

Guía de Estudio y Apuntes de Estadística

Estadística

Grado en Ingeniería en
Organización Industrial



Material docente en abierto de la Universidad Rey Juan Carlos

Depositado en: <https://burjcdigital.urjc.es/home>

Handle: <https://hdl.handle.net/10115/113498>

©2025 Autoras: Ana Medina-López y Anubis Brita Pérez Maldonado

Algunos derechos reservados. Este documento se distribuye bajo la licencia

"Atribución-Compartir-Igual 4.0 Internacional" de Creative Commons.

Disponible en: <https://creativecommons.org/licenses/by-sa/4.0/deed.es>

ÍNDICE

1. Presentación	4
2. Créditos	4
3. Objetivos de aprendizaje.....	4
4. Guía Docente	5
5. Profesoras	5
6. Horarios	5
7. Tutorías	6
8. Temario	6
9. Evaluación	8
10. Cronograma	11
11. Recomendaciones de estudio	12
12. Bibliografía básica	13
13. Bibliografía complementaria	13

TEMARIO

(Bloque Temático I – Estadística e Información)

Tema 1. Introducción a la Estadística	4
1.1 El conocimiento estadístico	
1.2 Las ciencias estadísticas. El método estadístico	
Tema 2. Técnicas de recogida de datos	19
2.1 Recogida de datos: objetivos y limitaciones	
2.2 Fuentes de información estadística	

(Bloque Temático II – Análisis Estadístico de una Variable)

Tema 3. Distribuciones de frecuencia. Medidas de posición, dispersión y forma...26
3.1 Variable estadística
3.2 Distribuciones de frecuencias
3.3 Momentos
3.4 Representaciones gráficas
3.5 Medidas de posición
3.6 Medidas de dispersión
3.7 Datos atípicos

(Bloque Temático III – Análisis Estadístico de Dos Variables)

Tema 4. Distribuciones bidimensionales. Tablas de contingencia.....	34
4.1 Tablas de correlación y contingencia	
4.2 Representación gráfica: nube de puntos	
4.3 Distribuciones marginales	
4.4 Distribuciones condicionadas	
4.5 Dependencia e independencia	

(Bloque Temático IV – Probabilidad)

Tema 5. Noción de probabilidad	43
5.1 Experimento aleatorio	
5.2 Sucesos y operaciones con sucesos	
5.3 Axiomas y propiedades	
Tema 6. Probabilidad condicionada e independencia	59
6.1 Probabilidad condicionada	
6.2 Teorema de Bayes	
6.3 Independencia	

(Bloque Temático V – Variables Aleatorias)

Tema 7. Variables aleatorias discretas	65
7.1 Funciones de probabilidad	
7.2 Parámetros (media, varianza, etc.)	
Tema 8. Variables aleatorias continuas	75
8.1 Función de densidad	
8.2 Función de distribución	
8.3 Parámetros	

(Bloque Temático VI – Inferencia Estadística)

Tema 9. Modelos de distribución de probabilidad.....	85
9.1 Modelos discretos	
9.2 Modelos continuos	
9.3 Distribución Normal	
9.4 Distribución Chi-cuadrado, t de Student y F de Snedecor	
9.5 Teorema Central del Límite	
Tema 10. Nociones de inferencia estadística.....	97
10.1 La inferencia y sus métodos	
10.2 Estimación puntual	
10.3 Intervalos de Confianza	



Asignatura Estadística

Grado en Ingeniería en Organización Industrial

Profesoras: Ana Medina-López y Anubis Brita Pérez-Maldonado

Material docente en abierto de la Universidad Rey Juan Carlos

Curso académico 2025-26

Licencia de publicación en abierto CC-BY-SA

“Algunos derechos reservados. Esta obra se distribuye bajo la licencia”

Guía de Estudio y Contenido de la asignatura Estadística

1. Presentación

La recopilación, organización, análisis e interpretación de datos es una parte esencial del trabajo de un ingeniero en organización industrial, cuyo desempeño profesional depende de la capacidad para fundamentar decisiones en información objetiva y fiable. Por ello, esta asignatura tiene como objetivo que el estudiante conozca y utilice los métodos y técnicas estadísticas más apropiados para obtener, procesar y analizar datos en contextos industriales y empresariales.

A lo largo del curso se introducen los principios fundamentales del razonamiento estadístico, tanto en su vertiente **descriptiva** como en la **inferencial**, de forma que los estudiantes puedan aplicar estas herramientas a situaciones reales de toma de decisiones dentro de la ingeniería y la gestión de procesos.

La asignatura se enmarca en la formación básica del grado y se imparte en el primer semestre del segundo curso. Se imparte en castellano.

Conocimientos previos recomendados: Matemáticas básicas.

2. Créditos: 6 ECTS

3. Objetivos de aprendizaje

Al finalizar la asignatura, el estudiante será capaz de:

- ☐ **Comprender los fundamentos teóricos de la estadística** y reconocer su utilidad en el análisis y la toma de decisiones en ingeniería y gestión de procesos industriales.
- ☐ **Identificar, recopilar y organizar datos relevantes** procedentes de diferentes fuentes, valorando su calidad, fiabilidad y adecuación al problema de estudio.

- ☐ **Aplicar técnicas de estadística descriptiva** para resumir e interpretar la información mediante medidas numéricas y representaciones gráficas, adaptándolas a contextos reales del ámbito industrial.
- ☐ **Formular y analizar fenómenos aleatorios**, comprender el significado de conceptos como suceso, probabilidad, independencia y probabilidad condicionada, y utilizarlos en situaciones prácticas.
- ☐ **Resolver problemas estadísticos mediante razonamiento lógico y cuantitativo**, identificando la técnica adecuada para cada situación.
- ☐ **Aplicar métodos de inferencia estadística** —estimación, intervalos de confianza y contrastes de hipótesis— para extraer conclusiones objetivas a partir de muestras de datos en entornos industriales.
- ☐ **Utilizar herramientas informáticas básicas y recursos docentes abiertos** para el tratamiento y análisis de datos, siguiendo criterios de precisión y rigor.
- ☐ **Desarrollar autonomía en el estudio**, integrando el trabajo teórico y práctico mediante la planificación del esfuerzo personal y el aprovechamiento de tutorías, actividades complementarias y sesiones de resolución de ejercicios.
- ☐ **Interpretar resultados estadísticos en términos técnicos y operativos**, y comunicar conclusiones de manera clara y estructurada, adecuadas a un entorno profesional de ingeniería.
- ☐ **Participar de forma activa en la resolución de casos prácticos**, aplicando los conocimientos adquiridos a situaciones industriales reales o simuladas, y reconociendo la importancia del análisis estadístico en la optimización de procesos productivos y organizativos.

4. Guía docente

Puede descargar la guía docente de la asignatura en el siguiente enlace:

<https://servicios.urjc.es/guiasdocentes/>

5. Profesoras

La profesora responsable de la asignatura es Ana Medina-López, profesora Titular de Universidad en el Departamento de Economía Financiera y Contabilidad. Doctora por la Universidad Rey Juan Carlos. Imparte esta asignatura en el Campus de Madrid-Vicálvaro, turno de mañana. Cuenta con cuatro quinquenios de docencia, un sexenio de investigación y cinco evaluaciones positivas en el programa Docencia.

La Profesora Anubis Brita Pérez-Maldonado, Ayudante Doctora adscrita al Departamento de Economía Financiera y Contabilidad. Imparte esta asignatura en el Campus de Móstoles, turno de tarde. En su trayectoria académica cuenta con un quinquenio de docencia y una evaluación positiva en el programa Docencia.

6. Horarios

Grado en Ingeniería en Organización Industrial (Vicálvaro), impartida por la profesora ana.medina@urjc.es:

Martes: de 13:00 a 14:00 horas, (de 18 de noviembre de 2025 a 9 de diciembre de 2025)

Miércoles: de 9:00 a 11:00 horas

Viernes: de 9:00 a 11:00 horas

Grado en Ingeniería en Organización Industrial (Móstoles), impartida por la profesora anubis.perez@urjc.es:

Lunes: 17:00 a 19:00

Martes: 19:00 a 20:00 (durante los meses de septiembre y noviembre)

Jueves: 18:00 a 20:00

7. Tutorías

El horario de tutorías se coordina directamente con las profesoras, quienes facilitarán la atención personalizada a través del correo electrónico institucional garantizando flexibilidad y disponibilidad para el estudiante: ana.medina@urjc.es y anubis.perez@urjc.es

Mediante cita previa vía Teams o presencial.

8. Temario

Bloque temático	Tema	Apartados
I.- Estadística e información	Tema 1. Introducción a la Estadística	1.1 El conocimiento estadístico y su aplicación en el ámbito económico y social. 1.2 Las ciencias estadísticas. El Método Estadístico.
	Tema 2. Técnicas de Recogida de Datos	2.1 Recogida de datos: objetivos y limitaciones. 2.2 Fuentes de Información Estadística (Pública y Privada)

II.- Análisis Estadístico de una variable	Tema 3. Distribuciones de Frecuencia. Medidas de posición, de dispersión y de forma.	3.1 Variable Estadística. 3.2 Distribuciones de frecuencias. 3.3 Momentos. 3.4 Representaciones gráficas. 3.5 Medidas de posición centrales: media aritmética, mediana y moda. 3.6 Medidas de posición no centrales: cuartiles, deciles y percentiles. 3.7 Medidas de dispersión absolutas: varianza y desviación típica. 3.8 Medidas de dispersión relativas: coeficiente de variación 3.9 Tipificación. 3.10 Medidas de asimetría y curtosis. 3.11 Datos atípicos y diagramas de caja
IV.- Análisis Comparativo de variables estadísticas	Tema 6. Números Índice	6.1 Concepto y tipología. 6.2 Números índices complejos. 6.3 Propiedades. 6.4 Cambio de base. Enlace de números índices. 6.5 Índices de precios e índices de cantidades. Índices en Economía. 6.6 Deflactación

V.- Introducción a la Probabilidad y a la Inferencia Estadística	Tema 7. Elementos básicos de probabilidad	7.1 Noción y determinación de la probabilidad 7.2 Probabilidad condicionada e independencia de sucesos
	Tema 8. Variable aleatoria	8.1 Distribución de probabilidad y función de distribución 8.2 V.A. discreta y V.A. continua 8.3 Esperanza, varianza y tipificación
	Tema 9. Modelos de distribución de probabilidad	9.1 Modelos discretos 9.2 Modelos continuos 9.3 Distribución Normal 9.4 Distribución Chi-cuadrado, t de Student y F de Snedecor 9.5 Teorema Central del Límite
	Tema 10. Nociones de Inferencia Estadística	10.1 La inferencia y sus métodos 10.2 Estimación puntual 10.3 Intervalos de Confianza

9. Evaluación

La evaluación se llevará a cabo según se indica a continuación:

Convocatoria ordinaria:

- ❓ Dos pruebas teóricas tipo test, cada una con una ponderación de 15%. La primera se realizará la tercera semana de noviembre y se evaluarán los temas del 1 al 6. La segunda se realizará en la fecha oficial de la evaluación ordinaria (enero de 2026) y se evaluarán los temas del 7 al 10.
- ❓ Dos pruebas prácticas, cada una con una ponderación de 25%. La primera se realizará la tercera semana de noviembre y se evaluarán los temas del 1 al 6. La segunda se realizará en la fecha oficial de la evaluación ordinaria (enero de 2026) y se evaluarán los temas del 7 al 10.
- ❓ Resolución en el aula de casos prácticos. Se realizarán 10 casos prácticos (uno por cada tema), con diez preguntas a responder. La ponderación total de la media de los diez casos prácticos será de 20%. No reevaluable en la convocatoria extraordinaria.

Convocatoria extraordinaria:

Los estudiantes que no hayan superado la asignatura en convocatoria ordinaria podrán presentarse al examen oficial en la fecha de la convocatoria extraordinaria para examinarse de las pruebas no superadas en convocatoria ordinaria.

Sistema de evaluación	Actividad de evaluación	Ponderación sobre la nota final	Reevaluable en convocatoria extraordinaria	Nota mínima	Contenidos	Fecha
Pruebas de evaluación orales o escritas realizadas de forma presencial sobre contenidos teóricos.	Prueba teórica. En las pruebas teóricas tipo test las preguntas bien contestadas suman 1 punto, las no contestadas suman 0 puntos y las contestadas erróneamente restan 0,2 puntos.	15%	Sí	5 (Libertatoria)	Temas 1 al 6	Se realizarán tantas pruebas como indique el docente, que informará sobre las fechas de las pruebas de evaluación continua las primeras semanas de clase, a lo largo del periodo lectivo.
Pruebas de evaluación orales o escritas realizadas de forma presencial sobre contenidos teóricos.	Prueba teórica. En las pruebas teóricas tipo test las preguntas bien contestadas suman 1 punto, las no contestadas suman 0 puntos y las contestadas erróneamente restan 0,2 puntos.	15%	Sí	5 (Libertatoria)	Temas 7 al 10	En la fecha oficial de la convocatoria ordinaria.





10. Cronograma Planificación Curso 2025-26

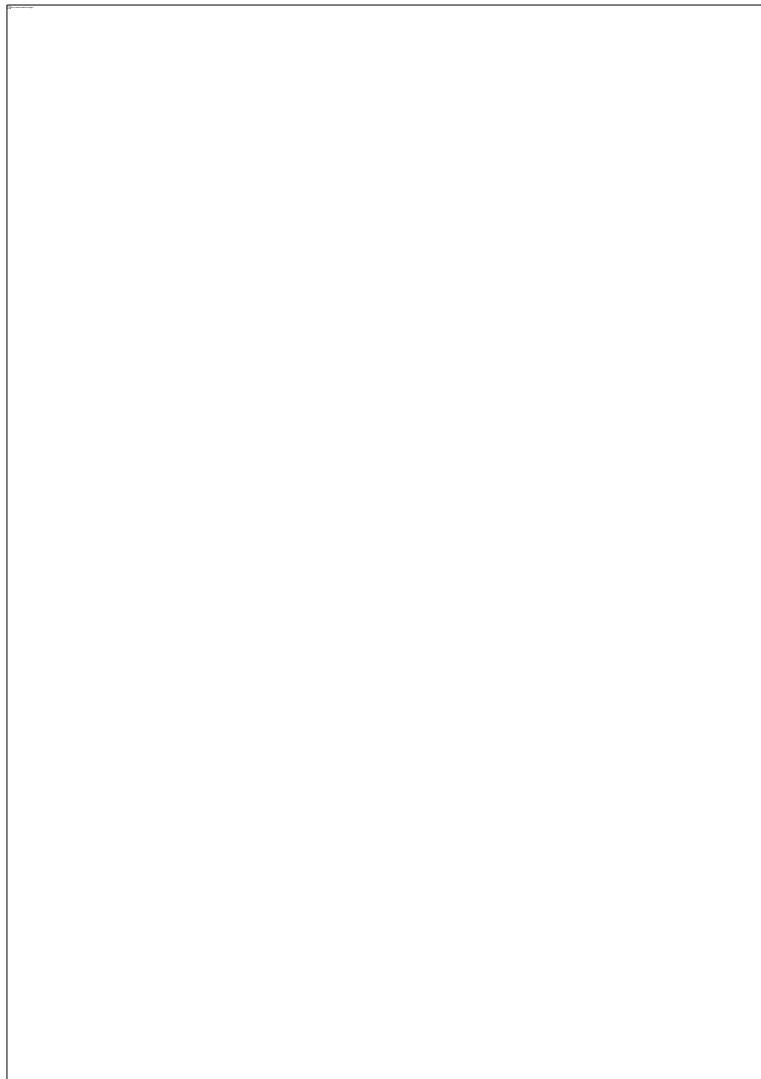
Primer Cuatrimestre

Grado en ingeniería en Organización Industrial (2140015)

Estadística

Vicálvaro

Móstoles



Enero 2026

Semana	Lunes	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo
				1	2	3	4
Período de evaluación ordinario	5	6	7	8	9	10	11
Período de evaluación ordinario	12	13	14	15	16	17	18
Período de evaluación ordinario	19 Prueba 2 (40%) Temas 7 al 10 Hora: 18:00- 20:00	20	21	22 Fin del período de evaluación ordinario	23 Inicio 2do cuatrimestre	24	25

11. Recomendaciones de estudio

La asignatura está diseñada para que el estudiante adquiriera las competencias teóricas y prácticas necesarias mediante una combinación equilibrada de clases presenciales, actividades dirigidas y trabajo autónomo. El tiempo total estimado de dedicación del alumnado es de **180 horas**, distribuidas en los distintos tipos de trabajo que se detallan a continuación.

1. Clases teóricas (26 horas)

Durante las sesiones teóricas se presentan los conceptos fundamentales de la asignatura, las definiciones, los modelos y los métodos esenciales. Estas clases proporcionan la base conceptual que permitirá al estudiante comprender en profundidad el resto de las actividades. Es recomendable asistir con los apuntes preparados y revisar el material entregado previamente por el profesorado.

2. Clases de resolución de ejercicios, problemas y casos prácticos (30 horas)

En estas sesiones se aplican los contenidos teóricos a situaciones reales o simuladas, mediante resolución de problemas, análisis de casos y ejercicios guiados. La participación activa del alumnado resulta crucial para consolidar los conocimientos y desarrollar habilidades de razonamiento cuantitativo y analítico.

3. Realización de pruebas (4 horas)

A lo largo del curso se programan evaluaciones que permiten medir el grado de comprensión de los contenidos, así como la capacidad del alumno para resolver ejercicios y aplicar técnicas estadísticas. Se recomienda repasar periódicamente los temas para llegar a estas pruebas con una preparación adecuada.

4. Tutorías académicas (30 horas)

El estudiante dispone de un amplio margen para consultas individuales o grupales con el profesorado. Las tutorías permiten aclarar dudas, reforzar conceptos y orientar el

estudio de manera personalizada. Se aconseja aprovechar este recurso de forma regular, especialmente al preparar exámenes o resolver ejercicios complejos.

5. Actividades relacionadas: jornadas, seminarios, etc. (18 horas)

Se incluyen actividades complementarias como asistencia a seminarios, jornadas técnicas o sesiones de refuerzo. Estas actividades acercan al estudiante a la aplicación real de la estadística en contextos empresariales e industriales, enriqueciendo la formación más allá del aula.

6. Trabajo autónomo del estudiante

El aprendizaje requiere una dedicación constante fuera del horario presencial. El tiempo estimado se distribuye en:

- ☐ **Preparación de clases teóricas (30 horas):** lectura previa de materiales, revisión de contenidos y elaboración de apuntes.
- ☐ **Preparación de prácticas, ejercicios y casos (30 horas):** resolución de ejercicios adicionales, práctica con datos reales y trabajo con herramientas informáticas.
- ☐ **Preparación de pruebas (12 horas):** repaso sistemático de los temas, realización de simulacros y resolución de dudas.

Total de horas de trabajo del estudiante: 180 horas

Esta organización temporal permite un aprendizaje progresivo y equilibrado, combinando estudio individual, trabajo guiado y actividades de aplicación práctica. Se recomienda al estudiante planificar su dedicación semanal desde el inicio del curso para mantener un ritmo constante y aprovechar al máximo los recursos disponibles.

12. Bibliografía básica

Fundamentos de Estadística. Autor: Daniel Peña. Editorial: Alianza Editorial, 2008.

Problemas de Estadística Autor: J. López de la Manzanara Barbero Editorial Pirámide. Madrid, 2008

Estadística para los Negocios y la Economía. Autor: Paul Newbold. Editorial: Prentice Hall, 2010.

Elementos Básicos de Estadística Económica y Empresarial Autor: A. M. Montiel, F. Rius y F. J. Barón Editorial: Prentice Hall, 2002.

Estadística descriptiva para economía y administración de empresas: cuestiones tipo test y ejercicios con Microsoft Excel. Autor: Arnaldos García, F., Díaz Delfa, M. T., Faura Martínez, U. y Molera Peris, L. Editorial: Thomson Paraninfo, Madrid.

Estadística aplicada: conceptos y ejercicios a través de Excel. Autor: Pérez López, César. Editorial: Garceta, 2012

Inferencia Estadística. Autor: Jesús Esteban García, y otros. Editorial: Garceta, 2010.



13. Bibliografía complementaria

Introducción a la Estadística Económica y Empresarial. Teoría y Práctica. Autor: Martín Pliego, F.J. Editorial: Thomson (2004)

Fuentes de Información Estadística Autor: Carlos González-Conde Llopis Editorial: UAM ediciones

Grado en Ingeniería en Organización Industrial**Tema 1. Introducción a la Estadística*****Bloque temático I: Estadística e información*****1.1 El conocimiento estadístico y su aplicación en el ámbito económico y social.****1.2 Las ciencias estadísticas. El Método Estadístico.****Introducción**

La estadística es una herramienta esencial en el análisis de datos y la toma de decisiones en entornos industriales, económicos y sociales. En el contexto de la Ingeniería en Organización Industrial, permite transformar grandes volúmenes de información en conocimiento útil para optimizar procesos, evaluar resultados y planificar estrategias.

El conocimiento estadístico abarca un conjunto de competencias que permiten medir, organizar, representar, analizar e interpretar datos. Su aplicación práctica se extiende desde el control de calidad en la producción hasta el análisis de mercados, la gestión de recursos humanos o la evaluación de políticas públicas. En una sociedad dinámica, donde los fenómenos no siguen patrones fijos, la estadística se convierte en una herramienta clave para comprender la realidad y actuar sobre ella con base empírica.

Una distinción fundamental en estadística es la que existe entre población y muestra. La población representa el conjunto total de elementos que se desea estudiar, mientras que la muestra es un subconjunto representativo que permite realizar inferencias sin necesidad de observar todos los casos. Esta diferencia es crucial para aplicar métodos estadísticos de forma eficiente y con rigor.

Asimismo, es importante diferenciar entre estadística descriptiva e estadística inferencial. La primera se ocupa de resumir y visualizar los datos observados mediante medidas como la media, la mediana o la desviación típica. La segunda permite generalizar los resultados muestrales a la población, estimar parámetros desconocidos y realizar contrastes de hipótesis, lo que resulta indispensable para la toma de decisiones bajo incertidumbre.

El método estadístico se estructura en varias fases: recolección de datos, organización y presentación, resumen e interpretación, y análisis formal o modelización. Estas etapas permiten pasar de la observación empírica a la comprensión analítica, y son aplicables a estudios de tiempos, análisis de fallos, satisfacción del cliente o eficiencia energética.

En este proceso, es esencial conocer los tipos de variables (cualitativas y cuantitativas) y las escalas de medición (nominal, ordinal, de intervalo y de razón), así como identificar posibles

sesgos que pueden distorsionar los resultados, como el sesgo de selección, de información o de confirmación.

Por último, se distinguen dos tipos de estudios estadísticos: los observacionales, que registran fenómenos sin intervenir, y los experimentales, donde se manipulan variables para evaluar efectos. Ambos enfoques son válidos y complementarios, siempre que se apliquen con rigor metodológico.

Así la estadística proporciona a los ingenieros en organización industrial las herramientas necesarias para convertir datos en decisiones estratégicas, mejorar procesos y contribuir al desarrollo económico y social con base en evidencia.

1.1 El conocimiento estadístico y su aplicación en el ámbito económico y social

El conocimiento estadístico constituye el conjunto de saberes, habilidades y competencias que permiten comprender, interpretar y aplicar conceptos y métodos estadísticos con el fin de analizar datos y tomar decisiones fundamentadas. En el contexto de la Ingeniería en Organización Industrial, este conocimiento resulta esencial para procesar la gran cantidad de información generada por los procesos productivos, logísticos y de gestión empresarial.

La Estadística permite medir, observar, analizar y extraer conclusiones a partir de los datos. Su aplicación abarca desde la planificación de la producción hasta el control de calidad, la optimización de recursos y la evaluación de políticas industriales o económicas. En una sociedad en constante cambio, los fenómenos económicos y empresariales no pueden regularse mediante leyes fijas; por ello, el análisis estadístico se convierte en una herramienta indispensable para la toma de decisiones basadas en evidencia empírica.

