# HERRAMIENTAS COMPUTACIONALES EN EL LABORATORIO

## PROFESOR: PABLO ALBELLA

*PRIMER CURSO DEL GRADO EN FÍSICA*


**1. INTRODUCCIÓN …. 1**

*Origen de la Estadística ….. 1*
*Lenguaje: algunas definiciones …. 1*
*Pensamiento crítico y estadístico …. 2*
*Ramas de la Estadística …. 3*

**2. ESTADÍSTICA DESCRIPTIVA …. 3**

*Representación gráfica …3*
*Medidas de centralización …. 12*
*Medidas de dispersión …. 13*

**3. PROBABILIDAD …. 18**

*Algunas definiciones …. 18*
*Conjuntos …. 18*
*Definición frecuentista de probabilidad …. 19*
*Definición axiomática de probabilidad …. 19*
*Propiedades de la probabilidad …. 19*
*Determinación sistemática de probabilidades …. 20*
*Técnicas para contar …. 20*
*Probabilidad condicionada …. 22*


[1] Se prestará especial atención a la distribución binomial, uniforme discreta, de Poisson, uniforme continua, distribución normal y normal estándar

# 1. INTRODUCCIÓN

La estadística es la ciencia que se encarga de la planificación de estudios y experimentos: obtener datos; y después organizar, resumir, presentar, analizar e interpretar esos datos para sacar conclusiones basadas en ellos.

Los conceptos y métodos estadísticos no solo son útiles sino a menudo indispensables para comprender el mundo que nos rodea. Proporcionan formas de obtener nuevas visiones sobre el comportamiento de muchos fenómenos que nos podemos encontrar en física[2]. La estadística nos permite hacer juicios inteligentes y tomar decisiones informadas ante preguntas desconcertantes que involucran incertidumbre y variación[3].

## 1.1 Origen de la Estadística

La palabra estadística se deriva del latín status (estado). Los métodos estadísticos se han desarrollado a lo largo de muchos siglos. Los primeros usos de la estadística consistieron en recopilaciones de datos y gráficos que describían diversos aspectos de un estado o un país. En 1662, John Graunt publicó información estadística sobre nacimientos y defunciones. Al trabajo de Graunt le siguieron estudios de tasas de mortalidad y enfermedad, tamaños de poblaciones y de tasas de ingresos y de desempleo[4]. Tanto hogares, gobiernos como negocios dependen de datos estadísticos para orientarse. Por ejemplo, las tasas de desempleo, de inflación, índices de consumo y tasas de natalidad y mortalidad se recopilan cuidadosamente de forma regular, sirviendo a los líderes de los negocios para tomar decisiones que afectan a contrataciones futuras, niveles de producción y expansiones hacia otros mercados.

## 1.2 Lenguaje: algunas definiciones

En este apartado damos definiciones formales de algunos conceptos básicos:

- *Datos:* Colecciones de observaciones, como medidas, géneros, o respuestas a encuestas.
- *Población:* Colección completa de todas las medidas o datos que se están considerando[5].
- *Censo:* Colección de datos de todos y cada uno de los miembros de una población[6].
- *Muestra:* Subcolección de miembros seleccionados de una población[7].
- *Variable:* Cualquier característica cuyo valor puede variar de un miembro a otro de una población[8].

  *Variable discreta:* el conjunto de posibles valores es, o bien finito (“numerable”), o bien se puede enumerar en una secuencia infinita.

  *Variable continua:* el conjunto de posibles valores consiste en un intervalo entero sobre la recta numérica (la colección de posibles valores no es numerable).

[2] Una de las ramas más importantes de la física es la llamada física estadística, que se encarga del estudio de las propiedades de la materia condensada en términos de las leyes físicas que gobiernan el movimiento de los átomos que la forman, utilizando teoría de probabilidad y métodos estadísticos.

[3] La presencia de incertidumbre o variación es un requisito para poder aplicar los métodos estadísticos. Si todas las medidas experimentales fueran iguales, una sola observación sería suficiente para obtener toda la información deseada.

[4] Para ampliar información, recomendamos recurrir al excelente libro de Hald [2003].

[5] Las medidas numéricas que describen características de una población se denominan parámetros.

[6] Limitaciones de tiempo, dinero u otros recursos escasos suelen hacer los censos poco prácticos o imposibles.

[7] Las medidas numéricas que describen características de una muestra se denominan estadísticos.

[8] Los datos resultan de realizar observaciones sobre una sola variable (conjunto de datos univariado) o simultáneamente sobre dos (bivariado) o más variables (multivariado).

## 1.3 Pensamiento crítico y estadístico

Cuando llevamos a cabo un estudio estadístico[9], debemos realizar un proceso de "preparación, análisis y conclusión" (ver esquema más abajo, adaptado de Triola [2012]). Comenzamos con una preparación que implica consideraciones sobre el contexto, la fuente de los datos y el método de muestreo. Continuamos construyendo gráficos adecuados, explorando los datos y ejecutando los cálculos necesarios para el método estadístico que estemos utilizando. Finalmente sacamos conclusiones determinando si los resultados tienen significancia estadística y práctica[10].

*Preparación*

1. Contexto
    a. ¿Qué significan los datos?
    b. ¿Cuál es el propósito del estudio?
2. Fuente de los Datos
    a. ¿Los datos provienen de una fuente con especial interés de forma que existan presiones para obtener resultados favorables a la fuente?
3. Método de Muestreo
    a. ¿Hubo o no hubo sesgo al recopilar los datos? [11]

*Análisis*

1. Representar los Datos
2. Explorar los Datos
    a. ¿Existen valores atípicos[12]?
    b. ¿Qué estadísticos[13] importantes resumen los datos?
    c. ¿Cómo están distribuidos los datos?
    d. ¿Faltan datos?
3. Aplicar métodos estadísticos
    a. Utilizar tecnología (software) para obtener resultados.

*Conclusión*

1. Significancia estadística
    a. ¿Tienen los resultados significancia estadística?
    b. ¿Tienen los resultados significancia práctica?

[14] Pensemos en histogramas, diagramas de barras, diagramas de dispersión, etc.
[15] Media, mediana (percentiles) y moda.
[16] Rango, varianza y desviación estándar.
[17] En este curso estudiaremos únicamente el muestreo y la estimación puntual.

### 1.4 Ramas de la Estadística

Los estudios estadísticos suelen producir tal número de datos que suele ser necesario resumirlos, para así poder describir sus características más importantes. LA ESTADÍSTICA DESCRIPTIVA —el arte de resumir datos— se ocupa de la representación gráfica[14] y del estudio de las *medidas de resumen* (como las de centralización[15], dispersión[16], etc). La Estadística Descriptiva también se encarga del análisis de las relaciones —las dependencias— entre variables, donde la *correlación* y la *regresión* son los principales temas.

Habiendo obtenido una muestra de una población, podemos (debemos) estar interesados en sacar conclusiones de la población a partir de la información obtenida de la muestra. LA ESTADÍSTICA INFERENCIAL es la rama de la Estadística que se ocupa de hacer generalizaciones válidas a partir de muestras. Los procedimientos de inferencia más importantes son la *estimación puntual*[17], los tests de hipótesis y la estimación por intervalos de confianza.

El puente que une ambas ramas es la PROBABILIDAD. Tanto la Estadística como la Probabilidad se ocupan de preguntas que involucran poblaciones y muestras, pero lo hacen de "forma inversa" una a la otra. En un problema de probabilidad, las propiedades de la población se dan por conocidas, y se hacen y responden preguntas sobre una muestra tomada de esa población. En un problema estadístico, el experimentador tiene acceso a las características de una muestra, y mediante técnicas inferenciales es capaz de sacar conclusiones sobre la población. La relación entre ambas disciplinas se puede resumir diciendo que la Probabilidad razona desde la población a la muestra (razonamiento deductivo), mientras que la Estadística Inferencial razona desde la muestra a la población (razonamiento inductivo). *Esto se ilustra en la Figura 1.*

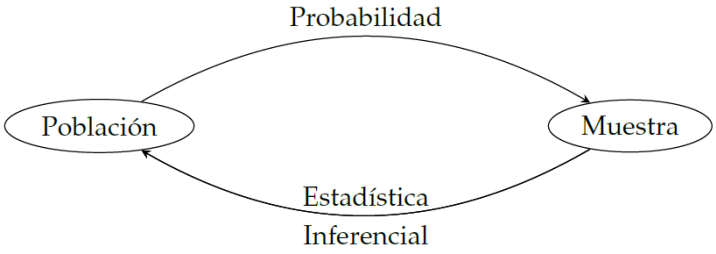



Figura 1: La relación entre la Probabilidad y la Estadística Inferencial. Adaptada de Devore [2014].

