$$

$$s_{xy} = \frac{182400}{14} - 170 \cdot 76.43 = 35.47$$

2.2. Dependencia o Correlación

En cursos de Análisis Matemático se estudian relaciones *determinísticas*, si dos variables x e y están relacionadas de esa manera, sabiendo el valor de x, podemos conocer exactamente el valor de y.

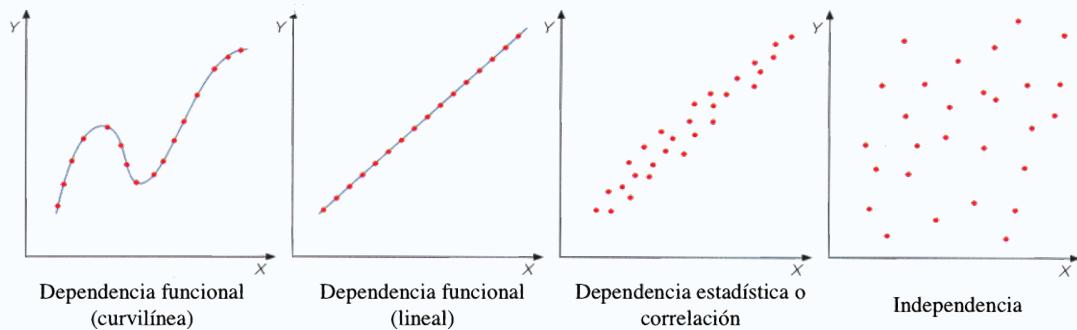
En muchas aplicaciones encontramos variables que parecen estar relacionadas, aunque no de manera determinística. Esto significa que, saber el valor de una variable (la variable explicativa) no nos permite conocer exactamente el valor de la otra (la variable respuesta), ya que ésta puede considerarse una variable aleatoria. Por ejemplo, sean ‘x’ la edad de un niño e ‘y’ su talla, sabemos que la talla depende de la edad, pero también depende de muchas otras condiciones; para una determinada edad, la talla puede considerarse como una variable aleatoria. Cuando ocurre esto decimos que las variables están **correlacionadas**, son estadísticamente dependientes.

Definición 2.7:

Entre dos variables estadísticas existe **dependencia estadística o correlación** cuando los valores que toma una de ellas están relacionados con los valores que toma la otra, pero no de manera exacta.

La relación existente entre dos variables queda reflejada en los diagramas de dispersión o nubes de puntos de la distribución bidimensional.

- Si los puntos de la nube se sitúan sobre una recta o una curva cuya expresión matemática podemos determinar, hablaremos de dependencia funcional entre las variables X e Y.
- Si los puntos de la nube se agrupan en torno a una posible recta, o curva, no muy definida pero reconocible, hablaremos de dependencia estadística o correlación entre las variables X e Y.
- Si los puntos de la nube no se agrupan en torno a ninguna curva, están completamente en desorden, hablaremos de independencia entre las variables X e Y.



https://yoquieroaprobar.es/_pdf/04140.pdf

2.3. Correlación lineal

El caso más simple que estudiaremos en este capítulo, es el modelo de **regresión lineal**.

El **coeficiente de correlación lineal de Pearson**, es un número *adimensional* que sirve para medir, de forma cuantitativa, la dependencia lineal de las variables X e Y.

Definición 2.8:

El **coeficiente de correlación lineal** de Pearson mide una tendencia lineal entre dos variables numéricas.

Es el método de correlación más utilizado y se necesita que:

- la tendencia debe ser de tipo lineal.
- no existan valores atípicos (outliers).
- las variables deben ser cuantitativas, si son cualitativas no podremos aplicar la correlación de Pearson.
- debe haber suficientes datos (algunos autores recomiendan tener más de 30 puntos u observaciones).

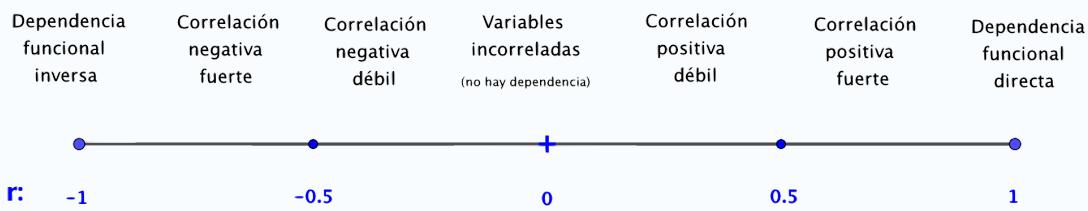
Es de gran ayuda, para observar la tendencia lineal, el disponer del diagrama de dispersión o nube de puntos correspondiente.

El **coeficiente de correlación lineal**, r , se define como:

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

- $-1 \leq r \leq 1$
- Si $|r| = 1 \rightarrow$ relación funcional
- Si $r = 0 \rightarrow$ variables incorreladas, no hay dependencia entre ellas.
- Si $r > 0 \rightarrow$ correlación positiva o directa; Si $r < 0 \rightarrow$ correlación negativa o inversa.

- Si $|r| \approx 1 \rightarrow$ correlación muy fuerte.



Coeficiente de correlación y diagrama de dispersión:

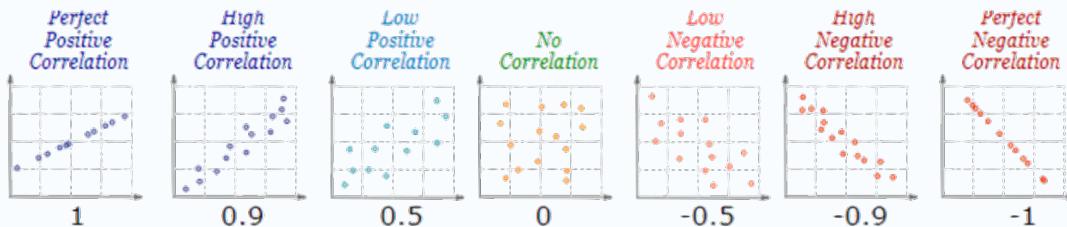
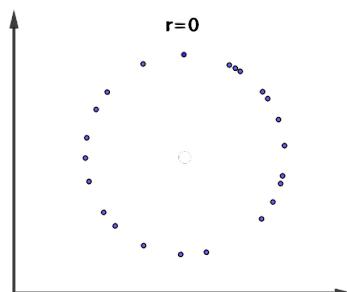


Imagen de <https://www.maximaformacion.es/blog-dat/que-es-la-correlacion-estadistica-y-como-interpretarla/>

Observación

No es suficiente el cálculo del coeficiente de correlación para negar la dependencia entre dos variables. Puede haber casos en que $r = 0$ y sin embargo existir un tipo de dependencia (aunque no lineal) entre las variables.



2.4. Regresión lineal

Teorema 2.4:

Dada la variable estadística bidimensional $(x_i, y_i; n_{ij})$ ((x_i, y_i, n_i) , en tablas de entrada simple), si el diagrama de dispersión indica una dependencia lineal y el coeficiente de correlación lineal confirma una correlación fuerte, la recta $y = a + bx$ que mejor se ajusta a la nube de puntos es la **recta de regresión**, de ecuación:

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

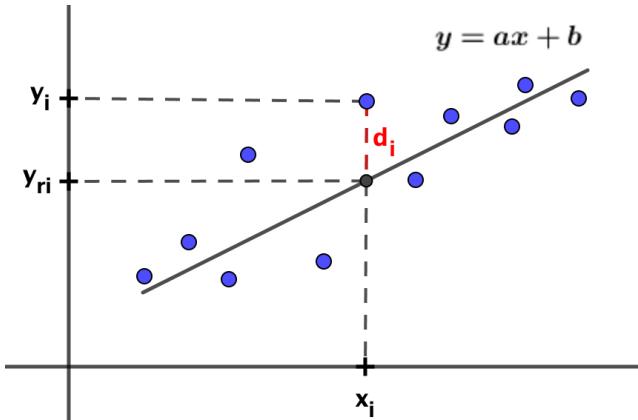
Demostración.

Para encontrar esta expresión usaremos el **método de los mínimos cuadrados**.

Variable estad. 2dim	x_i	y_i	n_i
...

$$y = a + bx$$

$$y_{ri} = a + bx_i; \quad d_i = y_i - y_{ri}$$



El *método de los mínimos cuadrados* consiste en encontrar los coeficientes a y b de la recta $y = a + bx$ tal que la suma de los cuadrados de las distancias, D , da cada punto a la recta sea **mínima**. Una vez encontrada esta recta, recibe el nombre de **recta de regresión**.

$D = d_1^2 + d_2^2 + \dots + d_N^2 = \sum_{i=1}^N d_i^2 = \sum_{i=1}^N [y_i - (a + bx_i)]^2$ Puesto que $D = D(a, b)$, por análisis matemático sabemos que el mínimo ha de verificar que las derivadas de D sean cero: $\frac{\partial D}{\partial a} = \frac{\partial D}{\partial b} = 0$

$$\begin{cases} \frac{\partial D}{\partial a} = -2 \sum_{i=1}^N [y_i - (a + bx_i)] = 0 \\ \frac{\partial D}{\partial b} = -2 \sum_{i=1}^N [y_i - (a + bx_i)] \cdot x_i = 0 \end{cases}$$

$$\rightarrow \begin{cases} \sum_{i=1}^N y_i - Na - b \sum_{i=1}^N x_i = 0 \\ \sum_{i=1}^N y_i \cdot x_i - a \sum_{i=1}^N x_i - b \sum_{i=1}^N x_i^2 = 0 \end{cases}$$

$$\text{Dividiendo por } N, \quad \begin{cases} \bar{y} - a - b\bar{x} = 0 \\ \frac{\sum_{i=1}^N x_i y_i}{N} - a\bar{x} - b\frac{\sum_{i=1}^N x_i^2}{N} = 0 \end{cases} \rightarrow \begin{cases} a = \bar{y} - b\bar{x} \\ s_{xy} + \bar{x}\bar{y} - a\bar{x} - b(s_x^2 + \bar{x}^2) \end{cases}$$

$$\text{Resolviendo el sistema, } \begin{cases} a = \bar{y} - \frac{s_{xy}}{s_x^2} \bar{x} \\ b = \frac{s_{xy}}{s_x^2} \end{cases}, \text{ con lo que la recta de regresión buscada es:}$$

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

Recta de regresión de Y sobre X.

A la pendiente, $\frac{s_{xy}}{s_x^2} = m_{YX}$, se le llama **coeficiente de regresión de Y sobre X**. Y es la variable dependiente (explicada) y X es la variable independiente (explicativa).

La recta de regresión de Y sobre X sirve para *hacer predicciones sobre la Y sabida la X*. Si lo que deseamos es *hacer predicciones sobre la X sabida la Y*, necesitaremos la recta de regresión de X sobre Y, que se obtiene de modo análogo a la anterior pero considerando ahora la X como variable dependiente (explicada) y la Y como variable independiente (explicativa). Es decir, intercambiando los papeles de X e Y:

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

Recta de regresión de X sobre Y.

A la pendiente, $\frac{s_{xy}}{s_y^2} = m_{XY}$, se le llama **coeficiente de regresión de X sobre Y**. X es la variable dependiente (explicada) e Y es la variable independiente (explicativa).

□

Definición 2.9:

Hay pues dos rectas de regresión:

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

Recta de regresión de Y sobre X.

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

Recta de regresión de X sobre Y.

La primera sirve para hacer predicciones sobre los valores de Y mientras que la segunda se usa para hacer predicciones sobre la X.

Teorema 2.5:

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \pm \sqrt{m_{YX} \cdot m_{XY}}$$

El signo de r coincide con el de s_{xy} .

Demostración.

Evidente, $\frac{s_{xy}}{s_y^2} = m_{XY}$; $\frac{s_{xy}}{s_x^2} = m_{YX}$, por lo que $m_{YX} \cdot m_{XY} = \frac{s_{xy}}{s_y^2} \cdot \frac{s_{xy}}{s_x^2} = r^2$

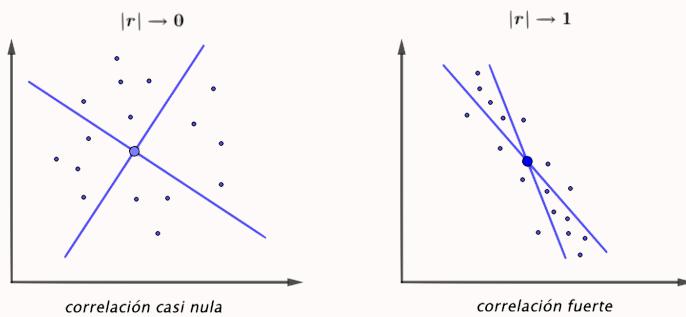
□

Teorema 2.6:

Propiedades de las dos rectas de regresión.

- Las dos rectas cortan en el punto (\bar{x}, \bar{y}) , llamado *centro de gravedad* de la nube de puntos (diagrama de dispersión).

- Las pendientes de las dos rectas m_{YX} y m_{XY} tienen el mismo signo, que a su vez es igual al signo del coeficiente de correlación lineal r . Dicho signo lo dicta el de la covarianza s_{xy} .
- En los casos en los que tenemos correlación perfecta (funcional), $r = \pm 1$, ($XY = XY$), las dos rectas de regresión coinciden. De no ser así, forman un determinado ángulo que es mayor cuando menor es la correlación, siendo máximo (90°) cuando las variables son incorreladas, $r = 0$, en este caso las rectas son perpendiculares ($x = \bar{x}$; $y = \bar{y}$).



2.5. Coeficiente de determinación

El coeficiente de determinación mide la bondad del ajuste de la recta de regresión

Definición 2.10:

El coeficiente de correlación lineal indica el grado de linealidad entre las dos variables, pero para analizar la bondad del ajuste de la recta de regresión se utiliza un parámetro nuevo llamado **coeficiente de determinación**, r^2 , indica el porcentaje de la variación de Y que puede ser explicada por X y, también, de la X respecto de la Y puesto que r^2 coincide para las dos rectas de regresión: $r^2 = \frac{s_{xy}^2}{s_x^2 \cdot s_y^2} = m_{XY} \cdot m_{YX}$

Como $0 \leq r^2 \leq 1$, la interpretación de esta medida es similar a la de r :

- Si $r^2 = 0$, significa que no existe tal relación lineal entre las variables (puede existir otro tipo de relación o no haber ninguna entre las dos variables, incorreladas). No tendría sentido utilizar las rectas de regresión para analizar la influencia de X sobre Y, ni para predecir el valor de Y, dado X.
- Si $r^2 = 1$, significa que el ajuste es perfecto (funcional). Las rectas pasan por todos los puntos de la nube, obviamente porque éstos están alineados y proporcionan toda la información sobre el comportamiento de Y en función de X, así como de X para Y, para la muestra considerada.
- Si $0 < r^2 < 1$, habrá una determinada correlación lineal entre las variables, mayor cuanto más se aproxime a uno el coeficiente de determinación. Como r^2 , indica el *porcentaje de la variación de Y que puede ser explicada por X o al revés*, un valor concreto de r^2 se

puede interpretar en los siguientes términos: si $r^2 = 0,92$, por ejemplo, significa que las rectas obtenidas explican en un 92% el comportamiento de una variable en función de la otra. El 8% restante de la variación puede deberse al azar o a la influencia de otras variables distintas.

Un coeficiente de determinación alto implica la relación lineal entre las variables es fuerte, pero eso no tiene por qué implicar que los cambios en una variable se expliquen por la otra variable. Así, si consideramos la variable (X,Y) en que X representa los ingresos mensuales de una muestra de 100 familias de una determinada ciudad e Y representa la asignación media asignada a sus hijos, el cambio brusco en los ingresos mensuales de una familia puede explicar un descenso de la asignación que dan a sus hijos, pero un descenso en la asignación a los hijos no explica el cambio en los ingresos de la familia (podría ser un castigo).

Si existe una relación causal entre dos variables, la correlación por sí misma no nos dice nada sobre la dirección de la causalidad.

Correlación no implica causalidad.

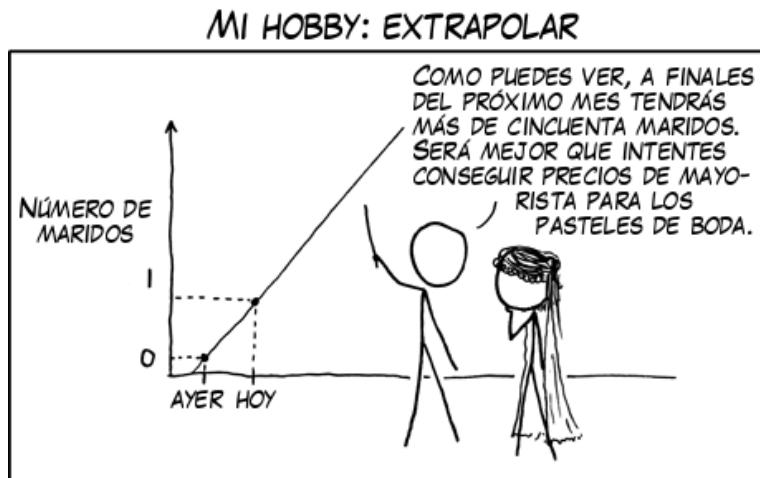
2.5.1. Valoración de las predicciones. Interpolación y extrapolación

Definición 2.11:

La recta de regresión nos permite predecir valores de una variable a partir de los de la otra. No obstante, hay que tener siempre presente que existen las siguientes limitaciones:

- Las predicciones realizadas a partir de una recta de regresión no son fiables si entre X e Y no hay un alto grado de correlación lineal, es decir, si r no es, en valor absoluto, cercano a 1.
- Las predicciones deben hacerse con valores próximos a los pares considerados. Las estimaciones obtenidas para valores próximos al centro de gravedad de la distribución son más fiables que las obtenidas para valores muy alejados de él.
- Los valores que están dentro del rango de la variable explicativa son interpolaciones; las que están fuera, extrapolaciones. Extrapolar siempre supone un riesgo añadido, la variable explicada puede tener una dependencia lineal con la variable explicativa en el rango de valores, pero no tenerlo o no ser lineal fuera de este rango.

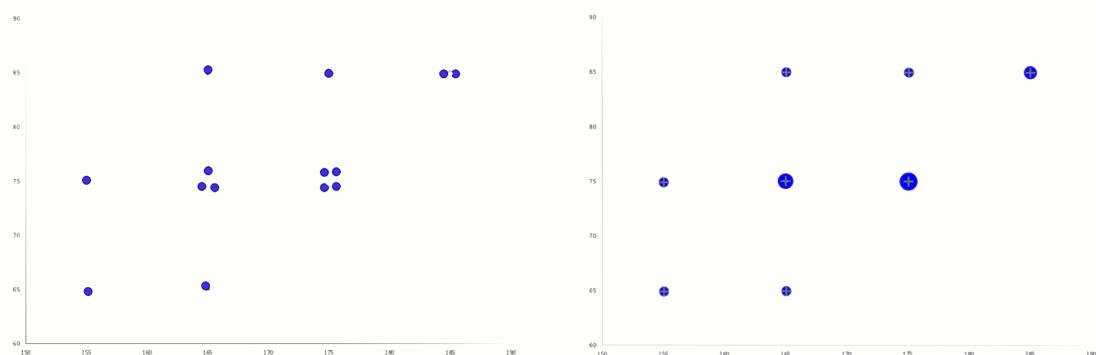
- La fiabilidad de una recta de regresión es mayor cuanto mayor sea el número de datos considerados para calcularla, N.
- Se considera buena la predicción si $r^2 > 75\%$ y muy buena a partir de $r^2 > 90\%$. Si $r^2 < 60\%$, las predicciones serán poco fiables.



Ejemplo 2.4:

Siguiendo con el ejemplo que nos ocupa en todo el tema:

$\downarrow Y / X \rightarrow$	[60,70] \rightarrow 65	[70,80] \rightarrow 75	[80,90] \rightarrow 85	n_{x_i}
[150,160] \rightarrow 155	1	1	0	2
[160,170] \rightarrow 165	1	3	1	5
[170,180] \rightarrow 175	0	4	1	5
[180,190] \rightarrow 185	0	0	2	2
n_{y_j}	2	8	4	14



$$\bar{x} = \frac{2380}{14} = 170; \quad s_x = \sqrt{\frac{405750}{14} - 170^2} = 9.06$$

$$\bar{y} = \frac{1070}{14} = 76.43; \quad s_y = \sqrt{\frac{82350}{14} - 76.43^2} = 6.37$$

$$s_{xy} = \frac{182400}{14} - 170 \cdot 76.43 = 35.47$$

Calcula el coeficiente de correlación lineal y el coeficiente de determinación. ¿Cómo es la dependencia de estas variables?

¿Cuál es el valor que debería tener la variable Y para un valor de X=77 Kg? ?Qué debería valer ya X si la Y=210 cm? ¿Cómo de fiables son estas predicciones?

Ala vista del diagrama de dispersión, intuimos que va a haber una correlación positiva y débil. Calculemos el coeficiente de correlación:

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{35.47}{9.06 \cdot 6.37} = 0.61 \rightarrow \text{correlación positiva y débil, no tiene demasiado sentido hacer predicciones.}$$

$r^2 = 0.61^2 = 0.38 = 38\%$, las predicciones explicarían el 38% de los valors, el 63% restante no.

Aunque la correlación indica que las predicciones no serán fiables, vamos a hacerlas ya que las piden:

— Recta de regresión de X sobre Y: $x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y})$.

$$x - 170 = \frac{35.47}{6.37^2}(y - 76.43) \rightarrow x = 0.87y + 103.19$$

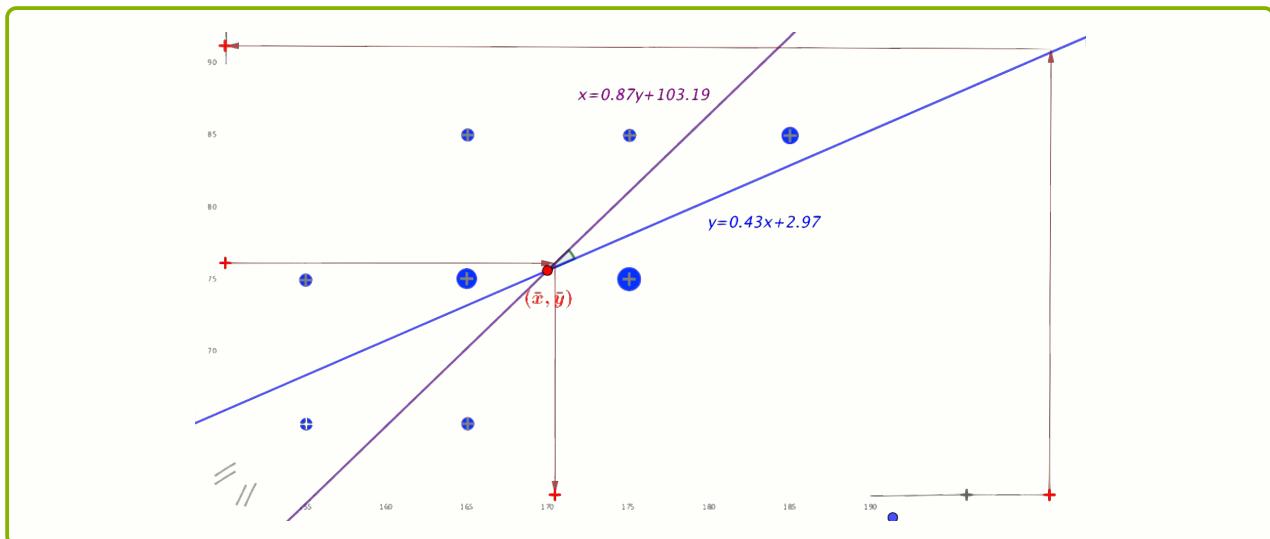
Para un $y = 77$ kg, obtenemos una $y = 170.18$ cm.

— Recta de regresión de Y sobre X: $y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$.

$$y - 76.43 = \frac{35.47}{9.06^2}(x - 170) \rightarrow y = 0.43x + 2.97$$

Para una $x = 210$ cm, obtenemos un $y = 77$ Kg.

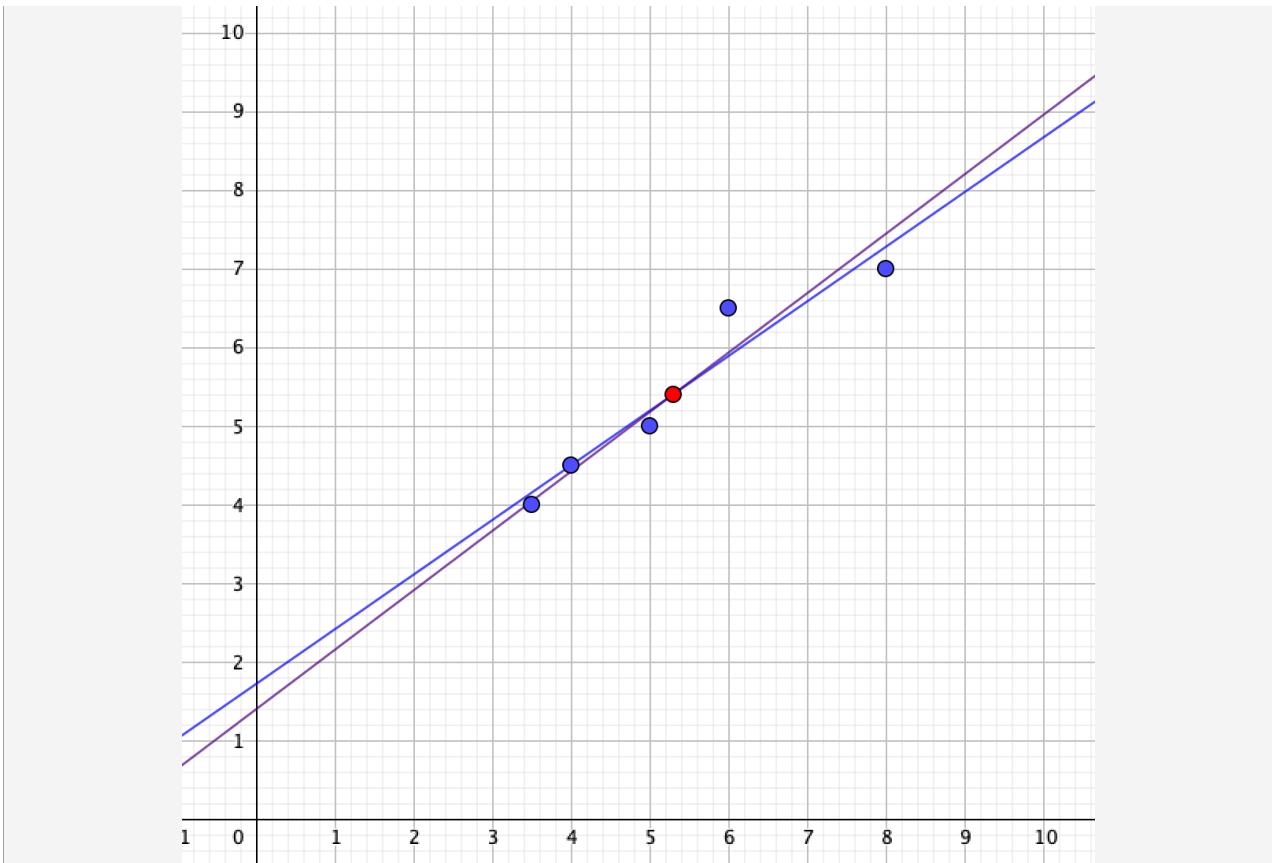
Ya hemos advertido que las previsiones no son fiables dado que la correlación es débil ($r=0.61$). Aún así, la segunda previsión es mucho menos fiable que la primera pues se trata de una peligrosa extrapolación, una altura de 210 cm está fuera de nuestro rango de valores, $[150,190]$. En la siguiente figura mostramos las dos rectas de regresión así como las predicciones efectuadas. Obsérvese como la segunda predicción es una extrapolación.



Ejercicio resuelto 2.1. Las notas obtenidas en Matemáticas y Física por cinco alumnos son:

Matemáticas (X)		6	4	8	5	3.5
Física (Y)		6.5	4.5	7	5	4

- Dibujar el diagrama de dispersión y comentar si se espera o no que haya correlación lineal.
- Calcular e interpretar los coeficientes de correlación lineal y de determinación.
- Escribir las dos rectas de regresión.
- Para una alumno de ese grupo que saque un 7.5 en matemáticas, ¿qué nota se espera que obtenga en física?
- Si una alumno obtiene un 9 en física, ¿cuál debería ser su nota en matemáticas?
- Qué fiabilidad merecen las predicciones anteriores.



— a) A la vista de la nube de puntos, se espera una correlación positiva (o directa) y fuerte.

Se han incluído en el dibujo las dos rectas de regresión y en cdg centro de gravedad, (\bar{x}, \bar{y}) , de la nube de puntos (punto rojo).

— b) y c). Cálculo de parámetros de la distribución bidimensional. Todos las frecuencias son 1, y $N=5$.

Con una sencilla hoja de cálculo obtenemos: $\Sigma x = 26.5$, $\Sigma y = 27$, $\Sigma x^2 = 153.25$, $\Sigma y^2 = 152.5$, $\Sigma xy = 152$, por lo que:

$$\bar{x} = 5.3; s_x = 1.6; \quad \bar{y} = 5.4; s_y = 1.16; \quad s_{xy} = 1.78$$

$$\text{De ahí: } r = 0.96; \quad r^2 = 0.92 = 92\%$$

Al ser $r = 0.96$, la correlación es directa (positiva) y fuerte, habrá un alto grado de fiabilidad en las predicciones (ojo con las extrapolaciones). Por ser $r^2 = 92\%$, el 92% de la nota de física es explicada por la de matemáticas y viceversa.

Las rectas de regresión son : $y = 0.695x + 1.715$; $x = 1.323y - 1.843$

— d), e) y f): predicciones.

$$x = 7.5 \rightarrow y = 0.695x + 1.715 \rightarrow y = 6.93$$

La predicción es altamente fiable dado el elevado coeficiente de correlación y el que se trata de una interpolación, $7.5 \in [3.5, 8]$

$$y = 9 \rightarrow x = 1.323y - 1.843 \rightarrow x = 10.06$$

La predicción es fiable por el elevado coeficiente de correlación pero arriesgada al tratarse de una extrapolación, $9 \notin [4, 7]$

De todos modos, el número de datos $N = 5$ es muy pequeño para garantizar que las predicciones tengan la fiabilidad necesaria.

2.6. Recta de Tukey

Definición 2.12:

La **recta de Tukey o recta mediana-mediana** permite ajustar una recta a una nube de puntos en algunos casos en los que el ajuste mínimo cuadrático produce resultados no muy buenos. La recta de Tukey o Mediana-Mediana es un método novedoso y práctico que puede venir a apoyar al que hemos visto de *mínimos cuadrados o recta de regresión*.

La recta de Tukey^a es un método ingenioso que no requiere de un fundamento matemático demasiado abstracto en su tratamiento.

Se utiliza cuando hay valores extraños (outliers) que, como sabemos, afectan mucho a las medias. Veremos como se calcula en el siguiente **ejemplo**.

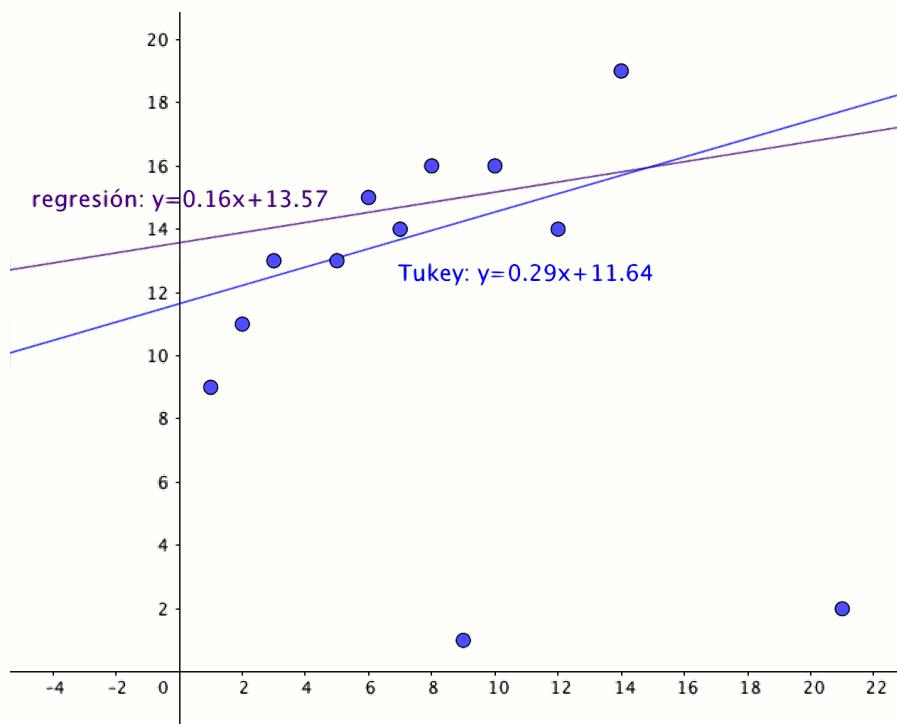
^aJohn Wilder Tukey (16 de junio de 1915 - 26 de julio de 2000). Estadístico estadounidense. Introdujo, además, los diagramas de cajas y bigotes.

Ejemplo 2.5:

Considera la siguiente distribución, dibuja la nube de puntos, la recta de regresión de Y sobre X y la recta de Tukey.

X	1	2	3	5	6	7	8	9	10	12	14	21
Y	9	11	13	13	15	14	16	1	16	14	19	2

Hemos puesto en rojo los valores *extraños*. Se ven a simple vista en el diagrama de dispersión.



Como se puede observar, la recta de regresión (morada) se ajusta muy mal al diagrama de dispersión. En cambio, la recta de Tukey se ajusta mejor, sobre todo si obviamos mentalmente los puntos que están muy aislados de la nube (los señalados en rojo en la tabla).

Se deja como ejercicio para el/la lector/a el encontrar la recta de regresión de Y sobre X:
 $y = 0.16x + 13.57$

Vamos con la recta de Tukey:

1. Se ordenan los datos en orden creciente de las abcisas (ya lo tenemos).
2. Se divide el conjunto de datos ordenados en tres grupos:

$$G1 = \{(1, 9), (2, 11), (3, 13), (5, 13)\}$$

$$G2 = \{(6, 15), (7, 14), (8, 16), (9, 1)\}$$

$$G3 = \{(10, 16), (12, 14), (14, 19), (21, 2)\}$$

3. Para cada grupo G_i se calculan las medianas de las abcisas (x) y de las ordenadas (y) del grupo, m_{x_i} y m_{y_i} .

$$G1 : \left\{ \begin{array}{l} \text{Abcisas G1: } 1, 2, 3, 5 \rightarrow m_{x_1} = 2.5 \\ \text{Ordenadas G1: } 9, 11, 13, 11 \rightarrow m_{y_1} = 12 \end{array} \right\} \Rightarrow P1(2, 5, 12)$$

$$G2 : \left\{ \begin{array}{l} \text{Abcisas G2: } 6, 7, 8, 9 \rightarrow m_{x_1} = 7.5 \\ \text{Ordenadas G2: } 1, 14, 15, 16 \rightarrow m_{y_1} = 14.5 \end{array} \right\} \Rightarrow P1(7, 5, 14.5)$$

$$G3 : \left\{ \begin{array}{l} \text{Abcisas G3: } 10, 12, 14, 21 \rightarrow m_{x_1} = 13 \\ \text{Ordenadas G3: } 2, 14, 16, 29 \rightarrow m_{y_1} = 15 \end{array} \right\} \Rightarrow P1(13, 15)$$

Obsérvese que el número de datos en este caso es $N = 12$, múltiplo de 3, y, por tanto, cada grupo está formado por 4 datos.

Si el número de datos N no es múltiplo de 3, puede ocurrir que:

- Sea múltiplo de 3 más 1; en este caso, el grupo G2 se deja con un dato más.
- Sea múltiplo de 3 más 2; en este caso, el grupo G2 se deja con un dato menos, se añaden un dato más a los grupos G1 y G3.

4. La recta de Tukey para por el baricentro del triángulo $P1P2P3$ y tiene por pendiente la de la recta que pasa por $P1$ y por $P3$.

$$\text{Baricentro: } B_{P1P2P3} = \frac{P1 + P2 + P3}{3} = (7.67, 13.83)$$

$$\text{Pendiente: } m_{P1P3} = \frac{15 - 12}{13 - 2.5} = 0.29$$

Luego, la recta de Tukey es: **$y = 0.29x + 13.83$**

Regresiones no lineales

El modelo de regresión lineal, estudiado para detectar dependencia lineal entre las variables, también se puede aplicar para detectar otros tipos de dependencia no lineales, como la exponencial o la potencial.

Teorema 2.7:

Dependencia exponencial

Si dos variables se sospecha que están relacionadas por un modelo exponencial: $y = ke^{ax}$, tomando logaritmos: $\ln y = ax + b$, con $k = e^b$ ($b = \ln k$).

Para encontrar una dependencia exponencial en una serie de datos (x_i, y_i, n_i) , buscaremos una dependencia lineal en la serie $(x_i, \ln y_i, n_i)$. Si la dependencia lineal es $Y = aX + b$, la dependencia exponencial será $Y = Ke^{aX}$, con $K = e^b$

Ejemplo 2.6:

Encontrar la dependencia exponencial para los siguientes datos (X habitantes del mundo, en millones; Y año):

X	1750	1800	1850	1900	1950	2000
Y	728	949	1171	1608	2516	5923

Cálculos:

$$\Sigma x = 11250; \Sigma y = 12895; \sigma x^2 = 21137500; \Sigma y^2 = 46799675; \Sigma x \cdot y = 24955950$$

$$\bar{x} = 1875; s_x = 85.39; \bar{y} = 2149.17; s_y = 1783.54; s_{xy} = 129631.25$$

$$r = 0.85; \text{ Recta de regresión de } Y \text{ sobre } X: y = 17.78x - 31183.33$$

Tomando logaritmos en la población (Y):

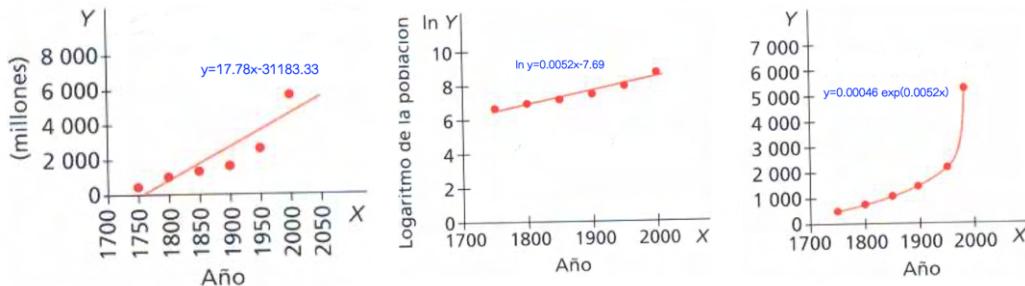
$$\text{Ahora, } \Sigma y = 43,33; \Sigma y^2 = 314.01; \Sigma x \cdot y = 81450.40$$

$$\text{Luego } \bar{y} = 7.22; s_y = 0.45; s_{xy} = 37.58$$

$$r = 0.98. \text{ Recta de regresión de } \ln Y \text{ sobre } X: \ln y = 0.0052x - 7.69$$

Por último, la dependencia exponencial será:

$$y = e^{0.0052x - 7.69} = e^{0.0052x} \cdot e^{-7.69} = 0.00046e^{0.0052x}$$



Ejemplo tomado de “https://yoquieroaprobar.es/_pdf/04140.pdf”

Teorema 2.8:**Dependencia potencial**

Si dos variables estadísticas X e Y están relacionadas por un modelo potencial: $y = k a^x$, con $k > 0$, al tomar logaritmos, se obtiene: $\ln y = \ln k + a \ln x$ de donde se deduce que las variables ($\ln X$, $\ln Y$) están relacionadas por un modelo lineal.

Para encontrar dependencia potencial en una lista de datos (x_i, y_i, n_i) , se aplica una regresión lineal a los datos: $(\ln x_i, \ln y_i, n_i)$.

Si la recta de regresión es $\ln Y = a \ln X + b$, entonces la dependencia potencial en los datos originales es: $y = k a^x$, con $k = e^b$.

Veremos un **ejemplo** que aclare todo esto.

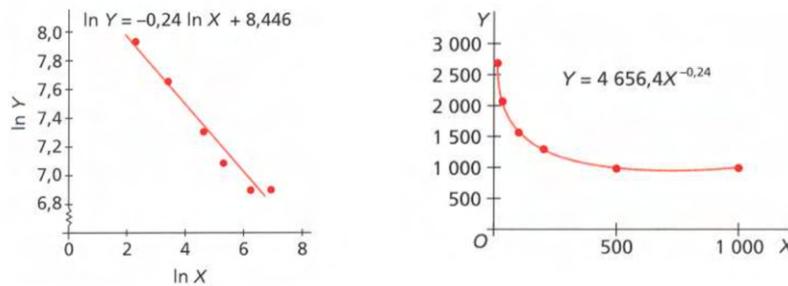
Ejemplo 2.7:

El científico inglés Fry Richardson midió la costa oeste de Gran Bretaña con “reglas” de distintos tamaños, y obtuvo los siguientes datos aproximados (principio de los fractales):

Calcula el coeficiente de correlación para estos datos y también para los datos $(\ln X, \ln Y)$. Compara los resultados. ¿Cuál es la ley que mejor describe la dependencia entre las variables X e Y ?

Tamaño de la regla (cm)	1000	500	200	100	30	10
Longitud de la costa (Km)	1000	1000	1200	1500	2100	2800

Para los datos (X, Y) , se encuentra (hágase) que $r = 0,698$. En cambio, para los datos $(\ln x, \ln y)$ tenemos que $r = 0,985$, mucho más elevado que el anterior, y la recta de regresión es: $\ln y = -0,24 \ln x + 8,446 \rightarrow y = e^{-0,24 \ln x + 8,446} = e^{8,446} x^{-0,24} = 4656,4 x^{-0,24}$



Ejemplo tomado de “https://yoquieroaprobar.es/_pdf/04140.pdf”

2.7. Ejercicios

Se deberían intentar resolver los ejercicios antes de ver su resolución.

Ejercicio 2.1.

Se ha hecho una encuesta a 12 personas que han tenido un accidente de tráfico, preguntando por el número de meses transcurridos e incluyendo el grupo de edad.

Las respuestas han sido:

Carmen, 35: [60, 70); Teresa, 15: [50, 60); Pilar, 12: [50, 60); Esther, 6: [20, 30); Juan, 8: [40, 50); Jacinto, 15: [30, 40); Jesús, 24: [50, 60); Marta, 12: [30, 40); José, 28: [40, 50); Andrés, 3: [20, 30); María Jesús, 20: [40, 50); Beatriz, 16: [30, 40)

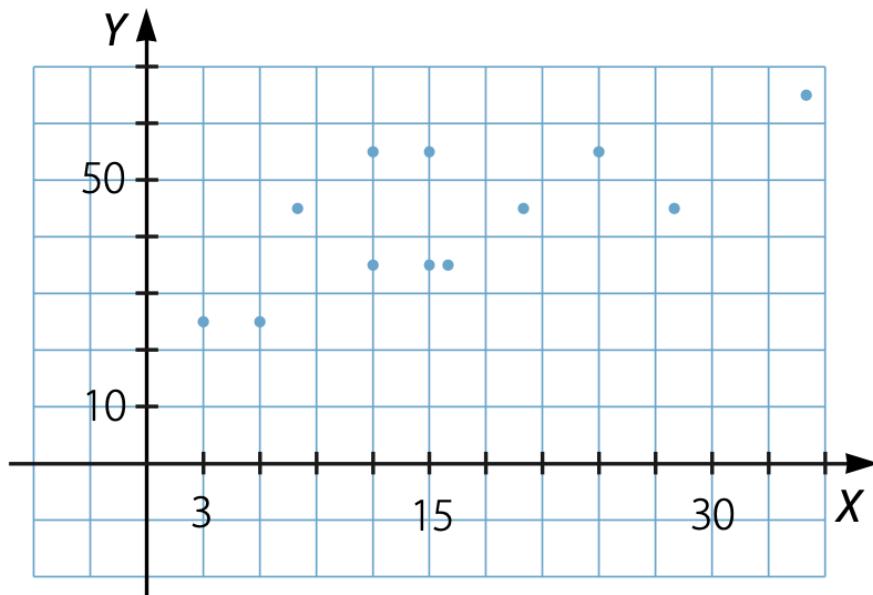
- a) Construye la tabla correspondiente a la variable bidimensional.

- b) *Representa el diagrama de dispersión.*
- c) *Estudia si hay correlación entre ambas variables, y determina su coeficiente de correlación lineal.*
- d) *Si la persona entrevistada tuviese 87 años, ¿cuántos meses es de esperar que hayan transcurrido desde su último accidente?*
- e) *Si a un encuestado le han transcurrido 18 meses desde su último accidente, ¿qué edad es de esperar que tenga?*
- f) *Como de fiables son tus predicciones.*

— a) Representamos los datos en una tabla de doble entrada:

Y \ X	3	6	8	12	15	16	20	24	28	35	total
[20, 30[→ 25	1	1									2
[30, 40[→ 35				1	1	1					3
[40, 50[→ 45			1				1		1		3
[50, 60[→ 55				1	1			1			3
[60, 70[→ 65										1	1
totales	1	1	1	2	2	1	1	1	1	1	12

— b) Diagrama de dispersión:



Se observa cierta correlación lineal positiva y no muy fuerte.

— c) Cálculos:

x_i	y_i	x_i^2	y_i^2	x_i y_i	f_1
25	3	625	9	75	1
25	6	625	36	150	1
35	12	1225	144	420	1
35	15	1225	225	525	1
35	16	1225	256	560	1
45	8	2025	64	360	1
45	20	2025	400	900	1
45	28	2025	784	1260	1
55	12	3025	144	660	1
55	15	3025	225	825	1
55	24	3025	576	1320	1
65	35	4225	1225	2275	1
		520	194	24300	4088
					9330
43,333333		16,166667	12,1335165	8,9053667	76,944444
\bar{x}	\bar{y}	s_x	s_y	s_{xy}	
r=	0,7120964				

$\bar{x} = 43,3$ años; $\bar{y} = 16.2$ meses; $s_x = 12.1$; $s_y = 8.9$; $s_{xy} = 76.9 \rightarrow r = 0.71$. Hay una correlación positiva y no muy fuerte.

— e) Para $x = 87$ años \rightarrow

Necesitamos la recta de regresión de Y sobre X:

$$y - 16.17 = \frac{76.94}{12.13^3} (x - 43.3) \rightarrow y = 0.52x - 6.48$$

$x = 87$ años $\rightarrow y = 38.76$, casi 39 meses.

— f) Para $y = 18$ meses \rightarrow

Necesitamos la recta de regresión de X sobre Y:

$$x - 43.33 = \frac{76.94}{8.91^2} (y - 16.17) \rightarrow x = 0.97y + 27.66$$

$y = 18$ meses $\rightarrow x = 45.12$, 45 años aproximadamente.

— g) Al no ser muy fuerte la correlación ($r = 0.71$), las estimaciones que hemos hecho no son muy fiables, siendo la primera (87 años \rightarrow 39 meses) mucho menos fiable a tratarse de una extrapolación (nuestros datos abarcan edades desde 20 hasta 70 años).

Ejercicio 2.2. La recta de regresión de una variable Y respecto de la variable X es $y = 0,3x + 1$. Los valores que ha tomado la variable x han sido $\{3, 4, 5, 6, 7\}$.

- a) Determina el valor esperado de y para el valor particular de $x = 3.5$.
- b) Si los valores de la variable Y utilizados para la regresión se multiplican por 10 y se dejan los mismos valores para la variable X, determina razonadamente la nueva recta de regresión.

— a) $x = 3.5 \rightarrow y = 0.3 \cdot 3.5 + 1 = 2.05$

$$\text{— b) } Y' : 10 \cdot y_i \rightarrow \bar{y}' = 10\bar{y} ; S'_{XY} = \frac{\sum_{i=1}^N x_i 10y_i n_i}{N} - \bar{x} \cdot 10\bar{y} = 10s_{xy}$$

$$y - 10\bar{y} = \frac{10s_{xy}}{s_x^2} (x - \bar{x}) \rightarrow y = 10 \frac{s_{xy}}{s_x^2} x + 10 \left(\bar{y} - \frac{s_{xy}}{s_x^2} \right) = 3x + 10$$

Ejercicio 2.3. Cien alumnos prepararon un examen de Matemáticas. Se representa por X el número de problemas hechos por cada alumno en la preparación, y por Y , la calificación obtenida. Sabiendo que las medias aritméticas de esas variables fueron $\bar{x} = 9.2$ e $\bar{y} = 9.5$, que el coeficiente de correlación entre esas variables fue 0.7 y que la desviación típica de la variable Y fue el doble que la de la variable X , calcula las ecuaciones de las rectas de regresión.

$$s_y = 2s_x \rightarrow r = \frac{s_{xy}}{s_x s_y} = \frac{s_{xy}}{2s_x^2} = 0.7 \Rightarrow \frac{s_{xy}}{s_x^2} = 1.4$$

$$y - \bar{y} = \frac{s_{xy}}{s_x^2} (x - \bar{x}) \Rightarrow y - 9.5 = 1.4 (x - 9.2)$$

$$\frac{s_{xy}}{s_y^2} = \frac{s_{xy}}{(2s_x)^2} = \frac{1}{4} \frac{s_{xy}}{s_x^2} = \frac{1}{4} \frac{1}{4} = 0.35$$

$$x - \bar{x} = \frac{s_{xy}}{s_y^2} (y - \bar{y}) \Rightarrow x - 9.2 = 0.35 (y - 9.5)$$

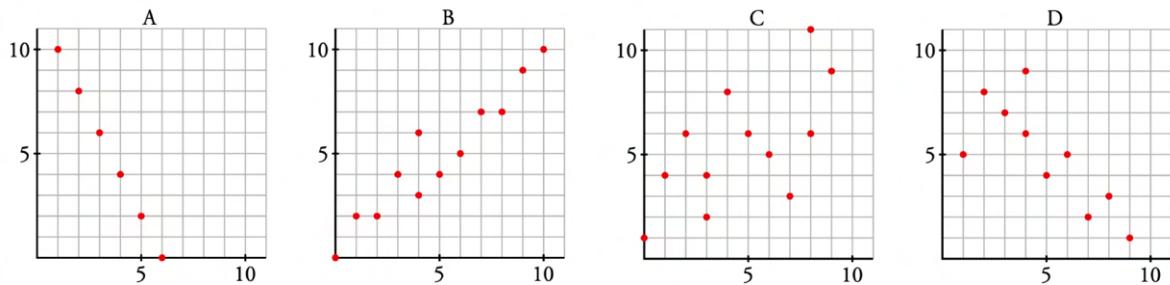
2.7.1. Problemas propuestos (con solución)

PB. 1. En cada uno de estos casos debes decir si, entre las dos variables que se citan, hay relación funcional o estadística (correlación) y, en este último caso, indicar si es positiva o negativa:

- a) En un conjunto de familias: estatura media de los padres – estatura media de los hijos.
- b) Entre los países del mundo respecto a España: volumen de exportación – volumen de importación.
- c) En los países del mundo: tasa de mortalidad infantil – médicos por cada 1 000 habitantes.
- d) En las viviendas de una ciudad: kw consumidos en un mes – coste del recibo de la luz.
- e) En los equipos de fútbol: posición al finalizar la liga – número de partidos perdidos.

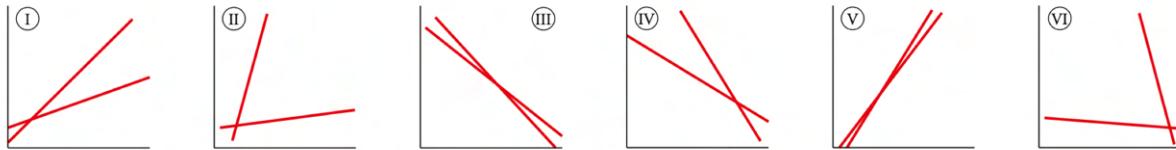
E+, E-, F+, F-

PB. 2. Dados los coeficientes de correlación: $r_1 = 0.64$; $r_2 = 0.95$; $r_3 = -1$; $r_4 = -0.76$; asocia cada uno de ellos a las siguientes nubes de puntos.



$$r_1 \leftarrow C; r_2 \leftarrow B; r_3 \leftarrow A; r_4 \leftarrow D$$

- PB. 3. Dados los coeficientes de regresión $r_1 = -0.9$; $r_2 = 0.99$; $r_3 = 0.6$; $r_4 = -0.2$; $r_5 = -0.5$; $r_6 = 0.1$ y las rectas de regresión de la siguiente figura, asocia por parejas.



$$r_1 \leftarrow IA; r_2 \leftarrow A; r_3 \leftarrow AI; r_4 \leftarrow III; r_5 \leftarrow II; r_6 \leftarrow I$$

- PB. 4. Un excursionista, en diez marchas distintas, toma las siguientes medidas: x – altura de lugar (en m); y – presión atmosférica (en mm Hg); z – número de pulsaciones en reposo.

x	0	184	231	481	730	911	1343	1550	1820	2184
y	760	745	740	720	700	685	650	630	610	580
z	73	78	75	78	83	80	89	80	85	92

Halla el coeficiente de correlación y la recta de regresión para la distribución x-y y para x-z y analiza los resultados.

$$r_{xy} = -0.08x + 759; r_{xz} = 0.87x + 74.8$$

- PB. 5. La recta de regresión de Y sobre X de una distribución bidimensional es $y = 1.6x - 3$. Sabemos que $\bar{x} = 10$ y $r = 0.8$.

Calcula \bar{y} y estima el valor de y para $x = 12$ y para $x = 50$.

¿Qué estimación te parece más fiable?

$$(x, y) \in r: \bar{y} = 13; y(12) = 16.2; y(50) = 77;$$

Primera estimación más fiable; 12 más próximo a \bar{x} que 50.

PB. 6. Considera la siguiente variable bidimensional:

x	1	2	3	4	5	6	7	8	9
y	5	6	8	11	1	13	14	14	17

Encuentra la recta de regresión de Y sobre X y la recta de Tukey.

$$\text{Regresión: } y = 1.43x + 2.74; \quad \text{Tukey: } y = 1.33x + 3.67$$

PB. 7. Dada la siguiente variable bidimensional, encuentra la recta de regresión de Y sobre X y la recta de Tukey.

x	1	2	3	4	5	6	7	8	9	10
y	7	8	6	8	1	10	9	7	8	11

$$\text{Regresión: } y = 0.32x + 5.33; \quad \text{Tukey: } y = 0.14x + 6.23$$

PB. 8. Para una variable bidimensional se conoce $r = -0.5$, $s_x = 2$; y $s_y = 3$. Razóna si alguna de las siguientes rectas de regresión de Y sobre X corresponde a estos datos:

- a) $y = -x + 2$; b) $y = 0.5x - 1$; c) $3x + 4y - 4 = 0$

$$\left(\text{Recta de regresión: } y = \frac{s_y}{s_x}x + \bar{y} - \frac{s_y}{s_x}\bar{x} \right) \quad (C)$$

PB. 9. Las rectas de regresión de Y sobre X y de X sobre Y en una distribución bidimensional, son las siguientes: $y = 0 - 91x - 5.88$; $x = 0.85y + 13.24$

¿Cuál es el coeficiente de correlación de Pearson de la distribución?

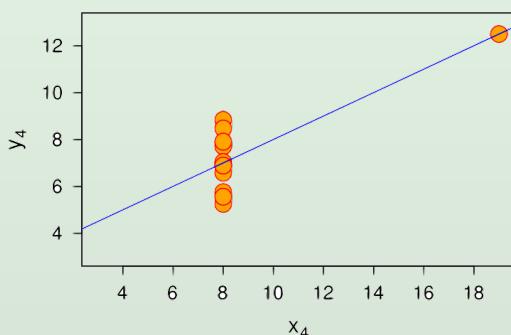
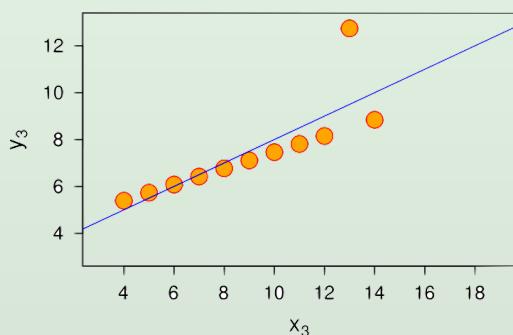
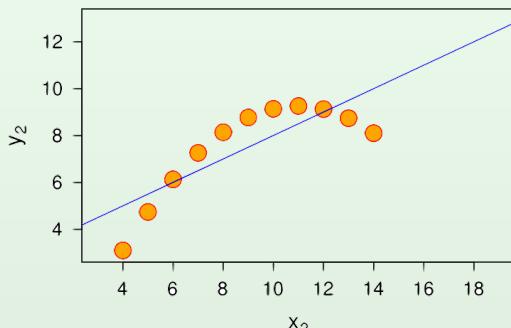
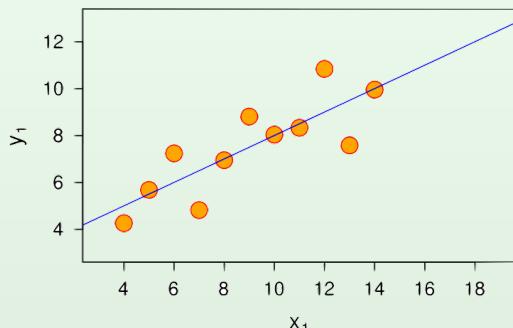
$$(Correlación:) \quad r = -0.98$$

2.8. Curiosidades

Cuarteto de Anscombe

La correlación y la regresión están relacionadas entre sí. Ahora bien, puede ocurrir que dos variables estén correlacionadas entre sí pero, en cambio, la relación entre ellas puede ser de tipo muy diferente en uno u otro caso. El estadístico inglés Frank Anscombe, ideó un conjunto de ejemplos de datos (conocido como el cuarteto de Anscombe)^a de tal manera que todos ellos tienen los mismos parámetros estadísticos (medias, desviaciones típicas, correlación y recta de regresión) pero, en cambio, la relación entre las variables es muy diferente entre sí: el primero

sería una distribución normal (lo que cabría esperar de los parámetros), la segunda es una función no lineal, y las otras dos son lineales pero afectadas cada una por un dato que se escapa de la relación de los demás. Los diagramas de dispersión y la recta de regresión correspondiente son:



En este conjunto se puede apreciar claramente que es muy importante no fiarse sólo de la correlación y regresión que haya entre dos variables (aunque sea fuerte, los cuatro ejemplos tienen una correlación de 0.816) para pensar que hay una relación lineal entre las mismas. Así, en primer lugar, debemos considerar el diagrama de dispersión para, luego, estudiar los valores de la correlación y de la regresión. Los dos últimos ejemplos también nos muestran la influencia que tienen en la recta de regresión los datos aislados que se encuentran lejos del resto.

Se pueden consultar las 4 distribuciones y sus parámetros coincidentes en :

["https://upload.wikimedia.org/wikipedia/commons/b/b6/Anscombe.svg"](https://upload.wikimedia.org/wikipedia/commons/b/b6/Anscombe.svg)

^ahttp://e-educativa.catedu.es/44700165/aula/archivos/repositorio/2000/2007/html/31_rectas_de_regresin_i.html

Correlación no implica causalidad

El ejemplo más clásico de que “**Correlación ni implica causalidad**”^a es el de los piratas y el calentamiento global. Este se basa en un estudio desarrollado nada menos que por **Bobby Henderson**, el creador de la “*Iglesia pastafafari*”. Su intención era combatir los argumentos de los creacionistas, un grupo muy dado a encontrar correlaciones donde no las hay y a concluir que hay una causa detrás.

Casualmente la causa que siempre encuentran es la misma, ‘Dios causa que ...’, de nuevo casualmente, coincide con lo que estaban intentando demostrar a priori. Para ilustrar el hecho de

que el que dos fenómenos se den al mismo tiempo no implica que uno cause el otro. Henderson representó la temperatura global de la Tierra en función del número de piratas en el mundo.

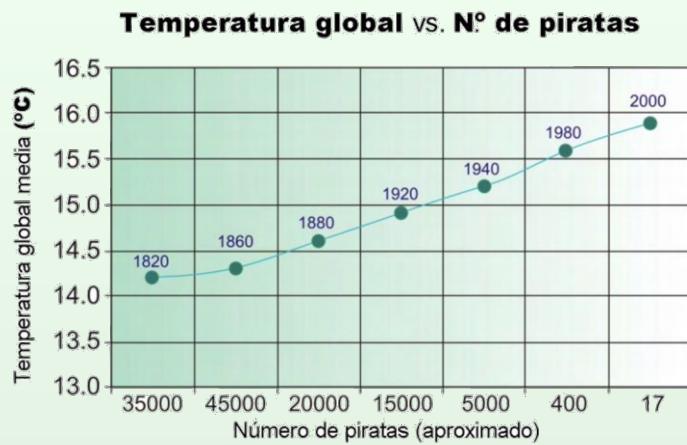
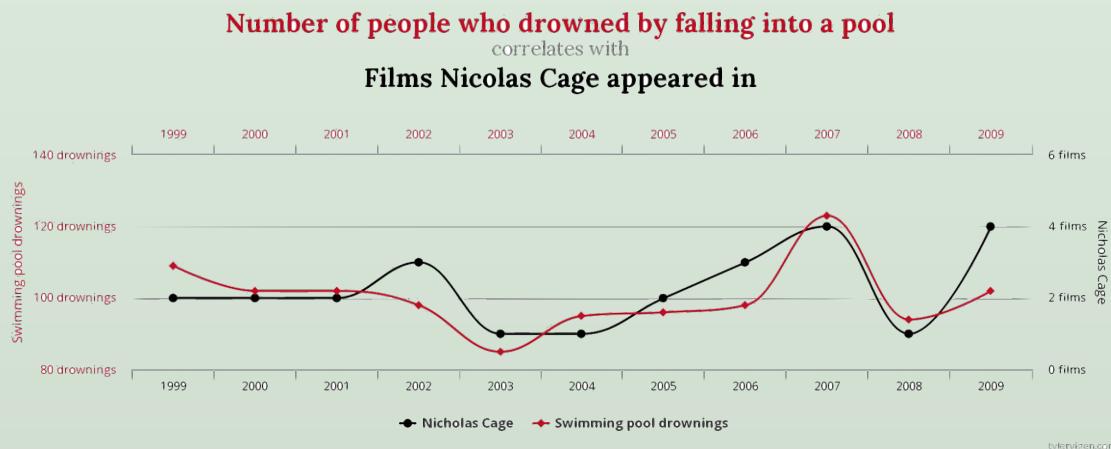


Imagen: PiratesVsTemp.svg: RedAndr / Osado (CC).

Claramente se aprecia que, a medida que el número de piratas se ha reducido, la temperatura de la atmósfera ha aumentado. Según los argumentos de los creacionistas, y otros grupos favorables a encontrar causas donde no las hay, esto significaría que la escasez de piratas es la verdadera causa del calentamiento global. No hay otra explicación. Por este motivo los seguidores de la religión de Henderson se disfrazan de piratas en el momento del culto, para combatir así el cambio climático.

Veamos otro ejemplo. La página web “**Spurious Correlations**” (*correlaciones espurias*) se dedica a buscar en distintas bases de datos correlaciones absurdas entre series de datos. Una de las más populares es la que aparece en la siguiente gráfica, que representa a través de los años tanto el número de ahogamientos en piscina producidos en los Estados Unidos como el número de películas realizadas por Nicolas Cage.



La correlación es clara. Cuantas más películas hace el bueno de Nicolas más gente muere ahogada. Lo mejor será que el pobre se retire y así ahorrará sufrimiento al mundo.

Dado que es difícil de creer que la gente se ahogue por culpa de Nicolas Cage, o que los piratas determinen la temperatura global, podemos concluir que estas correlaciones no implican que una cosa sea la causa de la otra. Veamos entonces la explicación canónica a estas gráficas. Que

dos fenómenos se den a la vez, o que uno preceda al otro, no implica que uno sea la causa del otro. Aunque observamos una correlación entre A (películas de Cage) y B (ahogamientos en piscina) eso no significa que las películas de Nicolas Cage provoquen que la gente quiera morir de una manera agónica a la vez que refrescante.

¿‘Y, si no es A la causa de B, por qué se dan los dos fenómenos a la vez de forma repetida? Bueno, en general, si hay una fuerte correlación entre los fenómenos A y B, tenemos cuatro posibilidades:

- Que A cause B (que los ahogamientos en piscinas hagan que el bueno de Nicolas quiera hacer más cine para animar a las familias).
- Que B cause A (yo mismo estuve tentado de ahogarme después de ver La búsqueda 2).
- Que haya un tercer fenómeno, C, que provocara tanto A como B (es complicado imaginar alguno, pero a lo mejor el Orden Mundial conspira para reducir la población humana tanto mediante el ahogamiento como mediante el aburrimiento).
- Puro y duro azar. Hay muchos datos en el mundo, así que si los comparamos todos más tarde o más temprano encontraremos este tipo de correlaciones que no significan nada.

Nadie duda de que la correlación no implica causalidad. Científicos de todos los campos dedican cantidades ingentes de tiempo a repetir experimentos para distinguir correlaciones importantes de correlaciones espurias. Incluso se ha observado que muchos experimentos científicos con grandes correlaciones tienen una probabilidad alta de ser puramente casuales. Eso ocurre porque en el mundo se realizan muchos experimentos continuamente. La probabilidad de que nunca se dé una correlación espuria es realmente baja y son precisamente las correlaciones inesperadas las que más interesan a la comunidad científica. El único remedio para evitar esto es la repetición de los experimentos. Sin embargo, todo esto no quiere decir que las correlaciones no tenga relevancia, o que no sean indicativas de causalidad. Tenemos que saber distinguir entre correlaciones más y menos probables. Tenemos que analizar cada caso cuantitativamente y averiguar cuál es la probabilidad de que un evento sea aleatorio para saber si debemos indagar más o no..

SACANDO CONCLUSIONES.

Los ejemplos que se muestran a continuación, subrayan la importancia de no lanzarse a sacar implicaciones de tipo causal tan pronto se tiene noticia de una correlación estadística.

Las estadísticas muestran que casi todos los accidentes de circulación se producen entre vehículos que ruedan a velocidad moderada. Muy pocos ocurren a más de 150 Km. por hora. ¿‘Significa esto que resulta más seguro conducir a gran velocidad?

No, de ninguna manera. Con frecuencia, las correlaciones estadísticas no reflejan causas y efectos. Casi todo el mundo circula a velocidad moderada, y como es natural, la mayoría de los accidentes se producen a estas velocidades.

Suele decirse que casi todos los accidentes de automóvil ocurren cerca de casa. ¿‘Significa esto que viajar por carretera, a muchos kilómetros de nuestra ciudad, es menos peligroso que callejear por nuestro barrio?

No. Las estadísticas reflejan, sencillamente, que se usa más el coche por los alrededores de nuestra residencia que por carreteras alejadas.

Otro estudio mostró que en cierta ciudad se produjo un súbito aumento de mortalidad por fallo cardíaco y un fuerte incremento en el consumo de cerveza. ¿Es posible que beber cerveza sea causa de que aumente la probabilidad de ataque al corazón?

No. En ambos casos el aumento fue debido a un veloz incremento de la población. Por igual causa, los ataques al corazón podrían ser atribuidos a cientos de otras cosas: aumento del consumo de café, de chicle, de partidas de tute, o de ver la televisión.

En 1984 murieron en España muchas más personas por accidente de tráfico que en 1960. ¿Basta esto para afirmar que era más peligroso viajar en 1984 que en 1960?

No. ...

Recientes estadísticas muestran que la tasa de natalidad es el doble que la tasa de mortalidad. ¿Será verdad, por tanto, que una de cada dos personas es inmortal?

No. ...

La probabilidad de tener un accidente de tráfico aumenta con el tiempo que te pases en la calle. Por tanto, cuanto más rápido circules, menor es la probabilidad de que tengas un accidente. ¿Es cierto?

No. ...

Diversos autores han advertido de las especiales consideraciones que hay que realizar al interpretar el significado de una correlación. La posible mutua dependencia de las dos variables analizadas de una tercera que no se tiene en cuenta invita a prestar una mayor atención a los resultados obtenidos con base en un estudio observacional.

Ejemplos:

— Existe una elevada correlación positiva y significativa entre las ventas anuales de chicle y la incidencia del crimen en los Estados Unidos de América.

Obviamente, no es lícito concluir que prohibiendo la venta de chicle podría reducirse el crimen, pues ambas variables dependen de una tercera: el tamaño de la población analizada.

— Existe una elevada correlación negativa y significativa entre el índice de mortalidad infantil de un país y el número de teléfonos móviles de ese país.

A nadie se le ocurriría que para disminuir la mortalidad infantil haya que vender más teléfonos móviles. En realidad, lo que carece de sentido es concluir que, dado que la correlación estadística existe, debe existir también una relación del tipo causa-efecto entre las variables analizadas.

Cum hoc ergo propter hoc (en latín “con esto, por tanto a causa de esto”) es una falacia (es decir, un argumento que parece válido, pero que no lo es) que se comete al inferir que dos o más eventos están conectados causalmente porque se dan juntos. La falacia consiste en inferir que existe una relación causal entre dos o más eventos por haberse observado una correlación estadística entre ellos. Esta falacia muchas veces se refuta mediante la frase “correlación no implica causalidad.”

^aArtículo de Daniel Manzano, obtuvo el primer premio del concurso DIPC de divulgación del evento Ciencia Jot Down 2016

RESUMEN: Distribuciones bidimensionales. Correlación y Regresión lineal

- ▷ Tablas de simple y de doble entrada:

x_i	i_i	n_i	X/Y	\dots	y_j	\dots	
:	:	:		:			
:	:	:	\leftrightarrow	x_i	n_{ij}		$\sum \rightarrow n_{x_i}$
:	:	:			:		
		N					$\sum \rightarrow n_{y_j}$

Cálculos, a partir de la tabla de entrada simple:

x_i	y_i	n_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$	$y_i \cdot n_i$	$y_i^2 \cdot n_i$	$x_i \cdot y_i \cdot n_i$
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
		N	$\sum \rightarrow \bar{x}$	$\sum \rightarrow s_x$	$\sum \rightarrow \bar{y}$	$\sum \rightarrow s_y$	$\sum \rightarrow s_{xy}$

- ▷ Coeficientes de correlación y de determinación:

$$r = \frac{s_{xy}}{s_x s_y}; \quad -1 \leq r \leq 1; \quad r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = m_{XY} \cdot m_{YX}$$

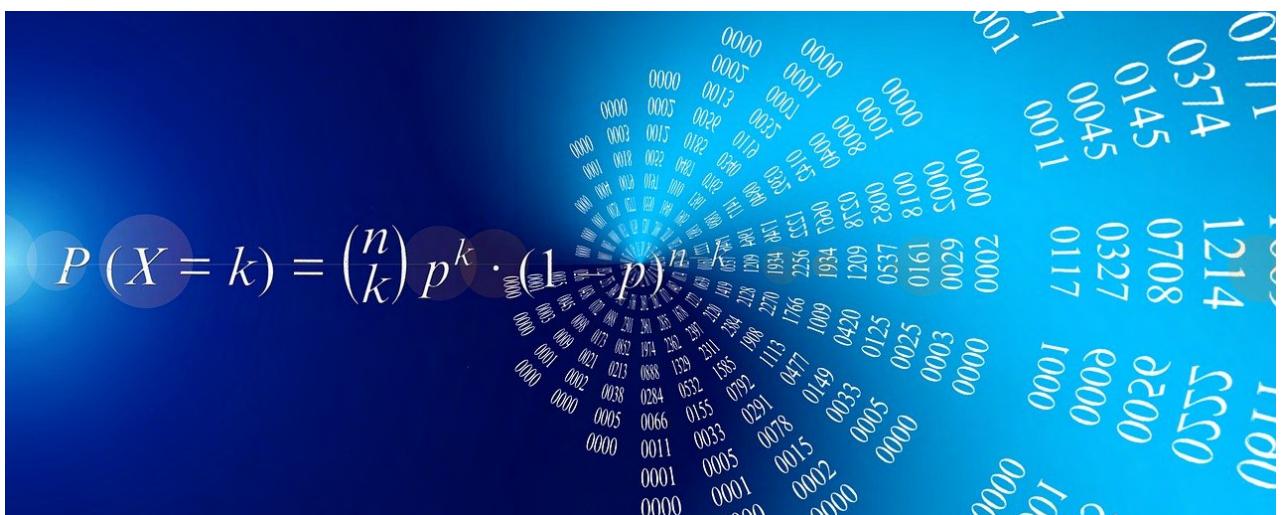
- ▷ Rectas de regresión:

$$\text{Y sobre X: } y - \bar{y} = \frac{s_{xy}}{s_x^2} (x - \bar{x}); \quad \text{X sobre Y: } x - \bar{x} = \frac{s_{xy}}{s_y^2} (y - \bar{y})$$

- ▷ Recta de Tukey (mediana - mediana)

Parte II

Probabilidad



Capítulo 3

Cálculo de probabilidades

3.1. Introducción

La teoría de la probabilidad es un modelo matemático que se ocupa de analizar los fenómenos aleatorios en contraposición a los fenómenos deterministas (aquellos en los cuales el resultado del experimento que se realiza, atendiendo a determinadas condiciones, produce un resultado único y previsible que se repetirá la cantidad de veces que éste vuelva a hacerse siempre y cuando se respeten las mismas condiciones).

El origen de la probabilidad reside en la necesidad del ser humano de anticiparse a los hechos y de predecir en cierta medida el futuro para lo que se construyen patrones y conexiones que intentan poner algo de orden en el caos.

El término probabilidad proviene probable, o sea, de aquello que es más posible que ocurra, y se entiende como el mayor o menor grado de posibilidad de que un evento *aleatorio, estocástico o de azar* ocurra.

Estas tres palabras proceden de nuestras tres culturas:

- *Aleatorio*, deriva del latín “aleatorius”, es ‘lo que no se puede predecir’.
- *Estocástico*, del griego “stochastikós” y significa ‘hábil en conjeturar’.
- *Azar*, del árabe hispánico “azzahr” y hacía alusión a una flor, la misma que se pintaba en una taba (hueso utilizado antiguamente para jugar a algo parecido a los dados), de ahí que finalmente la palabra azar se relacionara con la buena o mala ‘fortuna’.

La idea de probabilidad es uno de esos conceptos que cualquier ser humano tiene preaprendido. Todos tenemos conocimiento intuitivo de lo que supone que una cosa sea muy difícil que ocurra (acertar en la lotería) o de algo que sea más fácil que ocurra (lanzar una moneda y que salga cara). Otra cosa es la definición matemática. Desde el punto de vista formal, el concepto de probabilidad se puede abordar desde tres puntos de vista diferentes: **Bernouilli, Laplace y Kolmogorov**, como veremos en próximas secciones.

La importancia de la probabilidad radica en que, mediante este recurso matemático, es posible ajustar de la manera más exacta posible los imponderables debidos al azar en los más variados campos tanto de la ciencia como de la vida cotidiana.

La teoría de la probabilidad se aplica en áreas variadas del conocimiento, tanto en ciencias (estadística, matemática, física, química, astronomía, meteorología, medicina) como en ciencias sociales (sociología, psicología social, economía).

3.2. Sucesos

Definición 3.1:

Experimento aleatorio, estocástico o de azar es aquel que, aún repetido en análogas condiciones, tiene un *resultado impredecible*, siempre da resultados diferentes.

Son experimentos aleatorios:

- Lanzar una moneda al aire para observar si al caer sale cara o cruz.
- Sacar y observar el resultado al extraer una carta de una baraja.
- Lanzar un dado para observar la puntuación obtenida.

Controversia

Existe cierta controversia^a sobre si los fenómenos aleatorios existen realmente o simplemente surgen del desconocimiento de los factores que desencadenan el mismo o de las leyes físicas que lo rigen. Por ejemplo, si en el lanzamiento de un dado conociéramos exactamente la fuerza, altura al suelo y ángulo del lanzamiento, las dimensiones exactas del dado y las propiedades del suelo, se podría mediante complejos cálculos conocer el resultado final. Es por esto que algunas veces se define un fenómeno aleatorio como aquel en el que pequeños cambios en sus factores producen grandes diferencias en su resultado (teoría del caos).

Esto no quiere decir necesariamente que exista un completo determinismo científico, sino que en ocasiones el azar es consecuencia de la ignorancia de un suceso o de la incapacidad para procesar toda la información que se tiene.

Algunas propuestas realizadas desde la física, como la interpretación de Copenhague de la mecánica cuántica sostienen que a nivel atómico existen los fenómenos aleatorios genuinos.

Recientemente ha aparecido la propuesta de que algunos sistemas físicos, en concreto los sistemas macroscópicos caóticos podrían ser genuinamente no-computables aunque deterministas, eso implica que aun siendo deterministas no es posible calcular con seguridad su evolución futura, mostrando un comportamiento aparentemente aleatorio.

^aAngulo Bustíos, César (2011). '1'. Estadística. Universidad de Piura. ISBN 978-9972-48-137-6.

Definición 3.2:

Espacio muestral de un experimento aleatorio es ‘el conjunto de todos los resultados posibles de un experimento aleatorio’, se representa por E o Ω .

- Para el caso del lanzamiento de una moneda, $E = \{C, X\}$, esto es, sale cara o cruz.

- Para el resultado de extraer una carta de una baraja española, $E = \{1\text{oros}, 2\text{oros}, \dots, 12\text{bastos}\}$, es decir, una cualquiera de las cuarenta cartas que componen la baraja ($\{1, 2, 3, 4, 5, 6, 7, 10 = \text{sota}, 11 = \text{caballo}, 12 = \text{rey}; \text{oros, copas, espadas, bastos}\}$).
- Para el lanzamiento de una dado, $E = \{1, 2, 3, 4, 5, 6\}$

Teorema 3.1:

Tipos de espacios muestrales:

- Espacios muestrales **Discretos o numerables**
 - Espacios muestrales **finitos**.
Constan de un número finito de elementos, por ejemplo, '*lanzamiento de un dado*': $E = \{1, 2, 3, 4, 5, 6\}$
 - Espacios muestrales **infinitos numerables**.
Constan de un número infinito numerable de elementos, por ejemplo, '*lanzar un dado hasta obtener un cinco*': $E = \mathbb{N}$
- Espacios muestrales **continuos**, que siempre son infinitos no numerables.
Constan de un número infinito no numerable de elementos, por ejemplo, '*elección al azar de un número del intervalo [0,1]*': $E = [0, 1]$

Sería conveniente dar un vistazo al apéndice A.1 Conjuntos.

Definición 3.3:

Suceso aleatorio: es cada uno de los subconjuntos del espacio muestral.

- En el experimento de lanzar un dado, $E = \{1, 2, 3, 4, 5, 6\}$, son sucesos:

$$A = \{2, 4, 6\} : \text{'ha salido par'}$$

$$B = \{3, 6\} : \text{'es múltiplo de 3'}$$

$$C = \{1\} : \text{'sale el uno'}$$

etc

3.2.1. Tipos de sucesos

Definición 3.4:

Sucesos elementales: están formados por un solo elemento del espacio muestral.

Sucesos compuestos: formados por dos o más elementos del espacio muestral (o por ninguno).

Suceso seguro: es el que siempre se verifica, coincide con el espacio muestral E .

Suceso imposible: es el que nunca se verifica, se representa por \emptyset (se trata del suceso formado por ningún elemento del espacio muestral, el conjunto vacío – ver apéndice A.1 –)

Los sucesos son subconjuntos del espacio muestral, incluidos el propio E y el conjunto vacío \emptyset (subconjuntos impropios).

Ejemplo 3.1:

En el experimento de lanzar un dado, $E = \{1, 2, 3, 4, 5, 6\}$,

$\{1\}$ es un suceso elemental.

$\{1, 2, 3, 4, 5, 6\}$ es el suceso seguro.

$\{2, 4, 6\}$ es un suceso compuesto.

\emptyset es el suceso imposible. ($\{-8.73\}$)

El conjunto de *todos* los conjuntos de un espacio muestral recibe el nombre de **espacio de sucesos** y se designa por \mathcal{S}

En el experimento del lanzamiento de una moneda: $E = \{C, X\}$ y $\mathcal{S} = \{\emptyset, \{C\}, \{X\}; E\}$

Teorema 3.2:

Si en un experimento aleatorio el espacio muestral tiene n elementos, se puede hablar de hasta 2^n sucesos en el espacio de sucesos.

Al número de elementos de un conjunto X se le llama cardinal, $card(X)$. Entonces,

$$\text{Si } card(E) = n \Rightarrow card(\mathcal{S}) = 2^n$$

Demostración. Como los sucesos no son más que subconjuntos, probar este teorema consiste en probar que para un conjunto de n elementos, el conjunto de sus *partes* (subconjuntos) tiene 2^n elementos. Puede verse esta demostración en el apéndice A.1. \square

Definición 3.5:

Dado un suceso cualquiera de un experimento aleatorio, $A \subset E$, se llama **suceso contrario o complementario** de A , que denotaremos por A' , \bar{A} o A^C a aquel suceso que se verifica cuando no lo hace el suceso A .

Evidentemente, $(A')' = A$

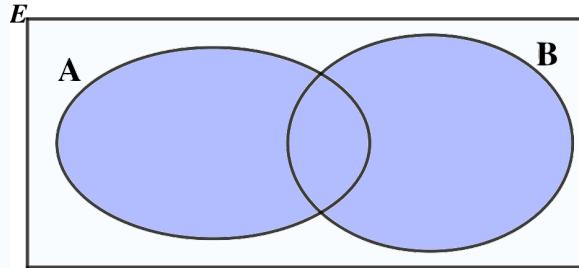
Al lanzar un dado, el suceso contrario a $A = \text{'ha salido 2 o 5'} = \{2, 5\}$, es $A' = \{1, 3, 4, 6\}$.

3.2.2. Operaciones con sucesos**Definición 3.6:**

Dados dos sucesos de un espacio muestral asociado a un experimento aleatorio, $A, B \subset E$, se definen:

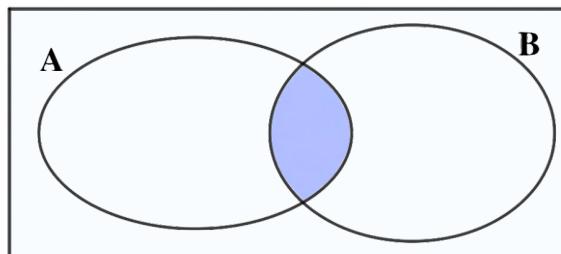
Suceso unión, $A \cup B$, es el suceso que se verifica siempre que lo hagan A o B o ambos.

$$A \cup B$$



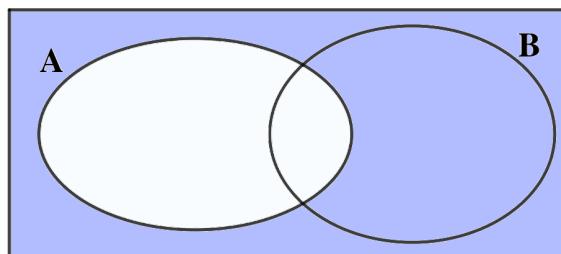
Suceso Intersección, $A \cap B$, es el suceso que se verifica siempre que lo hagan A y B simultáneamente.

$$A \cap B$$



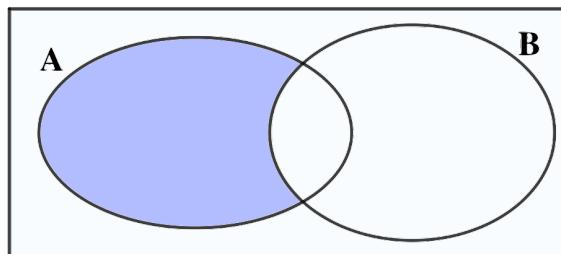
Suceso Complementario, A' , es el suceso que se verifica siempre que no lo haga A .

$$A'$$



Suceso diferencia, $A - B$ o $A \setminus B$, o $A \sim B$, es el suceso que se verifica siempre que lo haga A pero no lo haga B .

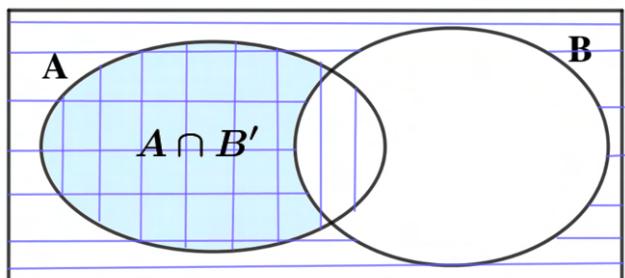
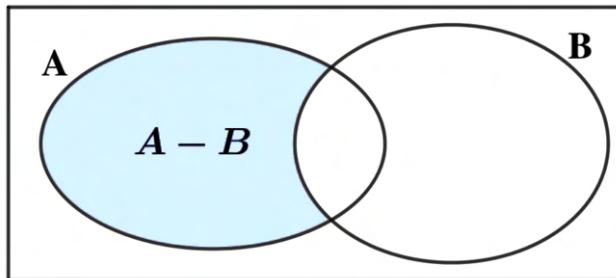
$$A - B = A \setminus B = A \sim B$$



Teorema 3.3:

$$A - B = A \cap B'$$

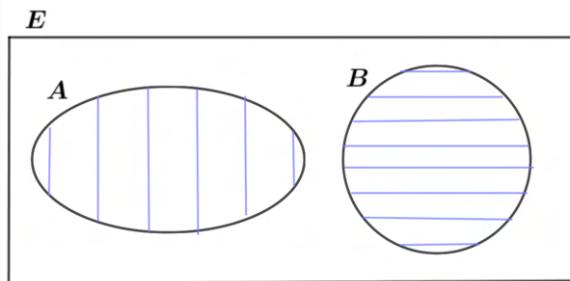
Demostración. Presentamos estos diagramas de Ven como prueba. En el de la derecha, A está rayado en vertical y B' en horizontal, por tanto, la intersección es la zona en que se cruzan las líneas.



□

Definición 3.7:

Dos sucesos A y B son **incompatibles** si no se verifican nunca simultáneamente, es decir,
 A y B incompatibles si
 $A \cap B = \emptyset$



Evidentemente, A y A' son incompatibles.

Si A y B sí pueden verificarse simultáneamente, $A \cap B \neq \emptyset$, decimos que A y B son **compatibles**.

Por ejemplo, los sucesos A =“sale par” y B =“sale impar” al lanzar un dado son sucesos incompatibles.

En muchos casos, para obtener el espacio muestral, resulta conveniente dibujar un “**diagrama de árbol**” (Ver apéndice B Combinatoria), sobre todo en **experimentos compuestos**, los que constan de dos o más partes, como el lanzamiento de dos monedas (o dados) donde, para analizar todos los posibles resultados, consideramos que primero lanzamos la primera moneda y luego la segunda.

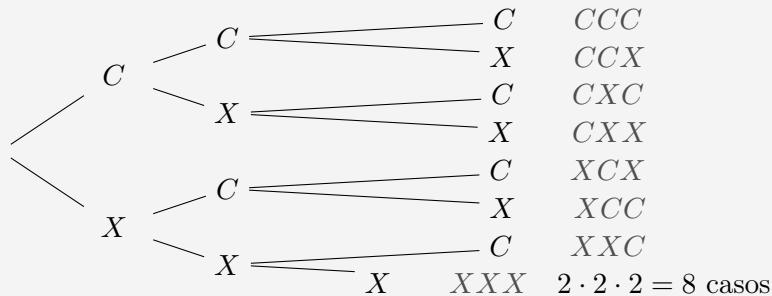
Teorema 3.4:

Propiedades de las operaciones con sucesos.

Sean A , B , y C sucesos. E el suceso seguro y \emptyset el seguro imposible, se cumple:

	Unión \cup	Intersección \cap
1. Asociativa	$A \cup (B \cup C) = (A \cup B) \cup C$	$A \cap (B \cap C) = (A \cap B) \cap C$
2. Conmutativa	$A \cup B = B \cup A$	$A \cap B = B \cap A$
3. Idempotente	$A \cup A = A$	$A \cap A = A$
4. Simplificación	$A \cup (A \cap B) = A$	$A \cap (B \cup A) = A$
5. Distributiva	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
6. Neutro	$A \cup \emptyset = A$	$A \cap E = A$
7. Absorción	$A \cup E = E$	$A \cap \emptyset = \emptyset$
8. Complementario	$A \cup A' = E$	$A \cap A' = \emptyset$
9. Leyes de Morgan	$(A \cup B)' = A' \cap B'$	$(A \cap B)' = A' \cup B'$

Ejercicio resuelto 3.1. Encuentra el espacio muestral asociado al experimento aleatorio consistente en lanzar tres monedas al aire. Si se define el suceso A =‘al menos una de las monedas es cara’, ¿de cuántos sucesos elementales consta el suceso A ? Describe el suceso B =‘salen dos caras’



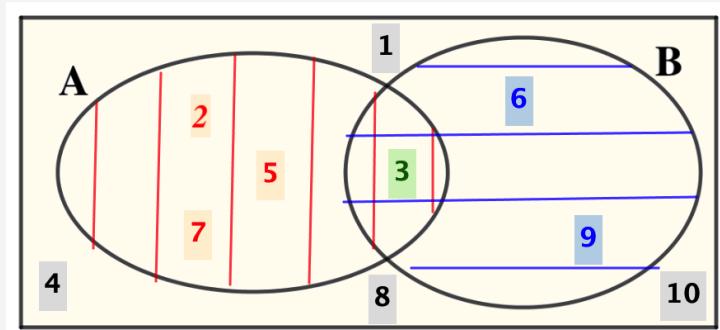
$$E = \{CCC, CCX, CXC, CXX, XCC, XCX, XXC, XXX\}$$

En muchas ocasiones, para determinar un suceso es más sencillo si se piensa en el suceso contrario: $A' = \text{'no sale ninguna cara'} = \text{'todas son cruces'} = \{XXX\}$. Entonces, $A = E - A' = \{CCC, CCX, CXC, CXX, XCC, XCX, XXC, \}$, el suceso A está formado por 7 elementos.