Principales tareas del conocimiento estadístico:

- Medir y reunir datos relevantes.
- Observar, organizar y representar la información.
- Analizar los resultados para formular conclusiones y decisiones.

Importancia en el ámbito económico y social

El conocimiento estadístico proporciona el soporte necesario para la gestión racional de los recursos económicos y sociales. En el ámbito empresarial, facilita el análisis de mercados, la estimación de la demanda, la valoración de riesgos y la evaluación del desempeño organizativo. En el ámbito público, contribuye al diseño y seguimiento de políticas económicas, sanitarias o medioambientales basadas en datos.

Componentes del conocimiento estadístico

- Comprensión de conceptos clave: media, mediana, desviación típica, probabilidad, correlación, etc.
- Capacidad para recopilar, organizar y representar datos de manera coherente.
- Interpretación crítica de resultados, identificando posibles errores o sesgos.

- Manejo de herramientas estadísticas manuales o digitales.
- Conciencia de los límites de los métodos estadísticos y sus supuestos.

El dominio de estas competencias permite transformar datos en información y, finalmente, en conocimiento útil para la toma de decisiones. Este proceso de transformación es fundamental para los ingenieros en organización industrial, que deben analizar indicadores de productividad, eficiencia energética, tiempos de proceso o calidad.

Ejemplos de aplicación práctica

- Índice de desempleo: relación entre población desempleada y población activa.
- Índice de precios al consumo (IPC): evolución del coste de la cesta de bienes de consumo.
- Contraste de hipótesis: evaluación de mejoras en un proceso productivo tras un cambio tecnológico.
- Análisis de correlación: relación entre productividad y horas de formación del personal.

En todos estos casos, el ingeniero industrial utiliza datos reales para extraer conclusiones que guíen decisiones de inversión, innovación o mejora continua.

1.2 Las ciencias estadísticas. El Método Estadístico

La Estadística es una rama de las matemáticas que se ocupa de reunir, organizar, analizar e interpretar datos numéricos con el fin de obtener conclusiones válidas y fundamentar la toma de decisiones. El método estadístico se aplica tanto a la descripción de datos (estadística descriptiva) como a la inferencia sobre poblaciones (estadística inferencial).

Fases del método estadístico

1. Recogida de datos: selección de fuentes fiables mediante censos, encuestas o registros administrativos.
2. Ordenación y presentación: uso de tablas, gráficos y resúmenes numéricos que faciliten la comprensión.
3. Resumen e interpretación: cálculo de medidas de tendencia central y dispersión.
4. Análisis formal o modelización: búsqueda de relaciones entre variables y contrastes de hipótesis.

Estas fases permiten pasar de la observación empírica a la comprensión analítica. En Ingeniería en Organización Industrial, el método estadístico es aplicable a estudios de tiempos y movimientos, análisis de fallos, control estadístico de procesos y estudios de satisfacción del cliente.

Tipos de variables y escalas de medición

- Variables cualitativas: expresan cualidades o categorías (e.g., tipo de combustible, nivel educativo).
- Variables cuantitativas: expresan magnitudes numéricas, pudiendo ser discretas o continuas.

- Escalas de medición:
 1. Nominal: clasifica sin orden (ej. colores, sectores económicos).
 2. Ordinal: establece un orden sin distancia cuantitativa (ej. niveles de satisfacción).
 3. De intervalo: distancias iguales, sin cero absoluto (ej. temperatura en °C).
 4. De razón: distancias iguales con cero absoluto (ej. peso, ingresos).

Tipos de sesgos

- Sesgo de selección: muestra no representativa.
- Sesgo de información: errores en la medición o registro.
- Sesgo de confusión: influencia de variables externas no controladas.
- Sesgo de publicación: preferencia por resultados significativos.
- Sesgo de confirmación: interpretación selectiva de resultados.
- Sesgo de supervivencia: atención solo a los casos exitosos.
- Sesgo de respuesta: respuestas condicionadas por presión social.
- Sesgo del observador: influencia de las expectativas del investigador.

Estudios observacionales y experimentales

En el ámbito de la investigación estadística, es fundamental distinguir entre dos tipos principales de estudios: observacionales y experimentales.

- ☐ Estudios observacionales: Se caracterizan por la ausencia de intervención directa del investigador. Su función es registrar y analizar fenómenos tal como ocurren en la realidad, sin modificar las condiciones. Por ejemplo, estudiar la relación entre la edad de los individuos y su uso del transporte público permite observar patrones sin alterar el comportamiento de los participantes.
- ☐ Estudios experimentales: En este caso, el investigador manipula deliberadamente una o más variables para observar el efecto que dicha intervención produce. Un ejemplo sería evaluar el impacto de un nuevo sistema de iluminación en la seguridad industrial, donde se introduce una modificación controlada para medir sus consecuencias.

Ambos enfoques pueden ser válidos, siempre que se apliquen correctamente. El método estadístico, cuando se utiliza de forma rigurosa, garantiza la validez de las conclusiones y la objetividad en la interpretación de los resultados. Estos principios son pilares fundamentales del análisis cuantitativo, especialmente en las ciencias aplicadas y en la ingeniería.

Diferencias entre población y muestra

En estadística, la población representa el conjunto completo de elementos o individuos que se desea estudiar, mientras que una muestra es un subconjunto representativo extraído de esa población. Por ejemplo, si se quiere analizar la calidad de tornillos producidos en una fábrica, la población serían todos los tornillos fabricados en un día, y la muestra serían los 50 tornillos seleccionados aleatoriamente para inspección. Trabajar con muestras permite ahorrar tiempo y recursos, aunque introduce incertidumbre que debe gestionarse mediante técnicas inferenciales.

Diferencias entre estadística descriptiva e inferencial

La estadística descriptiva se enfoca en resumir y visualizar los datos mediante medidas como la media, la mediana, la desviación típica y gráficos. No busca generalizar, sino describir lo observado. En cambio, la estadística inferencial utiliza los datos muestrales para hacer afirmaciones sobre la población, estimar parámetros desconocidos o realizar pruebas de hipótesis. Por ejemplo, calcular la media de defectos en una muestra es descriptivo; usar esa media para estimar la calidad del lote completo es inferencial.

Ejemplos reales de manipulación de gráficos

En el ámbito empresarial e industrial, los gráficos pueden ser manipulados para influir en la interpretación de los datos. Por ejemplo, alterar el eje vertical de un gráfico de barras puede exagerar diferencias mínimas entre procesos productivos. También se puede omitir información relevante o usar escalas no proporcionales para presentar resultados más favorables. Por eso, es fundamental desarrollar pensamiento crítico al interpretar visualizaciones estadísticas.

Etapas del conocimiento estadístico: enfoque descriptivo vs inferencial

Las etapas clásicas del análisis estadístico —recolección de datos, organización, resumen y presentación— están principalmente orientadas a la estadística descriptiva. Sin embargo, para tomar decisiones informadas en ingeniería, es necesario avanzar hacia la estadística inferencial, que permite generalizar resultados, evaluar riesgos y optimizar procesos bajo incertidumbre. La transición del enfoque descriptivo al inferencial marca el paso de observar datos a comprender y predecir fenómenos.

Grado en Ingeniería en Organización Industrial**Tema 2. Técnicas de recogida de datos*****Bloque temático I: Estadística e información*****2.1 Recogida de datos: objetivos y limitaciones.****2.2 Fuentes de Información Estadística (Pública y Privada)****Introducción**

En Ingeniería en Organización Industrial las decisiones no se toman al azar, se programan turnos en función de la demanda prevista, se reconfiguran líneas de producción según la tasa de defectos, se renegocian contratos logísticos cuando cambian los tiempos de entrega y se planifica la capacidad energética atendiendo a consumos reales. En todos estos escenarios la materia prima es el dato. Ahora bien, los datos no valen por sí mismos: su utilidad depende de cómo se han recogido, con qué precisión, en qué condiciones y para responder a qué preguntas. Este tema aborda, de manera conectada, tres cuestiones: cómo planificar la recogida de información, qué diseños de muestreo conviene emplear y de qué fuentes —públicas y privadas— podemos obtener datos fiables que permitan conclusiones sólidas. La idea-fuerza es sencilla: medición de calidad → análisis válido → decisión mejor.

2.1 Recogida de datos: objetivos y limitaciones

El objetivo de la recogida de datos es obtener información relevante y fiable sobre variables bien definidas para poder estimar parámetros, describir fenómenos, contrastar hipótesis o modelizar relaciones. Si pretendemos saber si un cambio de herramienta disminuye la tasa de defectos, necesitamos datos que reflejen ese fenómeno y no otro; si queremos comparar productividad entre turnos, necesitamos mediciones que no estén contaminadas por diferencias de producto o por paradas programadas.

La primera limitación es económica y temporal: medirlo todo (un censo) suele ser inviable y, aunque lo fuera, retrasaría decisiones urgentes. La segunda limitación son los errores de medición y registro: sensores descalibrados, duplicidades en hojas de máquina, campos en blanco o cambios de formato en el ERP. La tercera son los sesgos: seleccionar solo la parte “fácil” de la población, registrar en el turno con el personal más colaborador, o diseñar un cuestionario que sugiera respuestas. Por último, aparece la accesibilidad: hay variables cuya observación requiere permisos, formación o paradas que no siempre son posibles.

Planificar bien evita buena parte de estos problemas. Se empieza redactando una pregunta central (por ejemplo, “¿qué proporción de piezas presenta defecto de soldadura tras introducir el robot X?”) y descendiendo a objetivos operativos (precisión deseada, horizonte temporal, líneas afectadas). Se definen variables y unidades: qué es exactamente “defecto”, cómo se codifica, si el tiempo de ciclo se mide en segundos de máquina o en segundos de operador. Se

acota el ámbito geográfico y temporal: planta A y B, turnos de mañana y tarde, durante cuatro semanas para cubrir variabilidad. Se construye un marco muestral que enumere sin omisiones ni duplicados todos los elementos posibles (lotes, piezas, pedidos). Solo entonces tiene sentido fijar el tamaño muestral, elegir diseño de muestreo y asignar recursos (personal, instrumentos, TI y calendario). Esta secuencia no es burocracia: es lo que separa un estudio replicable de una colección de anécdotas.

La recogida de datos determina la validez del estudio. Su objetivo es obtener información relevante y fiable sobre las variables, minimizando el sesgo y los errores de medición, con un coste y tiempo razonables.

La recogida de datos constituye una fase crítica en cualquier estudio estadístico, ya que de ella depende directamente la validez de los resultados obtenidos. Su principal objetivo es obtener información relevante y fiable sobre las variables de interés, minimizando tanto el sesgo como los errores de medición, todo ello dentro de unos límites razonables de coste y tiempo.

Objetivos de la recogida de datos:

- ☐ **Estimar parámetros** como medias, proporciones o tasas que describan cuantitativamente la población.
- ☐ **Describir fenómenos** y analizar su variabilidad para comprender mejor su comportamiento.
- ☐ **Contrastar hipótesis** y evaluar posibles mejoras en procesos, productos o servicios.
- ☐ **Modelizar relaciones** entre variables con el fin de facilitar la toma de decisiones informadas.

Limitaciones frecuentes:

- ☐ **Coste y tiempo**, que pueden condicionar el tamaño muestral o la profundidad del estudio.
- ☐ **Errores de medición o registro**, que afectan la precisión de los datos.
- ☐ **Sesgos**, como los de selección, no respuesta o del observador, que pueden distorsionar los resultados.
- ☐ **Acceso restringido a la población**, lo que dificulta la representatividad de la muestra.

Aspectos a planificar

Antes de iniciar cualquier proceso de recogida de datos, es imprescindible definir con precisión los aspectos clave que guiarán la investigación. Una planificación adecuada permite optimizar recursos, minimizar errores y asegurar la relevancia de la información obtenida.

Los principales aspectos a considerar son:

Finalidad de la investigación: Determinar el propósito del estudio, ya sea descriptivo, explicativo, predictivo o evaluativo.

Caracteres o variables a recoger: Identificar qué atributos se van a medir, asegurando que sean pertinentes y medibles.

Ámbito geográfico y temporal: Delimitar el espacio físico y el periodo de tiempo en el que se desarrollará la investigación.

Población de estudio y marco muestral: Definir el conjunto de individuos u objetos de interés y establecer el listado o base de datos que servirá para seleccionar la muestra.

Tamaño muestral: Determinar el número de observaciones necesarias en función del nivel de precisión deseado y la variabilidad esperada en los datos.

Metodología y recursos: Establecer los métodos de recogida (encuestas, entrevistas, observación, etc.) y los recursos humanos, técnicos y financieros disponibles para llevar a cabo el estudio.

Diseños de muestreo: elegir bien para medir mejor

Si podemos medir una parte representativa en lugar de toda la población, ganamos tiempo y ahorramos costes sin perder capacidad inferencial. El diseño más básico es el **muestreo aleatorio simple**, donde todas las piezas o lotes tienen la misma probabilidad de ser elegidos. Es potente por su neutralidad, pero puede dejar poco representados subgrupos pequeños; por ejemplo, si el turno de noche produce poco, quizá apenas aparezca en la muestra por puro azar.

Cuando el proceso es continuo, resulta muy operativo el **muestreo sistemático**: se elige al azar el primer elemento y, a partir de ahí, se inspecciona cada k -ésimo. En una línea de ensamblaje donde cada pieza tarda lo mismo, es sencillo instruir “revisa cada 10.^a pieza”. Su punto débil es la **periodicidad**: si aparece un defecto cada 10 piezas, puede coincidir con el patrón y sesgar el resultado.

Si sospechamos que los subgrupos difieren entre sí, el **muestreo estratificado** es la opción prudente. Se dividen los datos en **estratos homogéneos** (turnos, materiales, máquinas) y se toma muestra en cada uno, normalmente **proporcional** a su tamaño. Así no perdemos representatividad y, además, suele mejorar la precisión a igualdad de n . La condición para que funcione es conocer bien la estructura de la población.

Cuando la población está **dispersa**, conviene el **muestreo por conglomerados**: se seleccionan algunos conglomerados completos (por ejemplo, varias plantas) y se observa todo dentro de ellos. Ahorra desplazamientos y coordinaciones, aunque, a igualdad de n , proporciona menor precisión que el estratificado.

Existen diseños **no probabilísticos** —conveniencia, cuotas, bola de nieve— que pueden ser útiles en exploratorios o para acceder a colectivos difíciles (p. ej., expertos en mantenimiento predictivo repartidos por filiales). Su límite es claro: sus resultados **no se generalizan** con garantías y, por tanto, no deberían sustentar decisiones de alto impacto sin respaldo adicional.

Tipos de muestreo

- ☐ Aleatorio simple: Todos con la misma probabilidad.
 - + Elimina sesgos.
 - Puede ignorar subgrupos. Ej.: 50 piezas al azar.
- ☐ Sistemático: Cada k -ésimo elemento tras un inicio aleatorio.
 - + Sencillo.

- Sesgo si hay patrones. Ej.: cada 10ª pieza.
- ☐ Estratificado: Muestra por estratos homogéneos.
 - + Representatividad.
 - Requiere conocer estructura. Ej.: turnos.
- ☐ Por conglomerados: Seleccionar conglomerados completos.
 - + Ahorra costes.
 - Menor precisión. Ej.: plantas completas.
- ☐ Conveniencia: Elementos accesibles.
 - + Rápido.
 - Alto sesgo. Ej.: primeros operarios.
- ☐ Cuotas: Fijar cuotas sin aleatoriedad.
 - + Mantiene proporciones.
 - Sesgo. Ej.: 40% H / 60% M.
- ☐ Bola de nieve: Participantes recomiendan a otros.
 - + Útil para expertos.
 - No representativo.

Cálculo del tamaño muestral

- ☐ Para proporciones: $n_0 = (Z^2 \cdot p \cdot q) / E^2$ (p desconocida $\rightarrow 0,5$).
- ☐ Corrección por población finita: $n = [N \cdot n_0] / [N + n_0 - 1]$.
- ☐ Para medias: $n_0 = (Z^2 \cdot \sigma^2) / E^2$.

Ejemplo (95%, $E=0,05$, $p=0,5$): $n_0 = (1,96^2 \cdot 0,5 \cdot 0,5) / 0,05^2 = 384,16 \rightarrow 385$. Si $N=2.000$: $n \approx 322$.

Técnicas de recogida de datos

Hay tres canales principales. El censo observa a toda la población; su ventaja es obvia, pero el coste y el tiempo lo reservan para operaciones puntuales o para magnitudes críticas en periodos acotados. Los registros administrativos u operativos (ERP, MES, SCADA, GMAO) generan datos de forma continua con gran trazabilidad; su limitación es que reflejan exactamente lo que el sistema captura: si un tipo de paro no está codificado, no existirá en la base de datos. Por último, la encuesta (entrevista o cuestionario) permite recoger información no observable por sensores —percepción de seguridad, usabilidad de un nuevo HMI, clima de equipo—, siempre que el instrumento esté bien diseñado.

Un buen cuestionario evita ambigüedades y dobles negaciones, no induce respuestas, ordena los bloques de lo general a lo específico y se pilota antes de lanzarlo. La regla práctica es “una idea por pregunta” y “un lenguaje que cualquier encuestado entienda sin esfuerzo”.

Una herramienta útil para medir actitudes es el diferencial semántico de Osgood. Se presenta un concepto (por ejemplo, “nuevo sistema de automatización del puesto P-120”) y se pide valorar en escalas bipolares de 1 a 7 pares como “eficiente–ineficiente”, “seguro–inseguro”, “sencillo–complejo” o “ruidoso–silencioso”. Aunque capta percepciones subjetivas, al codificarse numéricamente permite análisis comparables entre turnos, antigüedades o plantas.

En el entorno industrial conviven sensores y encuestas. Los primeros entregan datos en tiempo real, reducen el error humano y soportan análisis de alta frecuencia; requieren calibración,

mantenimiento y auditorías periódicas. Las segundas capturan contexto y valoración humana, y por tanto completan el cuadro. Lo sensato no es elegir entre una u otra, sino integrarlas: por ejemplo, explicar un aumento de vibración a través de un cambio de procedimiento reportado por los operarios.

La fiabilidad de lo medido se protege con metrología y proceso: calibraciones programadas, mantenimiento preventivo, verificaciones cruzadas con patrones, control de versiones en catálogos de códigos y una política clara de depuración. Un principio fundamental es el tratamiento de valores imposibles: si un tiempo de ciclo aparece negativo, eso no es “un valor raro”: es un error de captura que debe rastrearse a su origen (sensor, reloj, software, carga de datos) y corregirse. Si se repite, hay que rediseñar el procedimiento o implantar reglas de validación en el sistema.

- Censo: toda la población (coste alto).
- Registros: fuentes continuas (ERP/MES/SCADA).
- Encuesta: muestra representativa mediante entrevista o cuestionario.

Entrevista y cuestionario: pautas

- Redacción clara; evitar ambigüedades y dobles negaciones.
- Evitar preguntas que induzcan la respuesta.
- Orden lógico y prueba piloto.

Diferencial semántico de Osgood (1957)

Escalas bipolares (1–7) con adjetivos opuestos; convierte percepciones en datos cuantificables.

Sensores vs. encuestas

En el ámbito de la recolección de datos, tanto los sensores como las encuestas ofrecen ventajas y limitaciones que deben considerarse según el objetivo del estudio.

Los sensores permiten obtener datos en tiempo real con alta frecuencia y menor intervención humana, lo que reduce el riesgo de errores subjetivos. Sin embargo, requieren calibración periódica, mantenimiento técnico y pueden estar limitados por el entorno físico o la infraestructura disponible.

Las encuestas, por otro lado, son herramientas valiosas para captar percepciones, opiniones y comportamientos que no pueden ser medidos directamente por dispositivos. Aunque su frecuencia de aplicación suele ser menor y están sujetas a sesgos de respuesta o errores de diseño, siguen siendo fundamentales para comprender aspectos cualitativos de la realidad social.

La elección entre sensores y encuestas dependerá del tipo de información que se desea obtener, los recursos disponibles y el contexto específico de la investigación.

Fiabilidad y control de errores

Para garantizar la calidad de los datos y la robustez de los resultados estadísticos, es fundamental aplicar procedimientos sistemáticos de control. Entre ellos destacan:

- ☐ **Calibración** de los instrumentos de medición, asegurando que los valores registrados sean precisos y consistentes.
- ☐ **Mantenimiento preventivo** de los equipos utilizados, evitando fallos técnicos que puedan comprometer la recogida de datos.
- ☐ **Validación cruzada**, que permite verificar la coherencia de los datos mediante comparaciones entre diferentes fuentes o métodos.
- ☐ **Trazabilidad**, asegurando que cada dato pueda ser rastreado hasta su origen, lo que facilita auditorías y revisiones.

Además, hay que identificar y corregir errores evidentes en los datos:

- ☐ **Detección de datos imposibles**, como tiempos negativos o valores fuera de los rangos esperados.
- ☐ **Verificación y depuración** de los registros, eliminando o corrigiendo aquellos que no cumplen con los criterios de calidad establecidos.

2.2 Fuentes de información estadística (pública y privada)

Una **fuentes** es cualquier repositorio o documento que aporta datos útiles para responder una pregunta de investigación. Las **fuentes primarias** generan información original: un censo, una encuesta propia, un registro de producción, un parte de mantenimiento o una descarga directa

de un SCADA. Las **fuentes secundarias** reorganizan o interpretan datos previamente recogidos: un informe del INE, un anuario sectorial, un panel privado o un artículo técnico que integra varias bases.

En el trabajo académico conviene priorizar **fuentes públicas oficiales** por su trazabilidad: el **INE** y **INEbase** ofrecen, entre otras, la **EPA** (empleo), el **IPI** (producción industrial), el **IPC** (precios) o el **DIRCE** (estructura empresarial); **Eurostat** permite comparaciones homogéneas entre países de la UE; la **Agencia Tributaria** y los bancos centrales proporcionan series macro y fiscales; los portales **autonómicos y municipales** añaden granularidad local para movilidad, vivienda o comercio. Las **fuentes privadas** —**Statista**, **Nielsen**, **Ipsos**, informes del **Banco de España**, **IET/OMT**, **AENA**— aportan rapidez, foco sectorial o indicadores que la estadística pública no cubre con la misma frecuencia. En todos los casos hay que **documentar fecha de descarga, versión, definiciones y cambios metodológicos**, y, si una magnitud es crítica para una decisión, **triangular** con más de una fuente.

Definición y clasificación: primarias (datos originales) y secundarias (organizan/interpretan).

Fuentes públicas útiles:

1. INE e INEbase (EPA, IPC, IPI, Directorio Central de Empresas).
2. Eurostat (series comparables UE).
3. Agencia Tributaria; portales autonómicos y municipales (Comunidad de Madrid, Ayuntamiento de Madrid).
4. Fuentes privadas:
5. Statista, Nielsen, Ipsos; Banco de España; IET/OMT; AENA.

Notas de uso responsable de fuentes

- ☐ Citar fecha/versión y metodología; comprobar cambios de base.
- ☐ Priorizar fuentes oficiales para trabajos académicos y auditorías.

Aplicación a Ingeniería en Organización Industrial

Ejemplos: dimensionar una muestra de inspección de calidad; diseñar un cuestionario de seguridad; usar Eurostat e INE para planificar capacidad; combinar registros SCADA con encuestas de operarios para evaluar la aceptación de una automatización.

Una recogida de datos bien diseñada, con muestreo apropiado, medición fiable y fuentes contrastadas, sustenta decisiones robustas de mejora continua en la organización industrial.

Caso práctico integrado: rediseño del plan de inspección en una línea de componentes metálicos

La dirección de planta quiere reducir retrabajos por soldadura. La **pregunta central** es si la introducción de un nuevo robot reduce la **proporción** de piezas con defecto respecto al sistema anterior. Se decide medir durante cuatro semanas en los tres turnos y en dos materiales.

Población y marco. Todas las piezas producidas en la línea L-3 en ese periodo, clasificadas por turno y material. Se depuran duplicidades en el ERP y se aseguran códigos homogéneos de defecto.

