NOTAS BÁSICAS DE ESTADÍSTICA APLICADA A LA INVESTIGACIÓN SOCIAL

Antonio Caparrós Ruiz
Departamento de Economía Aplicada (Estadística y Econometría)
Universidad de Málaga

Julio del 2025

Prólogo

El presente material docente se basa principalmente en la experiencia adquirida y en los conocimientos elaborados y transmitidos durante la impartición de contenidos relacionados con la Estadística Aplicada a la Investigación Social. El principal objetivo de este documento es servir de apoyo bibliográfico en la iniciación del alumnado en el conocimiento elemental y en el uso de las herramientas básicas de Estadística Descriptiva, que sean de utilidad para el tratamiento de los datos y para la reducción de la información estadística, así como para la interpretación de documentos que contenga conceptos estadísticos.

Los tópicos tratados en esta obra se desarrollan en 8 temas, y abordan cuestiones que son comunes a programas de Estadística Descriptiva incluidos en los Planes de Estudios de titulaciones relacionadas con las Ciencias Sociales. Concretamente, los diversos temas se distribuyen en dos Bloques. Por un lado, el Bloque I está compuesto por 4 temas, que se centran en el análisis de estadísticas univariantes. En particular, en el tema 1 se realiza una presentación de nociones estadísticas básicas en un contexto donde se observa sólo una variable, con el fin de servir de fundamento para la elaboración de conceptos más complejos. En el tema 2 se proporciona la base teórica para obtener medidas de promedio y comprender la utilidad de las mismas como instrumentos de reducción estadística. Posteriormente, en el tema 3 se presentan las medidas de dispersión y las medidas de forma, que son herramientas de gran utilidad para comprender la distribución de los datos y complementar las conclusiones que se puedan obtener a través de las medidas de posición. Por último, el tema 4 está dedicado al estudio de la desigualdad desde un punto de vista estadístico.

En cuanto al Bloque II, también contiene 4 temas. El tema 5 aborda la tabulación y medidas de asociación lineal entre dos variables; es decir, es un enfoque bivariante que permite avanzar en el conocimiento de las interrelaciones de las variables estadísticas. Con este punto de partida, en el tema 6 se introduce el análisis de regresión que permite establecer una relación de causalidad de una variable dependiente con otras explicativas. El análisis se realiza a un nivel básico, por lo que se considera sólo la existencia de una variable explicativa. No obstante, con este sencillo planteamiento es posible cuantificar la relación entre dos variables y realizar predicciones, aunque hay que ser cautelosos y conscientes de la limitación del mismo, ya que no se abordan cuestiones relacionadas con la inferencia estadística. En el tema 6 también se presentan los componentes de las series temporales, y cómo ajustar una tendencia lineal para captar la evolución a largo plazo de una variable. En cuanto al tema 7, se introducen los números índices que son herramientas estadísticas diseñadas para observar la evolución de las variables. En este escenario, se explica la importancia del Índice de Precios al Consumo, y su utilidad para evaluar la inflación, observar la evolución del poder adquisitivo y deflactar de series monetarias. Finalmente, en el tema 8 se muestran de forma simplificada algunas fuentes estadísticas que puedan ser de utilidad para los análisis desarrollados en la Investigación Social.

El enfoque didáctico y metodológico pretende ser autocontenido. De tal forma, que el alumnado con escasos conocimientos estadísticos pueda realizar un aprendizaje adecuado de los mismos y avanzar en sus capacidades. El planteamiento es transmitir al alumnado que al igual que no existen barreras iniciales para empezar a aprender un idioma, tampoco existen obstáculos para iniciarse en el lenguaje de la Estadística. En este sentido, se prioriza la interpretación de los resultados en el aprendizaje y en el desarrollo de los contenidos, intentando minimizar en la medida de lo posible las demostraciones matemáticas con mayor dificultad. Además, como complemento de los temas desarrollados, se introducen anexos que exponen cómo se obtendrían algunos resultados mediante la utilización de la hoja de cálculo Excel.

Para finalizar, cabe señalar que el fin primordial de este documento es poner a disposición del estudiantado una obra accesible, con una exposición sencilla, que facilite el estudio de los conceptos estadísticos básicos, apoye las explicaciones realizadas en las clases y sea un punto de partida para un posterior aprendizaje de herramientas metodológicas más avanzadas.

TEMA 1. CONCEPTOS BÁSICOS

1.1 INTRODUCCIÓN

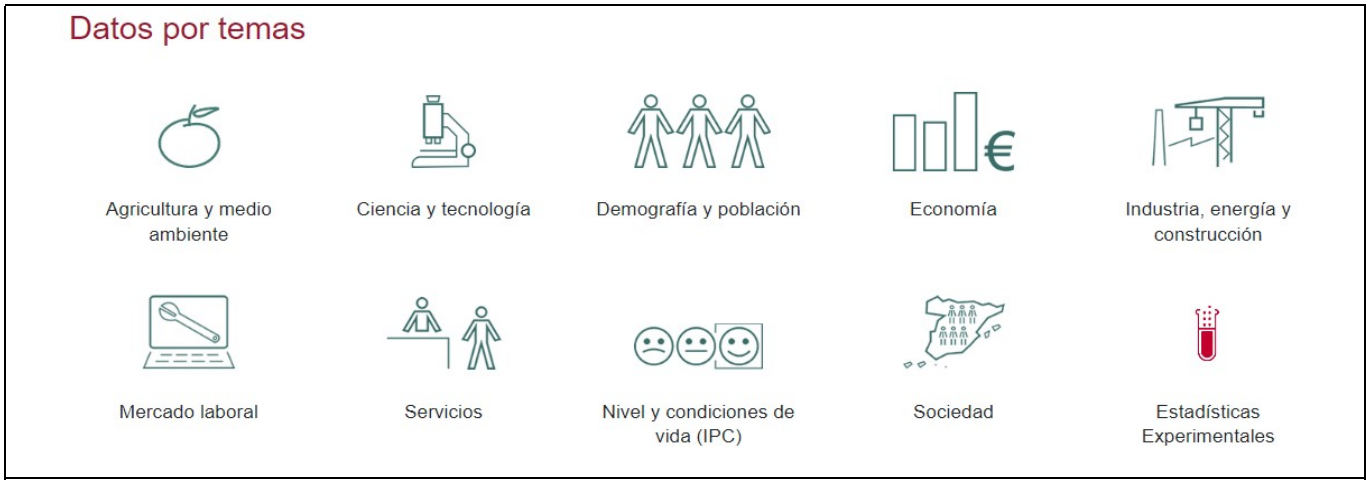
Los contenidos de este documento giran en torno a los postulados de la Ciencia Estadística y, en concreto, a los fundamentos de la Estadística Aplicada a la Investigación Social. Previamente, es necesario indicar cuáles son las acepciones de la palabra estadística. En primer lugar, se puede definir a la estadística (en minúscula) como una colección de datos numéricos presentados de manera ordenada y sistemática; y, en segundo lugar, se puede mencionar a la Estadística (en mayúscula) como la Ciencia que estudia la realidad utilizando grandes conjuntos de datos, con objeto de describir su comportamiento, detectar regularidades y predecir su evolución futura. Concretamente, los temas del presente material docente se ubican dentro de la rama de la Estadística conocida como Estadística Descriptiva.

Dentro del ámbito de la Estadística Descriptiva, este documento desarrolla algunas herramientas estadísticas que pueden utilizarse en la Investigación Social. El uso de estos instrumentos va a permitir reducir, analizar y tratar la información procedente de Estadísticas Sociales, lo cual es un punto de partida para la elaboración de informes que permitan interpretar la realidad social con objeto de planificar, detectar problemas y anticipar posibles soluciones a los mismos.

Existen diversos organismos que elaboran estadísticas de utilidad en el campo de la Investigación Social, algunos ejemplos son el Instituto Nacional de Estadística (INE) o el Ministerio de Trabajo y Economía Social (MITES).

En relación al INE, su información estadística puede estructurarse por temas como aparece en la figura 1.1.

Figura 1.1. Información estadística del INE por temas



Fuente: INE (2025a).

Por ejemplo, si se accede a la sección de "Sociedad" es posible encontrar estadísticas relacionadas con "Educación y Cultura", "Salud", "Seguridad y Justicia", "Análisis Sociales" y "Procesos electorales". Estas fuentes estadísticas permiten diseñar diversos tipos de información. Por ejemplo, si nos centramos en "Salud" aparecen estadísticas que se elaboran de forma periódica como se muestra en la tabla 1.1.

Tabla 1.1. Estadísticas de Salud elaboradas de forma periódica

Operaciones estadísticas que el INE elabora de forma periódica	Últimos datos
Estimación del número de defunciones semanales	Semana 18 2025
El empleo de las personas con discapacidad	Año 2023
Encuesta de morbilidad hospitalaria	Año 2023
Encuesta de salud de España	Año 2023
Estadística de defunciones según la causa de muerte	Provisionales 1S/2024 y año 2023
Estadística de profesionales sanitarios colegiados	Año 2024
Estadística del salario de las personas con discapacidad	Año 2022

Fuente: INE (2025b).

Más concretamente, si seleccionamos "Estadística del salario de las personas con discapacidad", los resultados más agregados se muestran en la tabla 1.2, donde se constata que el salario anual de las personas con diversidad funcional en el año 2022 representa el 81,3% del correspondiente a las personas sin diversidad funcional. Además, se observan significativas diferencias entre hombres y mujeres, siendo el salario anual de los hombres con diversidad funcional superior al de las mujeres en un 9,76%.

Tabla 1.2. Salario de las personas con diversidad funcional. Año 2022

	Valor	Ratio salarial
Total	22.040,7	81,3
Hombres	22.938,4	77,5
Mujeres	20.898,9	85,4

Valor: euros. Ratio salarial: porcentaje respecto al salario de las personas sin discapacidad

Fuente: INE (2025c).

Por otro lado, es posible obtener un catálogo de resultados correspondientes al periodo 2010-2022. Un ejemplo de estos aparece desarrollado en la tabla 1.3.

Tabla 1.3. Salario anual de las personas sin discapacidad y de las personas con discapacidad: 2010-2022

Fuente: INE (2025d).

Otras estadísticas de interés dentro de la sección de Salud son las incluidas dentro del apartado dedicado a "El empleo de las personas con discapacidad". En la tabla 1.4 se exponen algunos resultados generales para el año 2023, y en la figura 1.2, la evolución de la tasa de paro de este colectivo junto a la del total de trabajadores.

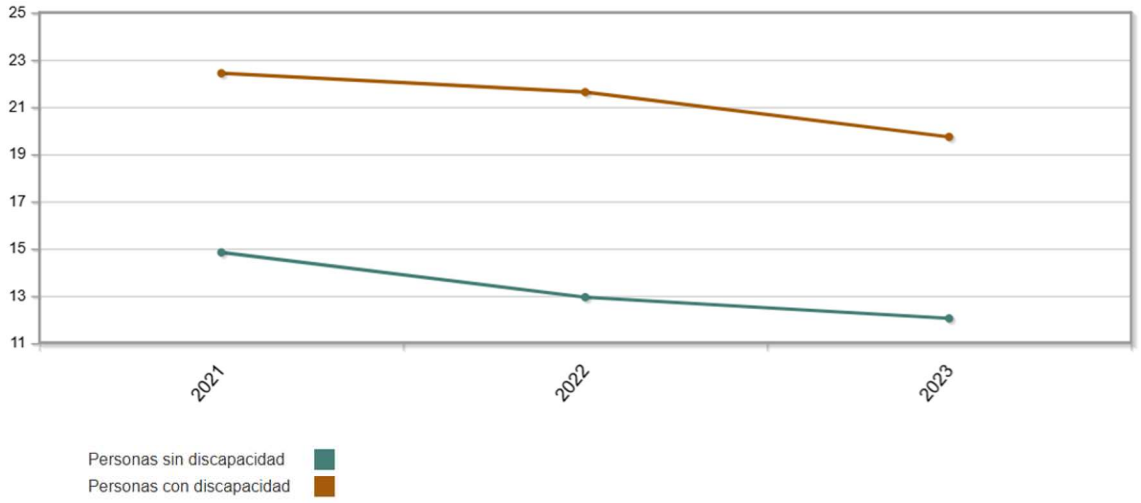
Tabla 1.4. Empleo de las personas con diversidad funcional. Año 2023

	Valor	Variación
Tasa de actividad	35,5	0,2
Tasa de empleo	28,5	0,8
Tasa de paro	19,7	-1,9

Valor en %. Variación: diferencia respecto a la tasa del mismo período del año anterior

Fuente: INE (2025e).

Figura 1.2. Tasa de paro (%) de las personas con diversidad funcional y sin diversidad funcional: 2021-2023



Fuente: INE (2025f).

Centrándonos en el portal del MITES, es posible encontrar la siguiente colección de fuentes estadísticas ofrecida en la tabla 1.5.

Tabla 1.5. Algunas estadísticas del MITES

Información estadística

- Mercado de trabajo
- Inmigración y emigración
- Políticas del mercado de trabajo. Formación profesional y medidas de apoyo al empleo
- Condiciones de trabajo y relaciones laborales
- Prestaciones de Seguridad Social y otra protección social
- Otra información

Fuente: MITES (2025a).

Por ejemplo, la sección dedicada a "Inmigración y emigración" contiene información relativa a los apartados que aparecen en la tabla 1.6.

Tabla 1.6. Secciones estadísticas del MITES dedicadas a Inmigración y Emigración

- Autorizaciones de trabajo a extranjeros
- Trabajadores extranjeros afiliados a la Seguridad Social en alta laboral
- Contratos registrados de trabajadores extranjeros
- Demandantes de empleo extranjeros
- Beneficiarios de prestaciones por desempleo extranjeros
- Españoles residentes en el extranjero retornados
- Extranjeros con certificado de registro o tarjeta de residencia en vigor
- Extranjeros con autorización de estancia por estudios en vigor
- Concesiones de nacionalidad española por residencia
- Visados expedidos en oficinas consulares
- Flujo de autorizaciones de residencia concedidas a extranjeros

Fuente: MITES (2025b).

Los datos propuestos en esta sección introductoria valgan como ejemplo de la existencia de un gran número de fuentes estadísticas en España relacionadas con temas relativos a la Investigación Social, que pueden ser la base para la realización de estudios que permitan identificar problemas y arrojar conclusiones de interés para la elaboración de políticas sociales. Esta presentación inicial se complementará con la exposición en el tema 8 de otras fuentes estadísticas relevantes.

En los siguientes epígrafes se muestran algunos conceptos básicos necesarios para iniciar el análisis estadístico de datos relativos a la Investigación Social.

1.2 POBLACIÓN Y MUESTRA

La Población es el colectivo de individuos o elementos que poseen ciertas características comunes y que son objeto de observación y estudio estadístico. Algunos ejemplos de Población podrían ser los siguientes:

a) Personas en riesgo de exclusión laboral (por ejemplo, jóvenes sin formación).

b) Ancianos en residencias.

c) Inmigrantes en centros de acogida.

d) Hogares desfavorecidos.

Un concepto asociado al del Población es el de Elemento. Un Elemento es cada uno de los entes o fenómenos que integran la población. En el caso de los ejemplos anteriores, el Elemento sería una persona en los tres primeros, y un hogar en el último.

La observación de los elementos de la población se puede realizar desde diversos puntos de vista. Si se observan todos los elementos de la población, se dice que la observación es exhaustiva. Un ejemplo de observación exhaustiva sería el Censo de Población. En la tabla 1.7 se presentan algunos resultados generales relativos al Censo de Población y Viviendas del año 2021.

Tabla 1.7. Resultados relativos al Censo de Población (2021)

	Número	Personas residentes
TOTAL	18.553.289	47.400.798
Hogares familiares	18.539.223	47.066.972
En alojamientos	2.607	7.199
En viviendas familiares convencionales	18.536.616	47.059.773
Según número de miembros		
- Una persona sola	5.001.166	5.001.166
- Dos personas	5.203.749	10.407.498
- Tres personas	3.837.982	11.513.946
- Cuatro personas	3.123.216	12.492.864
- Cinco o más personas	1.373.110	7.651.498
Establecimientos colectivos	14.066	333.826

Fuente: INE (2023).

Si sólo se observa una parte de la población nos encontramos con una Observación Parcial. La Observación Parcial puede ser de dos tipos:

a) Subpoblación:

Subconjunto de elementos de la población que presentan una cierta característica que los diferencia de los demás. Bajo un contexto nacional, un ejemplo de subpoblación es el "Directorio de Empresas y Establecimiento con Actividad Económica en Andalucía" donde la característica es que la empresa o el establecimiento pertenece a Andalucía.

b) Muestra:

Subconjunto de elementos de la población, que no presentan una característica que los diferencian del resto, sino que pretenden representar a toda la población. Un ejemplo de muestra son los hogares que componen la Encuesta de Población Activa.

1.3 VARIABLES Y ATRIBUTOS

La observación de la información estadística nos permite obtener datos sobre las propiedades, rasgos o cualidades de los elementos de una población. En definitiva, nos permite conocer los caracteres de los elementos de la población. Existen dos tipos de caracteres: variables y atributos.

Las variables son caracteres cuantitativos. La medición de una variable da lugar a valores numéricos y las variables pueden clasificarse como variables discretas y variables continuas.

Una variable es discreta si el número de valores que puede tomar es finito o infinito numerable. Por ejemplo, el número de contratos firmados por un trabajador durante su vida laboral.

Una variable es continua si puede tomar infinitos valores dentro de un intervalo: por ejemplo, el salario mensual puede estar comprendido entre 1400€ y 4000€ mensuales y, en principio, entre ambos extremos pueden existir infinitos valores; no obstante, los acuerdos sobre unidades monetarias transforman a dicha variable en discreta, ya que la fracción más pequeña en el caso del € es el céntimo.

Los atributos son caracteres cualitativos. Las diferentes categorías que presenta un atributo se denominan modalidades. Las modalidades deben ser exhaustivas y mutuamente excluyentes. Por ejemplo, si nos centramos en el carácter cualitativo "nacionalidad de los ciudadanos en España" habría cuatro modalidades: español, extranjero, doble nacionalidad y apátrida.

Las modalidades de un atributo pueden representarse mediante una escala nominal u ordinal, según sea el caso. En la escala nominal se asignan números a las categorías sin un orden lógico. Por ejemplo, en la tabla 1.8 aparece una escala nominal para el atributo "Nacionalidad".

Tabla 1.8. Escala nominal para el atributo "Nacionalidad"

Modalidad	Escala nominal
Español	1
Extranjero	2
Doble Nacionalidad	3
Apátrida	4

Fuente: Elaboración propia.

En la escala ordinal los valores numéricos que representan a las categorías cuentan con un orden lógico. Por ejemplo, en el cuestionario de la Encuesta de Condiciones de Vida (INE, 2025g) existe una pregunta sobre las dificultades que encuentran las personas para llegar a final de mes. Las modalidades recogidas para este atributo son: "con mucha facilidad", "con facilidad", "con cierta facilidad", "con cierta dificultad",

"con dificultad", "con mucha dificultad". Estas modalidades tienen una ordenación natural y se pueden recodificar con una escala ordinal donde los valores muestren un orden lógico (tabla 1.9).

Tabla 1.9. Escala ordinal para el atributo "Dificultades para llegar a fin de mes"

Modalidad	Escala ordinal
Con mucha facilidad	1
Con facilidad	2
Con cierta dificultad	3
Con dificultad	4
Con mucha dificultad	5

Fuente: Elaboración propia.

Las observaciones de las variables generan datos que pueden ser de varios tipos:

a) Datos de corte transversal:

Se observa un carácter para cada elemento de la muestra o población referido a un mismo periodo temporal. Por ejemplo, la renta media anual (€) por persona en cada Comunidad Autónoma publicada en el año 2024 (tabla 1.10) por la Encuesta de Condiciones de Vida. Hay que tener en cuenta que los datos de renta se refieren al año previo a la entrevista.

Tabla 1.10. Renta media anual (€) por persona por CCAA: año 2024

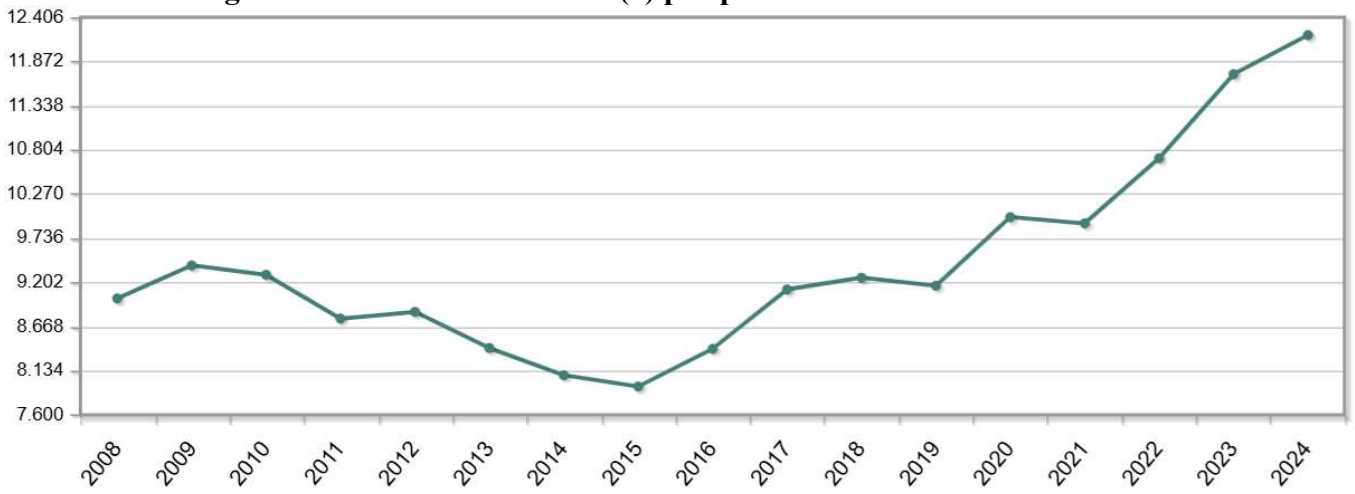
CCAA	Renta media (€)
Andalucía	12.191
Aragón	15.747
Principado de Asturias	16.201
Baleares	15.926
Canarias	13.372
Cantabria	14.708
Castilla y León	14.940
Castilla-La Mancha	12.357
Cataluña	16.546
Comunidad Valenciana	13.374
Extremadura	12.421
Galicia	14.558
Comunidad de Madrid	17.275
Región de Murcia	11.967
Comunidad Foral de Navarra	17.253
País Vasco	19.078
La Rioja	14.529
Ceuta	13.403
Melilla	12.745

Fuente: Elaboración propia a partir de datos del INE (2025h).

b) Datos o series temporales:

Se observa un carácter para un único elemento de la población en diversos periodos temporales. Por ejemplo, la renta media anual (€) por persona en Andalucía para los años 2008-2024 (figura 1.3).

Figura 1.3. Renta media anual (€) por persona en Andalucía: 2008-2024



Fuente: INE (2025h).

c) Datos de panel:

Se observa un carácter para varios elementos durante varios periodos. La renta media anual (€) por persona en cada Comunidad Autónoma para el periodo 2020-2024 (tabla 1.11).

Tabla 1.11. Renta media (€) por persona para cada CCAA: 2020-2024

CCAA	2020	2021	2022	2023	2024
Andalucía	9.990	9.915	10.703	11.719	12.191
Aragón	13.097	13.345	14.015	14.810	15.747
Principado de Asturias	12.786	12.861	13.777	15.432	16.201
Baleares	12.658	11.235	12.451	14.139	15.926
Canarias	9.935	10.161	10.716	12.177	13.372
Cantabria	12.748	12.848	13.811	14.162	14.708
Castilla y León	12.697	12.656	13.323	14.124	14.940
Castilla-La Mancha	10.485	10.257	11.037	11.913	12.357
Cataluña	14.170	14.159	14.692	15.830	16.546
Comunidad Valenciana	11.332	11.237	11.876	12.805	13.374
Extremadura	9.147	95.00	10.133	11.363	12.421
Galicia	11.469	11.453	12.352	13.147	14.558
Comunidad de Madrid	14.580	14.836	15.695	16.817	17.275
Región de Murcia	9.850	9.931	10.632	11.314	11.967
Comunidad Foral de Navarra	15.094	15.269	15.970	16.599	17.253
País Vasco	15.813	15.544	16.427	18.189	19.078
La Rioja	13.504	12.913	13.538	14.184	14.529
Ceuta	9.853	10.397	12.152	13.421	13.403
Melilla	11.427	12.012	13.089	13.854	12.745

Fuente: Elaboración propia a partir de datos del INE (2025h).

1.4 REDUCCIÓN ESTADÍSTICA Y DISTRIBUCIÓN DE FRECUENCIAS

En este epígrafe nos centramos en las variables y se exponen diversos métodos para agrupar y reducir la información estadística.

En primer lugar, si la variable toma pocos valores (k valores) estamos ante un caso de "Estadísticas con datos sin agrupar". Los k valores, ordenados de menor a mayor, se representan en una columna y, al lado, en otra columna se pone el número de veces que cada valor aparece repetido (frecuencias absolutas) dentro de la observación realizada en la población. A partir de aquí podemos definir los siguientes conceptos:

a) Distribución de frecuencias:

Conjunto de valores observados para una variable con sus frecuencias correspondientes. Simbólicamente viene dada por los pares (x_i, n_i) donde x_i son los valores de la variable y n_i son sus frecuencias. El subíndice "i" es una forma simplificada de representar un par genérico, como hay k pares: $i = 1 \dots k$.

b) Frecuencia absoluta:

La frecuencia absoluta se denomina n_i, y es el número de observaciones del valor x_i. Como hemos indicado anteriormente, el subíndice i puede tomar los valores i=1…k. La suma de todas las frecuencias absolutas equivale al tamaño de la población N (número total de observaciones):

$$N = \sum_{i=1}^{k} n_i$$

c) Frecuencia relativa:

La frecuencia relativa se denomina f_i, y es la proporción de observaciones con valor x_i. Se calcula como el cociente entre la frecuencia absoluta y el tamaño de la población:

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

Se puede demostrar fácilmente que la suma de todas las frecuencias relativas es igual a 1:

$$\sum_{i=1}^{k} f_i = 1$$

Demostración:

$$\sum_{i=1}^{k} f_i = \frac{\sum_{i=1}^{k} n_i}{N} = \frac{N}{N} = 1$$

La frecuencia relativa también se puede expresar en términos porcentuales, para ello basta con multiplicarla por 100:

$$p_i = 100 * f_i$$

La suma de todas las frecuencias relativas porcentuales equivale a 100:

$$\sum_{i=1}^{k} p_i = 100$$

Demostración:

$$\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} 100 * f_i = 100 \sum_{i=1}^{k} f_i = 100$$

d) Frecuencias acumuladas:

Las frecuencias acumuladas se designan por N_i para cada valor de la variable, son el resultado de sumar (acumular) las frecuencias absolutas correspondientes a dicho valor y a los valores inferiores. Por ejemplo, para un valor x_h, su frecuencia acumulada es igual $\sum_{i=1}^{h} n_i$. Las frecuencias acumuladas se pueden transformar en términos relativos dividiendo por $N \left(F_i = \frac{N_i}{N}\right)$ o en términos relativos porcentuales multiplicando por 100 ($P_i = \frac{N_i}{N} * 100$). Los conceptos anteriores asociados a la distribución de frecuencias se pueden resumir en la siguiente tabla 1.12.

Tabla 1.12. Distribución de frecuencias (datos no agrupados)

Variable	**Frecuencias absolutas**	**Frecuencias relativas**	**Frecuencias porcentuales**	**Frecuencias acumuladas**	**Frecuencias acumuladas relativas**	**Frecuencias acumuladas porcentuales**
$\boldsymbol{x_i}$	$\boldsymbol{n_i}$	$\boldsymbol{f_i = n_i/N}$	$\boldsymbol{p_i = f_i * 100}$	$\boldsymbol{N_i}$	$\boldsymbol{F_i}$	$\boldsymbol{P_i}$
x_1	n_1	f_1	p_1	$N_1 = n_1$	$F_1 = f_1$	$P_1 = p_1$
x_2	n_2	f_2	p_2	$N_2 = N_1 + n_2$	$F_2 = F_1 + f_2$	$P_2 = P_1 + p_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_i	f_i	p_i	$N_i = N_{i-1} + n_i$	$F_i = F_{i-1} + f_i$	$P_i = P_{i-1} + p_i$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k	p_k	$N_k = N$	$F_k = 1$	$P_k = 100$
	$\sum_{i=1}^{k} n_i = N$	$\sum_{i=1}^{k} f_i = 1$	$\sum_{i=1}^{k} p_i = 100$			

Fuente: Elaboración propia.

Las distribuciones de frecuencias absolutas se denominan "Estadísticas Primarias", mientras que las frecuencias relativas o acumuladas son "Estadísticas Derivadas".

A continuación, se presenta el Ejemplo 1.1 donde se analiza la distribución de frecuencias asociadas a la variable X: número de personas por hogar.

Ejemplo 1.1:

Distribución de frecuencias de datos no agrupados.
Variable X: número de personas por hogar.

x_i	n_i	$f_i = \frac{n_i}{N}$	$p_i = f_i * 100$	N_i	$F_i = \frac{N_i}{N}$	$P_i = F_i * 100$
1	10	0,20	20	10	0,20	20
2	12	0,24	24	22	0,44	44
3	14	0,28	28	36	0,72	72
4	12	0,24	24	48	0,96	96
5	2	0,04	4	50	1	100
	$\sum_{i=1}^{5} n_i = 50$	$\sum_{i=1}^{5} f_i = 1$	$\sum_{i=1}^{5} p_i = 100$			

En segundo lugar, si se observan muchos valores para la variable surgen las "Estadísticas con datos agrupados en intervalos". En estas estadísticas los valores observados están agrupados en intervalos del tipo $(L_{i-1}\text{-}L_i]$. Existe un convenio en que los intervalos sean abiertos por la izquierda, símbolo "(", y cerrados por la derecha, símbolo "]", es decir, en cada intervalo quedan excluidos los valores iguales al extremo inferior L_{i-1} e incluidos los valores iguales al extremo superior L_i.

En "Estadísticas de datos agrupados en intervalos" surgen conceptos adicionales:

a) Amplitud del intervalo:

La diferencia entre el extremo superior y el inferior del intervalo se denomina amplitud de intervalo:

$$a_i = L_i - L_{i-1}$$

b) Marca de clase:

Al punto central del intervalo se representa por:

$$x_i = \frac{L_{i-1} + L_i}{2}$$

c) Densidad de frecuencia:

Las densidades de frecuencias (o alturas de los intervalos) se obtienen dividiendo la frecuencia absoluta entre la amplitud del intervalo:

$$h_i = \frac{n_i}{a_i}$$

Si se añaden estos nuevos conceptos a la tabla 1.12 surge la tabla 1.13:

Tabla 1.13. Distribución de frecuencias (datos agrupados)

Intervalo	**Marca de clase**	**Frecuencia absoluta**	**Amplitud de intervalo**	**Densidad frecuencia**	**Frecuencia relativa**	**Frecuencia relativa porcentual**	**Frecuencia acumulada**	**Frecuencia relativa acumulada**	**Frecuencia relativa acumulada porcentual**
$L_{i-1} - L_i$	$x_i = \frac{L_{i-1}+L_i}{2}$	n_i	$a_i = L_i - L_{i-1}$	$h_i = \frac{n_i}{a_i}$	$f_i = \frac{n_i}{N}$	$p_i = f_i * 100$	N_i	F_i	P_i
$L_0 - L_1$	x_1	n_1	a_1	$h_1 = \frac{n_1}{a_1}$	f_1	p_1	$N_1 = n_1$	$F_1 = f_1$	$P_1 = p_1$
$L_1 - L_2$	x_2	n_2	a_2	$h_2 = \frac{n_2}{a_2}$	f_2	p_2	$N_2 = N_1 + n_2$	$F_2 = F_1 + f_2$	$P_2 = P_1 + p_2$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$L_{i-1} - L_i$	x_i	n_i	a_i	$h_i = \frac{n_i}{a_i}$	f_i	p_i	$N_i = N_{i-1} + n_i$	$F_i = F_{i-1} + f_i$	$P_i = P_{i-1} + p_i$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$L_{k-1} - L_k$	x_k	n_k	a_k	$h_k = \frac{n_k}{a_k}$	f_k	p_k	$N_k = N$	$F_k = 1$	$P_k = 100$
		$\sum_{i=1}^{k} n_i = N$			$\sum_{i=1}^{k} f_i = 1$	$\sum_{i=1}^{k} p_i = 100$			

Fuente: Elaboración propia.

A continuación, se muestran los Ejemplos 1.2 y 1.3 donde se presentan dos casos de distribuciones de frecuencias con datos agrupados en intervalos.

Ejemplo 1.2:

Distribuciones de frecuencias con datos agrupados en intervalos con amplitud constante.
Variable X: Salarios por hora en € de los trabajadores de una empresa.

$L_{i-1} - L_i$	$a_i = L_i - L_{i-1}$	x_i	n_i	f_i	p_i	N_i	F_i	P_i
10 – 15	5	12,5	5	0,125	12,5	5	0,125	12,5
15 – 20	5	17,5	9	0,225	22,5	14	0,35	35
20 – 25	5	22,5	12	0,300	30	26	0,65	65
25 – 30	5	27,5	14	0,350	35	40	1	100

Ejemplo 1.3:

Distribuciones de frecuencias con datos agrupados en intervalos con amplitud variable.
Variable X: Salarios por hora en € de los trabajadores de una empresa.

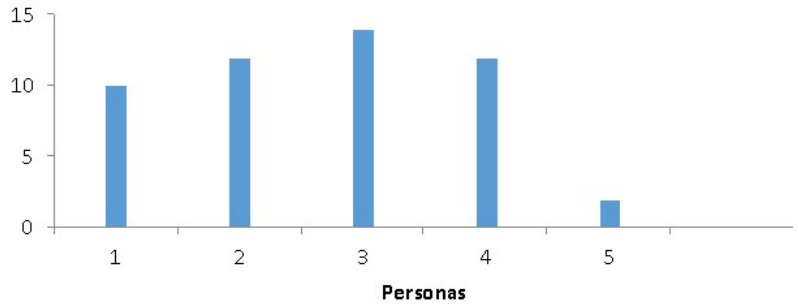
L_{i-1} - L_i	$a_i = L_{i-} - L_{i-1}$	x_i	n_i	$h_i = \frac{n_i}{a_i}$	f_i	p_i	N_i	F_i	P_i
0 – 2	2	1	2	1	0,2	20	2	0,2	20
2 – 4	2	3	5	2,5	0,5	50	7	0,7	70
4 – 10	6	7	3	0,5	0,3	30	10	1	100

1.5 REPRESENTACIONES GRÁFICAS.

Para la representación de estadísticas, relativas a variables, se puede utilizar el Diagrama de Barras, el Histograma y el Polígono de Frecuencias Acumuladas.