B = 'salen dos caras' = 'una de las monedas es cruz y las otras dos caras' = $\{XCC, CXC, CCX\}$, está formado por tres sucesos elementales.

Ejercicio resuelto 3.2. Una bolsa contiene 10 bolas numeradas del 1 al 10. La experiencia aleatoria consiste en extraer una bola y observar su número. Sean $A = \text{'obtener número primo'}$ y $B = \text{'obtener múltiplo de tres'}$, describe los sucesos $A, B, A \cap B, A \cup B, A', B', A - B, B \cap A', (A \cap B)'$

En muchas ocasiones, para determinar operaciones con sucesos son muy útiles los diagramas de Venn



En la figura (diagrama de Venn), aparece A dibujado con líneas verticales rojas, B con horizontales azules. Donde aparecen las rayas cruzadas es $A \cap B$, cualquier zona rayada es $A \cup B$. La zona rayada sólo verticalmente (rojo) es $A - B$, la rayada sólo horizontalmente (azul) $B - A$. El complementario de A, A' será toda la zona del rectángulo excepto la zona rayada con líneas verticales rojas. El complementario de la unión, $(A \cup B)'$, será toda la zona no rayada de ningún modo. Etc.

$$A = \text{'primo'} = \{2, 3, 5, 7\}; \quad B = \text{'3'} = \{3, 6, 9\}$$

$$A \cap B = \{3\}; \quad A \cup B = \{2, 5, 7, 3, 6, 9\}$$

$$A' = \{1, 4, 6, 8, 9, 10\}; \quad B' = \{1, 2, 4, 5, 7, 8, 10\}$$

$$A - B = \{2, 3, 5, 7\} - \{3, 6, 9\} = \{2, 5, 7\}$$

$$B \cap A' = \{3, 6, 9\} \cap \{1, 4, 6, 8, 9, 10\} = \{6, 9\} = B - A$$

$(A \cap B')' \rightarrow$ calculemos, previamente, $A \cap B'$, que sabemos que coincide con $A - B$ (Ver teorema 3.2.2) y tenemos su resultado: $A \cap B' = A - B = \{2, 5, 7\} \rightarrow$
 $\rightarrow (A \cap B')' = E - (A \cap B') = \{1, 3, 4, 6, 8, 9, 10\}$

Ejercicio resuelto 3.3. Se extraen dos cartas de una baraja española (sin reinserción)^a

Se consideran los sucesos $A =$ ‘las dos cartas son de copas’ y $B =$ ‘una carta es de copas y la otra es rey’. Calcula $A \cap B$ y $A \cup B^b$

Los casos en que se pueden extraer dos copas de una baraja son: la primera carta de copas puede ser una cualquiera de las diez que hay en la baraja. Extraída la primera, la segunda puede ser cualquier de las nueve copas que quedan. Total $10 \cdot 9 = 90$ casos. PERO, hemos contado cada caso dos veces, por ejemplo, si la primera es el ‘siete de copas’ y la segunda el ‘as de copas’ o si la primera es el ‘as’ y la segunda el ‘siete’; hay que dividir por 2, así, los diferentes casos en que se pueden dos copas de una baraja son $90/2=45$.

Si recordamos la combinatoria (ver apéndice B Combinatoria), tenemos un conjunto de 10 elementos (cartas que sean copas en la baraja española) de la que vamos a extraer dos, *sin que nos importe el orden y sin poder repetir* (sin reinserción). Tenemos

$$C_{10}^2 = \binom{10}{2} = \frac{10!}{(10-2)! 2!} = \frac{10 \cdot 9 \cdot 8!}{8! 2} = 45$$

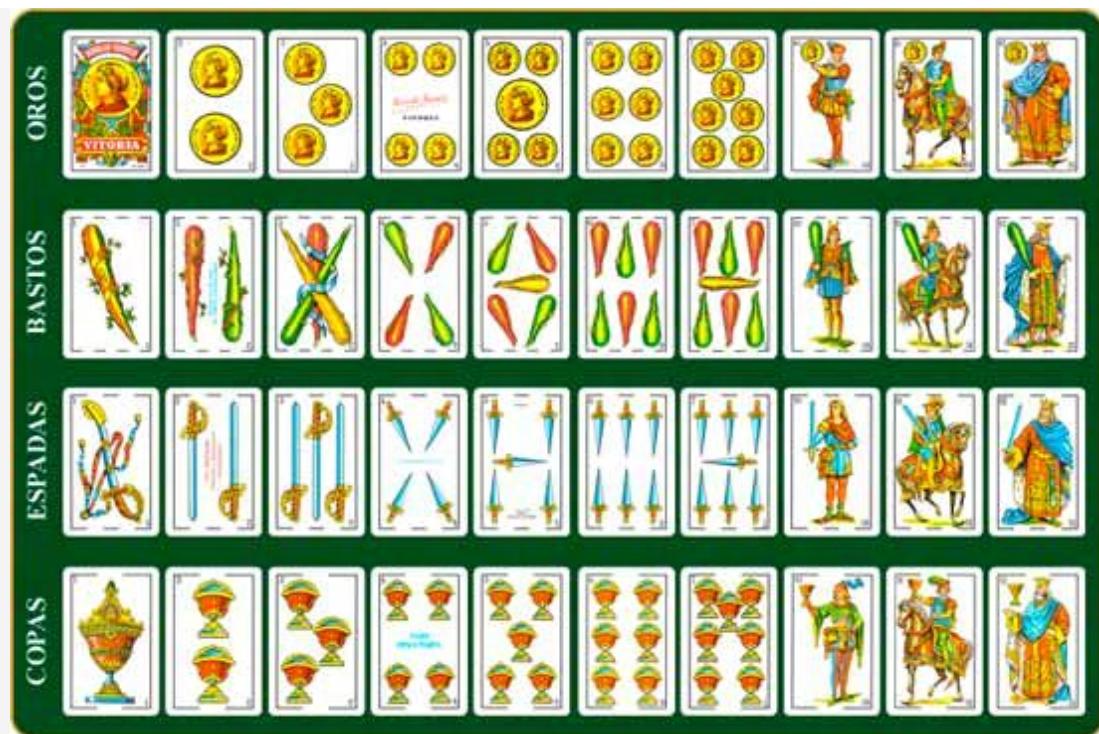
Esquemáticamente, si designamos por *or*, *co*, *es*, *ba* respectivamente a las cartas de oros, copas espadas y bastos (palos de la baraja) y por *n* al número de la misma, $n \in \{1, 2, 3, \dots, 9, 10\}$, el suceso pedido es:

$A = \{(n-co; m-co); \text{ con } n < m \in \{1, 2, 3, \dots, 9, 10\}\}$ (Al exigir $n < m$ evitamos que influya el orden y que haya repeticiones).

El suceso B, usando esta misma notación, sería:

$B = \{(n-co; rey-or); (n-co; rey-co); (n-co; rey-es); (n-co; rey-ba); n \in \{1, 2, \dots, 9\}\} \cup \{(rey-co; rey-or); (rey-co; rey-es); (rey-co; rey-ba)\}$, en total, $9 + 9 + 9 + 9 + 1 + 1 + 1 = 39$ casos.

Las distintas posibilidades de extraer dos cartas, sin reinserción, de una baraja española son, como se ha explicado anteriormente, $\frac{40 \cdot 39}{2} = \binom{40}{2} = 780$ casos posibles. (1or,2or), (1or,3or), ..., (caballo bastos,rey bastos)=(9-ba,rey-ba)



$\Rightarrow A \cap B = \{(n - co; rey - co); n \in \{1, 2, \dots, 9\}\}$, hay 9 casos.

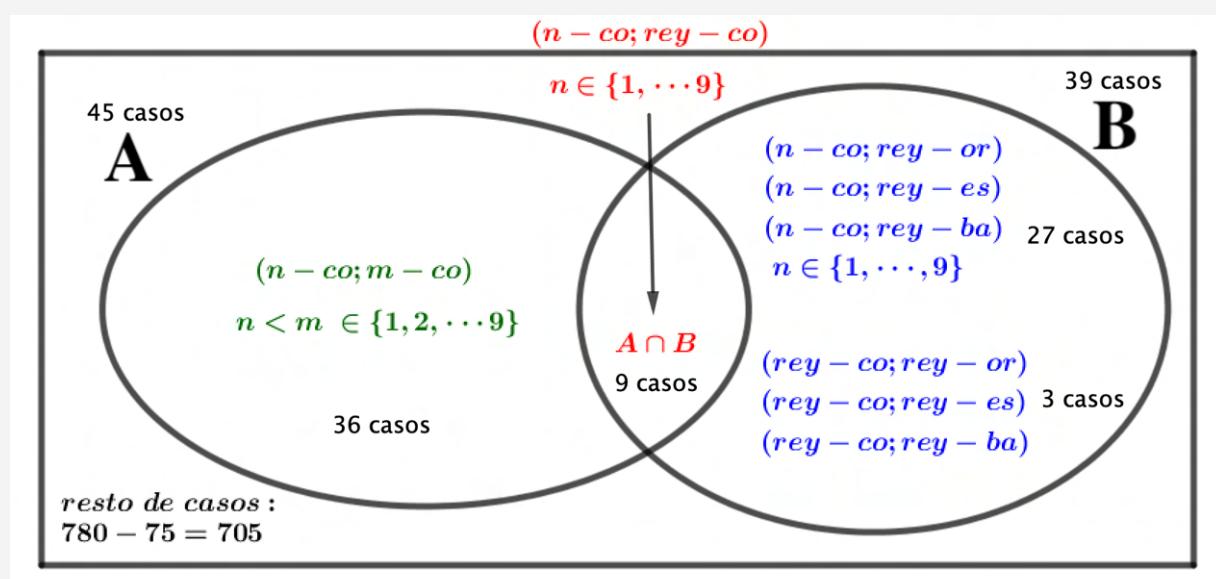
$\Rightarrow A \cup B = \{(n - co; m - co); n < m \in \{1, 2, \dots, 9\}\} \cup$

$\{(n - co; rey - co); n \in \{1, 2, \dots, 9\}\} \cup$

$\{(n - co; rey - or); (n - co; rey - es); (n - co; rey - ba); n \in \{1, \dots, 9\}\} \cup$

$\{(rey - or; rey - co); (rey - es; rey - co); (rey - ba; rey - co)\}$

Es decir, $A \cup B$ está formado por cualquier ‘pareja de cartas donde una de ellas ha de ser de copas y la otra ha de ser o bien copas o bien un rey’.



^aLa extracción de varias cartas de una baraja puede realizarse de dos modos distintos, **con reinserción** (devolviendo cada carta extraída de nuevo al mazo antes de la siguiente extracción) o **sin reinserción** (cada carta que se extrae no se vuelve a reinsertar en la baraja).

^bPara la resolución de este ejercicio, consideramos las cartas marcadas con los números de 1 al 10, es decir, la sota será en 8, el caballo el 9 y el rey el 10.

3.3. Probabilidad



Los inicios de la teoría de la probabilidad.

La teoría matemática de la probabilidad tuvo como uno de sus primeros puntos de partida el intentar resolver un problema particular concerniente a una apuesta de juego de dados entre dos personas. El problema al que nos referimos involucraba una gran cantidad de dinero y puede plantearse de la siguiente forma:^a

Dos jugadores escogen cada uno de ellos un numero del 1 al 6, distinto uno del otro, y apuestan 32 doblones de oro a que el número escogido por uno de ellos aparece en tres ocasiones antes que el número del contrario al lanzar sucesivamente un dado. Suponga que el numero de uno de los jugadores ha aparecido dos veces y el numero del otro una sola vez. ¿Cómo debe dividirse el total de la apuesta si el juego se suspende?

Este problema, que data de la Edad Media, si no antes, es el conocido como “el problema de la partida interrumpida” y fue propuesto por Antoine Gombaud, Caballero de Méré (1607-1684), escritor francés y matemático aficionado a Blaise Pascal (1623-1662) quien lo consultó con Pierre de Fermat (1601- 1665) e iniciaron un intercambio de cartas a propósito del problema. Esto sucede en el año 1654. Con ello se inician algunos esfuerzos por dar solución a éste y otros problemas similares que se plantean. Con el paso del tiempo se sientan las bases y las experiencias necesarias para la búsqueda de una teoría matemática que sintetice los conceptos y los métodos de solución de los muchos problemas particulares resueltos a lo largo de varios años.



Pierre de Fermat (1601-1665)



Blaise Pascal (1623-1662)

En el segundo congreso internacional de matemáticas, celebrado en la ciudad de Paris en 1900, el matemático David Hilbert (1862-1943) plantea 23 problemas matemáticos de importancia. Uno de estos problemas es el de encontrar axiomas o postulados a partir de los cuales se pueda construir una teoría matemática de la probabilidad. Aproximadamente treinta años después, en 1933, el matemático ruso A. N. Kolmogorov (1903-1987) propone ciertos axiomas que a la poste resultaron adecuados para la construcción de la teoría de la probabilidad.

^aEsto ocurrió en la sociedad francesa de 1650, donde el juego era un entretenimiento muy corriente y no demasiado sujeto a restricciones legales. Los juegos eran numerosos y cada vez más complicados, y se apostaban grandes cantidades de dinero.

Definición 3.8:

Frecuencia absoluta de un suceso A , N_A , es el número de veces que se verifica A cuando se realiza el experimento N veces.

Frecuencia relativa de un suceso es el cociente entre su frecuencia absoluta y el número N de veces que se realiza el experimento, N_A/N .

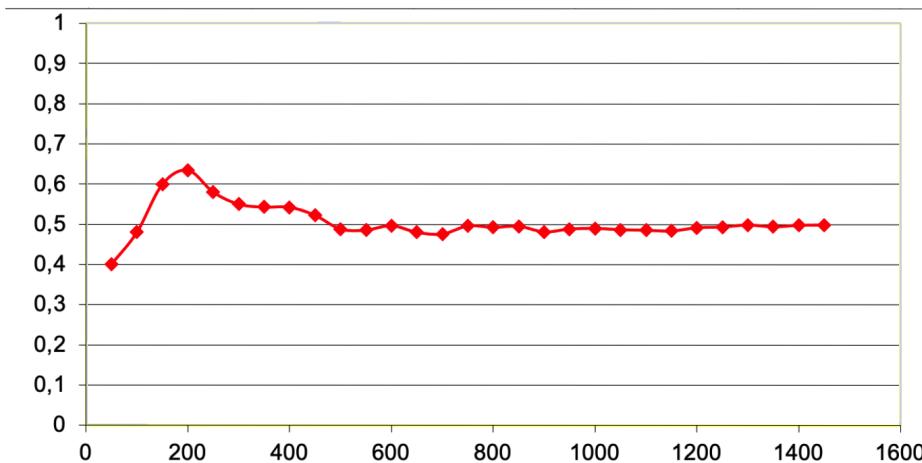
Ley de los grandes números: En un experimento aleatorio, al realizar un gran número de pruebas, N , la frecuencia relativa de un suceso A tiende a estabilizarse alrededor de un número fijo, $p(A)$, que se llama **probabilidad** de A :

$$p(A) = \lim_{N \rightarrow \infty} \frac{N_A}{N}$$

Esta es la *definición frecuentista o “a posteriori”* de la probabilidad. Para asignar probabilidades a sucesos se recurre a la experimentación.

Jacob Bernoulli, 1689, definió la probabilidad usando la ley de los grandes números de Chebishev.

“Ley de los grandes números”



Sucesivos lanzamientos de una moneda y frecuencia relativa del suceso 'a salido cara'

Ley de los grandes números

Que la frecuencia se va estabilizando cuando aumenta el número de experiencias es una *verdad empírica*, no demostrable pero sí reiteradamente comprobable. Su enunciado es el *principio básico del azar, la ley de los grandes números*. Aunque al principio parezca haber grandes fluctuaciones en la frecuencia relativa, a medida que aumente el número de experimentos éstas se van suavizando y la frecuencia relativa tiende a un número fijo, la probabilidad.

Ejemplo 3.2:

Probabilidad de que al lanzar una chincheta sobre la mesa quede con la punta hacia arriba o apoyada en la mesa.

A priori, no lo sabremos, tenemos que recurrir a la experimentación.

Si dejemos caer 100 chinchetas, y 23 de ellas caen con la punta hacia arriba, estimaremos la probabilidad de este suceso en 0.23. Si dejamos caer 1.000 y obtenemos 307 hacia arriba estimaremos su probabilidad en 0.307, siendo esta estimación más segura que la anterior. Y, así sucesivamente, vamos adquiriendo seguridad sobre el valor que asignamos a la probabilidad de ese suceso a medida que aumentemos el número de experimentos realizados. Del mismo modo deberíamos proceder si tuviéramos la sospecha de que, por ejemplo un dado o una moneda están trucados.



Esta definición de probabilidad tiene un inconveniente: es necesario realizar un gran número de pruebas para asignar como probabilidades el valor al que se aproximan las frecuencias relativas del suceso en estudio y además, siempre obtenemos un valor aproximado, en lugar del valor exacto de la proba-

bilidad. Es la *asignación de probabilidad “a posteriori”*. En muchas ocasiones no hay otra manera de hacerlo (horas de duración de una batería, nivel de riesgo en una operación, etc).

Definición 3.9:

Definición clásica de probabilidad. Regla de Laplace

Pierre Simon Laplace (1749-1827) formula la primera definición que se conoce de probabilidad:

“La probabilidad de un suceso A, que representaremos por $p(A)$, es el cociente entre el número de casos favorables a dicho suceso y el número de casos posibles”.

$$p(A) = \frac{\text{número de casos favorables}}{\text{número de casos posibles}} \left(= \frac{\text{favorables}}{\text{posibles}} \right)$$

Para aplicar esta definición es necesario que los sucesos elementales del espacio muestral han de ser igualmente probables (*equiprobables*).

Son casos favorables los elementos que componen el suceso A , y los casos posibles todos los resultados del experimento, es decir, todo el espacio muestral.

Así, si una experiencia aleatoria consta de n sucesos elementales y es razonable suponer que ninguno de ellos tiene más probabilidad de salir que los demás (equiproblables), la probabilidad de cada uno de ellos es $1/n$. Para un suceso que conste de k sucesos elementales, su probabilidad será k/n .

Esta definición de probabilidad, muchas veces trivial, es en ocasiones muy complicada (determinar si los sucesos son o no equiprobables). Es la *asignación de probabilidades “a priori”*.

Ejemplo 3.3:

Calcular la probabilidad de obtener suma 7 al lanzar dos dados.

El espacio muestral de lanzar dos dados y sumar las puntuaciones obtenidas es $E = \{2, 3, 4, \dots, 12\}$, pero, ¡CUIDADO!, todos los sucesos no son equiprobables. Por ejemplo, obtener suma 6 se puede hacer con dos ‘tres’, pero también con un ‘cuatro’ y un ‘dos’ (0 un 2 y un 4) o un ‘cinco’ y un ‘uno’; en cambio, suma 12 solo se obtiene si sales dos ‘seis’. Para solucionar este aparente problema, consideremos dos dados distintos, observemos los 36 resultados posibles (por cada uno de los resultados del primer dado, hay 6 del segundo) y sumemos las puntuaciones obtenidas.

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12



Ahora sí, vemos que de los 36 sucesos posibles, solo en 6 ocasiones la suma es siete, por lo que $p(\text{suma } 7) = 6/36$ (es la suma más probable).

Probabilidades a priori y a posteriori

Son **probabilidades a priori** aquellas que se pueden determinar de antemano, sin realizar ningún tipo de comprobación experimental, **en base a consideraciones teóricas**. Un ejemplo puede ser el de la probabilidad de obtener un cinco en el lanzamiento de un dado perfecto.

Son **probabilidades a posteriori** las que se determinan con posterioridad a la experimentación, **a través de la experiencia**. Es necesario estimar la probabilidad estudiando el valor límite al que se acercan las frecuencias relativas al realizar un gran número de pruebas en análogas condiciones. Es importante tener en cuenta, que las frecuencias relativas que se obtienen de un número reducido de pruebas no puede servir para realizar una estimación fiable de la probabilidad de un determinado suceso, es necesario un elevado número de pruebas, precisamente la garantía de que la definición de probabilidad a posteriori sea buena la da la Ley de los grandes números, $N \rightarrow \infty$.

Definición 3.10:

Definición Axiomática^a de probabilidad. Andrei Nicolaievich **Kolmogorov** (1903-1987).

Axiomas:

A1. La probabilidad de un suceso es un número positivo o nulo: $p(A) \geq 0$

A2. La probabilidad del suceso seguro es uno: $p(E) = 1$

Es decir,
$$\boxed{0 \leq p(A) \leq 1}$$

A3. Si A y B son dos sucesos *incompatibles* de un experimento aleatorio ($A \cap B = \emptyset$), $p(A \cup B) = p(A) + p(B)$

Propiedades:

P1. Para sucesos complementarios^b,
$$\boxed{p(A') = 1 - p(A)}$$

P2. La probabilidad del suceso imposible es cero: $p(\emptyset) = 0$

P3. Si $A \subset B \rightarrow p(A) \leq p(B)$

P4. En general, para sucesos compatibles,

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

^aUn axioma es un principio o afirmación matemática que se acepta sin demostración. Para que un conjunto de axiomas sea válido, es necesario que a partir de ellos no se llegue a afirmaciones contradictorias.

^bEn ocasiones, es más sencillo calcular la probabilidad del suceso contrario al que nos piden. En estos casos, esta propiedad resulta de una relevante utilidad.

$$p(A \cup P \cup C) = p(A) + p(P) + p(C) - p(A \cap P) - p(A \cap C) - p(P \cap C) + p(A \cap P \cap C)$$

— Los axiomas establecen que la probabilidad sea una “*medida*” en matemáticas.

— El segundo axioma fija la cantidad total de probabilidad. Pero podría ser cualquier número positivo, el tomar $p(E) = 1$ es por similitud con las frecuencias relativas y significa que la “cantidad de probabilidad” se dará en “tantos por uno”. En ocasiones, las probabilidades se dan en % (lo cual equivale a tomar $p(E) = 100$).

Ejemplo 3.4:

Hallar la probabilidad de que al lanzar tres monedas se obtenga al menos una cara.

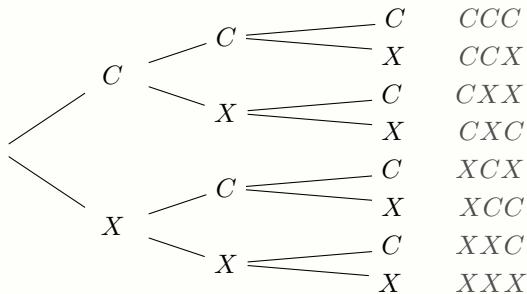
Este es un caso en que es mucho más fácil calcular la probabilidad del suceso contrario . (Propiedad P1).

Sea A =‘obtener al menos una cara en tres lanzamientos de una moneda’. Su suceso contrario será A' =‘salen tres cruces.’

Suponiendo monedas no cargadas y usando la *ley de Laplace*, $P(A') = 1/8$, ya que solo en una ocasión de las ocho posibles se obtienen tres cruces, XXX (primera moneda C,X; independientemente, la segunda puede ser C,X; lo mismo para el tercer lanzamiento, C,X).

Luego, $p(A) = 1 - 1/8 = 7/8$. En el siguiente diagrama de árbol pueden contabilizarse estos casos.

Distintos resultados en el lanzamiento de una monedas tres veces.



Ejemplo 3.5:

Al lanzar un dado de Laplace (sucesos elementales equiprobables; dado equilibrado, no lastrado), se consideran los sucesos A =‘sale un número impar’ y B =‘sale múltiplo de 3’. Calcular $p(A \cap B)$ y $p(A \cup B)$

$$A = \{1, 3, 5\}; B = \{3, 6\} \rightarrow A \cap B = \{3\}; A \cup B = \{1, 3, 5, 6\}$$

— Usando la ley de Laplace, $\left(\frac{\text{favorables}}{\text{posibles}} \right)$, tenemos:

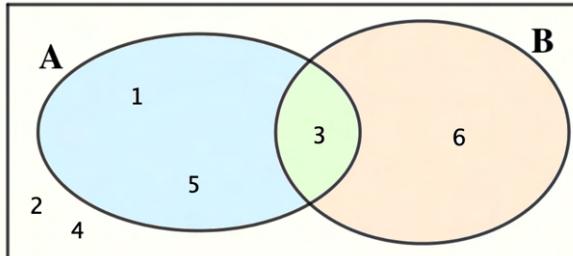
$$\Rightarrow p(A \cap B) = 1/6; \quad p(A \cup B) = 4/6$$

— Si usamos pa propiedad P5 de la axiomática de Kolmogorov,

$$\Rightarrow p(A \cup B) = p(A) + p(B) - p(A \cap B) = 3/6 + 2/6 - 1/6 = 4/6$$

— Simplemente, al observar la figura (diagrama de Venn)

$$\Rightarrow p(A \cup B) = 4/6$$

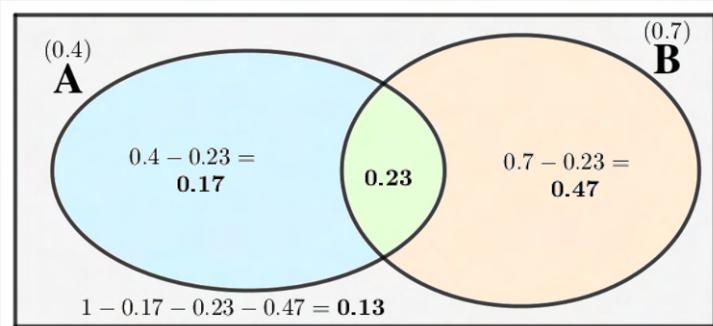


Ejercicio resuelto 3.4. Supóngase un determinado grupo de personas de entre las que se elige una al azar y en que se consideran los siguientes sucesos: $A = \text{'la persona elegida lleva gafas'}$, $B = \text{'la persona elegida tiene los ojos marrones'}$. Se sabe, que en este grupo se dan las siguientes probabilidades: $p(A) = 0.4$; $p(B) = 0.7$; $p(A \cap B) = 0.23$.

Calcula la probabilidad de que una persona elegida al azar,

- No lleve gafas.
- Use gafas o tenga los ojos marrones.
- Lleve gafas pero no tenga los ojos marrones.

Un **diagrama de Venn** nos ayudará a resolver el problema.



Colocando 0.23 en la intersección, la zona azul $A - B$ (gafas pero no ojos marrones) tendrá una probabilidad de $0.4 - 0.23 = 0.17$, análogamente encontramos que la probabilidad de que la persona tenga ojos marrones pero no lleve gafas ($B - A$, zona naranja) es 0.47 y que la probabilidad de que la persona elegida ni tenga ojos marrones ni use gafas, la zona del exterior, será 0.13, puesto que la probabilidad total ha de ser uno.

Sin más que observar el gráfico:

- $p(\text{no lleve gafas}) = 0.47 + 0.13 = 1 - 0.4 = 0.6$
- $p(\text{use gafas o tenga ojos marrones}) = 0.17 + 0.23 + 0.47 = 1 - 0.13 = 0.87$
- $p(\text{lleva gafas pero no tenga los ojos marrones}) = 0.17$

También podríamos resolver el problema sin necesidad de usar un diagrama de Venn.

- No llevar gafas es el suceso contrario a A , A' .

Así, $p(A') = 1 - p(A) = 1 - 0.4 = 0.6$

b) Usar gafas o tener ojos marrones es el suceso unión, por ello,

$$p(A \cup B) = p(A) + p(B) - p(A \cap B) = 0.4 + 0.7 - 0.23 = 0.87$$

c) Llevar gafas sin tener los ojos marrones es el suceso diferencia, $A - B$. Como los sucesos $A - B$ y $A \cap B$ son disjuntos (zona morada y zona verde), la probabilidad de su unión (el suceso A) es la suma de probabilidades (Axioma 3 de Kolmogorov):

$$(A - B) \cap (A \cap B) = \emptyset \rightarrow$$

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

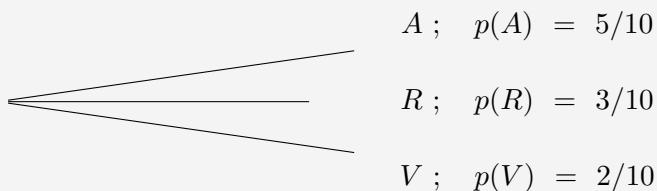
$$0.4 = p(A - B) + 0.23 \rightarrow p(A - B) = 0.17$$

Realmente, el uso del diagrama de Venn simplifica enormemente la resolución del problema.

Ejercicio resuelto 3.5. Una urna contiene 5 bolas azules, 3 rojas y 2 verdes. Consideremos el experimento consistente en extraer de la urna una bola al azar, calcular las siguientes probabilidades:

- a) Sale bola azul; b) la bola extraída no es roja; c) sale bola negra; d) sale bola azul o verde.

En este caso, los sucesos A='bola azul', R='bola roja' y V='bola verde' son mutuamente incompatibles por lo que podemos aplicar el axioma 3, pero, resolveremos el problema con la ayuda de un **diagrama de árbol**, lo que facilitará mucho la resolución.



a) $p(A) = 5/10$

b) $p(R') = 1 - 3/10 = 7/10 = 5/10 + 2/10$

c) $p(N) = 0/10 = 0$

d) $p(A \cup V) = 5/10 + 2/10 = 7/10 = p(R')$

Ejercicio resuelto 3.6. En un grupo de 40 alumnos hay 22 chicos y 18 chicas. Llevan gafas 8 chicos y 6 chicas.

- a) Elegido un alumno al azar, calcula la probabilidad de que sea chico y no lleve gafas.

- b) ¿Cuál es la probabilidad de elegir a una persona con gafas?

- c) Si sabemos que la persona que va a ser elegida es un chico, ¿cuál es ahora la probabilidad de que lleve gafas?

Vamos a ver un nuevo tipo de 'gráfico' que nos facilitará la resolución del problema, en esta ocasión usaremos una **tabla de contingencia**.

	chico	chica	
gafas	8	6	14
sin gafas	14	12	26
	22	18	40

Colocamos, en negrita, los datos del problema y rellenamos los huecos y totales de la tabla.

a) $p(\text{chico y gafas}) = 8/40$

b) $p(\text{gafas}) = 14/40$; c) $p(\text{gafas, sabiendo que es chico}) = 8/22$

Esto nos llevará al siguiente apartado, *probabilidad condicionada*.

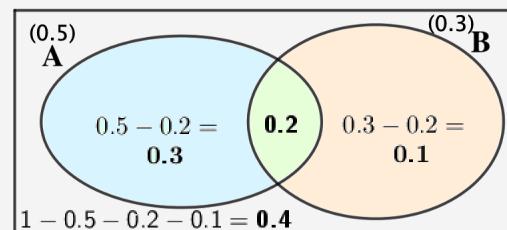
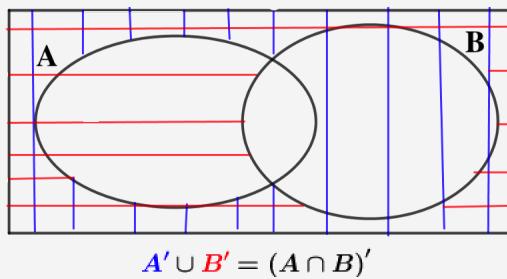
En múltiples ocasiones, un diagrama de Venn, un diagrama de árbol o una tabla de contingencia, que obviamente responda al enunciado del problema, nos facilitará mucho su resolución. Es la experiencia la que determinará que opción usar en cada momento. (En general, el diagrama de árbol se usará para sucesos incompatibles y para pruebas compuestas y el diagrama de Venn o tablas de contingencia se usará para sucesos compatibles).

En lo que queda de tema, por motivos didácticos, se intentará siempre el uso de estas ayudas y evitar, en lo posible, el excesivo uso de fórmulas.

Ejercicio resuelto 3.7. .

Se sabe que $p(A) = 0.5$, $p(B') = 0.7$ y que $p(A' \cup B') = 0.8$.

Calcular: $p(A \cup B)$; $p(A \cap B)$; $p(A - B)$ y $p(A' \cap B')$.



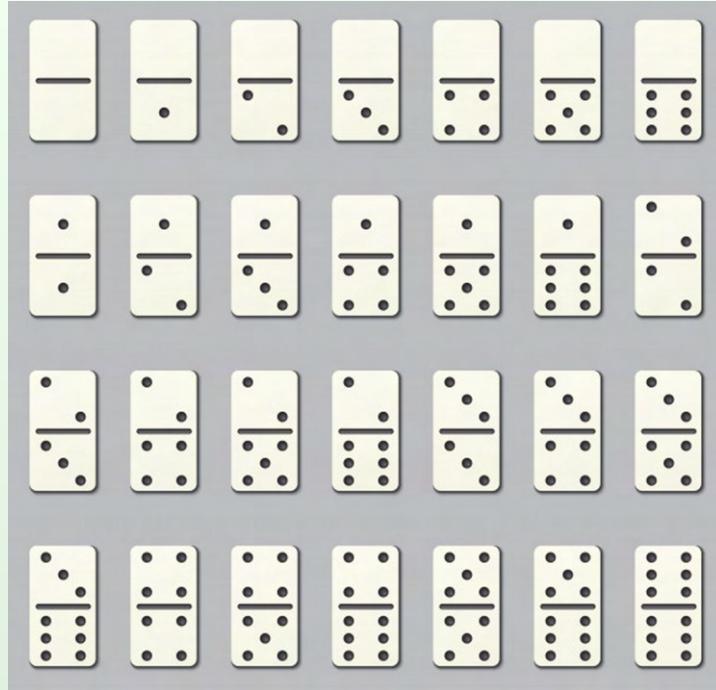
Como $p(B') = 0.7 \rightarrow P(B) = 1 - 0.7 = 0.3$

Hemos dibujado con verticales azules A' y con horizontales rojas B' , la unión está formada por todas las zonas rayadas de cualquier modo (la intersección donde aparecen las rayas cruzadas). Por ello, vemos que $A' \cup B' = (A \cap B)'$ (Leyes de Morgan), luego $p(A' \cup B') = p(A \cap B)' = 1 - p(A \cap B) = 0.8 \rightarrow p(A \cap B) = 0.2$

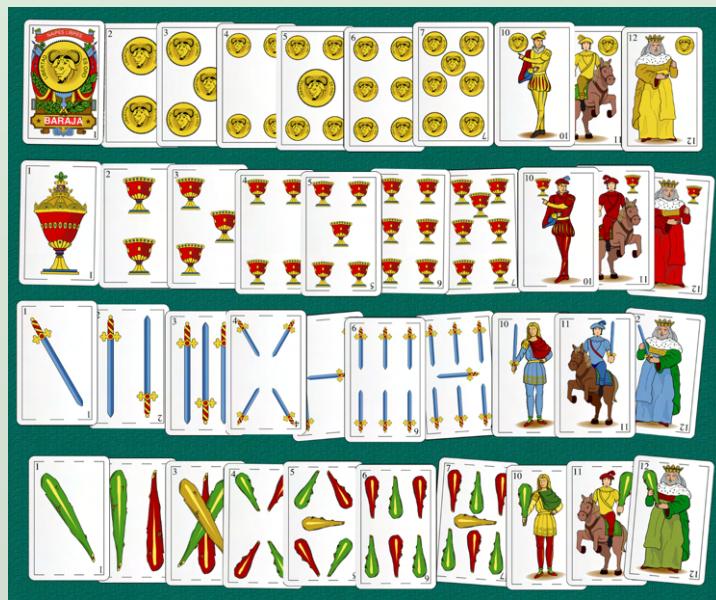
Averiguado esto, es fácil obtener las probabilidades que faltan y aparecen el el segundo diagrama.

Dominó y barajas (española y francesa)

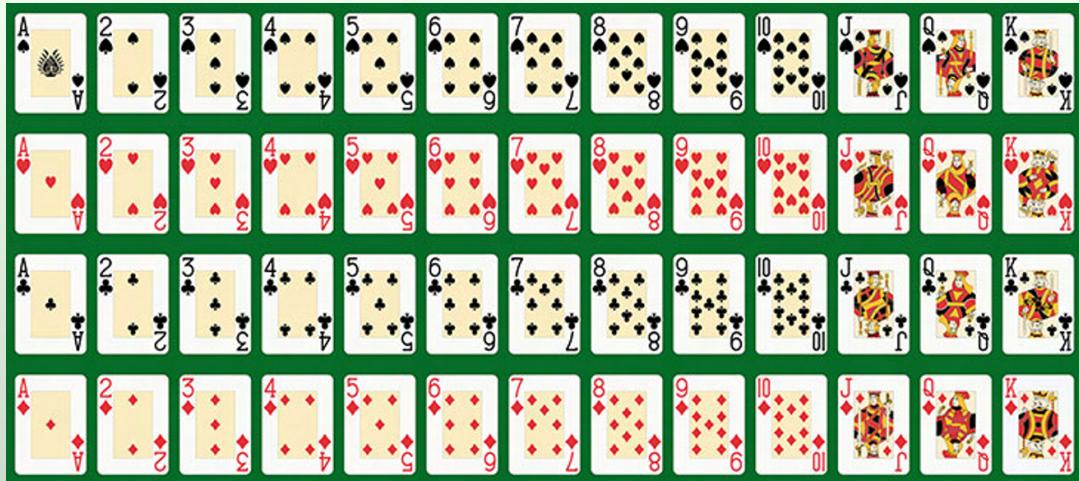
El **dominó** es un juego de mesa en el que se juegan y emplean unas fichas rectangulares, generalmente blancas por la cara y negras del revés. La cara blanca está dividida por dos cuadrados, cada uno de los cuales está numerado mediante disposiciones de puntos del cero al seis. Hay en total 28 fichas, $\{(0,0), (0,1), (0,2), \dots, (6,6)\}$



La **baraja española** es un mazo o conjunto de cuarenta naipes o cartas de la baraja. Los naipes están divididos en cuatro "familias", "pintas" o "palos". Los palos son "ros", "copas", "espadas" "bastos", a cada uno de los cuales le corresponde su iconografía característica. Cada palo tiene diez cartas: siete cartas numeradas del uno al siete, llamadas cartas numéricas y tres figuras (sota, caballo y rey) numeradas correlativamente del diez al doce.



La **baraja francesa** es un conjunto de naipes o cartas, formado por 52 unidades repartidas en cuatro palos: corazones, diamantes (que son rojas), tréboles y picas (que son negras). Cada palo está formado por 13 cartas, 10 numéricas (del 1 al 10) y tres figuras (sota-Jack, reina-Queen y rey-King: J, Q, K).



3.4. Probabilidad condicionada

La *probabilidad condicionada* es uno de los conceptos clave en Teoría de la Probabilidad. Hasta ahora, hemos introducido el concepto de probabilidad considerando que la única información sobre el experimento era el espacio muestral. Sin embargo, hay situaciones en las que se incorpora información suplementaria, como que haya ocurrido otro suceso, con lo que puede variar el espacio de resultados posibles y consecuentemente, sus probabilidades. En este contexto aparece el concepto de probabilidad condicionada.

El objetivo es analizar cómo afecta el conocimiento de la realización de un determinado suceso a la probabilidad de que ocurra cualquier otro.

Definición 3.11:

Se llama **probabilidad condicionada** de que ocurra el suceso B sabiendo que ha ocurrido el suceso A , y se denota por $P(B|A)$ al cociente entre $p(A \cap B)$ y $p(A)$ y *mide las veces que ocurre B de entre las que ocurre A* .

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

La probabilidad de un suceso condicionada a otro es el cociente entre la probabilidad de la intersección y la probabilidad del suceso que condiciona. El suceso que condiciona, su probabilidad, va en el denominador.

Teorema 3.5:

Como la intersección de sucesos es commutativa, $A \cap B = B \cap A$, se tiene:

$$p(A \cap B) = p(A) \cdot p(B|A); \quad p(A \cap B) = p(B) \cdot p(A|B)$$

Demostración. Para demostrar la primera expresión no hay más que despejar en la definición de probabilidad condicionada. En la segunda expresión se han intercambiado los papeles de B y A . \square

Teorema 3.6:

$$p(A \cap B \cap C) = p(A) \cdot p(B|A) \cdot p(C|A \cap B)$$

Demostración. $p(A \cap B \cap C) = p((A \cap B) \cap C) = p(A \cap B) \cdot p(C|A \cap B) = (\rightarrow)$

por el teorema anterior, $(\rightarrow) = p(A) \cdot p(B|A) \cdot p(C|A \cap B)$ \square

Definición 3.12:

Sucesos Dependientes e Independientes

Si $p(B|A) = p(B) \rightarrow$ se dice que A y B son **independientes**. El saber que ha ocurrido A no modifica la probabilidad de que ocurra B , el hecho de que ocurra B no se ve influido por la ocurrencia o no de A , son *independientes*.

Si $p(B|A) \neq p(B) \rightarrow$ se dice que A y B son **dependientes**

Teorema 3.7:

Si A es independiente de $B \Rightarrow B$ es independiente de A .

Demostración. A independiente $B \rightarrow p(A|B) = p(A); \frac{p(A \cap B)}{p(B)} = p(A) \rightarrow$

$\rightarrow \frac{p(A \cap B)}{p(A)} = p(B); p(B|A) = p(B) \rightarrow B$ independiente de A . \square

Teorema 3.8:

Para sucesos independientes: $\mathbf{p(A \cap B) = p(A) \cdot p(B)}$

Demostración. Evidentemente, al ser A y B independientes, $p(B|A) = p(B)$, por lo que al calcular, según la definición, la probabilidad de la intersección: $p(A \cap B) = p(A) \cdot p(B|A) = p(A) \cdot p(B)$ \square

También se puede tomar este hecho como comprobación de si dos sucesos son o no independientes:

$$\mathbf{A \text{ y } B \text{ son independientes} \iff p(A \cap B) = p(A) \cdot p(B)}$$

Atención: no se debe confundir sucesos **independientes** con sucesos **incompatibles**:

- Si $p(B|A) = p(B)$ → se dice que A y B son **independientes**.
- Si $p(A \cap B) = 0$ → se dice que A y B son **incompatibles**.

Podemos reconocer que nos preguntan por una probabilidad condicionada, $p(B|A)$, y no por una normal, $p(B)$ por el enunciado, que será de la forma:

- Calcula la probabilidad de B condicionada a A .
- Calcula la probabilidad de B *sabiendo* que ha ocurrido A .
- *Sabiendo* que ha ocurrido A , calcula la probabilidad de que ocurra B .

¡Ojo a los subjuntivos!

Ejemplo 3.6:

De una baraja de 40 cartas española, se extrae una carta al azar. Calcular:

- la probabilidad de que sea figura (sota, caballo o rey).
- la probabilidad de que sea rey.
- sabiendo que la carta extraída es figura, ¿cuál es la probabilidad de que sea rey?
- sabiendo que la carta extraída es de copas, ¿cuál es la prob. de que sea rey?
- ¿Son independientes los sucesos “la carta es rey” y “la carta es figura”? ¿Y los sucesos “la carta es rey” y “la carta es de copas”?

Llamamos F =“la carta es figura”; R =“la carta es rey”; Co =“la carta es de copas”.

a) $p(F) = (3 \cdot 4)/40 = 3/10 = 0.30$

b) $p(R) = 4/40 = 0.10$

c) $p(R|F) = 0.33$

— con fórmula, $p(R|F) = \frac{p(R \cap F)}{p(F)} = \frac{4/40}{12/40} = 4/12 = 1/3 = 0.33$

— directamente, de la definición: probabilidad sacar un rey (hay 4) sabiendo que el sorteo se efectúa entre las figuras (son 12), entonces, $p(R|F) = 4/12 = 1/3 = 0.33$. ¡Más fácil!

d) $p(R|Co) = 1(\text{el rey de copas})/10(\text{copas}) = 0.10$

e) Dependencia e independencia de los sucesos R con F y co :

— R y F : $p(R|F) = 0.33 \neq 0.10 = p(R) \rightarrow R$ y F son *dependientes*.

— R y Co : $p(R|Co) = 0.10 = p(R) \rightarrow R$ y Co son *independientes*.

Es decir, el hecho de saber o no que la carta a extraer va a ser de copas no afecta a la probabilidad de sacar un rey, en cambio; si sabemos que la extracción va a ser sobre las figuras, si cambia (aumenta en este caso) la probabilidad de sacar rey.

La información modifica la probabilidad.

3.5. Teorema de la probabilidad total y teorema de Bayes

Se llaman **pruebas compuestas o experiencias compuestas** a aquellas en las que se pueden distinguir dos o más etapas.

El desarrollo de los sucesos de una experiencia compuesta se puede representar con un **diagrama en árbol**.

El diagrama de árbol es una representación gráfica de los posibles resultados del experimento, el cual consta de una serie de pasos, donde cada uno de estos tiene un número infinito de maneras de ser llevado a cabo. Se utiliza en los problemas de conteo y probabilidad.

Para la construcción de un diagrama en árbol se partirá poniendo una rama para cada una de las posibilidades, acompañada de su probabilidad. Cada una de estas ramas se conoce como rama de primera generación.

En el final de cada rama de primera generación se constituye, un nudo del cual parten nuevas ramas conocidas como ramas de segunda generación, según las posibilidades del siguiente paso, salvo si el nudo representa un posible final del experimento (nudo final).

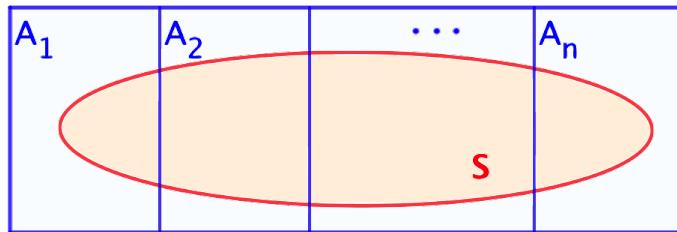
Hay que tener en cuenta que la suma de probabilidades de las ramas de cada nudo ha de dar siempre 1.

Las probabilidades de los sucesos finales (intersección de las ramas que conducen a ellos) se obtiene multiplicando las probabilidades de las ramas.

Definición 3.13:

Un **Sistema Completo de Sucesos** es una serie de sucesos A_1, A_2, \dots, A_n que cumplen:

$$A_1 \cup A_2 \cup \dots \cup A_n = \bigcup_{i=1}^n A_i = E; \quad A_i \cap A_j = \emptyset, \forall i \neq j; \quad p(A_i) \neq 0, \forall i$$



Teorema 3.9:

Teorema de la probabilidad total

Si S es un suceso cualquiera y $A_i, 1 \leq i \leq n$ es un sistema completo de sucesos, se cumple:

$$p(S) = p(S|A_1) \cdot p(A_1) + p(S|A_2) \cdot p(A_2) + \dots + p(S|A_n) \cdot p(A_n)$$

Demostración.

$A_1 \cup A_2 \cup \dots \cup A_n = E \rightarrow S = S \cap E = S \cap A_1 \cup S \cap A_2 \cup \dots \cup S \cap A_n$, todos disjuntos (al serlo los A_i), por lo que, por el Axioma 3 de Kolmogorov, $p(S) = \sum_{i=1}^n p(S \cap A_i) = p(A_1 \cap S) + p(A_2 \cap S) + \dots + p(A_n \cap S) = p(S|A_1) \cdot p(A_1) + p(S|A_2) \cdot p(A_2) + \dots + p(S|A_n) \cdot p(A_n)$. La intersección de sucesos es comutativa: $A_i \cap S = S \cap A_i, \forall i$

□

Este teorema proporciona una forma de obtener la probabilidad (total) de un suceso de la segunda (o última) etapa de la experiencia compuesta sin condicionar a ninguno de la primera. En un diagrama de árbol, basta con sumar las probabilidades de las ramas finales en que se verifica el suceso en cuestión. (Ver ejemplo siguiente).

Teorema 3.10:

Teorema de Bayes

Si S es un suceso cualquiera y $A_i, 1 \leq i \leq n$ es un sistema completo de sucesos, se cumple, que para cada suceso A_i ,

$$p(A_i|S) = \frac{p(A_i) \cdot p(S|A_i)}{p(S|A_1) \cdot p(A_1) + p(S|A_2) \cdot p(A_2) + \dots + p(S|A_n) \cdot p(A_n)}$$

Demostración.

$p(A_i|S) = \frac{p(A_i \cap S)}{p(S)} = \frac{p(S) \cdot p(A_i)}{p(S)}$, solo queda aplicar el teorema de la probabilidad total en el denominador.

La intersección de sucesos es comutativa: $A_i \cap S = S \cap A_i, \forall i$

□

Este teorema proporciona el cálculo de probabilidades ‘a posteriori’, es decir, conocer la probabilidad de un suceso de la primera etapa condicionado (sabiendo) que ha ocurrido determinado suceso de la segunda etapa. En un diagrama de árbol, basta con observar todos los sucesos que verifican el suceso de la segunda etapa (posibles) y aquel o aquellos de éstos que verifican el suceso de la primera etapa (favorables) y, ahora, aplicar la regla de Laplace. (Ver ejemplo siguiente).

Ejemplo 3.7:

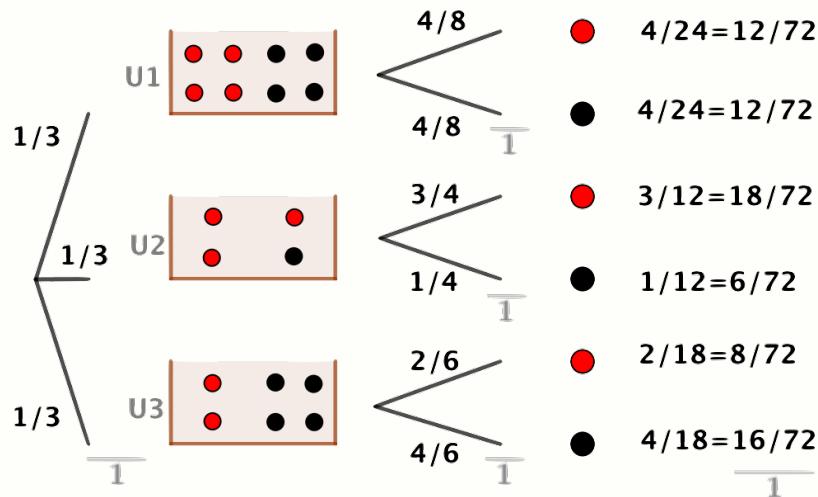
Tenemos tres urnas. La primera contiene 4 bolas rojas y 4 negras; la segunda, 3 rojas y 1 negra, y la tercera, 2 rojas y 4 negras. Elegimos una urna al azar y sacamos una bola de ella.

Calcula la probabilidad de que la bola extraída haya sido negra.

Sabiendo que la bola extraída es negra, ¿cuál es la probabilidad de que sea de la urna 3.

Tenemos una experiencia compuesta. En la primera parte elegimos, al azar, una urna (todas tienen la misma probabilidad de ser elegidas, cada una de ellas $p(U_i) = 1/3$). La

segunda parte consiste en extraer una bola de la urna, que puede ser roja o negra con probabilidades $p(R)$ y $p(N)$ distintas según la urna de la que se trate.



Las probabilidades de los sucesos compuestos se obtienen multiplicando las ramas que conducen a ellos: $p(\text{urna 1 y bola negra}) = p(U1 \cap N) = \frac{1}{3} \cdot \frac{4}{8} = \frac{4}{24}$. Una de cada tres veces vamos a la urna uno; una vez allí, cuatro de cada ocho veces sacamos bola negra, luego $\frac{1}{3} \cdot \frac{4}{8} = \frac{4}{24}$, es decir, cuatro de cada veinticuatro veces habremos ido a la urna 1 y sacado bola roja. Obtenemos las *probabilidades de los sucesos compuestos multiplicando las probabilidades de las ramas de donde provienen*.

Presentamos las probabilidades finales (de los sucesos compuestos) reducidas a común denominador para mayor claridad del problema.

Probabilidad de que la bola extraída haya sido negra: Si nos fijamos en el árbol, la bola extraída es negra en 12 de cada 72 veces (U1), también en 6 de cada 72 veces (U2) y en 16 de cada 72 veces, en total, en $12+6+16=34$ de cada 72 veces. Esto nos permite concluir que $p(N) = 34/72$

Hemos calculado la probabilidad de un suceso de la segunda etapa (la bola es negra) sin condicionar a ninguno de la primera etapa (U1, U2, U3): tenemos un **Teorema de la Probabilidad Total**, hemos, simplemente, *sumado todas las probabilidades que conducen a ese suceso*.

Usando fórmulas, $P(N) = p(U1) \cdot p(N|U1) + p(U2) \cdot p(N|U2) + p(U3) \cdot p(N|U3) = \frac{1}{3} \cdot \frac{4}{8} + \frac{1}{3} \cdot \frac{1}{4} + \frac{1}{3} \cdot \frac{4}{6} = \frac{34}{72}$

Probabilidad de haber ido a la urna 3 sabiendo que la bola extraída es negra. Los casos posibles de extraer bola negra son $12/72$, $6/72$ y $16/72$, de ellas solo en $16/72$ ocasiones hemos ido a la urna 3 (casos favorables), sin más que usar la regla de Laplace, concluimos que $p(U3|N) = \frac{16/72}{12/72 + 6/72 + 16/72} = \frac{16/72}{34/72} = \frac{16}{34} \approx 47\% > 33\%$ como decía la intuición, en la urna tres es más probable sacar bola negra que en las otras.

Hemos calculado una probabilidad de la primera etapa (U_3), condicionado a un suceso de la segunda etapa (N), esto es un **Teorema de Bayes**, tan solo hemos aplicado la regla de Laplace: *favorables/posibles*.

$$\text{Usando fórmulas, } P(U_3|N) = \frac{p(U_3) \cdot p(N|U_3)}{p(N)} = \frac{1/3 \cdot 4/3}{34/72} \approx 47\%$$

Como se puede comprobar, se pueden resolver problemas de ambos teoremas tan solo con un buen árbol y sabiendo que es lo que nos preguntan, huyendo de las fórmulas. Intentaremos resolver así la mayoría de los problemas.

Nótese que, como hemos comprobado en el árbol, todas las ramas que parten de un nodo, sus probabilidades suman UNO. Así como las probabilidades de las hojas finales.

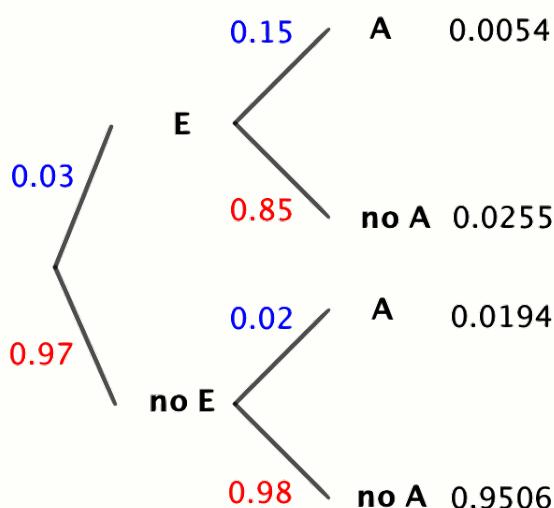
Presentamos algunos ejercicios resueltos más, usando todos los métodos: árbol, Venn y tablas contingencia. Unos son más sencillos y directos que otros, que requieren cálculos adicionales.

Ejemplo 3.8:

Ejercicio resuelto 3.8. La probabilidad de que una determinada enfermedad aparezca en un individuo es del 3 %. De entre los enfermos, el 15 % tienen antecedentes familiares y el 2 % de los que no están enfermos también tienen antecedentes familiares. Calcular:

- a) probabilidad de que una persona de esa ciudad no tenga antecedentes familiares.
- b) probabilidad de que una persona que no tenga antecedentes familiares (sabiendo que la persona no tiene antecedentes familiares), no padezca la enfermedad.
- c) tener la enfermedad y tener antecedentes familiares, ¿son sucesos incompatibles?, ¿y e independientes?

⇒ Árbol.



Llamamos E al suceso ‘la persona está enferma’, $\text{no } E$ a su suceso contrario, A a ‘la persona tiene antecedentes’ y $\text{no } A$ a su contrario.

Una persona cualquiera puede estar o no E y, después, puede que tenga o no A .

Hemos colocado en las ramas los valores conocidos, en azul. En rojo aparecen las probabilidades de los contrarios, sin más que recordar que en un árbol “las probabilidades de todas las ramas debe sumar uno”.

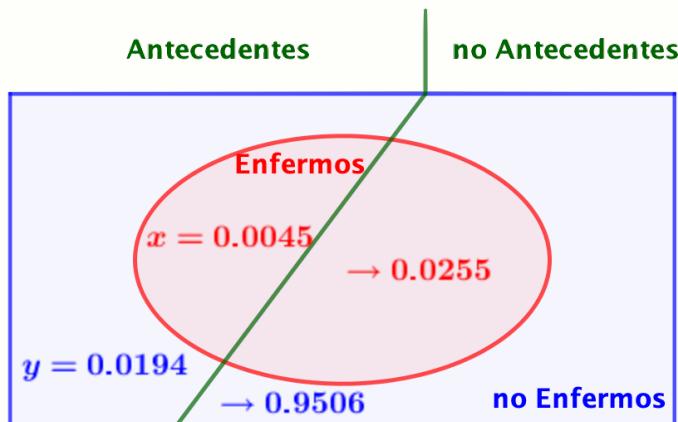
$$a) p(\text{no } A) = 0.0255 + 0.9506 = 0.9761$$

$$b) p(\text{no } E|\text{no } A) = \frac{f}{p} = \frac{0.9506}{0.0255 + 0.9506} = 0.9739$$

d) $p(\text{no } E|\text{no } A) = 0.9739 \neq 0.97 = p(\text{no } E) \rightarrow \text{no } E \wedge \text{no } A$ son dependientes, luego E y A también dependientes.

$p(\text{no } E \wedge \text{no } A) = 0.9506 \neq 0 \rightarrow \text{no } E \wedge \text{no } A$ son compatibles, y también E y A son compatibles. ($p(E \cap A) = 0.054 \neq 0$).

⇒ *Diagrama de Venn.*



El problema no es directamente abordable por un diagrama de Venn, a no ser que lo preparemos antes.

Puestos que enfermos hay el $3\% = 0.03$, dentro de la zona roja, E , haremos que el total sea 0,03.

Llamando $x = p(E \cap A) = p(E) \cdot p(A|E) = 0.03 \cdot 0.14 = 0.0045$. Así, en la zona roja que falta pondremos $0.03 - 0.0045 = 0.0255 = p(E \cap \text{no } A)$.

En la zona azul están los $\text{no } E$, en total han de ser 0.97 (100% - 3%). Llamando $y = p(\text{no } E \cap A) = p(\text{no } E) \cdot p(A|\text{no } E) = 0.97 \cdot 0.02 = 0.0194$. A la derecha pondremos $0.97 - 0.0194 = 0.9506 = p(\text{no } E \cap \text{no } A)$.

Ahora, ya podemos resolver el problema:

$$a) p(\text{no } A) = 0.0255 + 0.9506 = 0.9761$$

$$b) p(\text{no } E|\text{no } A) = \frac{0.9506}{0.0255 + 0.9506} = 0.9739$$

c) de modo análogo a lo resuelto en el árbol, los sucesos E y A son dependientes y compatibles.

⇒ *Tabla de contingencia.*

	Ant	no Ant	
E	$x=0.0045$	0.0255	0.03
no E	$y=0.0149$	0.9506	0.97
	0.0194	0.9769	

También necesitamos cálculos previos para abordar el problema mediante una tabla de contingencia.

$$p(A|E) = 0.15 = \frac{p(A \cap E)}{p(E)} = \frac{x}{0.03} \rightarrow x = 0.045$$

$$p(A|no\ E) = 0.02 = \frac{p(A \cap no\ E)}{p(no\ E)} = \frac{y}{0.97} \rightarrow y = 0.0194$$

Calculadas x e y , completamos las celdas adyacentes viendo las sumas parciales (en negrita en la tabla). Completamos la fila de abajo sin mas que sumar.

Con la tabla acabada, ya podemos resolver el problema:

a) $p(no\ A) = 0.9761$

b) $p(no\ E|no\ A) = \frac{0.9506}{0.9761} = 0.9739$

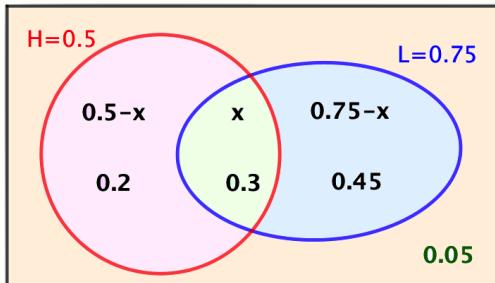
c) de modo análogo a lo resuelto en el árbol, los sucesos E y A son dependientes y compatibles.

Ejemplo 3.9:

Ejercicio resuelto 3.9. Una población está formada, a partes iguales, por hombres (H) y mujeres (M). La probabilidad de que una persona de esa población no lea ningún periódico ($no\ L$) es 0.25. Además, el porcentaje de individuos o bien lee algún periódico (L) o bien es hombre es del 95 %. Se elige una persona al azar, calcula:

- a) $p(H \cap L)$; b) $p(L|H)$; c) ¿ H y L son dependientes?, ¿compatibles?

⇒ **Diagrama de Venn.** (Llamamos $x = p(H \cap L)$)



$$p(L') = 0.25 \rightarrow p(L) = 0.75$$

$$p(H \cup L) = 0.95 \rightarrow p(H \cup L)' = 0.05$$

$$0.5 - x + x + 0.75 - x + 0.05 = 1$$

$$\Rightarrow x = 0.3$$

$$p(H \cap L) = 0.3 \neq 0 \rightarrow \text{Compatibles};$$

$$p(L|H) = \frac{0.3}{0.5} = 0.6 \neq 0.75 = p(L) \rightarrow \text{Dependientes}$$

⇒ **Tabla de contingencia.**

	H	M	
L	x	0.75-x	0.75
no L	0.3	0.45	
	0.5	0.5	1

Llamamos $x = p(H \cap L)$

$$p(H \cup L) = x + 0.75 - x + 0.5 - x = 0.95$$

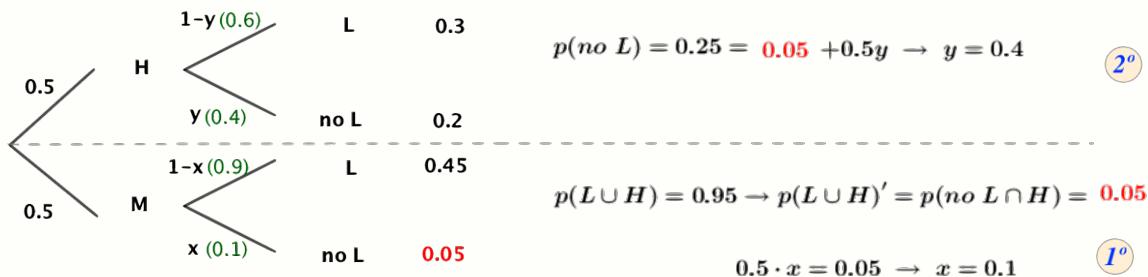
$$\rightarrow x = 0.3$$

Completamos la tabla.

$$p(H \cap L) = 0.3 \neq 0 \rightarrow \text{compatibles}.$$

$$p(L|H) = \frac{0.3}{0.5} = 0.6 \neq 0.75 = p(L) \rightarrow \text{dependientes.}$$

⇒ **Árbol.** El problema no directamente abordable desde un árbol, por lo que tendremos que hacer unos cálculos previos.



$$p(H \cap L) = 0.3 \neq 0 \rightarrow \text{compatibles.}$$

$$p(L|H) = 1 - y = 0.6 \neq 0.75 = 0.3 + 0.45 = p(L) \rightarrow \text{dependientes.}$$

⇒ **Teóricamente (fórmulas).**

$$p(H) = p(M) = 0.5; \quad P(\text{no } L) = 0.25 \rightarrow p(L) = 0.75$$

$$p(H \cup L) = p(H) + p(L) - p(H \cap L) \rightarrow 0.95 = 0.5 + 0.75 - p(H \cap L) \rightarrow$$

$$\rightarrow p(H \cap L) = 0.3 \neq 0 \rightarrow \text{compatibles.}$$

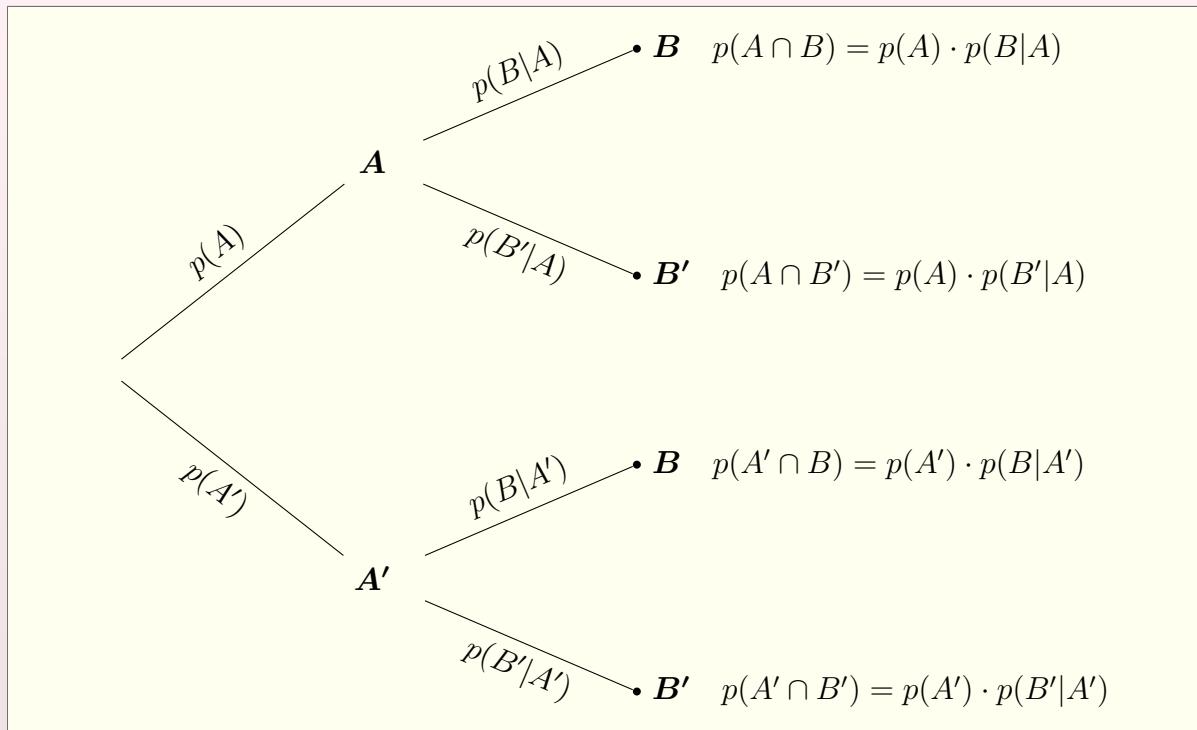
$$p(L|H) = \frac{p(H \cap L)}{p(H)} = \frac{0.3}{0.5} = 0.6 \neq 0.75 = p(L) \rightarrow \text{dependientes.}$$

Con estos ejercicios resueltos queremos poner de manifiesto que de las cuatro estrategias para la resolución de problemas de probabilidad (árbol, Venn, contingencia y fórmulas), en cada problema habrá una más adecuada que otra aunque, con los cálculos necesarios, todas pueden ser útiles. Es la experiencia, tras la realización de muchos ejercicios, la que dirá que método es mejor y más rápido usar. De todos modos, se recomienda que los ejercicios del siguiente apartado se intenten hacer usando distintas estrategias.

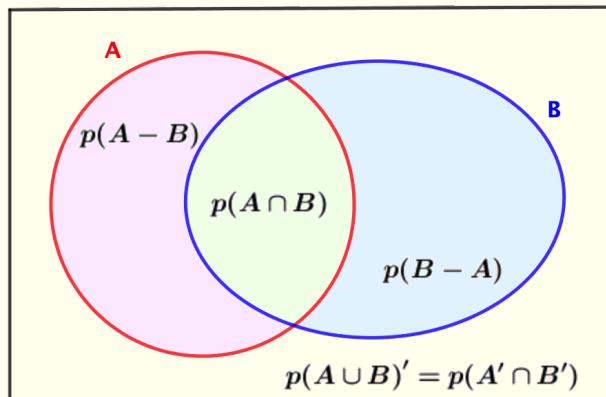
En el siguiente cuadro esquematizamos estas cuatro estrategias.

Estrategias para la resolución de los problemas de probabilidad

⇒ **Árbol.**



⇒ Diagrama de Venn.



⇒ Tabla de contingencia.

	suceso B	suceso B'	totales
suceso A	$p(A \cap B)$	$p(A \cap B')$	Σ
suceso A'	$p(A' \cap B)$	$p(A' \cap B')$	Σ
totales	Σ	Σ	1

⇒ Teóricamente (fórmulas).

Axiomática de Kolmogorov y definición y teoremas de probabilidad condicionada, probabilidad total y teorema de Bayes.

3.6. Ejercicios

Se deberían intentar resolver los ejercicios antes de ver su resolución.

Ejercicio 3.1.

Una urna 1 tiene tres bolas rojas y dos negras, una urna 2 tiene dos rojas y tres negras.

A. Nos dirigimos a una urna al azar y sacamos, también al azar, una bola, miramos su color, la devolvemos a la urna y, al azar, sacamos otra bola para observar su color. (Extracciones con reinserción).

B. Nos dirigimos a una urna al azar y sacamos dos bolas para observar sus colores. (Extracciones sin reinserción).

C. Nos dirigimos a la urna 1 y sacamos una bola, luego nos dirigimos a la urna 2 y sacamos una segunda bola, observando los colores de las bolas extraídas.

D. Nos dirigimos a la urna 1, sacamos una bola para observar su color y la introducimos en la urna 2. Luego, al azar, extraemos una bola de la urna 2 para observar su color.

Calcúlese, en cada caso:

a) Probabilidad de que las dos bolas sean negras.

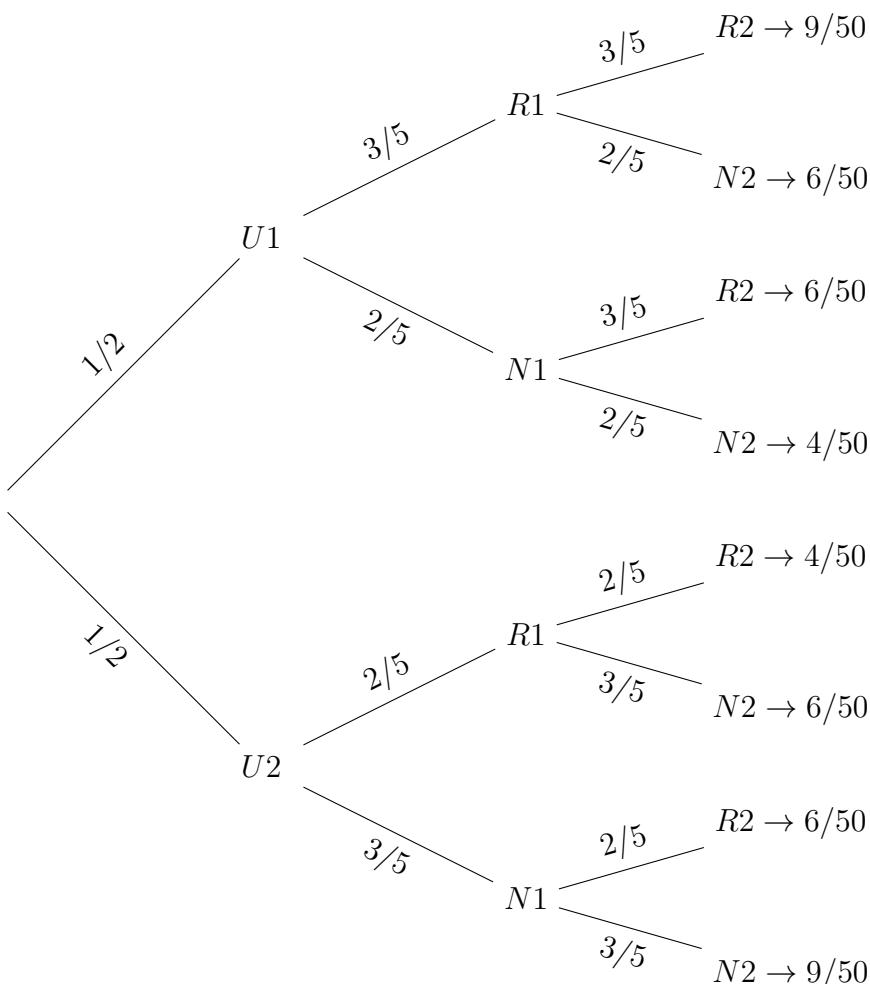
b) Probabilidad de que las bolas extraídas sean de distinto color.

c) Sabiendo que la bola extraída en segundo lugar es negra, probabilidad de que también lo haya sido la primera bola extraída.

Llamamos U_1 , U_2 a las urnas; R_1 , N_1 a los sucesos ‘la primera bola extraída es roja/negra’ y R_2 , N_2 a los sucesos ‘la segunda bola extraída es roja/negra’.

Presentamos un diagrama de árbol para cada caso. Se debería estar seguro de la disposición de las ramas y de las probabilidades de cada una de ellas. También se debe comprobar que todas las ramas (incluidas las finales –hojas–) suman uno.

A. Elegimos urna al azar y sacamos dos bolas con reinserción.

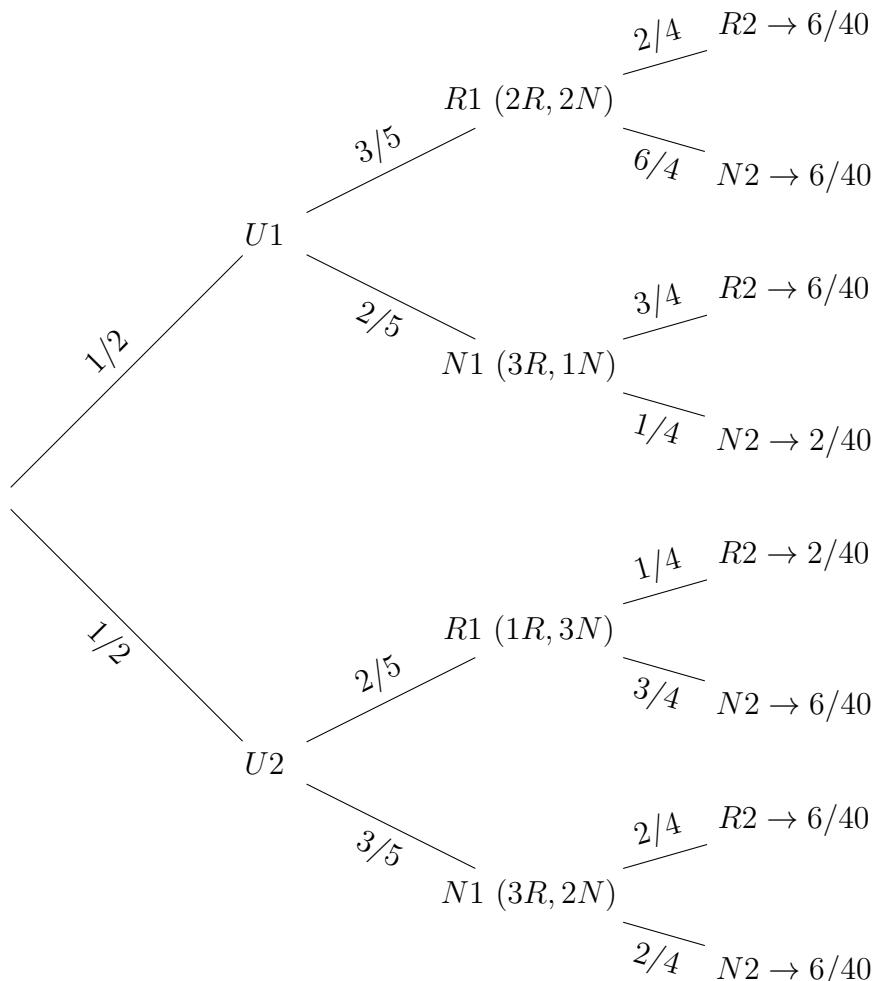


$$\Rightarrow a) \quad p(N1N2) = 4/50 + 9/50 = 13/50 = 26\%$$

$$\Rightarrow b) \quad p(\neq \text{colores}) = p(N1R2) + p(R1N2) = 6/50 + 6/50 + 6/50 + 6/50 = 24/50 = 48\%$$

$$\Rightarrow c) \quad p(N1|N2) = \frac{4/50 + 9/50}{6/50 + 4/50 + 6/50 + 9/50} = 13/25 = 52\%$$

B. Elegimos urna al azar y sacamos dos bolas sin reinserción.

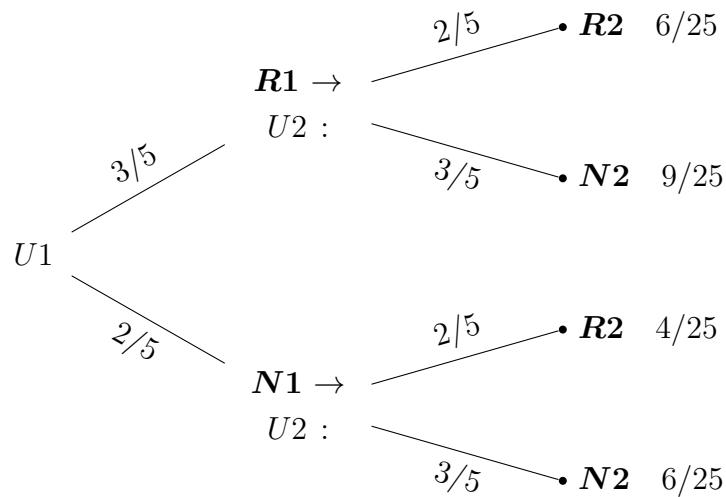


$$\Rightarrow a) \quad p(N1N2) = 2/40 + 6/40 = 9/40 = 20\%$$

$$\Rightarrow b) \quad p(\neq \text{colores}) = p(N1R2) + p(R1N2) = 6/40 + 6/40 + 6/40 + 6/40 = 24/40 = 60\%$$

$$\Rightarrow c) \quad p(N1|N2) = \frac{2/40 + 6/40}{6/40 + 2/40 + 6/40 + 6/40} = 8/20 = 40\%$$

C. Sacamos una bola de cada urna (primero de la U1 y luego de la U2).

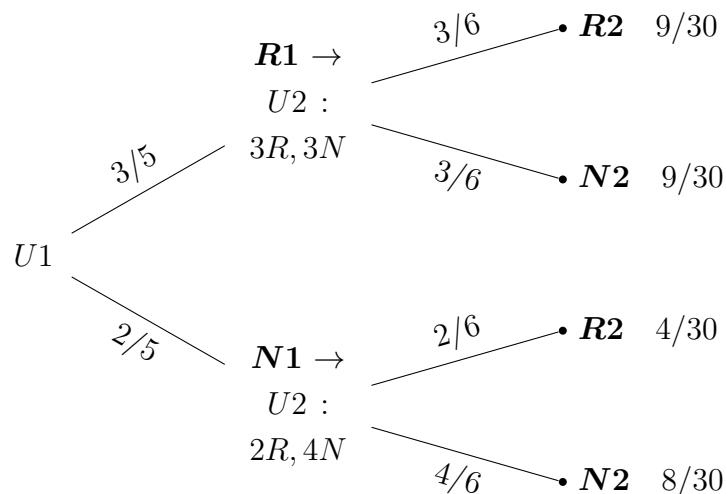


$$\Rightarrow a) \quad p(N1N2) = 6/25 = 24\%$$

$$\Rightarrow b) \quad p(\neq \text{colores}) = p(N1R2) + p(R1N2) = 9/25 + 4/25 = 52\%$$

$$\Rightarrow c) \quad p(N1|N2) = \frac{6/25}{9/25 + 6/25} = 6/15 = 40\%$$

D. Trasvasamos una bola (R1 o N1) de la urna U1 a la urna U2 y extraemos bola (R2 o N2).



$$\Rightarrow a) \quad p(N1N2) = 8/30 = 26.7\%$$

$$\Rightarrow b) \quad p(\neq \text{colores}) = p(N1R2) + p(R1N2) = 4/30 + 9/30 = 13/30 = 43.3\%$$

$$\Rightarrow c) \quad p(N1|N2) = \frac{8/30}{9/30 + 8/30} = 8/17 = 47.1\%$$

Ejercicio 3.2. De una baraja española de 40 cartas se extraen dos de ellas (sin reemplazo). Calcula la probabilidad de que:

- a) las dos sean reyes.
- b) una sea de copas y otra sea rey.
- c) al menos una de ellas sea de copas.

Este problemas lo resolveremos razonando.

$$a) p(\text{dos reyes}) = p(1^{\text{a}} \text{ rey y } 2^{\text{a}} \text{ rey}) = p(R1 \cap R2) = p(R1) \cdot p(R2|R1) = \frac{4}{40} \cdot \frac{3}{39} = 0.77\%$$

$$b) p(\text{una de copas y otra rey}) = p(1^{\text{a}} \text{ copas y } 2^{\text{a}} \text{ rey}) + p(1^{\text{a}} \text{ rey y } 2^{\text{a}} \text{ copas})$$

$$\text{Si hacemos } p(Co1 \cap R2) + p(R1 \cap Co2) = p(Co1) \cdot p(R2|Co1) + p(R1) \cdot p(Co2|R1)$$

En $p(Co1 \cap R2)$ se nos presenta el siguiente problema: que la segunda rey si la primera es copas, $p(R2|Co1)$, es algo delicado, pues la primera podría haber sido ser el rey de copas, con lo que quedarían $3/39$ y no $4/39$ reyes en la baraja.

Cuando saquemos la carta de copas vamos a distinguir si esa carta es de copas excluido el rey de copas o si es el rey de copas: $p(Co) = p(Co - RC) + p(RC)$, así,

$$p(Co1 \cap R2) = p(Co - RC \cap R2) + p(RC \cap R2) = \frac{9}{40} \cdot \frac{4}{39} + \frac{1}{40} \cdot \frac{3}{39}$$

Para el suceso $p(R1 \cap Co2)$ estamos en las mismas, al sacar $R1$ puede que sea el rey de copas u otro, faremos $R1 = RnoCo1 \cup RC$, distinguiremos así entre sacar un rey que no sea de copas (quedan $10/39$ copas en la baraja) o sacar el rey de copas (quedan $9/39$ copas en la baraja).

$$p(R1 \cap Co2) = p(RnoCo1 \cap Co2) + p(RCo1 \cap Co2) = \frac{3}{40} \cdot \frac{10}{39} + \frac{1}{40} \cdot \frac{9}{39} = \\ = \frac{(9 \cdot 4 + 1 \cdot 3) + (3 \cdot 10 + 1 \cdot 9)}{40 \cdot 39} = \frac{78}{1560} = 5\%$$

c) al menos una carta será de copas cuando lo sea la primera, o la segunda, o las dos. Es más sencillo pensar en el suceso contrario: '(al menos una es de copas)' '=' 'ninguna es de copas'.

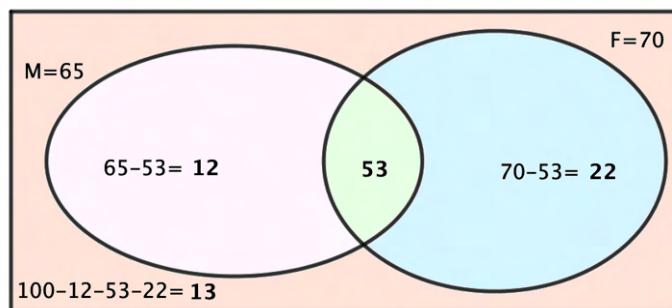
$$p(\text{ninguna Co}) = p(\text{noCo1} \cap \text{noCo2}) = p(\text{noCo1}) \cdot p(\text{noCo2}|\text{noCo1}) = \frac{30}{40} \cdot \frac{29}{39} = \frac{870}{1560}$$

$$\text{Por lo que } p(\text{al menos una es de Copas}) = 1 - \frac{870}{1560} = \frac{690}{1560} = 44.2\%$$

Ejercicio 3.3. Un 65 % de los alumnos de un centro han aprobado Matemáticas, un 70 % ha aprobado Filosofía, y un 53 % ha aprobado ambas materias. Si se elige al azar un estudiante, calcúlese la probabilidad de que:

- a) haya aprobado al menos una de las dos materias.
- b) haya suspendido ambas materias
- c) Si aprobó Matemáticas ¿Cuál es la probabilidad de haber aprobado Filosofía?

Resolvemos este ejercicio mediante un diagrama de Venn. Trabajamos en tantos por cien.



$$a) p(\text{aprobar alguna}) = p(M \cup F) = (12 + 53 + 22) \% = 87 \%$$

$$b) p(\text{suspender ambas}) = p(M \cup F)' = p(M' \cap F') = 1 - 87 \% = 13 \%$$

$$c) p(F|M) = 53/65 \approx 81.5 \%$$

Podría haberse intentado resolver el problema mediante una tabla de contingencia

	F	F'	
M	53		65
M'			
	70		100

Ejercicio 3.4.

Suponiendo que la riqueza es independiente del sexo, completar la siguiente tabla y calcular:

a) probabilidad de que sabiendo que una persona no es pobre, sea hombre.

	Rico	Pobre	Total
Hombre			0.607
Mujer			0.393
0.002			

Tenemos una tabla de contingencia, las suma total debe ser **1**.

Empezamos por llamar $x = p(\text{Rico} \cap \text{Hombre})$, al ser la riqueza independiente del sexo, tendremos que $p(\text{Rico}|\text{Hombre}) = x/0.607$ ha de coincidir con $p(\text{Rico}) = 0.002$. De ahí: $x/0.607 = 0.002 \rightarrow x = 0.0012$.

A partir de ahí, sin más que restar por filas, obtenemos $p(\text{Pobre} \cap \text{Hombre}) = 0.607 - 0 - 0.0012 = 0.6058$

Obtenemos la $p(\text{Pobre})$ sin más que restar en la columna final: $1 - 0.002 = 0.998$

Las dos probabilidades que faltan, se pueden obtener sin más que restar columnas. Con todo ellos, tenemos:

	Rico	Pobre	Total
Hombre	x=0.0012	0.6058	0.607
Mujer	0.0008	0.3922	0.393
	0.002	0.998	1

a) $p(Hombre|no Pobre) = p(Hombre|Rico) = 0.0012/0.002 = 60\%$

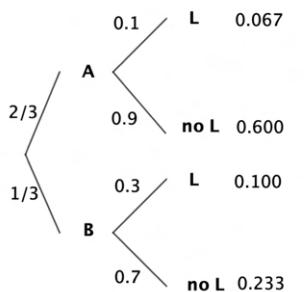
b) $p(Rico o Hombre) = p(Rico \cup Hombre) = 0.0008 + 0.0012 + 0.6058 = 0.6078$

Ejercicio 3.5. La ciudad A tiene el doble de habitantes que la ciudad B, pero un 30% de ciudadanos de B lee literatura, mientras que sólo un 10% de ciudadanos de A lee literatura.

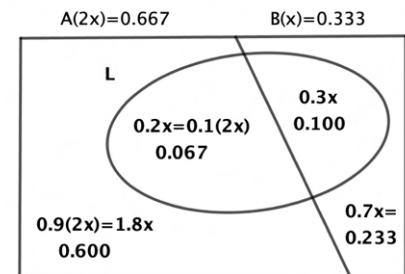
a) De un ciudadano se sabe sólo que vive en la ciudad A o en la ciudad B. Calcula de forma razonada que lea literatura.

b) Si nos presentan un ciudadano que vive en la ciudad A o en la ciudad B, pero del cual sabemos que lee literatura, calcula razonadamente la probabilidad de que sea de la ciudad B.

El problema es lo bastante sencillo como para resolverlo con cualquiera de las tres estrategias (árbol, Ven o tabla), tan solo hay que considerar que si en A hay doble población que B, sea x =población de B, luego $2x$ =población de A y $3x$ =población total. Con esto, $p(A)=2/3$ y $p(B)=1/3$.



	L	no L	
A	0.1 2/3	→ 0.9 2/3	2/3
B	0.3 1/3	→ 0.7 1/3	1/3
Σ_L		$\Sigma_{no L}$	1



a) $p(L) = 0.067 + 0.100 = 0.167$

b) $p(A|L) = \frac{0.067}{0.067 + 0.100} = 0.401$

Ejercicio 3.6. Escribo tres cartas y los tres sobres correspondientes. Introduzco cada carta en un sobre al azar, es decir sin mirar el destinatario. Averiguar razonadamente cuál es la probabilidad de que haya introducido solo una carta en el sobre correcto.

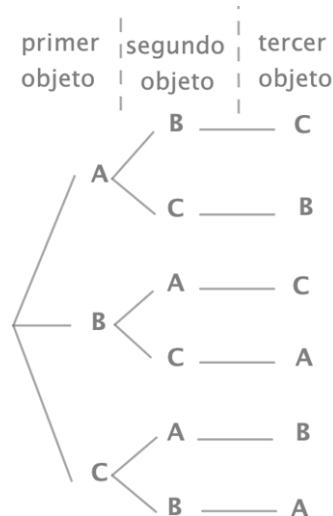
Resolveremos el problema por razonamiento. Si llamamos A,B,C a los sobres y a,b,c a las cartas, la situación ideal (acertarlos todos) sería Aa,Bb,Cc.

Supongamos que tenemos los sobres ordenados, A,B,C y vamos a repartir las cartas al azar. Las distintas formas de ordenar 3 objetos a,b,c es $3! = 6$ (ver apéndice B de Combinatoria, también es fácil de deducir mediante un esquema en árbol). Estos son los casos posibles.

