# ANÁLISIS DE DATOS AGROPECUARIOS

IVÁN RAMÍREZ-MORALES / BERTHA MAZON-OLIVO

Editorial UTMACH

REDES 2017

COLECCIÓN EDITORIAL

# Análisis de Datos Agropecuarios

Iván Ramírez-Morales
Bertha Mazon-Olivo

Coordinadores

Primera edición en español, 2018

Este texto ha sido sometido a un proceso de evaluación por pares externos con base en la normativa editorial de la UTMACH

Ediciones UTMACH

Gestión de proyectos editoriales universitarios

302 pag; 22X19cm - (Colección REDES 2017)

Título: Análisis de Datos Agropecuarios. / Iván Ramírez-Morales / Bertha Mazon-Olivo (Coordinadores)

ISBN: 978-9942-24-120-7

*Publicación digital*

**Título del libro:** Análisis de Datos Agropecuarios.

**ISBN:** 978-9942-24-120-7

**Comentarios y sugerencias:** editorial@utmachala.edu.ec

**Diseño de portada**: MZ Diseño Editorial

**Diagramación:** MZ Diseño Editorial

**Diseño y comunicación digital:** Jorge Maza Córdova, Ms.



Km. 5 1/2 Vía Machala Pasaje

www.utmachala.edu.ec

Machala - Ecuador

# Índice

# Introducción

El análisis de datos es un proceso complejo que trata de encontrar patrones útiles y relaciones entre los datos a fin de obtener información sobre un problema específico y de esta manera tomar decisiones acertadas para su solución.

Las técnicas de análisis de datos que son exploradas en el presente libro son actualmente utilizadas en diversos sectores de la economía. En un inicio, fueron empleadas por las grandes empresas a fin de incrementar sus rendimientos financieros.

El libro se basa en la aplicación de la especialización inteligente, de este modo, gracias al trabajo colaborativo, se combina al sector agropecuario con las tecnologías, matemáticas, estadística y las ciencias computacionales, para la optimización de los procesos productivos.

La idea de descubrir la información oculta en las relaciones entre los datos, incentiva a encontrar aplicaciones para el sector agropecuario, por ejemplo los obtenidos de una producción avícola, o los datos que se generan durante los procesos de fermentación, los parámetros físicos y químicos del suelo, del agua y de las plantas, los datos de sensores, de espectrometría, entre otros.

En la actualidad, este sector se ha mantenido con su producción habitual sin un destacado repunte ni diferenciación, a pesar de existir herramientas científicas que han permitido desarrollar dispositivos tecnológicos y sus aplicaciones.

Este libro ha sido el resultado de la sistematización de las experiencias individuales de un equipo humano con objetivos comunes y una historia académica multidisciplinar, cuyos hallazgos de investigación han sido publicados en revistas científicas y conferencias de alto impacto. El área temática sobre la que se centra este texto es en técnicas de extracción, procesamiento y análisis de datos del ámbito agropecuario, se combinan para entregar al lector una obra de calidad y alto valor científico.

Así, el presente libro está concebido desde diferentes puntos de vista de profesionales agrónomos, informáticos, electrónicos, matemáticos, estadísticos y empresarios. Todos buscan un objetivo en común: "descubrir el conocimiento oculto en los datos que proporcione una ventaja competitiva". Se aborda el ciclo completo del proceso de obtención de conocimiento a partir de datos crudos del sector agropecuario, con la finalidad de apoyar la toma de decisiones. Este ciclo involucra procesos de: selección de los datos (extracción, comunicación, almacenamiento), pre-procesamiento, transformación, aplicación de modelos y/o técnicas de análisis, presentación e interpretación de resultados. El enfoque temático del libro es el siguiente:

Capítulo 1: Ciencia de Datos en el sector Agropecuario.- En este capítulo se aborda una revisión desde los inicios del análisis de datos en el sector agropecuario hasta el progreso actual que se ha dado en esta área del conocimiento que se considera como la nueva revolución en la agricultura y la ganadería de precisión.

Capítulo 2: Obtención de datos en sistemas agropecuarios.- El enfoque del capítulo es la generación de datos crudos en los sistemas agropecuarios, aplicando métodos y técnicas básicas donde se registran información de: número de unidades producidas, cantidad de nutrientes, variables climáticas, muestreo y monitoreo de organismos vivos, entre otros.

Capítulo 3: Internet de las cosas (IoT).- Este capítulo aborda los sistemas de telemetría para obtención de datos y control de dispositivos, aplicando tecnologías como: redes de sensores inalámbricos (dispositivos electrónicos, sensores, actuadores y puertas de enlace), protocolos de comunicación, centros de procesamiento de datos (cloud computing) y aplicaciones IoT para el sector agropecuario.

Capítulo 4: Matemáticas aplicadas al sector agropecuario.- Este capítulo explica los procedimientos para la creación de modelos matemáticos determinísticos que representen procesos asociados al sector agropecuario, como una alternativa de solución en la ingeniería.

Capítulo 5: Estadística básica con datos agropecuarios.- El capítulo se enfoca en los atributos, escalas de medición de las variables, su influencia en la elección del procedimiento estadístico a desarrollar, así como, el papel de las medidas de resumen, estimación puntual y prueba de hipótesis en la investigación científica.

Capítulo 6: Estadística predictiva con datos agropecuarios.- El capítulo considera las principales técnicas de la estadística avanzada aplicada al sector agropecuario, con el propósito de establecer predicciones que permita tomar mejores decisiones.

Capítulo 7: Inteligencia de negocios en el sector agropecuario.- El capítulo comprende la obtención de conocimiento a partir de datos crudos con la finalidad de apoyar la toma de decisiones en empresas del sector agropecuario. Involucra procesos de extracción, transformación y almacenamiento de datos en nuevos almacenes (Data warehouse - Big Data), distribución y análisis de la información con técnicas: multidimensional OLAP y tableros de control (dashboards).

Capítulo 8: Inteligencia Artificial aplicada a datos agropecuarios.- El capítulo trata sobre las principales técnicas de machine learning aplicadas a los datos agropecuarios, entre éstas se destacan: las redes de neuronas artificiales, máquinas de soporte de vectores, vecinos más cercanos, análisis de componentes principales, entre otros.

# 01 Capítulo Ciencia de datos en el sector agropecuario

Iván Ramírez-Morales; Bertha Mazon-Olivo; Alberto Pan

En el año 2015 de acuerdo a la revista Fortune, 151 startups de tecnología fueron financiados con $ 976 millones para enfocarse en el análisis de datos agropecuarios para la producción de alimentos. Transnacionales como Monsanto, Dupont y Archer Daniels Midland están invirtiendo importantes sumas de dinero en programas de ciencia de datos agrícolas.

**Iván Ramírez-Morales:** Doctor en Medicina Veterinaria y Zootecnia por la Universidad Agraria de la Habana, Máster en Desarrollo Comunitario por la Universidad Nacional de Loja y Doctor en Tecnologías de la Información y de las Comunicaciones por la Universidade A Coruña, ha realizado varios cursos en Brasil, Japón, Perú y Argentina. Fue Oficial de Territorio del Programa Marco ART/PNUD de la ONU, y Director de Planificación del Gobierno Provincial de El Oro. Actualmente es Profesor Titular en la Universidad Técnica de Machala, su área de investigación se centra en el uso de tecnologías para el mejoramiento de la productividad agropecuaria, Cuenta a la fecha más de 15 publicaciones indexadas, varias de ellas en revistas de alto impacto en los índices de JCR y SJR.

**Bertha Mazon-Olivo:** Ingeniera en Sistemas y Magíster en Informática Aplicada por la Escuela Superior Politécnica de Chimborazo. Profesora Titular en la Universidad Técnica de Machala. Es estudiante del programa doctoral en Tecnologías de la Información y las Comunicaciones en Universidade da Coruña, España. Sus líneas de investigación son: Internet de las Cosas, Ciencia de Datos y Desarrollo de Aplicaciones Informáticas. Cuenta con varias publicaciones indexadas.

**Alberto Pan:** Director Técnico de Denodo y Profesor Asociado de la Universidad de A Coruña. Recibió una Licenciatura en Ciencias de la Computación en la Universidad de A Coruña en 1996 y un Ph.D. en Informática por la misma universidad en 2002. Sus intereses de investigación están relacionados con la extracción e integración de datos y la automatización de la web. Alberto ha dirigido varios proyectos a nivel nacional y regional en el campo de la integración de datos y acceso a la Web oculta. También es autor o coautor de numerosas publicaciones en revistas científicas y actas de congresos.

Este movimiento de inversiones tiene lógica, ya que en los últimos años han emergido un número creciente de maquinarias, equipos y sistemas de monitoreo y control para el sector agropecuario que generan datos, y estos datos requieren ser interpretados. Entre los dispositivos generadores de datos se encuentran los tractores, arados, sembradoras, cosechadoras, las redes de sensores, estaciones meteorológicas, los espectrómetros portátiles, drones, cámaras, imágenes de satélite, invernaderos y galpones inteligentes, entre otros.

Por este motivo el Departamento de Agricultura de los Estados Unidos ha invertido en su iniciativa OpenAg, que ha hecho públicos sus datos. Esta situación implica un necesario esfuerzo multidisciplinario entre profesionales de las ciencias de datos y del sector agropecuario, para dar respuesta a la demanda mundial generada y promover una producción de alimentos más eficiente.

Los datos que se pueden recolectar o generar, con el uso de las Tecnologías de la Información y Comunicación (TIC) en la agricultura, pueden ser (Bendre, Thool, & Thool, 2015; Wolfert, Ge, Verdouw, & Bogaardt, 2017): datos en campo (características físico-químicas del suelo, topografía, datos de productividad), datos del clima, datos derivados de la interpretación de imágenes de cámaras o satelitales (variabilidad espacial y/o temporal del cultivo), datos de mapas de rendimiento, datos de mapas con prescripciones de aplicación de insumos, datos históricos, otros datos internos o externos. La integración y análisis de estos datos, pueden servir para generar decisiones automáticas (que se ejecutan en dispositivos, robots o maquinaria) o apoyar las decisiones humanas.

La agricultura de precisión permite incrementar la eficiencia y productividad aplicando un enfoque científico durante todas las fases del ciclo de producción de plantas y animales. Se plantea que la próxima gran revolución en la agricultura de precisión será a través de la ciencia de datos.

La optimización de los procesos productivos cada vez van evolucionando al fusionar los métodos, técnicas y tecnologías de la Agricultura de Precisión con la Ciencia de Datos, permi-

tiendo disminuir el uso de recursos como: energía, agua, fertilizantes, plaguicidas, etc.; lo que conlleva a producir más alimentos con menos esfuerzo, costos e impactos ambientales.

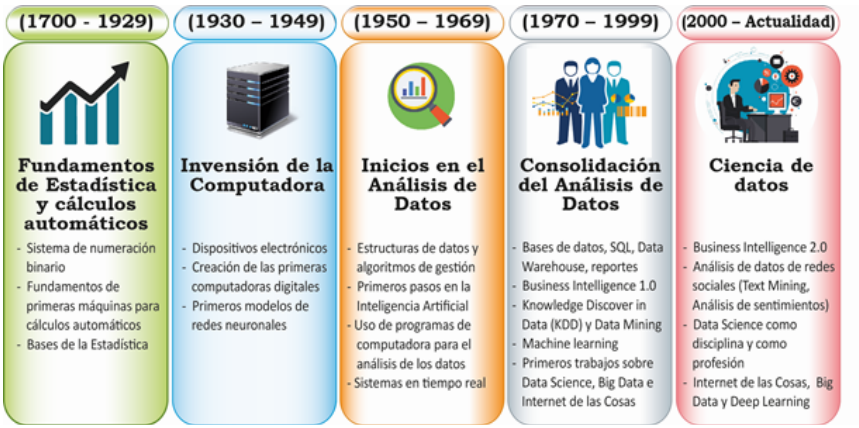
En la actualidad emergen nuevas técnicas y algoritmos para el procesamiento de los datos. La aplicación de estas técnicas involucra una diversidad de disciplinas.

Este capítulo se enfoca en la sistematización y comprensión de la evolución de los procesos de análisis de datos, los fundamentos teóricos de la Ciencia de Datos, los tipos de análisis de datos, las técnicas enfocadas al análisis de los datos agropecuarios aplicando las TIC, las herramientas y demás terminología asociada.

## Evolución de la ciencia de datos

La Ciencia de Datos es una disciplina relativamente joven, en comparación con otras disciplinas con las que guarda relación como la Estadística y las Ciencias de la Computación. Varios aportes relacionados con la evolución de Data Science (Capgemini, 2015; Jifa & Lingling, 2014; Press, 2013) fueron revisados para elaborar el resumen de hitos más relevantes que se describen a continuación (ver Imagen 1.1):

Imagen 1.1: Evolución de la Ciencia de Datos




Fuente: Elaboración propia

## Hitos importantes en la evolución de la ciencia de datos

### Fundamentos de estadística y cálculos automáticos (1700 – 1929)

A continuación se describen los hitos que establecieron las bases teóricas de la Estadística y de los cálculos automáticos, como los primeros aportes para la construcción de un computador.

- 1703: Leibniz G. propuso la utilización del sistema de numeración binario para cálculos sencillos. En la actualidad es la base de los cálculos automáticos que realiza un computador.
- (1791-1871): Charles Babbage es conocido como el padre de la computación. Diseñó e implementó parcialmente una máquina de diferencias mecánicas para calcular tablas de números. Además, diseñó una máquina analítica para ejecutar programas de tabulación o computación.
- (1815-1852): Ada Lovelace es conocida como la primera programadora de la historia que contribuyó con Babbage.
- 1847: George Boole siguió los pasos de Leibniz y propuso el Álgebra de Boole, que consiste en aritmética aplicada al sistema de numeración binario.
- 1749: Gottfried Achenwall introdujo el término alemán Statistik, que designaba originalmente el análisis de datos del Estado, es decir, la "ciencia del Estado". Sin embargo, no fue hasta el siglo XIX cuando el término Estadística adquirió el significado de recolectar y clasificar datos, concepto introducido por el militar británico Sir John Sinclair (1754-1835).
- 1805. Legendre propone los métodos de mínimos cuadrados para análisis de regresión que consiste en un proceso estadístico para estimar las relaciones entre variables.

- 1890: Herman Hollerith usó las tarjetas perforadas en el censo de Estados Unidos, en los siguientes años, fueron el medio para el ingreso y almacenamiento de datos.

### Invención de la computadora (1930 – 1949)

En estas dos décadas surgió definitivamente la invención de los dispositivos electrónicos y las primeras computadoras digitales.

- 1930's a 1940's: Se crearon unas pocas computadoras con fines militares, académicos e investigativos.
- 1936: Alan Turing propuso la teoría de "Máquina Universal de Turing" donde se estableció los principios del proceso de cómputo de cualquier computador digital.
- 1937: Claude Shannon a través de su Teoría de la Información, hizo posible la aplicación del álgebra de Boole en los dispositivos electrónicos, que fueron usados en la construcción de las primeras computadoras.
- 1943: Warren McCulloch y Walter Pitts, proponen los primeros modelos de redes neuronales, basadas en modelos matemáticos e informáticos.
- 1949, Von Neuman propone "La Arquitectura de Von Neuman" que buscó mejorar la arquitectura del computador ENIAC para dar origen al EDVAC y al primer computador comercial UNIVAC I, y sentó las bases del resto de computadoras digitales inventadas hasta la actualidad.

### Inicios en el análisis de datos (1950 – 1969)

Creación de los primeros modelo de datos, algoritmos de tratamiento y predicción e inicios de la inteligencia artificial.

- 1950's: En esta década surgen algoritmos de ordenación y búsqueda en estructuras de datos, el procesamiento por lotes, el almacenamiento temporal de datos.
- 1950: Primer modelo de predicción meteorológica propuesta por un equipo meteorólogos estadounidenses, empleando la computadora ENIAC.

- 1950: Turing dio los primeros pasos en la inteligencia artificial con su artículo "Computing Machinery and Intelligence", donde hizo la siguiente pregunta ¿puede pensar una máquina? El enfoque de Turing consistió en ver a la inteligencia artificial como una imitación del comportamiento humano.
- 1957 - Frank Rosenblatt diseñó la primera red neuronal para computadoras (el perceptrón), que simula los procesos mentales del cerebro humano.
- 1958: Hans Peter Luhn (de IBM), en el artículo "A Business Intelligence System", define la Inteligencia de Negocios, como: "la habilidad de aprender las relaciones de hechos presentados de forma que guíen las acciones hacia una meta deseada".
- 1959: Edsger Dijkstra propone el algoritmo Dijkstra de cálculo de la ruta más corta o mínima basada en la teoría de grafos para resolver problemas de transporte y logística.
- 1962: John Tukey escribió sobre "The Future of Data Analysis", donde argumenta la importancia del uso de programas de computadora en el análisis de los datos.
- 1960's: Surgimiento de sistemas en tiempo real, dando lugar al procesamiento de datos en tiempo real.
- 1965: Ingo Rechenberg, introdujo una técnica que llamó estrategia evolutiva y es el punto de partida de los algoritmos genéticos o computación evolutiva.
- 1969: Edgar Codd definió la teoría de las Base de datos relacionales (DBMS).

### Consolidación del análisis de datos (1970 – 1999)

La estadística tradicional es tratada con la tecnología informática, ejecutándose tareas como: almacenamiento, procesamiento y análisis de datos.

- 1970's: Auge de las primeras bases de datos y de las aplicaciones empresariales.

- 1974: Peter Naur, publica el libro "Concise Survey of Computer Methods" con los resultados de una encuesta de métodos informáticos para el procesamiento de datos utilizados en varias aplicaciones. Además, hace algunas definiciones importantes como los conceptos de dato y ciencia de datos. En el mismo año, Donald Chamberlin define el lenguaje estructurado de consultas en bases de datos (SQL).
- 1977: The International Association for Statistical Computing (IASC), vincula la estadística tradicional con la tecnología informática moderna y el conocimiento de expertos del dominio para convertir los datos en información y conocimiento.
- 1980s: Ralph Kimball y Bill Inmon, proponen el concepto de Data Warehouse y se crean los primeros sistemas de reportes.
- 1989: Gregory Piatetsky-Shapiro, organizó el primer workshop, relacionado con el descubrimiento del conocimiento en datos, titulado "Knowledge Discover in Data (KDD)", convirtiéndose en los siguientes años en un evento anual de ACM SIGKDD "Conference on Knowledge Discovery and Data Mining"[1] hasta la actualidad.
- 1989: Howard Dresner, populariza el término Business Intelligence (BI).
- 1990s: Proliferación de múltiples aplicaciones BI de escritorio dando lugar a Business Intelligence 1.0. También es la década del aprendizaje automático "Machine Learning"[2] que pasa de un enfoque basado en el conocimiento a un enfoque basado en datos. Los científicos comienzan a crear programas para que las computadoras analicen grandes cantidades de datos y obtengan conclusiones, o "aprendan", de los resultados.

[1] http://www.kdd.org/

[2] http://todobi.blogspot.com/2016/02/una-breve-historia-del-machine-learning.html

- 1996: en una reunión organizada en Japón por los miembros de International Federation of Classification Societies (IFCS), se volvió a mencionar la ciencia de datos en la conferencia titulada "Data science, classification, and related methods".
- 1996: Fayyad, Piatetsky-Shapiro y Smyth publicaron un trabajo titulado "From Data Mining to Knowledge Discovery in Databases" donde diferencian los conceptos de KDD y Data Mining. KDD se refiere al proceso general de descubrir el conocimiento útil a partir de datos, y la minería de datos se refiere a un paso particular en este proceso. La minería de datos es la aplicación de algoritmos específicos para extraer patrones de datos.
- 1997: El profesor C. F. Jeff Wu dio la conferencia inaugural titulada "Statistics = Data Science?" para su nombramiento en la Universidad de Michigan, donde hace una petición de renombrar a la estadística como ciencia de datos y a los estadistas como científicos de datos.
- 1999: J. Zahavi en su trabajo "Mining Data for Nuggets of Knowledge" critica a los métodos estadísticos convencionales de que sólo funcionan bien con pequeños conjuntos de datos; pero, las bases de datos actuales pueden involucrar millones de filas y decenas de columnas de datos introduciendo el término "Big Data", aduciendo que la escalabilidad es un gran problema en la minería de datos. En este mismo año, Kevin Ashton, en una conferencia en Procter & Gamble, habló por primera vez de Internet de las Cosas o "Internet of Things".

### Ciencia de datos (2000 – hasta la actualidad)

Uso generalizado de la analítica de datos, Inteligencia de Negocios, Big Data, Internet de las Cosas, Machine Learning y surgimiento de Deep Learning.

- 2000's – 2010's: Consolidación de las aplicaciones Business Intelligence 2.0 mediante plataformas BI (Oracle, SAP, IBM, Microsoft, Tableau, Qlik, etc.) que gestionan y analizan información de Data Warehouse y datos no estructurados. Crecimiento de la Web 2.0 o era de las

redes sociales como: Linkedin, Facebook, Twitter, Instagram; y es el comienzo de la explosión de datos.

- 2001: William Cleveland, presenta el trabajo "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics" como una disciplina independiente que se extiende de la estadística e incorpora avances en computación de datos.
- 2002: Lanzamiento de las revistas "Data Science Journal"[3] y en el 2003 "Journal of Data Science"[4] .
- 2004 - 2005: Primeros desarrollo de plataformas de Big Data e Internet de las Cosas (MapReduce, Hadoop, etc.).
- 2006 - Geoffrey Hinton acuña el término Deep Learning "aprendizaje profundo" para explicar nuevos algoritmos que permiten a las computadoras ver y distinguir objetos y texto en imágenes y videos.
- 2009: Mike Loukides escribió "What is Data Science?"[5] .
- 2012: Tom Davenport y Patil en su publicación "Data Scientist: The Sexiest Job of the 21st Century" en Harvard Business Review[6], menciona que Data Science se ha convertido en una opción muy atractiva de estudio de grados de Máster y/o PhD debido a que los científicos de datos son muy demandados y mejor pagados en muchas empresas que tienen que tratar con un gran volumen y diversidad de datos.

Cabe resaltar que desde finales de los noventa, los aportes sobre ciencia de datos, como: artículos científicos, conferencias y libros se han incrementado considerablemente, proponiéndose una variedad de métodos, técnicas y algoritmos para los diferentes tipos de análisis de datos. A la par, se han desarrollado varias herramientas informáticos y lenguajes de programación que sirven para hacer los cálculos más eficientes.

[3] https://datascience.codata.org/

[4] http://www.jds-online.com/journal/

[5] https://www.oreilly.com/ideas/what-is-data-science

[6] https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

# Fundamentos de la ciencia de datos

## Ciencia de datos (data science)

La ciencia de datos consiste en la aplicación de métodos científicos para construir algoritmos y sistemas que permiten detectar patrones y descubrir conocimiento útil para la toma de decisiones. Involucra procesos de integración y análisis de datos de distintas fuentes y en una variedad de formatos, a fin de construir modelos que ayudan a identificar y comprender fenómenos complejos.

Varios autores han contribuido en la definición de Ciencia de Datos. A continuación, se citan algunos:

- "La ciencia que trata con los datos, una vez que se han establecido, mientras que la relación de los datos con lo que representan se delega a otros campos y ciencias" (Naur, 1974).
- "La extracción de conocimiento útil de los datos para resolver problemas empresariales mediante un proceso sistemático con etapas bien definidas" (Provost & Fawcett, 2013).
- "Disciplina que crea sistemas y algoritmos para descubrir conocimiento, detectar patrones, generar información útil y/o realizar predicciones a partir de datos a gran escala" (Molina-Solana, Ros, Dolores Ruiz, Gomez-Romero, & Martin-Bautista, 2017).
- "Extracción de conocimiento accionable directamente de los datos a través de un proceso de descubrimiento o formulación y prueba de hipótesis" (NIST, 2015).

Según Gartner (2014) en su reporte gráfico "Hype Cycle for Emerging Technologies Maps the Journey to Digital Business"[7] , Data Science se muestra como una tecnología o disciplina emergente, de gran expectativa por la comunidad de científicos, profesionales, empresarios y personas que les interesa obtener un valor agregado de sus datos.

[7] https://www.gartner.com/newsroom/id/2819918

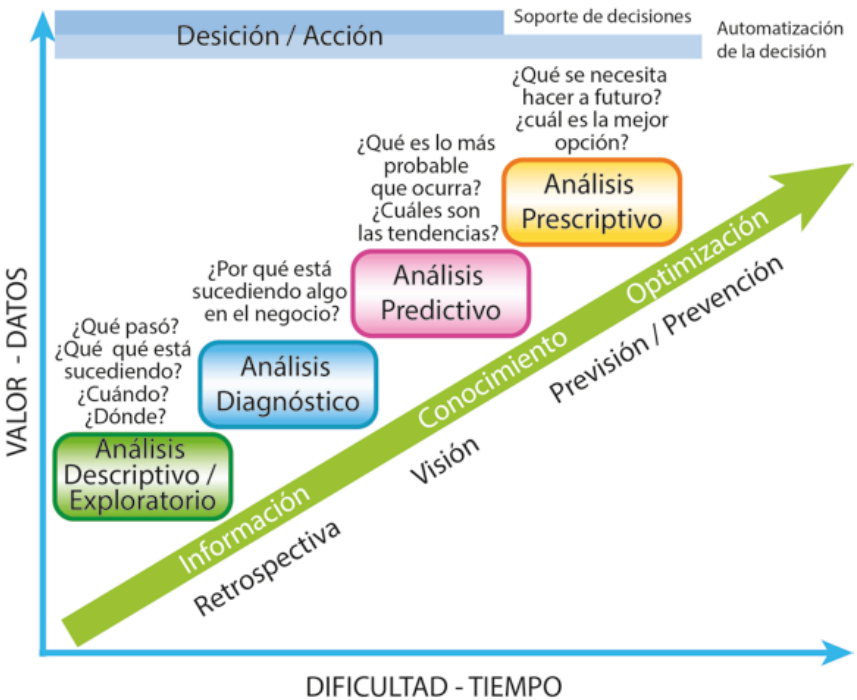
## Análisis de datos (data analytics)

El proceso analítico de datos es la síntesis del conocimiento a partir de la información (NIST, 2015). El análisis de datos se traduce como el proceso de obtención, transformación y modelado de datos, con el fin de determinar patrones de comportamiento que ayuden en la toma de decisiones.

### Tipos de análisis de datos

Según (Gartner, 2012; Loury, 2014; National Academi of Science, 2017; Sivarajah, 2017), los métodos analíticos de datos se clasifican en: Análisis descriptivo y exploratorio, Análisis Diagnóstico, Análisis Predictivo y Análisis Prescriptivo (ver Imagen 1.2).

Imagen 1.2: Tipos de análisis de datos




Fuente: Elaboración propia

A continuación se describen los tipos de análisis de datos:

- El análisis descriptivo y exploratorio. Implican el resumen y la descripción de patrones de conocimiento utilizando métodos de visualización de datos estáticos y dinámicos. Las fuentes de datos sin procesar (raw data) pueden ser conjuntos de datos (data sets) en distintos formatos: hoja de cálculo (xls, xlsx, ods, ect.), archivos de texto (CSV, TXT, XML, JSON, HTML, etc.), bases de datos relacionales, informes de sistemas transaccionales, etc. A menudo, este tipo de análisis sirve para crear informes de gestión que se ocupan de modelar comportamientos pasados o presentes; responden a las preguntas: ¿qué pasó en el negocio?, ¿qué ocurrió y qué está sucediendo?, ¿cuándo?, ¿dónde?

  Los análisis descriptivos son la base de información de una organización; es decir, son las principales aplicaciones de la inteligencia empresarial o de negocios (Mazon-Olivo et al., 2017). Los métodos más comunes de análisis descriptivo son: informes y consultas estáticas o personalizadas (reports & queries statics or Ad hoc), métodos estadísticos básicos (media, moda, desviación estándar, varianza, medición de frecuencia de eventos específicos, etc.), análisis multidimensional OLAP (On-Line Analytical Processing - procesamiento analítico en línea), tableros de control y cuadros de mandos (dashboards, scorecards) y otras técnicas de visualización de datos.

- Análisis diagnóstico. Consiste en sondear datos para certificar o rechazar proposiciones comerciales o hipótesis. Este tipo de análisis se basa en información descriptiva para comprender: ¿Por qué está sucediendo algo en el negocio?

- Análisis predictivo. Este tipo de análisis busca descubrir patrones y capturar relaciones entre los datos; pronosticar el resultado probable de una situación dada y generar un modelo estadístico de los datos actuales e históricos para determinar las posibilidades futuras, en base a técnicas de aprendizaje supervisado, no supervisado y

semi-supervisado; responden a las preguntas ¿Qué es lo más probable que ocurra a futuro? ¿Cuáles son las tendencias?

- Análisis prescriptivo. Este tipo de análisis se realiza para encontrar cuál es la solución entre una gama de variantes. Su tarea es optimizar recursos y aumentar la eficiencia operativa. Las soluciones prescriptivas ayudan a los analistas de la empresa, en la toma de decisiones mediante la determinación de acciones y la evaluación del impacto con respecto a los objetivos, los requisitos y las limitaciones del negocio. El análisis prescriptivo se basa en la aplicación de reglas de negocio, aprendizaje automático y procedimientos de modelado computacional, para intentar responder a la pregunta: ¿qué se puede hacer para que algo suceda?, ¿cuál es la mejor opción de decisión?, ¿qué acciones tomar?

Un tipo de análisis prescriptivo es el Análisis preventivo que consiste en tener la capacidad de tomar medidas preventivas sobre eventos que pueden influir indeseablemente en el desempeño de la organización, por ejemplo, identificar los posibles riesgos y recomendar estrategias de mitigación en el futuro.

### Campos y técnicas de análisis de datos

Las Técnicas de análisis de datos más prioritarias se resumen el Cuadro 1.1, están clasificadas por categorías, tipo de análisis, preguntas y campos específicos de acción. Los aportes de varios autores (National Academi of Science, 2017; Provost & Fawcett, 2013; Sivarajah, 2017) y algunos sitios web[8] fueron revisados  para elaborar este cuadro.

[3] https://www.nap.edu/read/23670/chapter/6

https://www.sv-europe.com/blog/10-reasons-organisation-ready-prescriptive-analytics/

http://www.healthcareimc.com/main/making-sense-of-analytics/

https://twitter.com/doug_laney/status/611172882882916352

**Categorías del análisis de datos:**

Retrospectiva (mirada al presente o pasado). Consiste en descubrir patrones de información o conocimiento en bases de datos (transaccionales estructuradas y no estructuradas o pre-procesadas como data warehouse) que registran hechos o transacciones de la organización. Se busca apoyar en la decisión a los mandos medios y estratégicos, con información procesada y organizada.

Visión (mirada al presente o pasado para predecir el futuro). Conlleva procesos de diagnóstico, predicción y de apoyo a la decisión en una organización; además del descubrimiento de patrones de información, la búsqueda de causas y efectos a posibles problemas encontrados, la comprobación de hipótesis, clasificación de datos en grupos de interés, identificación de tendencias, y/o se predicción de información faltante pasada, presente o futura.

Cuadro 1.1. Campos y técnicas de análisis de datos

| *Cat. | **Ta | Preguntas | Campo Específico | Técnicas |
|---|---|---|---|---|
| Retrospectiva | Análisis descriptivo y exploratorio | ¿Qué pasó en el negocio? | Descriptivo | Reportes y consultas estáticas mediante lenguaje estructurado de consultas (SQL) |
| | | ¿Qué ocurrió y qué está sucediendo? ¿Cuándo? ¿Dónde? | Descriptivo de apoyo a la decisión Inteligencia de Negocios | Consultas y reportes personalizados (Ad Hoc)<br>Análisis multidimensional (OLAP)<br>Indicadores clave de desempeño (KPI's)<br>Tableros de control y cuadros de mando (dashboards y scorecards)<br>Métodos estadísticos básicos |

| *Cat. | **Ta | Preguntas | Campo Específico | Técnicas |
|---|---|---|---|---|
| Visión | Análisis Diagnóstico | ¿Por qué está sucediendo algo en el negocio? | Descriptivo, exploratorio y apoyo a la decisión<br>Inteligencia de Negocios<br>Minería de datos descriptiva<br>Análisis situacional<br>Causa y efecto | Análisis OLAP con distintos niveles de detalle (slice and dice, drill and down, across)<br>Tableros de control y cuadros de mando (dashboards y scorecards)<br>Monitoreo automático y alertas<br>Clustering o segmentación, reglas de asociación, patrones secuenciales<br>Pruebas de hipótesis<br>Minería de textos, análisis de sentimientos u opiniones<br>Otras técnicas |
| | Análisis Predictivo | ¿Qué información que se desconoce?<br>¿Qué es lo más probable que ocurra a futuro?<br>¿Cuáles son las tendencias? | Predictivo y apoyo a la decisión<br>Minería de datos predictiva<br>Modelos de análisis de tendencias<br>Evaluación de probabilidad y de riesgos<br>Análisis de big data (datos a gran escala)<br>Análisis de datos no estructurados: imágenes, audios, videos, archivos .pdf, etc. | Aprendizaje de máquina (machine learning), aprendizaje profundo (deep learning)<br>Regresión lineal y no lineal<br>Árboles de decisión<br>Métodos bayesianos<br>Redes neuronales<br>Series temporales<br>Máquina de soporte vectorial (Support-vector machines)<br>Otras técnicas |

| *Cat. | **Ta | Preguntas | Campo Específico | Técnicas |
|---|---|---|---|---|
| Previsión / prevención | Análisis Prescriptivo | ¿Qué se necesita hacer a futuro? ¿Cuál es la mejor opción de decisión? ¿Qué acciones tomar? ¿Cómo actuar? | Previsión, recomendación<br>Automatización de la decisión<br>Formulación de estrategias<br>Planificación en base a escenarios o la mejor opción<br>Sistemas de recomendación | Métodos de simulación: simulación de eventos discretos, simulación de Monte Carlo, modelos estocásticos con incertidumbre<br>Modelos de optimización<br>Motor de reglas (rules engine) y Procesador complejo de eventos (complex event processor)<br>Programación Lineal y no lineal<br>Otras técnicas |

* CAT=Categoría, ** TA=Tipo de Análisis

Fuente: Elaboración propia

Previsión / prevención (de la visión a la actuación). Además de apoyar a la decisión, se pretende tomar acciones o recomendar estrategias para resolver problemas puntuales; dependiendo del caso, se busca automatizar la decisión; por ejemplo, en un sistema de telemetría aplicado a la agricultura que utiliza tecnologías de Internet de las Cosas, puede activar el riego en el momento oportuno, basándose en datos de sensores de humedad del suelo. Otros ejemplos son los sistemas de recomendación de fertilizantes para nutrición de plantas, según los requerimientos nutricionales de un cultivo, determinar tipo y dosis de fertilizantes a aplicar y el presupuesto. Otros casos, son los sistemas de predicción/prevención de riesgos en la producción (alertas de plagas, inundaciones, incendios, etc.), sistemas que ayudan a generar estrategias de optimización en el uso de recursos, etc.

## Disciplinas que se relacionan con la ciencia de datos

La ciencia de datos es una disciplina relativamente joven e interdisciplinaria; guarda relación con otras disciplinas (ver Imagen 1.3) como la Estadística, Matemática y las Ciencias de la Computación. En el contexto de las Ciencias de la computación, los campos específicos interrelacionados son: Inteligencia de Negocios, Minería de datos, Aprendizaje Automático, Inteligencia Artificial, Redes Neuronales, Bases de datos, Datos a Gran Escala (Big data), entre otros (Costa & Santos, 2017; Kamilaris, Kartakoullis, & Prenafeta-boldú, 2017; Leading Edge, 2015).

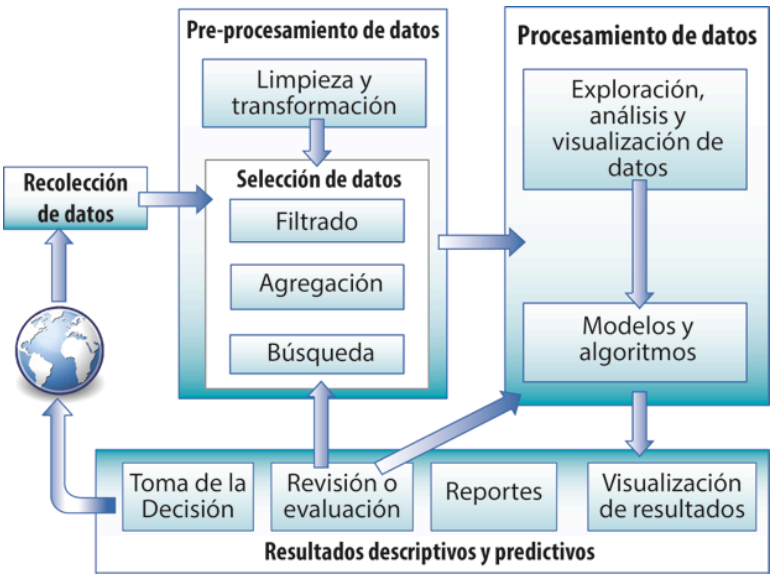
Imagen 1.3: Disciplinas que se relacionan con Ciencia de Datos


Fuente: Elaboración propia

## El ciclo de vida de la ciencia de datos

El ciclo de vida de la ciencia de los datos comprende un conjunto de procesos que transforman los datos crudos en conocimiento accionable (NIST, 2015).

Imagen 1.4: El ciclo de vida de la Ciencia de Datos




Fuente: (Larson & Chang, 2016; Provost & Fawcett, 2013),

En la Imagen 1.2, se observan las etapas del ciclo de vida de Data Science:

- Recolección de datos crudos del mundo real en diversidad de formatos.
- Pre-procesamiento de datos. Comprende tareas de edición, limpieza y transformación (data munging), ajuste de datos para detectar omisiones, verificación de legibilidad y consistencia para su codificación y almacenamiento, con el propósito de garantizar la calidad de datos (data quality). Seguidamente, ciertos datos son seleccionados aplicando filtros, agregaciones y búsqueda parcial con el fin de realizar tareas de procesamiento.
- Procesamiento de datos. Las tareas que se realizan son: exploración, análisis y visualización de los datos, creación de modelos y algoritmos confiables para obtener resultados.
- Obtención de Resultados. Éstos pueden ser descriptivos o predictivos y deben ser debidamente evaluados

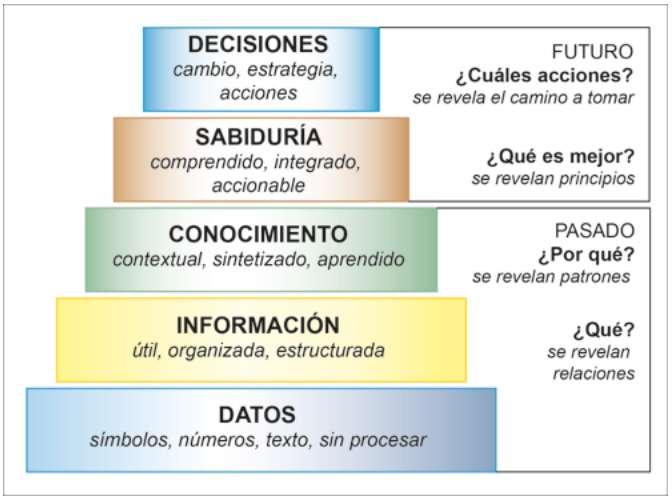
y/o revisados para generar conclusiones, nuevas teorías o tomar decisiones que implican cambios con afectación al mundo real.