El Diagrama de Barras permite la representación de las frecuencias absolutas o relativas (no acumuladas) para estadísticas correspondientes a datos sin agrupar. En el Diagrama de Barras se utiliza un diagrama cartesiano, donde en el eje de abscisas se ponen los valores de la variable X y en el de ordenadas las frecuencias, absolutas o relativas, pero sin acumular. En cada valor de la variable se eleva un segmento de altura igual a la frecuencia que le corresponde. A continuación, se representa el diagrama de barras para la distribución asociada al ejemplo 1.1 (Figura 1.4).

Figura 1.4. Diagrama de barras: Ejemplo1.1



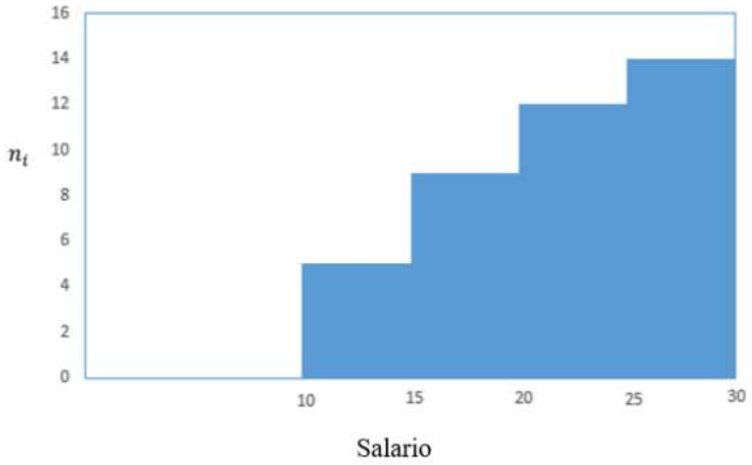
Fuente: Elaboración propia.

Con datos agrupados, la representación de las frecuencias absolutas o relativas (no acumuladas) se realiza a través del Histograma. Se pueden presentar dos casos:

a) Los intervalos tienen amplitudes constantes:

Se utiliza un diagrama cartesiano. En el eje de abscisas se ponen los límites de los intervalos y en el de ordenadas las frecuencias, absolutas o relativas. Sobre los intervalos se levantan rectángulos que tienen por base la amplitud del intervalo y por altura su frecuencia. En la figura 1.5 aparece el histograma para el ejemplo 1.2.

Figura 1.5. Histograma: Ejemplo 1.2

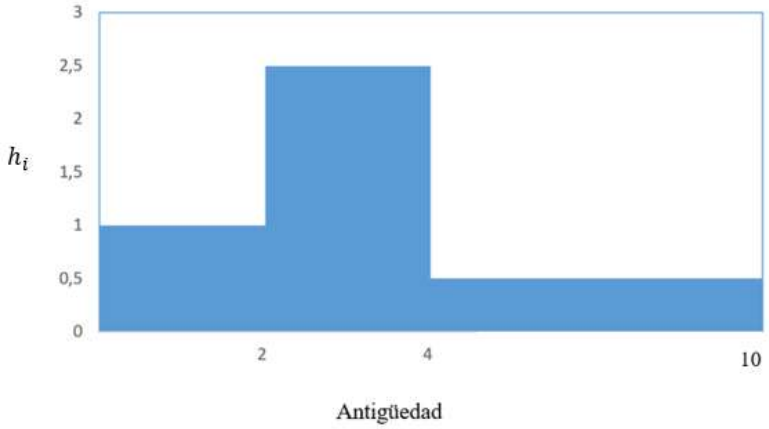


Fuente: Elaboración propia.

b) Los intervalos presentan amplitudes distintas:

Se utiliza un diagrama cartesiano. En el eje de abscisas se ponen los límites de los intervalos y en el de ordenadas las densidades de frecuencias. Sobre los intervalos se levantan rectángulos que tienen por base la amplitud del intervalo y por altura su densidad de frecuencia. En la figura 1.6 aparece el histograma para el ejemplo 1.3:

Figura 1.6. Histograma: Ejemplo 1.3

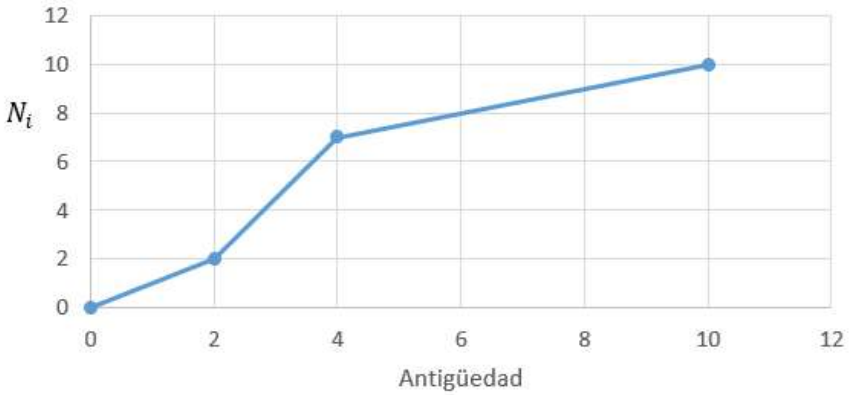


Fuente: Elaboración propia.

Para la representación de las frecuencias acumuladas, en el caso de estadísticas con datos agrupados, se utiliza el Polígono de Frecuencias Acumuladas. Ahora, en el eje horizontal se representan los límites de los intervalos y en el de ordenadas las frecuencias acumuladas. A continuación, partiendo desde el límite inferior del primer intervalo, se van uniendo los pares de puntos formados entre los límites superiores y sus

correspondientes frecuencias acumuladas. En la figura 1.7 se expone el Polígono de Frecuencias Acumuladas resultante del ejemplo 1.3.

Figura 1.7. Polígono de Frecuencias Acumuladas: Ejemplo 1.3



Fuente: Elaboración propia.

Las representaciones gráficas para los caracteres relativos a atributos se denominan Diagrama de Sectores y Diagrama de Rectángulos.

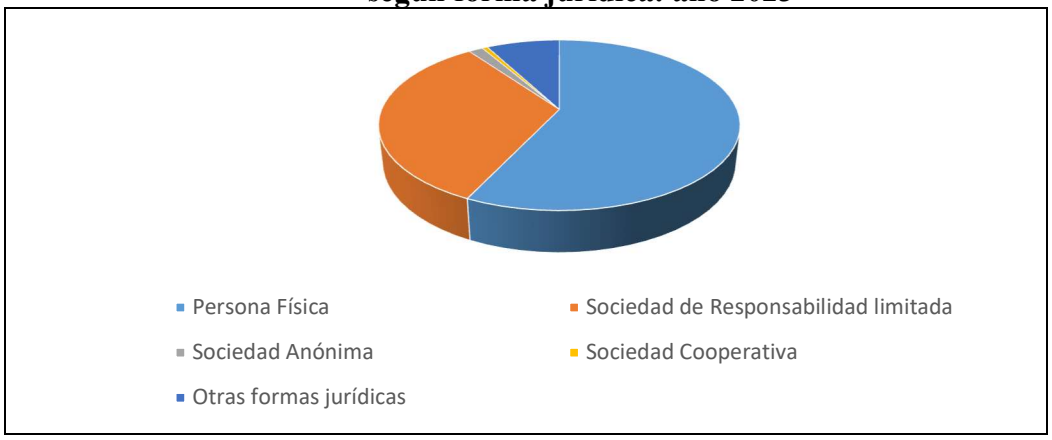
En el Diagrama de Sectores la representación consiste en dividir un círculo en tanto sectores como modalidades existan, asignando a cada modalidad del atributo un arco de círculo proporcional a su frecuencia. En la tabla 1.14 se expone la distribución del número de empresas en España según forma jurídica para el año 2023, y en la figura 1.8 se presenta su correspondiente Diagrama de Sectores.

Tabla 1.14. Número de empresas en España según forma jurídica: año 2023

Forma jurídica	**Número**	**Grados**
Persona Física	1.831.133	205,52
Sociedad de Responsabilidad Limitada	1.058.429	118,79
Sociedad Anónima	49.115	5,51
Sociedad Cooperativa	18.743	2,10
Otras Formas Jurídicas	250.160	28,08
Total	3.207.580	360

Fuente: Elaboración propia a partir de datos del Directorio Central de Empresas (INE, 2024a).

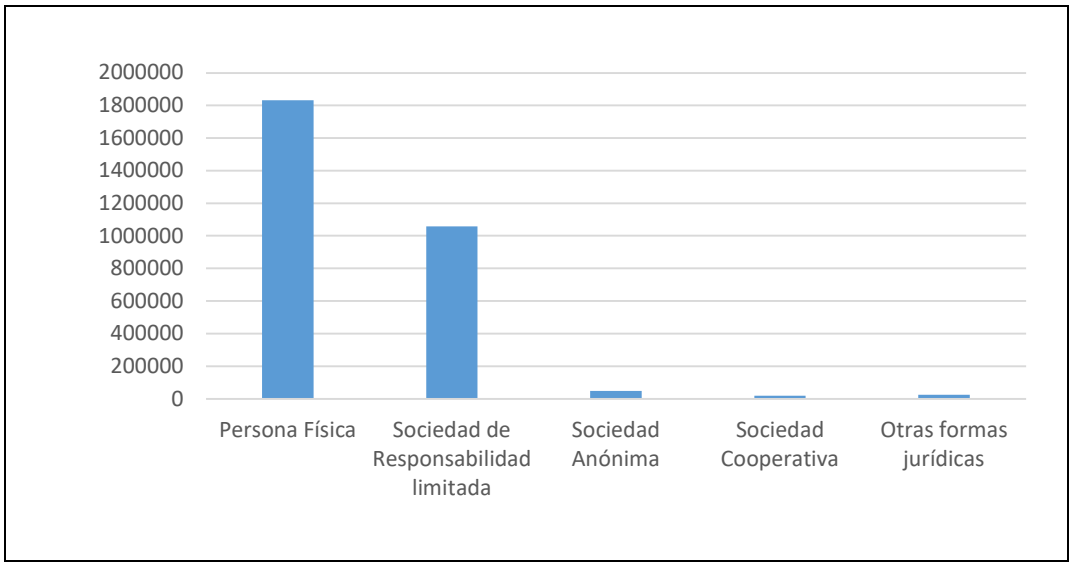
Figura 1.8. Diagrama de sectores correspondiente al número de empresas en España según forma jurídica: año 2023



Fuente: Elaboración propia a partir de datos del Directorio Central de Empresas (INE, 2024a).

En el Diagrama de Rectángulos se utiliza un diagrama cartesiano. En el eje de abscisas se ponen las modalidades del atributo y en el de ordenadas las frecuencias absolutas o relativas. Sobre los atributos se levantan rectángulos de amplitud constante y de altura su frecuencia. En la figura 1.9 aparece el Diagrama de Rectángulos asociado al número de empresas en España según forma jurídica para el año 2023.

Figura 1.9. Diagrama de rectángulos correspondiente al número de empresas en España según forma jurídica: año 2023

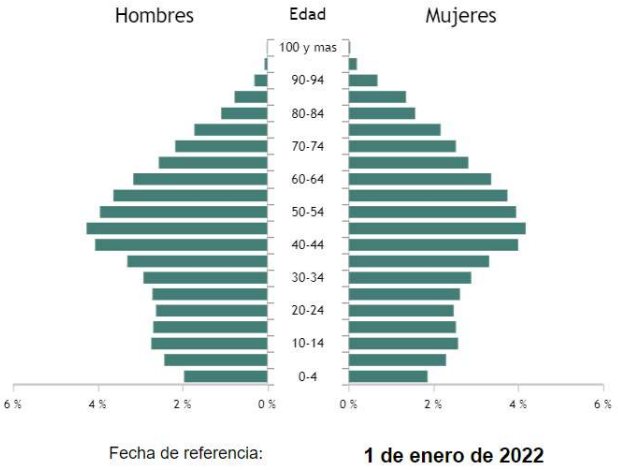


Fuente: Elaboración propia a partir de datos del Directorio Central de Empresas (INE, 2024a).

Por último, en esta sección dedicada a las representaciones gráficas, cabe realizar una mención especial a las Pirámides de Población que son representaciones gráficas de estadísticas mixtas, es decir de la variable "edad" junto al atributo "sexo". Las Pirámides de Población son dos histogramas que se rotan y se unen de forma que comparten el eje de ordenadas sobre el que se representa la variable edad. Uno representa la

distribución de frecuencias de los hombres y el otro el de las mujeres. Las diferentes formas que adopta informan sobre la estructura de edad de la población y su envejecimiento. A continuación, en la figura 1.10 se muestra la Pirámide de Población para España en el año 2022.

Figura 1.10. Pirámide de Población para España en 2022



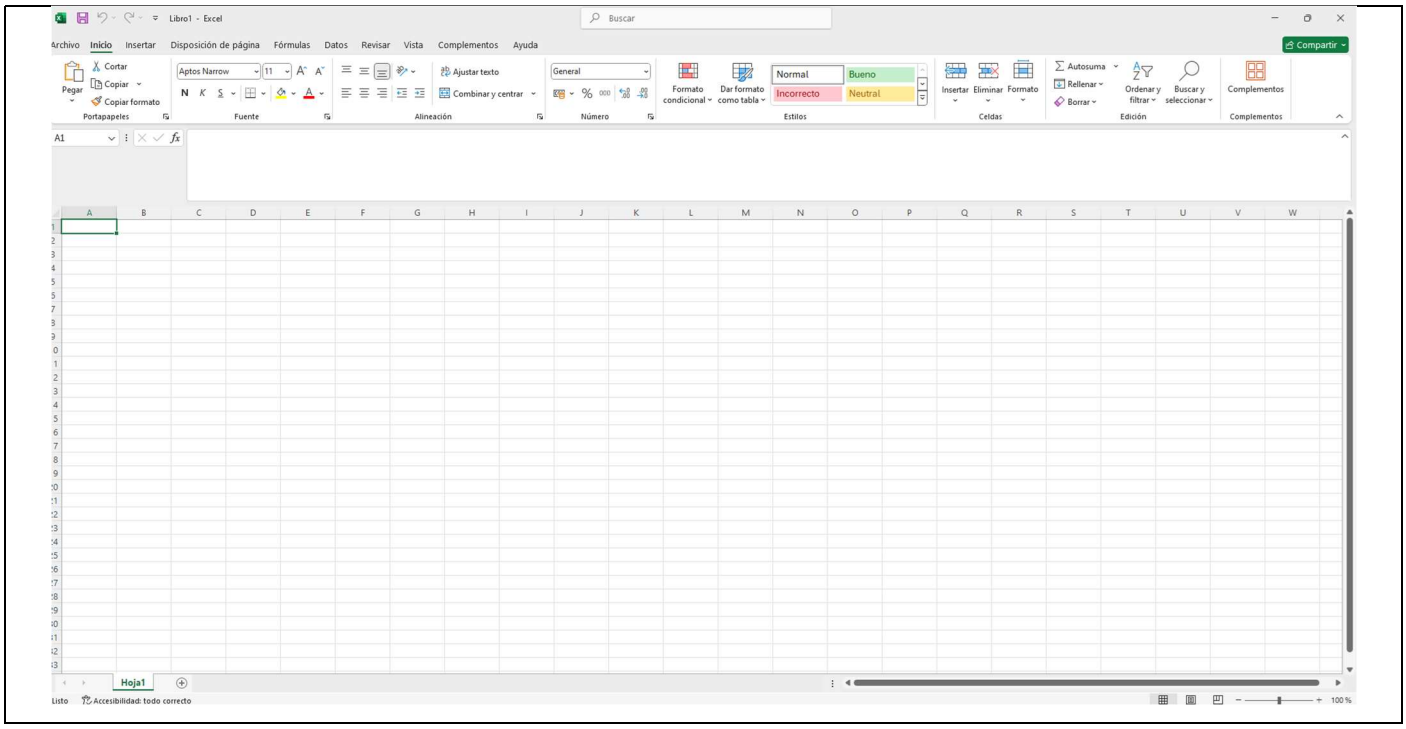
Fuente: INE (2024b).

Anexo 1.A. Ejemplos de obtención de distribuciones de frecuencias y representaciones gráficas con Excel

El programa Excel es un software en formato de hoja de cálculo que permite la organización y el tratamiento de grandes cantidades de datos desde diversos puntos de vista. Sus posibilidades son amplias y abarca desde cálculos matemáticos, estadísticos y lógicos sencillos hasta operaciones más complejas de análisis de datos. A lo largo de este documento, se van a presentar diversos anexos que muestran cómo se pueden desarrollar a través de la hoja de cálculo Excel algunos de los conceptos tratados.

Este primer anexo asociado al tema 1 está dedicado, por una parte, a replicar los ejemplos propuestos y, por otra parte, a elaborar algunos de los gráficos expuestos en el epígrafe 1.5. En primer lugar, a modo de breve introducción, es imprescindible mostrar inicialmente el formato que presenta la hoja de cálculo Excel, que consiste en la intersección entre filas con asignación numérica y columnas con asignación de letras. Un extracto de la pantalla inicial aparece en la figura 1.A.1.

Figura 1.A.1. Extracto de una hoja de cálculo de Excel



Fuente: Elaboración propia a partir del programa Excel.

En las casillas de las hojas de cálculo se introduce la información. Por ejemplo, si nos centramos en el Ejemplo 1.1, podríamos rellenar los datos relativos a x_i y a n_i, y generar posteriormente los valores del resto de columnas como se muestra en la figura 1.A.2 a través de operaciones sencillas.

Figura 1.A.2. Obtención de la distribución de frecuencias del Ejemplo 1.1 a partir de Excel

- Fórmulas:

	A	B	C	D	E	F	G
1	x_i	n_i	$f_i = \frac{n_i}{N}$	$p_i = f_i * 100$	N_i	$F_i = \frac{N_i}{N}$	$P_i = F_i * 100$
2	1	10	=B2/SUMA($B$2:$B$6)	=C2*100	=A2	=C2	=F2*100
3	2	12	=B3/SUMA($B$2:$B$6)	=C3*100	=A3+E2	=C3+F2	=F3*100
4	3	14	=B4/SUMA($B$2:$B$6)	=C4*100	=A4+E3	=C4+F3	=F4*100
5	4	12	=B5/SUMA($B$2:$B$6)	=C5*100	=A5+E4	=C5+F4	=F5*100
6	5	2	=B6/SUMA($B$2:$B$6)	=C6*100	=A6+E5	=C6+F5	=F6*100

- Resultados:

	A	B	C	D	E	F	G
1	x_i	n_i	$f_i = \frac{n_i}{N}$	$p_i = f_i * 100$	N_i	$F_i = \frac{N_i}{N}$	$P_i = F_i * 100$
2	1	10	0.2	20	1	0.2	20
3	2	12	0.24	24	3	0.44	44
4	3	14	0.28	28	6	0.72	72
5	4	12	0.24	24	10	0.96	96
6	5	2	0.04	4	15	1	100

Fuente: Elaboración propia a partir del programa Excel.

Como se puede observar en la figura 1.A.2, las operaciones entre casillas pueden realizarse a través de funciones o a través de sencillos operadores matemáticos. Por ejemplo, la casilla C2 de la primera fila se genera utilizando la siguiente fórmula: =B2/SUMA($B$2:$B$6) y la casilla D2 se crea con: =C2*100. Si marcásemos estas casillas, y las copiásemos en el resto de casillas usadas de las columnas C y D, aparecerían los resultados derivados de la aplicación de las fórmulas anteriores, variando la letra y el número según corresponda, siempre que no aparezcan acotados por el símbolo del $. Por ejemplo, el valor 0,24 de la casilla C3 se ha obtenido aplicando la fórmula: =B3/SUMA($B$2:$B$6).

El manejo de la hoja de cálculo es muy intuitivo, y la mejor forma de aprender su funcionamiento es empezar a practicar desde el principio a través de ejemplos sencillos. En este sentido, si se replicase el ejercicio anterior para los datos para los Ejemplo 1.2 y 1.3, respectivamente, se obtendrían los resultados que se exponen en las figuras 1.A.3 y 1.A.4.

Figura 1.A.3. Obtención de la distribución de frecuencias del Ejemplo 1.2 a partir de Excel

- Fórmulas:

	A	B	C	D	E	F	G	H	I
1	L_{i-1} - L_i	$a_i = L_i - L_{i-1}$	x_i	n_i	f_i	p_i	N_i	F_i	P_i
2	10 – 15	=15-10	=(10+15)/2	5	=(D2/SUMA($D$2:$D$5))	=E2*100	=D2	=E2	=H2*100
3	15 – 20	=20-15	=(20+15)/2	9	=(D3/SUMA($D$2:$D$5))	=E3*100	=D3+G2	=E3+H2	=H3*100
4	20 – 25	=25-20	=(25+20)/2	12	=(D4/SUMA($D$2:$D$5))	=E4*100	=D4+G3	=E4+H3	=H4*100
5	25 – 30	=30-25	=(30+25)/2	14	=(D5/SUMA($D$2:$D$5))	=E5*100	=D5+G4	=E5+H4	=H5*100

- Resultados:

	A	B	C	D	E	F	G	H	I
1	L_{i-1} - L_i	$a_i = L_i - L_{i-1}$	x_i	n_i	f_i	p_i	N_i	F_i	P_i
2	10 – 15	5	12.5	5	0.125	12.5	5	0.1	12.5
3	15 – 20	5	17.5	9	0.225	22.5	14	0.4	35
4	20 – 25	5	22.5	12	0.3	30	26	0.7	65
5	25 – 30	5	27.5	14	0.35	35	40	1	100

Fuente: Elaboración propia a partir del programa Excel.

Figura 1.A.4. Obtención de la distribución de frecuencias del Ejemplo 1.3 a partir de Excel

- Fórmulas:

	A	B	C	D	E	F	G	H	I	J
1	L_{i-1} - L_i	$a_i = L_{i-} - L_{i-1}$	x_i	n_i	hi	f_i	p_i	N_i	F_i	P_i
2	0 – 2	=2-0	=(0+2)/2	2	=(D2/B2)	=(D2/SUMA($D$2:$D$4)	=F2*100	=D2	=(H2/$H$4)	=I2*100
3	2 – 4	=4-2	=(4+2)/2	5	=(D3/B3)	=(D3/SUMA($D$2:$D$4)	=F3*100	=D3+H2	=(H3/$H$4)	=I3*100
4	4 – 10	=10-4	=(10+4)/2	3	=(D4/B4)	=(D4/SUMA($D$2:$D$4)	=F4*100	=D4+H3	=(H4/$H$4)	=I4*100

- Resultados:

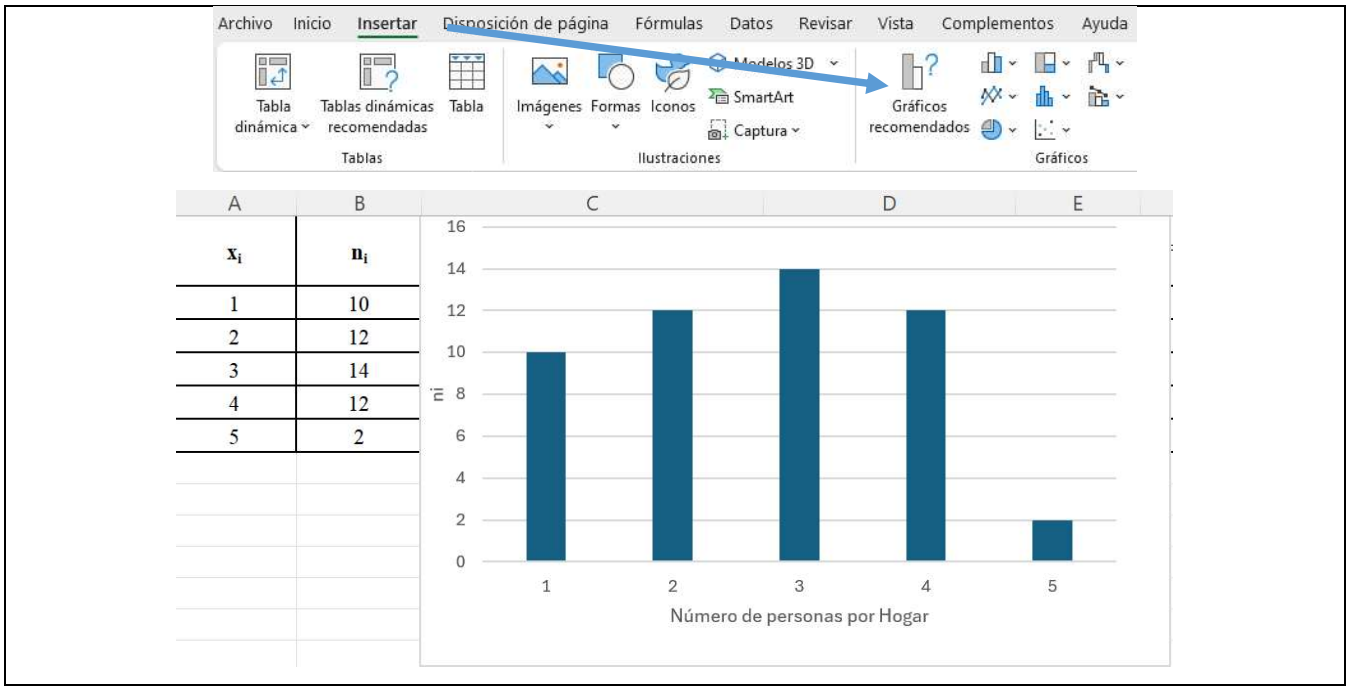
	A	B	C	D	E	F	G	H	I	J
1	L_{i-1} - L_i	$a_i = L_{i-} - L_{i-1}$	x_i	n_i	hi	f_i	p_i	N_i	F_i	P_i
2	0 – 2	2	1	2	1	0.2	20	2	0.2	20
3	2 – 4	2	3	5	2.5	0.5	50	7	0.7	70
4	4 – 10	6	7	3	0.5	0.3	30	10	1	100

Fuente: Elaboración propia a partir del programa Excel.

Otros de los objetivos de este anexo es mostrar cómo realizar algunas representaciones gráficas básicas con Excel. En primer lugar, se expone cómo obtener un Diagrama de Barras, utilizando los datos el Ejemplo 1.1. En la figura 1.A.5 se muestra el gráfico resultante de marcar las columnas x_i y n_i, y seguir la secuencia

"Insertar"→" Gráfico recomendado". Los títulos de los ejes, y otras cuestiones relativas al formato del gráfico se han realizado, haciendo clic en el gráfico y dirigiéndose en la barra de herramientas hacia la opción "Diseño de Gráfico".

Figura 1.A.5. Representación de un Diagrama de Barras con Excel: Ejemplo 1.1

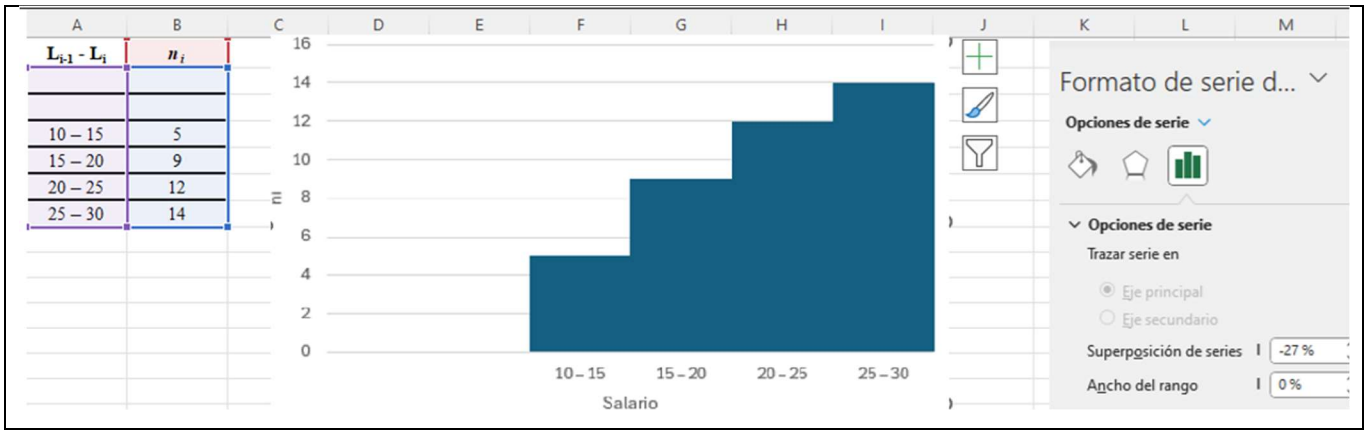


x_i	n_i
1	10
2	12
3	14
4	12
5	2

Fuente: Elaboración propia a partir del programa Excel.

En segundo lugar, nos centramos en la representación de un Histograma con intervalos de amplitud constante, siguiendo el Ejemplo 1.2. Aparte de los pasos iniciales indicados en el gráfico anterior, ahora hay que activar la opción "Formato de Serie" haciendo clic en alguno de los rectángulos representativos de los datos, y asignar un valor 0 en el "Ancho de rango" (Figura 1.A.6).

Figura 1.A.6. Representación de un Histograma con amplitud constante con Excel: Ejemplo 1.2

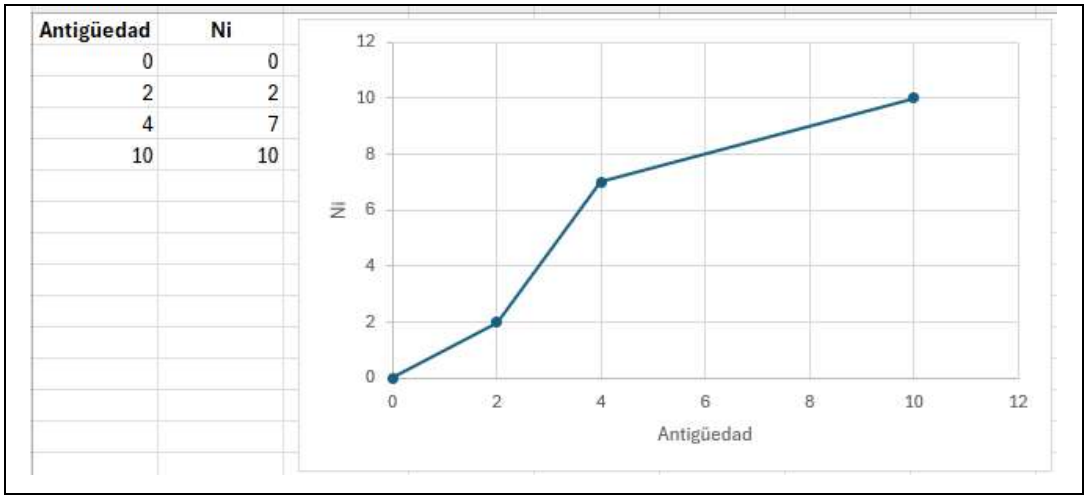


$L_{i-1} - L_i$	n_i
10 – 15	5
15 – 20	9
20 – 25	12
25 – 30	14

Fuente: Elaboración propia a partir del programa Excel.

En tercer lugar, en la figura 1.A.7 aparece la representación del polígono de frecuencia acumulada asociada al Ejemplo 1.3, correspondiente a una distribución de frecuencias de datos agrupados con intervalos variables. En este caso, en el eje de ordenadas se muestran las frecuencias acumuladas N_i, y en el eje X los valores máximos de los intervalos, excepto el primer punto que se crea con el límite inferior del primer intervalo, y un valor de 0 para N_i .

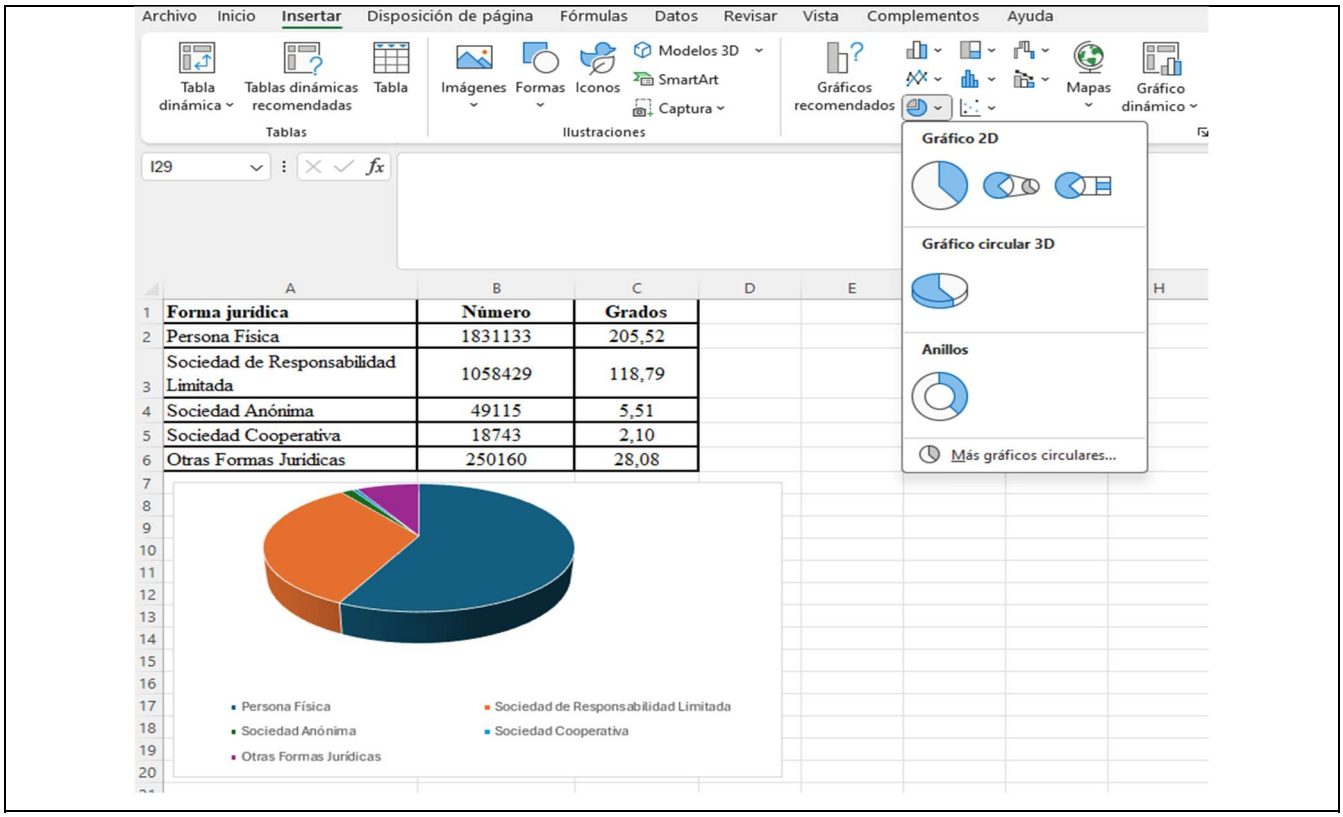
Figura 1.A.7. Representación de un Polígono de Frecuencias Acumuladas con Excel: Ejemplo 1.3



Fuente: Elaboración propia a partir del programa Excel.