## 2. ESTADÍSTICA DESCRIPTIVA

La estadística descriptiva puede dividirse en dos grandes áreas. Por un lado, la representación gráfica de los datos y por otro el resumen numérico de los datos mediante medidas de centralización y dispersión. Las principales cinco características de los datos son:

1. Centro: Valor representativo que indica dónde se encuentra el centro del conjunto de datos.
2. Variación: Medida de la cantidad que varían los valores de los datos.
3. Distribución: Naturaleza o forma de la dispersión de los datos en el rango de valores.
4. Valores atípicos: Valores de la muestra que se encuentran muy lejos de la gran mayoría de los otros valores de la muestra.
5. Tiempo: Cualquier cambio con el tiempo en las características de los datos.

[18] Existen varios algoritmos para determinar el número adecuado de clases en las que tenemos que dividir nuestros datos. Por ejemplo, según la llamada *regla de Sturges*, el número ideal de clases puede aproximarse por $1 + \log_2 n$, siendo n el número de datos en nuestra muestra

## 2.1 Representación gráfica

Muchas muestras de datos son tan grandes que para comprenderlas necesitamos organizar y resumir los datos, bien numéricamente en tablas, o bien visualmente en gráficas, como describiremos a continuación.

A la hora de organizar y resumir conjuntos muy grandes de datos, una distribución de frecuencias (o tabla de frecuencias) puede ser muy útil para entender la naturaleza de la distribución de los datos (ver Tablas 1 y 2).

Tabla 1: Números que salen al tirar un dado 100 veces.

| Número (clase) | Frecuencia absoluta | Frecuencia relativa | % |
|---|---|---|---|
| 1 | 16 | 0.16 | 16 |
| 2 | 19 | 0.19 | 19 |
| 3 | 12 | 0.12 | 12 |
| 4 | 12 | 0.12 | 12 |
| 5 | 23 | 0.23 | 23 |
| 6 | 18 | 0.18 | 18 |
| TOTAL | 100 | 1 | 100 |

Tabla 2: Números que salen al tirar un dado 1000 veces.

| Número (clase) | Frecuencia absoluta | Frecuencia relativa | % |
|---|---|---|---|
| 1 | 154 | 0.15 | 15 |
| 2 | 177 | 0.18 | 18 |
| 3 | 162 | 0.16 | 16 |
| 4 | 164 | 0.16 | 16 |
| 5 | 176 | 0.18 | 18 |
| 6 | 167 | 0.17 | 17 |
| TOTAL | 1000 | 1 | 100 |

*Distribución de frecuencias*: Una distribución de frecuencias (o tabla de frecuencias) muestra cómo están agrupados los datos en categorías (clases) mutuamente excluyentes, enumerando las categorías junto con el número (frecuencia) de valores en cada categoría.

*Distribución de frecuencias relativa*: Una variación de la distribución de frecuencias básica (absoluta) es la distribución de frecuencias relativa, en la que cada frecuencia se reemplaza por una frecuencia relativa (proporción) o un porcentaje.

$$Frecuencia\ relativa\ de\ una\ clase = \frac{frecuencia\ de\ la\ clase}{suma\ de\ todas\ las\ frecuencias}$$

$$Porcentaje\ de\ una\ clase\ = \frac{frecuencia\ de\ la\ clase}{suma\ de\ todas\ las\ frecuencias} \times 100\%$$

[18] Existen varios algoritmos para determinar el número adecuado de clases en las que tenemos que dividir nuestros datos. Por ejemplo, según la llamada *regla de Sturges*, el número ideal de clases puede aproximarse por $1 + \log_2 n$, siendo n el número de datos en nuestra muestra

De cara a construir tablas de frecuencias y sus representaciones gráficas correspondientes, necesitamos definir una serie de términos asociados a las clases:

*Límites inferiores*: de clase Son los números más pequeños que pueden pertenecer a las distintas clases.

*Límites superiores:* de clase Son los números más grandes que pueden pertenecer a las distintas clases.

*Fronteras de clase:* Son los números que se utilizan para separar las clases, pero sin los huecos creados por los límites. También se conocen como límites reales de clase.

*Marcas de clase:* Son los puntos intermedios de cada clase. Se obtienen promediando los límites inferior y superior de clase.

*Anchura de clase* Es la diferencia entre dos límites de clase consecutivos en una distribución de frecuencias.

Construimos distribuciones de frecuencias para (1) resumir conjuntos grandes de datos, (2) analizar los datos para ver su distribución e identificar valores atípicos, y (3) tener una base para construir gráficos (como *histogramas*, que introduciremos más adelante).

Para construir una distribución o tabla de frecuencias se puede utilizar el siguiente procedimiento:

1. Seleccionar el número de clases, habitualmente entre 5 y 2018.
2. Calcular la anchura de clase.

$$Anchura\ de\ clase\ \approx \frac{(valor\ máximo\ -\ valor\ mínimo)}{número\ de\ clases}$$

   Generalmente conviene redondear hacia arriba este valor.

3. Elegir el valor del primer límite inferior de clase bien utilizando el valor mínimo de los datos bien otro valor por debajo de él.
4. Utilizando el primer límite inferior y la anchura de clase, enumerar el resto de límites inferiores de clase, añadiendo la anchura de clase en cada caso.
5. Enumerar los límites inferiores de clase en una columna vertical y determinar entonces los límites superiores de clase.
6. Tomar cada valor individual de los datos y asignarlo a una clase concreta. LA FRECUENCIA (absoluta) de esa clase es el número de valores existente en ella.

[18] Existen varios algoritmos para determinar el número adecuado de clases en las que tenemos que dividir nuestros datos. Por ejemplo, según la llamada *regla de Sturges*, el número ideal de clases puede aproximarse por $1 + \log_2 n$, siendo n el número de datos en nuestra muestra

***Distribución de frecuencias cumulativa*:** Otra variación de la distribución de frecuencias es la llamada DISTRIBUCIÓN DE FRECUENCIAS CUMULATIVA, en la que la frecuencia de cada clase es la suma de las frecuencias de esa clase y todas las anteriores (ver Tabla 3).

Tabla 3: Números que salen al tirar un dado 100 veces.

| Número (clase) | Frecuencia cumulativa |
|---|---|
| 1 o menos | 16 |
| 2 o menos | 35 |
| 3 o menos | 47 |
| 4 o menos | 59 |
| 5 o menos | 82 |
| 6 o menos | 100 |

**EJEMPLO 1: RESUMEN CONSTRUCCIÓN DE TABLAS DE FRECUENCIAS VARIABLE DISCRETA**

Los datos que se dan a continuación corresponden a los pesos en Kg de 80 personas:

60 ; 66 ; 77 ; 70 ; 66 ; 68 ; 57 ;70 ; 66 ; 52 ; 75 ; 65 ; 69 ; 71 ; 58 ; 66 ; 67 ; 74 ; 61 ; 63 ; 69 ; 80 ; 59 ; 66 ; 70 ; 67 ; 78 ; 75 ; 64 ; 71 ; 81 ; 62 ; 64 ; 69 ; 68 ; 72 ; 83 ; 56 ; 65 ; 74 ; 67 ; 54 ; 65 ; 65; 69 ; 61 ; 67 ; 73 ; 57 ; 62 ; 67; 68 ; 63 ; 67 ; 71 ; 68 ; 76 ; 61 ; 62 ; 63 ; 76 ; 61 ; 67 ; 67 ; 64 ; 72 ; 64 ; 73 ; 79 ; 58 ; 67 ; 71 ; 68 ; 59 ; 69 ; 70 ; 66 ; 62 ; 63 ; 66 ;

1. Obtener una distribución de datos en intervalos de amplitud 5, siendo el primer intervalo [50; 55) ó 50-54, dependiendo de la notación.
2. Calcular el porcentaje de personas de peso menor que 65 Kg.
3. ¿Cuántas personas tienen peso mayor o igual que 70 Kg, pero menor que 85?