## Jerarquía del conocimiento en ciencia de datos

Hay varios enfoques sobre el proceso de transformación de los datos en información y conocimiento. García-Marco (2011) explica en su trabajo, la pirámide de la información como un constructo clave de la ciencia de la información; hace una crítica a otros modelos similares y presenta un modelo ampliado desde sus hallazgos y perspectiva. La jerarquía DICS, explicada en (García-Marco, 2011; Jifa & Lingling, 2014), consiste en la cadena de conversión: Dato –> Información –> Conocimiento –> Sabiduría. El cómo los datos se convierten en decisiones, tiene que ver con los procesos explicados en la sección del Ciclo de Vida de la Ciencia de Datos. Los datos que representan hechos reales no relacionados, son procesados para encontrar relaciones y generar información; el análisis de la información permite descubrir patrones muy útiles (conocimiento), que al ser presentados de manera oportuna a la persona adecuada y con la experiencia necesaria, pueden aprovechar para actuar con sabiduría y tomar las decisiones más acertada en beneficio de su organización. En la Imagen 1.5 se aprecia la pirámide del conocimiento, que va desde los datos hasta la toma de decisiones[9] .

[9] http://information4dummies.blogspot.com/2014/04/modelo-yo-conceptos-de-representacion.html

Imagen 1.5: La pirámide del conocimiento




Fuente: (Larson & Chang, 2016; Provost & Fawcett, 2013)

A continuación se explica la diferencia entre datos, información y conocimiento:

### Datos

Los datos son el recurso más abundante del planeta; se calcula que a nivel mundial, se generan a diario 2.5 Exabytes de datos (BBC News, 2014; Khoso, 2016). En todo el mundo, las empresas recopilan datos de sus transacciones diarias; los gobiernos recopilan regularmente datos de censos e informes de incidencias en los departamentos de policía; a diario en las redes sociales, millones de personas suben fotos, videos y envían mensajes de texto. Este diluvio de datos crece rápidamente y cada vez más, con las nuevas tecnologías como el Internet de las Cosas (IoT), las redes de sensores inalámbricas (WSN) y los objetos inteligentes (Smart Object). En el 2013, la cantidad total de datos en el mundo, fue de 4,4 Zetta bytes (Zb), y se estima unos 44 Zb para el 2020 y 163 Zb para el 2025 (BBC News, 2014; Khoso, 2016).

Los datos que se obtienen de diversas fuentes internas o externas de una organización y aquellos que aún no han sido procesados se los denomina datos crudos (raw data) o

datos en bruto. También, son considerados como la materia prima que se procesa para obtener información y conocimiento que es útil para la toma de daciones. Los datos representan características o atributos de un objeto o hecho de la vida real. A continuación se citan algunos aportes de autores respecto a la definición de dato:

- Naur P. (1974). "Los datos son una representación formal de hechos o ideas que pueden ser comunicados o manipulados por algún proceso".
- Según Flores (1981). "Los datos se describen además como una representación del mundo real".
- "El valor de los datos reside en su uso" National Science Foundation (NSF): Bits of power (1997).
- Collis (2009), "los datos se definen como una representación de hechos que pueden ser recogidos, registrados y utilizados como base para la toma de decisiones".
- Hernández (2014), define al dato como la "unidad mínima semántica que por sí sola no representa información".

Clases de datos. Los datos según su formato se clasifican en:

- Estructurados. Tiene un estructura bien definida, por ejemplo una base de datos relacional o almacén de datos (data warehouse).
- Semi-estructurados. Conjuntos de datos con un formato diferente a las bases de datos relacionales. Por ejemplo archivos de texto plano (CSV, JSON, XML, HTML, etc.), archivos de hojas de cálculo (XLS, XLSX, ODS, otras), reportes de sistemas informáticos, etc.
- No estructurados. Datos sin una estructura estándar, por ejemplo: imágenes, audios, videos, documentos de textos, etc.

Datos según el origen:

- Internos. Datos generados en la propia organización; por ejemplo datos de compras, producción, ventas, etc.
- Externos. Datos que se obtienen de sistemas externos.

Por ejemplo de sistemas gubernamentales (seguro social, servicios de recaudación de impuestos), redes sociales, datos en tiempo real provenientes de dispositivos inteligentes (Smart objects) utilizados en entornos de Internet de las Cosas (IoT), entre otros.

Datos según su valor:

- Cuantitativos (numéricos). Son de dos tipos: 1) discreto, corresponde a valores enteros, por ejemplo: número de plantas por hectárea en un cultivo de cacao, número de cajas de banano cosechadas por hectárea, etc.) y, 2) continuo, corresponde a valores numéricos con decimales. Por ejemplo: temperatura, humedad del suelo, etc.
- Cualitativos. O también conocidos como alfanuméricos, representan un texto con varios caracteres o un sólo carácter. Pueden combinarse letras, símbolos incluso números. Se clasifican en 1) ordinales, por ejemplo: fases fenológicas de un cultivo de banano y, 2) nominales por ejemplo: nombre del cultivo, nombre de un fertilizante, etc.
- Lógicos, representan un valor lógico que puede ser "verdadero" o "falso", "SI" o "NO". Por ejemplo ¿la solicitud ha sido aprobada? La respuesta puede ser Si o No

### Información

Este término se refiere a un conjunto de hechos que tienen un significado, un propósito y un formato adecuado para la toma de decisiones. La información[10] es el resultado del procesamiento y análisis de los datos. Los datos depurados, transformados, organizados, relacionados o clasificados se convierten en información. El concepto de información responde a preguntas como: ¿qué?, ¿quién?, ¿dónde? y ¿cuándo? La información es capaz de cambiar opiniones, pensamientos y criterios de acuerdo a la forma de ser percibida por el receptor (Guillén, López Ayuso, Paniagua, & Cadenas, 2015).

[9] http://infomedicsa.blogspot.com/2016/04/dikw-datos-informacion-conocimiento.html
http://mizrablogs.blogspot.com/2017/02/capitulo-ii-datos-en-medicina.html

## Conocimiento

Es la realidad de un objeto, captada y entendida por un sujeto. La información se transforma en conocimiento al ser interpretada de forma reflexiva y con base en la experiencia. El conocimiento responde a las preguntas ¿cómo?, ¿por qué?

Según Guillén et al. (2015), el conocimiento:

> *"...surge de ideas verificadas y validadas por convención"; es decir, "...es el resultado de procesar información y hallar ciertos patrones invariantes que generan un cuerpo coherente de juicios acerca del mundo".*

## Tipos de conocimiento

El conocimiento según el origen puede ser:

- Tácito. Cuando el conocimiento se basa en la experiencia personal, éste resulta sencillo de aplicarlo pero difícil explicarlo debido a que permanece en nuestro inconsciente de forma desarticulada.
- Explícito. Cuando el conocimiento es fácilmente explicado de manera escrita u oral; se encuentra estructurado y es fácil explicar y compartir con los demás. Ejemplo: libros, artículos científicos, etc.

Roiger (2017) presenta la clasificación del conocimiento según la forma de extracción:

- Evidente: Es aquella información que se puede obtener de forma sencilla. Por ejemplo a través de una consulta en una base de datos a través de un sistema informático, se puede verificar las facturas emitidas por la venta de un producto en una fecha determinada.
- Multi-dimensional: Es la información capaz de ser representada mediante varias perspectivas o vistas, brindando cierta estructura a los datos analizados. La técnica utilizada es el análisis mediante cubos OLAP. Por ejemplo: *Total de ventas de cajas de banano por, año, semestre y mes. Dónde: total ventas es la medida de la información, y las vistas son: producto=banano, año, semestre y mes*

- Oculto: Es aquella información que no se encuentra visible de manera superficial, es desconocida, por lo tanto, no evidente, pero útil si se aplican técnicas correctas para su descubrimiento. Entre las principales técnicas se encuentra las de Minería de Datos (Data Mining) y Estadística. Por ejemplo: En base a la información de ventas de camarón durante los últimos cinco años, determinar la proyección de ventas para el próximo año.
- Profundo: Es la información que se encuentra inmersa en los datos, pero con un grado de seguridad potencialmente alto, donde se debe poseer una especie de clave para entenderla e interpretarla. Se aplican técnicas como Aprendizaje de Máquina (Machine learning), Aprendizaje profundo (Deep Learning) y otras técnicas de inteligencia artificial.

**Sabiduría (Wisdom).**

Consiste en la capacidad de intuir, comprender e interpretar el conocimiento para planificar o tomar la mejor decisión posible a corto, mediano o largo plazo. Es la capacidad de emplear el juicio basado en principios para saber ¿cuál es la mejor decisión a tomar?

## El perfil del científico de datos (data scientist)

En la actualidad, el universo de los datos es abundante, crece cada vez y a pasos agigantados, trayendo consigo nuevas profesiones como el de científico de datos (data scientist). Al examinar la literatura científica, algunos autores han contribuido en la comprensión del perfil profesional del científico de datos. Davenport & Patil (2012) presentan el trabajo de científico de datos como el más atractivo del siglo XXI; además, describen el perfil y cómo las organizaciones pueden explotar su potencial. Dhar (2013) también describe el campo y las habilidades que debe tener un científico de datos. Provost y Fawcett (2013) propusieron relacionar el campo de la Ciencia de Datos con otros temas como Big Data y la toma de decisiones basada en datos e identificaron los principios fundamentales presentes en la Ciencia

de Datos. El Instituto Nacional de Estándares y Tecnología (NIST, 2015), define a un científico de datos como un profesional con suficiente conocimiento de las necesidades del negocio (empresa o institución), conocimiento del dominio del problema o del contexto de los datos, habilidades analíticas (estadística, matemática), manejo de herramientas de software e ingeniería de sistemas para administrar los procesos de datos según el ciclo de vida de la ciencia de los datos. Según Costa & Santos (2017), los conocimientos que debe saber el científico de datos son:

- Gestión de los datos. Características, esquemas y estructura de los datos, modelado de datos, administración de sistemas gestores de datos, manejo de grandes volúmenes de datos.
- Seguridad, privacidad y ética en el manejo de los datos.
- Teorías, métodos y herramientas computacionales. Programación de algoritmos, gestores de almacenamiento de datos, almacenes de datos (data warehouse), inteligencia artificial, aprendizaje de máquina (machine learning), aprendizaje profundo (Deep Learning).
- Capacidades personales y sociales. Perspicacia para los negocios, comunicación, emprendimiento, creatividad, trabajo interdisciplinario.
- Diseño de sistemas computacionales. Sistemas distribuidos, redes de datos, procesamiento paralelo/distribuido, manejo de servidores y computación en la nube, escalabilidad en sistemas distribuidos.
- Procesamiento de flujos de datos. Captura y extracción de datos, limpieza y transformación de datos, procesamiento de datos, minería de datos, visualización de datos y comunicación de resultados, análisis de datos (análisis descriptivo / exploratorio de datos, análisis de métodos predictivos /prescriptivos, análisis automatizado, etc.)
- Investigación de tópicos relacionados al campo de estudio. Estadística, matemática, aplicación del método científico, ciencias computacionales, dominio del problema relacionado con los datos, etc.

En el Cuadro 1.2, se aprecian algunas de las competencias del científico de datos, algunas de ellas definidas en los trabajos de: Davenport & Patil (2012), Dhar (2013), NIST (2015) y Costa & Santos (2017) .

Cuadro 1.2: Competencias del científico de datos

| Competencias Generales | Competencias específicas |
|---|---|
| Asegura flujos eficientes de datos | Realiza captura eficiente de los datos |
| | Limpia y transforma datos |
| | Asegura la calidad de los datos |
| | Integra datos de diversidad de fuentes y formatos |
| | Resuelve problemas basándose en el análisis de datos |
| | Extrae valor de los datos |
| | Respeta la privacidad, seguridad y ética de los datos |
| Identifica patrones y tendencias en los datos | Explora y visualiza datos o grandes volúmenes de datos |
| | Analiza datos aplicando técnicas y herramientas adecuadas |
| | Define y prueba hipótesis |
| | Realiza experimentos con los datos y responde a preguntas específicas |
| | Obtiene y valida conclusiones |
| | Encuentra patrones y realiza predicciones basado en conocimiento oculto en los datos |
| Diseña, construye, implementa y optimiza artefactos de datos | Crea y programa algoritmos para el tratamiento de datos |
| | Diseña e implementa plataformas hardware & software para gestión y análisis de datos |
| | Implementa herramientas de inteligencia de negocios (análisis, visualización de datos, minería de datos, etc.) |
| | Crea modelos estadísticos / matemáticos de análisis de datos |
| | Crea soluciones efectivas, escalables y robustas basadas en datos |
| Trabaja con diferentes datos | Maneja grandes volúmenes de datos |
| | Maneja datos de distintas estructuras y formatos |

| Competencias Generales | Competencias específicas |
|---|---|
| Comunica y disemina aportes a la ciencia de datos | Comunica las mejores prácticas de gestión y análisis de datos |
| | Realiza contribuciones con otros grupos relacionados con proyectos de datos |
| | Comunica, por escrito o verbalmente los resultados y hallazgos de sus de investigaciones relacionadas con datos |
| Contribuye en la gestión y mejora del rendimiento de un negocio | Comprende los requerimientos del usuario |
| | Identifica amenazas de la competencia basándose en los datos |
| | Define reglas y medidas para monitorear productos |
| | Sugiere nuevas estrategias y acciones a tomar para cambiar la dirección del negocio |
| | Contribuye a la mejora del rendimiento del negocio |
| | Informa y brinda soporte al personal de mandos medios y estratégicos del negocio |
| | Identifica tendencias y oportunidades |
| | Encuentra respuestas a preguntas importantes del negocio |
| | Presenta estrategias y recomienda acciones de mejora del rendimiento de la producción |
| | Advierte de posibles riesgos del negocio a la alta gerencia y aconseja las acciones a tomar |
| | Asiste en el replanteamiento de nuevos retos del negocio. |

## Herramientas para el científico de datos

Existen una variedad de herramientas tanto para la gestión de datos como para la visualización y analítica de datos. En el Cuadro 1.3 se muestra un listado de las herramientas que más presencia tienen en el mercado de la ciencia de datos.

Cuadro 1.3: Herramientas para el científicos de datos

| Tipo de Herramienta | Herramienta |
|---|---|
| Gestión de datos[11] | - Gestores de bases de datos relacionales: PostgreSQL, Oracle DB, Mysql, Microsoft SqlServer, Teradata, Netezza, etc.<br>- Gestores de bases de datos NoSQL: MongoDB, Couchbase, Amazon DynamoDB, Casandra, etc.<br>- Big data: Hadoop, Mahout, Hive, Spark, Kafka, Apache Cloudera[12], Apache Hortonwork[13] , etc.<br>- Motores de búsqueda: Elasticsearch, Microsoft Azure Search, Splunk, Google Search Appliance, etc. |
| Análisis y visualización de datos | - Extracción Transformación y Carga(ETL) e integración de datos: Pentaho Data Integration[14], Talend[15], Informatica[16], Denodo[17], etc.<br>- Inteligencia de Negocios: Microsoft PowerBI[18], Pentaho[19], QlikView[20], Tableau[21], Microstrategy[22] , Oracle BI[23], IBM Cognos[24], Eclipse BIRT, JasperReport, etc.<br>- Estadística y Minería de datos: R, Python, SPSS, Minitab, Orange, RapidMiner, Weka, Knime, etc.<br>- Aprendizaje de Máquina (machine learning)[25]: MatLab, Mathematica, SPSS, TensorFlow, Microsoft Azure Machine Learning, Apache Mahout, OpenCV, KNIME, R, Python, etc. |

11 https://db-engines.com/en/ranking/relational+dbms
12 https://www.cloudera.com/
13 https://es.hortonworks.com/
14 https://www.hitachivantara.com/en-us/products/big-data-integration-analytics/pentaho-data-integration.html
15 https://es.talend.com/
16 https://www.informatica.com/products/data-integration.html
17 https://www.denodo.com/en
18 https://powerbi.microsoft.com/es/
19 https://www.hitachivantara.com/go/pentaho.html
20 https://www.qlik.com/us/
21 https://www.tableau.com/
22 https://www.microstrategy.com/es
23 https://www.oracle.com/solutions/business-analytics/business-intelligence/index.html
24 https://www.ibm.com/products/cognos-analytics
25 https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico

## Campos de aplicación de la ciencia de datos

La ciencia de datos es un campo interdisciplinario y transversal, que puede aplicarse en todo ámbito, donde se requiere analizar datos como estrategia para descubrir el conocimiento oculto y aprovechable para tomar decisiones. La ciencia de datos está cambiando la manera de cómo nos ocupamos de: negocios, producción, marketing, publicidad, mejora de la calidad de los servicios, investigación científica, etc.

## Aplicación de la ciencia de datos en el sector agropecuario

En el caso del sector agropecuario, aunque un trabajador experimentado es capaz de supervisar personalmente los procesos productivos de animales y cultivos agrícolas, hay tareas que requieren de mucho esfuerzo si se realizaran manualmente; por ejemplo: detectar la presencia de enfermedades, estimar la producción en base a datos históricos, optimizar el uso de recursos, mejorar la calidad de la producción, pronosticar posibles riesgos de desastres naturales como sequías, inundaciones, heladas, etc. En la actualidad gracias al uso de tecnologías y la aplicación de la ciencia de datos, es posible el monitoreo de una unidad de producción agropecuaria (FAO, 2016), las 24 horas, 7 días a la semana, 365 días al año. La Agricultura de Precisión, el Internet de las Cosas y la Ciencia de Datos aplicadas eficientemente trae beneficios sustanciales, como el ahorro de tiempo y dinero a los agricultores, y la mejora de la producción.

En este capítulo se presentó un panorama general de lo que se puede hacer con la Ciencia de Datos. Primero se explica la evolución y los fundamentos teóricos, luego se explican los tipos de análisis de datos y las técnicas de análisis según el campo específico, seguido se determinan las disciplinas que guardan relación con la ciencia de datos, el ciclo de vida de data science, la jerarquía del conocimiento, el perfil del científico de datos y sus competencias y finalmente las herramientas que pueden utilizarse así como los campos de aplicación de data science en varios sectores como el agropecuario.

## Referencia bibliográfica


BBC News. (2014). Big Data: Are you ready for blast-off? Retrieved November 30, 2017, from http://www.bbc.com/news/business-26383058

Bendre, M. R., Thool, R. C., & Thool, V. R. (2015). Big Data in Precision Agriculture : Weather Forecasting for Future Farming. *In 2015 1st International Conference on Next Generation Computing Technologies* (pp. 4–5). Dehradun, India. http://doi.org/10.1109/NGCT.2015.7375220

Capgemini. (2015). A brief history of Data Science. Retrieved October 13, 2017, from https://whatsthebigdata.com/2015/02/17/history-of-data-science-infographic/

Collis, J., & Hussey, R. (2009). Business research: *A practical guide for undergraduate and postgraduate students*. (3rd ed.). Palgrave Macmillan: Hampshire. Retrieved from https://www.researchgate.net/publication/38177413_Business_research_A_practical_guide_for_undergraduate_and_postgraduate_students

Costa, C., & Santos, M. Y. (2017). The data scientist profile and its representativeness in the European e-Competence framework and the skills framework for the information age. *International Journal of Information Management*, (xxxx). http://doi.org/10.1016/j.ijinfomgt.2017.07.010

Davenport, T., & Patil, D. J. (2012). Data Scientist The Sexiest Job of the 21st Century Meet the people who can coax treasure out of messy, unstructured data. Retrieved May 17, 2018, from http://billsynnotandassociates.com.au/images/stories/documents/data_scientist.pdf

Dhar, B. V. (2013). Data Science and Prediction. *Communications of the ACM, 56(12), 64–73.* http://doi.org/10.1145/2500499

FAO. (2016). *Programa mundial del censo agropecuario 2020. Volumen 1. Programa, definiciones y conceptos.* Retrieved from http://www.fao.org/3/a-i4913s.pdf

Flores, I. (1981). *Data base architecture*. New York.: Van Nostrand Reihold Company.

García-Marco, F.-J. (2011). La Pirámide de la Información Revisitada: Enriqueciendo el Modelo Desde la Ciencia Cognitiva. *El Profesional de La Informacion*, 20(1), 11–24. http://doi.org/10.3145/epi.2011.ene.02

Gartner. (2012). Analytic Ascendancy Model. Retrieved December 12, 2017, from http://www.growwithfarm.com/evolving-analytics-from-descriptive-to-prescriptive/

Guillén, M. A., López Ayuso, B., Paniagua, E., & Cadenas, J. M. (2015). Una revisión de la Cadena Datos-Información-Conocimiento desde el Pragmatismo de Peirce. *Documentación de Las Ciencias de La Información*, *38*(Dic), 153–177. http://doi.org/10.5209/rev_DCIN.2015.v38.50814

Hernández Mendo, A., Castellano, J., Camerino, O., Jonsson, G., Villaseñor, Á., Lopes, A., & Anguera, M. T. (2014). Programas informáticos de registro , control de calidad del dato , y análisis de datos. *Psicología Del Deporte*, *23*(1), 111–121.

Jifa, G., & Lingling, Z. (2014). Data, DIKW, Big Data and Data Science. *Procedia Computer Science*, *31*, 814–821. http://doi.org/10.1016/j.procs.2014.05.332

Kamilaris, A., Kartakoullis, A., & Prenafeta-boldú, F. X. (2017). A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture*, *143*(January), 23–37. http://doi.org/10.1016/j.compag.2017.09.037

Khoso, M. (2016). How Much Data is Produced Every Day? Retrieved November 30, 2017, from http://www.northeastern.edu/level-blog/2016/05/13/how-much-data-produced-every-day/

Larson, D., & Chang, V. (2016). International Journal of Information Management A review and future direction of agile , business intelligence , analytics and data science. *International Journal of Infor-*

*mation Management*, 36(5), 700–710. http://doi.org/10.1016/j.ijinfomgt.2016.04.013

Leading Edge. (2015). Data science: the new monetization model for analytics industry. Retrieved December 4, 2017, from http://www.leadingedgeprovider.com/2016/07/data-science-the-new-monetization-model-for-analytics-industry/

Loury, J. (2014). Evolving Analytics: From Descriptive to Prescriptive. Retrieved December 11, 2017, from http://www.growwithfarm.com/evolving-analytics-from-descriptive-to-prescriptive/

Mazon-Olivo, B., Rivas, W., Pinta, M., Mosquera, A., Astudillo, L., & Gallegos, H. (2017). Dashboard para el soporte de decisiones en una empresa del sector minero. *Conference Proceedings - Universidad Técnica de Machala*, *1*, 1218–1229. Retrieved from http://investigacion.utmachala.edu.ec/proceedings/index.php/utmach/article/view/219/191

Molina-Solana, M., Ros, M., Dolores Ruiz, M., Gomez-Romero, J., & Martin-Bautista, M. J. (2017). Data science for building energy management: A review. *Renewable & Sustainable Energy Reviews*, *70*(December 2016), 598–609. http://doi.org/10.1016/j.rser.2016.11.132

National Academi of Science. (2017). Overview of Data Science Methods. In *Strengthening Data Science Methods for Department of Defense Personnel and Readiness Missions* (pp. 53–79). The National Academies Press. http://doi.org/10.17226/23670

Naur, P. (1974). *Concise Survey of Computer Methods*. Lund: Studentlitteratur.

NIST. (2015). NIST Special Publication 1500-1 NIST. Big Data Interoperability Framework : Volume 1 , Definitions. *National Institute of Standards and Technology*, *1*, 32. http://doi.org/10.6028/NIST.SP.1500-1

Press, G. (2013). A Very Short History Of Data Science. Retrieved October 10, 2017, from https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/

Provost, F., & Fawcett, T. (2013). *Data Science for Business. What you need to know about Data Mining and Data-Analytic thinking*. O'Reilly Media.

Roiger, R. (2017). *Data Mining: A Tutorial-Based Primer (Segunda Ed)*. United States of America: CRC Press.

Sivarajah, U. et al. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, *70*, 263–286. http://doi.org/10.1016/J.JBUSRES.2016.08.001

Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M. (2017). Big Data in Smart Farming – A review. *Agricultural Systems*, *153*, 69–80. http://doi.org/10.1016/j.agsy.2017.01.023

# 02 Capítulo Obtención de datos en sistemas agropecuarios

Salomón Barrezueta Unda ; Diego Villaseñor Ortiz

Desde mediados del siglo XX, con la aparición de las computadoras y con ello los programas informáticos, el tratamiento de la información sufrió una transformación que marcó el sendero de la agricultura moderna, tanto a nivel teórico como tecnológico. En este aspecto los sistemas agropecuarios, como parte del sector primario dependen de la disponibilidad y procesamiento de datos y de saber gestionar de forma oportuna la información.

La forma de ver y analizar los datos está en constante evolución, el modelo tradicional de proyectar la producción en función de supuestos, cambió y cada día toma más importancia la información primaria que se genera en los sistemas

**Salomón Barrezueta Unda:** Profesor titular de la Universidad Técnica de Machala (UTMach), es Ingeniero Agrónomo y Master en Gerencias y Administración de Empresas Agropecuarias por la UTMach. Tiene el grado de Doctor en investigación agraria y forestal obtenido en la Universidade da Coruña (España).

**Diego Villaseñor Ortiz:** Profesor Titular de la Universidad Técnica de Machala (UTMach), es Ingeniero Agrónomo, con Maestría en Ciencias Agronómicas con mención en suelos, obtenida en la Universidad de Concepción (Chile). Actualmente es parte del programa de doctorado en Ciencias del suelo y nutrición de plantas en la Universidad Estadual Paulista (Brasil).

agropecuarios para realizar sus propias proyecciones y toma de decisiones.

Sin embargo, hay otros tipos de información de gran relevancia que se relaciona con la adopción de la tecnología agropecuaria y cuya gestión no debe descuidarse como lo recomienda Palmieri & Rivas, (2007), y cita varios factores:

- El proceso de producción y sus actores (agricultores e intermediarios, etc.)
- Las dimensiones económicas, sociales y ambientales
- El precio y mercado
- El aprovechamiento del conocimiento tradicional
- La movilización de información interna.

## Importancia del uso de la información en sistemas agropecuarios

El proceso de modernización agrario inicia en la década de los 60 y se acelera con las exigencias de producción de los años 90, a partir de este momento las naciones buscan la seguridad alimentaria de su población en el marco de una agricultura sostenible (WCED, 1987), lo que generó nuevas estructurantes de conductas y modelos de trabajo en la sociedad rural (Chaparro, 2014; Urcola, 2012).

Es a partir de ese momento cuando aumentó de forma exponencial la cantidad de información accesible, que es potencialmente importante para la producción agropecuaria, más aún con la introducción de las Tecnologías de la información y comunicación (TICs), que han afectado la forma como se trabaja en las organizaciones dedicadas a la producción, investigación o innovación agropecuaria, generando una amplia gama de nuevas aplicaciones y también de complejos desafíos.

Cuando se menciona la relación TICs y sector agropecuario, es conveniente tener en cuenta que se habla de la aplicación de tecnologías en los procesos primarios, por lo que se debe realizar un estudio económico previo detallado de los posi-

bles efectos de la utilización masiva de las TICs, los cuales deben medirse a partir del impacto económico y social en cada sistema agrario (Albornoz, 2006).

Para Basso *et al.*, (2013) la agricultura del siglo XXI deberá desarrollarse en ambiente donde la correcta interpretación de los datos es el factor del éxito o el fracaso de los agricultores. En este contexto se advierte de un nuevo "paradigma agrario" (Urcola, 2012), donde la tecnología juega un papel fundamental en la toma de decisiones en los sistemas agrarios.

En este contexto, la generación de la información depende de los métodos para la toma de datos, la frecuencia, de la capacidad instalada en infraestructura y tecnología, que permitan aplicar la información en beneficio del agricultor. Un ejemplo es el registro del rendimiento de banano por lotes, una disminución de este indicar puede relacionar la una baja fertilidad del suelo y a inadecuada frecuencia de riego, que incide de directamente en el rendimiento.

En el Cuadro 2.1 se presentan algunas necesidades del productor y técnicos agropecuarios y cómo puede aportar la generación de información, en la toma de decisiones.

Cuadro 1.3: Herramientas para el científicos de datos

| Necesidades | Aporte de la información |
|---|---|
| Disponer de la multitud de datos ecológicos, biológicos, tecnológicos y económicos que representan a un agrosistema | Bases de datos organizadas por áreas o campos |
| Integración en un único marco conceptual que los formalice y relacione | Modelos conceptuales, de datos y matemático |
| Procesar según las leyes y metodologías de las disciplinas agrarias (agronomía, zootecnia, silvicultura, etc.) | Modelos de simulación agropecuarias que tratan cada uno de los aspectos de los agrosistemas y sistemas de información |
| Seleccionar las mejores alternativas de manejo, organización o comercialización, a partir de criterios productivos, económicos y ecológicos | Sistemas de soporte de decisiones |
| Transmitir la información en tiempo y forma adecuadas | Ofimática y telemática Agromática |

## La generación de información como soporte de decisiones en los sistemas agropecuarios

Los retos que enfrentan las naciones en vías de desarrollo con una alta dependencia de la agricultura, son: procesar la información que generen los sistemas agropecuarios con el fin de estimar el nivel de producción en un período de sequía, calcular el mínimo uso del recurso hídrico para evitar pérdidas por escorrentías y acelerar procesos de erosión de suelo o modelar las dosis de alimentos para el ganado con el objetivo de minimizar el desperdicio.

En Latinoamérica, cada día el uso de los datos agropecuarios para el desarrollo de aplicaciones informáticas se está extendiendo, no solo para el beneficio del agricultor que la genera, sino de toda una cadena de valor, que permite llevar un control seguro de los registros, con soluciones en servidores web para facilitar el uso en diferentes sistemas operativos; así mismo, su utilización no requiere de amplios conocimientos de informática, como es el caso del Sistema de Control de Recursos Filogenéticos (SISCORFI), una aplicación web que registra la ubicación de los bancos de semilla y disponibilidad de variedades vegetales mejoradas programa desarrollado en Cuba con el objetivo de proveer material genético certificado y eliminar posible introducción o comercialización de plantas portadoras de plagas o de baja productividad (Coronado-Hernández, 2015).

Con una correcta gestión de la información y adecuada aplicación de la tecnología en la agricultura es posible tener conocimiento de ofertas, demandas, ubicación de mercados y precios. Como es el caso desarrollado en la Pampa Argentina, con la incorporación de TICs, las cuales facilitan el contacto directo con los compradores, así mismo las fuentes de conocimiento se hacen accesibles para tomar decisiones (Coronado-Hernández, 2015).

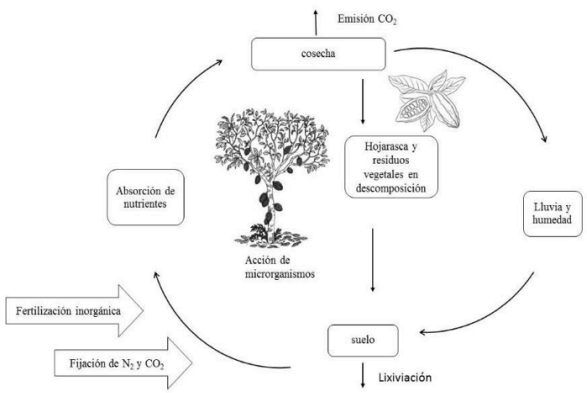
La gestión de información que se genera en los sistemas agrarios está enfocada en capturar, clasificar, preservar, recuperar, compartir y difundirla para quienes la reciban mejoren sus procesos productivos y administrativos.

## Variables en sistemas agropecuarios

La producción agropecuaria comprende muchas variables como los volúmenes de producción, tanto de carne en canal y de animales sacrificados, así como de productos vegetales específicos, precios, pesos, coeficientes técnicos e inventarios en general. Se desprenden de estas variables un conjunto enorme de datos que los agricultores o los organismos del estado no le dan el uso adecuado; como, por ejemplo, las ocurrencias de brotes epidemiológico en animales menores en determinados meses, lo que permite al estado ejecutar campañas de vacunación o erradicación de la enfermedad, o regular el precio interno de productos vegetales como el banano, en época de mayor oferta.

En la imagen 2.1 se presenta el sistema agrario del cacao, donde se distinguen varios elementos que son susceptibles a variaciones externas como el factor clima, población de microorganismos descomponedores o la intensidad de las labores agropecuarias. Por otro lado, la planta responde con una mayor o menor exigencia de nutrientes que incide en la cantidad de biomasa aérea que aporta la planta. Esto ocasiona que los datos generados tengan alto grado de variación estacional que inciden en la producción si no son interpretados de la forma adecuada.

Imagen 2.1. Ciclo nutricional del cacao y su interacción con el medio




Fuente: Barrezueta-Unda & Paz-González (2017)

En este aspecto el primer paso para ordenar los datos es clasificar el sistema agropecuario, el cual está en dependencia del uso de insumos, grado de conservación de la biodiversidad o del grado de tecnificación, factores que originan varios modelos de producción como la del cacao (Cerda *et al*., 2014) que se dividen en: agroforestal, asociado con sombra, asociado sin sombra temporal, monocultivo con sombra temporal o monocultivo sin sombra. Con un panorama claro del modelo y sistema agrario implantado, los datos se ordenan en función de lo detallado, debido a que la información que genere un monocultivo no es igual a un cultivo asociado.

Ahora, con la información disponible, es necesario que el técnico formule recomendaciones y para que los agricultores las adopten se debe conocer tanto el elemento humano involucrado en el cultivo (trabajadores agrarios) como los elementos biofísicos (suelo, aire, agua), las metas de los productores (umbral de producción) y de las restricciones que ellos enfrentan para lograr esas metas.

## Identificación y medición de los insumos variables

En muchos casos, en especial los pequeños agricultores, no registran todos los procesos que realizan en sus fincas; o, en caso contrario, llevan los registros de manera tradicional, es decir, en cuadernos y archivos físicos. Pons-Pérez *et al*., (2016) recomienda subir la información en repositorios o almacenar en la nube para proteger la información. Pero para que este proceso tenga efecto es necesario capacitar a los agricultores en la forma y detalle de registrar información de cada proceso sea estos por parcelas, lotes, cuarteles, etc. (Coronado-Hernández, 2015).

Para identificar qué variables proporcionan información pertinente en los sistemas agrarios, los técnicos agropecuarios y agricultores se deben familiarizar con las prácticas locales, realizando preguntas que pueden influir en la producción y en el resultado final. A continuación se da una lista de referencias que se puede considerar previo a la obtención de datos:

Preparación del terreno: ¿Es el mismo para todas las parcelas?

Siembra: ¿Se usa la misma semilla en todas las parcelas?; ¿Se emplea la misma cantidad de semilla?; ¿Es la misma la técnica de siembra?

Deshierbes/labranzas: ¿La cantidad de tiempo requerida para esta operación diferirá de una parcela a otra?; ¿Es la misma la técnica para todas las parcelas?

Selección de material germoplasma: ¿Se requiere para todos los tratamientos?; ¿Qué método utiliza?

Aplicación de pesticidas y fertilizante: ¿Son iguales para todos los lotes? Si las prácticas agronómicas no son idénticas para todos las parcelas, lotes o tratamientos hay que considerar cuáles de los siguientes tipos de insumos podrían ser afectados por las diferencias, y en qué magnitud.

Agroquímicos: ¿Qué fertilizantes, insecticidas, herbicidas, difieren en tipo o cantidad?

Semilla: ¿Difieren en tipo o cantidad?

Equipo: ¿Se necesita el mismo tipo de equipo?; ¿La calibración del equipo es la misma?; ¿Se necesita la misma cantidad de tiempo de operación del equipo?

Mano de obra: ¿Cuánto difiere la mano de obra debido a distintas operaciones de deshierbe, deshije, riego, densidad de siembra, preparación del terreno, etc.?; ¿Varía significativamente la mano de obra requerida con el tipo o cantidad de semilla o el fertilizante aplicado?; ¿Difiere entre tratamientos el tipo de mano de obra requerida?

## Conformación de un marco de trabajo para el registro de información

Según SAGARPA, (2013) para la obtención de datos de cada sistema agropecuario es necesario definir un marco de trabajo que contenga los siguientes criterios

- Población en estudio: los instrumentos de captación contemplan el registro de datos a partir de la identificación de las zonas productoras y de los agentes que interactúan en ellas, el tipo de cultivo y la variedad.
- Cobertura: se debe considerar la superficie, si el registro de datos es total o parcial y qué cultivos o actividad agropecuaria se incluye y excluye.
- Referencia temporal: la información debe estar referida a lapsos de tiempo que estén acordes al sistema agropecuario, como ejemplo el banano que se registra en función de cintas de colores que se coloca en los racimos para estimar el tiempo de cosecha (Imagen 2.2).

Imagen 2.2. Sistemas de colores para el control de las cintas en banano

| Semana | Color cinta | |
|---|---|---|
| 1 | | rojo |
| 2 | | marrón |
| 3 | | negro |
| 4 | | verde |
| 5 | | azul |
| 6 | | blanco |
| 7 | | amarillo |
| 8 | | lila o morado |

Fuente: Torres (2012)

- Revisión y validación de datos: Se debe analizar de manera recurrente la información en lapsos de tiempo que estén relacionados con las etapas fenológicas del cultivo o el desarrollo de los animales; si fuera el caso, considerando que todo el año se realizan actividades agropecuarias.
- Toma de decisiones con los datos obtenidos: Con la infor-

mación captada, el siguiente paso es la tabulación y procesamiento estadísticos de los datos para la toma de decisiones. Algunos autores recomiendan la difusión de resultados con el objetivo de conformar redes de información agropecuarias para su validación y desarrollo de la investigación.

## Procedimientos de muestreo en sistemas agropecuarios

A continuación, se detallan varios procedimientos para captar información en campo y centros de procesamiento de productos agrarios, con el fin de identificar variables que se puedan estructurar en indicadores de un sistema agropecuario.

Antes de definir qué se debe muestrear, es necesario establecer las técnicas de recolección de datos propuestas por USAID/OFDA, (2008):

Vuelos de reconocimiento a baja altura (Aviones o helicópteros): Se realiza desde aviones, helicópteros o globos aerostáticos lo que permite una rápida cobertura de la zona en estudio, identificar daños de plagas, usos del suelo y vías de acceso. La desventaja es el alto costo y la baja disponibilidad.

Evaluación terrestre: el desplazamiento por superficie, permite la apreciación cualitativa y cuantitativa de los daños y la toma de muestras. Su Desventaja es al cubrir zonas geográficas de difícil acceso.

Encuestas por muestreo sobre el terreno: Se utilizan técnicas de muestreo para la cuantificación de datos específicos de producción y daño, midiéndose a partir de submuestras. La modalidad de obtener la información es mediante la entrevista y encuestas a personas directamente afectadas. La desventaja es que los resultados en muchos de los casos son aproximados.

Otras técnicas: aerofotografía, imágenes satelitales y sistemas de sensores remotos.

## Datos meteorológicos

El clima es medido a través de los datos meteorológicos, los cuales son obtenidos en las estaciones agrometeorológicas de primer o segundo orden. El tipo de datos meteorológicos (humedad, temperatura, precipitación, etc.) y la frecuencia (semestral, bianual, anual, etc.) de las lecturas es una cuestión definida por la aplicación de dichos datos, como ejemplo determinar los períodos de mayor pluviosidad para la programación del sistema de riego.

Para las aplicaciones a nivel de predio o de actividad, es suficiente contar con los registros diarios de la estación agrometeorológica más cercana a la empresa o finca, para datos con alta variabilidad estacional como la lluvia y la evapotranspiración lo más conveniente es contar con un pluviómetro y un tanque de evaporación para un registro in situ.