Por último, la figura 1.A.8 representa un diagrama de sectores con los datos relativos a la tabla 1.14. En este caso, hay que activar la opción de "Gráfico circular".

Figura 1.A.8. Representación de un Diagrama de Sectores con Excel



Fuente: Elaboración propia a partir del programa Excel.

TEMA 2. MEDIDAS DE POSICIÓN

2.1 INTRODUCCIÓN

Un promedio o medida de posición es un valor de la variable, observado o no, que pretende ser representativo del conjunto de observaciones de la población. A los promedios también se les denomina medidas de tendencia central. Este tema está dedicado al estudio de los principales promedios que pueden utilizarse en la Estadística Aplicada a la Investigación Social.

2.2 MEDIA ARITMÉTICA, GEOMÉTRICA Y CUADRÁTICA.

2.2.1 MEDIA ARITMÉTICA

La media aritmética es el número que resulta de dividir la suma de todos los valores observados entre el número total de observaciones. Su fórmula es la siguiente:

$$\bar{X} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i} = \frac{\sum x_i n_i}{N}$$

La unidad de medida de la media aritmética, $\bar{X}$, es la misma que la de la variable.

Ejemplo 2.1:

Distribución de frecuencias con datos sin agrupar y con frecuencias unitarias.

X: Número de desempleados por hogar.

Número de desempleados (x_i)	$\bar{X} = \frac{\sum x_i}{N} = \frac{10}{5} = 2$ desempleados
0	
1	
2	
3	
4	

Ejemplo 2.2:

Distribución de frecuencias con datos sin agrupar y sin frecuencias unitarias.

X: Número de personas activas por hogar.

Número de personas activas (x_i)	**Número de hogares (n_i)**	$x_i n_i$	$\bar{X} = \frac{\sum x_i n_i}{N} = \frac{30}{15} = 2\ personas$
1	6	6	
2	4	8	
3	4	12	
4	1	4	

Ejemplo 2.3:

Distribución de frecuencias con datos agrupados.

X: Salario por hora (€).

Salario €/ hora	**Número de trabajadores (n_i)**	x_i	$x_i n_i$	$\bar{X} = \frac{\sum x_i n_i}{N} = \frac{3470}{54} = 64{,}26\ €/h$
15 - 35	9	25	225	
35 - 55	12	45	540	
55 - 75	14	65	910	
75 - 95	10	85	850	
95 - 115	9	105	945	

Como promedio algunas ventajas de la media aritmética son:

a) Utiliza toda la información disponible, es decir, hace uso de todos los valores de la distribución para su cálculo.

b) Es única y sencilla de calcular.

c) Tiene una expresión algebraica definida, lo que permite derivar propiedades.

Por el contrario, algunos de sus principales inconvenientes son:

a) Es muy sensible a la existencia de valores extremos o anómalos en la variable.

b) En el caso de datos agrupados en intervalos, el valor de la media aritmética depende de los intervalos elegidos.

c) No es posible su cálculo en distribuciones abiertas, es decir, aquellas donde el límite inferior del primer intervalo o el límite superior del último intervalo estén sin determinar. Esto ocurrirá siempre que no exista información adicional que permita cerrar los intervalos.

Las propiedades de la media aritmética son las siguientes:

1.- La suma de desviaciones de los valores de la variable respecto a la media vale cero:

$$\sum_{i=1}^{k} (x_i - \bar{X}) n_i = 0$$

Demostración:

$$\sum(x_i - \bar{X})n_i = \sum x_i n_i - \sum \bar{X} n_i = \bar{X}N - \bar{X}N = 0$$

2.- La media aritmética no varía si todas las frecuencias de la distribución se multiplican (o dividen) por una constante. Es decir, dada una distribución de frecuencias (x_i, n_i) con media aritmética igual a $\bar{X}$, si se multiplican (o dividen) las frecuencias por una constante b se genera otra distribución de frecuencias (x_i, bn_i) cuya media aritmética $\bar{X}'$ es igual a:

$$\bar{X}' = \bar{X}$$

Demostración:

$$\bar{X}' = \frac{\sum x_i\, bn_i}{bN} = \frac{b\sum x_i n_i}{bN} = \bar{X}$$

3.- Le afecta el cambio de origen.

Si a todos los valores de la variable se les suma (o resta) una constante C, la media aritmética queda aumentada (o disminuida) en esa constante:

$$x_i' = C + x_i \rightarrow \bar{X}' = C + \bar{X}$$

Demostración:

$$\bar{X}' = \frac{\sum(C + x_i)\, n_i}{N} = C\frac{\sum n_i}{N} + \frac{\sum x_i n_i}{N} = C + \bar{X}$$

Ejemplo 2.4:

La renta media mensual de un conjunto de hogares es de 1500 €, si todos los hogares reciben una subvención de 100 €, la nueva renta media es de 1600 €, ya que:

$$x_i' = x_i + 100 \rightarrow \bar{X}' = \bar{X} + 100 = 1500 + 100 = 1600€$$

4.- Le afecta el cambio de escala.

Si todos los valores de la variable se multiplican (o dividen) por una constante C, la media aritmética queda multiplicada (o dividida) por esa constante:

$$x_i' = C * x_i \rightarrow \bar{X}' = C * \bar{X}$$

Demostración:

$$\bar{X}' = \frac{\sum (Cx_i)\, n_i}{N} = C\frac{\sum x_i n_i}{N} = C\bar{X}$$

Ejemplo 2.5:

La renta media mensual de un conjunto de hogares es de 1500 €, si todos los hogares reciben una subvención que incrementa su renta en un 10%, la nueva renta media es de 1650 €.

El incremento del 10% implica multiplicar cada renta por 1,1:

$$\frac{X_1 - X_0}{X_0} * 100 = 10 \;\rightarrow\; X_1 = 1{,}1 * X_0$$

A continuación, se aplica la propiedad asociada al cambio de escala:

$$x_i' = 1{,}1 * x_i \rightarrow \bar{X}' = 1{,}1 * \bar{X} = 1{,}1 * 1500 = 1650€$$

5.- Si se produce un cambio simultáneo de origen (suma o resta de una cantidad "a") y de escala (multiplicación o división de una cantidad "b"), la nueva media aritmética es igual a:

$$x_i' = a + bx_i \rightarrow \bar{X}' = a + b\bar{X}$$

Demostración:

$$\bar{X}' = \frac{\sum (a + bx_i)\, n_i}{N} = a\frac{\sum n_i}{N} + b\frac{\sum x_i n_i}{N} = a + b\bar{X}$$

6.- Si a partir de un conjunto de valores se obtienen dos o más subconjuntos disjuntos, puede calcularse la media aritmética de todo el conjunto a partir de las medias aritméticas de cada subconjunto. La aplicación de la propiedad al caso de dos subconjuntos sería la siguiente:

Conocidos $\begin{cases} \bar{X}_1\ (media\ aritmética\ del\ grupo\ 1)\ ,\ \ N_1(tamaño\ del\ grupo\ 1) \\ \bar{X}_2\ (media\ aritmética\ del\ grupo\ 2)\ ,\ \ N_2(tamaño\ del\ grupo\ 2) \end{cases}$

$$\bar{X} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

Demostración:

$$\bar{X} = \frac{\sum x_i n_i}{N} = \frac{\sum x_i (n_{1,i} + n_{2,i})}{N} = \frac{N_1 \frac{\sum x_i n_{1,i}}{N_1} + N_2 \frac{\sum x_i n_{2,i}}{N_2}}{N} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

Ejemplo 2.6:

Las notas medias de dos grupos de Estadística Aplicada a la investigación Social son de 7 y 7,5 puntos respectivamente, si los tamaños de los grupos son de 70 y 80 alumnos, ¿Cuál es la nota media del total de los alumnos?

$$\bar{X} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2} = \frac{70 * 7 + 80 * 7{,}5}{150} = 7{,}26\ puntos$$

2.2.2 MEDIA ARITMÉTICA PONDERADA:

La media aritmética ponderada es el valor que resulta de dividir la suma de todas las observaciones, multiplicadas cada una por un peso o ponderación (ω_i). El peso recoge la importancia de esa observación entre la suma de ponderaciones:

$$\bar{X}_p = \frac{\sum_{i=1}^{n} x_i \omega_i}{\sum_{i=1}^{n} \omega_i}$$

La suma de las ponderaciones ha de ser igual a 100, o a 1 si éstas están expresadas en tanto por uno.

Ejemplo 2.7:

Una oposición consta de cuatro exámenes, cada uno de ellos con su ponderación. Obtenga la nota final obtenida por el opositor:

Notas (x_i)	**Peso (ω_i)**	$x_i\omega_i$
7,5	16,5	123,75
4	33,5	134
5,8	25,0	145
7,5	25,0	187,5

$$\bar{X}_p = \frac{\sum_{i=1}^{4} x_i \omega_i}{\sum_{i=1}^{4} \omega_i} = \frac{590{,}25}{100} = 5{,}90\ puntos$$

2.2.3 MEDIA GEOMÉTRICA

La Media Geométrica se utiliza como promedio para algunas variables expresadas en porcentajes como, por ejemplo, tipos de interés o tasas de variación. La expresión algebraica de la media geométrica con frecuencias unitarias es la siguiente:

$$G = \sqrt[N]{x_1 * x_2 * \dots * x_N} = \sqrt[N]{\prod_{i=1}^{N} x_i}$$

Un inconveniente de la media geométrica es que no tiene sentido calcularla cuando algún valor es 0 o negativo.

Ejemplo 2.8:

Se dispone de información sobre el tipo de interés hipotecario constituido para 5 viviendas constituidas en el mismo periodo temporal. Obtenga la media geométrica del tipo de interés.

Tipo de interés hipotecario (%)	$G = \sqrt[5]{\prod_{i=1}^{5} x_i} = 3{,}28\%$
3,10	
3,24	
3,30	
3,35	
3,4	

2.2.4 MEDIA CUADRÁTICA

La Media Cuadrática es el promedio adecuado cuando la variable toma valores positivos y negativos. Su expresión algebraica es:

$$C = \sqrt{\frac{x_1^2 n_1 + x_2^2 n_2 + \dots + x_k^2 n_k}{N}} = \sqrt{\frac{\sum_{i=1}^{k} x_i^2 n_i}{N}}$$

Ejemplo 2.9:

Se conocen los errores (en puntos porcentuales) cometidos en la predicción de la tasa de paro anual en España por la Comisión Europea para el periodo 2020-2024. Obtenga el error cuadrático medio.

Año	Error (puntos porcentuales)	
2020	0,43	$C = \sqrt{\frac{x_1^2+x_2^2+\cdots+x_5^2}{N}} = \sqrt{\frac{\sum_{i=1}^{5} x_i^2}{N}}$ =0,30 puntos porcentuales
2021	-0,29	
2022	0,33	
2023	0,08	
2024	-0,26	

2.3 MEDIANA Y MODA

2.3.1 MEDIANA

La Mediana es aquel valor de la variable que divide a la distribución en dos partes iguales; es decir, dejando a su izquierda y a su derecha el mismo número de observaciones. Tiene, por tanto, la misma unidad de medida que la variable.

Para el cálculo de la Mediana hay que ordenar los valores de la variable de menor a mayor. Después, se procede según el tipo de distribución.

1.- Para distribuciones de datos no agrupados con frecuencias unitarias**,** se pueden presentar dos casos:

a) Número de observaciones impar:

La mediana es el valor central una vez ordenados de menor a mayor.

Ejemplo 2.10:

Obtenga la mediana para el número de horas semanales dedicadas al cuidado de los hijos:

Horas semanales dedicadas al cuidado de hijos (x_i)	
23	$M_e = 29\ horas$
26	
29	
31	
32	

b) Número de observaciones par:

La mediana es la semisuma de los dos valores centrales, una vez ordenados éstos de menor a mayor.

Ejemplo 2.11:

Obtenga la mediana para el número de horas semanales dedicadas al cuidado de los hijos:

Horas semanales dedicadas al cuidado de hijos (x_i)	$Me = \frac{29 + 31}{2} = 30\ horas$
23	
26	
29	
31	
32	
34	

2.- Para distribuciones de datos no agrupados con frecuencias no unitarias, también se pueden presentar dos casos distintos:

a) Existe una frecuencia acumulada igual a N/2. La mediana será en ese caso la media aritmética entre el valor de la variable al que le corresponde N/2 y el siguiente.

Ejemplo 2.12:

Obtenga la Mediana para la siguiente distribución correspondiente al número de horas diarias dedicadas a las tareas de hogar de en un conjunto de hogares:

Número de horas (x_i)	**Número de hogares (n_i)**	N_i	$\frac{N}{2} = \frac{22}{2} = 11$ $Me = \frac{2 + 3}{2} = 2{,}5\ horas$
1	6	6	
2	5	11	
3	7	18	
4	4	22	

b) No existe una frecuencia acumulada igual a N/2. La mediana es el valor de la variable que corresponde a la primera frecuencia acumulada inmediatamente superior a N/2.

Ejemplo 2.13:

Obtenga la Mediana para la siguiente distribución correspondiente al número de personas activas en un conjunto de hogares:

Número de personas activas (x_i)	**Número de familias** (n_i)	N_i	$\frac{N}{2} = \frac{20}{2} = 10$ $Me = 2\ personas\ activas$
1	6	6	
2	8	14	
3	4	18	
4	2	20	

3.- Para distribuciones de datos agrupados se busca el intervalo mediano. El intervalo mediano es el primero cuya frecuencia acumulada sea igual o inmediatamente superior a N/2. Una vez identificado el intervalo mediano, la mediana se calcula mediante la siguiente fórmula:

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} * a_i$$

El subíndice i hace referencia a la fila donde se encuentra el intervalo mediano.

Ejemplo 2.14:

Obtenga el Salario Mediano:

Salario € / hora	**Número de trabajadores** (n_i)	N_i	$\frac{N}{2} = \frac{54}{2} = 27$ $Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} * a_i = 55 + \frac{27-21}{14} * 20 = 63{,}57$ €/hora
15 - 35	9	9	
35 - 55	12	21	
55 - 75	14	35	
75 - 95	10	45	
95 - 115	9	54	

Las principales ventajas de la mediana son:

a) Es única.

b) Los valores extremos de la distribución no afectan a la mediana.

c) Es posible su cálculo en el caso de intervalos abiertos, siempre que dicho intervalo no sea el intervalo mediano.

Las propiedades de la mediana respecto a los cambios de origen y escala son las siguientes:

1. La mediana se ve afectada por cambios de origen:

$$x_i' = C + x_i \rightarrow Me' = C + Me$$

2. La mediana se ve afectada por cambios de escala:

$$x_i' = C * x_i \rightarrow Me' = C * Me$$

2.3.2 MODA

La moda es el valor de la variable que se presenta mayor número de veces, o sea, el valor más frecuente. Tiene, por tanto, la misma unidad de medida que la variable.

Para el cálculo de la moda también hay que considerar el tipo de distribución:

1.- Para distribuciones de datos sin agrupar: la moda es aquel valor de la variable que tenga mayor n_i.

Ejemplo 2.15:

Obtenga el valor de la Moda para el número de personas activas por familia:

Número de personas activas (x_i)	**Número de familias** (n_i)	
1	6	$M_o = 2\ personas\ activas$
2	8	
3	4	
4	2	

2.- Para distribuciones con datos agrupados se busca el intervalo modal. Además, en este caso, es necesario distinguir entre distribuciones con intervalos de amplitud constante y distribuciones con intervalos de amplitud variable.

a) Intervalos de amplitud constante:

El intervalo modal es aquel que presenta mayor n_i, y la moda se calcula a través de la siguiente fórmula:

$$Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} * a_i$$

b) Intervalos de amplitud variable:

El intervalo modal es aquel que presenta mayor h_i, y la moda se obtiene aplicando la siguiente fórmula:

$$Mo = L_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} * a_i$$

Ejemplo 2.16:

Obtenga la moda para las siguientes distribuciones de salarios:

a) Intervalos de amplitud constante:

Salario Euros / hora	**Número de trabajadores (n_i)**	a_i	$Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} * a_i =$
15 - 35	9	20	$= 55 + \frac{10}{12 + 10} * 20 = 64{,}1\ €/hora$
35 - 55	12	20	
55 - 75	14	20	
75 - 95	10	20	
95 - 115	9	20	

b) Intervalos de amplitud variable:

Salario Euros / hora Intervalos	**Número de trabajadores (n_i)**	a_i	$h_i = \frac{n_i}{a_i}$	$Mo = L_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} * a_i =$
15 - 45	9	30	0,3	$= 45 + \frac{0{,}35}{0{,}30 + 0{,}35} * 10 = 50{,}38\ €/hora$
45 - 55	12	10	1,2	
55 - 95	14	40	0,35	
95 - 115	9	20	0,45	

La principal ventaja de la moda es que la existencia de valores anormalmente grandes o pequeños de la variable no afectan a su cálculo; mientras que su principal inconveniente es que pueden existir distribuciones con más de una moda. En este último caso, pierde sentido su utilización como promedio.

Las propiedades de la moda respecto a los cambios de origen y escala son las siguientes:

1. La moda se ve afectada por cambios de origen:

$$x_i' = C + x_i \rightarrow Mo' = C + Mo$$

2. La moda se ve afectada por cambios de escala:

$$x_i' = C * x_i \rightarrow Mo' = C * Mo$$

2.4 CUANTILES

Los Cuantiles son valores de la variable que dividen a la distribución en partes iguales, conteniendo cada una de ellas el mismo número de observaciones. Los Cuantiles más utilizadas son los Percentiles (o Centiles), los Cuartiles y los Deciles.

Los percentiles (P_j) son noventa y nueve valores de la distribución, que la dividen en cien intervalos dentro de los cuales se encuentran el 1% de las observaciones de la distribución. Así, por ejemplo, el percentil j (con j=1…99) deja por debajo de sí al j % de las observaciones.

Para distribuciones de datos sin agrupar, su cálculo es idéntico que para la mediana con la única diferencia que en lugar de utilizar $\frac{N}{2}$ se usará $\frac{jN}{100}$.

Ejemplo 2.17:
Para la siguiente distribución de personas activas, obtenga el P_{12} y el P_{67}:

Número de personas activas (x_i)	**Número de familias (n_i)**	**N_i**	$\frac{12N}{100} = \frac{12 * 20}{100} = 2{,}4;\ P_{12} = 1\ persona\ activa$ $\frac{67N}{100} = \frac{67 * 20}{100} = 13{,}40;\ P_{67} = 2\ persona\ activa$
1	6	6	
2	8	14	
3	4	18	
4	2	20	

Para distribuciones con datos agrupados, se busca el intervalo cuya frecuencia absoluta acumulada sea mayor o igual que $\frac{jN}{100}$, y el percentil se calcula mediante la siguiente fórmula:

$$P_j = L_{i-1} + \frac{\frac{jN}{100} - N_{i-1}}{n_i} * a_i$$

con j = 1, 2…99.

Ejemplo 2.18:

Obtenga el P_{37} para la siguiente distribución de salarios:

Salario Euros / hora Intervalos	**Número de trabajadores (n_i)**	N_i
15 - 45	9	9
45 - 55	13	22
55 - 95	12	34
95 - 115	10	44

$$\frac{jN}{100} = \frac{37 * 44}{100} = 16{,}28$$

$$P_{37} = L_{i-1} + \frac{\frac{37N}{100} - N_{i-1}}{n_i} * a_i = 45 + \frac{16{,}28-9}{13} * 10 = 50{,}6\ €/hora$$

Los Cuartiles Q_j son tres valores de la distribución, que la dividen en cuatro intervalos. Dentro de cada intervalo se encuentra el 25% de las observaciones de la distribución. Así, por ejemplo, el cuartil Q_1 deja por debajo de sí al 25% de las observaciones. Existe una equivalencia entre los cuartiles y los percentiles, por lo que para calcular los cuartiles se puede recurrir a los procedimientos aprendidos en el caso de los percentiles. En particular, la equivalencia es la siguiente:

$$Q_1 = P_{25};\ Q_2 = P_{50};\ Q_3 = P_{75}$$

En cuanto a los Deciles D_j son nueve valores que dividen a la distribución en 10 intervalos, que contiene cada uno el 10% de las observaciones. Las equivalencias entre Deciles y los Percentiles son las siguientes:

$$D_1 = P_{10};\ D_2 = P_{20};\ D_3 = P_{30}$$
$$D_4 = P_{40};\ D_5 = P_{50};\ D_6 = P_{60}$$
$$D_7 = P_{70};\ D_8 = P_{80};\ D_9 = P_{90}$$

Anexo 2.A Ejemplos de obtención de medidas de posición con Excel

En este anexo, en primer lugar, se presenta la figura 2.A.1 donde se muestran las operaciones que hay que realizar para obtener la media, la mediana, la moda y los percentiles 25 y 75 haciendo uso de los datos del Ejemplo 2.16 relativos a la distribución del salario con intervalos variables.

Figura 2.A.1. Obtención de promedios para la distribución con intervalos variables del Ejemplo 2.16

- Fórmulas:

	A	B	C	D	E	F	G
1	**Salario**	**Número de trabajadores**	**x_i**	**x_i n_i**	**a_i**	**h_i**	**N_i**
2	**€/ hora**						
3	15 - 45	9	=(15+45)/2	=C3*B3	=45-15	=(B3/E3)	=B3
4	45 - 55	12	=(45+55)/2	=C4*B4	=55-45	=(B4/E4)	=B4+G3
5	55 - 95	14	=(55+95)/2	=C5*B5	=95-55	=(B5/E5)	=B5+G4
6	95 - 115	9	=(95+115)/2	=C6*B6	=115-95	=(B6/E6)	=B6+G5
7							
8	**Promedios**						
9							
10	**Media**	**Mediana**	**Moda**				
11	=SUMA(D3:D73)/SUMA(B3:B6)	=55+((22-G4)*E5)/B5	=45+((F5/(F3+F5))*E4				
12							
13	**Percentil 25**	**Percentil 75**					
14	=45+((11-G3)*E5)/B5	=55+((33-G4)*E5)/B5					

- Resultados:

	A	B	C	D	E	F	G
1	**Salario**	**Número de**	**x_i**	**x_i n_i**	**a_i**	**h_i**	**N_i**
2	**€/ hora**	**trabajadores**					
3	15 - 45	9	30	270	30	0.3	9
4	45 - 55	12	50	600	10	1.2	21
5	55 - 95	14	75	1050	40	0.35	35
6	95 - 115	9	105	945	20	0.45	44
7							
8	**Promedios**						
9							
10	**Media**	**Mediana**	**Moda**				
11	65.11	57.86	50.38				
12							
13	**Percentil 25**	**Percentil 75**					
14	50.71	89.29					

Fuente: Elaboración propia a partir del programa Excel.

En segundo lugar, hay que resaltar que en la actualidad un gran número de la información estadística tabulada se corresponde con estadísticas provenientes de microdatos generados a través de encuestas. En los microdatos, cada elemento (individuo, hogar, empresa…) tiene asignado un identificador. En este contexto, supongamos un ejemplo simulado con 30 individuos donde se les pregunta por su edad. Las respuestas se tabularían en la hoja de cálculo según se muestra en la figura 2.A.2. A partir de dichos datos, se pueden calcular los promedios desarrollados en el tema anterior mediante el uso de funciones de Excel.

El rango de los datos en el ejemplo abarca las casillas comprendidas entre B2 y B31. Con esta información, las funciones correspondientes a cada promedio serían:

a) Media aritmética: =PROMEDIO(B2:B31).

b) Mediana: =MEDIANA(B2:B31).

c) Moda: =MODA.UNO(B2:B31).

d) Percentil j-ésima: =PERCENTIL.INC(B2:B31;j/100). En el ejemplo se calculan los percentiles 25 y 75.

Figura 2.A.2. Obtención de promedios con Excel

- Datos:

	A	B
1	**Identificador**	**Edad**
2	1	18
3	2	20
4	3	21
5	4	35
6	5	41
7	6	65
8	7	52
9	8	19
10	9	20
11	10	22
12	11	45
13	12	57
14	13	22
15	14	24
16	15	23
17	16	31
18	17	56
19	18	61
20	19	62
21	20	19
22	21	22
23	22	27
24	23	28
25	24	55
26	25	21
27	26	23
28	27	24
29	28	25
30	29	27
31	30	29

- Fórmulas:

C
Promedios
Media aritmética
=PROMEDIO(B2:B31)
Mediana
=MEDIANA(B2:B31)
Moda
=MODA.UNO(B2:B31)
Percentil 25
=PERCENTIL.INC(B2:B31;25/100)
Percentil 75
=PERCENTIL.INC(B2:B31;75/100)

- Resultados:

C
Promedios
Media aritmética
33.13
Mediana
26
Moda
22
Percentil 25
22.0
Percentil 75
44.00

Fuente: Elaboración propia a partir del programa Excel.

TEMA 3. MEDIDAS DE DISPERSIÓN Y FORMA

3.1 INTRODUCCIÓN

En el apartado anterior definíamos una serie de medidas de tendencia central cuyo objetivo era sintetizar y representar a la información estadística contenida en una distribución. Los estadísticos de variabilidad o dispersión que vamos a estudiar ahora muestran si los valores de las observaciones están próximos entre sí, están muy separados o cualquier otra situación intermedia.

Medir la representatividad de un promedio equivale a cuantificar la separación de los valores de la distribución respecto a dicha medida. En un principio, se distinguen entre medidas de dispersión absolutas y relativas. En los dos siguientes epígrafes se desarrollan ambas medidas.

3.2 MEDIDAS DE DISPERSION ABSOLUTA

Las medidas de dispersión absoluta son aquellas que dependen de la unidad de medida de la variable. A continuación, se exponen las siguientes medidas:

a) Recorridos.

b) Varianza

c) Desviación típica.

3.2.1 RECORRIDOS

El Recorrido es la diferencia entre el mayor y menor valor de una distribución:

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

Se trata de una primera aproximación a la medida de la dispersión de una distribución. El Recorrido está expresado en las mismas unidades de medida que la variable, y si los datos están agrupados en intervalos el recorrido será la diferencia entre el valor del límite superior del último intervalo y el valor del límite inferior del primer intervalo. En los Ejemplos 3.1 y 3.2 se muestra cómo obtener el Recorrido para una distribución con datos sin agrupar y para una distribución con datos agrupados, respectivamente.

Ejemplo 3.1:

Horas de estudio (x_i)	**Número de alumnos** (n_i)
3	20
4	25
5	14
6	12
9	10
10	9

$$R = x_{max} - x_{min} = 10 - 3 = 7\ horas$$

Ejemplo 3.2:

Salario € / hora Intervalos	**Número de trabajadores** (n_i)
15 - 35	9
35 - 55	12
55 - 75	14
75 - 95	10
95 - 115	9

$$R = x_{max} - X_{min} = 115 - 15 = 100\ €/\text{hora}$$

La principal ventaja del Recorrido como medida de variabilidad es su facilidad de cálculo, mientras que su principal inconveniente es que le afecta los valores extremos de la distribución.

Otros tipos de Recorridos son los Recorridos intercuantílicos que se obtienen a partir de los cuantiles. En concreto, se pueden definir los siguientes Recorridos:

a) Recorrido interpercentílico:

Es la diferencia entre el último y el primer percentil, y contiene el 98% de las observaciones:

$$R_c = P_{99} - P_1$$

b) Recorrido interdecílico;

Es la diferencia entre el último y el primer decil, e incluye al 80% de las observaciones:

$$R_D = D_{90} - D_{10} = P_{90} - P_{10}$$

c) Recorrido intercuartílico:

Es la diferencia entre el último y el primer cuartil. Por tanto, recoge al 50% de las observaciones:

$$R_Q = Q_3 - Q_1 = P_{75} - P_{25}$$

3.2.2 VARIANZA

La Varianza es la media aritmética de los cuadrados de las desviaciones de los valores observados con respecto a la media aritmética de la distribución:

$$S^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{X})^2 n_i}{N}$$

La unidad de medida de la Varianza es la unidad de medida de la variable al cuadrado.

A partir de la anterior fórmula se obtiene otra que es la que habitualmente se utiliza:

$$S^2 = \frac{\sum x_i^2 n_i}{N} - \bar{X}^2$$

Demostración:

$$S^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{X})^2 n_i}{N} = \frac{\sum x_i^2 n_i + \sum \bar{X}^2 n_i - 2\sum x_i \bar{X} n_i}{N} =$$

$$= \frac{\sum x_i^2 n_i + N\bar{X}^2 - 2N\bar{X}^2}{N} = \frac{\sum x_i^2 n_i}{N} - \bar{X}^2$$

Ejemplo 3.3:

Obtenga la Varianza a partir de la siguiente distribución de frecuencias de horas de estudio:

Horas de estudio (x_i)	$\boldsymbol{x_i^2}$	$\bar{X} = \frac{\sum x_i n_i}{N} = \frac{37}{6} = 6{,}17\ horas$ $S^2 = \frac{\sum x_i^2}{N} - \bar{X}^2 = \frac{267}{6} - (6{,}17)^2 = 6{,}43\ horas^2$
3	9	
4	16	
5	25	
6	36	
9	81	
10	100	

Ejemplo 3.4:

Obtenga la varianza a partir de la siguiente distribución de frecuencias de número de personas activas en una familia:

Número de personas activas (x_i)	**Número de familias** (n_i)	$x_i n_i$	x_i^2	$x_i^2 n_i$	$\bar{X} = \frac{\sum x_i n_i}{N} = \frac{24}{20} = 2{,}1\ personas\ activas$
1	6	6	1	6	$S^2 = \frac{\sum x_i^2 n_i}{N} - \bar{X}^2 = \frac{106}{20} - (2{,}1)^2 =$
2	8	16	4	32	$= 0{,}89\ (personas\ activas)^2$
3	4	12	9	36	
4	2	8	16	32	

Ejemplo 3.5:

Obtenga la varianza para la siguiente distribución de salarios:

Salario Euros / hora Intervalos	**Número de trabajadores** (n_i)	x_i	$x_i n_i$	x_i^2	$x_i^2 n_i$	$\bar{X} = \frac{\sum x_i n_i}{N} = \frac{3470}{54} = 64{,}26\ €/hora$
15 - 35	9	25	225	625	5625	
35 - 55	12	45	540	2025	24300	$S^2 = \frac{\sum x_i^2 n_i}{N} - \bar{X}^2 = \frac{260550}{54} - (64{,}26)^2 =$
55 - 75	14	65	910	4225	59150	$= 695{,}65\ (€/hora)^2$
75 - 95	10	85	850	7225	72250	
95 - 115	9	105	945	11025	99225	

La principal ventaja de la varianza es que utiliza toda la información de la distribución para su cálculo, mientras que sus principales inconvenientes son los siguientes:

a) No sirve para comparar dispersiones en distintas distribuciones que posean distintas medias o bien vengan expresadas en distintas unidades de medida.

b) No puede obtenerse en distribuciones con intervalos abiertos, a no ser que se proporcione información complementaria.

c) Al igual que la media aritmética, su valor depende del criterio de agrupación en intervalos.

d) Es muy sensible a los valores anormales de la variable.

Las propiedades de la varianza son las siguientes:

1. No le afecta el cambio de origen.

Si en la distribución de frecuencias se suman a todos los valores de la variable una constante, la varianza no varía:

$$x_i' = C + x_i \rightarrow (S^2)' = S^2$$

Demostración:

$$(S^2)' = \frac{\sum(x_i' - \bar{X}')^2 n_i}{N} = \frac{\sum((C+x_i) - (C + \bar{X}))^2 n_i}{N} = \frac{\sum(x_i - \bar{X})^2 n_i}{N} = S^2$$

2. Sí le afecta el cambio de escala.

Si en la distribución de frecuencias se multiplican todos los valores de la variable por una constante, la varianza queda multiplicada por el cuadrado de dicha constante:

$$x_i' = C * x_i \rightarrow (S^2)' = C^2 * S^2$$

Demostración:

$$S^{2\prime} = \frac{\sum(x_i' - \bar{X}')^2 n_i}{N} = \frac{\sum(C * x_i - C * \bar{X})^2 n_i}{N} = C^2 \frac{\sum(x_i - \bar{X})^2 n_i}{N} = C^2 * S^2$$

3. La varianza es no negativa, y valdrá cero cuando todos los valores de la variable sean iguales entre sí. $S^2 \geq 0$.

La varianza está expresada en las unidades de la variable elevada al cuadrado, lo cual dificulta su interpretación. Por ese motivo, se define la desviación típica o estándar.

3.2.3 DESVIACIÓN TÍPICA

La Desviación Típica es la raíz cuadrada de la varianza:

$$S = \sqrt{S^2}$$

Está expresada en las mismas unidades de la variable. Se conviene en tomar el valor positivo de la raíz, por lo cual $S \geq 0$.

Ejemplo 3.6:

Obtenga la desviación típica a partir de la siguiente distribución de frecuencias de horas de estudio:

Horas de estudio (x_i)	x_i^2	$\bar{X} = \frac{\sum x_i n_i}{N} = \frac{37}{6} = 6{,}17\ horas$
3	9	$S^2 = \frac{\sum x_i^2}{N} - \bar{X}^2 = \frac{267}{6} - (6{,}17)^2 = 6{,}43\ horas^2$
4	16	
5	25	
6	36	$S = \sqrt{\frac{\sum x_i^2}{N} - \bar{X}^2} = \sqrt{6{,}43} = 2{,}54\ horas$
9	81	
10	100	

Ejemplo 3.7:

Obtenga la desviación típica a partir de la siguiente distribución de frecuencias de número de personas activas:

Número de personas activas x_i	**Número de familias** n_i	$x_i n_i$	x_i^2	$x_i^2 n_i$	$\bar{X} = \frac{\sum x_i n_i}{N} = \frac{24}{20} = 2{,}1\ personas\ activas$ $S^2 = \frac{\sum x_i^2 n_i}{N} - \bar{X}^2 = \frac{106}{20} - (2{,}1)^2 = 0{,}89\ (personas\ activas)^2$ $S = \sqrt{\frac{\sum x_i^2 n_i}{N} - \bar{X}^2} = \sqrt{0{,}89} = 0{,}94\ personas\ activas$
1	6	6	1	6	
2	8	16	4	32	
3	4	12	9	36	
4	2	8	16	32	

Ejemplo 3.8:

Obtenga la desviación típica a partir de la siguiente distribución de frecuencias de salarios:

Salario Euros / hora Intervalos	**Número de trabajadores n_i**	x_i	$x_i n_i$	x_i^2	$x_i^2 n_i$
15 - 35	9	25	225	625	5625
35 - 55	12	45	540	2025	24300
55 - 75	14	65	910	4225	59150
75 - 95	10	85	850	7225	72250
95 - 115	9	105	945	11025	99225

$$\bar{X} = \frac{\sum x_i n_i}{N} = \frac{3470}{54} = 64{,}26\ €/hora$$

$$S^2 = \frac{\sum x_i^2 n_i}{N} - \bar{X}^2 = \frac{260550}{54} - (64{,}26)^2 = 695{,}65\ (€/hora)^2$$

$$S = \sqrt{\frac{\sum x_i^2 n_i}{N} - \bar{X}^2} = \sqrt{695{,}65} = 26{,}375\ €/hora$$

Las propiedades de la Desviación Típica respecto a los cambios de origen y escala son las siguientes:

1.-No le afecta el cambio de origen.