Como se trata de efectuar una distribución de datos agrupados, debemos obtener primero los intervalos correspondientes, situando los datos en sus lugares respectivos:

| Número (clase) | Frecuencia absoluta | Frecuencia cumulativa |
|---|---|---|
| [50-55) | 2 | 2 |
| [55-60) | 7 | 9 |
| [60-65) | 17 | 26 |
| [65-70) | 30 | 56 |
| [70-75) | 14 | 70 |
| [75-80) | 7 | 77 |
| [80-85) | 3 | 80 |
| | 80 | |

| Número (clase) | Frecuencia absoluta | Frecuencia cumulativa |
|---|---|---|
| 50-54 | 2 | 2 |
| 55-59 | 7 | 9 |
| 60-64 | 17 | 26 |
| 65-69 | 30 | 56 |
| 70-74 | 14 | 70 |
| 75-79 | 7 | 77 |
| 80-84 | 3 | 80 |
| | 80 | |

2. Para calcular el porcentaje de personas de peso menor que 65kg, lo más efectivo es calcular la tabla de frecuencias cumulativa a partir de las frecuencias absolutas:

| Número (clase) | Frecuencia absoluta | Frecuencia cumulativa |
|---|---|---|
| 50-54 | 2 | 2 |
| 55-59 | 7 | 9 |
| 60-64 | 17 | 26 |
| 65-69 | 30 | 56 |
| 70-74 | 14 | 70 |
| 75-79 | 7 | 77 |
| 80-84 | 3 | 80 |
| | 80 | |

Observando la columna de frecuencias acumuladas se deduce que existen N3 = 26 individuos cuyo peso es menor que 65 Kg, que en términos de porcentaje corresponden a:

$$\frac{26}{80} \cdot 100 = 32.5\%$$

3. Para calcular cuántas personas tienen peso mayor o igual que 70 Kg. pero menor que 85:

El número de individuos con peso comprendido entre 70 y 85 Kg. es: $n_5 + n_6 + n_7 = 14+7+3 = 24$
O lo que es equivalente a utilizar la frecuencia cumulativa: $N_7 - N_4 = 80 - 56 = 24$

**EJEMPLO 2: Crear tablas variable discreta**

Sobre un grupo de n = 21 personas se realizan las siguientes observaciones de sus pesos, medidos en kilogramos:

| $X \sim x_1 + x_2 + \ \ldots + x_{21}$ | | | | | | |
|---|---|---|---|---|---|---|
| 58 | 42 | 51 | 54 | 40 | 39 | 49 |
| 56 | 58 | 57 | 59 | 63 | 58 | 66 |
| 70 | 72 | 71 | 69 | 70 | 68 | 64 |

Agrupar los datos en una tabla estadística.

**Solución**

- Tenemos 21 observaciones o datos, una recomendación es agruparlo en *k* intervalos con $k \sim \sqrt{X} = \sqrt{21} \sim 4$ *o* $5$
- $l_0 = x_{min} = 39$ *y* $l_5 = x_{max} = 72$ .
- *A =* $l_5 - l_0 =$*33 por lo que* $a = \frac{A}{k} = \frac{33}{5} = 6.6$
- Entonces, tomamos *k* = 5 intervalos de longitud *a* = 6,6 comenzando por $l_0 = x_{min} = 39$ y terminando en $l_5 = x_{max} = 72$

Entonces, tomamos k = 5 intervalos de longitud a = 6,6 comenzando por $l_0 = x_{min} = 39$ y terminando en $l_5 = x_{max} = 72$

| Clases $i_{i-1}$ - $i_i$ | Marca $c_i$ | Frecuencia absoluta $f_i$ | Frecuencia relativa $f_i / n$ | *Frecuencia Cumulativa* $F_i$ |
|---|---|---|---|---|
| 39-45.6 | 42.3 | 3 | 0.1428 | 3 |
| 45.6-52.2 | 48.9 | 2 | 0.0952 | 5 |
| 52.2-58.8 | 55.5 | 6 | 0.2857 | 11 |
| 58.8-65.4 | 62.1 | 3 | 0.1428 | 14 |
| 65.4-72 | 68.7 | 7 | 0.3333 | 21 |
| | | n=21 | 1 | |

***HISTOGRAMAS**:* Si bien una distribución de frecuencias es una herramienta útil para resumir datos e investigar su distribución, una herramienta aún mejor es el *histograma*, que consiste en un gráfico que es más fácil de interpretar que una tabla de números.

UN HISTOGRAMA es un gráfico consistente en barras de la misma anchura adyacentes unas a otras (a menos que existan huecos en los datos). El eje horizontal representa clases de valores cuantitativos y el eje vertical representa frecuencias. Las alturas de las barras corresponden a los valores de las frecuencias, de forma que un histograma es básicamente un gráfico de una distribución de frecuencias.

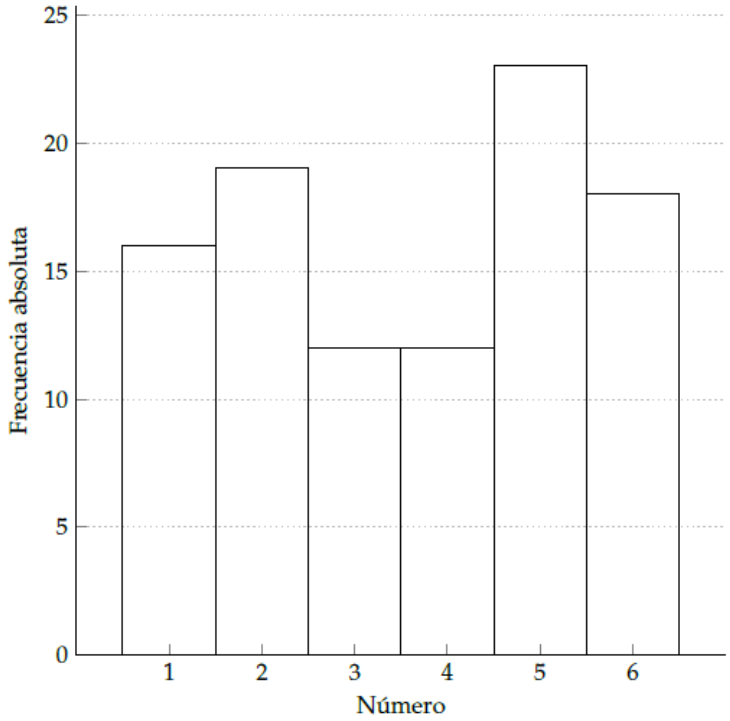



Figura 2: Histograma asociado a la Tabla 1.