Ahora veamos los casos favorables a acertar solo una carta con su sobre. Si se trata de la carta a, la solución sería Aa,Bc,Cb y no hay más. Solo hay, también, una sola forma de acertar la carta b (Ac, Bb, Ca) y una sola para la carta c (Ab,Ba,Cc). En total, 3 casos posibles.

Aplicando la regla de Laplace:

$$p(\text{acertar solo una}) = \frac{3}{6} = 50\%.$$



Formas de ordenar 3 objetos

Ejercicio 3.7. En un curso de 40 alumnos hay 24 que aprueban Historia, 20 aprueban Lengua y 27 aprueban Filosofía. Sabiendo que no hay ninguno que las suspenda todas, ¿cuál es la probabilidad de que al elegir a un alumno al azar, haya aprobado la Historia y la Lengua?. Sabemos también que las tres asignaturas las aprueban 10 alumnos, que 14 aprueban Filosofía e Historia y que 15 aprueban Lengua y Filosofía.

Si sabemos que un alumno ha aprobado la Historia, ¿cuál es la probabilidad de que apruebe Lengua?

Los sucesos ‘aprobar Lengua’ y ‘aprobar Historia’, ¿son incompatibles?, ¿son independientes?. Razona tus respuestas.

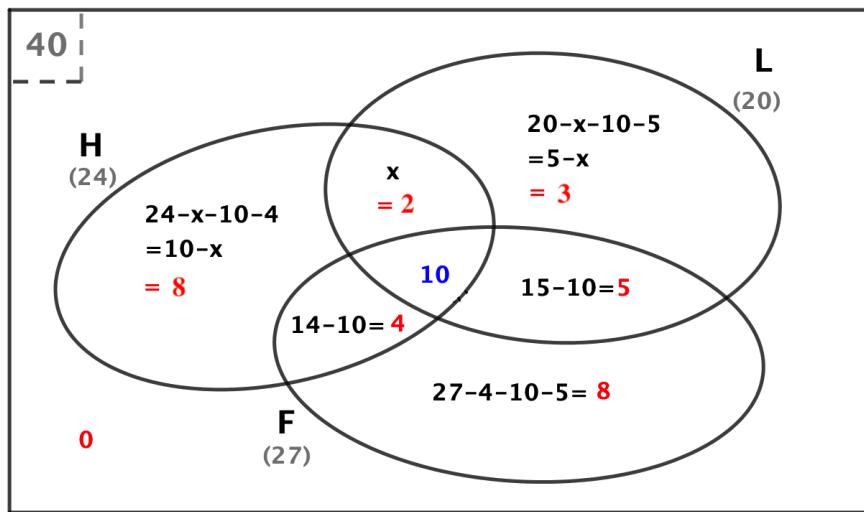
Usaremos un diagrama de Venn con tres conjuntos compatibles (que se intersecten entre ellos). Ojo a las intersecciones.

En este caso trabajaremos con números absolutos, dentro del rectángulo (referencial) han de estar los 40 alumnos. (Ver figura adjunta).

Empezamos colocando $10 = H \cap L \cap F$. Como en todo $F \cap H = 14$, fuera de la triple intersección colocaremos $14 - 10 = 4$. Lo mismo para $L \cap F$, esos 15 son los 10 de la triple intersección y 5 de fuera.

$L \cap H????$. Llamaremos x a los alumnos que han aprobado solo F y H, excluyendo a los 10 que también han aprobado F, así en $F \cap H$ harán $10 + x$.

Completamos ahora las zonas en que solo aparecen H, L o F. De este modo, en F hay $27 - 4 - 10 - 5 = 8$; en H, $6 - x$, y en L, $5 - x$. Como hay 0 fuera de los tres conjuntos, ya tenemos a todo el alumnado colocado. Sumándolos a todos hemos de obtener 40 $\rightarrow x = 2$



$$p(H \cup L) = (8 + 2 + 3 + 4 + 10 + 5)/40 = 32/40 = 80\%$$

$$p(L|H) = (10 + 2)/(24) = 50\%$$

$p(L|H) = 50\% = p(L) = 20/40 = 50\% \rightarrow H$ y L son independientes.

Ejercicio 3.8. En el palacio de rey Nebuzaradan hay 3 cámaras con dos cofres en cada una de ellas. En la primera cámara, cada cofre contiene un diamante, en la segunda hay una esmeralda dentro de cada cofre, y en la tercera, un cofre contiene un diamante y el otro una esmeralda. Un ladrón escoge una de las tres cámaras al azar y roba uno de los cofres. ¿Cuál es la probabilidad de que haya robado un diamante? Si cuando llega a su escondite descubre que ha robado una esmeralda, ¿cuál es la probabilidad de que dentro del segundo cofre de la misma cámara en que ha entrado hubiese un diamante?

$$p(D) = 1/6 + 1/6 + 1/6 = 0.5$$

$$p(C3|D) = \frac{1/6}{1/6 + 1/6 + 1/6} = 0.33$$



Ejercicio 3.9.

En 1912, durante su primer viaje a través del Atlántico, el transatlántico Titanic chocó contra un iceberg y se hundió. En la siguiente tabla tienes información sobre el número de mujeres que sobrevivieron y murieron en relación a su nivel económico:

Nivel económico	Murieron	Sobrevivieron
Alto	6	126
Medio	13	90
Bajo	107	101

¿Cuál es la probabilidad de que una mujer de nivel alto sobreviviera? ¿Y una de nivel medio? ¿Y una de nivel bajo?

¿Cuál es la probabilidad de que una mujer que sobrevivió a la catástrofe fuera de nivel alto? ¿Y de nivel medio? ¿Y de nivel bajo?

Primero, calculemos las sumas parciales:

Nivel económico	Murieron	Sobrevivieron	totales
Alto	6	126	132
Medio	13	90	103
Bajo	107	101	208
totales	126	317	443

San A, M, B los sucesos ‘la persona elegida es de nivel económico alto’, ‘medio’, ‘bajo’, y Mu, So los sucesos ‘la persona elegida murió’, ‘sobrevivió’.

$$p(So|A) = 126/132 = 95,5\%; \quad p(So|M) = 90/103 = 87.4\%; \quad p(So|B) = 101/208 = 48.6\%$$

$$p(A|So) = 126/317 = 39.7\%; \quad p(M|So) = 90/317 = 28.4%; \quad p(B|So) = 101/317 = 31.9\%$$

Ejercicio 3.10. En una bolsa de caramelos surtidos hay 10 caramelos de sabor naranja (N), 5 sabor a limón (L) y 3 con sabor a fresa (F). Todos tienen el mismo tamaño y hasta extraerlos de la bolsa no se sabe de qué sabor son. Se extraen tres caramelos al azar.

a) Calcular de forma razonada la probabilidad de extraer primero uno con sabor naranja, luego uno con sabor a fresa y, por último, uno con sabor a limón.

b) Calcular de forma razonada la probabilidad de que sean de tres sabores diferentes.

Aunque el problema parece estar pidiendo ser abordado por un diagrama de árbol, necesitaríamos $3 \times 3 \times 3 = 27$ hojas. Excesivo, así que lo intentamos resolver razonándolo. Cada caramelo, en cada una de las tres extracciones, lo denotaremos por N_i , L_i , F_i , donde con $i = 1, 2, 3$ indicamos el orden en la extracción.

$$a) p(N_1 \cap F_2 \cap L_3) = p(N_1) \cdot p(F_2 | N_1) \cdot p(L_3 | N_1 \cap F_2) = \frac{10}{18} \cdot \frac{3}{17} \cdot \frac{5}{16} = \frac{10 \cdot 3 \cdot 5}{18 \cdot 17 \cdot 16} = \frac{150}{4896} \approx 3\%$$

Sacar el primer caramelo de naranja tiene una probabilidad de $10/18$. Quedan en la bolsa 17 caramelos (9N, 5L, 3F). Sacar ahora uno de fresa tiene una probabilidad de $3/17$, quedando 16 caramelos en la

bolsa (9N, 5L, 2F), por lo que la probabilidad de sacar esta vez uno de limón es de $5/16$. En total, $(10 \cdot 3 \cdot 5)(18 \cdot 17 \cdot 16) \approx 3\%$.

b) $p(\text{distintos sabores})$. Las distintas formas de ordenar 3 objetos son $3! = 6$, todas ellas con las mismas probabilidades.

$$\rightarrow p(\text{distintos sabores}) = 6 \cdot p(\text{NFL}) = 6 \cdot \frac{150}{4896} \approx 18\%$$

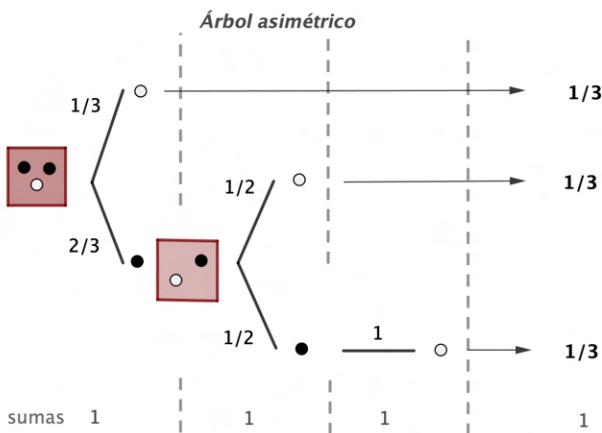
Hay que considerar los siguientes casos de ordenar los tres sabores:

- NFL, con probabilidad $(10 \cdot 3 \cdot 5)/(18 \cdot 17 \cdot 16)$
- NLF, con probabilidad $(10 \cdot 5 \cdot 3)/(18 \cdot 17 \cdot 16)$
- FNL, con probabilidad $(3 \cdot 10 \cdot 5)/(18 \cdot 17 \cdot 16)$
- LNF, con probabilidad $(5 \cdot 10 \cdot 3)/(18 \cdot 17 \cdot 16)$
- LFN, con probabilidad $(5 \cdot 3 \cdot 10)/(18 \cdot 17 \cdot 16)$

Todas las probabilidades son iguales (tanto numeradores como denominadores), luego:

$$\rightarrow p(\text{distintos sabores}) = 6 \cdot p(\text{NFL}) = 6 \cdot \frac{150}{4896} \approx 18\%$$

Ejercicio 3.11. Para elegir a un muchacho entre tres se prepara una bolsa con dos bolas negras y una bola blanca. Los tres van sacando, por orden, una bola que no devuelven. Quién saque la bola blanca gana. ¿Quién lleva más ventaja: el primero, el segundo o el tercero?



Cuando un jugador saca bola blanca, gana y acaba el juego (*árbol asimétrico*).

Primer jugador, $p(\text{Blanca}) = 1/3$

Segundo jugador, $p(\text{Blanca}) = 2/3 \cdot 1/2 = 1/3$

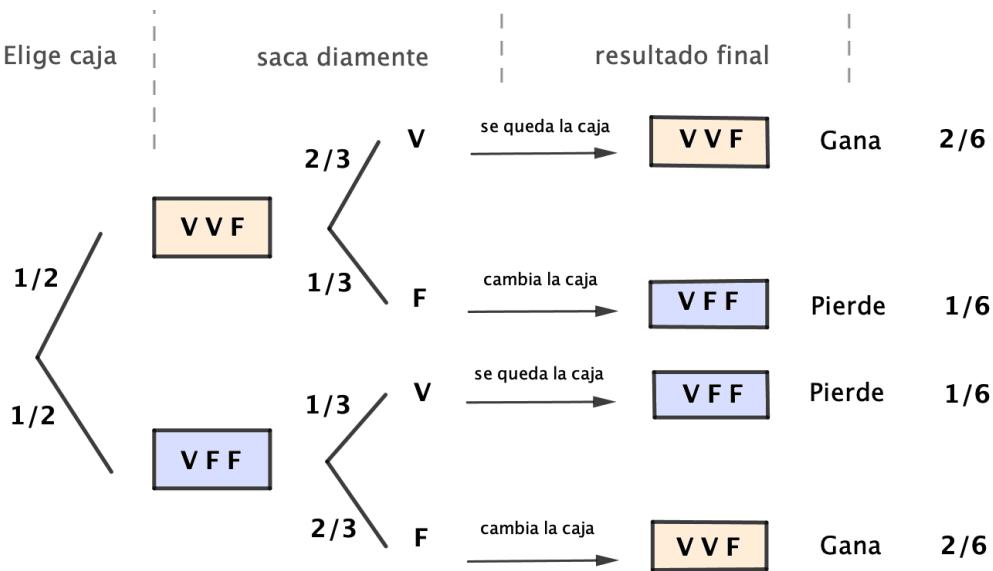
Tercer jugador, $p(\text{Blanca}) = 2/3 \cdot 1/2 \cdot 1 = 1/3$

Independientemente del turno de cada jugador, todos tienen la misma probabilidad de ganar.

Ejercicio 3.12. Mr. Stoneguy, un experimentado comerciante de diamantes, decide recomendar a su hijo permitiéndole escoger una de dos cajas. Cada caja contiene 3 diamantes. En una de las cajas 2 de los diamantes son reales y el otro es una excelente imitación; en la otra caja uno es real y los otros dos son imitaciones. Si el hijo escoge aleatoriamente entre las dos cajas, su oportunidad de conseguir 2 diamantes verdaderos es $1/2$, por lo que Mr. Stoneguy decide ayudarlo de la siguiente manera: permite a su hijo sacar un diamante de una de las cajas y examinarlo para ver si es un verdadero diamante; si el diamante

examinado es real, el hijo se queda con esa caja, y se queda con la otra caja en el otro caso. ¿Cuál es la probabilidad de que el hijo se quede con 2 verdaderos diamantes.

El hijo de Mr. Stoneguy, primero elige caja (con probabilidades $1/2$ para cada una de ellas) y, luego, saca un diamante. La probabilidad de que sea bueno es de $2/3$ en la caja uno y de $1/3$ en la dos. Si el diamante sacado es bueno, conserva la caja; si es malo, la cambia.



Decimos que el hijo ‘Gana’ si, al final, se queda con la caja que contiene dos diamantes buenos, así: $p(\text{Gana}) = 2/6 + 2/6 = 66.7\% > 50\%$, por lo que realmente Mr. Stoneguy sí ayuda a su hijo en la elección.

Ejercicio 3.13. En un congreso de 200 jóvenes se pasa una encuesta para conocer los hábitos en cuanto a contratar viajes por internet. Se observa que de los encuestados, 120 son hombres de los que 84 de ellos contratan los viajes por internet., mientras que solo 24 de las mujeres lo hace.

Elegido un congresista al azar, calcula la probabilidad de que:

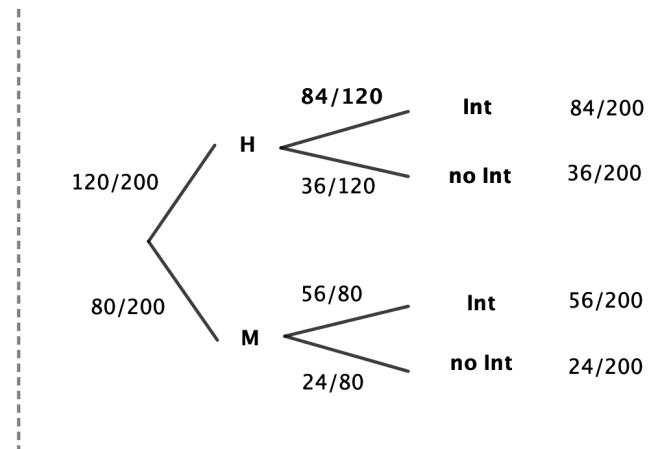
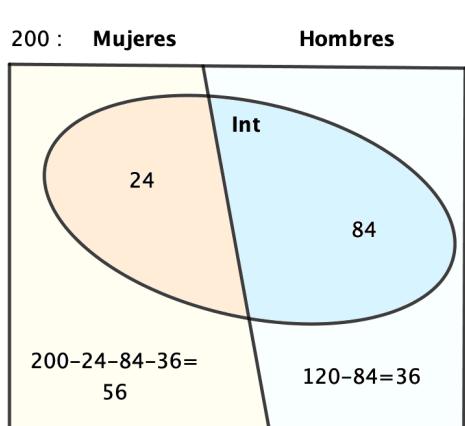
- No contrate sus viajes por internet.
- Sí use internet para contratar viajes , si la persona elegida es una mujer.
- Sea un hombre, sabiendo que contrata sus viajes por internet.

El problema parece dispuesto para ser atacado por una tabla de contingencia. Pondremos en negrita los datos conocidos e iremos averiguando los que faltan (se han numerado en el orden en que se han averiguado) :

- $p(\text{No internet}) = 92/200$; b) $p(\text{Internet} | \text{Mujer}) = 24/80$; c) $p(H | \text{Internet}) = 84/108$

	Usa Internet para los viajes	No usa internet para los viajes	
Hombre	84	(1) 120-84 = = 36	120
Mujer	24	(3) 80-24 = = 56	(2) 200-120 = = 80
	(4) 84+24 = = 108	(4bis) 36+56 = = 92	200

Podría, con unos pocos cálculos (sobre todo para el árbol), abordarse el problema mediante un diagrama de Venn o mediante un diagrama de árbol.



Ejercicio 3.14. Se dispone de tres monedas. La primera de ellas es una moneda ‘cargada’ de modo que la probabilidad de obtener cara es solo 0.4. La segunda moneda es una moneda ‘trucada’ que tiene dos cruces y la tercera también está ‘cargada’, pero de modo que la probabilidad de obtener cara es 0.6. Se pide:

- Escribir el espacio muestral asociado al experimento aleatorio del lanzamiento ordenado de estas tres monedas.
- Calcula la probabilidad de obtener exactamente dos cruces.
- Calcula la probabilidad del suceso $C_1X_2C_3$, es decir, cara en la primera moneda, crux en la segunda y cara en la tercera (CXC ordenadamente).
- Calcula la probabilidad de obtener al menos una cara.

Abordamos el problema sin estrategia de árbol, Ven o contingencia, directamente.

Usamos la notación CXC para indicar que la primera moneda ha salido cara, C; la segunda cruz, X y la tercera cara.

Para la primera moneda, $p(C)=0.4$ y $p(X)=0.6$. Para la segunda, $p(C)=0$ y $p(X)=1$. Para la tercera, $p(C)=0.6$ y $p(X)=0.4$.

a) $E = \{CXC, CX\bar{X}, \bar{X}XC, \bar{X}\bar{X}X\}$, la moneda central ‘trucada’ siempre sale \bar{X} .

b) Observando el espacio muestral anterior y teniendo en cuenta las probabilidades para cada moneda,

$$p(\text{dos cruces}) = p(C\bar{X}\bar{X}) + p(\bar{X}XC) = 0.4 \cdot 1 \cdot 0.4 + 0.6 \cdot 1 \cdot 0.6 = 0.16 + 0.36 = 0.52$$

$$c) p(CXC) = 0.4 \cdot 1 \cdot 0.6 = 0.24$$

d) El suceso contrario a ‘al menos una cara’ es ‘todas cruz’, así:

$$p(\text{al menos una cara}) = 1 - p(\text{todas cruz}) = 1 - p(\bar{X}\bar{X}\bar{X}) = 1 - 0.6 \cdot 1 \cdot 0.4 = 1 - 0.24 = 0.78$$

Ejercicio 3.15. Un dado está cargado de modo que la probabilidad de obtener las distintas caras es proporcional al cuadrado del número que estas tienen.

¿Cuál es la probabilidad de obtener un 5? ¿Cuál es la probabilidad de obtener suma 7 en dos lanzamientos?

$$p(1) = k; p(2) = k2^2 = 4k; p(3) = k3^2 = 9k; p(4) = k4^2 = 16k; p(5) = k5^2 = 25k; p(6) = k6^2 = 36k$$

$$\text{Como } \sum_{i=1}^6 p(i) = 1 \rightarrow k + 4k + 9k + 16k + 25k + 36k = 91k = 1 \rightarrow k = 1/91$$

La probabilidad de obtener un 5 es: $p(5) = 25k = 25/91$

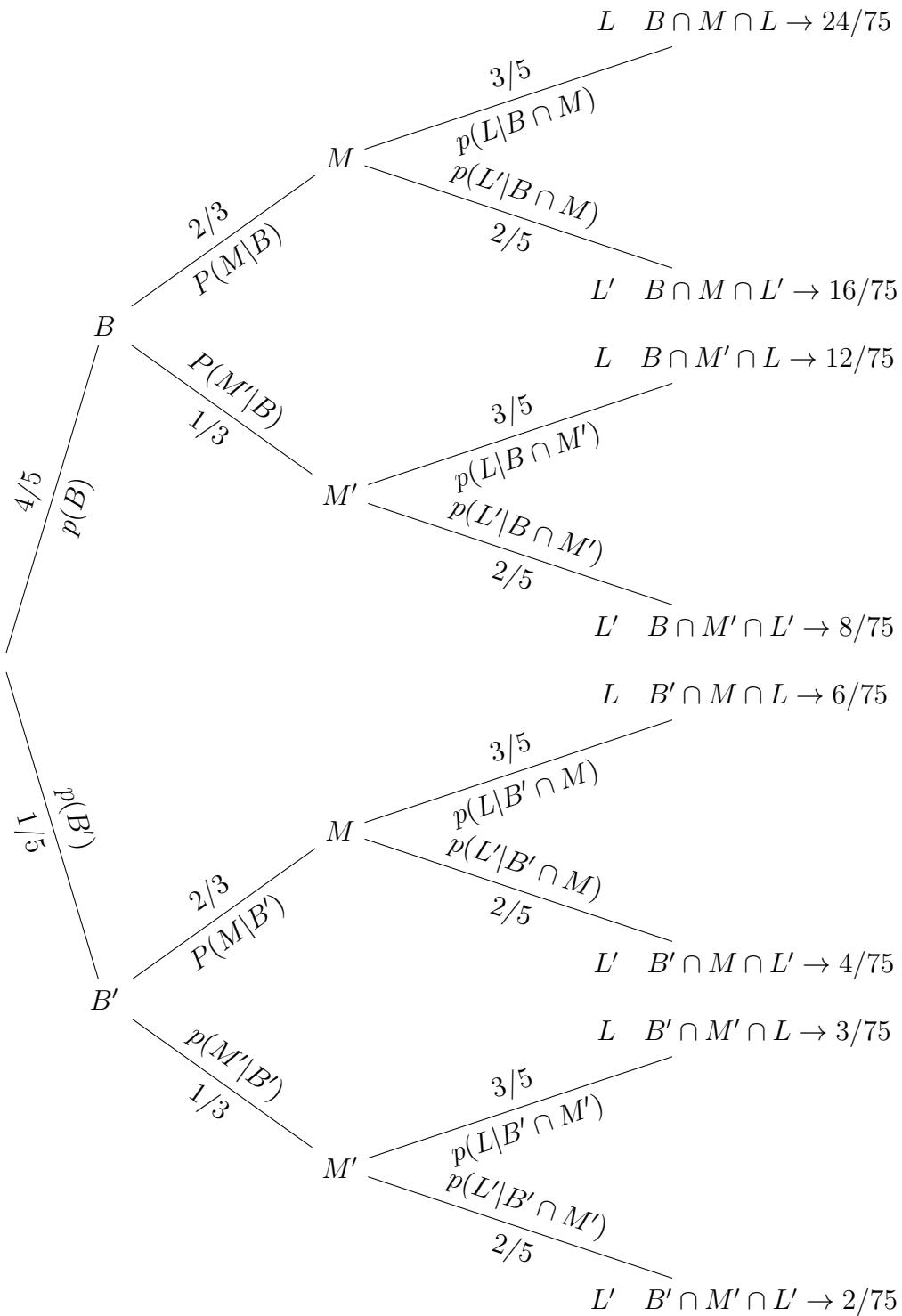
Para ver las probabilidades de las distintas sumas al lanzar dos dados podríamos construir una tabla de doble entrada, que va a ser largo, o usar el razonamiento.

Al lanzar dos dados se obtiene suma siete si salen (1,6), con la misma probabilidad que (6,1); o si sale (2,5) o (5,2); o si sale (3,4) o (4,3), así:

$$p(\text{suma } 7) = 2 \cdot p(1, 6) + 2 \cdot p(2, 5) + 2 \cdot p(3, 4) = 2 \cdot \frac{1}{91} \frac{36}{91} + 2 \cdot \frac{4}{91} \frac{25}{91} + 2 \cdot \frac{9}{91} \frac{16}{91} = \frac{560}{8281}$$

Ejercicio 3.16. En un centro de secundaria, aprueban Biología 4 de cada 5 alumnos, las Matemáticas las aprueban 2 de cada 3 alumnos y 3 de cada 5 alumnos aprueban Lengua. Elegido al azar un alumno matriculado de esas asignaturas en ese centro, calcula la probabilidad de que a) Suspenda esas tres asignaturas, b) Suspenda solo una de ellas y c) los sucesos ‘aprobar biología’ y ‘aprobar Matemáticas’, ¿son independientes?.

Llamaremos B, M, L a los sucesos aprobar Biología, Matemáticas y Lengua. llamaremos B', M' y L' a los sucesos suspender esas asignaturas. Resolvemos el problema mediante un diagrama de árbol. (Presentamos, ahora, toda la información en él).



$$a) p(\text{suspender las tres}) = p(B' \cap M' \cap L') = 2/75$$

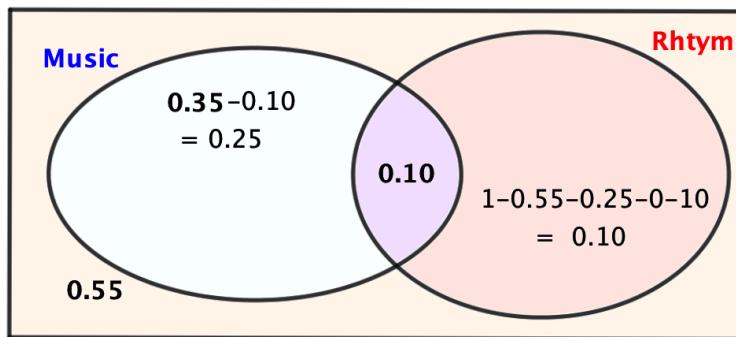
$$b) p(\text{suspender solo una}) = P(B' \cap M \cap L) + p(B \cap M' \cap L) + p(B \cap M \cap L') = \\ = 6/75 + 12/75 + 16/75 = 34/75$$

Ejercicio 3.17. Un estudio revela que el 10 % de los oyentes de radio sintoniza a diario las cadenas Music y Rhythm, que un 35 % sintoniza a diario Music y que el 55 % de los oyentes no escucha ninguna de las dos emisoras . Obtén:

- a) La probabilidad de que un oyente elegido al azar sintonice la cadena Rhythm.
- b) La probabilidad de que un oyente elegido al azar sintonice la cadena Rhythm pero no la Music.
- c) La probabilidad de que un oyente, del que sabemos que escucha Rhythm, escuche Music.

Usaremos, para la resolución, un diagrama de Venn.

Llamamos R y M a los sucesos ‘la persona escucha Rhytm’ y ‘la persona escucha Music’, respectivamente.

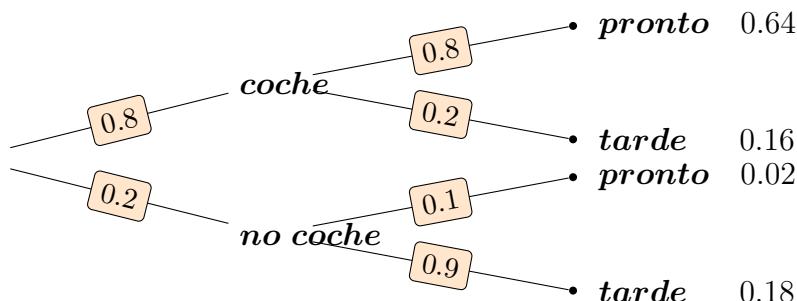


- a) $p(R) = 0.10 + 0.10 = 0.20$
- b) $p(R - M) = 0.10$
- c) $p(M|R) = 0.10/(0.10 + 0.10) = 0.50$

Ejercicio 3.18. A un alumno le lleva en coche a la facultad el 80 % de las veces un amigo.

Cuando le lleva en coche llega tarde el 20 % de los días. Cuando el amigo no le lleva, el alumno llega temprano a clase el 10 % de los días. Determinar:

- a) La probabilidad de que llegue pronto a clase y le haya llevado el amigo.
- b) La probabilidad de que llegue tarde a clase.
- c) Ha llegado pronto a clase. ¿Cuál es la probabilidad de que no le haya llevado su amigo?
- d) Llegar pronto a clase e ir en coche, ¿son sucesos independientes?



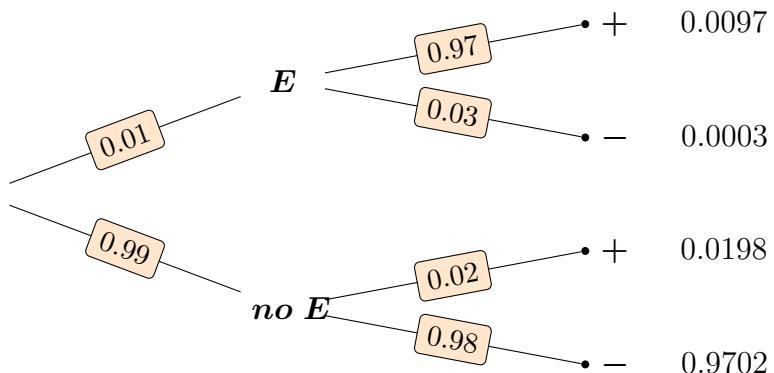
- a) $p(pronto \cap coche) = 0.64$
- b) $p((tarde) = 0.16 + 0.18 = 0.34$

c) $p(\text{no coche} \mid \text{pronto}) = 0.02/(0.64 + 0.02) \approx 0.03$

d) $p(\text{no coche}) = 0.2 \neq 0.03 = p(\text{no coche}|\text{pronto}) \rightarrow$ los sucesos ‘coche’ (o ‘no coche’) y ‘pronto’ no son independientes.

Ejercicio 3.19. El 1 % de la población de un determinado lugar padece una enfermedad. Para detectar esta enfermedad se realiza una prueba de diagnóstico. Esta prueba da positiva en el 97 % de los pacientes que padecen la enfermedad; en el 98 % de los individuos que no la padecen da negativa. Si elegimos al azar un individuo de esa población:

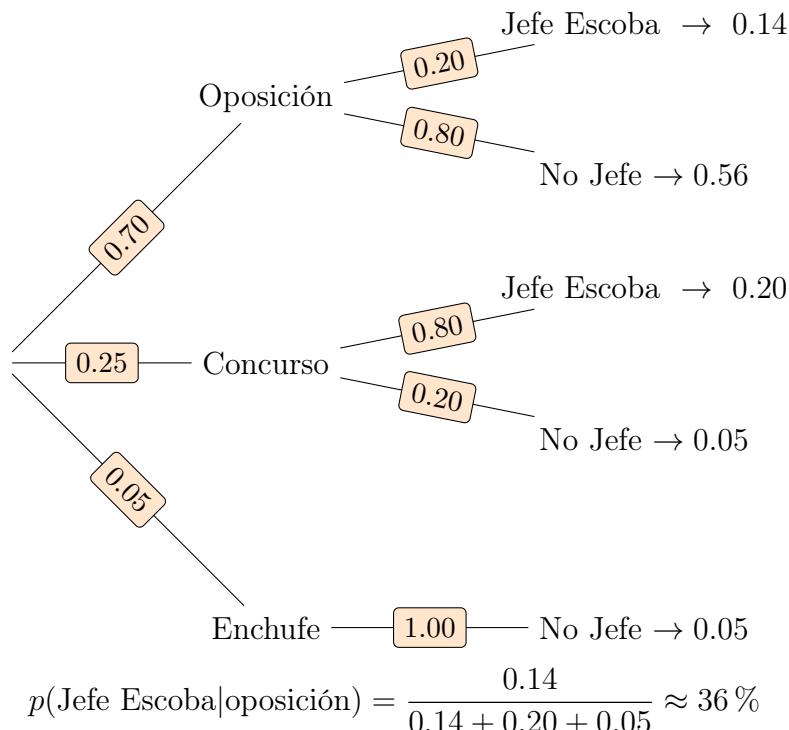
a) ¿Cuál es la probabilidad de que el individuo dé positivo y padezca la enfermedad? b) Si sabemos que ha dado positiva, ¿cuál es la probabilidad de que padezca la enfermedad?



a) $p(E \cap +) = 0.0097;$ b) $p(E|+) = 0.0097/(0.0097 + 0.0198) = 32.9\%$

Ejercicio 3.20. En cierto país, los ascensos de barrendero a jefe de escoba (JE) son muy disputados. Se puede acceder por tres conductos: por oposición, por concurso de méritos o por enchufe con el ministro de Limpieza Pública.

La probabilidad de que un opositor alcance la plaza es de 0.2. La probabilidad de que se obtenga la plaza si se concursa es 0,8. Todos los enchufados del ministro de Limpieza Pública consiguen puesto. Sabiendo que los aspirantes a jefes de escoba se reparten del siguiente modo: 70 % son opositores; 25 % concursan; 5 % consiguen el enchufe, calcular cuál es la probabilidad de que un cierto jefe de escoba alcance la plaza por oposición.

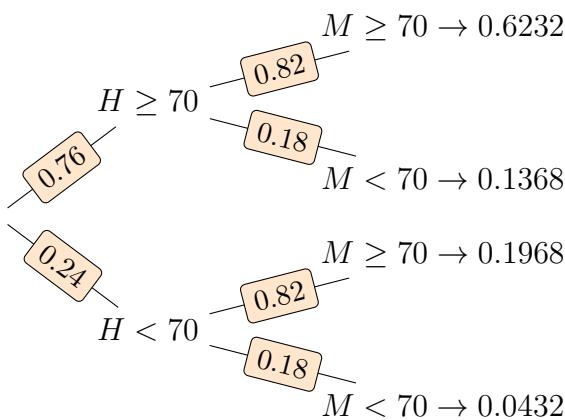


Ejercicio 3.21. Un hombre y una mujer de la misma edad se casan a los 20 años. Las probabilidades de que lleguen a los 70 años son 0.76 para el hombre y 0.82 para la mujer.

Se pregunta cuál es la probabilidad de que a los 70 años:

- a) Ambos estén vivos
- b) No viva ninguno.
- c) Viva solamente la mujer.
- d) Viva al menos uno de los dos.

Podemos resolver el problema con un árbol o con una tabla.



Hemos llamado $H \geq 70$ y $H < 70$ a que el hombre llegue vivo a los 70 años y de que no. Analogamente, para la mujer, $M \geq 70$ y $M < 70$.

	$H \geq 70$	$H < 70$	
$M \geq 70$	0.6232	0.1968	0.82
$M < 70$	0.1368	0.0432	0.18
	0.76	0.24	

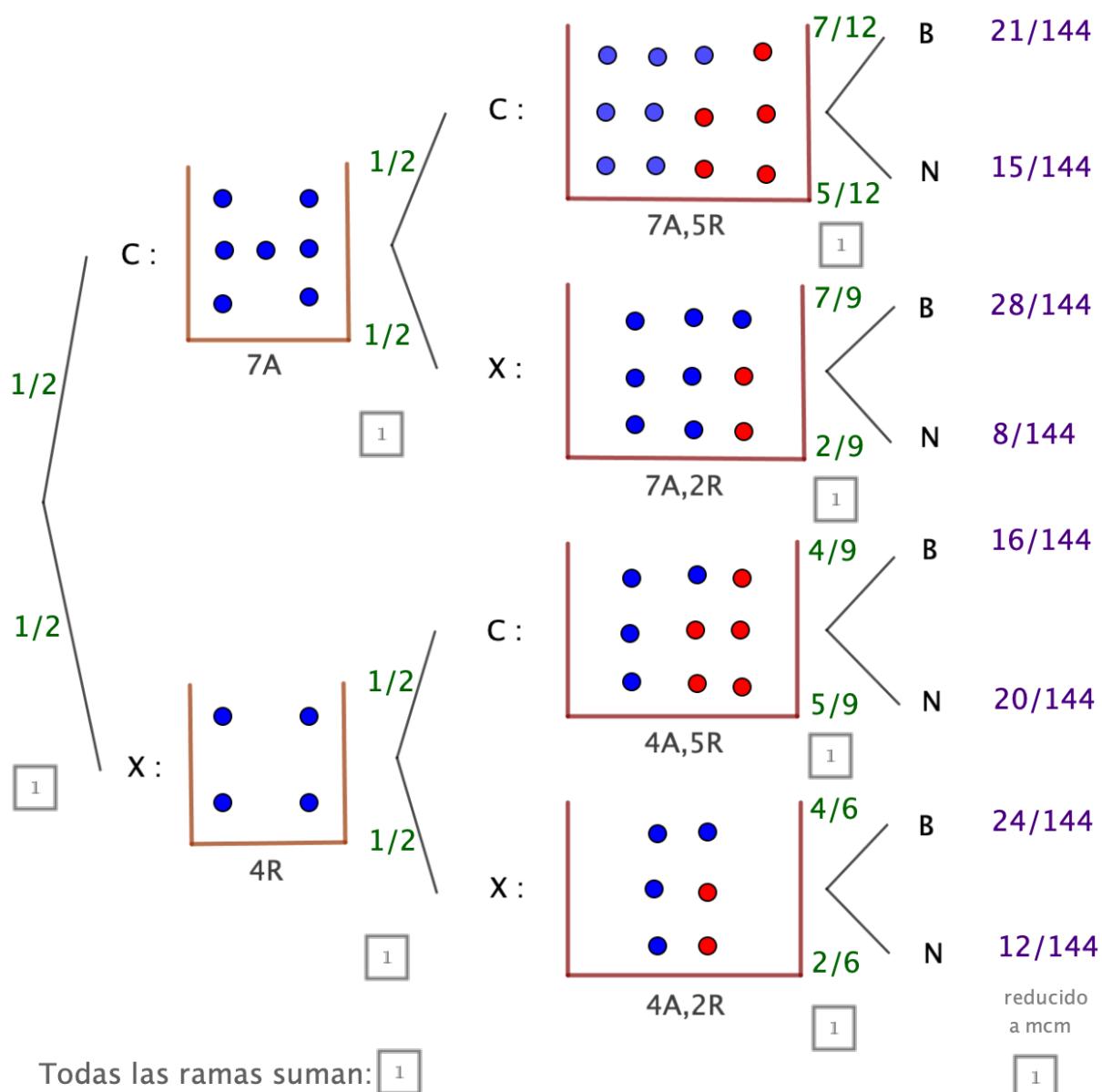
$$p(\text{ambos vivos}) = 0.6232$$

$$p(\text{solo vive mujer}) = 0.1968$$

$$p(\text{ninguno vivo}) = 0.0432$$

$$p(\text{almenos uno vivo}) = 1 - 0.0432 = 0.9568$$

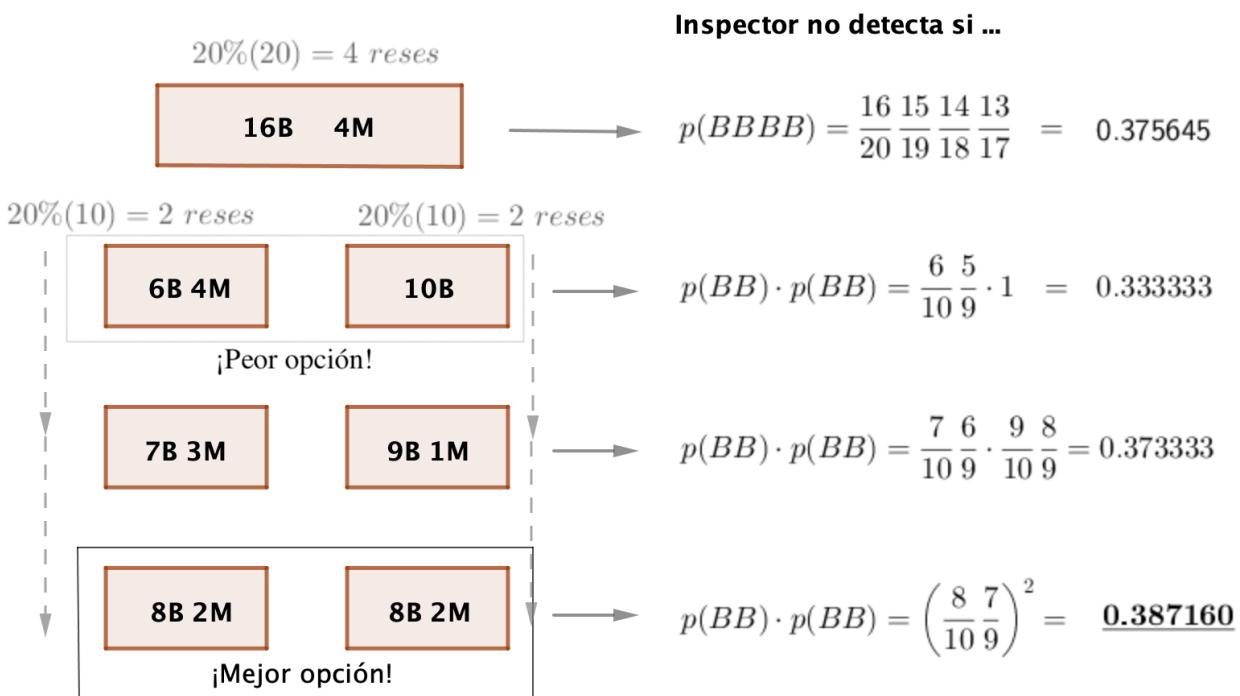
Ejercicio 3.22. Se lanza una moneda y si sale cara se ponen 7 bolas blancas en una urna y si sale cruz se ponen 4 blancas. Se vuelve a lanzar la moneda y se ponen 5 o 2 bolas negras, según se saque cara o cruz. Despues se saca una bola de urna así compuesta. ¿Cuál es la probabilidad de que la bola extraída sea negra? Si la bola ha sido negra, ¿cuál es la probabilidad de que hayan salido 2 veces cara?



$$p(N) = 15/144 + 8/144 + 20/144 + 12/144 = 55/144 = 38.2\%$$

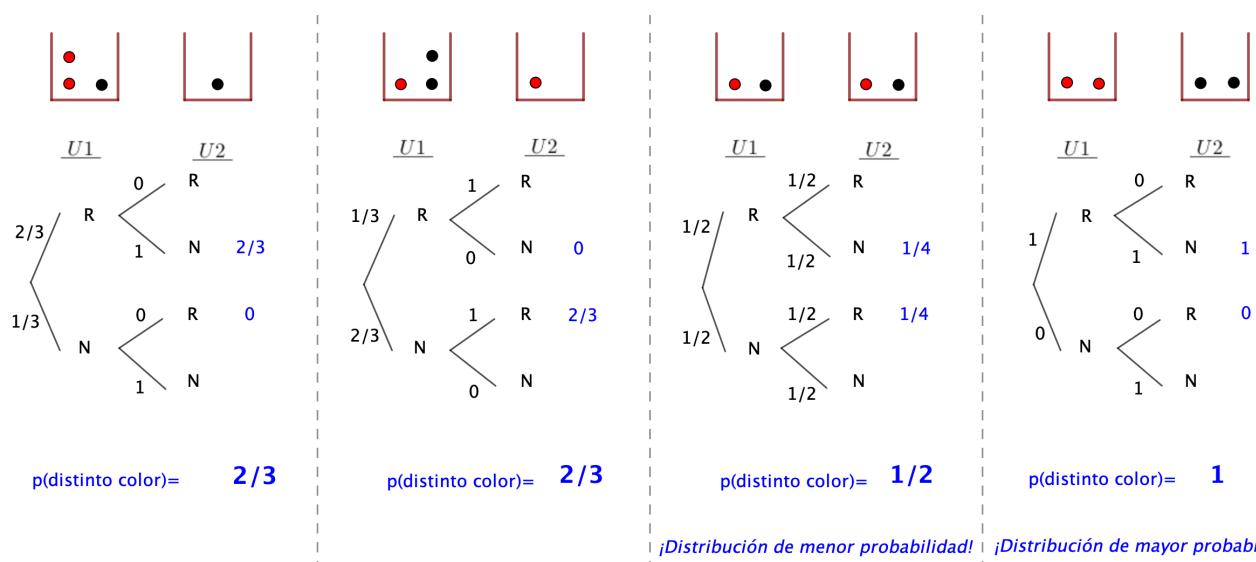
$$p(C_1 \cap C_2 | N) = 21/55 = 38.2\%$$

Ejercicio 3.23. Mr. Bandit, un bien conocido ranchero pero no bien conocido ladrón de ganado, tiene 20 cabezas de ganado listas para vender. Dieciséis de estas cabezas son suyas y consecuentemente llevan su propia marca. Las otras cuatro llevan marcas ajena. Mr. Bandit sabe que el inspector de marcas revisa el 20 % del ganado de cualquier cargamento. Él tiene dos camiones, uno de los cuales puede cargar a las 20 cabezas a la vez, y el otro puede cargar sólo 10 cabezas. Mr. Bandit considera 4 estrategias en su intento de llevar el ganado al mercado para venderlo sin que sea descubierto: 1) enviar en un solo cargamento las 20 cabezas, 2) enviar dos cargamentos de 10 cabezas cada uno, en donde las 4 cabezas robadas se encuentran en uno de los viajes, 3) se envían dos cargamentos de 10, uno con 3 cabezas robadas y el otro con una, y 4) se envían dos cargamentos de 10, cada uno con dos cabezas robadas. ¿Qué estrategia minimiza la probabilidad de que Mr. Bandit sea descubierto?



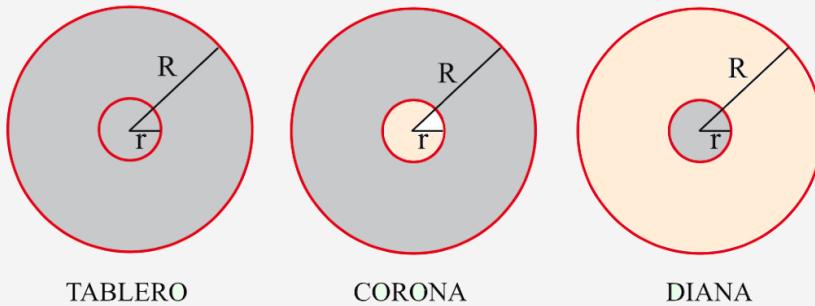
Ejercicio 3.24. Dispones de dos bolas rojas y dos bolas negras que disponer en dos urnas, con la condición de que nunca quede una urna sin bolas.

- a) Comprueba que solo hay 4 disposiciones posibles.
- b) En cada disposición posible de las 4 bolas en las dos urnas, te diriges a ellas y extraes una bola de cada urna. ¿Cuál es la probabilidad, en cada disposición, de extraer bolas de distinto color?
- c) ¿En qué disposición es mayor esta probabilidad? ¿Y menor?



Ejercicio 3.25. En la figura se muestra un tiro al blanco. El punto central del tiro al blanco se llama diana. Se sabe que una persona da al tablero 9 de cada 10 veces que lanza y que, si ha dado al tablero, da a la diana proporcionalmente a la superficie de esta. Sabiendo que $R = 4r$, determine las siguientes probabilidades:

$P(\text{corona})$; $P(\text{diana} / \text{tablero})$; $P(\text{tablero} / \text{corona})$; $P(\text{corona} / \text{no tablero})$; $P(\text{no tablero} / \text{diana})$; $P(\text{diana})$; $P(\text{tablero} / \text{diana})$; $P(\text{diana} / \text{no tablero})$; $P(\text{no tablero} / \text{corona})$; $P(\text{corona} / \text{tablero})$



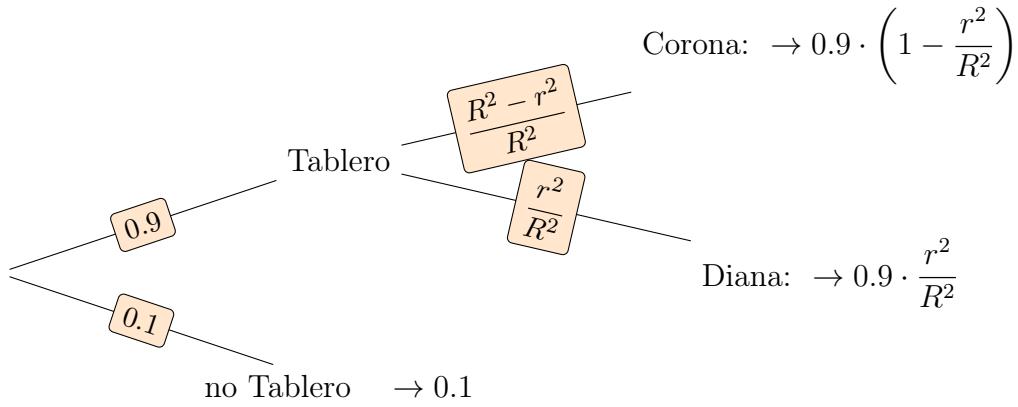
Donde la palabra ‘tablero’ representa el suceso “dar en el tablero”; la palabra ‘diana’, “dar en la diana”; etc.

Una vez impactado el tablero, la probabilidad de hacer diana o corona será proporcional a la fracción de las áreas de estas zonas (A_D , A_C) respecto a la del tablero (A_T)

$$A_T = \pi R^2; \quad A_C = \pi(R^2 - r^2); \quad A_D = \pi r^2$$

$$p(C|T) = \frac{\pi(R^2 - r^2)}{\pi R^2} = 1 - \frac{r^2}{R^2}; \quad p(D|T) = \frac{\pi(r^2)}{\pi R^2} = \frac{r^2}{R^2}$$

Representamos las distintas opciones y sus probabilidades en el siguiente árbol. Llamamos D al suceso ‘dar en la diana’, $D|T$ a ‘dar en la diana sabiendo que se ha dado en el tablero’, etc.



$$p(C) = 0.9 \cdot \left(1 - \frac{r^2}{R^2}\right)$$

$$p(D|T) = \frac{0.9(r^2/R^2)}{0.9} = r^2/R^2$$

$$p((T|C) = \frac{0.9(1 - r^2/R^2)}{0.9(1 - r^2/R^2)} = 1$$

$$p(C|no\ T) = (no\ T|D) = 0$$

$$p(D) = 0.9r^2/R^2$$

$$p(p(T|D) = 1$$

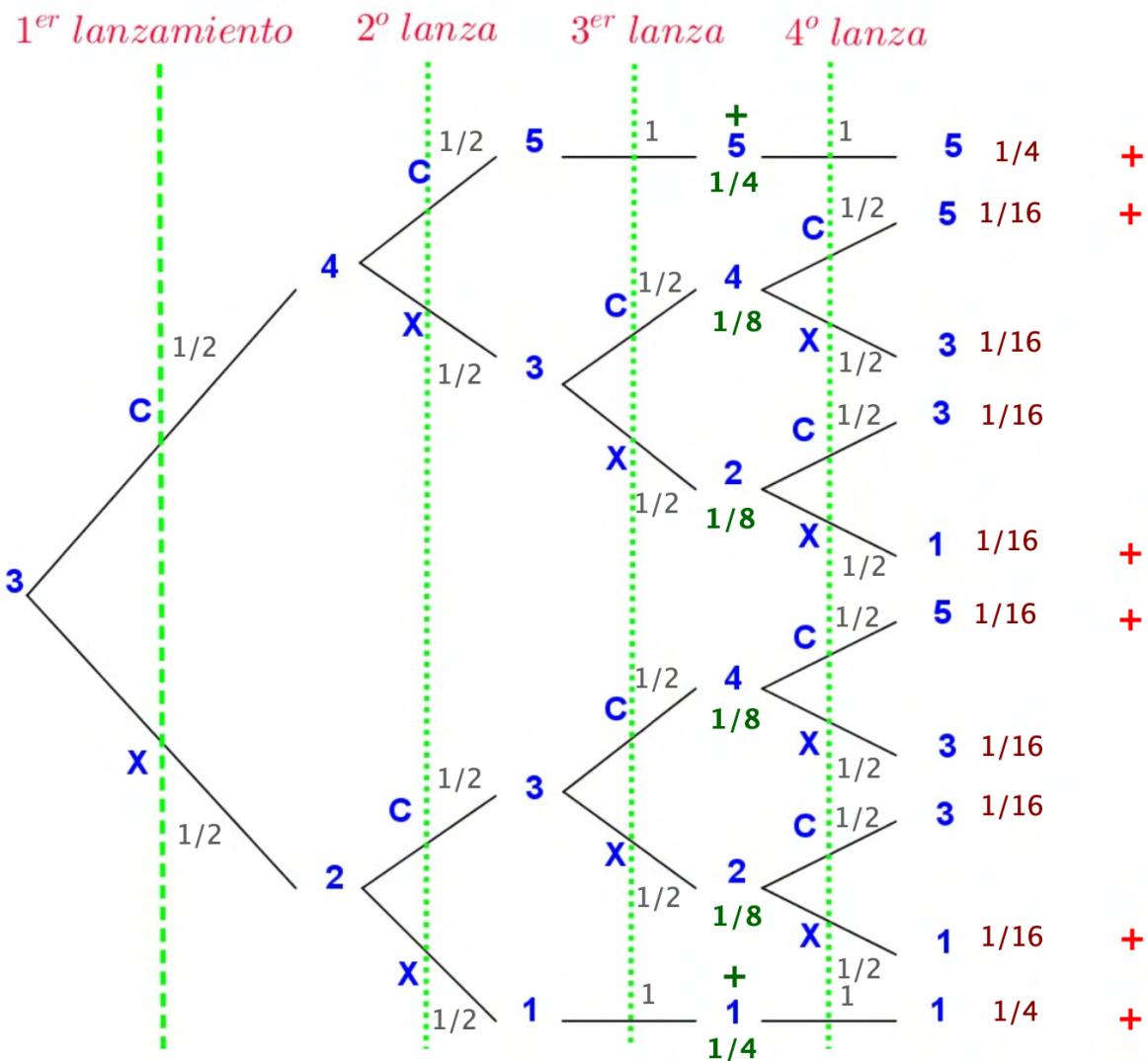
$$p(D|no\ T) = p(no\ T|C) = 0$$

$$p(C|T) = 1 - r^2/R^2$$

Ejercicio 3.26. Imagina cinco sillas alineadas 1, 2, 3, 4, 5 y que un individuo está sentado inicialmente en la silla central (número 3). Se lanza una moneda al aire y, si el resultado es cara, se desplaza a la silla situada a su derecha, mientras que si el resultado es cruz, se desplaza a la situada a su izquierda. Se realizan sucesivos lanzamientos (y los cambios de silla consecutivos correspondientes) teniendo en cuenta que si tras alguno de ellos llega a sentarse en alguna de las sillas de los extremos (1 o 5), permanecerá sentado en ella con independencia de los resultados de los lanzamientos posteriores. Se pide:

- a) Dibujar el diagrama de árbol para cuatro lanzamientos de moneda.
- b) La probabilidad de que tras los tres primeros lanzamientos esté sentado de nuevo en la silla central (3).
- c) La probabilidad de que tras los tres primeros lanzamientos esté sentado en alguna de las sillas de los extremos (1 o 5).
- d) La probabilidad de que tras los cuatro primeros lanzamientos esté sentado en alguna de las sillas de los extremos (1 o 5).

La siguiente figura ilustra la resolución del problema. Hemos colocado un + en los casos en que el individuo queda sentado en las silla 1 o 5 tras 3 lanzamientos y un + cuando queda sentado en las sillas 1 o 5 tras el cuarto.



$p(\text{silla } 3 \text{ tras } 3 \text{ lanzamientos}) = 0$, tras tres lanzamientos se encuentra en las sillas 1, 2, 4 o 5, nunca en la 3.

$$p(\text{silla } 1 \text{ o } 5 \text{ tras } 3 \text{ lanzamientos}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

$$p(\text{silla } 1 \text{ o } 5 \text{ tras } 4 \text{ lanzamientos}) = \frac{1}{4} + 4 \cdot \frac{1}{16} + \frac{1}{4} = \frac{3}{4}$$

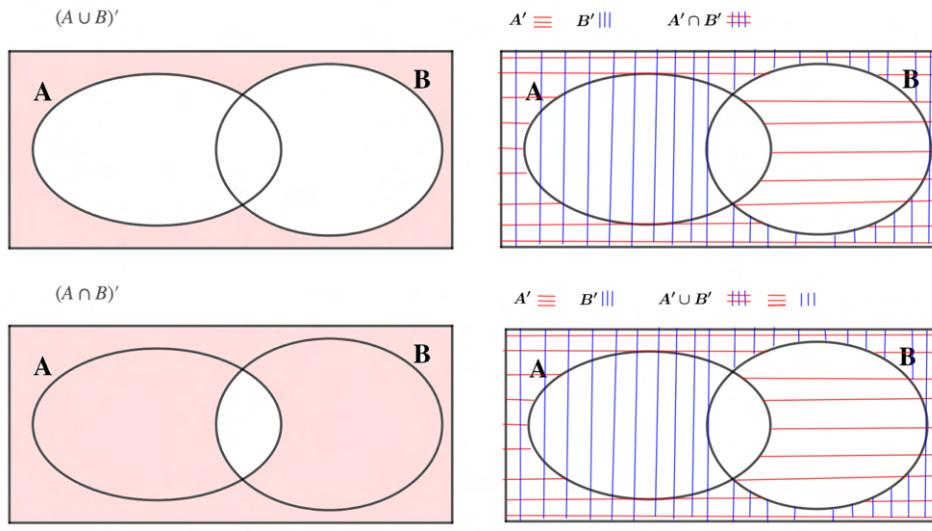
Los siguientes problemas ('machaca', sin enunciado literario) se pueden resolver aplicando las fórmulas vistas en el tema, pero es mucho más fácil resolverlos si nos ayudamos del correspondiente diagrama de Venn.

Ejercicio 3.27. Sean A y B dos sucesos de un espacio de probabilidad tales que: $P[A'] = 0.6$; $P[B] = 0.3$; $P[A' \cup B'] = 0.9$

a) ¿Son independientes A y B ?; ¿son incompatibles?

b) Calcula $P[A|B']$.

Aunque es sencillo volver a recorcar con un diagrama de Venn que $(A' \cup B')$ es el complementario de $(A \cap B)$ (es una de las leyes de Morgan), presentamos la siguiente figura que será de utilidad en los problemas 'machaca'.



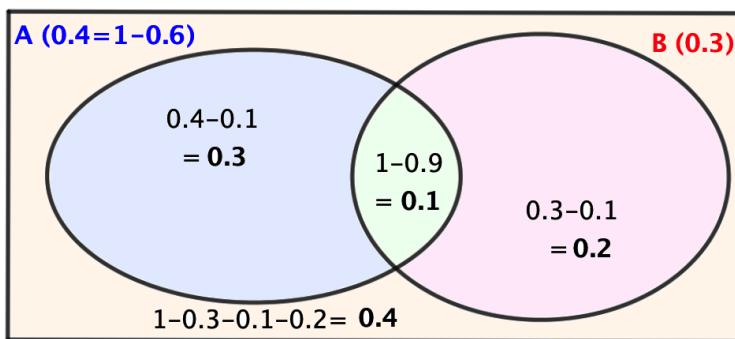
Leyes de Morgan

$$p(A') = 0.6 \rightarrow p(A) = 0.4; \quad p(A \cap B) = 1 - p(A' \cup B') = 1 - 0.9 = 0.1$$

$$\text{Si } P(A) = 0.4 \text{ y } P(A \cap B) = 0.1 \rightarrow p(A - B) = 0.4 - 0.1 = 0.3$$

$$\text{Análogamente, } p(B - A) = 0.3 \cdot 0.1 = 0.2$$

$$\text{La región externa, } p(A \cup B)' = 1 - 0.3 - 0.1 - 0.2 = 0.4$$



$$p(A) = 0.4; \quad p(A|B) = 0.1/0.3 = 0.33 \rightarrow p(A) \neq p(A|B) \rightarrow A \text{ y } B \text{ son dependientes.}$$

$$p(A \cap B) = 0.1 \neq 0 \rightarrow A \text{ y } B \text{ son compatibles.}$$

De la última figura, B' 'ha ocurrido' en $0.3+0.4=0.7$ 'ocasiones', de ellas, A 'ha ocurrido' en solo 0.3 'ocasiones', por lo que: $p(A|B') = 0.3/0.7 = 0.43$. Para facilitar el cálculo basta con

que en el diagrama, al saber que ha ocurrido B' (el suceso que condiciona), tapes con tu mano todo el suceso B . Te queda solo 0.3 y 0.4. De ellos, en 0.3 ocurre A .

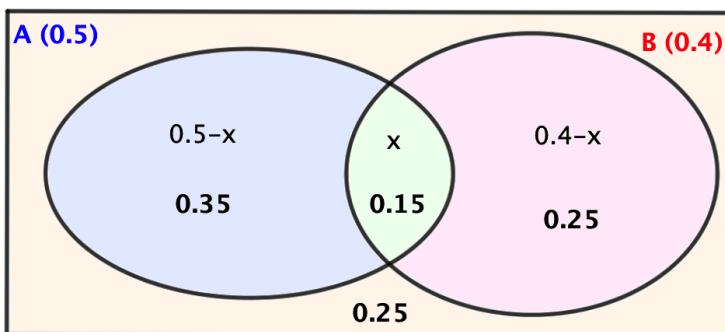
Con un razonamiento análogo, se puede calcular $p(A'|B)$ (si ocurre B , nos quedamos solo con B e ignoramos el resto, tenemos 0.1 -ocurre A^- y 0.2 -ocurre A'^-):

$$p(A'|B) = 0.2/(0.1 + 0.2) = 0.2/0.3 = 0.67 \quad \text{y} \quad p(A'|B') = 0.4/(0.3 + 0.4) = 0.4/0.7 = 0.57$$

Ejercicio 3.28. Sabiendo que $p(A) = 0.5$; $p(B') = 0.6$; $p(A' \cap B') = 0.25$, calcula: $p(A|B)$; $p(A'|B)$; $p(A|B')$; $p(A'|B')$. Calcula, también, $p(A \cap B|A \cup B)$.

$$p(B') = 0.6 \rightarrow p(B) = 0.4$$

O bien dibujándonos quién es $A' \cap B'$ o bien recordando la figura “Leyes de Morgan” del ejercicio 2.20, tenemos que $A' \cap B'$ representa la región externa a A y a B . Ponemos los datos en nuestro diagrama de Venn y nos vemos obligado a llamar $x = p(A \cap B)$, pues nos es desconocida. Con esto, $p(A - B) = 0.5 - x$ y $p(B - A) = 0.4 - x$. Puesto que la probabilidad total ha de ser 1 : $0.5 - x + x + 0.4 - x + 0.25 = 1 \rightarrow x = 0.15$ y ya podemos completar nuestro diagrama de Venn. ¡Ahora, que nos pregunten lo que quieran!



$$0.5 - x + x + 0.4 - x + 0.25 = 1 \rightarrow x = 0.15$$

Para calcular las probabilidades condicionadas a B , nos centramos solo en B y observamos que hay 0.15 y 0.25 posibilidades, en las primeras ocurre A y en las segundas no (ocurre B'). Para probabilidades condicionadas a B' nos centramos en lo que no es B (podemos tapar el conjunto B con nuestra mano) y observamos que quedan 0.35 y 0.25 posibilidades, en las primeras ocurre A y en las segundas A' .

$$p(A|B) = 0.15/(0.15 + 0.25) = 0.375; \quad p(A'|B) = 0.25/(0.15 + 0.25) = 0.625$$

$$p(A|B') = 0.35/(0.35 + 0.25) = 0.583; \quad p(A'|B') = 0.25/(0.35 + 0.25) = 0.417$$

Para calcular $p(A \cap B|A \cup B)$, puesto que sabemos que se ha verificado el suceso $A \cup B$, tenemos 0.35 (ocurre solo A) más 0.15 (ocurren A y B simultáneamente) más 0.25 (ocurre solo B), luego: $p(A \cap B|A \cup B) = 0.15/(0.35 + 0.15 + 0.25) = 0.12/0.75 = 0.200$

Ejercicio 3.29. Dos sucesos tienen la misma probabilidad de ocurrir, 0.5. La probabilidad de que ocurra uno de los sucesos sabiendo que ha ocurrido el otro es 0.3. ¿Cuál es la probabilidad de que no ocurra ninguno de los sucesos?

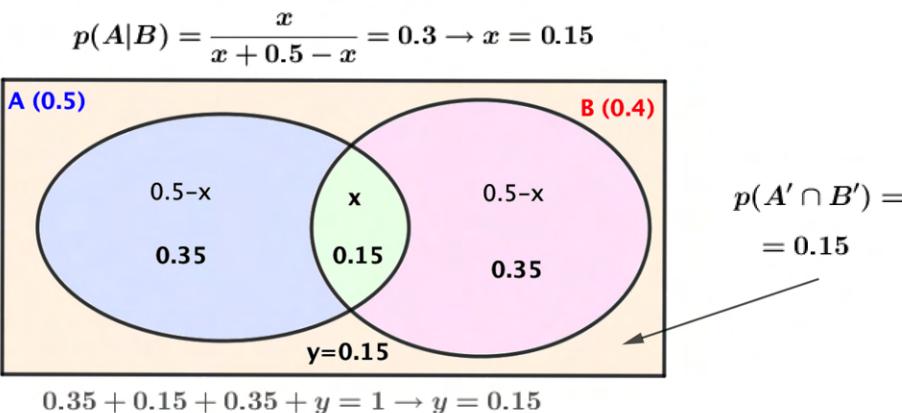
Tenemos que $p(A) = p(B) = 0.5$ y que, por ejemplo, $p(A|B) = 0.3 = \frac{p(A \cap B)}{p(B)} = \frac{p(A \cap B)}{0.5} \rightarrow p(A \cap B) = 0.5 \cdot 0.3 = 0.15$

Lo que es compatible con que $0.3 = p(B|A) = \frac{p(A \cap B)}{p(A)} = \frac{0.15}{0.5} = 0.3$.

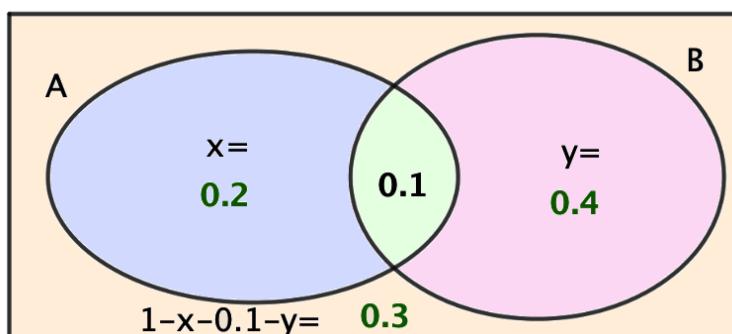
Ahora, $p(A \cup B) = p(A) + p(B) - p(A \cap B) = 0.5 + 0.5 - 0.15 = 0.85$

De donde, $p(A' \cap B') = p(A \cup B)' = 1 - p(A \cup B) = 1 - 0.85 = 0.15$

Con la ayuda de un diagrama de Venn.



Ejercicio 3.30. Dados dos sucesos A y B , sabemos que $p(A \cap B) = 0.1$; $p(A \cup B) = 0.7$ y que $p(A|B) = 0.2$. Calcula $p(A)$ y $p(B)$ y di si los sucesos A y B son independientes y/o incompatibles.



Hemos llamado $x = p(A - B)$ e $y = p(B - A)$.

$$p(A|B) = 0.2 \rightarrow \frac{0.1}{0.1 + y} = 0.2 \rightarrow y = 0.4$$

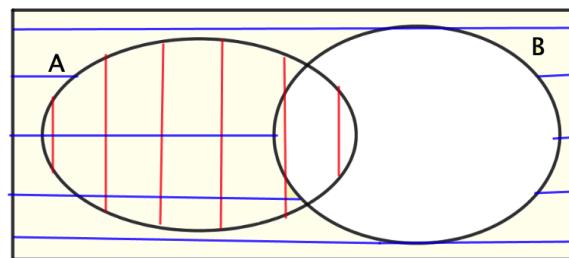
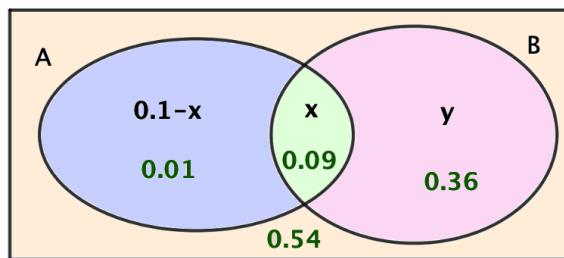
$$p(A \cup B) = 0.7 \rightarrow 0.7 = x + 0.1 + 0.4 \rightarrow x = 0.2$$

$$\text{Con esto, } p(A) = 0.2 + 0.1 = 0.3; \quad p(B) = 0.1 + 0.4 = 0.5$$

Como $p(A \cap B) = 0.1 \neq 0 \rightarrow A$ y B son compatibles.

Como $p(A) = 0.3 \neq 0.2 = p(A|B) \rightarrow A$ y B son dependientes.

Ejercicio 3.31. Sabemos que $p(B|A) = 0.9$; $p(A|B) = 0.2$ y $p(A) = 0.1$. Calcula $p(B)$ y di si los sucesos A y B son independientes y/o incompatibles. ¿Qué vale $p(A \cup B')$?



Llamamos $x = p(A \cap B)$, como $p(A) = 0.1 \rightarrow p(B - A) = 0.1 - x$. Llamamos $y = p(B - A)$.

$$P(B|A) = 0.9 \rightarrow \frac{x}{0.1 - x + x} = 0.9 \rightarrow x = 0.09$$

$$P(A|B) = 0.2 \rightarrow \frac{0.09}{0.09 + y} = 0.2 \rightarrow y = 0.36$$

La zona externa, $p(A \cup B)' = 1 - 0.01 - 0.09 - 0.36 = 0.64$. ¡Listo!

$$p(B) = 0.09 + 0.36 = 0.45$$

$p(A \cap B) = 0.09 \neq 0 \rightarrow$ compatibles.

$p(B|A) = 0.9 \neq 0.45 = p(B) \rightarrow$ dependientes.

En el dibujo de la derecha hemos representado A en vertical y B' en horizontal. Como buscamos la unión, estamos interesados en todo lo rayado, vertical u horizontalmente, la zona amarilla (si buscásemos la intersección nos interesaría la zona en que se cruzan las líneas, $A - B$).

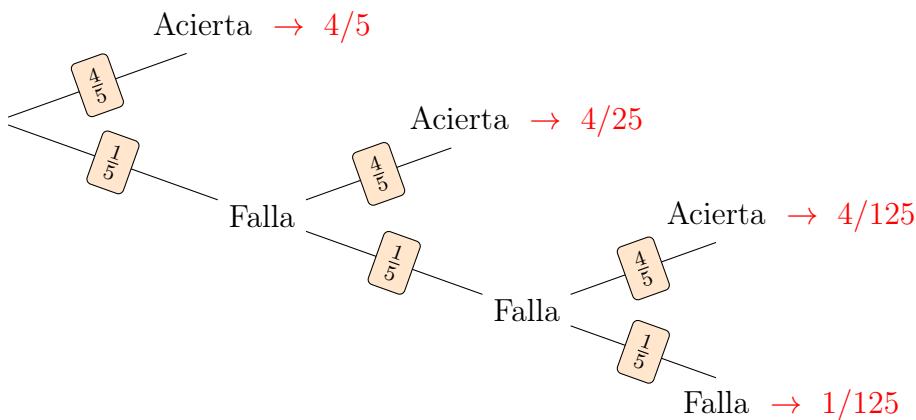
De este modo, $p(A \cup B') = 1 - 0.36 = 0.64$

Los siguientes problemas nos preparan para el siguiente tema de distribuciones de probabilidad. En concreto, para la distribución binomial de probabilidad.

Ejercicio 3.32. Un arquero dispone de 3 flechas en su carcaj. Su probabilidad de hacer diana es constante y vale 0.8; una vez conseguida una diana, acaba el juego.

¿Cuál es la probabilidad que el arquero haga diana al segundo intento? ¿Y de que haga diana en algún intento?

¿Cuál es la probabilidad de fallar todos sus intentos si dispone de 5 flechas en su carcaj?



$$p(\text{acerto 2 intento}) = 4/25$$

$$p(\text{algun acerto}) = 1 - p(\text{ningún acerto}) = 1 - 1/125 = 124/125$$

$$\text{Con cinco flechas: } p(\text{algun acerto}) = 1 - 1/5^5 = 3124/3125$$

Ejercicio 3.33. Un jugador de tenis tiene una probabilidad de 0.4, constante, de ganar una partida, si juega cuatro partidas, calcula la probabilidad de que gane más de la mitad.

Para resolver el problema con un árbol necesitaríamos $2 \cdot 2 \cdot 2 \cdot 2 = 24 = 16$ ramas, excesivo. Intentemos resolver el problema razonando.

En cada jugada, independientemente de jugadas anteriores, la probabilidad de ganar es $p(G) = 0.4$ y la de perder $p(P) = 0.6$. El jugador ganará más de la mitad de las partidas si gana 3 o 4 partidas.

— Ganar 3 partidas es tanto como perder una sola, ésta puede ser la primera, la segunda, la tercera o la cuarta.

Calculemos la probabilidad de perder la primera y ganar las restantes: $PGGG \rightarrow 0.6 \cdot 0.4 \cdot 0.4 \cdot 0.4 = 0.6 \cdot 0.4^3$. Pero la probabilidad de perder en la tercera de las partidas también es: $GGPG \rightarrow 0.4 \cdot 0.4 \cdot 0.6 \cdot 0.4 = 0.6 \cdot 0.4^3$. Así, para los cuatro casos:

$$p(\text{ganar 3 partidas}) = p(3) = 4 \cdot 0.6 \cdot 0.4^3$$

— Ganar la cuatro partidas tiene una probabilidad de $GGGG \rightarrow 0.4 \cdot 0.4 \cdot 0.4 \cdot 0.4 = 0.4^4$

— Finalmente, ganar más de la mitad de las partidas = $p(3) + p(4) = 4 \cdot 0.6 \cdot 0.4^3 + 0.4^4 = 0.1792 \approx 18\%$.

Ejercicio 3.34. Un alumno realiza un examen tipo test que consta de 5 preguntas. Cada una de las preguntas tiene tres posibles respuestas, de las que solo una es correcta. Si el alumno aprueba contestando correctamente tres o más preguntas, obtener de forma razonada la probabilidad de que apruebe si escoge las respuestas de cada una de las preguntas completamente al azar

En esta ocasión, el árbol es aún mayor, de $2^5 = 32$ ramas. Razonemos de modo análoga al problema anterior.

$$p(\text{acertar}) = p(A) = 1/3; \quad p(\text{fallar}) = p(F) = 2/3.$$

Aprueba si acierta tres o más: $p(\text{aprobar}) = p(3) + p(4) + p(5)$, donde con $p(i)$ indicamos la probabilidad de acertar i preguntas.

El suceso contrario, $p(\text{suspender}) = p(0) + p(1) + p(2)$, parece que tiene tantos cálculos como el suceso pedido, por lo que abordamos el problema directamente.

— Probabilidad de acertar 3 de las 5 preguntas:

$$\text{AAAFF} \rightarrow 1/3 \cdot 1/3 \cdot 1/3 \cdot 2/3 \cdot 2/3 = (1/3)^3 \cdot (2/3)^2$$

$$\text{AAFFA} \rightarrow 1/3 \cdot 1/3 \cdot 2/3 \cdot 2/3 \cdot 1/3 = (1/3)^3 \cdot (2/3)^2$$

Todas las formas de obtener 3 aciertos y dos fallos tienen la misma probabilidad, $(1/3)^3 \cdot (2/3)^2$.

¿Cuántas formas hay de acertar 3 de 5 preguntas? $\rightarrow C_5^3 = \binom{5}{2} = \frac{5!}{3! \cdot 2!} = 10$ (* Ver cálculo sin usar combinatoria al final del problema).

$$\text{Luego, } p(3) = 10 \cdot (1/3)^3 \cdot (2/3)^2$$

— Probabilidad de acertar 4 de las cinco preguntas $\rightarrow \text{AAAAF, AAAFA, AAFAA, AFAAA, FAAAA}$, 5 formas distintas ($C_5^1 = C_5^4 = 5$), todas ellas con la misma probabilidad, 4 aciertos y 1 fallo, $(1/3)^4 \cdot (2/3)$.

— Probabilidad de acertarlas todas, una sola forma ($C_5^5 = C_5^0 = 1$, 5 aciertos, 0 fallos), con probabilidad $(1/3)^5$

— Finalmente,

$$p(\text{aprobar}) = p(3) + p(4) + p(5) = 10 \cdot (1/3)^3 \cdot (2/3)^2 + 5 \cdot (1/3)^4 \cdot (2/3) + 1 \cdot (1/3)^5 = 0.2099 \approx 21\%$$

Las distintas formas de ordenar 3 A y 2 F son:

Si las F están juntas: FFAAA; AFFAA; AAFFA; AAAFF

Si las F dejan un espacio entre ellas: FAFAA; AFAFA; AAFAF

Si las F dejan dos espacios entre sí: FAAFA; AFAAF

Si dejan 3 espacios entre ellas: FAAAA

En total, 10 formas distintas

Es recomendable consultar el apéndice B de combinatoria.

Para acabar los problemas resueltos de probabilidad, presentamos los ‘tres problemas clásicos del caballero de Méré’ que dieron lugar a la teoría de la probabilidad.^a

^aAntoine Gombard, Caballero De Meré y experto jugador, planteó a Blaise Pascal tres problemas sobre apuestas. En 1654, Pascal y Pierre de Fermat (1601-1665) mantuvieron abundante correspondencia sobre estos problemas. Las soluciones que entre los dos encontraron sentaron las bases del Cálculo de Probabilidades.

Ejercicio 3.35. ¿Es ventajoso apostar por el resultado de obtener, al menos un 6, en una serie de 4 lanzamientos de un dado?

Suceso contrario, $p(\text{ningún 6 en 4 lanzamientos}) = (5/6)^4$,

luego $p(\text{algún 6 en 4 lanzamientos}) = 1 - (5/6)^4 = 0.5168 > 50\%$, sí es ventajoso.

Ejercicio 3.36. ¿Es ventajoso apostar por el resultado de obtener al menos un doble 6 en una serie de 24 lanzamientos con un par de dados?

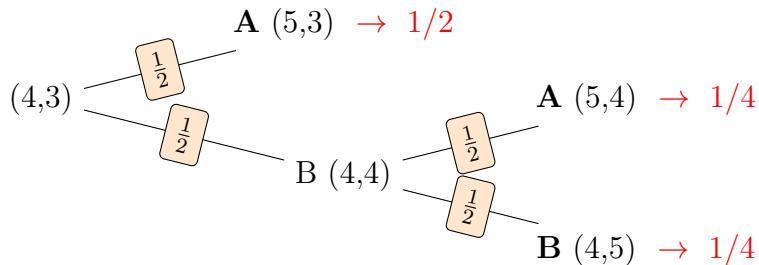
$p(\text{doble 6 al lanzar un par de dados}) = p(6, 6) = 1/36 \rightarrow p(\text{no - 6, 6}) = 35/36$

$p(\text{obtener ningón doble 6 en 24 lanzamientos}) = (35/36)^{24} = 0.5086$

Luego, $p(\text{obtener algún doble 6 en 24 lanzamientos}) = 1 - (35/36)^{24} = 0.4914 < 50\%$, no es ventajoso.

Ejercicio 3.37. La apuesta interrumpida: A y B apuestan, 32 escudos de oro cada uno, a cara o cruz lanzando una moneda. El primero que llegue a obtener 5 puntos gana la apuesta y se lleva todo el dinero (A gana si sale cara y B si sale cruz, el primero en llegar a 5 victorias gana). El juego se interrumpe por causas de fuerza mayor cuando A tiene 4 puntos y B 3 puntos. ¿Cómo deben repartirse el dinero?

Al principio los jugadores A y B están (4,3) (A gana 4 partidas y B gana 3). Ambos jugadores tienen la misma probabilidad de ganar ($1/2$). Si sale cara, gana A y se ponen (5,3); concluye la partida. Si gana B están (4,4) y hacen un segundo lanzamiento. Representamos la situación con el siguiente diagrama de árbol.



Probabilidad de ganar A: $p(A) = 1/2 + 1/4 = 3/4 = 75\%$, probabilidad de ganar B: $p(B) = 25\%$. Los jugadores repartirán la apuesta proporcionalmente a las posibilidades que tienen de ganar:

Para A: 75 % de 64 doblones = 48 doblones ; Para B: 25 % de 64 doblones = 16 doblones .

3.6.1. Problemas propuestos (con solución)

- PB. 1. Un aparato está formado por dos partes A y B. El proceso de fabricación es tal que la probabilidad de un defecto en A es 0,06 y la probabilidad de un defecto en B es 0,07. ¿Cuál es la probabilidad de que el producto no sea defectuoso?

0.8742

- PB. 2. La baraja española consta de diez cartas de oros, diez cartas de copas, diez cartas de espadas y diez cartas de bastos. Se extraen tres cartas. Averiguar razonadamente cuál es la probabilidad de que al menos una de las cartas de oros en los siguientes supuestos:

- No se devuelven las cartas después d la extracción.
- Después d cada extracción se devuelve la carta a la baraja antes de la extracción siguiente.

Píense en el suceso contrario. a) 0.4375; b) 0.4423

- PB. 3. La ciudad A tiene el triple de habitantes que la ciudad B. Un 10 % de habitantes de la ciudad A son alérgicos y un 30 % de habitantes de la ciudad B son alérgicos. Se selecciona un ciudadano sin saber de qué ciudad es. Deducir razonadamente cuál es la probabilidad de que sea alérgico.

Entre todos los habitantes alérgicos de ambas ciudades se selecciona un ciudadano. ¿Cuál es la probabilidad de que sea de la ciudad A?.

0.15; 0.50

PB. 4. En un aparato de radio hay presintonizadas tres emisoras A, B y C que emiten durante todo el día. La emisora A siempre ofrece música, mientras que la B y la C lo hacen la mitad de tiempo de emisión. Al encender la radio se sintoniza indistintamente cualquiera de las tres emisoras.

- a) Obtener de forma razonada la probabilidad de que al encender la radio escuchemos música.
- b) Si al poner la radio no escuchamos música, calcular de forma razonada cuál es la probabilidad de que esté sintonizada la emisora B.

sol

PB. 5. ¿Cuál es la probabilidad de obtener 12 al multiplicar los resultados de dos dados? ¿Y de que la diferencia sea 2?

Haz las tablas de doble entrada. 1/9; 2/9

PB. 6. Una fábrica tiene 3 máquinas que fabrican tornillos, la máquina A produce el 50 % del total de los tornillos de la fábrica, la B el 30 % y la C el 20 %. De la máquina A salen un 5 % de tornillos defectuosos, un 4 % de la B y un 2 % de la C.

- a) Calcula la probabilidad de que un tornillo elegido al azar resulte defectuoso.
- b) Si sabemos que el tornillo elegido es defectuoso, ¿cuál es la probabilidad de que haya sido fabricado por la máquina C?

0.042; 0.095

PB. 7. Sabiendo que $p(A) = 0.3$, $p(B) = 0.4$ y $p(A|B) = 0.2$, calcula:

- a) $p(A' \cup B)$, $p(B|A)$, $p(A' \cup B')$; $P(A \cap B|A \cup B)$;
- b) $p(A|B)$, $p(A'|B)$, $p(A|B')$, $p(A'|B')$
- c) ¿Son A y B Independientes?, ¿son incompatibles?

0.32/0.40; 0.08/0.30; 0.92; 0.08/0.62

0.08/0.40; 0.32/0.40; 0.22/0.60; 0.38/0.60

No; no.

- PB. 8. En cierto país, los trabajadores de la enseñanza se distribuyen así: un 56 % trabaja en enseñanza primaria, un 34 % en secundaria, y el resto en superior (universidades).

La enfermedad que más afecta a este colectivo es la neurosis angustiosa depresivo aguda (**NADA**, o N para abreviar). El 30 % de los que trabajan en primaria la padecen, así como el 40 % de los de secundaria y el 20 % de los que trabajan en superior.

- ¿Cuál es la probabilidad de que al elegir a un enseñante al azar padezca N?
- Sabiendo que el enseñante elegido padece N, ¿cuál es la probabilidad de que trabaje en secundaria?

0.306; 0.444

- PB. 9. Analizada la sangre de los habitantes de una determinada ciudad, resulta que el 40 % son del tipo A, el 35 % del tipo B y el resto del tipo C. Un cierto virus produce una epidemia y ataca de forma distinta a cada habitante. Se comprueba que el 0'03 % de los habitantes con sangre del tipo A son atacados por el virus, también lo son el 0'05 % de los de tipo B y el 0'07 % de los de tipo C. Elegimos un habitante al azar y resulta estar enfermo. ¿Cuál es la probabilidad de que su sangre sea del tipo C?

0.3723

- PB. 10. Un ladrón, al huir de la policía, puede hacerlo por tres calles A, B o C, con probabilidades 0.25, 0.60 y 0.15, respectivamente. Si huye por la calle A, la probabilidad de ser alcanzado por la policía es de 0.4 y es de 0.5 y 0.6 si lo hace por las calles B y C, respectivamente.

¿Cuál es la probabilidad de que la policía capture al ladrón?

Más tarde vemos a la policía conducir a su vehículo al ladrón esposado, ¿cuál es la probabilidad de que haya huido por la calle A?

0.49; 0.204

- PB. 11. En un examen, un alumno sólo ha estudiado 15 temas de los 25 que contiene el cuestionario. El examen consiste en contestar dos temas extraídos al azar del total de temas. Halla la probabilidad de que el alumno sepa los dos temas que le han tocado.

0.35

PB. 12. Se tiene una bolsa con 10 bolas rojas y 6 negras, de la que se extraen dos bolas. Halla la probabilidad de que ambas sean negras.

a) Con devolución a la bolsa de la primera bola extraída. b) Sin devolución.

(a) 9/64; (b) 1/8

PB. 13. En un sorteo hay 20 papeletas y 5 están premiadas. Si se compran dos papeletas, ¿cuál es la probabilidad de que ambas tengan premio?

1/19

PB. 14. La probabilidad de que un hombre fume es 0,6 y la de que una mujer sea fumadora es 0,3. En una fábrica hay un 75 % de hombre y un 25 % de mujeres. Tomamos una persona al azar. ¿Cuál es la probabilidad de que fume?

Una persona desconocida ha dejado un cigarrillo encendido y se ha producido un pequeño incendio. ¿Cuál es la probabilidad de que el causante fuera un hombre?.

0.525; 0.857

PB. 15. Un avión tiene 5 bombas. Se desea destruir un puente. La probabilidad de destruirlo de un bombazo es 1/5. ¿Cuál es la probabilidad de que se destruya el puente?

Árbol asimétrico: 0.67232

PB. 16. Laura y Javier se reparten los ejercicios que les ha propuesto su profesora. Laura se queda con el 45 % y Javier con el resto. Por otro lado, sabemos que Laura resuelve incorrectamente un 10 % de los ejercicios que intenta y Javier, un 8 %.

Halla la probabilidad de que al elegir la profesora un ejercicio al azar, esté mal resuelto.

Halla la probabilidad de que al elegir la profesora un ejercicio al azar, halla sido hecho por Javier, sabiendo que está mal resuelto.

0.089; 0.494

PB. 17. Se lanza una moneda y si sale cara se ponen 7 bolas blancas en una urna y si sale cruz se ponen 4 blancas. Se vuelve a lanzar la moneda y se ponen 5 o 2 bolas negras, según se saque cara o cruz. Después, se saca una bola de urna así compuesta. ¿Cuál es la probabilidad de que la bola extraída sea negra? Si la bola ha sido negra, ¿cuál es la probabilidad de que hayan salido 2 veces cara?

0.382; 0.273

- PB. 18. María y Laura idean el siguiente juego: cada una lanza un dado . Si en los dos dados sale el mismo número, gana Laura; si la suma de ambos es 7, gana María; y en cualquier otro caso hay empate .

- Calcule la probabilidad de que gane Laura, asociado al experimento .
- Probabilidad de que gane María .

Ambos tiene la misma probabilidad: $P(T)d = P(M) = 1/6$

- PB. 19. Una caja con una docena de huevos contiene dos de ellos ro- tos . Se extraen al azar sin reemplazamiento (sin devolverlos después y de manera consecutiva) cuatro huevos.

- Calcular la probabilidad de extraer los cuatro huevos en buen estado.
- Calcular la probabilidad de extraer, entre los cuatro hue- vos, exactamente un huevo roto.

a) 14/33; b) 16/33

- PB. 20. En un aula de dibujo hay 40 sillas, 30 con respaldo y 10 sin él . Entre las sillas sin respaldo hay 3 nuevas y entre las sillas con respaldo hay 7 nuevas .

- Tomada una silla al azar, ¿cuál es la probabilidad de que sea nueva?
- Si se coge una silla que no es nueva, ¿cuál es la probabi- lidad de que no tenga respaldo?

Tabla de contingencia. a) 1/4; b) 7/30

- PB. 21. En una clase hay 12 alumnos y 16 alumnas . El profesor saca consecutivamente a 4 diferentes a la pizarra. Se pide calcular:

- ¿Cuál es la probabilidad de que todos sean alumnas?
- Siendo la primera alumna, ¿cuál es la probabilidad de que sean alternativamente una alumna y un alumno?
- ¿Cuál es la probabilidad de que sean dos alumnas y dos alumnos?

a) 4/45; b) 22/195; c) 176/445

PB. 22. En un experimento aleatorio se consideran los sucesos A y B. La probabilidad de que no se verifique A es 0,1. La probabilidad de que no se verifique B es 0,4. La probabilidad de que no se verifique A ni B es 0,04. Hallar la probabilidad de que:

- a) Se verifique el suceso A o se verifique el suceso B.
- b) Se verifique el suceso A y se verifique el suceso B. ¿Son independientes los sucesos A y B?

a) 0,96; b) 0,54; c) Si.