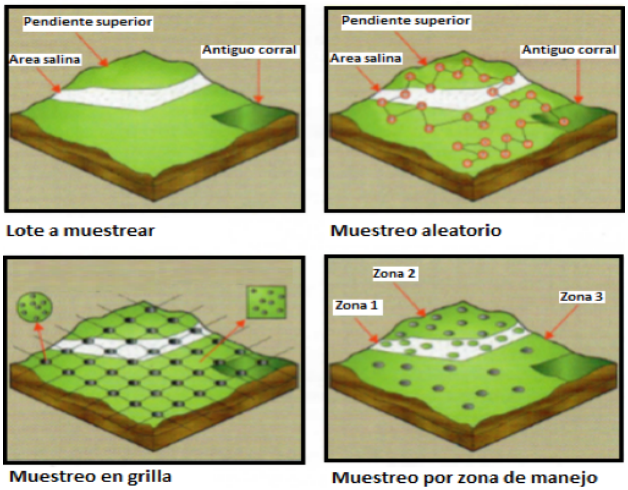
## Estructura de un plan de muestreo de suelos

La generación de la información del recurso suelo inicia con un análisis en profundidad de sus propiedades debido a que es el soporte del sistema de producción, lo cual incluye la planta y su manejo (Arévalo-Gardini *et al*., 2015). Por otra parte, para conocer el nivel adecuado de elementos disponibles en el suelo que la planta necesita, es necesario establecer un cronograma, donde se definan la frecuencia, tipo e intensidad del muestro de suelo.

El agricultor o técnico que realiza la interpretación de un análisis de suelos se enfrenta a una gran variedad de métodos que en algunos casos no son pertinentes al tipo de problemas productivo del cual necesita tomar decisiones. Se puede destacar, que la causa principal de errores en los análisis de suelos, es la forma como fue realizado el muestreo antes que los procedimientos analíticos de laboratorio. Una inadecuada estratificación del sitio donde se extrae la muestra puede incidir en que no tenga la suficiente representatividad de todo el terreno.

Para un buen muestreo se recomienda tomar muestras de suelo separadas de áreas de terreno con distinta topografía, geomorfología, tipo de suelo, vegetación natural o prácticas de manejo anteriores. Por lo tanto, una superficie extensa puede ser dividida en áreas que posean condiciones uniformes de manejo de suelo, o de acuerdo con la ubicación de cultivos anteriores (imagen 2.3). Se debe asignar un número de identificación permanente a cada área, además se recomienda mantener un mapa de las áreas muestreadas, registrando coordenadas con una unidad GPS y así preservar para futuras referencias.

Imagen 2.3. Esquema básico de muestreo de suelos




Fuente: Bruulsema *et al.*, (2013)

En la mayoría de los casos se deben tomar, al azar, por lo menos 15 a 20 submuestras para hacer una muestra compuesta mezclada que puede pesar de uno a varios kilos, la cual se homogeniza, para a continuación, tomar una porción de suelo, que es codificada y colocada en doble funda plástica para el traslado al laboratorio, tratando de evitar la contaminación de la muestra por manipulación inadecuada del operado (tocar con las manos), por lo que se recomienda el uso de guantes plásticos.

Para estudios agrológicos, se recomienda además de las consideraciones anteriores, tomar muestras representativas de los principales horizontes del perfil de suelo que serán utilizados por el sistema radicular de las plantas (Imagen 2.4). Para ello se sugiere realizar calicatas en los sectores específicos de la plantación, en función de la profundidad efectiva de las raíces. Esto determinará, por ejemplo la planificación del riego y la aplicación de fertilizantes que favorezcan el desarrollo de la planta, permitiendo que tanto las dosis de corrección (situaciones de déficit nutricional), como el cálculo de aporte potencial (situaciones de alta reserva nutricional), se realicen considerando también las características físicas de cada horizonte, principalmente densidad aparente y clase textural (Hirzel, 2008)

Imagen 2.4. Evaluación morfológica y muestreo de horizontes de suelos en una calicata representativa.

Fuente: Diego Villaseñor

Finalmente, se recomienda evitar el muestreo en los siguientes espacios del predio: a) al pie de cercas vivas o muertas, b) lugares de acumulación de material vegetal o estiércol, c) lotes de quemas recientes o acumulación de cenizas, d) terrenos que muestren señales de fertilización reciente, e) sitios cercanos a carreteras, guardarrayas o caminos vecinales, f) lugares cercanos a canales y g) sitios de pendientes pronunciadas o erosionadas.

Los tipos de muestra dependen del material y equipo a utilizar, en el caso del uso de la pala se debe cavar un hoyo en forma de "V" del ancho de una pala y la profundidad requerida según el cultivo (Imagen 2.5 a y b), tomar una tajada de suelo de 2 a 3 centímetros de espesor de la pared del hoyo (Imagen 2.5 c), con una navaja o cuchillo cortar las dos secciones laterales de la tajada de suelo colectado y eliminar (Imagen 2.5 e), depositar en un balde plástico limpio de impurezas como fertilizantes, cal, estiércol, cemento, etc. (Imagen 2.5 f), repetir esta operación para cada uno de los puntos elegidos como sitios de muestreo.

Imagen 2.5. Secuencia sugerida en la toma de muestras de suelos con pala


Fuente: Villalba (2012)

Otra forma de tomar muestras de suelo es con el uso de un barreno que se introduce de forma vertical al suelo y que se hace girar como si fuera un tornillo. El barreno tiene la capacidad de tomar la muestra en los 0 – 20 cm de suelo; obtenida la muestra se sigue igual procedimiento que la toma con la pala. Repetir esta operación para cada uno de los puntos elegidos como sitios de muestreo.

## Valores de referencia para la interpretación de análisis de suelos

Una vez que el proceso de muestreo concluye y los resultados son enviados al agricultor o al técnico de campo, es necesario la interpretación de los datos. Para esto, es necesario conocer los criterios de interpretación, los cuales se logran comparando valores de referencia con el tamaño de las partículas de la fracción mineral (textura). En el cuadro 2.2 se presentan dos suelos de distintas clases texturales los cuales se comparan con algunas propiedades químicas del suelo, lo que indica los rangos adecuados para el normal desarrollo de las plantas (Hirzel, 2008).

Cuadro 2.2 Propiedades químicas de suelo adecuadas para diferentes clases texturales.

| Propiedad | Unidad | Nivel óptimo de Clase Textural | |
|---|---|---|---|
| | | FA a FLA | FL a FY |
| Materia orgánica (M.O) | % | ≥1,5 | ≤1,5 |
| pH | | 6,2-7 | 5,8-6,8 |
| Conductividad eléctrica (C.E) | ds $m^{-1}$ | | ≤1,5 |
| Capacidad de intercambio catiónico (C.I.C) | cmol $kg^{-1}$ | 8-15 | 15-30 |
| Nitrógeno (N) Orgánico | mg $kg^{-1}$ | 15-30 | 20-40 |
| Fósforo (P) | mg $kg^{-1}$ | ≥15 | ≥20 |
| Potasio (K) Intercambiable | cmol $kg^{-1}$ | 0,3 - 0,5 | 0,4 - 0,6 |
| Calcio (Ca) Intercambiable | cmol $kg^{-1}$ | 7 - 10 | 8 - 12 |
| Magnesio (Mg) Intercambiable | cmol $kg^{-1}$ | 1 - 1,5 | 1,2 - 2 |
| Sodio (Na) Intercambiable | cmol $kg^{-1}$ | 0,03-0,3 | 0,05-0,6 |
| Suma de Bases | cmol $kg^{-1}$ | ≥8 | ≥10 |

| Propiedad | Unidad | Nivel óptimo de Clase Textural | |
|---|---|---|---|
| Relación Ca/CIC | % | 60 - 65 | 55 - 65 |
| Relación Mg/ CIC | % | 12 - 15 | 10 - 15 |
| Relación K/CIC | % | 2 - 3 | 3 - 4 |
| Azufre (S) | mg $kg^{-1}$ | ≥8 | ≥10 |
| Hierro (Fe) | mg $kg^{-1}$ | 2 - 4 | 2 - 10 |
| Manganeso (Mn) | mg $kg^{-1}$ | 1 - 2 | 2 – 5 |
| Zinc (Zn) | mg $kg^{-1}$ | 0,8 – 1,5 | 1 – 2 |
| Cobre (Cu) | mg $kg^{-1}$ | 0,5 - 1 | 0,5 – 1 |
| Boro (B) | mg $kg^{-1}$ | 0,8 – 1,5 | 1 - 2 |

Franco arenoso (FA), Franco limo arenoso (FLA), Franco limoso (FL), Franco arcilloso (FY)

Fuente: Adaptado de Pumisacho & Sherwood, (2002))

Uno de los problemas que tienen que resolver los técnicos agropecuarios son los valores de referencia a nivel local, para esto el apoyo de fuentes externas es indispensable para inferir sus resultados. En el cuadro 2.3 se presentan parámetros de referencia utilizados por el Instituto de Investigaciones Agropecuarias (INIAP) por niveles (alto, medio y bajo) para recomendar las enmiendas de fertilizantes. Se presentan estos parámetros en forma diferenciada, ya que en Ecuador, se carece de información general de suelos que distingan los niveles físico-químicos que podrían existir entre dos clases texturales distintas.

Cuadro 2.3. Propiedades químicas de suelo adecuadas según INIAP.

| Propiedad | Unidad | Rangos | | |
|---|---|---|---|---|
| | | Alto | Medio | Bajo |
| Materia orgánica (M.O) | % | < 3,0 | 3,0 – 5,0 | > 5,0 |
| Nitrógeno amoniacal (NH4) | mg $kg^{-1}$ | < 31,0 | 31,0 – 40,0 | > 40 |
| Fósforo (P) | mg $kg^{-1}$ | < 8,0 | 8,0 - 14 | > 14 |
| Zinc (Zn) | mg $kg^{-1}$ | < 3,1 | 3,1 – 7,0 | > 7,0 |
| Cobre (Cu) | mg $kg^{-1}$ | < 1,1 | 1,1 – 4,0 | > 4,0 |
| Hierro (Fe) | mg $kg^{-1}$ | < 20,0 | 20,0 – 40,0 | > 40,0 |
| Manganeso (Mn) | mg $kg^{-1}$ | < 5,1 | 5,1 – 15,0 | > 15,0 |
| Boro (B) | mg $kg^{-1}$ | < 0,20 | 0,20 – 0,49 | > 0,49 |
| Azufre (S) | mg $kg^{-1}$ | < 4,0 | 4,0 – 19,0 | > 19,0 |
| Cloro (Cl) | mg $kg^{-1}$ | < 17 | 17,0 – 32,9 | > 32,9 |
| Potasio (K) | cmol $kg^{-1}$ | < 0,2 | 0,20 – 0,38 | > 0,38 |
| Calcio (Ca) | cmol $kg^{-1}$ | < 5,1 | 5,1 – 8,9 | > 8,9 |
| Magnesio (Mg) | cmol $kg^{-1}$ | < 1,7 | 1,7 – 2,3 | > 2,3 |
| Sodio (Na) | cmol $kg^{-1}$ | < 0,5 | 0,5 – 1,0 | > 1,0 |
| Relación Aluminio (Al) + hidrógeno (H) | cmol $kg^{-1}$ | < 0,5 | 0,5 – 1,5 | > 1,5 |
| Aluminio (Al) | cmol $kg^{-1}$ | < 0,3 | 0,3 – 1,0 | > 1,0 |
| Conductividad Eléctrica (C.E) | ds $m^{-1}$ | < 2,0 | 2,0 – 4,0 | >4,0 – 8,0 |

cmol $kg^{-1}$ = miliequivalentes referidos al suelo en 100 cc de suelo en pasta saturada.
Fuente: Adaptado de Pumisacho & Sherwood, (2002)

La información generada de los análisis, también se puede enfocar en función de los umbrales máximos y mínimos de las plantas los cuales se contrastan con información publicada en textos científicos en los casos que no se disponga de información local o regional. En el cuadro 2.4 se describen los niveles nutricionales óptimos que necesita el cultivo del cacao para la región litoral del Ecuador; así como los métodos con los que se obtuvieron.

Cuadro 2.4. Valores óptimos de propiedades generales del suelo para el cultivo del cacao en Ecuador.

| Propiedades químicas | Óptimo |
|---|---|
| NH4 (mg $kg^{-1}$) –Kjeldahl- | >65 |
| P (mg $kg^{-1}$) –Olsen modificado/lectura espectrofotómetro de luz visible- | 12-25 |
| K (cmol $kg^{-1}$) –Olsen modificado/lectura espectrofotómetro de absorción atómica- | 0,3-1,2 |
| Ca (cmol $kg^{-1}$) -Olsen modificado/lectura espectrofotómetro de absorción atómica- | 4,0-18 |
| Mg (cmol $kg^{-1}$) -Olsen modificado/lectura espectrofotómetro de absorción atómica- | 0,9-4 |
| S (mg $kg^{-1}$) - Fosfato mono cálcico 0.008M/ Turbidimetría Ba CL2- | >22 |
| Zn (mg $kg^{-1}$) –Olsen modificado/lectura espectrofotómetro de luz visible- | 0,5-2,2 |
| Cu (mg $kg^{-1}$) –Olsen modificado/lectura espectrofotómetro de luz visible- | 1,8-5,9 |
| Fe (mg $kg^{-1}$) –Olsen modificado /lectura espectrofotómetro de luz visible- | 19-45 |
| Mn (mg $kg^{-1}$) –Olsen modificado/lectura espectrofotómetro de luz visible- | 3-6 |
| B (mg $kg^{-1}$) -CaH4(PO4)2.2H2O/lectura espectrofotómetro de luz visible- | 0,16-0,9 |
| Al (cmol $kg^{-1}$) -Cloruro de potasio 1N/lectura espectrofotómetro de luz visible- | 0,1-1,5 |
| ∑Bases (cmol $kg^{-1}$) | 15-30 |
| Ca/Mg (cmol $kg^{-1}$) | 2,6-8,0 |
| Mg/K (cmol $kg^{-1}$) | 7,5-15 |
| Ca+Mg/K (cmol $kg^{-1}$) | 27,5-55,0 |
| CIC (m $kg^{-1}$) –Acetato de amonio pH 7/lectura espectrofotómetro de luz visible- | 19,35 |
| CE (dS/m)-pasta de saturación/lectura Potenciómetro- | 2,00 |
| Materia orgánica (%) -Walkley y Black/Volumetria- | >3 |

| Propiedades químicas | Óptimo |
|---|---|
| Carbono orgánico (%) – en analizador elemental- | >2 |
| Nitrógeno total (%) –en analizador elemental- | >0,4 |
| C/N | >9,5 |
| Propiedades físicas | |
| pH –1:25 $H_2O$:suelo –potenciómetro- | 6,5-7 |
| Textura- Bouyouoc –Hidrómetro- | Franco; Franco-arenoso, Franco arcillo arenoso |
| Densidad aparente (Mg $m^3$) [7] | 1,25 |
| Profundidad suelo (metros) | 1,30-1,50 |

Fuente: Barrezueta-Unda & Paz-González (2017)

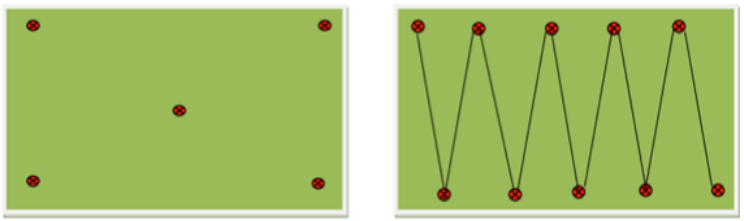
## Muestreo de productos agropecuarios en campo

Para establecer un método para la recolección de las muestras de material vegetal (frutos, hojas, raíces, etc.), se requiere considerar varios factores como el tipo del cultivo, la fuente de hídrica para riego, gradiente topográfica, textura de suelo, porcentaje de humedad ambiental, heliofanía, dirección del viento, fauna doméstica y silvestre, barreras naturales, barreras artificiales, colindancia de la parcela con zonas industriales, urbanas o rurales, potreros o establos y de la intensidad del ataque de plagas (SAGARPA, 2011).

SAGARPA, (2011) y Zaccagnini, (2015) recomiendan que el método de muestreo apropiado para realizar inferencias generalizadas de la población en superficies iguales o menores a 10 ha de plantas, es fijar cinco puntos (muestreo en aros), cuando se conoce la forma de la finca (Imagen 2.6 a). Cuando se desconocen las condiciones y/o características en que se realiza la producción, o cuando la unidad de producción tiene antecedentes de lotes con ataques severos de patógenos, es necesario realizar un recorrido en forma de zig-zag o en W (muestreo aleatorio simple sistemático) como se detalla en la imagen 2.6 b, con el propósito de abar-

car la totalidad de la finca o lote y que todas las unidades o elementos tengan la misma probabilidad se puedan incluir, con el fin de lograr una mayor representatividad y uniformidad de la unidades o elementos existentes dentro del área que se registra la información.

Imagen 2.6. Esquema de muestreo: A muestreo en aros, B muestreo aleatorio sistemático o tipo W.

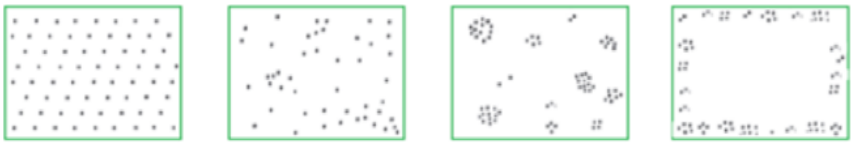


Fuente: SAGARPA, (2011)

## Monitoreo de plagas en campo

El monitoreo de plagas es un proceso en el cual influyen factores no controlados por el hombre como las condiciones meteorológicas, que modifican los hábitos y comportamiento de la plaga, por lo cual evaluar las características de su distribución en el cultivo la cual puede ser homogénea, al azar, agregada o periférica (Imagen 2.7). Es importante antes de establecer un plan de monitoreo que genere información confiable para la toma de decisiones para su control en campo (INTA, 2012).

Imagen 2.7. Tipo de distribución espacial de plagas en cultivos: homogénea, al azar, agregada y periférica



Fuente: INTA, (2012)

Un programa de monitoreo debe considerar un tamaño de muestra que refleje adecuadamente las densidades reales de plagas y sus enemigos naturales, presentes en el huerto.

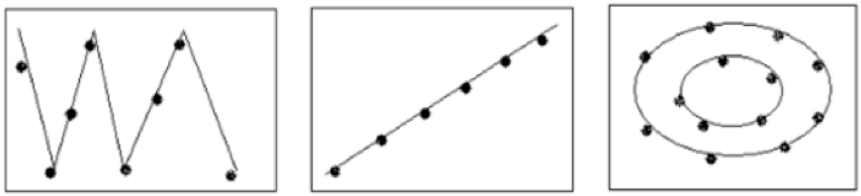
En general, mientras mayor es el tamaño de la muestra (mayor número de estructuras observadas), la estimación refleja mejor la densidad real de la plaga presente en el lote (Ripa & Larral, 2008).

Para llevar a cabo un proceso de monitoreo de plagas varios autores coinciden en los siguientes pasos:

- Registro de las características de la plaga encontrada.
- Observación y registro de factores, que modifican la densidad de las plagas, la susceptibilidad del cultivo y su capacidad de recuperación.
- Análisis de los datos obtenidos.
- Estimación de la tendencia de las poblaciones de las plagas.
- Toma de decisiones

En el caso de una distribución en el contorno del lote o finca, es conveniente aplicar el tratamiento solo en los bordes (Imagen 2.8). En una distribución agregada, se recomienda un tratamiento "tipo focos caliente" es decir controlar sólo el sector afectado a fin de evitar la desimanación de la plaga a la totalidad del lote.

Imagen 2.8. Recorrido de muestreo en cultivos permanentes y transitorios



Fuente: INTA, (2012)

En cuanto al recorrido de muestreo, existen varias opciones en tanto el mismo sea representativo de toda la parcela de observación. En general, se recomienda monitorear, al menos, el 1% de las plantas de un cuartel (con un mínimo de 10 plantas) y evaluar en terreno la efectividad de esta medición, aumentando la muestra en la medida que se detecte

variabilidad o carencia en la precisión. Por otra parte se debe considerar el costo y disponibilidad de personal capacitado (Ripa & Larral, 2008).

## Procedimiento de muestreo en área de postcosecha

El proceso de muestreo en área de empaques o acondicionamiento de los productos agropecuarios conocido como postcosecha se fundamenta en controles de calidad que tiene como finalidad prevenir la contaminación del producto de peligros físicos, químicos y biológicos, así como, evitar los daños de la mercancía por una mala manipulación.

La forma de muestreo y tamaño de muestra no es sólo el procedimiento de tomar un número determinado de muestras, su objetivo es suministrar información sobre la presencia o ausencia de microorganismos patógenos en los productos agrícolas, útiles para la aceptación o rechazo de dicho producto. Así, después del análisis de la muestra, se obtendrán resultados que se confrontan con determinados criterios, que permitan tomar acciones de control o prevención (SAGARPA, 2011).

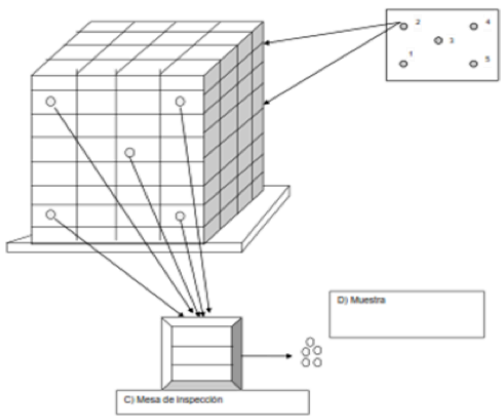
### Muestreo de patógeno en material y superficie de empaque

El procedimiento de muestreo en superficies de contacto y de los operarios dentro de instalaciones de empaque o de procesamiento en fresco de los vegetales obedecen a los siguientes criterios (SAGARPA, 2011):

- Lavar y sanitizar las manos antes de iniciar el muestreo.
- Usar guantes, cubre boca y bata durante todo el desarrollo del muestreo (realizar un cambio de guantes al tomar muestras diferentes o que provengan de otra parcela o lote, para evitar contaminación cruzada entre productos).
- Elegir una superficie que tenga contacto directo con el producto agropecuario

- En el caso de superficies vivas (manos de trabajadores), tomar un hisopo nuevo estéril
- Realizar un frotis en una superficie de 35 $cm^2$, recorriendo la superficie seleccionada en forma horizontal de izquierda a derecha con un lado de la cara de la esponja, alternado de otro recorrido horizontal de derecha a izquierda con el otro lado de la cara de la esponja, seguido de un frotis vertical (de arriba hacia abajo) alternando los recorridos con ambas caras de la esponja. Colocar la esponja en la bolsa correspondiente y cerrarla.
- Marcar y etiquetar la bolsa.
- Ubicar el sitio en donde se colectó
- Cuando el producto se encuentre en proceso de empaque o en bandas, la primera muestra se tomará en un momento determinado, el segundo muestreo transcurridos 30 minutos y un tercer muestreo 30 minutos después del segundo.
- Cuando el producto se encuentra empacado y en estiba, se tomarán muestras de varios puntos del contenedor con la finalidad de tener una muestra representativa (Imagen 2.9).

Imagen 2.9. Selección de muestras en carga consolidada




Fuente: FDA (2013)

Food and Drug Administration (FDA, 2003) recomienda (Cuadro 2.3) el tamaño de la muestra para productos agropecuarios utilizada para la detección de patagones (CAC/GL 33-1999), en paso de fronteras. Procedimiento recomendado en productos agropecuarios que son transportado en vías terrestres y marítima.

Cuadro 2.3. Número de muestras primarias por lote de producto agropecuario

| Peso (kg) del producto por lote | Número mínimo de muestra por kg |
|---|---|
| Productos envasados o al granel que no son homogéneos | |
| <50 | 3 |
| 50-500 | 5 |
| >500 | 10 |
| Número de unidad, cajas u otros recipientes del lote | Número mínimo de muestra por unidad |
| 1-25 | 1 |
| 26-100 | 5 |
| >100 | 10 |

Fuente: SAGARPA, (2013)

## Obtención de datos en empresas agropecuarias

Las empresas medianas y pequeñas del sector agropecuario carecen de un sistema que gestionen su información; como se detalló en los párrafos anteriores, todo sistema de producción genera datos que necesitan gestionarse.

La empresa agropecuaria representa un tipo definido de sistema social y económico y como tal posee ciertas características particulares derivadas especialmente de los subsistemas biológicos, de las tecnologías de producción específicas, del ecosistema en el cual se sustenta y de la identidad cultural del grupo social que vive y trabaja en ella.

El modelo de gestión de la información que se desarrolle para la empresa debe permitir la organización, almacenamiento, recuperación y procesamiento de los datos en todos los aspectos, lo que permite el diseño de un sistema agropecuario más ajustado a los objetivos del productor donde:

- La empresa tiene una ubicación geográfica, la cual define los suelos, clima, infraestructura regional, distancia a los mercados, etc.
- Las empresas suelen dividirse en departamentos, con funciones bien diferenciadas y limitadas, posibilitando la evaluación y diseño de cada uno por separado (administración, maquinarias y producción según tipo: ganadería, lechería, agricultura, silvicultura, forestación, horticultura, granja, etc.).
- Los intercambios (entradas y salidas) de los sistemas agropecuarios con el ambiente son de diferente naturaleza como: energía, materiales, dinero, información.

En este contexto, el flujo de información que generan las empresas agropecuarias debe ser bidireccional para que todos los departamentos estén integrados en el proceso de control y de toma de decisiones, como estar en conocimiento de los niveles de producción, flujo de ventas, productos rechazados por no conformidad y otros aspectos necesarios para implantar una cultura de gestión de la información.

## Referencia Bibliográfica


Albornoz, I. (2006). Software para el sector agropecuario. *Littec*, 1–39. Retrieved from http://www.littec.ungs.edu.ar/pdfespa?ol/DT 05-2006 Albornoz.pdf

Arévalo-Gardini, E., Canto, M., Alegre, J., Loli, O., Julca, A., & Baligar, V. (2015). Changes in soil physical and chemical properties in long term improved natural and traditional agroforestry management systems of cacao genotypes in peruvian amazon. *PLOS ONE*, *10*(7), e0132147. https://doi.org/10.1371/journal.pone.0132147

Barrezueta-Unda, S., & Paz-González, A. (2017). Estudio comparativo de la estructura elemental de materia orgánica de suelo y mantillo cultivados de cacao en El Oro, Ecuador. Revista Agroecosistemas, (3), 2–9

Basso, L. R., Pascale Medina, C., de Obschatko, E. S., & Preciado Patiño, J. (2013). *Agricultura inteligente: la iniciativa de la Argentina para la sustentabilidad en la producción de alimentos y energía*. (Ministerio de Agricultura, Ed.). Buenos Aires: IICA.

Bruulsema, T., P. Fixen, and G. Sulewski. 2013. "4R de La Nutrición de Las Plantas". *IPNI*, *Norcross-Estados Unidos*.

Chaparro, A. M. (2014, October 21). *Sostenibilidad de los sistemas de producción campesina en el proceso de mercados campesinos (Colombia)*. Universidad de Córdoba, Servicio de Publicaciones. Retrieved from http://www.tesisenred.net/handle/10803/283272

Cerda, R., Deheuvels, O., Calvache, D., Niehaus, L., Saenz, Y., Kent, J., … Somarriba, E. (2014). Contribution of cocoa agroforestry systems to family income and domestic consumption: looking toward intensification. *Agroforestry Systems*, *88*(6), 957–981. https://doi.org/10.1007/s10457-014-9691-8

Coronado-Hernández, H. (2015). Sistema de información para el control de procesos en la producción, poscosecha y análisis sensorial de café especial. *Revista Nova*, *1*(1), 1–8.

Cline, Marlin G. 1944. "Principles of soil sampling." *Soil Science 58* (4). journals.lww.com: 275.

Hirzel, J. 2008. "Diagnóstico Nutricional Y Principios de Fertilización En Frutales Y Vides." Colección Libros INIA-24. ISSN.

INTA. (2012). Monitoreo de plagas. In *Aplicación eficiente de fitosanitarios* (pp. 1–16). Buenos Aires, Argentina: Ediciones Instituto Nacional de Tecnologia Agropecuaria.

Krüger, H. (2006). Recursos naturales y medioambiente. Sostenibilidad del desarrollo agrario, 1–13.

Palmieri, V., & Rivas, L. (2007). Gestión de información para la innovación tecnológica agropecuaria. *COMUNIICA*, *3*(2), 17–26. Retrieved from http://infoandina.mtnforum.org/sites/default/files/publication/files/GIinnovacion07.pdf

Pons-Pérez, C., Molina-Concepción, O., Ruiz-Martínez, L., Medero-Vega, V., Sánchez-Socarras, P., & Rojan-Mirón, R. (2016). Las TIC herramientas para contribuir a la extensión agrícola y la innovación rural. *Revista Agricultura Tropical*, *2*(1), 77–83.

Pumisacho, M., & S. Sherwood. 2002. *El Cultivo de La Papa en Ecuador*. Instituto Nacional Autónomo de Investigaciones Agropecuarias.

Ripa, P., & Larral, R. (2008). Monitoreo de plagas y registros. In *Manejo de plagas en paltos y cítricos* (pp. 51–60). Santiago de Chile, Chile: SACH.

SAGARPA. (2011). *Manual técnico de muestreo de productos agrícolas y fuentes de agua para la detección de organismos patógenos*. México: SENASICA.

SAGARPA. (2013). *Diseño Conceptual de la Generación de Información Agropecuaria*. Mexico: SAGRAPA. Retrieved from http://infosiap.siap.gob.mx/opt/estadistica/normatividad/sistema/nsagarpa-siap-verde.pdf

Torres, S. (2012). *Guía práctica para el manejo de banano orgánico en el valle del Chira*. Piura, Peru: Swisscontact.

UNL. (1982). Análisis de los datos del sistema de información. In *Agromática* (pp. 1–10). Santa Fe, Argentina: UNL.

Urcola, M. (2012). Articulación de las "TIC" en el sector agrícola pampeano: la apropiación de la telefonía celular, las computadoras e internet entre los productores de una localidad del sur santafesino. *Revista Temas Y Debates*, *23*(1), 73–100.

USAID/OFDA. (2008). Evaluación de Daños y Análisis de Necesidades.

Sadzawka R., A, M.A. Carrasco R., R. Grez Z., M.L. Mora G., H. Flores P. & A. Neaman. 2006. Métodos de análisis de suelos recomendados para los suelos de Chile. Revisión 2006. Instituto de Investigaciones Agropecuarias, Serie Actas INIA N° 34, Santiago, Chile, 164 p.

Villalba, R. C. (2012). Toma de muestras de suelo. Retrieved May 2, 2018, from http://articulacionfeyalegriasenaroberto.blogspot.com/2012/10/clase-toma-de-muestra-de-suelos.html

Vidal, I. 2007. Fertirrigación, cultivos y frutales. Facultad de Agronomía, *Universidad de Concepción. Chillán, Chile.* 118 pp.

WCED. (1987). The Brundtland report: '*Our common future.*' Oxford University Press.

Zaccagnini, M. (2015). *Manual de Buenas prácticas para la conservación del suelo, la biodiversidad y sus servicios ecosistémicos.* (M. Zaccagnini, M. Wilson, & J. Oszust, Eds.) (1a ed.). Buenos Aires: PNUD-INTA. http://doi.org/10.13140/RG.2.1.1652.0724

# 03 Capítulo Internet de las cosas (IoT)

Dixys Hernández Rojas; Bertha Mazon-Olivo;
Carlos Escudero

El Internet de las Cosas (IoT) constituye un avance importante para la sociedad. Millones de usuarios, hombres y máquinas, participan a nivel mundial activamente en Internet tanto en su vida laboral como en la social y gracias a las tecnologías inalámbricas disponibles, han ampliado sus posibilidades de interacción en la red en cualquier lugar y momento. Por tanto, la tecnología sirve como herramienta de colaboración y toma de decisiones en un mundo en el que converge lo físico con lo digital.

**Dixys Hernández Rojas:** Ingeniero Electrónico y Máster en Electrónica por la U. Central Marta Abreu de Las Villas, Cuba. Docente e Investigador en algunas universidades de Cuba y Ecuador, Director / Ingeniero de proyectos en Grupo Arteche en México y en Goliath Consulting LLC, Irvine, USA. Actualmente es Profesor Titular y Director del Grupo de Investigación AutoMathTIC de la UTMACH. Sus intereses de investigación son IoT, WSN y desarrollo de Sistemas Embebidos. Cursa su doctorado en Universidade da Coruña, España. Cuenta con varias publicaciones.

**Bertha Mazon-Olivo:** Ingeniera en Sistemas y Magíster en Informática Aplicada por la Escuela Politécnica de Chimborazo. Profesora Titular en la Universidad Técnica de Machala. Es estudiante del programa doctoral en Tecnologías de la Información y las Comunicaciones en Universidade da Coruña, España. Sus líneas de investigación son: Internet de las Cosas, Ciencia de Datos y Desarrollo de Aplicaciones Informáticas. Cuenta con varias publicaciones indexadas.

**Carlos Escudero:** Máster de la Universidad de Vigo, España en 1991 y el Doctorado en Informática de la Universidad de La Coruña en 1998. Obtuvo dos becas para ser investigador antes y después de su doctorado en la Universidad Estatal de Ohio (1996 y 1998), durante 6 y 3 meses, respectivamente. Actualmente es Profesor Asociado (desde 2000) y Vicedecano del Gobierno de la Facultad de Informática de la Universidade da Coruña.

El IoT implica un escenario donde las “cosas”, típicamente dispositivos electrónicos inteligentes con sensores y actuadores distribuidos geográficamente, se encuentran identificados y conectados a Internet, que permiten el control y monitoreo remoto de situaciones críticas de un dominio de aplicación, incluso sin la interacción humana. Sin embargo, para poder detectar dichas situaciones es necesario comunicar, almacenar, analizar y procesar eficientemente la gran cantidad de información generada cada día. Una aplicación importante de IoT es la Agricultura Inteligente (Smart Agriculture) y se define como el uso de las Tecnologías de la Información y Comunicaciones en la gestión localizada de cultivos o parcelas agrícolas, basado en la existencia de variabilidad en campo, para aplicar el tratamiento adecuado en el momento justo.

Este capítulo tiene el propósito de acercar al lector hasta el punto de recolección de datos agropecuarios provenientes principalmente de sensores, presentes hoy en día en sistemas IoT. El mismo tiene un nivel básico e introductorio donde se abordan los fundamentos de IoT de forma global, evolución y definiciones ya establecidas por la comunidad científica y de forma específica el IoT en Agricultura de precisión. También serán tratadas las tecnologías más usadas y disponibles actualmente en este dominio de aplicación de la AGP, para continuar con los aspectos teóricos y prácticos involucrados en IoT para AGP y terminar con un recorrido por las arquitecturas y plataformas IoT actuales.

## Historia del internet de las cosas

El término Internet de las cosas es una extensión del ya conocido Internet, del cual hoy en día todos somos usuarios permanentes y lo conocemos como la gran autopista de la información, donde podemos encontrar prácticamente cualquier información que necesitamos gracias a la interconexión de millones de computadoras, bases de datos y usuarios alrededor de todo el mundo. Pero, ¿cuál fue su origen? ¿A quién le debemos este nombre?

Podríamos decir que la evolución del Internet que conocemos hoy en día comenzó en la década de los 70 a los 80, donde se crearon los primeros protocolos de comunicaciones que son la base de nuestro internet, por las principales universidades de Estado Unidos, como el Instituto Tecnológico de Massachusetts (MIT) y la Universidad de California, Los Ángeles UCLA), que convirtieron en los 90 a la red militar y académica ARPANET en el Internet actual.

Pero mucho antes, Nikola Tesla (1926) y Alan Turing (1950) anticiparon la conectividad global, la miniaturización tecnológica y la necesidad de inteligencia en sensores y equipos de comunicación, incluso en 1874 científicos franceses lograron transmitir información meteorológica desde la cima del monte Mont Blanc hasta París, constituyendo los primeros experimentos de telemetría hasta ahora registrados.

Continuando con la historia, con la aparición del internet la década de los 90 tuvo una gran actividad de desarrollo tecnológico creando cada vez más aplicaciones pensadas para internet. Fue así que en 1990 John Romkey, en el evento Interop, mostró al mundo el primer objeto (thing) conectado a Internet, una tostadora que podía ser encendida o apagada remotamente.

Luego en 1999 fue creado el centro Auto-ID dentro del MIT, por sus fundadores Sanjay Sarma, David Brock y Kevin Ashton, que lograron enlazar en Internet objetos a través de tarjetas RFID . Pero no fue hasta que el director ejecutivo de Auto-ID, Kevin Ashton en este mismo año y luego David L. Brock en el 2001, que acuñaron el término Internet de las Cosas en la historia del Internet.

## Definición de IoT

El concepto del Internet de las Cosas ha tenido múltiples definiciones desde 1999 hasta nuestros días, refiriéndose en sus inicios a solo cosas identificables vía RFID exclusivamente, añadiéndoles inteligencia y mayor ámbito.

Podemos decir que el Internet de las cosas actual sería el conjunto de objetos inteligentes, perfectamente auto-iden-

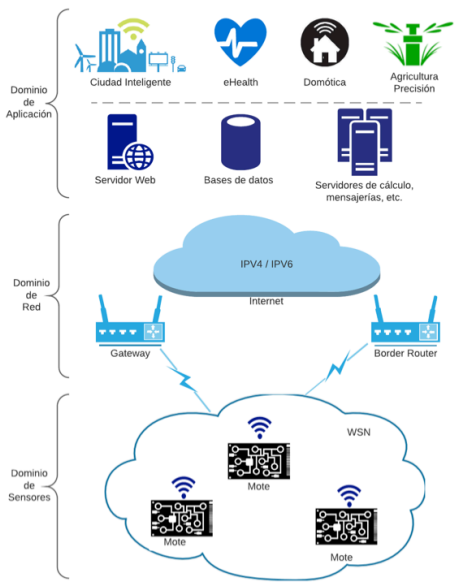
tificables, capaces de interactuar remotamente entre sí y con el resto de equipos conectados a través de Internet en tiempo real, incluso sin la interacción humana.

En el ámbito agropecuario, IoT lo se ve como el conjunto de sensores capaces de medir parámetros climáticos, suelo, agua, cultivos (tronco, fruto, clorofila, etc.) y otros, que envían la información a un servidor local o remoto (nube); proporcionando a los empresarios agropecuarios y clientes finales la posibilidad de monitorear su producción remotamente desde un terminal conectado a internet, sea este una PC, Tablet o smartphone.

## Arquitectura IoT

Actualmente existen varias arquitecturas de IoT y para este capítulo se ha propuesto la arquitectura de la Imagen 3.1 que ha sido adaptada de (Campoverde, Hernandez-Rojas, & Mazon-Olivo, 2015) y consta de 3 capas: Dominio de Aplicación, Dominio de Red y Dominio de sensores.