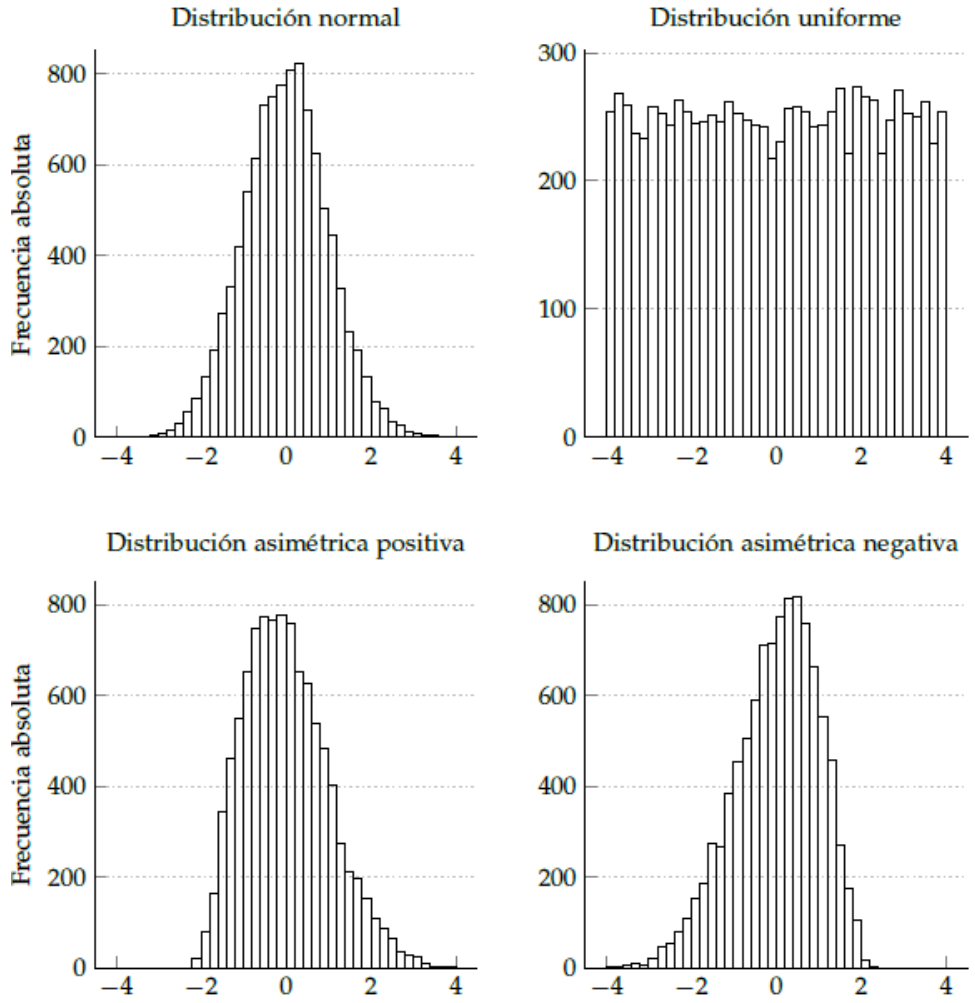



Figura 3: Histogramas de distribuciones comunes, todos ellos obtenidos a partir de 10000 datos. Una distribución normal tiene forma de "campana". En ella, (1) las frecuencias aumentan hasta un máximo y luego disminuyen, y (2) el gráfico tiene simetría, siendo la parte izquierda del histograma aproximadamente una imagen especular de la derecha. Con una distribución uniforme, los distintos valores posibles ocurren con aproximadamente la misma frecuencia, de forma que las alturas de las barras del histograma son aproximadamente iguales. Una distribución de datos está sesgada si no es simétrica y se extiende más hacia un lado que hacia otro. Datos sesgados hacia la derecha (también llamados positivamente asimétricos) tienen una cola derecha más larga, mientras que datos sesgados hacia la izquierda (o negativamente asimétricos) tienen una cola izquierda más larga.

Un HISTOGRAMA DE FRECUENCIAS RELATIVAS tiene la misma forma y eje horizontal que un histograma convencional, pero en el eje vertical aparecen las frecuencias relativas (como proporciones o porcentajes) en vez de las frecuencias absolutas.

Hemos visto que el histograma es una representación gráfica que nos proporciona una mejor comprensión de los datos. Otras representaciones gráficas útiles en Física son los diagramas de dispersión, series temporales, diagramas de barras, diagramas de tarta o polígonos de frecuencias. En el caso de querer representar frecuencias cumulativas, tenemos las llamadas *ojivas*.

*Diagramas de dispersión:* Un DIAGRAMA DE DISPERSIÓN (scatter plot en inglés) es un gráfico (x, y) de datos apareados con posible relación entre ellos[19] (ver Figura 4).

*Series temporales*: Un GRÁFICO DE SERIE TEMPORAL muestra datos cuantitativos que se han tomado en distintos momentos (tiempos), como mensualmente o anualmente (ver Figura 5).

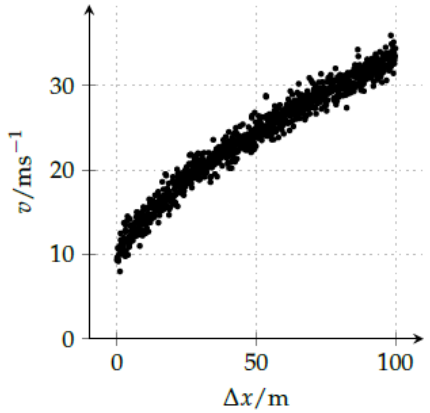



Figura 4: Diagrama de dispersión que muestra la velocidad de un móvil, v, en función del espacio recorrido, Δx.

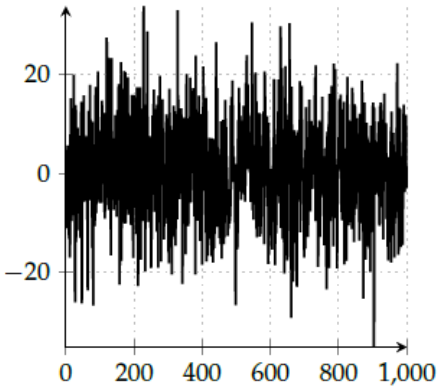



Figura 5: Diagrama de serie temporal.

***Diagramas de barras:*** Un DIAGRAMA DE BARRAS utiliza barras de la misma anchura para mostrar frecuencias de categorías de datos cualitativos. El eje vertical representa frecuencias absolutas o relativas. El eje horizontal identifica las distintas categorías. Las barras pueden estar o no separadas por pequeños huecos.

Un diagrama de barras múltiple tiene dos o más conjuntos de barras y se utiliza para comparar dos o más conjuntos de datos (ver Figura 6).

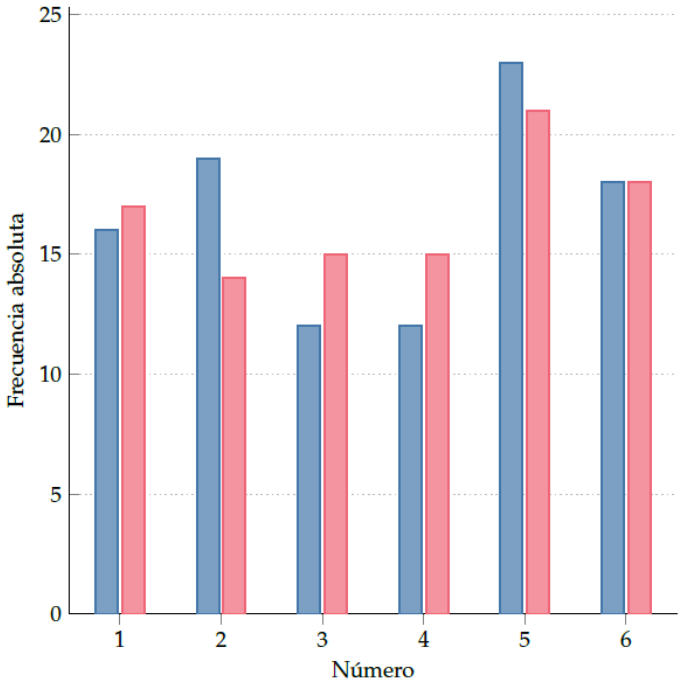



Figura 6: Diagrama de barras que muestra la distribución de frecuencias al lanzar 100 veces dos dados.

[19] Más adelante estudiaremos el llama-do coeficiente de correlación lineal, que nos proporciona información objetiva sobre la existencia o no de correlación entre dos variables.

***Diagramas de tarta:*** Un DIAGRAMA DE TARTA es un gráfico que representa datos categóricos como partes de un círculo, en las que el tamaño de cada parte es proporcional a la frecuencia de cada categoría (ver Figura 7).

En general no se recomienda representar datos con diagramas de tarta, en parte porque no tienen una escala apropiada (para una discusión en profundidad sobre el uso de diagramas de tarta, ver el enlace: http://www.perceptualedge.com/articles/08-21-07.pdf).

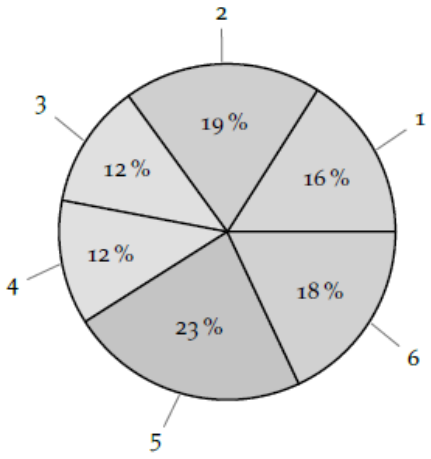



Figura 7: Diagrama de tarta asociado a la Tabla 1.

***Polígonos de frecuencias:*** Un POLÍGONO DE FRECUENCIAS utiliza segmentos de línea conectados a puntos situados directamente encima de las marcas (puntos intermedios) de cada clase. Un polígono de frecuencias es muy similar a un histograma, pero utiliza segmentos de línea en vez de barras (ver Figura 8).