PB. 23. En un experimento aleatorio, la probabilidad de un suceso A es dos veces la probabilidad de otro suceso B, y la suma de la probabilidad de A y la probabilidad del suceso contrario de B es 1,3 . Se sabe, además, que la probabilidad de la intersección de A y B es 0,18 . Calcular la probabilidad de que:

- a) Se verifique el suceso A o se verifique el suceso B.
- b) Se verifique el suceso contrario de A o se verifique el suceso contrario de B.
- c) ¿Son independientes los sucesos A y B?

a) 0,72; b) 0,72; c) Si.

PB. 24. De una baraja española de 40 cartas se retiran los oros y los ases . De las 27 cartas que quedan se extraen dos cartas al azar (sin devolver la primera). Calcula la probabilidad de los siguientes sucesos:

- a) Ambas son del mismo palo .
- b) Al menos una es una figura .
- c) Únicamente la segunda carta es una figura .

a) 4/39; b) 22/39; c) 3/13

PB. 25. En el último pedido de una fábrica de coches, el 7,5% de los coches tiene cierre centralizado y llantas de aleación. El 67,5% de los coches tienen cierre centralizado y no tienen llantas de aleación. El 87,5% de los coches no tiene llantas de aleación.

- a) ¿Qué porcentaje de coches tiene cierre centralizado?
- b) Entre los coches con cierre centralizado, ¿qué porcentaje tiene llantas de aleación?
- c) ¿Qué probabilidad hay de que un coche no tenga ni cierre centralizado ni llantas de aleación?

Diagrama de Venn. a) 75%; b) 10%; c) 20%

PB. 26. Los gerentes de unos grandes almacenes han comprobado que el 40 % de los clientes paga sus compras con tarjetas de crédito y el 60 % restante lo hace en efectivo . Ahora bien, si el importe de la compra es superior a 100 euros, la probabilidad de pagar con tarjeta pasa a ser 0,6. Si además sabemos que en el 30 % de las compras el importe es superior a 100 euros, calcular:

- a) Probabilidad de que un importe sea superior a 100 euros y abonado con tarjeta.
- b) Probabilidad de que un importe sea superior a 100 euros, sabiendo que fue abonado en efectivo.

(a) 0,18; (b) 0,22

PB. 27. Un ordenador personal está contaminado por un virus y tiene cargados dos programas antivirus que actúan independientemente uno de otro . El programa P1 detecta la presencia del virus con una probabilidad de 0,9 y el programa P2 detecta el virus con una probabilidad de 0,8.

- a) ¿Cuál es la probabilidad de que el virus no sea detectado por ninguno de los dos programas antivirus?
- b) ¿Cuál es la probabilidad de que un virus que ha sido detectado por el programa P1 sea también detectado por P2?

(a) 0,02; (b) 0,80

PB. 28. Una persona cuida de su jardín pero es bastante distraída y se olvida de regarlo a veces. La probabilidad de que se olvide de regar el jardín es $\frac{2}{3}$. El jardín no está en muy buenas condiciones, así que si se le riega tiene la misma probabilidad de progresar que de estropearse, pero la probabilidad de que progrese si no se le riega es de 0,25 . Si el jardín se ha estropeado, ¿cuál es la probabilidad de que la persona olvidara regarlo?

3/4

PB. 29. Un libro tiene 3 capítulos. El 98 % de las páginas del primer capítulo no tienen ningún error. El 93 % del segundo y el 95 % del tercero tampoco tienen nigún error.

El primer capítulo tiene 130 páginas, 153 el segundo y 180 el tercero.

Elegida una página al azar se observa que no contiene ningún error, ¿cuál es la probabilidad de que sea del capítulo 3?

0,3383

PB. 30. En una empresa hay 160 trabajadores. Elegido uno de ellos al azar se dan las siguientes probabilidades de que hablen idiomas:

0.0625 hablan inglés, francés y alemán; 0.175 hablan inglés y francés; 0.15625 inglés y alemán; 0.1375 francés y alemán; 0.375 francés; 0.36625 alemán y 0.425 inglés.

¿Cuántos trabajadores hablan un solo idioma? ¿Cuántos no hablan ninguno de los tres idiomas?

Venn. 120; 40)

PB. 31. En una urna hay 2 bolas blancas y 3 negras. Dos personas sacan, alternativamente, una bola cada uno sin reemplazamiento. Gana el primero que saca bola blanca. ¿Quién lleva ventaja?

Árbol asimétrico. Ventaja para el primero, $\frac{3}{5}$ a $\frac{2}{5}$.

PB. 32. En una casa hay tres llaveros, A, B y C con 5, 7 y 8 llaves respectivamente, de las cuales sola hay una en cada llavero que abre una determinada puerta. Se escoge un llavero al azar y, de él, también una llave al azar para abrir la puerta en cuestión.

¿Cuál es la probabilidad de acertar?

¿Cuál es la probabilidad de haber escogido el tercer llavero y que la llave no abra?

Si la llave escogida abre la puerta, ¿cuál es la probabilidad de que sea del llavero A?

131/840; 7/24; 56/131

PB. 33. En una población, el 40 % tiene el pelo castaño, el 25 % tiene los ojos castaños y el 15 % tiene el pelo castaño y los ojos castaños. Escogida una persona al azar, calcula la probabilidad de que:

- Si tiene los ojos castaños, también tenga el cabello castaño.
- Si tiene el pelo castaño, también tenga los ojos castaños.
- Tenga el pelo castaño o los ojos castaños.

%, 37.5%; 50%

PB. 34. Un hombre quiere abrir su puerta y tiene n llaves de las cuales sólo una abre la puerta deseada. Como no recuerda cual es la llave correcta, prueba las llaves al azar, descartando una llave, si no abre la puerta, esto es, no vuelve a probar con ella. ¿Qué es más probable, que acierte a abrir la puerta en el primer o en el segundo intento?

Igual, $1/n$.

- PB. 35. Colocamos en una bolsa 10 bolas numeradas en la forma siguiente: $-1, -2, -3, -4, -5, +1, +2, +3, +4, +5$. Tomamos al azar una de las bolas y anotamos el número obtenido. Sin devolver la bola a la bolsa tomamos otro número al azar. ¿Cuál es la probabilidad de que el signo del producto de los dos números que hemos obtenido sea positivo?

44.4 %

- PB. 36. Existe una prueba para el diagnóstico del cáncer que acierta en el 90 % de las ocasiones (es decir, cuando un individuo tiene cáncer, se le diagnostica cáncer con probabilidad 0.9 y, cuando no tiene, da como resultado que no tiene la maligna enfermedad con la misma probabilidad). Se sabe además que de cada 1000 habitantes, uno tiene cáncer. ¿Cuál es la probabilidad de que a un individuo que se le ha diagnosticado cáncer, tenga realmente la enfermedad? ¿Qué opinión le merece la prueba como dato aislado para un posible tratamiento?

% I ~

3.7. Curiosidades

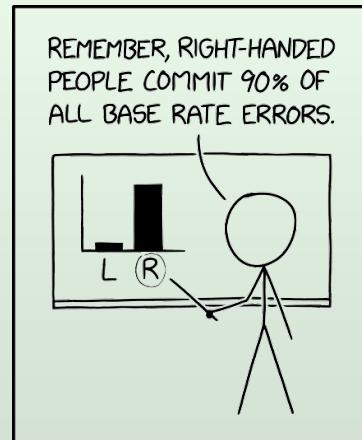
Errores frecuentes en el cálculo de probabilidades

Por ejemplo, si un matrimonio tiene tres hijos, ¿cuál es la probabilidad de que todos sean de un mismo sexo? Podríamos razonar de esta manera: “Por lo menos dos tienen que ser del mismo sexo. El tercero, o será igual o no será igual. Por consiguiente, la probabilidad de que los tres sean iguales es $1/2$ ”. Pero ahora veamos todas las combinaciones posibles; escribiendo V por varón y H por hembra, tendremos: VVV, VVH, VH \bar{V} , V \bar{H} H, H \bar{V} V, HV \bar{H} , HH \bar{V} , HHH. Únicamente en dos de estas ocho combinaciones (VVV y HHH) todos son iguales. Así pues, la probabilidad correcta de que los tres hijos sean de un mismo sexo es $2/8$, o sea de $1/4$.

Otra causa frecuente de equivocaciones en el cálculo de probabilidades consiste en suponer que ciertos sucesos tienen relación entre sí cuando en realidad no la tienen. Muchas personas se imaginan, por ejemplo, que si al tirar una moneda al aire y sale cara varias veces seguidas, lo más probable es que la próxima vez salga cruz. No hay tal. Por más veces que haya salido cara, la probabilidad de que en el próximo tiro salga cruz sigue siendo $1/2$. Muchos sistemas ridículos de jugar a la ruleta y otros juegos de azar se ba-

san en esta “falacia del jugador” que presume que los resultados previos influyen en los futuros (la momeda no guarda memoria de los resultados anteriores).

Algo parecido es el caso del individuo que se creía protegido cuando al viajar en avión, metía una bomba inofensiva en la maleta. Se hacía reflexión de que la probabilidad de que una persona llevara una bomba en un avión es pequeñísima; y la de que dos personas lleven sendas bombas, tiene que ser infinitesimal. La intuición de este buen hombre era decidida, pero su conocimiento de la estadística y la probabilidad puede calificarse de nulo.



Martin Gardner. Selecciones del Reader's Digest.

El problema de probabilidad de los taxis de colores (y su curiosa solución)

Los problemas de probabilidad son divertidos y muy difíciles a veces, especialmente cuando el resultado de los cálculos desafía el sentido común. Los economistas Tversky y Kahneman crearon un problema de este tipo, realmente interesante, de esos que van contra la intuición. Más o menos viene a decir lo siguiente:

“En una ciudad hay dos compañías de taxi: azules y verdes. Un 15 % de los taxis son azules y un 85 % son verdes. En un accidente nocturno, un testigo asegura que vió un taxi azul. Se sabe que gracias a unas pruebas independientes que ese testigo es capaz de identificar correctamente el color de un taxi el 80 % de las veces. ¿De color era el taxi?”

Casi todo el mundo que intenta resolver el problema cree que el taxi era seguramente azul. Pero el taxi era probablemente verde.

Tal y como aprendimos en CSI lo fiable son los datos y las pruebas, no los testigos, que pueden confundir los colores de los coches, sobre todo en una noche oscura y bajo una luz amarilla.

Ateniéndonos únicamente a los datos, como haría Grissom, el cálculo de probabilidades permite darse cuenta de que, en función de las cantidades de taxis y teniendo en cuenta la fiabilidad del testigo, la probabilidad de que el taxi fuera verde es del 59 por ciento, frente sólo a un 41 por ciento de que fuera azul.

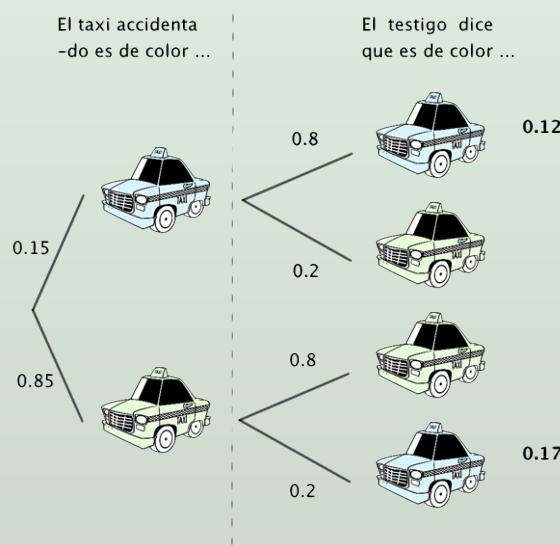
Lo que sucede es que la gente otorga subjetivamente un alto peso a la fiabilidad del 80 % del testigo, que es alta pero no perfecta. En cambio, la gran desproporción de taxis de un

color y otro hace que algo improbable (que el taxi involucrado fuera azul, aun habiendo muchos menos taxis de ese color) siga siendo improbable, y que el testimonio carezca de valor.

El cálculo de estas probabilidades dista de ser trivial y se hace mediante una tabla o un árbol.

¿La moraleja? Que cuanto más improbable sea un hecho por su propia naturaleza, más carente de valor será la fiabilidad de un testigo que diga haberlo visto suceder, a menos que sea infalible.

De @ALVY, en microsiervos.com, (20/02/2008)



Puesto que el testigo asegura que el taxi accidentado es Azul, calculemos la probabilidad de que el taxi sea Azul/Verde, sabiendo que el testigo asegura que es Azul: $P(A|dice\ A)$ y $p(V|dice\ A)$:

$$P(A|dice\ A) = \frac{0.15 \cdot 0.8}{0.15 \cdot 0.8 + 0.85 \cdot 0.2} = \mathbf{0.41}$$

$$P(V|dice\ A) = \frac{0.85 \cdot 0.2}{0.15 \cdot 0.8 + 0.85 \cdot 0.2} = \mathbf{0.59}$$

Es más probable (60 % frente 40 %, aprox.) que el taxi accidentado sea de color Verde.

La falacia del fiscal

“Señoría, tras hallar la sangre de la acusada en la escena del crimen todo queda más claro”, afirmaba el fiscal mientras sostenía su firme mirada ante la jueza. Poco después y tras un leve carraspeo, declaró: “no hay dudas de su culpabilidad”.^a

El fiscal, volvió a tomar asiento y repasó los papeles que tenía delante... la probabilidad de hallar el tipo de sangre de la acusada en la escena del crimen, siendo esta la culpable, era de casi un 98 %, y no llegaba al 100 % porque había una pequeña probabilidad de que la sangre no fuese de la persona que había cometido el crimen.

Pero el fiscal se estaba equivocando en algo, estaba cometiendo la conocida **falacia del fiscal**. Pero ¿en qué se equivocaba? ¿En qué consiste esta falacia?

¿Me dejas que te cuente?

Probabilidad Condicionada

Para poder entender en que se equivocaba el fiscal es importante que conozcamos un concepto básico, la probabilidad condicionada.

Es fácil ser consciente de que determinados sucesos cambian el curso de los acontecimientos y, por tanto, la probabilidad de aquello que nos interesa.

Empezando por un ejemplo muy sencillo, la probabilidad de que, al lanzar dos dados de 6 caras, la suma de los resultados sea 3 es de $2/36$ (de las 36 posibles formas en las que caerán los dados, 2 de ellas tendrán una suma de 3 en los casos: $2+1$ o $1+2$). Sin embargo, si sabemos que en el primero de los dados nos ha salido un 1, la probabilidad aumenta a $1/6$ porque ya solo depende de que en el segundo dado nos salga un 2.

Cuando estamos en este tipo de circunstancias hablamos de probabilidad condicionada y lo expresamos como la probabilidad de que suceda un evento A dado que ha sucedido cierto evento B o, en notación matemática $P(A | B)$.

Cabe mencionar que aquello a lo que condicionamos, lo que hemos llamado evento B, no siempre será algo que haya pasado antes si no que puede hacer referencia a determinadas circunstancias que hacen que cambié todo. Por ejemplo, la probabilidad de ingresar en UCI por la COVID-19 cambia según la franja de edad en la que te encuentras y, por tanto, estamos hablando de probabilidades condicionadas a tus circunstancias, no a algo que ya haya pasado.

Una cuestión importante cuando hablamos de probabilidad condicionada es entender que se trata, al fin y al cabo, de una probabilidad para el evento A. ¿Y qué quiero decir con esto?, pues que nos estamos centrando en la probabilidad del evento A, aunque sea bajo unas condiciones concretas B. Y aquí, el orden de los factores sí altera el resultado, es decir, la probabilidad de que pase A dadas las condiciones B no serán nunca las mismas que la probabilidad del evento B bajo las condiciones A.

A este error es a lo que llamamos en probabilidad la '*falacia del fiscal*' y es un error de interpretación mucho más común de lo que pensamos.

Una falacia común

Volviendo al ejemplo de los dos dados, hemos visto que la probabilidad de que sumen 3 bajo la condición de que el primero de ellos había dado 1 era de $1/6$. Pero, si condicionamos a que la suma sea 3, sabemos que el primer dado solo ha podido dar como resultado 1 o 2, por tanto, la probabilidad de que haya salido un 1 en esas circunstancias sería de $1/2$, muy distinta de la primera.

Puede parecer obvio en este ejemplo pero, la cuestión es que, confundir estas probabilidades es un error típico, por ejemplo, en la detección de enfermedades.

Después de un positivo

Imaginad que os hacen una prueba para ver si sufrís una determinada enfermedad. Toda prueba de este tipo tiene asociada una probabilidad de dar positivo bajo la condición de sufrir realmente la enfermedad. Este valor, conocido como **sensibilidad**, suele ser bastante alto. Pongamos que en nuestro ejemplo es de 98 %. También tiene una probabilidad de dar negativo cuando no se tiene la enfermedad. Este valor es conocido como **especificidad** y también suele ser alta. Pongamos que es de un 95 % en este caso

Ante estas condiciones, obtener un positivo puede suponer un drama. Puede parecer que la probabilidad de tener la enfermedad es del 0.98, pero... recapitulemos.

En este caso, el evento de interés (A) es “tener la enfermedad” y lo que ya es conocido (B) es “haber dado positivo”. Buscamos entonces la probabilidad $P(\text{tener la enfermedad} | \text{haber dado positivo})$.

Sin embargo, la sensibilidad hace referencia a la probabilidad de dar positivo dado que se tiene la enfermedad, esto es: $P(\text{haber dado positivo} | \text{tener la enfermedad}) = 0.98$ y, como ya hemos visto, no tiene porque ser la misma que la anterior.

Pero seguro que ahora os ha surgido una pregunta ¿podemos calcular la primera en función de la segunda?

Esto es lo que se denomina el problema de la probabilidad inversa y la respuesta es sí. De hecho, aquí aparece uno de los teoremas que más me gustan y que, como ya sabéis, es el teorema de Bayes.

A vueltas con el teorema de Bayes

En el Teorema de Bayes que aparecen dos elementos además de las probabilidades condicionadas. El primero es $P(A)$ que, en el caso de la enfermedad representa la probabilidad de sufrirla (con positivo o sin positivo). Este valor recibe el nombre de **incidencia de la enfermedad** y si estamos hablando de una enfermedad rara, será muy bajita, pongamos de 0.001. Por otra parte, $P(B)$ es la probabilidad de que la prueba sea positiva, sin importar si se está enfermo o no. Esta probabilidad, si bien no la sabemos directamente, se puede calcular como:

$P(\text{positivo} | \text{enfermedad}) \cdot P(\text{enfermedad}) + P(\text{positivo} | \text{No enfermedad}) \cdot P(\text{No enfermedad})$ (este es el conocido *Teorema de la Probabilidad Total*).

Pues bien, vayamos a los números:

- $P(\text{enfermedad}) = 0.001$,
- $P(\text{No enfermedad}) = 0.999$;
- $P(\text{positivos} | \text{enfermedad}) = 0.98$, la ‘*sensibilidad*’;
- $P(\text{positivo} | \text{No enfermedad}) = 1 - \text{‘especificidad’} = 1 - 0.95 = 0.05$

Usando entonces el teorema de Bayes, se llega a una probabilidad de sufrir la enfermedad habiendo obtenido un positivo de 0.02. Un valor muy bajo que no tiene nada que ver con la seguridad que en un inicio nos alarmó.

Pues bien, ahora que hemos entendido que es esto de la probabilidad condicionada y la probabilidad inversa, es hora de volver con nuestro fiscal.

Volviendo al juicio

La probabilidad que nuestro fiscal venía manejando era la de haber hallado ese tipo de sangre si la acusada era realmente culpable. Sin embargo, la que le debía interesar realmente, era la de que la acusada fuera culpable dada la única prueba disponible: una muestra de sangre tipo 0–.

Hemos visto que, usando el Teorema de Bayes como en el caso de la enfermedad, el fiscal podía dar la vuelta a la probabilidad que sí que tenía: $P(\text{tipo de sangre } 0- \text{ en la escena del crimen}$

| la acusada cometió el crimen)=0.98 para convertirla en la que realmente le interesaba. Solo necesitaba un valor inicial para la probabilidad de culpabilidad, $P(A)$.

Juguemos con los números

Supongamos que empezamos por creer que la acusada es inocente y damos una probabilidad muy baja a su culpabilidad, de 0.01, por ejemplo. Ahora solo nos falta el denominador, $P(B)$, que en este caso es $P(\text{tipo de sangre } 0- \text{ en la escena del crimen})$. Utilizando de nuevo el teorema de la probabilidad total, podemos calcularla como:

$$P(\text{tipo } 0- \mid \text{la acusada es culpable}) \cdot P(\text{culpable}) + P(\text{tipo } 0- \mid \text{no culpable}) \cdot P(\text{no culpable})$$

Ya sabemos que la primera probabilidad que aparece es la que manejaba el fiscal y tiene un valor de 0.98, la segunda es la de culpable que hemos prefijado en 0.01. Después tenemos la probabilidad de haber hallado ese tipo de sangre si no tenemos ni idea de quien cometió el crimen. Asumimos que esa probabilidad es la misma que en la población general, que para el grupo 0- es de 0.07. Por último, tenemos la probabilidad de nos ser culpable que será $1 - 0.01 = 0.99$.

Combinando todos estos valores y utilizando el Teorema de Bayes nos queda que la probabilidad de que la acusada sea culpable es de 0.12, mucho menor que la probabilidad que manejaba el fiscal. Cabe destacar que, cuanto mas raro sea el tipo de sangre, es decir, más específica sea la prueba, mayor será la probabilidad de que sea culpable.

No solo un ejemplo

El error que cometía nuestro fiscal y que ya hemos dicho que se conoce como la *falacia del fiscal o Prosecutor's Fallacy*, en inglés, no es solo un ejemplo de juguete. Uno de los casos más famosos en los que esta falacia llevó a la cárcel, injustamente, a una persona, fue en el juicio contra Sally Clark. Sally estaba acusada de haber asesinado a sus dos hijos pequeños. Los dos pequeños de la familia Clark habían sufrido muerte súbita, un suceso muy triste pero que se produce de forma natural en los primeros meses de vida de algunos bebés.

Durante el juicio, se incurrió en la falacia del fiscal al considerar lo probable que era que los niños hubiesen muerto si la madre había sido la culpable, pero no la probabilidad de culpabilidad de la madre que se veía considerablemente reducida si se consideraban todas las posibles causas de muerte y, sobre todo, que la muerte súbita podía tener que ver con la genética de los bebés y, por tanto, ser más común entre hermanos.

Así pues, solo me queda decir que no os dejéis llevar por las apariencias, y que siempre penséis con claridad en cual es el evento del que queréis calcular su probabilidad, no vayamos a darle la vuelta.

!Gracias por leer hasta aquí!

Artículo del blog de Anabel Forte (anabelforte.com), doctora en Matemáticas y en Ciencias y Técnicas Estadísticas por la Universitat de València.

^a<http://anabelforte.com/2021/04/18/falacia-fiscal-probabilidad/>

RESUMEN: Cálculo de probabilidades

- ▷ Definiciones a posteriori y a priori de probabilidad:

$$p(A) = \lim_{N \rightarrow \infty} \frac{N_A}{N} \quad \text{Ley grandes números ;} \quad p(A) = \frac{\text{favorables}}{\text{posibles}} \quad \text{Ley de Laplace}$$

- ▷ Axiomática de Kolmogorov:

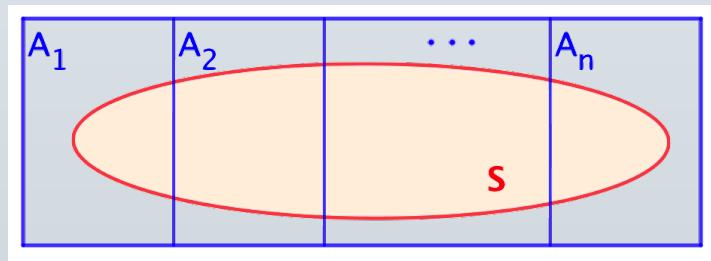
$$p(\emptyset) = 0 \leq p(A) \leq 1 = p(E) ; \quad p(A') = 1 - p(A) ; \quad p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

- ▷ Probabilidad condicionada:

$$p(A|B) = \frac{p(A \cap B)}{p(B)} ; \quad \begin{cases} A, B \text{ independientes si } p(A|B) = p(A) \\ A, B \text{ incompatibles si } p(A \cap B) = \emptyset \end{cases}$$

A y B independiente $\leftrightarrow p(A \cap B) = p(A) \cdot p(B)$

- ▷ Teoremas de la probabilidad total y de Bayes:



$$p(S) = p(S|A_1) \cdot p(A_1) + p(S|A_2) \cdot p(A_2) + \dots + p(S|A_n) \cdot p(A_n)$$

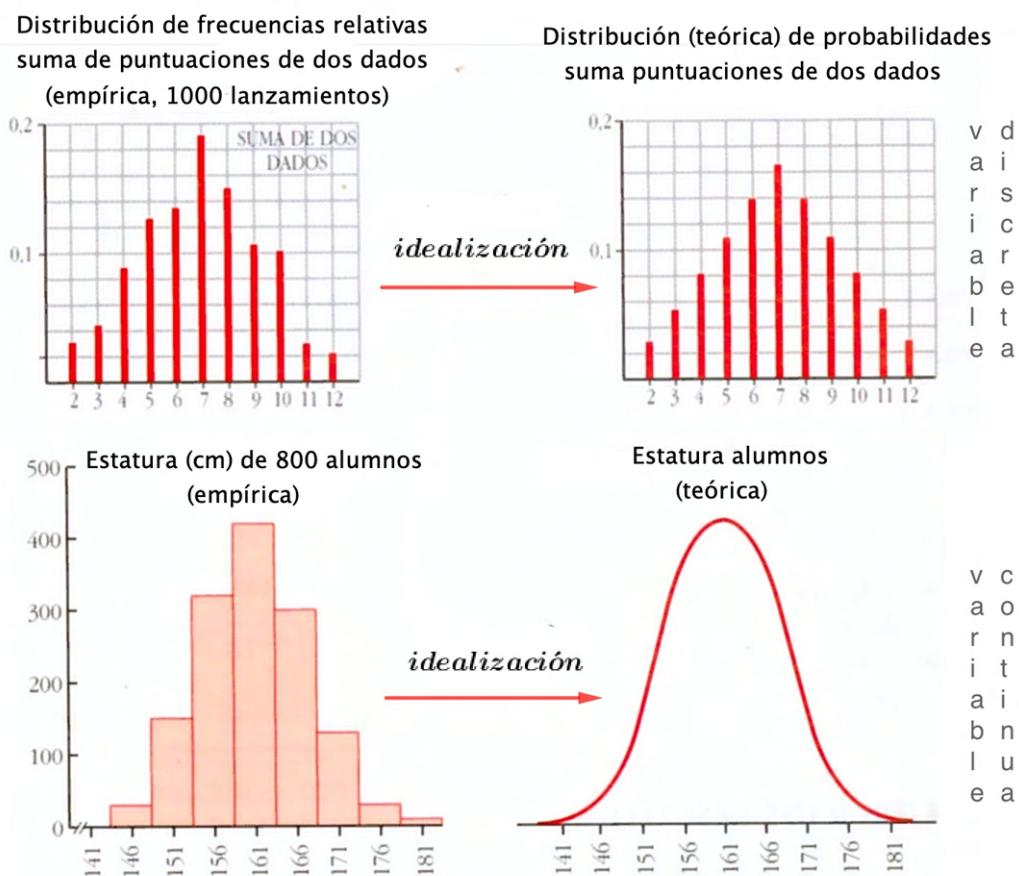
$$p(A_i|B) = \frac{p(A_i) \cdot p(S|A_i)}{p(S)}$$

Capítulo 4

Distribuciones de Probabilidad

Las distribuciones de probabilidad son idealizaciones matemáticas de las distribuciones estadísticas (frecuencias relativas).

Si los valores que toma una variable real x dependen del azar, se dice que x es una *variable aleatoria*, v.a. Por ejemplo, los valores que toma la variable x que representa la suma de puntos obtenidos al lanzar un dado, $\{2, 3, \dots, 12\}$.



A cada valor x_i le corresponderá un valor $p_i = p(x = x_i)$, probabilidad de que ocurra x_i . Diremos que los valores (x_i, p_i) constituyen la *distribución de probabilidad* de la v.a. x .

La distribución de probabilidad de una variable aleatoria es una función que asigna a cada suceso definido sobre la variable la probabilidad de que dicho suceso ocurra. La distribución de probabilidad está definida sobre el conjunto de todos los sucesos. Puede decirse que tiene una relación estrecha con las distribuciones de frecuencia. De hecho, una distribución de probabilidades no es más que una frecuencia teórica (modelo matemático, idealización).

La distribución de probabilidad queda determinada por la *función de distribución*, cuyo valor en cada x real es la probabilidad de que la variable aleatoria sea menor o igual que x .

4.1. Variable aleatoria

Definición 4.1:

Sea E el espacio muestral de un determinado experimento aleatorio. Se llama **variable aleatoria**, v.a., X , a toda aplicación $\mathbf{X} : E \rightarrow \mathbb{R}$

Es decir, a la aplicación que asigna a cada suceso del espacio muestral un número real.

- Si el conjunto imagen, recorrido, $X(E)$ es finito, tenemos una **v.a. discreta**.
- Si el conjunto imagen, recorrido, $X(E)$ es un intervalo (que no sea un solo punto), tenemos una **v.a. continua**.

Ejemplo 4.1:

Consideremos el experimento aleatorio de lanzar dos dados.

El espacio muestral es $E = \{(1, 1); (1, 2); (1, 3); \dots; (1, 6); (2, 1); (2, 2); \dots; (6, 6)\}$

Consideramos la v.a. X tal que a cada elemento de E le asocie su suma, es decir,

$$X(1, 1) = 2; X(1, 2) = 3; \dots; X(2, 2) = 4; \dots; X(6, 6) = 12$$

$$\text{En este caso, } X(E) = \{2, 3, \dots, 12\}$$

4.2. V.A. Discreta

4.2.1. Función de probabilidad

Definición 4.2:

X es la v.a. discreta asociada al espacio muestral E , x_i es un valor de la v. a. ($X = x_i \in X(E)$), la probabilidad de que ocurra x_i , de que la variable aleatoria X tome el valor x_i la denotamos por: $P(X = x_i) = p_i$

La aplicación: $P : X(E) \rightarrow \mathbb{R} : x_i \rightsquigarrow P(X = x_i) = P(x_i) = p_i$ se llama **función de probabilidad o ley de probabilidad de la v.a. X** .

Ejemplo 4.2:

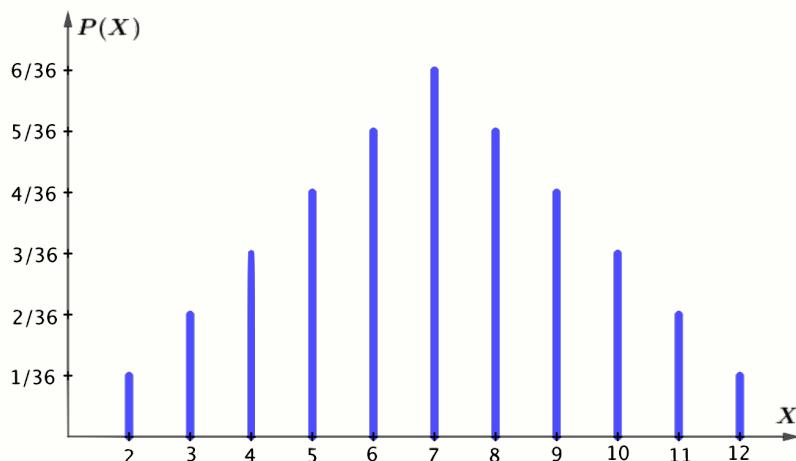
Considerando la v.a. del ejemplo anterior, y ayudándonos de una tabla de doble entrada donde representamos los 36 casos posibles en el lanzamiento de dos dados y la suma de sus puntuaciones, tenemos:

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12



$P(X = 2) = 1/36 = P(X = 12)$; $P(X = 3) = 2/36 = P(X = 11)$; $P(X = 4) = 3/36 = P(X = 10)$; $P(X = 5) = 4/36 = P(X = 9)$; $P(X = 6) = 5/36 = P(X = 8)$; $P(X = 7) = 6/36$

Nótese que $P(2) + P(3) + \dots + P(12) = 36/36 = 1$

**4.2.2. Función de distribución****Definición 4.3:**

Consideremos ordenada la v.a.: $X(E) : x_1 < x_2 < \dots < x_n$, la función:

$$F : \mathbb{R} \rightarrow \mathbb{R} / x \rightsquigarrow F(x) = P(X \leq x)$$

se llama **función de distribución** de la v.a. X

Si $x_k \leq x \leq x_{k+1} \rightarrow F(x) = P(x_1) + P(x_2) + \dots + P(x_k)$, es decir, $F(x)$ es la suma de las probabilidades de los sucesos elementales x_i tales que $x_i \leq x$.

La función de distribución es el concepto idealizado de la frecuencia relativa acumulada de una distribución estadística. F es la ‘probabilidad acumulada’.

Teorema 4.1:

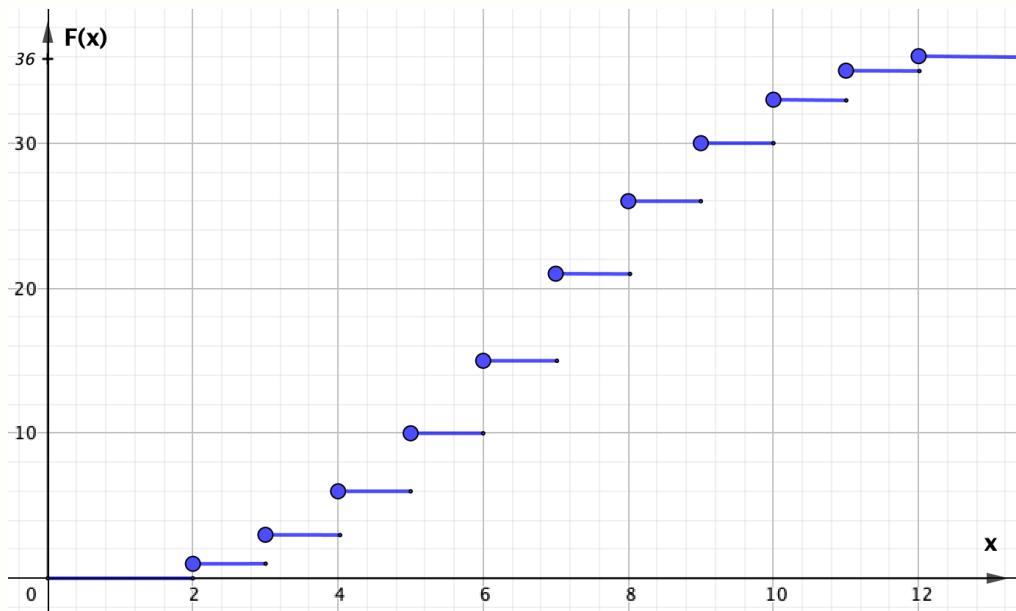
Propiedades de la función de distribución:

- 1) F es creciente;
- 2) F es escalonada;
- 3) $P[x_i < x \leq x_j] = F(x_j) - F(x_i)$

Ejemplo 4.3:

La función de distribución de los ejemplos anteriores es:

$$P(X = 2) = 1/36 = P(X = 12); P(X = 3) = 2/36 = P(X = 11); P(X = 4) = 3/36 = P(X = 10); P(X = 5) = 4/36 = P(X = 9); P(X = 6) = 5/36 = P(X = 8); P(X = 7) = 6/36$$



4.2.3. Esperanza matemática, varianza y desviación típica

variable estadística discreta	X f	$x_1 \quad x_2 \quad \cdots \quad x_n$	$f_1 \quad f_2 \quad \cdots \quad f_n$	idealización \rightarrow	X P	$x_1 \quad x_2 \quad \cdots \quad x_n$	$p_1 \quad p_2 \quad \cdots \quad p_n$	variable aleatoria diccreta
-------------------------------------	------------	--	--	-------------------------------	------------	--	--	-----------------------------------

Definición 4.4:

Al igual que la distribución estadística tenía \bar{x} , s_x^2 , s_x , la **distribución de probabilidad** también tendrá **esperanza matemática**, **valor esperado o media**, que ahora denotaremos por μ , y **varianza y desviación típica**, σ^2 , σ

$$E(x) = \mu = x_1 P(x_1) + x_2 P(x_2) + \cdots + x_n P(x_n) = \sum_{i=1}^n x_i P(x_i)$$

$$\sigma^2 = (x_1 - \mu)^2 P(x_1) + (x_2 - \mu)^2 P(x_2) + \cdots + (x_n - \mu)^2 P(x_n) = \sum_{i=1}^n (x_i - \mu)^2 P(x_i)$$

$$\sigma = \sqrt{\sum_{i=1}^n (x_i - \mu)^2 P(x_i)}$$

Notación: para distribuciones estadísticas: \bar{x} ; s_x , en distribuciones de probabilidad μ ; σ

Teorema 4.2:

Propiedad: se cumple que

$$\sigma^2 = \sum_{i=1}^n x_i^2 P(x_i) - \mu^2$$

Si la v.a. X representa la ganancia o pérdida en un juego de azar, entonces la esperanza matemática representa la ganancia o pérdida media por jugada. Se dice que un **juego es justo** o equitativo si la **esperanza es cero**.

Ejemplo 4.4:

Continuando con nuestro ejemplo de la suma de puntuaciones al lanzar dos dados,

$$\mu = 2 \cdot 1/36 + 3 \cdot 2/36 + \cdots + 7 \cdot 6/36 + \cdots + 12 \cdot 1/36 = 7$$

Haciendo los cálculos necesarios, se obtiene: $\sigma^2 = 54.83$; $\sigma = 2.42$

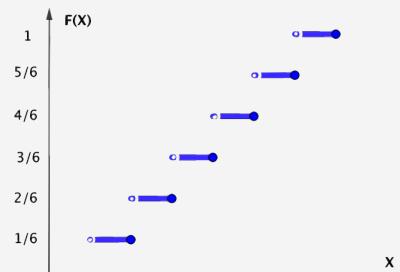
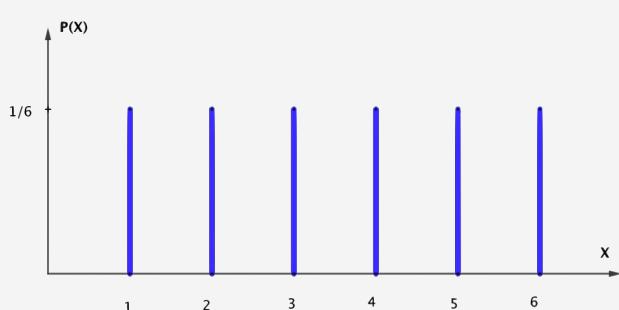
Alrededor de los $2/3$ de las veces que se lance el dado se obtendrán puntuaciones comprendidas en $\mu \pm \sigma = 7 \pm 2.42 \approx (5, 9)$, es decir, alrededor del 67 % de las veces, los dados sumarán 5, 6, 7, 8 o 9. (*Interpretación conjunta de μ y σ*).

Ejercicio resuelto 4.1. Considera la variable aleatoria X que describe las puntuaciones obtenidas la lanzar un dado.

Describe X y representa su función de probabilidad y su función de distribución.

Calcula la esperanza matemática y la desviación típica de las puntuaciones obtenidas al lanzar un dado.

$X = \{1, 2, 3, 4, 5, 6\}; P(X = x_i) = p(x_i) = 1/6, \forall i = \{1, \dots, 6\}$. Todas las puntuaciones tienen la misma probabilidad, $1/6$, de salir.



x_i	1	2	3	4	5	6
p_i	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

Haciendo los cálculos:

$$\mu = 3.5; \sigma = 3.9$$

Ejemplo 4.5:

Es costumbre penalizar los errores en exámenes tipo test, lo cual es obvio para que ningún estudiante opte a tener un determinado número de aciertos si contesta al azar. En el caso de exámenes de preguntas con tres respuestas, cada acierto se contrarresta con dos fallos. Veamos por qué.

Cuando se responda al azar, la puntuación debería ser cero. Esta debería ser su calificación esperada, su esperanza matemática.

En este caso, al contestar al azar, $p(\text{acerto}) = 1/3$; $p(\text{fallo}) = 2/3$. Si otorgamos 1 punto por cada acierto, deberemos penalizar (restar) con x puntos por cada fallo con objeto de que la esperanza matemática de su calificación sea cero:

$$\mu = E(X) = 1 \cdot \frac{1}{3} + x \cdot \frac{2}{3} = 0 \rightarrow x = -\frac{1}{2}, \text{ por cada error restaremos 0.5 puntos (un acierto se compensa con dos fallos).}$$

4.3. Distribución binomial

Distribuciones de probabilidad de v.a. discreta (idealizaciones o modelos matemáticos que representen distribuciones estadísticas) hay muchísimas. La más importante y famosa de ellas es la *distribución binomial*, pero veamos unas más sencillas:

Distribución uniforme (discreta).

Definición 4.5:

Sea X una v.a. discreta que toma los valores $\{x_1, x_2, \dots, x_n\}$ donde la probabilidad de tomar cualquiera de ellos es la misma en todos los casos: $p(X = x_i) = \frac{1}{n}, \forall i$.

X se distribuye como una v.a. discreta **uniforme**, es la distribución más sencilla.

$$\text{Se cumple que: } E[X] = \mu = \frac{1}{n} \sum_{i=1}^n x_i; \quad \text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (x_i - E[X])^2$$

Si la v.a. discreta toma los valores $1, 2, \dots, n$ con probabilidades $p(x = i) = 1/n, \forall i = 1, 2, \dots, n$, se tiene que $\mu = (n+1)/2$ y $\sigma^2 = (n^2 - 1)/12$

Ejemplo: Resultados al lanzar un dado. (Ejercicio resuelto 4.1)

Distribución de Bernouilli.

Un experimento que sólo admite 2 resultados posibles excluyentes:

- Suceso A (representa el éxito) con probabilidad $P(A) = p$.
- Suceso A' (representa el fracaso, contrario a A) con probabilidad $P(A') = 1 - p = q$,

recibe el nombre de *prueba de Bernouilli*.

Definición 4.6:

Consideremos la v.a. discreta X asociada al experimento que asocia el valor 1 al suceso A, éxito, con probabilidad p y el valor 0 al suceso A' , fracaso (0 éxitos), con probabilidad q . Esta variable recibe el nombre de *variable de Bernouilli* y se denota por $X \sim Ber(p)$.

La distribución de probabilidad (función de probabilidad) es: $P(X = 1) = p$ y $P(X = 0) = 1 - p = q$, con $p + q = 1$ y su media, varianza y desviación típica son: $\mu = p$, $\sigma^2 = p \cdot q$, $\sigma = \sqrt{p \cdot q}$

Ejemplos: Estudiar los resultados de lanzar una moneda perfecta o trucada, el sexo de una persona, el que una determinada pieza fabricada sea defectuosa, etc. Todos ellos corresponden a experimentos con dos resultados posibles e incompatibles y no necesariamente de igual probabilidad, son distribuciones de Bernouilli.

Distribución Binomial.**Definición 4.7:**

Supongamos que se realizan n pruebas de Bernouilli sucesivas e independientes. Entonces, a la variable aleatoria discreta $X = \text{"número de veces que ocurre el suceso A (éxito) en las } n \text{ pruebas"}$ se la denomina **variable binomial** de parámetros n y p y se denota

por $X \sim B(n, p)$ donde p es la probabilidad de éxito en cada prueba de Bernouilli (ha de ser constante en todas las n pruebas).

La variable binomial X se puede considerar como la suma de n variables independientes de Bernouilli, es decir $X = X_1 + X_2 + \dots + X_n$ con $X_i \sim Ber(p) \quad \forall i = 1, 2, \dots, n$

La v.a. definida toma los valores $\{0, 1, 2, \dots, n\}$ con probabilidades:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} p^k q^{n-k}; \quad q = 1 - p; \quad k = 0, 1, 2, \dots, n$$

y su media, varianza y desviación típica son:

$$\mu = n p; \quad \sigma^2 = n p q; \quad \sigma = \sqrt{n p q}$$

Los coeficientes de la binomial son números combinatorios $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

. Distribución binomial:

- Consideramos un experimento aleatorio *dicotómico*, es decir, solo consideramos dos posibilidades mutuamente excluyentes: se verifica el suceso A al que llamamos *éxito* o no se verifica, se verifica pues A' , contrario a A y le llamamos *fracaso*.
- La probabilidad de que ocurra A , $p(A) = p = cte$ en el experimento y todas las veces que se realice la experiencia (las sucesivas *pruebas son independientes* unas de otras). Por tanto, también la probabilidad de fracaso será constante, $p(A') = 1 - p(A) = 1 - p = q = cte$.
- Realizamos el experimento n veces y nos preguntamos por la probabilidad de tener k éxitos. $k \in \{0, 1, 2, \dots, n\}$. X es la v.a. discreta binomial que describe el número de éxitos.

La **función de probabilidad de la v.a. binomial** es:

$$f(X = k) = p(X = k) = \binom{n}{k} p^k q^{n-k}$$

De aquí, la **función de distribución de la v.a. binomial** es:

$$F(X = k) = \sum_{x_i \leq k} \binom{n}{x_i} p^{x_i} q^{n-x_i} \quad F(x < 0) = 0; \quad F(x \geq n) = 1$$

Como la distribución binomial de pende de dos parámetros, n (número de experimentos) y p (probabilidad de éxito), es costumbre hablar de ella en la forma $B(n, p)$

Teorema 4.3:

Se demuestra que en una v.a. Binomial, $X \in B(n, p)$:

$$\begin{aligned}\mu &= \mathbf{E}(\mathbf{X}) = \mathbf{n} \mathbf{p} \\ \sigma^2 &= \mathbf{n} \mathbf{p} \mathbf{q} \\ \sigma &= \sqrt{\mathbf{n} \mathbf{p} \mathbf{q}}\end{aligned}$$

$$P(X = x) = F(x) - F(x - 1)$$

Demostración. Demostración del valor de la media en una $X \sim B(n, p)$:

$$\mu = E(X) = \sum_{x_0}^n x \binom{n}{x} p^x (1-p)^{n-x} = 0 + \sum_{x_1}^n x \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x_1}^n x \binom{n}{x} p^x q^{n-x}$$

$$\text{Como } \binom{n}{x} = \frac{n!}{x! (n-x)!} = \frac{n}{x} \frac{(n-1)!}{(x-1)! (n-x)!} = \frac{n}{x} \frac{(n-1)!}{(x-1)! [(n-1)-(x-1)]!} = \frac{n}{x} \binom{n-1}{x-1}$$

$$\text{Luego, } E(X) = \sum_{x_1}^n x \frac{n}{x} \binom{n-1}{x-1} p^x q^{n-x} = n \sum_{x=1}^n \binom{n-1}{x-1} p^x q^{n-x}$$

$$\text{Cambio: } k = x - 1 \rightarrow \begin{cases} x = 1 \rightarrow k = 0 \\ x = n \rightarrow k = n - 1 \end{cases}$$

$$E(X) = n \sum_{k=0}^{n-1} \binom{n-1}{k} p^{k+1} q^{n-(k+1)} = np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k q^{(n-1)-k}$$

$$\text{Binomio Newton: } (a+b)^n = \sum_{k=0}^n n \binom{n}{k} a^k b^{n-k}$$

$$E(X) = \mu = n p (p + 1 - p)^{n-1} = n p$$

□

Son ejemplo de distribuciones binomiales: número de caras al lanzar 100 monedas, número de hijas en 500 familias con 3 hijos, número de familias con un solo hijo en una población de 1000 familias, número de accidentes de tráfico si han circulado 10000 automóviles, número de reacciones negativas ante un fármaco administrado a 5000 pacientes, etc

Ejemplo 4.6:

Se lanzan 6 dados, calcular la probabilidad de obtener exactamente dos cuatros.

El experimento es similar a lanzar un dado 6 veces y observar cuantas veces sale 4.

— Tenemos una experiencia dicotómica, sale 4 (éxito) o no sale 4 al lanzar un dado.

- La probabilidad de éxito es $p = 1/6$; $q = 1 - p = 5/6$ es la de fracaso. Estas probabilidades son constantes todas las veces que se lance el dado (veces que se realice la experiencia)
- Realizamos el experimento $n = 4$ veces y nos preguntamos por la probabilidad de que salgan dos cuatros $k = 2$, tener 2 éxitos.

Tenemos una $B(6, 1/6)$, luego, $f(2) = P(X = 2) = \binom{6}{2} (1/6)^2 (5/6)^4 = 0.20$

El valor esperado de éxitos (veces que sale el 4) en estos 6 experimentos es: $E(X) = \mu = 6 \cdot 1/6 = 1$, por término medio saldrá un 4 de cada 6 lanzamientos, y la desviación típica es $\sigma = \sqrt{6 \cdot 1/6 \cdot 5/6} = 0.91$

Sería conveniente repasar los ejercicios resueltos del tema 3 de probabilidad, 32, 33 y 34, así como el apéndice B de combinatoria.

Dos 4 en seis lanzamientos, ¿cuántos casos hay?. Representamos por 4 el éxito y por _ el fracaso, cuando sale cualquier número que no sea el cuatro.

$44___$; $_44___$; $__44__$; $___44_\$; $____44$ Los 4 salen juntos.

$4__4___$; $_4__4__$; $__4__4_\$; $___4__4$ Les separa una posición.

$4___4__$; $_4___4_\$; $__4___4__$ Les separan dos posiciones.

$4____4__$; $_4____4__$ Les separan tres posiciones.

$4_____4$ Les separan cuatro posiciones.

En total tenemos 15 formas distintas de obtener dos 4 (éxitos) en seis lanzamientos, $15 = \binom{6}{2}$.

Todas ellas tienen la misma probabilidad:

$$p(4___4__) = \frac{1}{4} \frac{5}{6} \frac{5}{6} \frac{5}{6} \frac{1}{4} \frac{5}{6} = \left(\frac{1}{4}\right)^2 \left(\frac{5}{6}\right)^2 ; \quad p(44___4__) = \frac{1}{4} \frac{1}{4} \frac{5}{6} \frac{5}{6} \frac{5}{6} \frac{5}{6} = \left(\frac{1}{4}\right)^2 \left(\frac{5}{6}\right)^2$$

En total, la probabilidad de obtener $x = 2$ éxitos en los $n = 6$ lanzamientos es:

$$p(x = 6) = \binom{6}{2} (1/6)^2 (5/6)^4 = 15 (1/6)^2 (5/6)^4 = 0.20$$

Teorema 4.4:

Propiedades de la distribución Binomial:

1. La distribución Binomial se puede obtener como suma de n variables aleatoria independientes Bernouilli con el mismo parámetro p .

2. Si tenemos dos variables aleatorias que se distribuyen según una Binomial con el mismo parámetro p , es decir, con la misma probabilidad de éxito, $X \rightarrow B(n, p)$ e $Y \rightarrow B(m, p)$, entonces siempre se verifica $X + Y \rightarrow B(n + m, p)$. Si no tienen la misma probabilidad no se pueden sumar.
3. Sea X una variable aleatoria e Y otra variable aleatoria que verifican que $X \rightarrow B(n, p)$ e $Y = X/n$, entonces se verifica $Y \rightarrow B(1, p/n)$ y además su esperanza y varianza son $E[Y] = p$ y $\sigma^2 = pq/n$.

Ejercicio resuelto 4.2. En una celebración *muy numerosa* la gente se dispone en mesas de 5 comensales, hay tantas mujeres como hombres. En una de estas mesas, ¿cuál es la probabilidad de que hayan más mujeres que hombres? ¿Por qué decimos que la fiesta es *muy numerosa*?

Solución: $p(2^a \text{ chica}) = \begin{cases} p(2^a \text{ chica} | 1^a \text{ chica}) = \frac{N/2 - 1}{N - 1} \\ p(2^a \text{ chica} | 1^o \text{ chico}) = \frac{N/2}{N - 1} \end{cases}$ Son distintas, la diferencia

es $\frac{1}{N - 1} \rightarrow 0$ si $N \gg 1$ Es decir, si la fiesta es ‘muy numerosa’, $N \gg 1$, entonces la probabilidad de elegir una chica de un grupo de 5 es prácticamente constante y podemos considerar que tenemos una *distribución binomial $B(5, 1/2)$* , ya que hay tanto chico (fracaso) como chica (éxito), por lo que $p = q = 1/2$.

En un grupo de 5 personas (mesa), habrán más chicas si tenemos 3, 4 o 5 éxitos en la $B(5, 1/2)$, $F(3) = p(x \geq 3) = p(3) + p(4) + p(5)$

$$p(3) = \binom{5}{3} (1/2)^3 (1/2)^2; \quad p(4) = \binom{5}{4} (1/2)^4 (1/2)^1; \quad p(5) = \binom{5}{5} (1/2)^5 (1/2)^0;$$

Para calcular los coeficientes, número combinatorios, podemos acudir a la fila 5 del triángulo de Tartaglia (Ver B.6 “números combinatorios” en apéndice B de combinatoria)

			1	1		
		1	2	1		
	1	3	3	1		
	1	4	6	4	1	
1	5	10	10	5	1	
			$\binom{5}{3}$	$\binom{5}{4}$	$\binom{5}{5}$	

$$F(3) = p(X \geq 3) = [10 + 5 + 1] (1/2)^5 = \frac{16}{2^5} = \frac{1}{2} = 50\%$$

Además, el número esperado (media) de chicas por mesa es, evidentemente, $E(X) = \mu = 5 \cdot \frac{1}{2} = 2.5$; con $\sigma = \sqrt{5 \frac{1}{2} \frac{1}{2}} = 1.12$

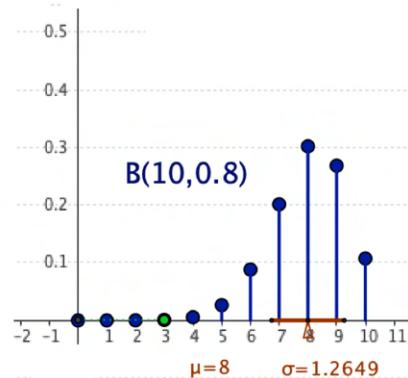
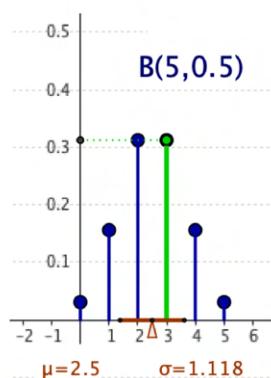
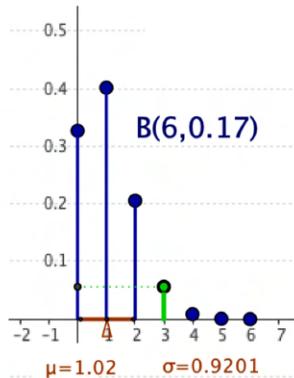
Ejercicio resuelto 4.3. En una $B(10, 0.8)$, calcular $F(X = 9) = p(X \leq 9)$

$$F(X = 9) = p(X \leq 9) = p(8) + p(7) + \dots + p(1) + p(0) = 1 - [p(9) + p(10)]$$

En este caso, es más fácil (menos cálculos) pensar en el suceso contrario.

$$F(x) = 1 - \binom{10}{9} 0.8^9 0.2^1 - \binom{10}{10} 0.8^{10} 0.2^0 = 0.597$$

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$



Existen muchas más distribuciones de probabilidad de variable discreta, pero la binomial es la más importante. De todos modos, vamos a ver una distribución discreta más (también muy importante). Además, en determinadas condiciones (como veremos) sirve como aproximación de la binomial.

Distribución de Poisson.

Definición 4.8:

Una variable aleatoria discreta X se dice que sigue una **distribución de Poisson** de parámetro λ si X toma todos los valores $0, 1, 2, \dots, n$, con probabilidades en cada caso:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Se denota por $X \rightsquigarrow P(\lambda)$ y se cumple que $E(X) = \mu = \lambda$; $\sigma = \sqrt{\lambda}$.

La distribución de Poisson sirve para modelizar el número X de eventos que ocurren aleatoriamente en el tiempo o en una región. Algunos ejemplos de experimentos en los cuales la variable aleatoria puede ser modelizada con distribución de Poisson son:

- El número de llamadas recibidas en una centralita durante un tiempo determinado.
- El número de bacterias por volumen de fluido.
- El número de llegadas de clientes a una caja de pago de un supermercado en un tiempo determinado.
- El número de piezas defectuosas que produce una máquina durante cierto día.
- El número de accidentes de tránsito en un cruce dado durante un tiempo establecido.
- El número de árboles de determinada especie distribuidos aleatoriamente en un área de bosque.

Algunos de estos ejemplos son *procesos temporales*, interesa conocer cuántas veces ocurre un evento en un intervalo de tiempo, y otros son *procesos espaciales*, interesa conocer cuántos “puntos” hay en un volumen o un área.

Definición 4.9:

Un proceso temporal de Poisson es cuando se cumplen con las siguientes características:

- Invarianza: las condiciones no cambian en el tiempo.
- Falta de memoria: lo que sucede en el intervalo de tiempo $[0, t_1)$ no influye en lo que suceda en el intervalo $[t_2, t_3)$, para $t_1 < t_2 < t_3$.
- Sucesos aislados: la probabilidad de que en un intervalo de tiempo muy corto ocurra más de una vez el evento es despreciable comparada con la probabilidad de que ocurra una vez o ninguna.

En estas condiciones, si X_t es la v.a. que mide el número de veces que ocurre el evento en un intervalo de tiempo de longitud t , se demuestra que que X_t es una variable aleatoria discreta cuya función de probabilidad viene dada por:

$$P(x) = \frac{(c \cdot t)^x}{x!} e^{-c \cdot t}; \quad x = 0, 1, 2, \dots$$

X_t es una distribución de Poisson con parámetro $\lambda_t = c \cdot t$, con c una constante positiva que indica la cantidad de veces que ocurre el evento de estudio por unidad de tiempo. c se llama *tasa de ocurrencia del proceso*.

Ejemplo 4.7:

Los clientes a un mostrador de un negocio se distribuyen con una distribución de Poisson a una tasa de 5 por hora. Si queremos saber cuál es la probabilidad de que no lleguen más de tres clientes en una hora, definimos la v.a. X_1 = “cantidad de clientes que llegan al mostrador en una hora”. Entonces $X_1 \sim P(\lambda_1)$, pues $\lambda_1 = 5 \cdot 1$. De esto modo, la probabilidad pedida es: $P(X_1 \leq 3) = F(3) = \sum_{k=0}^3 \frac{5^k}{k!} e^{-5} = 0.2650$

Si queremos calcular la probabilidad de que lleguen al menos 6 clientes en dos horas, no podemos utilizar la v.a. X_1 antes definida, tendremos que *redefinirla*, ya que el intervalo de tiempo ahora es de 2 horas. Luego, X_2 = “cantidad de clientes que llegan al mostrador en dos horas”, $X_2 \sim P(\lambda_2)$, siendo ahora $\lambda_2 = 5 \cdot 2 = 10$. El cálculo de la probabilidad pedida es: $P(X_2 \geq 6) = 1 - P(X_2 < 6) = 1 - P(X_2 \leq 5) = 1 - F(5) = 0.9329$

Si lo que queremos calcular la probabilidad de que lleguen exactamente 5 clientes en media hora, $X_{1/2}$ = “cantidad de clientes que llegan al mostrador en media hora”, $X_{1/2} \sim P(2.5)$ y $P(X_{1/2} = 5) = \frac{2.5^5}{5!} e^{-2.5} = 0.0668$

Definición 4.10:

Se denomina *proceso espacial de Poisson* cuando cumple con las siguientes condiciones:

- Homogeneidad espacial: la probabilidad de que un punto este en una región dada, sólo depende del tamaño de esa región (área o volumen) y no de su forma o posición.
- No interacción: lo que ocurre en una región es independiente de lo que ocurre en otra, si no se superponen.

La v.a. X_a que mide el número de “puntos” en una región de área o volumen a , es distribución de Poisson con parámetro $\lambda_a = c \cdot a$, donde c es *la tasa de ocurrencia del proceso*.

Ejemplo 4.8:

La distribución de plantas de cierta especie en una zona sigue un proceso de Poisson con una tasa de 5 plantas por metro cuadrado. Si deseamos calcular la probabilidad de no hallar plantas en un área cuadrada de 1 metro de lado, definimos la v.a. X_1 = “número de plantas en una región cuadrada de área 1 m^2 ”, donde $X_1 \sim P(\lambda_1)$ con $\lambda_1 = 5 \cdot 1$. Es decir, $X_1 \sim P(5)$ y la probabilidad pedida es $P(X_1 = 0) = \frac{5^0}{0!} e^{-5} = 0.0067$.

Teorema 4.5:**Aproximación de Poisson a la binomial**

Si $X \sim B(n, p)$, se demuestra que cuando n es grande y p pequeño, es válida la siguiente aproximación:

$$P(X = k) = \binom{n}{k} p^k q^{n-k} \approx \frac{\lambda^k}{k!} e^{-\lambda}; \quad \text{con } k = 0, 1, \dots, n \wedge \lambda = np$$

Es decir, $X \approx P(np)$. Esta aproximación es aceptable si $p \leq 0.05$ y $n \geq 20$.

Ejemplo 4.9:

Un peso muy bajo en el nacimiento, menor a 1500 g, es una de las causas de mortalidad infantil. En determinada población, el porcentaje de niños con muy bajo peso al momento de nacer es de 1.2 %. Si consideramos 200 nacimientos en un hospital de esa población, ¿cuál es la probabilidad de que el número de recién nacidos con muy bajo peso en ese grupo sea mayor a 3?

X = “número de niños con muy bajo peso entre los 200 nacimientos de un hospital”, $X \sim B(200, 0.012)$ entonces:

$$P(X > 3) = 1 - P(X \leq 3) = 1 - \sum_{k=0}^{200} \binom{200}{k} 0.012^k (1-0.012)^{200-k} = 1 - 0.7795 = 0.2205$$

Como $p = 0.012 \leq 0.05$ y $n = 200 \geq 20$, se puede usar la aproximación de Poisson a la Binomial y, de este modo, facilitar los cálculos:

$$X \approx P(\lambda = np) = P(200 \cdot 0.012) = P(2.4) \rightarrow$$

$$p(X > 3) = 1 - p(X \leq 3) = 1 - e^{-2.4} \left[\frac{2.4^0}{0!} + \frac{2.4^1}{1!} + \frac{2.4^2}{2!} + \frac{2.4^3}{3!} \right] = 1 - 0.7787 = 0.2213$$

4.4. Variable Aleatoria Continua

Definición 4.11:

Una variable aleatoria continua es aquella que puede tomar cualquier valor (al menos teóricamente) entre 2 fijados.

En estos casos no es posible, como en el caso discreto, asignar una probabilidad positiva a cada valor de la variable de forma que la suma de esas probabilidades sea la unidad y, por tanto, el tratamiento probabilístico de este tipo de variables es distinto al de las variables discretas.

Anteriormente vimos cómo asociar a una variable aleatoria discreta una distribución de probabilidad. Para ello asignábamos a cada uno de los valores del recorrido de la variable aleatoria X la probabilidad de que X tomara ese valor.

Sin embargo, en el caso de una variable aleatoria continua no podemos proceder de igual manera. Considera el experimento consistente en escoger al azar una persona y la variable aleatoria que asigna a cada una su peso. Esta variable puede tomar, en principio, cualquier valor dentro de un intervalo de \mathbb{R} , por lo que hemos de distribuir la probabilidad entre infinitos valores. En consecuencia, *la probabilidad* de que una variable aleatoria continua *tome un valor determinado es*, en general, *cero*. Buscaremos una alternativa para describir las probabilidades asociadas a este tipo de variables.

Lo que nos interesa en variable continua es la probabilidad de que X tome valores en un determinado intervalo.



4.4.1. Función de distribución y función densidad

Definición 4.12:

Si $X(E) = [a, b]$, definimos la **función de distribución** para una v.a. continua como:

$$F(X) = P(X \leq x)$$

Definición 4.13:

Supuesta F derivable, $F'(x) = f(x)$, entonces $F(x) = \int_a^x f(x) dx$.

A $f(x)$ le llamamos **función densidad** de X .

Teorema 4.6:

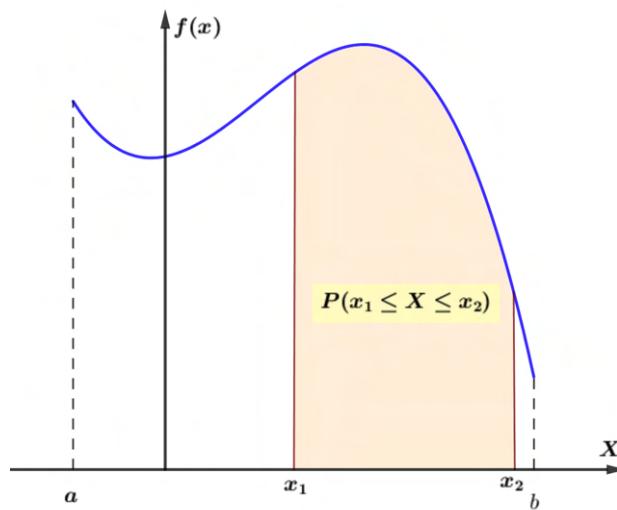
Se cumple que:

$$P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x) dx$$

Geométricamente, la probabilidad de que X tome valores entre dos dados x_1 y x_2 es el área encerrada por la función densidad $f(x)$ entre los valores $X = x_1$ y $X = x_2$.

Obviamente:

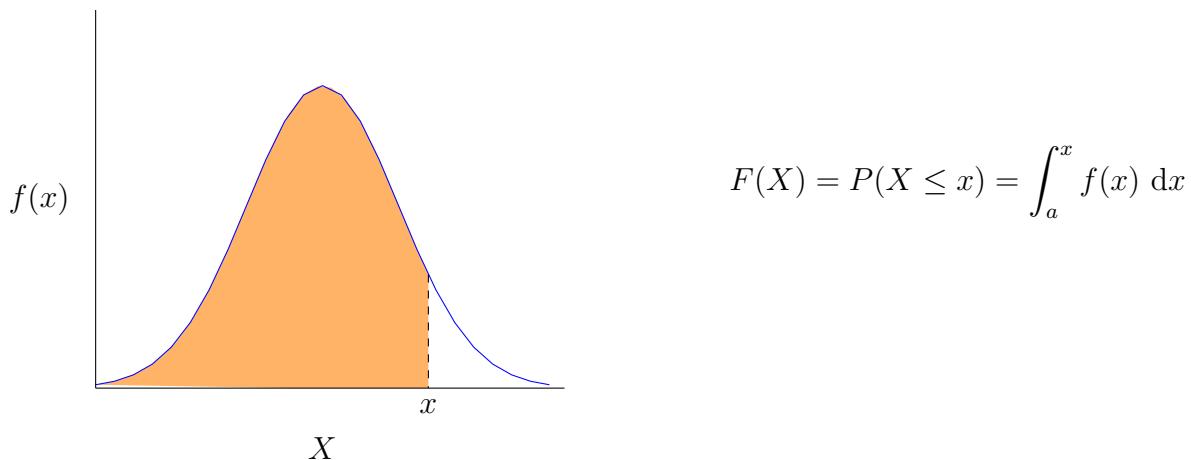
$$P(x_1 \leq X \leq x_2) = P(x_1 < X \leq x_2) = P(x_1 \leq X < x_2) = P(x_1 < X < x_2)$$



$$F(B) = P(X \leq b) = P(a \leq X \leq b) = \int_a^b f(x) \, dx = 1$$

Teorema 4.7:

$$\int_{-\infty}^{+\infty} f(x) \, dx = 1$$



4.4.2. Esperanza matemática y desviación típica de una v.a. continua

Definición 4.14:

Par X v.a. continua, se definen

$$\text{Esperanza matemática } \mu = E(X) = \int_{-\infty}^{+\infty} x \cdot f(x) \, dx$$

$$\text{Varianza } \sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 \cdot f(x) \, dx \quad \text{Desviación típica } \sigma = \sqrt{\sigma^2}$$

Si X toma valores en $[a,b]$, los límites de integración son ‘a’. y ‘b’.

En las distribuciones más usuales, los valores de estas integrales suelen venir tabulados (en forma de tablas).

4.5. Distribución Normal

Existen multitud de distribuciones de variable aleatoria continua, la más importante es la distribución Normal. Veremos antes algunas más sencillas.

Distribución uniforme (continua).

Esta distribución es la más sencilla de las distribuciones continuas, surge al considerar una variable aleatoria que toma valores *equiprobables* en un intervalo finito y su nombre se debe al hecho de que la densidad de probabilidad de esta variable aleatoria es uniforme (constante) sobre todo el intervalo de definición.

Definición 4.15:

Se dice que una v.a. X se distribuye según una **distribución uniforme** en el intervalo $[a,b]$, $X \sim U[a, b]$, si la probabilidad de que tome cualquier subintervalo de valores es proporcional a la longitud de dicho subintervalo.

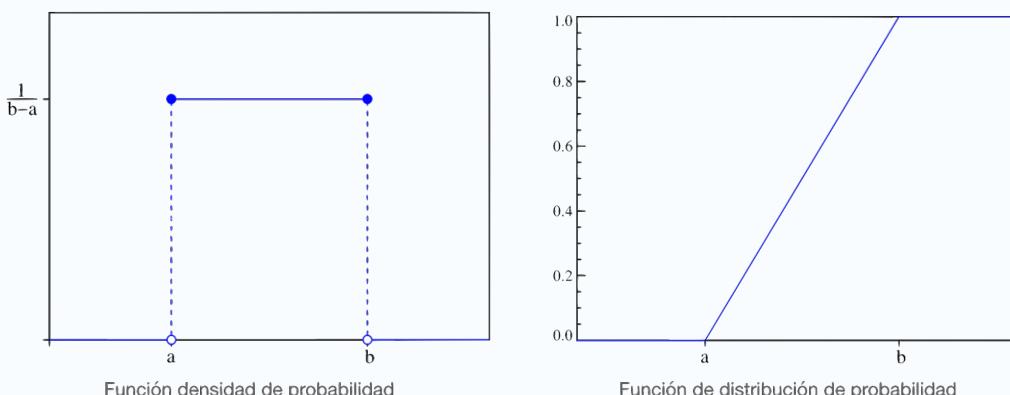
Función densidad: $f(x) = \frac{1}{b-a}; \quad a < x < b; \quad 0 \text{ en cualquier otro caso.}$

Función de distribución: $F(X) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x > b \end{cases}$

Esperanza matemática: $E(X) = \frac{a+b}{2}$

Desviación típica: $\sigma = \sqrt{\frac{(b-a)^2}{12}}$

Distribución UNIFORME de probabilidad de variable aleatoria continua



Como en todas las v.a. continuas, el intervalo de definición de la distribución uniforme puede ser abierto, semiabierto o cerrado.

Ejemplo 4.10:

Elegimos un número real al azar entre $[2, 6]$ y sea X = “número seleccionado”. Calcula la probabilidad de que el número elegido sea menor que 5 y calcula la esperanza matemática (el número esperado).

$$X \sim U([2, 6]) \rightarrow P(X \leq 5) = \int_2^5 f(x) \, dx = \int_2^5 \frac{1}{6-2} \, dx = \int_2^5 \frac{1}{4} \, dx = \frac{1}{4} \left[x \right]_2^5 = \frac{3}{4} = 75\% \\ E(X) = \frac{2+6}{2} = \frac{8}{2} = 4$$

Distribución exponencial .

Esta distribución suele ser modelo de aquellos fenómenos aleatorios que miden el tiempo que transcurre entre dos sucesos. Por ejemplo, entre la puesta en marcha de un cierto componente y su fallo o el tiempo que transcurre entre dos llegadas consecutivas a un cajero automático.

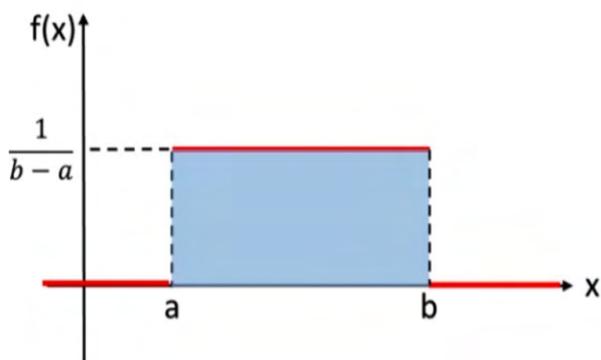
Definición 4.16:

Sea X una v.a. continua que puede tomar valores $x \geq 0$, decimos que X sigue una **distribución exponencial** de parámetro λ (y se nota $X \sim e^\lambda$) si su **función de densidad** está dada por:

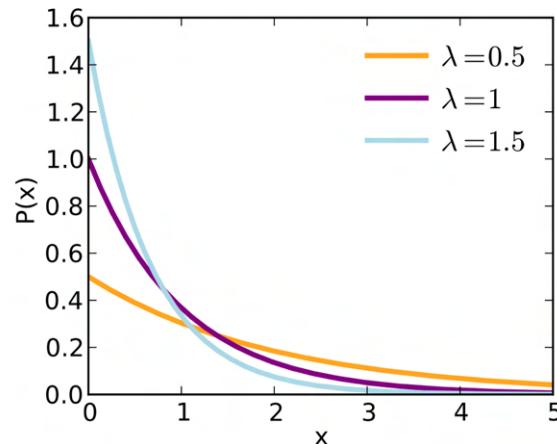
función densidad: $f(x) = \lambda e^{-\lambda x}; \quad x \geq 0; \quad 0 \text{ en cualquier otro caso.}$

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

Esperanza matemática: $E(X) = \frac{1}{\lambda} \quad \quad \quad \text{Desviación típica:} \quad \sigma = \frac{1}{\lambda}$



Distribución uniforme



Distribución exponencial

Ejemplo 4.11:

El tiempo de vida media de una determinada bombilla sigue una distribución exponencial con media de 100 horas.

¿Cuál es la probabilidad de que una bombilla dure más de 30 horas? ¿Y más de 150?

Si cogemos 50 lámparas, ¿Cuántas se espera que duren por lo menos 30 horas?

$$\text{Como } E(X) = 100 = \frac{1}{\lambda} \rightarrow \lambda = \frac{1}{100} \Rightarrow X \sim \exp(1/100)$$

$$P(X > 30) = 1 - P(x \leq 30) = 1 - F(30) = 1 - (1 - e^{-30/100}) = e^{-30/100} = 0.7408 \approx 74\%$$

$$P(X > 150) = \dots = e^{-150/100} = 0.2231 \approx 22\%$$

$n=50$ bombillas, cada una de ellas con la probabilidad de durar más de 30 horas (éxito) igual a $p=0.7408$ forman una distribución binomial $B(50, 0.7408)$, así, el número esperado de éxito de esas bombillas será de $E(X) = n p = 50 \cdot 0.7408 \approx 37$ bombillas.

También podríamos haber razonado así: como la duración de las bombillas es independiente de la bombilla elegida, si el que una dure más de 30 horas tiene una probabilidad del 74% (aprox.), de 50 bombillas durarán más de 30 horas $\rightarrow 50 \cdot 74\% \approx 37$ bombillas.

Distribución Normal o de Laplace-Gauss .

En estadística y probabilidad se llama distribución normal, de Gauus o de Laplace-Gauss a la distribución de probabilidad de variable aleatoria continua que con más frecuencia aparece en estadística.

La gráfica de su función densidad tiene forma acampanada y es simétrica respecto de la media. Esta curva se conoce con el nombre de “campana de Gauss” y a su gráfica se le llama ‘gaussiana’.

La importancia de esta distribución radica en que permite calcular numerosos fenómenos naturales, sociales y psicológicos. Algunos ejemplos asociados a fenómenos que siguen una distribución normal son:

- caracteres morfológicos de individuos como la estatura, peso, etc,
- caracteres fisiológicos como el efecto de determinado fármaco,

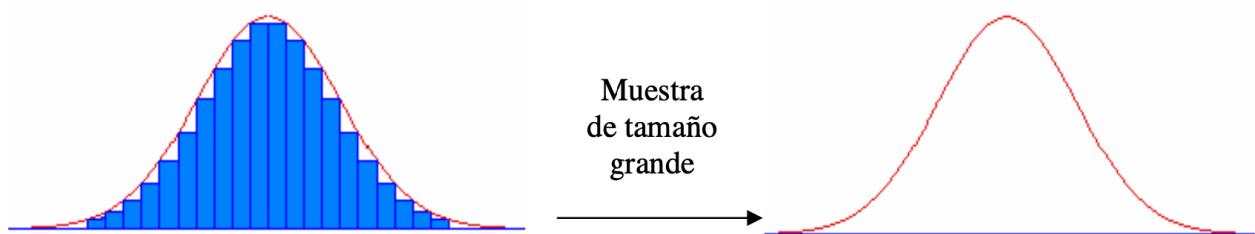
- caracteres sociológicos como el consumo de cierto producto por un grupo de individuos,
- caracteres psicológicos como el cociente intelectual,
- nivel de ruido en telecomunicaciones,
- errores cometidos al medir determinadas magnitudes,
- etc.

La distribución normal es la más extendida en estadística y muchos test estadísticos están basados en ella.

La distribución normal fue presentada por primera vez por Abraham de Moivre en un artículo del año 1733 en el contexto de cierta aproximación de la distribución binomial para grandes valores de n . Su resultado fue ampliado por Laplace en su libro ‘Teoría analítica de las probabilidades’ (1812) y, en la actualidad, se llama teorema de De Moivre-Laplace

El nombre de Gauss se ha asociado a esta distribución porque la usó con profusión cuando analizaba datos astronómicos y algunos autores le atribuyen su descubrimiento.

El nombre de “campana” viene de Esprit Jouffret que usó el término ‘bell surface’ por primera vez para referirse a esta curva. El nombre *distribución normal* fue otorgado, independientemente, por Charles S. Peirce, Francis Galton y Wilhelm Lexis por 1875.



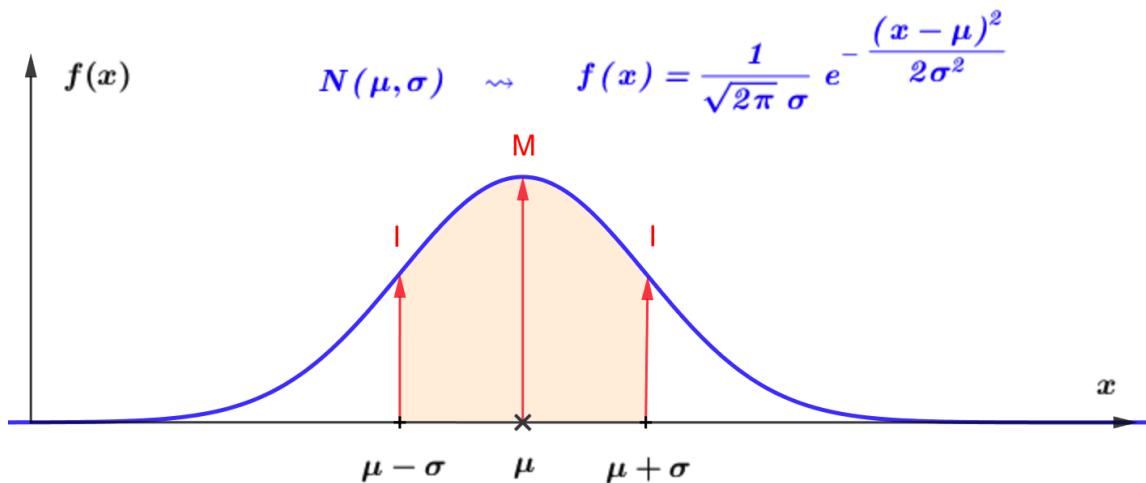
Muchas distribuciones estadísticas pueden逼近arse a una distribución normal.

En muchas ocasiones encontraremos que un conjunto de valores se corresponde con una distribución normal o muy逼近ada a la normal. Además, veremos (próximo tema) que, en determinadas circunstancias, el proceso de ‘muestreo’ garantiza la normalidad aunque la población de la que se extrae la muestra no sea normal. Esto aumenta más aún la importancia del papel de la normal.

Definición 4.17:

Una v.a. continua \mathbf{X} se distribuye según una **Normal** de media μ y desviación típica σ , que denotamos por $\mathbf{X} \rightsquigarrow N(\mu, \sigma)$ si su función de densidad de probabilidad es:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
 $x \in \mathbb{R}; \quad \mu \in \mathbb{R}; \quad \sigma > 0$



Curva normal, ‘campana de Gauss’.

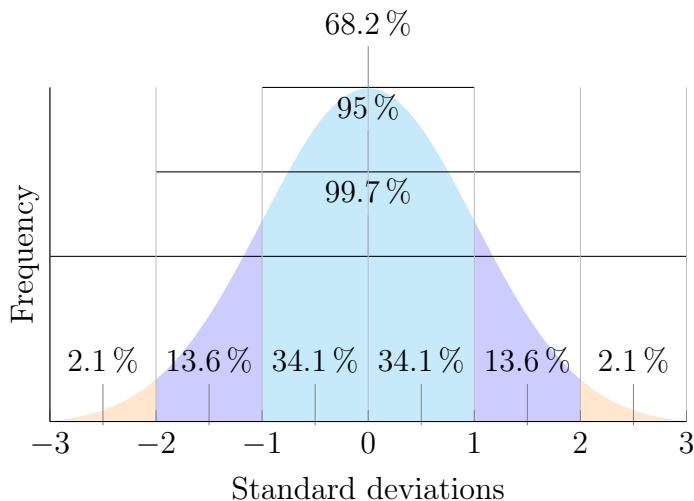
Teorema 4.8:

Propiedades de la ‘campana de Gauss’ (de la Normal).

1. El eje OX ($y=0$) es Asíntota Horizontal de la curva normal.
2. Presenta un **máximo** en $x = \mu$ $M_{ax} \left(\mu, \frac{1}{\sqrt{2\pi}\sigma} \right)$
3. Hay dos puntos de **inflexión** en $x = \mu \pm \sigma$
4. La curva normal es **simétrica** respecto de la recta $x = \mu$: $f(x - \mu) = f(x + \mu)$
5. En la curva normal **coinciden media, mediana y moda**.
6. El área total bajo la curva es uno: $\int_{-\infty}^{+\infty} f(x) dx = 1$
7. Si $Y = aX + B$ y $X \rightsquigarrow N(\mu, \sigma)$ $\Rightarrow Y \rightsquigarrow N(a\mu + b, a\sigma)$
8. Si $X_1 \rightsquigarrow N(\mu_1, \sigma_1)$ y $X_2 \rightsquigarrow N(\mu_2, \sigma_2)$ $\Rightarrow Y = X_1 + X_2 \rightsquigarrow N(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$
9. **Teorema central del límite** (próximo tema):

Si X_1, X_2, \dots, X_n son n v.a. independientes con media μ y desviación típica σ , entonces:

$$\sum_{i=1}^n X_i \rightsquigarrow N(n\mu, \sigma\sqrt{n}); \quad \frac{\sum_{i=1}^n X_i}{n} \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$



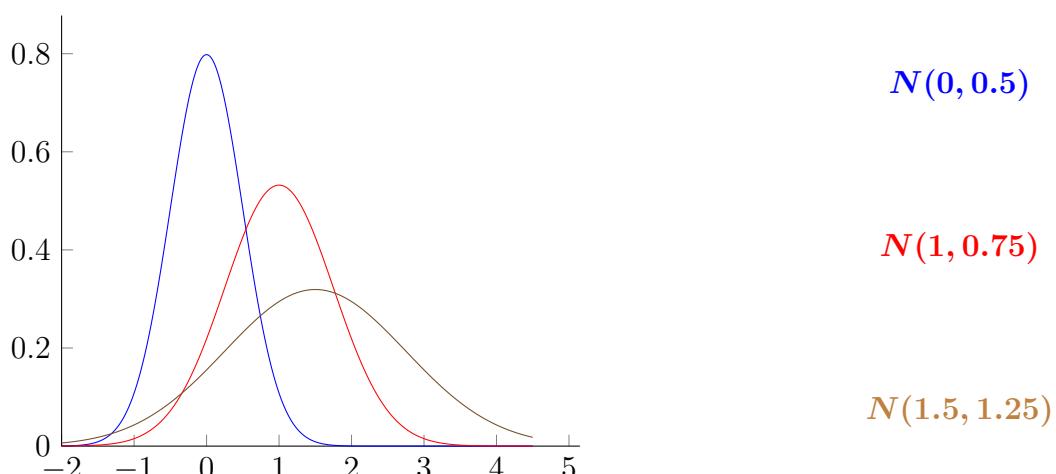
En una distribución normal, la probabilidad de que la variable se aleje menos de una desviación típica de la media es del 68.2 %.

En estadística descriptiva decíamos, en la sección *1.4.5 interpretación conjunta de media y desviación típica*, que los individuos de la distribución que se alejan menos de una desviación típica de la media eran alrededor de los 2/3 de la población, es decir, del 67% de la población. Ello es debido a que muchas distribuciones estadísticas se aproximan a una distribución teórica Normal.

Alejarse menos de dos desviaciones típicas tiene una probabilidad del 95 % y tres desviaciones típicas del 99.7 %.

$$P(\mu-\sigma \leq x \leq \mu+\sigma) = 68.2\%; \quad P(\mu-2\sigma \leq x \leq \mu+2\sigma) = 95\%; \quad P(\mu-3\sigma \leq x \leq \mu+3\sigma) = 99.7\%$$

A mayor desviación típica, los datos son más dispersos respecto de la media. Por ello, las curvas normales son más achataadas a mediada que aumenta la desviación típica indicando la mayor dispersión de los datos. Los distintos valores de la media hacen que las curvas normales estén centradas en distintos puntos.

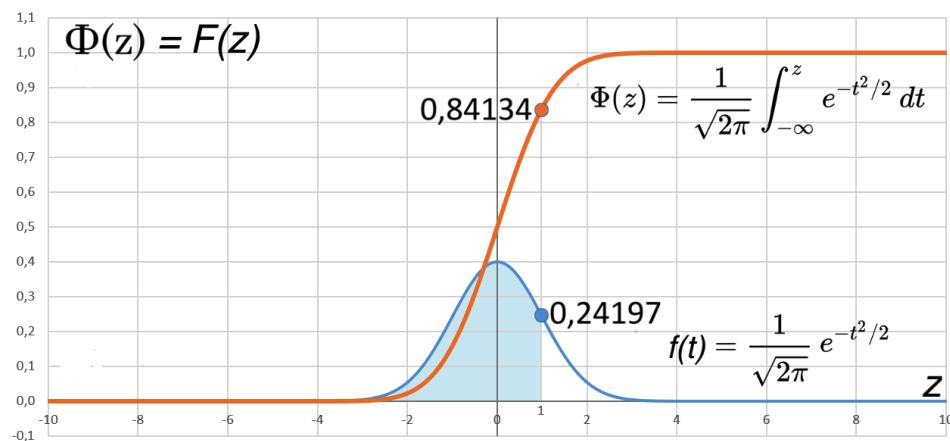


4.5.1. Normal standard (típica o tipificada)

Definición 4.18:

$$\text{Si } \mu = 0 \text{ y } \sigma = 1 \rightarrow N(0, 1) : f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$$\text{Ahora, } F(x) = P(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt \rightarrow \text{TABLAS}$$



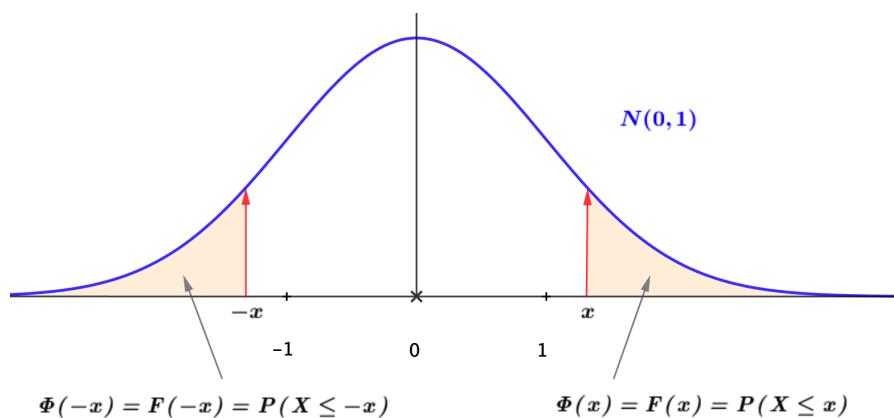
Sería muy complicado elaborar tablas que proporcionen la probabilidad de cada uno de los casos distintos que se pueden presentar con las distribuciones normales; lo que sí existe es una alternativa sencilla que evita estos problemas cuando tenemos un conjunto de valores que tiende a tomar un comportamiento de tipo normal. Y para ello sólo utilizamos un “miembro” de la familia de distribuciones normales: aquella cuya $\mu = 0$ y una $\sigma = 1$. Esta distribución se conoce como **Distribución Normal Estándar**, para ella si está tabulada la función de distribución.

todas las distribuciones pueden convertirse a la estándar por medio del proceso que llamaremos ‘tipificación de la variable’ y así poder calcular probabilidades en cualquier distribución normal $N(\mu, \sigma)$.