Si en la distribución de frecuencias sumamos a todos los valores de la variable una constante, la Desviación Típica no varía:

$$x_i' = C + x_i \rightarrow S' = S$$

2.-Sí le afecta el cambio de escala.

Si en la distribución de frecuencias multiplicamos todos los valores de la variable por una constante, la Desviación Típica queda multiplicada por el valor absoluto de dicha constante. Los cambios de escala de las variables en Ciencias Sociales suelen ser positivos, por lo que el valor absoluto de la constante coincidirá con la constante:

$$x_i' = C * x_i \rightarrow S' = C * S$$

3.3 MEDIDAS DE DISPERSIÓN RELATIVA

La varianza y la desviación típica son medidas de dispersión absoluta. Eso implica que no pueden usarse para comparar dispersiones de dos distribuciones con distintas unidades de medida, o bien, si aun teniendo la misma unidad de medida, las medias aritméticas son diferentes. Por ello, es necesario hacer uso de medidas de dispersión relativas, una de las más utilizadas es el Coeficiente de Variación de Pearson, que se define como el cociente entre la desviación típica y la media aritmética:

$$CV = \frac{S}{\bar{X}}$$

El CV representa el número de veces que la desviación típica contiene a la media, luego a mayor valor del CV menor representatividad de la media. El CV es adimensional (es decir, carece de unidades de medida), y alcanza su valor mínimo en 0 cuando la desviación típica es cero.

El CV tiene dos utilidades:

1.- Determinar si una media aritmética es representativa de su distribución. Para ello se usan los siguientes límites:

a) Si $CV < 0{,}5$, la media aritmética es representativa.

b) Si $0{,}5 \leq CV < 1$, hay que cuestionarse la representatividad de la media aritmética.

c) Si $CV \geq 1$, la media no es representativa.

2. Comparar la representatividad de la media aritmética de una distribución con respecto a otra.

Ejemplo 3.9:

Se tiene la siguiente distribución de la renta mensual en cientos de euros de las familias de dos distritos de una ciudad:

Distrito A		**Distrito B**	
Renta (10^2 €)	**Número de familias**	**Renta (10^2 €)**	**Número de familias**
5-10	90	5-10	10
10-20	50	10-20	80
20-30	30	20-30	70
30-40	25	30-40	25
40-50	20	40-50	10
50-100	5	50-100	5

Indique cuál de los dos distritos tiene una media más representativa, el distrito A o el distrito B.

Distrito A

Renta (10^2 €)	$\mathbf{n_i}$	$\mathbf{x_i}$	$\mathbf{x_i n_i}$	$\mathbf{x_i^2}$	$\mathbf{x_i^2 n_i}$
5-10	90	7,5	675	56,25	5062,5
10-20	50	15	750	225	11250
20-30	30	25	750	625	18750
30-40	25	35	875	1225	30625
40-50	20	45	900	2025	40500
50-100	5	75	375	5625	28125

$$\bar{X}_A = \frac{\sum x_i n_i}{N} = \frac{4325}{220} = 19{,}66 * 10^2 €$$

$$S_A^2 = \frac{\sum x_i^2 n_i}{N} - \bar{X}_A^2 = \frac{134312{,}5}{220} - 19{,}66^2 = 223{,}996 * 10^4 €^2$$

$$S_A = \sqrt{S_A^2} = 14{,}97 * 10^2 €$$

$$CV_A = \frac{S_A}{\bar{X}_A} = \frac{14{,}97}{19{,}66} = 0{,}76$$

Distrito B

Renta (10^2 €)	n_i	x_i	$x_i n_i$	x_i^2	$x_i^2 n_i$
5-10	10	7,5	75	56,25	562,5
10-20	80	15	1200	225	18000
20-30	70	25	1750	625	43750
30-40	25	35	875	1225	30625
40-50	10	45	450	2025	20250
50-100	5	75	375	5625	28125

$$\bar{X}_B = \frac{\sum x_i n_i}{N} = \frac{4725}{200} = 23{,}625 * 10^2€$$

$$S_B^2 = \frac{\sum x_i^2 n_i}{N} - \bar{X}_B^2 = \frac{141312{,}5}{200} - 23{,}625^5 = 148{,}42 * 10^4€^2$$

$$S_B = \sqrt{S_B^2} = 12{,}18 * 10^2€$$

$$CV_B = \frac{S_B}{\bar{X}_B} = \frac{12{,}18}{23{,}62} = 0{,}52$$

$CV_A > CV_B$, lo que implica que la media de la distribución B es más representativa que la media de la distribución A; es decir, la distribución B es menos dispersa en términos relativos que la distribución A. En cualquier caso, ambos coeficientes de variación son mayores que 0,5, luego en ambas distribuciones nos cuestionamos la representatividad de la media.

Las propiedades del Coeficiente de Variación respecto a los cambios de origen y escala son las siguientes:

1. Le afecta los cambios de origen:

Si en la distribución de frecuencias se suman a todos los valores de la variable una constante, el Coeficiente de Variación varía.

Demostración:

$$x_i' = x_i + C \rightarrow \begin{cases} \bar{X}' = \bar{X} + C \\ S' = S \end{cases} \rightarrow CV' = \frac{S}{\bar{X} + C}$$

Si C es un número positivo la representatividad de la media aritmética aumenta; mientras que disminuye si C es negativa.

2. No le afecta los cambios de escala:

Si en la distribución de frecuencias se multiplican todos los valores de la variable por una constante positiva, el Coeficiente de Variación no varía.

Demostración:

$$x_i' = x_i * C \rightarrow \begin{cases} \bar{X}' = C * \bar{X} \\ S' = C * S \end{cases} \rightarrow CV' = \frac{C * S}{C * \bar{X}} = CV$$

3.4 TIPIFICACIÓN DE VARIABLES

Una variable X, de media $\bar{X}$ y varianza S_x^2 queda tipificada si se le aplica la siguiente transformación:

$$Z = \frac{X - \bar{X}}{S_x}$$

De esta manera, cada valor concreto de la nueva variable será:

$$z_i = \frac{x_i - \bar{X}}{S_x}$$

La variable tipificada procede de la variable X, a la que se le ha aplicado simultáneamente un cambio de origen (se le ha restado la media $\bar{X}$) y un cambio de escala (se ha multiplicado por $\frac{1}{S_x}$). La nueva variable, Z, es adimensional, es decir, carece de unidades de medida. Además, se cumple que:

a) La media de una variable tipificada es cero:

$$\bar{Z} = 0$$

Demostración:

$$\bar{Z} = \frac{\sum z_i n_i}{N} = \frac{\sum \left(\frac{x_i - \bar{X}}{S_x}\right) n_i}{N} = \frac{1}{S_x} \frac{\sum (x_i - \bar{X}) n_i}{N} = \frac{1}{S_x} * \frac{0}{N} = 0$$

b) La varianza de una variable tipificada es 1:

$$S_z^2 = 1$$

Demostración:

$$S_z^2 = \frac{\sum z_i^2 n_i}{N} - \bar{Z}^2 = \frac{\sum \left(\frac{x_i - \bar{X}}{S_x}\right)^2 n_i}{N} = \frac{1}{S_x^2}\frac{\sum (x_i - \bar{X})^2 n_i}{N} = \frac{1}{S_x^2} * S_x^2 = 1$$

La utilidad de la variable tipificada es que permite comparar observaciones individuales correspondientes a distribuciones diferentes.

Ejemplo 3.10:

Haciendo uso de los resultados del Ejemplo 3.9, si un individuo del distrito A tiene una renta de 2500 euros y otro del distrito B tiene una renta de 2700 euros, ¿cuál está en mejor posición relativa dentro de su distrito?

$$z_A = \frac{x_A - \bar{X}_A}{S_A} = \frac{25 - 19{,}66}{14{,}97} = 0{,}3567$$

$$z_B = \frac{x_B - \bar{X}_B}{S_B} = \frac{27 - 23{,}625}{12{,}18} = 0{,}2771$$

$z_A > z_B$, el individuo del distrito A está en mejor posición relativa.

3.5 MEDIDAS DE FORMA: ASIMETRÍA Y CURTOSIS

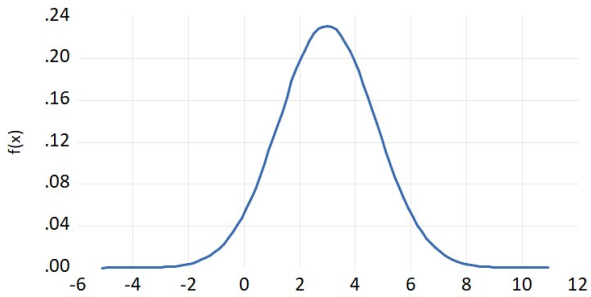
3.5.1 ASIMETRÍA

Una distribución es simétrica cuando a la derecha y a la izquierda de la media aritmética existen el mismo número de observaciones. En caso contrario, la distribución es asimétrica. Será asimétrica positiva o hacia la derecha si las frecuencias más altas se encuentran en el lado izquierdo de la distribución, mientras que en el derecho hay frecuencias más pequeñas. Presentará asimetría negativa o hacia la izquierda en caso contrario.

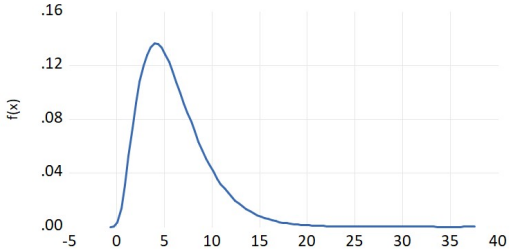
La Asimetría puede estudiarse de forma gráfica o analítica (es decir, a través de medidas estadísticas). En la figura 3.1 se muestra un ejemplo de los tipos de asimetría.

Figura 3.1. Ejemplos de tipos de asimetría

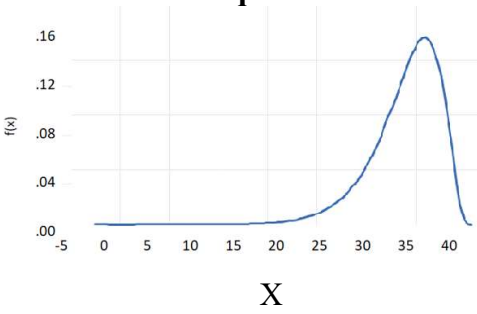
Distribución simétrica



Distribución con asimetría positiva o hacia la derecha



Distribución con asimetría negativa o hacia la izquierda



Fuente: Elaboración propia.

En distribuciones campaniformes unimodales, existe una relación entre los valores que toman la media aritmética $\bar{X}$ y la moda M_o, y el tipo de asimetría de la distribución. En concreto, ocurre lo siguiente:

a) Cuando la distribución es simétrica: $\bar{X} = M_o$.

b) Si la distribución es asimétrica a la derecha: $\bar{X} > M_o$.

c) Si la distribución es asimétrica a la izquierda: $\bar{X} < M_o$.

A partir de estas relaciones podemos usar una medida de asimetría relativa conocida como Coeficiente de Asimetría de Pearson:

$$A_p = \frac{\bar{X} - M_o}{S}$$

El Coeficiente de asimetría de Pearson A_p es adimensional y a través de su valor podemos deducir lo siguiente:

a) Si $A_p > 0$, existe asimetría hacia la derecha.

b) Si $A_p < 0$, existe asimetría hacia la izquierda.

Otra medida de asimetría es el Coeficiente de Asimetría de Fisher, que se calcula con la siguiente expresión:

$$\gamma_1 = \frac{m_3}{S^3} = \frac{\frac{\sum(x_i - \bar{X})^3 n_i}{N}}{S^3}$$

donde m_3 es el momento de orden 3 con respecto a la media. Los momentos son valores que caracterizan a la distribución, y pueden ser con respecto al origen o con respecto a la media. Sus fórmulas de cálculo aparecen en la tabla 3.1.

Tabla 3.1 Momentos de una distribución de frecuencias

Momentos con respecto al origen	**Momentos con respecto a la media**
$a_r = \frac{\sum_{i=1}^{k} x_i^r n_i}{N}$, r=0,1,2…	$m_r = \frac{\sum_{i=1}^{k} (x_i - \bar{X})^r n_i}{N}$, r=0,1,2…
Ejemplo de momento con respecto al origen $a_1 = \bar{X}$	**Ejemplo de momento con respecto a la media** $m_2 = S^2$

Fuente: Elaboración propia.

El coeficiente de asimetría de Fisher γ_1 es adimensional, y a partir de sus valores podemos deducir el tipo de distribución de la variable:

a) Si $\gamma_1 > 0$ entonces la distribución es asimétrica a la derecha.

b) Si $\gamma_1 < 0$ entonces la distribución es asimétrica a la izquierda.

Hay que señalar que tanto el coeficiente de asimetría de Pearson como el de Fisher no se ven alterados por los cambios de origen y/o escala que se produzcan en la variable.

Ejemplo 3.10:

Estudie la asimetría de la siguiente distribución relativa al número de personas activas de los hogares:

Número de personas activas (x_i)	**Número de familias** (n_i)	$x_i n_i$	$x_i^2 n_i$	$(x_i - \bar{X})^3 n_i$
1	6	6	6	-7,99
2	8	16	32	-0,01
3	4	12	36	2,92
4	2	8	32	13,72

a) Coeficiente de asimetría de Pearson:

$$\bar{X} = \frac{\sum x_i n_i}{N} = \frac{42}{20} = 2{,}1 \; personas \; activas$$

$$Mo = 2 \; personas \; activas$$

$$S = \sqrt{\frac{\sum x_i^2 n_i}{N} - \bar{X}^2} = \sqrt{\frac{106}{20} - 2{,}1^2} = 0{,}94 \; personas \; activas$$

$$A_p = \frac{\bar{X} - M_o}{S} = \frac{2{,}1 - 2}{0{,}94} = 0{,}106$$

$A_p > 0$, existe asimetría positiva o hacia la derecha.

b) Coeficiente de asimetría de Fisher:

$$\gamma_1 = \frac{m_3}{S^3} = \frac{0{,}432}{0{,}94^3} = 0{,}52$$

$\gamma_1 > 0$, existe asimetría positiva o hacia la derecha.

Ejemplo 3.11:

Estudie la asimetría de la siguiente distribución relativa al salario de un conjunto de trabajadores:

Salario Euros / hora Intervalos	**Número de trabajadores (n_i)**	**N_i**	$\boldsymbol{h_i = \frac{n_i}{a_i}}$	**x_i**	$\boldsymbol{x_i n_i}$	$\boldsymbol{x_i^2 n_i}$	$\boldsymbol{(x_i - \bar{X})^3 n_i}$
15 - 45	9	9	0,3	30	270	8.100	-389524,696
45 - 55	12	21	1,2	50	600	30.000	-41397,55
55 - 95	14	35	0,35	75	1050	78.750	13543,0634
95 - 115	9	44	0,45	105	945	99.225	571261,056

a) Coeficiente de asimetría de Pearson:

$$\bar{X} = \frac{\sum x_i n_i}{N} = \frac{2865}{44} = 65{,}11\ €/h$$

$$S^2 = \frac{\sum x_i^2 n_i}{N} - \bar{X}^2 = 671{,}483\ (€/h)^2$$

$$S = \sqrt{671{,}483} = 25{,}91\ €/h$$

$$Mo = L_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} * a_i = 45 + \frac{0{,}35}{0{,}35 + 0{,}3} * 10 = 50{,}38\ €/hora$$

$$A_p = \frac{\bar{X} - M_o}{S} = 0{,}57$$

Existe asimetría positiva o hacia la derecha, ya que $A_p > 0$.

b) Coeficiente de asimetría de Fisher:

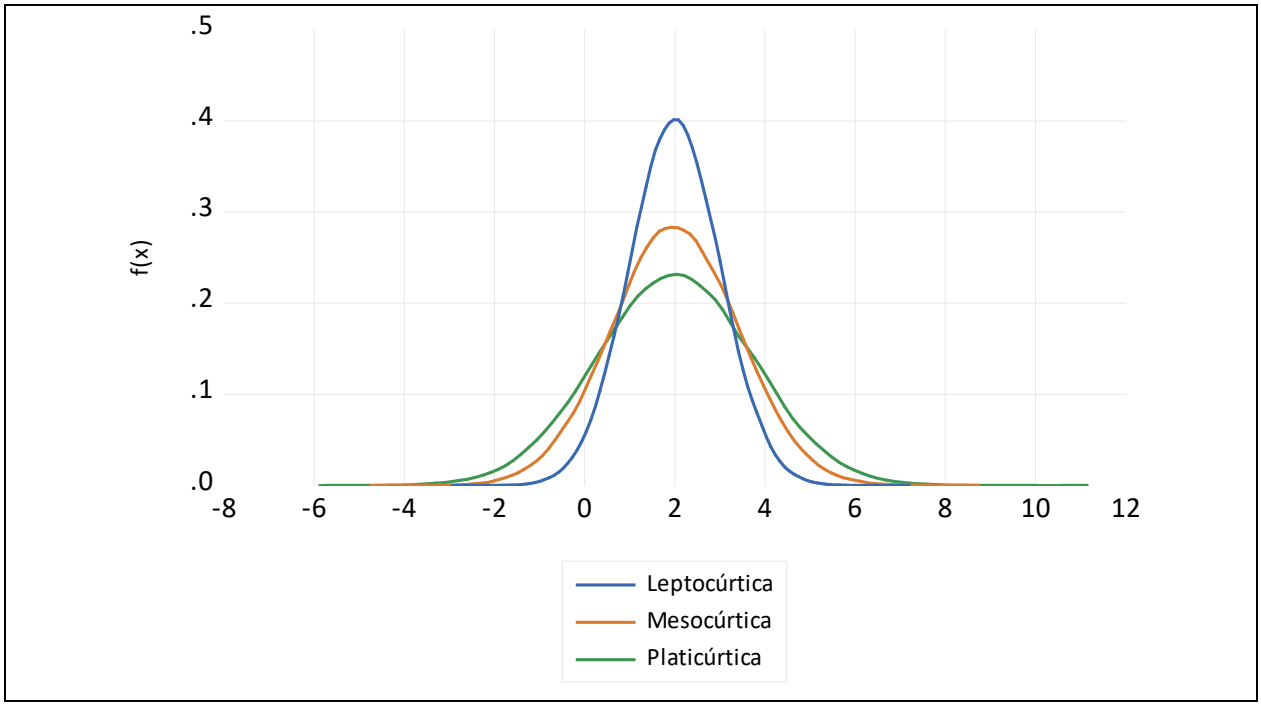
$$\gamma_1 = \frac{m_3}{S^3} = \frac{3497{,}31}{25{,}91^3} = 0{,}2$$

Existe asimetría positiva o hacia la derecha, ya que $\gamma_1 > 0$.

3.5.2 CURTOSIS

La curtosis indica si la distribución es más o menos apuntada en la zona central. En la figura 3.2 aparecen las distintas situaciones que pueden encontrarse.

Figura 3.2. Ejemplos de tipos de curtosis



Fuente: Elaboración propia.

Para medir el apuntamiento se utiliza el Coeficiente de Apuntamiento de Fisher:

$$\gamma_2 = \frac{m_4}{S^4} - 3$$

donde m_4 es el momento de orden 4 con respecto a la media.

El Coeficiente de Apuntamiento de Fisher es adimensional, es decir, carece de unidad de medida. En función de sus valores podemos deducir lo siguiente:

a) Si $\gamma_2 > 0$, la distribución es leptocúrtica o apuntada.

b) Si $\gamma_2 < 0$, la distribución es platicúrtica o aplastada.

c) Si $\gamma_2 = 0$, la distribución es mesocúrtica o normal.

El Coeficiente de Apuntamiento de Fisher no le afecta los cambios de origen y/o escala de la variable.

Ejemplo 3.12:

Estudie el apuntamiento de la siguiente distribución:

Número de personas activas (x_i)	**Número de familias** (n_i)	$x_i n_i$	$x_i^2 n_i$	$(x_i - \bar{X})^4 n_i$
1	6	6	6	8,78
2	8	16	32	0
3	4	12	36	2,62
4	2	8	32	26,06

$$\bar{X} = \frac{\sum x_i n_i}{N} = \frac{42}{20} = 2{,}1\ personas\ activas$$

$$m_4 = \frac{\sum (x_i - \bar{X})^4 n_i}{N} = \frac{37{,}46}{20} = 1{,}87$$

$$S^2 = \frac{\sum x_i^2 n_i}{N} - \bar{X}^2 = \frac{106}{20} - (2{,}1)^2 = 0{,}89\ personas\ activas^2\ ;\ \ S = \sqrt{S^2} = 0{,}94\ personas\ activas$$

$$\gamma_2 = \frac{m_4}{S^4} - 3 = \frac{1{,}87}{0{,}94^2} - 3 = -0{,}60.$$

Como $\gamma_2 < 0$, la distribución es platicúrtica.

Ejemplo 3.13:

Estudie el apuntamiento de la siguiente distribución:

Salarios Euros / hora Intervalos	**Número de trabajadores** n_i	x_i	$x_i n_i$	$x_i^2 n_i$	$(x_i - \bar{X})^4 n_i$
15 - 45	9	30	270	8.100	13676212,1
45 - 55	12	50	600	30.000	625516,98
55 - 95	14	75	1050	78.750	133940,897
95 - 115	9	105	945	99.225	22787603,5

$$\bar{X} = \frac{\sum x_i n_i}{N} = \frac{2865}{44} = 65{,}11\ €/h;\ \ S^2 = \frac{\sum x_i^2 n_i}{N} - \bar{X}^2 = 671{,}483\ (€/h)^2\ ;\ S = \sqrt{671{,}483} = 25{,}91\ €/h$$

$$m_4 = \frac{\sum (x_i - \bar{X})^4 n_i}{N} = 845983{,}489;\ \ \gamma_2 = \frac{m_4}{S^4} - 3 = -1{,}12$$

Como γ_2<0, la distribución es platicúrtica.

Anexo 3.A Ejemplo de obtención de medidas de dispersión y de forma con Excel

En este anexo 3.A se muestra las fórmulas que hay que aplicar para obtener las medidas dispersión y medidas de forma expuestas en el tema 3, suponiendo que los datos están tabulados de forma que aparezca una observación para cada elemento. Además, se indica cómo se construiría una variable tipificada. Las expresiones algebraicas necesarias en Excel para cada uno de los conceptos explicados y adaptadas al ejemplo usado en la figura 2.A.2 del anexo 2.A son:

a) Medidas de dispersión absolutas.

1. Recorrido: =MAX(B2:B31)-MIN(B2:B31)

2. Recorrido interpercentílico: =PERCENTIL.INC(B2:B31;0.99)-PERCENTIL.INC(B2:B31;0.01)

3. Recorrido interdecílico: =PERCENTIL.INC(B2:B31;0.9)-PERCENTIL.INC(B2:B31;0.1)

4. Recorrido intercuartílico: =PERCENTIL.INC(B2:B31;0.75)-PERCENTIL.INC(B2:B31;0.25)

5. Varianza: =VAR.P(B2:B31)

6. Desviación estándar: =DESVEST.P(B2:B31)

b) Medidas de dispersión relativas.

Coeficiente de variación: =(DESVEST.P(B2:B31)/ PROMEDIO(B2:B31))

c) Medidas de forma.

1. Coeficiente de asimetría de Pearson:

=(PROMEDIO(B2:B31)-MODA.UNO(B2:B31))/(DESVEST.P(B2:B31))

2. Coeficiente de asimetría de Fisher: =COEFICIENTE.ASIMETRIA.P(B2:B31)

3. Coeficiente de curtosis de Fisher: =CURTOSIS(B2:B31)

Estos resultados estadísticos se ofrecen en la figura 3.A.1 junto a una columna que muestra los valores obtenidos tras tipificar la variable edad: $= \left(\frac{Valor - \text{DESVEST.P(B2:B31)}}{\text{PROMEDIO(B2:B31)}}\right)$.

Figura 3.A.1. Obtención de medidas de dispersión con Excel

	A	B	C	D	E	F	G	H
1	**Identificador**	**Edad**		**Medidas de dispersión absoluta**				**Variable tipificad**
2	1	18						-0.99359008
3	2	20						-0.86227861
4	3	21		**Recorrido**				-0.79662288
5	4	35		47				0.12255737
6	5	41		**Recorrido interpercentílico**				0.51649176
7	6	65		45.84				2.09222933
8	7	52		**Recorrido interdecílico**				1.23870481
9	8	19		37.5				-0.92793435
10	9	20		**Recorrio intercuartílico**				-0.86227861
11	10	22		22				-0.73096715
12	11	45		**Varianza**				0.77911469
13	12	57		231.982222				1.56698347
14	13	22						-0.73096715
15	14	24		**Desvición Estándar**				-0.59965569
16	15	23		15.2309626				-0.66531142
17	16	31						-0.14006556
18	17	56		**Medida de dispersión relativa**				1.50132774
19	18	61						1.8296064
20	19	62		**Coeficiente de variación de Pearson**				1.89526213
21	20	19		0.460				-0.92793435
22	21	22						-0.73096715
23	22	27		**Medidas de forma**				-0.40268849
24	23	28						-0.33703276
25	24	55		**Coeficiente de asimetría de Pearson**				1.43567201
26	25	21		0.73096715				-0.79662288
27	26	23		**Coeficiente de asimetría de Fisher**				-0.66531142
28	27	24		0.9128935				-0.59965569
29	28	25		**Coeficiente de curtosis de Fisher**				-0.53399995
30	29	27		-0.66367899				-0.40268849
31	30	29						-0.27137703

Fuente: Elaboración propia a partir del programa Excel.

TEMA 4. MEDIDAS DESIGUALDAD

4.1 INTRODUCCIÓN

Las medidas de concentración ponen de manifiesto el mayor o menor grado de igualdad o equidistribución en el reparto del total de los valores de la variable (por ejemplo, rentas o salarios).

A partir de una distribución de rentas con N individuos, cuyas rentas son $x_1 \leq x_2 \leq \cdots \leq x_N$, es posible estudiar hasta qué punto la suma total de rentas $\sum_{i=1}^{N} x_i$ está equitativamente repartida.

Las dos situaciones extremas son:

a) Concentración máxima.

De los N individuos, sólo uno percibe el total de la renta y los demás nada:

$$x_1 = x_2 = \cdots = x_{N-1} = 0 \quad \text{y} \quad x_N \neq 0$$

b) Concentración mínima o equidistribución.

Todos los individuos perciben la misma cantidad:

$$x_1 = x_2 = \cdots = x_N$$

Evidentemente, las dos situaciones anteriores son casos extremos, y la mayoría de las distribuciones van a presentar un grado de concentración que se sitúa entre ambos casos extremos. En este sentido, el objetivo es proporcionar instrumentos que nos indiquen cómo es la desigualdad o la concentración. En los dos epígrafes siguientes se presentan un instrumento gráfico que se denomina Curva de Lorenz, y una medida cuantitativa que se conoce como Indice de Gini.

4.2 CURVA DE LORENZ

Para la obtención de la Curva de Lorenz, en primer lugar, es necesario que los valores de la variable x_i estén ordenados de menor a mayor (si la estadística es con datos agrupados, los cálculos se hacen con la marca de clase). A continuación, se forman las siguientes columnas:

a) Los productos $x_i n_i$ son las rentas percibidas por los n_i rentistas.

b) Las frecuencias acumuladas: N_i.

c) Los totales acumulados V_i, que contienen la renta total percibida por los N_i rentistas. Estos totales V_i se calculan como:

$V_1 = x_1 n_1$

$V_2 = x_1 n_1 + x_2 n_2$

…

$$V_k = x_1 n_1 + x_2 n_2 + \cdots + x_k n_k = \sum_{i=1}^{k} x_i n_i$$

V_k también se expresa como V.

d) La columna de frecuencias relativas acumuladas porcentuales:

$$P_i = \frac{N_i}{N} * 100$$

e) Finalmente, expresaremos cada V_i en porcentajes de V y los denominamos como Q_i:

$$Q_i = \frac{V_i}{V} * 100$$

A partir de las anteriores columnas se puede elaborar la siguiente tabla 4.1.

Tabla 4.1. Obtención de estadísticas derivadas para cálculo de Curva de Lorenz

Variable	Marca de clase	Frec. Abs.	Renta total	Frec acum.	Totales renta acumulados	Frecuencias relativas porcentuales acumuladas	Totales de renta acumulada relativos (en porcentaje)
$\boldsymbol{L_{i-1} - L_i}$	$\boldsymbol{x_i}$	$\boldsymbol{n_i}$	$\boldsymbol{x_i n_i}$	$\boldsymbol{N_i}$	$\boldsymbol{V_i}$	$P_i = \frac{N_i}{N} * 100$	$Q_i = \frac{V_i}{V} * 100$
$L_o - L_1$	x_1	n_1	$x_1 n_1$	N_1	$V_1 = x_1 n_1$	P_1	Q_1
$L_1 - L_2$	x_2	n_2	$x_2 n_2$	N_2	$V_2 = x_1 n_1 + x_2 n_2$	P_2	Q_2
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
$L_{k-1} - L_k$	x_k	n_k	$x_k n_k$	N_k	$V_k = x_1 n_1 + x_2 n_2 + \cdots + x_k n_k$	$P_k = 100$	$Q_k = 100$

Fuente: Elaboración propia.

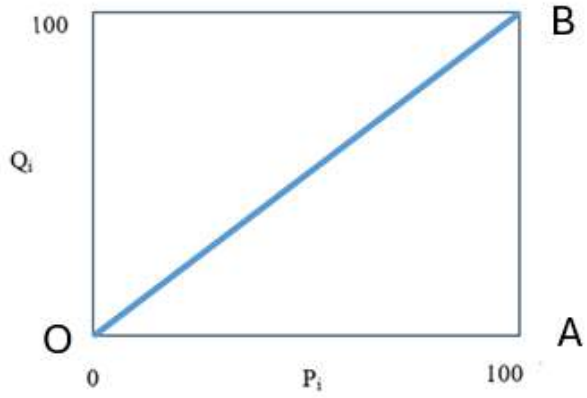
La Curva de Lorenz resulta de la representación de los puntos (P_i, Q_i) en unos ejes de coordenadas y unirlos entre sí. Para la representación, P_i se pone en el eje de abscisas y Q_i en el eje de ordenadas. Tanto el eje de abscisas como el de ordenadas tienen un recorrido de 0 a 100.

La curva de Lorenz puede presentar dos casos extremos:

a) Concentración mínima.

Si, $P_i = Q_i$, la curva de Lorenz coincidirá con la diagonal "OB" del cuadrado, y significaría que hay concentración mínima o equidistribución (figura 4.1).

Figura 4.1. Curva de Lorenz: Caso de concentración mínima

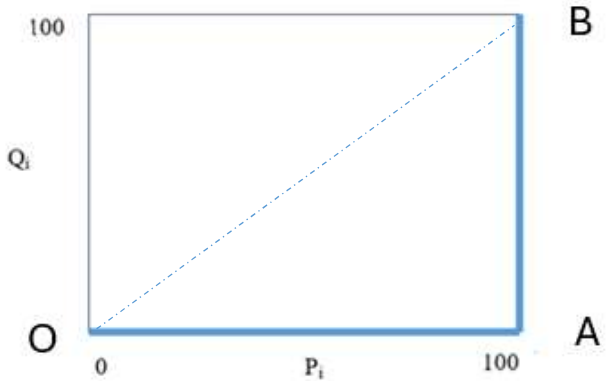


Fuente: Elaboración propia.

b) Concentración máxima.

Todos los individuos tienen renta igual a cero, excepto uno que recibe toda la renta. En este caso, el gráfico se corresponde con la figura 4.2.

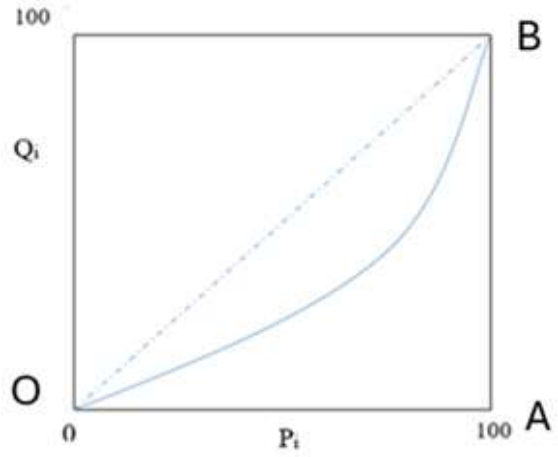
Figura 4.2. Curva de Lorenz: Caso de concentración máxima



Fuente: Elaboración propia

El resto de casos serán situaciones intermedias como en la figura 4.3. Cuanto más próxima esté la curva a la diagonal OB, más equitativa será la distribución:

Figura 4.3. Curva de Lorenz: Situaciones intermedias



Fuente: Elaboración propia

4.3 INDICE DE GINI

El Índice de Gini (I_G) es el cociente entre el área formada entre la curva de Lorenz con la diagonal y el área formada por el triángulo OAB. Es una medida adimensional, y su valor está comprendido entre 0 y 1:

$$0 \leq I_G \leq 1$$

El valor 0 surge cuando hay equidistribución y el valor 1 cuando hay máxima concentración. Cuanto más cercano a 1 esté el I_G mayor será la concentración. Por el contrario, conforme se aproxime el I_G a 0 la desigualdad irá disminuyendo.

El I_G se calcula aplicando la siguiente fórmula:

$$I_G = \frac{\sum_{i=1}^{k-1} P_i Q_{i+1} - \sum_{i=1}^{k-1} Q_i P_{i+1}}{10000}$$

donde P_i y Q_i son los elementos utilizados para la obtención de la curva de Lorenz.