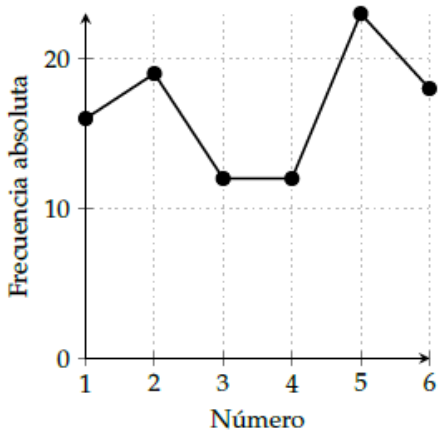



Figura 8: Polígono de frecuencias asociado a la Tabla 1.

***Ojivas*:** La OJIVA es un polígono de frecuencias cumulativas. Las ojivas son útiles para determinar cuántas observaciones se encuentran por debajo de un valor concreto, como se muestra en la Figura 9. Una ojiva utiliza fronteras de clase a lo largo del eje horizontal y frecuencias cumulativas en el eje vertical.

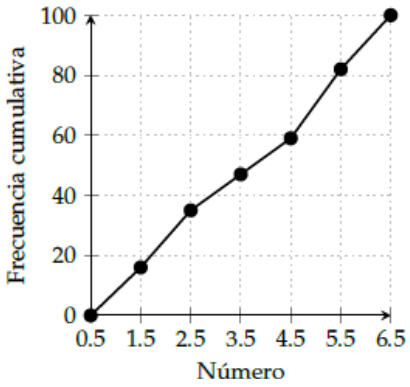



Figura 9: Ojiva asociada a la Tabla 3.

## 2.2 Medidas de centralización

Representar los datos visualmente nos permite obtener impresiones preliminares y comprender muchas características de los datos, en especial cómo están distribuidos. Para llevar a cabo un análisis más formal, necesitamos extraer distintas medidas de resumen. En esta sección nos interesamos por las medidas de centralización, en particular la *media, mediana y moda*.

*Media:* La MEDIA (aritmética) es generalmente la medida más importante, familiar y útil de un conjunto de datos. Se calcula su-mando todos los datos y dividiendo por el número de datos[20], y se suele denotar por $\bar{x}$ en caso de tener una muestra de datos, o por $\mu$ en caso de ser la media de una población completa.

$$\bar{x} = \frac{\sum x}{n}$$

$$\mu = \frac{\sum x}{N}$$

- Las medias de muestras obtenidas de la misma población suelen variar menos que otras medidas centrales.
- La media de un conjunto de datos utiliza todos los valores.
- Una desventaja de la media es que un solo valor atípico (*outlier*) puede hacer variar la media sustancialmente.

[20] Notar que el denominador es n − 1 en vez de n, debido a varias razones. Se puede mostrar que si dividimos por n tendemos a subestimar el valor de la desviación estándar. En realidad, sólo tenemos n − 1 valores independientes (grados de libertad). En caso de tener los datos resumidos en una tabla de frecuencias, la desviación estándar puede calcularse según la expresión:

$$\bar{x} = \frac{\sum(f \cdot x)}{\sum f}$$

, donde *f* y *x* representan la frecuencia y marca (punto intermedio) de cada clase. Si asignamos distintos pesos, *w*, a cada dato podemos computar una media ponderada:

$$\bar{x} = \frac{\sum(w \cdot x)}{\sum w}$$

*Mediana:* La MEDIANA de un conjunto de datos es la medida central que es el valor medio, siempre que los datos estén ordenados. Se suele denotar por $\tilde{x}$ en el caso de una muestra de datos o por $\tilde{\mu}$ en caso de ser la mediana de una población entera. Para calcularla, lo primero que hay que hacer es ordenar los datos (de menor a mayor) para después:

1. Si el número de datos es impar, la mediana es el número ubicado en el medio exacto de la lista ordenada.
2. Si el número de datos es par, la mediana se calcula computando la media entre los dos valores medios de la lista ordenada.

- La mediana no cambia demasiado al incluir unos pocos valores extremos (atípicos).
- La mediana no utiliza todos los valores disponibles.

La mediana divide el conjunto de datos en dos partes del mismo tamaño. Con el fin de obtener medidas de localización más finas, podemos dividir los datos en más partes. Los cuartiles dividen los datos en cuatro partes iguales (conteniendo cada parte el 25 %de los datos aproximadamente), mientras que los percentiles lo hacen en cien partes iguales (conteniendo cada parte aproximadamente el 1 % de los datos)[21].

*Moda:* La MODA de un conjunto de datos es el valor que ocurre con mayor frecuencia. Un conjunto de datos puede tener una moda, más de una moda, o ninguna:

- Cuando dos valores ocurren con la misma frecuencia máxima, cada uno de ellos es una moda y el conjunto de datos es BIMODAL.
- Cuando más de dos valores ocurren con la misma frecuencia máxima, cada uno de ellos es una moda y el conjunto de datos se dice que es MULTIMODAL.
- Si ningún valor se repite, entonces decimos que no hay moda.

## 2.3 Medidas de dispersión

Ya hemos dicho previamente que la variación de los datos es probablemente el tema más importante de la Estadística. En esta sección vamos a hablar de conceptos tan importantes como el rango, la desviación estándar y la varianza. Una medida de centralización como la media nos proporciona información parcial sobre un conjunto de datos o distribución. Distintas muestras o poblaciones pueden tener medias idénticas, pero diferir entre sí de otras maneras importantes.

***Rango:*** El RANGO de un conjunto de datos no es más que la diferencia entre los valores máximo y mínimo. Cuanto mayor es el rango mayor es la variabilidad de un conjunto de datos. Como el rango utiliza los valores máximo y mínimo, es muy sensible a va-lores extremos, por lo que no es tan útil como las otras medidas de dispersión.

[21] A menos que el número de datos sea un múltiplo de cien, hay que tener cuidado a la hora de obtener percentiles.

***Desviación estándar:*** La DESVIACIÓN ESTÁNDAR es la medida de dispersión más utilizada en Estadística. La desviación estándar de una muestra, s, es una medida de cuánto se desvían los datos de la media. Se calcula según la expresión[22]:

$$s = \sqrt{\frac{(x - \bar{x})^2}{n - 1}}$$

- La desviación estándar es una medida de cuánto se desvían los datos de la media.
- El valor de la desviación estándar suele ser positivo (nunca negativo). Sólo es cero cuando todos los datos son iguales. Cuanto mayor es s mayor es la variación.
- El valor de la desviación estándar puede aumentar drásticamente al incluir uno o más valores atípicos (*outliers*).
- La desviación estándar se mide en las mismas unidades que los datos originales.

***REGLA DE ORO DEL RANGO***: Para poder entender e interpretar valores de desviación estándar podemos recurrir a la llamada regla de oro del rango. Ésta se basa en el hecho de que, para muchos conjuntos de datos, la gran mayoría de los datos (sobre el 95 %) se encuentran dentro de dos desviaciones estándar de la media, de forma que la desviación estándar de la muestra, s, se puede estimar a partir del rango según la expresión:

$$s \approx \frac{rango}{4}$$

, donde rango = (valor máximo) − (valor mínimo). En caso de conocer la desviación estándar, la regla de oro del rango nos permite obtener estimaciones de *los valores mínimo y máximo típicos:*

$$valor\ mínimo\ típico\ = \bar{x} - 2s$$

$$valor\ máximo\ típico\ = \bar{x} + 2s$$

***Varianza:*** A partir de la desviación estándar podemos calcular la VARIANZA, definida como el cuadrado de la desviación estándar:

- *varianza muestral:* $s^2$ = cuadrado de la desviación estándar muestral *s*
- *varianza poblacional:* $\sigma^2$ = cuadrado de la desviación estándar poblacional $\sigma$

La varianza se mide en las unidades de los datos originales *al cuadrado*.

Como la desviación típica, el valor de la varianza puede aumentar drásticamente al incluir uno o más valores atípicos (*outliers*).

La varianza siempre es mayor o igual que cero, siendo nula únicamente cuando todos los datos son iguales.

El hecho de que la varianza no tenga las mismas unidades que los datos originales hace que su interpretación sea menos intuitiva que la de la desviación típica.