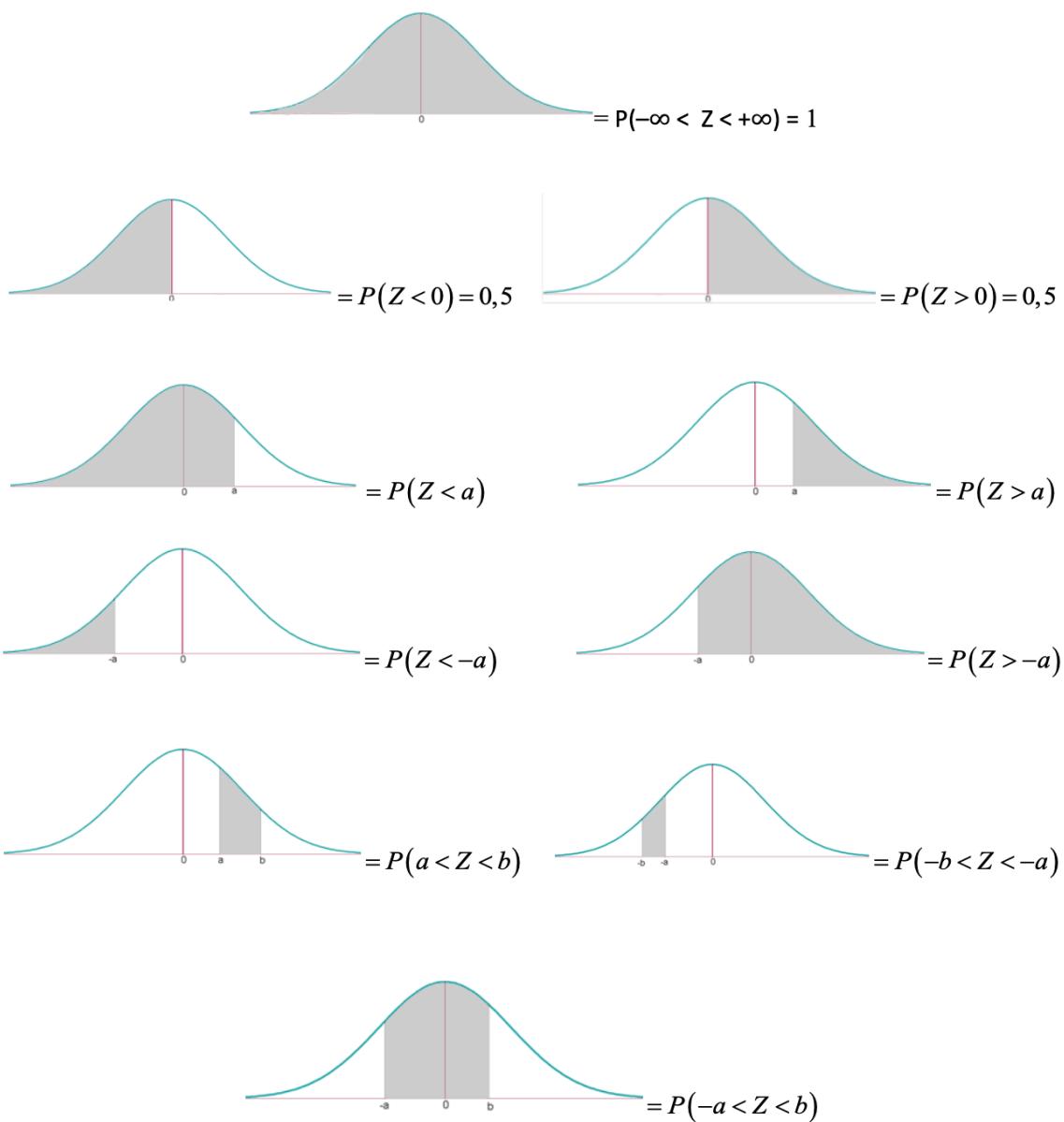
Teorema 4.9:

Propiedades de la $N(0,1)$

1. Es simétrica respecto de $x = 0$: $f(-x) = f(x)$
2. Tiene un $M_{ax} = (0, 0.4)$
3. Tienen inflexiones en $x = \pm 1$
4. $y = 0$ es su Asíntota Horizontal
5. ¡Importante!: $F(-x) = 1 - F(x)$



el área bajo la curva es 1 → $\Phi(-x) = 1 - \Phi(x)$

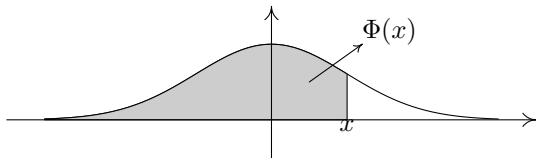


Simetrías en el cálculo de probabilidades en $N(0,1)$

**Tabla de la función de distribución Φ
de una normal $N(0, 1)$ para $x \geq 0$**

$$\Phi(x) = P[X \leq x] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{t^2}{2}\right) dt$$

Si $x < 0 \implies \Phi(x) = 1 - \Phi(-x)$



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511966	0.515953	0.519939	0.523922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555670	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621720	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802337	0.805105	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823814	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.879000	0.881000	0.882977
1.2	0.884930	0.886861	0.888768	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903200	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935745	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959070	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965620	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	0.998605
3.0	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999
3.1	0.999032	0.999065	0.999096	0.999126	0.999155	0.999184	0.999211	0.999238	0.999264	0.999289
3.2	0.999313	0.999336	0.999359	0.999381	0.999402	0.999423	0.999443	0.999462	0.999481	0.999499
3.3	0.999517	0.999534	0.999550	0.999566	0.999581	0.999596	0.999610	0.999624	0.999638	0.999651
3.4	0.999663	0.999675	0.999687	0.999698	0.999709	0.999720	0.999730	0.999740	0.999749	0.999758
3.5	0.999767	0.999776	0.999784	0.999792	0.999800	0.999807	0.999815	0.999822	0.999828	0.999835
3.6	0.999841	0.999847	0.999853	0.999858	0.999864	0.999869	0.999874	0.999879	0.999883	0.999888
3.7	0.999892	0.999896	0.999900	0.999904	0.999908	0.999912	0.999915	0.999918	0.999922	0.999925
3.8	0.999928	0.999931	0.999933	0.999936	0.999938	0.999941	0.999943	0.999946	0.999948	0.999950
3.9	0.999952	0.999954	0.999956	0.999958	0.999959	0.999961	0.999963	0.999964	0.999966	0.999967
4.0	0.999968	0.999970	0.999971	0.999972	0.999973	0.999974	0.999975	0.999976	0.999977	0.999978

Cálculo de probabilidades en $N(0,1)$

Como se observa, las probabilidades de que la variable z tome un determinado valor menor (o menor o igual) que k está tabulada en $[0, +\infty[$, en realidad hasta $k = 4.09$. Para valores mayores esta probabilidad es prácticamente 1 (con precisión hasta las millonésimas). Para calcular probabilidades para $z < 0$ o para z comprendida entre dos valores, $a < z < b$, usaremos las propiedades de simetría de la curva normal.

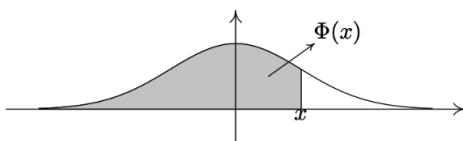
Ejemplo 4.12:

Directamente e las tablas, $P(z < 1,37) = 0.914657$, y a la inversa, $P(z < k) = 0.698468 \rightarrow k = 0.52$, como puede observarse en la siguiente figura.

Tabla de la función de distribución Φ de una normal $N(0,1)$ para $x \geq 0$

$$\Phi(x) = P[X \leq x] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{t^2}{2}\right) dt$$

$$\text{Si } x < 0 \implies \Phi(x) = 1 - \Phi(-x)$$



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511966	0.515953	0.519939	0.523922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555670	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621720	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802337	0.805105	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823814	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.879000	0.881000	0.882977
1.2	0.884930	0.886861	0.888768	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903200	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935745	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959070	0.959941	0.960796	0.961636	0.962462	0.963273

Ejercicio resuelto 4.4. En una $N(0,1)$, calcular las siguientes probabilidades:

- a) $p(z < 2.05)$; b) $p(z \geq 1.27)$; c) $p(z \leq -0.36)$; d) $p(z > -2.75)$
- e) $p(0.23 < z \leq 2.17)$; f) $p(-1.74 < z < -0.92)$; g) $p(-1.42 < z \leq 1)$

Nótese que en las tablas que estamos usando llaman a la función de distribución $F(z) = \Phi(z)$, usaremos esta notación.

Las desigualdades no importan que sean o no estrictas ($<$ ó \leq) ya que en v.a. continua la probabilidad de que la variable tome un valor concreto es cero:

$$p(z \leq k) = p(z < k) + p(z=k)^0; \quad p(z \geq k) = p(z > k); \quad p(a \leq z \leq b) = p(a < z < b)$$

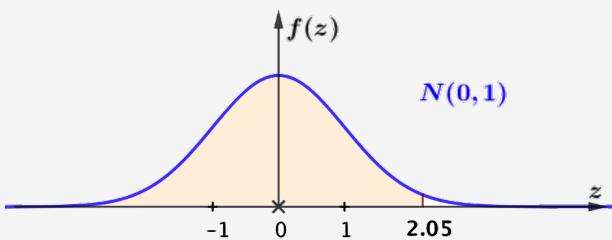
Acompañaremos cada cálculo con un dibujo y, de este modo, no hará falta memorizar nada. Representaremos por Φ el valor de la función de distribución, como en las tablas ($\Phi(z) = F(z)$).

Recordar: la curva (campana de Gauss) es simétrica y el área total que encierra es 1. Con la tabla, el cálculo de probabilidades se reduce al cálculo de áreas.

$$a) \quad p(z < 2.05) =$$

$$= \Phi(2.05) =$$

$$= 0.979818$$

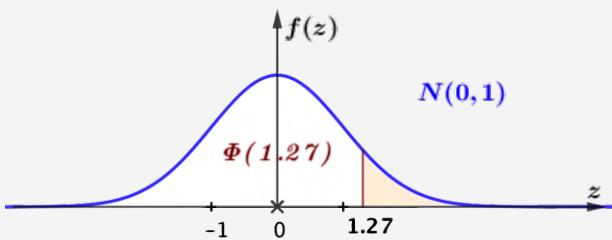


$$b) \quad p(z \geq 1.27) =$$

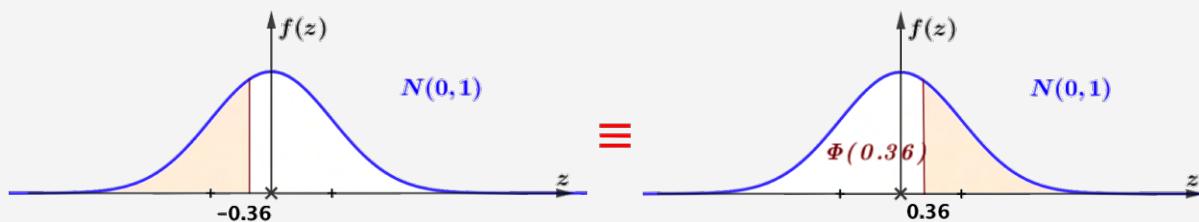
$$= 1 - \Phi(1.27) =$$

$$= 1 - 0.897958 =$$

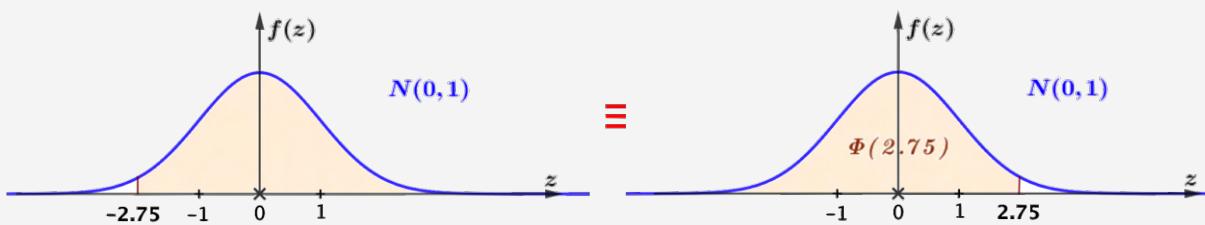
$$= 0.102042$$



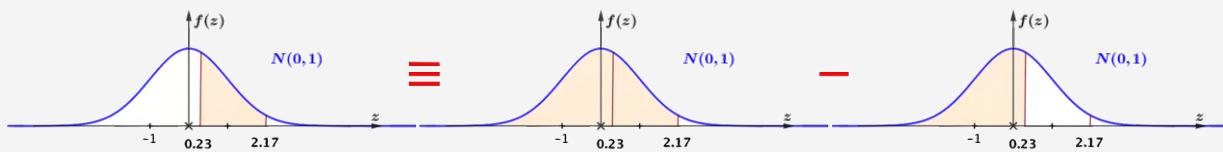
$$c) \quad p(z \leq -0.36) = 1 - \Phi(0.36) = 1 - 0.640576 = 0.359424$$



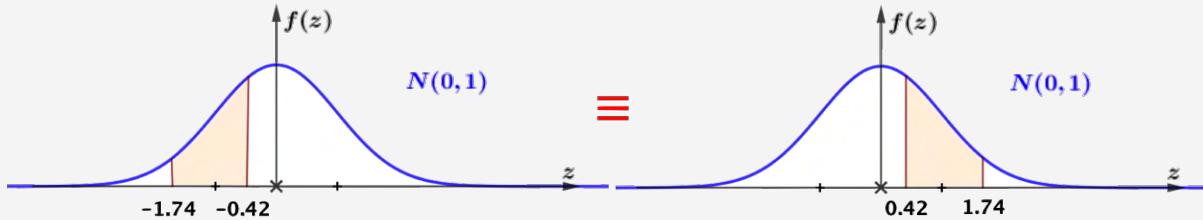
$$d) \quad p(z > -2.75) = \Phi(2.75) = 0.997029$$



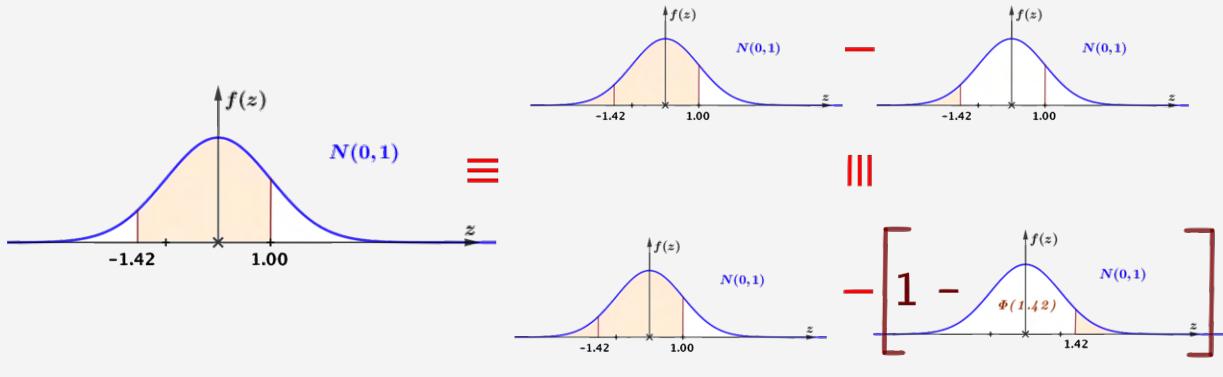
$$e) \quad p(0.23 < z < 2.17) = \Phi(2.17) - \Phi(0.23) = 0.984997 - 0.590954 = 0.394043$$



$$f) \quad p(-1.74 < z < -0.42) = \Phi(1.74) - \Phi(0.42) = 0.959070 - 0.662757 = 0.296313$$



$$g) \quad p(-1.42 < z \leq 1.00) = \Phi(1) - [1 - \Phi(1.42)] = 0.841345 - [1 - 0.922196] = 0.763541$$



Cálculo del valor de z a partir de su probabilidad asociada, uso de la tabla en sentido inverso.

La tabla normal se emplea también en sentido contrario para hallar la abscisa (el valor de z) correspondiente a una probabilidad determinada. Esto es, igual que se sabe que $P(z < 1) = 0.8413$, en sentido contrario la pregunta sería: ¿Cuánto debe valer z para que $P(z < k) = 0.8413$? La respuesta es evidente: el valor debe ser $k = 1$.

Si el valor de probabilidad no figura en la tabla tomaremos el más cercano. También se puede interpolar. Así, para $p(z < k) = 99.5\% = 0.9950$, tomaremos $z = 2.575$, intermedio entre 2.57 y 2.58, con valores de probabilidad respectivos son 0.9949 y 0.9951; k es el promedio de los dos valores.

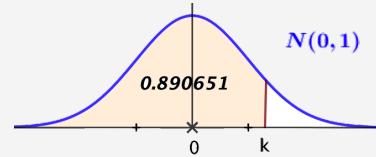
Ejercicio resuelto 4.5. Calcula los siguientes valores de k para que:

- a) $p(z < k) = 0.890651$; b) $p(z < k) = 0.359424$; d) $p(z > k) = 81\%$;
e) $p(z > k) = 0.15$; f) $p(-k < z < k) = 95\%$
-

a) $p(z < k) = 0.890651 > 0.5 \rightarrow k > 0$

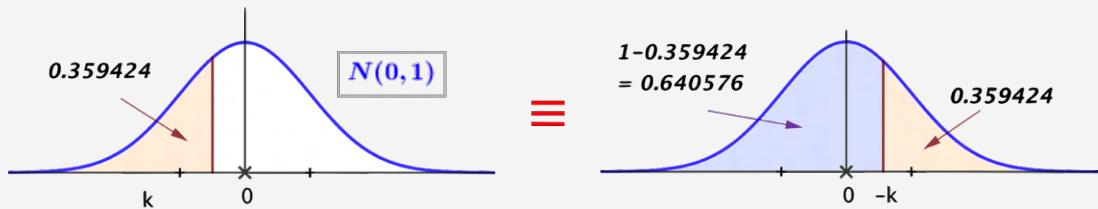
Leyendo en la tabla a la inversa:

$$\begin{array}{c} 3 \\ \hline 1.2 | \leftarrow 0.890651 \\ \uparrow \end{array}$$



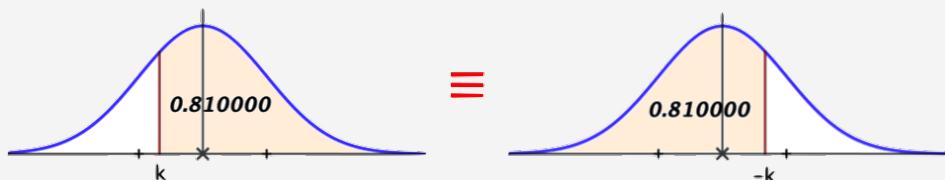
$$\Phi(k) = 0.890651 \rightarrow k = 1.23$$

b) $p(z < k) = 0.359424 < 0.5 \rightarrow k < 0$



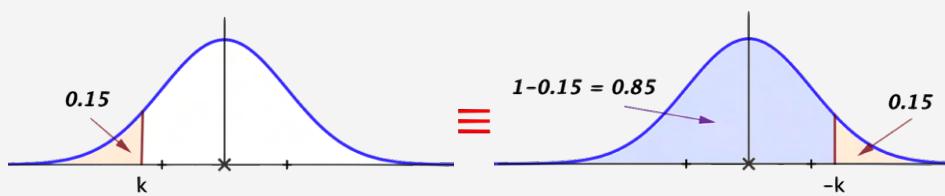
$$\Phi(-k) = 1 - 0.359424 = 0.640576 \quad (\text{leyendo la tabla al revés}) \rightarrow -k = 0.36 \rightarrow k = -0.36$$

c) $p(z > k) = 81\% = 0.810000 > 0.5 \rightarrow k < 0$



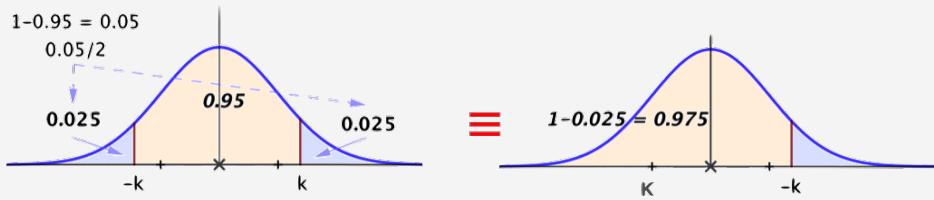
$$\Phi(-k) = 0.810000 \quad (\text{leyendo la tabla al revés}) \rightarrow -k = 0.88 \rightarrow k = -0.88$$

e) $p(z > k) = 0.150000 < 0.5 \rightarrow k < 0$

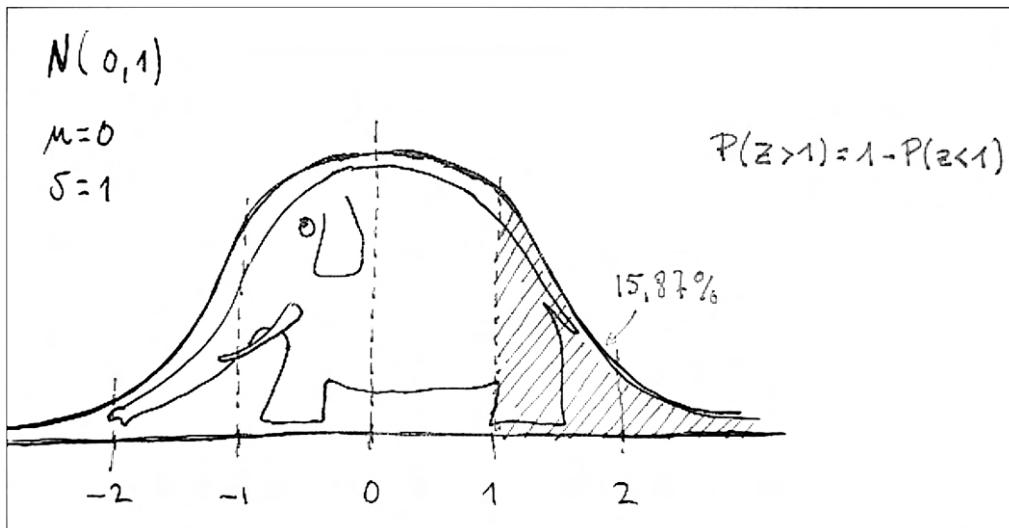


$$\Phi(-k) = 1 - 0.15 = 0.850000 \quad (\text{leyendo la tabla al revés}) \rightarrow -k = 1.04 \rightarrow k = -1.04$$

f) $p(-k < z < k) = 95\%$



$$(\text{leyendo la tabla al revés}) \quad \Phi(k) = 0.975000 \rightarrow k = 1.96$$



4.5.2. Cálculo de probabilidades en una $N(\mu, \sigma)$. Tipificación

Para poder calcular probabilidades en una $N(\mu, \sigma)$, hemos de transformarla en una $N(0, 1)$ de la que disponemos tablas. Para ello necesitamos:

- trasladar la $N(\mu, \sigma) \rightarrow N(0, \sigma)$
- contraer/dilatar para que $N(0, \sigma) \rightarrow N(0, 1)$

Todo ello se consigue, de un solo paso, con el siguiente cambio de variable (*tipificación*):

Definición 4.19:

$$x \in N(\mu, \sigma) \rightarrow z \in N(0, 1) :$$

$$z = \frac{x - \mu}{\sigma}$$

A este cambio de variable se le llama **tipificación**. Lo usamos para pasar de una $N(\mu, \sigma)$ a una $N(0, 1)$.

Con ello: $p(x < k) = p\left(z < \frac{k - \mu}{\sigma}\right)$; $p(x \geq k) = p\left(z \geq \frac{k - \mu}{\sigma}\right)$;

$$p(k_1 \leq x < k_2) = p\left(\frac{k_1 - \mu}{\sigma} \leq z < \frac{k_2 - \mu}{\sigma}\right); \quad \text{etc}$$
Ejemplo 4.13:

En una $N(7, 2.3)$, calcula $p(x < 8.45)$

$$x = 8.45 \rightarrow z = \frac{8.45 - 7}{2.3} = 0.63 > 0, \text{ leyendo directamente en las tablas,}$$

$$p(x < 8.45) = p(z < 0.63) = \Phi(0.63) = 0.735653 \approx 74\%$$

Ejercicio resuelto 4.6. En una $N(50, 5)$, calcula: $p(x < 56)$; $p(x > 48)$; $p(48 < x < 56)$
¿Cuál es el percentil 75 de esta distribución?

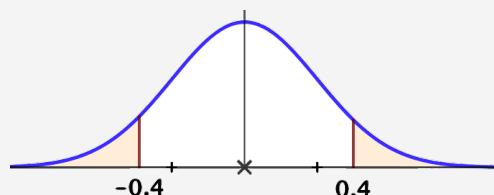
$$N(50, 5) \rightarrow \mu = 50; \sigma = 5 \rightarrow \text{tipificación: } z = \frac{x - 50}{5} \rightarrow N(0, 1)$$

- $p(x < 56) = p\left(z < \frac{56 - 50}{5}\right) = p(z < 1.2) = \Phi(1.2) = 0.884930$
- $p(x < 48) = p\left(z < \frac{48 - 50}{5}\right) = p(z < -0.4); z < 0, \text{ buscamos la situación simétrica:}$

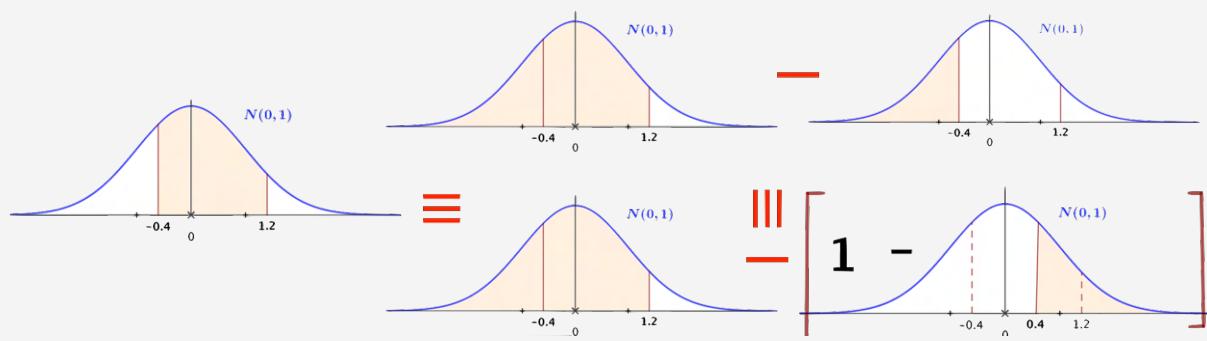
$$p(x < 48) = p(z < -0.4) = p(z > 0.4) =$$

$$= 1 - p(z < 0.4) = 1 - \Phi(0.4) =$$

$$1 - 0.655422 = 0.344578$$



$$— p(48 < x < 56) = (\text{tipificación}) = p\left(\frac{48 - 50}{5} < z < \frac{56 - 50}{5}\right) = p(-0.4 < z < 1.2)$$



$$p(48 < x < 56) = p(-0.4 < z < 1.2) = \Phi(1.4) - [\Phi(-0.4)] = 0.884930 - [1 - 0.655422] = 0.540352$$

— Percentil 75: $P_{75} = k \leftrightarrow P(x < k) = p\left(z < \frac{k-50}{5}\right) = 0.75 = 0.750000 \rightarrow$ (leyendo las tablas al revés) $\rightarrow P_{75} : \frac{k-50}{5} = 0.675 \rightarrow P_{75} = k = 53.375$

Ejercicio resuelto 4.7. Las alturas de 800 estudiantes se distribuyen normalmente con media 173 cm y desviación típica 11 cm. Aproximadamente, ¿cuántos alumnos tienen una altura)

a) De 175 cm; b) menor a 175 cm; c) entre 175 y 190 cm; d) ¿qué altura hay que tener para estar entre los 10 % más altos?

— a) $p(x = 175) = 0$ La probabilidad de que una v.a. continua tome un valor concreto es cero.

$$— b) p(x < 175) = p\left(z < \frac{175 - 173}{11}\right) = p(z < 0.18) = \Phi(0.18) = 0.571424$$

Del total de los 800 alumnos, $800 \cdot 0.571424 \approx 457$ de ellos miden menos de 175 cm .

$$— c) p(175 < x < 190) = p\left(\frac{175 - 173}{11} < z < \frac{190 - 173}{11}\right) = p(0.18 < z < 1.55) = \Phi(1.55) - \Phi(0.18) = 0.939429 - 0.571424 = 0.368005$$

Del total de los 800 alumnos, $800 \cdot 0.368005 \approx 294$ de ellos tienen alturas comprendidas entre 175 y 190 cm.

— Estar entre el 10 % de los más altos es lo mismo que dejar por detrás, en altura, al 90 % de los alumnos del centro; es decir, estar en el percentil 90: $P_{90} = k \leftrightarrow p(x < k) = 90\%$

$$p(x < k) = p\left(z < \frac{k - 173}{11}\right) = 0.900000 \rightarrow$$
 (leyendo las tablas al revés) $\rightarrow \frac{k - 173}{11} = 1.28 \Rightarrow k = 187.1\text{cm} = P_{90}$ Para estar entre el 10 % de los más altos de ese centro hay que medir más de 187.1cm.

Ejercicio resuelto 4.8. En una determinada ciudad, sus habitantes se distribuyen, en edades, como una media de 40 años. Se sabe que el 3,35 % de los habitantes de esa ciudad tienen más de 60 años.

¿Cuál es la desviación típica? ¿Qué porcentaje de la población es menor de 25 años?

$$a) p(x > 60) = p\left(z > \frac{60 - 40}{\sigma}\right) = 0.0335 \text{ (leyendo las tablas al revés)} \frac{20}{\sigma} = 1.835 \rightarrow \sigma = 10.9$$

$$\begin{aligned} b) \quad & \text{En una } N(40, 10.9), \quad p(x < 25) = p\left(z < \frac{25 - 40}{10 - 9}\right) = p(z < -1.38) = (\text{simetría}) \\ & = p(z > 1.38) = 1 - \Phi(1.38) = 1 - 0.916207 = 0.083793 \approx 8.37\% \end{aligned}$$

4.6. Aproximación de la Binomial por una Normal

Teorema 4.10:

El **Teorema central del límite** dice que si X_1, X_2, \dots, X_n son n v-a. independientes con medias $\mu_1, \mu_2, \dots, \mu_n$ y desviaciones típicas $\sigma_1, \sigma_2, \dots, \sigma_n$, entonces, la variable $Y = X_1 + X_2 + \dots + X_n$, “cuando **n es grande**”, se aproxima a una distribución **normal** con media $\mu_1 + \mu_2 + \dots + \mu_n$ y desviación típica $\sqrt{\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2}$

$$Y = X_1 + X_2 + \dots + X_n \longrightarrow N\left(\mu_1 + \mu_2 + \dots + \mu_n, \sqrt{\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2}\right)$$

Teorema 4.11:

Aproximación de la binomial a la normal

Por el teorema central de límite (anterior), ensayo es un éxito, y toma el valor cero si podemos considerar una variable X que es un fracaso, es decir:

sigue una distribución binomial $B(n, p)$, como una suma de n variables independientes X_i con media p y varianza $p(1 - p)$, pues suponemos que X_i toma el valor uno si el

x_i	p_i	$x_i p_i$	$x_i^2 p_i$
1	p	p	p
0	1-p	0	0
Σ	1	p	p

Calculamos la media y la desviación típica:

$$\mu = \sum_{i=1}^2 x_i p_i = 1 \cdot p + 0 \cdot (1 - p) = 0;$$

$$\sigma^2 = \sum_{i=1}^2 x_i^2 p_i - \mu^2 = 1^2 \cdot p + 0^2 \cdot (1 - p) - 0^2 = p - p^2 = p(1 - p) = p q$$

Por tanto, como consecuencia del teorema central del límite, y si n es grande, podemos aproximar una variable aleatoria X que sigue una distribución binomial $B(n, p)$ por una distribución normal cuya media y desviación típica son:

$$\mu = \mu_1 + \mu_2 + \dots + \mu_n = p + p + \xrightarrow{n-\text{veces}} + p = n p$$

$$\sqrt{\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2} = \sqrt{pq + pq + \xrightarrow{n-\text{veces}} + pq} = \sqrt{n p q}$$

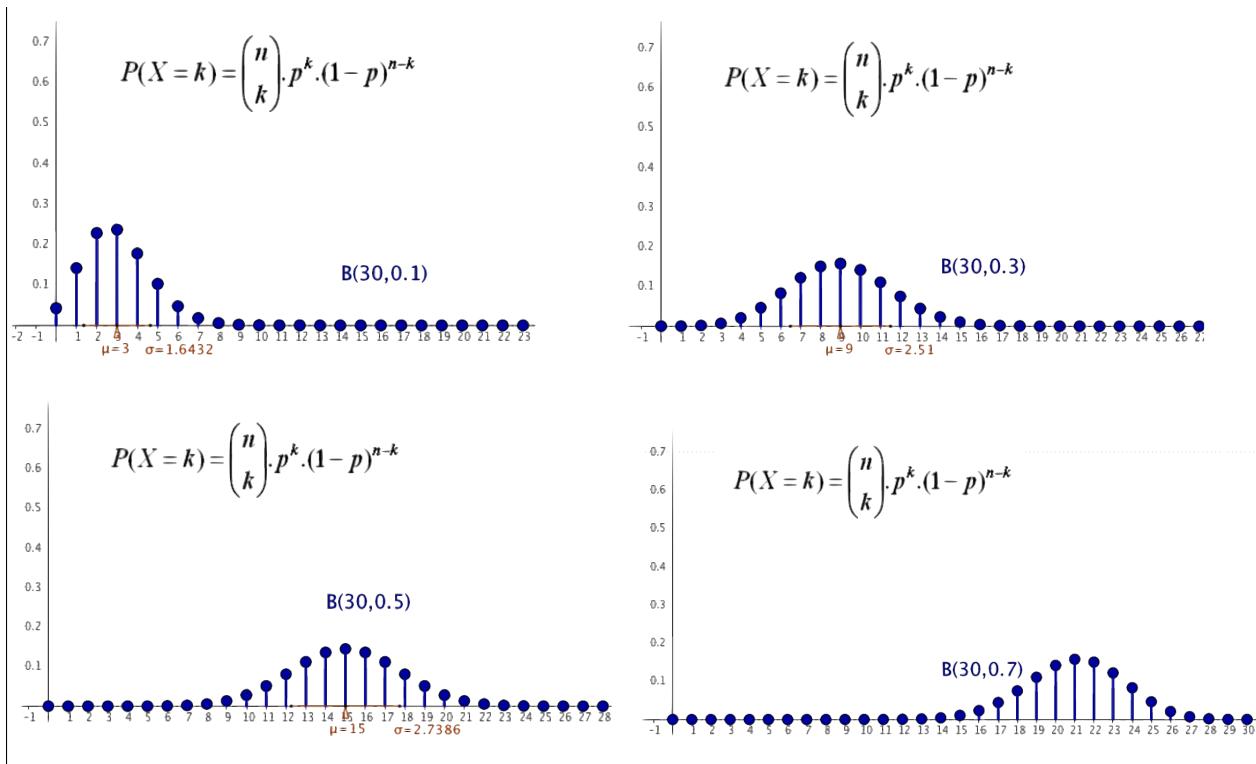
$$X \sim B(n, p) \quad n \text{ "grande"} \rightarrow X \sim N(\mu = np, \sigma = \sqrt{npq})$$

Validez de la aproximación de la binomial a la normal

Abraham de Moivre demostró que esta aproximación es buena si n es mayor que 30 y p no está próximo a cero ni a uno. En general se considera que la aproximación es buena si:

$$B(n, p) \approx N(np, \sqrt{npq}) \quad \text{si } n \geq 30 \wedge np \geq 5 \wedge n(1-p) \geq 5$$

Si no se cumplen estas condiciones NO se puede usar la aproximación.



En el caso de que podamos aproximar la Binomial por una Normal, hemos de tener en cuenta que estamos pasando de una variable discreta (binomial), X , a una variable continua (normal), que llamaremos X' y es necesario efectuar una **"corrección por continuidad"** (Como $p(x = k)$ sí tiene sentido en v.a. discreta pero no en continua, $p(x' = k) = 0$, cuando hagamos la aproximación $B \approx N$ tomaremos para k el intervalo $x' \in]k - 0.5, k + 0.5[$)

Al igual que llamábamos x a la v.a. normal y z a la v.a. normal tipificada, ahora también haremos una distinción entre los nombres que damos a las variables: si x es la v.a. discreta, cuando aproximemos por una distribución de probabilidad de v.a. continua usaremos la variable x' .

Corrección por continuidad

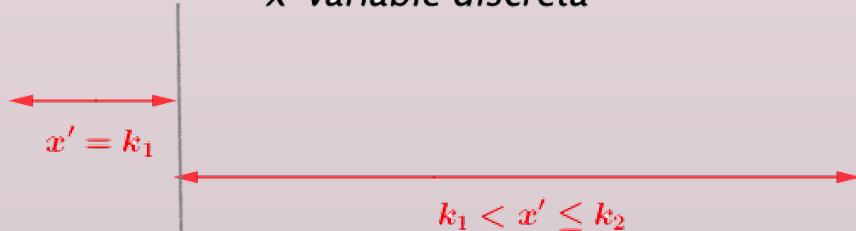
Es importante tener en cuenta que en el caso continuo la probabilidad asociada a un valor concreto de la variable es nula, $p(x = k) = 0$, por tanto, cuando se approxima una distribución discreta (x') a una distribución normal (x), que es continua, se debe utilizar la **corrección de continuidad** de Fisher o corrección de medio intervalo. Esta corrección consiste en considerar cada valor discreto x' como un intervalo de amplitud 1 de la forma $[x - 0.5, x + 0.5]$. Así pues:

- $p(x = k) = p(k - 0.5 < x' < k + 0.5)$
- $p(x \leq k) = p(x' < k + 0.5)$ (tomamos el valor de k)
- $p(x < k) = p(x' < k - 0.5)$ (no tomamos el valor de k)
- $p(x \geq k) = p(x' \geq k + 0.5)$ (tomamos el valor de k)
- $p(x > k) = p(x' > k - 0.5)$ (no tomamos el valor de k)
- $p(k_1 < x < k_2) = p(k_1 + 0.5 < x' < k_2 - 0.5); \quad p(k_1 < x \leq k_2) = p(k_1 + 0.5 < x' < k_2); \quad etc$

x variable continua



x variable discreta



$$B(n, p) \approx N(np, \sqrt{npq}) \quad \text{si} \quad n \geq 30 \wedge np \geq 5 \wedge n(1-p) \geq 5$$

$$x \in B(n, p) \rightsquigarrow x' \in N(\mu, \sigma) \rightsquigarrow z \in N(0, 1)$$

En la aproximación $B \approx N$, 1º) ‘corrección por continuidad’ y 2º) ‘tipificación de la variable’.

Ejemplo 4.14:

Es un hecho comprobado que cuando tenemos una distribución binomial $B(n; p)$, a medida que n crece, es difícil hacer uso de las fórmulas y/o tablas.

Por ejemplo, si tiramos un dado 500 veces, calcular la probabilidad de obtener entre 100 y 165 cincos (ambos inclusive).

Llamamos ‘éxito = obtener cinco’ entonces $p = 1/6$ y ‘fracaso = no obtener cinco’ y $q = 5/6$.

Tenemos una $B(500; 1/6)$, y nos piden $p(100 \leq x \leq 165)$.

Es inviable aplicar las tablas (pues repetimos el experimento 500 veces) y tampoco las fórmulas pues habría que hacer 65 cálculos difíciles, del tipo $p(x = 137) = \binom{500}{137} (1/6)^{137} (5/6)^{363}$

Afortunadamente, como $n \geq 30$, $np = 83.33 \geq 5$ y $nq = 8.33 \geq 5$, se puede aproximar la binomial por una normal:

$$\mu = np = 83.33; \sigma = \sqrt{npq} = 8.33 \rightarrow X \in B(500, 1/6) \approx X' \in N(83.33, 8.33)$$

$$\text{Así, } p(100 \leq x \leq 165) = (\text{corrección por continuidad}) = p(100 - 0.5 \leq x' \leq 165 + 0.5) = (\text{tipificación}) = p\left(\frac{99.5 - 83.33}{8.33} < z < \frac{165.5 - 83.33}{8.33}\right) = p(1.94 < z < 9.86) = \Phi(9.86) - \Phi(1.94) = (\text{con la precisión de las tablas que tenemos}) = 1 - 0.973810 = 0.02619$$

Ejercicio resuelto 4.9. Una moneda se lanza 400 veces. Calcula la probabilidad de que el número de caras: a) Sea mayor que 200; b) Esté entre 180 y 220, ambos inclusive.

Llamando ‘éxito’ al obtener cara al lanzar la moneda, $p = q = 0.5$ y realizamos el experimento $n = 400$ veces. Tenemos una $B(400, 0.5)$

$$- a) P(x > 200) = \sum_{k=201}^{400} \binom{400}{k} 0.5^k 0.5^{400-k}$$

$$- b) P(180 \leq x \leq 220) = \sum_{k=180}^{200} \binom{400}{k} 0.5^k 0.5^{400-k}$$

En ambos casos hay que calcular muchos términos, siendo todos ellos bastante complicados.

Como $n = 400 > 30$ y $nq = pq = 200 \cdot 0.5 = 100 \geq 5$, afortunadamente podemos usar la aproximación normal de esta binomial: $\mu = np = 200$; $\sigma = \sqrt{nqp} = \sqrt{100} = 10$

$$B(400, 0.5) \approx N(200, 10)$$

En ambos casos hemos de hacer la corrección por continuidad, así,

$$p(x > 200) = p(x' > 200 + 0.5) \quad y \quad p(180 \leq x \leq 220) = p(179.5 < x' < 220.5)$$

$$\begin{aligned}
 - a) \quad p(x > 200) &= p(x' > 200.5) = \text{(tipificación)} = p\left(z > \frac{200.5 - 200}{10}\right) = \\
 &p(z > 0.05) = 1 - \Phi(0.05) = 1 - 0.519939 = 0.419939 \\
 - b) \quad p(180 \leq x \leq 220) &= p(179.5 < x' < 220.5) = \text{(tipificación)} \\
 &= p\left(\frac{179.5 - 200}{10} < z < \frac{220.5 - 200}{10}\right) = p(-2.05 < z < 2.05) = 2\Phi(2.05) - 1 = \\
 &2(0.979818) - 1 = 0.8040364 \quad (El/la lector/a debería hacer los dibujos correspondientes)
 \end{aligned}$$

Ejercicio resuelto 4.10. Cierta enfermedad tiene una probabilidad muy baja de ocurrir, $p = 1/100\,000$. Calcular la probabilidad de que en una ciudad con 800 000 habitantes haya más de 5 personas con dicha enfermedad. Calcular el número esperado de habitantes que la padecen.

$$B(800000, 0.00001) \rightarrow p(x > 5) = 1 - p(x \leq 5) = 1 - \sum_{k=0}^5 \binom{800000}{k} 0.00001^k 0.99999^{800-k},$$

cálculos excesivos.

Se cumplen las condiciones para aproximar por una normal (compruébese, $np = 8$), por lo que haciendo primero la corrección por continuidad y después la tipificación de la variable, tenemos:

$$\mu = np = 800000 \cdot 0.00001 = 8; \quad \sigma = \sqrt{npq} = 2.83$$

$$B(800000, 0.00001) \approx N(8, 2.83)$$

$$p(x > 5) = p(x' > 5.5) = p\left(x' > \frac{5.5 - 8}{2.83}\right) = p(z > -0.88) = \text{(simetría)} = p(z < 0.88) = \Phi(0.88) = 0.810570 \quad (El/la lector/a debería hacer los dibujos correspondientes)$$

Para la $B(800000, 0.00001)$, el valor esperado de habitantes que padecen la enfermedad es $E(X) = \mu = np = 8$.

4.7. Ejercicios

Ejercicio 4.1. De los estudiantes universitarios españoles, uno de cada 5 abandona sus estudios. Se seleccionan 5 estudiantes universitarios españoles al azar, de modo independiente

a) ¿Cuál es la probabilidad de que uno o ninguno de dichos estudiantes abandonen sus estudios?

b) ¿Qué es más probable, que todos abandonen sus estudios, o que ninguno lo haga? Razone la respuesta de modo numérico.

$$p = p(abandono) = 1/5 = 0.2 : \text{"éxito"}; \quad q = 1 - p = 0.8; \quad B(5, 0.2)$$

$$p(x \leq 1) = p(x = 0) + p(x = 1) = \binom{5}{0} 0.2^0 0.8^5 + \binom{5}{1} 0.2^1 0.8^4 = 0.7373$$

$$p(x = 0) = \binom{5}{0} 0.2^0 0.8^5 = 0.8^5 > p(x = 5) = \binom{5}{5} 0.2^5 0.8^0 = 0.2^5$$

Luego: $p(\text{no abandone ninguno}) > p(\text{abandonen todos})$

Ejercicio 4.2. Un juego de ruleta tiene 25 casillas numeradas del 1 al 25. Un jugador gana si sale 2 o múltiplo de 2.

a) Si juega 100 veces, calcule la probabilidad de que gane exactamente 50 veces.

b) Si juega 200 veces, calcule la probabilidad de que gane entre 90 y 110 veces, ambos valores incluidos.

— a) 1, 2, 3, ..., 24, 25 → 12 de 25 son pares

$$p = p(\text{"ha salido par"}) = 12/25 : \text{'éxito'}; \quad q = 1 - p = 13/25; \quad B(100, 12/25)$$

$$p(x = 50) = \binom{100}{50} (12/25)^{50} (13/25)^{50}, \text{ cálculo excesivo.}$$

Como $n = 100 > 30$; $np = 100 \cdot 12/25 = 48 \geq 5 \wedge nq = 100 \cdot 13/25 = 52 \geq 5$, usaremos la aproximación de la binomial por la normal.

$$\mu = np = 48; \quad \sigma = \sqrt{npq} = 5 \quad B(100, 12/25) \approx N(48, 5)$$

$$p(x = 50) = p(49.5 < x' < 50.5) \stackrel{(1)}{=} p\left(\frac{49.5 - 48}{5} < z < \frac{50.5 - 48}{5}\right) \stackrel{(2)}{=} p(0.3 < z < 0.5) =$$

$$\Phi(0.50) - \Phi(0.30) = 0.691462 - 0.617911 = 0.073551$$

Donde, en (1) hemos hecho la ‘corrección por continuidad’ y en (2) hemos la ‘tipificación de la variable’: $x \in B(100, 12/25) \rightsquigarrow x' \in N(48, 5) \rightsquigarrow z \in N(0, 1)$

— b) Si juega 200 veces → $n = 200; \mu = 200 \cdot 12/25 = 96; \sigma = \sqrt{npq} = 7.07 \quad B(200, 12/25) \approx N(96, 7.07)$

$$p(90 \leq x \leq 110) = p(89.5 < x' < 110.5) = p(-0.92 < z < 2.05) = \Phi(2.05) - [1 - \Phi(0.92)] = 0.979818 - [1 - 0.821214] = 0.801032$$

También, en este caso, hemos hecho la corrección por continuidad y, posteriormente, la tipificación de la variable.

Ejercicio 4.3. La probabilidad de que una persona escriba un mensaje de Twitter sin faltas de ortografía es 0,75. Se sabe además que una persona escribe a lo largo del día 20 mensajes de Twitter.

A partir de esta información, responde a las siguientes cuestiones.

a) ¿Cuál es la probabilidad de que exactamente la mitad de los mensajes escritos en un día, es decir 10, no tengan faltas de ortografía?

b) ¿Cuál es la probabilidad de que ningún mensaje de los 20 escritos en un día tenga faltas de ortografía?

c) ¿Cuál es la probabilidad de que 18 o más mensajes de los 20 escritos en un día sí tengan faltas de ortografía?

Convengamos en llamar éxito a ‘no cometer faltas de ortografía en un tuit’, así, $p = 0.75$; $q = 0.25$. Escribimos 20 tuits, por lo que estamos frente a una $B(20, 0.75)$

$$\text{— a)} p(x=10) = \binom{20}{10} 0.75^{10} 0.25^{10} = 0.0099$$

— b) No cometer ninguna falta en los 20 tuits supone tener 20 éxitos:

$$p(x=20) = \binom{20}{20} 0.75^{20} 0.25^0 = 0.0032$$

— c) Que 18 o más de los 20 tuits tengan faltas de ortografía es tener solo 2, 1 o 0 éxitos, así:

$$p(x \leq 2) = p(2) + p(1) + p(0) = \binom{20}{2} 0.75^2 0.25^{18} + \binom{20}{1} 0.75^1 0.25^{19} + \binom{20}{0} 0.75^0 0.25^{20} = 1.61 \times 10^{-9} \approx 0$$

Ejercicio 4.4. El 30 % de los habitantes de un determinado pueblo ve un concurso de televisión. Desde el concurso se llama por teléfono a 10 personas del pueblo elegidas al azar. Calcular la probabilidad de que, de las 10 personas elegidas, estuvieran viendo el concurso de televisión:

a) Tres o menos personas.

b) Ninguna de las 10 personas a las que se ha llamado.

Si suponemos un pueblo con muchos habitantes ($p \approx \text{cte}$) tenemos una distribución binomial $B(10, 0.3)$, donde éxito es que la persona elegida estuviese viendo el concurso de TV.

$$\text{— a)} p(x \leq 3) = p(3) + p(2) + p(1) + p(0) = \sum_{k=0}^3 \binom{10}{k} 0.3^k 0.7^{10-k} = 0.6496$$

$$\text{— b)} p(x=0) = \binom{10}{0} 0.3^0 0.7^{10} = 0.0282$$

Ejercicio 4.5. Un estudiante universitario de matemáticas ha comprobado que el tiempo que le cuesta llegar desde su casa a la universidad sigue una distribución normal de media 30 minutos y desviación típica 5 minutos.

a) ¿Cuál es la probabilidad de que tarde menos de 40 minutos en llegar a la universidad?

b) ¿Cuál es la probabilidad de que tarde entre 20 y 40 minutos?

c) El estudiante, un día al salir de su casa, comprueba que faltan exactamente 40 minutos para que empiece la clase ¿Cuál es la probabilidad de que llegue tarde a clase?

Tenemos una $N(30, 5)$:

- a) $p(x < 40) = p\left(z < \frac{40 - 30}{5}\right) = p(z < 2) = \Phi(2) = 0.9772$
- b) $p(20 < x < 40) = p(-2 < z < 2) = 2\phi(2) - 1 = 0.9544$
- c) $p(x > 40) = p(z > 2) = 1 - \Phi(2) = 0.0288$

Ejercicio 4.6. Las calificaciones de un examen en una clase siguen una distribución normal de media $\mu = 20$ y desviación típica $\sigma = 10$: Calcula:

- a) La probabilidad de que un alumno obtenga una calificación entre 15 y 25.
- b) La calificación que sólo superan o igualan el 20 % de los alumnos.

$N(20, 10)$

— a) $p(15 \leq x \leq 25) = p\left(\frac{15 - 20}{10} < z < \frac{25 - 20}{10}\right) = p(-0.5 < z < 0.5) = 2\Phi(0.5) - 1 = 0.3830$

— b) $p(x \geq k) = 0.2 \Leftrightarrow p(x < k) = 0.8 \rightarrow$ tipificando, $p\left(\frac{k - 20}{10}\right) = 0.8$,

leyendo las tablas al revés, $\frac{k - 20}{10} = 0.84 \rightarrow k = 28.4$

Ejercicio 4.7. El peso de un grupo de personas sigue una distribución normal de media 54.3 kg i desviación típica de 6.5 kg.

- a) ¿Cuál es el porcentaje de personas con peso superior a 57 kg?
- b) ¿Qué porcentaje de personas pesan entre 50 i 57 kg?
- c) Si se elige una persona al azar que está dentro del 70 % de las personas que menos pesan, com máximo, ¿cuántos kilos debería pesar?

$N(54.3, 6.5)$

— a) $p(x > 57) = p\left(z > \frac{57 - 54.3}{6.5}\right) = p(z > 0.42) = 1 - p(z < 0.42) = 1 - \Phi(0.42) = 0, .372$

— b) $p(50 < x < 57) = p\left(\frac{50 - 54.3}{6.5} < z < \frac{57 - 54.3}{6.5}\right) = p(-0.66 < z < 0.42) =$

$$\Phi(0.42) - \Phi(-0.66) = \Phi(0.42) - [1 - \Phi(0.66)] = 0.4082$$

— c) $p(x < k) = p\left(z < \frac{k - 54.3}{6.5}\right) < 0.7000$, leyendo la tabla al revés,

$$\frac{k - 54.3}{6.5} = 0.525 \rightarrow k = 57.7 \text{ kg.}$$

Ejercicio 4.8. El número de horas de vida de una determinada bacteria (tipos A) se distribuye según una normal de media 110 horas i desviación típica de 0.75 horas. Calcula la probabilidad que, eligiendo al azar una bacteria:

- a) su número de horas de vida sobrepase las 112.25 horas.
- b) su número de horas de vida sea inferior a 109.25 horas.
- c) De otra bacteria (tipo B) se sabe que el número de horas de vida se distribuye según una normal de media 110 horas, pero se desconoce su desviación típica. Experimentalmente se ha comprobado que la probabilidad que una bacteria tipos B viva más de 125 horas es 0.1587. Calcula la desviación típica de la distribución del número de horas de vida de las bacterias tipo B.

$$N_A(110, 0.75)$$

— a) $p(x > 112.25) = p\left(z > \frac{112.25 - 110}{0.75}\right) = p(z > 3) = 1 - p(z < 3) = 1 - \Phi(3.00) = 0.0013$

— b) $p(x < 109.25) = p\left(z < \frac{109.25 - 110}{0.75}\right) = p(z < -1) = p(z > 1) = 1 - p(z < 1) = 1 - \Phi(1) = 0.1587$

— c) $N_B(110, \sigma_B) \rightarrow p(x > 125) = p\left(z > \frac{125 - 110}{\sigma_B}\right) = p\left(z > \frac{15}{\sigma_B}\right) = 0.1587$

$$p\left(z < \frac{15}{\sigma_B}\right) = 1 - p\left(z > \frac{15}{\sigma_B}\right) = 1 - 0.1587 = 0.8413$$

leyendo las tablas al revés, $\frac{15}{\sigma_B} = 1.00 \rightarrow \sigma_B = 15$

Ejercicio 4.9. Las notas de Matemáticas de 500 alumnos presentados al examen de EBAU tienen una distribución normal con media 6,5 y desviación típica 2.

- a) Calcule la probabilidad de que un alumno haya obtenido más de 8 puntos.
- b) ¿Cuántos alumnos obtuvieron notas menores de 5 puntos?
- c) ¿Qué nota hay que sacar para estar en el 25 % de las mejores notas?

$$N(6.5, 2); \quad n = 500$$

— a) $p(x > 8) = p\left(\frac{8 - 6.5}{2}\right) = p(z > 0.75) = 1 - p(z < 0.75) = 1 - \Phi(0.75) = 0.2266$

— b) $p(x < 5) = p\left(\frac{5 - 6.5}{2}\right) = p(z < -0.75) = p(z > 0.75) = 1 - p(z < 0.75) = 1 - \Phi(0.75) = 0.2266 ; \quad 0.2266 \cdot 500 \approx 113$ de los 500 alumnos tienen notas menores a 5.

— c) $p(x > k) = 25 \% \leftrightarrow p(x < k) = 75 \%$

$$p(x < k) = p\left(z < \frac{k - 6.5}{2}\right) = \Phi\left(\frac{k - 6.5}{2}\right) = 0.7500$$

leyendo las tablas al revés, $\frac{k - 6.5}{2} = 0.675 \rightarrow k = 7.85$

Ejercicio 4.10. Se estima que el 40 % de los alumnos que comienzan un grado de ingeniería acaban obteniendo el grado. Si se elige al azar a 5 alumnos que comenzaron una ingeniería, calcule:

- a) La probabilidad de que los 5 alumnos obtengan el grado de ingeniero.
- b) La probabilidad de que como máximo 2 obtengan el grado de ingeniero.
- c) La media y la desviación típica de la distribución.

$$B(5, 0.4)$$

— a) $p(x = 5) = \binom{5}{5} 0.4^5 0.6^0 = 0.01024$

— b) $p(x \leq 2) = p(0) + p(1) + p(2) = \sum_{k=0}^2 \binom{5}{k} 0.4^k 0.6^{5-k} = 0.68256$

— c) $\mu = np = 2; \quad \sigma = \sqrt{np(1-p)} = 1.09$

Ejercicio 4.11. Cierta tipo de batería dura un promedio de tres años, con una desviación típica de 0,5 años. Suponiendo que la duración de las baterías es una variable normal:

- a) ¿Qué porcentajes de las baterías se espera que duren entre 2 y 4 años?
- b) Si una batería lleva funcionando tres años, ¿cuál es la probabilidad de que dure menos de 4,5 años?

$$N(3, 0.5)$$

— a) $p(2 < x < 4) = (\text{tipificando}) = p(-2 < z < 2) = 2\Phi(2) - 1 = 0.9544$

El 95 %, aproximadamente, de las baterías duran entre 2 y 4 años.

— b) Nos preguntan por una **probabilidad condicionada**:

$$p(x < 4.5 | x > 3) = \frac{p((x < 4.5) \cap (x > 3))}{p(x > 3)} = \frac{p(3 < x < 4.5)}{p(x > 3)} =$$

$$\text{Tipificando, } = \frac{p(0 < z < 3)}{p(z > 0)} = \frac{\Phi(3) - \Phi(0)}{1 - \Phi(0)} = \frac{0.9987 - 0.5}{1 - 0.5} = 0.9974$$

La probabilidad de que dure menos de 4,5 años si lleva funcionando 3 años es de 99.74 %.

Ejercicio 4.12. Una gran empresa debe reponer las batas de sus 1000 operarios. Se sabe que la talla media es de 170 cm, con una desviación típica de 3 cm. Las batas se confeccionan en tres tallas válidas para estaturas entre 155 y 165 cm, 165 y 175 cm y, finalmente, entre 175 y 185 cm. ¿Cuántas batas de cada talla ha de adquirir?

Deberá adquirir 48 batas de talla pequeña, 905 de talla media y 48 de talla grande.

$$N(170.3)$$

— $p(155 < x < 165) = (\text{tipificando}) = p(-5 < z < -1.67) = p(1.67 < z < 5) = \Phi(5) - \Phi(1.67) = 1 - 0.9525 = 0.0475$

$0.0475 \cdot 1000 = 48$ batas pequeñas.

— $p(165 < x < 175) = p(-1.67 < z < 1.67) = 2\Phi(1.67) - 1 = 0.9050$

$0.9050 \cdot 1000 = 905$ batas medianas.

— $p(175 < x < 185) = p(1.67 < z < 5) = \Phi(5) - \Phi(1.67) = 1 - 0.9525 = 0.0475$

$0.0475 \cdot 1000 = 48$ batas grandes.

4.7.1. Problemas propuestos

PB. 1. $B(10, 0.4) \rightarrow p(x = 0); p(x = 3); p(x = 5); p(x = 10); \mu; \sigma$

0.00605; 0.215; 0.2007; 0.000105; 4; 1.55

PB. 2. $N(0, 1) \rightarrow p(z \leq 0.84); p(z < 1.5); p(z < 2); p(z < 1.87); p(z = 2.35); p(z \leq 0); p(z < 4); p(z = 1.23)$

0.7996; 0.9332; 0.9772; 0.9693; 0.9906; 0.5; 1; 0

PB. 3. $n(0, 1) \rightarrow p(z \leq k) = 0.7019; p(z > -k) = 0.8997$

0.53; 1.28

PB. 4. $N(173, 6) \rightarrow p(x < 173); p(x \geq 180.5); p(161 < x \leq 170); p(x = 171); p(x > 191); p(x < 155)$

0.5; 0.1056; 0.3269; 0.8716; 0.2857; 0; 0.0013; 0.0013

PB. 5. $B(100, 0.1) top(x = 10); p(x < 2); p(5 < x < 15)$

0, 0.0023; 0.8664

PB. 6. Se sabe que 2 de cada 8 habitantes de una ciudad utiliza el transporte público para ir a su trabajo. Se hace una encuesta a 140 de esos ciudadanos. Determinar:

- Número esperado de ciudadanos que no van a su trabajo en transporte público.
- Probabilidad de que el número de ciudadanos que van al trabajo en transporte público esté entre 30 y 45.

a) 35; b) 0.8374

PB. 7. Un estudio de un fabricante de televisores indica que la duración media de un televisor es de 10 años, con una desviación típica de 0,7 años. Suponiendo que la duración media de los televisores sigue una distribución normal:

- a) Calcula la probabilidad de que un televisor dure más de 9 años.
- b) Calcula la probabilidad de que un televisor dure entre 9 y 11 años.

a) 0.9222; b) 0.8444

PB. 8. En un examen, al que se presentaron 2 000 estudiantes, las puntuaciones se distribuyeron normalmente, con media 72 y desviación típica 9.

- a) ¿Cuántos estudiantes obtuvieron una puntuación entre 60 y 80?
- b) Si el 10% superior de los alumnos recibió la calificación de sobresaliente, ¿qué puntuación mínima habrá que tener para recibir tal calificación?

a) 1438 alumnos; b) 84 puntos

PB. 9. Se ha aplicado un test de fluidez verbal a 500 alumnos de primero de ESO de un centro de secundaria. Se supone que las puntuaciones obtenidas se distribuyen según una Normal de media 80 y desviación típica 12. Se pide:

- a) ¿Qué puntuación separa el 25% de los alumnos con mayor fluidez verbal?
- b) ¿A partir de qué puntuación se encuentra el 25% de los alumnos con mayor fluidez verbal?

a) 71.96; b) 88.04

PB. 10. Una persona que desea encontrar trabajo se presenta a dos entrevistas en las empresas A y B. En la entrevista de la empresa A obtiene una puntuación de 9, con una media de puntuación de 7 para la totalidad de los candidatos y una varianza de 4. En la entrevista de la empresa B obtiene una puntuación de 8, con una media de puntuación de 6 para la totalidad de los candidatos y una desviación típica de 1,5. ¿En qué entrevista ha obtenido esa persona una mejor puntuación relativa?

Esta mejor situado en la empresa B

En la empresa A, esta situado sobre un 84.13% de los candidatos.

En la empresa B, esta situado sobre un 90.82% de los candidatos.

Hubiese bastado con comparar las puntuaciones típicas.

- PB. 11. Tiramos una moneda perfecta 100 veces. Hacemos la predicción de que saldrán un número de caras comprendido entre 44 y 56. Calcula la probabilidad de no acertar.

0.1936

- PB. 12. El 90 % de los miembros de un club pasan sus vacaciones en la playa. Calcule una aproximación, obtenida utilizando la tabla normal, de la probabilidad de que, de un grupo de 60 miembros, 50 o menos vayan a ir a la playa a pasar sus vacaciones.

0.0655

- PB. 13. Se conoce, por estudios previos, que la proporción de reses que enfermarán después de suministrarles una determinada vacuna es del 2 %. Una granja tiene 600 reses que son vacunadas.

- a) Determina el número esperado de reses que no enfermarán.
- b) Halla la probabilidad de que el número de reses que enferman sea, como máximo, 20.
- c) Determina la probabilidad de que el número de reses que no enferman sea, como mínimo, 590.

a) 588; b) 0.9934; c) 0.3300

- PB. 14. La probabilidad de que deje de fumar un paciente, que se ha sometido a un régimen médico riguroso, es de 0,8. Se eligen 100 pacientes, que se han sometido a dicho régimen, ¿cuál es la probabilidad de que hayan dejado de fumar entre 74 y 85 pacientes, ambos inclusive?

0.8621

- PB. 15. Un alumno hace un examen tipo test que consta de 4 preguntas. Cada una de las preguntas tiene tres posibles respuestas, de las cuales solo una es correcta. Si un alumno aprueba contestando correctamente a dos o más preguntas, obtener de forma razonada la probabilidad de que apruebe si responde al azar a cada una de las preguntas.

- PB. 16. En una cierta prueba, el 35 por ciento de la población examinada obtuvo una nota superior a 6, el 25 por ciento entre 4 y 6, y el 40 por ciento inferior a 4. Suponiendo que las notas siguen una distribución normal, hállese la nota media y la desviación típica. ¿Qué porcentaje de la población tiene una nota que se diferencia de la media en menos de dos unidades?

$$N(4.79; 3.13) — 47.14\%$$

- PB. 17. Si una bombilla fluorescente presenta un 90% de posibilidades de tener una vida útil de al menos 800 horas, seleccionando 20 bombillas fluorescentes de este tipo, justificar si las siguientes afirmaciones son ciertas:
- Al seleccionar exactamente 18 bombillas fluorescentes, más del 30% tienen una vida útil de al menos 800 horas.
 - La probabilidad de que dos bombillas fluorescentes o menos NO tengan una duración de al menos 800 horas es menor que 0,7.
 - El valor esperado de bombillas con una vida útil de al menos 800 horas si se toma una muestra de 100 bombillas fluorescentes es igual a 10

$$E - V - F$$

- PB. 18. El consumo de azúcar en un determinado país, calculado en Kg por persona y año, varía según una distribución normal de media 15 y desviación típica 5.
- ¿Qué porcentaje de personas de ese país consumen menos de 10 Kg de azúcar al año? (1 punto)
 - ¿Cuál es el porcentaje de personas del país cuyo consumo anual de azúcar es superior a 25 Kg?

$$a) 15.87\%; b) 2.28\%$$

- PB. 19. Supongamos que en una población de Extremadura tienen una estatura que se distribuye según una normal de media 170 cm y desviación típica 10 cm.
- ¿Qué porcentaje de habitantes miden entre 170 y 185 cm?
 - ¿A partir de qué altura están el 33% de los habitantes más altos?

$$a) 43.32\%; b) 33\%$$

- PB. 20. En una determinada población de árboles, el 20 % tienen más de 30 años. Si se eligen 40 árboles al azar, calcule la probabilidad de que solamente 4 de ellos tengan más de 30 años.

¿Qué condición se debe exigir al problema para que se pueda considerar que la población de árboles siga una distribución binomial.

0.0475; La población de árboles debe ser muy numerosa

- PB. 21. En un bombo tenemos 10 bolas idénticas numeradas del 0 al 9 y cada vez que hacemos una extracción devolvemos la bola al bombo. Si hacemos 5 extracciones, calcula la probabilidad de que el 7 salga menos de dos veces.

0.08146

- PB. 22. Se sabe que dos poblaciones distintas X e Y se distribuyen según una Normal de media 25. Además $p(x \geq 27) = p(y > 30) = 0.1587$. Calcular sus respectivas varianzas.

4 y 25, respectivamente

- PB. 23. La distribución del número de rapes capturados por los barcos pesqueros que salen a faenar en una cierta zona se ajusta a una normal de media 220. Se sabe que, tomando un barco al azar la probabilidad de que capture más de 250 es 0.1587.

Calcula la desviación típica de la distribución, así como el número de rapes que un barco debe capturar para estar en el percentil 96.

30; 273 rapes

- PB. 24. La distribución del número de rapes capturados por los barcos pesqueros que salen a faenar en una cierta zona se ajusta a una normal de media 220. Se sabe que, tomando un barco al azar la probabilidad de que capture más de 250 es 0,1587. Calcula la desviación típica de la distribución. Calcula, también, el número de rapes que un barco debe capturar para estar en el percentil 96.

0.1815; 273 defectuosos

- PB. 25. El tiempo de duración de las bombillas de una cierta marca, medido en horas, sigue una distribución normal de media μ y desviación típica σ . Se sabe que el 69.50 % de las bombillas duran menos de 5061.2 horas, y que el 16.60 % de las bombillas duran

más de 5116.4 horas. ¿Cuál es la probabilidad de que una bombilla de esta marca dure entre 5061.2 y 5116.4 horas?

$$N(5000, 120) \rightarrow \text{probabilidad pedida es } 0.139$$

- PB. 26. Un tirador de dardos acierta ocho de cada diez lanzamientos. Utilizando la aproximación de la binomial a la normal, encuentra la probabilidad de que de 50 lanzamientos acierte 45.

$$0.0287$$

- PB. 27. La probabilidad de que un cierto tipo de piezas de una máquina sea defectuoso es del 6 %. En un almacén se han recibido 2000 piezas. ¿Cuántas habrá defectuosas por término medio? ¿Cuál será la desviación típica?

$$120; 10.6$$

- PB. 28. Se elige al azar una familia de seis hijos y se observa el número de hijos varones. Calcula la probabilidad de que la familia tenga:
- a) Dos hijos varones; b) Los mismos hijos que hijas. c) Alguna hija.

$$a) 0.2344; b) 0.3125; c) 0.9844$$

- PB. 29. La balanza de una frutería comete errores en cada pesada que se distribuyen según una normal de media 0 gramos y desviación típica 5 gramos:
- a) ¿Cuál es la probabilidad de que en una pesada la balanza marque un peso superior en 10 gramos al verdadero?
 - b) ¿Qué porcentaje de veces la balanza cometerá un error de más de 8 gramos a favor del cliente?
 - c) ¿Cuál es la probabilidad de que el error sea de menos de 8 gramos?

$$a) 0.0228; b) 0.0548; c) 0.8904$$

- PB. 30. Una universidad pública recibe 800 solicitudes de acceso para uno de los grados en los que la oferta de plazas se reduce a 120. Sabiendo que la nota final, de un solicitante, después de las pruebas de acceso sigue una distribución normal de media 7.3 y desviación típica 0.7, calcula la nota mínima para obtener una de las plazas ofertadas.

PB. 31. Una máquina que expende bebidas está regulada de modo que la cantidad de líquido que echa está distribuida normalmente con media de 200 ml y desviación típica de 15 ml

- a) ¿Qué porcentaje de los vasos se llenarán con más de 224 ml?
- b) Si usamos 6 vasos de 224 ml de capacidad, ¿cuál es la probabilidad de que se derrame líquido únicamente en uno de los vasos?

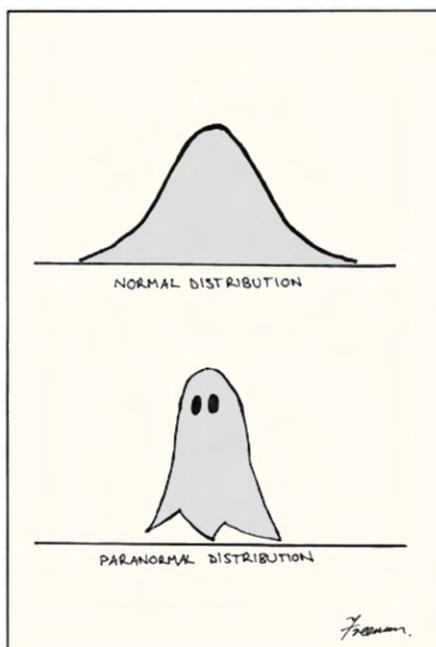
$$\text{a) } 0.0548; \text{ b) } B(6, 0.0548), P(x=1) = 0.2487$$

PB. 32. En el proceso de fabricación de una pieza intervienen dos máquinas: la máquina A produce un taladro cilíndrico y la máquina B secciona las piezas con un grosor determinado. Ambos procesos son independientes.

El diámetro del taladro producido en A, en mm, es $N(23,0.5)$ y el grosor producido por B es $N(11.5,0.4)$.

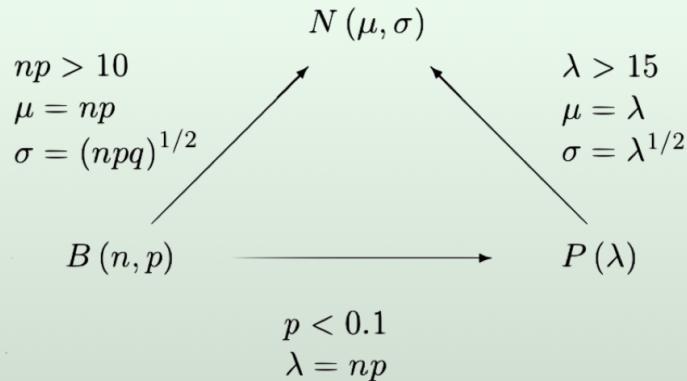
- a) Calcula el porcentaje de piezas que tienen un taladro de entre 20.5 y 24 mm.
- b) Encuentra el porcentaje de piezas que tienen un grosor de entre 10.5 y 12.7 mm.
- c) Suponiendo que solo son válidas las piezas cuyas medidas son las dadas en los apartados a) y b), calcula el porcentaje de piezas aceptables que se consiguen.

$$\text{a) } 0.9772; \text{ b) } 0.9925; \text{ c) independientes, } p(A \cup B) = p(A) \cdot p(B) = 0.9669; \text{ aprox. } 97\%$$



4.8. Curiosidades

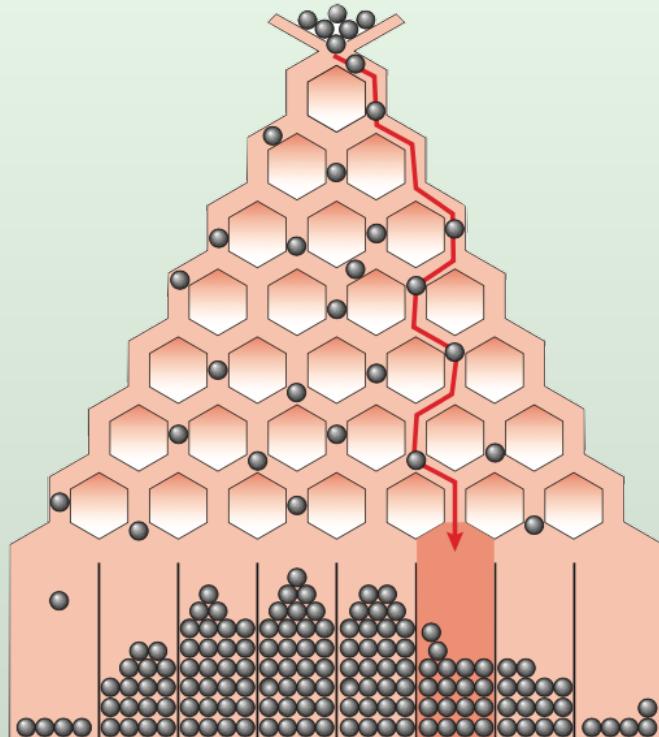
Aproximaciones entre distintas distribuciones



El aparato de Galton

El aparato, tablero o máquina de Galton es un dispositivo inventado por Francis Galton para demostrar el teorema del límite central, en particular que, con una muestra lo suficientemente grande, la distribución binomial se aproxima a la distribución normal.

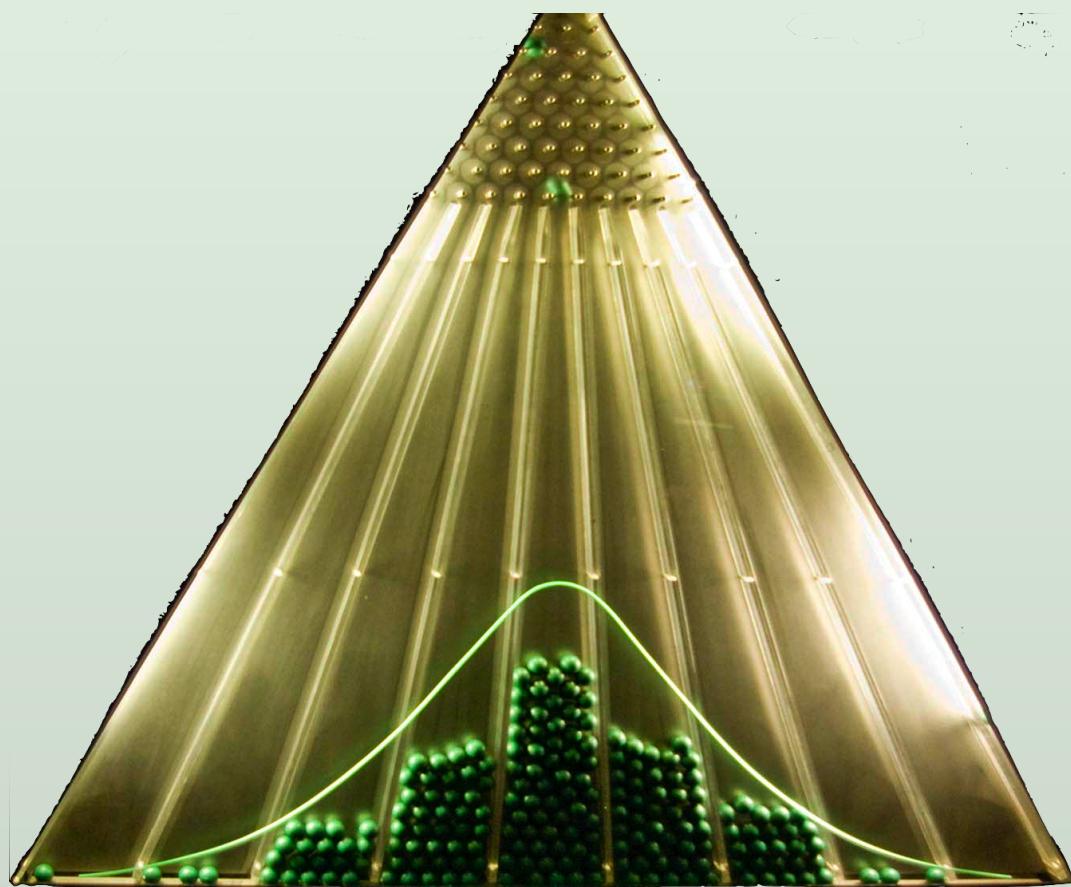
La máquina consta de un tablero vertical con varias filas de clavos. Las bolitas caen desde la parte superior, botando aleatoriamente y van depositándose, a medida que caen, en los casilleros de la parte inferior. Formando una superficie de campana.



La x cantidad de bolitas chocarán con el primer clavo teniendo una probabilidad de $1/2$ de ir a la izquierda o hacia la derecha, y a medida que continúan va teniendo más caminos a donde ir, es decir más posibilidades para que las bolitas se desvíen. A lo largo de esta estructura, las bolitas toman caminos aleatorios hasta caer en alguno de los canales colocados en la base

Si observamos el recorrido de una bola en el aparato de Galton . En cada bifurcación la bola puede ir a la izquierda con probabilidad p o a la derecha con probabilidad $q=1-p$. La variable aleatoria que toma valor 0 si cae a la izquierda o 1 si cae a la derecha se llama de Bernouilli y la variable X que da el número de unos al finalizar el experimento (lugares a la derecha) se denomina binomial. Los posibles valores de esta variable dependen del número de pisos que tiene el aparato de Galton.

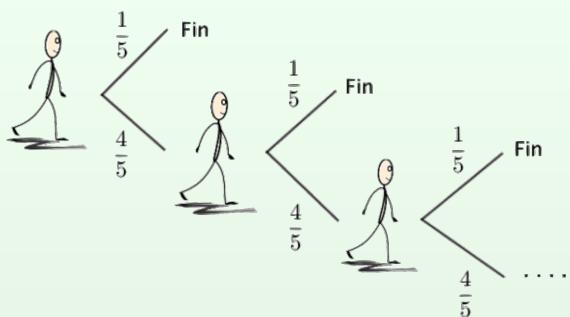
Si lanzamos una detrás de otra n bolitas al final del recorrido caerán en los distintos canales formando una determinada figura que si repites el experimento varias veces verás que caprichosamente se vuelve a formar casi la misma figura. Es decir estamos obteniendo una aproximación de la distribución de probabilidad. Al final, tendrán mayores probabilidades los canales interiores que los exteriores, formándose la conocida distribución normal.



* Suma de una serie con el valor esperado (sumar derivando)

Un señor avanza un paso y lanza una moneda al aire. Si sale cara se detiene y acaba el juego, si sale cruz da otro paso y vuelve a lanzar la moneda.

La probabilidad de cara es $1/5$ y la de cruz $4/5$.



Evidentemente, la probabilidad de que el Sr. dé n pasos es:

$$p(n) = (4/5)^{n-1} (1/5) = 4/4 (4/5)^{n-1} (1/5) = 1/4 (4/5)^n$$

Como en 1 de cada 5 ocasiones el juego se detiene, es de sentido común esperar que el 'número de pasos esperado' que dará el Sr. antes de pararse será 5. Calculémoslo:

$$E(n) = \sum_{n=1}^{\infty} n p(n) = \sum_{n=1}^{\infty} n \frac{1}{4} \left(\frac{4}{5}\right)^n = \frac{1}{4} \sum_{n=1}^{\infty} n \left(\frac{4}{5}\right)^n$$

Si, como hemos dicho, $E(n) = 5$, necesariamente, la serie ha de valer $\sum_{n=1}^{\infty} n \left(\frac{4}{5}\right)^n = 20$

Un caso más general de serie sería $S = \sum_{n=1}^{\infty} n a^n$, con $|a| < 1$.

La suma de una Progresión Geométrica de razón menor que uno en valor absoluto es:

$$S_{PG} = \sum_{n=1}^{\infty} a^n = \frac{1}{1-a}; \quad |a| < 1; \quad S_{PG} = S_{PG}(a)$$

Derivando esta expresión,

$$\text{— por una parte: } \frac{dS_{PG}}{da} = \sum_{n=1}^{\infty} n a^{n-1} \frac{a}{a} = \frac{1}{a} \sum_{n=1}^{\infty} n a^n = \frac{1}{a} S$$

$$\text{— por otra parte: } \frac{dS_{PG}}{da} = -\frac{1}{(1-a)^2} (-1) = \frac{1}{(1-a)^2}$$

$$\text{De ambas expresiones: } \frac{1}{(1-a)^2} = \frac{1}{a} S \rightarrow S = \frac{1}{(1-a)^2}$$

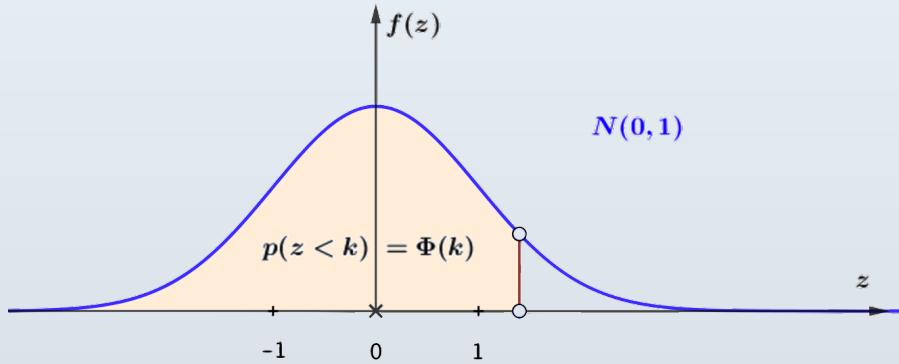
En nuestro caso, para $a = 4/5 \rightarrow S = 20$!!!!!

RESUMEN: Distribuciones de probabilidad

- ▷ Distribución binomial.

$$B(n, p) \quad : \quad p(X_k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- ▷ Distribución normal.



$$N(\mu, \sigma) \rightarrow \text{TIPIFICACIÓN } x \in N(\mu, \sigma) \rightarrow z = \frac{x - \mu}{\sigma} \in N(0, 1)$$

- ▷ La binomial se aproxima a la normal.

$$n \geq 30 \quad \wedge \quad np \geq 5 \quad \wedge \quad n(1-p) \geq 5$$

$$B(np) \approx N \left(\mu = np, \sigma = \sqrt{np(1-p)} \right)$$

$$x \in B(n, p) \xrightarrow{\text{Corrección por continuidad}} x' \in N(\mu, \sigma) \xrightarrow{\text{Tipificación}} z \in N(0, 1)$$

Parte III

Estadística Inferencial



Capítulo 5

Distribuciones muestrales. Estimación

5.1. Introducción

La estadística inferencial es la parte de la estadística que comprende los métodos y procedimientos que por medio de la inducción determina propiedades de una población estadística, a partir de una parte de esta, de una *muestra*. Su objetivo es obtener conclusiones útiles para hacer deducciones sobre una totalidad, *población*, basándose en la información numérica de la muestra.

Se dedica a la generación de los modelos, inferencias y predicciones asociadas a los fenómenos en cuestión teniendo en cuenta la aleatoriedad de las observaciones. Se usa para modelar patrones en los datos y extraer inferencias acerca de la población bajo estudio.

Método de la inferencia estadística:

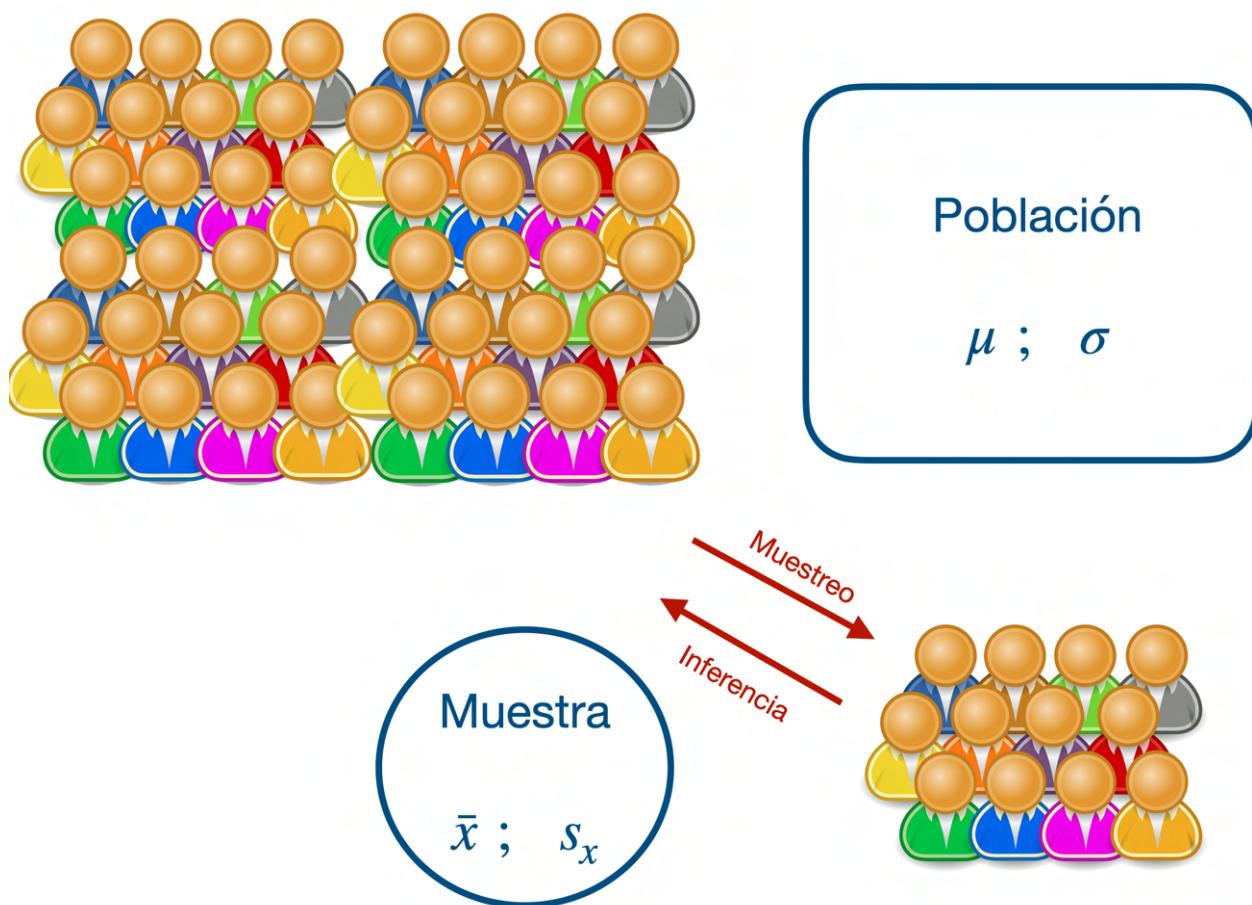
- **Planteamiento del problema:** un problema de inferencia estadística suele iniciarse con una fijación de objetivos o algunas preguntas del tipo: ¿Cuál será la media de esta población respecto a tal característica? ¿Se parecen estas dos poblaciones? ¿Hay alguna relación entre...? En el planteamiento se definen con precisión la población, la característica a estudiar, las variables, etc.
- **Elaboración de un modelo:** en caso de establecer un modelo teórico, se replantea el procedimiento y se llega a una conclusión lógica. Los posibles modelos son distribuciones de probabilidad.
- **Extracción de la muestra:** se usa alguna técnica de muestreo o un diseño experimental para obtener información de una pequeña parte de la población. *Teoría del muestreo*.

- **Tratamiento de los datos:** en esta fase se eliminan posibles errores, se depura la muestra, se tabulan los datos y se calculan los valores que serán necesarios en pasos posteriores, como la media muestral, la varianza muestral.

Los métodos de esta etapa están definidos por la estadística descriptiva (Tema 1) .

- **Estimación de los parámetros:** con determinadas técnicas se realiza una predicción sobre cuáles podrían ser los parámetros de la población.
- **Contraste de hipótesis:** los contrastes de hipótesis son técnicas que permiten simplificar el modelo matemático bajo análisis. Frecuentemente el contraste de hipótesis recurre al uso de estadísticos muestrales.
- **Conclusiones:** se critica el modelo y se hace un balance. Las conclusiones obtenidas en este punto pueden servir para tomar decisiones o hacer predicciones.

El estudio puede comenzar de nuevo a partir de este momento, en un proceso cíclico que permite conocer cada vez mejor la población y características de estudio.



En este tema, y el siguiente, se nos van a plantear las tres situaciones que mostramos en los siguientes ejemplos:

Situación 1. Las alturas de los alumnos de una facultad tiene media de $\mu = 175$ cm y desviación típica $\sigma = 5$. ¿Cuál es la probabilidad de que la estatura media de uno de los grupos de 30 alumnos de esa facultad esté entre 174 y 176 cm?

$$\mu = 175 \longrightarrow P[\bar{x} \in (174, 176)]$$

Situación 2. La estatura media de un grupo de alumnos de la facultad es de $\bar{x} = 175$ cm. ¿Cuál es la probabilidad de que la altura media de todos los alumnos de la facultad esté entre 174 y 176 cm?

$$\bar{x} = 175 \longrightarrow P[\mu \in (174, 176)]$$

Situación 3. Aseguramos que la media de las alturas de los alumnos de la facultad es $\mu = 175$ cm, para comprobarlo escogemos una muestra formada por 30 alumnos y calculamos su media que resulta ser de $\bar{x} = 176.3$ cm. ¿Es razonable admitir la hipótesis de que $\mu = 175$ cm?

- ▷ En la situación 1, a partir de la población podemos deducir el comportamiento de una muestra.
- ▷ En la situación 2, a partir de una muestra pretendemos inferir el valor de un parámetro de la población.
- ▷ En la situación 3, pretendemos contrastar una hipótesis acerca de la población a partir de la información obtenida de una muestra (tema 6).