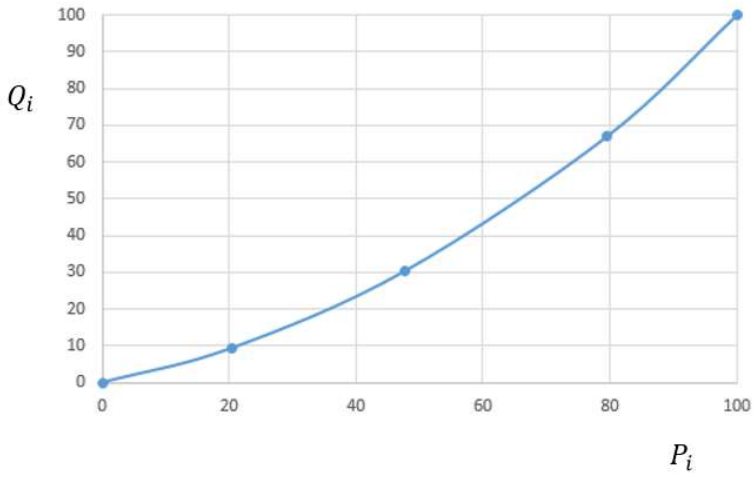
Al Índice de Gini le afecta el cambio de origen, pero no el de escala. Para el caso del cambio de origen, si todos los valores de la distribución se suman por una cantidad positiva, el I_G disminuye y se reduce la concentración. Por el contrario, si se resta una cantidad a cada valor de la distribución, el I_G aumenta y crece la desigualdad. Por ejemplo, una subvención de igual cuantía a todos los ciudadanos implicaría una reducción de la desigualdad; mientras que un impuesto de suma fija (todos los ciudadanos son gravados por la misma cantidad monetaria) aumentaría la desigualdad.

En el ejemplo 4.1 se presenta un ejercicio para la obtención de la curva de Lorenz y el índice de Gini, a partir de una distribución de salarios.

Ejemplo 4.1:

Salario Euros / hora Intervalos	**Número de trabajadores (n_i)**	N_i	x_i	$x_i n_i$	V_i	P_i	Q_i	$P_i Q_{i+1}$	$Q_i P_{i+1}$
15 - 45	9	9	30	270	270	20,45	9,42	621,13	449,79
45 - 55	12	21	50	600	870	47,73	30,37	3198,48	2415,52
55 - 95	14	35	75	1050	1920	79,55	67,02	7954,55	6701,37
95 - 115	9	44	105	945	2865	100	100	-	-

Curva de Lorenz:



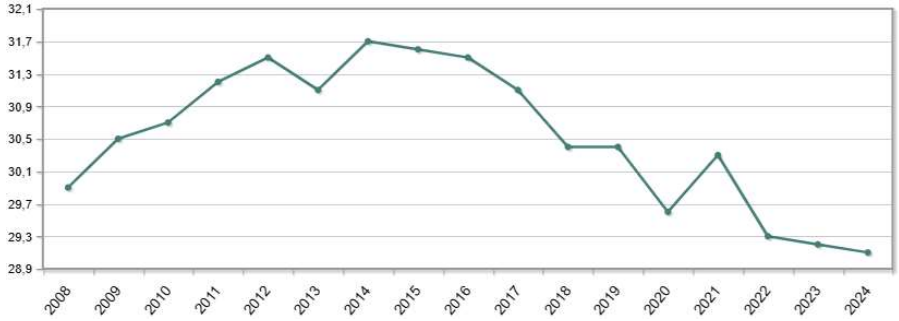
Índice de Gini:

$$I_G = \frac{\sum P_i Q_{i+1} - \sum Q_i P_{i+1}}{10000} = \frac{11774{,}16 - 9565{,}87}{10000} = 0{,}22$$

Existe poca desigualdad en la distribución de salarios ya que I_G está cercano a 0.

En la página web del INE, es posible obtener información sobre la evolución del índice de Gini, obtenida a partir de la Encuesta de Condiciones de Vida (INE, 2025i). En la figura 4.4 aparece el índice de Gini (multiplicado por 100) y se constata que a partir del año 2014 la desigualdad, en términos generales, ha ido disminuyendo; en particular, se constata que en el año 2024 la concentración es inferior a la registrada en el año 2008.

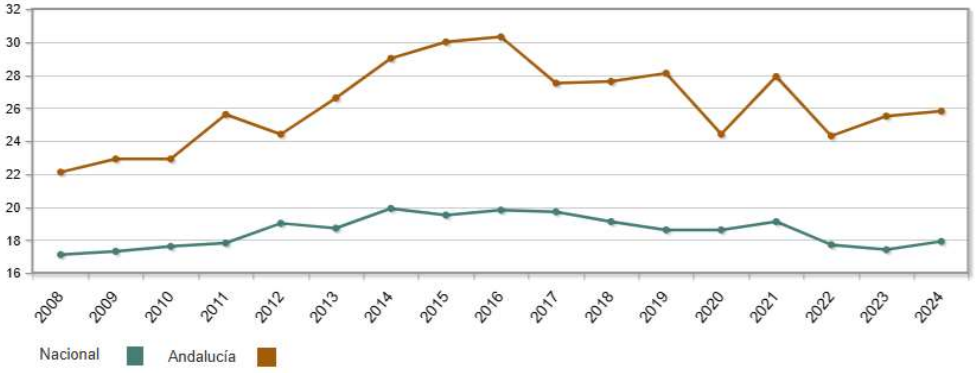
Figura 4.4. Índice de Gini para España: 2008-2023



Fuente: Encuesta de Condiciones de Vida (INE, 2025i).

En la información estadística proporcionada por el INE también es posible encontrar otros indicadores relacionados con la desigualdad y complementarios del I_G. Un ejemplo es la tasa de riesgo de pobreza, que es el porcentaje de personas que están por debajo del umbral de pobreza. El umbral de pobreza se define como "el 60% de la mediana de los ingresos por unidad de consumo de las personas". Los ingresos por unidad de consumo se calculan dividiendo el total de los ingresos del hogar entre el número total de miembros del hogar, ponderando al primer adulto por "1", "0,5" a los demás adultos y "0,3" a los menores de 14 años. En la figura 4.5 se presenta la tasa de riesgo de pobreza para España y Andalucía para el periodo 2008 y 2024. En dicho gráfico se observa que la tasa de riesgo de pobreza es superior en Andalucía para todos los años considerados, siendo del 25,8% en el año 2024, es decir, 7,9 puntos porcentuales superior a la registrada en España para el mismo año.

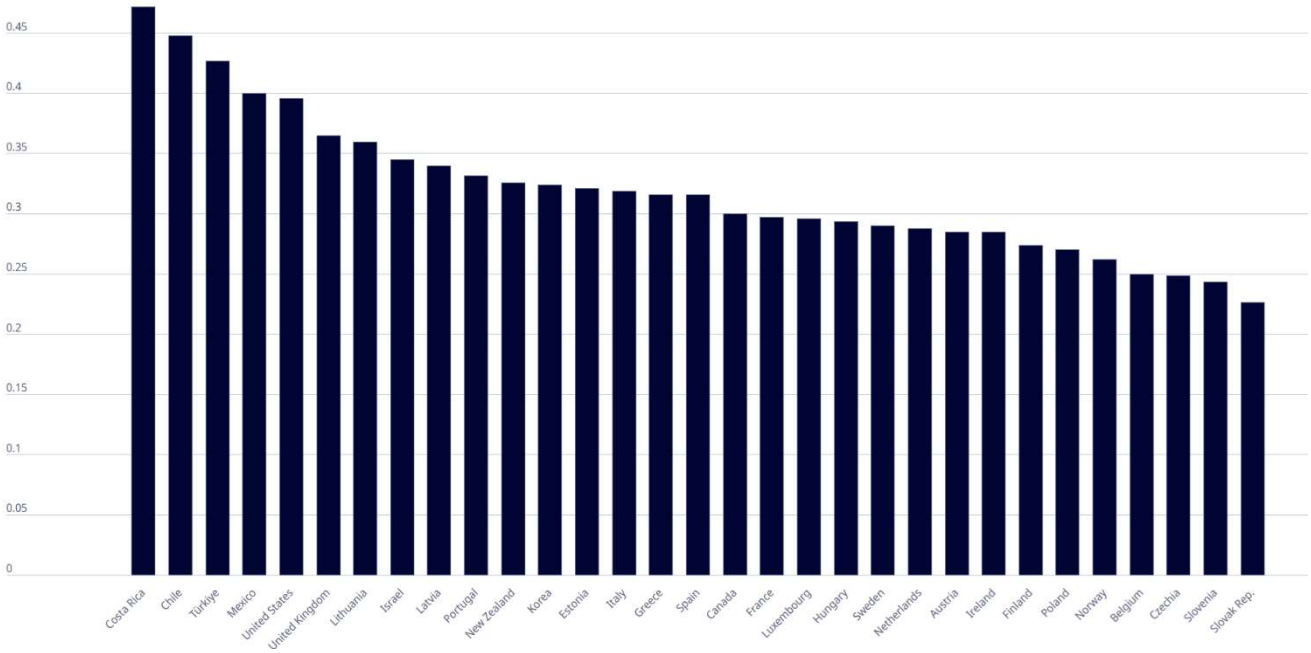
Figura 4.5 Tasa de riesgo de pobreza en España y en Andalucía: 2008-2024



Fuente: Encuesta de Condiciones de Vida (INE, 2025j).

Finalmente, a nivel internacional es posible observar en la figura 4.6, los valores del índice de Gini que proporciona la OCDE para un conjunto de países para el año 2022. En dicho gráfico, se puede observar que el valor más alto se alcanza para Costa Rica con un índice de Gini del 0,47; mientras que la cifra más baja se observa para la República Eslovaca con un valor del 0,23. España se sitúa en una posición intermedia con un valor situado en el 0,32.

Figura 4.6. Índice de Gini para países de la OCDE (2022)



Fuente: OCDE (2025).

Anexo 4.A. Ejemplo de obtención del Índice de Gini con Excel

En este anexo 4.A se muestran las operaciones que hay que realizar con Excel para obtener el índice de Gini haciendo uso de los datos del ejemplo 4.1. Las fórmulas aplicadas y los resultados numéricos se exhiben en la figura 4.A.1

Figura 4.A.1. Obtención del índice de Gini con Excel

- Fórmulas:

	A	B	C	D	E	F	G	H	I	J
1	Salario	Número de trabajadores	N_i	x_i	$x_i n_i$	V_i	P_i	Q_i	P_iQ_{i+1}	Q_iP_{i+1}
2	Euros / hora	n_i								
3	Intervalos									
4	15 - 45	9	=+B4	=+(15+45)/2	=+D4*B4	=+E4	=+(C4/$C$7)*100	=+(F4/$F$7)*100	=+G4*H5	=+H4*G5
5	45 - 55	12	=+B5+B4	=+(45+55)/2	=+D5*B5	=+E5+E4	=+(C5/$C$7)*100	=+(F5/$F$7)*100	=+G5*H6	=+H5*G6
6	55 - 95	14	=+B6+B5+B4	=+(55+95)/2	=+D6*B6	=+E6+E5+E4	=+(C6/$C$7)*100	=+(F6/$F$7)*100	=+G6*H7	=+H6*G7
7	95 - 115	9	=+B7+B6+B5+B4	=+(95+115)/2	=+D7*B7	=+E7+E6+E5+E4	=+(C7/$C$7)*100	=+(F7/$F$7)*100	.	.
8										
9									INDICE DE GINI	=+(SUMA(I4:I6)-SUMA(J4:J6))/10000

- Resultados:

	A	B	C	D	E	F	G	H	I	J
1	Salario	Número de trabajadores	N_i	x_i	$x_i n_i$	V_i	P_i	Q_i	P_iQ_{i+1}	Q_iP_{i+1}
2	Euros / hora	n_i								
3	Intervalos									
4	15 - 45	9	9	30	270	270	20.45	9.42	621.13	449.79
5	45 - 55	12	21	50	600	870	47.73	30.37	3198.48	2415.52
6	55 - 95	14	35	75	1050	1920	79.55	67.02	7954.55	6701.57
7	95 - 115	9	44	105	945	2865	100	100	.	.
8										
9									INDICE DE GINI	0.22

Fuente: Elaboración propia a partir del programa Excel.

PROPUESTAS DE EJERCICIOS PARA BLOQUE I

1. Suponga que se desea analizar al colectivo de personas mayores en residencias, ponga un ejemplo de "variable" y otro de "atributo" que se puedan estudiar para ese colectivo.

2. Complete la siguiente tabla correspondiente a la distribución del número de años de experiencia laboral (X) de un conjunto de trabajadores.

$\mathbf{X}$	$\mathbf{n_i}$	$\mathbf{x_i}$	$\mathbf{a_i}$	$\mathbf{N_i}$	$\mathbf{h_i}$
0-2	20				
2-4	60				
4-8	5				
8-20	2				

3. Complete la información de la siguiente distribución sobre la renta del hogar mensual de un conjunto de 10 personas. Además, obtenga la media aritmética y estudie su representatividad, obtenga la moda y analice el tipo de asimetría, y estudie la concentración.

Renta $(\mathbf{10^3}€)$	$\mathbf{x_i}$	$\mathbf{n_i}$	$\mathbf{x_i n_i}$	$\mathbf{N_i}$	$\mathbf{V_i}$	$\mathbf{P_i}$	$\mathbf{Q_i}$
1-3		3					
3-5							
-	6	2					

4. Si en una distribución, se cumple que $P_i = Q_i$ (para todo i), ¿qué puede afirmar sobre la concentración de la distribución? ¿cómo sería la curva de Lorenz?

5. Dada una variable X, indique cómo se obtiene su variable tipificada, explique su utilidad e indique si tiene unidad de medida. ¿Cómo le afecta a la variable tipificada los cambios de escala y de origen de la variable X?

6. ¿Cuál son las ventajas del Coeficiente de Variación de Pearson sobre la Desviación Típica como medida de dispersión?

7. ¿Qué porcentaje de observaciones son excluidas en el cálculo del Recorrido Interpercentílico?

8. ¿Qué tipo de asimetría existe en una distribución campaniforme cuando $\bar{X} < M_o$? Apoye su respuesta con un gráfico.

9. A partir de una distribución de salarios mensuales en el primer empleo para mujeres tituladas en Ciencias Sociales:

Salarios mensuales (10^3 €)	**Mujeres (n_i)**
0,7 - 1	20
1 - 1,5	77
1,5 - 2	47
2 - 2,5	14
2,5 - 3	6

a) Obtenga la media aritmética y estudie su representatividad.

b) Estudie la asimetría de la distribución sabiendo que $\sum(x_i - \bar{X})^3 n_i = 13{,}192$.

c) ¿Cuál es el salario mínimo del 20% de trabajadoras con salarios más altos?

d) ¿Cuál es el salario máximo del 30% de trabajadoras con salarios más bajos?

e) Obtenga el número de mujeres con salarios superiores a la moda.

f) Estudie analíticamente la concentración de la distribución de salarios.

g) La media aritmética de los salarios mensuales en el primer empleo de los varones titulados en Ciencias Sociales es igual a $1{,}60 * 10^3$€, y su desviación típica es igual a $0{,}621 * 10^3$€. Si una mujer recibe un salario de 1600 € y un varón un salario de 1700 €, ¿cuál tiene una mejor posición relativa en su distribución?

h) Si la población de varones titulados en Ciencias Sociales asciende a 150, ¿cuál es el salario mensual medio en el primer empleo de los/as titulados/as en Ciencias Sociales (hombres y mujeres)?

i) Para las mujeres, ¿qué ocurriría con la representatividad de la media aritmética en los siguientes escenarios?

(1) Los salarios mensuales aumentasen en un 5%.

(2) Los salarios mensuales aumentasen en 20 €.

10. Un conjunto de varones son encuestados sobre cuál sería el ingreso mínimo vital (en 10^2 €) necesario para llegar a fin de mes:

Ingreso (10^2 €)	**Personas (n_i)**
4-8	16
8-16	49
16-20	10
20-30	7

a) ¿Cuál es la población objeto de estudio? Obtenga la columna de frecuencias acumuladas (N_i) y proporcione una interpretación para N_2.

b) Obtenga la media aritmética y estudie su representatividad (Nota: $\sum x_i^2 n_i = 15247$).

c) Obtenga el ingreso vital más frecuente, y estudie la asimetría de la distribución.

d) Obtenga e interprete la mediana.

e) ¿Qué porcentaje de individuos tienen un ingreso vital superior a 900 €?

f) ¿Tiene sentido utilizar $\frac{\sum(x_i-\bar{X})n_i}{N}$ como una medida de dispersión con respecto a la media aritmética? Razone la respuesta.

g) Obtenga el recorrido interdecílico. ¿Es una medida adimensional de la dispersión de una variable? Razone la respuesta.

11. Para un conjunto de hogares se conoce el importe anual de las ayudas por asistencia social (X, en 10^2 €).

Importe (10^2 €)	**Hogares (n_i)**
4-10	63
10-15	62
15-20	56
20-40	195
40-50	77

a) ¿Cuál es la variable objeto de estudio? Obtenga la columna de frecuencias acumuladas porcentuales (P_i) y proporcione una interpretación para P_2.

b) Obtenga la mediana e interprétela. Razone la respuesta.

c) Obtenga la media aritmética y estudie su representatividad.

d) Estudie la concentración de la distribución.

e) ¿Qué ocurriría con la desigualdad en los siguientes escenarios:

(1) Los importes aumentan en un 10%.

(2) Los importes aumentan en 25 €.

f) ¿Cómo cambia la media aritmética, la varianza, la desviación típica y el coeficiente de asimetría de Fisher en cada uno de los escenarios anteriores?

g) ¿Cuál es el porcentaje de hogares por encima del importe medio?

TEMA 5. ESTADÍSTICA DESCRIPTIVA BIVARIANTE

5.1 INTRODUCCIÓN

Hasta ahora todo el análisis se ha centrado en una única variable, es decir, se ha realizado un análisis univariante. Este tipo de enfoque tiene su propio interés, ya que nos permite obtener conclusiones relevantes sobre distintas características de la distribución asociada a la variable objeto de estudio. No obstante, habitualmente el interés se centra en más de una variable y en las posibles interrelaciones que puedan existir entre ellas. En este tema se realiza una aproximación a esta última cuestión mediante la presentación del análisis estadístico descriptivo de dos variables. Esto nos dispone para abordar cuestiones como el concepto de independencia estadística o la asociación lineal entre dos variables, que van a poner las bases para el tema 6 donde se cuantifica mediante el análisis de regresión la relación lineal entre las variables y se procede a la realización de predicciones.

5.2 TABLAS DE DOBLE ENTRADA. SÍNTESIS NUMÉRICA.

La representación estadística de los datos correspondientes a dos variables se conoce como tabla de doble entrada, y nos permite realizar una primera síntesis numérica de la información. En la tabla de doble entrada se va a distinguir, igual que en el caso de una sola variable, entre datos sin agrupar o datos agrupados en intervalos.

a) Distribuciones con datos sin agrupar:

Existen dos variables X e Y. Los valores de X son: $x_1, x_2, \dots, x_k$; y los valores de Y son: $y_1, y_2, \dots y_h$. Para cada par de valores (x_i, y_i) se registran n_{ij} observaciones (Tabla 5.1).

Tabla 5.1. Tabla de doble entrada para distribuciones de datos sin agrupar

X \ Y	y_1	y_2	…	y_j	…	y_h
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1h}
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2h}
…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{ih}
…	…	…	…	…	…	…
x_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{kh}

Fuente: Elaboración propia.

n_{ij} se denomina frecuencia absoluta bidimensional. La suma de todas las frecuencias absolutas bidimensionales es el tamaño de la población:

$$\sum_{i=1}^{k}\sum_{j=1}^{h} n_{ij} = N$$

La frecuencia relativa bidimensional f_{ij}, es el cociente entre la frecuencia absoluta bidimensional y el tamaño de la población:

$$f_{ij} = \frac{n_{ij}}{N}$$

y se cumple que:

$$\sum_{i=1}^{k}\sum_{j=1}^{h} f_{ij} = 1$$

Ejemplo 5.1:

Supongamos la siguiente distribución bivariante correspondiente a la edad (X) y al número de años trabajados (Y) para 20 trabajadores:

X (edad)	**Y (años trabajados)**
20	1
20	2
25	3
25	3
30	5
30	6
30	6
35	10
35	11
40	11
40	12
40	15
40	20
40	30
50	7
50	9
50	12
50	15
60	10
60	30

A partir de la tabulación anterior es posible generar la siguiente tabla de doble entrada de frecuencias absolutas bidimensionales:

X\Y	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**9**	**10**	**11**	**12**	**15**	**20**	**30**
20	1	1	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	2	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	1	2	0	0	0	0	0	0	0	0
35	0	0	0	0	0	0	0	0	1	1	0	0	0	0
40	0	0	0	0	0	0	0	0	0	1	1	1	1	1
50	0	0	0	0	0	0	1	1	0	0	1	1	0	0
60	0	0	0	0	0	0	0	0	1	0	0	0	0	1

b) Distribuciones con datos agrupados en intervalos:

En las distribuciones con datos agrupados, los valores de X y de Y se recogen en intervalos: k intervalos para X, y h para Y (tabla 5.2).

Tabla 5.2. Tabla de doble entrada para datos agrupados en intervalos (en términos absolutos)

X \ Y	$(L_0\text{-}L_1)_y$	$(L_1\text{-}L_2)_y$	**…**	$(L_{j-1}\text{-}L_j)_y$	**…**	$(L_{h-1}\text{-}L_h)_y$
$(L_0\text{-}L_1)_x$	n_{11}	n_{12}	…	n_{1j}	…	n_{1h}
$(L_1\text{-}L_2)_x$	n_{21}	n_{22}	…	n_{2j}	…	n_{2h}
…	…	…	…	…	…	…
$(L_{i-1}\text{-}L_i)_x$	n_{i1}	n_{i2}	…	n_{ij}	…	n_{ih}
…	…	…	…	…	…	…
$(L_{k-1}\text{-}L_k)_x$	n_{k1}	n_{k2}	…	n_{kj}	…	n_{kh}

Fuente: Elaboración propia.

Como hemos señalado anteriormente, la distribución de frecuencias se puede expresar en términos relativos dividiendo cada n_{ij} entre N (tabla 5.3).

Tabla 5.3. Tabla de doble entrada para datos agrupados en intervalos (en términos relativos)

X \ Y	$(L_0\text{-}L_1)_y$	$(L_1\text{-}L_2)_y$	**…**	$(L_{j-1}\text{-}L_j)_y$	**…**	$(L_{h-1}\text{-}L_h)_y$
$(L_0\text{-}L_1)_x$	f_{11}	f_{12}	…	f_{1j}	…	f_{1h}
$(L_1\text{-}L_2)_x$	f_{21}	f_{22}	…	f_{2j}	…	f_{2h}
…	…	…	…	…	…	…
$(L_{i-1}\text{-}L_i)_x$	f_{i1}	f_{i2}	…	f_{ij}	…	f_{ih}
…	…	…	…	…	…	…
$(L_{k-1}\text{-}L_k)_x$	f_{k1}	f_{k2}	…	f_{kj}	…	f_{kh}

Fuente: Elaboración propia.

Ejemplo 5.2:

A continuación, se muestra una tabla de doble entrada para la distribución de la edad (X) de 100 trabajadores y sus salarios por hora (Y, €), tanto en términos absolutos como relativos:

* En términos absolutos:

X \ Y	**6-10**	**10-30**	**30-50**
18 – 25	36	4	0
25 - 40	2	30	1
40 - 65	0	6	21

* En términos relativos:

X \ Y	**6-10**	**10-30**	**30-50**
18 – 25	0,36	0,04	0
25 - 40	0,02	0,30	0,01
40 - 65	0	0,06	0,21

5.3 DISTRIBUCIONES MARGINALES Y CONDICIONADAS.

A partir de la distribución bivariante es posible obtener las distribuciones marginales de cada una de las variables consideradas individualmente. El procedimiento de cálculo de las distribuciones marginales se presenta a continuación.

a) Para hallar la distribución marginal de X hay que sumar las frecuencias asociadas a cada uno de los valores observados de X. En definitiva, según la exposición realizada, hay que sumar los valores de cada fila de frecuencias. En la tabla 5.4 se muestra dicho procedimiento para datos agrupados en intervalos.

Tabla 5.4. Distribución marginal de X

X	$\boldsymbol{n_{i.}}$	$\boldsymbol{f_{i.}}$
$L_0 - L_1$	$n_{1.} = \sum_{j=1}^{h} n_{1,j}$	$f_{1.} = \frac{n_{1.}}{N}$
$L_1 - L_2$	$n_{2.} = \sum_{j=1}^{h} n_{2,j}$	$f_{2.} = \frac{n_{2.}}{N}$
. . .	. . .	. . .
$L_{k-1} - L_k$	$n_{k.} = \sum_{j=1}^{h} n_{k,j}$	$f_{k.} = \frac{n_{k.}}{N}$
	$N = \sum_{i=1}^{k} n_{i.}$	$\sum_{i=1}^{k} f_{i.} = 1$

Fuente: Elaboración propia.

b) Para hallar la distribución marginal de Y hay que sumar las frecuencias asociadas a cada uno de los valores observados de Y; es decir, hay que sumar los valores de cada columna de frecuencias (tabla 5.5).

Tabla 5.5. Distribución marginal de Y

Y	$\boldsymbol{n_{.j}}$	$\boldsymbol{f_{.j}}$
$L_0 - L_1$	$n_{.1} = \sum_{i=1}^{k} n_{i,1}$	$f_{.1} = \frac{n_{.1}}{N}$
$L_1 - L_2$	$n_{.2} = \sum_{i=1}^{k} n_{i,2}$	$f_{.2} = \frac{n_{.2}}{N}$
. . .	. . .	. . .
$L_{h-1} - L_h$	$n_{.h} = \sum_{i=1}^{k} n_{i,h}$	$f_{.k} = \frac{n_{.h}}{N}$
	$N = \sum_{j=1}^{h} n_{.j}$	$\sum_{j=1}^{h} f_{.j} = 1$

Fuente: Elaboración propia.

Otro tipo de distribuciones que podemos obtener a partir de la distribución bivariante son las distribuciones condicionadas**.** Estas distribuciones se generan tras imponer alguna condición sobre los valores de X o de Y. Por ejemplo, si se desea obtener la distribución de X condicionada a que Y tome un valor comprendido en el intervalo $(L_{j-1} - L_j)$, las frecuencias absolutas de esa distribución son precisamente las que aparecen en la columna j. En cambio, si se obtiene la distribución de Y condicionada a que X esté en el intervalo $(L_{i-1} - L_i)$, las frecuencias absolutas son las que se muestran en la fila i.

Las distribuciones marginales y condicionadas son distribuciones unidimensionales, por lo que a partir de ellas podemos analizar sus características: medidas de posición central, dispersión, forma o concentración.

Ejemplo 5.3:

A partir de la distribución bivariante que aparece en el ejemplo 5.2 obtenga las distribuciones marginales de X e Y, la distribución de la edad condicionada a que el salario se encuentre comprendido entre 30 y 50 euros, y la distribución del salario condicionada a que la edad de los trabajadores esté entre 25 y 40 años.

Distribución marginal de X

X (Edad)	$n_{i.}$	$f_{i.}$
18-25	40	0,40
25-40	33	0,33
40-65	27	0,27
Total	100	1

Distribución marginal de Y

Y (Salario/hora)	$n_{.j}$	$f_{.j}$
6-10	38	0,38
10-30	40	0,40
30-50	22	0,22
Total	100	1

Distribución de X condicionada a que Y esté comprendida en el intervalo 30-50

X (Edad)	$n_{i,3}$	$f_{i,3}$
18-25	0	0
25-40	1	0,045
40-65	21	0,955
Total	22	1

Distribución de Y condicionada a que X esté comprendida en el intervalo 25-40

Y (Salario/hora)	$n_{2,j}$	$f_{2,j}$
6-10	2	0,06
10-30	30	0,91
30-50	1	0,03
Total	33	1

5.4 DEPENDENCIA ESTADÍSTICA Y COVARIACIÓN. DIAGRAMAS DE DISPERSIÓN

En este epígrafe se analiza cómo verificar la posible dependencia que existe entre dos variables X e Y. Se dice que dos variables son estadísticamente independientes si los valores que toma una no están influenciados por los valores de la otra. Para comprobar la independencia o no de dos variables hay que contrastar si se cumple la siguiente condición:

$X\ e\ Y$ son estadísticamente independientes si: $f_{i,j} = f_{i.} * f_{.j}\ \ ,\forall\ \ i,j$

Ejemplo 5.4:

Verifique si las variables edad y salario de los trabajadores son o no estadísticamente independientes:

X \ Y	**6-10**	**10-30**	**30-50**	$\mathbf{f_{i.}}$
18 - 25	0,36	0,04	0	0,40
25 - 40	0,02	0,30	0,01	0,33
40 - 65	0	0,06	0,21	0,27
$\mathbf{f_{.j}}$	0,38	0,40	0,22	

Se observa, por ejemplo, que:

$$f_{11} \neq f_{1.} * f_{.1}$$

$$0{,}36 \neq 0{,}40 * 0{,}38$$

por lo que se puede afirmar que la edad y el salario son dos variables que no son independientes estadísticamente.

El concepto de Covariación está asociado al de dependencia estadística. De hecho, se dice que dos variables presentan Covariación si existe alguna relación de dependencia entre ellas. Las relaciones de dependencia pueden adoptar diversas formas:

a) Dependencia causal unilateral:

Cuando la ocurrencia de una variable influye sobre la ocurrencia de otra, pero no al contrario. La variable que ejerce influencia (X) se suele denominar: explicativa, independiente, causa o exógena; mientras que la variable influida (Y) se denomina variable explicada, dependiente, efecto o variable endógena. Por ejemplo, si X es el Nº de Curriculum Vitae enviados a empresas e Y el Nº de ofertas laborales recibidas, se podría establecer la siguiente relación de causalidad:

X (Nº de CV enviado a empresas) →Y (Nº de ofertas recibidas)

b) Interdependencia:

En este caso, los valores de las dos variables son el resultado de una relación de equilibrio donde ambas variables interactúan. Por ejemplo, en los mercados de bienes existe una relación de causalidad bidireccional entre los precios y las cantidades demandadas:

X (precio de un bien) → Y (cantidad demandada de un bien)

Y (cantidad demandada de un bien) →X (precio de un bien)

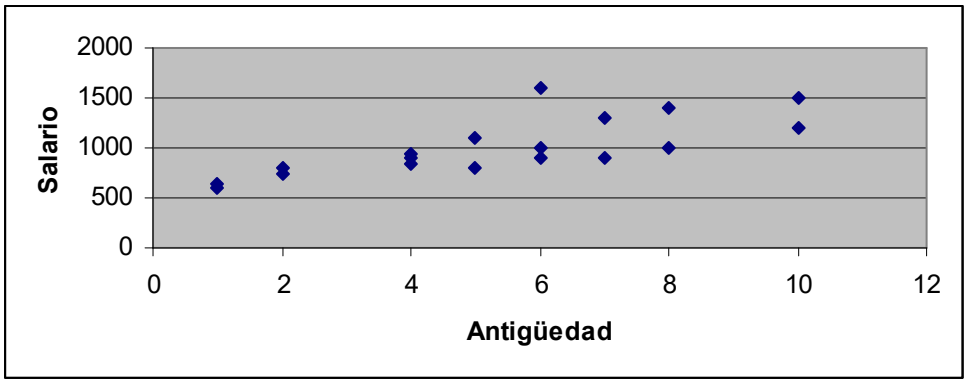
c) Dependencia indirecta:

Las variables X e Y está relacionadas a través de una tercera variable Z:

$$Z \text{ (renta del padre)} \rightarrow \begin{Bmatrix} X\ (renta\ del\ \ hijo) \\ Y\ \ (renta\ \ de\ la\ hija) \end{Bmatrix}$$

Un método gráfico para atisbar el tipo de covariación es el diagrama de dispersión, donde las variables X e Y son representadas en un eje de coordenadas. Por ejemplo, en la figura 5.1, se observa una relación lineal positiva entre la antigüedad en la empresa y el salario para un conjunto de pares de valores.

Figura 5.1 Caso simulado: Trabajadores según salario y antigüedad



Fuente: Elaboración propia.

5.5 COVARIANCIA Y COEFICIENTE DE CORRELACIÓN LINEAL

5.5.1 COVARIANZA

La Covarianza (que vamos a denotar por $S_{x,y}$) es una medida estadística para cuantificar la relación lineal entre dos variables. La fórmula de la covarianza es la siguiente:

$$S_{x,y} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{h}(x_i - \bar{X})(y_j - \bar{Y})n_{ij}}{N}$$

En el caso de frecuencias unitarias, tendríamos pares de datos:

X_i	Y_i
x_1	y_1
x_2	y_2
.	.
.	.
.	.
x_N	y_N

y la fórmula de la covarianza sería igual a:

$$S_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})}{N}$$

En los casos anteriores, para facilitar el cálculo, se suelen utilizar las siguientes expresiones que son equivalentes a las anteriores:

a) Para frecuencias no unitarias:

$$S_{x,y} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{h}(x_i - \bar{X})(y_j - \bar{Y})n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{h} x_i y_j n_{ij}}{N} - \bar{X}\bar{Y}$$

b) Para frecuencias unitarias:

$$S_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})}{N} = \frac{\sum_{i=1}^{N} x_i y_i}{N} - \bar{X}\bar{Y}$$

La unidad de medida de la covarianza $S_{x,y}$ es igual al producto de la unidad de X por la unidad de Y. Además, $S_{x,y}$ no está acotada, pudiendo tomar valores entre $-\infty$ e ∞. En función de los valores que alcance $S_{x,y}$ se pueden señalar las siguientes relaciones entre X e Y:

$$\begin{cases} S_{xy} > 0 \Rightarrow \text{Relación lineal positiva o directa} \\ S_{xy} < 0 \Rightarrow \text{Relación lineal negativa o inversa} \\ S_{xy} = 0 \Rightarrow \text{No hay relación lineal} \end{cases}$$

Las propiedades de la covarianza respecto a los cambios de origen y escala son las siguientes:

1. No le afecta los cambios de origen.

Si a todos los valores de X e Y se les suma una constante, la covarianza no varía:

$$\begin{cases} x_i' = x_i + c \\ y_j' = y_j + d \end{cases} \rightarrow S_{x',y'} = S_{x,y}$$

Demostración:

$$S_{x',y'} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{h}(x_i' - \bar{X}')(y_i' - \bar{Y}')n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{h}((x_i + c) - (\bar{X} + c))((y_j + d) - (\bar{Y} + d))n_{ij}}{N} =$$

$$= \frac{\sum_{i=1}^{k}\sum_{j=1}^{h}(x_i - \bar{X})(y_j - \bar{Y})n_{ij}}{N} = S_{x,y}$$

2. Le afecta los cambios de escala.

Si a todos los valores de X e Y se les multiplica por una constante (c y d, respectivamente), la covarianza se ve multiplicada por dichas constantes:

$$\begin{cases} x_i' = x_i * c \\ y_j' = y_j * d \end{cases} \rightarrow S_{x',y'} = c * d * S_{x,y}$$

Demostración:

$$S_{x',y'} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{h}(x_i' - \bar{X}')(y_i' - \bar{Y}')n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{h}((x_i * c) - (\bar{X} * c))((y_j * d) - (\bar{Y} * d))n_{ij}}{N} =$$

$$= c * d * \frac{\sum_{i=1}^{k}\sum_{j=1}^{h}(x_i - \bar{X})(y_j - \bar{Y})n_{ij}}{N} = c * d * S_{x,y}$$

A continuación, se presenta dos ejemplos sobre cómo obtener la covarianza.