Imagen 3.1. Arquitectura del Internet de las Cosas




Fuente: Adaptado de (Campoverde *et al.*, 2015)

## Dominio de aplicación

En esta capa se encuentra la infraestructura de comunicación, almacenamiento y procesamiento de datos, así como las herramientas de análisis y presentación de la información al usuario. La infraestructura puede estar formada desde un servidor físico o virtualizado a un Centro de Procesamiento de Datos (CPD) complejo que involucra un conjunto de recursos físicos, lógicos y humanos para el control de los procesos y datos en el contexto de IoT. La virtualización de los recursos físicos y disponibilidad en internet se conoce como computación en la nube o Cloud Computing (Im, Kim, & Kim, 2013; Suciu *et al.*, 2015; Suciu, Halunga, Vulpe, & Suciu, 2013; Wang, Bi, & Xu, 2014). A continuación, se describen algunas de las principales funciones de esta capa:

- Recolección de datos crudos. El CPD se comunica con la capa Dominio de Red mediante el internet y usa un protocolo de comunicación para recolectar los datos crudos (Ali *et al.*, 2016; Gitzel, Turrin, & Maczey, 2015; Karkouch, Mousannif, Al Moatassime, & Noel, 2016). Existen varios protocolos de comunicación, por ejemplo: MQTT, CoAP, REST, XMPP, etc. (Al-Fuqaha, Guizani, Mohammadi, Aledhari, & Ayyash, 2015; Mijovic, Shehu, & Buratti, 2016). MQTT es muy popular por su bajo consumo de ancho de banda y bajo consumo de recursos.
- Pre-procesamiento y almacenamiento de datos. Consiste en la limpieza y transformación de datos para luego ser almacenados en sistemas gestores de bases de datos, y/o pasar a un sistema de cálculo o simplemente ser monitoreados y controlados en tiempo real (Cai, Xu, Jiang, & Vasilakos, 2016; Kambatla, Kollias, Kumar, & Grama, 2014; Moniruzzaman & Hossain, 2013; Wolfert, Ge, Verdouw, & Bogaardt, 2017).
- Monitoreo y control. Los datos de sensores de la WSN se presentan en un tablero de control (dashboard IoT) visual para que el usuario comprenda el estado actual de la zona o área que está vigilando. Un dashboard IoT además de monitorear sensores también puede controlar

actuadores de la WSN, por ejemplo, encender o apagar una bomba, abrir o cerrar una electroválvula.

- Aplicaciones o dominios IoT. Software con interfaz web o móvil que interactúa con el usuario y a su vez con los componentes IoT. Las aplicaciones IoT también se las conoce como Smart: ciudades inteligentes (Smart Cities), hogar y edificio inteligente (Smart Home and Building), cuidado y salud inteligente (Smart Healthcare), agricultura inteligente o Agricultura de Precisión (Smart Agriculture o Precision Agriculture), etc. (Botta, de Donato, Persico, & Pescapé, 2015; Shaikh, Zeadally, & Exposito, 2015; O Vermesan & Friess, 2014; Ovidiu Vermesan & Friess, 2015).
- Análisis de datos. Procesa los datos crudos obtenidos de la WSN y los combina con datos extraídos de los sistemas transaccionales para obtener información útil que ayude la toma de decisiones. La Estadística, Inteligencia de Negocios, Minería de Datos, Inteligencia Artificial, Machine Learning son algunas de las disciplinas aplicables para análisis de datos en el contexto de IoT.

El sistema que coordina y gestiona todos los componentes del dominio de aplicación de internet de las cosas se le conoce como Plataforma IoT.

## Dominio de red

Comprende componentes de pre-procesamiento y comunicación entre la Red de Sensores Inalámbrica (WSN) y la plataforma IoT. Los componentes IoT de esta capa son:

- Gateway o Micro data center. Es un dispositivo con características de un mini computador que además de coordinar la comunicación con la WSN y con la plataforma IoT, se encarga de obtener los datos crudos de los dispositivos IoT o motes, luego realizar un pre-procesamiento y almacenamiento temporal y seguidamente enviarlos a la plataforma IoT mediante un protocolo de comunicación.

- Red de comunicación con la WSN. Comprende la tecnología que hace posible la comunicación entre un mote y un Gateway IoT. Ejemplos de tecnologías WSN son: Zigbee, Bluetooth Low Energy (BLE), LoRa, Sigfox, etc., las cuales serán explicadas con más detalle en la sección de Tecnologías de Comunicación.
- Red de comunicación con la Plataforma IoT. Comprende la tecnología de comunicación del Gateway IoT con la plataforma IoT, normalmente es de LAN (Local Area Network), WAN (Wide Area Network) o MAN (Metropolitan Area Network); Estas tecnologías pueden ser: Wi-Fi, Ethernet, Wi-Max, LoRa-WAN, etc.

Algunos autores como (Aazam & Huh, 2015; Ai, Peng, & Zhang, 2017; Khan, Parkinson, & Qin, 2017) hablan de computación en la niebla (Fog Computing) y computación en el borde (Edge Computing), a continuación se explican estos nuevos conceptos:

- Fog Computing. Extensión de cloud computing y sus servicios al borde de la red; es decir, consiste en descentralizar de la cloud, el almacenamiento de datos, el procesamiento, los servicios y las aplicaciones para llevarlo a un entorno localizado. Esta responsabilidad por lo general es delegada a un gateway IoT de tipo Micro Data Center, transmitiendo a la cloud sólo los mensajes y datos de contexto global.
- Edge Computing. La computación de este tipo va más allá de ser localizada a un Gateway o micro data center, cada dispositivo de la red desempeña la función de procesar los datos más cercanos y de decidir qué datos debe enviar al dispositivo de nivel superior (Gateway IoT o Cloud).

## Dominio de sensores

En esta capa se ubican las redes de sensores inalámbricas (WSN) y los dispositivos (motes) IoT que integran transductores, sensores y actuadores. En las secciones subsiguientes se explican con más detalle.

## Transductores, sensores y actuadores

En el argot de los sistemas embebidos y la telemetría existen tres términos que tienden a ser confundidos, estos son: transductores, sensores y actuadores dada la similitud y naturaleza que los envuelve. Según la definición de diccionarios, un transductor es un dispositivo que transfiere un tipo de energía en otra diferente. Ejemplos comunes tenemos a los micrófonos, capaces de transformar sonido en impulsos eléctricos y en el caso contrario a los altavoces o parlantes que convierten impulsos eléctricos en sonido. También tenemos a los focos de luz incandescentes que emiten luz cuando pasa corriente por un filamento. Otro ejemplo sería el motor que convierte energía eléctrica en energía mecánica o de movimiento. ¿Conoce algún otro ejemplo?

Por otro lado tenemos a los sensores, que según el diccionario es un dispositivo que puede detectar cambios de estímulos físicos y lo convierten en una señal que puede ser medida o guardada. Ejemplo de sensores los tenemos en el cuerpo humano, que logran detectar luz, sonidos, cambios químicos, presión y temperatura. ¿Puede identificarlos? ¿Cuál será la señal de salida de estos sensores?

Una acotación importante en este momento para distinguir los conceptos y términos anteriores sería que un sensor puede ser usado por sí solo para medir algo, pero un transductor necesita además del elemento de sensado un circuito eléctrico asociado. Es decir, un transductor contiene un sensor y la mayoría de los sensores deben ser transductores.

El actuador es un dispositivo que conmuta una señal eléctrica o mueve algo y utiliza energía para lograr un movimiento o conmutación. Por tanto podemos decir que un actuador es un tipo específico de transductor. De los ejemplos anteriores de transductores, cuáles serían actuadores?

### Sensores

Los sensores típicamente convierten estímulos físicos en señales eléctricas analógicas o digitales y pueden ser clasi-

ficados de acuerdo al tipo de estímulo en: acústicos, eléctricos, magnéticos, ópticos, térmicos y mecánicos.

De acuerdo a la clasificación anterior, los tipos de sensores más comunes por la variable a medir son:

- Temperatura: Termistores, Termostatos, Termocuplas y RTD (Resistance Temperature Detectors)
- Sonido: Micrófonos e Hidrófonos
- Luz: LDR (Light Dependant Resistor), Fotodiodos, Fototransistores y Celdas solares
- Fuerza/Presión: Celdas extensiométricas (Strain Gauge), Interruptores de presión y celdas de carga
- Posición: Potenciómetros, LDVT (Linear Variable Differential Transformer), Reflectivos y Encoders
- Velocidad: Tacogeneradores y acelerómetros

## Actuadores

Los actuadores son dispositivos capaces de conseguir el movimiento de algo por medio de una energía o simplemente conmutar una corriente o un voltaje para que otro dispositivo pueda generar una acción en su entorno de un proceso dado. En función de esta energía los actuadores pueden ser clasificados en neumáticos, hidráulicos y eléctricos y en función del movimiento conseguido pueden ser lineales o rotatorios.

Por medio de los actuadores, un sistema automatizado puede abrir o cerrar una esclusa, activar o desactivar una electroválvula para dejar pasar agua, encender o apagar una bomba de agua, controlar el ángulo y altura de boquillas o dispensadores. Abrir escotillas de sembradoras, ajustar la cantidad de fertilizantes, dosificar el alimento de animales y muchas más aplicaciones. ¿Puede mencionar otras aplicaciones agropecuarias donde un actuador ayuda a automatizar el proceso?

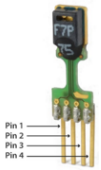
En el Cuadro 3.1, muestra algunos de los sensores y actuadores comerciales más usados en Agricultura de Precisión,

donde cada fila del cuadro responde a un tipo de sensor específico y las columnas a) y b) representan ejemplos o modelos representativos con la idea que constituya a su vez una guía inicial en la búsqueda del sensor idóneo para proyectos agropecuarios.

En la fila de sensores de temperatura tenemos en a) el sensor MCP9700A que es un sensor de temperatura, cuyo voltaje de salida es proporcional a la temperatura en un rango de -40°C (100 mV) hasta 125°C (1.75 V) con una sensibilidad de 10 mV/°C. Este sensor puede ser usado para conocer la temperatura ambiental. En b) tenemos un PT -1000, que es un sensor de temperatura para el suelo principalmente, con un rango de -50 a 300°C. El mismo constituye un sensor resistivo, donde la resistencia de salida varía entre 920 y 1200 ohms aproximadamente y a 0°C la resistencia es de 1 Kohm.

Cuadro 3.1: Transductores comerciales agrupados según la variable a medir

| Tipos de Sensores | a) | b) |
|---|---|---|
| Temperatura |  |  |
| Humedad del suelo |  |  |
| Humedad ambiental |  |  |

| Tipos de Sensores | a) | b) |
|---|---|---|
| Radiación |  |  |
| Dendrómetros |  |  |
| Sensores combinados |  |  |
| Actuadores para riego: electroválvulas |  |  |
| Actuadores electromagnéticos |  |  |
| Actuadores de movimiento |  |  |

En los sensores de humedad del suelo, en a) se muestra un sensor Watermark of Irrometer, con un rango de medición de 0 a 200 cb, donde el valor de resistencia de salida es proporcional a la tensión de agua en la tierra. En b) se muestra un Tensiómetro de Irrometer, que permite medir y leer directamente la humedad del suelo. Este sensor no es afectado por la salinidad del suelo, por ende no necesita calibraciones adicionales.

En la fila de sensores de humedad ambiental, tenemos en a) al sensor 808H5V5 cuyo voltaje de salida es proporcional a la humedad relativa en la atmósfera y el valor se entrega en porciento de humedad relativa (0 a 100 %RH). En b) se muestra un sensor de humedad condensada, Leaf Wetness Sensor (LWS). Aquí el voltaje de salida es inversamente proporcional a la humedad condensada en el sensor.

Otro parámetro importante en agricultura es la radiación solar, los sensores de este tipo se muestra en a) un sensor resistivo de luminosidad (LDR), donde la resistencia de salida es proporcional a la intensidad de luz que incide sobre la celda. En b) se encuentra el sensor SQ-110, utilizado para medir radiaciones solares y el voltaje de salida es proporcional a la intensidad de la luz, en el rango visible del espectro (410 a 655 nm).

Los dendrómetros son unos sensores interesantes que permiten medir el diámetro del tronco de una planta o del fruto para conseguir un seguimiento efectivo del cultivo. En a) tenemos al sensor resistivo Ecomatik DC2, con un rango de 0 a 20 Kohm, que mide el diámetro del tronco de una planta (Trunk Diameter Dendometer). En b) se muestra al Fruit Diameter dendometer (Ecomatik DF), presenta el mismo rango de resistencia de salida que el de tronco, pero se lo utiliza principalmente para medir el diámetro de los frutos.

En el Cuadro 3.1 también se ha incluido una fila para transductores que combinan más de un sensor como es el caso de a) que puede ser utilizado para una estación meteorológica, incluye un pluviómetro, anemómetro y dirección del viento. El anemómetro brinda una salida digital cuya fre-

cuencia es proporcional a la velocidad del viento. La dirección del viento (Wind vane) brinda una salida resistiva, donde la resistencia es proporcional a un ángulo, es decir brinda 16 posiciones que indicaría la dirección actual del viento. El pluviómetro, brinda una salida digital que se activa un interruptor cuando el nivel del agua ha llenado el recipient del sensor (0.28 mm) aproximadamente. En b) en cambio tenemos al sensor SHT75 de Sensirion que es un sensor digital de humedad y temperatura, con un rango de -40°C a +123.8 °C y la humedad de 0 a 100 %RH. La salida de este sensor es una palabra digital en formato $I^2C$[1].

Las tres últimas filas del Cuadro 3.1 contienen algunos actuadores usados en diversas aplicaciones, siendo una de ellas la agricultura de precisión. Siendo el control de riego una problemática de automatización en este dominio de aplicación. En esta fila en a) se muestra una típica electroválvula, utilizada principalmente para controlar el paso del agua a las zonas de riego y en b) se muestra la imagen de un interruptor de presión para riego.

Luego se ha incluido una fila de actuadores electromagnéticos que generalmente constituyen un paso intermedio de los actuadores de fuerza final, como son los contactores y relés mostrados en a) y b) respectivamente. Estos actuadores permiten a su vez accionar motores o bombas de agua, mostrado en a) y vástagos lineales como CAHB-20/21 mostrado en b) de la última fila.

## Wireless Sensor Network

Según la arquitectura IoT, analizada anteriormente en el epígrafe Arquitectura IOT, la base de esta arquitectura está formada por los transductores, sensores y actuadores que interactúan con el proceso según el dominio de aplicación analizado en el epígrafe anterior. Estos transductores nece-

[1] $I^2C$: Inter-Integrated Circuit. Es un bus de datos serial que permite interconectar circuitos integrados y partes de un circuito electrónico donde cada uno dispone de una dirección específica.

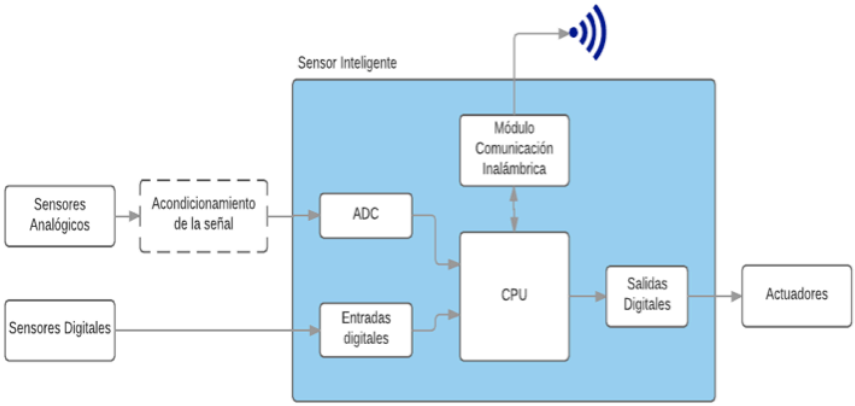
sitan de circuitos adicionales que permitan convertir las señales analógicas medidas en digital y a su vez poder transmitirlas hacia el servidor, para que un usuario en cualquier parte del mundo pueda observar. Estos circuitos han evolucionado junto con los avances de la electrónica y las telecomunicaciones. Actualmente los transductores son conectados a dispositivos pequeños, típicamente alimentados por baterías, con ciertas limitaciones de recursos, pero con la suficiente capacidad de procesar esa señal y transmitirla o recibir comandos remotos para los actuadores. Estos dispositivos son denominados "Transductores inteligentes" (smart sensors o smart transducers) y en el argot IoT son llamados también "motes".

En la mayoría de las aplicaciones IoT, existen muchos motes interconectados en una red, es decir una red de sensores. Actualmente el medio de comunicación más usado es el inalámbrico, por lo que estas redes reciben el nombre de Redes de Sensores Inalámbricos o inglés Wireless Sensors Network (WSN). En este epígrafe vamos a introducirnos en las WSN para conocer los principales smart transducers comerciales disponibles en el mercado, que pueden ser usados en proyectos agropecuarios. También revisaremos las principales tecnologías inalámbricas usadas en las WSN actuales.

Smart Transducers

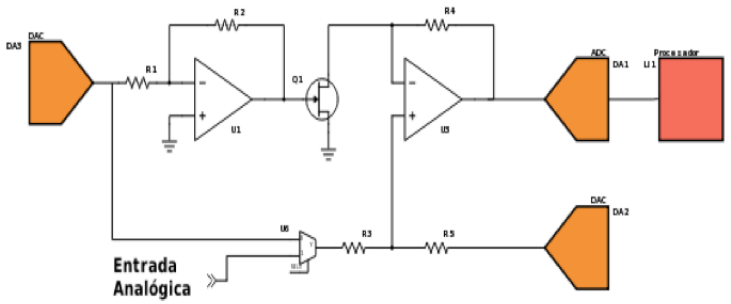
En la Imagen 3.2 se muestra el esquema general de un smart transducer. Donde podemos observar que a través de las entradas y salidas analógicas y digitales del procesador se conectan los sensores y actuadores. En un mote el procesador debe tener la capacidad de realizar un procesamiento primario de la información y encapsular el dato dentro del protocolo de comunicación que usa la WSN para comunicarse con el servidor o la cloud.

Imagen 3.2. Diagrama general de un Transductor Inteligente




El módulo de comunicaciones muchas veces (tendencia actual) aparece integrado junto con el procesador y el bloque de acondicionamiento de la señal es opcional en dependencia del sensor y la variable a medir. Este bloque es necesario porque hay que ajustar los niveles de la señal de salida del sensor con la entrada analógica del mote, ya que estas entradas analógicas pertenecen a un convertidor análogo - digital (ADC) encargado de digitalizar la variable medida para su posterior procesamiento y transmisión. Estos ADC por lo general tienen un rango de 0 a 5VDC de entrada y las salidas de los sensores manejan rangos de mV a su salida. Por tanto está señal analógica debe ser ajustada al rango de entrada del ADC para conseguir la mayor precisión posible. El acondicionamiento de las señales analógicas se caracteriza por el uso de amplificadores operacionales con configuraciones básicas, diferenciales o incluso con amplificadores de ganancia programable inteligentes como se muestra en la Imagen 3.3.

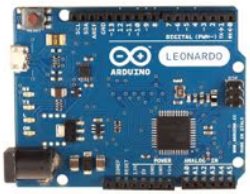
Imagen 3.3. Amplificador de ganancia programable.




Fuente: Adaptado de (Rodríguez Arias & Hernández Rojas, 1999)

En el Cuadro 3.2 se muestran algunos de los smart sensors Open Hardware[2] disponibles comercialmente.

Cuadro 3.2: Smart sensors comerciales

| Smart sensor | Imagen | Características |
|---|---|---|
| Arduino[3] |  | -Microcontroller: ATmega32u4<br>-Operating Voltage: 5V<br>-Input Voltage: 7-12V<br>-Digital I/O Pins: 20<br>-PWM Channels:7<br>-Analog Input Channels:12<br>-SRAM: 2.5 KB (ATmega32u4)<br>-La configuración de conectores y señales se han convertido en un estándar. Decenas de shield o tarjetas con aplicaciones específicas han adoptado este estándar y se denominan Arduino compatible. |

[2] Open Hardware: Hardware libre, significa que los diagramas de los circuitos y sus especificaciones son de acceso público y pueden ser replicados sin costos ni regalías.

[3] https://www.arduino.cc/

| Smart sensor | Imagen | Características |
|---|---|---|
| Raspberry Pi[4] |  | - Quad Core 1.2GHz Broadcom BCM2837 64bit CPU<br>- 1GB RAM<br>- BCM43438 wireless LAN and Bluetooth Low Energy (BLE) on board<br>- 40-pin extended GPIO<br>- 4 USB 2 ports<br>- 4 Pole stereo output and composite video port<br>- Full size HDMI<br>-CSI camera port for connecting a Raspberry Pi camera<br>- Micro SD port for loading your operating system and storing data<br>- Upgraded switched Micro USB power source up to 2.5A |
| Beaglebone[5] |  | - Processor: AM335x 1GHz ARM® Cortex-A8<br>- 512MB DDR3 RAM<br>- 4GB 8-bit eMMC on-board flash storage<br>- 3D graphics accelerator<br>- NEON floating-point accelerator<br>- 2x PRU 32-bit microcontrollers<br>- ConnectivityUSB client for power & communication, USB host<br>- Ethernet, HDMI<br>- 2x 46 pin headers<br>- Software Compatibility: Debian, Android, Ubuntu, Cloud9 IDE on Node.js w/ BoneScript library |

[4] https://www.raspberrypi.org/

[5] https://beagleboard.org/black

| Smart sensor | Imagen | Características |
|---|---|---|
| Raspberry Pi[4] |  | - Microcontroller: ATmega1281<br>- Frequency: 14.74 MHz<br>- SRAM: 8 kB<br>- EEPROM: 4 kB<br>- FLASH: 128 kB<br>- SD card: 2 GB<br>- Weight: 20 g<br>- Temperature range:[-10 °C, +65 °C]<br>-7 analog inputs, 8 digital I/O<br>- 2 UARTs, 1 I2C, 1 SPI, 1 USB<br>- Battery voltage: 3.3-4.2 V<br>- USB charging: 5 V - 480 mA<br>- Solar panel load: 6-12 V - 330 mA |

## Tecnologías de comunicación

Como hemos mencionado anteriormente la arquitectura IoT está basada en redes WSN donde sensores inteligentes intercambian información entre ellos y son capaces de enviar datos de telemetría hacia el servidor gracias a los módulos de comunicación inalámbricas que poseen, muchas veces integradas con el procesador en un solo chip. Entre las tecnologías inalámbricas más usadas tenemos a: Zigbee, BLE, Lora, Sigfox las cuales son detalladas en el Cuadro 3.3.

[6] http://www.libelium.com/products/waspmote/hardware/

| Tecnología | Logotipo | Características |
|---|---|---|
| Wi-fi[7] |  | - Wi-Fi: Wireless Fidelity<br>- Estándar inicial IEEE 802.11, seguidos por 802.11a, 802.11b, 802.11g, 802.11n, 802.11ac, 802.11sd,802.11ah (Wifi-Halow)<br>- Frecuencias: 2.4GHZ, 5.4GHz, 60Ghz, 900 MHz.<br>- Modos de conexión: infraestructura: dispositivo con un punto de acceso a red (Access Point), ad-hoc: red virtual entre dispositivos sin el uso de un AP físico, wifi-direct: conexión entre dispositivos que negocian cuál de ellos actuará como AP, simulando una red Wifi. |
| Zigbee[8] |  | - Estándar IEEE 802.15.4<br>- Bajo consumo<br>- Frecuencias: 868 MHz en Europa, 915 MHz en USA y 2.4 GHz en todo el mundo<br>- Puede formar redes tipo estrella, árbol y malla<br>- La red está formada por un coordinador, routers y dispositivos finales.<br>- Aplicaciones: Domótica, Energía Inteligente, Ciudades inteligentes, entre otros.<br>- Especificaciones: Zigbee PRO, Zigbee RF4CE y Zigbee IP |

[7] http://www.wi-fi.org/

[8] http://www.zigbee.org/

| Tecnología | Logotipo | Características |
|---|---|---|
| Bluetooth Low Energy [9] |  | - Bajo consumo<br>- Permite crear redes de área personal (PAN)<br>- Frecuencia: 2.4GHz<br>- Utiliza Frequency Hopping para evitar interferencias con otras tecnologías que usan la misma banda de frecuencia.<br>- Estándar IEEE 802.15<br>- la última especificación es la 5.0<br>- Ha sido integrado en los teléfonos inteligentes.<br>- Velocidad de transferencia de 1 Mhz<br>- Redes tipo estrella. |
| LoRa[10] |  | - Lora: Long range (largo alcance)<br>- Plataforma inalámbrica de bajo consumo<br>-Frecuencias: 868 MHz en Europa y 915 MHz en USA<br>- Protocolos: Lora y LoRaWan<br>- Velocidad baja: decenas de Kbps<br>- Grandes distancias de cobertura ( algunos kilómetros) |

[9] https://www.bluetooth.com/

[10] https://www.lora-alliance.org/

| Tecnología | Logotipo | Características |
|---|---|---|
| LTE |  | - LTE: Long Term Evolution (Evolución a largo plazo)<br>- Estándar creado por el 3GPP[11]<br>- Versión actual es LTE Advanced Pro<br>- Velocidades de transmisión de 3Gps, latencias de 2ms<br>- Tecnología IP de extremo a extremo<br>- Red formada por nodos eNB como estaciones bases que dispone de funcionalidad de control embebidas, evitan el uso de controladores de red (RNC)<br>-Enfocada a aplicaciones IoT<br>- Conocida como 4.5G |



## Gateway y border routers

En los epígrafes anteriores se ha dado una introducción importante al dominio de los sensores (ver Imagen 3.1), es decir los principales sensores inteligentes comerciales y tecnologías de comunicación inalámbricas predominantes en las WSN actuales. Por tanto ya podemos subir al próximo dominio de red, encargado de garantizar que los datos provenientes de los sensores lleguen a un servidor o cloud computing que soporte la aplicación específica de IoT.

Es necesario recordar que en Internet el protocolo aún predominante es el IPV4, con su limitante en cantidad de direcciones IP disponibles. El IoT por concepto requiere que cada dispositivo disponga de una dirección IP única, es aquí donde se requiere la implementación de IPV6 a nivel de nodo sensor, para que los millones de dispositivos IoT puedan garantizar el requerimiento mencionando. Por tanto

[11] http://www.3gpp.org/

podemos intuir un pequeño problema de compatibilidad de dispositivos IoT IPV6 sobre redes IPV4 predominantes en internet, siendo éste uno de los grandes retos encargado a este dominio de IoT.

En este domino de red, tenemos como actores principales a los Gateway y Border routers que de una u otra manera dan solución y soporte a este dominio. Los border routers como su nombre lo indica son ruteadores que permiten conectar una red con otra, en este caso la WSN con Internet. Estos ruteadores se encuentran en el borde de la WSN, es decir el punto de conexión de comunicación extremo de la WSN y a ello se debe su nombre por encontrarse en el "borde" de la red.

Los gateways (pasarelas en español, aunque este es un anglicismo técnico completamente aceptado) permiten la conexión y comunicación entre dispositivos de una misma red o diferentes redes, traduciendo el protocolo de una red al nuevo protocolo que usan en la otra que deseamos conectarnos.

En cambio IoT necesita de un nuevo tipo de gateway que combine funciones del gateway y border router convencional y que brinde además seguridad, conectividad estable, traslación de protocolos, filtrado y procesamiento del dato, capacidad de almacenamiento y análisis y gestión de acceso de los motes.

Los gateway IoT actuales están basados en Windows o Linux y pueden ser implementados en diferentes plataformas de hardware, desde un teléfono inteligente, una Raspberry Pi o en plataformas propietarias más robustas como es el caso de Meshlium[12] , un gateway IoT que permite conectar motes de Libelium con diferentes plataformas de cloud como se muestra en la Imagen 3.4.

[12] http://www.libelium.com/products/meshlium/

Imagen 3.4. Meshlium, un gateway IoT de Libelium.

## Cloud computing y plataformas Iot

### Cloud computing

#### Concepto de cloud computing

Se conoce como Cloud computing o simplemente la cloud, como el acceso ubicuo a la red bajo demanda a un conjunto de recursos informáticos configurables como: redes, servidores, almacenamiento, aplicaciones y servicios (NIST, 2013). También las podemos definir como el conjunto de aplicaciones y servicios que se encuentran ejecutándose en una red distribuida de recursos virtualizados con acceso utilizando protocolos de internet y estándares de red (Pizzolli *et al.*, 2016; Sosinsky, 2011).

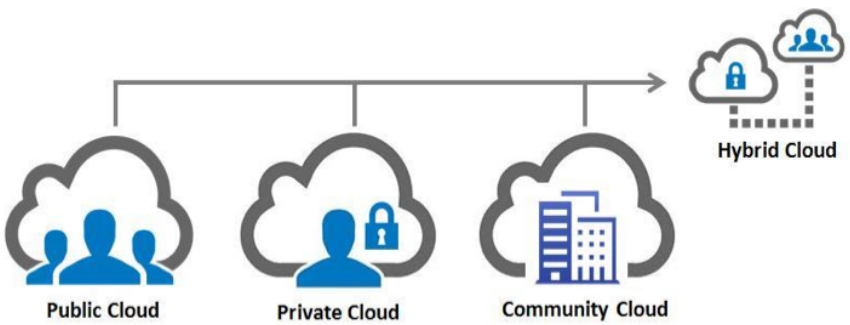
#### Modelos de despliegue de cloud computing

En la Imagen 3.5 se muestra un esquema de los tipos de modelos de despliegue de una cloud:

- Nube pública. Libre acceso desde cualquier parte del mundo con posibles restricciones.

- Nube privada. Se implementa dentro de las instalaciones de una empresa (on-premise) y es de su uso exclusivo
- Nube híbrida. Combinación de una nube pública y privada al mismo tiempo.
- Nube Comunidad. Para una organización de propósito común.

Imagen 3.5. Modelos de despliegue de Cloud Computing




## Modelo de servicio de cloud computing

En el Cuadro 3.4 Se presenta los modelos de servicios de cloud computing.

Cuadro 3.4: Modelos de servicio de cloud computing

| Tipo de Servicio | Descripción del servicio |
|---|---|
|   | SaaS (software como servicio), infraestructura, plataforma y software de aplicación listo para su uso, en donde todo aspecto de la nube es abstracto para el usuario. Ejemplos: Google Drive, OneDrive, Dropbox, etc. |
|   | PaaS (plataforma como servicio). Infraestructura y plataforma predefinidas a partir de las cuales el usuario puede implementar su aplicación mediante herramientas especificadas por el proveedor del servicio. Ejemplos: Microsoft Azure, Google Cloud Platform, Ecuahosting. |

| Tipo de Servicio | Descripción del servicio |
|---|---|
|  | IaaS (infraestructura como servicio). Provee un entorno de virtualización de recursos físicos para que el usuario sea el encargado de definir una infraestructura que se ajuste a sus necesidades, los servicios IaaS más populares son: OpenStack[13] y CloudStack[14] como alternativas de cloud IaaS open source; alternativas de pago son: IBM Cloud[15] , Amazon Web Services[16] , Microsoft Azure[17] , Google Cloud Platform[18] , entre otras. |



## Plataformas IoT

Las plataformas IoT son sistemas computacionales de proveedores externos o desarrollados a medida, los cuales han sido creados para recibir datos de sensores, almacenarlos en sus sistemas de bases de datos y ofrecer servicios adicionales de procesamiento, análisis de datos, monitoreo de la WSN y control de actuadores. Las plataformas más destacadas actualmente son:

- Thingspeak: https://thingspeak.com/
- IBM Bluemix: http://www.ibm.com/cloud-computing/bluemix/
- Amazon: http://aws.amazon.com/es/iot/
- Carriots: https://www.carriots.com/
- Adafruit IO: https://io.adafruit.com/
- Thingworx: http://www.thingworx.com/
- Temboo: https://temboo.com/
- Thethings: https://thethings.io/
- IoTMach: http://iotmach.utmachala.edu.ec/

[13] https://www.openstack.org/
[14] https://cloudstack.apache.org/
[15] https://www.ibm.com/cloud-computing/
[16] https://aws.amazon.com/es/
[17] https://azure.microsoft.com/es-es/
[18] https://cloud.google.com/

IoTMach es una plataforma IoT creada por docentes y estudiantes del grupo de investigación AutoMathTIC de la Unidad Académica de Ingeniería Civil de la Universidad Técnica de Machala, que además de monitorear y controlar redes WSN a través de dashboards creados dinámicamente dispone de herramientas con inteligencia de negocios dedicadas a la agricultura, planeamiento y control de riego entre otras funcionalidades.

Este capítulo ha hecho un recorrido por los aspectos más importantes del IoT que permiten al lector iniciar sus primeros pasos en la automatización ya sea de agricultura como de cualquier aplicación. Los conceptos, estructuras y plataformas indicadas en este capítulo podrán ser utilizadas en nuevos dominios IoT, a diferencia de los sensores que siempre estarán relacionados con las variables a medir del dominio en cuestión. IOT es un campo nuevo de investigación, pero con un crecimiento acelerado, para un estudio más profundo y actualización continua se recomienda seguir la bibliografía recomendada.

## Referencias Bibliográficas


Aazam, M., & Huh, E. N. (2015). Fog computing micro datacenter based dynamic resource estimation and pricing model for IoT. In *Proceedings - International Conference on Advanced Information Networking and Applications, AINA* (Vol. 2015–April, pp. 687–694). http://doi.org/10.1109/AINA.2015.254

Ai, Y., Peng, M., & Zhang, K. (2017). Edge cloud computing technologies for internet of things: A primer. *Digital Communications and Networks*. http://doi.org/10.1016/j.dcan.2017.07.001

Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., & Ayyash, M. (2015). Internet of Things: A Survey on Enabling Technologies, Protocols and Applications. *IEEE Communications Surveys & Tutorials, PP*(99), 1–1. http://doi.org/10.1109/COMST.2015.2444095

Ali, M. I., Ono, N., Kaysar, M., Shamszaman, Z. U., Pham, T.-L., Gao, F., … Mileo, A. (2016). Real-time data analytics and event detection for IoT-enabled communication systems. *Web Semantics: Science, Services and Agents on the World Wide Web, 42*. http://doi.org/http://dx.doi.org/10.1016/j.websem.2016.07.001

Botta, A., de Donato, W., Persico, V., & Pescapé, A. (2015). Integration of Cloud Computing and Internet of Things: A Survey. *Future Generation Computer Systems, 56*, 684–700. http://doi.org/10.1016/j.future.2015.09.021

Cai, H., Xu, B., Jiang, L., & Vasilakos, A. (2016). IoT-based Big Data Storage Systems in Cloud Computing: Perspectives and Challenges. *IEEE Internet of Things Journal, PP*(99), 1. http://doi.org/10.1109/JIOT.2016.2619369

Campoverde, A., Hernandez-Rojas, D., & Mazon-Olivo, B. (2015). Cloud computing con herramientas open-source para Internet de las cosas. *Maskana, 6*, 173–182. Retrieved from http://dspace.ucuenca.edu.ec/handle/123456789/23826

Gitzel, R., Turrin, S., & Maczey, S. (2015). A Data Quality Dashboard for Reliability Data, 90–97. http://doi.org/10.1109/CBI.2015.24

Im, J., Kim, S., & Kim, D. (2013). IoT mashup as a service: Cloud-based mashup service for the internet of things. *Proceedings - IEEE*

*10th International Conference on Services Computing, SCC 2013*, 462–469. http://doi.org/10.1109/SCC.2013.68

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of Parallel and Distributed Computing*, *74*(7), 2561–2573. http://doi.org/10.1016/j.jpdc.2014.01.003

Karkouch, A., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Data quality in internet of things: A state-of-the-art survey. *Journal of Network and Computer Applications, 73,* 57–81. http://doi.org/10.1016/j.jnca.2016.08.002

Khan, S., Parkinson, S., & Qin, Y. (2017). Fog computing security: a review of current applications and security solutions. *Journal of Cloud Computing*, *6*(1), 19. http://doi.org/10.1186/s13677-017-0090-3

Mijovic, S., Shehu, E., & Buratti, C. (2016). Comparing Application Layer Protocols for the Internet of Things via Experimentation. In *2016 IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI)*. Bologna, Italy: IEEE. http://doi.org/10.1109/RTSI.2016.7740559

Moniruzzaman, A., & Hossain, S. (2013). Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. *International Journal of Database Theory and Application*, *6*(4), 1–14. Retrieved from http://www.sersc.org/journals/IJDTA/vol6_no4/1.pdf

NIST. (2013). NIST Cloud Computing Standards Roadmap. *National Institute of Standar and Technology. Special Publication 500-291 V2*, 1–102. http://doi.org/10.6028/NIST.SP.500-291r2

Pizzolli, D., Cossu, G., Santoro, D., Capra, L., Dupont, C., Charalampos, D., … Cascata, D. (2016). Cloud4IoT : a heterogeneous , distributed and autonomic cloud platform for the IoT, 476–479. http://doi.org/10.1109/CloudCom.2016.80

Rodríguez Arias, S., & Hernández Rojas, D. (1999). Amplificador de rango dinámico programable con auto-diagnóstico en tiempo real.

Shaikh, F. K., Zeadally, S., & Exposito, E. (2015). Enabling Technologies for Green Internet of Things. *IEEE Systems Journal, PP(*99), 1–12. http://doi.org/10.1109/JSYST.2015.2415194

Sosinsky, B. (2011). *Cloud Computing Bible*. Indianapolis: Wiley Publishing, Inc.

Suciu, G., Halunga, S., Vulpe, A., & Suciu, V. (2013). Generic platform for IoT and cloud computing interoperability study. ISSCS 2013 - *International Symposium on Signals, Circuits and Systems*. http://doi.org/10.1109/ISSCS.2013.6651222

Suciu, G., Suciu, V., Martian, A., Craciunescu, R., Vulpe, A., Marcu, I., … Fratu, O. (2015). Big Data, Internet of Things and Cloud Convergence ??? An Architecture for Secure E-Health Applications. *Journal of Medical Systems*, *39*(11). http://doi.org/10.1007/s10916-015-0327-y

Vermesan, O., & Friess, P. (2014). *Internet of Things Applications - From Research and Innovation to Market Deployment*. *River Publishers*. Retrieved from https://books.google.com.br/books?id=kw-2doAEACAAJ

Vermesan, O., & Friess, P. (2015). *Building the Hyperconnected Society. River Publishers*. http://doi.org/978-87-93237-99-5

Wang, C., Bi, Z., & Xu, L. Da. (2014). IoT and cloud computing in automation of assembly modeling systems. *IEEE Transactions on Industrial Informatics*, 10(2), 1426–1434. http://doi.org/10.1109/TII.2014.2300346

Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M. (2017). Big Data in Smart Farming – A review. *Agricultural Systems*, 153, 69–80. http://doi.org/10.1016/j.agsy.2017.01.023

# 04 Capítulo Matemáticas aplicadas al sector agropecuario

Bladimir Serrano; Carlos Loor; Eduardo Tusa

Este capítulo explica los procedimientos para la creación de modelos matemáticos que representen procesos asociados al sector agropecuario, como una alternativa de solución en la ingeniería. Se hace una breve explicación introductoria

**Bladimir Serrano:** Ingeniero Civil, Máster Universitario en Ingeniería Computacional y Matemática, de la Universidad Rovira y Virgili, donde obtuvo una mención en Modelación y Simulación de Eventos Discretos, Docente de la UTMACH, desde Mayo del 2009, dictando las cátedras de Álgebra Lineal, Cálculo Diferencial e Integral, Ecuaciones Diferenciales, Estadística Descriptiva e Inferencial, Fisica I y Fisica II, en las Carreras de Ingeniería Civil, Ingeniería de Sistemas, Ingeniería Acuícola, Ingeniería Agronómica y Medicina Veterinaria, hasta febrero del 2017.

**Carlos Loor:** Ingeniero en Electricidad Especialización Electrónica de la Escuela Superior Politécnica del Litoral. Magíster en Educación Superior de la Universidad Tecnológica San Antonio de Machala. Máster Universitario en Ingeniería Electromecánica de la Universidad Politécnica de Madrid (UPM), donde obtuvo una mención en Métodos Numéricos. Actualmente, es candidato a PhD en Diseño, Fabricación y Gestión de Proyectos Industriales en la UPM. Es docente de la Unidad Académica de Ingeniería Civil de la Universidad Técnica de Machala, donde ha impartido las asignaturas de Mecánica Técnica I (Estática), Mecánica Técnica II (Dinámica) y Ecuaciones Diferenciales y Cálculo Integral.