5.2. Intervalos característicos

5.2.1. Intervalos característicos en $N(0,1)$

Definición 5.1:

En una $N(\mu, \sigma)$, llamamos **intervalo característico** de probabilidad p a un intervalo centrado en la media $(\mu - k, \mu + k)$, de modo que

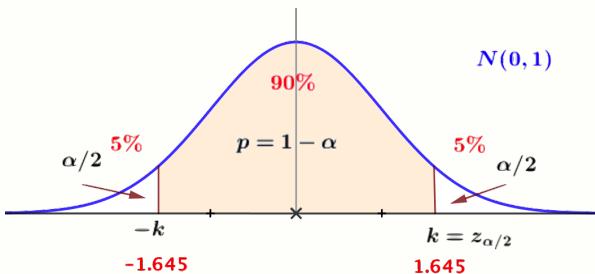
$$P[X \in (\mu - k, \mu + k)] = P(\mu - k < x < \mu + k) = p$$

p es el llamado **nivel de confianza** y se denota como $p = 1 - \alpha$, siendo α la **significación o riesgo**. k es el **valor crítico** del intervalo y se le suele llamar $k = z_{\alpha/2}$. A la amplitud del intervalo se le llama “margen de error”. Su mitad, la semiamplitud del intervalo, es el “error máximo admisible”.

Ejemplo 5.1:

En una $N(0,1)$, calcula el intervalo característico con un nivel de confianza del 90 %

(con una significación del
 $10\% = \alpha = 1 - p = 1 - 0.9$)



Una confianza del 90 % o una significación del 10 % supone unas colas en la distribución, a derecha e izquierda de la normal del 5 %. Buscamos pues: $p(-k < z < k) = 90\% \rightarrow p(z < k) = 95 \rightarrow$, mirando las tablas al revés, $k = z_{\alpha/2} = 1.645\%$ y el intervalo solicitado es $(-1.645, 1.645)$.

Teorema 5.1:

Los principales valores críticos en una normal $N(0, 1)$ son (convendría memorizarlos por su frecuente uso) .

Principales valores críticos $N(0,1)$		
$p = 1 - \alpha$ nivel de confianza	$\alpha/2$ significación	$z_{\alpha/2}$ valor crítico
0.90 (90 %)	0.050	1.645
0.95 (95 %)	0.025	1.960
0.99 (99 %)	0.005	2.575

Ejercicio resuelto 5.1. En una $N(0,1)$ calcula los intervalos característicos con un nivel de confianza del 95 % y del 99 %. (demostración del teorema anterior)

$$\text{--- } p = 95\% \rightarrow \alpha = 1 - p = 5\% \rightarrow \alpha/2 = 0.025$$

$$p(-k < z < k) = 95\% \rightarrow p(z < k) = 97.5\% = 0.9750 \rightarrow k = 1.96 : (-1.96, 1.96)$$

$$\text{--- } p = 99\% \rightarrow \alpha = 1 - p = 1\% \rightarrow \alpha/2 = 0.005$$

$$p(-k < z < k) = 99\% \rightarrow p(z < k) = 99.5\% = 0.9950 \rightarrow k = 2.575 : (-2.575, 2.575)$$



Ejercicio resuelto 5.2. Calcula el valor crítico ($k = z_{\alpha/2}$) para una significación $\alpha = 0.002$

$$\alpha = 0.002 \rightarrow \alpha/2 = 0.001 \rightarrow p(z < k) = 0.999 \Rightarrow k = 3.09$$

El intervalo característico sería $(-3.09, 3.09)$

5.2.2. Intervalos característicos en una $N(\mu, \sigma)$

Definición 5.2:

$X \sim N(\mu, \sigma)$, queremos encontrar un intervalo centrado en la media, $(\mu - k, \mu + k)$, de modo que

$$P(X \in (\mu - k, \mu + k)) = P(\mu - k < x < \mu + k) = p = 1 - \alpha ,$$

es decir, en el que esté el $p\% = (1 - \alpha)\%$ de la población. Este es el llamado **intervalo característico**.

Tipificando la variable, $Z = \frac{X - \mu}{\sigma}$, en la $N(0, 1)$, el intervalo característico correspondiente a un nivel de confianza $p = 1 - \alpha$ es $(-z_{\alpha/2}, z_{\alpha/2})$, ahora, en $N(\mu, \sigma)$:

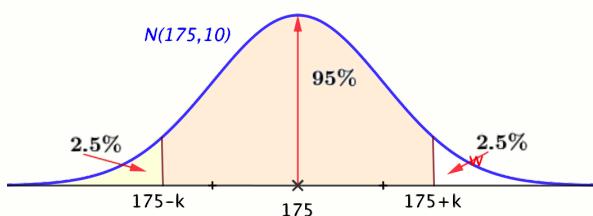
$$\begin{aligned} -z_{\alpha/2} < z < z_{\alpha/2} &\rightarrow -z_{\alpha/2} < \frac{X - \mu}{\sigma} < z_{\alpha/2} \rightarrow \\ &\rightarrow z_{\alpha/2} \cdot \sigma < X - \mu < z_{\alpha/2} \cdot \sigma \rightarrow \mu - z_{\alpha/2} \cdot \sigma < X < \mu + z_{\alpha/2} \cdot \sigma \end{aligned}$$

El intervalo característico es: $(\mu - z_{\alpha/2} \cdot \sigma, \mu + z_{\alpha/2} \cdot \sigma)$

Ejemplo 5.2:

En una $N(175, 10)$, calcula el intervalo característico para un nivel de confianza del 95 %.

$$p = 95\% \rightarrow \alpha = 5\% \rightarrow \alpha/2 = 2.5\% = 0.0025$$



$$p(175 - k < x < 175 + k) = 95\% \rightarrow p(x < 175 + k) = 97.5\%$$

Tipificando: $p(x < 175 + k) = p\left(\frac{175 + k - \mu}{\sigma} < \frac{k}{\sigma}\right) = p\left(\frac{k}{\sigma}\right) = 97.5\% = 0.9750$

Leyendo las tablas al revés: $\frac{k}{\sigma} = 1.96 \rightarrow k = 19.6$

El intervalo característico buscado es: $(175 - 19.6, 175 + 19.6) = (155.4, 194.6)$

Mucho más rápido, si hemos aprendido (memorizado) los valores críticos más usuales (90 %, 95 % y 99 %), tenemos $z_{\alpha/2}$ (95 %) = 1.96 y el intervalo buscado (intervalo de confianza al 95 %) es:

$$(\mu - z_{\alpha/2} \cdot \sigma, \mu + z_{\alpha/2} \cdot \sigma) = (175 \pm 1.96 \cdot 10) = (155.4, 194.6)$$

Ejercicio resuelto 5.3. Calcula el intervalo de confianza al 99 % en una $N(20, 3)$

Para $p = 99\% \rightarrow z_{\alpha/2} = 2.575 \rightarrow$

$$(\mu - z_{\alpha/2} \cdot \sigma, \mu + z_{\alpha/2} \cdot \sigma) = (20 \pm 2.575 \cdot \sqrt{3}) = (12.275, 27.725)$$

5.3. Estimación de las medias muestrales. Teorema central del límite

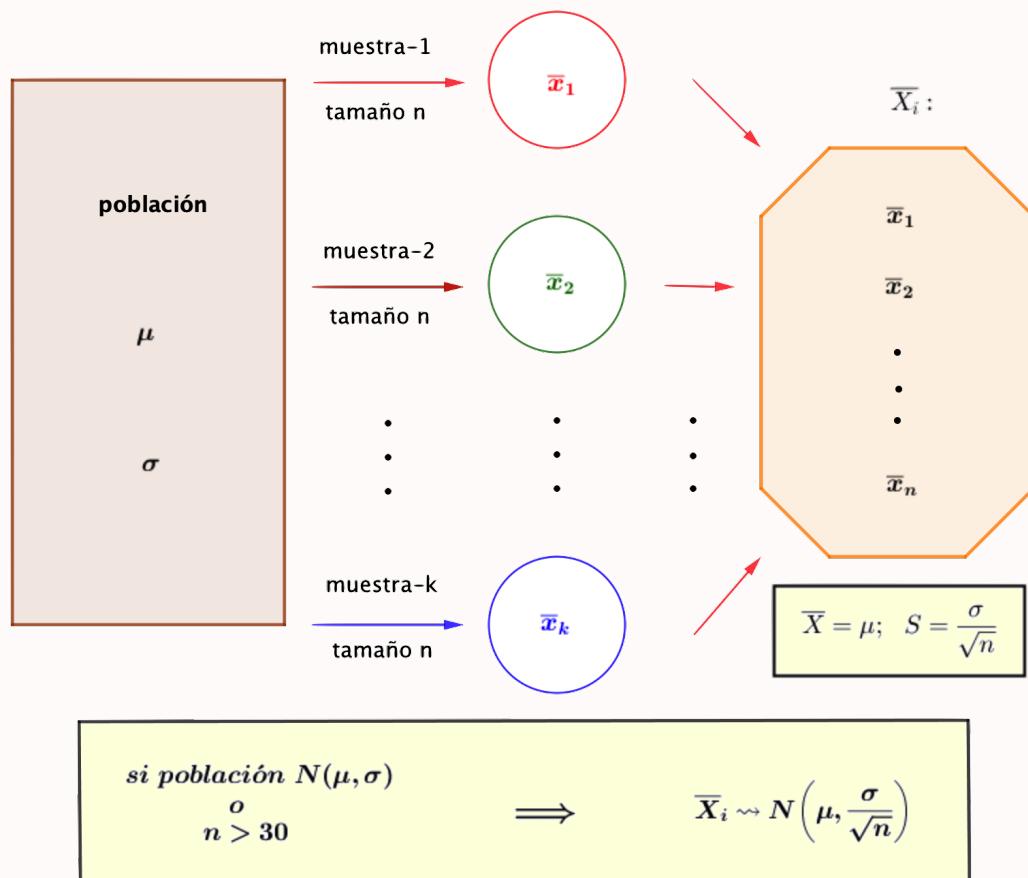
Teorema 5.2:

Dada una población de media μ y desviación típica σ , no necesariamente ‘normal’, tomamos todas las muestras de tamaño n y calculamos su media, \bar{X} .

Para cada muestra que extraigamos, calculamos su media \bar{x}_i . Considerándolas todas, tendremos una distribución de medias muestrales $X : \bar{x}_1, \bar{x}_2, \dots, \bar{x}_i, \dots, \bar{x}_j, \dots$. Su media es \bar{X} .

El **teorema central del límite** asegura que esta distribución de medias muestrales es tal que:

- Tiene la misma media que la población, $\bar{X} = \mu$.
- Su desviación típica depende del tamaño de las muestras, es $s_X = \frac{\sigma}{\sqrt{n}}$, disminuye al aumentar n .
- Si, o bien la población de partida es normal o bien el tamaño de las muestras en $n > 30$, la distribución de las medias muestrales es normal: $\bar{X} \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.



σ desconocida

Si σ (desviación típica poblacional) es desconocida pero $n \geq 30$, se toma como desviación típica de la población la “cuasi-desviación típica muestral”:

$$\sigma_{n-1} = \sqrt{\frac{n}{n-1}} s_n$$

siendo s_n la desviación típica muestral.

$$s_{n-1} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

es la cuasidesviación típica muestral (s_{n-1}^2 es la cuasivarianza muestral).

s_{n-1} es un estimador de la varianza poblacional σ_N (realmente lo es de la cuasivarianza, σ_{N-1} , pero prácticamente coinciden para $N \gg 1$). Al dividir por $n-1$ en vez de por n la cuasivarianza aumenta, es mayor que la varianza. Se trata, pues, de hacer una *sobreestimación* para la desviación típica poblacional para compensar el error de haber tomado una muestra.

Por tanto, la distribución de las medias muestrales, es una distribución de tipo “normal”, siempre que la población de procedencia lo sea, o incluso si no lo es, siempre que el tamaño de las muestras sea 30 o mayor.

La desviación típica de la distribución de medias, σ/\sqrt{n} , mide el grado de variabilidad de las medias muestrales. Cuanto menor sea, más ajustadas a la media de la población serán las medias que obtengamos de una muestra. De su propia definición, es fácil darse cuenta de que cuanto mayor es el tamaño de la muestra, menor es este grado de variabilidad, y por tanto más similar a la media de la población será la media obtenida de la muestra. Es decir, cuanto mayor es el valor de n , mejor es la aproximación “normal”.

Teorema 5.3:

Consecuencias del teorema central del límite:

► Control de las medias muestrales

Si $N \rightsquigarrow N(\mu, \sigma)$, podemos calcular la probabilidad de que una media determinada esté en un intervalo concreto.

► Control de la suma de individuos de una muestra

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} \rightarrow \sum_{i=1}^n x_i = n \cdot \bar{X} \text{ y tendremos que}$$

$$\text{si } N \rightsquigarrow N(\mu, \sigma) \rightarrow \sum_{i=1}^n x_i \rightsquigarrow N\left(n\mu, n\frac{\sigma}{\sqrt{n}}\right) \sim N(n\mu, \sigma\sqrt{n})$$

Podremos calcular la probabilidad de que la suma de todos los elementos de la muestra esté en un determinado intervalo.

► Inferir la media de la población a partir de la muestra

A partir de una muestra concreta podremos sacar conclusiones válidas para la población (esto lo veremos más adelante, en el apartado intervalo de confianza para la media').

Ejemplo 5.3:

Una compañía aérea sabe que el equipaje de sus pasajeros tiene como media 25 kg, con una desviación típica de 6 kg. Si uno de sus aviones transporta a 50 pasajeros, el peso medio de los equipajes de dicho grupo estará en la distribución muestral de medias:

$$\text{Población: } N(25, 6) \rightarrow \text{Muestra-avión: } N\left(25, \frac{6}{\sqrt{50}}\right) = N(25, 0.84)$$

La probabilidad de que el peso medio para estos pasajeros sea superior a 26 kg sería:

$$p(\bar{X} > 26) = p\left(z > \frac{26 - 25}{0.84}\right) = p(z < 1.18) = 0.1190 = 11.90\%$$

Ejemplo 5.4:

Si una población sigue una distribución $N(5; 0.5)$ y elegimos todas las muestras de tamaño 100.

¿Cuál es el valor de la suma de los elementos de la muestra? ¿A qué valor se aproxima la desviación típica?

La distribución de las sumas de los elementos de las muestras, como $n = 100 > 30$ y según el TCL (teorema central del límite), seguirá una distribución normal:

$$\sum x_i \in N(n\mu, \sigma\sqrt{n}) = N(5 \cdot 100, 0.5 \cdot \sqrt{100}) = N(500, 50)$$

Ejercicio resuelto 5.4. El peso, en Kg, de los soldados de un cuartel sigue una $N(69,8)$. Las guardias diarias se organizan mediante grupos de 12 soldados elegidos, cada día, al azar de entre los soldados del cuartel.

- a) Calcula la probabilidad de que la media de los pesos de los soldados de una determinada guardia sea mayor de 75 Kg.
- b) Encuentra el intervalo característico de la media de los pesos de la guardia para una probabilidad del 95 %.
- c) Calcula la probabilidad de que la suma de los pesos de los soldados de la guardia sea menor de 750Kg.
- d) ¿Cuál es la probabilidad de que un soldado de la guardia, elegido al azar, pese más de 75 kg?

Por el th. Central del límite, como la población es normal, las medias de los pesos de los soldados, así como la suma de ellos, de una guardia (muestra tamaño 12) también será normal:

$$\bar{X} \rightsquigarrow (N(69,8/\sqrt{12}) = N(69,2.31); \quad \sum X_i \rightsquigarrow N(12 \cdot 69,8 \cdot \sqrt{12}) = N(828,27.71)$$

$$\begin{aligned} - a) \quad \bar{X} \rightsquigarrow N(69,2.31) \rightarrow p(\bar{X} > 75) &= P\left(\frac{75 - 69}{2.31}\right) = p(z > 2.60) = \\ &= 1 - \Phi(2.60) = 0.0047 \end{aligned}$$

$$- b) \quad p = 95\% \rightarrow z_{\alpha/2} = 1.96 \rightarrow 69 \pm 1.96 \cdot 2.31 \rightarrow (64.47, 73.53)$$

$$\begin{aligned} - c) \quad \sum X_i \rightsquigarrow N(828,27.71) \rightarrow p\left(\sum x_i < 750\right) &= p\left(\frac{650 - 828}{27.71}\right) = \\ &= p(z < -2.81) = p(z > 2.81) = 1 - \Phi(2.81) = 0.0025 \end{aligned}$$

Un soldado de la guardia (12 soldados elegidos al azar), elegido al azar, x_i , es realmente un soldado del cuartel elegido al azar, un elemento de la población:

$$d) \quad N(69,8) \rightarrow p(x_i > 75) = p\left(\frac{75 - 69}{8}\right) = p(z > 0.75) = 1 - \Phi(0.75) = 0.2266$$

5.3.1. Estimación por intervalos: intervalos de confianza para la muestra

Definición 5.3:

Con el th. Central del límite hemos visto que, conocidos μ y σ de una población, podemos calcular la probabilidad de que la media de una muestra extraída de la población (o de la suma de todos sus elementos) esté en un determinado intervalo. Para que esto sea posible necesitamos que, o bien la población de partida sea Normal, o bien que el tamaño de la muestra sea $n \geq 30$.

Nuestra pregunta ahora es la contraria: tenemos una población de μ desconocida, pero de la que sí sabemos su σ y tenemos una muestra de tamaño n , en las condiciones de que la población sea normal o $n \geq 30$. Se puede demostrar que $\mu \approx \bar{x}$ y que $\sigma \approx s_{n-1}$. A esto se le llama **estimación puntual**.

ACLARACIÓN CALCULADORA ESTADÍSTICA: Introducidos los valores de nuestra distribución n ; x_i , n_i , con una calculadora, p.e., obtenemos los valores: \bar{x} , s_n , s_{n-1} . Pues bien:

- Si los datos pertenecen a la población: $\mu = \bar{x}$ y $\sigma = s_n$
- pero, si los datos pertenecen a una muestra: $\mu \approx \bar{x}$ y $\sigma \approx s_{n-1}$

La calculadora proporciona la *desviación típica* s_n y *cuasidesviación típica* s_{n-1} (varianza y cuasivarianza):

$$s_n = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}; \quad s_{n-1} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} \quad \rightarrow \quad s_{n-1} = \sqrt{\frac{n-1}{n}} s_n$$

s_{n-1}^2 es un estimador insesgado de σ_{N-1}^2 y no de σ_N^2 , pero no hay que preocuparse demasiado. a efectos prácticos es casi lo mismo cuando la población es grande. Para estimar la varianza de una población grande usaremos la cuasivarianza muestral.

El conocimiento de unos valores aproximados μ y de σ nos dice poco, para mayor precisión usaremos la **estimación por intervalos de confianza** para determinar estos parámetros.

Definición 5.4:

Estimación por intervalos.

De una población desconocida, extraemos una muestra de tamaño **n** y calculamos \bar{x} , s_n , s_{n-1} .

Con un nivel de confianza del $100 \mathbf{p\%} = 100(1 - \alpha)\%$, la media poblacional estará en el intervalo:

$$\mu \in \left(\bar{x} - z_{\alpha/2} \cdot \frac{s_{n-1}}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{s_{n-1}}{\sqrt{n}} \right)$$

Cuanto mayor sea el tamaño de la muestra (a mayor n), más eficacia de la estimación ya que el intervalo será más pequeño (margen de error menor) y mayor será el nivel de confianza (α menor).

Tamaño de la muestra (n), longitud del intervalo (margen de error) y nivel de confianza ($p = 1 - \alpha$) son tres variables relacionadas. Fijadas dos de ellos, podemos determinar la tercera.

Definición 5.5:

Intervalo de confianza para la muestra.

Tenemos una población con σ conocida, pero de la cual desconocemos μ . Deseamos encontrar un intervalo con un nivel de confianza $p = 1 - \alpha$ para la media μ . Tomamos una muestra de n elementos y media \bar{x} :

Al 100 $p\% = 100(1 - \alpha)\%$ de confianza, la media poblacional μ tendrá un valor comprendido en el intervalo:

$$\mu \in \left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

Si de la población desconocemos σ , tomamos s_{n-1} de la muestra.

Ejemplo 5.5:

Para estimar la media de las alturas de los alumnos de una universidad se talla, al azar, a 300 de ellos obteniendo una media de 173 cm y una cuasidesviación típica de 12.3 cm.

- Da una estimación puntual de la media y desviación típica de las alturas de los alumnos de la universidad.
- Determina un intervalo para la media de los alumnos de la universidad con una significación del 5 %.

$$n = 300; \quad \bar{x} = 173 \text{ cm}; \quad s_{n-1} = 12.3 \text{ cm}$$

- La estimación puntual para la media es $\mu = \bar{x} = 173$ cm y para la desviación típica $\sigma = s_{n-1} = 12.3$ cm
- Intervalo de confianza con significación $\alpha = 5\% = 1 - p$; $p = 95\%$, el valor característico es $z_{\alpha/2} = 1.96$.

Como desconocemos la desviación típica de la población, tomaremos como desviación típica poblacional la cuasidesviación típica muestral: $\sigma = s_{n-1} = 13.2$

$$\mu : \bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 173 \pm 1.96 \cdot \frac{12.3}{\sqrt{300}} = 173 \pm 1.4$$

El intervalo pedido es $\mu \in (171.6, 174.4)$

Ejemplo 5.6:

En esta misma universidad se sabe que la desviación típica de los pesos de sus alumnos es de 13 kg. Para estimar su media se ha tallado a 100 de ellos obteniendo una media de 66.5 kg.

— Determina un intervalo de confianza del 99 % para la media de los pesos de los alumnos de la universidad.

$$n = 100; \quad \bar{x} = 66.5 \text{ kg}; \quad \sigma = 13$$

Intervalo de confianza con significación $p = 99\%$ $\alpha = 1\% = 1 - p$, el valor característico es $z_{\alpha/2} = 2.575$.

$$\mu : \bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 66.5 \pm 2.575 \cdot \frac{13}{\sqrt{100}} = 66.5 \pm 3.35$$

El intervalo pedido es $\mu \in (63.15, 69.85)$

Ejercicio resuelto 5.5. *Para determinar la nota media del examen de matemáticas de los alumnos de segundo de bachillerato de España en un determinado año, se toma una muestra de 500 de ellos obteniéndose que la nota media de éstos es de 5.43. Determinar un intervalo de confianza al 90 % de la nota media de matemáticas de todos los alumnos de España sabiendo que:*

- a) la cuasivarianza de la nota de los 500 alumnos muestreados es 4.28.
- b) la desviación típica en el examen de matemáticas de todos los alumnos de España, en esta convocatoria, es 2.23.

$$n = 500; \quad \bar{x} = 5.43; \quad p = 90\% \rightarrow z_{\alpha/2} = 1.645$$

$$a) \quad \sigma = s_{n-1} = \sqrt{4.28} = 2.07; \quad b) \quad \sigma = 2.23$$

$$- a) \quad \mu : \bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 5.43 \pm 1.645 \cdot \frac{2.07}{\sqrt{500}} = 5.43 \pm 0.34$$

El intervalo pedido es $\mu \in (5.09, 5.77)$

$$— b) \mu : \bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 5.43 \pm 1.645 \cdot \frac{2.23}{\sqrt{100}} = 5.43 \pm 0.37$$

El intervalo pedido es $\mu \in (5.06, 5.80)$

Distribución muestral de la diferencia de muestras

Supongamos dos poblaciones que sigan distribuciones normales $N(\mu_x, \sigma_x)$ y $N(\mu_y, \sigma_y)$ de las que tomamos muestras de tamaños n_x y n_y , respectivamente (si los tamaños son mayores a 30, no hace falta exigir que las distribuciones sean normales), en estas condiciones:

La **distribución muestral de la diferencia de medias** sigue una distribución normal de media $\mu_x - \mu_y$ y de desviación típica $\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$, es decir:

$$N(\mu_x, \sigma_x) \rightarrow n_x; N(\mu_y, \sigma_y) \rightarrow n_y \rightarrow \bar{x} - \bar{y} \in N \left(\mu_x - \mu_y, \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \right)$$

La variable tipificada viene dada por la expresión:

$$z = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

Ejemplo 5.7:

El responsable de la sede central de una empresa afirma que las edades de sus empleados siguen una distribución normal con una media de 41 años y una desviación típica de 5 años. Por otro lado, el responsable de una sede de las sucursales de dicha empresa en otro país, ha determinado que sus empleados también tienen edades que se ajustan a una distribución normal con una media de 39 años y desviación típica de 3 años.

Con el fin de hacer un estudio comparativo se seleccionan muestras de 40 personas de cada sede de la empresa.

- a) Determina la distribución para la diferencia de las medias muestrales.
- b) ¿Cuál es la probabilidad de que los empleados de la sede central tengan una media de edad de al menos 3 años mayor que los de la sucursal extranjera?

$$— a) \bar{x} - \bar{y} \in N \left(\mu_x - \mu_y, \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \right) = N(41 - 39, \sqrt{(5^2/40) + (3^2/40)}) = N(2.00, 0.92)$$

$$— b) p(\bar{x} - \bar{y} \geq 3) = p(z < (3 - 2)/0.92) = p(z > 1.09) = 1 - \Phi(1.09) = 0.1379$$

Ejercicio resuelto 5.6. Para estudiar la influencia del tabaco en el peso de los recién nacidos se realiza un estudio en un hospital en que se consideran dos grupos de 300 futuras madres no fumadoras y 220 fumadoras. En el grupo de mujeres no fumadoras los recién nacidos presentan una media de pesos de 3.6 kg con una desviación típica de 0.5, en el grupo de las fumadoras la media de pesos es de 3.2 kg con desviación típica de 0.8 kg.

a) Determinar, con un nivel de confianza del 95 %, como influye que la madre sea fumadora en el peso de su hijo al nacer.

b) Determina el error máximo admisible cometido en la estimación anterior.

$$\bar{x} = 3.6 \text{ kg}; \sigma_x = 0.5; n_x = 300; \quad \bar{y} = 3.2 \text{ kg}; \sigma_y = 0.8; n_y = 220$$

$$\bar{x} \in N(3.6, 0.5); \quad \bar{y} \in N(3.2, 0.8) \Rightarrow$$

$$\bar{x} - \bar{y} \in N((3.6 - 3.2), \sqrt{(0.5^2/300) + (0.8^2/220)}) = N(0.400, 0.061)$$

$$1 - \alpha = 95\% \rightarrow z_{\alpha/2} = 1.96$$

$$IC(95\%) : (\bar{x} - \bar{y}) \pm \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \rightarrow 0.4 \pm 1.96 \cdot 0.061 = 0.4 \pm 0.124$$

$$IC(95\%) : (0.276, 0.524)$$

Con un nivel de confianza del 95 % el peso de un bebe de madre no fumadora supera en 276 g, como mínimo, y 524 g como máximo al de una madre fumadora.

El error máximo admisible es la mitad de la amplitud del intervalo (ya lo tenemos calculado):

$$E = z_{\alpha/2} \cdot \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} = 0.124 \text{ g.}$$

5.3.2. Relación entre nivel de confianza, error admisible y tamaño de la muestra

Definición 5.6:

Llamaremos **error admisible** a la mitad del margen de error (longitud del intervalo de confianza).

$$\text{Al } 100 p\% = 100 (1 - \alpha)\% \longrightarrow \mu \in \left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

El margen de error es $2 z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$, por lo que el *error admisible es*:

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \longrightarrow n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2$$

- Al aumentar el nivel de confianza, $p \uparrow \Rightarrow n \uparrow$
- Al disminuir el error admisible, $E \downarrow \Rightarrow n \uparrow$

Ejemplo 5.8:

La desviación típica de las alturas de los alumnos de una determinada universidad es de $\sigma = 6.17$. ¿Qué tamaño ha de tener la muestra a extraer para estimar la media de las alturas con un error menor de 0.5 cm y un nivel de confianza del 95 %?

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 1.96 \cdot \frac{6.17}{\sqrt{n}} < 0.5 \rightarrow \left(1.96 \cdot \frac{6.17}{0.5} \right)^2 < (\sqrt{n})^2 \rightarrow n > 584.982 \rightarrow n = 585 \text{ alumnos.}$$

Ejercicio resuelto 5.7. ¿Cuál es el error máximo admisible que cometemos al dar como intervalo de confianza (1.63, 1.93)?

El error máximo admisible es la mitad de la amplitud del intervalo (margen de error), así que: $E = \frac{1.93 - 1.63}{2} = 0.15$

5.3.3. Ejercicios de estimación de las medias muestrales

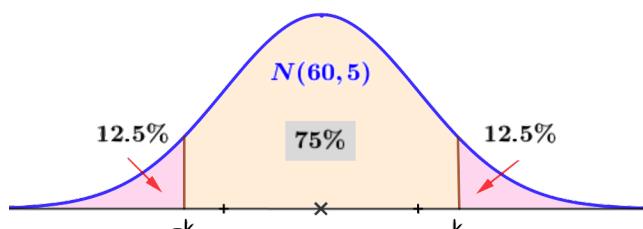
Ejercicio 5.1. En una $N(60, 5)$, determina los intervalos característicos al 75 %, al 95 % y al 99.8 %.

$$p = 75\% \rightarrow \alpha = 25\%$$

$$p(-k < z < j) = 0.75 \rightarrow p(z < z_{\alpha/2}) = 0.875 \rightarrow z_{\alpha/2} = 1.15$$

Intervalo característico al 75 %, IC(75 %):

$$\mu \pm z_{\alpha/2} \cdot \sigma \rightarrow 60 \pm 1.15 \cdot 5 \Rightarrow (54.25, 65.75)$$



$$\text{Para } p = 95\% \rightarrow z_{\alpha/2} = 1.96 \rightarrow IC(95\%) : 60 \pm 1.96 \cdot 5 \Rightarrow (50, 2, 69.8)$$

$$\text{Para } p = 99.8\% \rightarrow \alpha = 0.2\% \rightarrow z_{\alpha/2} : p(z < z_{\alpha/2}) = 0.999 \rightarrow z_{\alpha/2} = 3.09$$

$$IC(99.8\%) : 60 \pm 3.09 \cdot 5 \Rightarrow (44.55, 75.45)$$

Ejercicio 5.2. Los parámetros de una variable son: $\mu = 25$, $\sigma = 3$. Nos disponemos a extraer una muestra de $n = 100$ individuos.

- a) Halla el intervalo característico para las medias muestrales con un nivel de confianza del 95 %.
- b) Para esa muestra, calcula $p(24 < x < 26)$.
- c) Realiza el mismo resultado para la población, suponiendo que es normal.

$$\mu = 25; \sigma = 3 \rightarrow n = 100 : \bar{x} \in N\left(25, \frac{3}{\sqrt{100}}\right) = N(25, 0.3)$$

— a) $IC(95\%) : \bar{x} \pm z_{\alpha/2} \cdot \sigma / \sqrt{n} = 25 \pm 1.96 \cdot 0.3 \Rightarrow (19.12, 30.88)$

— b) $N(25, 0.3) : p(24 < \bar{x} < 26) = p\left(\frac{24 - 25}{0.3} < z < \frac{26 - 25}{0.3}\right) = p(-3.33 < z < 3.33) = \Phi(3.33) - \Phi(-3.33) = \Phi(3.33) - [1 - \Phi(3.33)] = 2\Phi(3.33) - 1 = 2(0.99955) - 1 = 0.9991$

— c) Suponemos la población normal:

$$N(25, 3) \rightarrow p(24 < \bar{x} < 26) = p\left(\frac{24 - 25}{3} < z < \frac{26 - 25}{3}\right) = p(-0.33 < z < 0.33) = 2\Phi(0.33) - 1 = 0.2586$$

Ejercicio 5.3. De una variable estadística, conocemos la desviación típica, $\sigma = 7.5$, pero desconocemos la media, μ . Para estimarla, extraemos una muestra de tamaño $n = 80$ cuya media obtenemos: $\bar{x} = 35$. Estima la media de la población mediante un intervalo de confianza del 99 %.

¿Cuál es el margen de error cometido? ¿Cuál es el error admisible?

Si queremos reducir el error admisible a la mitad manteniendo la misma significación, ¿cuál debe ser el tamaño de la muestra?

$$\sigma = 7.5; \mu = ? \rightarrow n = 80 : \bar{x} = 35 \rightarrow p = 99\%; \quad p = 99\% \rightarrow \alpha = 1\% \rightarrow z_{\alpha/2} = 2.576$$

$$IC(99\%), \sigma \text{ conocida} \quad \mu \in \bar{x} \pm z_{\alpha/2} \cdot \sigma / \sqrt{n} \rightarrow \mu \in 35 \pm 2.576 \cdot 7.5 / \sqrt{80} \Rightarrow (32.84, 37.16)$$

Margen de error cometido: $37.16 - 32.84 = 4.32$; Error admisible: $4.32/2 = 2.16$

Reducción del error admisible a la mitad aumentando el tamaño de la muestra:

$$E = 2.16/2 = 1.08; \quad p = 99\% \rightarrow z_{\alpha/2} = 2.576 \rightarrow E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$$1.08 = 2.576 \cdot \frac{7.5}{\sqrt{n}} \rightarrow n \geq \sqrt{\frac{2.576 \cdot 7.5}{1.08}} = 320.0123 \rightarrow n = 321$$

Ejercicio 5.4. Sabemos que la desviación típica de las notas de matemáticas las pruebas EBAU de los alumnos de la comunidad valenciana es 2.3. Queremos estimar la nota media de los alumnos de un instituto con un error menor a 1 punto. Para ello, tomamos una muestra de 10 alumnos del centro, seleccionados al azar, que se han presentado a la EBAU. ¿Con qué nivel de confianza podremos realizar la estimación?

$$\sigma = 2.3; \quad \bar{x} = ?; \quad E = 1 \text{ punto}; \quad n = 10$$

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \rightarrow z_{\alpha/2} = \frac{E \sqrt{n}}{\sigma} = \frac{1 \cdot \sqrt{10}}{2.3} = 1.375$$

$$p(z < 1.375) = 0.9154 \rightarrow \alpha/2 = 1 - 0.9154 = 0.0846 \rightarrow \alpha = 2 \cdot 0.0846 = 0.1692 \rightarrow p = 1 - \alpha = 1 - 0.1692 = 0.8308$$

El nivel de confianza es del 83 % (aproximadamente).

Ejercicio 5.5. En una distribución $N(25, 8)$, tomamos muestras de tamaño 64.

a) ¿Cuál es la distribución de las medias de las muestras?

b) ¿Cuál es la probabilidad de extraer una muestra cuya media esté comprendida entre 24 y 26?

$$— \text{ a) } N(25, 8) \rightarrow n = 64 > 30 \rightarrow \bar{x} \in N(25, 8/\sqrt{64}) = N(25, 1)$$

$$— \text{ b) } p(24 < \bar{x} < 26) = p\left(\frac{24 - 25}{1} < z < \frac{26 - 25}{1}\right) = p(-1 < z < 1) = \Phi(1) - \Phi(-1) = \Phi(1) - [1 - \Phi(1)] = 2\Phi(1) - 1 = 2(0.8413) - 1 = 0.6826$$

Ejercicio 5.6. Se sabe que el cociente intelectual de los alumnos de una universidad se distribuye según una ley normal de media 100 y varianza 729.

a) Halla la probabilidad de que una muestra de 81 alumnos tenga un cociente intelectual medio inferior a 109.

b) Halla la probabilidad de que una muestra de 36 alumnos tenga un cociente intelectual medio superior a 109.

$$\sigma^2 = 729; \quad \sigma = 27; \quad \mu \in N(100, 27) \rightarrow n = 81 > 30 \rightarrow \bar{x} \in N(100, 27/\sqrt{81}) = N(27, 3)$$

$$— \text{ a) } p(\bar{x} < 109) = p(z < (109 - 100)/3) = p(z < 3) = 0.99875$$

$$— \text{ b) } n = 36 \rightarrow \bar{x} \in N(100, 27/\sqrt{36}) = N(27, 1.8)$$

$$p(\bar{x} < 109) = p(z < (109 - 100)/1.8) = p(z < 0.5) = 0.6915$$

Ejercicio 5.7. En una muestra de 50 jóvenes, encontramos que la dedicación media diaria de ocio es de 400 minutos y su desviación típica de 63 minutos. Calcula el intervalo de confianza de la media de la población al 95 % de nivel de confianza.

$n = 50; \bar{x} = 400; s = 63$ IC(95 %) para μ :

$$IC(95\%) \rightarrow z_{\alpha/2} = 1.96 \rightarrow \mu \in 400 \pm 1.96 \cdot \frac{63}{\sqrt{50}} \Rightarrow \mu \in 3491(382.54, 417.46)$$

Ejercicio 5.8. Las notas en un cierto examen se distribuyen normal con media $\mu = 5.3$ y desviación típica $\sigma = 2.4$. Halla la probabilidad de que un estudiante tomado al azar tenga una nota:

- a) Superior a 7.
- b) Inferior a 5.
- c) Comprendida entre 5 y 7.

Tomamos al azar 16 estudiantes. Halla la probabilidad de que la media de sus notas:

- d) Sea superior a 7.
- e) Sea inferior a 5.
- f) Esté comprendida entre 5 y 7.
- g) La suma de sus notas sea mayor que 100.
- h) Halla k para que el intervalo $(5.3 - k; 5.3 + k)$ contenga al 95 % de las notas.
- i) Halla b para que el intervalo $(5.3 - b; 5.3 + b)$ contenga al 95 % de las notas medias de las muestras de 16 individuos.

$$N(5.3, 2.4)$$

— a) $p(x > 9) = p(z > (9 - 5.3)/2.4) = p(z > 0.71) = 1 - \Phi(0.71) = 0.2388$

— b) $p(x < 5) = p(x < (5 - 5.3)/2.4) = p(<< -0.13) = p(z > 0.13) = 1 - \Phi(0.13) = 0.4483$

— c) $p(5 < x < 7) = p(-0.13 < z < 0.71) = p(z < 0.71) - p(z < -0.13) = p(z < 0.71) - p(z > 0.13) = p(z < 0.71) - [1 - p(z < 0.13)] = \Phi(0.71) + \Phi(0.13) - 1 = 0.3129$

$$n = 16 \rightarrow N(5.3, 2.4/\sqrt{16}) = N(5.3, 0.6)$$

— d) $p(\bar{x} > 7) = p(z > (7 - 5.3)/0.6) = p(z > 2.83) = 1 - Phi(2.83) = 0.0023$

— e) $p(\bar{x} < 5) = p(z < (5 - 5.3)/0.6) = p(z < -0.5) = 1 - Phi(0.5) = 0.3085$

— f) $p(5 < \bar{x} < 7) = p(-0.5 < z < 2.83) = \Phi(2.83) + \Phi(0.5) - 1 = 0.6892$

$$\sum x_i \in N(16 \cdot 5.3, 2.4 \cdot \sqrt{16}) = N(84.8, 9.6)$$

- g) $p\left(\sum_{i=1}^{16} x_i > 100\right) = p(z > (100 - 84.8)/9.6) = p(z > 1.58) = 1 - \Phi(1.58) = 0.0571$
- h) $(5.3 - k, 5.3 + k) \supset 95\% N(5.3, 2.4) : k = z_{\alpha/2} \cdot \sigma = 1.96 \cdot 2.4; k = 4.704$
- i) $(5.3 - b, 5.3 + b) \supset 95\% N(5.3, 0.6) : b = z_{\alpha/2} \cdot \sigma = 1.96 \cdot 0.3; k = 1.176$

Ejercicio 5.9. El peso de los paquetes recibidos en un almacén se distribuye normal con media 300 kg y desviación típica 50 kg. ¿Cuál es la probabilidad de que 25 de los paquetes, elegidos al azar, excedan el límite de carga del montacargas donde se van a meter, que es de 8200 kg?

$$N(300, 50); n = 25 : \sum x_i \in N(300 \cdot 25, 50 \cdot \sqrt{25}) = N(7500, 250)$$

$$p(\sum x_i > 8200) = p(z > (8200 - 7500)/250) = p(z > 2.8) = 1 - \Phi(2.8) = 0.0026$$

Ejercicio 5.10. Se supone que el peso medio de las sandías de cierta variedad sigue una distribución normal con $\mu = 6$ kg y $\sigma = 1$ kg. Si empaquetamos las sandías en cajas de 8 unidades:

- a) Halla la probabilidad de que la media de los pesos de las sandías de una caja sea menor que 5.5 kg.
- b) Calcula la probabilidad de que entre las 8 sandías de una de las cajas pesen más de 50 kg.

$$N(6, 1); n = 8 \rightarrow \bar{x} \in N(6, 1/\sqrt{8}) = N(1, 0.35); \sum x_i \in N(6 \cdot 8, 1 \cdot \sqrt{8}) = N(48, 2.83)$$

$$— a) p(\bar{x} < 5.5) = p(z < (5.5 - 6)/0.35) = p(z < -1.43) = 1 - \Phi(1.43) = 0.0764$$

$$— b) p(\sum x_i > 50) = p(z > (50 - 48)/2.83) = p(z > 0.71) = 1 - \Phi(0.71) = 0.2389$$

Ejercicio 5.11. Se ha tomado una muestra aleatoria de 100 individuos a los que se ha medido el nivel de glucosa en sangre, obteniéndose una media muestral de 110 mg/cm³. Se sabe que la desviación típica de la población es de 20 mg/cm³.

- a) Obtén un intervalo de confianza, al 90 %, para el nivel de glucosa en sangre en la población.
- b) ¿Cuál es el error máximo que se ha cometido?

$$n = 100; \bar{x} = 110; \sigma = 20$$

$$\mu \in N(110, 20/\sqrt{100}) = N(110, 2)$$

$$— a) p = 90\% \rightarrow \alpha = 10\% \rightarrow p(z < z_{\alpha/2}) = 0.95 \rightarrow z_{\alpha/2} = 1.96$$

$$IC(90\%) : 110 \pm 1.96 \cdot \frac{20}{\sqrt{100}} = 110 \pm 3.92 : \mu \in (106.08, 113.92)$$

— b) Error máximo cometido: $E = \frac{113.92 - 106.8}{2} = 3.92$

Ejercicio 5.12. El peso, en kg, de los jóvenes entre 16 y 20 años de una cierta ciudad es una variable aleatoria, x , que sigue una distribución normal con $\sigma = 5$.

Si se desea que la media de la muestra no difiera en más de 1 kg de la media de la población, con probabilidad 0.95, ¿cuántos jóvenes se deberían tomar en la muestra?

$$\sigma = 5; n = 25 : \bar{x} \in N(\mu, 5/\sqrt{25}) = N(\mu, 1)$$

$$E > 1; p = 95\% \rightarrow z_{\alpha/2} = 1.96 \rightarrow n?$$

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 1.96 \cdot 5/\sqrt{n} < 1 \rightarrow 9.8 < \sqrt{n} \Rightarrow n > 96.04$$

Se deberían tomar 97 jóvenes en la muestra.

5.4. Estimación de la proporción

Definición 5.7:

Supongamos que estamos interesados en conocer si los individuos de una población poseen o no una determinada característica. Habrá una proporción de individuos, p , de la población que sí la tengan y otra, $q = 1 - p$, de que no la tengan.

Tomamos muestras M_i , de tamaño n , que sometidas a este estudio resultarán ser $B(n, p_i)$. Evidentemente, por término medio tendremos $n \cdot p_i$ individuos que sí tienen la característica buscada (éxitos). La proporción de ellos será $\frac{n \cdot p_i}{n} = p_i$, proporción de los individuos de la muestra M_i que tienen la característica buscada.

De las k muestras extraídas de la población tendremos las proporciones $\{p_1, p_2, \dots, p_k\}$ que formarán la **distribución de las proporciones muestrales** o distribución muestral de la proporción, de valor esperado o media \hat{p}

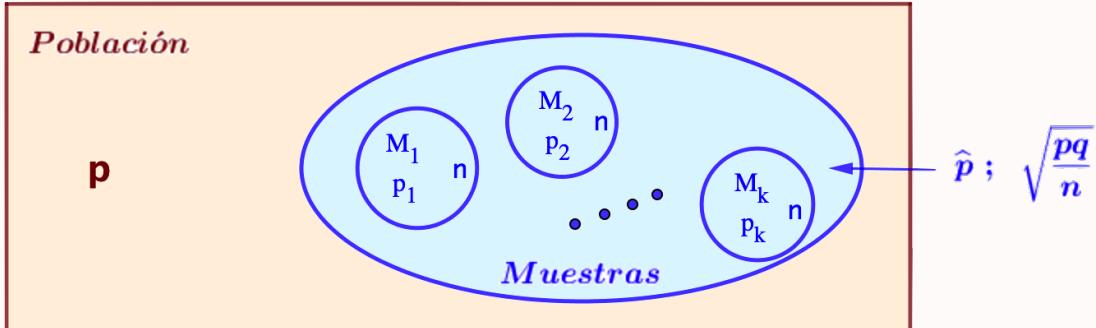
Teorema 5.4:

Distribución de las proporciones muestrales: $\{p_1, p_2, \dots, p_k\}$, todas de tamaño n , extraídas de una población en que la proporción en p

- La media \hat{p} de la distribución de las muestras de la proporción coincide con la proporción p de la población.
- La desviación típica de las medias de las proporciones es $\frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}}$

- Si $n \geq 30$ y también $np \geq 5 \wedge nq \geq 5$, $B \approx N$, y la distribución de las proporciones muestrales es normal:

$$\hat{p} = N\left(p, \sqrt{\frac{pq}{n}}\right)$$



Si en algún caso, la proporción de la población es desconocida, la aproximaremos por \hat{p} (media de la distribución de las proporciones muestrales), siempre que $n \geq 100$.

Ejemplo 5.9:

Una vacuna inmuniza al 85% de los que la reciben. Si se toma una muestra de 30 personas, ¿cuál es la distribución de sus medias muestrales? ¿Y si se toman 1000?

Se estudia un grupo de 1000 vacunados comprobando que 873 de ellos son inmunes a la enfermedad. Escribe los valores de p y de \hat{p} .

Calcula la probabilidad de que de los 1000 vacunados haya menos de 800 inmunes. Más de 860. Entre 840 y 860.

$$p = 0.85 \rightarrow q = 0.15; \quad n = 30 : \quad \hat{p} = p = 0.85; \quad \sigma_{\hat{p}} = \sqrt{\frac{0.85 \cdot 0.15}{30}} = 0.065$$

$$n = 30 \rightarrow \hat{p} \in N(0.85, 0.065)$$

Es decir, las proporciones de tamaño 30 que vayamos encontrándose distribuirán normalmente con media 0.85 y desviación típica 0.065

$$n = 100 \rightarrow \hat{p} \in N(0.85, \sqrt{0.85 \cdot 0.15 / 1000}) = N(0.85, 0.011)$$

$$873 \text{ de } 1000 \text{ da } \hat{p} = 0.873; \quad p = 0.85$$

$$n = 100 : N(0.85, 0.011) \rightarrow p(\hat{p} < 840/1000) = p(z < (0.84 - 0.85)/0.011) = p(z < -0.91) = p(z > 0.91) = 1 - \Phi(0.91) = 0.1814$$

$$p(\hat{p} > 860/1000) = p(z > (0.86 - 0.85)/0.011) = p(z > 0.91) = 1 - \Phi(0.91) = 0.1814$$

$$p(840/1000 < \hat{p} < 860/1000) = p(-0.91 < z < 0.91) = 2\Phi(0.91) - 1 = 0.6372$$

Ejercicio resuelto 5.8. El 10% de los paquetes de arroz de una determinada marca contiene menos arroz de lo indicado. Analizamos 400 de estos paquetes, ¿cuál es la distribución de las proporciones de envases con menos peso del indicado?

Calcula la probabilidad de que en las bolsas analizadas haya más de 50 con peso inferior al indicado.

$$p = 0.1; q = 0.9; n = 40 > 30 \rightarrow \hat{p} \in N(0.1, \sqrt{0.1 \cdot 0.9/400}) = N(0.1, 0.015)$$

$$p(\hat{p} > 50/400) = p(\hat{p} > 0.125) = p(z > (0.125 - 0.1)/0.015) = p(z > 1.67) = 1 - \Phi(1.67) = 0.0475$$

Hay una probabilidad del 4.75% de que en la muestra de 500 paquetes haya más de 50 con peso menor del indicado.

5.4.1. Intervalo de confianza para una proporción

Teorema 5.5:

Para estimar la proporción poblacional p , se extrae una muestra de tamaño n y se obtiene \hat{p} .

El intervalo de confianza para p con un nivel de confianza del $100(1 - \alpha)\%$ es:

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

$$IC(\alpha) : p \in \left(\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right)$$

$$\text{El error admisible es : } E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Ejemplo 5.10:

De una muestra de 300 personas mayores de 18 años en una gran ciudad, se encontró que 104 de ellas leían algún periódico regularmente. Hallar, con un nivel de confianza del 95%, un intervalo para estimar la proporción de lectores de periódicos de esa ciudad entre los mayores de 18 años.

A la vista de los resultados del problema anterior, si se desea conseguir una cota del error de 0.01, con el mismo nivel de confianza del 95 %, ¿a cuántos individuos debe tener la muestra?

$$n = 300; \quad \hat{p} = 104/300 = 0.347; \quad 95\% : z_{\alpha/2} = 1.96$$

$$p \in 0.347 \pm 1.96\sqrt{(0.347 \cdot (1 - 0.347)/300)} = 0.347 \pm 0.027$$

$$IC(95\%) : (0.340, 0.374)$$

$$E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = 1.96 \cdot \sqrt{0.347(1 - 0.347)/n} < 0.01 \rightarrow n > 1.96^2 \cdot 0.347 \cdot (1 - 0.347)/0.01^2 = 8710.6 \rightarrow n = 8711$$

Ejercicio resuelto 5.9. Se lanza una moneda 200 veces obteniéndose 113 caras. Da una estimación de la probabilidad de obtener cara con un nivel de confianza del 90 %.

La proporción de caras en los 200 lanzamientos es $\hat{p} = 113/200 = 0.565$

$$\text{Al } 90\%, z_{\alpha/2} = 1.645 \rightarrow p \in N(0.565, \sqrt{0.565 \cdot 0.435/200}) = N(0.565, 0.035)$$

$$p \in \hat{p} \pm z_{\alpha/2} \cdot \sqrt{pq/n} = 0.565 \pm 1.645 \cdot 0.035 = 0.565 \pm 0.058$$

$$IC(90\%) : (0.507, 0.623)$$

Es decir, entre $0.507 \cdot 200$ y $0.623 \cdot 200$, entre 101 y 125 caras se obtendrán, en los 200 lanzamientos, con un nivel de confianza del 90 %

5.4.2. Ejercicios de estimación de las proporciones

Ejercicio 5.13. ¿De qué tamaño habría que elegir una muestra para estimar la proporción de ciudadanos a los que les gusta el fútbol con un nivel de confianza del 95 % y un error inferior a 0.05, si en una muestra de 25 ciudadanos 15 de ellos respondieron que les gustaba el fútbol?

Para $n = 25 \rightarrow \hat{p} = 15/25 = 0.6; \hat{q} = 1 - \hat{p} = 0.4$.

Al 95 % $\rightarrow \alpha = 5\% \rightarrow z_{\alpha/2} = 1.96$, deseamos $E < 0.05 \rightarrow n ?$

$$z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} < E \quad (\hat{p} > 0, \hat{q} > 0, n > 0)$$

$$n > \frac{z_{\alpha/2}^2 \hat{p} \hat{q}}{E^2} = \frac{1.96^2 \cdot 0.6 \cdot 0.4}{0.05^2} = 368.7936 \rightarrow n \geq 369 \text{ ciudadanos.}$$

Ejercicio 5.14. Para saber qué proporción de alumnos de la ESO practican deportes se selecciona una muestra de 100 alumnos, de ellos contestan afirmativamente 75. ¿Cuál es el intervalo de confianza para la proporción de los alumnos deportistas, con un nivel de confianza del 95%

$$\hat{p} = 75/100 = 0.75, \quad q = 1 - p = 0.25; \quad n = 100$$

Para $1 - \alpha = 95\% \rightarrow \alpha = 5\% : P(z < z_{\alpha/2}) = 1 - \frac{\alpha}{2} = 1 - 0.025 = 0.975$ Leyendo en las tablas de la normal al revés, $z_{\alpha/2} = 1.96$.

Hemos repetido el cálculo de $z_{\alpha/2}$ para recordar conocimientos del tema anterior.

$$IC(95\%) : \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p} \hat{q}}{n}} = 0.75 \pm 1.96 \sqrt{\frac{0.75 \cdot 0.25}{100}} = 0.75 \pm 0.08$$

El intervalo de confianza al 95 % de los alumnos de la ESO que hacen deporte es $(0.67, 0.83)$

Ejercicio 5.15. En una encuesta hecha a un total de 100 votantes elegidos al azar en su ciudad, se indica que el 55 % volvería a votar por el alcalde actual.

a) Calcular un intervalo de confianza al 99 % y otro al 99.73 % para la proporción de votantes favorables al alcalde actual.

b) ¿Cuáles deben ser los tamaños muestrales en el sondeo para tener, con los mismos niveles de confianza, la certeza de que el alcalde actual salga reelegido por mayoría absoluta, en el caso de arrojar la encuesta los mismos resultados?

c) ¿Qué efecto que tendría sobre el intervalo de confianza el aumento, o la disminución, del nivel de confianza?

— a) Calculemos, previamente, el valor de $z_{\alpha/2}$ para un nivel de confianza del $1 - \alpha = 0.9973 \rightarrow \alpha = 0.027$, por ello, $p(z < z_{\alpha/2}) = 1 - \frac{0.027}{2} = 0.99865$. Leyendo las tablas de la normal típica al revés, obtenemos $z_{\alpha/2} = 3.00$

Es sabido que para $1 - \alpha = 99\% \rightarrow z_{\alpha/2} = 2.575$

$$IC(\alpha) : \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p} \hat{q}}{n}}; \quad \hat{p} = 0.55; \quad n = 100$$

$$IC(99\%) : 0.55 \pm 2.575 \sqrt{\frac{0.55 \cdot 0.45}{100}} = 0.55 \pm 2.575 \cdot 0.0497 = 0.55 \pm 0.128$$

$$IC(99.73\%) : 0.55 \pm 3 \sqrt{\frac{0.55 \cdot 0.45}{100}} = 0.55 \pm 3 \cdot 0.0497 = 0.55 \pm 0.149$$

Al 99 % : $(0.422, 0.678)$ y al 99.73 % $(0.401, 0.699)$

— b) La mayoría absoluta la obtendrá si el error es menor de 5 %, ya que $\hat{p} = 55\%$. Con un error mayor el alcalde podría perder la mayoría absoluta ($0.55 - E < 0.50$ si $E > 0.05$).

$$z_{\alpha/2} \sqrt{\frac{\hat{p} \hat{q}}{n}} < E \rightarrow n > \frac{z_{\alpha/2}^2 \hat{p} \hat{q}}{E^2}$$

$$n(99\%) > \frac{2.575^2 \cdot 0.55 \cdot 0.45}{0.05^2} = 656.4 \rightarrow n \geq 657$$

$$n(99.73\%) > \frac{3^2 \cdot 0.55 \cdot 0.45}{0.05^2} = 891 \rightarrow n \geq 892$$

— c) Al aumentar el nivel de confianza, aumenta la amplitud del intervalo. Es decir, cuanta más precisión se quiera tener en la estimación, mayor será el error máximo admisible.

Ejercicio 5.16. En un saco mezclamos bolas blancas y bolas negras poniendo 14 blancas por cada bola negra. Extraemos una muestra de 100 bolas.

- a) ¿Cuál es la probabilidad de que la proporción de bolas negras esté entre 0.05 y 0.1?
- b) Halla un intervalo para el 99 % de las proporciones de las muestras de tamaño 100.

— a) Proporción bolas negras: $\hat{p} = 1/15$; $\hat{q} = 14/15$; $n = 100$ $\hat{p} \in N(0.067, 0.025)$

$$p(0.050 < \hat{p} < 0.100) = p(-0.68 < z < 1.32) = \Phi(1.32) + \Phi(0.68) - 1 = 0.6583$$

— b) $IC(99\%)$, $z_{\alpha/2} = 2.575$: $0.067 \pm 2.575 \cdot 0.025 \rightarrow (0.03, 0.131)$

Ejercicio 5.17. La probabilidad de que un recién nacido sea chica es 0.515. Si han nacido 184 bebés, ¿cuál es la probabilidad de que haya 100 o más chicas? Halla el intervalo característico correspondiente al 95 % para la proporción de chicas en muestras de 184 recién nacidos.

Proporción de chicas: $p = 0.515$; muestra: $n = 184 \rightarrow \hat{p} \in N(0.515, \sqrt{0.515 \cdot (1 - 0.515)/184}) = N(0.515, 0.037)$

La proporción de 100 chicas entre 184 bebés es $100/184 = 0.534$

$$p(\hat{p} > 0.534) = p[z > (0.534 - 0.515)/0.037] = p(z > 0.51) = 1 - \Phi(0.51) = 0.3050$$

$$IC(95\%) \rightarrow z_{\alpha/2} = 1.96 \rightarrow 0.515 \pm 1.96 \cdot 0.037 = 0.515 \pm 0.072$$

El intervalo pedido es: (443, 587)

Una forma **más precisa** de resolver este y otros problemas es: número de chicas en 184 bebés es una $B(184, 0.515) \rightarrow p(x > 100)$ son muchos cálculos.

Estamos en condiciones de aproximar por la normal:

$$B(184, 0.515) \approx B(184 \cdot 0.515, \sqrt{194 \cdot 0.515 \cdot (1.0.515)}) = N(94.76, 6.78)$$

Ahora viene la **precisión**, al pasar de la binomial a la normal hemos de hacer la *corrección por continuidad*.

$$p(x \geq 100) = p(x' > 99.5) = p(z > (99.5 - 94.76)/6.78) = p(z > 0.70) = 1 - \Phi(0.70) = 0.2420$$

Ejercicio 5.18. Se desea estimar la proporción, p , de individuos zurdos de una ciudad a través del porcentaje observado en una muestra aleatoria de individuos, de tamaño n .

a) Si el porcentaje de zurdos en la muestra es igual al 30 %, calcula el valor de n para que, con un nivel de confianza del 95 %, el error cometido en la estimación sea inferior al 3 %.

b) Si el tamaño de la muestra es de 100 personas, y el porcentaje de zurdos en la muestra es del 35 %, determina, usando un nivel de significación del 1 %, el correspondiente intervalo de confianza para la proporción de zurdos en la población.

— a) $p \leftarrow n$, $\hat{p} = 0.3$; $\hat{q} = 0.7$ $1 - \alpha = 95\% \rightarrow z_{\alpha/2} = 1.96$

$$E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \hat{q}}{n}} \rightarrow n = \frac{z_{\alpha/2}^2 \hat{p} \hat{q}}{E^2}$$

$$E < 0.03 \rightarrow n > \frac{1.06^2 \cdot 0.3 \cdot 0.7}{0.03^2} = 896.37$$

La muestra ha de ser de al menos $n = 897$ individuos.

— b) $n = 100$; $\hat{p} = 0.35$ ($\hat{q} = 0.65$) $\alpha = 1\% \rightarrow 1 - \alpha = 99\% \rightarrow z_{\alpha/2} = 2.575 \Rightarrow IC$

$$p \in \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \hat{q}}{n}} = 0.35 \pm 2.75 \cdot \sqrt{0.35 \cdot 0.65 / 100} = 0.35 \pm 0.12$$

$IC(99\%) : (23, 47)$

Ejercicio 5.19. En una muestra de 100 noticias en la web de una facultad, se observan que aparecen 6 con faltas de ortografía.

a) Estima la proporción real de noticias con faltas, con un nivel de confianza del 99 %.

b) ¿Cuál es el error máximo cometido al hacer la estimación anterior?

c) ¿De qué tamaño tendríamos que coger la muestra, con un nivel de confianza del 99 %, para obtener un error inferior a 0.05 ?

— a) $n = 100$; $\hat{p} = 6/100 = 0.06$; $\hat{q} = 0.94$; al 99 %, $z_{\alpha/2} = 2.575$

$$p \in \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \hat{q}}{n}} = 0.06 \pm 2.575 \cdot \sqrt{0.06 \cdot 0.94 / 100}$$

$IC(99\%) : (0.00, 0.12)$

— b) $E = z_{\alpha/2} \cdot \sqrt{\hat{p} \hat{q}/n} = 2.575 \cdot \sqrt{0.06 \cdot 0.94/100} = 0.06$

— c) En las misma condiciones, para $E < 0.05$:

$$E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \hat{q}}{n}} \rightarrow n = \frac{z_{\alpha/2}^2 \hat{p} \hat{q}}{E^2}$$

$$E < 0.05 \rightarrow n > \frac{2.575^2 \cdot 0.06 \cdot 0.94}{0.05^2} = 149.58$$

Para un error inferior al 5% hay que muestrear 150 noticias.

Ejercicio 5.20. Un estudio realizado por una compañía de seguros de automóviles establece que una de cada cinco personas accidentadas es mujer. Si se contabilizan, por término medio, 169 accidentes cada fin de semana:

a) ¿Cuál es la probabilidad de que, en un fin de semana, la proporción de mujeres accidentadas supere el 25%?

b) ¿Cuál es la probabilidad de que, en un fin de semana, la proporción de hombres accidentados supere el 85%?

c) ¿Cuál es, por término medio, el número esperado de hombres accidentados cada fin de semana?

— a) $n = 169; \hat{p} = 1/5 = 0.2; \hat{q} = 0.8 \rightarrow p \in N(p, \sqrt{pq/n}) = N(0.20, 0.03)$

$$p(p > 0.25) = p(z > (0.25 - 0.20)/0.03) = p(z > 1.67) = 1 - \Phi(1.67) = 1.0.0525 = 4.75\%$$

— b) ‘hombres accidentados’ = (‘mujeres accidentadas’) $\rightarrow N(0.80, 0.03)$

$$p(p > 0.85) = p(z > (0.85 - 0.80)/0.03) = p(z > 1.67) = 1 - \Phi(1.67) = 1.0.0525 = 4.75\%$$

Número de hombres accidentados cada semana: $p \cdot n = 0.8 \cdot 169 = 132.5$, aproximadamente 133 hombres accidentados semanalmente.

Ejercicio 5.21. A partir de una muestra de tamaño 400, se estima la proporción de individuos que leen el periódico en una gran ciudad. Se obtiene una cota de error de 0,0392 con un nivel de confianza del 95%.

Calcula la proporción, \hat{p} , obtenida en la muestra.

$$E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} (1 - \hat{p})}{n}} ; \quad E = 0.0392; \quad n = 400; \quad 1 - \alpha = 95\% \rightarrow z_{\alpha/2} = 1.96$$

$$E^2 = z_{\alpha/2}^2 \cdot \frac{\hat{p} \hat{q}}{n} \rightarrow \hat{p} (1 - \hat{p}) = \frac{E^2}{z_{\alpha/2}^2} \cdot n = \frac{0.0392^2}{1.96^2} \cdot 400 = 0.16$$

$$\hat{p} - \hat{p}^2 = 0.16 \rightarrow \hat{p}^2 - \hat{p} + 0.16 = 0 \rightarrow \begin{cases} p = 0.8 \\ p = 0.2 \end{cases}$$

Como no tenemos más información, ambos resultados son válidos.

Ejercicio 5.22. Con una muestra de 500 individuos, se ha estimado, con un nivel de confianza del 90 %, que la estatura media de los trabajadores de una cierta fábrica está entre 174.3 cm y 175.1 cm.

- a) Si la desviación típica de la población es desconocida, averigua la media, \bar{x} , y la desviación típica, s_x , de la muestra.
- b) ¿Cuál sería el intervalo si la muestra fuera de tamaño la cuarta parte y se mantuviese el nivel de confianza?

— a) La media muestral será el punto medio del intervalo de confianza: $\bar{x} = (174.3 + 175.1)/2 = 174.7$ cm.

La mitad del amplitud del intervalo es el error máximo admisible: $E = (175.1 - 174.3)/2 = 0.4 = z_{\alpha/2} \cdot \frac{s_x}{\sqrt{n}} = 1.96 \cdot \frac{s_x}{\sqrt{500}} \rightarrow s_x = 5.4$

— b) Si $n = 500/4 = 125$ trabajadores, entonces $E = z_{\alpha/2} \cdot \frac{s_x}{\sqrt{n}} = 1.96 \cdot \frac{5.44}{\sqrt{125}} = 0.8$

El intervalo de confianza sería $174.7 \pm 0.8 \rightarrow (173.9, 175.5)$

Ejercicio 5.23. Supongamos que, a partir de una muestra aleatoria de tamaño 25, se ha calculado el intervalo de confianza para la media de una población normal, obteniéndose una amplitud de 4. Si el tamaño de la muestra hubiera sido 100, permaneciendo invariables todos los demás valores que intervienen en el cálculo, ¿cuál habría sido la amplitud del intervalo?

$$E = 4; n = 25 \rightarrow \text{si } n = 100 \rightarrow E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{100}} = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{4 \cdot 25}} = \frac{1}{\sqrt{4}} \left(z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{25}} \right) = \frac{1}{2} \cdot (4) = 2$$

Ejercicio 5.24. a) Un fabricante de medicamentos afirma que cierta medicina cura una enfermedad de la sangre en el 80 % de los casos. Los inspectores de sanidad utilizan el medicamento en una muestra de 100 pacientes y deciden aceptar dicha afirmación si se

curan 75 o más. Si lo que afirma el fabricante es realmente cierto, ¿cuál es la probabilidad de que los inspectores rechacen dicha afirmación?

b) Supongamos que en la muestra se curan 60 individuos. Determina, con una confianza del 95 %, cuál es el error máximo cometido al estimar que el porcentaje de efectividad del medicamento es del 60 %.

— a) $p = 0.8 \rightarrow q = 0.2; n = 100 \rightarrow \hat{p} \in N\left(p, \sqrt{\frac{\hat{p}\hat{q}}{n}}\right) = N(0.8, \sqrt{0.8 \cdot 0.2/100}) = N(0.80, 0.04)$

$$p(\hat{p} < 75\%) = p(z < (0.75 - 0.8)/0.04) = p(z < -1.25) = p(z > 1.25) = 1 - \Phi(1.25) = 0.1056$$

— b) si $\hat{p} = 0.6 \rightarrow \hat{q} = 0.4$; al 95 %: $z_{\alpha/2} = 1.96$

$$E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96 \cdot \sqrt{0.6 \cdot 0.4/100} = 0.096 = 9.6\%, \text{ aproximadamente 10 personas.}$$

5.5. Problemas tipo de distribuciones muestrales

Problemas tipo de distribuciones muestrales

1-a) Datos: $\bar{x}; n, \sigma$; pregunta: IC para μ

$$\mu \in N\left(\bar{x}, \frac{\sigma}{\sqrt{n}}\right) \longrightarrow \text{al } (1 - \alpha)\% : IC(\alpha) : \mu \in \bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Si se desconoce σ se usará s_{n-1} muestral.

1-b) Datos: $\mu; n, \sigma$; pregunta: IC para \bar{x}

$$\bar{x} \in N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \longrightarrow \text{al } (1 - \alpha)\% : IC(\alpha) : \bar{x} \in \mu \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

2) Datos: n, σ, α ; pregunta: tamaño de la muestra n

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \longrightarrow n \geq \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$$

3) Datos: n , σ , E ; pregunta: nivel de confianza $(1 - \alpha)\%$

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \longrightarrow z_{\alpha/2} = \frac{E \sqrt{n}}{\sigma} \longrightarrow p(z \geq z_{\alpha/2}) = \alpha/2 \longrightarrow \alpha$$

4-a) Datos: n , \hat{p} ; α ; pregunta IC para la proporción p

$$IC(\alpha) : p \in \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \hat{q}}{n}}$$

4-b) Datos: n , p ; α ; pregunta IC para la proporción muestral \hat{p}

$$IC(\alpha) : \hat{p} \in p \pm z_{\alpha/2} \cdot \sqrt{\frac{pq}{n}}$$

5.6. Ejercicios

5.6.1. Problemas propuestos

- PB. 1. En un gran supermercado se ha obtenido que el número medio de toneladas descargadas diariamente en los últimos 100 días ha sido igual a 10. Determine el intervalo, con un nivel de confianza del 95 %, en el que estará la media si la desviación típica es igual a 6.

(8.824; 11.176)

- PB. 2. El peso de los niños varones a las diez semanas de vida se distribuye según una Normal con desviación típica de 87 g. ¿Cuántos datos son suficientes para estimar, con una confianza del 95 %, el peso medio de esa población con un error no superior a 15 g?

130 niños

PB. 3. En el juzgado de una determinada ciudad se presentaron en el año 2005 un total de 5500 denuncias. Se seleccionó una muestra aleatoria de un 5 % de ellas. Entre las denuncias seleccionadas se determinó que 55 habían sido producidas por violencia doméstica. Determinar, justificando la respuesta:

- a) La estimación puntual que podríamos dar para el porcentaje de denuncias por violencia doméstica en esa ciudad en el año 2005.
- b) El error máximo que cometeríamos con dicha estimación puntual con un nivel de confianza del 99 %.

$$\text{a) } \hat{p} = 0.062 \quad \text{b) } E = 0.062$$

PB. 4. Se desea determinar el porcentaje de jóvenes entre 14 y 19 años que necesitan llevar gafas en cierto instituto. ¿Qué tamaño de muestra debemos escoger para que, al tomar el porcentaje muestral como aproximación del porcentaje poblacional, cometamos un error máximo del 10 %, con un nivel de confianza del 95 %?

$$\text{d) } n = 0.0026$$

PB. 5. Se supone que las alumnas universitarias de una determinada ciudad sigue una determinada distribución Normal de media 1.65 m y desviación típica 10 cm. Se toma una muestra al azar de 100 de esas alumnas y se calcula su media. ¿Cuál es la probabilidad de que esa media sea mayor de 1.66 m?

$$\text{e) } 0.1587$$

PB. 6. La duración de las baterías de un determinado modelo de teléfono tiene una distribución normal de media 34.5 horas y desviación típica 6.9 horas. Se toma una muestra aleatoria simple de 36 teléfonos móviles.

- a) ¿Cuál es la probabilidad de que la duración media de las baterías de la muestra esté comprendida entre 32 y 33.5 horas?
- b) ?Y de que sea mayor de 38 horas?

$$\text{f) } \text{a) } 0.1772; \text{ b) } 0.0012$$

PB. 7. En cierta población humana, la media muestral X de una característica se distribuye mediante una distribución Normal. La probabilidad de que X sea menor o igual que 75 es 0.58 y la de que X sea mayor que 80 es 0.04. Hallar la media y la desviación típica de X . (Tamaño muestral $n = 100$)

Media muestral y poblacional: 74.35; desviación típica muestral: 3.22

- PB. 8. Se desea hacer un estudio de mercado para conocer el precio medio de los libros de texto. Para ello se selecciona una muestra aleatoria cogiendo 121 libros de texto, encontrando que tienen un precio medio de 23 euros. Si sabemos que los precios de los libros de texto siguen una distribución Normal con desviación típica de 5 euros, encontrar un intervalo de confianza al 98.8 % para el precio medio de los libros de texto.

(21.86; 24.14)

- PB. 9. La duración de la baterías de cierto teléfono móvil se puede aproximar por una distribución Normal con una desviación típica de 5 meses. Se toma una muestra aleatoria simple de 10 baterías y se obtienen las siguientes duraciones (en meses):
 33, 34, 26, 37, 30, 39, 26, 31, 36, 19
 Hallar un intervalo de confianza al 95 % para ese modelo de batería.

(28,34.2)

- PB. 10. En una encuesta se pregunta a 10 000 personas cuántos libros leen al año, obteniéndose una media de 5 libros. Se sabe que la población tiene una distribución normal con desviación típica 2.
- Hallar un intervalo de confianza al 80 % para la media poblacional.
 - Para garantizar un error de estimación de la media poblacional no superior a 0.25 con un nivel de confianza del 95 %, ¿a cuántas personas como mínimo sería necesario entrevistar?

a) (4.97; 5.02) ; b) 246 personas de tamaño muestral

- PB. 11. Un fabricante de pilas alcalinas sabe que el tiempo de duración, en horas, de las pilas que fabrica sigue una distribución Normal de media desconocida y varianza 3600. Con una muestra de su producción, elegida al azar, y un nivel de confianza del 95 % ha obtenido para la media el intervalo de confianza (372.6; 392.2).
- Calcule el valor que obtuvo para la media de la muestra y el tamaño muestral utilizado.
 - ¿Cuál sería el error de su estimación, si hubiese utilizado una muestra de tamaño 225 y un nivel de confianza del 86.9 %?

a) media 382.4 horas, $n=144$; b) $E=6.04$

- PB. 12. Se hizo una encuesta aleatoria entre 130 estudiantes universitarios de los cuales 85 eran mujeres, sobre el número de horas que estudian diariamente fuera del aula, obteniéndose una media de 3.4 horas.
- Si la desviación típica es de 1.1 horas, obtener un intervalo de confianza, al 98 %, para la media del número de horas que estudian diariamente fuera del aula los estudiantes universitarios.
 - Obtener un intervalo de confianza al 90 % para la proporción de mujeres entre los estudiantes universitarios.

a) $(3.175; 3.625)$; b) Proporción mujeres universitarias: $(58.49\%, 72.26\%)$

- PB. 13. A una muestra aleatoria de 300 estudiantes de bachillerato de determinada provincia se les preguntó si utilizaban habitualmente la bicicleta para acudir a su instituto. Sabiendo que se obtuvo 90 respuestas afirmativas, determinar justificando la respuesta:
- El intervalo de confianza al 95 % para el porcentaje de estudiantes de bachillerato de esa provincia que utilizan habitualmente la bicicleta para acudir al instituto.
 - El error máximo que cometeríamos, con una confianza del 95 %, si estimamos que dicho porcentaje es del 30 %.

a) $(0.2481; 0.3518)$; b) $E=0.0518$

- PB. 14. En una universidad se toma al azar una muestra de 100 alumnos y se encuentra que han suspendido todas las asignaturas 10 alumnos. Se pide hallar:
- Con un nivel de confianza del 95 %, un intervalo para estimar el porcentaje de alumnos que aprueba al menos una asignatura.
 - A la vista del resultado anterior, se pretende repetir la experiencia para conseguir una cota de error del 0.03 con el mismo nivel de confianza del 95 %. ¿Cuántos individuos ha de tener la muestra?

a) $(0.84; 0.96)$; b) 385 alumnos

- PB. 15. En una muestra de 600 personas de una ciudad se observa que 30 son inmigrantes.
- Determinar el intervalo de confianza de nivel 0.95 para el porcentaje de inmigrantes en la ciudad.

- b) Si se quiere estimar el porcentaje de inmigrantes con un error máximo de 0.02, ¿cuál es el tamaño de la muestra que habría que considerar si se usa un nivel de significación del 1 %?

$$\text{a)} (0.0325; 0.0674); \text{ b)} 788 \text{ personas}$$

- PB. 16. Se quiere estimar la media del consumo, en litros de leche por persona al mes. Sabiendo que dicho consumo sigue una normal con desviación típica de 6 litros:

- a) ¿Qué tamaño muestral se necesita para estimar el consumo medio con un error menor de 1 litro y con un nivel de confianza del 96 %?
- b) Si la media del consumo mensual de leche por persona fuese igual a 21 litros, hallar la probabilidad de que la media de una muestra de 16 personas sea mayor que 22 litros.

$$\text{a)} 5 \text{ personas}; \text{ b)} 0.2546$$

- PB. 17. En una población, una variable aleatoria sigue una ley Normal de media desconocida y desviación típica 9. ¿De qué tamaño, como mínimo, debe ser la muestra con la cual se estime la media poblacional con un nivel de confianza del 97 % y un error máximo admisible igual a 3?

$$n > 43$$

- PB. 18. Una compañía aérea sabe que el equipaje de sus pasajeros tiene como media 25 kg. con una desviación típica de 6 kg. Si uno de sus aviones transporta a 50 pasajeros, el peso medio de los equipajes de dicho grupo estará en la distribución muestral de medias, ¿cuál?. Calcula la probabilidad de que el peso medio para estos pasajeros sea superior a 26 kg

$$N(25; 0.84); 11.9\%$$

- PB. 19. La masa de las manzanas de una cosecha se distribuyen normalmente con media 125 g y una desviación típica de 20 g.

- a) ¿Cuál es la probabilidad de que una manzana elegida al azar pese más de 130 g?
- b) ¿Cuál es la probabilidad de que el peso medio en una muestra de 25 peras sea mayor de 130 g?

$$\text{a)} 0.4013; \text{ b)} 0.1056$$

- PB. 20. En unas elecciones, el 52 % de la población votó al candidato A. Si antes de las elecciones se hubiese hecho un sondeo en una muestra de 500 habitantes, ¿cuál hubiese sido la probabilidad de obtener menos de un 50 % de votos para ese candidato, suponiendo que se ha mantenido la intención de voto?

0.1814

- PB. 21. Al 75 % de los jóvenes de una ciudad les gusta el cine. Si seleccionamos 25 muestra de 100 jóvenes cada una, ¿en cuántas cabe esperar que el porcentaje de jóvenes cinéfilos esté comprendido entre el 70 % y el 80 %? ¿Y si las muestras fueses de 1000 jóvenes?

19; 25

- PB. 22. El ascensor de cierto edificio puede transportar una carga máxima de 300 kg.

- a) Si el peso en kilogramos de los usuarios de ese ascensor tiene distribución N (63, 12), ¿cuáles la probabilidad de que un grupo aleatorio de cuatro de ellos sobrepase el peso límite?
 b) Se sabe que el 64.8 % de las veces que el ascensor es usado por un grupo de 4 personas, el peso total de los usuarios no excede cierto peso x ; ¿Cuál es el valor de x_M ?

a) 0.0228; b) 261.12 kg

- PB. 23. Sabemos que las bolsas de azúcar producidas en una fábrica tienen una media de 500 gramos de peso y una desviación típica de 35 gramos. Dichas bolsas se empaquetan en cajas de 100 unidades. Calcula la probabilidad de que una caja pese más de 51 kilogramos.

0.0021

- PB. 24. Las especificaciones de un fabricante de botes de pintura dicen que el peso de los botes sigue una distribución normal de media 1 kg de pintura y una desviación estándar de 0.1 kg.

- a) ¿Cuál es la media y la desviación estándar de la media muestral de los pesos de una muestra aleatoria simple de 20 botes?
 b) Se ha comprado un lote del que se ha tomado una muestra de 20 botes y en el que la media de los pesos obtenidos es de 0.98 kg. Construye un intervalo de confianza del 95 % para la media.

(a) $N(1; 0.022)$; b) $(0.937; 1.023)$

- PB. 25. Se quiere conocer la permanencia media de pacientes en un hospital, con el fin de estudiar una posible ampliación del mismo. Se tienen datos referidos a la estancia, expresada en días, de 800 pacientes, obteniéndose los siguientes resultados: $\bar{x} = 8.1$ días; $s_x = 9$ días. Se pide obtener un intervalo de confianza del 95 % para la estancia media.

 $(7.476; 8.723)$

- PB. 26. Se hizo una encuesta aleatoria entre 130 estudiantes universitarios, de los cuales 85 eran mujeres, sobre el número de horas que estudian diariamente fuera del aula, obteniéndose una media de 3.4 horas.
- Si la desviación típica es de 1,1 horas, obtener un intervalo de confianza, al 98 %, para la media del número de horas que estudian diariamente fuera del aula los estudiantes universitarios.
 - Obtener un intervalo de confianza, al 90 %, para la proporción de mujeres entre los estudiantes universitarios.

(a) $(3.175; 3.625)$; b) $(0.6123; 0.6957)$

- PB. 27. De una muestra aleatoria de 2100 personas de una población hay 630 que leen un determinado diario. Calcular el intervalo de confianza para la proporción poblacional para un nivel de confianza del 99 %.

 $(0.274, 0.326)$

- PB. 28. Tomada al azar una muestra de 60 alumnos de la universidad se encontró que un tercio hablaban el idioma inglés. Hallar, con un nivel de confianza del 90 %, un intervalo para estimar la proporción de alumnos que hablan el idioma inglés entre los alumnos de la universidad.

 $(0.23, 0.43)$

- PB. 29. Supongamos que queremos estudiar la producción media de leche al día de un determinado tipo de vacas con un error menor que 0.5 litros y un nivel de confianza del 0.95 %. Si de estudios anteriores sabemos que la desviación típica es de 1.5 litros, ¿qué tamaño de muestra debemos tomar?

58 = u

- PB. 30. Queremos determinar el porcentaje de estudiantes que necesitan gafas. De un estudio realizado hace tres años sabemos que el 65 % de ellos usaban gafas.
- ¿Qué tamaño de muestra debemos coger para cometer un error máximo del 5 % con un nivel de riesgo del 5 %?
 - Si no tenemos información previa, ¿qué tamaño de muestra debemos tomar?

$$z_{0.05} = u \quad z_{0.025} = u$$

- PB. 31. Una máquina fabrica bombillas que tienen una duración media de 700 horas y una desviación típica de 150 horas. ¿Cuál es la probabilidad de que la media de duración en una muestra de 100 bombillas sea menor o igual a 650 horas?

$$0.0004$$

- PB. 32. Una población de un tipo de plantas tiene una talla media de 15 cm y desviación típica de 2.5 cm. Se toma al azar una muestra de 45 plantas. ¿Cuál es la probabilidad de que la media de las tallas de la muestra sea superior a 12.5 cm?

I

- PB. 33. En la elección para formar parte del consejo escolar, un alumno ha recibido el 50 % de los votos favorables. Si se elige una muestra de 40 alumnos que han votado.
- ¿Cuál es la distribución que sigue la proporción de votantes que han votado?
 - Halla la probabilidad de que más del 40 % de la muestra le votasen.

$$N(0.5; 0.0225)$$

- PB. 34. Los paquetes recibidos en una oficina de correos tienen un peso medio de 20 kg con una desviación típica de 5 kg. Calcula la probabilidad de que el peso de 50 paquetes elegidos al azar supere el límite de seguridad del ascensor, que es de 1000 kg.

$$0.0$$

- PB. 35. Las consultas de un médico de cabecera duran una media de 8 minutos, con una desviación típica de 2.3 minutos. Si una tarde tiene citados 32 pacientes, ¿cuál es la probabilidad de que los atienda en menos de 4 horas?

0.1093

- PB. 36. Uno de los principales fabricantes de televisores compra piezas a dos compañías. Las piezas de la compañía A tienen una vida media de 7.2 años con una desviación típica de 0.8 años, mientras que las de la compañía B tienen una vida media de 6.7 años con una desviación típica de 0.7. Determina la probabilidad de que una muestra aleatoria de 34 piezas de la compañía A tenga una vida media de al menos un año más que la de una muestra aleatoria de 40 piezas de la compañía B.

0.0023

- PB. 37. Tomada al azar una muestra de 500 personas de una determinada comunidad, se encontró que 300 leían la prensa regularmente. Hallar, con una confianza del 90 %, un intervalo para estimar la proporción de lectores entre las personas de esa comunidad.

(0.564; 0.636)

- PB. 38. En una muestra aleatoria de 600 coches de una ciudad, 120 son de color blanco. Construya un intervalo de confianza de la proporción de coches de color blanco con un nivel de confianza del 98 %.

(0.162; 0.238)

- PB. 39. El peso de los paquetes enviados por una determinada empresa de transportes se distribuye según una normal, con una desviación típica de 0.9 Kg. En un estudio realizado con una muestra aleatoria de 9 paquetes, se obtuvieron los siguientes pesos en kilos: 9.5; 10; 8.5; 10.5; 12.5; 10.5; 12.5; 13; 12.

- Halla un intervalo de confianza, al 90 %, para el peso medio de los paquetes enviados por esa empresa.
- Calcula el tamaño mínimo que debería tener una muestra, en el caso de admitir un error máximo de 0.3 Kg, con un nivel de confianza del 90 %.

a) $I = (10.23; 11.77)$; b) $n \text{ mínimo} = 25$

PB. 40. En un país se sabe que la altura de la población se distribuye según una normal cuya desviación típica es igual a 10 centímetros.

a) Si dicha media fuera de 170 centímetros, calcular la probabilidad de que la media muestral, de una muestra de 64 personas, difiera menos de un centímetro de la media de la población.

b) ¿Cuál es el tamaño muestral que se debe tomar para estimar la media de la altura de la población con un error menor de 2 centímetros y con un nivel de confianza del 95 %?

$$\text{a) } P(169 < x < 171) = 0.5762; \text{ b) } n_{\text{mínimo}} = 97$$

PB. 41. Para conocer la audiencia de uno de sus programas (proporción de televidentes que lo prefieren), una cadena de TV ha encuestado a 1000 personas elegidas al azar obteniendo una proporción muestral del 33 % de personas favorables a ese programa. Calcule el error máximo de estimación, por medio de un intervalo de confianza, con un nivel del 92 %.

$$E = 0.026$$

PB. 42. Se va a tomar una muestra aleatoria de 600 recién nacidos en este año en una ciudad para estimar la proporción de varones entre los recién nacidos de esa ciudad, mediante un intervalo de confianza con un nivel del 95 %. ¿Cuál será el error de estimación a ese nivel si se observan 234 varones en la muestra?

$$E = 0.039$$

PB. 43. Se sabe que el peso X de la grasa corporal en adultos que no hacen ejercicio sigue una distribución con media de 24.3 kg y desviación típica de 2.4. En cambio, el peso Y de la grasa en adultos que hacen ejercicio regularmente se distribuye con una media de 20.1 kg y desviación típica de 1.7. Si se eligen en ambas poblaciones muestras aleatorias de 50 personas, ¿cuál es la probabilidad de que la diferencia de la grasa corporal medias sea mayor de 3 kg?

$$0.9981$$

PB. 44. Halla el intervalo de confianza al nivel del 90 % para la diferencia de salarios medios de los trabajadores y las trabajadoras de una gran empresa cuando se ha elegido una muestra de 40 hombres y 35 mujeres, siendo el salario medio de los hombres 1051

euros, y el de las mujeres, 1009 euros, y las desviaciones típicas, de 90 y 78 euros, respectivamente.

(10.19; 73.81)

- PB. 45. Se va a tomar una muestra aleatoria de 600 recién nacidos en este año en una ciudad para estimar la proporción de varones entre los recién nacidos de esa ciudad, mediante un intervalo de confianza con un nivel del 95 %. ¿Con qué proporción estimada será máxima la amplitud de ese intervalo? ¿Cuál es la amplitud máxima?

$$\text{Ayuda: } \frac{dp}{dE} = 0 \leftarrow p_{\max}; \quad z_{\alpha/2} \text{ y } n \text{ constantes.}$$

$$p = 0.5, \text{Amplitud máxima} = 2E = 0.08$$

- PB. 46. Se desea estimar la proporción de adultos que leen un determinado diario local por medio de un intervalo de confianza. Obtenga el tamaño mínimo de la muestra que garantice, aún en la situación más desfavorable, un error de la estimación inferior a 0.03, con un nivel de confianza del 95 %.

$$\text{Caso más desfavorable para } p = 0.5 \text{ (ver ayuda problema anterior); } n = 1068$$

5.7. Curiosidades

Estadística inferencial

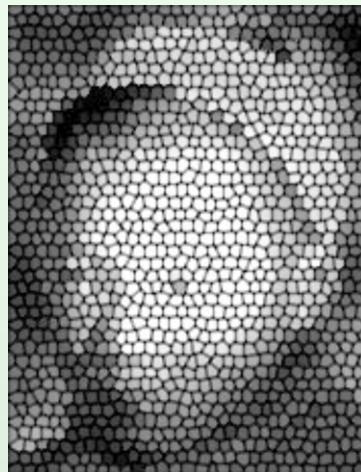
En general, casi nunca se puede tratar con poblaciones al completo. Ya sea porque la población a estudiar es muy grande, ya sea por motivos económicos, de falta de personal cualificado, por que el estudio es destructivo (horas de duración de una bombilla), o para una mayor rapidez en la recogida y presentación de los datos. Lo que se suele hacer es estudiar tan sólo una muestra de la población.

En consecuencia, deberemos contentarnos con utilizar muestras, que sean capaces de revelarnos algo acerca de la población de las que han sido extraídas. De la forma de elegirlas, y las condiciones que han de verificar forma parte de la “teoría del muestreo”.

La Estadística inferencial se ocupa de extender o extrapolar a toda una población, informaciones obtenidas de una muestra, así como de la toma de decisiones.

Al trabajar con muestras, hay que diferenciar los valores observados en la muestra, que llamaremos **estadísticos**, de los valores reales correspondientes a la población, que llamaremos **parámetros poblacionales**.

Observa desde muy de cerca la imagen de la izquierda. Observar esa imagen de esta manera, es equivalente a tomar una muestra de una población. En principio solo tienes en tu mente un conjunto de datos, que no te dicen nada.



Sin embargo, si te alejas unos 5 metros y observas de nuevo la imagen, empezarás a extraer más información, y posiblemente adivines que representa esta imagen. Habrás hecho una inferencia de los datos muestrales, para tener una imagen del conjunto.

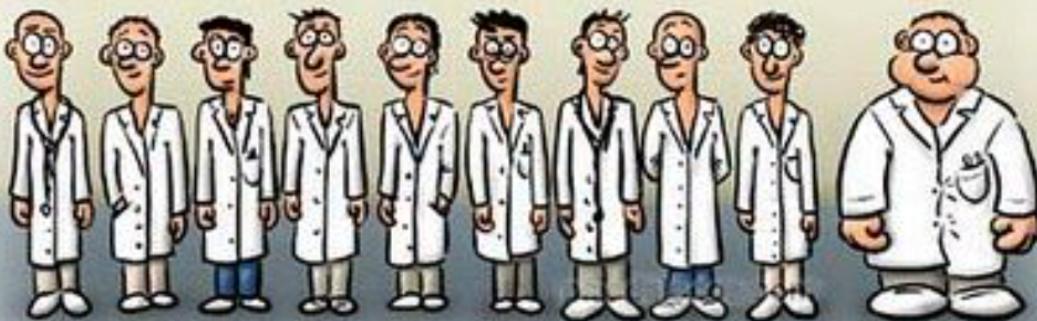
Esta es en resumidas cuentas el objeto de las técnicas de la estadística inferencial: obtener muestras e inferir datos sobre la población.