Diseño. Para garantizar comparabilidad entre turnos se adopta **muestreo estratificado proporcional** (los turnos producen cantidades distintas). Se fija una precisión de **± 5 puntos porcentuales** al 95 % y, como no hay datos previos fiables, se usa ($p=0,5$). El cálculo entrega ($n_0=384,16$). Como la producción prevista son unas **2.000 piezas/día**, si el estudio durara un solo día aplicaríamos corrección y obtendríamos **$n=322$** ; repartido por cuatro semanas, la muestra diaria es moderada y operativa. Se añade un 10 % por incidencias.

Medición. Se definen criterios de aceptación y se forma a los inspectores. Se registran variables auxiliares (máquina, operador, lote) para poder analizar patrones. Paralelamente, se instala un sensor que registra parámetros de corriente del robot y temperatura de baño: no sustituyen a la inspección visual, pero ayudan a explicar cambios.

Encuesta complementaria. A los operarios y mandos se les pasa un breve cuestionario con un **Osgood** sobre el nuevo robot (eficiente–ineficiente, seguro–inseguro, sencillo–complejo, interferente–fluido en el flujo de trabajo). Con ello se captan barreras de adopción que los datos de calidad no revelan por sí solos.

Depuración y análisis. Se bloquea en el sistema la posibilidad de tiempos negativos y se revisan valores fuera de rango. Se calcula la proporción de defectos por turno y material, con intervalos de confianza; si procede, se contrasta la **diferencia de proporciones** antes/después. Se cruzan resultados con parámetros de sensor y con percepciones de operarios para identificar causas técnicas o organizativas.

Decisión. Si la reducción es estadísticamente significativa y económicamente relevante, se estandariza el robot; si hay diferencias por turno, se revisa formación y procedimiento; si el Osgood detecta “complejidad alta” y “seguridad percibida baja”, se ajusta el HMI o la protección. Todo el proceso queda **documentado** (método de muestreo, fórmulas, versiones de catálogo de defectos, fechas y fuentes), de manera que la auditoría de calidad pueda replicarlo.

La recogida de datos implica el **diseño del estudio estadístico** que conecta la pregunta con la decisión. Elegir adecuadamente el **muestreo**, justificar el **tamaño muestral** con cálculos claros, combinar **sensores** con **encuestas** cuando proceda, blindar la **fiabilidad** con metrología y buenas prácticas, y apoyarse en **fuentes trazables** son los ingredientes de un análisis que se sostiene en el tiempo. Con estos elementos, el ingeniero en organización industrial puede traducir datos en mejoras reales de productividad, calidad y seguridad.

Grado en Ingeniería en Organización Industrial**Bloque II - Análisis Estadístico de una variable****Tema 3: Distribuciones de frecuencia. Medidas de posición, de dispersión y de forma**

3.1 Variable Estadística.**3.2 Distribuciones de frecuencias.****3.3 Momentos.****3.4 Representaciones gráficas.****3.5 Medidas de posición centrales: media aritmética, mediana y moda.****3.6 Medidas de posición no centrales: cuartiles, deciles y percentiles.****3.7 Medidas de dispersión absolutas: varianza y desviación típica.****3.8 Medidas de dispersión relativas: coeficiente de variación****3.9 Tipificación.****3.10 Medidas de asimetría y curtosis.****3.11 Datos atípicos y diagramas de caja****Introducción**

El análisis estadístico de una variable es el primer paso para comprender el comportamiento de los datos en procesos industriales, logísticos y organizativos. Este tema se centra en las herramientas fundamentales de la estadística descriptiva, que permiten resumir, visualizar y caracterizar conjuntos de datos mediante medidas numéricas y representaciones gráficas.

Se comienza con la definición de variable estadística, que puede ser cualitativa o cuantitativa, y cuya naturaleza determina el tipo de análisis aplicable. A partir de los datos recogidos, se construyen distribuciones de frecuencia, que organizan la información en tablas y permiten observar patrones, concentraciones y dispersión.

El estudio de los momentos estadísticos y las representaciones gráficas (como histogramas, polígonos de frecuencia o diagramas de caja) facilita la interpretación visual de los datos. Las medidas de posición centrales —media, mediana y moda— describen el valor típico o representativo, mientras que las medidas no centrales —cuartiles, deciles y percentiles— permiten analizar la distribución en segmentos.

Las medidas de dispersión, tanto absolutas (como la varianza y la desviación típica) como relativas (como el coeficiente de variación), indican el grado de variabilidad de los datos respecto a su centro. La tipificación permite estandarizar variables para compararlas en distintas escalas.

Finalmente, se abordan las medidas de forma, como la asimetría y la curtosis, que describen la simetría y la concentración de los datos, así como la identificación de datos atípicos mediante herramientas como el diagrama de caja, fundamentales para detectar anomalías en procesos industriales.

Este conjunto de técnicas permite al ingeniero en organización industrial transformar datos en información útil para la mejora continua, el control de calidad, la toma de decisiones y la optimización de recursos.

1. Variable estadística

Definición: Una variable estadística es la característica que se estudia en una población.

Población: Todo el conjunto de individuos o elementos que tienen características comunes.

Herramienta estadística: censo.

Muestra: sub-conjunto representativo de la población. **Herramienta:** encuesta.

Una variable puede ser:

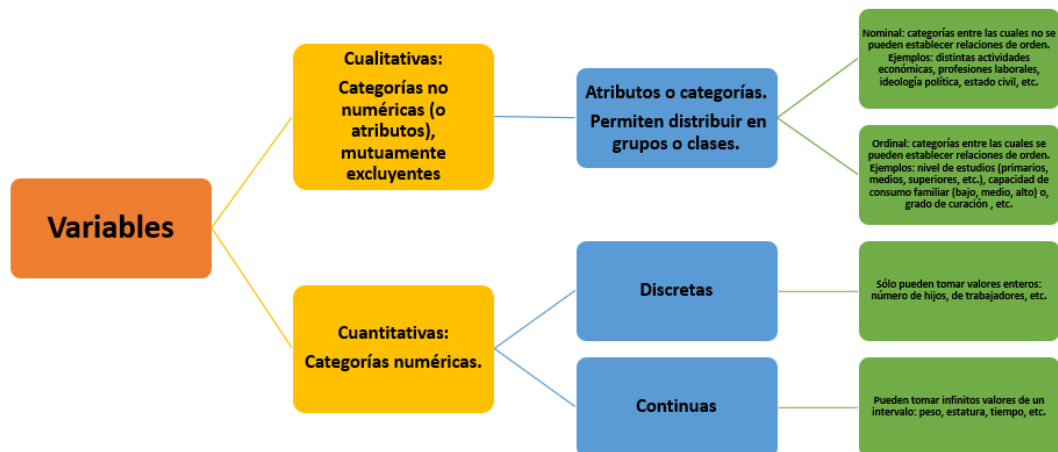
☐ **Cualitativa:**

- o *Nominal:* No tienen orden (ej. estado civil, profesión).
- o *Ordinal:* Tienen orden (ej. nivel educativo, nivel de ingresos).

☐ **Cuantitativa:**

- o *Discreta:* Toman valores enteros (ej. número de hijos).
- o *Continua:* Toman valores dentro de un intervalo (ej. peso, altura).

Clasificación:



La **población** es el conjunto completo de individuos con características comunes. Se estudia mediante censos. La **muestra** es un subconjunto representativo de la población, estudiado mediante encuestas.

2. Distribuciones de frecuencias

Datos no agrupados:

- ☐ **Valor de la variable (xi):** Diferentes valores que toma la variable.
- ☐ **Frecuencia absoluta (ni):** Número de veces que aparece cada valor.
- ☐ **Frecuencia relativa (fi):** Proporción de cada valor respecto al total: $f_i = n_i / N$
- ☐ **Frecuencia acumulada absoluta (Ni):** Suma de frecuencias hasta ese valor.
- ☐ **Frecuencia acumulada relativa (Fi):** $F_i = N_i / N$

Ejemplo: Se realizó una encuesta en relación al número de visitas al médico durante un año y se obtuvieron las siguientes respuestas: 5, 9, 10, 13, 1, 2, 7, 13, 10, 7, 7, 7, 10, 1, 1, 5, 5, 10, 2, 7, 10, 10, 5, 9. Represente los datos en una tabla de distribución de frecuencias.

x_i	n_i	$f_i = \frac{n_i}{N}$	N_i	$F_i = \frac{N_i}{N}$
1	3	$=3/24= 0,125$	3	$=3/24= 0,125$
2	2	$=2/24= 0,083$	$3+2 = 5$	$=5/24 = 0,208$
5	4	0,167	$5+4 =9$	0,375
7	5	0,208	14	0,583
9	2	0,083	16	0,667
10	6	0,25	22	0,917
13	2	0,083	24	1
	$\sum n_i = 24 = N$	$\sum f_i = 0,999 \approx 1$		

Datos agrupados:

- ☐ **Intervalos o clases:** Rango de valores agrupados.
- ☐ **Amplitud del intervalo (ci):** Diferencia entre extremos del intervalo.
- ☐ **Marca de clase ($\bar{x}_i$):** Promedio de los extremos del intervalo.
- ☐ **Densidad de frecuencia (di):** $d_i = n_i / ci$

3. Momentos

Respecto al origen:

Se calculan como:

Donde:

x_i : valores de la variable.

0: el origen se coloca en el cero.

n_i : frecuencia absoluta de x_i .

N: número total de datos.

r: orden del momento.

Los más utilizados son los de orden 1 y 2 (se emplearán en el cálculo de la media y de la varianza).

Respecto a la media (momentos centrales):



Donde:

x_i : valores de la variable.

$\bar{x}$: media aritmética.

n_i : frecuencia absoluta de x_i .

N : número total de datos.

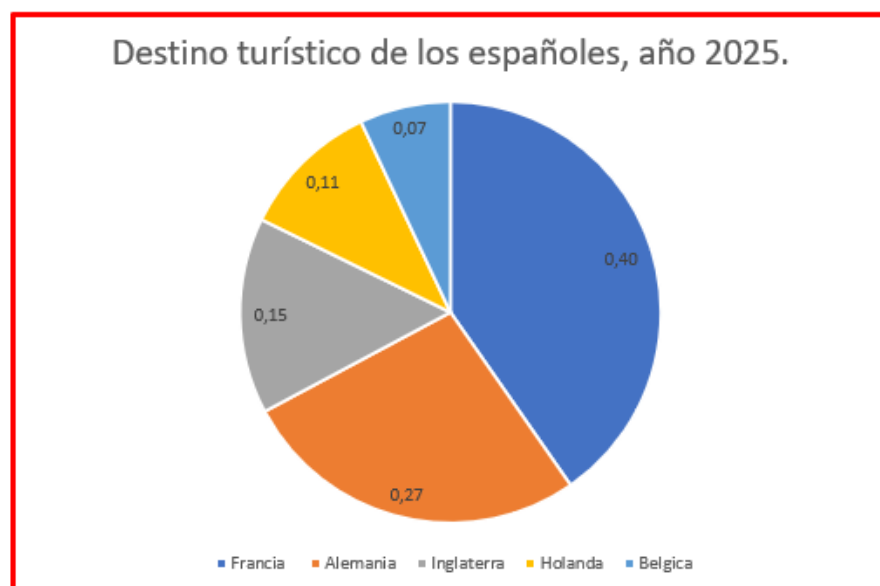
r : orden del momento.

Los más empleados son de orden 2 (varianza), 3 y 4.

4. Representaciones gráficas

Variables cualitativas:

- ☐ Diagramas de barras y sectorial (o de tarta).

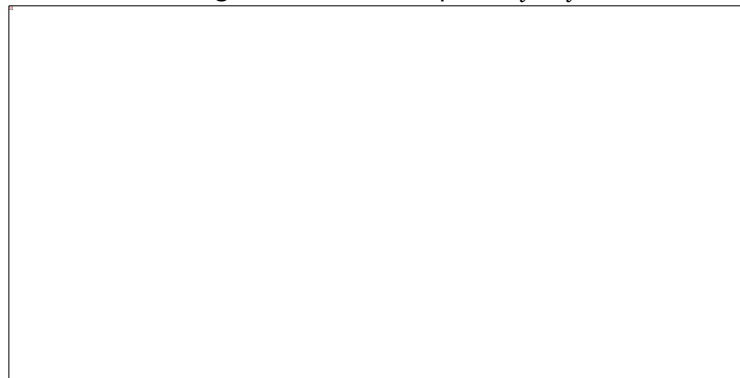


Variables cuantitativas discretas en distribuciones no agrupadas:

- ☐ Diagrama de barras para n_i y/o f_i .

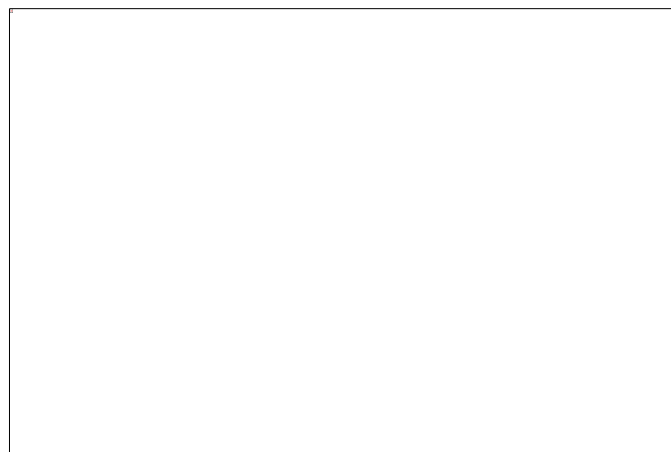


- ☐ Diagrama escalonado para N_i o F_i .



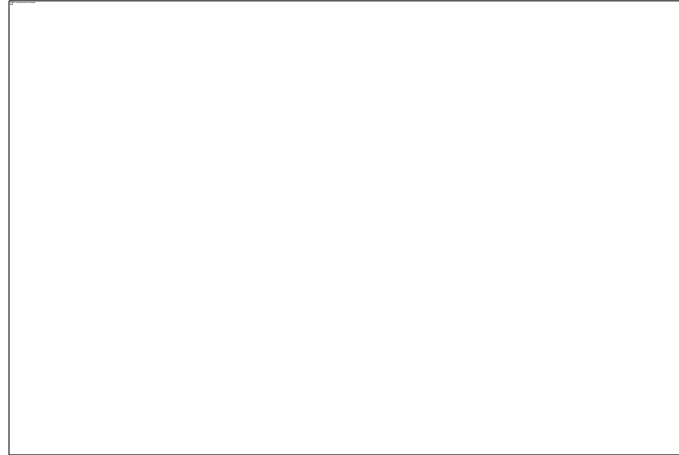
Variables cuantitativas en distribuciones agrupadas:

- ☐ Histograma de frecuencias.



- ☐ Polígono de frecuencias para d_i o n_i

- ☐ Polígono de frecuencias acumulado para N_i o F_i .



5. Medidas de posición centrales

Media aritmética:

Media aritmética ($\bar{x}$):



Ventajas:

En su cálculo intervienen todos los valores de la distribución.

Es única.

Siempre es calculable (salvo en intervalos abiertos).

Inconvenientes:

Es muy sensible a valores extremos (muy bajos o muy altos).

Mediana:

- ☐ Valor central de la distribución ordenada.
- ☐ Si N es impar: posición $(N+1)/2$
- ☐ Si N es par: promedio de los dos valores centrales.

Moda:

- ☐ Valor más frecuente.

- ☐ Puede no ser única (bimodal, multimodal).

6. Medidas de posición no centrales

- ☐ **Cuartiles (Q1, Q2, Q3):** Dividen en 4 partes la distribución.
- ☐ **Deciles:** Dividen en 10 partes.
- ☐ **Percentiles:** Dividen en 100 partes.

7. Medidas de dispersión absolutas

Varianza:

- ☐ $s^2 = \sum ((x_i - \bar{x})^2 * f_i)$
- ☐ No puede ser negativa. A mayor varianza, mayor dispersión.

Desviación típica:

- ☐ $s = \sqrt{s^2}$
- ☐ Tiene las mismas unidades que la variable.

Rango:

- ☐ $R = x_{\max} - x_{\min}$

Recorrido intercuartílico:

- ☐ $RI = Q3 - Q1$

8. Medidas de dispersión relativas

Coeficiente de variación de Pearson:

- ☐ $CV = s / \bar{x}$
- ☐ Adimensional. Útil para comparar distribuciones con distintas unidades.

9. Tipificación

- ☐ **Z-score:**
 - ☐ $Z_i = (x_i - \bar{x}) / s$
 - ☐ Indica cuántas desviaciones típicas se aleja un valor de la media.

10. Medidas de forma

Asimetría:

- ☐ **Distribución simétrica:** media = mediana = moda.
- ☐ **Distribución asimétrica positiva:** cola a la derecha.
- ☐ **Distribución asimétrica negativa:** cola a la izquierda.

Coeficiente de asimetría de Fisher:

$$\square \quad g_1 = m_3 / s^3$$

Curtosis:

- ☐ **Leptocúrtica:** más apuntada que la normal.
- ☐ **Mesocúrtica:** igual que la normal.
- ☐ **Platicúrtica:** menos apuntada.

Coeficiente de curtosis de Fisher:

$$\square \quad g_2 = m_4 / s^4 - 3$$

11. Datos atípicos y diagramas de caja

Los datos atípicos son valores que se alejan significativamente del resto de la distribución. Se identifican comúnmente mediante diagramas de caja (boxplots), que muestran:

- ☐ Mediana
- ☐ Cuartiles
- ☐ Valores extremos
- ☐ Posibles outliers

Estos diagramas permiten visualizar la dispersión, simetría y presencia de valores atípicos en una distribución.

Grado en Ingeniería en Organización Industrial**Tema 4. Distribuciones bidimensionales. Tablas de contingencia.*****Bloque temático III: Análisis estadístico de dos variables***

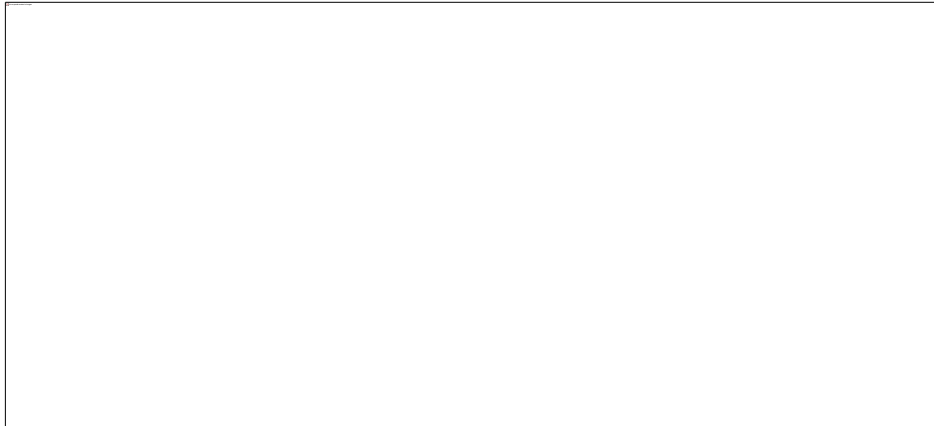
4.1 Tablas de correlación y contingencia.**4.2 Representación gráfica: nube de puntos.****4.3 Distribuciones marginales.****4.4 Distribuciones condicionadas.****4.5 Dependencia e independencia estadística.****4.1. Tablas de correlación y contingencia**

Cuando se estudia una población o conjunto de datos, a menudo interesa analizar simultáneamente dos características o variables de cada individuo. A este tipo de estudio se le denomina análisis bidimensional. Por ejemplo, podríamos analizar el salario y la antigüedad en la empresa de un grupo de trabajadores, la renta y el gasto en un conjunto de familias, o el número de empleados y la producción en una empresa. En todos estos casos, cada individuo se caracteriza por un par de valores (x, y) que forman lo que se conoce como variable estadística bidimensional.

Los datos bidimensionales pueden presentarse de dos formas:

- Sin agrupar: cuando se dispone de los n pares de observaciones individuales $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.
- Agrupados: cuando las observaciones se resumen en una tabla de frecuencias, indicando cuántas veces aparece cada combinación de valores (x_i, y_j) .

Cuando se analizan simultáneamente dos variables estadísticas, se habla de distribución bidimensional. En este caso, cada individuo o elemento de la población se caracteriza por dos valores, uno para cada variable. Las tablas de contingencia o de correlación permiten organizar los datos observados según las combinaciones posibles de ambas variables.

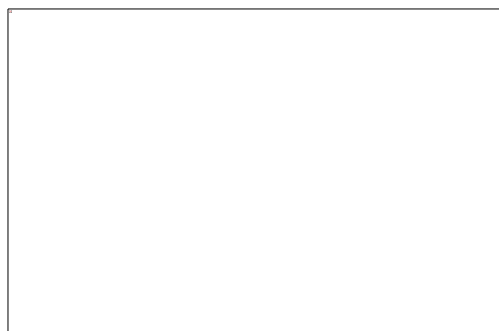


Si ambas variables son cualitativas o discretas, se construye una tabla de contingencia, en la que cada celda representa el número de individuos que presentan simultáneamente una modalidad de la primera variable y una de la segunda. Si una o ambas variables son cuantitativas, suele denominarse tabla de correlación, donde los valores se agrupan por intervalos o clases numéricas.

El objetivo principal del análisis bidimensional es detectar si existe algún tipo de relación o dependencia entre las variables. Pueden darse distintos grados de relación: desde la independencia total hasta una relación funcional exacta, en la que cada valor de X determina un único valor de Y .

Distribución conjunta

La distribución conjunta describe cómo se combinan los valores de las dos variables. Se representa mediante una tabla de doble entrada, donde las filas corresponden a los valores de X y las columnas a los valores de Y . Cada celda contiene la frecuencia n_{ij} , que indica cuántos individuos presentan simultáneamente el valor x_i de X y el valor y_j de Y . Si se dividen todas las frecuencias por el total N , se obtienen las frecuencias relativas f_{ij} , cuya suma total es 1.



Cuando las variables son cuantitativas, esta tabla se denomina tabla de correlación; si son cualitativas, se llama tabla de contingencia. La distribución conjunta permite identificar cómo se

asocian los valores de una variable con los de la otra y es el punto de partida para obtener distribuciones marginales y condicionadas.

Distribuciones marginales

Las distribuciones marginales son las distribuciones individuales de cada variable dentro de una distribución bidimensional. Se obtienen sumando las frecuencias de la tabla conjunta por filas (para X) o por columnas (para Y). De esta forma, se obtiene la frecuencia total con que aparece cada valor de una variable, independientemente de los valores de la otra.

Por ejemplo, la distribución marginal de X se obtiene sumando todas las frecuencias correspondientes a cada fila, mientras que la de Y se obtiene sumando las frecuencias de cada columna. El total de las frecuencias marginales siempre coincide con el número total de observaciones N. A continuación, las distribuciones marginales de x y de y:

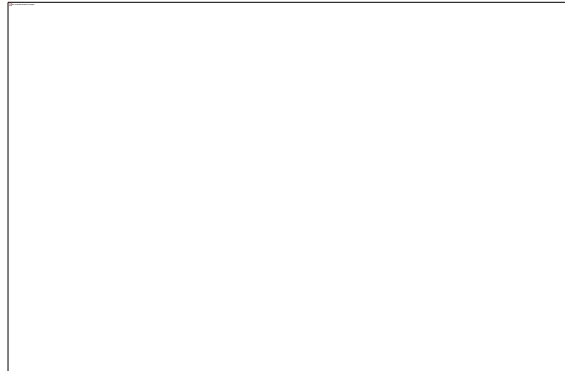
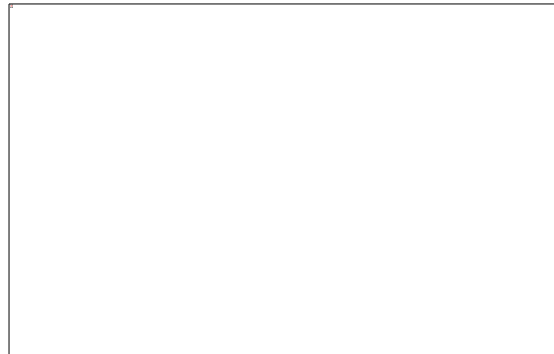
x_i	$n_{i\cdot}$	$f_{i\cdot}$	
x_1	$n_{1\cdot}$	$f_{1\cdot}$	
x_2	$n_{2\cdot}$	$f_{2\cdot}$	
...	...	...	
x_i	$n_{i\cdot}$	$f_{i\cdot}$	
...	...	...	
x_h	$n_{h\cdot}$	$f_{h\cdot}$	
Total	N	1	

Distribuciones condicionadas

Las distribuciones condicionadas muestran cómo se distribuye una variable cuando se fija un valor concreto de la otra. Por ejemplo, podemos estudiar la distribución de X condicionada a un valor determinado de Y, o viceversa. Matemáticamente, las frecuencias condicionadas se obtienen dividiendo las frecuencias conjuntas n_{ij} por la frecuencia marginal correspondiente.