[22] Notar que el denominador es n − 1 en vez de n, debido a varias razones. Se puede mostrar que si dividimos por n tendemos a subestimar el valor de la desviación estándar. En realidad, sólo tenemos n − 1 valores independientes (grados de libertad). En caso de tener los datos resumidos en una tabla de frecuencias, la desviación estándar puede calcularse según la expresión:

$$s = \frac{\sum[f \cdot (x - \bar{x})^2]}{n - 1}$$

, donde f y x representan de nuevo la frecuencia y marca de cada clase. Para calcular la desviación estándar de una población, denotada por $\sigma$, utilizamos la expresión:

$$\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}}$$

***Desviaciones de la media:*** Cuando hablamos de variabilidad, lo primero que se nos ocurre son DESVIACIONES DE LA MEDIA, $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$. Podríamos *sumar* todas las desviaciones para obtener una medida única de la variabilidad. Por desgracia esa suma siempre es cero:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = \sum_{i=1}^{n} x_i - n\bar{x} = \sum_{i=1}^{n} x_i - n\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = 0$$

Podemos tomar el valor absoluto de las desviaciones, para evitar que los términos positivos y negativos se cancelen. De esa forma obtenemos la llamada desviación absoluta media[23]:

$$\text{desviación absoluta media} = \frac{\sum|x - \bar{x}|}{n}$$

***Regla Empírica 68 – 95 – 99.7*:** La conocida como REGLA EMPÍRICA establece que, para conjuntos de datos con una distribución aproximadamente en forma de campana (normal), se aplican las siguientes propiedades:

- Aproximadamente el 68% de todos los valores caen dentro de 1 desviación estándar de la media.
- Aproximadamente el 95% de todos los valores caen dentro de 2 desviaciones estándar de la media.
- Aproximadamente el 99.7% de todos los valores caen dentro de 3 desviaciones estándar de la media.

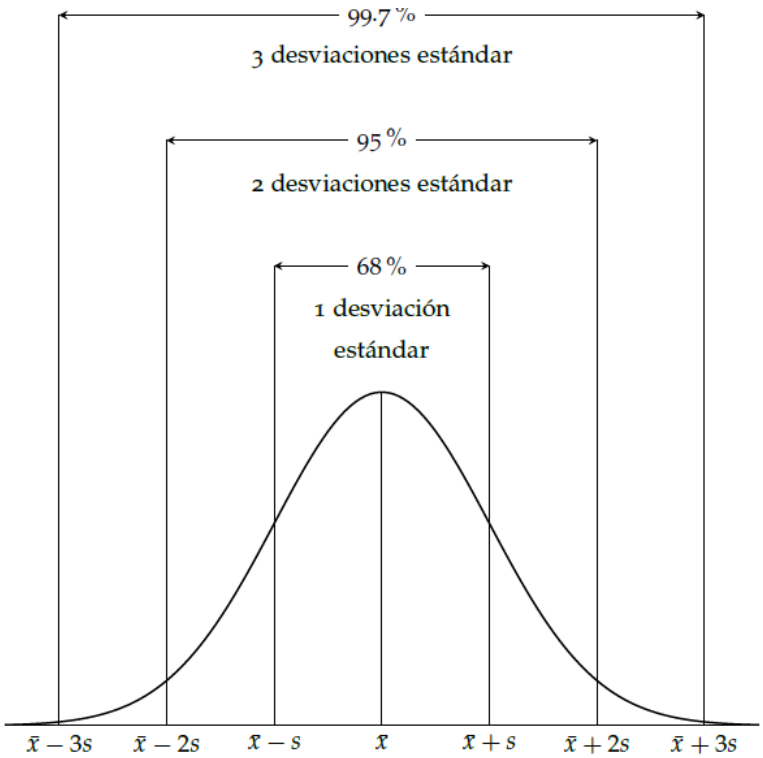



Figura 11: La regla empírica.

***Coeficiente de variación:*** En caso de querer comparar la variación de dos conjuntos de datos diferentes[24] podemos utilizar el COEFICIENTE DE VARIACIÓN (CV), que describe la desviación estándar relativa a la media:

$$CV = \frac{s}{\bar{x}} \cdot 100\% \ (muestra)$$

$$CV = \frac{\sigma}{\mu} \cdot 100\% \ (población)$$

[23] En la práctica no se utiliza la desviación absoluta media debido a complicaciones algebraicas a la hora de aplicar los métodos de la estadística inferencial (la función valor absoluto no es derivable).

[24] Bien porque sus medias sean sustancialmente diferentes, bien porque utilicen diferentes escalas o unidades de medida.

EJERCICIO 1: Medidas centralización en variable continua

Se ha hecho un estudio del tiempo que tarda en aparecer una burbuja en un líquido que contiene nanopartículas de oro, al ser iluminado con un láser de alta potencia a 88 muestras distintas, obteniéndose la tabla de datos adjunta.

| Intervalos (s) | $f_i$ |
|---|---|
| 38-43 | 7 |
| 44-49 | 8 |
| 50-55 | 15 |
| 56-61 | 25 |
| 62-67 | 18 |
| 68-73 | 9 |
| 74-79 | 6 |

| Intervalos (s) | $f_i$ |
|---|---|
| [38-44) | 7 |
| [44-50) | 8 |
| [50-56) | 15 |
| [56-62) | 25 |
| [62-68) | 18 |
| [68-74) | 9 |
| [74-80) | 6 |

a) Halla la media aritmética, moda, mediana y el cuartil $Q_1$.
b) Halla el rango y la desviación típica.

**Construimos una tabla de frecuencias** para calcular la media estadística, moda, mediana y desviación típica.

| Intervalos (s) | $x_i$ | $f_i$ | $F_i$ | $xi \cdot fi$ | $xi^2 \cdot fi$ |
|---|---|---|---|---|---|
| 38-43 | 41 | 7 | 7 | 287 | 11767 |
| 44-49 | 47 | 8 | 15 | 376 | 17672 |
| 50-55 | 53 | 15 | 30 | 795 | 42135 |
| 56-61 | 59 | 25 | 55 | 1475 | 87025 |
| 62-67 | 65 | 18 | 73 | 1170 | 76050 |
| 68-73 | 71 | 9 | 82 | 639 | 45369 |
| 74-79 | 77 | 6 | 88 | 462 | 35574 |
| $\Sigma$ | | 88 | | 5204 | 315592 |

a) **Media aritmética**: $\bar{x} = \frac{\sum x_i \cdot f_i}{N} = \frac{5204}{88} = 59.14$

**Moda Mo**: buscamos la clase modal, la de mayor fi y aplicamos la fórmula: $M_o = L_i + c \cdot \frac{D_1}{D_1 + D_2}$

donde:

$L_i$ = Límite inferior de la clase modal,

*C* = Amplitud de los intervalos,

$D_1$ = Diferencia entre la frecuencia absoluta de la clase modal y la de la clase anterior.

$D_2$ = Diferencia entre la frecuencia absoluta de la clase modal y la de la clase siguiente.

**La clase modal se corresponde con la mayor frecuencia $f_i$ =25, es decir 56-61**

Aplicamos la fórmula de $M_o$ y queda: $M_o = 56 + 6 \cdot \frac{25-15}{(25-15)+(25-18)} = 59.52$

**Mediana M**

Intervalo de la clase mediana $\frac{N}{2} = \frac{88}{2} = 44$ buscamos $F_i > 44$, encontramos que se corresponde con el intervalo 56-61. Ahora aplicamos la fórmula: $M = L_i + c \cdot \frac{\frac{N}{2} - F_{i-1}}{f_i}$ donde:

$L_i$ = Límite inferior de la clase mediana,

*C* = Amplitud de los intervalos,

*N* = Número total de datos.

$F_{i-1}$ = Frecuencia absoluta acumulada de la clase anterior a la clase mediana.