La siguiente image, medio pixelada a la izquierda, pasa desapercibida si te alejas lo suficiente (tu cerebro hace la inferencia).

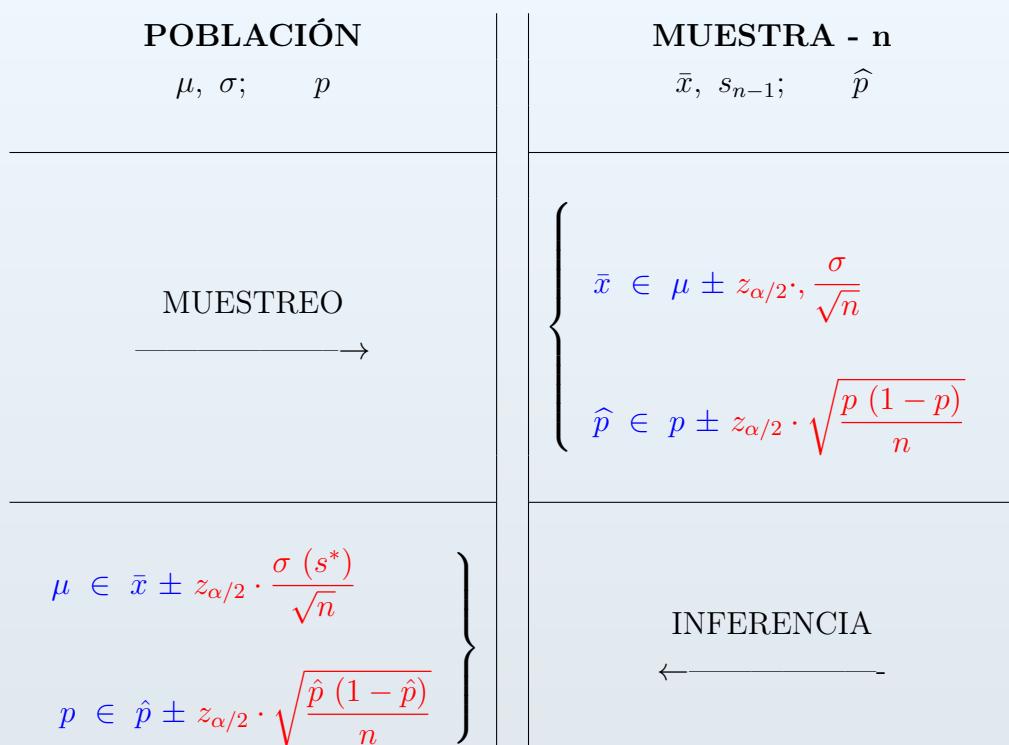




**Nueve de cada diez
especialistas en nutrición
recomiendan la comida
baja en calorías**

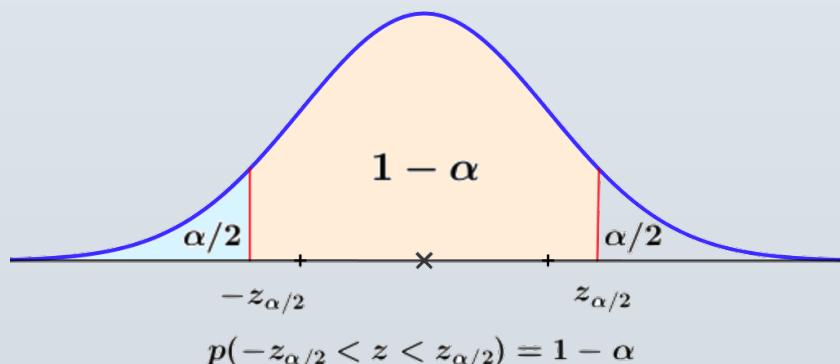


RESUMEN ‘‘Distribuciones muestrales’’



s^* : si σ es desconocido usaremos s muestral.

En rojo aparecer el error máximo admisible.



- ▷ Distribuciones de las sumas de los elementos de una muestral y de las diferencias de las medias de dos muestras:

$$\sum x_i \sim N(n\mu, \sigma\sqrt{n}); \quad \bar{x} - \bar{y} \sim N \left(\mu_x - \mu_y, \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \right)$$

Capítulo 6

Contraste de hipótesis

6.1. Introducción

En el tema anterior hemos visto como estimar la media o proporción de una población mediante un intervalo, a partir de los datos de una muestra.

En este tema abordaremos la toma de decisiones, es decir, plantearemos determinadas hipótesis sobre los parámetros (μ , p) de una población y a partir de los datos de una muestra decidiremos si podemos o no aceptar la hipótesis inicial.¹

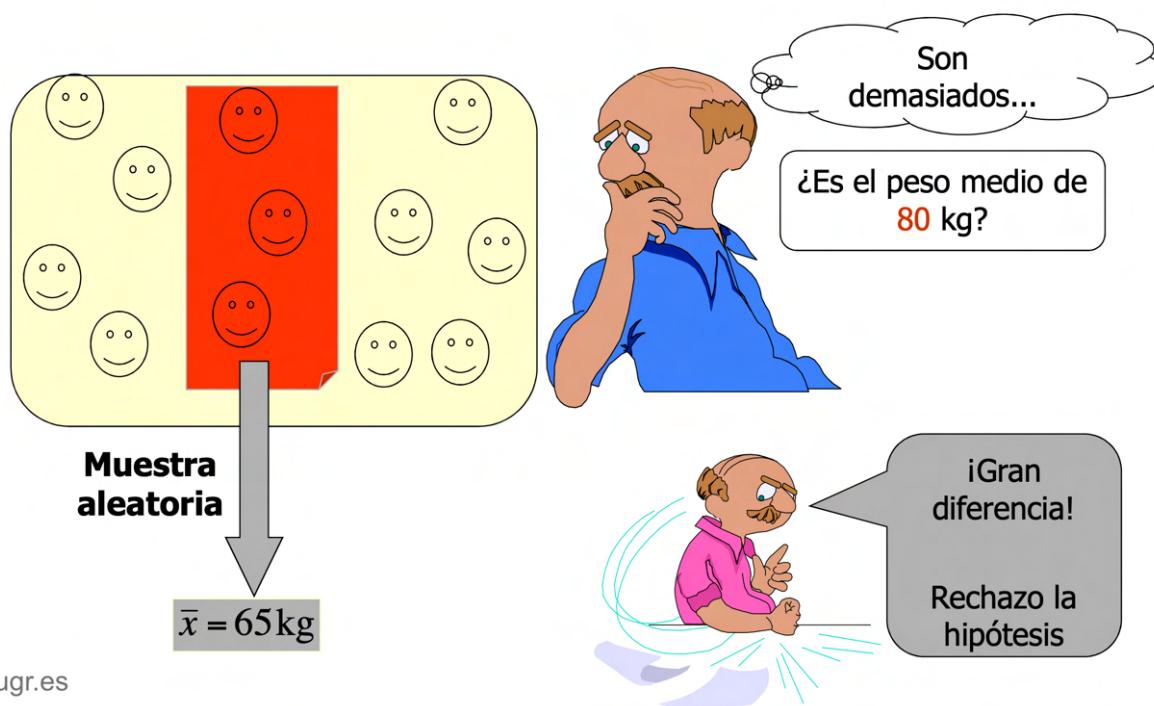
Una hipótesis es, en estadística, una afirmación acerca de una población. Su verdad o falsedad podría determinarse con exactitud si tuviésemos la oportunidad de evaluar a todos los individuos que la componen. Como esto es imposible en general, el criterio para aceptar o rechazar una hipótesis estadística se basa en un razonamiento de tipo probabilístico: a través del estudio de una o varias muestras se determina la probabilidad de que los resultados obtenidos sean compatibles con la hipótesis establecida. Si resulta altamente improbable que, de ser cierta la hipótesis, se hayan obtenido dichos resultados la rechazaremos. Si no es así, lo más que podemos decir es que no existen razones para rechazar dicha hipótesis.

Ejemplos:

- Hace algunos años, la media de estatura de los españoles adultos varones era de 170 cm y su desviación típica 10 cm. Pasado el tiempo, un muestreo realizado a 36 adultos da una medida de 173 cm. ¿Puede afirmarse que esa diferencia de 3 cm es debida al azar o realmente la estatura media ha aumentado?

¹La confección de este tema está basada en los apuntes del profesor Francisco Ocaña Peinado, de la Universidad de Granada y de los profesores Jorge Escribano Colegio e Inmaculada Niña Granada, del colegio Inmaculada niña de Granada.

- Supongamos que, respecto a una determinada ley, el 52 % de los ciudadanos está en contra. Pasado el tiempo, una encuesta realizada a 500 personas indica que los ciudadanos en contra han descendido hasta el 48 %. ¿Ha cambiado realmente la opinión pública o tal resultado es debido al azar?



6.2. Elementos de un contraste de hipótesis

6.2.1. Hipótesis

Definición 6.1:

Se llaman **hipótesis estadísticas** a las conjeturas que se hacen sobre la población.

Un **test de hipótesis o contraste de hipótesis** es el procedimiento estadístico mediante que investiga la verdad o falsedad de una hipótesis acerca de una población.

Estas hipótesis se formulan, normalmente, sobre la media poblacional μ o la proporción poblacional p .

Llamaremos **hipótesis nula**, H_0 , a la hipótesis que se formula y se quiere contrastar para rechazarla o no hacerlo, es decir, la que mantendremos salvo que los datos muestren de forma evidente su falsedad.

Llamaremos **hipótesis alternativa**, H_1 , a cualquier otra hipótesis que sea diferente de la ya formulada y que sea contraria a H_0 , de forma que el rechazo de H_0 suponga la aceptación de H_1 y el no rechazo de H_0 , la no aceptación de H_1 .

Ejemplo 6.1:

- ▷ Decidir la inocencia o culpabilidad de una persona en un país en el que se sigue el principio de presunción de inocencia:

Como se quiere evitar condenar a una persona inocente, sólo se hará cuando haya una fuerte evidencia de su culpabilidad, cuando esté demostrada ésta. En caso de duda, se primará la inocencia frente a la culpabilidad. Por tanto, en la terminología propuesta sería: $H_0 : \text{Inocente}$; $H_1 : \text{Culpable}$.

- ▷ Decidir si un alumno sabe o no la asignatura de Estadística, y por tanto aprueba o suspende la asignatura:

Desde el punto de vista del profesorado, un estudiante no sabe la asignatura mientras no demuestre lo contrario; es decir, el examen ha de presentar pruebas suficientes de que conoce la asignatura. En general, en caso de duda o de falta de datos, se primará el suspenso frente al aprobado. Por tanto, en la terminología propuesta sería: $H_0 : \text{El estudiante no sabe la asignatura (suspende)}$; $H_1 : \text{El estudiante sí sabe la asignatura (aprueba)}$.

Teorema 6.1:

Observaciones:

Sobre la metodología de los tests de hipótesis hay que tener en cuenta que:

- ▷ No sirven para demostrar H_0 .
- ▷ Sirven para decidir que, a partir de los datos de la muestra, o no puede rechazarse H_0 , o sí debe rechazarse y aceptar H_1 .

6.2.2. Errores

Con el método del contraste de hipótesis podemos cometer dos tipos de errores:

Definición 6.2:

	Rechazar H_0	No rechazar H_0
H_0 cierta	Error de tipo I	Decisión correcta
H_1 cierta	Decisión correcta	Error de tipo II

Siguiendo con los ejemplos anteriores:

Ejemplo 6.2:

- Decidir la inocencia o culpabilidad de una persona en un estado en el que se sigue el principio de presunción de inocencia: $H_0 : \text{Inocente}$; $H_1 : \text{Culpable}$.

	Rechazar H_0	No rechazar H_0
H_0 cierta	Error de tipo I Se condena a un inocente	Decisión correcta
H_1 cierta	Decisión correcta	Error de tipo II Se absuelve a un culpable

- Decidir si un alumno sabe o no la asignatura de Estadística, y por tanto aprueba o suspende la asignatura: $H_0 : \text{El estudiante no sabe la asignatura (suspende)}$; $H_1 : \text{El estudiante sí sabe la asignatura (aprueba)}$.

	Rechazar H_0	No rechazar H_0
H_0 cierta	Error de tipo I Se aprueba a quien no sabe	Decisión correcta
H_1 cierta	Decisión correcta	Error de tipo II Se suspende a quien sí sabe

Riesgos al tomar decisiones

Ejemplo 1: Se juzga a un individuo por la *presunta* comisión de un delito

- H_0 : Hipótesis nula
 - Es inocente
- H_1 : Hipótesis alternativa
 - Es culpable



Tipos de error al tomar una decisión

		Realidad	
		Inocente	Culpable
Veredicto	Inocente	OK	Error Menos grave
	Culpable	Error Muy grave	OK

Riesgos al contrastar hipótesis

Ejemplo 2: Se cree que una nueva dieta ofrece buenos resultados

Ejemplo 3: Parece que hay una incidencia de enfermedad más alta de lo normal

- **H_0 : Hipótesis nula**
 - (Ej.1) Es inocente
 - (Ej.2) La nueva dieta no tiene efecto
 - (Ej.3) La incidencia es normal
- **H_1 : Hipótesis alternativa**
 - (Ej.1) Es culpable
 - (Ej.2) La nueva dieta es útil
 - (Ej. 3) La incidencia es anormal



Tipos de error al contrastar hipótesis

		Realidad	
Decisión	H_0 Cierta	H_0 Falsa	
No Rechazo H_0 (Retener H_0)	Correcto <i>La dieta no tiene efecto y así se decide</i> Probabilidad $1 - \alpha$ «Especificidad del test»	Error de tipo II <i>La dieta sí tiene efecto pero no se percibe.</i> Probabilidad β «Falso negativo»	
Rechazo H_0 Acepto H_1	Error de tipo I <i>La dieta no tiene efecto pero se decide que sí.</i> Probabilidad α «Falso positivo»	Correcto <i>La dieta tiene efecto y el experimento lo confirma.</i> Probabilidad $1 - \beta$ «Sensibilidad del test»	

6.2.3. Nivel de significación y potencia

Definición 6.3:

Llamaremos **nivel de significación**, α , a la probabilidad de cometer un error de tipo I, es decir, $\alpha = p$ (Rechazar H_0 | H_0 es cierta) .

Llamaremos **potencia del contraste** al valor de $1 - \beta$, siendo β la probabilidad de cometer un error de tipo II, es decir, $\beta = p$ (No rechazar H_0 | H_0 es falsa)

Teorema 6.2:

Lo ideal sería minimizar α y β , pero esto no puede hacerse simultáneamente pues si disminuye uno aumenta el otro y viceversa.

Así, si un examen es muy exigente se disminuye α , es decir, la probabilidad de aprobar a un estudiante que no sabe; sin embargo, se aumenta β , la probabilidad de suspender a un estudiante que sí sabe. Pero si el examen es poco exigente disminuye la probabilidad de suspender a un alumno que si sabe la asignatura (β), pero aumenta la de aprobar a uno que no sabe lo suficiente (α).

La única manera de disminuir los dos tipos de errores a la vez es aumentando el tamaño de la muestra (preguntar muchas cosas, para tener más datos sobre lo que sabe o no el estudiante)

En general, se fija de antemano un **nivel de confianza** ($1 - \alpha$), que asegure un error de tipo I admisible (haciendo mínima la probabilidad de “condenar a un inocente” o “aprobar a quien no sabe”) y de entre todos los contrastes con dicho nivel de confianza se elige el de mayor potencia. (El estudio de la potencia de un test se escapa al nivel de este curso, así que daremos por hecho que los contrastes de este tema cumplen esa condición).

6.2.4. Región de aceptación y región crítica

Definición 6.4:

Formuladas ya las hipótesis nula y alternativa, el siguiente paso es un criterio para decidir si rechazamos o no la hipótesis nula (no aceptamos o sí la hipótesis alternativa), es decir, ¿con cuál de las dos hipótesis nos quedamos?

Una vez formulada la hipótesis nula, es necesario que las evidencias sean muy fuertes para rechazarla; es decir, puede que haya cambios debidos al azar, en cuyo caso el cambio no es significativo, y no cambiamos, pero puede que los cambios sean debidos a otras causas. En este último caso es cuando el cambio es significativo y rechazaremos.

Por lo tanto, lo primero que debemos hacer es fijar un cierto intervalo dentro del cual es normal que puedan haber cambios, es decir, una región tal que si el parámetro (en nuestro caso media o proporción) se mantiene en dicho intervalo, no rechazamos H_0 , nos seguimos quedando con H_0 , pues esas pequeñas variaciones son debidas al azar. Ese intervalo o región se denomina **región de aceptación**, y será mayor o menor dependiendo del **nivel de confianza** que precisemos, $1 - \alpha$.

La región que quede fuera de la región de aceptación indica que en este caso los cambios no se pueden atribuir al azar, y por tanto hemos de rechazar H_0 y aceptar H_1 . Tal región se llama **región crítica** o de rechazo.

6.2.5. Tipos de contraste: bilateral o de colas

Definición 6.5:

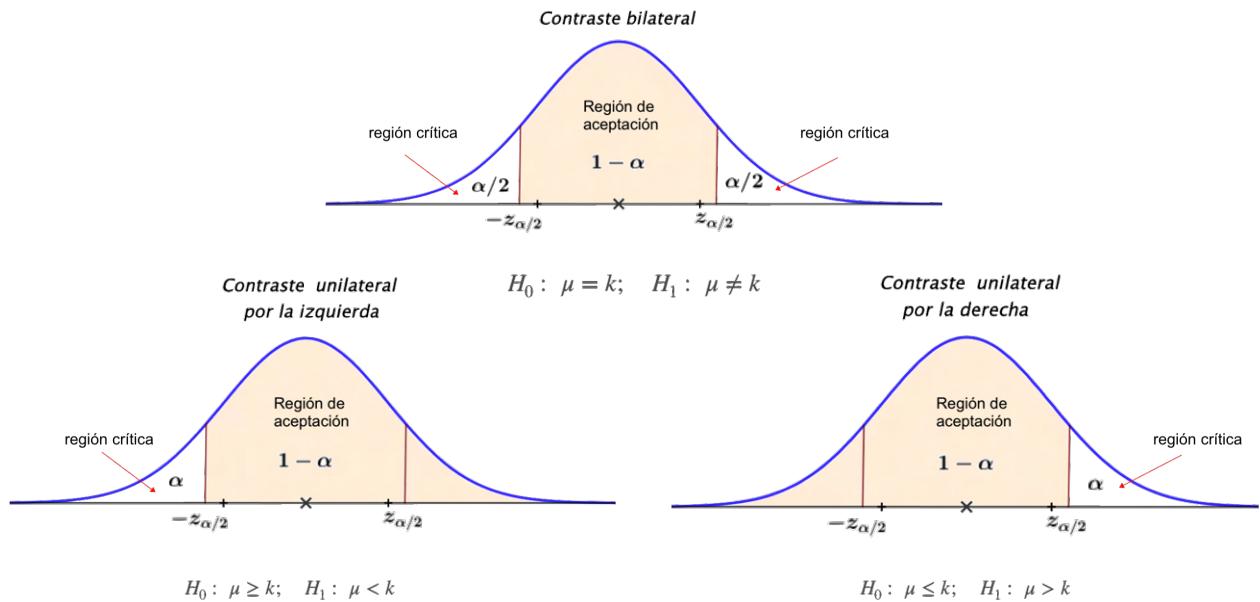
Distinguiremos entre dos tipos de contraste o test, que determinan la región de aceptación y la región de rechazo:

- **Contraste Bilateral (o de dos colas):**

En este caso la región de rechazo o región crítica está formada por los dos extremos fuera del intervalo. Dicho caso se presenta cuando la hipótesis nula es del tipo $H_0 : \mu = k$ ó $p = k$ y la hipótesis alternativa, por tanto, es del tipo $H_1 : \mu \neq k$ ó $p \neq k$

- **Contraste Unilateral (o de una cola):**

En este caso la región de rechazo o región crítica está formada por sólo uno de los extremos fuera del intervalo. Dicho caso se presenta cuando la hipótesis nula es del tipo $H_0 : \mu \geq k$ ó $p \geq k$ y la hipótesis alternativa, por tanto, es del tipo $H_1 : \mu < k$ ó $p < k$. (El sentido de las desigualdades pueden cambiar).



6.3. Metodología general de un test de hipótesis

Definición 6.6:

Algoritmo a seguir en un contraste de hipótesis:

1. Enunciar la hipótesis nula H_0 y la alternativa H_1 . Ambas hipótesis deben ser mutuamente excluyentes.

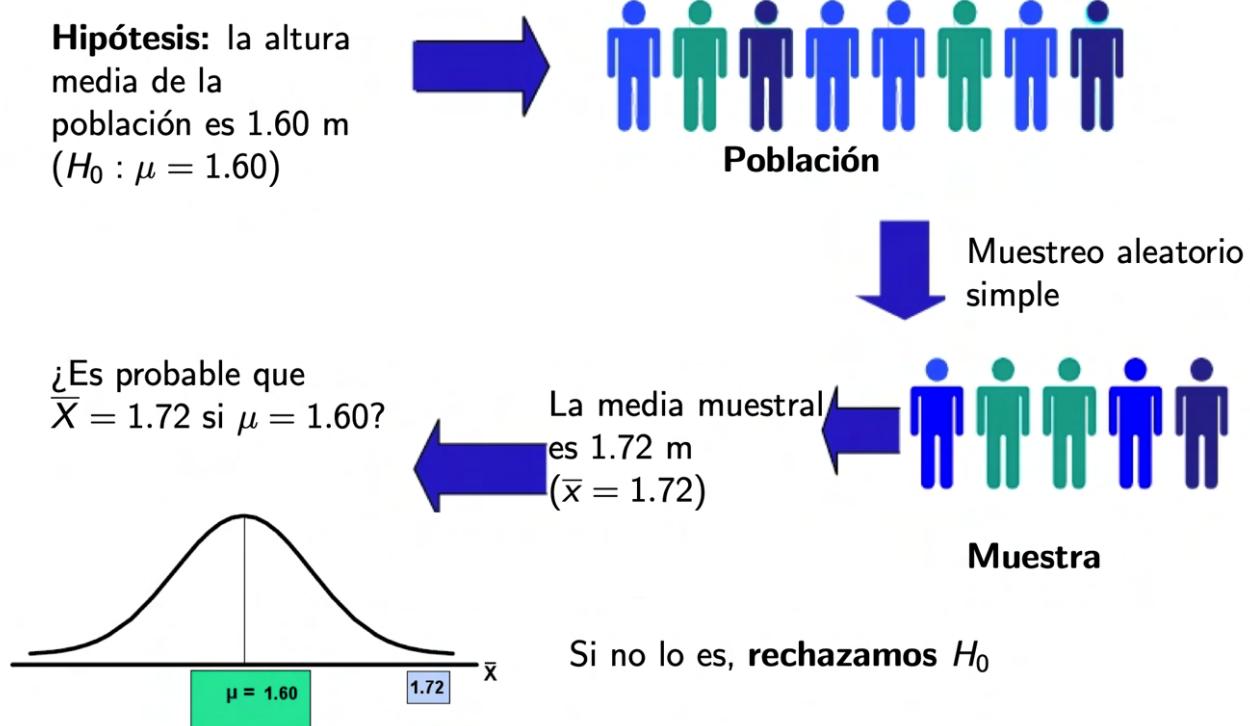
Una vez enunciadas, se analizará si el contraste es bilateral (la hipótesis alternativa es del tipo \neq) o si se trata de un contraste unilateral (la hipótesis alternativa es del tipo $>$ ó $<$).

2. Se elige un estadístico cuya distribución muestral es conocida. En nuestro caso será la media o la proporción muestral. Este estadístico se llama estadístico de contraste.
3. Determinar, a partir del nivel de confianza, $1 - \alpha$, o del de significación, α , el valor de $<_\alpha$ para contrastes bilaterales o el de $z_{\alpha/2}$ para contrastes unilaterales, y con dichos valores se construyen las regiones de aceptación correspondientes:

$$(-z_{\alpha/2}, z_{\alpha/2}) : \text{ c. bilaterales; } (-z_\alpha, +\infty) \text{ o } (-\infty, z_\alpha) : \text{ c. unilaterales}$$

4. Calcular el valor concreto del estadístico de contraste a partir de la muestra.
5. Aplicar el test, es decir, dependiendo de si el estadístico de contraste cae en la región de aceptación no rechazar la hipótesis nula H_0 y, si cae fuera, rechazar H_0 y aceptar la hipótesis alternativa H_1 .

Proceso del contraste de hipótesis



En los siguientes apartados veremos los tipos de test de hipótesis más habituales son el ‘test para la media de una población’, el ‘test para la proporción de una población’ y, como ampliación, el ‘test de comparación de medias de dos poblaciones’.

6.4. Contraste de hipótesis para la media poblacional

(p es la proporción de la población con una característica determinada).

Teorema 6.3:

Planteamiento para el contraste de hipótesis para la media de una población.

Contraste bilateral

c. unilateral derecha

c. unilateral izquierda

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

$$\begin{cases} H_0 : \mu < \mu_0 \\ H_1 : \mu \geq \mu_0 \end{cases}$$

$$\begin{cases} H_0 : \mu > \mu_0 \\ H_1 : \mu \leq \mu_0 \end{cases}$$

Si la población de partida es normal (o $n \geq 30$), $N(\mu, \sigma)$, entonces, la distribución de las medias muestrales es $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

Si la hipótesis nula es cierta, entonces

$$\bar{x} \rightsquigarrow N(\mu, \sigma/\sqrt{n}) \text{ y, tipifiando,}$$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0, 1) \text{ (estadístico de contraste).}$$

Si desconocemos la desviación típica de la población, σ , usaremos la desviación típica de la muestra, s ($n \geq 30$).

Ejemplo 6.3:

Se cree que el tiempo medio de ocio que dedican al día los estudiantes de Bachillerato sigue una distribución normal de media 350 minutos y desviación típica 60 minutos. Para contrastar esta hipótesis, se toma una muestra aleatoria formada por 100 alumnos, y se observa que el tiempo medio de ocio es de 320 minutos. Con un nivel de significación del 10 %, ¿se contradice la afirmación inicial?

- 1.- Formulación de hipótesis: $H_0 : \mu = 350$; $H_1 : \mu \neq 350$, bilateral.
- 2.- Estadístico de contraste (media muestral): $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0, 1)$
- 3.- Región de aceptación: $\alpha = 10\% \rightarrow z_{\alpha/2} = 1.645 \Rightarrow$; aceptación en $(-1.645, 1.645)$
- 4.- Valor del estadístico de contraste para la muestra: $z = \frac{320 - 350}{60/\sqrt{100}} = -5$
- 5.- Decisión: $z = -5 \notin (-1.645, 1.645) \rightarrow$ se rechaza la hipótesis nula: *hay evidencias estadísticas significativas para suponer que el tiempo medio diario de ocio de los alumnos de Bachillerato no es de 350 minutos.*

Ejercicio resuelto 6.1. Una encuesta, realizada a 64 empleados de una fábrica, concluyó que el tiempo medio de duración de un empleo en la misma era de 6.5 años con una desviación típica de 4. ¿Sirve esta afirmación para aceptar, con un nivel de significación del 1 %, que el tiempo medio de empleo en esa fábrica es menor o igual que 6?

- 1.- Formulación de hipótesis: $H_0 : \mu \leq 6$; $H_1 : \mu \geq 6$, unilateral derecha.
- 2.- Estadístico de contraste (media muestral): $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0, 1)$
- 3.- Región de aceptación: $\alpha = 0.1 = 10\% \rightarrow p(z < z_\alpha) = 0.9 \Rightarrow z_\alpha = 1.28$; aceptación en $(-\infty, 1.28)$

- 4.- *Valor del estadístico de contraste para la muestra: $z = \frac{6.5 - 6}{4/\sqrt{64}} = 1$*
- 5.- *Decisión: $z = 1 \in (-\infty, 1.28)$ → no se puede rechazar la hipótesis nula: no hay evidencias estadísticas significativas que indiquen que el tiempo medio de empleo en dicha fábrica no sea inferior a 6 años.*

Observaciones:

- ▷ En la práctica, la muestra se toma después de haber formulado las hipótesis, con el fin de que el resultado de la muestra no influya en el planteamiento de éstas.
- ▷ Al disminuir el nivel de significación, α , aumenta la región de aceptación y por tanto es posible que una hipótesis que se rechace con un nivel de significación del 10% no se pueda rechazar a un nivel de significación del 5%.
- ▷ Cuanto más ‘fuera’ de la región de aceptación se encuentre nuestro estadístico de contraste, con mayor confianza podremos rechazar la hipótesis nula y por tanto mayor seguridad tendremos en que nuestra decisión es la correcta. De la misma manera, cuanto más ‘dentro’ de la región de aceptación se encuentre, mayor seguridad tendremos a la hora de no rechazar la hipótesis nula.

6.5. Contraste de hipótesis para la proporción poblacional

Teorema 6.4:

Planteamiento para el contraste de hipótesis para la proporción de una población.

Contraste bilateral

c. unilateral derecha

c. unilateral izquierda

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases} \quad \begin{cases} H_0 : p < p_0 \\ H_1 : p \geq p_0 \end{cases} \quad \begin{cases} H_0 : p > p_0 \\ H_1 : p \leq p_0 \end{cases}$$

Si la población de partida es normal (o $n \geq 30$), la distribución de las proporciones muestrales es $\hat{p} \sim N \left(p_0, \sqrt{\frac{p_0 (1 - p_0)}{n}} \right)$

Si la hipótesis nula es cierta, entonces

$$\hat{p} \sim N \left(p_0, \sqrt{\frac{p_0 (1 - p_0)}{n}} \right) \text{ y, tipificando,}$$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 (1 - p_0)}{n}}} \rightsquigarrow N(0, 1) \text{ (estadístico de contraste).}$$

Las consideraciones realizadas anteriormente siguen siendo válidas en este caso.

Ejemplo 6.4:

El ayuntamiento de una ciudad afirma que el 65 % de los accidentes juveniles de los fines de semana son debidos al alcohol. Un investigador decide contrastar dicha hipótesis, para lo cual toma una muestra formada por 35 accidentes y observa que 24 de ellos han sido debidos al alcohol. Con un nivel de confianza del 99 %, ¿qué podemos decir sobre la afirmación del ayuntamiento?

1- Hipótesis: $H_0 : p = 0.65$; $H_1 : p \neq 0.65$, bilateral

2- Estadístico de contraste para la proporción: $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 (1 - p_0)}{n}}}$

3- Zona de aceptación para la significación requerida: $1 - \alpha = 0.99 \rightarrow z_{\alpha/2} = 2.575$, zona de aceptación en $(-2.575, 2.575)$

4- Cálculo del estadístico de contraste: a partir de la muestra, $\hat{p} = 24/35 = 0.686$

$$z = \frac{0.686 - 0.65}{\sqrt{\frac{0.65 \cdot 0.35}{35}}} = 0.44$$

5- Decisión: como $z = 0.44 \in (-2.575, 2.575)$, no podemos rechazar la hipótesis nula, no hay evidencias estadísticas significativas que indiquen que la afirmación del ayuntamiento no sea correcta.

Ejercicio resuelto 6.2. *Un investigador, utilizando información de anteriores comicios, sostiene que, en una determinada zona, el nivel de abstención en las próximas elecciones es del 40 % como mínimo. Se elige una muestra aleatoria de 200 individuos para los que se concluye que 75 estarían dispuestos a votar.*

Determinar, con un nivel de significación del 1 %, si se puede admitir como cierta la afirmación del investigador.

1- Hipótesis: $H_0 : p \geq 0.4$; $H_1 : p < 0.4$, unilateral izquierda

2- Estadístico de contraste para la proporción: $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 (1 - p_0)}{n}}}$

3- Zona de aceptación para la significación requerida: $\alpha = 0.01 \rightarrow p(z < z_\alpha) = 0.99$, mirando las tablas de la normal típica al revés, $z_\alpha = 2.33$, zona de aceptación en $(-2.33, \infty)$

4- Cálculo del estadístico de contraste: a partir de la muestra, $\hat{p} = 125/200 = 0.375$. ¡Ojo!: 75 dispuestos a votar son 125 que se abstendrán.

$$z = \frac{0.625 - 0.4}{\sqrt{\frac{0.4 \cdot 0.6}{200}}} = 6.5$$

5- Decisión: como $z = 6.5 \in (-2.33, \infty)$, no podemos rechazar la hipótesis nula, es decir, no hay evidencias estadísticas significativas que indiquen que el investigador no tiene razón.

6.6. C.H. para las medias de dos poblaciones

Contraste de hipótesis para la diferencia de medias

Teorema 6.5:

Consideramos dos distribuciones normales: $N(\mu_x, \sigma_x)$ y $N(\mu_y, \sigma_y)$

Tenemos que contrastar si las medias son iguales (hipótesis nula), $\mu_x = \mu_y$, con lo que el planteamiento es:

$$H_0 : \mu_x = \mu_y; \quad H_1 : \mu_x \neq \mu_y \quad \text{contraste bilateral}$$

Para un nivel de confianza α , la región de aceptación es $(-z_{\alpha/2}, z_{\alpha/2})$

Tomamos una muestra en cada distribución, de tamaños n_x en $N(\mu_x, \sigma_x)$ y n_y en $N(\mu_y, \sigma_y)$

Sabemos (tema anterior) que $\bar{x} - \bar{y} \rightsquigarrow N\left(\mu_x - \mu_y, \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}\right)$

Si H_0 se cumple, $\bar{x} - \bar{y} \in N\left(0, \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}\right)$ y, al tipificar, obtenemos como estadístico de contraste: $z = \frac{\bar{x} - \bar{y} - 0}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \rightsquigarrow N(0, 1)$

Al igual que pasa con el test de una media, en caso de desconocer las desviaciones típicas de las poblaciones, y si las muestras son suficientemente grandes, se pueden sustituir por las desviaciones típicas muestrales.

Ejemplo 6.5:

A los 100 alumnos de una clase se les separa en dos grupos: aquellos que practican habitualmente un deporte y los que no practican ninguno, formando cada grupo 60 y 40 alumnos, respectivamente. Les medimos la altura, obteniendo para el primer grupo una media de 1.80 m. y una desviación típica 0.08 m., y para el segundo grupo una media de 1.76 m. y una desviación típica de 0.10 m. Suponiendo que la variable aleatoria altura sigue una distribución normal en los dos grupos, ¿es posible afirmar, con un nivel de confianza del 95 %, que hay diferencia de altura entre los alumnos que practican algún deporte y los que no?

$$N(1.80, 0.08) \rightarrow n_x = 60; \quad N(1.76, 0.10) \rightarrow n_y = 40; \quad 1 - \alpha = 95\%; \quad \alpha = 5\%$$

1- Hipótesis: $H_0 : \mu_x = \mu_y$; $H_1 : \mu_x \neq \mu_y$, contraste bilateral.

2- Estadístico de contraste:
$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \rightsquigarrow N(0, 1)$$

3- Para $\alpha = 0.05 \rightarrow z_{\alpha/2} = 1.96$; zona de aceptación: $(-1.96, 1.96)$

4- Cálculo del estadístico de contraste:
$$z = \frac{1.80 - 1.76}{\sqrt{\frac{0.08^2}{60} + \frac{0.10^2}{40}}} = 2.105$$

5- Como $z = 2.105 \notin (-1.96, 1.96)$ rechazamos la hipótesis nula, es decir, existen diferencias estadísticamente significativas entre la altura media de los chicos que practican deporte y la de los que no.

Ejercicio resuelto 6.3. A fin de determinar si existen diferencias significativas entre dos grupos de estudiantes, realizamos el mismo examen a 30 estudiantes del primer grupo y a 35 del segundo, obteniendo media 5.5 y desviación típica 0.5, para el primer grupo, y 5.2 de media y 1 de desviación típica para el segundo grupo. ¿Qué conclusión se obtiene con un nivel de significación del 1 %?

$\mu_x = 5.5, \sigma_x = 0.5 \rightarrow n_x = 30; m_y = 5.2, \sigma_y = 1 \rightarrow n_y = 35$, contraste bilateral.

1- Hipótesis: $H_0 : \mu_x = \mu_y; H_1 : \mu_x \neq \mu_y$

2- Estadístico de contraste: $z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \rightsquigarrow N(0, 1)$

3- Para $\alpha = 0.01 \rightarrow z_{\alpha/2} = 2.575$; zona de aceptación: $(-2.575, 2.575)$

4- Cálculo del estadístico de contraste: $z = \frac{5.5 - 5.2}{\sqrt{\frac{0.5^2}{30} + \frac{1^2}{35}}} = 1.56$

5- Decisión: como $z = 1.56 \in (-2.575, 2.575)$, no podemos rechazar la hipótesis nula, es decir, no existen diferencias estadísticamente significativas entre ambos cursos.

6.7. Ejercicios

Ejercicio 6.1. La altura en cm de las cañas producidas por una determinada variedad en cada cosecha es una variable aleatoria que sigue una ley normal con desviación típica $\sigma = 16$ cm. Para contrastar si la altura media de las cañas de la última cosecha es de 170 cm, se ha tomado una muestra aleatoria de 64 de estas cañas y se han medido sus longitudes, resultando como media muestral $\bar{x} = 166$ cm.

¿Son suficientes estos datos para rechazar que la altura media de las cañas de la última cosecha es de 170 cm, a un nivel de significación $\alpha = 0.05$?

$H_0 : \mu = 170; H_1 : \mu \neq 170$ Bilateral

$\alpha = 0.05; N(170, 16); n = 64; \bar{x} = 166$

$\alpha = 0.05 \rightarrow z_{\alpha/2} = 1.96$ Región de aceptación: $(-1.96, 1.96)$

Estadístico de contraste: $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{166 - 170}{16/\sqrt{64}} = -2$

Como $-2 \notin (-1.96, 1.96)$, rechazamos la hipótesis nula y aceptamos la alternativa. Con lo cual, aceptamos que la altura de las cañas no miden 170 cm, al nivel de significación 0'05 (pudiendo haber cometido un error del tipo I).

Ejercicio 6.2. Un comerciante ha observado durante un largo periodo de tiempo que sus beneficios semanales se distribuyen según una ley normal con una media de 5000 euros y una desviación típica de 520 euros. A finales del año pasado se abrió un supermercado frente a su comercio y él cree que su beneficio semanal medio ha disminuido desde entonces. Para contrastar esta suposición, ha tomado una muestra aleatoria de 16 semanas del año actual y ha encontrado que el beneficio semanal medio de esa muestra es de 4700 euros. ¿Puede afirmarse, a un nivel de significación $\alpha = 0.01$, que estos datos avalan la creencia del comerciante?

$$H_0 : \mu \geq 500; \quad H_1 : \mu < 5000 \text{ Unilateral derecha.}$$

$$\alpha = 0.01; \quad N(5000, 520); \quad n = 16; \quad \bar{x} = 4700$$

$\alpha = 0.01 \rightarrow p(z > -z_\alpha) = p(z < z_\alpha) = 1 - \alpha = 0.99 \rightarrow z_\alpha = 2.33$, por lo que la región de aceptación es $(-2.33, +\infty)$

$$\text{Estadístico de contraste: } z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{4700 - 5000}{520/\sqrt{16}} = -2.30$$

Como $-2.3 \in (-2.33, +\infty)$, se encuentra en la zona de aceptación. Por lo tanto, no podemos rechazar la hipótesis nula, es decir, no se puede afirmar, al nivel 0.01, que los datos de la muestra apoyen la creencia de que el nuevo supermercado ha disminuido el beneficio semanal medio del comerciante.

Ejercicio 6.3. Sólo el 75 % de los alumnos de un centro de enseñanza realizan correctamente un test psicotécnico que lleva utilizándose mucho tiempo. Para tratar de mejorar este resultado, se modificó la redacción del test, y se propuso a un grupo de 120 alumnos de ese centro, elegidos al azar. De los 120 alumnos a los que se le pasó el nuevo test, lo realizaron correctamente 107. ¿Podemos afirmar que la nueva redacción del test ha aumentado la proporción de respuestas correctas, a un nivel de significación $\alpha = 0.025$?.

$$H_0 : p \leq 0.75; \quad H_1 : p > 0.75 \text{ Unilateral izquierdo.}$$

$$\alpha = 0.025 \quad p = 0.75 \quad n = 120 > 30; \quad \hat{p} = 107/120$$

$\alpha = 0.025 \rightarrow z_\alpha : / p(z < z_\alpha) = 1 - \alpha = 0.975 \Rightarrow z_\alpha = 1.96$ y la región de aceptación es $(-\infty, 1.96)$

$$\text{Estadístico de contraste: } z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{\sqrt{n}}}} = \frac{107/120 - 0.75}{\sqrt{0.75 \cdot 0.25/120}} = 3.58$$

Como $3.58 \notin (-\infty, 1.96)$, el estadístico de contraste se encuentra en la zona de rechazo o región crítica. Por lo tanto, rechazamos la hipótesis nula y aceptamos la alternativa: podemos afirmar que la nueva redacción del test ha aumentado la proporción de respuestas correctas, a un nivel confianza del 97.5 % (significación de 2.5 %).

Ejercicio 6.4. El peso en vacío de los envases fabricados por una empresa, según su método usual, es una variable aleatoria que sigue una ley normal con media 20 gramos y una desviación típica de 1 gramo.

Se desea contrastar si un nuevo proceso de fabricación no aumenta dicho peso medio. Para ello, se eligen al azar 25 envases fabricados por la nueva técnica y se encuentra que la media de su peso en vacío es de 20.5 gramos. ¿Se puede afirmar, a un nivel de significación del 2%, que el nuevo proceso ha aumentado el peso medio de los envases?

$$H_0 : \mu \leq 20; \quad H_1 : \mu > 20 \quad \text{Unilateral derecho.}$$

$$\alpha = 0.02; \quad N(20, 1); \quad n = 25; \quad \bar{x} = 20.5$$

$$\alpha = 0.02 \rightarrow p(z > z_\alpha) = 1 - \alpha = 0.98 \rightarrow z_\alpha = 2.06, \text{ la región de aceptación es } (-\infty, 2.06)$$

$$\text{Estadístico de contraste: } z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{20.5 - 20}{1/\sqrt{25}} = 2.5$$

Como $2.5 \notin (-\infty, 2.06)$, rechazamos la hipótesis nula y aceptamos la alternativa. Por lo tanto, a la vista de los datos obtenidos en la muestra, se puede afirmar, al nivel $\alpha = 0.02$, que el nuevo proceso ha aumentado el peso medio de los envases (podemos estar cometiendo un error de tipo II).

Ejercicio 6.5. En unas elecciones municipales de una ciudad, el 42% de los votantes dieron su voto al partido A. En una encuesta realizada un año después a 500 personas con derecho a voto, sólo 184 votarían al partido A. Con estos datos, ¿puede afirmarse que ha disminuido la proporción de votantes a ese partido? Responder a la pregunta anterior con niveles de significación α del 1%, del 2.5% y del 0.1%

$$H_0 : p \geq 0.42; \quad H_1 : p < 0.42 \quad \text{Contraste unilateral derecho.}$$

$$\alpha = 0.01; \quad \alpha = 0.025; \quad \alpha = 0.001; \quad p = 0.42 \quad n = 500 > 30; \quad \hat{p} = 184/500$$

$$\alpha = 0.01 \rightarrow z_\alpha/p(z > -z_\alpha) = p(z < z_\alpha) = 1 - \alpha = 0.99 \rightarrow z_\alpha = 2.33$$

$$\alpha = 0.025 \rightarrow z_\alpha/p(z > -z_\alpha) = p(z < z_\alpha) = 1 - \alpha = 0.975 \rightarrow z_\alpha = 1.96$$

$$\alpha = 0.001 \rightarrow z_\alpha/p(z > -z_\alpha) = p(z < z_\alpha) = 1 - \alpha = 0.999 \rightarrow z_\alpha = 3.09$$

$$\text{Estadístico de contraste: } z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{184/500 - 0.42}{\sqrt{[184/500 \cdot (500 - 184)/500]/120}} = -2.355$$

Decisiones:

Para $\alpha = 0.01 \rightarrow -2.355 \notin (-2.33, +\infty)$

Para $\alpha = 0.025 \rightarrow -2.355 \notin (-1.96, +\infty)$

Para $\alpha = 0.001 \rightarrow -2.355 \in (-3.09, +\infty)$

En el primer y segundo caso ($\alpha = 0.01 \wedge \alpha = 0.025$) rechazamos la hipótesis nula y aceptamos la alternativa, ha disminuido la proporción de votantes. En el tercer caso ($\alpha = 0.001$) el estadístico de contraste cae dentro de la región de aceptación y no podemos rechazar la hipótesis nula, no podemos afirmar que haya disminuido la proporción de votantes con una significación del 0.001.

Resumiendo, los datos permiten afirmar que ha disminuido la proporción de votantes al partido A a los niveles 0.025 y 0.01, pero no ha disminuido la proporción al nivel 0.001.

Ejercicio 6.6. *Se sabe que la longitud en cm de una determinada especie de coleópteros sigue una distribución normal de varianza 0.25 cm^2 . Capturados 6 ejemplares de dicha especie, sus longitudes (en cm) fueron: 2.75; 1.72; 2.91; 2.6; 2.64; 3.34*

¿Se puede aceptar la hipótesis de que la población tiene una longitud media de 2.656 cm? Usar $\alpha = 0.05$.

$$H_0 : \mu = 2.656; \quad H_1 : \mu \neq 2.656 \quad \text{Bilateral}$$

$$\alpha = 0.05; \quad N(2.656, \sqrt{0.25}) = N(2.656, 0.5); \quad n = 6; \quad \bar{x} = (2.75 + 1.72 + 2.91 + 2.6 + 2.64 + 3.34)/6 = 2.66$$

$$\alpha = 0.05 \rightarrow z_{\alpha/2} = 1.96 \quad \text{Región de aceptación: } (-1.96, 1.96)$$

$$\text{Estadístico de contraste: } z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{2.66 - 2.566}{0.5/\sqrt{6}} = 0.019$$

Como $0.019 \in (-1.96, 1.96)$, el valor observado del estadístico de prueba se encuentra en la región de aceptación. Por lo tanto, no rechazamos la hipótesis nula, la longitud media de los coleópteros es de 2.656 al nivel de significación 0'05, pudiendo haber cometido un error del tipo I.

Ejercicio 6.7. *El 40 % de los escolares de cierto país suelen perder al menos un día de clase a causa de gripes y catarros. Sin embargo, un estudio sobre 1000 escolares revela que en el último curso hubo 450 en tales circunstancias. Las autoridades defienden que el porcentaje del 40 % para toda la población de escolares se ha mantenido. Contrastar, con un nivel de significación del 5 %, la hipótesis defendida por las autoridades sanitarias frente a que el porcentaje ha aumentado, como parecen indicar los datos, explicando claramente a qué conclusión se llega.*

$$H_0 : p \leq 0.4; \quad H_1 : p > 0.4 \quad \text{Bilateral derecha.}$$

$$\alpha = 5 \% \quad p = 0.4 \quad n = 1000, \quad \hat{p} = 450/1000 = 0.45$$

$$\alpha = 0.05 \rightarrow z_{\alpha}/p(z > z_{\alpha}) = 1 - \alpha = 0.95 \Rightarrow z_{\alpha} = 1.645 \text{ con lo que la región de aceptación es } (1.645, +\infty).$$

Estadístico de contraste: $z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{\sqrt{n}}}} = \frac{0.45 - 0.4}{\sqrt{0.4 \cdot 0.6/1000}} = 3.227$

Como $z = 3.227 \notin (-1.96, 1.96)$ cae en la zona de rechazo. Por lo tanto, rechazamos la hipótesis nula y aceptamos la alternativa. Con lo cual, con una probabilidad del 5 % de equivocarnos, afirmamos que más del 40 % de los alumnos falta un día a clase por la gripe (podemos estar cometiendo un error del tipo II).

Ejercicio 6.8. Una de las entradas a cierta ciudad sufría constantemente retenciones de tráfico, de forma que el tiempo de espera en la cola formada por el semáforo allí instalado seguía una distribución Normal de media 10 minutos y desviación típica 4 minutos. Con el fin de descongestionar ese punto y bajar la media de tiempo de espera, se habilitó una vía de acceso auxiliar. Transcurrida una semana se hizo un estudio sobre 36 vehículos y se obtuvo que el tiempo medio de espera en el citado semáforo fue de 8.5 minutos. Las autoridades municipales mostraron su satisfacción y dijeron que la medida había funcionado, pero la opinión pública, sin embargo, defiende que la situación sigue igual. Suponiendo que la desviación típica se ha mantenido:

- Plantee un test para contrastar la hipótesis defendida por la opinión pública frente a la de los responsables municipales. Si se concluye que la media de tiempo de espera bajó y realmente no lo hizo, ¿cómo se llama el error cometido?
- ¿A qué conclusión se llega con un nivel de significación del 5 %?
- ¿A qué conclusión se llega con un nivel de significación del 1 %?

$$H_0 : \mu \geq 10; \quad H_1 : \mu < 10 \text{ unilateral derecho.}$$

$$\alpha = 5 \% \wedge 1 \% ; \quad N(10, 4); \quad n = 36, \quad \delta x = 8.5$$

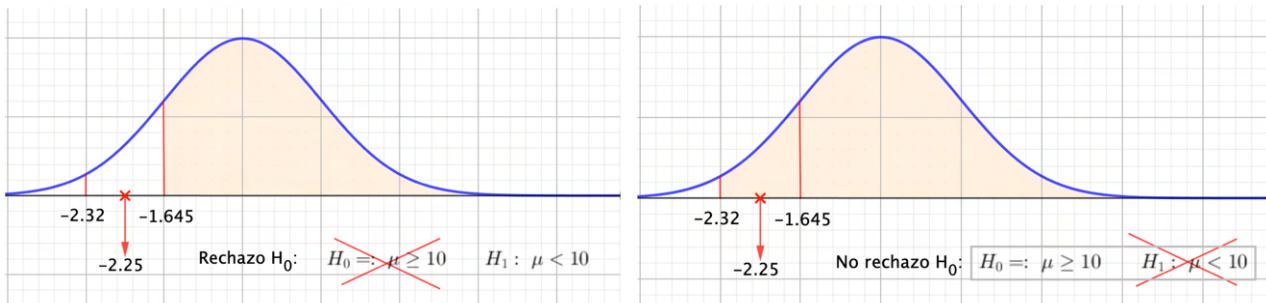
— a) Para una significación α , buscamos el $z\alpha$ tal que $p(z < -z\alpha) = p(z > z\alpha) = 1 - \alpha$, lo cual determina la región de aceptación de la hipótesis nula: $(-z\alpha, +\infty)$

$$\text{Seguidamente calculamos el estadístico de contraste: } z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{8.5 - 10}{4/\sqrt{36}} = -2.25$$

La decisión consiste en no rechazar la hipótesis nula si $z \in (-z\alpha, +\infty)$ (a riesgo de cometer un error de tipo I) o sí rechazar la hipótesis nula y aceptar la hipótesis alternativa si $z \notin (-z\alpha, +\infty)$ (corremos el riesgo de equivocarnos con esta decisión y cometer un error de tipo II).

- Para $\alpha = 0.05 \rightarrow z\alpha = 1.645 \Rightarrow z = -2.25 \notin (-1.645, +1.645)$ y rechazamos la hipótesis nula, no podemos asegurar $\mu \geq 10$ ($\mu < 10$), y las autoridades tienen razón (no ha aumentado el tiempo de espera).
- Para $\alpha = 0.01 \rightarrow z\alpha = 2.32 \Rightarrow z = -2.25 \in (-2.32, +2.32)$ y no rechazamos la hipótesis nula ($\mu \geq 10$), la opinión pública tiene razón.

Por lo tanto, rechazamos la hipótesis nula al nivel del 5% y no a rechazamos al nivel del 1%. Con lo cual, se puede afirmar, al nivel 0.05, que las autoridades llevan razón y el nivel de espera no supera los 10 minutos; sin embargo los conductores llevan razón y tienen que esperara más de 10 minutos al nivel del 1%.



Ejercicio 6.9. El alcalde de una ciudad prometió, en su programa electoral, oponerse a la construcción de una central de tratamiento de ciertos residuos, puesto que en aquel momento sólo un 10% de los ciudadanos estaban a favor de la central de tratamiento de residuos. En los últimos días se ha encuestado a 100 personas de las cuales 14 están a favor de la central. El alcalde afirma sin embargo que el porcentaje de ciudadanos a favor sigue siendo del 10% o incluso ha disminuido. ¿Tiene razón el alcalde con un nivel de significación del 2%?

$$H_0 : p \leq 0.1; \quad H_1 : p > 0.1 \text{ Unilateral izquierda}$$

$$\alpha = 0.02; \quad p = 0.1; \quad n = 100, \quad \hat{p} = 14/100 = 0.14$$

$\alpha = 0.02 \rightarrow p(z < z_\alpha) = 1 - \alpha = 0.98 \Rightarrow z_\alpha = 2.06$ con lo que la región de aceptación de la hipótesis nula es $(-\infty, 2.06)$

$$\text{Estadístico de contraste: } z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{\sqrt{n}}}} = \frac{0.14 - 0.1}{\sqrt{0.1 \cdot 0.9/100}} = 1.333$$

Decisión: como el valor crítico, $z = 1.333 \in (-\infty, 2.06)$, se encuentra en la zona de aceptación, no rechazamos la hipótesis nula (no aceptamos la alternativa). Con lo cual, para un nivel de significación 0.02, el alcalde acierta en que el porcentaje de ciudadanos a favor sigue siendo del 10% o incluso ha disminuido.

Ejercicio 6.10. Una máquina de envasado automático llena en cada saco una cierta cantidad de determinado producto. Se seleccionan 20 sacos, se pesa su contenido y se obtienen los siguientes resultados (en kilos): 49; 50; 49; 50; 50; 50; 49; 50; 50; 50; 49; 50; 50; 51; 52; 48; 50; 51; 51

A partir de esta información y suponiendo que la variable, peso de cada saco, se distribuye normalmente con desviación típica 1 kg:

- a) ¿Se puede admitir que el peso medio de los sacos que llena la máquina es de aproximadamente 51 kg? (Usar $\alpha = 0.01$)
- b) ¿Se puede admitir que el peso medio de los sacos que llena la máquina es menor de 50 kg? (Usar $\alpha = 0.05$)

$n = 20$; $\bar{x} = 50$, calculada de los datos del problema.

Peso de los sacos $N(\mu, 1)$

— a) $H_0 : \mu = 51$; $H_1 : \mu \neq 51$; $\alpha = 0.01$, bilateral.

$$\text{Estadístico de contraste: } z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{50 - 51}{1/\sqrt{20}} = -4.47$$

Zona de aceptación: $\alpha = 0.01 \rightarrow z_{\alpha/2} = 2.57 \Rightarrow (-2.57, 2.57)$

El valor observado del estadístico de prueba $z = -4.47$ se encuentra en la zona de rechazo ($z = -4.47 \notin (-2.57, 2.57)$), por lo tanto, tomamos la decisión de rechazar la hipótesis nula y aceptamos la alternativa, el peso de la máquina de envasado no es de 51 Kg para este nivel de significación.

— b) $\mu \geq 50$; $H_1 : \mu < 50$; $\alpha = 0.05$, unilateral derecho.

$$\text{Estadístico de contraste: } z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{50 - 50}{1/\sqrt{20}} = 0$$

Zona de aceptación: $\alpha = 0.05 \rightarrow z_\alpha / p(z < -z_\alpha) = p(z > -z_\alpha) = 1 - \alpha = 0.95 \rightarrow z_\alpha = 1.645 \Rightarrow (-1.645, 1.645)$

El valor observado del estadístico de prueba $z = 0$ se encuentra en la zona de aceptación ($z = 0 \notin (-1.645, 1.645)$), por lo tanto, tomamos la decisión de no rechazar la hipótesis nula, el peso de la máquina de envasado no es menor 50 Kg para este nivel de significación.

Ejercicio 6.11. El consumo de cierto producto sigue una distribución normal con varianza 300. A partir de una muestra de tamaño 25 se ha obtenido una media muestral igual a 180.

a) Halle un intervalo de confianza al 95 % para la media del consumo.

b) ¿Se podría afirmar que el consumo medio de este producto no llega a 200? (Usar $\alpha = 0.05$)

— a) $(1 - \alpha) \cdot 100\% = 95\% \rightarrow z_{\alpha/2} = 1.96$

$$\text{El intervalo característico es: } \mu \in \bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 180 \pm 1.96 \cdot \frac{\sqrt{300}}{\sqrt{25}} \Rightarrow (173.21, 186.79)$$

— b) $H_0 : \mu \geq 200$; $H_1 : \mu < 200$ Unilateral derecha.

Para $\alpha = 0.05$, $z_\alpha = 1.64$ y la región de aceptación es $(-1.645, +\infty)$

$$\text{El estadístico de contraste es } z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{180 - 300}{\sqrt{300}/\sqrt{25}} = -5.773$$

Como el estadístico de contraste está fuera de la región de aceptación, está en la región de rechazo o crítica ($z = -5.773 \notin (-1.645, +\infty)$), rechazamos la hipótesis nula y aceptamos la alternativa, con lo que Se podría afirmar que el consumo medio de este producto no llega a 200 (con un nivel de significación del 5%).

Ejercicio 6.12. Las autoridades educativas publican en un estudio que el 25 % de los estudiantes de Bachillerato de una cierta comunidad autónoma tienen ordenador portátil. A partir de una muestra aleatoria de tamaño 300 se ha obtenido que sólo 70 de ellos tienen ordenador portátil. ¿Se podría asegurar que las autoridades dicen la verdad? (Usar $\alpha = 0.06$)

$$H_0 : p = 0.25; \quad H_1 : p \neq 0.25 \text{ Bilateral.}$$

$$n = 300; \quad \hat{p} = 70/300 = 0.233; \quad \alpha = 0, .06$$

Para $\alpha = 0.06 \rightarrow p(-z_{\alpha/2} < z < z_{\alpha/2}) = 1 - \alpha = 0.94 \rightarrow p(z < z_{\alpha/2}) = 0.97 \Rightarrow z_{\alpha/2} = 1.88$, con lo que la región de aceptación es: $(-1.88, 1.88)$

Calculamos, ahora, el estadístico de contraste:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.233 - 0.25}{\sqrt{0.25 \cdot 0.75/300}} = -0.664$$

Decisión: como $z = -0.664 \in (-1.88, 1.88)$, estamos en la región de no rechazo, no podemos refutar que el 25 % de los estudiantes tiene un portátil en casa con un nivel de significación 0.06, pudiendo haber cometido un error del tipo I.

Ejercicio 6.13. Los alumnos de preescolar tienen una estatura que es una variable aleatoria de media desconocida y desviación típica 16 cm. Si seleccionamos una muestra aleatoria de 100 de tales alumnos y obtenemos una estatura media de 95 cm,

a) ¿Se puede afirmar que la estatura media de los alumnos de preescolar es menor de 95 cm? (Usar $\alpha = 0.01$).

b) ¿Se puede afirmar que la estatura media de los alumnos de preescolares mayor de 100 cm? (Usar $\alpha = 0.05$)

$$N(\mu, 16) \quad n = 100 \quad \bar{x} = 95$$

— a) $H_0 : \mu \geq 95; \quad H_1 : \mu < 95$ Unilateral derecha.

$\alpha = 0.01 \rightarrow z_\alpha : p(z > -z_\alpha) = p(z < z_\alpha) = 1 - \alpha = 0.99 \rightarrow z_\alpha = 2.33$, con lo que la zona de aceptación es: $(-2.33, +\infty)$

El estadístico de contraste es $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{95 - 95}{16/\sqrt{100}} = 0$

Decisión: $z = 0 \in (-2.33, +\infty)$, no podemos rechazar la hipótesis nula, podemos afirmar al nivel 0.01, que los alumnos miden 95 cm o más.

— b) $H_0: \mu \leq 100$; $H_1: \mu > 100$ Unilateral izquierda.

$\alpha = 0.05 \rightarrow z_\alpha: p(z < z_\alpha) = 1 - \alpha = 0.95 \rightarrow z_\alpha = 1.645$, con lo que la zona de aceptación es: $(-\infty, 1.645)$

El estadístico de contraste es $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{95 - 100}{16/\sqrt{100}} = -3.125$

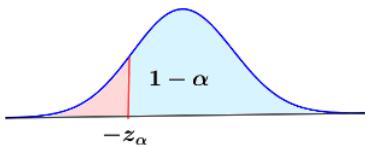
Decisión: $z = -3.125 \in (-\infty, 1.645)$, no podemos rechazar la hipótesis nula, con una probabilidad de equivocarnos del 5 %, afirmamos que los estudiantes miden igual o menos de 95 cm.

Ejercicio 6.14. La conclusión de un contraste de hipótesis realizado con un nivel de significación igual a 0.1 ha sido “aceptar la hipótesis nula H_0 ”. ¿Cuál habría sido la conclusión para un nivel de significación igual a 0.05?

La conclusión sería seguir aceptando la hipótesis nula, puesto que con el nivel de significación del 0.05, la región de aceptación aumenta y, por tanto, el estadístico seguirá estando en dicha región.

Contraste unilateral derecho

$$\begin{aligned} H_0: \mu \geq \bar{x} \quad (p \geq \hat{p}); \\ H_1: \mu < \bar{x} \quad (p < \hat{p}) \end{aligned}$$

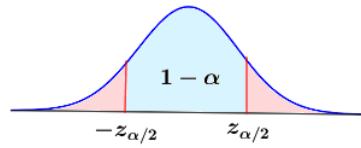


Estadístico de contraste:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad \text{para la media de una población}$$

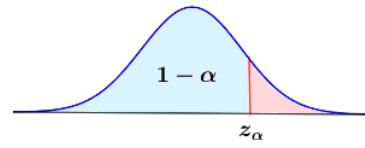
Contraste bilateral

$$\begin{aligned} H_0: \mu = \bar{x} \quad (p = \hat{p}); \\ H_1: \mu \neq \bar{x} \quad (p \neq \hat{p}) \end{aligned}$$



Contraste unilateral izquierdo

$$\begin{aligned} H_0: \mu \leq \bar{x} \quad (p \leq \hat{p}); \\ H_1: \mu > \bar{x} \quad (p > \hat{p}) \end{aligned}$$



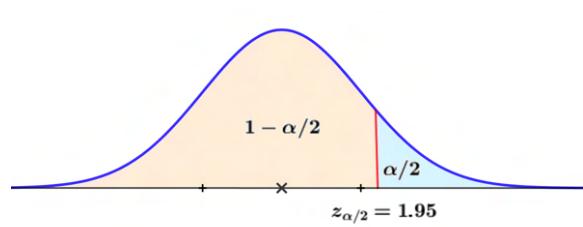
$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \quad \text{para la proporción de una población}$$

Ejercicio resuelto 6.4. En un test de hipótesis para estudiar si el cociente intelectual medio de los estudiantes de una universidad es 113, hemos seleccionado una muestra aleatoria de 180 estudiantes, obteniendo una media de 115. La zona de aceptación obtenida ha sido el intervalo (111.98, 114.02). Por tanto, hemos rechazado la hipótesis. Si $\sigma = 7$, ¿cuál es la probabilidad de haber rechazado la hipótesis, cuando en realidad era verdadera? ¿Cómo se llama este tipo de error?

La semiamplitud del intervalo es $z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

$$\frac{114.02 - 11.98}{2} = z_{\alpha/2} \cdot \frac{7}{\sqrt{180}} \rightarrow z_{\alpha/2} = 1.95$$

$$p(z < 1.95) = 1 - \frac{\alpha}{2} = (\text{tablas}) = 0.9774 \rightarrow \alpha = 0.0512$$



El error que consiste en rechazar H_0 cuando esta es verdadera se llama error de tipo I. La probabilidad de cometerlo es precisamente α , el nivel de significación. En este caso concreto: $\alpha = 0.0512$

6.7.1. Problemas propuestos

PB. 1. En un determinado instituto aseguran que las notas obtenidas por sus alumnos en las pruebas de acceso a la Universidad tienen una media igual o superior a 7 puntos. Pero la media obtenida en una muestra aleatoria de 80 alumnos en los últimos exámenes fue de 6.89 puntos. Si sabemos que la varianza es igual a 4.84, ¿podemos considerar, con un nivel de significación del 1 %, que la afirmación hecha por el instituto es cierta?

$$\text{no rechazamos } H_0 : \mu \geq 7$$

PB. 2. En una determinada región, el número semanal de accidentes de tráfico producido durante el año pasado siguió una distribución normal de media 3.2 y desviación típica 1.3. Se ha llevado a cabo una campaña de prevención contra los accidentes de tráfico y la media semanal de accidentes en las 40 semanas siguientes ha sido de 3.05. Admitiendo que la desviación típica no ha variado, ¿podemos afirmar, con un nivel de significación del 5 %, que la campaña no ha tenido éxito (es decir, que el número de accidentes no ha disminuido con respecto al año anterior)?

$$\text{no rechazamos } H_0 : \mu \geq 3.2$$

PB. 3. Un fabricante garantiza a un laboratorio farmacéutico que sus máquinas producen comprimidos con un diámetro de 25 mm. Una muestra de 100 comprimidos dio como media de los diámetros 25.18 mm. Suponiendo que el diámetro de los comprimidos es una variable aleatoria con distribución normal, de desviación típica 0.89 mm, se desea contrastar, con un nivel de significación del 5 %, si el diámetro medio que afirma el fabricante es correcto. Para ello:

- a) Plantea la hipótesis nula y la hipótesis alternativa del contraste.
- b) Realiza el contraste al nivel de significación indicado.

$H_0 : \mu = 25$; $H_1 : \mu < 25$; (a) rechazamos H_0 : $\mu < 25$

- PB. 4. Un laboratorio ha preparado un elevado número de dosis de cierta vacuna. Se conoce que el peso de dichas dosis se distribuye normalmente, con desviación típica de 0.10 mg. El peso medio de las dosis ha de ser de 0.70 mg. Se requiere la máxima precisión en el peso de las dosis. Por ello, se elige una muestra de 200 dosis y se comprueba su peso medio, que resulta ser de 0.66 mg. Realiza un contraste de hipótesis, con un nivel de significación de 0.05, para decidir si se debe retirar las dosis producidas, o bien la diferencia de peso medio es debida al azar.

rechazamos H_0 , se deben retirar las dosis producidas.

- PB. 5. El concejal de cultura de una determinada localidad afirma que el tiempo medio dedicado a la lectura por los jóvenes entre 15 y 30 años, residentes en dicha localidad, es, como mucho, de 8 horas semanales. Tomando una muestra aleatoria de 100 jóvenes entre 15 y 30 años, se obtuvo que la media de horas semanales que dedicaban a leer era de 8.3, con una desviación típica igual a 1. Con un nivel de significación del 5 %, ¿podemos aceptar la afirmación del concejal?

rechazamos H_0 , no podemos dar por cierta la afirmación del concejal

- PB. 6. Se afirma que, en una determinada localidad, el 20 % de las familias tienen dos o más hijos. Tomando una muestra aleatoria de 120 familias, había dos o más hijos en 22 de ellas. A un nivel de significación de 0.1, ¿podemos rechazar la afirmación?

solo rechazamos H_0 , $d = 0.2$

- PB. 7. El 15 % de los empleados de una gran empresa se declara fumador. Después de llevar a cabo una campaña contra el tabaco durante un año, se quiso comprobar si ésta había sido efectiva. Se hizo una encuesta a 85 empleados elegidos al azar, obteniéndose que 11 de ellos seguían fumando. A un nivel de significación del 0.01, ¿podemos considerar que la proporción de fumadores no ha variado después de la campaña?

no rechazamos H_0 , no hay suficiente evidencia para aceptar que la proporción haya variado

- PB. 8. Hemos realizado 300 lanzamientos con un dado, que sospechamos que está trucado, y hemos obtenido un seis en 71 ocasiones. Con un nivel de significación del 1 %, contrasta la hipótesis de que la probabilidad de obtener seis no es mayor de 1/6.

rechazamos H_0 , aceptamos H_1 : $p < 0.9$

- PB. 9. Hace un año, 3 de cada 10 familias de una determinada población realizaba sus compras habituales en hipermercados. En una encuesta realizada este año entre 105 familias de la localidad escogidas al azar, 34 de ellas afirman que compran habitualmente en hipermercados. Con un nivel de significación del 5 %, contrasta la hipótesis de que el porcentaje no ha aumentado (es decir, que permanece igual o menor que el del año anterior).

no rechazamos H_0 , $d \geq 0.3$

- PB. 10. En las elecciones a la alcaldía de cierta localidad, que se celebraron hace un año, el partido que ganó obtuvo el 57 % de los votos. Recientemente se ha realizado una encuesta, escogiendo al azar a 160 vecinos (mayores de 18 años), 88 de los cuales afirmaban que seguían a favor del alcalde. ¿Podemos considerar, a un nivel de significación de 0.01, que el alcalde no obtendría menor número de votos si se repitieran ahora las elecciones?

no rechazamos H_0 , $d \leq 0.57$

- PB. 11. En un contraste de hipótesis para comprobar si la media de edad de los asistentes a una exposición era de 20 años, hemos seleccionado una muestra aleatoria de 100 asistentes, obteniendo una media de 20.5 años. La zona de aceptación obtenida ha sido el intervalo (19.59,20.41) y sabemos que la desviación típica es $\sigma = 2.1$. Por tanto, hemos rechazado la hipótesis.

¿Cómo se llama la probabilidad de habernos equivocado es la decisión, es decir, de haber rechazado la hipótesis, cuando en realidad era cierta? ¿Influye el tamaño de la muestra en la probabilidad de cometer este tipo de error?

el nivel de significación. No depende del tamaño de la muestra.

Error tipo I: rechazar H_0 siendo cierta. Probabilidad cometer error tipo I es α ,

- PB. 12. La nota media en unas oposiciones celebradas el año pasado fue de 4.35 con una desviación típica de 2.5 puntos. Este año se han vuelto a convocar unas oposiciones similares. Con un nivel de significación de 0.01, y suponiendo que la desviación típica sigue siendo la misma, queremos contrastar la hipótesis de que la media no ha variado. Para ello, vamos a extraer una muestra aleatoria de 100 exámenes. Así, la zona de aceptación será el intervalo (3.71,4.99). Si al final la media real fuera de 3 puntos y hubiéramos aceptado H_0 (siendo falsa), ¿qué tipo de error habríamos

cometido? ¿Cómo influye el tamaño de la muestra en la probabilidad de cometer este tipo de error?

A mayor tamaño de la muestra, menor es la probabilidad de cometerlo.
Error de tipo II: aceptar H_0 siendo falsa.

PB. 13. En una autoescuela afirman que el porcentaje de sus alumnas y alumnos que obtienen el permiso de conducir la primera vez que se examinan es del 53%. Para contrastar esta hipótesis, vamos a seleccionar una muestra aleatoria de 65 alumnas y alumnos de esa autoescuela. La zona de aceptación de la hipótesis que hemos considerado para la proporción es el intervalo (0.41,0,65). Supongamos que hemos obtenido una proporción muestral que cae fuera de la zona de aceptación y que, por tanto, rechazamos la hipótesis.

- a) ¿Cómo se llama el error que cometíramos equivocándonos en esta decisión; es decir, rechazando H_0 , si en realidad fuera cierta?
- b) ¿Influye el tamaño de la muestra en la probabilidad de cometer este tipo de error?

b) No influye el tamaño de la muestra.
La probabilidad de cometer este error es α , el nivel de significación.
a) Si rechazamos H_0 siendo cierta cometemos error tipo I.

PB. 14. A principio de año, un estudio en cierta ciudad indicaba que un 15% de los conductores utilizaba el móvil con el vehículo en marcha. Con el fin de investigar la efectividad de las campañas que se han realizado desde entonces para reducir estos hábitos, recientemente se ha hecho una encuesta a 120 conductores y 12 hacían uso indebido del móvil.

- a) Plantea un test para contrastar que las campañas no han cumplido su objetivo, frente a que sí lo han hecho, como parecen indicar los datos. ¿A qué conclusión se llega con un nivel de significación del 4%?
- b) Calcula un intervalo de confianza del 96% para la proporción de conductores que usan indebidamente el móvil después de las campañas.

a) No se rechaza H_0 : $p \geq 0.15$, b) $(0.04,0.16)$

PB. 15. Un directivo de cierta empresa de material eléctrico afirma que la vida media de cierto tipo de bombillas es de 1500 horas. Otro directivo de la misma empresa afirma que la vida media de dichas bombillas es igual o menor de 1500 horas. Elegida una

muestra aleatoria simple de 81 bombillas de dicho tipo, vemos que su vida media ha sido de 1450 horas. Suponiendo que la vida de las bombillas sigue una distribución normal con desviación típica igual a 180 horas:

a) ¿Es compatible la hipótesis $H_0 : \mu = 1500$, frente a la hipótesis $H_1 : \mu \neq 1500$ con una confianza del 99 %, con el resultado experimental $\bar{x} = 1450$?

b) ¿Es compatible la hipótesis $H_0 : \mu \leq 1500$, frente a la hipótesis $H_1 : \mu > 1500$ con una confianza del 99 %, con el resultado experimental $\bar{x} = 1450$?

a) Si, rechazamos H_0 ; b) rechazamos H_0 y aceptamos H_1

- PB. 16. Una empresa eléctrica fabrica focos que tienen una duración que está distribuida aproximadamente en forma normal con una media de 800 horas y una desviación estándar de 40 horas. Pruebe la hipótesis de que $\mu = 800$ horas en contraposición de la alternativa de que $\mu \neq 800$ horas si una muestra aleatoria de 30 focos tiene una duración promedio de 788 horas. Utilice un nivel de significación de 0.04.

No se rechaza H_0 , es decir, los focos tienen una duración promedio de 800 horas.

- PB. 17. Un fabricante de cigarros afirma que el contenido promedio de nicotina no excede de 3.5 miligramos, con una desviación standar de 1.4 milímetros. Para una muestra de 8 cigarros se tiene un contenido promedio de nicotina de 4.2 miligramos .¿Está esto de acuerdo con la afirmación del fabricante ?. Use nivel de significación de 0.05.

Se acepta H_0 es decir, es correcta la afirmación del fabricante.

- PB. 18. Se sabe que la desviación típica de las notas de cierto examen es 2.4. Para una muestra de 36 estudiantes se obtuvo una nota media de 5.6. ¿Sirven estos datos para confirmar la hipótesis de que la nota media del examen fue 6, con un nivel de confianza del 95 %?

No se rechaza H_0 , la nota media del examen fue de 6, con un nivel de confianza del 95 %

- PB. 19. Debido a la futura fusión de dos entidades de ahorro, un estudio preliminar estima que como máximo un 5 % de los clientes causará baja en la nueva entidad resultante. Un analista de mercados sospecha que la proporción de bajas será mayor y para contrastarlo realiza una encuesta a 400 clientes, elegidos al azar, sobre su intención de seguir operando con la nueva entidad después de la fusión. De ellos 370 contestan que seguirán operando con la nueva entidad.

Formula un test para contrastar la hipótesis de que la proporción es la que se formula en el estudio preliminar frente a la del analista. ¿A qué conclusión se llega con un nivel de significación del 5 %?

Comenta que son los errores de tipo I y tipo II en este caso.

cuando realmente sí la tiene.

Error de tipo II (no rechazar H_0 siendo falsa): afirmar que el analista no tiene razón

que causaría baja es la que cree el analista, cuando realmente no es cierto

Con un riesgo de 5 %, el analista tiene razón.

No se rechaza H_0 : $d \geq 0.05$.

PB. 20. En un hospital se observó que los pacientes abusaban del servicio de urgencias, de forma que un 30 % de las consultas podían perfectamente haber esperado a concertar una cita con el médico de cabecera, porque no eran realmente urgencias. Puesto que esta situación ralentizaba el servicio, se realizó una campaña intensiva de concienciación. Transcurridos unos meses se ha recogido información de 120 consultas al servicio, de las cuales sólo 30 no eran realmente urgencias:

a) Hay personal del hospital que defiende que la campaña no ha mejorado la situación. Plantee un test para contrastar esta hipótesis frente a que sí la mejoró. Si se concluye que la situación no ha mejorado y realmente sí lo hizo, ¿cómo se llama el error cometido?

b) ¿A qué conclusión se llega en el test planteado en el apartado anterior con un nivel de significación del 1 %?

por lo menos el 30 % de los pacientes continúan yendo a urgencias.

a) Error tipo II; b) no rechazamos H_0 : $p \leq 0.30$,

PB. 21. En una determinada provincia, la nota media en matemáticas de los alumnos de 2º de Bachillerato del curso pasado fue de 5.8, con una desviación típica de 2.3 puntos.

Con un nivel de significación de 0.05, queremos contrastar la hipótesis de que la media no ha variado. Para ello, vamos a extraer una muestra aleatoria de tamaño 100. Así, la zona de aceptación será el intervalo (5.35,6.25). Si al final la media real fuera de 5 puntos, ¿cuál es la probabilidad de obtener una media muestral que nos lleve a cometer un error de tipo II?

zona de aceptación, 0.0643

Probabilidad cometer error tipo II es la probabilidad de obtener una \bar{x} que esté en la

6.8. Curiosidades

¿Por qué respecto a la hipótesis nula se habla de “no rechazo” y no de “aceptación”?

Esta es una característica general de la búsqueda del conocimiento científico. Ilustraremos la situación con un ejemplo.

En un juego se trata de descubrir un animal oculto y los participantes (los científicos) deben hacer preguntas sobre características de dicho animal con el propósito de descubrirlo. Uno de ellos tiene como hipótesis (nula) que el animal es una paloma y para contrastar su hipótesis pregunta: ¿el animal en cuestión tiene alas? Se le responde que sí, que efectivamente el animal tiene alas.

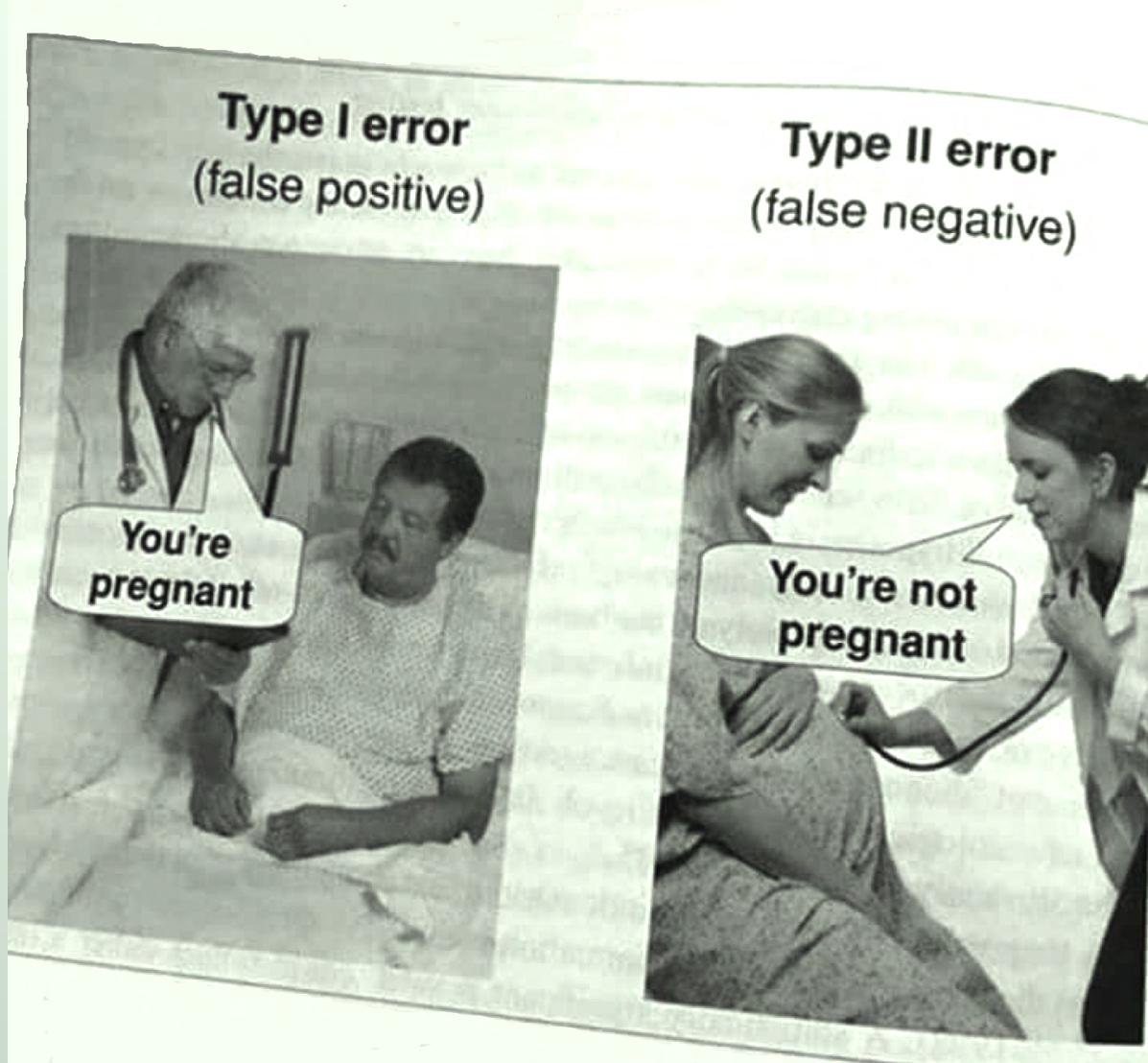
¿Qué conclusión puede sacarse hasta el momento? ¿Podría decirse entonces que su hipótesis es cierta y que el animal en cuestión es necesariamente una paloma? Sabemos que la respuesta es no. La evidencia (tener alas) es compatible con su hipótesis, pero ello no significa que sea verdadera, pues existen muchas otras hipótesis que son igualmente compatibles con dicha evidencia, por ejemplo, que el animal sea una mariposa, o que sea un murciélago.

El científico hace una nueva pregunta (observación): ¿el animal es vertebrado? Y como resultado de la observación se le responde que no. ¿Qué pasa con su hipótesis ahora? ?El animal podría ser una paloma? No. Ahora rechazamos la hipótesis de forma contundente para replantearla de tal manera que sea compatible con los nuevos hechos: tiene alas y no es vertebrado. La nueva hipótesis de trabajo puede ser: el animal es un mosquito.

Todas las verdades en la ciencia son de carácter transitorio, de forma que una afirmación sobre la naturaleza es verdadera porque no ha podido demostrarse que es falsa. Sin embargo, pueden existir muchas hipótesis compatibles con los hechos (no solo la nuestra), por esta razón el no rechazo de una hipótesis no implica su veracidad. No ocurre lo mismo cuando los hechos contradicen la hipótesis, así, si el animal no es vertebrado, estamos muy seguros de que no es una paloma.

“55 respuestas a dudas estadísticas”. Roberto Behar Gutiérrez y Pere Grima Cintas. Ediciones Díaz de Santos, S. A., Madrid 2004.

Chiste: errores en el contraste de hipótesis



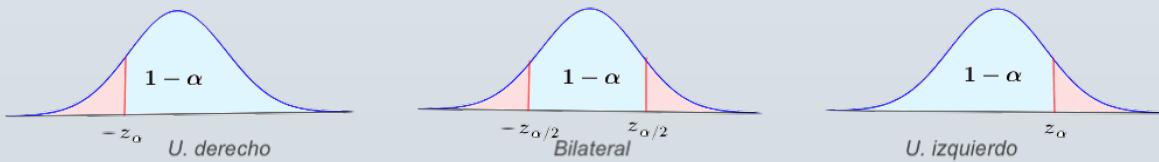
RESUMEN: Contraste de Hipótesis

▷ Test para la media

Contraste	H_0	H_1	Estadístico de contraste	Región de aceptación
Bilateral	$\mu = \bar{x}$	$\mu \neq \bar{x}$	$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$	$(-z_{\alpha/2}, z_{\alpha/2})$
U. izquierdo	$\mu \leq \bar{x}$	$\mu > \bar{x}$		$(-\infty, z_{\alpha})$
U. derecho	$\mu \geq \bar{x}$	$\mu < \bar{x}$		$(-z_{\alpha}, +\infty)$

▷ Test para una proporción

Contraste	H_0	H_1	Estadístico de contraste	Región de aceptación
Bilateral	$p = \hat{p}$	$p \neq \hat{p}$	$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$	$(-z_{\alpha/2}, z_{\alpha/2})$
U. izquierdo	$p \leq \hat{p}$	$p > \hat{p}$		$(-\infty, z_{\alpha})$
U. derecho	$p \geq \hat{p}$	$p < \hat{p}$		$(-z_{\alpha}, +\infty)$

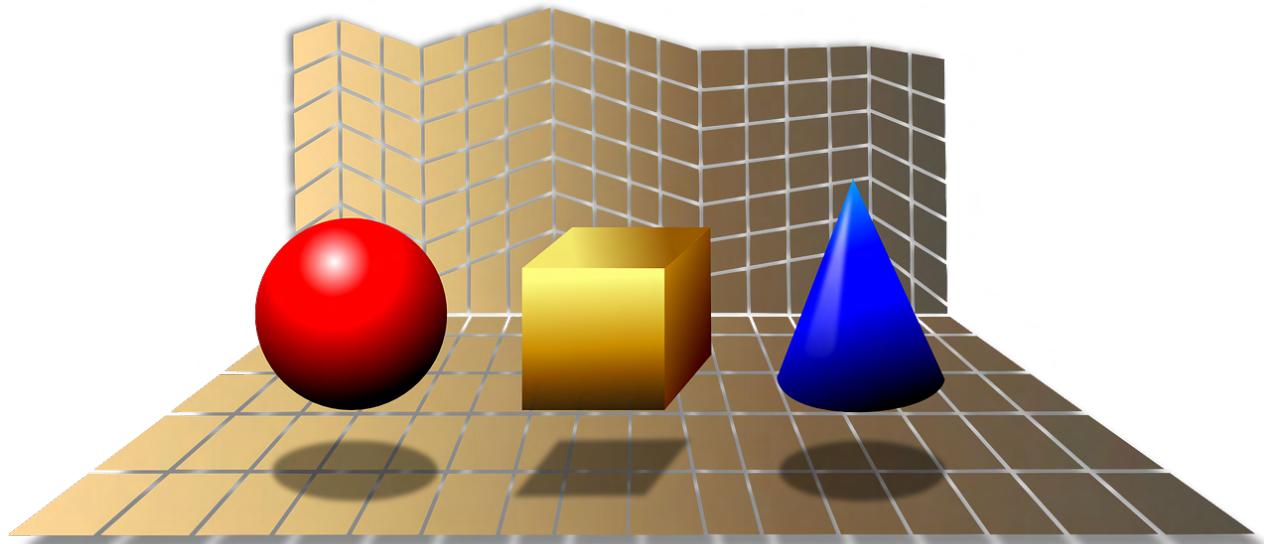


▷ Test para la comparación de medias

Contraste	H_0	H_1	Estadístico de contraste	Región de aceptación
Bilateral	$\bar{x} = \bar{y}$	$\bar{x} \neq \bar{y}$	$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$	$(-z_{\alpha/2}, z_{\alpha/2})$

Parte IV

Apéndices



Apéndice A

Ideas básicas de la teoría de conjuntos

Un conjunto es un “muchos” que puede ser pensado como uno.

George CANTOR

A.1. Conjuntos

Definición A.1:

Un **conjunto** es una colección de objetos con características similares considerada en sí misma como un objeto.

Cada uno de los objetos que forman el conjunto recibe el nombre de **elemento** del conjunto.

Un conjunto puede definirse por *extensión* (nombrando a todos sus elementos) o por *comprensión* (dando una propiedad que nos permita discernir si un objeto dado es o no un elemento del conjunto). Cuando el conjunto se define por extensión y ésta sea larga, se procede a definirlo *por recurrencia* o mediante una expresión generalista.

Los conjuntos se designan por letras mayúsculas del alfabeto latino, A , B , C , \dots . Los elementos de un conjunto se dan entre llaves y se suelen representar por letras minúsculas del alfabeto latino, $\{a, b, c, \dots\}$.

Nótese que en los conjuntos no es importante el orden en que se den sus elementos.

Definición A.2:

El conjunto que no contiene ningún elemento se le llama **conjunto vacío** y se representa por el símbolo $\emptyset = \{\}$.

Este conjunto se define como una necesidad para cuadrar toda la teoría de conjuntos.

Definición A.3:

Relación de pertenencia Se dice que un elemento a *pertenece* a un conjunto A si es de él, se representa así: $a \in A$. En caso contrario, se dice que no pertenece, $a \notin A$

Ejemplo A.1:**Ejemplos**

- Conjunto de los resultados que se obtienen al lanzar un dado: $A = \{1, 2, 3, 4, 5, 6\}$; $3 \in A$, $9 \notin A$.
- Conjunto de los números naturales: $\mathbb{N} = \{1, 2, 3, \dots\}$; $5 \in \mathbb{N}$, $2.47 \notin \mathbb{N}$.
- El conjunto de los múltiplos naturales de 7 es $\dot{7} = \{7, 14, 21, \dots\} = \{7n, \forall n \in \mathbb{N}\}$; $91 \in \dot{7}$, $163 \notin \dot{7}$.
- Los números naturales, enteros, racionales, reales y complejos se designan, respectivamente, por las letras \mathbb{N} , \mathbb{Z} , \mathbb{Q} , \mathbb{R} , \mathbb{C} .
- La expresión $\mathbb{R} \sim \{0, 1\}$ indica en conjunto de todos los números reales a excepción del número 0 y del 1.

A.1.1. Subconjuntos**Definición A.4:**

Dado un conjunto A , a cualquier conjunto B formado por cualquier número de elementos de A se le llama **subconjunto** de A .

Entre los posibles subconjuntos de A están, obviamente, el conjunto vacío \emptyset y el propio conjunto A . Estos subconjuntos reciben el nombre de ‘subconjuntos impropios’.