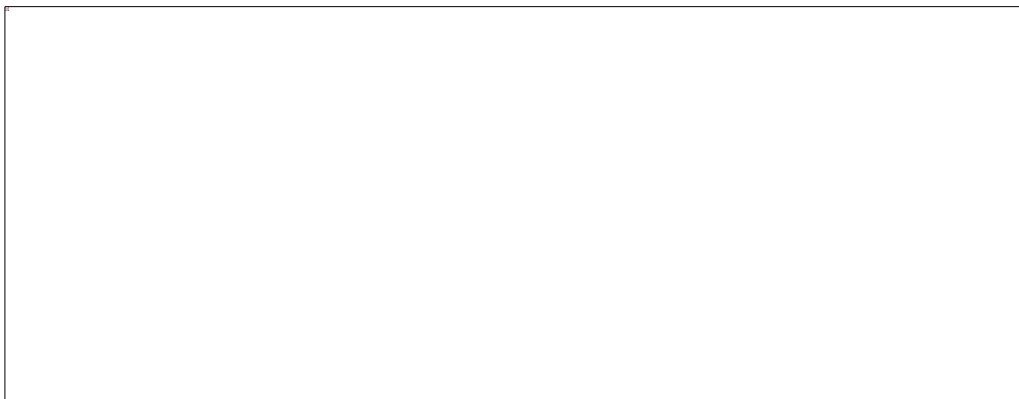
De esta forma, la distribución de $X/Y = y_j$ se calcula como $f_{i/j} = n_{ij} / n_{\cdot j}$, mientras que la distribución de $Y/X = x_i$ se obtiene como $f_{j/i} = n_{ij} / n_{i\cdot}$. Cada una de estas distribuciones condicionadas se interpreta como una distribución unidimensional, pero restringida al subconjunto de observaciones donde la variable condicionante toma un valor fijo.

Distribución condicionada de X/ y_j

Distribución condicionada de $Y/X = x_i$ 

4.2. Representación gráfica: nube de puntos

Cuando ambas variables son cuantitativas, la forma más útil de representar los datos es la nube de puntos. Cada punto del gráfico representa un individuo, cuyas coordenadas son los valores observados de las dos variables. Este tipo de representación permite detectar visualmente relaciones o tendencias: por ejemplo, si existe una relación lineal creciente o decreciente entre las variables.



4.3. Distribuciones marginales

A partir de una distribución bidimensional se pueden obtener las distribuciones marginales, que muestran cómo se distribuye cada variable individualmente. En una tabla de contingencia, se

obtienen sumando por filas (para una variable) o por columnas (para la otra). Las frecuencias marginales totales deben coincidir con el tamaño total de la muestra (n).

Distribución marginal de la X

--

Distribución marginal de la Y

--

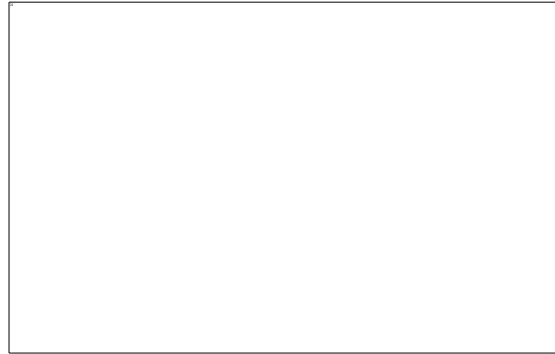
4.4. Distribuciones condicionadas

Una distribución condicionada muestra cómo se distribuye una variable cuando se fija una modalidad concreta de la otra variable. Por ejemplo, si analizamos la variable 'consumo energético' condicionada al tipo de maquinaria, obtendremos la distribución del consumo dentro de cada tipo de máquina.

Distribución condicionada de X/ y_j

x_i/y_j	$n_{i/j}$	$f_{i/j}$
x_1	n_{1j}	$f_{1/j}$
x_2	n_{2j}	$f_{2/j}$
...	...	...
x_i	n_{ij}	$f_{i/j}$
...	...	...
x_h	n_{hj}	$f_{h/j}$
Total	$n_{.j}$	1

Distribución condicionada de $Y/X = x_i$



4.5. Dependencia e independencia estadística

Dos variables son estadísticamente independientes si el valor de una no aporta información sobre la otra.

La variable Y es independiente de la variable X , cuando las “ h ” distribuciones condicionadas ($Y/X = x_i$) son iguales entre sí, e iguales a su vez a la distribución marginal de Y .

La definición es similar para la independencia de X .

Cuando Y es independiente de X , la variable X no afecta a la variable Y . La independencia es recíproca, si X es independiente de Y , también Y lo es de X .

Cuando esta condición no se cumple, las variables son dependientes, lo que indica la existencia de algún tipo de relación o asociación. En el caso de variables cuantitativas, esta relación se puede analizar más adelante mediante el coeficiente de correlación lineal.

Condición Necesaria de Independencia Estadística:

$S_{xy} = 0$ Covarianza igual a cero

La covarianza es una medida del grado de asociación lineal entre las dos variables. Su signo indica el tipo de relación: positiva si ambas aumentan o disminuyen conjuntamente, y negativa si una crece cuando la otra disminuye. Si la covarianza es próxima a cero, se considera que no existe relación lineal significativa entre las variables.

Es importante destacar que la covarianza solo mide asociaciones lineales. Si la relación entre las variables es no lineal, la covarianza puede ser cero incluso cuando existe dependencia. Por tanto, su interpretación debe hacerse siempre considerando la forma de la relación observada en los datos.

Condición Suficiente de Independencia Estadística (las frecuencias relativas condicionadas son iguales al producto de sus correspondientes frecuencias relativas marginales):



En el polo opuesto de la independencia estadística se encuentra la dependencia funcional, que se establece a partir de una fórmula matemática.

Decimos que la variable Y depende funcionalmente de la variable X , si para cada valor x_i de X , le corresponde un único valor y_j a la Y .

De manera análoga se define la dependencia funcional de la variable X , respecto de la variable Y .

Si la dependencia funcional se da en los dos sentidos decimos que es recíproca o biunívoca.

Lo habitual en estadística es trabajar con variables cuyo grado de dependencia está comprendido entre los dos extremos (independencia estadística y dependencia funcional).

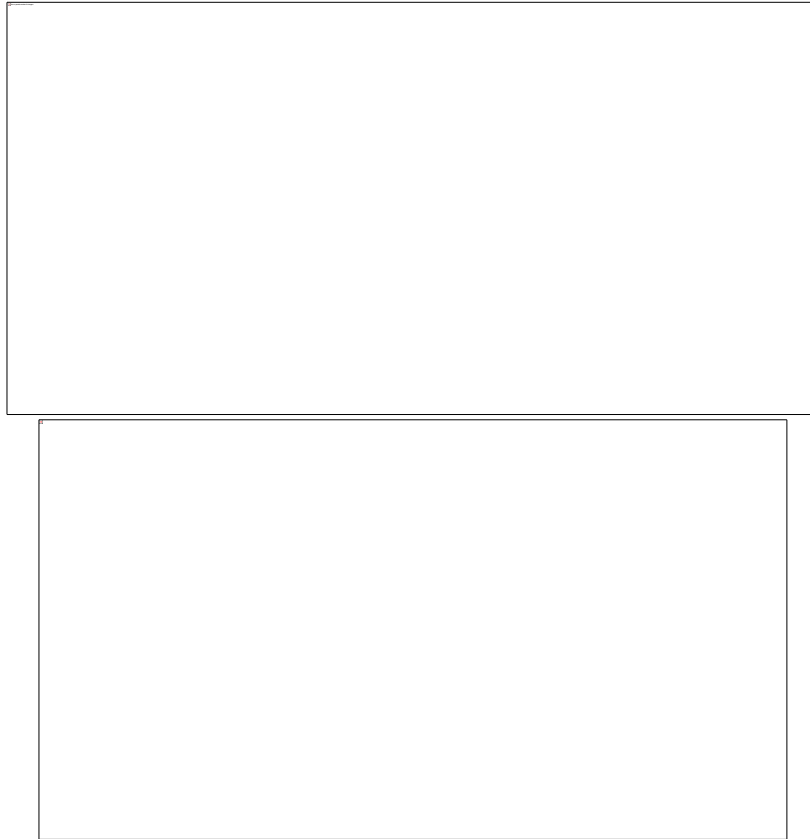
Se analiza la relación entre las variables, tanto en lo que respecta a la forma funcional que resulta más adecuada, como al grado de intensidad de la misma.

Momentos

En las distribuciones bidimensionales, los momentos permiten resumir la información sobre la localización y la dispersión de los datos. Pueden calcularse con respecto al origen o con respecto a la media.

Entre los momentos más importantes se encuentran:

- a_{11} = media de X ($\bar{x}$)
- a_{12} = media de Y ($\bar{y}$)
- m_{21} = varianza de X
- m_{12} = varianza de Y
- m_{11} = covarianza entre X e Y



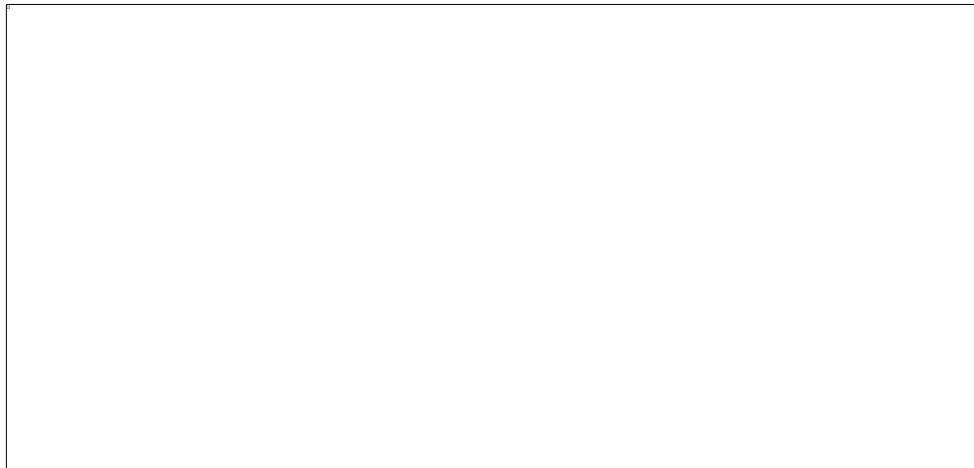
b) En observaciones agrupadas por frecuencias: $[(x_i, y_j); n_{ij}], i = 1 \dots h; j = 1 \dots k$

- Momentos con respecto al origen:

$$a_{rs} = \frac{\sum_{i=1}^h \sum_{j=1}^k x_i^r \cdot y_j^s \cdot n_{ij}}{N} = \sum_{i=1}^h \sum_{j=1}^k x_i^r \cdot y_j^s \cdot f_{ij}$$

- Momentos con respecto a la media:

$$m_{rs} = \sum_{i=1}^h \sum_{j=1}^k (x_i - \bar{x})^r \cdot (y_j - \bar{y})^s \cdot f_{ij}$$



Los puntos en los cuadrantes I y III aportan sumandos + a la covarianza, mientras que los puntos en los cuadrantes II y IV aportan sumandos negativos. Por lo tanto:

Una relación lineal positiva entre las variables, como la indicada en el gráfico, conllevará a más puntos en los cuadrantes I y III y menos en los cuadrantes II y IV y la covarianza será positiva. En el caso de una relación lineal perfecta todos los puntos se situarán en los cuadrantes I y III y la covarianza alcanzará su valor más alto.

De manera análoga, una covarianza negativa nos informará de una relación lineal negativa entre las variables. Valores altos en una de las variables se corresponderán con valores bajos en la otra.

Una covarianza grande en valor absoluto indicará un mayor grado de asociación lineal entre las variables. La covarianza positiva indicará relación lineal positiva y una covarianza negativa indicará relación lineal negativa.

La covarianza será próxima a cero cuando no exista relación entre ambas variables, o existiendo la relación sea marcadamente no lineal.

Errores frecuentes y cómo evitarlos

- Confundir distribución conjunta con distribución marginal. Recuerda que la conjunta considera ambas variables a la vez.
- Olvidar normalizar las frecuencias al construir distribuciones condicionadas. Siempre deben dividirse por la marginal correspondiente.
- Interpretar la independencia solo de forma visual: es necesario comprobarla mediante el producto de frecuencias o análisis estadístico.
- Representar la nube de puntos sin escalas adecuadas o sin unidades, lo que puede inducir a errores de interpretación.
- Suponer correlación lineal a partir de una simple tendencia: existen dependencias no lineales que requieren otras medidas.

Bibliografía

Agresti, A. (2019). Statistical Methods for the Social Sciences. Pearson.

José Javier Muruzábal Irigoyen, Dionisio Pérez Esteban (2024): Estadística y Probabilidad en la Ingeniería. 2ª Edición, Ibergarceta Publicaciones S.L. 9788419034403.

Grado en Ingeniería en Organización Industrial**Tema 5. La vinculación entre variables estadísticas: regresión y correlación*****Bloque temático III: Análisis estadístico de dos variables***

5.1 Planteamiento general del problema de la regresión.**5.2 Regresión lineal. Estimación por MCO.****5.3 El coeficiente de determinación y el coeficiente de correlación.****5.4 Regresión no lineal.****Introducción**

En la práctica estadística, los fenómenos reales suelen depender de múltiples factores. Sin embargo, en muchos casos nos centramos en estudiar la relación entre dos variables: una dependiente (Y) y una independiente o explicativa (X). Esta relación no suele ser exacta, sino aproximada, debido a la naturaleza aleatoria de los datos o a la presencia de otras variables no observadas. El objetivo del análisis de regresión es encontrar una expresión matemática que describa lo mejor posible esa relación.

Los modelos estadísticos no alcanzan a recoger todas las posibles variables explicativas de un cierto fenómeno porque contienen un comportamiento aleatorio que se deriva de la naturaleza humana del tipo de situaciones que analizan, además de incurrir en errores en la toma de datos, ya sean censos o encuestas.

En estos casos, las variables mantienen una dependencia no exacta, no matemática que requiere un tratamiento estadístico, que da nombre al tipo de vinculación a estudiar y que se concreta en dos procedimientos:

- a) Buscar y determinar aquella estructura de dependencia (aproximada) que mejor exprese el tipo de relación existente entre la variable explicada (Y) y la variable explicativa (X), que se conoce como REGRESIÓN.
- b) Calcular el grado de dependencia existente entre las variables (X e Y) mediante el análisis de CORRELACIÓN.

En este tema vamos a abordar ambas cuestiones, ciñéndonos al caso más sencillo de la regresión con una única variable explicativa.

5.1. Planteamiento general del problema de la regresión

Formalmente, la regresión busca una función $f(X)$ tal que los valores observados de Y puedan aproximarse por $\hat{Y} = f(X)$. En el caso más simple, se estudia la regresión lineal, donde se supone

que existe una relación lineal entre ambas variables: $\hat{Y} = a + bX$, siendo 'a' el término independiente y 'b' el coeficiente de regresión o pendiente.

A partir de la variable bidimensional (X,Y) con distribución de frecuencias conjunta $[(x_{ij}, y_{ij}); n_{ij}] (i=1...h; j=1...k)$ llamamos regresión de Y (*variable dependiente o explicada*) sobre X (*variable independiente o explicativa*), que denotamos como $R(Y/X)$, a la función que explica el comportamiento de Y para cada valor de X .

Siempre habrá una lógica que imponga que una de las variables sea la explicada y la otra la explicativa. En el caso del consumo y la renta de las familias, lo lógico es pensar que la renta determine el consumo y no al revés.

Dicha función se puede establecer siguiendo dos criterios, lo que permite que se hable de regresión de tipo I, por un lado, y regresión de tipo II o Mínimo Cuadrática, por otro. Veamos cada uno de los procedimientos:

5.2 Regresión lineal. Estimación por MCO.

Estudiamos dos tipos de regresión, llamadas regresión tipo 1 y tipo 2 y suelen referirse a diferentes enfoques dependiendo de cómo tratamos las variables. La regresión lineal estándar, que es la que solemos usar en estadística aplicada, asume que una variable es la independiente (X) y la otra es la dependiente (Y), y minimizamos los errores en la variable Y . Eso se llama regresión tipo 1.

Por otro lado, existe la regresión tipo 2 (a veces llamada regresión de mínimos cuadrados ortogonales), que se utiliza cuando ambas variables tienen errores de medición y no queremos asumir que solo una de ellas tiene variabilidad. En ese caso, minimizamos la distancia perpendicular a la recta. Esto se usa cuando ambas variables son simétricas en importancia o tienen errores comparables.

En cuanto a la utilidad: la regresión tipo 1 es la más común y es útil cuando tienes una variable claramente dependiente. La tipo 2 la utilizas cuando no puedes distinguir qué variable depende de cuál o ambas son medidas con cierto error.

REGRESIÓN DE TIPO I DE Y SOBRE X O TAMBIÉN $R_I(Y/X)$:

La distribución bidimensional de frecuencias se observa desde la perspectiva de la distribución condicionada de Y para cada valor de X , con el objetivo de asignar a la variable explicada (Y) el mejor valor posible. Esto nos lleva a leer la tabla de correlación a partir de las filas que la conforman, en cada una de las cuales los valores de la X están fijos y varían los de la Y . Su representación gráfica quedaría así,

Distribución bidimensional leída como
suma de distribuciones condicionadas:

Si tuviéramos que escoger qué valor de Y tomar como representante de los disponibles para cada uno de los valores de la X, un criterio lógico sería usar la media de los existentes.

$\bar{y}/(X=x_i)$ o simplificando $\bar{y}/x_i \quad \forall i = 1..h$

Representación gráfica de las medias condicionadas

El valor de Y seleccionado debería ser aquel que minimizara la suma de los errores cometidos en la estimación $e_{(j/i)}$, que vendrían dados, para cada x_i , por la diferencia entre los valores observados y_j y los estimados $\hat{y}/x_i$:

Debiendo, además, considerar el cuadrado de dichos errores para evitar la compensación de los errores positivos con los negativos.

Errores cometidos en la estimación

Con lo que la regresión $R_I (Y/X)$ estará formada por los puntos:

$(x_1, \bar{y}/x_1)$

$(x_2, \bar{y}/x_2)$

$(x_3, \bar{y}/x_3)$

$\vdots$

$(x_h, \bar{y}/x_h)$

Para valores de X distintos a los existentes en la distribución bidimensional de frecuencias, la regresión de tipo I no proporciona ninguna estimación para la Y. Es decir, la regresión de tipo I sólo permite estimar el comportamiento de Y para los valores de X presentes en la distribución de frecuencias. Esto representa una limitación de este método.

En consecuencia, cuando se trata de distribuciones bidimensionales de frecuencias es preferible emplear la regresión de tipo II de Y sobre X que veremos a continuación.

REGRESIÓN DE TIPO II DE Y SOBRE X O TAMBIÉN $R_{II} (Y/X)$ o mínimo cuadrática ordinaria (MCO):

Partiendo de la distribución conjunta de frecuencias, en lugar de buscar la mejor estimación de Y en cada distribución condicionada, por separado de las demás, lo que se hace es considerar la nube de puntos en su conjunto, seleccionando la función matemática que mejor se ajuste a ella. Es decir, se trata de obtener la ecuación de la función que pase lo más cerca posible de todos los puntos disponibles. Por lo tanto, podemos considerar dos etapas:

En primer lugar, representaremos la nube de puntos y elegiremos el tipo de función que mejor se adapte a los datos, basándonos sólo en la gráfica.

En segundo lugar, una vez escogido el tipo de función más idóneo (recta, parábola, hipérbola, etc.), habrá que determinar cuál de las infinitas que hay en el plano se aproxima más a todos los puntos. Esto requiere estimar los parámetros de la función correspondiente, mediante algún criterio, siendo el de los mínimos cuadrados ordinarios (MCO) el más habitual.

Ejemplos de funciones de ajuste

Dada la variable estadística bidimensional (X,Y) definida a partir de N pares de observaciones (x_i, y_i) repetidos o no, la regresión lineal supone que la función matemática que mejor se ajusta a la nube de puntos sea una recta:

$$\hat{y}_i = a + b \cdot x_i \quad \forall i = 1 \dots N$$

Donde:

$\hat{y}_i$ = valor teórico de la variable y , para cada x_i

a = término independiente u ordenada en el origen. Valor que toma la variable dependiente cuando la variable independiente toma el valor cero.

b = es el coeficiente de regresión y constituye la pendiente de la recta.

Una vez establecida la forma de la función, el objetivo será obtener los valores de los estadísticos a y b

Representación de la nube de puntos

Para cada valor de x_i , se tendrán dos valores, el correspondiente a la nube de puntos (real) y_i y el teórico $\hat{y}_i$ (que se obtendrá por el ajuste lineal por MCO).

Para cada x_i se tendrá una diferencia entre los dos valores de y , (el real y el teórico), que llamaremos error o residuo e_i tal que:

$$e_j = y_j - \hat{y}_j$$

El método MCO consiste en determinar los valores de ***a*** y ***b***, de forma que los residuos sean mínimos.

Regresión Lineal

Cálculo de los parámetros *a* y *b*:

El objetivo será minimizar la expresión:

Condición necesaria:

Al estar igualadas a cero se puede dividir a ambos lados por (-2) para obtener:

Desarrollamos y dividimos por *N* para obtener:

Se tendrá un sistema de dos ecuaciones con dos incógnitas (***a y b***), de dónde podremos despejar:

Podremos calcular los valores ajustados y los errores:

Obsérvese, que el signo del coeficiente de regresión (***b***, ***pendiente de la recta***) coincide con el de la covarianza (***m₁₁*** o ***S_{xy}***).

Una vez estimada la recta de regresión, estamos en condiciones de poder predecir cualquier valor ***y₀*** de la variable ***Y***, a partir de otro valor ***x₀*** de ***X***:

Una covarianza positiva determinará una pendiente positiva, en tanto que una covarianza negativa producirá una pendiente negativa.

Covarianza y coeficiente de regresión no nulos

En el caso de que la covarianza fuera cero también se anularía el coeficiente de regresión y la recta de regresión sería una paralela al eje de abscisas, indicando que el valor que toma la variable ***Y*** en la regresión lineal no dependería de ***X***, o lo que es lo mismo, que ***Y*** no dependería linealmente de ***X***:

Covarianza (***S_{xy}*** o ***m₁₁***) y coeficiente de regresión (***b***) nulos

En este caso ***X*** e ***Y*** serán independientes.

Una vez establecida la forma con la que se establece la dependencia entre las variables X e Y a través de la regresión lineal, cabe preguntarse por el grado de intensidad de dicha dependencia, lo que se conoce como correlación.

Comentarios sobre la Regresión tipo I y tipo II

Ahora, sobre el tema de los errores y cómo tratarlos: en un buen ajuste de regresión esperamos que los residuos (es decir, las diferencias entre los valores observados y los valores predichos) sean aleatorios. Eso significa que deben cumplir condiciones como la homocedasticidad, es decir, que la varianza de los errores sea constante a lo largo de todas las observaciones. También deben ser independientes y seguir aproximadamente una distribución normal.

Si estas condiciones se cumplen, podemos confiar más en el ajuste. En cambio, si vemos patrones en los residuos (por ejemplo, que se ensanchan o se reducen de manera sistemática), eso nos dice que el modelo no está capturando bien la variabilidad. Así que siempre hay que revisar esos supuestos.