Ejemplo 5.5:

La distribución de salarios por hora (variable Y, en €) según el intervalo de edad (variable *X*, en años) de un grupo de 100 trabajadores se presenta en la siguiente tabla. Obtenga la covarianza e interprétela.

Y / **X**	**6-10**	**10-30**	**30-50**
18 - 25	36	4	0
25 - 40	2	30	1
40 - 65	0	6	21

$$\sum_{i=1}^{k}\sum_{j=1}^{h} x_i y_j n_{ij} = 21{,}5 \cdot 8 \cdot 36 + 21{,}5 \cdot 20 \cdot 4 + 21{,}5 \cdot 40 \cdot 0 + 32{,}5 \cdot 8 \cdot 2 + 32{,}5 \cdot 20 \cdot 30 + 32{,}5 \cdot 40 \cdot 1 +$$
$$+52{,}5 \cdot 8 \cdot 0 + 52{,}5 \cdot 20 \cdot 6 + 52{,}5 \cdot 40 \cdot 21 = 79632$$

X (Edad)	$\boldsymbol{n_{i.}}$	$\boldsymbol{x_i}$	$\boldsymbol{x_i n_{i.}}$
18-25	40	21,5	860
25-40	33	32,5	1072,5
40-65	27	52,5	1417,5

Y (Salario/hora)	$\boldsymbol{n_{.j}}$	$\boldsymbol{y_j}$	$\boldsymbol{y_j n_{.j}}$
6-10	38	8	304
10-30	40	20	800
30-50	22	40	880

$$\bar{X} = \frac{\sum x_i n_{i.}}{N} = \frac{3350}{100} = 33{,}5\ años$$

$$\bar{Y} = \frac{\sum y_j n_{.j}}{N} = \frac{1984}{100} = 19{,}84\ €/hora$$

$$S_{x,y} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{h} x_i y_j n_{ij}}{N} - \bar{X}\bar{Y} = \frac{79632}{100} - (33{,}5 * 19{,}84) = 131{,}68\ años * (€/hora)$$

$S_{xy} > 0 \Rightarrow$ Relación lineal positiva o directa entre la edad (X) y el salario (Y).

Ejemplo 5.6:

Se conocen los salarios (Y, €/hora) y la edad de un grupo de 6 trabajadores ($X, años$). Obtenga e interprete la covarianza.

$\boldsymbol{x_i}$	$\boldsymbol{y_i}$	$\boldsymbol{y_i * x_i}$
30	10	300
33	15	495
35	18	630
40	22	880
42	25	1050
50	30	1500

La distribución de frecuencias es unitaria, por lo que hay que aplicar la siguiente fórmula:

$$S_{x,y} = \frac{\sum x_i y_i}{N} - \bar{X} * \bar{Y} = \frac{4855}{6} - \left(\frac{230}{6}\right) * \left(\frac{120}{6}\right) = 42{,}5\ años * (€/h)$$

$S_{xy} > 0 \Rightarrow$ Relación lineal positiva o directa entre la edad (X) y el salario (Y).

5.5.2 COEFICIENTE DE CORRELACIÓN LINEAL

El Coeficiente de Correlación Lineal $r_{x,y}$ es una medida estadística adimensional para cuantificar la relación lineal entre dos variables, cuya fórmula es la siguiente:

$$r_{x,y} = \frac{S_{x,y}}{S_x * S_y}$$

El Coeficiente de Correlación Lineal siempre tiene el mismo signo que la covarianza, ya que el denominador siempre es positivo. Además, $r_{x,y}$ es adimensional y puede tomar valores entre -1 y 1:

$$-1 \leq r_{xy} \leq 1$$

Los valores extremos de $r_{x,y}$ indican lo siguiente:

$$\begin{cases} r_{xy} = 1 \Rightarrow \text{Relación lineal perfecta positiva o directa} \\ r_{xy} = -1 \Rightarrow \text{Relación lineal perfecta negativa o inversa} \\ r_{xy} = 0 \Rightarrow \text{No hay relación lineal} \end{cases}$$

En los casos situados entre los valores extremos ocurre lo siguiente:

a) Correlación lineal positiva (entre 0 y 1) indica que hay relación lineal directa o positiva entre las variables, y a medida que se acerca a 1, más fuerte es la relación lineal directa o positiva.

b) Correlación lineal negativa (entre -1 y 0) indica que hay relación lineal inversa o negativa entre las variables, y a medida que se acerca a -1, más fuerte es la relación lineal inversa o negativa.

Las propiedades de $r_{x,y}$ respecto al cambio de origen y/o escala de la variable son las siguientes:

1. Cambio de origen: No le afecta.

Si a todos los valores de X e Y se les suma una constante (c y d, respectivamente), el coeficiente de correlación no varía:

$$\begin{cases} x_i' = x_i + c \\ y_j' = y_j + d \end{cases} \rightarrow r_{x',y'} = r_{x,y}$$

Demostración:

$$r_{x',y'} = \frac{S_{x',y'}}{S_{x'} * S_{y'}} = \frac{S_{x,y}}{S_x * S_y} = r_{x,y}$$

2. Cambio de escala: No le afecta.

Si a todos los valores de X e Y se les multiplica una constante (c y d, respectivamente), el coeficiente de correlación lineal no varía:

$$\begin{cases} x_i' = x_i * c \\ y_j' = y_j * d \end{cases} \rightarrow \; r_{x',y'} = c * d * r_{x,y}$$

Demostración:

$$r_{x',y'} = \frac{S_{x',y'}}{S_{x'} * S_{y'}} = \frac{c * d * S_{x,y}}{c * S_x * d * S_y} = r_{x,y}$$

A continuación, en los ejemplos 5.7 y 5.8, se obtienen los coeficientes de correlación lineal para los datos asociados a los ejemplos 5.5 y 5.6, respectivamente.

Ejemplo 5.7:

Con los datos del ejemplo 5.5 obtenga el coeficiente de correlación lineal:

X (Edad)	$n_{i.}$	x_i	$x_i n_{i.}$	x_i^2	$x_i^2 n_{i.}$
18-25	40	21,5	860	462,25	18490
25-40	33	32,5	1072,5	1056,25	34856,25
40-65	27	52,5	1417,5	2756,25	74418,75

Y (Salario/hora)	$n_{.j}$	y_j	$y_j n_{.j}$	y_j^2	$y_j^2 n_{.j}$
6-10	38	8	304	64	2432
10-30	40	20	800	400	16000
30-50	22	40	880	1600	35200

$$S_x^2 = \frac{127765}{100} - 33{,}5^2 = 155{,}4\ años^2$$

$$S_x = 12{,}47\ años$$

$$S_y^2 = \frac{53632}{100} - 19{,}84^2 = 146{,}69\ €^2$$

$$S_y = 12{,}11\ €$$

$$r_{x,y} = \frac{S_{x,y}}{S_x * S_y} = \frac{131{,}68\ años * euros}{12{,}47\ años * 12{,}11\ euros} = 0{,}87$$

Al ser positivo hay una relación lineal directa o positiva. Además, al no estar muy alejada de 1, se puede afirmar que la relación lineal positiva es fuerte.

Ejemplo 5.8:

Con los datos del ejemplo 5.6 obtenga el coeficiente de correlación lineal:

$\mathbf{x_i}$	$\mathbf{y_i}$	$\mathbf{y_i * x_i}$	$\mathbf{x_i^2}$	$\mathbf{y_i^2}$
30	10	300	900	100
33	15	495	1089	225
35	18	630	1225	324
40	22	880	1600	484
42	25	1050	1764	625
50	30	1500	2500	900

$$S_x = \sqrt{\frac{\sum x_i^2}{N} - \bar{X}^2} = \sqrt{\frac{9078}{6} - 1469{,}44} = 6{,}6\ años$$

$$S_y = \sqrt{\frac{\sum y_i^2}{N} - \bar{Y}^2} = \sqrt{\frac{2658}{6} - 400} = 6{,}56\ €/h$$

$$r_{x,y} = \frac{S_{x,y}}{S_x * S_y} = \frac{42{,}5}{6{,}6 * 6{,}56} = 0{,}98$$

Existe una fuerte relación lineal positiva, ya que $r_{x,y}$ está muy cercano a 1.

Anexo 5.A. Ejemplo de obtención de la Covarianza y del Coeficiente de Correlación Lineal con Excel

En este anexo se explica cómo obtener la Covarianza y el Coeficiente de correlación lineal usando como ejemplo los datos anuales correspondientes al número de mujeres activas y mujeres ocupadas (en 10^3 de personas) en España para el periodo 2018-2024. En concreto, las funciones que hay que utilizar adaptadas al rango de datos del ejemplo son:

a) Covarianza: =COVARIANCE.P(A2:A8;B2:B8)
b) Coeficiente de correlación lineal: =COEF.DE.CORREL(A2:A8;B2:B8)

Tabla 4.A.1. Obtención de la covarianza y el coeficiente de correlación lineal con Excel

- Fórmulas:

	A	B	C
1	**Activas**	**Ocupadas**	**Covarianza**
2	10600.3	8795.7	
3	10753.7	9033.7	=COVARIANCE.P(A2:A8;B2:B8)
4	10624.4	8772.7	
5	10942.2	9100.3	**Coeficiente de Correlación Lineal**
6	11082.8	9432.9	
7	11384.9	9805.5	=COEF.DE.CORREL(A2:A8;B2:B8)
8	11512	10052.5	

- Resultados:

	A	B	C	D
1	**Activas**	**Ocupadas**	**Covarianza**	
2	10600.3	8795.7		
3	10753.7	9033.7	152298.0424	
4	10624.4	8772.7		
5	10942.2	9100.3	**Coeficiente de Correlación Lineal**	
6	11082.8	9432.9		
7	11384.9	9805.5	0.99	
8	11512	10052.5		

Fuente: Elaboración propia.

Los resultados nos revelan la existencia de una asociación lineal positiva entre el número de mujeres activas y el número de mujeres ocupadas, ya que la covarianza es positiva. Además, la intensidad de la relación lineal es muy fuerte al estar el coeficiente de correlación lineal muy cercano a 1. Este resultado era esperado, ya que las personas activas están ocupadas o paradas; no obstante, la relación no es directa ya que la evolución del número de ocupados/as y parados/as no sigue un patrón determinista.

TEMA 6. ANALISIS DE REGRESIÓN Y SERIES TEMPORALES

6.1 INTRODUCCIÓN

Este tema 6 aborda, en primer lugar, el análisis de regresión entre una variable X y una variable Y (epígrafes 6.2 y 6.3). Dicho análisis es un procedimiento usado para cuantificar la relación entre variables y para realizar predicciones. En particular, nos centramos en la dependencia unilateral entre dos variables X e Y, donde X es la variable independiente o causa e Y es la variable dependiente o efecto. El objetivo del análisis de regresión es establecer una relación funcional entre las variables X e Y del tipo: $Y = f(X)$. La estimación de esa relación funcional permitirá conocer la influencia de X sobre Y, así como realizar previsiones de Y cuando X alcance determinados valores. En segundo lugar, se realiza una breve introducción a las series temporales, es decir, variables estadísticas que se observan a lo largo del tiempo, con el fin de presentar sus componentes desde un punto de vista del análisis clásico. Este enfoque también nos permitirá realizar predicciones de la variable observada (epígrafes 6.4 y 6.5).

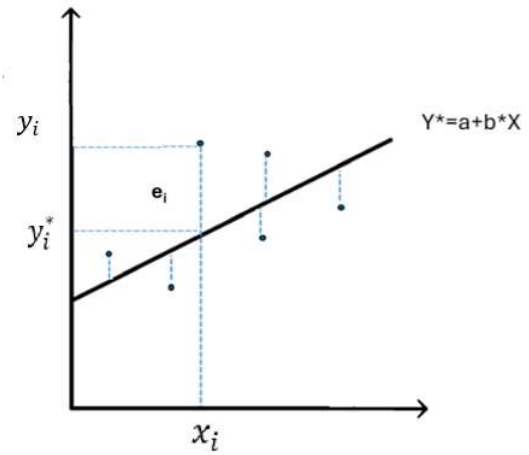
6.2 LA LÍNEA DE REGRESIÓN

El análisis de regresión va a consistir en obtener la línea ideal (denominada línea de regresión) que mejor ajuste los puntos en el diagrama de dispersión, lo que coloquialmente se denomina nube de puntos. En este enfoque básico se supone que la línea de regresión es una línea recta:

$$y_i^* = a + bx_i$$

Para centrar un poco los conceptos, supongamos que dado un conjunto de parejas de datos (x_i, y_i) se conoce esa línea de regresión que, de forma muy simplificada, se representa en la figura 6.1.

Figura 6.1. Ejemplo simulado de línea de regresión



Fuente: Elaboración propia.

En la anterior figura podemos comprobar que para un valor observado x_i existe un valor y_i observado y otro predicho por la línea de regresión y_i^*, este último valor también se conoce como valor ajustado. De esta forma, podemos deducir que cada valor observado y_i se puede descomponer en dos partes y_i^* y e_i. El segundo componente, e_i, se denomina error o residuo, y es la diferencia entre el valor observado de y_i y el valor ajustado y_i^*:

$$y_i = y_i^* + e_i$$
$$e_i = y_i - y_i^* \rightarrow \text{Error o residuo.}$$

En el epígrafe siguiente se expone el procedimiento para elegir los parámetros a y b de la recta de tal forma que la línea de regresión realice el mejor ajuste a la nube de puntos. En concreto, para obtener estimaciones de los valores de "a" y "b" se aplica el método de los mínimos cuadrados.

6.3 REGRESIÓN LINEAL Y AJUSTE MÍNIMO-CUADRÁTICO. PREDICCIÓN

El método del ajuste por mínimos cuadrados consiste en elegir los parámetros "a" y "b" que minimizan la suma de los cuadrados de los errores, es decir, si para una muestra de N observaciones tuviéramos el escenario que se muestra en la tabla 6.1:

Tabla 6.1. Ejemplo de obtención de valores ajustados y residuos en una línea de regresión

$\mathbf{Y_i}$	$\mathbf{X_i}$	$\mathbf{y_i^* = a + bx_i}$	$\mathbf{e_i = Y_i - Y_i^*}$
y_1	x_1	$y_1^* = a + bx_1$	$e_1 = y_1 - y_1^* = y_1 - (a + bx_1)$
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
y_N	x_N	$y_N^* = a + bx_N$	$e_N = y_N - y_N^* = y_N - (a + bx_N)$

Fuente: Elaboración propia.

Se elegirían como valores de a y b aquellos que hagan mínima la siguiente expresión:

$$\underset{a,b}{\text{Min}} \sum_{i=1}^{N} e_i^2$$

Se puede demostrar que las condiciones de primer orden resultantes de minimizar la expresión anterior son:

$$\sum_{i=1}^{N} (y_i - a - bx_i) = 0 \rightarrow \sum_{i=1}^{N} e_i = 0$$

$$\sum_{i=1}^{N} (y_i - a - bx_i) * x_i = 0 \rightarrow \sum_{i=1}^{N} e_i x_i = 0$$

Resolviendo este sistema se obtiene que los valores de a y b se calculan a través de las siguientes expresiones:

$$b = \frac{S_{x,y}}{S_x^2}$$

$$a = \bar{Y} - b\bar{X}$$

donde:

$S_{x,y}$: Covarianza entre X e Y.

S_x^2: Varianza de X.

$\bar{Y}$: Media aritmética de Y.

$\bar{X}$: Media aritmética de X.

La interpretación de los parámetros a y b es la siguiente:

a) El parámetro "a" se incluye para flexibilizar la recta y no obligarla a pasar por el origen de coordenadas. En términos matemáticos, es el valor de Y^* cuando X es igual a cero, pero en la mayoría de los casos carece de interpretación socio-económica.

b) El parámetro "b" muestra en cuantas unidades varía Y, por término medio, cuando la variable X aumenta en una unidad. La variable Y puede aumentar o disminuir, dependiendo del signo del coeficiente b, que a su vez depende del signo de la covarianza $S_{x,y}$. El coeficiente b será positivo cuando exista una relación lineal positiva entre X e Y, y negativo cuando la relación lineal sea negativa.

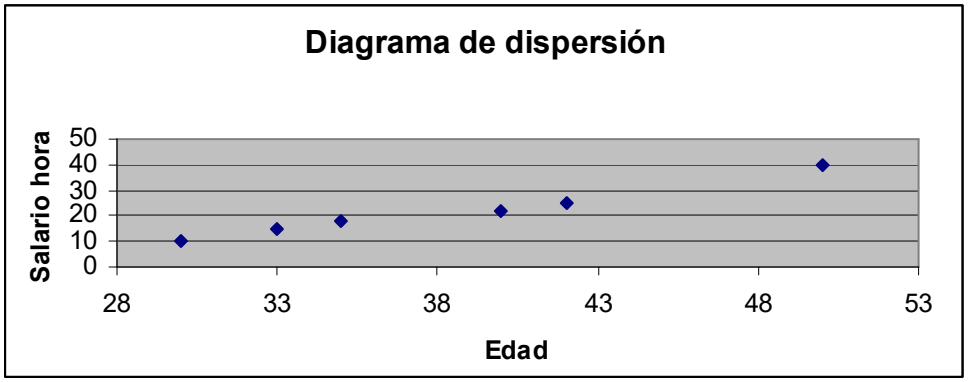
Ejemplo 6.1:

A continuación, se presentan los salarios (Y, €/hora) y la edad (X, años) de un grupo de 6 trabajadores, usados en los Ejemplos 5.6 y 5.8:

y_i	x_i
10	30
15	33
18	35
22	40
25	42
30	50

Se pide:

a) Obtenga el diagrama de dispersión:



b) Obtenga la línea de regresión.

y_i	x_i	x_i^2	$x_i y_i$
10	30	900	300
15	33	1089	495
18	35	1225	630
22	40	1600	880
25	42	1764	1050
30	50	2500	1500

$\bar{X} = \frac{\sum x_i n_i}{N} = 38{,}33\ años$

$\bar{Y} = \frac{\sum y_i n_i}{N} = 20$ €/h

$S_{x,y} = 42{,}5\ años * €$

$S_x^2 = 43{,}56\ (años)^2$

$b = \frac{S_{x,y}}{S_x^2} = 0{,}98$

$a = \bar{Y} - b * \bar{X} = -17{,}41$

$Y_i^* = -17{,}41 + 0{,}98 * X_i$

c) Interprete el valor de b.

$b = 0{,}98$: si aumenta en un año la edad entonces el salario €/hora se incrementa en 0,98 € por término medio.

* Coeficiente de Determinación:

Otro tema de interés relacionado con el análisis de regresión es valorar la bondad del ajuste, es decir, comprobar cómo de bien se ajusta la línea de regresión a la nube de puntos. Para ello, se utiliza el Coeficiente de Determinación, R^2. El Coeficiente de Determinación se define como la proporción de variancia de Y que es explicada por la variancia de la variable dependiente ajustada Y^*:

$$R^2 = \frac{S_{Y_*}^2}{S_Y^2}$$

Se puede demostrar que, si la línea de regresión contiene un término constante, la varianza de Y, S_Y^2 , es igual a la varianza explicada por la regresión $S_{Y^*}^2$ más la varianza residual $S_e^2 = \frac{\sum_{i=1}^{N} e_i^2}{N}$, es decir:

$$S_Y^2 = S_{Y^*}^2 + S_e^2$$

De lo anterior se deduce que el coeficiente de determinación se puede expresar como:

$$R^2 = \frac{S_{Y_*}^2}{S_Y^2} = 1 - \frac{S_e^2}{S_Y^2}$$

El coeficiente de determinación R^2 es adimensional, y toma valores entre 0 y 1:

$$0 \leq R^2 \leq 1$$

Cuanto más se acerque a 1 el coeficiente de determinación R^2 mejor será la bondad del ajuste. Si $R^2 = 1$, el ajuste es perfecto, si $R^2 = 0$ el ajuste es muy malo ya que el modelo no explica ninguna variación de la variable Y.

Si en el modelo de regresión lineal sólo existe una variable X que explique a Y puede establecerse una relación entre el coeficiente de determinación R^2 y el coeficiente de correlación lineal $r_{x,y}$ que facilita el cálculo del primero. La relación es la siguiente:

$$R^2 = r^2$$

Ejemplo 6.2:

Para el modelo del ejemplo 6.1 obtenga el coeficiente de determinación e interprételo:

$$R^2 = r^2 = \frac{S_{x,y}^2}{S_x^2 * S_y^2} = 0{,}96$$

Interpretación: El 96% de las variaciones de la variable dependiente son explicadas por el modelo.

* Predicción:

Por último, otras de las utilidades del análisis de regresión es la predicción, que consiste en obtener valores predichos de Y usando la relación estimada con el análisis de regresión. Para ello es necesario previamente proponer un valor para X que puede ser simulado o predicho mediante otras técnicas.

La capacidad predictiva del modelo o la fiabilidad de las predicciones será alta si su coeficiente de determinación está cercano a 1.

Así, con el modelo del ejemplo anterior se puede predecir el salario que recibiría un trabajador con una edad concreta, por ejemplo, ¿cuál sería el salario por hora en € predicho para un trabajador con 37 años?

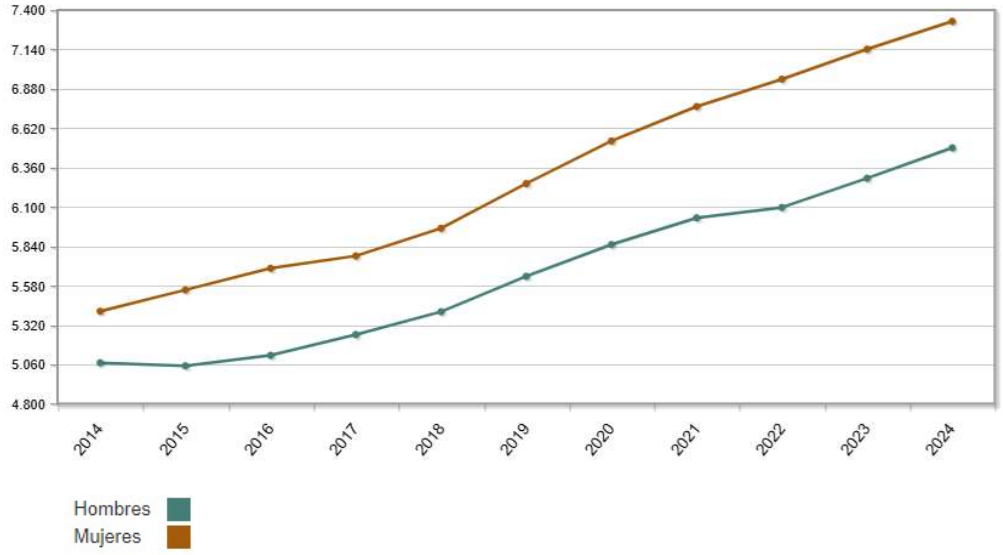
$$Y_i^* = -17{,}41 + 0{,}98 * 37 = 18{,}85€/h$$

y como $R^2 = 0{,}96$ se puede afirmar que la predicción es fiable.

6.4 ANALISIS CLÁSICO DE SERIES TEMPORALES. COMPONENTES.

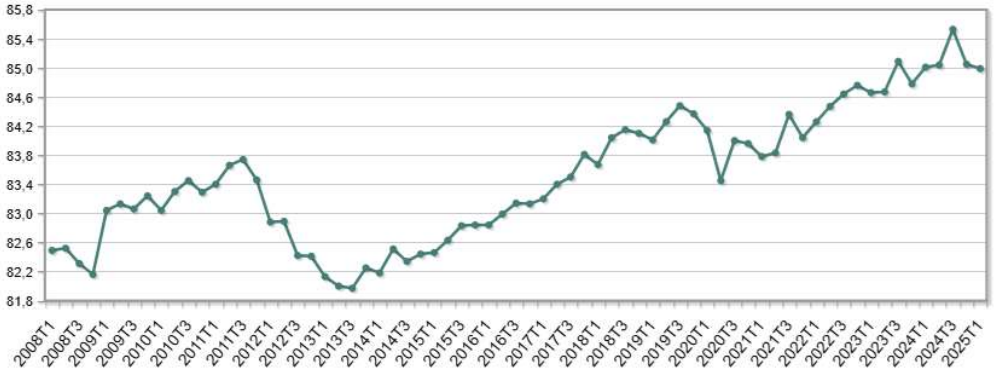
En las series temporales, la información relativa a las variables socioeconómicas está referida a un periodo temporal (mes, trimestre, año…). La observación a lo largo del tiempo de esas variables genera una sucesión de valores ordenados según el parámetro tiempo. Los ejemplos de estadísticas expresadas en series temporales en el ámbito socioeconómico son muy numerosos, a continuación, se presentan algunos de ellos (Figuras 6.2-6.4).

Figura 6.2. Población con estudios superiores según género (miles de personas)



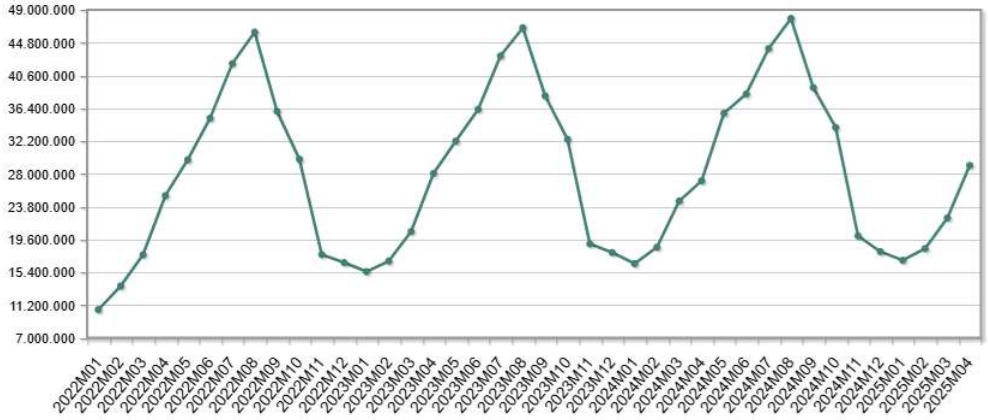
Fuente: INE (2025k).

Figura 6.3 Tasa de salarización en España



Fuente: INE (2025l).

Figura 6.4. Número de pernoctaciones en España



Fuente: INE (2025m).

Bajo un enfoque clásico de series temporales, la evolución de una variable Y_t puede ser explicada por las variaciones de las siguientes componentes: tendencia (T_t), ciclo (C_t), estacionalidad (E_t) y componente irregular (I_t). Esto conlleva que la variable Y_t se pueda expresar como una función de esas componentes:

$$Y_t = f(T_t, C_t, E_t, I_t)$$

Las hipótesis más utilizadas sobre cómo se interrelacionan los componentes son:

(a) Hipótesis aditiva:

$$Y_t = T_t + C_t + E_t + I_t$$

(b) Hipótesis multiplicativa:

$$Y_t = T_t * C_t * E_t * I_t$$

En la hipótesis aditiva se supone que los componentes son independientes; mientras que, en la hipótesis multiplicativa se asume que están interrelacionados, está última hipótesis es la más usual.

El significado de cada una de las componentes de la serie temporal es el siguiente:

a) Tendencia:

Refleja los movimientos a largo plazo de la variable originados, por ejemplo, por cambios tecnológicos, institucionales, sociales y demográficos.

b) Ciclo:

Muestra las fluctuaciones correspondientes al medio plazo, asociadas con el ciclo económico.

c) Estacionalidad:

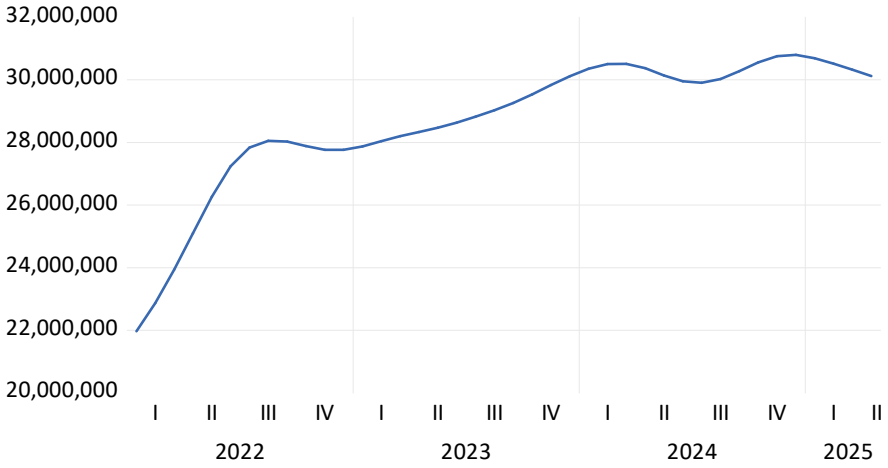
Representa las variaciones en el corto plazo, con una periodicidad inferior al año generadas, por ejemplo, por factores climáticos, tipo de estructura productiva o festividades.

d) Componente irregular:

Se corresponde con los cambios en el muy corto plazo que quedan fuera del control del analista, por ejemplo, una huelga.

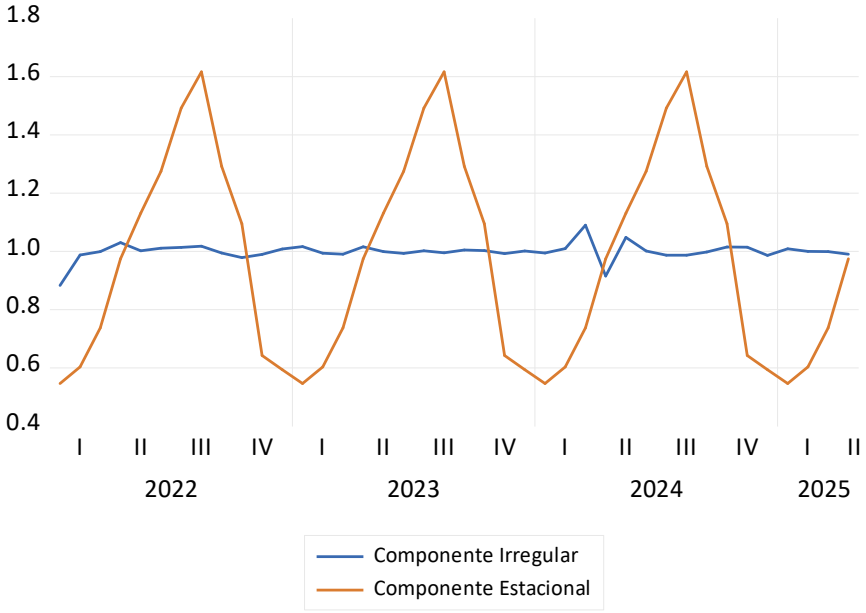
Dada una serie temporal, y mediante técnicas estadísticas ajenas al contenido del texto, es posible obtener de forma individualizada algunos de sus componentes. Por ejemplo, en la figura 6.5 se ofrece el componente tendencia-ciclo para la serie número de pernoctaciones en España, y en la figura 6.6 el componente estacional y el componente irregular, suponiendo una hipótesis multiplicativa, y usando el programa EViews, que es software usado para la estimación de modelos econométricos.

Figura 6.5. Componente tendencia y ciclo para el número de pernoctaciones en España



Fuente: Elaboración propia a partir del programa EViews.

Figura 6.6. Componente estacional y componente irregular del Nº de pernoctaciones en España



Fuente: Elaboración propia a partir del programa EViews.

6.5 COMPONENTE TENDENCIA. PREDICCIÓN

Una de las aplicaciones del análisis de regresión, consiste en suponer que la variable Y_t evoluciona en función de la variable tiempo, t, es decir:

$$Y_t = f(t)$$

En definitiva, bajo la perspectiva del análisis de regresión la variable X es sustituida por la variable tiempo. Si se supone que $f(t)$ es lineal, el componente de tendencia es una línea recta:

$$a + b * t$$

De tal forma, que:

$$Y_t^* = a + b * t$$

Todos los resultados obtenidos previamente para obtener los valores de "a" y "b" se pueden aplicar también ahora:

$$b = \frac{S_{t,y}}{S_t^2}$$

$$a = \bar{Y} - b\bar{X}$$

En este análisis, la variable t es generada asignando valores naturales consecutivos a cada una de las fechas (por ejemplo, a la primera fecha se le asigna el valor 1, 2 a la segunda, etc.). Además, hay que indicar que este enfoque tiene sentido cuando los datos son anuales o están desestacionalizados (es decir, datos con una periodicidad inferior al año sin el componente estacional).

Ejemplo 6.3:

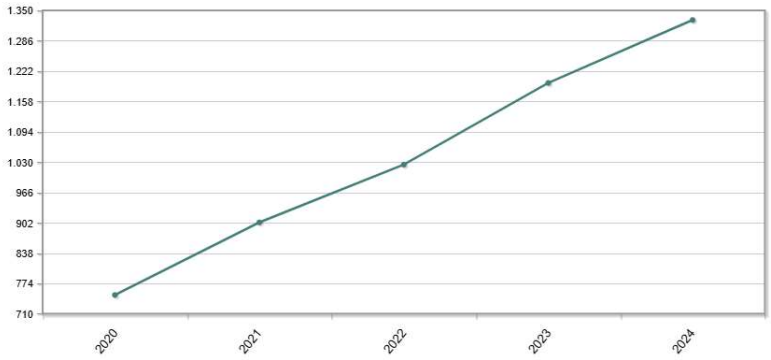
A partir de la serie correspondiente al número de ocupados procedentes de América Latina en España (Tabla 6.2 y Figura 6.7) ajuste una tendencia lineal, valore la bondad del ajuste y obtenga el valor predicho para el año 2025.