Para indicar que B es un subconjunto de A se escribe $B \subset A$ y se lee “ B está contenido o incluido en A ”. Este símbolo se puede usar al revés, entonces $A \supset B$ y se lee “ A incluye o contiene a B ”. Es un error escribir $B \in A$, el símbolo \in se reserva solo para la pertenencia de elementos a conjuntos, pero si puede usarse tanto $a \in A$ como $\{a\} \subset A$. Al incluir un elemento entre llaves indicamos que se trata de un conjunto *unitario*, formado por un único elemento. Si un conjunto C no es subconjunto de A se escribe $C \not\subset A$.

Es obvio que $\emptyset \subset A$ y que $A \subset A$. (En realidad, deberíamos escribir $A \subseteq A$).

Definición A.5:

Se llama **cardinal** de un conjunto A , $card(A)$, al número de elementos que forman el conjunto A .

Así, en el ejemplo anterior, $card(A) = 6$; $card(\mathbb{N}) = \infty$, etc.

Definición A.6:

Dado un conjunto A , el nuevo conjunto formado por todos los posibles subconjuntos de A se llama conjuntos de las **partes** de A y se representa por $\mathcal{P}(A)$.

En $\mathcal{P}(A)$ se incluyen los subconjuntos ‘impropios’ \emptyset y el propio A .

Teorema A.1:

Si un conjunto está formado por n elementos, el número de subconjuntos que se pueden formar a partir de A , incluyendo tanto el propio A como al conjunto vacío \emptyset , es decir, el número de elementos de las partes de A es 2^n , es decir:

$$\boxed{\text{Si } \text{card}(A) = n \rightarrow \text{card}(\mathcal{P}(A)) = 2^n}$$

Demostración. La demostración de este teorema se hace usando combinatoria¹ y el binomio de Newton².

Si el conjunto A tiene n elementos, el número de subconjuntos con k elementos es igual al número combinatorio $C(n, k) = \binom{n}{k}$. Un subconjunto de A puede tener 0 elementos como mínimo (\emptyset), y n como máximo (A), y por lo tanto:

$$\text{card}(\mathcal{P}(A)) =^1 \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n} =^2 (1+1)^n = 2^n$$

□

Ejemplo A.2:

En el conjunto $D = \text{resultados obtenidos al lanzar un dado ‘quinielístico’}$, $X = \{1, X, 2\}$, que tiene 3 elementos, podemos considerar hasta $8 = 2^3$ subconjuntos posibles. Así:

- Subconjuntos de cero elementos: $\{\} = \emptyset$
- Subconjuntos de un elemento: $\{1\}, \{X\}, \{2\}$
- Subconjuntos de dos elementos: $\{1, X\}, \{X, 2\}, \{1, 2\}$
- Subconjuntos de tres elementos $\{1, X, 2\} = Q$

Total, 8 subconjuntos.

Teorema A.2:

La relación de inclusión cumple las siguientes propiedades:

- Si $C \subset B \wedge B \subset A \rightarrow C \subset A$
- Si $B \subset A \wedge A \subset B \rightarrow A = B$

¹ver apéndice B ‘Combinatoria’

²Binomio de Newton: $(a+b)^n = \binom{n}{0}a^n b^0 + \binom{n}{1}a^{n-1}b^1 + \binom{n}{2}a^{n-2}b^2 + \cdots + \binom{n}{n}a^0 b^n$

- $\forall A, \emptyset \subset A$

Nota:

El símbolo \wedge equivale a la conjunción (lógica) “y”.

El símbolo \vee equivale a la disyunción (lógica) “o” .

Definición A.7:

Si $B \subset A$, se llama **complementario** de B respecto de A al subconjunto de A formado por todos los elementos que no estén en B . El complementario de un subconjunto se representa por uno de estos símbolos: B' , B^C , \bar{B} ; $\neg B$

El complementario siempre hace referencia a un ‘todo’ o conjunto de referencia E , así, el complementario de B será lo que le falta a B para ser ‘todo’, estará formado por todos los elementos (del conjunto referencial E) excepto los del propio B , esquemáticamente: $B' = E \sim B$, donde con \sim queremos representar lo que acabamos de decir (coger todos los elementos de E excepto - \sim - los de B)

Obviamente, $(A')' = A; E' = \emptyset; \emptyset' = E$

Ejemplo A.3:

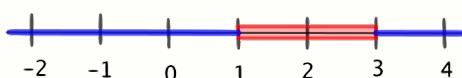
a) $E = \{1, 2, 3, 4, 5, 6\} \wedge B = \{1, 3\} \rightarrow B' = \{2, 4, 5, 6\}$

b) $\mathbb{R}^+ =]0, +\infty[\rightarrow (\mathbb{R}^+)' =]-\infty, 0] = \mathbb{R}^-$ y el 0

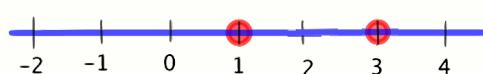
c) $[1, 3]^C = \mathbb{R} \sim [1, 3]$

No confundir con $\mathbb{R} \sim [1, 3]$ con $\mathbb{R} \sim \{1, 3\}$

$\mathbb{R} \sim [1, 3]$

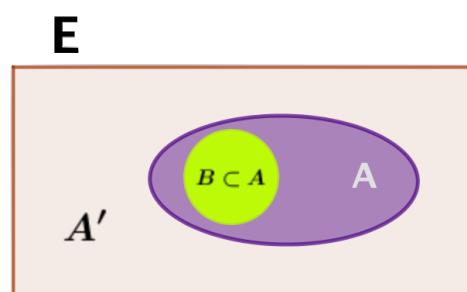


$\mathbb{R} \sim \{1, 3\}$



A.1.2. Diagramas de Venn

Los diagramas de Venn son esquemas usados en la teoría de conjuntos. Estos diagramas muestran ‘colecciones’ (conjuntos) de ‘objetos’ (elementos) por medio de líneas cerradas. La línea cerrada exterior abarca a todos los elementos bajo consideración, es el conjunto universal o referencial E .



Los diagramas de Venn fueron ideados hacia 1880 por John Venn.

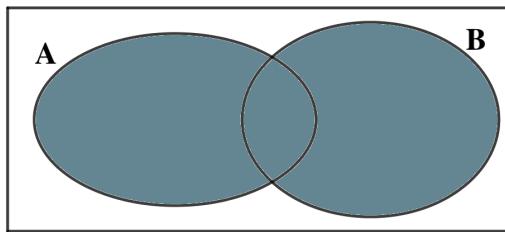
A.2. Operaciones con conjuntos

A.2.1. Unión e Intersección de conjuntos

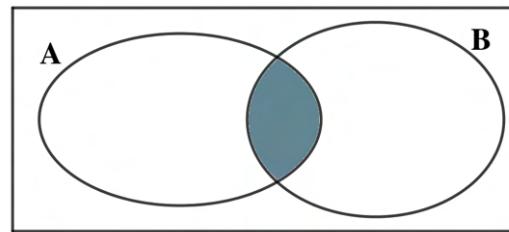
Definición A.8:

La **unión** de dos conjuntos A y B , que se denota por $A \cup B$, es el conjunto formado por todos los elementos que pertenecen a A o a B , los elementos que pertenecen a cualquiera de los dos conjuntos.

La **intersección** de dos conjuntos A y B , que se denota por $A \cap B$, es el conjunto formado por todos los elementos que pertenecen a A y a B simultáneamente, es decir, los elementos comunes a A y a B .



$$A \cup B \\ A \cup B = \{x / x \in A \vee x \in B\}$$



$$A \cap B \\ A \cap B = \{x / x \in A \wedge x \in B\}$$

Chiste


 \cap

 $=$

 \cup

 $=$


Teorema A.3:

Propiedades de la unión e intersección de conjuntos.

- $A \cup B = B \cup A$
- $A \cup \emptyset = A$
- $A \cup E = E$
- $A \cup A' = E$
- $A \cap B = B \cap A$
- $A \cap \emptyset = \emptyset$
- $A \cap E = A$
- $A \cap A' = \emptyset$

Definición A.9:

Si $A \cap B = \emptyset \rightarrow$ se dice que A y B son **disjuntos**

Ejemplo A.4:

a) Sean $A = \{a, A, e\}$; $B = \{A, B\} \rightarrow$

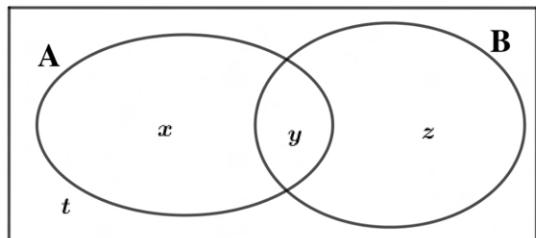
$$A \cup B = \{a, e, A, B\}; A \cap B = \{A\}.$$

b) Sean $\dot{3} = \{3, 6, 9, 12, \dots\}$; $\dot{5} = \{5, 10, 15, \dots\} \rightarrow$

$$A \cup B = \{3, 5, 6, 9, 10, \dots\}; A \cap B = \{15, 30, \dots\} \text{ (mcm)}.$$

Teorema A.4:

$$\text{card}(A \cup B) = \text{card}(A) + \text{card}(B) - \text{card}(A \cap B)$$



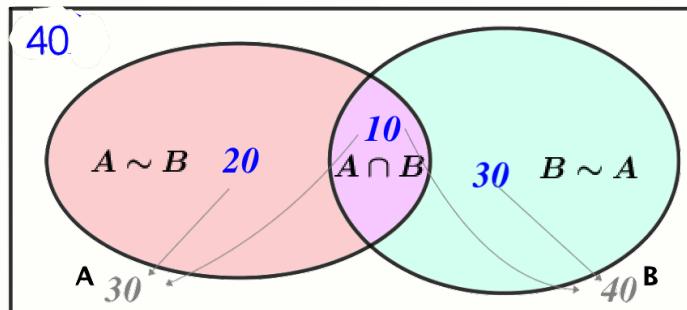
$$\text{card}(A) = x + z; \text{ card}(B) = y + z$$

$$\begin{aligned} \text{card}(A \cup B) &= x + y + z = \\ &= (x + z) + (y + z) - z \end{aligned}$$

Ejemplo A.5:

Se sabe que, de los 100 alumnos de bachillerato (conjunto referencial), a 30 les gusta la Anatomía (cardinal del conjunto A), a 40 la Biología (cardinal del conjunto B) y a 10 les gustan ambas asignaturas (cardinal de la intersección). ¿A cuántos alumnos les gustan alguna de estas asignaturas ($\text{card}(A \cup B)$)? ¿A cuántos de ellos no les gusta ninguna ($\text{card}(A \cup B)'$)?

El siguiente diagrama ilustra la solución y da un ejemplo que aclara el teorema anterior.



De los 30 alumnos que les gusta Anatomía, a 10 de ellos también les gusta la biología (pues les gustan las dos asignaturas); quedan $30 - 10 = 20$ alumnos a los que solo les gusta la Anatomía. Un razonamiento análogo nos llevará a concluir que son $40 - 10 = 30$ los alumnos a los que les gusta solo la biología. En total tendremos 20 alumnos que les gusta solo A, 30 que les gusta solo B y 10 que les gustan ambas, total 60 alumnos de los 100 a los que les gusta alguna de estas dos asignaturas. Hay, pues, 40 a los que no les gustan ninguna de ellas.

$$\text{card}(A \cup B) = 60; \quad \text{card}((A \cup B)') = 40$$

A.2.2. Diferencia de conjuntos

Definición A.10:

Dados dos conjuntos A y B , se llama conjunto **diferencia**, y se representa por $A \sim B$, al conjunto formado por todos los elementos de A excluidos los que pertenecen a B .

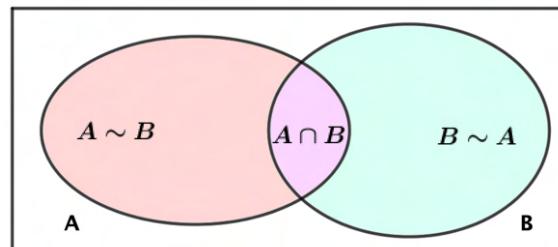
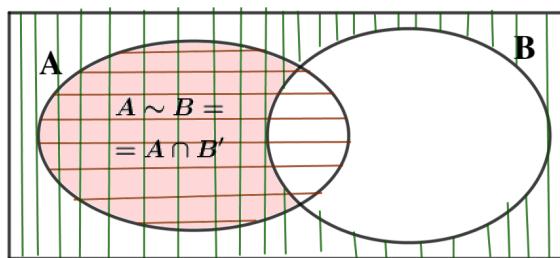
$$A \sim B = \{x / x \in A \wedge x \notin B\}$$

$$\text{Análogamente: } B \sim A = \{x / x \in B \wedge x \notin A\}$$

$$\text{Obviamente } A \sim B \neq B \sim A$$

Teorema A.5:

$$A \sim B = A \cap B'; \quad A \cup B = (A \sim B) \cup (A \cap B) \cup (B \sim A)$$



Demostración. En la primera figura, hemos dibujado el conjunto A con líneas horizontales y el B' (todo lo que no es B), con líneas verticales. La intersección es la sección común en que se cruzan las líneas. (La unión sería todo lo que esté rayado de cualquier forma, vertical, horizontal o cruzado).

En la segunda figura, hemos dibujado en rojo el conjunto $A \sim B$, en azul el $B \sim A$ y en morado en $A \cap B$. La unión $A \cup B$ está formada por todas las secciones coloreadas. Los tres conjuntos son disjuntos entre sí. \square

A.2.3. Producto cartesiano de dos conjuntos

Definición A.11:

El producto cartesiano de dos conjuntos A y B , que denotamos por $A \times B$, es el conjunto formado por todos los pares ordenados de elementos de A y B (el primer elemento del par ha de ser de A y el segundo de B).

$$A \times B = \{(a, b) / a \in A \wedge b \in B\}$$

Ejemplo A.6:

a) $A = \{1, 2, 3\}; B = \{a, b\} \rightarrow$

$$A \times B = \{(1, a), (1, b), (2, a), (2, b), (3, a), (3, b)\}$$

b) $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$ es el conjunto de los puntos del plano. \mathbb{R}^3 representa al espacio tridimensional, \mathbb{R}^n un espacio de n dimensiones.

A.3. Propiedades combinadas de las operaciones con conjuntos

Teorema A.6:

- Asociativas:

$$A \cup (B \cup C) = (A \cup B) \cup C \quad A \cap (B \cap C) = (A \cap B) \cap C$$

- Distributivas:

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \quad A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

- Leyes de Morgan

$$(A \cup B)' = A' \cap B'$$

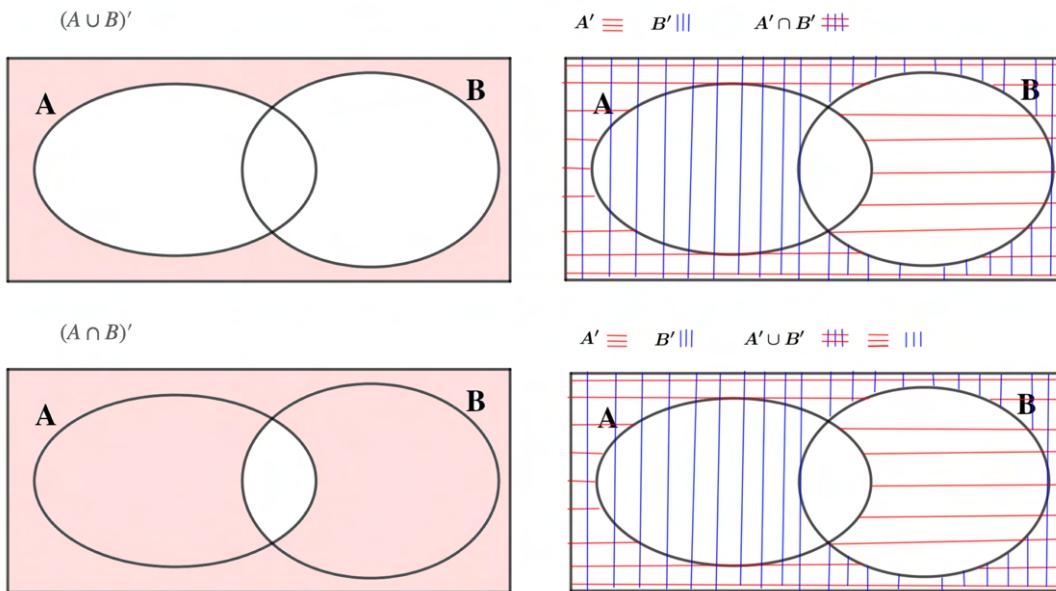
$$(A \cap B)' = A' \cup B'$$

“El complementario de la unión es la intersección de complementarios”.

“El complementario de la intersección es la unión de complementarios”.

Demostración. .

Mediante diagramas de Venn, es fácil demostrar estos teoremas.



Leyes de Morgan y diagramas de Venn.

□

Ejemplo A.7:

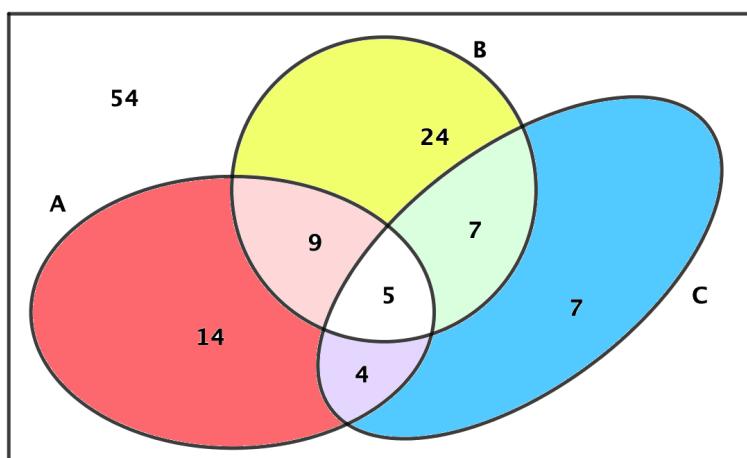
Si M representa el conjunto de los habitantes de Madrid y C el conjunto de los nacidos en Cataluña, entonces:

- $M \cup C$ representa el conjunto de las personas que viven en Madrid o que han nacido en Cataluña.
- $(M \cup C)'$ representa a las personas que no viven en Madrid y que no han nacido en Cataluña.
- M' son las personas que no viven Madrid, y C' aquellos que no han nacido en Cataluña.
- $M' \cap C'$ serán las personas que no viven en Madrid y que tampoco han nacido en Cataluña. Es evidente que $(M \cup C)' = M' \cap C'$.
- Igualmente: $M \cap C$ representa el conjunto de las personas que viven en Madrid y que han nacido en Cataluña.
- $(M \cap C)'$ representa a las personas que o no viven en Madrid o no han nacido en Cataluña.
- M' son las personas que no viven Madrid, y C' aquellos que no han nacido en Cataluña.
- $M' \cup C'$ serán las personas que no viven en Madrid o que no han nacido en Cataluña. Es evidente que $(M \cap C)' = M' \cup C'$.

Ejercicio resuelto A.1. En una ciudad se editan tres revistas A, B y C. Se ha preguntado a un grupo de personas sobre la lectura o no de esos periódicos, obteniéndose los siguientes resultados:

- Lectores de A: 32; lectores de B: 45; lectores de C: 23.
- Leen A y B 14 personas, A y C 9 y B y C 12.
- Leen las tres periódicos 5 personas.
- No leen ninguna revista 54 personas.

¿A cuántas personas se les ha pasado la encuesta?.



$$54 + 14 + 9 + 24 + 4 + 5 + 7 + 7 = 124 \text{ personas encuestadas}$$

Apéndice B

Combinatoria. Técnicas de recuento

La combinatoria es la parte de las matemáticas que se ocupan de la resolución de problemas de elección y disposición de los elementos de un conjunto, atendiendo a ciertas reglas.

B.1. Principio de multiplicación

Definición B.1:

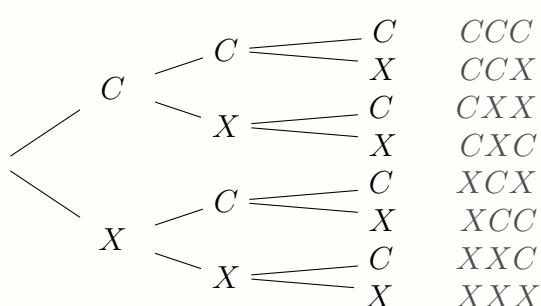
Si un principio de selección se puede separar en r etapas, de modo que el resultado de cada una de ellas no influya en la siguiente, y en cada uno de estas etapas se obtienen respectivamente n_1, n_2, \dots, n_r resultados, entonces el procedimiento global tiene

$$\prod_{i=1}^r n_i = n_1 \cdot n_2 \cdots n_r \text{ resultados.}$$

Veamos algunos ejemplos en los que nos ayudamos con un *diagrama de árbol*:

Ejemplo B.1:

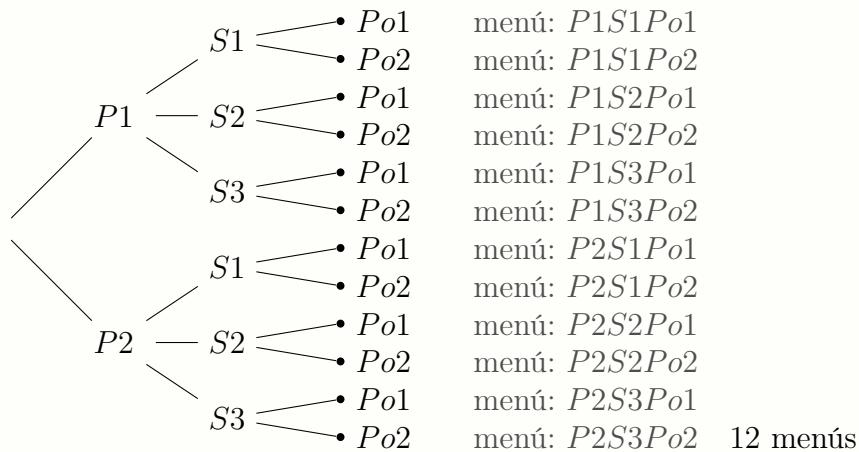
Distintos resultados en el lanzamiento de una monedas tres veces.



$$2 \cdot 2 \cdot 2 = 8 \text{ posibilidades}$$

Ejemplo B.2:

En un restaurante el menú se pueden elegir entre dos primeros platos (P), tres segundos (S) y dos postres (Po). ¿Cuántos menús diferentes se pueden pedir?



$$2 \cdot 3 \cdot 2 = 12 \text{ posibilidades}$$

El principio del palomar

También llamado principio de Dirichlet o principio de las cajas, establece que si n palomas se distribuyen en m palomares, y si $n > m$, entonces al menos habrá un palomar con más de una paloma. A manera de ejemplo: si se toman trece personas, al menos dos habrán nacido el mismo mes.

Aunque el principio del palomar puede parecer una observación trivial, se puede utilizar para demostrar resultados inesperados. Por ejemplo, hay por lo menos 2 personas en Guatemala con el mismo número de pelos en la cabeza.

Demostración: la cabeza de una persona tiene en torno a 150.000 cabellos y tener un millón de pelos requeriría de una cabeza gigante (nadie tiene un millón de pelos en la cabeza). Asignamos un palomar por cada número de 0 a 1.000.000 y asignamos una paloma a cada persona que irá al palomar correspondiente al número de pelos que tiene en la cabeza. Como en Guatemala hay más de un millón de personas, habrá al menos dos personas con el mismo número de pelos en la cabeza.



B.2. Permutaciones

Definición B.2:

Se llama **factorial** de un número natural $n \in \mathbb{N}$ y se denota como $n!$:

$$n! = n \cdot (n - 1) \cdot (n - 2) \cdots 3 \cdot 2 \cdot 1$$

Por convenio, se define $0! = 1$

Definición B.3:

Se llama **permutaciones de n elementos**, P_n , todos ellos distintos, al número de ordenaciones de esos elementos.

$$P_n = n! = n \cdot (n - 1) \cdot (n - 2) \cdots 3 \cdot 2 \cdot 1$$

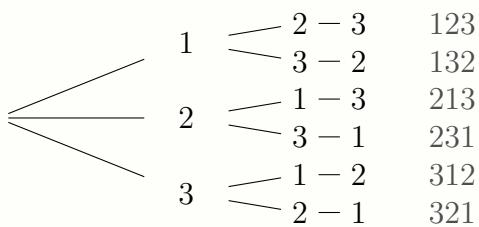
Esto es obvio. De los n elementos a ordenar, el primero lo podemos elegir de n formas distintas. Por cada una de ellas quedan $n - 1$ elementos de los que podemos escoger el segundo a ordenar de $n - 1$ formas distintas. Para el tercer elemento, hay 2 escogidos y quedan $n - 2$ por escoger, lo podemos hacer de $n - 2$ formas distintas; etc. Cuando queden solo dos elementos, ya hay $n - 2$ elegidos, tenemos 2 formas distintas de hacerlo; para cada una de estas elecciones quedará solo 1 elemento a escoger que podremos hacer de 1 sola forma. Usando el principio de multiplicación, tendremos: $P_n = n \cdot (n - 1) \cdot (n - 2) \cdots 3 \cdot 2 \cdot 1$.

Ejemplo B.3:

¿Cuántos números distintos, de tres cifras y ninguna de ellas repetidas, podemos obtener con los dígitos $\{3, 5, 7\}$

Se trata de obtener todas las ordenaciones posibles de tres elementos, por tanto, la solución es $P_3 = 3! = 6$ números distintos.

Ilustramos estos números con un diagrama de árbol.



El primer dígito del número puede ser de los tres dígitos; elegido éste, el segundo puede ser cualquiera de los dos que quedan y ya solo quedaría una posibilidad para el tercer dígito. Por el principio de superposición, $3 \cdot 2 \cdot 1 = 6$ números distintos.

Ejemplo B.4:

¿De cuantas maneras distintas se pueden colocar 10 libros distintos en el balde de una estantería?

El primer libro, de derecha a izquierda, por ejemplo, puede ser cualquiera de los 10; el segundo, cualquiera de los 9 restantes, ...

$$P_{10} = 10! = 3628800 \text{ formas de ordenación distintas.}$$

B.2.1. Permutaciones con repetición**Definición B.4:**

Para calcular el número de ordenaciones posibles de n elementos de los cuales hay n_1 repetidos, n_2 repetidos, etc; de modo que $n_1 + n_2 + \dots + n_r = n$ se usan las **permutaciones con repetición**.

$$PR_n^{n_1, n_2, \dots, n_r} = \frac{n!}{n_1! \cdot n_2! \cdots n_r!}$$

Ejemplo B.5:

¿De cuantas maneras distintas se pueden colocar 10 libros en el balde de una estantería, si hay 5 de ellos iguales entre sí y también 3 de ellos iguales?

Se trata de ordenar 10 objetos de los cuales hay 5, 3, 1, 1 repetidos ($5 + 3 + 1 + 1 = 10$).

$$PR_{10}^{5,3,1,1} = \frac{10!}{5! \cdot 3!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 3 \cdot 2} = 5040 \text{ ordenaciones.}$$

Ejemplo B.6:

Con los dígitos 1, 1, 1, 2, 2, 3, ¿cuántos números distintos con las seis cifras pueden formarse ?

$$PR_6^{3,2,1} = \frac{6!}{3! \cdot 2!} = 60 \text{ números distintos.}$$

Si los dígitos fuesen distintos, 1, 1, 1, 2, 2, 3, tendríamos un total de $6!$ ordenaciones distintas.

Pero, ocurre que las ordenaciones 1, 1, 1, 2, 2, 3 y 1, 1, 1, 2, 2, 3 (se ha alterado el orden de los unos) son el mismo número, y así hay $3!$ posibilidades distintas, por lo que al resultado anterior $6!$, hay que dividirlo entre estas $3!$ ordenaciones que conducen al mismo número.

Lo mismo para el dígito 2, hay $2!$ formas inicialmente distintas que conducen al mismo número, 1, 1, 1, 2, 2, 3 y 1, 1, 1, 2, 2, 3, por lo que hay que dividir el resultado anterior $6!/3!$ entre $2!$ obteniéndose el resultado final $6!/(3! \cdot 2!) = PR_6^{3,2} (= PR_6^{3,2,1})$.

B.2.2. Permutaciones circulares

Definición B.5:

Las **permutaciones circulares** son un caso particular de las permutaciones.

Se utilizan cuando los elementos se han de ordenar “en círculo”, (por ejemplo, los comensales en una mesa), de modo que el primer elemento que “se sitúe” en la mesa determina el principio y el final de muestra, quedando tan solo $n - 1$ elementos que ordenar:

$$PC_n = (n - 1)!$$

Al ordenar n elementos en una posición circular, la posición $1, 2, 3, \dots, n$ es la misma que la $3, \dots, n, 1, 2$, solo ha habido una rotación. La forma de solucionar esto es elegir un elemento, fijarlo en un lugar y ordenar los $n - 1$ restantes.

Ejemplo B.7:

¿De cuántas formas pueden sentarse 5 comensales en una mesa circular?

Obviamente se tratará de $PC_5 = (5 - 1)! = 4! = 24$ disposiciones distintas.

B.3. Variaciones

Definición B.6:

Se llaman **Variaciones de n elementos (distintos) tomados de r en r** sin repetir ninguno de los tomados, luego $r < n$, al número de ordenaciones distintas que se pueden conseguir.

$$V_n^r = n \cdot (n - 1) \cdot (n - 2) \cdots (n - r + 1) = \frac{n!}{(n - r)!}$$

Si de una muestra de tamaño n queremos extraer una muestra **ordenada y sin repetición** de tamaño r , el primer elemento lo podemos elegir entre cualquiera de los n que forman la muestra. Escogido éste, el segundo puede ser cualquiera de los $(n-1)$ restantes $[n - 2 + 1]$. El tercero, cualquiera de los $(n - 2)$ que aún quedan $[n - 3 + 1]$. Así, cuando lleguemos a extraer el r -ésimo elemento, quedarán $[n - r + 1]$ $(n - r + 1)$ posibilidades.

Ejemplo B.8:

¿Cuántas números de cinco cifras no repetidas se pueden formar con los dígitos del 1 al 9?

Tenemos que extraer una muestra de 5 elementos ordenados y no repetidos de una muestra de 9, tenemos:

$$V_9^5 = \frac{9!}{(9-5)!} = \frac{9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4!}{4!} = 15120 \text{ números.}$$

Ejemplo B.9:

De los 20 miembros de un club, hay que elegir a un presidente, un vicepresidente y un tesorero. ¿De cuántas formas distintas puede hacerse?

Hay que tomar 3 elementos de entre un grupo de 20, los elementos a tomar han de ser distintos (un amigo no puede ocupar dos cargos) e importa en orden en que los elijamos (el primero será el presidente, el segundo el vicepresidente y el tercer elegido será el secretario), tenemos:

$$V_{20}^3 = \frac{20!}{(20-3)!} = 20 \cdot 19 \cdot 18 = 6840 \text{ formas posibles de elección.}$$

El presidente del club puede ser cualquiera de los 20 amigos. Elegido el presidente, podemos elegir al vicepresidente entre cualquiera de los 19 amigos restantes. Una vez elegidos estos dos cargos, solo quedan 18 amigos de entre los que elegir al secretario. por el principio de multiplicación, las posibilidades son $20 \cdot 19 \cdot 18 = 6840$.

B.3.1. Variaciones con repetición**Definición B.7:**

Se llaman **Variaciones con repetición de n elementos (distintos) tomados de r en r** al numero de ordenaciones distintas que se pueden conseguir. Obsérvese que **puede ocurrir que $r > n$** .

$$VR_n^r = n^r$$

Para una población de tamaño n y la elección de una muestra de tamaño r donde importa el orden de extracción pero podemos repetir los elementos extraídos procederemos del siguiente modo:

El primer elemento escogido puede ser una cualquiera de los n que forman el conjunto de la población elegible. Puesto que se pueden repetir los elementos extraídos, el segundo elegido también puede ser uno de os n elementos cualquiera de la población, Y también para el tercero, y para el cuarto, y ...

Aplicando el principio de multiplicación tendremos $n \cdot n \dots^{r\text{-veces}} \cdot n = n^r$.

Ejemplo B.10:

¿Cuántos números de tres cifras se pueden formar con los dígitos impares 1,3,5,7,9?

Ahora sí podemos repetir elementos en la extracción, 337 es un número de tres cifras e importa el orden, 337 no es el mismo numero que 733. Tenemos, pues: $VR_5^3 = 5^3 = 125$ números.

Ejemplo B.11:

¿Cuántas quinielas futbolísticas hay que rellenar para estar seguros de acertar un pleno al 15?

Una quiniela futbolística es una lista ordenada de 15 elementos elegidos entre las 3 posibilidades $\{1, X, 2\}$. Se trata de elegir 15 elementos, ordenados, de entre 3 posibles. Obviamente se puede repetir y tenemos:

$$VR_3^{15} = 3^{15} = 14348907 \text{ quinielas distintas.}$$

B.4. Combinaciones

Definición B.8:

Las distintas formas de elegir un número de elementos entre varios se llaman **combinaciones**.

Las combinaciones de n elementos tomados de r en r , donde ahora no importa el orden y tampoco podemos repetir, es decir $r < n$, son:

$$C_n^r = \binom{n}{r} = \frac{V_n^r}{r!} = \frac{n!}{(n-r)! \cdot r!}$$

Sabemos que elegir a r elementos distintos de un grupo de n son V_n^r . Ahora bien, si en estas elecciones no importa el orden en que se hayan extraído los elementos, tendremos que dividir entre las posibles ordenaciones de r elementos que conduciran a la misma extracción, $P_r = r!$, por ello $C_n^r = \frac{V_n^r}{r!}$.

Ejemplo B.12:

Si de los 20 miembros de un club de un ejemplo anterior, hay que elegir tres de ellos para que acudan a una determinada reunión, ¿de cuántas formas distintas puede hacerse.

De un conjunto de 20 elementos, tenemos que escoger a 3 de ellos de modo que no podemos repetir ningún elemento escogido y no importa el orden en que los escojamos. Tenemos:

$$C_{20}^3 = \binom{20}{3} = \frac{20!}{(20-3)! \cdot 3!} = \frac{20 \cdot 19 \cdot 18}{3 \cdot 2 \cdot 1} = 1140 \text{ formas.}$$

Ahora se trata de escoger tres elementos distintos de un total de 20 de ellos. Por lo que sabemos se debería tratar de V_{20}^3 , pero, en este caso, no importa el orden en que estos tres elementos hayan sido elegidos (ABC forman el mismo grupo que BAC, p.e.). Por ello habrá que dividir entre las posibles formas que habrán salido de estos tres elegidos, P_{20}^3 y así, obtendremos $C_{20}^3 = V_{20}^3/P_{20}^3$.

Ejemplo B.13:

¿Cuántas primitivas distintas hay que llenar para asegurarse un acierto completo?

Se trata de escoger 6 números distintos (no se puede repetir) de entre 49 posibles y no importa el orden en que se estraigan ('123456' sería la misma extracción que '625413').

Se trata pues de:

$$C_{49}^6 = \binom{49}{6} = \frac{49!}{(49-6)! \cdot 6!} = \frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44 \cdot 43!}{\cancel{(49-6)!} \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = \frac{49 \cdot 47 \cdot 46 \cdot 45^3 \cdot 44}{5 \cdot 3} = 13983816$$

primitivas distintas.

B.4.1. Combinaciones con repetición

Definición B.9:

Se llaman **Combinaciones con repetición** de n elementos tomados de r en r , a las formas de elegir, sin importar el orden pero si pudiendo repetir, a r elementos de entre n .

$$CR_n^r = \binom{n+r-1}{r}$$

Ejemplo B.14:

En una bodega hay cinco tipos diferentes de botellas. ¿De cuántas formas se pueden elegir cuatro botellas?

Hay que elegir 4 elementos de entre 5 tipos diferentes, p.e., AEIOU. Una posible elección podría ser 'AAEE', que sería la misma elección de botellas que la 'AEAE', luego tenemos que no importa el orden (combinaciones) y sí podemos repetir (con repetición) por lo que se trata de 'combinaciones con repetición de 5 elementos tomados de 4 en 4':

$$CR_5^4 = \binom{5+4-1}{4} = \binom{8}{4} = \frac{8 \cdot 7 \cdot 6^2 \cdot 5}{4 \cdot 3 \cdot 2 \cdot 1} = 70 \text{ elecciones.}$$

Ejemplo B.15:

Un ascensor con 10 personas se detiene en 15 pisos.

a) ¿De cuántas formas pueden bajarse las personas?

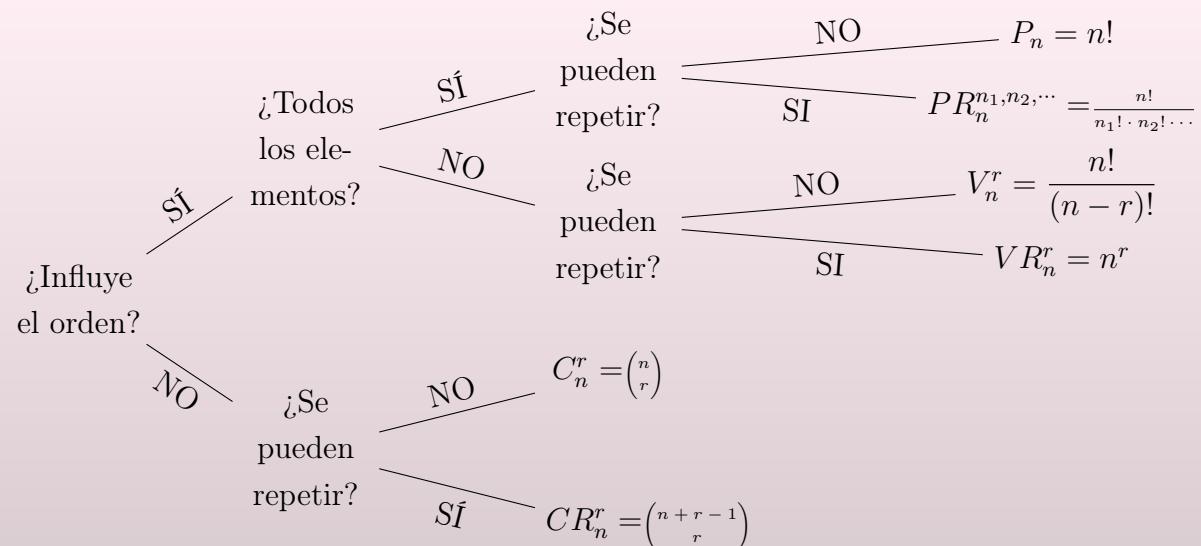
b) ¿Y si no puede bajar del ascensor más de una por piso?

En el primer caso tenemos $n = 15$ objetos (pisos 1, 2, 3, ...) de los cuales queremos escoger $r = 10$. Por ejemplo, la solución 1111111112 indicaría que 9 personas bajan en el piso 1º y 1 en el 2º. Pero 1111121111 sería la misma situación: podemos repetir pero no importa el orden, se trata de: $CR_{15}^{10} = 1961256$

En el segundo caso, al no poder repetir tenemos $C_{15}^{10} = 3003$

B.5. Resumen

Combinatoria



Ejercicio resuelto B.1. .

- ¿De cuántas formas distintas pueden colocarse en línea para una fotografía 5 amigos?
- Se va a programar un torneo de ajedrez para los 10 integrantes de un club. ¿Cuántos partidos se deben programar si cada integrante jugará con cada uno de los demás sin partidos de revancha?
- ¿Cuántos números de 5 cifras se pueden formar usando solo los dígitos impares 1, 3, 5, y 7?

- d) ¿Cuántas fichas tiene un dominó?
e) ¿Cuántos números de 6 cifras pueden formarse usando dos 1, un 3, un 5 y dos 7?
f) En una carrera con 10 atletas, ¿de cuántas formas pueden obtenerse las medallas de oro, plata y bronce?

a) $\rightarrow P_5 = 5! = 120$

b) $\rightarrow C_{10}^2 = \binom{10}{2} = 45$

c) $\rightarrow VR_4^5 = 4^5 = 1024$

d) $\rightarrow CR_7^2 = \binom{7+2-1}{7} = \binom{8}{7} = 28$

e) $\rightarrow PR_6^{2,1,1,2} = \frac{6!}{2! \cdot 1! \cdot 1! \cdot 2!} = 315$

f) $\rightarrow V_{10}^3 = \frac{10!}{(10-3)!} = 720$

B.6. Números combinatorios

Definición B.10:

Se define el **número combinatorio n sobre r** , con $r \leq n$, y se denota por $\binom{n}{r}$, como:

$$\boxed{\binom{n}{r} = C_n^r = \frac{n!}{r! \cdot (n-r)!}}$$

Indican las combinaciones, sin repetición, de n -elementos tomados de r en r o el número de subconjuntos de r -elementos de un conjunto de n -elementos.

Teorema B.1:

Propiedades de los números combinatorios.

$$\binom{n}{r} = \binom{n}{n-r}; \quad \binom{n}{0} = \binom{n}{n} = 1; \quad \binom{n}{r} + \binom{n}{r+1} = \binom{n+1}{r+1}$$

Ejemplo B.16:

Disposición de los números combinatorios en el **Triángulo de Tartaglia**

	1		1	
	1		2	1
	1		3	3
	1		4	6
...
			4	1
		

Teorema B.2:**Binomio de Newton**

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

Los coeficientes son los números de la fila n-ésima de Tartaglia.

Chiste

Chiste de **Mario** en el diario **Público** publicado el 22/10/2008.

Apéndice C

Tablas distribución Binomial y Normal

Chiste

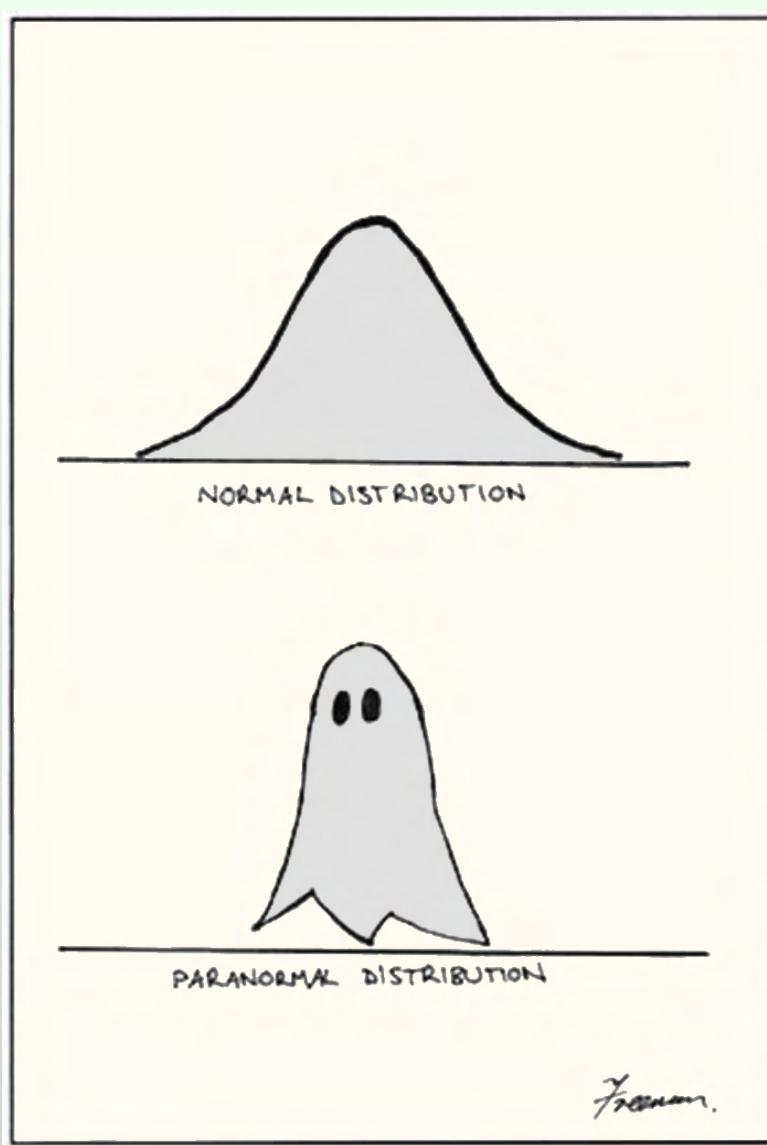


Tabla de probabilidades puntuales de la distribución $Binomial(n,p)$

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

n	k	p													
		0,01	0,05	0,10	0,15	1/6	0,20	0,25	0,30	1/3	0,35	0,40	0,45	0,50	
5	0	0,9510	0,7738	0,5905	0,4437	0,4019	0,3277	0,2373	0,1681	0,1317	0,1160	0,0778	0,0503	0,0313	
	1	0,0480	0,2036	0,3281	0,3915	0,4019	0,4096	0,3955	0,3602	0,3292	0,3124	0,2592	0,2059	0,1563	
	2	0,0010	0,0214	0,0729	0,1382	0,1608	0,2048	0,2637	0,3087	0,3292	0,3364	0,3456	0,3369	0,3125	
	3	0,0000	0,0011	0,0081	0,0244	0,0322	0,0512	0,0879	0,1323	0,1646	0,1811	0,2304	0,2757	0,3125	
	4	0,0000	0,0000	0,0005	0,0022	0,0032	0,0064	0,0146	0,0284	0,0412	0,0488	0,0768	0,1128	0,1563	
	5	0,0000	0,0000	0,0000	0,0001	0,0001	0,0003	0,0010	0,0024	0,0041	0,0053	0,0102	0,0185	0,0313	
6	0	0,9415	0,7351	0,5314	0,3771	0,3349	0,2621	0,1780	0,1176	0,0878	0,0754	0,0467	0,0277	0,0156	
	1	0,0571	0,2321	0,3543	0,3993	0,4019	0,3932	0,3560	0,3025	0,2634	0,2437	0,1866	0,1359	0,0938	
	2	0,0014	0,0305	0,0984	0,1762	0,2009	0,2458	0,2966	0,3241	0,3292	0,3280	0,3110	0,2780	0,2344	
	3	0,0000	0,0021	0,0146	0,0415	0,0536	0,0819	0,1318	0,1852	0,2195	0,2355	0,2765	0,3032	0,3125	
	4	0,0000	0,0001	0,0012	0,0055	0,0080	0,0154	0,0330	0,0595	0,0823	0,0951	0,1382	0,1861	0,2344	
	5	0,0000	0,0000	0,0001	0,0004	0,0006	0,0015	0,0044	0,0102	0,0165	0,0205	0,0369	0,0609	0,0938	
	6	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0014	0,0018	0,0041	0,0083	0,0156	
7	0	0,9321	0,6983	0,4783	0,3206	0,2791	0,2097	0,1335	0,0824	0,0585	0,0490	0,0280	0,0152	0,0078	
	1	0,0659	0,2573	0,3720	0,3960	0,3907	0,3670	0,3115	0,2471	0,2048	0,1848	0,1306	0,0872	0,0547	
	2	0,0020	0,0406	0,1240	0,2097	0,2344	0,2753	0,3115	0,3177	0,3073	0,2985	0,2613	0,2140	0,1641	
	3	0,0000	0,0036	0,0230	0,0617	0,0781	0,1147	0,1730	0,2269	0,2561	0,2679	0,2903	0,2918	0,2734	
	4	0,0000	0,0002	0,0026	0,0109	0,0156	0,0287	0,0577	0,0972	0,1280	0,1442	0,1935	0,2388	0,2734	
	5	0,0000	0,0000	0,0002	0,0012	0,0019	0,0043	0,0115	0,0250	0,0384	0,0466	0,0774	0,1172	0,1641	
	6	0,0000	0,0000	0,0000	0,0001	0,0001	0,0004	0,0013	0,0036	0,0064	0,0084	0,0172	0,0320	0,0547	
8	7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0005	0,0006	0,0016	0,0037	0,0078	
	0	0,9227	0,6634	0,4305	0,2725	0,2326	0,1678	0,1001	0,0576	0,0390	0,0319	0,0168	0,0084	0,0039	
	1	0,0746	0,2793	0,3826	0,3847	0,3721	0,3355	0,2670	0,1977	0,1561	0,1373	0,0896	0,0548	0,0313	
	2	0,0026	0,0515	0,1488	0,2376	0,2605	0,2936	0,3115	0,2965	0,2731	0,2587	0,2090	0,1569	0,1094	
	3	0,0001	0,0054	0,0331	0,0839	0,1042	0,1468	0,2076	0,2541	0,2731	0,2786	0,2787	0,2568	0,2188	
	4	0,0000	0,0004	0,0046	0,0185	0,0260	0,0459	0,0865	0,1361	0,1707	0,1875	0,2322	0,2627	0,2734	
	5	0,0000	0,0000	0,0004	0,0026	0,0042	0,0092	0,0231	0,0467	0,0683	0,0808	0,1239	0,1719	0,2188	
	6	0,0000	0,0000	0,0000	0,0002	0,0004	0,0011	0,0038	0,0100	0,0171	0,0217	0,0413	0,0703	0,1094	
9	7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0012	0,0024	0,0033	0,0079	0,0164	0,0313	
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0002	0,0007	0,0017	0,0039		
	9	0	0,9135	0,6302	0,3874	0,2316	0,1938	0,1342	0,0751	0,0404	0,0260	0,0207	0,0101	0,0046	0,0020
	1	0,0830	0,2985	0,3874	0,3679	0,3489	0,3020	0,2253	0,1556	0,1171	0,1004	0,0605	0,0339	0,0176	
	2	0,0034	0,0629	0,1722	0,2597	0,2791	0,3020	0,3003	0,2668	0,2341	0,2162	0,1612	0,1110	0,0703	
	3	0,0001	0,0077	0,0446	0,1069	0,1302	0,1762	0,2336	0,2668	0,2731	0,2716	0,2508	0,2119	0,1641	
	4	0,0000	0,0006	0,0074	0,0283	0,0391	0,0661	0,1168	0,1715	0,2048	0,2194	0,2508	0,2600	0,2461	
	5	0,0000	0,0000	0,0008	0,0050	0,0078	0,0165	0,0389	0,0735	0,1024	0,1181	0,1672	0,2128	0,2461	
10	6	0,0000	0,0000	0,0001	0,0006	0,0010	0,0028	0,0087	0,0210	0,0341	0,0424	0,0743	0,1160	0,1641	
	7	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0012	0,0039	0,0073	0,0098	0,0212	0,0407	0,0703	
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0009	0,0013	0,0035	0,0083	0,0176	
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0003	0,0008	0,0020	

Tabla de probabilidades puntuales de la distribución $Binomial(n,p)$ (Continuación)
 $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

n	k	p												
		0,01	0,05	0,10	0,15	1/6	0,20	0,25	0,30	1/3	0,35	0,40	0,45	0,50
10	0	0,9044	0,5987	0,3487	0,1969	0,1615	0,1074	0,0563	0,0282	0,0173	0,0135	0,0060	0,0025	0,0010
	1	0,0914	0,3151	0,3874	0,3474	0,3230	0,2684	0,1877	0,1211	0,0867	0,0725	0,0403	0,0207	0,0098
	2	0,0042	0,0746	0,1937	0,2759	0,2907	0,3020	0,2816	0,2335	0,1951	0,1757	0,1209	0,0763	0,0439
	3	0,0001	0,0105	0,0574	0,1298	0,1550	0,2013	0,2503	0,2668	0,2601	0,2522	0,2150	0,1665	0,1172
	4	0,0000	0,0010	0,0112	0,0401	0,0543	0,0881	0,1460	0,2001	0,2276	0,2377	0,2508	0,2384	0,2051
	5	0,0000	0,0001	0,0015	0,0085	0,0130	0,0264	0,0584	0,1029	0,1366	0,1536	0,2007	0,2340	0,2461
	6	0,0000	0,0000	0,0001	0,0012	0,0022	0,0055	0,0162	0,0368	0,0569	0,0689	0,1115	0,1596	0,2051
	7	0,0000	0,0000	0,0000	0,0001	0,0002	0,0008	0,0031	0,0090	0,0163	0,0212	0,0425	0,0746	0,1172
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0030	0,0043	0,0106	0,0229	0,0439
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0005	0,0016	0,0042	0,0098
11	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	
	0	0,8953	0,5688	0,3138	0,1673	0,1346	0,0859	0,0422	0,0198	0,0116	0,0088	0,0036	0,0014	0,0005
	1	0,0995	0,3293	0,3835	0,3248	0,2961	0,2362	0,1549	0,0932	0,0636	0,0518	0,0266	0,0125	0,0054
	2	0,0050	0,0867	0,2131	0,2866	0,2961	0,2953	0,2581	0,1998	0,1590	0,1395	0,0887	0,0513	0,0269
	3	0,0002	0,0137	0,0710	0,1517	0,1777	0,2215	0,2581	0,2568	0,2384	0,2254	0,1774	0,1259	0,0806
	4	0,0000	0,0014	0,0158	0,0536	0,0711	0,1107	0,1721	0,2201	0,2384	0,2428	0,2365	0,2060	0,1611
	5	0,0000	0,0001	0,0025	0,0132	0,0199	0,0388	0,0803	0,1321	0,1669	0,1830	0,2207	0,2360	0,2256
	6	0,0000	0,0000	0,0003	0,0023	0,0040	0,0097	0,0268	0,0566	0,0835	0,0985	0,1471	0,1931	0,2256
	7	0,0000	0,0000	0,0000	0,0003	0,0006	0,0017	0,0064	0,0173	0,0298	0,0379	0,0701	0,1128	0,1611
	8	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0011	0,0037	0,0075	0,0102	0,0234	0,0462	0,0806
12	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0012	0,0018	0,0052	0,0126
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0021	0,0054
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0005	
	0	0,8864	0,5404	0,2824	0,1422	0,1122	0,0687	0,0317	0,0138	0,0077	0,0057	0,0022	0,0008	0,0002
	1	0,1074	0,3413	0,3766	0,3012	0,2692	0,2062	0,1267	0,0712	0,0462	0,0368	0,0174	0,0075	0,0029
	2	0,0060	0,0988	0,2301	0,2924	0,2961	0,2835	0,2323	0,1678	0,1272	0,1088	0,0639	0,0339	0,0161
	3	0,0002	0,0173	0,0852	0,1720	0,1974	0,2362	0,2581	0,2397	0,2120	0,1954	0,1419	0,0923	0,0537
	4	0,0000	0,0021	0,0213	0,0683	0,0888	0,1329	0,1936	0,2311	0,2384	0,2367	0,2128	0,1700	0,1208
	5	0,0000	0,0002	0,0038	0,0193	0,0284	0,0532	0,1032	0,1585	0,1908	0,2039	0,2270	0,2225	0,1934
	6	0,0000	0,0000	0,0005	0,0040	0,0066	0,0155	0,0401	0,0792	0,1113	0,1281	0,1766	0,2124	0,2256
	7	0,0000	0,0000	0,0000	0,0006	0,0011	0,0033	0,0115	0,0291	0,0477	0,0591	0,1009	0,1489	0,1934
	8	0,0000	0,0000	0,0000	0,0001	0,0001	0,0005	0,0024	0,0078	0,0149	0,0199	0,0420	0,0762	0,1208
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0015	0,0033	0,0048	0,0125	0,0277
10	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0005	0,0008	0,0025	0,0068	0,0161
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0029
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0002

Tabla de probabilidades puntuales de la distribución $Binomial(n,p)$ (Continuación)
 $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

n	k	p												
		0,01	0,05	0,10	0,15	1/6	0,20	0,25	0,30	1/3	0,35	0,40	0,45	0,50
20	0	0,8179	0,3585	0,1216	0,0388	0,0261	0,0115	0,0032	0,0008	0,0003	0,0002	0,0000	0,0000	0,0000
	1	0,1652	0,3774	0,2702	0,1368	0,1043	0,0576	0,0211	0,0068	0,0030	0,0020	0,0005	0,0001	0,0000
	2	0,0159	0,1887	0,2852	0,2293	0,1982	0,1369	0,0669	0,0278	0,0143	0,0100	0,0031	0,0008	0,0002
	3	0,0010	0,0596	0,1901	0,2428	0,2379	0,2054	0,1339	0,0716	0,0429	0,0323	0,0123	0,0040	0,0011
	4	0,0000	0,0133	0,0898	0,1821	0,2022	0,2182	0,1897	0,1304	0,0911	0,0738	0,0350	0,0139	0,0046
	5	0,0000	0,0022	0,0319	0,1028	0,1294	0,1746	0,2023	0,1789	0,1457	0,1272	0,0746	0,0365	0,0148
	6	0,0000	0,0003	0,0089	0,0454	0,0647	0,1091	0,1686	0,1916	0,1821	0,1712	0,1244	0,0746	0,0370
	7	0,0000	0,0000	0,0020	0,0160	0,0259	0,0545	0,1124	0,1643	0,1821	0,1844	0,1659	0,1221	0,0739
	8	0,0000	0,0000	0,0004	0,0046	0,0084	0,0222	0,0609	0,1144	0,1480	0,1614	0,1797	0,1623	0,1201
	9	0,0000	0,0000	0,0001	0,0011	0,0022	0,0074	0,0271	0,0654	0,0987	0,1158	0,1597	0,1771	0,1602
	10	0,0000	0,0000	0,0000	0,0002	0,0005	0,0020	0,0099	0,0308	0,0543	0,0686	0,1171	0,1593	0,1762
	11	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0030	0,0120	0,0247	0,0336	0,0710	0,1185	0,1602
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0008	0,0039	0,0092	0,0136	0,0355	0,0727	0,1201
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0010	0,0028	0,0045	0,0146	0,0366	0,0739
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0007	0,0012	0,0049	0,0150	0,0370
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0013	0,0049	0,0148
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0003	0,0013	0,0046
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011
	18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002
	19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	20	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Tabla de probabilidades acumuladas de la distribución $Binomial(n,p)$

$$P(X \leq k) = \sum_{i=0}^k \binom{n}{i} p^i (1-p)^{n-i}.$$

n	k	p												
		0,01	0,05	0,10	0,15	1/6	0,20	0,25	0,30	1/3	0,35	0,40	0,45	0,50
5	1	0,9990	0,9774	0,9185	0,8352	0,8038	0,7373	0,6328	0,5282	0,4609	0,4284	0,3370	0,2562	0,1875
	2	1,0000	0,9988	0,9914	0,9734	0,9645	0,9421	0,8965	0,8369	0,7901	0,7648	0,6826	0,5931	0,5000
	3	1,0000	1,0000	0,9995	0,9978	0,9967	0,9933	0,9844	0,9692	0,9547	0,9460	0,9130	0,8688	0,8125
	4	1,0000	1,0000	1,0000	0,9999	0,9999	0,9997	0,9990	0,9976	0,9959	0,9947	0,9898	0,9815	0,9688
	5	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
6	1	0,9985	0,9672	0,8857	0,7765	0,7368	0,6554	0,5339	0,4202	0,3512	0,3191	0,2333	0,1636	0,1094
	2	1,0000	0,9978	0,9842	0,9527	0,9377	0,9011	0,8306	0,7443	0,6804	0,6471	0,5443	0,4415	0,3438
	3	1,0000	0,9999	0,9987	0,9941	0,9913	0,9830	0,9624	0,9295	0,8999	0,8826	0,8208	0,7447	0,6563
	4	1,0000	1,0000	0,9999	0,9996	0,9993	0,9984	0,9954	0,9891	0,9822	0,9777	0,9590	0,9308	0,8906
	5	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9998	0,9993	0,9986	0,9982	0,9959	0,9917	0,9844
	6	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
7	1	0,9980	0,9556	0,8503	0,7166	0,6698	0,5767	0,4449	0,3294	0,2634	0,2338	0,1586	0,1024	0,0625
	2	1,0000	0,9962	0,9743	0,9262	0,9042	0,8520	0,7564	0,6471	0,5706	0,5323	0,4199	0,3164	0,2266
	3	1,0000	0,9998	0,9973	0,9879	0,9824	0,9667	0,9294	0,8740	0,8267	0,8002	0,7102	0,6083	0,5000
	4	1,0000	1,0000	0,9998	0,9988	0,9980	0,9953	0,9871	0,9712	0,9547	0,9444	0,9037	0,8471	0,7734
	5	1,0000	1,0000	1,0000	0,9999	0,9999	0,9996	0,9987	0,9962	0,9931	0,9910	0,9812	0,9643	0,9375
	6	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9998	0,9995	0,9994	0,9984	0,9963	0,9922
	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
8	1	0,9973	0,9428	0,8131	0,6572	0,6047	0,5033	0,3671	0,2553	0,1951	0,1691	0,1064	0,0632	0,0352
	2	0,9999	0,9942	0,9619	0,8948	0,8652	0,7969	0,6785	0,5518	0,4682	0,4278	0,3154	0,2201	0,1445
	3	1,0000	0,9996	0,9950	0,9786	0,9693	0,9437	0,8862	0,8059	0,7414	0,7064	0,5941	0,4770	0,3633
	4	1,0000	1,0000	0,9996	0,9971	0,9954	0,9896	0,9727	0,9420	0,9121	0,8939	0,8263	0,7396	0,6367
	5	1,0000	1,0000	1,0000	0,9998	0,9996	0,9988	0,9958	0,9887	0,9803	0,9747	0,9502	0,9115	0,8555
	6	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9996	0,9987	0,9974	0,9964	0,9915	0,9819	0,9648
	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9998	0,9998	0,9993	0,9983	0,9961
	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
9	1	0,9966	0,9288	0,7748	0,5995	0,5427	0,4362	0,3003	0,1960	0,1431	0,1211	0,0705	0,0385	0,0195
	2	0,9999	0,9916	0,9470	0,8591	0,8217	0,7382	0,6007	0,4628	0,3772	0,3373	0,2318	0,1495	0,0898
	3	1,0000	0,9994	0,9917	0,9661	0,9520	0,9144	0,8343	0,7297	0,6503	0,6089	0,4826	0,3614	0,2539
	4	1,0000	1,0000	0,9991	0,9944	0,9910	0,9804	0,9511	0,9012	0,8552	0,8283	0,7334	0,6214	0,5000
	5	1,0000	1,0000	0,9999	0,9994	0,9989	0,9969	0,9900	0,9747	0,9576	0,9464	0,9006	0,8342	0,7461
	6	1,0000	1,0000	1,0000	1,0000	0,9999	0,9997	0,9987	0,9957	0,9917	0,9888	0,9750	0,9502	0,9102
	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9996	0,9990	0,9986	0,9962	0,9909	0,9805
	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999	0,9999	0,9997	0,9992	0,9980
	9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Tabla de probabilidades acumuladas de la distribución $Binomial(n,p)$ (Continuación)

$$P(X \leq k) = \sum_{i=0}^k \binom{n}{i} p^i (1-p)^{n-i}.$$

n	k	p												
		0,01	0,05	0,10	0,15	1/6	0,20	0,25	0,30	1/3	0,35	0,40	0,45	0,50
10	1	0,9957	0,9139	0,7361	0,5443	0,4845	0,3758	0,2440	0,1493	0,1040	0,0860	0,0464	0,0233	0,0107
	2	0,9999	0,9885	0,9298	0,8202	0,7752	0,6778	0,5256	0,3828	0,2991	0,2616	0,1673	0,0996	0,0547
	3	1,0000	0,9990	0,9872	0,9500	0,9303	0,8791	0,7759	0,6496	0,5593	0,5138	0,3823	0,2660	0,1719
	4	1,0000	0,9999	0,9984	0,9901	0,9845	0,9672	0,9219	0,8497	0,7869	0,7515	0,6331	0,5044	0,3770
	5	1,0000	1,0000	0,9999	0,9986	0,9976	0,9936	0,9803	0,9527	0,9234	0,9051	0,8338	0,7384	0,6230
	6	1,0000	1,0000	1,0000	0,9999	0,9997	0,9991	0,9965	0,9894	0,9803	0,9740	0,9452	0,8980	0,8281
	7	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9996	0,9984	0,9966	0,9952	0,9877	0,9726	0,9453
	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9996	0,9995	0,9983	0,9955	0,9893
	9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9997	0,9990
	10	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
11	1	0,9948	0,8981	0,6974	0,4922	0,4307	0,3221	0,1971	0,1130	0,0751	0,0606	0,0302	0,0139	0,0059
	2	0,9998	0,9848	0,9104	0,7788	0,7268	0,6174	0,4552	0,3127	0,2341	0,2001	0,1189	0,0652	0,0327
	3	1,0000	0,9984	0,9815	0,9306	0,9044	0,8389	0,7133	0,5696	0,4726	0,4256	0,2963	0,1911	0,1133
	4	1,0000	0,9999	0,9972	0,9841	0,9755	0,9496	0,8854	0,7897	0,7110	0,6683	0,5328	0,3971	0,2744
	5	1,0000	1,0000	0,9997	0,9973	0,9954	0,9883	0,9657	0,9218	0,8779	0,8513	0,7535	0,6331	0,5000
	6	1,0000	1,0000	1,0000	0,9997	0,9994	0,9980	0,9924	0,9784	0,9614	0,9499	0,9006	0,8262	0,7256
	7	1,0000	1,0000	1,0000	1,0000	0,9999	0,9998	0,9988	0,9957	0,9912	0,9878	0,9707	0,9390	0,8867
	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9994	0,9986	0,9980	0,9941	0,9852	0,9673
	9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9998	0,9993	0,9978	0,9941
	10	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9995
	11	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
12	1	0,9938	0,8816	0,6590	0,4435	0,3813	0,2749	0,1584	0,0850	0,0540	0,0424	0,0196	0,0083	0,0032
	2	0,9998	0,9804	0,8891	0,7358	0,6774	0,5583	0,3907	0,2528	0,1811	0,1513	0,0834	0,0421	0,0193
	3	1,0000	0,9978	0,9744	0,9078	0,8748	0,7946	0,6488	0,4925	0,3931	0,3467	0,2253	0,1345	0,0730
	4	1,0000	0,9998	0,9957	0,9761	0,9636	0,9274	0,8424	0,7237	0,6315	0,5833	0,4382	0,3044	0,1938
	5	1,0000	1,0000	0,9995	0,9954	0,9921	0,9806	0,9456	0,8822	0,8223	0,7873	0,6652	0,5269	0,3872
	6	1,0000	1,0000	0,9999	0,9993	0,9987	0,9961	0,9857	0,9614	0,9336	0,9154	0,8418	0,7393	0,6128
	7	1,0000	1,0000	1,0000	0,9999	0,9998	0,9994	0,9972	0,9905	0,9812	0,9745	0,9427	0,8883	0,8062
	8	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9996	0,9983	0,9961	0,9944	0,9847	0,9644	0,9270
	9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9995	0,9992	0,9972	0,9921	0,9807
	10	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9997	0,9989	0,9968
	11	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999	0,9998
	12	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Tabla de probabilidades acumuladas de la distribución $Binomial(n,p)$ (Continuación)

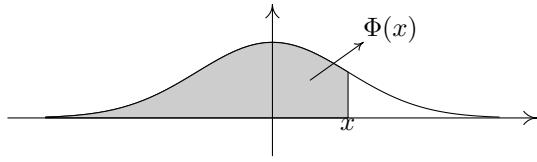
$$P(X \leq k) = \sum_{i=0}^k \binom{n}{i} p^i (1-p)^{n-i}.$$

n	k	p												
		0,01	0,05	0,10	0,15	1/6	0,20	0,25	0,30	1/3	0,35	0,40	0,45	0,50
20	0	0,8179	0,3585	0,1216	0,0388	0,0261	0,0115	0,0032	0,0008	0,0003	0,0002	0,0000	0,0000	0,0000
	1	0,9831	0,7358	0,3917	0,1756	0,1304	0,0692	0,0243	0,0076	0,0033	0,0021	0,0005	0,0001	0,0000
	2	0,9990	0,9245	0,6769	0,4049	0,3287	0,2061	0,0913	0,0355	0,0176	0,0121	0,0036	0,0009	0,0002
	3	1,0000	0,9841	0,8670	0,6477	0,5665	0,4114	0,2252	0,1071	0,0604	0,0444	0,0160	0,0049	0,0013
	4	1,0000	0,9974	0,9568	0,8298	0,7687	0,6296	0,4148	0,2375	0,1515	0,1182	0,0510	0,0189	0,0059
	5	1,0000	0,9997	0,9887	0,9327	0,8982	0,8042	0,6172	0,4164	0,2972	0,2454	0,1256	0,0553	0,0207
	6	1,0000	1,0000	0,9976	0,9781	0,9629	0,9133	0,7858	0,6080	0,4793	0,4166	0,2500	0,1299	0,0577
	7	1,0000	1,0000	0,9996	0,9941	0,9887	0,9679	0,8982	0,7723	0,6615	0,6010	0,4159	0,2520	0,1316
	8	1,0000	1,0000	0,9999	0,9987	0,9972	0,9900	0,9591	0,8867	0,8095	0,7624	0,5956	0,4143	0,2517
	9	1,0000	1,0000	1,0000	0,9998	0,9994	0,9974	0,9861	0,9520	0,9081	0,8782	0,7553	0,5914	0,4119
	10	1,0000	1,0000	1,0000	1,0000	0,9999	0,9994	0,9961	0,9829	0,9624	0,9468	0,8725	0,7507	0,5881
	11	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9991	0,9949	0,9870	0,9804	0,9435	0,8692	0,7483
	12	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9987	0,9963	0,9940	0,9790	0,9420	0,8684
	13	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9991	0,9985	0,9935	0,9786	0,9423
	14	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9997	0,9984	0,9936	0,9793
	15	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9985	0,9941
	16	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9987
	17	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998
	18	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
	19	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
	20	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

**Tabla de la función de distribución Φ
de una normal $N(0, 1)$ para $x \geq 0$**

$$\Phi(x) = P[X \leq x] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{t^2}{2}\right) dt$$

Si $x < 0 \implies \Phi(x) = 1 - \Phi(-x)$



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511966	0.515953	0.519939	0.523922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555670	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621720	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802337	0.805105	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823814	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.879000	0.881000	0.882977
1.2	0.884930	0.886861	0.888768	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903200	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935745	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959070	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965620	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	0.998605
3.0	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999
3.1	0.999032	0.999065	0.999096	0.999126	0.999155	0.999184	0.999211	0.999238	0.999264	0.999289
3.2	0.999313	0.999336	0.999359	0.999381	0.999402	0.999423	0.999443	0.999462	0.999481	0.999499
3.3	0.999517	0.999534	0.999550	0.999566	0.999581	0.999596	0.999610	0.999624	0.999638	0.999651
3.4	0.999663	0.999675	0.999687	0.999698	0.999709	0.999720	0.999730	0.999740	0.999749	0.999758
3.5	0.999767	0.999776	0.999784	0.999792	0.999800	0.999807	0.999815	0.999822	0.999828	0.999835
3.6	0.999841	0.999847	0.999853	0.999858	0.999864	0.999869	0.999874	0.999879	0.999883	0.999888
3.7	0.999892	0.999896	0.999900	0.999904	0.999908	0.999912	0.999915	0.999918	0.999922	0.999925
3.8	0.999928	0.999931	0.999933	0.999936	0.999938	0.999941	0.999943	0.999946	0.999948	0.999950
3.9	0.999952	0.999954	0.999956	0.999958	0.999959	0.999961	0.999963	0.999964	0.999966	0.999967
4.0	0.999968	0.999970	0.999971	0.999972	0.999973	0.999974	0.999975	0.999976	0.999977	0.999978

Apéndice D

El problema del borracho *

Enunciado

Un borracho parte de un farol dando pasos de igual longitud hacia ambos lados. ¿Cuál es la probabilidad de que después de N pasos vuelva al farol?

D.1. Esquema general de la resolución

En la Sección D.2 se analiza el problema desde una perspectiva general: se encuentra la probabilidad de que el borracho llegue, tras N pasos, cada uno con probabilidad p de avanzar hacia la derecha, a una posición m . Posteriormente, se hace $m = 0$ para dar respuesta a la pregunta del enunciado y se analizan los límites y el caso particular $p = \frac{1}{2}$, presentando gráficamente los principales resultados.

D.2. Resolución del problema

Sea un *borracho unidimensional*, que queda representado matemáticamente por una única variable: su posición, $m \in \mathbb{Z}$. Se ha supuesto, sin pérdida de generalidad (pues corresponde a tomar, arbitrariamente, un origen), que el borracho parte de $m_0 = 0$.

El borracho da N pasos. Un *paso* es un proceso en que se modifica la posición del borracho. Tras cada paso, la posición del borracho podrá, bien incrementarse en una unidad (lo que se denominará *dar un paso a la derecha*), bien decrementarse en una unidad (*dar un paso a la izquierda*)¹. Estos sucesos son obviamente *mutuamente excluyentes* y *complementarios* (en cada paso se da uno y sólo uno de ellos). Por tanto, si la probabilidad, en cada uno de los pasos, de que éste sea hacia la derecha es p , necesariamente la probabilidad de que el paso sea a la izquierda es $q \equiv 1 - p$, de modo que $p + q = 1$. Esto también implica que, si tras los N pasos, n_1 han sido a la derecha, el número de pasos a la izquierda habrá sido $n_2 \equiv N - n_1$.

¹Se han tomado los pasos de igual longitud, tal y como indica el enunciado.

Con estas definiciones previas, se está ya en disposición de resolver el problema.

D.2.1. Distribución binomial

El objetivo de esta sección es calcular la probabilidad, $W_N(n_1)$, de que, tras N pasos, n_1 hayan sido hacia la derecha y los restantes, $N - n_1$, hacia la izquierda.

Considérese, en primer lugar, la probabilidad de que se dé una secuencia concreta de pasos a izquierda y derecha, de modo que se cumpla la restricción anterior². Como cada paso es un evento independiente, y n eventos independientes, $\{S_i\}_{i=1}^n$, cumplen que

$$P\left(\bigcap_{i=1}^n S_i\right) = \prod_{i=1}^n P(S_i), \quad (\text{D.1})$$

entonces la probabilidad buscada, para una combinación como la descrita, no es más que $p^{n_1}(1-p)^{N-n_1}$, donde se han tenido en cuenta las ligaduras en q y n_2 . Sin embargo, no existe una única combinación de pasos a izquierda y derecha que resultan en n_1 pasos totales a derecha, sino que existen multitud de ellas.

Para cuantificarlas, es posible imaginar que existen tantas combinaciones posibles como maneras distintas hay de elegir n_1 elementos indistinguibles de un total de N . Esta cantidad no es más que³ el número combinatorio $\binom{N}{n_1}$. Por tanto, como todas las combinaciones son equiprobables, y se conoce la probabilidad de una de ellas y el número total de combinaciones que cumplen el requisito establecido, es directo afirmar que la probabilidad buscada es:

$$W_N(n_1) = \binom{N}{n_1} p^{n_1}(1-p)^{N-n_1} \quad (\text{D.2})$$

D.2.2. Distribución en función de la posición final

Si la posición inicial es $m_0 = 0$, y se toma como criterio de signos positivo hacia la derecha, siendo los pasos de igual longitud, la posición tras los N pasos será:

$$m = n_1 - n_2 = n_1 - (N - n_1) = 2n_1 - N \quad (\text{D.3})$$

Pueden comprobarse las siguientes propiedades intuitivas a partir de esta ecuación:

²Esto es, no se exige sólo que el número total de pasos a la derecha sea n_1 , sino, además, que se den en un orden concreto.

³Deducir esto es relativamente sencillo. Dados N elementos, de los que hay que elegir n_1 , tenemos N posibilidades para el primero, $N - 1$ para el segundo, ..., y $(N - n_1 + 1)$ para el n_1 -ésimo. Cada elección es independiente y por tanto esto da $N(N - 1) \cdots (N - n_1 + 1) \equiv \frac{N!}{(N - n_1)!}$ posibilidades distintas. Sin embargo, como los elementos son indistinguibles, es indiferente haber elegido los n_1 elementos en cualquier orden. Cada combinación única se ha tenido en cuenta $n_1!$ veces (el número de permutaciones de n_1 elementos). Por tanto, el número real de combinaciones de N elementos de los que se escogen n_1 es $\frac{N!}{(N - n_1)!n_1!} \equiv \binom{N}{n_1} = \binom{N}{N - n_1}$, número al que se llama *coeficiente binomial*.

1. Como $0 \leq n_1 \leq N$, entonces $-N \leq m \leq N$. En N pasos, el borracho no puede alejarse más de N posiciones de la inicial.
 2. Como $2n_1$ es par $\forall n_1 \in \mathbb{N}$, entonces m y N son ambos pares o impares. Con un número par de pasos, sólo es posible acceder a las posiciones pares como estado final. La afirmación análoga para los impares es igualmente cierta.
- a) Como corolario a este resultado, para que $m = 0$, N ha de ser par.

Despejando de la ecuación D.3, $n_1 = \frac{N+m}{2}$. Sustituyendo esto último en la ecuación D.2, y llamando $P(x = m|N)$ a la probabilidad de que la posición tras N pasos sea m , se llega al resultado⁴:

$$P(x = m|N) = W_N \left(n_1 = \frac{m+N}{2} \right) = \frac{N!}{\left(\frac{N+m}{2}\right)! \left(\frac{N-m}{2}\right)!} p^{\frac{N+m}{2}} (1-p)^{\frac{N-m}{2}} \quad (\text{D.4})$$

Las propiedades 1 y 2 deducidas anteriormente garantizan que los números $\frac{N+m}{2}, \frac{N-m}{2} \in \mathbb{N}$, y por tanto los factoriales están bien definidos. Aunque no se pide explícitamente, puede ser interesante comprobar que, en el caso extremo, cuando $m = N$, el prefactor (i.e., el coeficiente binomial) toma como valor la unidad y, entonces, $P(x = N|N) = p^N$. Sólo existe una posibilidad: que todos los movimientos sean hacia la derecha. El análisis para $m = -N$ es del todo análogo, obteniéndose $(1-p)^N$ como resultado. Con $0 \leq p \leq 1$, $P(x = \pm N|N)$ definen sendas sucesiones monótonas y decrecientes, con límite 0 cuando $N \rightarrow \infty$. Esto indica que, *a mayor número de pasos, la probabilidad de que el borracho llegue cualquiera de los extremos cae exponencialmente*.

D.2.3. Probabilidad de volver al farol: caso particular $m = 0$

La probabilidad de que el borracho vuelva a la posición inicial puede ser obtenida fácilmente, sin más que sustituir $m = 0$ en la ecuación D.4. Teniendo en cuenta, como se ha deducido previamente (propiedad 2.1 de la Sección D.2.1), que N ha de ser par, se tiene la probabilidad dada por la ecuación D.5.

$$P(x = 0|N) = \frac{N!}{\left(\frac{N}{2}\right)! \left(\frac{N}{2}\right)!} p^{N/2} (1-p)^{N/2} = \binom{N}{\frac{N}{2}} [p(1-p)]^{\frac{N}{2}} \quad (\text{D.5})$$

Es posible, previo a la representación gráfica, analizar ciertas propiedades del resultado obtenido.

- Si se analiza la dependencia de la probabilidad obtenida con p , al ser $P(x = 0|N)$ una función monótona y creciente de $p(1-p)$, los máximos de P se corresponden con los máximos de $p(1-p)$. Esta última es una parábola invertida con vértice en $p = \frac{1}{2}$, por lo que se puede

⁴También se ha calculado, con álgebra elemental, $N - n_1 = N - \frac{m+N}{2} = \frac{N-m}{2}$.

concluir que *la probabilidad de que el borracho vuelva a la posición inicial se maximiza cuando ambos pasos, a izquierda y derecha, son igualmente probables*. Resulta razonable que, de no ser así, existirá un sesgo que favorecerá que la posición final sea aquella cuyos pasos son más probables⁵.

- El factor $\binom{N}{\frac{N}{2}}$, que da el número de combinaciones posibles que devuelven al borracho a la posición inicial, crece⁶ con el número, N , de pasos. Por otro lado, al ser $p(1-p) < 1$, el factor $[p(1-p)]^{\frac{N}{2}}$ cae con N : cada una de las posibles combinaciones es menos probable. Por tanto, parece complicado deducir si $P(x = 0|N)$ crecerá o decrecerá con N . Por ello, se dará respuesta a este interrogante a partir de la representación gráfica, o bien, a partir del análisis de un caso particular.

Caso particular: ambos sucesos equiprobables ($p = \frac{1}{2}$)

En el caso particular $p = \frac{1}{2}$, esto es, el borracho tiene igual probabilidad de moverse a izquierda o derecha en cada paso, la ecuación D.5 se simplifica a:

$$P_{p=\frac{1}{2}}(x = 0|N) = \binom{N}{\frac{N}{2}} \frac{1}{2^N} \quad (\text{D.6})$$

Esta expresión, además de ser más simple, permite analizar el caso límite $N \rightarrow \infty$. Si se tiene en cuenta que $\binom{N}{\frac{N}{2}}$ es el coeficiente binomial central, el elemento central de la fila N del triángulo de Pascal (donde N es par), y que 2^N es la suma de los elementos de la N -ésima fila, la probabilidad obtenida no es más que la razón entre estos dos números. Esta razón disminuye⁷ con $N \rightarrow \infty$. Según habíamos analizado, $p = \frac{1}{2}$ daba la máxima probabilidad a un N fijo. Por tanto, es una cota superior de la probabilidad de que el borracho vuelva al origen, para cada N .

Como la cota superior decrece con N , este razonamiento permite concluir que, *a mayor número de pasos N , menor es la probabilidad de que el borracho vuelva al origen*.

D.2.4. Análisis gráfico de las ecuaciones D.4 y D.5

En primer lugar, se pretende observar gráficamente cómo se distribuye la probabilidad de que el borracho acabe, tras N pasos, en una posición final m . En este análisis, se han fijado N y p , y se ha graficado la ecuación D.4 en función de m . La propiedad 2 de la Sección D.2.2

⁵De hecho, el valor esperado de n_1 vendrá dado, según la expresión del valor esperado de una variable binomial, por $\bar{n}_1 = Np$. Utilizando la ecuación D.3, $\bar{m} = N(2p - 1)$.

⁶Esto puede deducirse de la identificación de los coeficientes binomiales con los valores del *triángulo de Pascal* o *triángulo de Tartaglia*. Este coeficiente binomial recibe el nombre de *coeficiente binomial central*, y se corresponde con los valores sobre el eje de simetría del triángulo. Estos valores crecen con N , lo que demuestra la afirmación precedente.

⁷Aunque no se demuestra explícitamente este resultado, porque excede el objetivo de este análisis, puede “comprobarse”: para la fila $N = 2$, dicho cociente es $\frac{2}{2^2} = 0.5$; para la fila $N = 4$, $\frac{6}{2^4} = 0.375$; para la fila $N = 16$, $\frac{12870}{2^{16}} \approx 0.196$.

establece que m y N son ambos pares o impares. Se han tomado N pares para exemplificar. La probabilidad de que el borracho acabe en m impares es, por tanto, nula. Para la claridad de la representación gráfica, se han representado los puntos de m par, y se han unido los resultados con líneas. En todo caso, ha de tenerse en cuenta que la distribución es discreta y sólo está definida para $m \in \mathbb{N}$, aunque se hayan introducido las líneas para facilitar la visualización.

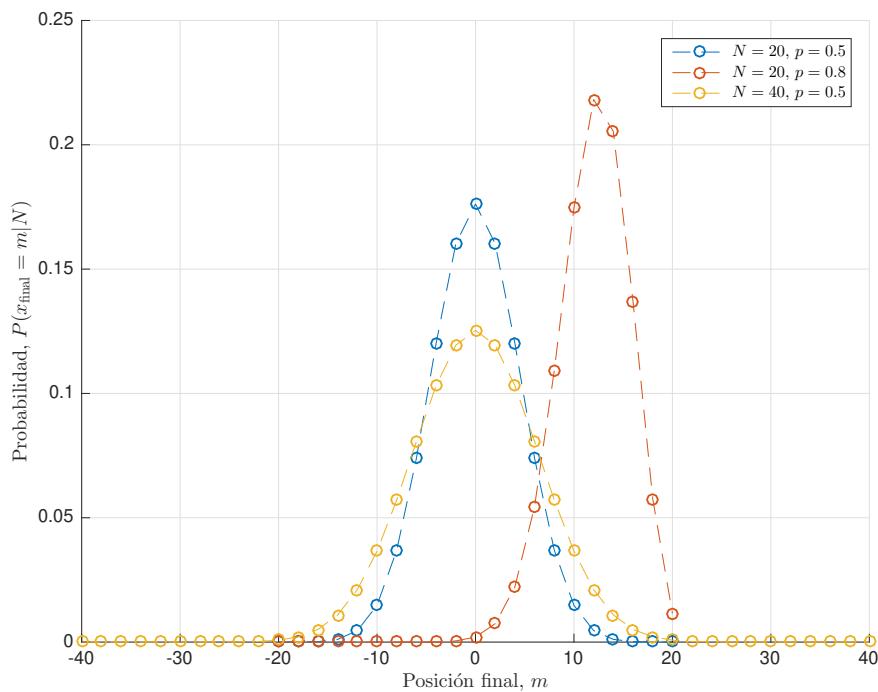


Figura D.1: Probabilidad de que el borracho, tras N pasos, acabe en una posición m . Se han estudiado varios casos, variando tanto p como N , que se encuentran detallados en la leyenda.

En la Figura D.1 se muestran diversas distribuciones de probabilidad de que el borracho acabe en una posición m .

- La curva azul ha sido generada con $N = 20$ y pasos a cada lado igualmente probables, $p = 0.5$. Tal y como sería razonable, en estas condiciones la distribución de probabilidad es par ($P(m) = P(-m)$) y tiene máximo en $m = 0$: la posición final más probable es el origen. Sin embargo, los valores de m cercanos a 0 tienen también probabilidades elevadas, y los valores extremos de m (cercaos a $\pm N$) son altamente improbables (corresponden a [casi] todos los pasos del borracho en un mismo sentido).
- Manteniendo fijo el número de pasos, $N = 20$, se ha variado la probabilidad p de que el paso sea hacia la derecha, haciendo este suceso más probable ($p > 0.5$). Consecuentemente, toda la gráfica se ha distorsionado, desplazándose hacia la derecha. Con $p = 0.8$, el valor esperado de los pasos a derecha, n_1 , es $Np = 16$. Así, 4 pasos son hacia la izquierda y el valor esperado de m sería 12, lo que concuerda a la perfección con la línea roja de la gráfica.
- Con $p = 0.5$ pero duplicando N , se obtiene, de nuevo, una curva con simetría par. Sin embargo, ésta se ha suavizado (presenta un pico menor): aunque el valor más probable sigue siendo $m = 0$, éste es ahora menos probable. Esto está en concordancia con lo deducido en la Sección

D.2.3: la probabilidad de volver al origen cae a medida que N toma valores mayores. Al mismo tiempo, la distribución ha crecido en anchura (ahora son posibles valores más lejanos de m), pero no en anchura relativa (pese a que N se ha duplicado, la anchura de la distribución ha crecido en un factor menor).

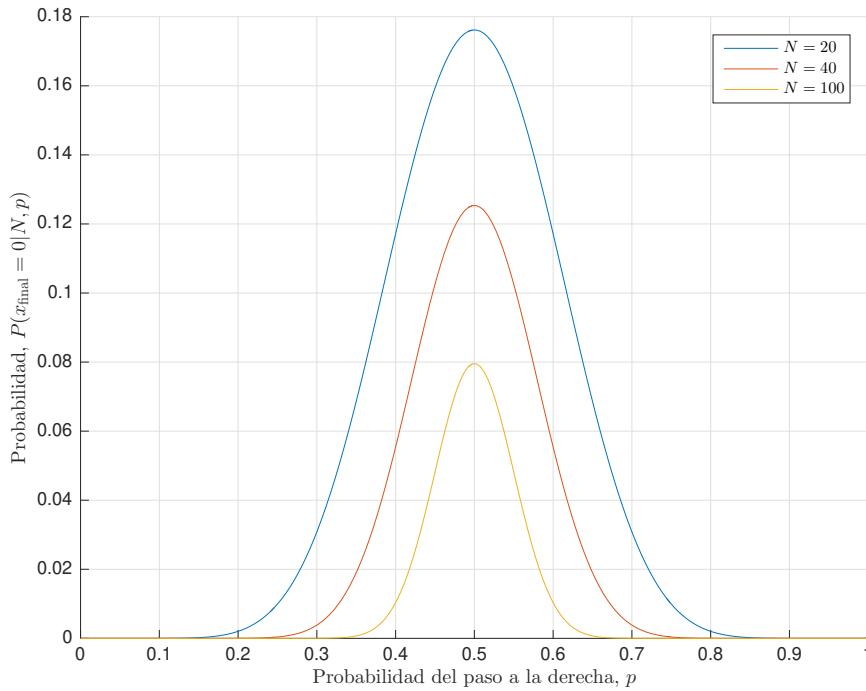


Figura D.2: Probabilidad de que el borracho, tras N pasos y con probabilidad p de que cada paso sea a la derecha, vuelva a la posición inicial.

Por otro lado, se ha analizado la ecuación D.5 manteniendo N constante en función de p (que puede tomar cualquier valor real entre 0 y 1). Para varios N , se muestran las curvas de probabilidad de que el borracho vuelva a la posición inicial en función de p en la Figura D.2.

Se observa, tal y como se había apuntado, que la máxima probabilidad de que el borracho vuelva se da cuando $p = \frac{1}{2}$ (entonces, el valor esperado de pasos a izquierda y derecha coinciden); y que, a mayores N , cae la probabilidad de que el borracho regrese al origen (no obstante, sigue siendo, en todo caso, el resultado más probable $\forall N$). Además, es también interesante señalar que, a medida que N se incrementa, la anchura de la distribución de probabilidad en función de p decrece: es decir, a mayores N , la ventana de valores de p que permiten que la probabilidad de regreso del borracho a su posición inicial tenga valores no despreciables se hace más pequeña. Esto es razonable, ya que, cuando N aumenta, la dispersión de valores en torno al valor esperado también aumenta (ver Figura D.1), y es necesario restringirse a p más cercanos a $\frac{1}{2}$ para que la probabilidad de que el borracho vuelva al origen sea alta.