La regresión tipo 1 y tipo 2 se usan en contextos distintos según las variables y sus errores. Para asegurarnos de que el ajuste es bueno, miramos la distribución de los residuos y comprobamos esas condiciones de homocedasticidad e independencia.

La diferencia principal es la siguiente: la regresión tipo 1 es la que normalmente usamos, donde consideramos que una variable es la independiente (X) y la otra es la dependiente (Y). En este caso, lo que hacemos es minimizar los errores solo en la dirección de Y , es decir, tratamos de ajustar una recta que prediga lo mejor posible el valor de Y a partir de X .

En cambio, la regresión tipo 2 se usa cuando las dos variables son tratadas de manera simétrica, es decir, no consideramos que una dependa de la otra de forma clara, sino que ambas pueden tener errores de medición. En este caso, lo que minimizamos es la distancia perpendicular a la recta, no solo en una dirección. Se utiliza cuando no podemos asignar claramente cuál es la variable independiente y cuál la dependiente, o cuando ambas variables tienen incertidumbre similar.

Tipo 1, minimizamos errores en una variable (la dependiente); tipo 2, minimizamos errores en ambas direcciones por igual.

La ecuación parece la misma. En ambos casos terminas con una recta que se escribe como “ Y igual a a más b por X ”. La diferencia está en cómo se calculan esos coeficientes a y b . Con la regresión tipo 1, lo que hacemos es minimizar las diferencias verticales, es decir, los errores en la variable Y . Eso nos da una línea que es la mejor posible en términos de ajustar Y en función de X .

En cambio, con la regresión tipo 2, lo que minimizamos es la distancia perpendicular a la recta. Esto tiene en cuenta los errores en ambas variables, tanto en X como en Y . Así que aunque la forma de la ecuación final sea igual, los coeficientes resultantes serán ligeramente distintos porque se ajustan de una manera diferente.

La diferencia es la forma de calcular esa recta. La tipo 1 ajusta minimizando los errores en Y, y la tipo 2 ajusta minimizando los errores en ambas variables de forma equilibrada. Así que la diferencia está en el método, aunque el resultado final sea una ecuación parecida.

La forma más conocida, donde el coeficiente b se calcula como la covarianza de X y Y dividida por la varianza de X (o equivalentemente, la razón entre la covarianza y la desviación típica de X) es la regresión tipo 1, que es la que ya conocemos: calculamos la covarianza entre las dos variables y la dividimos por la varianza de la variable independiente para obtener la pendiente b. Luego calculas la ordenada al origen a con la media de los valores.

La regresión tipo 2, en cambio, usa un método distinto, porque minimiza esas distancias perpendiculares, y la fórmula para b ahí es un poco más compleja y no se reduce a esa razón tan simple. Pero la de la tipo 1, es la más habitual.

En la regresión tipo 2, también conocida como regresión de mínimos cuadrados ortogonales, la fórmula del parámetro b es un poquito distinta porque toma en cuenta la variabilidad de ambas variables por igual.

En este caso, el coeficiente b se obtiene como el cociente entre la covarianza de X y Y y la media geométrica de las varianzas de X y Y. Es decir, b es igual a la covarianza de X y Y dividida por la raíz cuadrada del producto de la varianza de X por la varianza de Y.

De esa manera, no estás favoreciendo ni a la variable X ni a la Y, sino que estás teniendo en cuenta que ambas pueden tener errores de medición y estás minimizando las distancias ortogonales a la recta.

Esa es la diferencia, en la tipo 1 usas la varianza de X en el denominador, y en la tipo 2 usas esta especie de media de las varianzas de ambas variables para hacer el ajuste más simétrico.

En la mayoría de los casos prácticos, la regresión tipo 1 es la más recomendada y la más usada, sobre todo si tienes claro que una variable es la independiente y la otra es la dependiente. Es sencilla, robusta y es la que la mayoría de la gente conoce y entiende sin problemas.

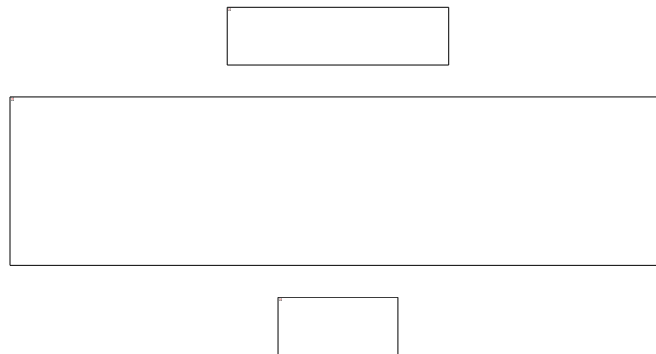
Ahora bien, si estás en una situación en la que ambas variables tienen errores de medición similares o no puedes decir claramente cuál depende de cuál, la regresión tipo 2 te puede venir mejor porque trata a las dos variables de manera más equilibrada. Pero en general, si no hay un motivo especial para usar la tipo 2, la tipo 1 suele ser la opción estándar y la más recomendable.

5.3. El coeficiente de determinación y el coeficiente de correlación

Se buscará una medida que recoja dicha información, a partir del siguiente razonamiento: si todos los puntos de la nube estuvieran sobre la recta de regresión, el grado de dependencia sería el máximo posible. Por el contrario, cuanto más se alejen los puntos de la función, es decir, cuanto mayores sean los residuos, menor será la intensidad de la relación.

Esto indica que podemos apoyarnos en los residuos, es decir, para medir el grado de dependencia. Veamos cómo:

El coeficiente de determinación (R^2)



Se puede demostrar que la varianza total de la variable Y se puede descomponer en la suma de la varianza residual y de la varianza debida a la regresión:

$$\sigma^2_Y = \sigma^2_{\text{regresión}} + \sigma^2_{\text{residual}}$$

$$\sigma^2_Y = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2 = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Indicando que cuanto mejor sea el ajuste, mayor será la parte de la varianza total explicada por la regresión y menor debida a los residuos

$$R^2 = \frac{\sigma^2_{\text{regresión}}}{\sigma^2_Y} = \frac{\frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Con lo que obtenemos el coeficiente de determinación (R^2): expresado en tanto por cien, representa el porcentaje de la variabilidad de la variable dependiente que viene explicada por el modelo de regresión:

En el supuesto de regresión lineal simple:

Informa sobre la proporción que de la varianza total supone la varianza debido a la regresión. Se cumple que:

Un R^2 cercano a cero indicará un ajuste pobre de los datos a la regresión, mientras que un R^2 próximo a 1 informará de un ajuste muy bueno.

El coeficiente de correlación lineal entre las variables X e Y (r):

Es una medida adimensional, cuyo signo viene dado por el de la covarianza, que al elevarlo al cuadrado se iguala al coeficiente de determinación y cuyo valor está acotado entre -1 y +1, es decir $-1 \leq r \leq 1$.

En el caso de que $r=0$, indica que no hay dependencia lineal, esto no implica que no pueda existir dependencia no lineal entre las variables.

Diferencia entre el coeficiente de determinación (R^2) y el coeficiente de correlación (r):

Mientras que (R^2) mide el grado de la bondad del ajuste de la regresión lineal o no lineal a la nube de puntos correspondiente, es decir, la eficacia del modelo planteado para explicar la variable Y a partir de otra u otras variables, (r) valora el grado de dependencia lineal entre dos variables.

Coeficiente de correlación lineal (r)		Coeficiente de determinación (R^2)
$r = -1$	Correlación lineal perfecta negativa	$R^2 = 1$
$-1 < r < 0$	Correlación lineal negativa	$0 < R^2 < 1$
$r = 0$	No hay dependencia lineal	$R^2 = 0$
$0 < r < 1$	Correlación lineal positiva	$0 < R^2 < 1$
$r = 1$	Correlación lineal perfecta positiva	$R^2 = 1$

La correlación mide el grado de asociación entre dos variables. Mientras que la regresión se centra en describir la forma de la relación (explicando cómo cambia Y cuando cambia X), la correlación cuantifica la intensidad y dirección de esa relación.

El análisis de correlación se basa en el estudio de la covarianza entre X e Y. Si la covarianza es positiva, significa que ambas variables tienden a aumentar o disminuir conjuntamente; si es negativa, una aumenta cuando la otra disminuye. La correlación estandariza esta medida para hacerla independiente de las unidades de medida, utilizando el coeficiente de correlación lineal de Pearson (r).

Coeficiente de regresión. Interpretación

En la regresión lineal simple, se estima una recta de la forma $\hat{Y} = a + bX$ mediante el método de los mínimos cuadrados ordinarios (MCO). Este método busca los valores de 'a' y 'b' que minimizan la suma de los cuadrados de los errores o residuos (diferencias entre los valores observados y los predichos).

Coeficiente de regresión (b): indica el cambio promedio en Y asociado a una variación de una unidad en X. Si $b > 0$, la relación entre X e Y es directa (a mayor X, mayor Y). Si $b < 0$, la relación es inversa. El coeficiente de regresión se calcula como:

$$b = \text{Cov}(X, Y) / \text{Var}(X)$$

Término independiente (a): representa el valor esperado de Y cuando X es cero, y se calcula como:

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

La interpretación de la recta de regresión permite realizar estimaciones y analizar la tendencia general de los datos.

Interpretación de los coeficientes de determinación y de correlación lineal.

Una vez estimada la regresión, es necesario evaluar la calidad del ajuste. Para ello se utilizan el coeficiente de determinación (R^2) y el coeficiente de correlación lineal (r).

Coeficiente de determinación (R^2): mide la proporción de la variabilidad total de Y que es explicada por la variable X mediante el modelo lineal. Toma valores entre 0 y 1. Cuanto más próximo a 1, mejor es el ajuste del modelo. Se calcula como:

$$R^2 = \text{Suma de cuadrados explicada} / \text{Suma total de cuadrados} = 1 - (\text{Suma de cuadrados de los residuos} / \text{Suma total de cuadrados})$$

Coeficiente de correlación lineal (r): mide la intensidad y dirección de la relación lineal entre las variables. Su valor está comprendido entre -1 y 1. Si $r = 1$, existe una relación lineal positiva perfecta; si $r = -1$, una relación lineal negativa perfecta; y si $r = 0$, no existe relación lineal.

La relación entre ambos coeficientes es directa: en la regresión lineal simple, $R^2 = r^2$.

Errores frecuentes y cómo evitarlos

- ☐ Confundir correlación con causalidad: una alta correlación no implica que una variable cause cambios en la otra.
- ☐ Utilizar la regresión lineal sin comprobar la linealidad de los datos. Antes de aplicar el modelo, conviene observar la nube de puntos.
- ☐ Ignorar los supuestos de homocedasticidad y normalidad de los residuos. Su incumplimiento puede invalidar las inferencias.
- ☐ Extrapolar fuera del rango observado: las predicciones solo son válidas dentro del intervalo de valores de X analizado.
- ☐ Interpretar R^2 sin contexto: un valor bajo no implica necesariamente un mal modelo si los datos son muy dispersos por naturaleza.

Comentarios sobre las condiciones del modelo de regresión

En lo que se refiere a comprobar las condiciones del modelo, como la homocedasticidad, la linealidad, la normalidad de los residuos, eso se hace de la misma manera tanto para la regresión tipo 1 como para la tipo 2.

Al final, lo que estamos haciendo es evaluar si el modelo ajusta bien a los datos y si los residuos se comportan como deberían: que tengan varianza constante, que no haya patrones raros y que sigan más o menos una distribución normal. Esto se comprueba igual en ambos tipos de regresión.

Así que en ese sentido, las herramientas de diagnóstico son las mismas.

Comprobar la linealidad de los residuos es bastante sencillo y se suele hacer con gráficos. La idea es que si el modelo es lineal, los residuos no deberían mostrar ningún patrón sistemático cuando los representas frente a los valores ajustados o frente a las variables independientes.

En otras palabras, lo que haces es dibujar un gráfico de dispersión de los residuos contra los valores predichos. Si ves que los residuos están dispersos de manera aleatoria, sin formar ninguna curva ni ningún patrón, entonces la suposición de linealidad es razonable. Si en cambio

ves que los residuos dibujan una forma curva o algún tipo de estructura, eso te indica que el modelo no es realmente lineal y quizá necesites transformaciones o un modelo más complejo. Simplemente trazas ese gráfico de residuos y compruebas que no haya un patrón visible.

En el eje horizontal puedes poner los valores ajustados, es decir, los valores predichos por tu modelo, y en el eje vertical representas los residuos, que son esas diferencias entre el valor observado y el valor predicho.

Si al hacer este gráfico ves que los residuos están distribuidos de forma aleatoria alrededor de cero, sin formar ningún patrón, entonces puedes asumir que la linealidad es adecuada. En caso contrario, no.

Comprobar la normalidad de los residuos es también un paso bastante común y se puede hacer de una forma muy visual. Normalmente se utiliza un gráfico Q-Q (cuantil-cuantil), que es una herramienta que te permite comparar la distribución de los residuos con la de una distribución normal teórica.

Lo que haces es representar los cuantiles de los residuos frente a los cuantiles de una distribución normal. Si los puntos caen aproximadamente sobre una línea recta, eso indica que los residuos se distribuyen de forma cercana a la normalidad. Si ves desviaciones grandes, especialmente en las colas o una forma de S muy marcada, eso sugiere que los residuos no son del todo normales. Esa línea recta de referencia, cuando haces un gráfico Q-Q, es simplemente la diagonal que representa la distribución normal teórica. Es decir, en el gráfico Q-Q pones en el eje X los cuantiles teóricos de una distribución normal y en el eje Y los cuantiles de los residuos. La línea recta que trazas es la que pasa por la diagonal (normalmente es una línea de 45 grados), que indica cómo se distribuirían los residuos si fueran perfectamente normales.

También puedes usar pruebas estadísticas como la prueba de Shapiro-Wilk o la de Kolmogórov-Smirnov para tener un criterio más numérico.

5.4 Regresión no lineal

5.4.1 Función exponencial

$$\hat{y}_i = A \cdot B^{x_i} \quad \forall i = 1 \dots N$$

Tomando logaritmos:

$$\ln(\hat{y}_i) = \underbrace{\ln(A)}_a + x_i \cdot \underbrace{\ln(B)}_b \quad \forall i = 1 \dots N$$

De donde se obtiene:

$$\ln(\hat{y}_i) = a + b \cdot x_i \quad \forall i = 1 \dots N$$

$$a = \overline{\ln(y)} - b \cdot \bar{x} \quad \text{y} \quad b = \frac{S_{x \ln y}}{S_x^2}$$

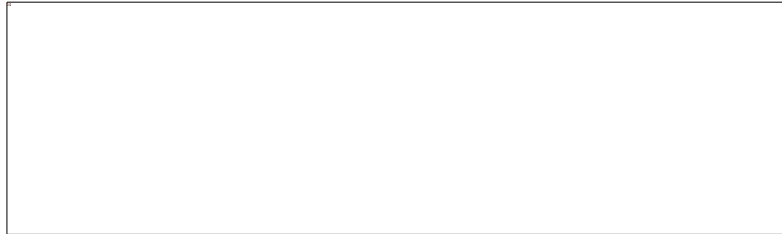
Una vez calculados a y b se tomarán antilogaritmos:

$$A = e^a \quad B = e^b$$

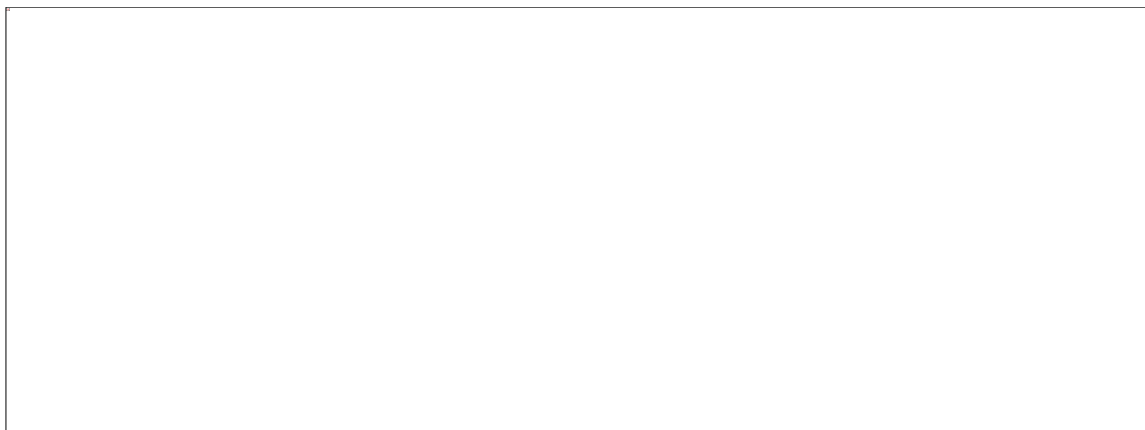
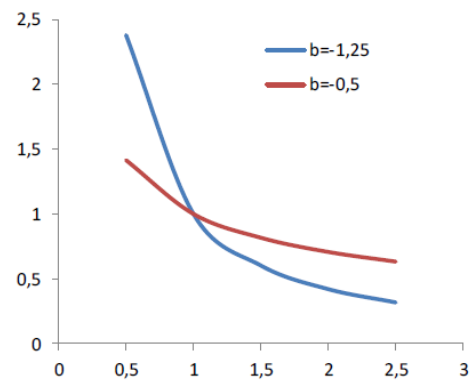
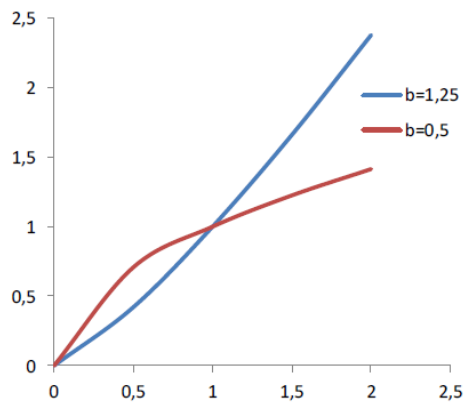
$$\hat{y}_i = e^{a+b \cdot x_i} \quad \forall i = 1 \dots N$$

b : mide, aproximadamente la tasa de variación $\hat{y}_i$ de ante un cambio unitario en x_i

5.4.2 Función potencial



Será creciente cuando b sea positivo y decreciente para valores de b negativos, convexa si $|b| > 1$ y cóncava si $0 < |b| < 1$





5.4.3 Función hiperbólica equilátera o recíproca:



Bibliografía

Agresti, A. (2019). Statistical Methods for the Social Sciences. Pearson.

José Javier Muruzábal Irigoyen, Dionisio Pérez Esteban (2024): Estadística y Probabilidad en la Ingeniería. 2ª Edición, Ibergarceta Publicaciones S.L. 9788419034403.

Grado en Ingeniería en Organización Industrial**Tema 6. Números Índice****Bloque temático IV: Análisis comparativo de variables estadísticas****6.1 Concepto y tipología.****6.2 Números índices complejos.****6.3 Propiedades.****6.4 Cambio de base. Enlace de números índices.****6.5 Índices de precios e índices de cantidades. Índices en Economía.****6.6 Deflactación****Introducción**

Los **números índice** son herramientas estadísticas que permiten medir la evolución relativa de una magnitud a lo largo del tiempo o entre distintas situaciones. Se utilizan ampliamente en el análisis económico, industrial y logístico para comparar precios, cantidades, costes o niveles de producción.

A modo de idea inicial para comprender el concepto de número índice podemos decir que un número índice expresa el cambio relativo de una variable entre un período actual (t) y un período base (0), normalmente en base 100:

- ☐ Si $I_t = 120$, la magnitud ha aumentado un 20 % respecto al año base.
- ☐ Si $I_t = 85$, ha disminuido un 15 %.

6.1 Concepto y tipología

Los números índices son instrumentos que facilitan la comparación de datos de distintas magnitudes a lo largo del tiempo o entre diferentes situaciones. Se usan en campos como la economía y la gestión empresarial.

Los números índice son una forma de resumir y comparar variaciones relativas en variables económicas, de producción, de costes, o de eficiencia en el tiempo. En el campo de la ingeniería industrial, nos sirven para medir cómo cambian los precios de materiales, cómo evoluciona la productividad de ciertos procesos, o cómo se ajustan los costes de producción a lo largo de diferentes periodos.

Los números índice como herramienta en el ámbito de la ingeniería en organización industrial nos permiten analizar tendencias, ajustar datos a la inflación o a cambios de eficiencia, y tomar

decisiones basadas en datos concretos. Veremos ejemplos específicos, como por ejemplo índices de precios de componentes industriales, índices de productividad en una línea de ensamblaje, o comparativas de costes a lo largo del tiempo.

a) Índices simples

Miden la evolución de **una sola variable**.

- ☐ **Ventajas:** fáciles de calcular e interpretar.
- ☐ **Limitaciones:** no reflejan la evolución conjunta de varios elementos.

Cuando hablamos de números índice simples, nos referimos a los tipos más básicos y directos. Por ejemplo, están los índices de precios simples, que comparan el precio de un único producto en un momento determinado con el precio de ese mismo producto en un periodo base. También tenemos los índices de cantidad simples, que hacen lo mismo, pero con la cantidad producida o consumida de un bien.

Un número índice simple se centra en una sola variable y nos dice cómo ha cambiado esa variable de un periodo a otro. Se pueden construir índices más complejos o ponderados, pero lo básico es empezar por esos índices simples.

Ejemplo resuelto 1: Índice simple de precios

Supongamos que el precio de un producto era de 50€ en 2020 y ha subido a 75€ en 2024. Calculemos el índice simple de precios tomando 2020 como período base:

$$I = \left(\frac{X_t}{X_0} \right) \times 100 = \left(\frac{75}{50} \right) \times 100 = 150$$

Esto indica que el precio del producto ha aumentado un 50% respecto al período base.

Índices simples en la Organización industrial:

- ☐ **Precio del cobre:** impacto en componentes eléctricos.
- ☐ **Coste de electricidad:** seguimiento en procesos intensivos.
- ☐ **Coste de transporte:** presión de fletes en distribución.

b) Índices complejos

Agrupan **varios bienes o servicios** en un único indicador, usando **ponderaciones**.

Principales fórmulas:

- ☐ **Laspeyres** (pondera con cantidades del año base)
- ☐ **Paasche** (pondera con cantidades del año actual)
- ☐ **Fisher** (media geométrica de Laspeyres y Paasche)

Índices complejos en la organización industrial:

- ☐ **Índice de costes de materias primas** (Laspeyres): acero, cobre, polímeros.
- ☐ **Índice de coste logístico total** (Fisher): transporte, almacenaje, última milla.
- ☐ **Índice de producción agregada** (Laspeyres de cantidades): familias de productos.

Reglas prácticas para construir índices

- ☐ Elegir un **año base** adecuado (por ejemplo, hace tres años).
- ☐ **Homogeneizar unidades**: €/t, €/km, €/MWh, etc.
- ☐ Definir **ponderaciones coherentes**: consumos físicos, horas-máquina, etc.
- ☐ Documentar el **alcance** del índice: qué costes incluye (fletes, seguros, almacenamiento, etc.).

6.2 Números índices complejos.

También llamados números índices compuestos o ponderados, se utilizan cuando queremos resumir el cambio de un conjunto de bienes o variables en lugar de uno solo. Es decir, ya no estamos mirando un único producto, sino un conjunto de ellos, y le damos a cada uno un peso o una ponderación según su importancia.

Entre los tipos más conocidos están el índice de Laspeyres y el índice de Paasche. El índice de Laspeyres usa las **cantidades del periodo base** para ponderar los precios, mientras que el índice de Paasche usa las **cantidades del periodo actual** para ponderar. Ambos son útiles, pero tienen enfoques un poco distintos. Y luego también hay otros como el índice de Fisher, que es una media geométrica de ambos y se considera un índice ideal porque combina las ventajas de los dos anteriores.

Ejemplo resuelto 2: Índice compuesto de precios con ponderaciones

Se analizan los precios de dos productos (A y B) en dos períodos. En el período base (2020), los precios y cantidades eran los siguientes:

- Producto A: Precio = 10€, Cantidad = 100 unidades.
- Producto B: Precio = 20€, Cantidad = 50 unidades.