Tabla 6.1. Nº de ocupados en España procedentes de América Latina (miles de personas)

Año	Trabajadores (miles)
2020	749,5
2021	902,7
2022	1024,6
2023	1196,9
2024	1329,8

Fuente: Elaboración propia a partir de datos de la EPA (2025n).

Figura 6.6. Representación gráfica del Nº de ocupados en España procedentes de América Latina (miles de personas)



Fuente: Elaboración propia a partir de datos de la EPA (2025n).

*** Ajuste de regresión: Tendencia lineal.**

Año	Trabajadores (Y_i, miles)	t	t*y	t^2	Y_i^2
2020	749,5	1	749,5	1	561750,25
2021	902,7	2	1805,4	4	814867,29
2022	1024,6	3	3073,8	9	1049805,16
2023	1196,9	4	4787,6	16	1432569,61
2024	1329,8	5	6649	25	1768368,04

$\bar{t} = 3\ años; \bar{Y} = 1040{,}7 * 10^3\ ocupados$

$S_{t,y} = 290{,}96\ años * 10^3\ ocupados$

$S_t = 1{,}41\ años$

$S_y = 205{,}98 * 10^3\ ocupados$

$$b = \frac{S_{t,y}}{S_t^2} = 146{,}35$$

$$a = \bar{Y} - b * \bar{t} = 601{,}65$$

$$Y_t^* = a + b * t = 601{,}65 + 146{,}35 * t$$

Interpretación de "b": Cuando transcurre un año el número de ocupados procedentes de América Latina aumenta en 146,35 miles de personas, por término medio.

* Bondad del ajuste:

$$R^2 = r^2 = \frac{S_{t,y}^2}{S_t^2\ S_y^2} = 0{,}99$$

El 99% de las variaciones de la variable dependiente son explicadas por el modelo, la bondad del ajuste es muy buena.

* Predicción para el año 2025:

$$Y_{2025}^* = a + b * 6 = 601{,}65 + 146{,}35 * 6 = 1479{,}75 * 10^3\ ocupados$$

La predicción es fiable ya que la bondad del ajuste es alta.

Anexo 6.A. Ejemplo de obtención de una línea de Regresión con Excel

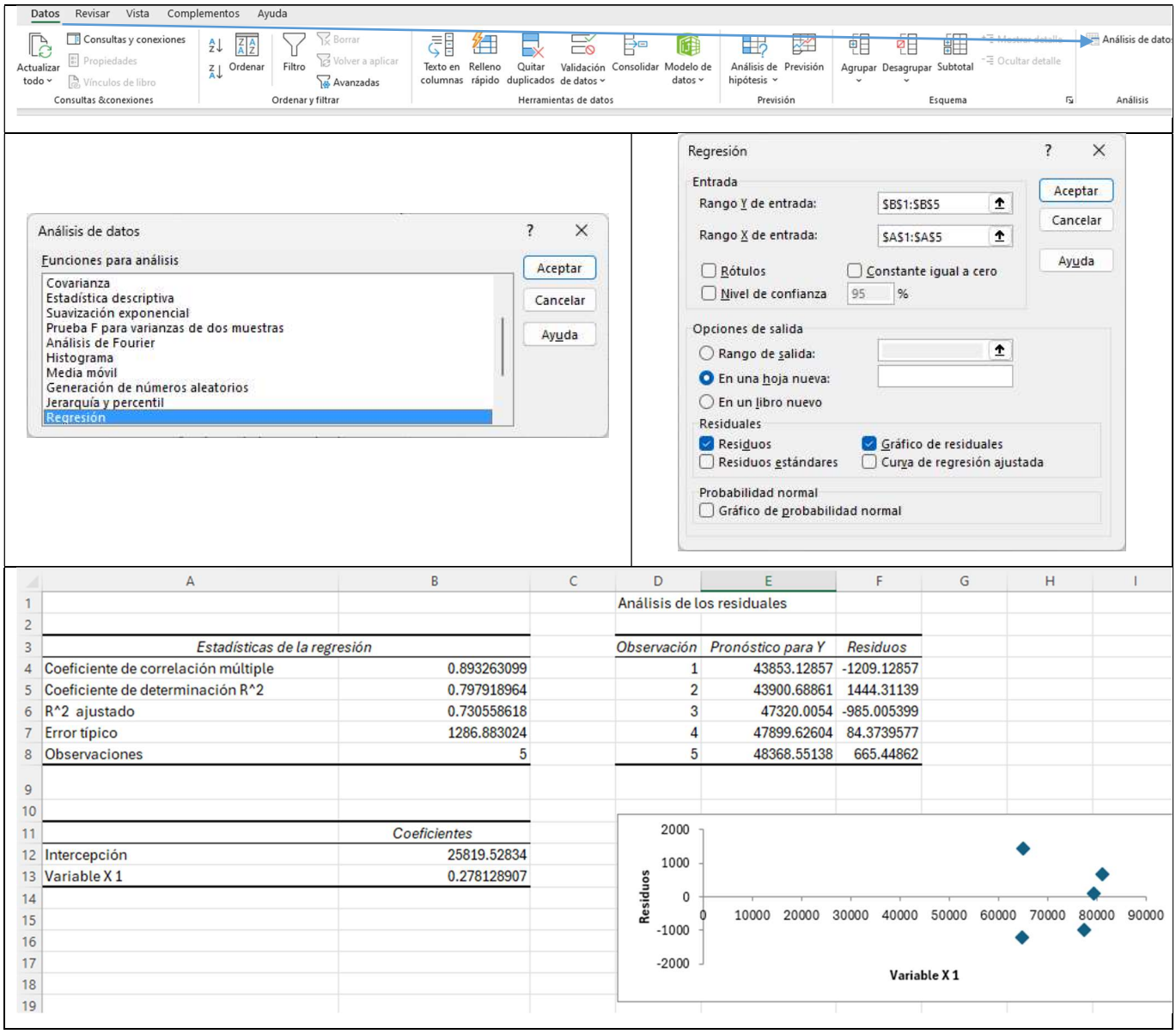
En este Anexo 6.A se plantea un análisis de regresión donde se relaciona el número de mujeres egresadas en másteres en España (X) con el número de mujeres matriculadas en el Doctorado (Y), el periodo de análisis abarca desde el curso 2017-2018 hasta el curso 2022-2023 (Ministerio de Ciencia, Innovación y Universidades, 2025). Establecer una relación entre ambas variables es mucho más complejo que el ejercicio que aquí se realiza, y abordar dicha complejidad nos llevaría al campo de la Econometría. Por ello, es necesario tener presente que este ejemplo es útil sólo a nivel ilustrativo para observar cómo se puede obtener una línea de regresión haciendo uso del programa Excel.

Una vez introducidos los datos en Excel, hay que hacer clic en "Datos", posteriormente en "Análisis de Datos", activar la opción "Regresión", y rellenar la información que se requiere. Esto nos permite obtener las estimaciones de los parámetros "a" y "b", el coeficiente de determinación R^2, y los gráficos de los residuos. Esta secuencia y un fragmento rediseñado de los resultados aparece en la figura 6.A.1. La línea de regresión estimada resultante es igual a:

$$y_i^* = 25819{,}528 + 0{,}278 * x_i$$

Los resultados revelan que un aumento del número de egresadas en másteres de 1000 personas, generaría un aumento del número de doctorandas de 278 personas, por término medio. La bondad del ajuste es buena, ya que el coeficiente de determinación es igual a 0,73 ($R^2 = 0{,}73$); es decir, el 73% de las variaciones del número de doctorandas son explicadas por el modelo. En cuanto a otros resultados, en la figura 6.A.1 aparecen dos columnas "Pronóstico para Y" que representa los valores ajustados por la línea de regresión Y_i^* y "Residuos" que equivale a $e_i = Y_i - Y_i^*$.

Figura 6.A.1. Línea de regresión con Excel



Fuente: Elaboración propia con Excel.

TEMA 7. MEDIDAS DE VARIACIÓN

7.1 INTRODUCCIÓN

En este tema se introducen una serie de instrumentos estadísticos que nos permiten analizar la evolución de variables expresadas como series temporales. En el primer lugar, en el epígrafe 7.2 se estudian las medidas estadísticas denominada "números índices simples" y "números índices en cadena". Posteriormente, el epígrafe 7.3 se dedica al caso particular del IPC (Indice de Precios al Consumo), que es útil para introducir los conceptos de inflación y deflación de magnitudes monetarias. Por último, se presenta de forma breve los problemas que surgen en la construcción de números índices.

7.2 INDICES SIMPLES E INDICES EN CADENA

Los números índices se introducen en Estadística para estudiar las fluctuaciones de una magnitud simple o compleja en función de uno de sus valores que se toma como término de comparación o referencia (base). Si las observaciones se corresponden con series temporales, las fluctuaciones se presentan al pasar de una unidad temporal a otra. La comparación se realiza por cociente, por lo que los números índices son adimensionales (carecen de unidad de medida). Los números índices simples son aquellos en los que se trata de estudiar la evolución de una única variable, es decir, de una única serie temporal. Según la base del índice sea fija o variable, se distinguen entre índices simples de base fija (se conocen como índices simples), e índices simples en cadena (se conocen como índices en cadena).

7.2.1 INDICES SIMPLES DE BASE FIJA

Para la obtención de un índice simple de base fija para un periodo t tomando como base el periodo 0, I_0^t, se divide el valor de la variable en el periodo t entre el valor en el periodo 0 (de referencia) y se multiplica por 100:

$$I_0^t = \frac{Y_t}{Y_0} * 100$$

En la expresión anterior, Y_0 representa el valor de la variable en el periodo de referencia (base), es decir, aquél con el que se quiere comparar. Con esta definición, el índice en el periodo considerado como base es igual a 100. La interpretación del índice simple de base fija en términos de tasa de variación es la siguiente: $(I_0^t - 100)$ es la tasa de variación porcentual de la variable Y en el periodo t con respecto al periodo 0. Esto se puede demostrar si desarrollamos la expresión matemática detrás de la tasa de variación TV_0^t:

$$TV_0^t = \left(\frac{Y_t - Y_0}{Y_0}\right) * 100 = \frac{Y_t}{Y_0} * 100 - 100 = I_0^t - 100$$

Ejemplo 7.1:

Se dispone de la serie de datos de ocupados en España con nacionalidad extranjera para el periodo 2020-2024.

Año	Ocupados extranjeros (miles de personas)
2020	2321,1
2021	2421,6
2022	2709,9
2023	3024,9
2024	3234

Fuente: INE (2025n).

Obtenga una serie de índices simples con base fija en el año 2020, e interprete el valor correspondiente al 2024.

Año	Ocupados extranjeros (miles de personas)	Indice: I_0^t (Base 2020)
2020	2321,1	100,00
2021	2421,6	104,33
2022	2709,9	116,75
2023	3024,9	130,32
2024	3234	139,33

$$I_{2020}^{2020} = \frac{Y_{2020}}{Y_{2020}} * 100 = 100 \qquad I_{2020}^{2021} = \frac{Y_{2021}}{Y_{2020}} * 100 = 104{,}33 \qquad I_{2020}^{2022} = \frac{Y_{2022}}{Y_{2020}} * 100 = 116{,}75$$

$$I_{2020}^{2023} = \frac{Y_{2023}}{Y_{2020}} * 100 = 130{,}32 \qquad I_{2020}^{2024} = \frac{Y_{2024}}{Y_{2020}} * 100 = 139{,}33$$

$$TV_{2020}^{2024} = \left(\frac{Y_{2024} - Y_{2020}}{Y_{2020}}\right) * 100 = I_{2020}^{2024} - 100 = 39{,}33\ \%$$

El número de ocupados extranjeros aumentó en un 5,33% entre los años 2020 y 2024,

7.2.2 INDICES EN CADENA

Los índices en cadena son índices donde el periodo base es cambiante, de forma que las comparaciones son siempre de un valor con respecto al que le precede en el tiempo. El índice en cadena para el periodo t, denominado IC_t es el resultado de dividir el valor de la variable en el periodo t entre el valor del periodo anterior t-1 y multiplicar por 100:

$$IC_t = \frac{Y_t}{Y_{t-1}} * 100$$

$IC_t - 100$ es la tasa de variación porcentual de la variable en el periodo t con respecto al periodo t-1:

$$TV_{t-1}^{t} = \left(\frac{Y_t - Y_{t-1}}{Y_{t-1}}\right) * 100 = \frac{Y_t}{Y_{t-1}} * 100 - 100 = IC_t - 100$$

Ejemplo 7.2:
A partir de los datos de ocupados en España con nacionalidad extranjera para el periodo 2020-2024, obtenga la serie de índices en cadena e interprete el valor correspondiente al 2024.

Año	Ocupados extranjeros (miles de personas)	Indice en cadena
2020	2321,1	-
2021	2421,6	104,33
2022	2709,9	111,91
2023	3024,9	111,62
2024	3234	106,91

$IC_{2021} = \frac{Y_{2021}}{Y_{2020}} * 100 = 104{,}33$ $\qquad$ $IC_{2022} = \frac{Y_{2022}}{Y_{2021}} * 100 = 111{,}91$

$IC_{2023} = \frac{Y_{2023}}{Y_{2022}} * 100 = 111{,}62$ $\qquad$ $IC_{2024} = \frac{Y_{2024}}{Y_{2023}} * 100 = 106{,}91$

Interpretación de $IC_{2024} = \frac{Y_{2024}}{Y_{2023}} * 100 = 106{,}91$.

$$TV_{2023}^{2024} = \left(\frac{Y_{20240} - Y_{2023}}{Y_{2023}}\right) * 100 = IC_{2024} - 100 = 6{,}91\%$$

El número de ocupados extranjeros se incrementó en un 6,91% entre los años 2023 y 2024.

Finalmente, otro tema de interés es la relación que existe entre los índices simples y los índices en cadena. En concreto, un índice en cadena se puede expresar en función de índices simples:

$$IC_t = \frac{Y_t}{Y_{t-1}} * 100 = \frac{\frac{Y_t}{Y_0} * 100}{\frac{Y_{t-1}}{Y_0} * 100} * 100 = \frac{I_0^t}{I_0^{t-1}} * 100$$

Por consiguiente, el índice simple se puede calcular como:

$$I_0^t = \frac{IC_t * I_0^{t-1}}{100}$$

7.3 INDICES DE PRECIOS AL CONSUMO

El Índice de Precios de Consumo (IPC) es una medida estadística de la evolución del conjunto de precios de los bienes y servicios que consume la población residente en viviendas familiares en España. La periodicidad de este índice de precios es mensual. A partir de 1 de enero de 2002 la metodología del IPC se renovó completamente. En el nuevo sistema se utiliza la información procedente de la Encuesta Continua de Presupuestos Familiares (ECPF) para determinar la composición de la cesta de la compra y las ponderaciones de los bienes y servicios. En particular, se tienen en cuenta todas aquellas parcelas de gasto que superan el 0,3 por mil del gasto total. Además, en el nuevo sistema se producen revisiones anuales de las ponderaciones de los bienes y servicios que componen la cesta de la compra del IPC, y revisiones quinquenales que implican un cambio de base del índice.

El último cambio de base del IPC ha sido en 2021. Algunos cambios destacables en la configuración de la cesta de la base 2021 fueron la incorporación de las mascarillas higiénicas y las suscripciones a los periódicos on-line. Por otro lado, algunos ejemplos de artículos que desaparecieron de la cesta fueron el reproductor de imagen, el reproductor portátil o el compact disc o el DVD. En la tabla 7.1 se muestran las ponderaciones de los distintos bienes en la cesta de la compra para los años 2024 y 2025 expresadas en tanto por mil.

Tabla 7.1 Ponderaciones de los bienes en el IPC: años 2024 y 2025

Ponderaciones	**2024**	**2025**
Índice general	1000,00	1000,00
Alimentos y bebidas no alcohólicas	191,603	185,386
Bebidas alcohólicas y tabaco	38,495	37,719
Vestido y calzado	39,465	39,680
Vivienda, agua, electricidad, gas y otros combustibles	119,965	121,584
Muebles, artículos del hogar	53,457	52,756
Sanidad	57,867	57,182
Transporte	143,783	143,931
Comunicaciones	33,535	32,609
Ocio y cultura	85,888	85,624
Enseñanza	18,778	18,723
Restaurantes y hoteles	139,304	147,513
Otros bienes y servicios	77,859	77,294

Fuente: Elaboración propia a partir de datos del INE (INE, 2025ñ).

En la tabla anterior se constata que las cuatro categorías de bienes con más peso en el IPC son: "Alimentos y bebidas no alcohólicas", "Transporte", "Vivienda, agua, electricidad, gas y otros combustibles", y "Restaurantes y hoteles". Por otro lado, la categoría con menor peso es la "Enseñanza".

El IPC es un instrumento adecuado para medir la inflación y para deflactar series de valores monetarios. También se utiliza como punto de referencia en la revisión de los contratos de arrendamiento de inmuebles, primas de seguros, en la negociación salarial, y en la actualización de pensiones.

A continuación, nos centramos en los conceptos de inflación y deflación de valores monetarios.

* Inflación:

La inflación es la subida generalizada y persistente de los precios de los bienes y servicios que se producen y se consumen. La inflación en España se mide a través del IPC. En particular, la tasa de inflación entre dos momentos de tiempo es la tasa de variación del IPC entre esos dos instantes:

$$\pi_{t_0}^{t_1} = \left(\frac{IPC_{t_1} - IPC_{t_0}}{IPC_{t_0}}\right) * 100$$

A partir de esta definición general, surgen las siguientes tasas:

a) Tasa de inflación intermensual.

Registra la variación de los precios en un mes, con respecto al mes anterior, y se calcula como:

$$\pi_{t-1}^{t} = \left(\frac{IPC_t - IPC_{t-1}}{IPC_{t-1}}\right) * 100$$

b) Tasa de inflación interanual.

Mide la variación de los precios en un mes con respecto al mismo mes del año anterior:

$$\pi_{t-12}^{t} = \left(\frac{IPC_t - IPC_{t-12}}{IPC_{t-12}}\right) * 100$$

c) Tasa de inflación acumulada.

Cuantifica la variación de los precios en un mes con respecto a diciembre del año anterior y se obtiene de la siguiente forma:

$$\pi_{diciembre,año\ anterior}^{t} = \left(\frac{IPC_t - IPC_{diciembre,año\ anterior}}{IPC_{diciembre,año\ anterior}}\right) * 100$$

Cada mes el INE informa sobre las variaciones del IPC mensual, y anual. Por ejemplo, en la tabla 7.2 aparece la información para el mes de mayo del 2025:

Tabla 7.2. Variación mensual y anual del IPC: mayo 2025

	Variación mensual	Variación anual
Indice general	0,1	2,0
Inflación subyacente	0,1	2,2

Fuente: INE (2025ñ)

La primera fila de la tabla 7.2 registran las tasas de inflación intermensual e interanual, respectivamente. Por otro lado, la segunda fila hace referencia a la inflación subyacente. Dicho concepto mide la variación de los precios excluyendo los precios de la energía y de los alimentos no elaborados; es decir, no tiene en cuenta los componentes más volátiles de la cesta de la compra.

Si nos centramos en el Índice General (es decir el dato más agregado del IPC), en la tabla 7.3 se recoge su evolución durante el periodo 2024:01-2025:05.

Tabla 7.3. Evolución del IPC entre 2024:01-2025:05

Año/Mes	IPC
2024:01	113,404
2024:02	113,807
2024:03	114,674
2024:04	115,472
2024:05	115,776
2024:06	116,212
2024:07	115,660
2024:08	115,707
2024:09	115,009
2024:10	115,726
2024:11	116,010
2024:12	116,534
2025:01	116,733
2025:02	117,191
2025:03	117,260
2025:04	117,997
2025:05	118,077

Fuente: Elaboración propia a partir de datos del INE (2025ñ).

A partir de estos datos, es posible obtener los diversos tipos de tasas de inflación que hemos mencionado previamente. Por ejemplo, el cálculo de la tasa de variación intermensual de la inflación para mayo del 2025 es el siguiente:

$$\pi_{abril,2025}^{mayo,2025} = \left(\frac{IPC_{mayo,2025} - IPC_{abril,2025}}{IPC_{abril,2025}} - 1\right) * 100 = \left(\frac{118{,}077 - 117{,}997}{117{,}997}\right) * 100 = 0{,}07\%$$

Interpretación: Los precios aumentaron un 0,07% entre abril y mayo del 2025.

En esta interpretación se entiende que los precios son los correspondientes a la cesta de bienes y servicios incorporada en la elaboración del IPC.

En segundo lugar, la tasa de variación interanual de la inflación para mayo del 2025 es igual a:

$$\pi_{mayo,2024}^{mayo,2025} = \left(\frac{IPC_{mayo,2025} - IPC_{mayo,2024}}{IPC_{mayo,2024}}\right) * 100 = \left(\frac{118{,}077 - 115{,}776}{115{,}776}\right) * 100 = 1{,}99\%$$

Interpretación: Los precios aumentaron un 1,99% entre mayo del 2024 y mayo del 2025.

Por otro lado, el cálculo de la tasa de inflación acumulada para mayo del 2025 sería la siguiente:

$$\pi_{diciembre,2024}^{mayo,2025} = \left(\frac{IPC_{mayo,2025} - IPC_{diciembre,2024}}{IPC_{diciembre,2024}}\right) * 100 = \left(\frac{118{,}077 - 116{,}534}{116{,}534}\right) * 100 = 1{,}32\%$$

Interpretación: Los precios han registrado un aumento del 1,32% entre diciembre del 2024 y mayo del 2025.

Finalmente, hay que señalar que el INE también ofrece el IPC en términos anuales. En la tabla 7.4 se muestra el IPC anual para el periodo 2016-2024.

Tabla 7.4. IPC: 2016-2024

Año	IPC
2016	93,222
2017	95,046
2018	96,638
2019	97,314
2020	97
2021	100
2022	108,391
2023	112,219
2024	115,333

Fuente: Elaboración propia a partir de datos del INE (INE, 2025ñ).

A partir de esta información, por ejemplo, es posible obtener:

* Tasa de inflación interanual para el 2024:

$$\pi_{2023}^{2024} = \left(\frac{IPC_{2024} - IPC_{2023}}{IPC_{2023}}\right) * 100 = \left(\frac{115{,}333 - 112{,}219}{112{,}219}\right) * 100 = 2{,}77\%$$

Interpretación: Los precios aumentaron un 2,77% entre los años 2023 y 2024.

* Tasa de inflación para un intervalo superior al año, por ejemplo, entre 2016 y 2024:

$$\pi_{2016}^{2024} = \left(\frac{IPC_{2024} - IPC_{2016}}{IPC_{2016}}\right) * 100 = \left(\frac{115{,}333 - 93{,}222}{93{,}222}\right) * 100 = 23{,}72\%$$

Interpretación: Entre los años 2016 y 2024, los precios crecieron un 23,72%.

* Deflación de valores monetarios:

La operación de deflación de series monetarias consiste en eliminar el efecto que los cambios en los precios de los bienes (la inflación) tiene sobre las series de valores monetarios. Cuando queremos conocer la evolución de una serie de valores a lo largo del tiempo (como por ejemplo una serie de salarios) lo habitual es que aparezcan expresados en unidades monetarias de cada periodo, es decir, en unidades monetarias corrientes. Esto implica que los valores no son directamente comparables puesto que el efecto de la inflación modifica la capacidad de compra del dinero, lo que se conoce como poder adquisitivo. Por este motivo, es interesante obtener una serie sin el efecto de la inflación, a este proceso se le denomina deflación de valores monetarios y su finalidad es la obtención de una serie de valores expresados en las unidades monetarias de un solo periodo, es decir, en unidades monetarias constantes. El resultado final del proceso de deflación será una serie de valores expresados en términos reales, que nos permita conocer la evolución del poder adquisitivo.

Para obtener una serie de valores monetarios (salarios, rentas de alquileres, prestaciones…) en términos reales o deflactada hay que dividir la serie expresada en términos corrientes por el IPC y multiplicarla por 100:

$$Y_t^d = \frac{Y_t}{IPC_t} * 100$$

La nueva serie vendrá dada en unidades monetarias (por ejemplo, euros) del periodo base del IPC. Si se quiere que la serie esté expresada en unidades monetarias de un periodo diferente al correspondiente a la base del IPC, es necesario realizar el oportuno cambio de base del IPC a ese periodo antes de realizar la operación de deflación.

Ejemplo 7.3:

Los salarios medios que una empresa ha pagado mensualmente a sus empleados durante el periodo 2016-2021 aparecen en la tabla adjunta. Además, se informa del IPC correspondiente a esos años.

Año	**Salarios (€ corrientes)**	**IPC (Base: 2021)**
2016	950	93,222
2017	987	95,046
2018	1039	96,638
2019	1113	97,314
2020	1200	97
2021	1250	100,000

a) Obtenga la serie de los salarios en euros constantes del 2021.

Año	**Salarios (€ corrientes)**	**IPC (Base: 2021)**	**Salarios € del 2021**
2016	950	93,222	1019,07
2017	987	95,046	1038,44
2018	1039	96,638	1075,14
2019	1113	97,314	1143,72
2020	1200	97	1237,11
2021	1250	100,000	1250

b) Obtenga la serie de los salarios en euros constantes del 2016.

Año	**Salarios (€ corrientes)**	**IPC (Base: 2021)**	**IPC (Base: 2016)**	**Salarios € del 2016**
2016	950	93,222	100	950
2017	987	95,046	101,956	968,06
2018	1039	96,638	103,664	1002,27
2019	1113	97,314	104,389	1066,20
2020	1200	97	104,052	1153,27
2021	1250	100,000	107,271	1165,27

c) Obtenga el incremento de los salarios en términos reales entre los años 2016 y 2021, e interprete el resultado:

Año	**Salarios € del 2021**	$\boldsymbol{I_{2016}^{2021}}$	$\boldsymbol{I_{2016}^{2021} - 100}$
2016	1019,07	100	-
2021	1250	122,66	22,66

$$TV_{2016}^{2021} = I_{2016}^{2021} - 100 = 22{,}66\%$$

Los salarios en términos reales aumentaron en un 22,66% entre los años 2016 y 2021.

d) Obtenga la serie de índices en cadena para los salarios en términos reales. e interprete el dato correspondiente al año 2021:

Año	Salarios € del 2021	IC_t
2016	1019,07	-
2017	1038,44	101,90
2018	1075,14	103,53
2019	1143,72	106,38
2020	1237,11	108,16
2021	1250	101,04

$TV_{2020}^{2021} = IC_{2021} - 100 = 1{,}04\%$. Los salarios en términos reales aumentaron en un 1,04% entre los años 2020 y 2021.

7.4 PROBLEMAS EN LA CONSTRUCCIÓN DE NÚMEROS ÍNDICES

En esta sección se señalan de forma esquemática algunos de los problemas que pueden encontrarse en la elaboración de números índices:

a) Selección de variables:

En la elaboración de un índice complejo es importante seleccionar las variables adecuadas que permitan que las conclusiones que se extraigan sean representativas del fenómeno objeto de estudio. Por ejemplo, como se ha mencionado anteriormente, en la elaboración del IPC es necesario seleccionar cuáles son los productos que representan a la cesta del consumidor y sus ponderaciones.

b) Selección de los lugares y tiempos de observación:

Tras seleccionar las variables que conforman el índice, hay que obtener los datos numéricos que se van a utilizar para su cálculo. En esta segunda etapa es necesario seleccionar los lugares y los tiempos de observación. Por ejemplo, en la observación de los precios de los productos hay que decidir si se observa el precio de producción o el de venta, y además a qué periodo corresponde ese precio (precio de un día, de una semana, de una quincena…). En el caso del IPC, tradicionalmente la recogida de precios se ha realizado en los establecimientos con mayor afluencia de público y mediante visitas personales. No obstante, en los últimos años se está introduciendo el procedimiento "Scanner Data", que consiste en usar las bases de datos de las empresas como sustitutas de la visita personal.

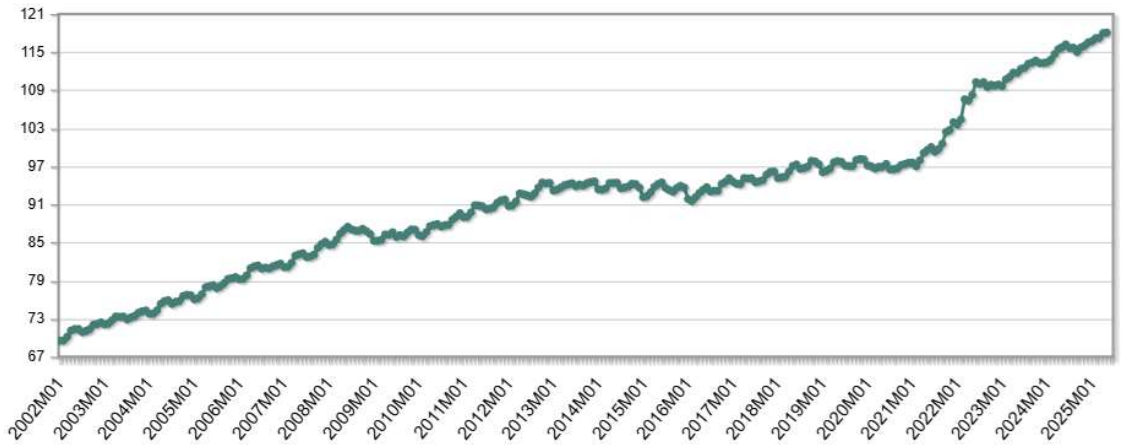
c) Selección del tiempo tomado como base de referencia:

El tiempo base debe de ser normal, es decir, no hay que elegir un periodo con un valor anómalo. Además, no debe de estar muy alejado del tiempo actual para que la comparación tenga utilidad. En este sentido, cabe recordar que el año base del IPC cambia cada cinco años, con el fin de recoger los cambios en los patrones de consumo de bienes y servicios de los individuos.

d) Enlace de índices nuevos con los antiguos:

La renovación de un índice genera el problema de la ruptura de continuidad en la serie temporal definida por el índice. Para solucionar este problema se procede al enlace de índices. En este sentido, el INE genera coeficientes de enlace calculados como el cociente del IPC en un mismo periodo para dos bases diferentes. Esto permite, por ejemplo, obtener en el INE una serie del IPC con base en el año 2021 desde el año 2002:

Figura 7.1. Evolución del IPC (Base 2021): 2002:01-2025:06



Fuente: INE (2025ñ).

e) Cambio de base:

Por último, cabe señalar que para realizar algunas comparaciones a veces conviene modificar o cambiar la base de índices ya calculados. Basta simplemente con hacer igual a 100 la cifra correspondiente al tiempo que se desee tomar como nueva base y transformar proporcionalmente la serie.

Anexo 7.A. Obtención de índices simples y en cadena, y deflación de series monetarias

Este anexo 7.A se exponen las figuras 7.A.1 y 7.A.2 que replican los resultados de los Ejemplos 7.1 y 7.3 con Excel. En particular, se muestran las fórmulas que habría que utilizar y los posteriores resultados numéricos.

Tabla 7.4.1. Ejemplo de obtención de índices simples y en cadena con Excel

- Fórmulas

	A	B	C
1	**Año**	**Ocupados**	
2	2020	2321.1	
3	2021	2421.6	
4	2022	2709.9	
5	2023	3024.9	
6	2024	3234	
7			
8			
9			
10	**Año**	**Indice (Base 2020)**	**Indice (Base 2020) menos 100**
11	2020	=+(B2/$B$2)*100	
12	2021	=+(B3/$B$2)*100	=+B12-100
13	2022	=+(B4/$B$2)*100	=+B13-100
14	2023	=+(B5/$B$2)*100	=+B14-100
15	2024	=+(B6/$B$2)*100	=+B15-100
16			
17	**Año**	**Indice en Cadena**	**Indice en Cadena menos 100**
18	2021	=+(B12/B11)*100	=+B18-100
19	2022	=+(B13/B12)*100	=+B19-100
20	2023	=+(B14/B13)*100	=+B20-100
21	2024	=+(B15/B14)*100	=+B21-100

- Resultados

10	**Año**	**Indice (Base 2020)**	**Indice (Base 2020) menos 100**
11	2020	100.00	
12	2021	104.33	4.33
13	2022	116.75	16.75
14	2023	130.32	30.32
15	2024	139.33	39.33
16			
17	**Año**	**Indice en Cadena**	**Indice en Cadena menos 100**
18	2021	104.33	4.33
19	2022	111.91	11.91
20	2023	111.62	11.62
21	2024	106.91	6.91

Fuente: Elaboración propia con Excel.

Tabla 7.4.2. Ejemplo de deflación de series monetarias con Excel

- Fórmulas:

	A	B	C	D	E
1	**Obtención de la serie de salarios en euros del 2021**				
2					
3		**Salarios**	**IPC**		
4	**Año**	**(€ corrientes)**	**(Base 2021)**	**Salarios (€ del 2021)**	
5	2016	950	93.222	=+(B5/C5)*100	
6	2017	987	95.046	=+(B6/C6)*100	
7	2018	1039	96.638	=+(B7/C7)*100	
8	2019	1113	97.314	=+(B8/C8)*100	
9	2020	1200	97	=+(B9/C9)*100	
10	2021	1250	100	=+(B10/C10)*100	
11					
12	**Obtención de la serie de salarios en euros del 2016**				
13					
14		**Salarios**	**IPC**	**IPC**	
15	**Año**	**(€ corrientes)**	**(Base 2021)**	**(Base 2016)**	**Salarios (€ del 2021)**
16	2016	950	93.222	=+(C16/$C$16)*100	=+(B16/D16)*100
17	2017	987	95.046	=+(C17/$C$16)*100	=+(B17/D17)*100
18	2018	1039	96.638	=+(C18/$C$16)*100	=+(B18/D18)*100
19	2019	1113	97.314	=+(C19/$C$16)*100	=+(B19/D19)*100
20	2020	1200	97	=+(C20/$C$16)*100	=+(B20/D20)*100
21	2021	1250	100	=+(C21/$C$16)*100	=+(B21/D21)*100

- Resultados:

	A	B	C	D	E
1	**Obtención de la serie de salarios en euros del 2021**				
2					
3		**Salarios**	**IPC**		
4	**Año**	**(€ corrientes)**	**(Base 2021)**	**Salarios (€ del 2021)**	
5	2016	950	93.222	1019.07	
6	2017	987	95.046	1038.44	
7	2018	1139	96.638	1178.63	
8	2019	1313	97.314	1349.24	
9	2020	1476	97	1521.65	
10	2021	1622	100	1622.00	
11					
12	**Obtención de la serie de salarios en euros del 2016**				
13					
14		**Salarios**	**IPC**	**IPC**	**Salarios**
15	**Año**	**(€ corrientes)**	**(Base 2021)**	**(Base 2016)**	**(€ del 2016)**
16	2016	950	93.222	100.000	950.000
17	2017	987	95.046	101.957	968.059
18	2018	1139	96.638	103.664	1098.738
19	2019	1313	97.314	104.390	1257.789
20	2020	1476	97	104.053	1418.512
21	2021	1622	100	107.271	1512.061

Fuente: Elaboración propia con Excel.