**Eduardo Tusa:** Ingeniero Electrónico (Magna Cum Laude) con una Subespecialización en Matemáticas de la Universidad San Francisco de Quito. Su cuarto año de formación de pregrado fue realizado en la Universidad de Illinois en Urbana - Champaign, USA. Máster en Visión, Imagen y Robótica (con distinción) de la Universidad de Borgoña (Francia), la Universidad de Girona (España) y la Universidad Heriot-Watt (Reino Unido). Actualmente, es doctorando en la especialidad de Señales, Imágenes, Voz, Telecomunicaciones en la Universidad Grenoble Alpes a través del Instituto Nacional de Investigación en Ciencia y Tecnología para el Ambiente y la Agricultura (IRSTEA, por sus siglas en Francés). Es docente de la Universidad Técnica de Machala, donde ha impartido las asignaturas de Programación en MATLAB, Informática, Nuevas Tecnologías de la Información y Comunicación, Cálculo Integral, Ecuaciones Diferenciales, Matemática Avanzada, Probabilidad y Estadística

de los modelos matemáticos, ilustrando la idea de modelos compartimentales para posteriormente, exponer los tipos de modelos matemáticos. A continuación, se revisa los principales aspectos de las ecuaciones diferenciales como herramientas para el planteamiento de modelos matemáticos determinísticos. Los modelos matemáticos se presentan de forma analítica y gráfica a través de implementaciones computacionales en el programa MATLAB.

## 1. Introducción

En la actualidad, se ha incrementado la necesidad de introducir los modelos y herramientas matemáticas en nuestras investigaciones. La utilización e interpretación adecuada de estas técnicas permiten la toma de decisiones óptimas para favorecer el desarrollo de los sistemas productivos. El carácter integral en la solución de las tareas científicas y económicas, así como la eficiencia de los métodos utilizados para influir sobre los objetivos de trabajo, exigen una alta preparación del especialista para emitir criterios con altos niveles de fiabilidad en los procesos agrícolas.

Por otra parte, el aumento progresivo de la población mundial, junto a la creciente necesidad de garantizar la alimentación de ésta, sumado a los cambios climáticos; han conllevado al constante desarrollo de la investigación agrícola. Para realizar estudios y proyecciones futuras sobre procesos agrícolas, se hace imprescindible conocer: ¿Cómo lograr mayores niveles de eficacia en el proceso de toma de decisiones?, ¿Qué métodos matemáticos permiten analizar datos con el fin de obtener conclusiones científicas? ¿Cómo fortalecer las investigaciones de los procesos agropecuarios utilizando la Matemática Aplicada? Esta disciplina en las ciencias agropecuarias permite brindar criterios y herramientas básicas para manejar problemas, recurriendo incluso, a la utilización de nuevas tecnologías con el fin de hacerle frente a objetos de estudio altamente complejos.

El presente capítulo tiene como punto de partida, la estructuración de los modelos matemáticos, su representación y sus tipos, para centrarnos principalmente en la generación

de modelos determinísticos mediante la resolución analítica de las ecuaciones diferenciales y su programación en MATLAB (Moore 2014).

## 2. Modelos matemáticos

Un primer acercamiento al modelamiento matemático parte desde su principal propósito. Los modelos constituyen representaciones, patrones, descripciones o analogías que persiguen la visualización de un determinado objeto de estudio que no puede ser valorado directamente, o del cual se desprende un conjunto de postulados, datos o inferencias reproducidas mediante expresiones matemáticas (Fowler and Fowler 1997). Un modelo matemático puede ser concebido como una representación en términos matemáticos del comportamiento de dispositivos y objetos reales (Fishwick 2007).

Un modelo matemático puede representar una simplificación o abstracción de un sistema real que existe en el universo. Así, el modelamiento matemático puede apreciarse en el crecimiento y decaimiento de las poblaciones de animales y seres humanos. Por ejemplo, la industria pesquera debería estar interesada en el efecto de la pesca sobre el crecimiento poblacional de los peces con la finalidad de no agotar este recurso. Inicialmente, se puede asumir un comportamiento exponencial de la población, y más adelante se pueden incorporar otros efectos, como los periodos de reproducción de ciertas especies.

La modelización matemática puede resumirse en tres pasos:

1. La construcción del modelo, es la transformación del sistema no matemático en el lenguaje matemático.

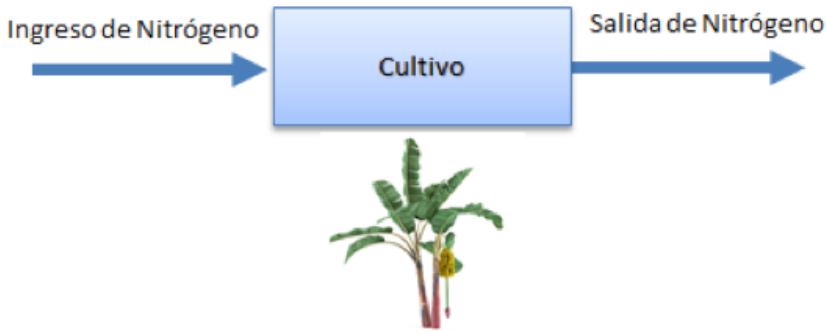
2. El análisis del modelo, consiste en el estudio sistemático del modelo.

3. La interpretación del análisis matemático, es la aplicación de los resultados del estudio matemático al sistema real.

## 2.1. Representación del modelo matemático

El marco del modelo compartimental es una manera natural que permite la formulación de procesos que poseen entradas y salidas a lo largo del tiempo. Como ejemplo, el compartimento del cultivo de banano (Ver Imagen 4.1) (Nomura *et al*. 2017) después de ser fertilizado con urea, donde básicamente la entrada es la cantidad de nitrógeno suministrada al cultivo por efecto del fertilizante.

Imagen 4.1. Diagrama compartimental para el cultivo de banano




Fuente: Elaboración propia.

Para el ejemplo anterior, la ley de equilibrio establece la relación entre la tasa neta de variación de nitrógeno dentro del cultivo y la diferencia entre la tasa de variación de nitrógeno que entra al cultivo menos la tasa de variación de nitrógeno que sale del cultivo.

Otro ejemplo que se ilustra en las Imágenes 4.2, 4.3, 4.4. establece las relaciones sobre la oxitetraciclina como un antibiótico prescrito en medicina veterinaria para tratar problemas respiratorios, hemoparasitosis, problemas diarreicos, etc (Fedeniuk 1998). Dentro del animal, el fármaco se propaga desde el tracto gastrointestinal hasta el torrente sanguíneo, desde donde es extraído por los riñones, para finalmente ser excretado por la orina. Las Imágenes 4.2, 4.3, 4.4. establecen los compartimentos respectivos para el tracto gastrointestinal, el torrente sanguíneo y los riñones del animal; donde la entrada y la salida es la oxitetraciclina.

La Imagen 4.2 relaciona la tasa de variación de oxitetraciclina en el tracto gastrointestinal igual a la tasa de variación de oxitetraciclina que ingresa en el tracto gastrointestinal menos la tasa de variación de oxitetraciclina que sale del tracto gastrointestinal.

Imagen 4.2. Diagrama compartimental para el tracto intestinal del ganado


Fuente: Elaboración propia.

La tasa de variación de oxitetraciclina en el torrente sanguíneo es igual a la tasa de variación de oxitetraciclina que ingresa en el torrente sanguíneo menos la tasa de variación de oxitetraciclina que sale del torrente sanguíneo.

Imagen 4.3. Diagrama compartimental para el torrente sanguíneo del caballo


Fuente: Elaboración propia.

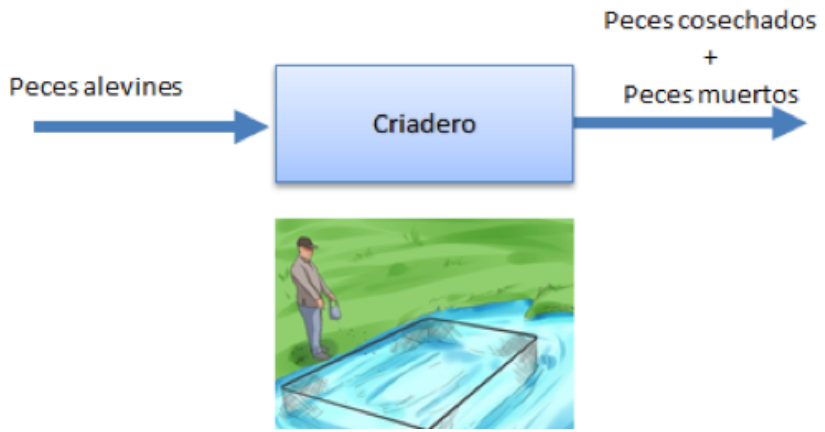
La tasa de variación de oxitetraciclina en el tracto urinario es igual a la tasa de variación de oxitetraciclina que ingresa al tracto urinario menos la tasa de variación de oxitetraciclina que sale del tracto urinario.

Imagen 4.4. Diagrama compartimental para el tracto urinario


Fuente: Elaboración propia.

Otro ejemplo que merece el análisis compartimental se basa en las dinámicas presentes en los criaderos de peces que representan modelos poblacionales. Los peces se cosechan a una tasa constante por semana, teniendo en cuenta la tasa de mortalidad o aglomeración y la tasa de natalidad per-cápita. De este modo, la Imagen 4.5. ilustra la relación de la tasa de cambio de la población de peces en el criadero como la diferencia entre la tasa de cambio de natalidad de la población de peces menos la tasa de cambio de mortalidad y la tasa de cambio de cosecha.

Imagen 4.5. Diagrama compartimental para el criadero de peces

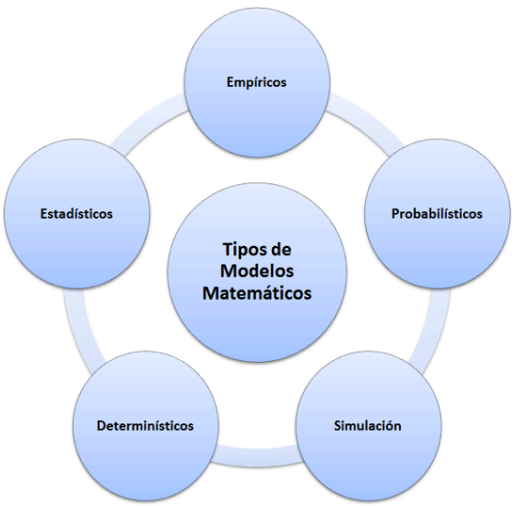



Fuente: Elaboración propia.

## 2.2. Tipos de modelos matemáticos

Existe un sinnúmero de formas de acercarse a un objeto de estudio a partir de diferentes aproximaciones o tipologías que caracterizan un modelo matemático. La Imagen 4.6. presenta los tipos de modelos matemáticos descritos por (Barnes and Fulford 2011).

Imagen 4.6. Tipos de modelos matemáticos




Fuente: Elaboración propia.

Por un lado, los modelos empíricos son las aproximaciones más básicas porque resultan de ajustar una curva a través de un conjunto de datos con la finalidad de predecir resultados para los cuales no existen datos (Austin et al. 1998). Los modelos poblacionales además de ser ajustados mediante aproximaciones exponenciales, puede recurrir a expresiones polinómicas. La desventaja de esta aproximación es que el modelo no puede extenderse a otros problemas relacionados debido a que está limitado a sus datos de ajuste.

El modelamiento mediante procesos estocásticos o probabilísticos (Bock 1996), estiman la probabilidad de un resultado predecible basado en los datos disponibles. Este tipo de modelamiento permite predecir la incertidumbre del resultado, lo cual es relevante para el estudio de fenómenos

que presentan un alto grado de variabilidad. Estos modelos son de importancia en los modelos de pequeñas poblaciones para la predicción de tasas de reproducción en un intervalo de tiempo.

Los modelos de simulación están vinculados a los programas computacionales que aplican un conj unto de reglas con el fin de generar datos que emulan un resultado real considerando diferentes escenarios (Bagni, Berchi, and Cariello 2002). Usualmente, los ingenieros utilizan estos modelos para la identificación de problemas que podrían surgir durante el uso o elaboración de un dispositivo. Si bien las simulaciones proporcionan modelos muy cercanos a la realidad, no implica necesariamente que sean los mejores modelos, los cuales se caracterizan por su simplicidad.

Los modelos determinísticos ignoran la variación aleatoria y formulan expresiones matemáticas que describen relaciones fundamentales entre las variables del problema (Gurney and Nisbet 1998). Por ejemplo, un modelo poblacional determinístico tiene como propósito la obtención de una ecuación diferencial que relacione las tasas de nacimiento y de mortalidad, con el tamaño de la población en un tiempo determinado.

Los modelos estadísticos se basan en pruebas de hipótesis que permiten categorizar un conjunto de datos empíricos (Montgomery and Runger 2010). Estas categorías se ajustan a determinadas distribuciones particulares asociadas a descriptores estadísticos como la media y la desviación estándar, con la finalidad de predecir resultados futuros. En términos de modelos poblacionales, una especie podría ser evaluada si a una muestra de su población se alimenta con un balanceado de cierto tipo (categoría A), mientras que otra muestra se alimenta con productos orgánicos (categoría B). Los resultados indicarían un porcentaje de margen de error con el que se ha realizado la predicción.

El modelamiento es una herramienta muy útil, así como también; es un marco de referencia para la investigación, el debate y el planeamiento; lo que proporciona una fuente

valiosa de información para la toma de decisiones. En este capítulo de libro, se dará énfasis a los modelos determinísticos representados a través de las ecuaciones diferenciales.

## 3. Ecuaciones diferenciales

La naturaleza se encuentra sometida a constantes cambios que pueden ser apreciados en diferentes campos de estudio. En la agricultura, se puede observar las variaciones que sufre un cultivo de ciclo corto en su tamaño durante el transcurso del tiempo, así como la variación interna de su cantidad de nitrógeno después de su fertilización. Existen muchos problemas en las ciencias agropecuarias que son formulados matemáticamente para determinar una función desconocida que debe satisfacer cierta ecuación, en la que figuran dicha función y sus derivadas. La ecuación que contiene la función desconocida y varias de sus derivadas, es una ecuación diferencial (Hinrichsen and Pritchard 2005).

La modelación matemática por ecuaciones diferenciales es una aproximación determinística que en la actualidad se ha constituido en la herramienta fundamental para cumplir con los objetivos planteados por muchos investigadores. Su aspiración es comprender el comportamiento de ciertos fenómenos que involucran cambios descritos por ecuaciones que relacionan magnitudes variantes en el tiempo (Greefrath 2011).

Retomando el ejemplo del modelo poblacional, la suposición de que la tasa de crecimiento posee un ritmo proporcional al tamaño de su población resulta razonable para una población de animales en condiciones ideales de ambiente ilimitado, nutrición adecuada, ausencia de depredadores, inmunidad ante enfermedades. Si representamos las variables que están relacionadas en este fenómeno, encontramos el tiempo, $t$, como la variable independiente y al número de habitantes, $N$, como la variable dependiente. Bajo estas condiciones, la rapidez de crecimiento de la población está descrita por la siguiente ecuación (1)

$$\frac{dN}{dt} = kN$$

siendo $k$, la constante de proporcionalidad. La ecuación (1) representa una ecuación diferencial, ya que contiene una función desconocida $N(t)$ que varía con el tiempo, y su derivada $^{dN}/_{dt}$ que representa la rapidez o tasa de crecimiento de la población. De esta manera, una ecuación diferencial está constituida por una función desconocida y algunas de sus derivadas (Zill 2016).

### 3.1. Clasificación de las ecuaciones diferenciales

Las ecuaciones diferenciales se clasifican en ecuaciones diferenciales ordinarias (EDO) y ecuaciones diferenciales en derivadas parciales (EDP). Una ecuación diferencial se dice ordinaria (EDO), si la función desconocida es función de una sola variable independiente. La ecuación (2) ilustra un ejemplo a continuación

$$\frac{dy}{dx} = x^2 - x$$

Una ecuación diferencial en derivadas parciales (EDP), es aquella cuya función desconocida depende de dos o más variables independientes. La ecuación (3) es un ejemplo de este tipo de ecuación

$$\frac{\partial^2 y}{\partial t^2} - \frac{\partial^2 y}{\partial x^2} = 0$$

### 3.2. Orden de una ecuación diferencial

El orden de una ecuación diferencial corresponde al orden de la mayor derivada que figura en la ecuación. Por ejemplo, la ecuación (1) es una ecuación diferencial de primer orden al igual que la ecuación (2). La ecuación (3) representa una ecuación diferencial de segundo orden.

### 3.3. Solución de una ecuación diferencial

La solución de una ecuación diferencial, está constituida por la expresión algebraica de la función desconocida *y = f(x)*. Esta función satisface la ecuación diferencial de forma idéntica, para todo x dentro del intervalo de interés. Si consideramos a

$$f(x) = c_1 sen(2x) + c_2 cos(2x)$$

una solución de la ecuación diferencial

$$\frac{d^2y}{dx^2} + 4y = 0$$

donde $c_1$ y $c_2$ son dos constantes cualesquiera, la forma de verificar esta solución es la siguiente:

1. Lo primero que debemos obtener es la segunda derivada de la función de la ecuación (4). La primera y segunda derivada de la función solución, son respectivamente

$$\frac{dy}{dx} = 2c_1 cos(2x) - 2c_2 sen(2x)$$

$$\frac{d^2y}{dx^2} = -4c_1 sen(2x) - 4c_2 cos(2x)$$

2. En segunda instancia, se sustituye en la ecuación (7) en la ecuación diferencial (5), resultando la siguiente identidad:

$$-4c_1 sen(2x) - 4c_2 cos(2x) + 4(c_1 sen(2x) + c_2 cos(2x)) = 0$$

$$-4c_1 sen(2x) - 4c_2 cos(2x) + 4c_1 sen(2x) + 4c_2 cos(2x) = 0$$

### 3.4. Ecuaciones diferenciales separables de primer orden

La solución general para este tipo de ecuación está dada por la siguiente expresión

$$F(x)dx + G(y)dy = 0$$

A continuación, se aplica las técnicas de integración de la siguiente manera

$$\int F(x)dx + \int G(y)dy = c$$

donde c es una constante arbitraria. La solución con valores iniciales se puede obtener a partir de la condición inicial $y(x_0)=y_0$, con la siguiente expresión matemática:

$$\int_{x_0}^{x} F(x)dx + \int_{y_0}^{y} G(y)dy = 0$$

Un ejemplo se puede ilustrar a través de la ecuación diferencial descrita a continuación

$$\frac{dy}{dx} = \frac{xe^x}{2y}$$

Esta expresión se puede reescribir de la forma señalada en la ecuación (8) aplicando el principio de separación de variables

$$2ydy - xe^x dx = 0$$

Aplicando el operador integración al lado izquierdo de la ecuación (12), se obtiene la solución en la ecuación (13)

$$\int 2ydy - \int xe^x dx = c$$

$$y^2 - xe^x + e^x = c$$

$$y^2 = xe^x - e^x + c$$

- La ecuación diferencial (11) puede ser resuelta mediante la utilización de lenguajes de programación como MATLAB (San Martín Cuenca and Tusa Jumbo 2015). La Imagen 4.7. presenta el código implementado en MATLAB

para la resolución de la ecuación (11). El comando *ode45* (Bober 2013; Quarteroni, Saleri, and Gervasio 2014) implementa el método numérico de Runge-Kutta (Runge 1895; Kutta 1901) con un paso de tiempo variable con la finalidad de realizar un cálculo eficiente. El comando se expresa a continuación con los siguientes parámetros

$$[x,y] = ode45(@fname, xspan, y0)$$

- *fname* es el nombre de la función. En la Imagen 4.7., la función se llama *f*.
- *xspan* es el vector que define el límite inicial y final de la integración.
- *y0* es el vector de condiciones iniciales.
- *x* es el valor de la variable independiente en la que se calcula el vector de soluciones y. Este vector no es necesariamente igual a xspan porque ode45 genera pasos más pequeños cuando el problema cambia rápidamente y pasos más grandes cuando es relativamente constante.
- y es el vector solución

La Imagen 4.8. presenta la gráfica de la solución de la ecuación diferencial (11) utilizando el comando *plot(x,y)*.

Imagen 4.7. Implementación de la ecuación diferencial (11) en MATLAB

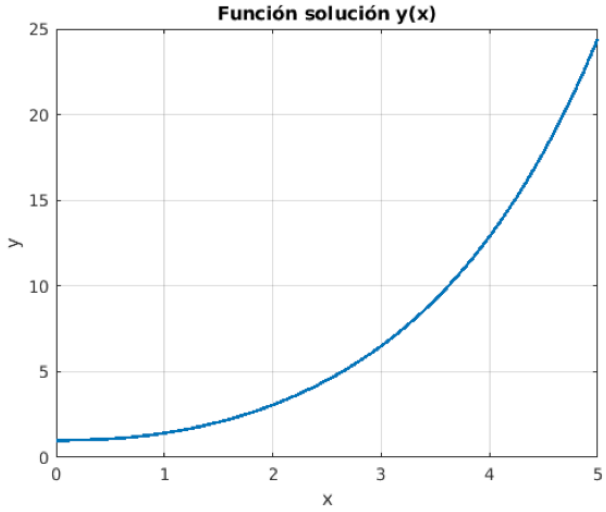
```
N = 100;                % Número de puntos
x0 = 0;                 % Condición inicial en x
y0 = 1;                 % Condición inicial en y
xmax = 5;               % Ancho de la ventana
x = linspace(0,1,N);    % Dominio de la función

% Ecuación diferencial
f = @(x,y) [x*exp(x)/(2*y)];
% Comando para resolver ecuaciones diferenciales
[xsol, ysol] = ode45(f, [x0,xmax], y0);

% Comandos de graficación de la función solución
plot(xsol, ysol,'LineWidth',2)
grid on;
title('Función solución y(x)')
xlabel('x')
ylabel('y')
```

Fuente: Elaboración propia.

Imagen 4.8. Gráfica de la solución de la ecuación diferencial (11) en MATLAB




Fuente: Elaboración propia.

## 3.5. Ecuaciones diferenciales lineales de primer orden

Si se retoma el ejemplo de la oxitetraciclina suministrada como medicamento veterinario, se puede considerar las constantes de proporcionalidad que se asocian con las tasas a las que se difunde la tetraciclina desde el tracto gastrointestinal hasta el torrente sanguíneo para luego ser eliminada por la orina. Los valores son de 0.72 mg/h y 0.15 mg/h respectivamente, suponiendo que la cantidad inicial de oxitetraciclina en el tracto gastrointestinal es 10 mg, y no hay antibiótico en el torrente sanguíneo y el tracto urinario.

Se asume que $x(t)$ representa la cantidad de oxitetraciclina en el tracto intestinal, mientras que $y(t)$ representa la cantidad del mismo medicamento en el torrente sanguíneo; ambas como funciones del tiempo.

Si consideramos que la cantidad de oxitetraciclina que sale del tracto gastrointestinal es proporcional a la concentración de medicamento, se procede a plantear el siguiente modelo

$$\frac{dx}{dt} = -k_1 x$$

La cantidad de medicamento presente en el torrente sanguíneo está dada por

$$\frac{dy}{dt} = k_1 x - k_2 y$$

donde $k_1$ y $k_2$ son las constantes de proporcionalidad dadas por 0.72 mg/h y 0.15 mg/h, respectivamente. Las condiciones iniciales a considerarse son $x(0)=10\ mg$ y $y(0)=0\ mg$. La ecuación diferencial (14) puede reescribirse de la forma dada por la ecuación (8)

$$\frac{1}{x}dx + k_1 dt = 0$$

Aplicando el operador integración al lado izquierdo de la ecuación (16), se obtiene la solución general en la ecuación (17)

$$\int \frac{1}{x}dx + \int k_1 dt = c$$

$$ln(x) + k_1 t = c$$

De la ecuación (17), se despeja solución general, $x(t)$, de la ecuación diferencial (14)

$$e^{ln(x)+k_1 t} = e^c$$

$$xe^{k_1 t} = c_0$$

$$x(t) = c_0 e^{-k_1 t}$$

Reemplazando la condición inicial $x(0)=10\ mg$ en la solución general, obtenemos una solución particular de la ecuación diferencial

$$x(t) = 10e^{-k_1 t}$$

Ahora, se procede a reemplazar la función *x(t)* en la ecuación diferencial (15)

$$\frac{dy}{dt} = 10k_1e^{-k_1t} - k_2y$$
$$\frac{dy}{dt} + k_2y = 10k_1e^{-k_1t}$$

La ecuación (20) posee la forma de la ecuación diferencial lineal de primer orden

$$\frac{dy}{dt} + p(t)y = q(t)$$

Este tipo de ecuaciones diferenciales se resuelven calculando el factor integrante

$$R(t) = e^{\int p(t)dt}$$

de modo que la ecuación diferencial puede reescribirse de la siguiente forma

$$\frac{d}{dt}\left(R(t)y(t)\right) = R(t)q(t)$$

El factor integrante de la ecuación diferencial (20) está dada por la siguiente expresión

$$R(t) = e^{\int k_2dt} = e^{k_2t}$$

Se reescribe la ecuación (20) de la forma expresada en la ecuación (23)

$$\frac{d}{dt}\left(e^{k_2t}y(t)\right) = e^{k_2t}10k_1e^{-k_1t}$$

Se aplica la operación de integración a ambos lados de la ecuación (25)

$$\int \frac{d}{dt}\left(e^{k_2 t}y(t)\right)dt = \int 10k_1 e^{(k_2-k_1)t}dt$$

$$e^{k_2 t}y(t) = \frac{10k_1}{k_2-k_1}e^{(k_2-k_1)t} + c$$

Se despeja la solución general $y(t)$ de la siguiente forma

$$y(t) = \frac{10k_1}{k_2-k_1}e^{-k_1 t} + ce^{-k_2 t}$$

Se reemplaza la condición inicial $y(0)=0$ mg para obtener el valor de la constante c

$$c = -\frac{10k_1}{k_2-k_1}$$

Finalmente, la solución particular de la ecuación diferencial (20) se presenta a continuación

$$y(t) = \frac{10k_1}{k_2-k_1}\left(e^{-k_1 t} + e^{-k_2 t}\right)$$

Cabe señalar, que la ecuación diferencial (14) resuelta por separación de variables, es una ecuación diferencial lineal de primer orden y puede resolverse calculando el factor integrante. La Imagen 4.9. presenta el código en MATLAB para la resolución de las ecuaciones diferenciales (14) y (15), mientras la Imagen 4.10., presenta las gráficas de las curvas resultantes.

Imagen 4.9. Implementación de las ecuaciones diferenciales (14) y (15) en MATLAB

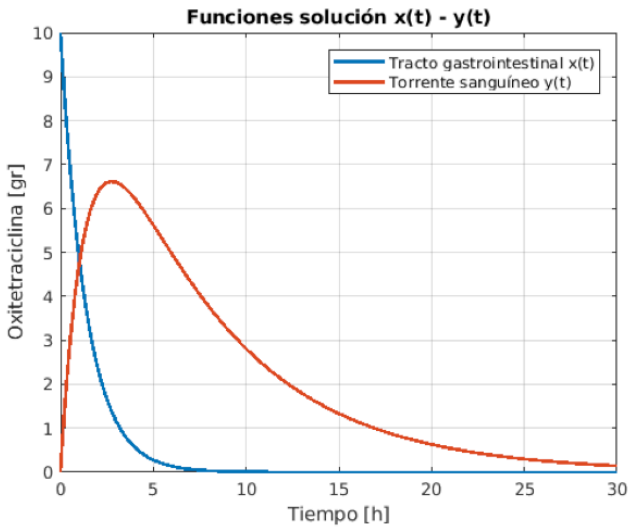
```
k1 = 0.72;                % Constante de proporcionalidad 1
k2 = 0.15;                % Constante de proporcionalidad 2
N = 1000;                 % Número de puntos
t0 = 0;                   % Condición inicial en t
x0 = 10;                  % Condición inicial en x
y0 = 0;                   % Condición inicial en y
tmax = 30;                % Ancho de la ventana
t = linspace(0,1,N);      % Dominio de la función

% Ecuaciones diferenciales
f = @(t,x) [-k1*x(1);k1*x(1)-k2*x(2)];
% Comando para resolver ecuaciones diferenciales
[tsol, xsol] = ode45(f, [t0,tmax], [x0, y0]);

% Comandos de graficación de la función solución
plot(tsol,xsol(:,1),tsol,xsol(:,2),'LineWidth',2)
grid on;
title('Funciones solución x(t) - y(t)')
xlabel('Tiempo [h]')
ylabel('Oxitetraciclina [gr]')
legend('Tracto gastrointestinal x(t)','Torrente sanguíneo y(t)')
```

Fuente: Elaboración propia.

Imagen 4.10. Gráfica de las soluciones de las ecuaciones diferenciales (14) y (15) en MATLAB




Fuente: Elaboración propia.

De la Imagen 4.10. se desprenden algunas observaciones. Primero, las curvas soluciones responden al comportamiento de combinaciones lineales de funciones exponenciales decrecientes. Segundo, la función que representa la cantidad de oxitetraciclina en el tracto gastrointestinal, $x(t)$, muestra que los 10 gr del medicamento han sido consumidos en alrededor de 7 horas. Esta estimación brinda una información puntual sobre la frecuencia con la cual se debe suministrar el medicamento. Tercero, la cantidad de oxitetraciclina en el torrente sanguíneo, $y(t)$, alcanza un valor máximo de 6.5 gr después de alrededor de 4 horas. Posteriormente, la cantidad del medicamento desaparece del torrente sanguíneo en alrededor de 30 horas. Cabe señalar que la cantidad de oxitetraciclina en el tracto gastrointestinal y en el torrente sanguíneo llegan a ser iguales a 4.7 gr, aproximadamente, después de alrededor de una hora que el medicamento ha ejercido su acción sobre el organismo del animal.

### 3.6. Ecuaciones diferenciales no lineales

Una ecuación diferencial ordinaria es no lineal cuando posee funciones no lineales de la variable dependiente o de sus derivadas, como por ejemplo; $sen(y)$ o $ey'$ que no pueden estar presentes en una ecuación lineal, como por ejemplo:

$$(1-y)y' + 2y = e^x$$

$$\frac{d^2y}{dx^2} + sen(y) = 0$$

$$\frac{d^4y}{dx^4} + y^2 = 0$$

La ecuación diferencial (11) que se resolvió analíticamente por separación de variables, posee términos no lineales. Sin embargo, no todas las ecuaciones diferenciales no lineales poseen un método analítico único para su resolución. Retomando el ejemplo de los criaderos de peces que representan modelos poblacionales (Borrelli and Coleman 1998), una

buena aproximación está dada por la ecuación diferencial ordinaria descrita a continuación

$$\frac{dy}{dt} = k_1 y - k_2 y^2 - k_3$$

donde $k_1$ es una constante de proporcionalidad a la tasa de cambio de la población, $k_2$ describe la tasa de aglomeración, $k_3$ representa la tasa de cosecha y la condición inicial es $y(x_0)=y_0$. Si se asume que no existe aglomeración, $k_2=0$, la ecuación (33) se reduce a una ecuación diferencial ordinaria lineal de primer orden:

$$\frac{dy}{dt} = k_1 y - k_3$$

Esta ecuación diferencial se resuelve mediante factor integrante como sigue

$$R(t) = e^{\int -k_1 dt} = e^{-k_1 t}$$

Multiplicando ambos lados de la ecuación diferencial y reescribiendo los términos del lado izquierdo, se obtiene lo siguiente

$$\frac{d}{dt}\left(y(t)e^{-k_1 t}\right) = -e^{-k_1 t} k_3$$

Se integra ambos lados de la ecuación diferencial

$$\int \frac{d}{dt}\left(y(t)e^{-k_1 t}\right) dt = \int -e^{-k_1 t} k_3 dt$$

La solución general está dada por la ecuación (38)

$$y(t)e^{-k_1 t} = \frac{k_3}{k_1} e^{-k_1 t} + c$$

$$y(t) = \frac{k_3}{k_1} + ce^{k_1 t}$$

Reemplazado la condición inicial $y(0)=y_0$, se obtiene una solución particular

$$y(t) = \frac{k_3}{k_1} + \left( y_0 - \frac{k_3}{k_1} \right) e^{k_1 t}$$

Otro caso interesante, es cuando se elimina el término de la cosecha, $k_3=0$, y se mantiene el término que describe la aglomeración, de modo que la ecuación diferencial (33) resulta en una ecuación diferencial no lineal como sigue a continuación

$$\frac{dy}{dt} = k_1 y - k_2 y^2$$

Esta ecuación diferencial se puede resolver claramente mediante separación de variables

$$\frac{dy}{k_1 y - k_2 y^2} = dt$$

Si se integra ambos lados de la ecuación, se obtiene la siguiente expresión

$$\int \frac{dy}{y(k_1 - k_2 y)} = \int dt$$

Se aplica la integración de fracciones parciales en el lado izquierdo de la ecuación diferencial (42) y la solución general se aprecia en la ecuación (43)

$$\int \frac{1}{k_1 y} dy + \int \frac{k_2}{k_1(k_1 - k_2 y)} dy = \int dt$$

$$\frac{1}{k_1} ln(y) - \frac{1}{k_1} ln(k_1 - k_2 y) = t + c$$

$$\frac{1}{k_1} ln\left(\frac{y}{k_1 - k_2 y}\right) = t + c$$

$$y(t) = \frac{k_1 e^{k_1(t+c)}}{1 + k_2 e^{k_1(t+c)}}$$

Reemplazado la condición inicial $y(0)=y_0$, se obtiene el valor de la constante c descrito a continuación

$$c = \frac{1}{k_1} ln\left(\frac{y_0}{k_1 - k_2 y_0}\right)$$

Y la ecuación (45) representa una solución particular de la ecuación diferencial

$$y(t) = \frac{k_1 y_0 e^{k_1 t}}{k_1 - k_2 y_0 (e^{k_1 t} - 1)}$$

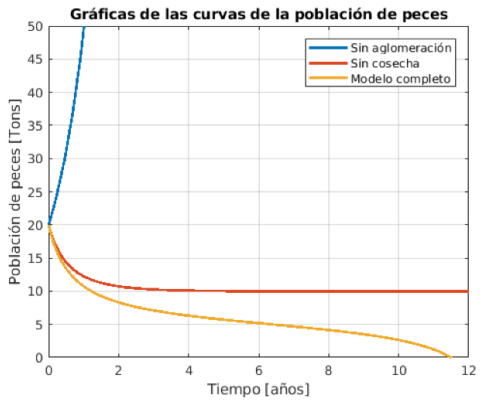
La Imagen 4.11. presenta la implementación de la ecuación diferencial (33) que no ha sido desarrollada analíticamente, utilizando el lenguaje de programación en MATLAB para su resolución. La Imagen 4.12. presenta los dos casos particulares que se han descrito previamente y el modelo poblacional completo.

Imagen 4.11. Implementación de las ecuaciones diferenciales (33), (34) y (40) en MATLAB

```
k1 = 1;                  % Constante de proporcionalidad 1
k2 = 1/10;               % Constante de proporcionalidad 2
k3 = 3;                  % Constante de proporcionalidad 3
N = 1000;                % Número de puntos
t0 = 0;                  % Condición inicial en t
y0 = 20;                 % Condición inicial en y
tmax = 12;               % Ancho de la ventana
t = linspace(0,1,N);     % Dominio de la función
% Ecuaciones diferenciales
yd1  = @(t,y) [k1*y-k3];
yd2 = @(t,y) [k1*y-k2*y.^2];
yd3 = @(t,y) [k1.*y-k2*y.^2-k3];
% Comando para resolver ecuaciones diferenciales
[t1, y1]  = ode45(yd1, [t0,tmax], y0);
[t2, y2] = ode45(yd2, [t0,tmax], y0);
[t3, y3] = ode45(yd3, [t0,tmax], y0);
% Comandos de graficación de la función solución
plot(t1,y1,t2,y2,t3,y3,'LineWidth',2)
axis([0 tmax 0 50])
grid on;
title('Gráficas de las curvas de la población de peces')
xlabel('Tiempo [años]')
ylabel('Población de peces [Tons]')
legend('Sin aglomeración','Sin cosecha','Modelo completo')
```

Fuente: Elaboración propia.

Imagen 4.12. Gráfica de las soluciones de las ecuaciones diferenciales (33), (34) y (40) en MATLAB




Fuente: Elaboración propia.

La Imagen 4.12. presenta la curva del modelo matemático que omite el término que describe la aglomeración: $k_1=1$, $k_2=0$ y $k_3=3$. Este modelo se dispara exponencialmente alcanzando una población de peces de 50 toneladas en menos de un año, partiendo de que inicialmente existían 20 toneladas. Este modelo describe una situación de recursos pesqueros ilimitados que está lejos de nuestra realidad. Bajo la misma condición inicial, la curva que omite el término de cosecha de peces, $k_1=1$, $k_2=0.1$ y $k_3=0$; tiende a mostrar un comportamiento de decrecimiento poblacional, alcanzando un valor estable de 10 toneladas de peces a partir de los 3 años en adelante. Cuando se considera el modelo completo con los efectos de aglomeración y cosecha, $k_1=1$, $k_2=0.1$ y $k_3=3$; se aprecia un decrecimiento más rápido de la población de peces llegando a una desaparición de estos recursos marinos al cabo de 11 años. Este último modelo matemático aproxima de mejor manera, una realidad de recursos marinos limitados que debe ser considerada por la pequeña y mediana industria pesquera.

## 4. Conclusiones

Se ha presentado detalladamente un enfoque matemático - conceptual a través del cual, se pueden ilustrar soluciones analíticas, gráficas y computacionales; de un pequeño grupo de problemas presentes en el campo de las ciencias agropecuarias. Los modelos determinísticos ofrecen una aproximación simple que permite la comprensión general de los elementos esenciales que actúan en el problema mediante el establecimiento de relaciones entre variables. Las ecuaciones diferenciales resuelven problemas que poseen comportamientos dinámicos implícitos en su naturaleza cambiante en el tiempo. Se han abordado principalmente ecuaciones diferenciales ordinarias de primer orden, tanto lineales como no lineales. El lector puede validar sus resultados a través del programa MATLAB que implementa diferentes métodos numéricos para la resolución de ecuaciones diferenciales. De esta manera, se brinda herramientas matemáticas importantes para la consolidación de una investigación portadora de una profunda fundamentación científica que orienten la modelización de los objetos de estudio en diferentes áreas profesionales.

## Referencia Bibliográfica


Austin, E. J., J. Willock, I. J. Deary, G. J. Gibson, J. B. Dent, G. Edwards-Jones, O. Morgan, R. Grieve, and A. Sutherland. 1998. "Empirical Models of Farmer Behaviour Using Psychological, Social and Economic Variables. Part I: Linear Modelling." *Agricultural Systems 58* (2): 203–24.

Bagni, Raul, Roberto Berchi, and Pasquale Cariello. 2002. "A Comparison of Simulation Models Applied to Epidemics." *Journal of Artificial Societies and Social Simulation 5* (3). http://jasss.soc.surrey.ac.uk/5/3/5.html.

Barnes, B., and G. R. Fulford. 2011. Mathematical Modelling with Case Studies: *A Differential Equations Approach Using Maple and MATLAB, Second Edition*. CRC Press.

Bober, William. 2013. *Introduction to Numerical and Analytical Methods with MATLAB® for Engineers and Scientists*. CRC Press.