En 2024, los precios han cambiado a:

- Producto A: Precio = 12€.
- Producto B: Precio = 25€.

Cálculo del índice de Laspeyres:

Cálculo del índice de Paasche:



Asumiendo cantidades actuales: 90 unidades para A y 60 para B:

Ambos índices muestran un aumento de precios de aproximadamente 22-23%, pero el índice de Paasche refleja un mayor peso en los cambios recientes.

6.3 Propiedades.

1. Existencia o determinación: no debe tomar un valor indeterminado, ni cero, ni infinito, además sólo puede ser positivo.
2. Identidad: si la situación actual coincide con la situación base, el índice ha de tomar el valor 1:

$$I_0^0 = I_t^t = 1$$

3. Inversión: si se intercambian las situaciones base y actual el número índice resultante es igual al inverso del inicial:

$$I_t^0 = \frac{1}{I_0^t} \Leftrightarrow I_0^t \cdot I_t^0 = 1$$

4. Circular (o encadenamiento): si se consideran tres situaciones, “0”, “h” y “t”, se debe cumplir:

$$I_0^t = I_0^h \cdot I_h^t \Leftrightarrow I_0^h \cdot I_h^t \cdot I_t^0 = 1$$

Esta propiedad permite realizar enlaces técnicos y cambios de base.

5. Homogeneidad: Un número índice complejo no debe alterarse a causa de un cambio en las unidades de medida de alguna de las modalidades que lo definen.

6.4 Cambio de base. Enlace de números índices.

Encadenar series se realiza cuando tienes datos con distintas bases de referencia y quieres crear una serie homogénea a lo largo del tiempo. Básicamente, lo que haces es ajustar las series que tienen diferentes bases para que todas se refieran a una base común.

Por ejemplo, si tienes una serie de índices con base en 2020 y otra con base en 2022, y quieres que todo quede en la base 2022, lo que haces es tomar el valor del índice en el año en que coinciden, dividir la serie antigua por ese valor común y así la “reescalas” a la nueva base.

En otras palabras, al encadenar o enlazar series ajustas las diferentes partes de la serie para que todas reflejen la misma base temporal. De ese modo, obtienes una serie homogénea y continua, lo que te permite comparar datos de manera coherente a lo largo del tiempo.

Una vez que tienes claro el proceso de encadenar, lo que haces básicamente es ir enlazando esos tramos de series de manera que, al final, tienes una serie que parece como si siempre hubiera tenido la misma base, aunque en realidad hayas tenido que hacer esos ajustes intermedios. El resultado de encadenar series es una base de datos unificada en en una sola para que tus datos sean consistentes a lo largo del tiempo.

6.5 Índices de precios e índices de cantidades. Índices en Economía.

Ejemplo resuelto: Cálculo del IPC

Supongamos una economía con una cesta de consumo que incluye dos productos:

- Producto X: Precio 2020 = 10€, Precio 2024 = 12€, Cantidad = 10 unidades.
- Producto Y: Precio 2020 = 15€, Precio 2024 = 18€, Cantidad = 5 unidades.

El IPC se calcula como un índice ponderado de precios y cantidades de ambos productos:

El IPC es 120, lo que indica un incremento del 20% en el nivel general de precios respecto al período base (en nuestro ejemplo 2020).

Para más información sobre el IPC – Consultad la página web oficial del INE

<https://www.ine.es/ss/Satellite?L=0&c=Page&cid=1254735893337&p=1254735893337&pageName=ProductosYServicios%2FPYSLayout>

<https://www.ine.es/dyngs/AYU/index.htm?cid=118>

Otros índices:

1. Índice de costes de producción industrial (ICPI), elaborado por el INE, refleja las variaciones de los precios de los bienes producidos por la industria. Muy usado en presupuestación y revisión de precios en contratos.
2. Índice de costes logísticos, (Construido internamente por empresas), integra transporte, almacenamiento, mantenimiento, combustible y personal. Permite evaluar la eficiencia global del sistema logístico.

Categoría	Ejemplo de índice	Finalidad
Precios y costes	ICPI, Laspeyres, Paasche	Evaluar inflación industrial
Productividad	OEE, PTF, productividad laboral	Medir eficiencia operativa
Producción y capacidad	IPI, utilización de capacidad	Planificación de recursos
Calidad	Cp, Cpk, DPMO	Control de procesos
Financieros	Coste unitario, ROS, VAI	Control de rentabilidad
Medioambientales	Intensidad energética, CO ₂	Eficiencia y sostenibilidad
Logísticos	Puntualidad, carga media	Eficiencia del transporte

6.6 Deflatación

Deflactar se aplica cuando trabajas con datos económicos o industriales a lo largo del tiempo. Básicamente, deflactar significa ajustar los valores monetarios para eliminar el efecto de la inflación y poder comparar esos valores en términos reales.

Por ejemplo, si tienes una serie de costes o precios a lo largo de varios años, esos valores nominales pueden haber subido simplemente porque ha habido inflación. Deflactar consiste en usar un índice de precios, como por ejemplo el IPC, para convertir esos valores nominales en valores reales. De esa forma, estás viendo los datos en precios constantes y puedes comparar de un año a otro sin que la inflación distorsione la comparación.

Podríamos decir que deflactar una serie es “quitarle” el efecto de la inflación para que puedas ver los valores en términos reales y hacer comparaciones más precisas a lo largo del tiempo. La deflatación permite ajustar valores nominales para eliminar el efecto de la inflación y trabajar con valores reales.

Deflactar una serie es sencillo en términos numéricos. Lo que haces es tomar el valor nominal de la serie en un año determinado y dividirlo por el índice de precios de ese mismo año. Luego multiplicas ese resultado por el índice de precios del año base que hayas elegido.

Por ejemplo, si tienes un valor nominal en 2024 y quieres llevarlo a precios constantes del año 2020, dividirías el valor de 2024 por el índice de precios de 2024 y luego multiplicarías por el índice de precios de 2020. De esa forma, obtienes el valor ajustado por la inflación.

Así que matemáticamente: valor nominal dividido por el índice de precios del año correspondiente, y luego multiplicado por el índice del año base. Así eliminas el efecto de la inflación.

Ejemplo resuelto: Ajuste de salarios

El salario de una persona fue 30.000€ en 2024. (decimos que 30.000€ es el salario nominal)

Sabiendo que el índice de precios al consumo (IPC) de 2024 ha sido 120, podemos calcular el salario real de esta persona referenciado a 2020.

Para ello tomaremos 2020 como año base para el cálculo (IPC = 100).

Calculemos el salario real:

$$\text{Salario_real} = (\text{Salario_nominal} / \text{IPC}) \times 100 = (30,000 / 120) \times 100 = 25.000\text{€}.$$

El salario real ajustado por la inflación en 2024 es 25.000€, lo que refleja una pérdida de poder adquisitivo.

Grado en Ingeniería en Organización Industrial**Tema 7. Nociones de probabilidad****Bloque temático V: Introducción a la Probabilidad y a la Inferencia Estadística****7.1 Noción y determinación de la probabilidad****7.2 Probabilidad condicionada e independencia de sucesos****Introducción**

En este tema se presentan las nociones fundamentales de la probabilidad, una herramienta básica para analizar fenómenos aleatorios. En muchas situaciones reales —como prever el resultado de una encuesta, estimar el riesgo de una inversión o calcular la fiabilidad de un proceso industrial— no se puede anticipar con certeza el resultado de un experimento. Por ello, la probabilidad permite cuantificar la incertidumbre asociada a los distintos resultados posibles y formular conclusiones basadas en el razonamiento estadístico.

7.1 Noción y determinación de la probabilidad

La **probabilidad** es una medida numérica que expresa el **grado de posibilidad de que ocurra un suceso aleatorio**. Permite cuantificar la incertidumbre asociada a los fenómenos en los que el resultado no se puede predecir con certeza, aunque sí se conozcan los posibles resultados.

En otras palabras, la probabilidad asigna un número entre **0 y 1** a cada suceso:

- ☐ **0** representa la imposibilidad de que ocurra.
- ☐ **1** representa la certeza absoluta de que ocurrirá.
- ☐ Los valores intermedios indican distintos niveles de posibilidad o riesgo.

Por ejemplo, al lanzar una moneda equilibrada, la probabilidad de obtener “cara” es 0,5, ya que ambos resultados son igualmente posibles.

Formas de determinar la probabilidad

Existen diferentes enfoques para determinar la probabilidad de un suceso, según el tipo de información disponible y el contexto del problema:

1. Enfoque clásico o a priori

Se aplica cuando todos los resultados posibles son igualmente probables.

La probabilidad se define como el cociente entre el número de casos favorables y el número total de casos posibles:

donde:

- ☐ $n(A)$: número de resultados favorables al suceso (A).
- ☐ $n(\Omega)$: número total de resultados posibles en el experimento.

Ejemplo: al lanzar un dado, la probabilidad de obtener un número par es

2. Enfoque frecuentista o empírico

Se basa en la **observación repetida** del fenómeno.

La probabilidad de un suceso se aproxima mediante la frecuencia relativa con que ocurre, cuando el experimento se repite un gran número de veces:

$$P(A) \approx \frac{N_A}{n}$$

donde:

- ☐ N_A : número de veces que ocurre el suceso A.
- ☐ n : número total de repeticiones del experimento.

A medida que el número de repeticiones aumenta, la frecuencia relativa se estabiliza en torno al valor teórico de la probabilidad.

3. Enfoque subjetivo o bayesiano

Se utiliza cuando la probabilidad refleja un **grado de creencia personal o razonada** basado en información previa, experiencia o juicio experto.

Es común en contextos en los que no se puede realizar un experimento repetido (por ejemplo, estimar la probabilidad de que un proyecto empresarial tenga éxito).

Este enfoque está vinculado al **Teorema de Bayes**, que permite actualizar las probabilidades conforme se dispone de nueva información.

Propiedades básicas de la probabilidad

Para cualquier suceso A dentro del espacio muestral Ω :

1. $0 \leq P(A) \leq 1$
2. $p(\Omega)=1$
3. $P(\emptyset)=0$
4. Si A y B son incompatibles, $p(A \cup B)=P(A)+P(B)$

Estas propiedades aseguran que la probabilidad cumpla las reglas lógicas del razonamiento matemático.

7.2 Probabilidad condicionada e independencia de sucesos

Para comprender la probabilidad condicionada y la independencia de sucesos, es necesario recordar primero los **elementos básicos de un fenómeno aleatorio**, que constituyen la base sobre la que se definen y relacionan ambos conceptos.

Elementos de un fenómeno aleatorio

Un fenómeno aleatorio es aquel cuyo resultado no puede predecirse con exactitud, aunque sí se conozcan todos los posibles resultados.

El conjunto de todos los resultados posibles se denomina *espacio muestral* y se representa por la letra griega Ω (omega). Cada subconjunto de Ω se llama suceso y puede representar un evento concreto dentro del fenómeno aleatorio (por ejemplo, obtener cara al lanzar una moneda).

- ☐ El suceso imposible es aquel que nunca puede ocurrir: su probabilidad es cero.

Suceso imposible: no ocurrirá nunca $S_i=\emptyset$; $P(\emptyset)=0$

- ☐ Suceso incompatible (disjuntos): dos sucesos (S_i) y (S_j) son incompatibles o mutuamente excluyentes si nunca pueden ocurrir a la vez ($S_i \cap S_j=\emptyset$)
Los sucesos incompatibles o mutuamente excluyentes son aquellos que no pueden suceder al mismo tiempo.

- ☐ La unión de dos sucesos recoge todos los casos en que ocurre uno, otro o ambos.

Unión: $S_i \cup S_j = \forall \omega \in \Omega \text{ tal que } \omega \in S_i \text{ o } \omega \in S_j$; ocurrirá siempre que ocurran S_i o S_j .

- ☐ La intersección representa los casos en los que ocurren simultáneamente ambos sucesos.
- ☐ Intersección: $S_i \cap S_j = \{\omega \in \Omega \mid \omega \in S_i \text{ y } \omega \in S_j\}$: ocurrirá siempre que ocurran simultáneamente S_i y S_j .
- ☐ El suceso complementario agrupa los casos en los que *no* ocurre un suceso dado.

Complementario: $\bar{S}_i = \{\omega \in \Omega \mid \omega \notin S_i\}$: ocurrirá cuando no ocurra S_i .

En probabilidad, las operaciones entre sucesos siguen reglas formales, cuando se tienen elementos comunes se cumple:

$$P(S_1 \cup S_2) = P(S_1) + P(S_2) - P(S_1 \cap S_2)$$

Ahora bien, si S_1 y S_2 son disjuntos $P(S_1 \cap S_2) = 0$ y

$$P(S_1 \cup S_2) = P(S_1) + P(S_2)$$

La probabilidad del suceso complementario de S_i es:

$$P(\bar{S}_i) = 1 - P(S_i) \text{ (muy importante)}$$

Probabilidad condicionada

En muchas situaciones, el cálculo de la probabilidad de un suceso depende de que ya se conozca información previa sobre la ocurrencia de otro suceso. En estos casos, se utiliza la probabilidad condicionada, que permite actualizar la probabilidad de un evento teniendo en cuenta que otro ha sucedido.

La probabilidad condicionada de A dado que ha ocurrido B, se interpreta como *la probabilidad de que ocurra A cuando sabemos que B ya ha tenido lugar*. En esta relación:

- ☐ **A** es el *suceso condicionado* (el que queremos calcular).
- ☐ **B** es el *suceso condicionante* (el que sabemos que ha ocurrido).

La probabilidad condicionada $P(A|B)$ representa la probabilidad de que ocurra A dado que ha ocurrido B. “A” será el suceso condicionado y “B” el suceso condicionante.

Se calcula como:

, siempre que $P(B) > 0$

Esta expresión significa que la probabilidad de A dado B es igual a la proporción de los casos en los que ocurren ambos sucesos (A y B) respecto al total de los casos en los que ocurre B. Es decir,

se restringe el espacio muestral original únicamente a las situaciones donde B sucede, y dentro de ese subconjunto se calcula la frecuencia relativa de A.

Donde:

$P(A \cap B)$ = probabilidad de la intersección de sucesos.

$P(A/B)$ es la fracción del espacio muestral de "B" en el que ocurre A, mide la probabilidad de que suceda A dentro del contexto de que B ya ha ocurrido.

Regla del producto:

A partir de la definición anterior, se obtiene una relación muy útil llamada regla del producto:

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

Esta regla permite expresar la probabilidad de que ocurran ambos sucesos conjuntamente de dos formas equivalentes, dependiendo de cuál se considere el condicionante.

Si los sucesos A y B son independientes, conocer que ha ocurrido B no cambia la probabilidad de A, y por tanto:

$$P(A|B) = P(A)$$

En cambio, si están relacionados, el hecho de que ocurra uno modifica la probabilidad del otro. La probabilidad condicionada es la base de muchos desarrollos posteriores, como el teorema de Bayes, los árboles de probabilidad y los modelos de decisión estadística.

Teorema de Bayes

El **Teorema de Bayes** es una de las herramientas más importantes de la probabilidad, ya que permite **actualizar la probabilidad de un suceso** cuando se dispone de nueva información. En otras palabras, sirve para **invertir el condicionamiento**: conocer la probabilidad de B dado A y utilizarla para calcular la probabilidad de A dado B.

Matemáticamente, el teorema se expresa como:

donde:

- ☐ $P(A|B)$: es la **probabilidad posterior**, es decir, la probabilidad de que ocurra A sabiendo que B ha ocurrido.

- $P(B|A)$: es la **probabilidad condicionada directa** o **verosimilitud**, la probabilidad de que ocurra B si A ha sucedido.
- $P(A)$: es la **probabilidad a priori** de A, antes de conocer que B ha ocurrido.
- $P(B)$: es la **probabilidad total** de que ocurra B, calculada considerando todas las posibles causas o sucesos que pueden originarlo.

Interpretación conceptual

El Teorema de Bayes permite **revisar una creencia inicial (probabilidad a priori)** cuando se observa un nuevo hecho o evidencia (el suceso B). El resultado, $P(A|B)$, es una **probabilidad actualizada** que combina la información previa con la nueva evidencia.

Por ejemplo, en contextos médicos, si A representa “tener una enfermedad” y B representa “dar positivo en una prueba diagnóstica”, el teorema permite calcular la **probabilidad de tener la enfermedad sabiendo que la prueba ha dado positiva**, usando los datos conocidos sobre la fiabilidad de la prueba y la frecuencia de la enfermedad en la población.

Cuando existen varios sucesos que pueden explicar B, el denominador $P(B)$ se obtiene mediante la **fórmula de la probabilidad total**:

$$P(B)=P(B | A_1) \cdot P(A_1)+P(B | A_2) \cdot P(A_2)+\cdots +P(B | A_n) \cdot P(A_n)$$

De este modo, el Teorema de Bayes permite calcular probabilidades posteriores en situaciones con varias hipótesis posibles.

El Teorema de Bayes constituye la base del razonamiento probabilístico moderno, de los métodos bayesianos y de muchos modelos utilizados en estadística, inteligencia artificial, medicina y economía, entre otros campos. Su valor radica en que ofrece un marco lógico para aprender de la experiencia y actualizar la información conforme se dispone de nuevos datos.

Sucesos independientes: pueden ocurrir a la vez y no se afectan mutuamente.

Dos eventos A y B son independientes si:

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

Diremos que dos sucesos (A y B) son independientes cuando la probabilidad de la intersección es igual al producto de las probabilidades individuales. La ocurrencia de A no modifica la probabilidad de B y viceversa.

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

$$P(B)=P(B|A)yP(A)=P(A|B)$$

Sucesos dependientes:

Es cuando el hecho de que ocurra B, modifica la probabilidad de A:

$$P(B) \neq P(B|A), P(A \cap B) \neq P(A) \cdot P(B)$$

Observación: un caso particular se produce cuando los sucesos A y B son disjuntos o mutuamente excluyentes, entonces, ambos son dependientes en probabilidad.

Si A y B son disjuntos $\Rightarrow P(A \cap B) = 0$, por lo tanto: $P(A|B) = 0$, siempre que:

$$P(A) > 0 \text{ y } P(B) > 0$$

No se debe confundir independencia de sucesos con sucesos disjuntos.

¡IMPORTANTE!

- ☐ Disjuntos $\neq$ Independientes
- ☐ En ingeniería: disjuntos = categorías excluyentes
- ☐ Independientes = procesos sin influencia mutua

Sucesos disjuntos: no pueden ocurrir a la vez.

Dos sucesos A y B son disjuntos si: $A \cap B = \emptyset$

Propiedad clave: $P(A \cup B) = P(A) + P(B)$

¿Cómo se calcula la probabilidad de sucesos que no son disjuntos ni independientes?

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

Teorema de la probabilidad total**Teorema de la probabilidad total y de Bayes**

Lista de verificación

Pregunta	Acción
¿Pueden ocurrir a la vez?	NO → Disjuntos
¿Un suceso afecta al otro?	NO → Independientes
¿Son categorías excluyentes?	Usar suma
¿Son procesos separados?	Usar producto
¿Hay influencia (turno, máquina, proveedor)?	Usar condicionada o Bayes

Frases clave para informes técnicos

- “Los fallos son mutuamente excluyentes → aplicamos suma de probabilidades.”
- “Las líneas operan independientemente → probabilidad conjunta = producto.”
- “El turno condiciona la tasa de defectos → usar probabilidad condicionada.”
- “Los eventos no son independientes → no se puede factorizar.”

Ejemplos prácticos en Organización Industrial

Ejemplo 1: Tipo de fallo en una máquina (DISJUNTOS)

Contexto: Una máquina puede fallar por 3 causas:

- ☐ A: Fallo eléctrico
- ☐ B: Fallo mecánico
- ☐ C: Fallo de software

Datos históricos (500 fallos):

Causa	Nº de fallos	Probabilidad
Eléctrico (E)	200	0,4
Mecánico (M)	250	0,5
Software (S)	50	0,1

¿Son disjuntos?

SÍ → Una máquina no puede fallar por dos causas a la vez ($E \cap M = \emptyset$),

$$P(E \cup M) = P(E) + P(M)$$

¿Son independientes?

NO → No tiene sentido: si ya falló por eléctrico, no puede fallar por mecánico. Pero si lo calculamos: $P(E \cap M) = 0$; $P(E) = 0,4$; $P(M) = 0,5 \Rightarrow 0 \neq 0,2$ (no se cumple la igualdad)

Conclusión: Disjuntos pero NO independientes → Usar probabilidad total para predecir fallos.

Conceptos clave (repaso rápido)

Concepto	Definición	Fórmula clave
$E \cap M = \emptyset$ $P(E \cup M) = P(E) + P(M)$ Disjuntos	No pueden ocurrir a la vez	
$P(E \cup M) \neq P(E) \cdot P(M)$ Independientes	Uno no afecta la probabilidad del otro	

Ejemplo 2: Inspección de calidad en dos líneas (INDEPENDIENTES)

Contexto:

- Línea 1: produce tornillos
- Línea 2: produce tuercas
- A: “tornillo en buen estado”, A^c : tornillo defectuoso → $P(A^c) = 0,02$
- B: “tuerca en buen estado”, B^c : tuerca defectuosa → $P(B^c) = 0,03$

¿Son independientes?

SÍ → Fallos en líneas separadas, sin interacción física ni de control.



		Tuercas		
		B		
Tornillos	A	0,9506	0,0294	0,98
		0,0194	0,0006	0,02
		0,97	0,03	1,0

Verificación:

$$P(A \cap B) = P(A) \cdot P(B) = 0,02 \cdot 0,03 = 0,0006$$

Ejemplo 3: Turno y tipo de producto (NI DISJUNTOS NI INDEPENDIENTES)

Contexto: Fábrica con 3 turnos.

- M: Turno mañana
- T: Turno tarde
- N: Turno noche
- P: Producto tipo “Premium”

Datos (1000 unidades):

Turno	Premium (P)	Estándar (E)	Total
Mañana (M)	120	280	400
Tarde (T)	80	220	300
Noche (N)	40	260	300
Total	240	760	1.000
Turno	Premium (P)	Estándar (E)	Total
Mañana (M)	0,120	0,280	0,4
Tarde (T)	0,80	0,220	0,3
Noche (N)	0,40	0,260	0,300
Total	0,240	0,760	1

¿Son disjuntos M y P

NO → Puede producirse Premium en mañana.

¿Son independientes M y P?

$$P(M) = 0,4$$

$$P(P) = 0,24$$

$$P(M \cap P) = 0,12 \neq 0,4 \cdot 0,24 = 0,096 \text{ NO son independientes}$$

Conclusión:

El turno afecta la probabilidad de producir Premium → dependientes.

Ejemplo 4: Estado de dos sensores en una cadena (INDEPENDIENTES)

Contexto: Cadena de montaje con 2 sensores de posición:

- S1: Sensor 1 activo
- S2: Sensor 2 activo
- Cada sensor falla con $P(F) = 0,01$, independientemente.

¿Son independientes S_1 y S_2?

Sí → Diseñados con circuitos separados.

		S1		
		F		
S2	F	0,0001	0,0099	0,01
		0,0099	0,9801	0,99
		0,01	0,99	1,0

Aplicación en control:

Probabilidad de que ambos fallen (parada de línea):

$$P(F1 \cap F2) = 0,0001$$

Ejemplo 5: Modo de transporte y retraso (DEPENDIENTES)

Contexto: Envíos desde almacén:

- T: Transporte por camión
- A: Transporte por avión
- R: Retraso > 2 días

Modo	Total envíos	Retrasos	% Retraso
Camión (T)	800	160	20%
Avión (A)	200	10	5%
	1000	170	

¿Son disjuntos T y R?

NO → Puede haber retraso en camión.

¿Son independientes T y R?



$$P(T) = 800/1000 = 0,8$$

$$P(R) = 170/1000 = 0,17$$

$$P(T / R) = 160/170=0,94$$

$$P(R/T)= 160/800 = 0,2$$

En sucesos independientes:

Conclusión: El modo de transporte influye en el retraso → usar probabilidad condicionada.