$f_i$ = Frecuencia absoluta de la clase mediana. $M = 56 + 6 \cdot \frac{44-30}{25} = 59.36$

**Cuartil $Q_1$**

$Q_1 = \frac{N}{4} = 22$ buscamos $F_i > 22$, encontramos el valor 30 que se corresponde con el intervalo 50-55. Ahora aplicamos la formula: $Q_1 = L_i + c \cdot \frac{\frac{N}{4} - F_{i-1}}{f_i} = 50 + 6 \cdot \frac{22-15}{15} = 52.8$

**$Q_1$= 52.8**

**b) Rango, Varianza y desviación típica**

**Rango** : 80 – 38 =42

**Varianza:** $\sigma^2 = \frac{\sum(x^2) \cdot f_i}{N} - x^2 = \frac{315592}{88} - 59.14^2 = 88.73$

**Desviación típica:** $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x^2) \cdot f_i}{N} - x^2} = 9.42$

# 3. PROBABILIDAD

MUCHOS RAZONAMIENTOS ESTADÍSTICOS dependen de la teoría de probabilidades, a través de modelos de azar. El término probabilidad se refiere al estudio de la aleatoriedad y la incertidumbre. En cualquier situación en la que puede ocurrir uno de una serie de resultados posibles, la disciplina de la probabilidad proporciona métodos para cuantificar las posibilidades, o probabilidad, asociadas con los diversos resultados.

### 3.1 Algunas definiciones

En el estudio de un cierto fenómeno aleatorio, llamaremos:

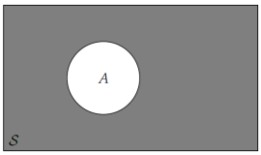
- *Suceso elemental*: Cada uno de los posibles resultados de un experimento aleatorio.
- *Espacio muestral*: Conjunto de todos los posibles resultados (contiene todos los sucesos elementales). Lo denotaremos por *S*.
- *Suceso:* Subconjunto del espacio muestral.
  - **Sucesos independientes:** Dos sucesos, A y B, son independientes cuando la probabilidad de que suceda A no se ve afectada porque haya sucedido o no B.
  - **Sucesos dependientes:** Dos sucesos, A y B, son dependientes cuando la probabilidad de que suceda A se ve afectada porque haya sucedido o no B.

### 3.2 Conjuntos

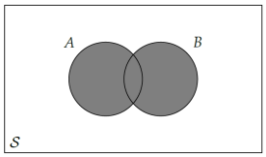
Un suceso es un conjunto, por lo que podemos utilizar relaciones y resultados de teoría de conjuntos básica para estudiar sucesos.

*Suceso contrario (o complementario): de A* Se denota por $\bar{A}$ y es el conjunto de todos los resultados no contenidos en *A*.
*Suceso unión*: La unión de dos sucesos *A* y *B* se denota por $A \cup B$ y se lee como "*A* o *B*". Consiste en el suceso que contiene todos los resultados que están bien en *A*, bien en *B* o en ambos (es decir, todos los casos que están en al menos uno de los dos sucesos).
*Suceso intersección:* Se denota por $A \cap B$ y se lee "*A* y *B*". Es el suceso que se verifica si lo hacen a la vez *A* y *B*.

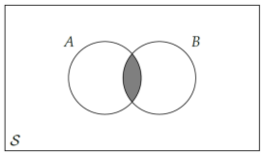
- El suceso imposible (o nulo) es aquel que no contiene ningún resultado en absoluto, y se denota por ∅. Cuando $A \cap B = \emptyset$, se dice que A y B son mutuamente excluyentes o sucesos incompatibles.



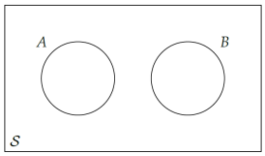
La región sombreada es $\bar{A}$.



La región sombreada es $A \cup B$.



La región sombreada es $A \cap B$.



Sucesos incompatibles.

### 3.3 Definición frecuentista de probabilidad

Imagina un experimento que pueda ser repetido de forma idéntica e independiente tantas veces como queramos[25]. Si contamos el número de veces que el suceso *A* ocurre, entonces la probabilidad *P(A)* se puede aproximar por:

$$P(A) = \frac{número\ de\ veces\ que\ ocurre\ A}{número\ de\ veces\ que\ se\ repite\ el\ experimento}$$

[25] Ejemplos de tales experimentos pueden ser lanzar una moneda o un dado.

A la hora de utilizar esta definición, es importante tener en cuenta que lo que obtenemos es una aproximación de la probabilidad. A medida que el número total de observaciones aumenta, las correspondientes aproximaciones tienden a acercarse al valor exacto de la probabilidad[26].

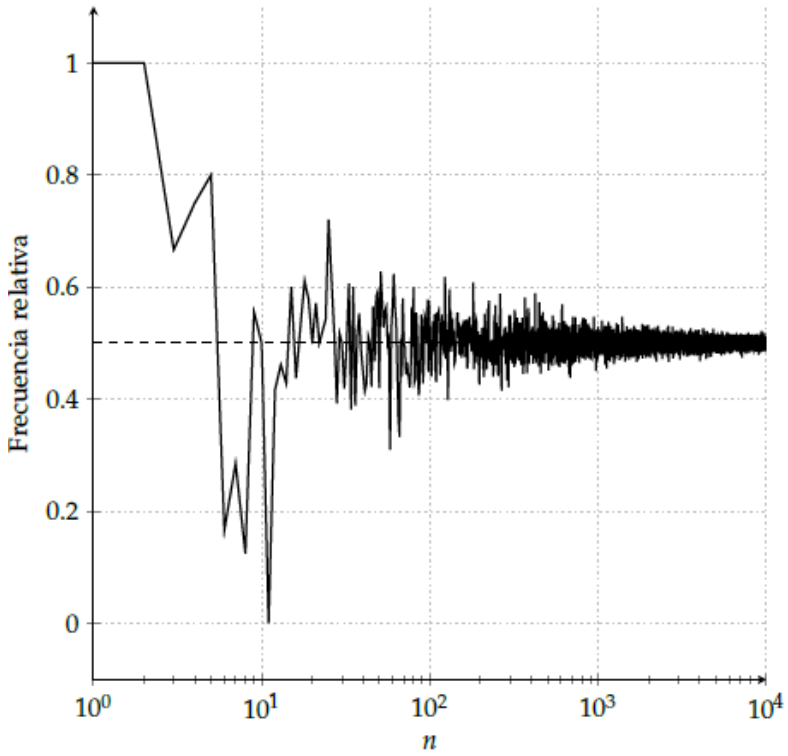



Figura 17: Comportamiento de la frecuencia relativa del suceso "cara" al lanzar una moneda n veces, con una fluctuación inicial y una estabilización a largo plazo.

### 3.4 Definición axiomática de probabilidad

El objetivo de la probabilidad es asignar a cada suceso *A* un número, *P(A)*, que llamamos probabilidad del suceso *A*, que de una medida precisa de la posibilidad de que ocurra *A*. Los axiomas de la probabilidad son:

*Axioma 1.* Para cualquier suceso *A, P(A) ≥ 0*.

*Axioma 2. P(S) = 1*.

*Axioma 3.* Si $A_1$, $A_2$, $A_3$,. . . es una colección infinita de sucesos incompatibles, entonces

$$P(A1 \cup A2 \cup A3 \cup \ldots) = \sum_{i=1}^{\infty} P(A_i)$$

### 3.5 Propiedades de la probabilidad

- *P(∅) = 0*, lo que implica que la propiedad contenida en el *Axioma 3* es válida para una colección *finita* de sucesos incompatibles.
- Para cualquier suceso *A*: $\boldsymbol{P(A) + P(\overline{A}) = 1}$, por lo que $\boldsymbol{P(\overline{A}) = 1 - P(A)}$.
- Para cualquier suceso *A*: ***P(A) ≤ 1***.
- Para cualquiera dos eventos *A* y *B*: $\boldsymbol{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$

[26] Este resultado se conoce como LA LEY DE LOS GRANDES NÚMEROS: "*A medida que un experimento se repite una y otra vez, la probabilidad de frecuencia relativa de un suceso tiende a aproximarse a la probabilidad real.*"

## 3.6 Determinación sistemática de probabilidades

La probabilidad de un suceso *A*, compuesto por una serie finita de sucesos elementales $E_1$, $E_2$, $E_3$,. . ., se puede calcular a partir de la probabilidad de cada uno de los sucesos elementales $E_i$ , con el requisito $\sum P(E_i) = 1$ :

$$P(A) = \sum_{todos\ los\ E_i\ en\ A} P(E_i)$$

*Sucesos igualmente probables:* En muchos experimentos en los que tenemos un número *N* de posibles resultados, es razonable asignar una misma probabilidad a todos los *N* sucesos elementales. En este caso la probabilidad de cada resultado viene dada por[27] *1/N*. Consideremos ahora un suceso *A*, siendo *N(A)* el número de resultados contenidos en *A* (número de casos favorables). Entonces

$$P(A) = \frac{N(A)}{N}$$

Es decir, cuando los posibles resultados son igualmente probables, el cálculo de probabilidades se reduce a contar: determinar tanto el número de resultados N(A) en A como el número total de casos posibles *N* en *S*, y calcular su cociente.