Bock, Hans H. 1996. "Probabilistic Models in Cluster Analysis." *Computational Statistics & Data Analysis 23* (1): 5–28.

Borrelli, Robert L., and Courtney S. Coleman. 1998. "Differential Equations: A Modeling Perspective." John Wiley and Sons. http://www.sidalc.net/cgi-bin/wxis.exe/?IsisScript=uccma.xis&method=post&formato=2&cantidad=1&expresion=mfn=002946.

Fedeniuk, Ricky Wayne. 1998. "Oxytetracycline Degradation in Model Meat Processing Systems." http://ecommons.usask.ca/handle/10388/etd-10212004-001432.

Fishwick, Paul A. 2007. *Handbook of Dynamic System Modeling*. CRC Press.

Fowler, A. C., and Anna C. Fowler. 1997. *Mathematical Models in the Applied Sciences*. Cambridge University Press.

Greefrath, Gilbert. 2011. "Using Technologies: New Possibilities of Teaching and Learning Modelling – Overview." In *Trends in Teaching and Learning of Mathematical Modelling*, edited by Gabriele Kaiser, Werner Blum, Rita Borromeo Ferri, and Gloria Stillman, 1:301–4. International Perspectives on the Teaching and Learning of Mathematical Modelling. Dordrecht: Springer Netherlands.

Gurney, William, and R. M. Nisbet. 1998. *Ecological Dynamics*. Oxford University Press.

Hinrichsen, Diederich, and Anthony J. Pritchard. 2005. *Mathematical Systems Theory I: Modelling, State Space Analysis, Stability and Robustness*. Vol. 48. Springer Berlin.

Kutta, W. 1901. "Beitrag Zur Näherungweisen Integration Totaler Differentialgleichungen." http://www.citeulike.org/group/1448/article/813805.

Montgomery, Douglas C., and George C. Runger. 2010. *Applied Statistics and Probability for Engineers*. John Wiley & Sons.

Moore, Holly. 2014. *MATLAB for Engineers*. 4th ed. Upper Saddle River, NJ, USA: Prentice Hall Press.

Nomura, Edson Shigueaki, Francine Lorena Cuquel, Erval Rafael Damatto Junior, Eduardo Jun Fuzitani, and Ana Lúcia Borges. 2017. "Fertilization with Nitrogen and Potassium in Banana Cultivars 'Grand Naine', 'FHIA 17' and 'Nanicão IAC 2001' Cultivated in Ribeira Valley, São Paulo State, Brazil." *Acta Scientiarum. Agronomy 39* (4): 505–13.

Quarteroni, Alfio, Fausto Saleri, and Paola Gervasio. 2014. *Scientific Computing with MATLAB and Octave*. Springer Science & Business Media.

Runge, Carl. 1895. "Über Die Numerische Auflösung von Differentialgleichungen." *Mathematische Annalen* 46 (2). Springer: 167–78.

San Martín Cuenca, Hugo Dennys, and Eduardo Alejandro Tusa Jumbo. 2015. "Fundamentos de Programación Para Ciencias E Ingeniería." Machala: Ecuador. http://repositorio.utmachala.edu.ec/handle/48000/6748.

Zill, Dennis G. 2016. *Differential Equations with Boundary-Value Problems*. Nelson Education.

05 Capítulo

# Estadística básica con datos agropecuarios

Irán Rodríguez Delgado; Bill Serrano;
Diego Villaseñor Ortiz

En la historia de la humanidad la estadística ha jugado un rol determinante en el procesamiento y análisis de datos en las áreas del conocimiento; y su contribución en la formulación de nuevas teorías ha sido decisiva, especialmente en las ciencias agropecuarias. Sin embargo, el desarrollo actual de nuevos y sofisticados softwares estadísticos que simplifican los procesos de análisis exigen a estudiantes, profesores e investigadores, además de una sólida base teórica-práctica, una actualización constante, enfocada en el conocimiento de los procedimientos de análisis de datos encaminados a apoyar la toma de las mejores decisiones.

**Irán Rodríguez Delgado:** Ingeniero Agrónomo (1992) Universidad Central de Las Villas, Cuba Magister en Agricultura Sostenible (2009) Universidad de Cienfuegos, Cuba; Investigador Agregado (2009) Instituto de Investigaciones de la Caña de Azúcar, Cuba; Profesor Titular (2015) Universidad Técnica de Machala. Autor de cuatro libros y 17 artículos publicados.

**Bill Serrano:** Ingeniero Agrónomo e Ingeniero en Gestión Empresarial, Magister en Administración de Empresas y estudiante doctoral en Análisis Económico y Estrategia Empresarial en la Universidad A Coruña. Fue Gerente de Almacén y Jefe Comercial Zonal en ICESA, Gerente de producto en ICESA y COMPTECO. Actualmente Profesor Titular en la Universidad Técnica de Machala.

**Diego Villaseñor Ortiz:** Profesor Titular de la Universidad Técnica de Machala (UTMach), es Ingeniero Agrónomo, con Maestría en Ciencias Agronómicas con mención en suelos, obtenida en la Universidad de Concepción (Chile). Actualmente es parte del programa de doctorado en Ciencias del suelo y nutrición de plantas en la Universidad Estadual Paulista (Brasil).

En el presente capítulo se profundiza sobre los tipos de variables, sus atributos y escalas de medición, como elementos indispensables a tener en cuenta en la elección del procedimiento estadístico a desarrollar, además se detalla todo lo relacionado con las medidas de resumen de datos y los elementos asociados con la estimación puntual de parámetros y los intervalos de confianza, unido a la explicación de los pasos para la ejecución de algunas pruebas estadísticas paramétricas y no paramétricas. Adicionalmente, se establecen las pautas para la construcción e interpretación de tablas de contingencia bidimensionales.

## La estadística como ciencia

### Concepto

La definición y conceptualización de la *estadística* como ciencia es muy amplia y diversa, y se encuentra asociada con la experiencia de cada profesional en su área del conocimiento. La *estadística*, según Steel & Torrie (1985) comenzó en sus inicios como una aritmética estatal de apoyo a los gobernantes para la recaudación de impuestos y para Barnett (1991) es la ciencia encargada de estudiar la forma en que se emplea la información y ofrecer el procedimiento ante situaciones prácticas que envuelven experimentos aleatorios. Johnson y Kuby (2012) definen a la estadística como el lenguaje universal de las ciencias, ya que es aquella que involucra información numérica y gráfica donde se resume su comportamiento y a partir de la cual se efectúa la interpretación en cualquiera área del conocimiento.

Batanero (2001) señala que la *estadística* se ha dividido clásicamente en dos segmentos; *estadística descriptiva*, la que permite realizar resúmenes del conjunto de datos con el objetivo de caracterizar y describir las variables objeto de estudio, sin extender sus resultados a una población; y la *estadística inferencial*, la cual estudia los resúmenes de datos con referencia a un modelo de distribución probabilístico y su finalidad es inferir el comportamiento de la población a partir de los resultados en la muestra. Sin embargo,

plantea que, en la actualidad esta segmentación se considera como una división simplista y lo más común es definirla como *análisis de datos*.

La estadística es la ciencia que estudia el conjunto de métodos, procedimientos y modelos utilizados para recolectar, organizar, clasificar, procesar, resumir, representar y analizar datos extraídos de una población o muestra representativa de la población de estudio, con el objetivo de realizar estimaciones válidas y obtener conclusiones necesarias para la toma de decisiones.

Por otro lado, Salcedo (2013) plantea que la *estadística*, en su vinculación con la investigación, le ofrece al profesional en formación, herramientas fundamentales que le permiten desarrollar competencias investigativas, al obtener conocimientos, habilidades y destrezas encaminadas a que puedan utilizarlas en la búsqueda de soluciones a situaciones problemáticas creadas en su entorno.

Desde nuestra *perspectiva* la estadística es una herramienta necesaria en la investigación científica que posibilita aplicar diferentes procedimientos en cada momento, en su tránsito por la línea de investigación, encaminados a realizar una interpretación adecuada de los procesos estudiados con un fundamento matemático, con la intención de apoyar la toma de las mejores decisiones.

Por consiguiente, y al tomar en cuenta que toda línea de investigación transita por diferentes niveles de la investigación (exploratorio, descriptivo, relacional, explicativo, predictivo y aplicativo) se entiende que, en cada momento dentro del proceso investigativo, al realizar el análisis de datos, se aplican diferentes procedimientos estadísticos (Supo, 2017). En el nivel exploratorio no se utiliza la estadística como herramienta, ya que solo se identifican y descubren nuevos problemas, es fenomenológico e interpretativo; en este nivel solamente se realiza investigación cualitativa, la cual precede a la investigación cuantitativa; en el descriptivo el procedimiento estadístico es univariado y solo se realiza una descripción o caracterización de una o más variables objeto

de estudio, ya sea numérica o categórica, sin la intensión de compararlas o buscar una relación o asociación entre ellas (no es inferencial ya que no es necesario tomar una muestra); o sea que, aunque se estudien dos o más variables (que es lo que más ocurre), la intensión en este nivel es describir cada variable de forma individual, o sea, cuantificar su frecuencia y no compararlas o determinar su grado de correlación o asociación. En el nivel descriptivo inicia la investigación cuantitativa. En el nivel relacional (es bivariado e inferencial) se cuantifica la relación o el grado de asociación entre dos variables, pero no demuestra relaciones de causalidad. En el nivel explicativo se buscan relaciones de causalidad entre dos o más variables (estudios de causa-efecto). En el nivel predictivo se busca predecir el comportamiento de las variables y se construyen modelos predictivos. En el nivel aplicativo se interviene y se realiza estadística para el control de calidad en los procesos. Por lo tanto, la segmentación que se realice de la estadística debe ser enfocada en el nivel investigativo donde se encuentre dentro de la línea de investigación.

## Otros conceptos básicos

Para un mejor entendimiento del texto que se aborda se expresan varios conceptos, los cuales constituyen una base necesaria para la posterior comprensión de los diferentes ejercicios prácticos que se analizan y explican.

Datos: constituyen valores finales de medición, recolectados de la aplicación de instrumentos de medición sobre el fenómeno estudiado y la base del desarrollo de la investigación.

Población o universo: conjunto completo de individuos o elementos acotados en un tiempo y en un espacio determinado, que poseen alguna característica común observable o medible que se desea estudiar o analizar (puede ser finita cuando se conoce el número total de elementos que la componen o infinita si no se conoce). Se encuentra constituida por los objetos (tangibles o intangibles) que tienen

en común la característica de interés (la variable en estudio). Generalmente es poco común realizar estudios con el total de la población, ya que se presentan inconvenientes tales como el incremento del tiempo para la recolección de datos y su análisis, además del aumento de los costos.

Población de estudio: conjunto de unidades de estudio que cumplen criterios de selección y a partir de la cual se efectúa el cálculo del tamaño muestral. Son todos los resultados posibles de medir la característica de interés en cada objeto del universo.

Tamaño poblacional (N): número de elementos que conforman una población finita.

Parámetro: valor numérico que describe o resume todos los resultados posibles de una característica de interés en la población estudiada. Normalmente se denotan con letras griegas. Según Canavos (1988) un parametro "es una caracterización numérica de la distribución de la población de manera que describe, parcial o completamente, la función de densidad de probabilidad de la característica de interés". Pardo y Ruíz (2005) lo definen como "un valor numérico que describe una característica poblacional".

Individuo: personas u objetos que contienen cierta información que se desea estudiar.

Muestra: subconjunto representativo de elementos de una población de estudio utilizado para realizar el análisis de datos y extrapolar las conclusiones obtenidas a dicha población.

Estadígrafo o estadístico: función definida sobre los valores numéricos que resumen los datos de una muestra (valor variable).

Unidad muestral: cada elemento o entidad que conforma la muestra.

Tamaño muestral (n): número de elementos de la población de estudio que conforman la muestra.

Constante: características de la población que no cambian ni en su estado ni expresión.

## Las variables y su clasificación

El conocimiento acerca del origen, características y propiedades de las variables constituye uno de los elementos más importantes en cualquier proceso investigativo que se desarrolle en los diferentes campos del saber, ya que las distintas escalas de medición que alcancen, definirán el procesamiento estadístico a realizar sobre los datos obtenidos, lo cual redundará en una correcta interpretación de la información generada y por consiguiente la toma de decisiones eficientes y eficaces.

La medición u observación de las variables generan datos, ya sean numéricos o categóricos, los cuales pueden alcanzar diferentes valores y atributos, constituyéndose en uno de los elementos más importante en el desarrollo de la investigación científica.

Todo proceso de investigación queda determinado por el número y naturaleza de las variables que se incluyan en el estudio; cuanto mayor sea el número de variables introducidas y controladas por el investigador, mayor será la significación matemática de los resultados que se generen en dicha investigación.

Dominar todo lo relacionado con las variables y su clasificación es básico a la hora de lograr un entendimiento claro del análisis estadístico que se debe aplicar, el cual puede ser diferente en función del valor final de medición obtenido y del tipo de distribución probabilística que presenten los datos.

En cualquier campo del conocimiento científico es muy usual trabajar con individuos diferentes unos de otros y para poder estudiarlos es necesario otorgarles un valor, lo cual es precisamente el papel de las variables.

El desarrollo de esta temática se centrará específicamente en los valores finales de medición (VFM) de variables per-

tenecientes a sistemas de producción agropecuarios y en ejemplos de investigaciones desarrolladas por estudiantes y profesores en las carreras de Medicina Veterinaria y Zootecnia, Ingeniería Agronómica, Ingeniería Acuícola y Economía Agropecuaria, pertenecientes a la Unidad Académica de Ciencias Agropecuarias de la Universidad Técnica de Machala, provincia de El Oro, Ecuador.

## Variable estadística. Definición

Es una propiedad, atributo o característica con respecto a la cual los individuos o elementos de una muestra, o de un grupo poblacional (sujetos u objetos) se diferencian en algo verificable y cuya variación puede ser observada o medida en las unidades de estudio (Hernández Sampier *et al.*, 2010), por lo que pueden obtenerse valores finales de medición diferentes en uno u otro, o modificarse en el propio sujeto u objeto en el transcurso de la investigación.

Desde el punto de vista investigativo una variable es una característica observable o medible en las unidades de estudio, de las cuales se generan los datos o valores finales de medición.

## Tipo de variables. Clasificación

La clasificación de las diferentes variables que se presentan en la investigación científica depende de varios criterios, complementados unos con otros y asociados de forma general a su forma de expresión.

## Naturaleza de los datos

El origen natural de los datos obtenidos en cualquier proceso investigativo genera diferentes posibilidades de análisis, el cual puede ser diferente para cada variable. Es por ello, que las variables por su naturaleza se clasifican en:

1. Cuantitativas o métricas: son aquellas cuya magnitud puede ser medida en términos numéricos, o sea, que invo-

lucran una medición numérica. Pueden ser continuas y discretas o discontinuas.

Continuas: variable cuyos valores son posibles dentro de cualquier intervalo. En principio pueden alcanzar infinitos valores fraccionados.

**Propiedades**

– Puede asumir un número incontable de valores.

– Alcanza un número infinito de valores entre dos puntos fijos en función de la precisión que se utilice en el estudio.

– Nunca puede ser medida con exactitud; el valor observado depende en gran medida de la precisión de los instrumentos de medición.

– Con una variable continua se presenta inevitablemente un error de medida, por ejemplo, la estatura de una persona (1,67 m; 1,675 m; 1,6758 m), en los cuales siempre se puede presentar un valor intermedio asociado con la cantidad de decimales que se utilicen.

– Son las que se obtienen de mediciones. Pueden ser representadas con números enteros (cuando se redondean) o fraccionarios.

Ejemplo: se obtuvo el peso de tomates en gramos y se alcanzaron los siguientes valores: 80,5 g y 80,6 g; sin embargo, entre los valores encontrados, si agregamos otro decimal, puede existir otro valor que puede ser 80,55 g.

Discretas o discontinuas: son aquellas variables que solo pueden alcanzar un determinado conjunto de valores dentro de su distribución de datos, los cuales serían discontinuos o enteros, pero nunca fraccionados.

**Propiedades**

– Entre las categorías de la variable no se puede introducir una modalidad intermedia, únicamente aquellos datos que pertenecen al conjunto.

– Los valores que toma esta variable se encuentran dentro de un conjunto numerable o finito de puntos.

– Puede asumir un número contable de valores. No pueden tomar valores intermedios.

– Son aquellos valores finales que se obtienen de efectuar un conteo.

Ejemplo: el número de árboles en un agroecosistema determinado, cuyas categorías pueden ser 10, 30 o 50; solo alcanzan cifras exactas ya que no puede existir un árbol y medio o 10,5 árboles.

2. Cualitativas o no métricas: son propiedades o atributos que no pueden ser medidos y solamente cuando se asocian a una frecuencia pueden tratarse de forma estadística. Pueden ser dicotómicas o politómicas.

Dicotómicas: son aquellas que tienen dos opciones de respuestas, o sea, dos categorías. Ejemplo: el sexo de animales, cuyas categorías son machos y hembras.

Politómicas: son aquellas que tienen más de dos opciones de respuesta o categorías. Ejemplo: índice de infestación de una plaga en una granja agrícola determinada, cuyas categorías pueden ser alta, media o baja.

No se debe confundir la clasificación descrita anteriormente con los tipos de investigación que se desarrollan en cualquier área del conocimiento, ya que cuando se refiere a investigación cualitativa es aquella donde no se utiliza la estadística como herramienta y pertenece al nivel investigativo exploratorio; y la investigación cuantitativa es aquella que emplea la estadística como herramienta y pertenece a los niveles descriptivos, relacional, explicativo, predictivo y aplicativo.

## Escalas de medición

Los valores finales que se obtienen luego de medir una variable presentan distintos atributos dentro de los cuales se encuentran el orden, la distancia y el origen, los cuales

brindan información sobre el tipo de variable en cuestión y condicionan el tipo de análisis estadístico a realizar. Restringido a ello, las variables por sus escalas de medición se clasifican en:

1. Categóricas: reciben este nombre porque sus VFM son categorías. Pueden ser nominales u ordinales.

Nominales: aquellas que no tienen ningún atributo o algún orden en particular.

**Propiedades**

- Caracteriza (describe, identifica, nombra, nomina) a un sujeto u objeto, de una muestra o población, en una categoría, sin que exista un orden implícito entre ellas.
- No tiene magnitud ni intervalo.
- Lo que estudia o representa la variable solo puede agruparse en categorías exhaustivas y mutuamente excluyentes.
- Una categoría de esta variable no es más que la otra, no existe un orden jerárquico, solo son diferentes.
- A cada una de las categorías de la variable se le asignan atributos que pueden ser tanto nombres como números (cuando se utilizan tienen un carácter simbólico).
- Representan el nivel más bajo de medición.
- Con la información generada por este tipo de variable no pueden realizarse las operaciones aritméticas habituales (suma, resta, multiplicación y división).

Ejemplo: variedades de soya (INIAP 305, INIAP 308, INIAP 310, entre otras).

Ordinales: son aquellas que cuentan con un orden en sus categorías como único atributo.

**Propiedades**

- Ordena o clasifica a los sujetos u objetos según posean más, menos o la misma cantidad de la variable que se mide.

– Tiene magnitud, pero no intervalo.

– Define categorías al establecer una relación mayor o menor que, o de igualdad/desigualdad.

– Es posible establecer un orden lógico entre ellas, lo que constituye su único atributo.

– Categorías que conllevan a una jerarquía, una es más o menos que la otra, aunque no permite cuantificar la distancia entre una categoría y otra.

– No se conoce la diferencia real de la magnitud entre las categorías de la variable ya que no es cuantificable o medible.

– No se pueden realizar con estas variables las operaciones aritméticas habituales (suma, resta, multiplicación y división).

Ejemplo: nivel de infestación de una plaga que puede ser leve, moderado o intenso, sin embargo, no se conoce la magnitud de la diferencia que se presenta entre uno u otro nivel.

2. Numéricas: reciben este nombre porque sus VFM son unidades, o sea números. Pueden ser de intervalo o de razón.

Intervalo: escala métrica que conserva las características de orden de la escala ordinal y se le agrega el atributo distancia.

**Propiedades**

– Incluye en sus VFM el cero absoluto, es decir que el cero es simplemente arbitrario o relativo y en realidad no significa ausencia de la variable, sino que es un nivel más de medición de la variable en cuestión.

– Tiene intervalos iguales y medibles. No tiene un origen real, por lo que puede asumir valores negativos.

– No solo indica que las temperaturas 15°C y 30°C son distintas y que 30°C es mayor que 15°C (orden), sino que, además, agrega una nueva información al plantear que 30°C es cualitativamente tan distinto de 15°C como lo es 15°C de 0°C (distancia).

- Son más informativas que las variables ordinales y nominales.
- No permiten multiplicación o división.

Ejemplo: la temperatura en una región determinada (puede ser 0˚C, -10˚C, 22˚C).

Razón: escala métrica que conserva las características de orden y distancia de la escala de intervalo y se le agrega el atributo origen.

**Propiedades**

- Tiene intervalos constantes entre valores, además de un origen real.
- El cero significa la real ausencia de la variable, aunque no del individuo.
- Tanto cero metros, como cero cantidad de vástagos en una planta, significa ausencia de altura y de ahijamiento, y no interesa que el primer valor corresponda a un individuo inexistente y el segundo a una planta que existe.
- Son las que mayor cantidad de información ofrecen.
- Permiten realizar las operaciones aritméticas habituales como suma, resta, división y multiplicación.

Ejemplo: las variables altura de la planta a los 60 días (cm) o peso de cerdos al sacrificio (kg).

Es importante significar que el análisis estadístico que se desarrolla sobre variables nominales no es el mismo que se aplica sobre variables ordinales, aunque si es igual para las variables de intervalo o de razón, a excepción del coeficiente de variación, el cual no puede ser calculado en variables donde el 0 es un valor más de variable (intervalo). Sin embargo, aunque una variable numérica de intervalo es de forma teórica diferente a una numérica de razón, en la práctica se utiliza el mismo tipo de técnica estadística.

Una variable cualitativa puede ser dicotómica o politómica medida en escala nominal u ordinal (categóricas) y una

variable cuantitativa puede ser discreta o continua medida en escala de intervalo o de razón (numéricas).

## Relación causa-efecto

En el ámbito de la investigación científica la experimentación surge cuando el investigador manipula una o varias variables (Montero y León, 2005) con la finalidad de detectar su influencia en otras variables medidas u observadas; por lo que dominar sus características y propiedades constituye un elemento importante que facilita el logro de una interpretación correcta del proceso estudiado. Su objetivo es demostrar relaciones de causalidad. De acuerdo al papel que juegan en el problema o propósito de la investigación y en el diseño experimental las variables se clasifican en variables dependientes (VD), variables independientes (VI) y variables intervinientes.

1. Dependientes: conocidas también como las variables de medida, exógenas, de respuesta, de estudio o de resultado.

### Características

- Depende del valor que asuman otros fenómenos o variables independientes.
- Su variabilidad está condicionada por la VI y por otras variables intervinientes.
- La VD es aquella que es observada o medida para determinar el efecto de la causa de variación manipulada por el investigador (VI).
- Es la variable que se desea caracterizar o explicar y en muchos casos optimizar en función de la modificación del o los factores de estudio (VI).
- En los estudios investigativos pueden observarse o medirse una o varias VD, debido a que la manipulación de una o dos VI pueden influir en varias particularidades de la unidad muestral.

– En general las respuestas se representan con la letra Y ($Y_1$, $Y_2$,....Ym). La VD se ubican en el eje de ordenadas (eje Y).

Ejemplos: rendimiento (t $ha^{-1}$) de un cultivo en la cosecha, peso de camarones (g) a los cuatro meses de edad, cantidad de leche (L) producida diariamente por cada vaca en un hato ganadero.

2. Independientes: conocidas también como variables manipuladas o controladas por el investigador, variables explicativas, exógenas o regresoras. Constituye el o los factores de estudio en una investigación experimental.

**Características**

– Es aquella propiedad o característica que se supone es la causa de variación del fenómeno estudiado.

– Es aquella cuyo valor no depende de otra variable, sino del criterio del investigador al estructurar su diseño de investigación.

– Los cambios en los valores o atributos de este tipo de variable determinan cambios en los valores de otra.

– En investigación experimental se denomina de esta manera a la variable que el investigador modifica en función del estudio que desarrolla y es aislada de cualquier otro factor.

– La VI es la que el investigador escoge para establecer los grupos en el estudio, aunque normalmente se utilizan uno o dos factores, ya que estudiar simultáneamente tres o más en un experimento imposibilita realizar una interpretación correcta de la influencia de cada uno en los resultados finales.

– La VI o variables explicativas se representan con la letra X ($X_1$, $X_2$,....Xp) y se ubican en el eje de las abscisas (eje X).

Ejemplo: se necesita conocer el efecto de la fertilización con nitrógeno (N) en el cultivo de la caña de azúcar, para lo cual se estudian varias dosis (40 kg $ha^{-1}$ de N, 60 kg $ha^{-1}$ de N y 80

kg $ha^{-1}$ de N). El factor de estudio es la fertilización nitrogenada y los niveles del factor de estudio constituyen las diferentes dosis utilizadas, que en experimentación se denominan tratamientos.

3. Intervinientes: los resultados de las variables de estudio (VD) pueden ser afectadas por los valores o la interposición de otras variables controladas o no por el investigador durante la investigación. Estas variables permiten determinar los indicadores de variabilidad.

**Características**

- Es aquella que determina las relaciones entre dos o más variables.
- Por su condición se interpone entre la VI y la variable dependiente (VD).

Las variables intervinientes pueden ser confusoras, intermedias o de control.

Confusoras: propia de estudios observacionales en los cuales el investigador no interviene, su aparición puede intensificar o antagonizar la relación aparente entre el problema y una posible causa. Influye sobre la VI y la VD. Ejemplo: variación genotípica de las plantas.

Intermedias: aparecen de manera inesperada, y por tanto es metodológicamente incontrolable su acción entre el factor causal y el efecto. Su naturaleza es aleatoria, no se conoce su distribución antes de efectuar la recolección de datos. Casi siempre es numérica y se denominan covariables. Ejemplos: precipitación y temperatura promedio en un periodo determinado.

Control: posee fuerte influencia sobre la VD y ningún efecto sobre la VI. Se identifica en el momento de la planeación de la investigación. En estudios observacionales su control se realiza mediante los criterios de exclusión y en los experimentales mediante la aplicación de la técnica de bloqueo. Aunque pueda tener algún tipo de influencia en la VD no se estudia como factor causal. Ejemplo: variación de la ferti-

lidad del suelo en un estudio sobre el comportamiento de diferentes variedades evaluadas a partir de la producción de biomasa al momento de la cosecha.

Por otro lado, si en un estudio experimental determinado, se necesita conocer el efecto de la fertilización en el rendimiento agrícola de un cultivo comercial, en la medida que se introduzcan cambios deliberados en las variables de entrada y se controle más de una variable, mayor será el poder predictivo y explicativo del objetivo de estudio, el cual se encamina a identificar las posibles causas de variación en las variables de salida, de tal manera que si desea explicar el efecto de la aplicación de diferentes fórmulas de fertilizantes (VI) en el rendimiento agrícola del cultivo (VD) se deben introducir en el estudio la medición de variables intervinientes tales como, temperatura, humedad relativa, luminosidad, actividad de los microrganismos en el suelo, profundidad del manto freático, entre otras. El modelo representado en la Imagen 5.1. idealiza un proceso o sistema, con variables controlables e incontrolables, que transforma alguna entrada (a menudo un material) en una salida que tiene una o más respuestas observables (Montgomery, 1991).

Imagen 5.1. Esquema que muestra el papel de los diferentes tipos de variables cuando se estudian relaciones de causalidad en un proceso investigativo.


Fuente: Modificado de Montgomery (1991).

## Categorización de variables

Las variables numéricas, ya sean discretas o continuas, pueden transformarse en categóricas al perder sus atributos de medición, a lo cual se le denomina categorización de la variable y se utiliza cuando se necesita realizar algún diagnóstico de una situación determinada, lo que puede ayudar a la toma de decisiones.

La transformación de una variable es posible siempre que sea en una de menor jerarquía y de forma general se presenta el inconveniente de la perdida de información, por lo que en la práctica siempre es conveniente, cuando sea el caso, medir las variables en la forma que ofrezcan la mayor información posible.

Ejemplo: en una clínica veterinaria se obtuvo el peso en kilogramos (variable numérica de razón) de 30 animales tratados en una semana y con fines de diagnóstico se agrupan de acuerdo al peso en: raza pequeña (hasta 5 kg de peso), raza mediana (a partir de 5 y hasta 20 kg) y raza grande (más de 20 kg), por lo que se define una nueva variable categórica ordinal ya que perdió sus atributos de origen y distancia, y posteriormente se agruparon en función de su padecimiento a *Babesia canis*, construyéndose dos categorías, enfermos con *B. canis* y no enfermos con *B. canis* (se define una nueva variable categórica nominal al perder el atributo orden) (anexo 1).

## Medidas de resumen de datos

En esta sección se analizan algunos de los fundamentos teóricos de las diferentes medidas utilizadas para resumir datos, relacionados con sus características, propiedades, ventajas, desventajas e importancia práctica, lo cual constituye la base teórica que permite realizar una correcta interpretación de los resultados obtenidos mediante el uso de un procesador estadístico, ya sea el Statistical Package for the Social Science (SPSS) en su versión 24 de prueba para Windows (se utiliza para el desarrollo de cada procedimiento estadístico descrito) u otro programa de preferencia. Aunque es importante

profundizar en estos conocimientos en la amplia bibliografía existente (Aguirre y Vizcaino, 2010; Castañeda, 2010; Johnson y Kuby, 2012; Lind *et al.*, 2015) y en bases de datos de acceso abierto disponibles, entre las que se encuentran: http://www.fao.org/faostat/es/#data/QC/visualize; http://www.ecuadorencifras.gob.ec/ y http://gel.eppo.int).

Los fenómenos biológicos no suelen ser constantes por lo que lo primero que se debe conocer en una población o muestra dada son sus parámetros o estadísticos correspondientes, elementos que permiten realizar una descripción adecuada de una variable determinada.

## Medidas de tendencia central

Las medidas de tendencia central son aquellas que facilitan obtener información sobre el conjunto de datos que se analiza; permiten conocer cuan agrupados se encuentran los valores que ha tomado la variable estudiada respecto al valor medio o promedio. Indican hacia donde apunta en general la distribución de datos y permiten identificar los valores más representativos.

Los principales métodos utilizados para ubicar el punto central de una distribución de datos son la media aritmética, la mediana y la moda.

## Media aritmética

Aunque existen varias medias, como la ponderada, hipergeométrica, cuadrática y armónica, la media aritmética es la más utilizada entre todas las medidas de resumen de datos. Se representa por la letra X con una barra horizontal encima ($\overline{X}$) para los datos muestrales y por la letra griega mu ($\mu$) para distribuciones de datos poblacionales. Solamente puede calcularse en datos numéricos. Se define como la sumatoria de todas las puntuaciones de una distribución de datos, dividida por el número total de casos.

### Propiedades

– Es única, o sea, que cada conjunto de datos posee una sola media.

– Representa un valor alrededor del cual oscilan todos los valores de la variable medida, es el valor medio de todos los datos, por lo que también se le denomina promedio.

– Tiene la ventaja de ser utilizada en procedimientos estadísticos como la comparación de medias de varios conjuntos de datos.

– Es apropiada para variables numéricas medidas en escala de razón.

– Es la única medida donde la suma de las desviaciones de cada valor respecto a la media es igual a cero. Se considera un punto de equilibrio en el conjunto de datos (Lind *et al*., 2004).

– Para su cálculo se utilizan todos los valores de la serie de datos, por lo que no se pierde ninguna información.

La media presenta la desventaja de que su valor puede estar influenciado o afectado por valores extremos o atípicos, denominados outliers en inglés (Milton, 1994). Según Maronna (1995) la media es muy sensible a valores extremos, por lo que no es robusta. Los valores de la distribución de datos pueden ser muy pequeños o muy grandes; al alejarse en exceso del resto de la serie de datos pueden condicionar en gran medida el valor de la media o promedio, por lo que puede perder representatividad. El investigador puede optar por realizar los cálculos y tenerlos en cuenta o no, aunque debe realizar la aclaración.

## Mediana

Es el valor central de los datos, es decir, supuesta la muestra ordenada en forma ascendente o descendente, es el valor de la serie de datos que divide en dos partes iguales a la población o muestra y se sitúa justamente en el centro de la mues-

tra (50% de valores son inferiores y otro 50% son superiores). Cuando el número de casos es impar la mediana es el valor que se encuentra en la posición central de la distribución de datos y cuando el número de casos evaluados es par, la mediana se obtendrá del valor medio de las dos observaciones que se encuentran en el centro del conjunto de datos.

**Propiedades**

– Es única y siempre existe.

– Puede determinarse en variables numéricas de intervalo o de razón y en las categóricas ordinales (Lind *et al.*, 2004).

– No presenta el problema de estar influenciada por valores extremos ya que no depende del valor que toma la variable, sino del orden de las mismas, por ello es adecuado su uso en distribuciones de datos asimétricas.

– Es mejor utilizar la mediana que la media cuando se trata de un conjunto de datos en el cual existen valores extremos o sesgados, o sea, en distribuciones asimétricas, ya que proporciona una medida de tendencia central más exacta.

**Moda**

Es el valor o la categoría de una variable que se presenta u ocurre con la mayor frecuencia, o sea, el que más se repite. Es una medida de centralización que tiene sentido estudiar en una variable cualitativa o cuantitativa. Para determinar la moda no necesita realizar ningún cálculo.

**Propiedades**

– Se puede utilizar en cualquier escala de medición.

– Si bien a simple vista no se observa la centralidad de la moda, debemos indicar que, en un grupo normal, la mayoría de los datos se encuentran cercanos a un punto central, por lo que se presume que el dato que más se repite estará cercano a este punto.

– No existe moda si todos los valores son diferentes o si se presentan el mismo número de veces.

– Para distribuciones simétricas unimodales, la media, la mediana y la moda corresponden al mismo valor.

– Pueden existir uno o más valores modales. Si se presentan dos valores con la mayor frecuencia sería una distribución de datos bimodal y si existen más de dos valores se define como multimodal, aunque en este caso su valor pierde representatividad y resulta muy difícil realizar interpretaciones.

## Medidas de posición

Representadas por los cuantiles y definidos como un valor observado de la variable en la muestra por debajo del cual se encuentra una frecuencia acumulada $k$, o sea, que el número de valores menores o iguales a él constituyen la proporción $p$ del número total de observaciones en la muestra. Se determinan mediante un método que obtiene la ubicación de los valores que dividen un conjunto de observaciones en partes iguales.

Dentro de los cuantiles se encuentran los percentiles, cuartiles y deciles.

### Percentiles

Un percentil de orden $k$ es igual a un cuantil de orden k/100, o sea, que para obtenerlo se divide la distribución de datos en 100 partes iguales.

### Cuartiles

Se determinan mediante la división de la distribución de datos en cuatro partes iguales, obteniéndose cuatro grupos con frecuencias similares (25%) y tres puntos de división denominados cuartiles (Anderson *et al*., 2008).

– Primer cuartil ($Q_1$) es igual al percentil 25 ($P_{25}$).

– Segundo cuartil ($Q_2$) es igual al percentil 50 ($P_{50}$). Al ser el valor que parte la distribución de datos en dos siempre será igual a la mediana.

– Tercer cuartil ($Q_3$) es igual al percentil 75 ($P_{75}$).

**Deciles**

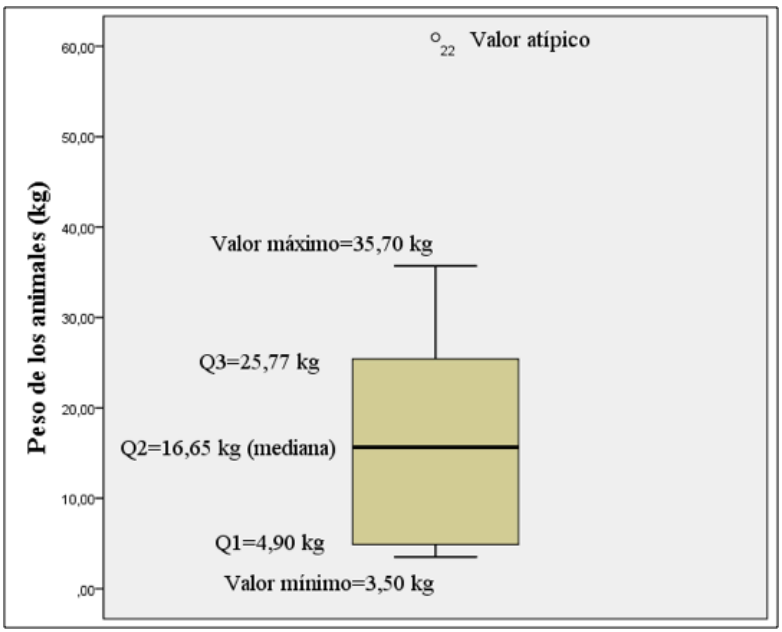
Para obtenerlo se divide entre 10 la distribución total de datos. En cualquier distribución de datos que se estudie el valor del Decil 5 ($D_5$) sería el mismo que el $P_{50}$ (mediana) y que el $Q_2$.

La utilización práctica de las medidas de posición se centra en conocer los porcentajes de casos que se encuentran por debajo o por encima de un punto dentro del conjunto de datos.

Ejemplo: se desea conocer los valores que representan las diferentes posiciones en la distribución de datos de la variable peso de los animales (kg) atendidos en una clínica veterinaria en una semana (anexo 1), para lo cual se elaboró un diagrama de cajas y sesgos con la utilización del SPSS.

Procedimiento estadístico: abrir la base datos con el SPSS>desplazarse en la barra de menú y seleccionar gráficos>generador de gráficos>aceptar>en galería se elige el gráfico que se desea elaborar, en este caso diagrama de cajas y de las tres opciones que se muestran se marca la opción de diagrama de caja 1-D por tratarse de una sola variable y se traslada hacia la vista previa del gráfico, y se arrastra la variable peso de los animales hacia el eje X>aceptar. Se genera el diagrama de cajas (Box plot) (Tukey, 1977), que puede ser editado al ser seleccionado y dar doble clic. En el visor de resultados del SPSS se muestra el diagrama de cajas y sesgos solicitado (Imagen 5.2.).

Imagen 5.2. Diagrama de cajas (Box plot) que muestra el valor de los cinco puntos que lo integran y la posición del valor atípico dentro de la distribución de datos.




Interpretación: primeramente, se observan los valores, mínimo (3,50 kg) y máximo (35,70 kg), de la distribución, que indican un amplio rango de valores (32,20 kg), además, se tiene que el 25% de los datos se encuentran por debajo de 4,9 kg ($Q_1$), el 50% se encuentran por debajo y por encima de 16,65 kg (valor mediano o $Q_2$) y el 75% se encuentran por debajo de 25,75 kg. Se evidencia una mayor dispersión de datos hacia la parte superior ya que el sesgo es más estirado, además se muestra el valor atípico, el cual se encuentra en la posición 22 de la base de datos y que corresponde a 61,0 kg.

## Medidas de dispersión o variabilidad

Son las medidas que permiten conocer la dispersión o variabilidad de todos los datos recolectados (Lind *et al*., 2004). Se utilizan para analizar la representatividad de las medidas de tendencia central (Gorgas *et al*., 2011). Los valores de las medidas de variabilidad se incrementan cuando los datos

son más dispersos y disminuyen cuando se encuentren más agrupados alrededor del punto central.