Grado en Ingeniería en Organización Industrial**Tema 8. Variables aleatorias****Bloque temático V: Introducción a la Probabilidad y a la Inferencia Estadística****8.1 Distribución de probabilidad y función de distribución****8.2 V.A. discreta y V.A. continua****8.3 Esperanza, varianza y tipificación****Introducción**

En el estudio de la probabilidad y la inferencia estadística, las variables aleatorias (V.A.) constituyen un concepto fundamental que permite modelizar fenómenos inciertos de manera cuantitativa. Este tema aborda las herramientas necesarias para describir y analizar el comportamiento de sistemas donde el resultado no es determinista, sino que depende del azar.

Comenzaremos explorando las distribuciones de probabilidad y las funciones de distribución, que permiten caracterizar cómo se distribuyen los posibles valores de una variable aleatoria. A continuación, se distinguirán las variables aleatorias discretas, aquellas que toman valores enteros, y de las variables aleatorias continuas, que son aquellas que pueden asumir cualquier valor dentro de un intervalo.

Finalmente, se introducen conceptos clave como la esperanza matemática (media esperada), la varianza, que mide la dispersión de los valores respecto a la media, y la tipificación, que permite estandarizar variables para facilitar su análisis comparativo.

Este tema proporciona las bases teóricas necesarias para comprender fenómenos industriales y organizativos desde una perspectiva probabilística, facilitando la toma de decisiones bajo incertidumbre.

8.1 Distribución de probabilidad y función de distribución

Una **distribución de probabilidad** describe cómo se asignan probabilidades a los posibles valores de una variable aleatoria. En el caso de variables discretas, se utiliza una **función de probabilidad** que asigna a cada valor una probabilidad específica. Para variables continuas, se emplea una **función de densidad**, que permite calcular probabilidades sobre intervalos mediante integración.

La **función de distribución acumulada (FDA)**, por otro lado, indica la probabilidad de que la variable aleatoria tome un valor menor o igual a un cierto número. Esta función es creciente y está acotada entre 0 y 1, siendo una herramienta clave para el análisis probabilístico.

8.2 Variable Aleatoria Discreta y Variable Aleatoria Continua

- **Variable Aleatoria Discreta:** Toma valores finitos o contables. Ejemplos típicos incluyen el número de defectos en una línea de producción o el número de clientes que llegan a una tienda en una hora. Se representa mediante una función de probabilidad $P(X=x_i)$.
- **Variable Aleatoria Continua:** Puede tomar cualquier valor dentro de un intervalo. Ejemplos incluyen el tiempo de espera en una cola o la temperatura de un horno industrial. Se describe mediante una función de densidad $f(x)$, y las probabilidades se calculan como áreas bajo la curva.

Ambos tipos de variables permiten modelar fenómenos industriales, logísticos y organizativos, facilitando la toma de decisiones basada en datos.

8.3 Esperanza, varianza y tipificación

Esperanza matemática (media esperada): Es el valor promedio ponderado por las probabilidades. Representa el valor esperado de la variable aleatoria en múltiples repeticiones del experimento.

- Discreta: $E[X] = \sum x_i$
- Continua: $E[X] = \int x f(x) dx$

Varianza: Mide la dispersión de los valores respecto a la media. Una varianza alta indica mayor incertidumbre.

- $Var(X) = E[X - E[X]]^2$

Tipificación: Consiste en transformar una variable aleatoria en una nueva con media 0 y varianza 1. Esto se logra mediante la fórmula:

- $Z = \frac{X - E[X]}{\sqrt{Var(X)}}$

Esta técnica es útil para comparar variables con diferentes escalas y para aplicar distribuciones estándar como la normal.

Variables aleatorias discretas y continuas

Una variable aleatoria discreta X es una función que asigna valores numéricos a los resultados de un experimento aleatorio, donde los posibles valores son contables (finitos o infinitamente numerables).

Función de Masa de Probabilidad (PMF): $P(X=x_i)=P(x_i), \sum P(x_i)=1$

Variables aleatorias discretas:

La función de distribución $F(X)$ es una probabilidad acumulada.

$$F(X)=P(\xi \leq X)=P(-\infty < \xi \leq X)$$

Propiedades de la función de distribución:

$$1) F(\infty)=1$$

$$2) F(-\infty)=P(\xi \leq -\infty)=0$$

$$3) 0 \leq F(X) \leq 1$$

La función de distribución tiene que estar siempre entre 0 y 1 ya que es una probabilidad.

4) La función de distribución es una función no decreciente. Siempre será positiva.

$$F(x_2)=F(x_1)+P(x_1 < \xi \leq x_2)$$

Al tratarse de una probabilidad, $P(x_1 < \xi \leq x_2)$, siempre será positiva, por tal motivo $F(x_2)$ siempre será mayor que $F(x_1)$, es decir, nunca será decreciente.

5) La función de distribución es siempre continua por la derecha. La función de distribución puede ser discontinua por la izquierda.

MUY IMPORTANTE Consecuencias de lo anterior

Cuando hablemos de variables aleatorias continuas, su función de distribución será continua por la derecha y por la izquierda. Esto implica que los límites que dependían de un punto son todos cero, por lo que habrá de trabajar de otra manera. **Concretamente trabajaremos con intervalos.**

Ejemplo en Ingeniería Industrial:

Número de productos defectuosos en una muestra de 100 unidades inspeccionadas (control de calidad).

$$X \sim B(n=100, p=0,02)$$

$$P(X=2) = \binom{100}{2} \cdot (0,02)^2 \cdot (0,98)^{98}$$

Variable Aleatoria Continua: una variable aleatoria continua X toma valores en un intervalo continuo del conjunto de los reales. Se describe mediante una función de densidad de probabilidad (FDP) $= f(x)$

Propiedades:

$$\int_{-\infty}^{+\infty} f(x) dx = 1$$

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

Ejemplo en Ingeniería Industrial:

Tiempo de ciclo de una máquina (en minutos)

$$X \sim \text{Normal}(\mu = 12,5 \text{ min}; \sigma = 0,8 \text{ min})$$

La probabilidad de que el ciclo dure entre 12 y 13 minutos será:

$$P(12 \leq X \leq 13) = (P(z_i = 0,625) - P(z_i = -0,625))$$

Variables aleatorias continuas (VAC):

Nunca debemos plantear que una variable aleatoria continua tome valores puntuales, siempre van a ser intervalos.

Si la función de distribución es continua, la $P(\xi = x_i) = 0$. Si la probabilidad es cero, necesitamos otra forma de trabajar. La solución es trabajar con intervalos.

$$F(x) = P(\xi \leq x) = \int_{-\infty}^x f(x) dx$$

$$P(a < \xi \leq b) = F(b) - F(a) = \int_{-\infty}^b f(x) dx - \int_{-\infty}^a f(x) dx = \int_a^b f(x) dx \text{ Regla de Barrow}$$

$$P(a < \xi \leq b) = P(a < \xi < b) = P(a \leq \xi \leq b) = P(a \leq \xi < b)$$

Da igual si entran o no los extremos ya que:

$$P(\xi = a) = 0 \text{ y } P(\xi = b) = 0$$

$$F(\infty) = 1 = \int_{-\infty}^{+\infty} f(x) dx$$

Momentos, esperanzas, varianzas y tipificación

Momentos con respecto al origen:

$$\text{VAD: } \alpha_r = \sum_{x_i} x_i^r \cdot P(\xi=x_i)$$

$$\text{VAC: } \alpha_r = \int_{-\infty}^{\infty} x_i^r \cdot f(x) \cdot dx$$

Para r =0

$$\text{VAD: } \alpha_0 = \sum_{x_i} x_i^0 \cdot P(\xi=x_i)=1$$

$$\text{VAC: } \alpha_0 = \int_{-\infty}^{\infty} x_i^0 \cdot f(x) \cdot dx=1$$

Para r = 1 (Valor esperado, valor medio, Esperanza matemática)

VAD:

$$\alpha_1 = \sum_{x_i} x_i \cdot P(\xi=x_i)=E(\xi)$$

VAC:

$$\alpha_1 = \int_{-\infty}^{\infty} x_i \cdot f(x) \cdot dx=E(\xi)$$

Para r =2

$$\text{VAD: } \alpha_2 = \sum_{x_i} x_i^2 \cdot P(\xi=x_i)$$

$$\text{VAC: } \alpha_2 = \int_{-\infty}^{\infty} x_i^2 \cdot f(x) \cdot dx$$

Significado práctico de la esperanza representa el valor promedio a largo plazo si se repite el experimento muchas veces. Es el centro de masa de la distribución.

Ejemplo: Tiempo promedio de procesamiento por orden en un centro de distribución.

X: tiempo en horas, $X \sim \text{Exponencial}(\lambda=0,5)$

$$E(X) = \frac{1}{\lambda} = 2 \text{ horas}$$

Interpretación: En promedio, cada orden tarda 2 horas en procesarse.

Propiedades de la esperanza matemática

1) La esperanza de una constante (C), será la misma constante:

$$E(C)=C$$

2) La esperanza de una V.A. se ve afectada por cambios de origen:

$$E(C+\xi)=C+E(\xi)$$

3) La esperanza de una V.A. se ve afectada por cambios de escala:

$$E(C \cdot \xi)=C \cdot E(\xi)$$

4) Si ξ_1 y ξ_2 son dos variables aleatorias, entonces:

$$E(\xi_1 \pm \xi_2)=E(\xi_1) \pm E(\xi_2)$$

5) Si ξ_1 y ξ_2 son independientes: $E(\xi_1 \cdot \xi_2)=E(\xi_1) \cdot E(\xi_2)$

Fórmulas generadoras de momentos con respecto al valor esperado:

$$\text{VAD: } \mu_r = \sum_{x_i} (x_i - \alpha_1)^r \cdot P(\xi = x_i)$$

$$\text{VAC: } \mu_r = \int_{-\infty}^{\infty} (x_i - \alpha_1)^r \cdot f(x) \cdot dx$$

Para r = 0

$$\text{VAD: } \mu_0 = \sum_{x_i} (x_i - \alpha_1)^0 \cdot P(\xi = x_i) = 1$$

$$\text{VAC: } \mu_0 = \int_{-\infty}^{\infty} (x_i - \alpha_1)^0 \cdot f(x) \cdot dx = 1$$

Para r = 2

$$\text{VAD: } \mu_2 = \sum_{x_i} (x_i - \alpha_1)^2 \cdot P(\xi = x_i) = \text{Varianza}(\xi)$$

$$\text{VAC: } \mu_2 = \int_{-\infty}^{\infty} (x_i - \alpha_1)^2 \cdot f(x) \cdot dx = \text{Varianza}(\xi)$$

Relación entre la varianza y los momentos con respecto al origen:

$$\mu_2 = \alpha_2 - \alpha_1^2 \quad (\text{Válido para VAC y VAD})$$

La varianza mide la dispersión alrededor de la media.

Ejemplo:

Desviación en el tiempo de entrega de un proveedor

$$X \sim N(5, 1) \rightarrow \sigma = 1 \text{ día} \rightarrow \text{Var}(X) = 1^2 = 1 \text{ día}^2$$

Interpretación: Los tiempos de entrega oscilan en promedio ± 1 día respecto a los 5 días esperados.

Propiedades de la varianza:

1) La varianza es igual a la esperanza de los cuadrados menos el cuadrado de la esperanza:

$$Var(\xi) = E(\xi^2) - [E(\xi)]^2$$

2) La varianza siempre es cero o positiva, siendo nula si y solo si la V.A. ξ , es una constante:

$$Var(\xi) \geq 0$$

3) La varianza no se ve afectada por un cambio de origen:

$$Var(\xi + k) = Var(\xi)$$

4) La varianza de una V.A. se ve afectada por cambios de escala:

$$Var(k \cdot \xi) = k^2 \cdot Var(\xi)$$

Desviación típica:

$$\sigma = \sqrt{Var(\xi)}$$

Coeficiente de variación:

Se emplea para comparar la dispersión entre distintas variables aleatorias:

$$CV = \frac{\sqrt{Var(\xi)}}{|E(\xi)|}$$

Tipificación:

En ocasiones interesa que la distribución de probabilidad tenga esperanza 0 y varianza 1.

Hemos visto en estadística descriptiva que siempre podemos realizar un cambio simultáneo de escala y de origen, al que llamamos tipificación:

$$Z_i = \frac{x_i - \bar{x}}{S_x}$$

$$\text{En variables aleatorias: } \xi^* = \frac{\xi - \alpha_1}{\sqrt{Var(\xi)}} \begin{cases} E(\xi^*) = 0 \\ Var(\xi^*) = 1 \end{cases}$$

Desigualdad de Tchebychev

Sea ξ una variable aleatoria de media μ y desviación típica σ , **conocidas**. La desigualdad de Tchebychev permite acotar inferiormente la probabilidad de que la variable aleatoria tome valores dentro del intervalo $(\mu - k\sigma; \mu + k\sigma)$:

$$P(\mu - k\sigma < \xi < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

Ello para cualquier valor de k y cualquier variable aleatoria.

Análogamente, también se puede obtener una cota superior de probabilidad para el suceso complementario:

$$P(|\xi - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Interpretación:

Al menos $1 - \frac{1}{k^2}$ de los datos están a menos de k desviaciones estándar de la media, sin importar la forma de la distribución, por esta razón tiene sentido utilizar la herramienta cuando la distribución de probabilidad de la variable aleatoria sea desconocida.

Aplicaciones prácticas de la desigualdad de Tchebychev a ingeniería en organización industrial

1. Control de Calidad: Límite Superior de Productos Fuera de Especificación

Escenario

Una línea de ensamblaje produce piezas con diámetro nominal = 50,0 mm

Desviación estándar del proceso: = 0,6 mm

Especificación: $50 \pm 1,8$ mm $\rightarrow [48,2; 51,8]$ mm

Queremos garantizar que el porcentaje de piezas fuera de especificación sea bajo, sin suponer normalidad.

Aplicación de Tchebychev:

Distancia desde la media a los límites: 1,8 mm

$$k = \frac{1,8}{\sigma} = \frac{1,8}{0,6} = 3$$

$$P(|X - 50| \geq 1,8) \leq \frac{1}{3^2} = \frac{1}{9} \approx 0,11 (11,11\%)$$

Conclusión Práctica:

A lo sumo el 11,1% de las piezas estarán fuera de $50 \pm 1,8$ mm, incluso si la distribución es sesgada o multimodal.

Útil para certificar robustez del proceso ante auditorías (ISO 9001, IATF 16949).

2. Gestión de Inventarios: Stock de seguridad sin suponer normalidad

Escenario

Tiempo de reabastecimiento de un proveedor:

= 7 días, = 2 días

Queremos cubrir el 95% de los casos con stock de seguridad.

Aplicación:

Queremos calcular: $P(X > \mu + k\sigma) \leq 0,05$

Usamos Tchebychev para conseguir un límite unilateral conservador así:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \Rightarrow P(X > \mu + k\sigma) \leq \frac{1}{2k^2}$$

$$\text{Queremos } \frac{1}{2k^2} \leq 0,05 \rightarrow k^2 \geq 10 \rightarrow k \geq \sqrt{10} \approx 3,16$$

El Stock de seguridad = $k \cdot \sigma = 3,16 \cdot 2 \approx 6,3$ días.

Resultado: Cubrimos al menos el 95% de los retrasos con 6,3 días de inventario de seguridad, sin asumir distribución normal.

Ventajas de la desigualdad de Tchebychev en ingeniería en organización industrial

Ventaja	Impacto
No requiere normalidad	Ideal para datos reales (sesgados, multimodales)
Conservador	Evita subestimar riesgos
Fácil de calcular	Solo necesita y
Auditable	Base matemática sólida para certificaciones

La Desigualdad de Tchebychev es una herramienta robusta y práctica en Ingeniería Industrial cuando no se puede asumir normalidad, pero se necesita garantizar niveles mínimos de desempeño en procesos, inventarios, calidad y costos. Se recomienda utilizarla como cota superior de riesgo en reportes de KPIs, auditorías y planes de mejora continua. Algunas aplicaciones prácticas en Ingeniería en Organización Industrial son:

Concepto	Aplicación práctica
Variable discreta	Recuento de defectos, número de órdenes diarias, fallos por turno
Variable continua	Tiempos de ciclo, costes unitarios, distancias de transporte
Esperanza	Planificación de capacidad, estimación de costos promedio
Varianza	Control de procesos, gestión de inventarios.
Momentos	Análisis de riesgos, optimización robusta



El uso riguroso de variables aleatorias y sus momentos permite a los ingenieros industriales modelar la incertidumbre, predecir comportamientos promedio, controlar variabilidad y tomar decisiones basadas en datos en sistemas productivos y logísticos complejos.

Referencias:

Montgomery, D. C. (2019). Introduction to Statistical Quality Control (8th ed.). Hoboken, NJ: John Wiley & Sons. ISBN: 978-1-119-39730-4.

Ross, S. M. (2014). Introduction to Probability Models (11th ed.). San Diego, CA: Academic Press (Elsevier). ISBN: 978-0-12-407948-9.

ISO 9001:2015, Sistemas de gestión de la calidad – Requisitos. Organización Internacional de Normalización (ISO). <https://www.iso.org/iso-9001-quality-management.html>

IATF 16949:2016

Sistema de gestión de la calidad para la industria automotriz. Requisitos específicos para organizaciones que fabrican piezas de producción y servicio. Publicada por la International Automotive Task Force (IATF) en colaboración con ISO. Documento oficial integrado con ISO 9001

Grado en Ingeniería en Organización Industrial**Tema 9. Modelos de distribución de probabilidad****Bloque temático V: Introducción a la Probabilidad y a la Inferencia Estadística**

9.1 Modelos discretos**9.2 Modelos continuos****9.3 Distribución Normal****9.4 Distribución Chi-cuadrado, t de Student y F de Snedecor****9.5 Teorema Central del Límite****Introducción**

El análisis estadístico en ingeniería requiere comprender cómo se distribuyen los datos en torno a ciertos valores esperados. Para ello, los modelos de distribución de probabilidad ofrecen herramientas matemáticas que permiten describir el comportamiento de fenómenos aleatorios, tanto discretos como continuos.

Este tema profundiza en los principales modelos utilizados en la práctica industrial y organizativa. Se inicia con los modelos discretos, como la distribución binomial o de Poisson, útiles para contar eventos en procesos de producción o logística. Luego se abordan los modelos continuos, como la distribución exponencial o normal, fundamentales para analizar tiempos, medidas físicas o variaciones en calidad.

La distribución normal ocupa un lugar central por su aplicabilidad en numerosos contextos y por ser la base de muchos métodos inferenciales. A continuación, se estudian distribuciones derivadas como la Chi-cuadrado, la t de Student y la F de Snedecor, esenciales para contrastes de hipótesis y análisis de varianza.

Finalmente, se presenta el Teorema Central del Límite, que justifica el uso de la distribución normal en muestras grandes, independientemente de la distribución original de los datos. Este teorema es clave para la inferencia estadística y para la toma de decisiones basada en datos en entornos industriales.

1. Distribución de Bernoulli

La distribución de Bernoulli modela un experimento aleatorio con solo dos posibles resultados: éxito (1) o fracaso (0). Es la base para entender fenómenos binarios en procesos industriales.

$$\xi \sim B(1, p)$$

Formulación de la V.A.:

$$A \rightarrow 1P(\xi=1)=p$$

$$\bar{A} \rightarrow 0P(\xi=0)=q, p+q=1$$

Función de cuantía:

$$P(\xi=x)=p^x \cdot q^{1-x}$$

con $x=0; x=1$

$$\text{Esperanza: } E(\xi)=p$$

$$\text{Varianza: } Var(\xi)=p \cdot q$$

$$\text{Desviación típica: } \sigma(\xi)=\sqrt{p \cdot q}$$

Aplicación a la ingeniería de organización industrial:

Inspección de calidad en una línea de producción

Una fábrica produce piezas metálicas para automóviles. En el control de calidad, se realiza una inspección aleatoria de cada pieza al final de la línea de montaje. Cada pieza se clasifica como:

Defectuosa (fracaso, codificado como $x=0$)

No defectuosa (éxito, codificado como $x=1$)

La probabilidad de que una pieza sea no defectuosa es “p” (por ejemplo, $p = 0,98$ si el proceso tiene un 2% de defectos). Entonces, la variable aleatoria x sigue una distribución de Bernoulli con parámetro “p”:

$$X \sim B(1; 0,98)$$

Función de probabilidad:

$$P(X=1)=0,98, P(X=0)=0,02$$

2. Distribución Binomial

$$\xi \sim B(n, p)$$

Fenómeno aleatorio consistente en la repetición “n” veces de un fenómeno dicotómico.

$$\text{Función de cuantía: } P(\xi=x) = \binom{n}{x} \cdot p^x \cdot q^{(n-x)} = \frac{n!}{(n-x)! \cdot x!} \cdot p^x \cdot q^{(n-x)}$$

$$\text{Esperanza: } E(\xi) = n \cdot p$$

$$\text{Varianza: } Var(\xi) = n \cdot p \cdot q$$

$$\text{Desviación típica: } \sigma(\xi) = \sqrt{n \cdot p \cdot q}$$

Propiedad aditiva:

Sean: $\xi_1 \sim B(n_1, p)$ y $\xi_2 \sim B(n_2, p)$, se define: $\gamma \sim B(n_1 + n_2, p)$

Aplicación a la ingeniería de organización industrial:

Control de calidad por muestreo en la recepción de lotes (aceptación por muestreo).

Una empresa recibe un lote de 10.000 tornillos de un proveedor. Para decidir si acepta o rechaza el lote completo sin inspeccionarlo todo (lo cual sería costoso), aplica un plan de muestreo:

- Se extrae una muestra aleatoria de $n = 50$ tornillos.
- Se define como éxito que un tornillo esté dentro de especificaciones (no defectuoso), con probabilidad “p” (dada por la calidad del proveedor).
- Se acepta el lote si el número de tornillos defectuosos en la muestra es menor o igual a dos (2).

La variable aleatoria X : número de tornillos defectuosos en la muestra sigue una distribución binomial:

$$X \sim B(50, p)$$

Donde: p es la proporción real de defectuosos en el lote.

Función de probabilidad:

$$P(X=k) = \binom{50}{k} \cdot p^k \cdot q^{(50-k)}, k=0,1,2..$$

Si $p = 0,02$ y se decide que el criterio para aceptar el lote es que $X \leq 2$

Se puede hallar la probabilidad de aceptar el lote calculando la probabilidad mediante la función de probabilidad o empleando tablas

3. Distribución de Poisson

$$\xi \sim P(\lambda), \text{ con } \lambda = n \cdot p$$

Se determinó como el límite de una distribución $B(n, p)$, cuando $n \rightarrow \infty$ y $p \rightarrow 0$. En la práctica:

$$n > 30$$

$$p \leq 0,1$$

Aparición de un suceso concreto de manera continuada, siempre y cuando:

- ☐ El proceso sea estable.
- ☐ Los sucesos aparecen aleatoriamente, de forma independiente.
- ☐ Se pueden dar dos situaciones:
- ☐ Tiempo entre la ocurrencia de dos fenómenos consecutivos.
- ☐ Número de sucesos en un intervalo de longitud fija.

Función de cuantía:

$$P(\xi=x) = \frac{\lambda^x}{x!} \cdot e^{-\lambda}, x=0,1, \dots, n$$

$$\text{Esperanza: } E(\xi) = n \cdot p = \lambda$$

$$\text{Varianza: } Var(\xi) = n \cdot p = \lambda$$

$$\text{Desviación típica: } \sigma(\xi) = \sqrt{n \cdot p}$$

Propiedad aditiva:

La suma de V.V.A.A. independientes también se distribuye como una V.A. binomial:

Sean: $\xi_1 \sim P(\lambda_1)$ y $\xi_2 \sim P(\lambda_2)$, se define: $\gamma = \xi_1 + \xi_2$, entonces:

$$\gamma \sim P(\lambda_1 + \lambda_2)$$

Aplicación a ingeniería en organización industrial

Modelado de fallos en una máquina (Mantenimiento predictivo)

Una fábrica tiene una máquina crítica que procesa piezas de alta precisión. Históricamente, se observa que falla en promedio 1,8 veces por turno de 8 horas debido a causas aleatorias (desgaste, suciedad, vibraciones, etc.).