TEMA 8. ALGUNAS FUENTES ESTADÍSTICAS EN LA INVESTIGACIÓN SOCIAL

8.1 INDICADORES SOCIALES

Como se indicó al inicio del tema 1 una gran parte de la producción estadística realizada en España es elaborada por el Instituto Nacional de Estadística (INE). En relación a los indicadores sociales, la base de datos del INE ofrece información relevante dentro de la sección "Sociedad" sobre los siguientes temas: "Educación y Cultura", "Salud", "Seguridad y Justicia" y "Análisis Sociales".

Los contenidos elaborados de forma periódica que podemos encontrar en la sección de "Sociedad" se muestran en Tabla 8.1.

Tabla 8.1 Algunos contenidos de la sección Sociedad del INE

Educación y Cultura

Operaciones estadísticas que el INE elabora de forma periódica	Últimos datos
Encuesta de inserción laboral de titulados universitarios	Año 2019
Encuesta sobre la participación de la población adulta en las actividades de aprendizaje	Año 2022
Encuesta de financiación y gastos de la enseñanza privada	Curso 2020/2021
Encuesta sobre el gasto de los hogares en educación	Curso 2019-2020
Encuesta de transición educativa-formativa e inserción laboral	Año 2019

Salud

Operaciones estadísticas que el INE elabora de forma periódica	Últimos datos
Estimación del número de defunciones semanales	Semana 18 2025
El empleo de las personas con discapacidad	Año 2023
Encuesta de morbilidad hospitalaria	Año 2023
Encuesta de salud de España	Año 2023
Estadística de defunciones según la causa de muerte	Provisionales 1S/2024 y año 2023
Estadística de profesionales sanitarios colegiados	Año 2024
Estadística del salario de las personas con discapacidad	Año 2022

Seguridad y justicia

Operaciones estadísticas que el INE elabora de forma periódica	Últimos datos
Estadística de condenados: Adultos	Año 2023
Estadística de condenados: Menores	Año 2023
Estadística de nulidades, separaciones y divorcios	Año 2023
Estadística de violencia doméstica y violencia de género	Año 2024

Análisis Sociales

Operaciones estadísticas que el INE elabora de forma periódica	Últimos datos
La vida de las mujeres y hombres en Europa	Edición 2020
Mujeres y hombres en España	Edición 2024

Fuente: INE (2025o).

Por otro lado, dentro del contexto de los indicadores sociales, en la página web del INE es posible observar la evolución los Indicadores de la Agenda 2030 para el Desarrollo Sostenible. En concreto, son 231 indicadores distribuidos entre 17 objetivos y 169 metas, cuya actualización es una operación estadística programada de forma anual. En la figura 8.1 se presentan los 17 objetivos propuestos por la Agenda 2030.

Figura 8.1. Objetivos de la Agenda 2030 para el Desarrollo Sostenible

Fuente: INE (2025p).

Si accedemos al objetivo número 1 "Poner fin a la pobreza en todas sus formas y en todo el mundo" aparecen una serie de metas a alcanzar para el año 2030, que aparecen plasmadas en la tabla 8.2. Dentro de cada meta, se incorporan sus correspondientes indicadores. Por ejemplo, en la Meta 1.3 "Implementar a nivel nacional sistemas y medidas apropiados de protección social para todos, incluidos niveles mínimos, y, de aquí a 2030, lograr una amplia cobertura de las personas pobres y vulnerables" se incluye el siguiente indicador 1.3.1. "Proporción de la población cubierta por sistemas o niveles mínimos de protección social, desglosada por sexo, distinguiendo entre los niños, los desempleados, los ancianos, las personas con discapacidad, las mujeres embarazadas, los recién nacidos, las víctimas de accidentes de trabajo, los pobres y los vulnerables", que a su vez contiene 4 subindicadores. En la tabla 8.3 se exponen algunos resultados sobre estos subindicadores relativos a la situación de partida (año 2015) y el último año con información actualizada.

Tabla 8.2. Metas incluidas dentro del objetivo "Igualdad de Género" de la Agenda 2030

Meta 1.1. Erradicar para todas las personas y en todo el mundo la pobreza extrema (actualmente se considera que sufren pobreza extrema las personas que viven con menos de 1,25 dólares de los Estados Unidos al día).
Meta 1.2. Reducir al menos a la mitad la proporción de hombres, mujeres y niños de todas las edades que viven en la pobreza en todas sus dimensiones con arreglo a las definiciones nacionales.
Meta 1.3. Implementar a nivel nacional sistemas y medidas apropiados de protección social para todos, incluidos niveles mínimos, y, de aquí a 2030, lograr una amplia cobertura de las personas pobres y vulnerables.
Meta 1.4. Garantizar que todos los hombres y mujeres, en particular los pobres y los vulnerables, tengan los mismos derechos a los recursos económicos y acceso a los servicios básicos, la propiedad y el control de la tierra y otros bienes, la herencia, los recursos naturales, las nuevas tecnologías apropiadas y los servicios financieros, incluida la microfinanciación.
Meta 1.5. Fomentar la resiliencia de los pobres y las personas que se encuentran en situaciones de vulnerabilidad y reducir su exposición y vulnerabilidad a los fenómenos extremos relacionados con el clima y otras perturbaciones y desastres económicos, sociales y ambientales.
Meta 1.a. Garantizar una movilización significativa de recursos procedentes de diversas fuentes, incluso mediante la mejora de la cooperación para el desarrollo, a fin de proporcionar medios suficientes y previsibles a los países en desarrollo, en particular los países menos adelantados, para que implementen programas y políticas encaminados a poner fin a la pobreza en todas sus dimensiones.
Meta 1.b. Crear marcos normativos sólidos en los planos nacional, regional e internacional, sobre la base de estrategias de desarrollo en favor de los pobres que tengan en cuenta las cuestiones de género, a fin de apoyar la inversión acelerada en medidas para erradicar la pobreza.

Fuente: INE (2025p).

Tabla 8.3. Información relativa a los subindicadores de la Meta 1.1 de la Agenda 2030

Subindicador 1.3.1.1. Proporción de personas mayores que reciben una pensión	Último periodo: 2023 **86,56%** Unidad: Porcentaje Dato base: 86,67 Periodo base: 2015
Subindicador 1.3.1.2. Proporción de personas con discapacidad que reciben prestaciones	Último periodo: 2023 **39,59%** Unidad: Porcentaje Dato base: 46,04 Periodo base: 2015
Subindicador 1.3.1.3. Proporción de desempleados que reciben prestaciones	Último periodo: 2024 **72,60%** Unidad: Porcentaje Dato base: 55,80 Periodo base: 2015
Subindicador 1.3.1.4. Proporción de trabajadores cubiertos en caso de lesiones laborales	Último periodo: 2022 **98,36%** Unidad: Porcentaje Dato base: 83,68 Periodo base: 2015

Fuente: INE (2025p).

8.2 LA PRODUCCIÓN ESTADÍSTICA NACIONAL, AUTONÓMICA Y LOCAL

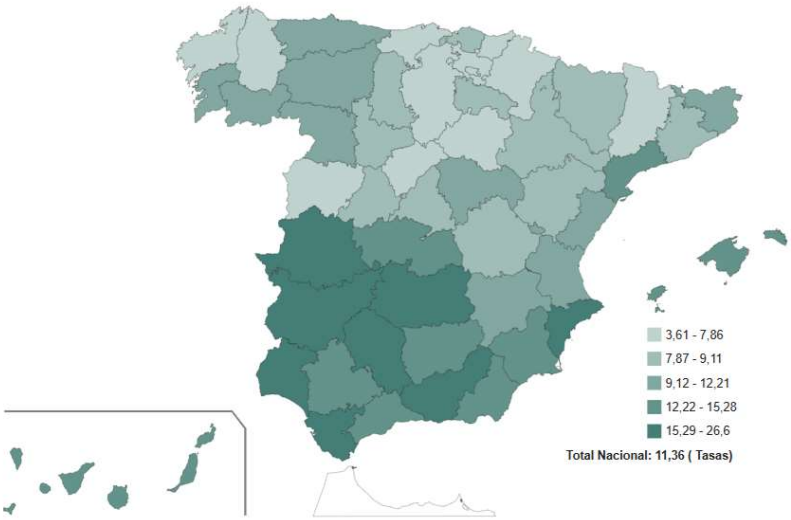
El INE también elabora estadísticas con ámbito autonómico y provincial. Si se accede a la sección dedicada a Estadísticas Territoriales, aparece un glosario de estadísticas que se pueden desglosar a nivel de comunidad autónoma y provincia, y en algunos casos hasta nivel municipal (figura 8.2).

Figura 8.2 Glosario de Estadísticas Territoriales del INE

Fuente: INE (2025q).

Por ejemplo, si se accede a la opción "Mercado laboral" se puede obtener un gráfico con la representación de las tasas de paro según provincia para el primer trimestre del 2025 (figura 8.3).

Figura 8.3 Tasa de paro según provincia



Fuente: INE (2025r).

8.3 ESTADÍSTICAS DEMOGRÁFICAS

Las estadísticas demográficas generadas por el INE se incluyen en la sección denominada "Demografía y Población". En esa sección es posible encontrar información sobre el Padrón, Cifras de Población y Censos Demográficos.

Las estadísticas incluidas en la sección Padrón que se elaboran de forma periódica se recogen en la tabla 8.4.

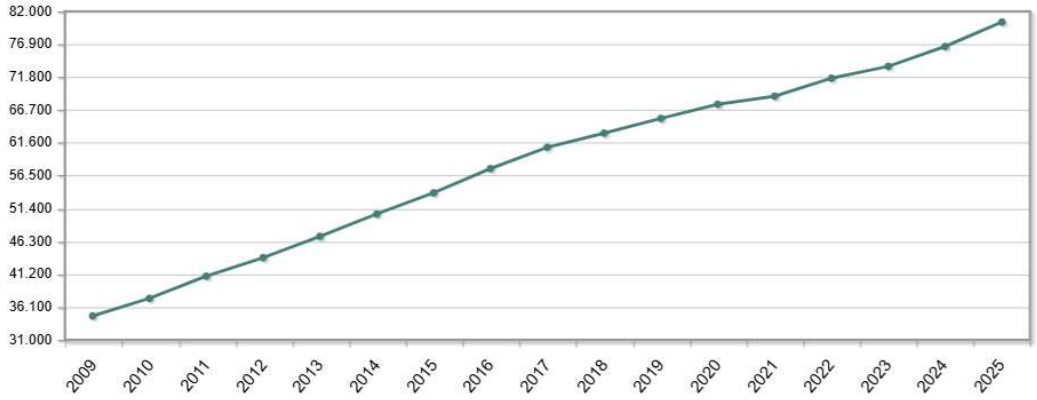
Tabla 8.4 Algunos contenidos de la sección Padrón del INE

Operaciones estadísticas que el INE elabora de forma periódica	Últimos datos
Cifras oficiales de población de los municipios españoles: Revisión del Padrón Municipal	01/01/2024
Relación de municipios y sus códigos por provincias	2025-01-01
Estadística del Padrón de españoles residentes en el extranjero	2025-01-01

Fuente: INE (2025s).

Por ejemplo, en la sección Estadísticas del Padrón de españoles residentes en el extranjero, es posible conocer los españoles residentes en el extranjero. según provincia de origen. En la figura 8.4 se presenta el caso de Málaga.

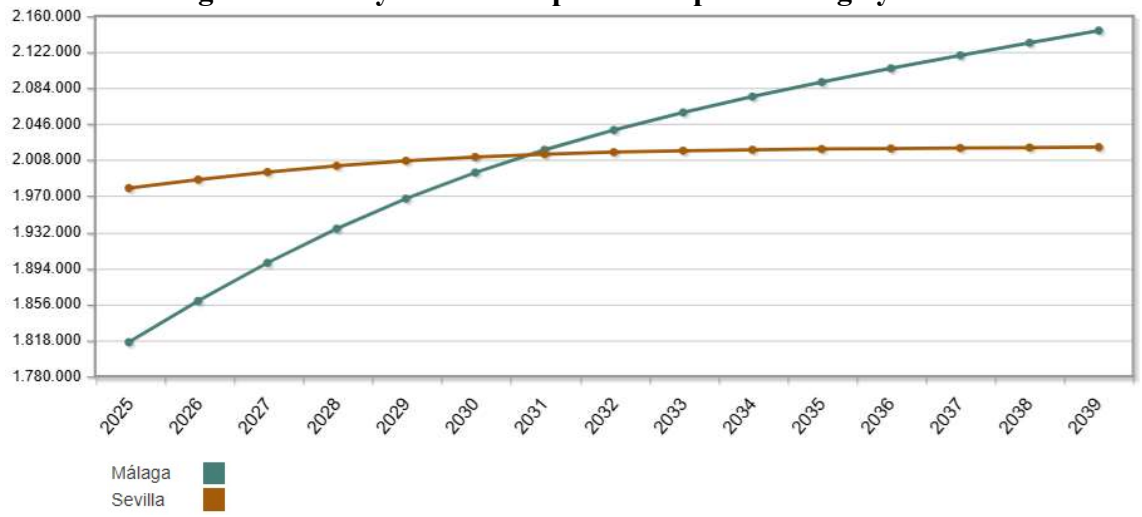
Figura 8.4. Población española residente en el extranjero: Málaga



Fuente: INE (2025t).

En cuanto a las estadísticas incluidas en "Cifras de Población y Censos Demográficos", una de sus peculiaridades es que nos permite obtener información de la población a un nivel más desagregado. Por ejemplo, en la sección de "Proyecciones de Población" se proporciona una simulación de la población que residirá en España durante los próximos años a nivel provincial. En figura 8.5 se muestra dicha proyección para las provincias de Málaga y Sevilla hasta el año 2039.

Figura 8.5. Proyecciones de población para Málaga y Sevilla



Fuente: INE (2025v).

Por otro lado, en el apartado dedicado a las "Proyecciones de hogares" se encuentran simulaciones futuras del número de hogares en España y su composición también a nivel provincial. En concreto, en la tabla 8.5 se presentan las correspondiente a Málaga desde el año 2024 hasta el 2039.

Tabla 8.5. Proyecciones de hogares y composición para Málaga: 2025-2039

	Total	1 persona	2 personas	3 personas	4 personas o más
2025	720.380	202.935	206.725	141.558	169.162
2026	741.414	211.338	214.393	144.770	170.913
2027	761.251	219.425	221.804	147.724	172.298
2028	779.738	227.312	228.923	150.307	173.196
2029	796.629	234.873	235.561	152.535	173.660
2030	811.873	241.904	241.784	154.424	173.761
2031	825.684	248.477	247.645	156.026	173.536
2032	838.417	254.835	253.200	157.355	173.027
2033	850.328	261.051	258.524	158.454	172.299
2034	861.388	266.932	263.581	159.396	171.479
2035	871.931	272.672	268.487	160.224	170.548
2036	882.094	278.311	273.268	160.949	169.566
2037	891.947	283.841	277.932	161.608	168.566
2038	901.533	289.278	282.451	162.208	167.596
2039	910.811	294.550	286.802	162.765	166.694

Fuente: INE (2025w).

Por última, otras estadísticas disponibles son las relativas a los Censos de Población, que permiten conocer de forma detallada algunas características relativas a personas, hogares, edificios o viviendas. Por ejemplo, a partir del Censo del 2021, es posible generar tablas como la 8.6 donde se muestra la distribución de las ocupadas en España según ocupación.

Tabla 8.6. Distribución de los trabajadores según ocupación y sexo: Censo del 2021

Sexo	Mujer
Ocupación a 1 dígito CNO-11	Personas
Total	8.704.134
1 - Directores y gerentes	237.156
2 - Técnicos y profesionales científicos e intelectuales	2.033.196
3 - Técnicos; profesionales de apoyo	801.687
4 - Empleados contables, administrativos y otros empleados de oficina	1.179.444
5 - Trabajadores de los servicios de restauración, personales, protección y vendedores	2.369.211
6 - Trabajadores cualificados en el sector agrícola, ganadero, forestal y pesquero	107.280
7 - Artesanos y trabajadores cualificados de las industrias manufactureras y la construcción (excepto operadores de instalaciones y maquinaria)	238.143
8 - Operadores de instalaciones y maquinaria, y montadores	136.446
9 - Ocupaciones elementales	1.347.978
No consta	253.587

Fuente: INE (2025y).

8.4 OTRO TIPO DE ESTADISTICAS

En esta sección se hace referencia brevemente a otras bases de datos de relevancia para la Investigación Social dentro del contexto andaluz. En particular, en el Instituto de Estadística y Cartografía de Andalucía (IECA) es posible encontrar un conjunto de indicadores sociales para la región andaluza agrupados por temas. Por ejemplo, aparece información sobre:

1. Población, familias y hogares.

2. Bienestar Social.

3. Ciencias, Tecnología y Sociedad Digital.

4. Mercado de trabajo.

5. Tejido empresarial.

6. Economía y macromagnitudes económicas.

7. Turismo.

8. Movilidad.

9. Uso del Tiempo.

10. Medio Ambiente y Cambio Climática.

11. Administración Pública.

Es interesante destacar que IECA pone a disposición de los investigadores los microdatos de un conjunto de encuestas de interés social (Tabla 8.7) agrupados por los siguientes temas: Población, Sociedad, Mercado de Trabajo y Transporte.

Tabla 8.7. Microdatos de interés científico puestos a disposición del público por IECA

Actividad estadística	Periodicidad	Serie disponible	Última actualización
Población			
Estadísticas de Población de Andalucía Basadas en Registros Administrativos (EPABRA)	Anual	2016-2023	14 de enero de 2025
Estadísticas longitudinales de biografías reproductivas en Andalucía	Puntual	2002-2021	22 de octubre de 2024
Estadísticas longitudinales de supervivencia y longevidad en Andalucía	Puntual	2002-2021	14 de febrero de 2024
Sociedad			
Encuesta social 2024. Consumo digital. Hábitos de la población andaluza	Puntual	2024	10 de junio de 2024
Encuesta social 2023. Consumo y sostenibilidad. Hábitos y actitudes de la población andaluza	Puntual	2023	25 de octubre de 2023
Encuesta social 2022. Relaciones sociales. Hábitos y actitudes de la población andaluza	Puntual	2022	24 de octubre de 2022
Encuesta social 2021. Digitalización y uso de datos personales. Capacidades y actitudes de la población andaluza	Puntual	2021	10 de marzo de 2022
Encuesta social 2020. Hábitos turísticos de la población andaluza	Puntual	2020	17 de diciembre de 2020
Encuesta social 2020. Hábitos y condiciones de vida de la población andaluza durante el estado de alarma	Puntual	2020	6 de agosto de 2020
Encuesta social 2019. Conciliación en el hogar: hábitos y actitudes de la población andaluza	Puntual	2019	1 de abril de 2020
Encuesta social 2018. Hogares y medio ambiente en Andalucía	Puntual	2018	27 de enero de 2020
Encuesta social 2010 y 2018. Panel de educación y transiciones al mercado laboral en Andalucía	Puntual	2010 y 2018	5 de diciembre de 2019
Encuesta social 2018. Educación y transiciones al mercado laboral en Andalucía	Puntual	2018	2 de mayo de 2019
Encuesta social 2017. Movilidad social en Andalucía	Puntual	2017	10 de octubre de 2018
Encuesta social 2011. Movilidad en las regiones urbanas de Andalucía	Puntual	2011	6 de marzo de 2013
Encuesta social 2010. Educación y hogares en Andalucía	Puntual	2010	2012
Encuesta social 2008. Hogares y medio ambiente en Andalucía	Puntual	2008	2012
Encuesta social 2007. Una visión de Andalucía	Puntual	2007	21 de marzo de 2011
Dependencia y solidaridad en las redes familiares	Puntual	2005	2005
Mercado de trabajo			
Inserción laboral de los egresados en universidades públicas de Andalucía	Anual	Cursos 2011-2012 a 2022-2023	10 de junio de 2025
Inserción laboral de los egresados de formación profesional en Andalucía	Anual	Cursos 2011-2012 a 2022-2023	12 de junio de 2025
Transporte			
Movilidad de la población en Andalucía a partir de información de teléfonos móviles	-	Desde el 14 de febrero de 2020 al 28 de mayo de 2021	Semanal

Fuente: IECA (2025).

PROPUESTAS DE EJERCICIOS PARA BLOQUE II

1. Si desea observar la intensidad de la relación lineal entre dos variables X e Y, ¿qué utilizaría el coeficiente de correlación lineal r_{xy} o la covarianza S_{xy}? Razone la respuesta. Además, si $r_{x,y} = 0$, ¿significa que X e Y son variables estadísticamente independientes?

2. Para un grupo de estudiantes se conoce el Nº de horas semanales de uso del móvil (Y) y el Nº de horas semanales de estudio (X).

X (horas de estudio)	**Y (horas de uso del móvil)**		
	10-20	**20-30**	**30-40**
0-10	2	4	7
10-20	3	4	3
20-30	9	3	1

a) Obtenga las distribuciones marginales de X e Y, ¿son las variables independientes? Razone la respuesta
b) Obtenga la covarianza y el coeficiente de correlación lineal.
c) Si todos los estudiantes incrementasen en 5 horas su dedicación al estudio, ¿cambiarían los resultados del apartado b? ¿y si aumentasen sus horas de estudios en un 10%?

3. Para un conjunto de individuos se dispone de información sobre su edad (X, en años) y el tiempo dedicado a las tareas del hogar (Y, en minutos).

Edad (X, años)	**Tiempo dedicado al hogar (Y, minutos)**		
	0-60	**60-180**	**180-300**
30-40	199	201	31
40-50	209	218	45
50-60	217	212	52

a) Obtenga las distribuciones marginales de X e Y, ¿son las variables independientes? Razone la respuesta.
b) Obtenga la distribución de la edad condicionada a que el número de minutos diarios dedicados a las tareas del hogar sea igual o inferior a 180 minutos, y para esta distribución obtenga la media aritmética.
c) Obtenga la distribución del tiempo dedicado al hogar condicionada a que la edad sea superior a 40 años, y para esta distribución obtenga la mediana.
d) Indique si existe relación lineal entre X e Y, su tipo y su intensidad. [Nota: $\frac{\sum\sum x_i y_j n_{ij}}{N} = 4120{,}34$; $\bar{X} = 45{,}36;\ \bar{Y} = 90{,}45;\ S_x = 8{,}109;\ S_y^2 = 4116{,}77$].
e) Si el tiempo dedicado al hogar se expresase en horas ¿cambiaría la intensidad de la relación lineal obtenida en el apartado d)? Razone la respuesta.

4. A partir de la Encuesta de Condiciones de Vida, se conoce para el periodo 2020-2024, el porcentaje de hogares en riesgo de pobreza (X) y el porcentaje de hogares con baja intensidad del trabajo (Y). Por baja

intensidad del trabajo se entiende hogares en los que sus miembros trabajaron menos del 20% del total de su potencial de trabajo durante el año de referencia.

Año	X	Y
2020	21,0	9,9
2021	21,7	11,7
2022	20,4	8,7
2023	20,2	8,5
2024	19,7	8,2

a) Obtenga e interprete la covarianza y el coeficiente de correlación lineal.

b) Obtenga los parámetros a y b de la línea de regresión: $Y_i^* = a + bX_i$. Interprete "b" y valore la bondad del ajuste.

c) Obtenga los parámetros a y b de la línea de regresión: $Y_i^* = a + bt$, siendo t una variable que recoge el tiempo y es igual a 1 en el año 2020. Obtenga la predicción para el 2025, ¿es fiable?

5. Para las variables X "porcentaje de personas con teléfono móvil (en puntos porcentuales)" e Y "porcentaje de personas que usan internet (en puntos porcentuales)" se dispone para el periodo 2016-2020 (N=5) de la siguiente información: $\sum x_i y_i = 43768{,}36;\ \bar{X} = 97{,}54;\ \bar{Y} = 89{,}7;\ S_x^2 = 1{,}54\,;\ S_y^2 = 13{,}97.$

a) Obtenga la línea de regresión: $Y_i^* = a + bX_i$.

b) Interprete el valor obtenido para el parámetro "b", y obtenga e interprete el coeficiente de determinación.

c) Si en el 2021, el porcentaje de personas con teléfono móvil es igual a 99,5 ¿cuál es la predicción del porcentaje de personas que usan internet en 2021? ¿Es fiable dicha predicción? Razone la respuesta.

6. Para las variables X "porcentaje de personas que usan internet (en puntos porcentuales)" e Y "porcentaje de personas que compran por internet (en puntos porcentuales)" se dispone para el periodo 2016-2020 (N=5) de la siguiente información: $\sum x_i n_i = 21580{,}62;\ \bar{X} = 89{,}7;\ \bar{Y} = 47{,}88;\ S_x = 3{,}73\,;\ S_y = 5{,}84.$

a) Obtenga la línea de regresión: $Y_i^* = a + bX_i$.

b) Interprete el valor obtenido para el parámetro "b", y obtenga la varianza residual (S_e^2).

c) Si en el 2021, el porcentaje de personas que usan internet es igual a 95, ¿cuál es la predicción del porcentaje de personas que compran por internet en el 2021? ¿Es fiable dicha predicción?

7. Se dispone de la siguiente información del gasto medio anual por hogar en alimentación (X, 10^3€) y del gasto medio anual en restaurantes en hoteles (Y, 10^3€) para el periodo 2017-2020 (N=4):

$$\sum x_i y_i = 45{,}42;\ \sum x_i = 17{,}19;\ \sum y_i = 10{,}64;\ \sum x_i^2 = 74{,}04;\ \sum y_i^2 = 29{,}44$$

a) Obtenga la línea de regresión: $Y_i^* = a + bx_i$. Además, interprete el valor de "b".

b) Obtenga el coeficiente de determinación e interprételo.

8. Para la serie correspondiente al Nº de ocupados de personas con discapacidad funcional se obtienen los índices en cadena (IC_t) y los índices simples con base en el año 2021 (I^t_{2021}), ¿cómo interpretaría $IC_{2023} - 100 = 4\%$ y $I^{2023}_{2021} - 100 = 4{,}6\%$, respectivamente?

9. Si en el año 2024, la pensión media de la Seguridad Social en España ascendió a 1260 €, ¿cuál sería la pensión media actualizada para el año 2025 (es decir, la que mantiene el poder adquisitivo), si se prevé un crecimiento interanual de la inflación del 2,3% para el año 2025?

10. En la siguiente tabla se muestra datos del gasto medio anual en vivienda, agua, electricidad y gas para el periodo 2017-2020, y del IPC correspondiente a este tipo de gasto.

Año	**Gasto medio (€, corrientes)**	**IPC (Base 2016)**
2017	8774,12	95,046
2018	9180,75	96,638
2019	9441,17	97,314
2020	9621,46	97

a) Obtenga la serie del gasto medio deflactada (€, constantes), y su tasa de variación para el periodo 2017-2020. Interprete dicha tasa.

b) Obtenga la serie de índices en cadena del gasto medio (€, constantes), e interprete el dato correspondiente para el año 2020.

c) Obtenga e interprete la tasa de inflación interanual para el año 2019.

d) Si en el año 2021, el gasto medio (€, corrientes) registró un crecimiento interanual del 2%, y el IPC registró un crecimiento interanual del 5%, ¿cuál es el gasto medio deflactado (€, constantes) en el 2021?

11. Se proporciona información sobre los índices simples (base 2020) del gasto medio anual en alimentación en € corrientes y sobre el IPC para el periodo 2020-2023.

t	$\boldsymbol{I^t_{2020}}$	**IPC (Base 2021)**
2020	100	97
2021	106,84	100
2022	117,83	108,39
2023	124,43	112,22

a) Obtenga la serie del gasto medio anual en alimentación en términos reales, y las tasas de variación interanuales en términos reales. (Nota: El gasto medio anual en alimentación en € corrientes fue de 4286 € en el año 2020).

b) Obtenga e interprete la tasa de inflación entre los años 2020 y 2023.

Referencias

IECA (2025). Instituto de Estadística y Cartografía de Andalucía/Difusión y reutilización/Microdatos: https://www.juntadeandalucia.es/institutodeestadisticaycartografia/dega/microdatos (último acceso, 13/06/2025).

INE (2023). Notas de Prensa: Censo de Población y Viviendas 2021. Instituto Nacional de Estadística, Madrid.

INE (2024a). INE/Economía/Empresas/Explotación estadística del Directorio Central de Empresas: https://www.ine.es/jaxiT3/Tabla.htm?t=39375&L=0 (último acceso: 26/06/2024).

INE (2024b). INE/Demografía y Población/Padrón/Estadística del Padrón Continuo: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177012&menu=ultiDatos&idp=1254734710990 (último acceso: 29/05/2024).

INE (2025a). Instituto Nacional de Estadística: https://www.ine.es (último acceso: 29/05/2025).

INE (2025b). INE/Sociedad: https://www.ine.es/dyngs/INEbase/es/categoria.htm?c=Estadistica_P&cid=1254735971047 (último acceso: 29/05/2025).

INE (2025c). INE/Sociedad/Salud/El salario de las personas con discapacidad: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176911&menu=ultiDatos&idp=1254735573175 (último acceso: 29/05/2025).

INE (2025d). INE/Sociedad/Salud/El salario de las personas con discapacidad: https://www.ine.es/jaxi/Tabla.htm?tpx=8527&L=0 (último acceso: 29/05/2025).

INE (2025e). INE/Sociedad/Salud/El empleo de las personas con discapacidad: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736055502&menu=ultiDatos&idp=1254735573175 (último acceso: 29/05/2025).

INE (2025f). INE/Actividad/Mercado de trabajo/ El empleo de las personas con discapacidad/ El empleo de las personas con discapacidad. Serie 2014-2022/Resultados Nacionales: https://www.ine.es/jaxi/Tabla.htm?tpx=71965&L=0 (último acceso: 29/05/2025).

INE (2025g). Encuesta de Condiciones de Vida. Instituto Nacional de Estadística, Madrid.

INE (2025h). INE/Condiciones de Vida/Encuesta de Condiciones de Vida/Resultados por CCAA: https://www.ine.es/jaxiT3/Tabla.htm?t=9947&L=0 (último acceso: 3/06/2025).

INE (2025i). INE/Nivel y condiciones de vida/Condiciones de Vida/Encuesta de Condiciones de Vida: https://www.ine.es/jaxiT3/Tabla.htm?t=59970&L=0 (último acceso: 7/06/2025).

INE (2025j. INE/Nivel y condiciones de vida/Condiciones de Vida/Encuesta de Condiciones de Vida: https://www.ine.es/jaxiT3/Tabla.htm?t=9963&L=0 (último acceso: 7/06/2025).

INE (2025k). INE/Mercado laboral/Actividad, ocupación y paro/Encuesta de Población Activa: https://www.ine.es/jaxiT3/Tabla.htm?t=65947&L=0 (último acceso: 10/06/2025).

INE (2025l). INE/Mercado laboral/Actividad, ocupación y paro/Encuesta de Población Activa: https://www.ine.es/jaxiT3/Tabla.htm?t=65191&L=0 (último acceso: 10/06/2025).

INE (2025m). INE/Servicios/Hostelería y Turismo/ Hoteles: https://www.ine.es/jaxiT3/Tabla.htm?t=2009&L=0 (último acceso: 10/06/2025).

INE (2025n). INE/Mercado laboral/Actividad, ocupación y paro/Encuesta de Población Activa: https://www.ine.es/jaxiT3/Tabla.htm?t=65963&L=0 (último acceso: 13/06/2025).

INE (2025ñ). INE/Nivel y Condiciones de Vida (IPC)/Índice de precios de consumo y vivienda: https://www.ine.es/dyngs/INEbase/operacion.htm?c=Estadistica_C&cid=1254736176802&menu=resultados&idp=1254735976607 (último acceso: 13/06/2025).

INE (2025o). INE/Sociedad: https://www.ine.es/dyngs/INEbase/categoria.htm?c=Estadistica_P&cid=1254735971047 (último acceso: 13/06/2025).

INE (2025p). INE/Objetivos de Desarrollo Sostenible: https://www.ine.es/dyngs/ODS/es/index.htm (último acceso: 23/06/2025).

INE (2025q). INE/Estadísticas Territoriales: https://www.ine.es/dynInfo/Infografia/Territoriales/index.html (último acceso: 13/06/2025).

INE (2025r). INE/ Estadísticas territoriales/Mercado de Trabajo/Tasa de paro: https://www.ine.es/dynInfo/Infografia/Territoriales/galeriaCapitulo.html?capitulo=4337 (último acceso: 13/06/2025).

INE (2025s). INE/Demografía y Población/Padrón: https://www.ine.es/dyngs/INEbase/categoria.htm?c=Estadistica_P&cid=1254734710984 (último acceso: 13/06/2025).

INE (2025t). INE/ Demografía y población/Padrón/Estadísticas del Padrón de españoles en el extranjero: https://www.ine.es/jaxi/Tabla.htm?tpx=24260&L=0 (último acceso: 13/06/2025).

INE (2025v). INE/Demografía y población/Cifras de población y censos demográficos/Proyecciones de población: https://www.ine.es/dynt3/inebase/es/index.htm?padre=6675&capsel=6681 (último acceso: 13/06/2025).

INE (2025w). INEbase/Demografía y población/Cifras de población y censos demográficos/Proyecciones de hogares: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176954&menu=resultados&idp=1254735572981 (último acceso: 13/06/2025).

INE (2025y). INEbase/Demografía y población/Cifras de población y censos demográficos/Censos de población y vivienda: https://www.ine.es/dyngs/INEbase/operacion.htm?c=Estadistica_C&cid=1254736177108&menu=resultados&idp=1254735572981#_tabs-1254736195867 (último acceso: 13/06/2025).

Ministerio de Ciencia, Innovación y Universidades (2025). Estadísticas y Datos Abiertos/Sistema Integrado de Estudiantes Universitarios/Estudiantes Universitarios: https://www.ciencia.gob.es/Ministerio/Estadisticas/SIIU/Estudiantes.html (último acceso: 11/06/2025).

MITES (2025a). Ministerio de Trabajo y Economía Social/ Estadísticas y Análisis:

https://www.mites.gob.es/es/estadisticas/index.htm (último acceso: 29/05/2025).

MITES (2025b). Ministerio de Trabajo y Economía Social/ Estadísticas y Análisis/Inmigración y Emigración: https://www.mites.gob.es/es/estadisticas/Inmigracion_emigracion/index.htm (último acceso: 29/05/2025).

OECD (2025). OECD/Data: https://data.oecd.org/inequality/income-inequality.htm (último acceso: 9/06/2025).