Las medidas de dispersión permiten conocer, como los valores de los datos se distribuyen a través del eje X mediante un valor numérico que representa el promedio de la variabilidad de los datos, por lo que facilitan calificar la confiabilidad de la medida de tendencia central.

Dentro de las medidas de dispersión o variabilidad se encuentran el rango, la varianza, la desviación típica o estándar, el error estándar o típico de la media y el coeficiente de variación.

## Rango

Conocido también como amplitud total, mide el recorrido total de los valores en la muestra. Se denota con la letra R o como AT. Es el límite dentro del cual se encuentran comprendidos todos los valores de la serie de datos; se obtiene al determinar la diferencia entre el número menor y el mayor (Garriga *et al*., 2010). Cuanto mayor es el rango o amplitud de los datos, más dispersos se encuentran alrededor de la media aritmética, sin considerar la afectación de posibles valores extremos (Sokal y Rohlf, 1994).

### Propiedades

- Sus unidades son las mismas que las unidades de las variables.
- El rango muestral no es una buena medida de dispersión, ya que para su determinación solamente utiliza dos observaciones, o sea los valores extremos (máximo y mínimo), por lo que puede estar influenciado por estos valores.
- El rango aumenta con el número de observaciones o se queda igual; pero nunca disminuye.
- Puede dar indicios de la variabilidad que presenta la distribución de datos.

### Varianza

Es la sumatoria de las diferencias cuadráticas de n puntuaciones con respecto a su media aritmética, o sea, mide el promedio de las desviaciones al cuadrado de las observaciones respecto a la media aritmética; por lo tanto, expresa la variabilidad de la distribución de datos alrededor de la media y nunca será negativa. Se denota por $S^2$ para los datos muéstrales y por sigma cuadrado ($\sigma^2$) para datos poblacionales. Se expresa en el cuadrado de la unidad de medida utilizada.

Cuando se calcula la varianza en poblaciones de datos en el denominador de la fórmula se utiliza el total de observaciones (N), aunque es el caso menos común, sin embargo, cuando interesa estimar la varianza poblacional y se utilizan datos que provienen de una muestra (caso más común), en el denominador de la fórmula se le resta uno al tamaño de la muestra (n-1), que serían los grados de libertad; con lo que se busca aplicar una pequeña medida de corrección que hace a la varianza más representativa y un estimador no sesgado de la varianza de la población.

## Desviación típica o estándar

Se define como la raíz cuadrada positiva de la varianza; es el promedio de la desviación de las puntuaciones con respecto a la media. Se expresa en las unidades originales de medición de la distribución de datos por lo que es más fácil de interpretar que la varianza. Se denota por S para los datos muéstrales y por sigma ($\sigma$) en datos poblacionales. Se calcula al determinar la raíz cuadrada de la varianza.

### Características

- La desviación típica o estándar origina como resultado un valor numérico que representa el promedio de la diferencia que se presenta entre los datos y la media aritmética.
- Su utilización es muy importante para evaluar el área que queda por debajo de una curva de distribución nor-

mal, que se relaciona con la probabilidad de casos que pertenecen a una población o conjunto de datos.

– No se recomienda su uso cuando la media aritmética no es la medida adecuada de tendencia central.

– Mientras mayor es la desviación típica o estándar, mayor será la dispersión de los datos alrededor de la media aritmética.

Se suele preferir a la desviación típica o estándar, puesto que se expresa en las mismas unidades que la media, mientras que la varianza se expresa en las unidades de la variable al cuadrado.

## Coeficiente de variación

El coeficiente de variación (CV) es la relación que se presenta entre la desviación típica o estándar de una población o muestra y su media aritmética. Es también denominado Coeficiente de variación de Pearson y su fórmula es:

Para datos poblacionales

$$CV = \frac{\sigma}{\mu} \text{x } 100$$ Ecua. 1

Para datos poblacionales

$$CV = \frac{S}{\bar{X}} \text{x} 100$$ Ecua. 2

### Propiedades

– Es adimensional debido a que no se expresa en unidades, las cuales se simplifican al dividir la desviación típica o estándar entre la media aritmética. Se expresa en porcentaje (%), lo que garantiza una mejor interpretación.

– Indica la variabilidad o dispersión relativa de los datos de la variable analizada alrededor de la media aritmética.

– Si el valor del CV aumenta existe mayor heterogeneidad de los valores de la variable en cuestión; y si disminuye se presenta mayor homogeneidad.

– Es la única medida que no es generada por el SPSS por lo que debe ser calculada de forma manual con la utilización de la fórmula descrita anteriormente.

– Se utiliza para comparar la dispersión o variabilidad de los datos entre variables medidas en diferentes unidades.

– Valores mayores a 20% indican la posible presencia de errores experimentales, de muestreo o imprecisiones en los instrumentos de medición utilizados, aunque lo recomendable es que sea lo más bajo posible.

## Medidas de distribución

Describen la forma en que se reúnen los datos de acuerdo a la frecuencia en que se encuentran dentro de la distribución. Permiten conocer la forma en que se agrupan o separan los valores en relación con su representación gráfica, aunque su utilidad radica en la posibilidad de identificar las características de la distribución sin necesidad de generar dicho gráfico. Sus principales medidas son el coeficiente de asimetría de Fisher y el coeficiente de Curtosis.

## Coeficiente de asimetría de fisher

Es un coeficiente adimensional, que no tiene unidades de medida y que se aplica a distribuciones de datos unimodales. Se denomina también como sesgo. Se denota de diferentes formas, entre las que se encuentran As, g1, alfa3, entre otros.

Para su determinación se utiliza la fórmula:

$$A_s = \frac{\bar{X} - Mo}{S} \qquad \text{Ecua. 3}$$

### Propiedades

– Medida de forma o apuntamiento que permite identificar si las frecuencias de datos se distribuyen de forma uniforme alrededor de la media aritmética, la cual constituye su eje de asimetría.

– A medida que el valor del coeficiente se aleje más de 0 indica una separación mayor de la aglomeración de los datos respecto a la media aritmética.

Puede ser negativa, simétrica o positiva

Asimetría negativa: cuando la mayoría de los datos se encuentran agrupados en el lado derecho de la media aritmética y la menor cantidad de datos se distribuyen al lado izquierdo, aunque con mayor dispersión o sesgo. Cuando el valor de la moda es superior a la media (As<0) la asimetría es negativa (Garriga *et al*., 2010).

Curva o distribución simétrica: cuando se presentan de forma aproximada, la misma cantidad de valores a ambos lados de la media aritmética. Cuando la media y la moda coincide (el numerador de la fórmula se convierte en 0 y el valor de $A_S$=0) nos encontramos ante una distribución simétrica.

Asimetría positiva: es cuando la mayoría de los datos se encuentran agrupados al lado izquierdo de la media aritmética y la menor cantidad de datos, pero más dispersos, se distribuyen al lado derecho. Si la media es mayor que la moda se obtiene una $A_S$>0 y la distribución de datos se establece con asimetría positiva.

## Coeficiente de curtosis

### Propiedades

– Medida de apuntalamiento que proporciona el grado de concentración que muestran las frecuencias de los valores en el punto medio de la distribución de datos.

– Se denota como g2.

Puede ser:

Leptocúrtica: es cuando se presenta una alta concentración de frecuencias de valores en el punto central de la distribución de datos. La distribución de datos es leptocúrtica cuando g2>0.

Mesocúrtica: se denomina así cuando existe una concentración de datos alrededor de la media aritmética aproximadamente igual a la distribución normal. La distribución de datos es mesocúrtica cuando g2=0, sin embargo, debido a que es difícil encontrar valores iguales a cero, se aceptan valores cercanos a ± 0,5 para definirla como una distribución normal (Distribución de Gauss).

Platicúrtica: es cuando existe una baja concentración de frecuencias de valores en relación al punto central de la distribución de datos y una mayor cantidad de valores alejados de este punto. La distribución es platicúrtica cuando g2<0.

Cuando g1 y g2 alcancen valores entre ±0,5 la curva de distribución de datos se denomina como normal, criterio definitorio a la hora de elegir el procedimiento estadístico a desarrollar, ya que puede conducir a emitir conclusiones sesgadas.

## Descripción de datos

La descripción de datos constituye una de las principales funciones de la estadística, la misma puede ser realizada por medio del cálculo de las medidas de resumen de datos, tablas y gráficos, en los cuales se muestra la forma en que se comporta o descubrir patrones de distribución ocultos en la información recolectada previamente. Sin embargo, no todos los procedimientos estadísticos son realmente útiles para las diferentes escalas de medida: lo que se debe tener en cuenta en el momento de realizar un análisis descriptivo. En el cuadro 5.1. se presentan las distintas medidas descriptivas para las diferentes escalas en que se miden las variables, así como tablas de frecuencias, tablas de contingencia (para su elaboración se requieren dos o más variables categóricas) y técnicas gráficas descriptivas recomendadas.

Cuadro 5.1. Alternativas para resumir los datos de forma tabular y gráfica en función de las escalas de medición de las variables analizadas.

| Escala de medición | TF | TC* | Estadísticos descriptivos | | | | Gráficos recomendados |
|---|---|---|---|---|---|---|---|
| | | | Tendencia central | Posición | Dispersión | Distribución | |
| Nominal u ordinal | Si | Si | Solo y mediana | No | No | No | Barras y circular |
| Intervalo | Si | No | Si | Si | Menos el CV | Si | Histograma, diagrama de cajas y sesgos |
| Razón | Si | No | Si | Si | Si | Si | Histograma, diagrama de cajas y sesgos |

Notación: TF=Tabla de Frecuencia. TC=Tabla de Contingencia. *La TC se realiza cuando se cuenta con dos o más variables categóricas medidas en escala ordinal o nominal.

## Descripción de variables numéricas

Debido a la mayor información que generan las variables numéricas, es posible realizar análisis más complejos con los datos que generan estas variables, utilizándose para su descripción una gran variedad de medidas de resumen.

## Estadísticos descriptivos

Son estadísticos utilizados para describir las características de variables numéricas y su determinación mediante el uso de un software estadísticos es algo sencillo; sin embargo, la presentación de los resultados y su interpretación correcta es el elemento fundamental a tener en cuenta en cualquier proceso investigativo que se desarrolle.

Para las variables en los que sus VFM son números, las medidas de resumen de datos (tendencia central, posición,

dispersión y de forma, son las que mejor describen las características de la base de datos generada.

A continuación, mediante un ejemplo práctico se desarrolla el procedimiento estadístico para el cálculo de los estadísticos descriptivos y se explica la forma de realizar la interpretación asociada con el proceso que se investiga.

Ejemplo: se necesita realizar una descripción del comportamiento del peso (kg) de los 30 animales atendidos durante una semana en una clínica veterinaria (anexo 1).

Procedimiento estadístico: abrir la base de datos con el SPSS y dentro del visor de datos se activa la pestaña analizar>estadísticos descriptivos>frecuencias>se traslada la variable de la ventana izquierda donde se encuentra hacia la ventana de análisis ubicada en el lado derecho del cuadro de diálogo>clic izquierdo en la pestaña estadísticos y se despliega un cuadro de dialogo donde se muestran todas las opciones dentro de las que se activan las medidas de tendencia central (media, mediana y moda), posición (dentro valores percentiles se selecciona cuartiles), dispersión (desviación estándar, varianza, rango, mínimo, máximo y media de error estándar) y distribución (asimetría y curtosis)>continuar>clic en la opción gráficos y se marca el gráfico que para una variable numérica es un histograma, además se puede marcar mostrar curva de la distribución normal>continuar>aceptar. En el visor de resultados del SPSS se muestra la tabla de estadísticos los cuales se interpretan en función de la variable objeto de estudio (Cuadro 5.2.).

Cuadro 5.2. Estadísticos descriptivos que resumen las características de la variable peso de los animales (kg).

| | | |
|---|---|---|
| N | Válido | 30 |
| | Perdidos | 0 |
| Media | | 18,5067 |
| Error estándar de la media | | 2,35220 |
| Mediana | | 15,6500 |
| Moda | | 4,90[a] |
| Desviación estándar | | 12,88351 |
| Varianza | | 165,985 |
| Asimetría | | 1,158 |
| Curtosis | | 2,545 |
| Rango | | 57,50 |
| Mínimo | | 3,50 |
| Máximo | | 61,00 |
| Percentiles | 25 | 4,9000 |
| | 50 | 15,6500 |
| | 75 | 25,7750 |

a. Existen múltiples modas. Se muestra el valor más pequeño.

## Interpretación

Los valores diferentes en las medidas de tendencia central son un indicio que evidencia que la distribución de datos no es similar a la distribución normal. La $\overline{X}$=18,51 kg, se encuentra afectada por un valor extremo; el valor de la Me es 15,65 kg y constituye una mejor medida de tendencia central que la media aritmética, ya que no se encuentra afectada por el valor extremo presente en los datos, sin embargo, se deduce que se presenta de forma general un sobrepeso generalizado en los animales considerados en el estudio. Los valores que más se repiten son el 4,9 y el 27,3 (dos veces cada uno), por lo que es una distribución bimodal, aunque el programa

muestra siempre el valor modal más pequeño. La desviación típica o estándar (S) =12,88 kg, evidencia la dispersión de los valores del peso alrededor de la media aritmética y se observa que el valor máximo se encuentra a más de tres S de la media. Se muestra un rango bastante amplio (57,5 kg), afectado por el valor atípico. La g1 positiva indica que los datos están sesgados hacia el lado derecho de la distribución. La g2 positiva expresa que es una distribución Leptocúrtica. El CV=69,6% muestra la alta dispersión relativa de los datos. El $P_{25}$ ($Q_1$) y el $P_{75}$ ($Q_3$) indican que el 25% y el 75% del total de datos se encuentran por debajo de 4,9 y 25,77 kg respectivamente; y el $P_{50}$ ($Q_2$) denota que 15,65 kg parte la distribución de datos en dos partes iguales.

## Descripción de variables categóricas

Para variables categóricas ya sean nominales u ordinales los principales procedimientos estadísticos que se pueden utilizar en su análisis descriptivo son las tablas de distribución de frecuencias (frecuencia absoluta referida al recuento de los datos y frecuencia relativa concerniente al porcentaje que abarcan), la moda y las tablas de contingencia, en la que se pueden comparar variables cualitativas politómicas y en otros casos, asociar las categorías de dos o más variables categóricas. Los gráficos sugeridos son el de sectores y de barras.

## Tablas de distribución de frecuencias

Son tablas que muestran la distribución absoluta y relativa de las categorías de una variable objeto de estudio, ya sea categórica o numérica. Es más apropiado y generalizado su empleo para variables categóricas, ya sean nominales u ordinales, aunque se utilizan en ocasiones para describir variables numéricas. Se entiende por frecuencia a la cantidad de veces que una categoría o un valor es observado o medido en un conjunto de datos y distribución de frecuencia es el método estadístico que se utiliza para describir dicho conjunto de datos. Según Morales (2012) una distribución de frecuencia es una tabla resumida donde se colocan los datos

divididos en grupos ordenados numéricamente, denominados clases o categorías.

Es preciso significar que, en el procedimiento estadístico desarrollado con el SPSS para el caso de variables categóricas, se deben establecer en el visor de variables las etiquetas de valor para cada categoría de dicha variable, sin embargo, es erróneo realizar la determinación de los estadísticos descriptivos (medidas de resumen) ya que los resultados obtenidos son falsos y solo se relacionan con el valor asignado a la etiqueta de la variable.

Ejemplo 1: en 30 animales tratados durante una semana en una clínica veterinaria se determinó el peso de inicio en kg y se necesita conocer la distribución de frecuencias (anexo 1).

Procedimiento estadístico: abrir la aplicación del software estadístico y acceder a la base de datos>buscar en la barra de menú y dar clic en la opción analizar>estadísticos descriptivos>frecuencias>se selecciona la variable a describir y se desplaza para el cuadro en blanco de la derecha>dar clic en gráfico que se desea obtener>se selecciona gráfico de barras (elegido en este caso) o circular por ser una variable categórica>continuar>aceptar.

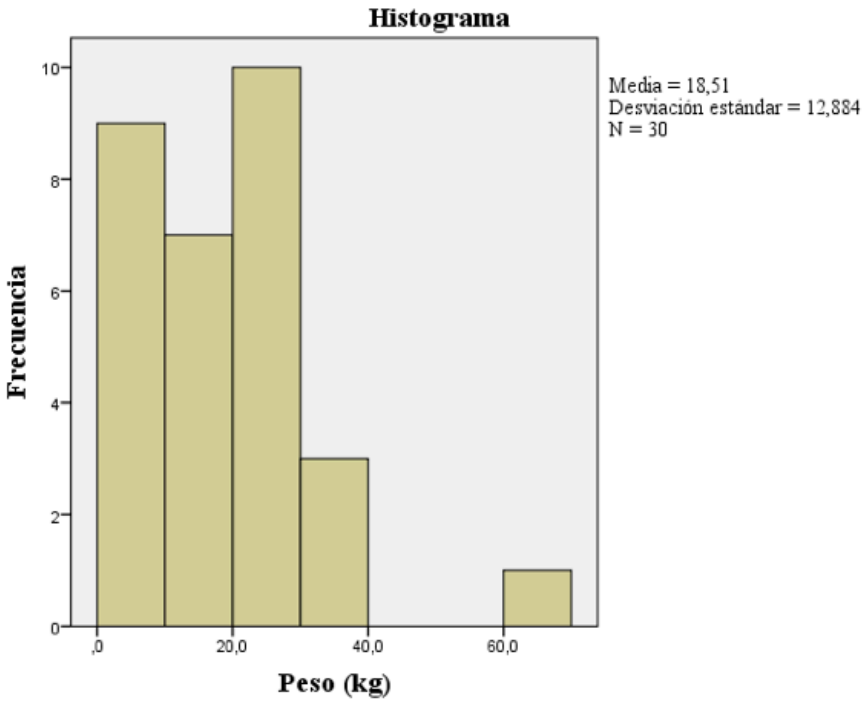
En el visor de resultados del SPSS se muestra la tabla de frecuencias obtenida (Cuadro 5.3.) y el gráfico de barras generado (Imagen 5.3.).

Cuadro 5.3. Tabla de distribución de frecuencias para la variable peso de los animales en kg.

| | | Frecuencia | Porcentaje (%) | Porcentaje válido (%) | Porcentaje acumulado (%) |
|---|---|---|---|---|---|
| Válido | 3,5 | 1 | 3,3 | 3,3 | 3,3 |
| | 3,7 | 1 | 3,3 | 3,3 | 6,7 |
| | 3,9 | 1 | 3,3 | 3,3 | 10,0 |
| | 4,2 | 1 | 3,3 | 3,3 | 13,3 |
| | 4,6 | 1 | 3,3 | 3,3 | 16,7 |
| | 4,8 | 1 | 3,3 | 3,3 | 20,0 |

| | | Frecuencia | Porcentaje (%) | Porcentaje válido (%) | Porcentaje acumulado (%) |
|---|---|---|---|---|---|
| | 4,9 | 2 | 6,7 | 6,7 | 26,7 |
| | 7,8 | 1 | 3,3 | 3,3 | 30,0 |
| | 12,3 | 1 | 3,3 | 3,3 | 33,3 |
| | 13,8 | 1 | 3,3 | 3,3 | 36,7 |
| | 14,8 | 1 | 3,3 | 3,3 | 40,0 |
| | 14,9 | 1 | 3,3 | 3,3 | 43,3 |
| | 15,3 | 1 | 3,3 | 3,3 | 46,7 |
| | 15,6 | 1 | 3,3 | 3,3 | 50,0 |
| | 15,7 | 1 | 3,3 | 3,3 | 53,3 |
| | 22,5 | 1 | 3,3 | 3,3 | 56,7 |
| | 22,6 | 1 | 3,3 | 3,3 | 60,0 |
| | 22,9 | 1 | 3,3 | 3,3 | 63,3 |
| | 23,2 | 1 | 3,3 | 3,3 | 66,7 |
| | 23,6 | 1 | 3,3 | 3,3 | 70,0 |
| | 25,1 | 1 | 3,3 | 3,3 | 73,3 |
| | 25,4 | 1 | 3,3 | 3,3 | 76,7 |
| | 26,9 | 1 | 3,3 | 3,3 | 80,0 |
| | 27,3 | 2 | 6,7 | 6,7 | 86,7 |
| | 32,1 | 1 | 3,3 | 3,3 | 90,0 |
| | 34,9 | 1 | 3,3 | 3,3 | 93,3 |
| | 35,7 | 1 | 3,3 | 3,3 | 96,7 |
| | 61,0 | 1 | 3,3 | 3,3 | 100,0 |
| | Total | 30 | 100,0 | 100,0 | |

Imagen 5.3. Histograma de frecuencias para la variable peso de los animales en kg.




Interpretación: los resultados muestran que los valores 4,9 y 27,3 kg se repiten dos veces y cada uno representa el 6,7% del total, además, se observan los valores mínimos (3,5 kg) y máximos (61,0 kg), que evidencian que existe una amplitud total de 57,5 y una alta dispersión de los valores.

Ejemplo 2: en 30 animales tratados durante una semana en una clínica veterinaria se necesita conocer la distribución de la variable tamaño de los animales, los cuales fueron agrupados por su peso en tres categorías (anexo 1).

Procedimiento estadístico: acceder a la base de datos y abrirla con el software estadístico>buscar en la barra de menú y dar clic en la opción analizar>estadísticos descriptivos>frecuencias>se selecciona y se desplaza para el cuadro en blanco de la derecha la variable a describir>dar clic en gráfico que se desea obtener, en este caso se solicita un grá-

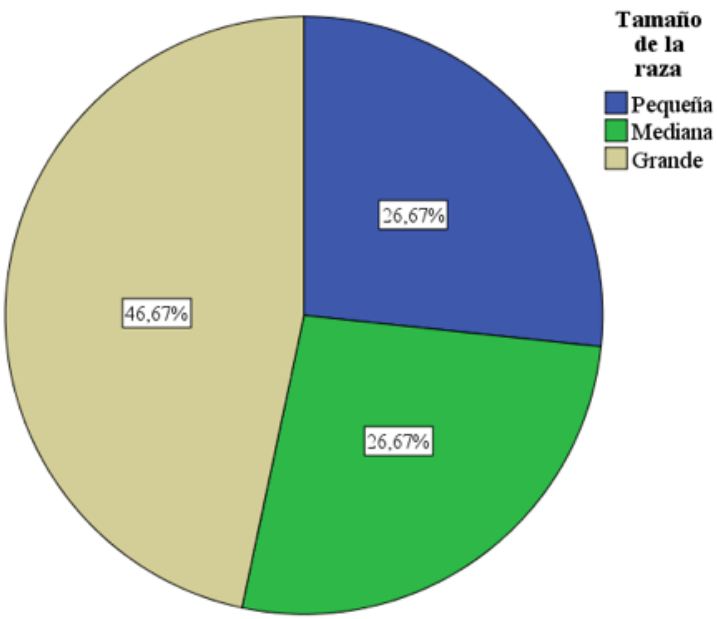
fico circular por ser una variable categórica, aunque en su defecto se puede solicitar un gráfico de barras>continuar>aceptar.

En el visor de resultados del SPSS se muestra la tabla de frecuencias obtenida (Cuadro 5.4.) y el gráfico circular generado (Imagen 5.4).

Cuadro 5.4. Tabla de distribución de frecuencias para la variable ordinal tamaño de la raza de animales.

| | Categorías de la variable | Frecuencia | Porcentaje (%) | Porcentaje válido (%) | Porcentaje acumulado (%) |
|---|---|---|---|---|---|
| Válido | Pequeña | 8 | 26,7 | 26,7 | 26,7 |
| | Mediana | 8 | 26,7 | 26,7 | 53,3 |
| | Grande | 14 | 46,7 | 46,7 | 100,0 |
| | Total | 30 | 100,0 | 100,0 | |

Imagen 5.4. Gráfico circular que muestra la distribución de frecuencias para la variable tamaño de la raza de animales.



Interpretación: los resultados muestran que el 26,67% de los animales se encuentran dentro de las categorías de pequeña (hasta 5 kg de peso) y mediana (a partir de 5 y hasta 20 kg), sin embargo, predominan los animales grandes, con más de 20 kg, los cuales representan el 46,7 % de la muestra analizada.

## Tablas de contingencia (TC)

Las TC se definen como un arreglo donde se cruzan dos o más variables; es utilizada para clasificar las observaciones de acuerdo con dos o más categorías presentes en cada variable (Lind *et al*., 2004), pueden ser categóricas por sus escalas de medición (nominales u ordinales) o cualitativas por su naturaleza (dicotómicas o politómicas).

Cuando se cruzan dos variables se tienen de TC bidimensionales, de dos vías o de doble entrada (Aguilera, 2005) y cuando se cruzan tres variables o más se obtienen TC multidimensionales (Aguilera, 2006).

Las TC bidimensionales pueden ser de 2x2, cuando se realizan estudios en que se efectúan comparaciones entre dos variables con dos categorías cada una (es la más sencilla y las más utilizada), de 3x2, cuando se refiere a una variable con tres categorías y otra con dos categorías (análisis bivariado correspondiente al nivel relacional) o en general de axb (cualquier combinacion) siendo “a” el número de categorías de la variable representada en las filas y “b” el número de categorías de la variable representada en las columnas. En cada casilla de una TC bidimensional se encuentra la frecuencia observada del objeto en las categorías correspondientes a cada variable.

Las TC utilizadas para análisis descriptivo proporcionan información resumida que facilita la descripción de las categorías de las variables implicadas; sin embargo, su finalidad no es realizar comparaciones que permitan arribar a conclusiones de una población, sino que pueden generar posibles hipótesis en función del comportamiento de los datos.

Para generar una TC bidimensional se necesita vincular las categorías de una de las variables en las columnas y las categorías de la otra variable en las filas; además, para que un caso sea incluido debe contar con un valor válido en cada variable.

Es posible construir diferentes tipos de TC bidimensionales, lo cual se encuentra condicionado a la intensión que el investigador expresó en el propósito del estudio. Si se necesita comparar dos grupos, entonces una variable es fija y la otra es aleatoria y las TC bidimensionales formadas se elaboran en función de las categorías que presenten cada una. Si se pretende asociar dos categorías de dos variables, entonces las dos variables son aleatorias y se trabaja con una TC de 2x2 (presenta cuatro casillas o núcleos) y los marginales constituyen el total para las columnas y filas.

La variable es fija cuando la distribución de sus datos se conoce antes de realizar la recolección de la información; es la que determina la orientación de los porcentajes dentro de la TC y se coloca habitualmente en las columnas. La variable es aleatoria cuando la distribución de sus datos se desconoce al iniciar el estudio, o sea, se conoce una vez realizada la recolección de la información.

Para una mejor comprensión del procedimiento a desarrollar para elaborar TC bidimensional se establece el siguiente ejemplo práctico.

Ejemplo: se desea estudiar la enfermedad de *Babesia canis* en perros atendidos en una clínica veterinaria determinada, y la posible influencia del lugar de tenencia de los animales en el incremento del padecimiento de la patología. Para desarrollar el estudio se eligieron dos grupos de animales (diseño de casos y controles) en función de su padecimiento a *B. canis* (variable fija con dos categorías conocidas antes de realizar la investigación, animales enfermos con *B. canis* (casos) y animales no enfermos con *B. canis*, que constituye el grupo control); además con el dueño de la mascota se determinó el lugar de tenencia (variable aleatoria desconocida antes de realizar el estudio y que puede presentar tres categorías, en la casa-terraza, en la terraza-patio y en el patio-calle (anexo 2). Se necesita determi-

nar la frecuencia de los diferentes lugares de tenencia de los animales dentro de los grupos conformados (casos y controles).

Procedimiento estadístico: acceder a la base de datos con el SPSS y seleccionar en la barra de menú la opción analizar>estadísticos descriptivos>tablas cruzadas>se selecciona y se traslada para columnas a la variable fija (*B. canis*) y se selecciona y se desplaza para filas a la variable aleatoria (lugar de tenencia)>clic en la sección casillas y se elige dentro de las tres opciones de porcentajes que muestra el cuadro de dialogo a columna, debido a que se necesita conocer cómo se distribuye el factor de riesgo (lugar de tenencia) dentro de los grupos con presencia o no de *B. canis* (VD)>continuar>aceptar.

En el visor de resultados del SPSS se muestra la TC generada en la que se observa la frecuencia de la variable lugar de tenencia de los animales dentro de los grupos conformados (Cuadro 5.5.).

Cuadro 5.5. Tabla de contingencia bidimensional que muestra la frecuencia absoluta y relativa de la variable lugar de tenencia dentro de cada grupo (diseño de casos y controles) en relación con la presencia o no de *B. canis*.

| Variable aleatoria | Categorías | Recuento y porcentaje | *Babesia canis* | | Total |
|---|---|---|---|---|---|
| | | | Casos (con *B. canis*) | Controles (sin *B. canis*) | |
| Lugar de tenencia | Casa-terraza | Recuento | 5 | 8 | 13 |
| | | % dentro de *B. canis* | 25,0% | 40,0% | 32,5% |
| | Terraza-patio | Recuento | 7 | 7 | 14 |
| | | % dentro de *B. canis* | 35,0% | 35,0% | 35,0% |
| | Patio-calle | Recuento | 8 | 5 | 13 |
| | | % dentro de *B. canis* | 40,0% | 25,0% | 32,5% |
| Total | | Recuento | 20 | 20 | 40 |
| | | % dentro de *B. canis* | 100,0% | 100,0% | 100,0% |

Interpretación: la frecuencia de los diferentes lugares de tenencia dentro del grupo de animales infestados con *B. canis* muestra que la mayor proporción (40,0%) de los afectados se mantienen en patio-calle, un 35% dentro de terraza-patio y un 25% dentro de patio-calle, por lo que será posible establecer la hipótesis que indica que podrían existir diferencias significativas entre ellas, sin embargo, para cumplir con este propósito se debe aplicar la prueba no paramétrica denominada Chi-cuadrado de independencia (para variables categóricas), la cual se desarrolla en la sección "Prueba Chi-cuadrado".

## Estimación estadística de parámetros

### Estimación puntual

En investigación científica se estudian fenómenos y se efectúan mediciones de una o más variables aleatorias perteneciente a una población, las cuales presentan una distribución probabilística que puede ser conocida o desconocida, y le corresponden algunos parámetros que la caracterizan como son la media poblacional, la varianza poblacional, la proporción poblacional, entro otros. Estos parámetros generalmente son desconocidos (a veces no es factible ni económico estudiar a cada individuo de la población), por ello, cuando se estudia una variable se toma una muestra aleatoria de esa variable y se efectúan estimaciones válidas y confiables de dichos parámetros.

Con la finalidad de realizar estimaciones aproximadas de los parámetros poblaciones con la precisión y confiabilidad que el problema investigado requiere, se utilizan estimadores puntuales (un valor concreto); que constituyen estadísticos que tienen el objetivo de acercarse lo más posible al verdadero valor del parámetro poblacional (Horra, 2003), entre los que se encuentran:

El estadístico o estadígrafo media muestral $\overline{X}$, es un estimador de la varianza poblacional $\mu$ (se simboliza como $\overline{X} = \hat{\mu}$ ).

El estadístico varianza muestral $S^2$, es un estimador de la varianza poblacional $\sigma^2$ (se simboliza como $S^2=\hat{\sigma}^2$).

El estadístico proporción muestral $p$, es un estimador de la proporción poblacional $P$ (se simboliza como $p=\hat{P}$)

Según Solarte *et al.* (2009) los estimadores deben cumplir algunos requisitos para que sean buenos estimadores:

Insesgados: el estimador puntual es insesgado (ausencia de sesgo) cuando el valor medio obtenido en diferentes muestras es igual al parámetro que se estima. Esta condición garantiza que la distribución muestral del estimador se encuentre alrededor del parámetro poblacional objetivo.

Insesgados de mínima varianza: el estimador insesgado de mínima varianza de un parámetro es el estimador que tiene la varianza más pequeña de entre todos los estimadores insesgados.

Consistentes: un estimador insesgado es consistente si la probabilidad de que el valor estimado coincida con el parámetro (su varianza tiende a 0) lo que ocurre a medida que el tamaño de la muestra aumenta.

Eficientes: al comparar dos estimadores, es más eficiente aquel que proporciona una estimación con una menor variabilidad (Rodrigo y Molina, 2011).

En la estadística inferencial se generalizan las conclusiones obtenidas en la muestra hacia su correspondiente población; sin embargo, para que las conclusiones sean válidas a nivel poblacional es necesario que la muestra sea representativa (Casas, 1996; Solarte *et al*., 2009). Por ello, en cualquier experimento que se realice para estimar la media poblacional u otro parámetro, se debe tener en cuenta el cálculo del tamaño de la muestra, el cual no puede ser muy grande, ya que se pierde tiempo y se incrementan los costos, o muy pequeño, donde la inferencia estadística carece de validez externa.

Según Supo (2017) para el cálculo del tamaño de la muestra se debe tener en cuenta:

El nivel investigativo: define el tipo de objetivo estadístico a emplear, en el nivel descriptivo, se utiliza el objetivo estimar y en el nivel relacional, se usa comparar.

La escala de medición de la variable de estudio: la fórmula es diferente en función de si la variable es categórica o numérica.

Marco muestral: es diferente la fórmula a emplear para poblaciones donde se conoce el marco muestral (poblaciones finitas) o es desconocido (poblaciones infinitas).

A continuación, se establecen las fórmulas que se utilizan para el cálculo del mínimo tamaño muestral en función de los criterios descritos anteriormente.

Para estimar parámetros categóricos en una población infinita o desconocida

$$n = \frac{Z_{1-\alpha}^2 * p * q}{d^2}$$ Ecua. 4

Para estimar parámetros categóricos en una población finita o conocida

$$n = \frac{N * Z_{1-\alpha}^2 * p * q}{d^2 * (N-1) + Z_{1-\alpha}^2 * p * q}$$ Ecua. 5

Para estimar parámetros numéricos (promedio) en una población infinita o desconocida

$$n = \frac{Z_{1-\alpha}^2 * S^2}{d^2}$$ Ecua. 6

Para estimar parámetros numéricos (promedio) en una población finita o conocida

$$n = \frac{N * Z_{1-\alpha}^2 * S^2}{d^2 * (N-1) + Z_{1-\alpha}^2 * S^2}$$ Ecua. 7

Para comparar proporciones en dos grupos basados en una variable categórica

$$n = \frac{\left[Z_{1-\alpha} * \sqrt{2p(1-p)} + Z_{1-\beta} * \sqrt{p_1(1-p_1) + p_2(1-p_2)}\right]^2}{(p_1 - p_2)^2} \quad p = \frac{(p_1 - p_2)}{2}$$ Ecua. 7

Para comparar promedios en dos grupos basados en una variable numérica

$$n = \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2 * (S_1^2 + S_2^2)}{(X_1 + X_2)^2}$$ Ecua. 8

**Donde:**

n=tamaño de la muestra.

N=tamaño de la población.

$Z_{1-\alpha/2}$= valor obtenido en la tabla de la distribución normal estandarizada para una confiabilidad de 1-α, valor que es fijado previamente por el investigador en relación con la rigurosidad del estudio que desarrolla. De forma convencional se trabaja en la mayoría de las investigaciones con una probabilidad de error (α)=0,05 (error de tipo I) y un nivel de confianza del 95% a dos colas (el valor tipificado es 1,96).

$\beta$=error de tipo II. 1-$\beta$ es el poder estadístico (normalmente 0,80) y el valor tipificado de $Z_{1-\beta}$ =0,84.

$p$=probabilidad de ocurrencia.

$q$=complemento de p.

$d$=error máximo de estimación permitido (precisión). Es fijada por el investigador y lleva la misma unidad de medida que la variable de estudio.

$S$=desviación típica o estándar obtenida en estudios preliminares o en un estudio piloto.

En la práctica es común trabajar con estimaciones puntuales, y aunque son útiles presentan limitaciones, entre las que se encuentran.

- No permiten establecer con claridad la exactitud de la estimación.

- Pueden no presentar la confiabilidad necesaria.

Por lo tanto, y por muy eficiente que sea un estimador puntual, es poco probable que estime con exactitud el verdadero valor del parámetro poblacional (Depool y Monasterio, 2013); por esta causa, existen otros procedimientos de estimación de parámetros que resuelven este problema, y que de denomina estimación de parámetros por intervalos de confianza.

## Estimación por intervalos de confianza

Un intervalo es un par de números reales A y B, con A<B, los cuales constituyen un intervalo, formado por todos los números reales encontrados entre A y B; por lo tanto, el intervalo proporciona dos extremos (límite inferior y límite superior) entre los cuales se debe encontrar la media poblacional, con nivel de confiabilidad predefinido (1–α x100). La estimación de parámetros por intervalos de confianza (rango de valores) se utiliza para la media, proporción, varianza y razón de varianzas (Depool y Monasterio, 2013), su finalidad es proporcionar, a partir de los datos (es aleatorio), una región donde se encuentre el verdadero parámetro poblacional; sin embargo, existe una probabilidad de que no quede dentro del intervalo de confianza (Sáez, 2012).

La estimación por intervalos de confianza es recomendable realizarla cuando n≥30 y se utiliza un procedimiento diferente para su determinación si se conoce o no la desviación estándar poblacional (σ).

### Caso con varianza conocida

Cuando la varianza de la población es conocida (menos utilizada) y los datos siguen una distribución normal los límites inferior y superior del intervalo de confianza para la media poblacional (μ) se determina mediante la fórmula:

$$\bar{X} \mp \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}$$ Ecua. 9

**Donde**

$\bar{}$=Estimación puntual que se realiza a partir de una muestra de tamaño n.

Primero se le resta la expresión que aparece para buscar el LI y se suma para buscar el LS (forma abreviada de escribir la fórmula).

$\sigma$=Desviación típica o estándar de la varianza poblacional (conocida).

n=Tamaño de muestra.

$Z_{\alpha/2}$=Percentil de la distribución normal estandarizada, del orden 1-α/2, el cual depende de la confiabilidad de la estimación, con este valor se busca el percentil correspondiente de 1-α/2.

**Fórmula para el intervalo de confianza**

$$\mu \in \left[\bar{X} - \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}} ; \bar{X} + \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}\right]$$ Ecua. 10

El intervalo contiene al verdadero parámetro poblacional μ con una confiabilidad de (1-α)x100%.

**Caso con varianza desconocida**

Cuando la varianza de la población es desconocida (caso más común) y los datos siguen una distribución normal los límites inferior y superior del intervalo de confianza para la media poblacional (μ) se determina mediante la fórmula:

$$\bar{X} \mp \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2};n-1}$$ Ecua. 11

La diferencia radica en que no se utiliza la varianza poblacional y se emplea la varianza muestral, en específico la desviación típica o estándar de la muestra (S).

**Fórmula para el intervalo de confianza**

$$\mu \in \left[\bar{X} - \frac{S}{\sqrt{n}}\, t_{\frac{\alpha}{2}:n-1} \; ; \; \bar{X} + \frac{S}{\sqrt{n}}\, t_{\frac{\alpha}{2}:n-1}\right] \qquad \text{Ecua. 12}$$

Se utiliza el percentil de la distribución T-Student del mismo orden 1-α/2 y con n-1 grados de libertad y se determina que el intervalo de confianza contiene al verdadero parámetro poblacional (μ) con una confiabilidad de (1-α) x100%.

Para una mejor comprensión del procedimiento a desarrollar para calcular los intervalos de confianza se establece el siguiente ejemplo práctico.

Ejemplo: se desea estudiar el comportamiento de las precipitaciones (mm) promedio por día con lluvia en el periodo menos lluvioso (junio-noviembre de 2017) en la finca Santa Inés de la Universidad Técnica de Machala. Para desarrollar el estudio se utilizaron los datos de precipitaciones por día durante el periodo mencionado, existentes en la estación meteorológica ubicada en la propia finca (anexo 3). Se necesita determinar el intervalo de confianza para estimar la media poblacional (μ) de precipitaciones promedio por día con una confiabilidad del 95%.