## 3.7 Técnicas para contar

Existen muchos experimentos en los que construir una lista de todos los posibles resultados (para poder contarlos) es prohibitivo, por lo que se hace necesario el uso de algunas técnicas para contar.

*Regla del producto para pares ordenados:* La primera técnica para contar se aplica a cualquier situación en la que un conjunto (suceso) consiste en pares ordenados de objetos y queremos contar el número de dichos pares. Un par ordenado implica que, si $O_1$ y $O_2$ son dos objetos, entonces el par ($O_1$ , $O_2$) es distinto al par ($O_2$, $O_1$).

> *Si el primer elemento u objeto de un par ordenado se puede seleccionar de $n_1$ maneras, y para cada una de esas $n_1$ maneras el segundo elemento se puede seleccionar de $n_2$ maneras, entonces el número de pares es $n_1 n_2$.*

*Regla del producto para k-tuplas*

Suponer un conjunto que consiste en una colección ordenada de k objetos (k-tuplas), y que hay $n_1$ posibles elecciones para el primer elemento; para cada elección del primer elemento, hay $n_2$ posibles elecciones del segundo elemento;. . . ; para cada posible elección de los primeros $k - 1$ elementos, hay $n_k$ posibles elecciones del elemento k-ésimo. Entonces hay $n_1 n_2 \ldots n_k$ k-tuplas posibles.

[27] Sea $p = P(E_i)$ para todo *i*: $1 = \sum_{i=1}^{N} P(E_i) = \sum_{i=1}^{N} p = p \cdot N$ , de manera que $p = 1/N$

*Diagrama de árbol:* Una forma de representar todas las posibilidades en problemas de conteo y probabilidad es mediante DIAGRAMAS DE ÁRBOL. Comenzando en un punto en el extremo izquierdo del diagrama, dibujamos una línea recta hacia la derecha para cada posible primer elemento de un par. Ahora para cada rama de primera generación construimos otro segmento para cada elección posible de un segundo elemento del par.

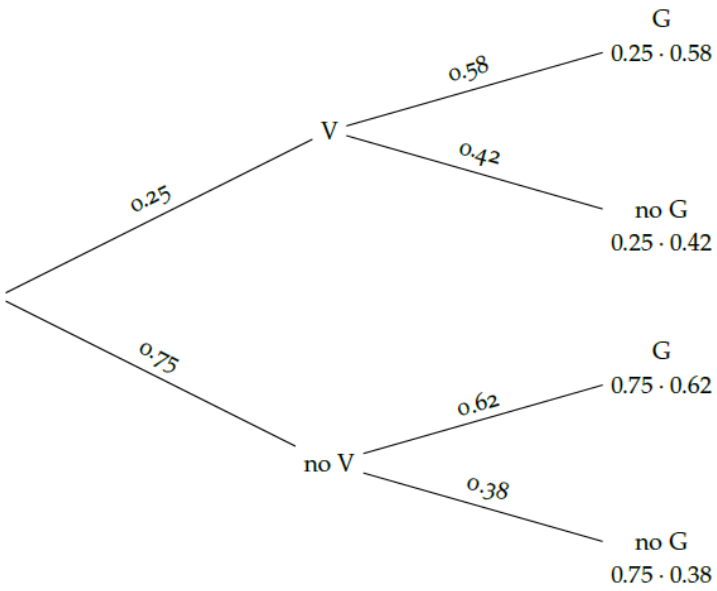



Figura 18: Diagrama de árbol correspondiente al siguiente enunciado: "*En una universidad en que la población estudiantil es muy numerosa, y basándose en la experiencia de años anteriores, se sabe que el 25% de los estudiantes son vegetarianos (V); entre los vegetarianos, el 58% llevan gafas; entre los no vegetarianos, el 62% llevan gafas.*"

*Regla del factorial:* Se utiliza para encontrar el número total de maneras en las que n elementos distintos[28] se pueden reordenar (teniendo en cuenta el orden), denotándose por $n! = n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot 1$. Mencionar también que, por definición, $0! = 1$.

*Permutaciones y Combinaciones:* Considerar un grupo (conjunto) de n objetos distintos. ¿Cuántas formas hay de elegir un subconjunto de k elementos a partir del grupo? En función de que el orden de los elementos juegue o no papel, se distinguen:

- *Permutaciones:* Una permutación es un subconjunto ordenado. El número de permutaciones de tamaño *k* que se pueden formar a partir de n objetos se denota por $P_{k,n}$ y se calcula según la expresión:

$$P_{k,n} = \frac{n!}{(n-k)!}$$

- *Combinaciones:* Una combinación es un subconjunto no ordenado. El número de combinaciones posibles se puede denotar por $C_{k,n}$ pero es muy común la notación $\binom{n}{k}$, que se lee "*n* en *k*", y se calcular según la expresión:

$$C_{k,n} = \binom{n}{k} = \frac{P_{k,n}}{k!} = \frac{n!}{k!\,(n-k)!}$$

[28] Entendiendo por distintos que existe alguna característica que diferencia un objeto concreto de cualquier otro.

## 3.8 Probabilidad condicionada

El número que asignamos como probabilidad a un suceso varía según la información que tenemos sobre él. En el momento en que tenemos más información, el número de casos posibles es más reducido y la probabilidad puede cambiar.

Dados dos sucesos *A* y *B*, llamamos suceso *A* CONDICIONADO AL SUCESO *B* (denotado por $A|B$) a aquel suceso que se verifica cuando se verifica *A* SABIENDO QUE ya se ha verificado B.

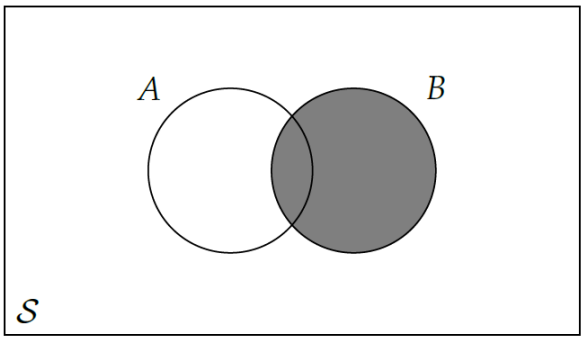



Figura 19: Motivación de la definición de probabilidad condicionada.

La definición de probabilidad condicionada da lugar al resultado conocido como LA REGLA DE MULTIPLICACIÓN:

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

La PROBABILIDAD CONDICIONADA de *A* sabiendo que *B* ya ha ocurrido se define como:

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

A menudo la probabilidad de que ocurra A no depende de otro suceso B, de forma que $P(A|B) = P(A)$. En ese caso se dice que *A* y *B* son sucesos independientes[29].

[29] En otras palabras, *A* y *B* son independientes si y solo si $P(A \cap B) = P(A) \cdot P(B)$ , lo que se conoce como la regla de multiplicación para $P(A \cap B)$.

## Referencias

1. D. C. Baird. Experimentation: An Introduction to Measurement Theory and Experiment Design. Prentice-Hall, Inc., 3rd edition, 1995.

2. George Casella and Roger L. Berger. Statistical Inference. Duxbury, 2nd edition, 2002.

3. Jay L. Devore. Probability and Statistics for Engineering and the Sciences. CENGAGE Learning, 9th edition, 2014.

4. David Freedman, Robert Pisani, and Roger Purves. Statistics. W. W. Norton & Company, 4th edition, 2007.

5. Anders Hald. A History of Probability and Statistics and Their Applications before 1750. John Wiley & Sons, 2003.

6. Darrel Huff. How To Lie With Statistics. W. W. Norton & Company, 3rd edition, 1993.

7. John R. Taylor. An Introduction to Error Analysis. University Science Books, 2nd edition, 1997.

8. Mario F. Triola. Elementary Statistics. Pearson, 12th edition, 2012.