Se define la variable aleatoria:

$$X = \text{número de fallos de la máquina en un turno}$$

Si los fallos ocurren de forma independiente y a una tasa constante (λ), entonces:

$$X \sim P(\lambda), \text{ con } \lambda = 1,8$$

Función de probabilidad:

$$P(X=k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}, k=0,1, \dots, n$$

Aplicación práctica en Organización Industrial:

Planificación del mantenimiento preventivo:

Calcular la probabilidad de 0 fallos en un turno:

$$P(X=0) = \frac{\lambda^0}{0!} \cdot e^{-\lambda} = e^{-1,8} = 0,165 = 16,5\% \text{ de turno sin fallos.}$$

También se podría hallar la probabilidad de que ocurran en un turno 3 o más fallos, (lo cual indicaría un nivel crítico de funcionamiento de la máquina). Esto daría una probabilidad de 22,3% y se podría establecer que si se supera el 20% de fallos en la máquina se tendría que realizar un mantenimiento correctivo de la máquina.

4. Distribución Uniforme

$$\xi \sim U[a, b]$$

Función de densidad: $f(x) = \begin{cases} \frac{1}{b-a}; & \text{si } a \leq x \leq b \\ 0; & \text{resto} \end{cases}$

Esperanza: $E(\xi) = \mu = \alpha_1 = \frac{a+b}{2}$

Varianza: $\sigma^2 = \alpha_2 - \alpha_1^2 = \frac{(b-a)^2}{12}$

Desviación típica: $\sigma = + \sqrt{\frac{(b-a)^2}{12}}$

Aplicación a ingeniería en organización industrial

Control de tolerancias en el mecanizado de piezas (Fabricación de precisión)

Una empresa fabrica ejes cilíndricos en un torno. El diámetro nominal debe ser 50 mm, pero debido a variaciones en el material, herramienta y ajuste, el diámetro real D varía aleatoriamente dentro de un rango de tolerancia especificado:

$$49,95 \leq D \leq 50,05 \text{ mm}$$

Se supone que, por el control del proceso, el diámetro sigue una distribución uniforme continua en ese intervalo (sin sesgo hacia valores superiores o inferiores).

$$D \sim U[49,95; 50,05]$$

Función de densidad:

$$f(x) = \begin{cases} \frac{1}{50,05-49,95} = \frac{1}{0,1}; & \text{si } 49,95 \leq D \leq 50,05 \\ 0; & \text{resto} \end{cases}$$

Esperanza: $E(D) = \mu = \alpha_1 = \frac{100}{2} = 50$

Varianza: $\sigma^2 = \alpha_2 - \alpha_1^2 = \frac{(0,1)^2}{12} = 0,00083 \text{ mm}^2$

Desviación típica: $\sigma = 0,0289 \text{ mm}$

Aplicación práctica en Organización Industrial:

Si el cliente exige que los diámetros de los ejes se encuentren dentro del intervalo $49,98 \leq D \leq 50,02mm$, se puede calcular la probabilidad de que las piezas estén fuera de especificación.

Se puede calcular la tasa de rechazo de las piezas y en caso de que supere un límite preestablecido se procedería al mantenimiento de la máquina.

Conclusión:

La distribución uniforme continua es ideal para modelar variables con límites físicos fijos y sin tendencia (tolerancias, tiempos de ciclo en procesos automáticos, errores de medición calibrados).

5. Distribución Normal

$$\xi \sim N(\mu, \sigma)$$

Función de densidad:

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Propiedad aditiva:

Sean: $\xi_i \sim N(\mu_i, \sigma_i)$ independientes.

$$\omega = a + b_1 \cdot \xi_1 + b_2 \cdot \xi_2 + \dots + b_n \cdot \xi_n$$

Entonces:

$$\omega \sim N\left(a + \sum_{i=1}^n b_i \cdot \mu_i; \sqrt{\sum_{i=1}^n b_i^2 \cdot \sigma_i^2}\right)$$

Aplicación a ingeniería en organización industrial

Ejemplo: Control de calidad del tiempo de llenado de botellas en una línea de envasado

Una planta de refrescos en Madrid llena botellas de 500 ml. El sistema de llenado automático está diseñado para que el volumen medio sea exactamente 500 ml.

Datos históricos (tras muestrear 10 000 botellas):

Media $\mu = 500,3$ ml

Desviación típica $\sigma = 2,1$ ml

La distribución del volumen por botella sigue una distribución normal:

$$N \sim N(\mu=500,3\text{ml}, \sigma=2,1\text{ml})$$

Preguntas típicas que resuelve un ingeniero industrial usando la normal:

1. ¿Qué porcentaje de botellas están por debajo del límite legal de 495 ml?

Tipificamos:

$$Z = (495 - 500,3) / 2,1 = -2,52$$

$$P(Z < -2,52) = 0,0059 \text{ (empleando la } N(0,1)) \rightarrow 0,59\% \text{ de las botellas}$$

→ Solo 59 botellas de cada 10 000 están fuera de especificación inferior.

2. ¿Cuál es la probabilidad de que una botella tenga entre 498 y 502 ml? (rango aceptable interno)

$$Z_1 = (498 - 500,3)/2,1 = -1,10$$

$$Z_2 = (502 - 500,3)/2,1 = +0,81$$

$$\begin{aligned} P(-1,10 < Z < 0,81) &= P(Z < 0,81) - P(Z < -1,10) \\ &= 0,7910 - 0,1357 \\ &= 0,6553 \rightarrow 65,53\% \end{aligned}$$

Distribuciones derivadas de la normal

1. Distribución Chi cuadrado de Pearson

Es una distribución que no procede de la realidad.

Su forma depende de los grados de libertad (n).

Es asimétrica a la derecha y se encuentra tabulada.

Tiende a hacerse simétrica a medida que aumenta n .

Definición:

Sean: $\xi_i \sim N(0,1)$ independientes con $i=1,2, \dots, n$

Siendo n los grados de libertad

$$\chi_n^2 = \xi_1^2 + \xi_2^2 + \dots + \xi_n^2 = \sum_{i=1}^n \xi_i^2$$

Esperanza: $E(\chi_n^2) = n$

Varianza: $Var(\chi_n^2) = 2n$

Desviación típica: $\sigma = \sqrt{2n}$

Propiedad aditiva:

Sean: $\xi_1 \sim (\chi_n^2)$; $\xi_2 \sim (\chi_m^2)$; independientes, entonces se define:

$$\gamma = \xi_1 + \xi_2$$

$$\gamma \sim (\chi_r^2), \text{ donde } r = n + m$$

Relación de la Ji-cuadrado con la normal:

Si n es grande: $\xi \sim (\chi_n^2) \sim N(\sqrt{2n-1}, 1)$

Aplicación a ingeniería en organización industrial

Es el estadístico que permite crear intervalos de confianza para la varianza y para la razón de varianzas en poblaciones con comportamiento normal (se verá su aplicación en el tema 10).

2. Distribución t de Student

$$\xi \sim (t_n)$$

Esperanza (para $n > 2$): $E(\xi) = 0$

$$\text{Varianza: } \text{Var}(\xi) = \frac{n}{n-2}$$

Aplicación a ingeniería en organización industrial

Es el estadístico que permite crear intervalos de confianza para la media y para la diferencia de medias en poblaciones con comportamiento normal y varianza poblacional desconocida (se verá su aplicación en el tema 10).

3. Distribución F de Fisher – Snedecor

$$\xi \sim F(m, n)$$

No tiene correspondencia con la realidad.

No depende de la varianza de las variables que la integran.

No es simétrica.

Sólo toma valores en abscisas positivas.

Se emplea en inferencia a la hora de comparar varianzas de poblaciones normales.

Si una V.A. ($\xi = x$) sigue una $F(m, n)$, entonces ($\xi = 1/x$) sigue una $F(n, m)$

Definición: Sean: $\xi_i \sim N(0; \sigma)$, independientes, $con i=1, \dots, m$ y

$\gamma_j \sim N(0; \sigma)$, independientes, $con j=1, \dots, n$

Entonces:

$$F(m, n) = \frac{\left(\frac{\xi_1^2 + \xi_2^2 + \dots + \xi_m^2}{m} \right)}{\left(\frac{\gamma_1^2 + \gamma_2^2 + \dots + \gamma_n^2}{n} \right)} = \frac{\frac{\chi_m^2}{m}}{\frac{\chi_n^2}{n}}$$

Si $m=1$, entonces: $F(1, n) = t_n^2$

Aplicación a ingeniería en organización industrial

Es el estadístico que permite crear intervalos de confianza para la razón de varianzas de dos poblaciones con comportamiento normal (se verá su aplicación en el tema 10).

Teorema central del límite

Sea la V.A. ξ que se obtiene como resultado de la acción conjunta de otras VVAA:

$$\xi = \xi_1 + \xi_2 + \dots + \xi_n$$

Cumpléndose:

1. Las VVAA son independientes entre sí.
2. Con media (μ_i) y varianza (σ_i^2) finitas (no necesariamente iguales).
3. Con n suficientemente grande (en la práctica $n > 30$). Pudiendo ser VVAA distintas.

Entonces:

$$\xi \sim N\left(\sum_{i=1}^n \mu_i; \sqrt{\sum_{i=1}^n \sigma_i^2}\right)$$

Aplicación a ingeniería en organización industrial

Ejemplo: Control de calidad en una línea de producción de piezas mecánicas

En una fábrica de tornillos para automóviles, se mide el diámetro de cada tornillo para asegurar que cumpla con especificaciones (por ejemplo, un diámetro nominal de 10 mm con tolerancia $\pm 0,1$ mm). El proceso de medición introduce variabilidad aleatoria debido a factores como desgaste de herramientas, variaciones en materia prima o fluctuaciones en la temperatura.

Supongamos que el diámetro individual de cada tornillo X_i sigue una distribución no necesariamente normal (por ejemplo, uniforme entre 9,85 mm y 10,15 mm, o incluso sesgada por defectos ocasionales). Los ingenieros no conocen la distribución exacta de X_i , pero recolectan muestras diarias.

Aplicación del Teorema Central del Límite (TCL):

El TCL establece que, para una muestra aleatoria de tamaño “n” suficientemente grande (típicamente $n \geq 30$), la distribución del promedio muestral $\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^n X_i$ es aproximadamente normal, independientemente de la distribución original de X_i , con:

- Media: $E(\bar{X}) = \mu$ (media poblacional del diámetro, digamos $\mu = 10$ mm).

- Varianza: $Var(\bar{X}) = \frac{\sigma^2}{n}$ (donde σ^2 es la varianza poblacional).

Paso a paso para llegar a la aproximación:

1. Estimar parámetros a partir de datos: De una muestra grande, calcula $\bar{x}$ (aproximación de μ) y S_1^2 (aproximación de σ^2).

2. Aplicar TCL: Para ($n = 50$) tornillos muestreados al azar:

$$\bar{X} = N\left(\bar{x}, \frac{S_1^2}{50}\right)$$

3. Cálculo de probabilidades para control de procesos: empleando Excel o tablas.

Si un promedio muestral cae fuera de los límites establecidos por los protocolos de control de calidad, se activa una alarma para ajustar la línea de producción (evitando desperdicios o rechazos masivos).

Grado en Ingeniería en Organización Industrial**Tema 10. Nociones de Inferencia Estadística****Bloque temático V: Introducción a la Probabilidad y a la Inferencia Estadística**

10.1 La inferencia y sus métodos**10.2 Estimación puntual****10.3 Intervalos de Confianza****Introducción**

La inferencia estadística es el conjunto de métodos que permite extraer conclusiones sobre una población a partir del análisis de una muestra. En el contexto de la ingeniería, donde las decisiones deben basarse en datos limitados pero representativos, la inferencia proporciona herramientas rigurosas para estimar parámetros, evaluar hipótesis y controlar procesos.

Este tema introduce los fundamentos de la inferencia, comenzando por sus principales métodos: la estimación puntual, que busca obtener un valor único como mejor aproximación del parámetro poblacional, y los intervalos de confianza, que ofrecen un rango dentro del cual se espera que se encuentre dicho parámetro con una determinada probabilidad.

Estas técnicas son esenciales en la toma de decisiones bajo incertidumbre, como en el control de calidad, la mejora de procesos, la evaluación de proveedores o el diseño de experimentos. El dominio de estos conceptos permite a los ingenieros en organización industrial aplicar el pensamiento estadístico en entornos reales, optimizando recursos y minimizando riesgos.

10.1 La inferencia y sus métodos

La inferencia estadística es el proceso mediante el cual se obtienen conclusiones sobre una población a partir del análisis de una muestra representativa. Dado que en la práctica industrial rara vez se dispone de datos completos de toda la población (por razones de coste, tiempo o accesibilidad), la inferencia permite tomar decisiones fundamentadas con un grado controlado de incertidumbre.

Métodos principales de inferencia:

1. Estimación

Se refiere a la obtención de valores aproximados de parámetros poblacionales (como la media o la proporción) a partir de los datos muestrales. Existen dos tipos:

- o Estimación puntual: proporciona un único valor como mejor estimación del parámetro.
- o Estimación por intervalos: ofrece un rango de valores dentro del cual se espera que se encuentre el parámetro con cierta confianza.

2. Contraste de hipótesis

Consiste en verificar si una afirmación sobre un parámetro poblacional es compatible con los datos observados. Se formula una hipótesis nula (H_0) y una alternativa (H_1), y se evalúa si los datos proporcionan suficiente evidencia para rechazar H_0 .

3. Análisis de regresión y correlación

Permite estudiar relaciones entre variables, muy útil en ingeniería para modelar procesos, prever comportamientos y optimizar recursos.

4. Análisis de varianza (ANOVA)

Utilizado para comparar medias entre varios grupos y determinar si las diferencias observadas son estadísticamente significativas.

Población: conjunto completo de elementos que se desea estudiar.

Muestra: subconjunto representativo de la población, usado para hacer inferencias sobre ella.

Estimación: Proceso de usar datos muestrales para calcular valores aproximados de parámetros desconocidos de la población.

- Puntual: un solo valor estimado (ej. media muestral).
- Por intervalo: un rango de valores con un nivel de confianza (ej. intervalo de confianza del 95%). Lo que veremos en este tema.

Aplicación en Ingeniería en Organización Industrial

En entornos industriales, la inferencia estadística se aplica en:

- ☐ **Control de calidad:** para estimar la proporción de productos defectuosos.
- ☐ **Evaluación de proveedores:** para verificar si cumplen con los estándares establecidos.
- ☐ **Optimización de procesos:** mediante el análisis de datos de producción.

- ☐ **Diseño de experimentos:** para mejorar productos o procesos con base en evidencia estadística.

10.2 Estimación puntual

La **estimación puntual** es uno de los métodos fundamentales de la inferencia estadística. Consiste en utilizar los datos de una muestra para calcular un único valor que sirva como mejor aproximación del parámetro desconocido de una población (por ejemplo, su media, proporción o desviación típica).

¿Qué es un estimador?

Un **estimador** es una función matemática aplicada a los datos muestrales que produce una estimación puntual. Por ejemplo:

- ☐ Para estimar la **media poblacional** se utiliza la **media muestral**.
- ☐ Para estimar la **proporción poblacional**, se usa la **proporción muestral**.
- ☐ Para estimar la **varianza poblacional**, se emplea la **varianza muestral**.

Propiedades deseables de un buen estimador:

1. **Insesgado:** El valor esperado del estimador coincide con el parámetro poblacional.
2. **Eficiente:** Tiene la menor varianza posible entre todos los estimadores insesgados.
3. **Consistente:** A medida que el tamaño de la muestra aumenta, el estimador se aproxima al valor real del parámetro.
4. **Suficiente:** Utiliza toda la información relevante contenida en la muestra.

Aplicación en Ingeniería en Organización Industrial

La estimación puntual es clave en múltiples procesos industriales:

- ☐ **Control de calidad:** Estimar la proporción de productos defectuosos en una línea de producción.
- ☐ **Gestión de inventarios:** Estimar la demanda media de un producto para optimizar el stock.
- ☐ **Evaluación de proveedores:** Estimar el porcentaje de entregas puntuales o dentro de especificaciones.
- ☐ **Análisis de tiempos de proceso:** Estimar la duración media de una operación para mejorar la planificación.



Estimación puntual

Se emplearán los siguientes estadísticos para realizar la estimación puntual:

Parámetro poblacional	Estimador
$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^N X_i}{N}$ Media,	
$\hat{\sigma}^2 = S_1^2 = \frac{S^2}{N-1}$ Varianza,	
$\hat{p} = \frac{X}{N} = \frac{\text{número de éxitos}}{\text{número de pruebas}}$ Proporción, p	

Intervalos de confianza

Intervalo: conjunto de valores numéricos entre los que esperamos encontrar el verdadero valor del parámetro poblacional desconocido.

Precisión: se refiere a la longitud del intervalo. A menor longitud mayor precisión y viceversa.

Nivel de confianza γ :

$$\gamma = 1 - \alpha$$

= nivel de significación. El nivel de significación se repartirá en las colas del estadístico de tal forma que el nivel de confianza sea centrado.

Precisión y nivel de confianza:

Son elementos contrapuestos, si se aumenta el nivel de confianza se reduce la precisión (aumenta la longitud del intervalo), mientras que si se aumenta la precisión es a costa de disminuir el nivel de confianza.

Hay una excepción, aumentando el tamaño de la muestra (N), se puede:

Aumentar el nivel de confianza manteniendo la precisión, o

Aumentar la precisión manteniendo el nivel de confianza.

Pasos a seguir para obtener intervalos de confianza:

1. Identificar el caso que aplica. Ejemplos de posibles casos: calcular un intervalo de confianza para la media conocida la varianza de la población.

2. Seleccionar el estadístico que aplica según el caso (ver formulario). Para el caso anterior, si la muestra es grande, el estadístico será $N(0,1)$.
3. Seleccionar el nivel de confianza " γ " o el nivel de significación " α ".
4. Obtener los extremos inferior y superior del estadístico.
5. Construir el intervalo de confianza según el caso.

Caso 1: Intervalos de confianza para la media poblacional, conocida la varianza poblacional.

Estadístico: $N(0,1)$, de donde se obtendrá $\frac{z_{\alpha}}{2}$.

$$\mu \in [\bar{x} \pm \frac{z_{\alpha}}{2} \cdot \frac{\sigma}{\sqrt{N}}]_{\gamma}$$

La precisión de la estimación depende de $\frac{z_{\alpha}}{2}$, σ y N .

Se llama error de estimación o error muestral al radio del intervalo de confianza:

$$\varepsilon = \frac{z_{\alpha}}{2} \cdot \frac{\sigma}{\sqrt{N}}$$

De donde se puede despejar N :

$$N = \frac{\frac{z_{\alpha}^2}{2} \cdot \sigma^2}{\varepsilon^2}$$

Caso 2: Intervalos de confianza para la media poblacional, con varianza poblacional desconocida.

Se pueden dar dos casos dependiendo del tamaño muestral.

Para muestras pequeñas (**N 30**):

Estadístico "t de Student" de donde se obtendrá: $t_{N-1; \frac{\alpha}{2}}$

$$\mu \in [\bar{x} \pm t_{N-1; \frac{\alpha}{2}} \cdot \frac{S_1}{\sqrt{N}}]_{\gamma}$$

La precisión de la estimación depende de $t_{N-1; \frac{\alpha}{2}}$, S_1 y N .

Observación: para **N 30** se empleará $\frac{z_{\alpha}}{2}$

$$\varepsilon = t_{N-1; \frac{\alpha}{2}} \cdot \frac{S_1}{\sqrt{N}}$$

De donde se puede despejar "N" según:

$$N = \frac{(t_{N-1, \frac{\alpha}{2}})^2 \cdot S_1^2}{\varepsilon^2}$$

Caso 3: Intervalos de confianza para la varianza poblacional.

En poblaciones con comportamiento normal se cumple:

$$\frac{N \cdot S^2}{\sigma^2} = \chi_{N-1}^2$$

El estadístico será " χ_{N-1}^2 ", de donde se obtendrán $K_{1, \frac{\alpha}{2}}$ y $K_{2, \frac{\alpha}{2}}$

$$\sigma^2 \in \left[\frac{N \cdot S^2}{K_{2, \frac{\alpha}{2}}}; \frac{N \cdot S^2}{K_{1, \frac{\alpha}{2}}} \right]$$

Caso 4: Intervalos de confianza para dos poblaciones.

Caso 4.1: Diferencia de medias en poblaciones normales con varianza conocida. Siendo N y M los tamaños muestrales de las poblaciones X e Y.

Estadístico: $N(0,1)$, de donde se obtendrá $Z_{\frac{\alpha}{2}}$

$$(\mu_x - \mu_y) \in [(\bar{X}_N - \bar{Y}_M) \pm z_{\frac{\alpha}{2}} \cdot \sqrt{(\frac{\sigma_x^2}{N} + \frac{\sigma_y^2}{M})}]$$

Caso 4.2: Diferencia de medias en poblaciones normales con varianzas desconocidas pero iguales (muestras independientes). Siendo N y M los tamaños muestrales de las poblaciones X e Y.

Estadístico: $t_{N+M-2, \frac{\alpha}{2}}$

$$(\mu_x - \mu_y) \in [(\bar{X}_N - \bar{Y}_M) \pm t_{N+M-2, \frac{\alpha}{2}} \cdot \sqrt{(\frac{N \cdot S_x^2 + M \cdot S_y^2}{N+M-2}) \cdot (\frac{1}{N} + \frac{1}{M})}]$$

Caso 4.3: Diferencia de medias para dos poblaciones normales muestras dependientes. Siendo N el tamaño muestral.

Estadístico: $t_{N-1, \frac{\alpha}{2}}$

$$(\mu_x - \mu_y) \in [(\bar{X} - \bar{Y}) \pm t_{N-1, \frac{\alpha}{2}} \cdot \frac{S_{1d}}{\sqrt{N-1}}]$$

S_{1d} la cuasi desviación típica muestral de la variable aleatoria diferencia, para la muestra dependiente.

Caso 4.4: Intervalos de confianza para el cociente de varianzas de muestras independientes.

Estadístico: $F_{(N-1),(M-1),\frac{\alpha}{2}}$

$$\frac{\sigma_x^2}{\sigma_y^2} \in \left[\frac{(M-1) \cdot N \cdot S_x^2}{K_2(\frac{\alpha}{2}) \cdot (N-1) \cdot M \cdot S_y^2}; \frac{(M-1) \cdot N \cdot S_x^2}{K_1(\frac{\alpha}{2}) \cdot (N-1) \cdot M \cdot S_y^2} \right]_{\gamma}$$

Caso 5: Intervalos de confianza en poblaciones no normales.

Se deben emplear muestras grandes (**N 30**):

Caso 5.1: Intervalos para la media

Estadístico: $N(0,1)$ de donde se obtendrá $Z_{\frac{\alpha}{2}}$

$$\mu \in \left[\bar{x} \pm Z_{\frac{\alpha}{2}} \cdot \frac{S_1}{\sqrt{N}} \right]_{\gamma}$$

Caso 6.1: Intervalos de confianza para la proporción dicotómica.

En muestras grandes (**N 30**) ocurre:

$$\hat{p} = N(\bar{p}; \sqrt{\frac{\bar{p} \cdot \bar{q}}{N}})$$

Estadístico: $N(0,1)$ de donde se obtendrá $Z_{\frac{\alpha}{2}}$

$$p \in \left[\bar{p} \pm Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{p} \cdot \bar{q}}{N}} \right]_{\gamma}$$

Caso 6.2: Intervalos de confianza para la diferencia de proporciones dicotómicas.

En muestras grandes (**N 30**) ocurre:

$$\hat{p} = N(\bar{p}; \sqrt{\frac{\bar{p} \cdot \bar{q}}{N}})$$

Estadístico: $N(0,1)$ de donde se obtendrá $Z_{\frac{\alpha}{2}}$

$$(p_x - p_y) \in \left[(\bar{p}_x - \bar{p}_y) \pm Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{p}_x \cdot \bar{q}_x}{N} + \frac{\bar{p}_y \cdot \bar{q}_y}{M}} \right]_{\gamma}$$