Procedimiento estadístico: acceder a la base de datos mediante el paquete estadístico SPSS y seleccionar en la barra de menú la opción analizar>clic en comparar medias>se selecciona la prueba T para una muestra>se selecciona la variable precipitaciones y se traslada para variables de prueba que se encuentra el lado derecho del cuadro de diálogo, el valor de prueba se deja en 0 que lo trae el paquete estadístico por defecto>clic en opciones y se mantiene el 95% que es el porcentaje de confiabilidad con el cual se calcula el intervalo de confianza>continuar>aceptar.

En el visor de resultados del SPSS se expone la prueba de muestra única donde se observan los límites inferior y superior del intervalo de confianza para la estimación de la media poblacional (μ) de precipitaciones promedio por día con lluvias (Cuadro 5.6.).

Cuadro 5.6. Prueba de muestra única que muestra los límites del intervalo de confianza para la estimación de la variable precipitaciones promedio por día con lluvias con una confiabilidad del 95%.

| | Valor de prueba = 0 | | | | | |
|---|---|---|---|---|---|---|
| | t | gl | Sig. (bilateral) | Diferencia de medias | 95% de intervalo de confianza de la diferencia | |
| | | | | | Inferior | Superior |
| Precipitaciones por día con lluvia | 2,718 | 89 | ,008 | 3,0367 | ,817 | 5,257 |

Interpretación: se estima, con una confiabilidad del 95%, que el verdadero parámetro poblacional (μ) de precipitaciones promedio por día con lluvia en la granja Santa Inés en el periodo junio-noviembre de 2017 se encuentra entre el límite inferior de 0,817 mm y el superior de 5,257 mm.

## Prueba de hipótesis estadística

### Conceptualización y caracterización

Una prueba de hipótesis es un procedimiento estadístico utilizado para verificar la veracidad o falsedad de una hipótesis a partir de la información que ofrece una muestra aleatoria tomada en su población correspondiente. Bajo este contexto se presenta la inferencia estadística, en la cual los valores obtenidos de una muestra aleatoria representativa son utilizados para estimar parámetros de la población de estudio donde fueron obtenidos.

Es importante acotar que una muestra aleatoria es aquella donde cada elemento de la población tiene la misma probabilidad de ser seleccionado. Una de las técnicas utilizadas para la selección de la muestra es el empleo de una tabla de números aleatorios (anexo 4) (Development Core Team, 2008), con la cual se garantiza realizar afirmaciones válidas para toda la población (Solarte *et al*., 2009).

Por otro lado, es preciso señalar que las hipótesis estadísticas pueden existir en cualquier línea y nivel de la investigación, cuestión que define el análisis estadístico a realizar; aunque en cualquier línea o nivel de la investigación existe la posibilidad que se presenten estudios sin hipótesis estadística.

Sin embargo, si un estudio tiene o no hipótesis estadística se relaciona con el *enunciado del estudio*, y si este es susceptible de ser calificado de verdadero o falso, es posible emitir un juicio de valor, y se puede definir una hipótesis; por otro lado, si el *enunciado del estudio* no es susceptible de ser calificado de verdadero o falso, no se puede emitir un juicio de valor, por tanto, no es posible establecer una hipótesis y lo que se realiza es una estimación puntual con sus correspondientes intervalos de confianza.

Ejemplo 1: la aplicación al suelo de extractos de la macroalga *Ascophyllum nodosum* (L.) Le Jolis incrementa el rendimiento agrícola del cultivo de arroz (enunciado que puede ser calificado de verdadero o falso (se emite un juicio de valor), por tanto, es un estudio donde se puede establecer el procedimiento de la prueba de hipótesis ya que puede o no existir un efecto en la producción de arroz.

Ejemplo 2: se desea estudiar la prevalencia de *B. canis* en perros atendidos en un sector de la Ciudad (este enunciado no puede ser calificado de verdadero o falso, por lo tanto, es un estudio donde no se puede definir una prueba de hipótesis y se realiza una estimación puntual, o sea se determina la proporción de perros afectados por *B. canis* en la población estudiada a partir de una muestra representativa y una confiabilidad predefinida.

Por lo tanto, desde el punto de vista de la intención analítica expresada en el enunciado del estudio se pueden presentar estudios donde el procedimiento estadístico sea Prueba de Hipótesis o una Estimación Puntual.

## Significancia estadística

En el recorrido de una línea de investigación se transita por diferentes momentos (niveles de la investigación), y si en cada uno de ellos se plantean estudios con hipótesis, estas pueden ser *empíricas* (son aquellas que se presentan en los niveles exploratorio, descriptivo y relacional; y se originan a partir de la experiencia del investigador) o *racionales* (aquellas que se presenta en los niveles explicativo, predictivo y aplicativo; y tiene su origen en el conocimiento previo, o sea, en los antecedentes investigativos).

Desde el punto de vista estadístico no hay diferencia entre hipótesis empírica o hipótesis racional, la diferencia se presenta en la forma en que se comprobará dicha hipótesis.

En el nivel exploratorio las hipótesis se comprueban sin procedimientos estadísticos, ya que este nivel es cualitativo y no se realiza uso de la estadística.

Si el estudio es cuantitativo y el investigador establece una hipótesis, esta debe ser verificada; y para ello se necesita desarrollar un procedimiento estadístico denominado *significancia estadística* (Fisher, 1954), en el cual se utilizan herramientas estadísticas y cuenta con cinco pasos, los cuales se describen a continuación:

### 1. Planteamiento de hipótesis

Consiste en plantear las hipótesis estadísticas en la que se definen la hipótesis del investigador denotada convencionalmente como $H_1$ (llamada hipótesis alternativa), ya que en realidad es el planteamiento que se desea demostrar y constituye la base de la investigación; y la hipótesis nula, expresada como $H_0$, la cual es la negación de la hipótesis que plantea el investigador y siempre contiene la igualdad. Ambas hipótesis son excluyentes y de forma convencional

se coloca primero la nula y después la alternativa, aunque puede ser de manera contraria. En realidad, lo básico es que la hipótesis del investigador es una proposición que se plantea en función del enunciado del estudio y para ser probada se desarrolla el proceso investigativo. Una vez concluido dicho proceso la hipótesis del investigador puede ser aceptada o rechazada.

Ejemplo: $H_1$: El rendimiento agrícola de la caña de azúcar es distinto cuando se aplican diferentes dosis de compost al suelo.

$H_0$: El rendimiento agrícola de la caña de azúcar es igual cuando se aplican diferentes dosis de compost al suelo.

En este ejemplo lo que se prueba es si verdaderamente la aplicación de alguna dosis de compost influye en el incremento del rendimiento agrícola de la caña de azúcar.

2. Establecer el nivel de significación

Según Fisher (1954), el nivel de significancia estadística equivale a la magnitud del error que se está dispuesto a cometer al aceptar la hipótesis del investigador cuando en realidad es falsa (error de tipo I). Se denota por la letra del alfabeto griego Alfa (α). El error de tipo II se produce cuando se rechaza la hipótesis del investigador cuando en realidad es verdadera. Se denota por la letra del alfabeto griego Beta ($\beta$).

De forma generalizada se establece un nivel de significancia (α)=0,05, que representa el 5%; el cual constituye la máxima cantidad de error que estamos dispuestos a aceptar para dar por válida y quedarnos con la hipótesis planteada por el investigador ($H_1$). Su complemento (95%) se denomina nivel de confiabilidad.

Alfa debe ser establecido o fijado previamente por el investigador para realizar la prueba estadística (convencionalmente se predefine 0,05); sin embargo, en investigaciones donde se necesite una mayor confiabilidad se establece 1% (α=0,01) o menos; y en otras puede ser mayor a 0,05; lo cual depende de las condiciones y lugar en que se desarrolle el

estudio, área del conocimiento, además del error de tipo I que el especialista en el tema investigado esté dispuesto a aceptar para dar por válida su hipótesis.

### Seleccionar y desarrollar la prueba estadística

Para el desarrollo de cualquier proceso investigativo se debe definir la estrategia metodológica y el diseño del estudio, donde se plasmará la idea de investigación, línea de investigación, propósito y enunciado del estudio, población de estudio (aquella que cumple criterios de selección), objetivos del estudio, tamaño de la muestra (debe ser representativa de la población de estudio), unidad muestral (sujeto u objeto donde se efectuará la observación o medición de la o las variables), tipo de muestreo (el más utilizado para la selección de la muestra es el muestreo aleatorio simple, aunque existen otros) y por último efectuar la recolección de datos, la cual permitirá realizar el contraste de la hipótesis del investigador a partir de la selección y desarrollo de la prueba estadística adecuada.

Cuando se trata de escoger que procedimiento estadístico se utilizará para concluir si la hipótesis del investigador es aceptada o rechazada, se debe tener en cuenta que estas pueden ser *paramétricas* o *no paramétricas*, lo cual depende de seis criterios o requisitos a cumplir, estos son:

1. Nivel investigativo.
2. Tipo de estudio.
3. Diseño de la investigación.
4. Objetivo estadístico.
5. Escalas de medición de las variables.
6. Distribución de la variable aleatoria (comportamiento aleatorio de los datos).

En el cuadro 5.7. se identifican las pruebas paramétricas y no paramétricas que se pueden realizar a partir de los criterios descritos anteriormente, cuando el objetivo estadístico es comparar. Sin embargo, cuando la variable numérica, no

cumple con al menos uno de los requisitos exigidos para efectuar pruebas paramétricas (sección "Distribución de la variable aleatoria"), se aplican las pruebas correspondientes a las variables ordinales, denominadas alternativas no paramétricas.

Cuadro 5.7. Principales pruebas estadísticas a desarrollar en función del tipo de estudio y las características de la variable aleatoria.

| Tipo de estudio | | Variable aleatoria (es la que se mide) | | | |
|---|---|---|---|---|---|
| | | Nominal dicotómica | Nominal politómica | Ordinal | Numérica |
| Transversal (muestras independientes) | Variable fija (conformación de grupo) | Pruebas no paramétricas | | | Pruebas paramétricas |
| | Un grupo | Chi-cuadrado (Bondad de Ajuste) | | | T de Student para una muestra |
| | Dos grupos | Chi-cuadrado de Homogeneidad | | U de Mann-Whitney | T de Student para muestras independientes |
| | Más de dos grupos | Chi-cuadrado de Homogeneidad | Análisis de correspondencia | H de Kruskal- Wallis | ANOVA de un factor INTERsujetos |
| Longitudinal (muestras relacionadas) | Dos medidas | Mc Nemar | Q de Cochran | Wilcoxon | T de Student para muestras relacionadas |
| | Más de dos medidas | Q de Cochran | | Friedman | ANOVA de medidas repetidas |

Fuente: Modificado de Supo, 2017. Nota: ANOVA=Análisis de varianza.

### 4. Lectura del p-valor

El cálculo o estimación del error de tipo I se conoce como *p-valor* o valor-p, constituye la magnitud del error que tiene un límite máximo de tolerancia denominado nivel de significancia (α). El *p-valor* es la probabilidad de equivocación, es la cuantificación del error y por tanto se debe efectuar su lectura.

El *p-valor* se puede obtener directamente con la utilización de un software estadístico (SPSS u otro de preferencia) para cualquiera de los procedimientos estadísticos que se pretenda desarrollar en un trabajo investigativo y se calcula para comprobar que la magnitud del error se encuentra por debajo del nivel de significación, lo cual garantiza la aceptación de la hipótesis del investigador. Denotado como el valor de significación (sig.) asintótica, cuando se utiliza SPSS, es el punto de referencia para realizar el contraste de hipótesis.

Por lo tanto, la lectura de la significación (sig.) asintótica (*p-valor*) obtenida es comparada con el nivel de significación fijado previamente por el investigador para realizar la prueba y se utiliza para tomar la decisión en la prueba estadística realizada.

### 5. Tomar la decisión

La decisión a tomar se encuentra relacionada con el *p-valor* calculado en la prueba estadística y el nivel de significación fijado. El *p-valor* al ser un valor de probabilidad varía entre 0 y 1.

Si el *p-valor* obtenido es menor al alfa predefinido para realizar la prueba (α=0,05 u otro) se presenta evidencia estadística que permite dar por válida la hipótesis del investigador. Si se obtiene un *p-valor* mayor al nivel de significancia se rechaza la hipótesis del investigador.

¿Qué sucede si el p-valor es igual a alfa? Esto no debe ocurrir ya que el *p-valor* es una variable numérica continua y el SPSS lo muestra con seis o más decimales (doble clic encima del valor), aunque normalmente se tiende a redondear las

cifras, pero como se trata de tomar una decisión se pueden aumentar los decimales.

Sin embargo, si se obtiene un *p-valor*=0,5000000 se rechaza la hipótesis del investigador ya que no fue posible probar la hipótesis para un $\alpha<0,05$.

### Distribución de la variable aleatoria

El procedimiento estadístico a seguir en cualquier trabajo de investigación científica no se encuentra condicionado al deseo o al conocimiento específico del estudiante o profesional que lo ejecute, sino a los seis criterios descritos en la sección anterior y que es preciso conocer; ya que, de no tenerse en cuenta, pueden conducir a que se generan interpretaciones erróneas y emitir conclusiones incorrectas. Uno de estos criterios es la distribución de la variable aleatoria, la cual debe cumplir con los siguientes requisitos para poder realizar *pruebas paramétricas*.

1. Variable de estudio numérica: medida en escala de intervalo o de razón.

2. Independencia de datos: se deben obtener puntuaciones diferentes debido a que provienen de sujetos u objetos ubicados en diferentes unidades experimentales, no influenciadas unas con otras. El cumplimiento de este requisito se garantiza en el diseño experimental y en la toma de muestras; de no cumplirse se aplica una prueba T de Student para muestras relacionadas (cuando se comparan dos grupos) o un ANOVA de un factor intrasujetos (medidas repetidas) en el caso de comparar tres o más grupos.

3. Normalidad de los datos: los valores de la VD objeto de análisis, en cada una de las poblaciones que se estudian, deben presentar una distribución similar a la distribución normal o gaussiana. Estadísticamente se comprueba con una prueba de normalidad, la cual presenta dos opciones para su interpretación; el Test de Kolmogorov-Smirnov, utilizada cuando el tamaño de la muestra es mayor que cinco casos o sujetos por grupo (corrección de significación de Lilliefors), o la prueba de Shapiro-Wilks, la cual debe ser utili-

zada para el caso de contar con muestras menores de cinco casos o sujetos por grupo. Cuando se cumple este requisito se aplican pruebas paramétricas y cuando no se cumple se pueden utilizar varias alternativas (ver sección "Prueba de normalidad de datos").

4. Homogeneidad de varianzas: la variabilidad de los datos de la variable dependiente, entre los grupos que se comparan, debe ser aproximadamente igual. La prueba estadística utilizada para verificar el cumplimiento o no de este requisito es el Test de Levene (ver sección "Prueba de homogeneidad de varianzas").

### Prueba de normalidad de los datos

Para demostrar la existencia de normalidad en la distribución de datos de una variable se debe seguir el procedimiento para realizar una prueba de hipótesis (sección "Significancia estadística") y desde el punto de vista estadístico constituye un paso necesario cuando se trabaja con variables dependientes numéricas, aunque dicho procedimiento no es el mismo cuando se tiene solamente un grupo o cuando se trabaja con dos o más grupos.

### Prueba de normalidad para una sola muestra (un grupo)

Ejemplo: se determinó el peso de 30 animales atendidos durante una semana en una clínica veterinaria (anexo 1) y se necesita realizar una descripción del comportamiento de la distribución de datos respecto a su aproximación con la distribución normal.

### Procedimiento estadístico

1. Planteamiento de hipótesis

$H_0$: La distribución de la variable peso de los animales es similar a la distribución normal.

$H_1$: La distribución de la variable peso de los animales es diferente a la distribución normal.

2. Establecimiento de nivel de significación: $\alpha=0,05$.

3. Selección y desarrollo de la prueba estadística

Debido a que se necesita conocer la distribución de la variable peso de los animales se aplica la prueba no paramétrica de Kolmogorov-Smirnov para una muestra.

## Pasos en SPSS

Se elige la base de datos definida y se abre con el SPSS>se busca en la barra de menú la pestaña analizar>pruebas no paramétricas>cuadros de diálogos antiguos>clic en K-S de una muestra...> trasladar la variable peso de los animales a lista de variables de prueba>aceptar.

En el visor de resultados del SPSS se muestran los resultados de la Prueba de Kolmogorov-Smirnov de la variable peso (kg) de los animales (Cuadro 5.8).

Cuadro 5.8. Prueba de Kolmogorov-Smirnov para una muestra (variable peso de los animales).

| | | |
|---|---|---|
| N | | 30 |
| Parámetros normales[a,b] | Media | 18,5067 |
| | Desviación estándar | 12,88351 |
| Máximas diferencias extremas | Absoluta | ,122 |
| | Positivo | ,121 |
| | Negativo | -,122 |
| Estadístico de prueba | | ,122 |
| Sig. asintótica (bilateral) | | ,200c,d |

a. La distribución de prueba es normal.

b. Se calcula a partir de datos.

c. Corrección de significación de Lilliefors.

d. Esto es un límite inferior de la significación verdadera.

4. Lectura del *p-valor*

El valor de la significación asintótica bilateral obtenido (*p-valor*) es de 0,200; mayor a 0,05 por lo que no se rechaza hipótesis nula.

5. Tomar la decisión

Mediante la prueba de Kolmogorov-Smirnov para una muestra realizada y con una probabilidad de error del 20,0% se concluye que la distribución de la variable peso de los animales es similar a la distribución normal.

### Prueba de normalidad para dos o más grupos

Ejemplo: se necesita conocer el porcentaje de supervivencia de camarones a los cuatro meses de manejo en las diferentes piscinas de la camaronera ya que se sospecha que dicha variable puede cambiar en relación con el laboratorio de procedencia de las larvas. Se efectuaron las mediciones y los datos se organizaron y tabularon en Microsoft Excel (anexo 6). Se requiere determinar si las distribuciones de datos de la variable supervivencia (%) para cada laboratorio de procedencia de las larvas cumplen con el requisito de normalidad de sus datos.

### Procedimiento estadístico

1. Planteamiento de hipótesis

$H_0$: La distribución de datos de la variable supervivencia de los camarones para cada laboratorio de procedencia es similar a la distribución normal.

$H_1$: La distribución de datos de la variable supervivencia de los camarones para cada laboratorio de procedencia no es similar a la distribución normal.

2. Establecimiento de nivel de significación: $\alpha$=0,05.

3. Selección y desarrollo de la prueba estadística: Prueba de normalidad

## Pasos en el SPSS

Abrir la base de datos con el SPSS>buscar en la barra de menú y dar clic en la opción analizar>estadísticos descriptivos>explorar>trasladar la variable de estudio a lista de dependientes y los grupos formados a lista de factores>clic en la pestaña gráficos> en descriptivos se deshabilita la opción de tallo y hojas ya que es innecesaria en esta prueba y se selecciona la opción gráficos de normalidad con pruebas>continuar>aceptar.

En el visor de resultados del SPSS se muestran los resultados de la prueba de normalidad (Cuadro 5.9.).

Cuadro 5.9. Prueba de normalidad de datos.

| Variable | Laboratorio de procedencia | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|---|
| | | Estadístico | gl | Sig. | Estadístico | gl | Sig. |
| Supervivencia (%) | DOBLE T | ,298 | 4 | . | ,926 | 4 | ,572 |
| | AQUALAB | ,162 | 5 | ,200* | ,971 | 5 | ,884 |
| | REYDAMAR | ,205 | 5 | ,200* | ,933 | 5 | ,618 |
| | NUTRIAGRO | ,202 | 5 | ,200* | ,933 | 5 | ,619 |
| | GENESIS | ,329 | 4 | . | ,895 | 4 | ,406 |

*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

4. Lectura del *p-valor*

Teniendo en cuenta que se presentan dos grupos con cuatro *gl* se escoge la prueba de Shapiro-Wilk para efectuar la interpretación. Se observa que para todos los grupos el p-valor fue mayor que 0,05 por lo que no se rechaza la hipótesis nula.

5. Tomar la decisión

A través de la prueba de Shapiro-Wilk y con una probabilidad de error por encima de 40,6% para todos los grupos

se concluye que los datos siguen una distribución similar a la distribución normal y se justifica la utilización de pruebas paramétricas de cumplirse los demás requisitos.

La principal ventaja de utilizar una distribución de datos similar a la distribución normal o gaussiana radica en el cumplimiento del supuesto que indica que el 99% de los valores que la conforman se encuentran diseminados a tres desviaciones estándar de distancia de la media aritmética (Solarte *et al*., 2009).

En caso del no cumplimiento del requisito de normalidad de datos se deben seguir las siguientes alternativas; eliminar los valores atípicos (cuando existan), transformar los datos (se debe comprobar que la transformación de los datos garantiza el cumplimiento del requisito) o aplicar las pruebas estadísticas no paramétricas descritas en la Tabla 5; aunque es importante acotar que en caso de comparar dos grupos se debe utilizar la prueba U de Mann Whitney; y si se presentan tres grupos o más se utiliza la prueba H de Kruskal-Wallis; aunque para saber entre que grupos se presentan las diferencias significativas se utiliza la U de Mann-Whitney (pruebas post-hoc), debiéndose ajustar el error, es decir, se debe dividir el error fijado ($\alpha$=0,05) por el número de contrastes que se realicen (tres en este caso).

### Prueba de homogeneidad de las varianzas

Las varianzas de los grupos comparados serán homogéneas cuando la magnitud de su variabilidad no sea diferente. La verificación de la homogeneidad o heterodasticidad en las varianzas entre los grupos objeto de estudio se ejecuta dentro del mismo procedimiento para realizar el ANOVA, aunque en este caso el procedimiento se explicará de forma específica.

### Dos grupos

La aplicación y explicación de esta prueba cuando se trabaja con dos grupos, se efectuará en la sección “Prueba T de Student para muestras independientes”, debido a que al

realizar la prueba T Student para muestras independientes el software estadístico genera una tabla por defecto donde muestra de forma simultánea los resultados de la Prueba de Levene de igualdad de varianzas y de la prueba T Student para la igualdad de medias.

### Más de dos grupos

Ejemplo: en un estudio observacional desarrollado en la granja Santa Inés de la Universidad Técnica de Machala con la finalidad de conocer la influencia del manejo del sistema productivo en la compactación del suelo, se efectuó muestreo de suelo a una profundidad de 0-15 cm en áreas dedicadas a pasto, cacao, banano, maíz y bosque; y se determinó mediante análisis de laboratorio, entre otras determinaciones físicas, la densidad aparente del suelo (g $cm^3$) (anexo 7). Se necesita establecer si existe diferencia significativa en la densidad aparente del suelo según el sistema productivo implementado. Objetivo del estudio: Contrastar la hipótesis de igualdad de varianzas en la densidad aparente del suelo en los diferentes cultivos a una profundidad entre 0-15 cm.

### Procedimiento estadístico

1. Planteamiento de hipótesis

$H_0$: Se asumen varianzas homogéneas para la densidad del suelo en los diferentes sistemas productivos a una profundidad de 0-15 cm.

$H_1$: No se asumen varianzas homogéneas para la densidad del suelo en los diferentes sistemas productivos a una profundidad de 0-15 cm.

2. Establecimiento de nivel de significación: $\alpha$=0,05.

3. Selección y desarrollo de la prueba estadística

Prueba de homogeneidad de varianzas. Test de Levene

## Pasos en el SPSS

Analizar>comparar medias>ANOVA de un factor>se traslada la variable de interés que es numérica para el recuadro de Lista de dependientes y para Factor la variable independiente (grupos)>clic en opciones y se selecciona la pestaña Prueba de homogeneidad de las varianzas, la cual es la que interesa en este ejemplo>continuar>aceptar.

En el visor de resultados del SPSS se muestra la tabla con la prueba de homogeneidad de varianzas para la variable densidad aparente del suelo a una profundidad de 0-15 cm (Cuadro 5.10.).

Cuadro 5.10. Prueba de homogeneidad de varianzas (Test de Levene) para la variable dependiente densidad aparente del suelo.

| Estadístico de Levene | gl1 | gl2 | Sig. |
|---|---|---|---|
| 2,023 | 4 | 10 | ,167 |

4. Lectura del *p-valor*

Se obtuvo un *p-valor* de 0,167 mayor que 0,05, por lo que se acepta la hipótesis nula que indica que se asumen varianzas homogéneas en los sistemas de producción estudiados.

5. Tomar la decisión

Al concluir la prueba de Levene y con una probabilidad de error del 16,7% se concluye que se presenta homogeneidad en las varianzas en función de la variable densidad aparente del suelo en los cinco sistemas productivos evaluados.

Es preciso aclarar que el procedimiento explicado anteriormente se encuentra implícito dentro de la prueba estadística de ANOVA de un factor intersujetos, por lo que es innecesario realizarlo de manera independiente, ya que cuando esta prueba se ejecuta, puede ser seleccionada la opción que permite la verificación de este requisito.

El requisito de homogeneidad de varianzas normalmente se incumple cuando el tamaño de la muestra de cada grupo es diferente, debiéndose utilizar alternativas como Brown-Forsythe y Welch, las cuales son estadísticos que ajus-

tan los *gl* de los residuos (son capaces de eliminar los problemas del no cumplimiento de la homogeneidad de las varianzas) y controlan el error de tipo I. Aunque Welch tiene mayor potencia al detectar mejor el efecto en caso de que exista. Sin embargo, es aceptado continuar con el procedimiento de ANOVA sin tener en cuenta lo descrito anteriormente, ya que es una prueba muy robusta que mantiene el error de tipo I en un 5% (Serra, 2017).

Cuando se cuente con variables numéricas que cumplen con los requisitos de normalidad e independencia de datos y homogeneidad en las varianzas se deben aplicar pruebas paramétricas; y cuando se trabaje con variables categóricas nominales (dicotómicas o politómicas), u ordinales; o cuando las variables numéricas no cumplan con al menos uno de los requisitos descritos en la sección "Distribución de la variable aleatoria" se deben emplear pruebas no paramétricas, las cuales, constituyen alternativas que no presentan la robustez de las pruebas paramétricas, sin embargo, alcanzan un grado de confiabilidad aceptable.

### Pruebas paramétricas

Se aplican cuando los datos cumplen con los requisitos analizados en la sección "Distribución de la variable aleatoria". Las pruebas paramétricas presentan mayor capacidad para detectar una relación real o verdadera entre dos variables, cuando esta existe. Pueden ser para una muestra, para dos o más muestras independientes y para dos o más muestras relacionadas.

### Prueba T de Student para muestras independientes

Es una prueba de contraste de hipótesis que se utiliza para comparar medias de dos grupos que no presenten relación entre ellos (poblaciones independientes) y la variable objeto de estudio es numérica.

El test determina si los grupos presentan o no diferencia significativa en relación con la variable numérica analizada.

Ejemplo: se necesita determinar si la cáscara de banano

maduro desechada en el proceso de industrialización de la fruta presenta un efecto diferente en la producción de leche en ganado vacuno, al compáralo con el efecto del banano verde. Para el desarrollo del estudio se formaron dos grupos de vacas reproductoras con características homogéneas, conformados por 18 animales cada uno y de forma conjunta se les brindó pastoreo por un tiempo de cinco horas y alimentación en los corrales a base de 47,0 kg de palmiste; 9,0 kg de gallinaza; 34,4 kg de caramelo de banano, 331,8 kg de raquis de banano, 10,8 kg de carbonato de calcio y 2,7 kg de sal común, durante 61 días. Adicionalmente, y de forma separada se le brindó al grupo control 222,4 kg de banano verde y al grupo experimental 204,2 kg de cáscara de banano maduro (diseño de casos y controles). La recolección de la información (producción de leche en kilogramos), se efectuó diariamente durante ocho semanas (Castro *et al*., 2018). La base de datos con los promedios semanales se muestra en el anexo 5.

### Procedimiento estadístico

1. Planteamiento de hipótesis

$H_0$: La producción de leche de vaca es igual en los grupos de casos y controles.

$H_1$: La producción de leche de vaca es diferente en los grupos de casos y controles.

2. Establecimiento de nivel de significación: $\alpha$=0,05.

3. Selección y desarrollo de la prueba estadística: para realizar el contraste de la hipótesis se desarrolla la Prueba T para muestras independientes. Lo que se necesita conocer es si la diferencia entre los grupos es significativa en relación con la producción de leche.

### Pasos en el SPSS

Buscar la matriz de datos y abrirla con el SPSS>seleccionar en la barra de menú la opción analizar>comparar medias>se elige prueba T para muestras independientes, se traslada la

variable a contrastar, en este caso la producción de leche y la variable grupos se traslada a variable se agrupación>clic en definir grupos>se coloca 1 para el grupo 1 y 2 para el grupo 2 debido a que solamente se cuenta con dos grupos>clic en opciones y se define el porcentaje de intervalo de confianza, en este caso es 95% ya que se estableció un alfa de 0,05>continuar>aceptar.

En el visor de resultados del SPSS se muestra la prueba T para muestras independientes que se genera, en la cual se exponen los valores de significación obtenido en el Test de Levene de igualdad de varianzas y en la prueba T para igualdad de medias (Cuadro 5.11.).

Cuadro 5.11. Prueba T para muestras independientes (comparación de casos y controles) donde se muestran los resultados de la prueba de Levene y la prueba t para la igualdad de medias.

| Producción de leche (kg) | Prueba de Levene de igualdad de varianzas | | Prueba t para la igualdad de medias | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | F | Sig. | t | gl | Sig. (bilateral) | Diferencia de medias | Diferencia de error estándar | 95% de intervalo de confianza de la diferencia | |
| | | | | | | | | Inferior | Superior |
| Se asumen varianzas iguales | ,003 | ,959 | ,530 | 286 | ,596 | ,10618 | ,20018 | -,28783 | ,50019 |
| No se asumen varianzas iguales | | | ,530 | | ,596 | ,10618 | ,20018 | -,28783 | ,50019 |

4. Lectura del *p-valor*

Debido a que el *p-valor* obtenido en la prueba de Levene (0,959) es mayor a 0,05 se asumen varianzas homogéneas y se escoge la fila superior para efectuar la interpretación

estadística de la prueba T Student para muestras independientes, en la que se obtiene un *p-valor* de 0,596, el cual es mayor a 0,05; por lo tanto, se acepta la hipótesis nula que indica que la producción de leche es igual entre los grupos estudiados.

5. Tomar la decisión

A partir de la prueba T para muestras independientes realizada y con una probabilidad de error de 59,6% se concluye que la producción de leche en vacas reproductoras no presenta diferencias estadísticamente significativas entre el grupo de casos, alimentados con un complementa a base de cáscara de banano maduro desechada y el grupo control, alimentados de forma complementaria con banano verde no comercializado. El estudio realizado indica que la utilización del residuo formado por la cáscara del banano maduro permite disminuir los costos de producción de leche y reutilizar desechos que provocan contaminación en el ambiente.

## Análisis de varianza

El análisis de varianza (ANOVA por sus siglas en inglés) es una herramienta estadística que se utiliza para conocer la varianza que se presenta en uno o más factores de estudio (variables independientes), definidas a partir del criterio de conformación de grupos utilizado por el investigador) con respecto a una o varias variables numéricas medidas en la investigación (variables dependientes).

El ANOVA es una extensión de la prueba T Student para muestras independientes y tiene la ventaja que permite realizar comparaciones de medias en estudios con más de dos grupos sin que el error de tipo I se incremente.

A continuación, se mencionan los diferentes tipos de procedimientos para realizar un ANOVA, los cuales se asocian con el número de factores que se estudian (uno o más de uno) y si las observaciones realizadas son independientes en cada grupo o son medidas repetidas dentro de los grupos.

– ANOVA de un factor intersujetos (entre grupos independientes).

– ANOVA de un factor intrasujetos (medidas repetidas o relacionadas dentro del grupo).

– ANOVA factorial intersujetos (dos o más factores entre grupos independientes).

– ANOVA factorial intrasujetos (medidas repetidas o relacionadas dentro del grupo).

– ANOVA factorial de diseño mixto (grupos independientes y medidas repetidas).

A continuación de efectuará un análisis teórico de los elementos asociados con la ejecución manual y el desarrollo de un ANOVA de un factor intersujetos; y una posterior interpretación de los resultados una vez aplicado el procedimiento estadístico con ayuda del paquete estadístico SPSS. Los demás tipos de ANOVAS mencionados anteriormente no se abordarán en este capítulo.

### ANOVA de un factor intersujetos o entresujetos

Es un procedimiento matemático que permite determinar la presencia o no de diferencias significativas entre las medias de los grupos o tratamientos analizados en función de la variable objeto de estudio. Comprueba la variabilidad asociada a las condiciones de los grupos (varianza del modelo) y la variabilidad debida al azar, o debida a las condiciones intrínsecas de cada sujeto u objeto que se encuentra dentro de cada tratamiento establecido en las unidades experimentales (variación de cada sujeto respecto a la media de su grupo, llamada varianza residual).

– Suma de cuadrados total (SCT). La variabilidad total definida por medio de la SCT se calcula mediante la sumatoria de la diferencia que existe entre cada uno de los datos obtenidos y la gran media elevado al cuadrado, o a través de la sumatoria de los valores de los efectos atribuibles a las condiciones de los grupos y los efectos no atribuibles o residuales, una vez que han sido determinados (SCT=SCTr+SCE).

Estos se calculan a partir de:

– Suma de cuadrados de los tratamientos (SCTr). Denominada también suma de cuadrados del modelo. Es utilizada para conocer la variabilidad total que es explicada por el modelo y se calcula mediante la sumatoria de la diferencia existente entre las medias de cada uno de los grupos o tratamientos y la gran media elevada al cuadrado, y cada resultado es multiplicado por el número de casos presentes en cada grupo o tratamiento. Los grados de libertad (*gl*) de la SCTr se corresponden con el número de grupos o tratamientos de comparación menos uno (k-1). La media cuadrática para los tratamientos se obtiene de dividir la SCTr y sus *gl*.

– Suma de cuadrados de los errores (SCE). Denominada también suma de cuadrados residual. Utilizada para determinar la variabilidad residual dentro de cada grupo no atribuible al modelo (puede ser causada por variables extrañas, condición específica de los individuos u otras) y se calcula mediante la sumatoria de la diferencia que se presenta entre todos los datos obtenidos y la media aritmética del grupo o tratamiento de donde proviene la observación elevado al cuadrado. Los *gl* (número de observaciones que son libres de variar) de la SCE se obtienen de la resta de los gl totales (número total de sujetos u objetos menos uno) y los gl del modelo. La media cuadrática para los residuos se obtiene de dividir la SCE y sus gl.

El estadístico F calculado se obtiene de la división de los cuadrados medios de los tratamientos y de los errores y es el valor que se utiliza para comparar con el F de la tabla y determinar si existe diferencia significativa o no entre los grupos o tratamientos, aunque con la ayuda del software se obtiene el valor de significación de la prueba, el cual se compara con el alfa predefinido para realizar la prueba y es suficiente para realizar el contraste de hipótesis y tomar la decisión.

Ejemplo: en un estudio observacional desarrollado en la granja Santa Inés de la Universidad Técnica de Machala con la finalidad de conocer la influencia del manejo del sistema productivo en el pH del suelo, se efectuó muestreo de

suelo a una profundidad de 0-15 cm en áreas dedicadas a pasto, cacao, banano, maíz y bosque; y se determinó el pH en agua (anexo 8). Se necesita establecer si existe diferencia significativa en el pH del suelo según el sistema productivo implementado. Objetivo del estudio: contrastar la hipótesis que plantea que las medias de pH del suelo son diferentes según el tipo de cultivo establecido.

**Procedimiento estadístico**

1. Planteamiento de hipótesis

$H_0$: Las medias de pH del suelo en los sistemas productivos son iguales.

$H_1$: Las medias de pH del suelo en sistemas productivos son diferentes.

2. Establecimiento del nivel de significación: $\alpha$=0,05.

3. Selección y desarrollo de la prueba estadística: debido a que se cuenta con cinco grupos y la variable dependiente es numérica se aplica la prueba paramétrica ANOVA de un factor intersujetos.

Primeramente, se identifican los tipos de variables estudiadas, en este caso la VI (sistema productivo), la cual constituye el factor de estudio y cuenta con cinco versiones (pasto, cacao, banano, maíz y bosque) y la VD (pH del suelo a la profundidad de 0-15 cm).

Una vez cargada la base de datos en el visor de datos del SPSS (vista de datos y vista de variables) se selecciona dentro de la barra de menú y se da clic en la pestaña analizar>comparar medias>ANOVA de un factor>se traslada a la lista de dependientes la variable pH del suelo y a factor los grupos objeto de estudio o de comparación, en este caso, la variable sistemas de producción, clic en la pestaña opciones>se selecciona los descriptivos, la prueba de homogeneidad de varianzas para comprobar el cumplimiento del requisito de homogeneidad de varianza; además, se seleccionan las pruebas Brown-Forsythe y Welch, las cuales ajustan los *gl* y constituyen alternativas al Ratio F que permiten utilizar la

prueba paramétrica en caso de presentarse un incumplimiento del requisito de homogeneidad de varianzas>continuar>seleccionar la pestaña post-hoc> se selecciona la prueba de rangos múltiples en función de las características de los grupos, en este caso Duncan que se utilizan cuando se cumple el supuesto de homogeneidad y se presentan pocos grupos y Games-Howell para cuando no se asuman varianzas homogéneas>continuar>aceptar.

En la tabla ANOVA de un factor intersujetos presentada en el visor de resultados del SPSS se muestran las fuentes de variación relacionadas con los valores para los inter-grupos o entre los grupos (varianza del modelo) y en los intra-grupos o dentro de grupos, denominada como varianza residual (Cuadro 5.12.).

Cuadro 5.12. ANOVA de un factor intersujetos para la variable pH del suelo.

| Fuentes de variación | Suma de cuadrados | gl | Media cuadrática | F | Sig. |
|---|---|---|---|---|---|
| Entre grupos | 2,339 | 4 | ,585 | 3,612 | ,045 |
| Dentro de grupos | 1,619 | 40 | ,162 | | |
| Total | 3,959 | 14 | | | |

4. Lectura del *p-valor*

El *p-valor* obtenido en la prueba realizada (0,045) es menor que el nivel de significación fijado de antemano para realizar la prueba (α=0,05) por lo que se acepta la hipótesis del investigador.

5. Tomar la decisión

Una vez concluida la prueba de ANOVA de un factor intersujetos y con una probabilidad de error de 4,5% se concluye que las medias de pH del suelo son diferentes en al menos un sistema productivo; aunque se desconoce entre qué sistema productivo se presentan dichas diferencias o las igualdades; por lo que es necesario revisar el resultado de las pruebas de comparaciones múltiples, denominadas pruebas post-hoc.

## Pruebas post-hoc

Las pruebas post-hoc (a posteriori), constituyen alternativas que se encuentran implementadas en el paquete estadístico SPSS (18 pruebas en total; 14 para cuando se asumen varianzas homogéneas y cuatro para el caso que no se asumen varianzas homogéneas), son aquellas que permiten conocer entre que grupos o tratamientos de comparación se presentan las diferencias o similitudes (Lizasoain & Joaristi, 2003). Cuando no se presentan diferencias significativas entre los grupos o tratamientos no es necesario interpretar la prueba post-hoc y se concluye el procedimiento estadístico, aunque puede existir un análisis de datos desde el punto aritmético.

Las pruebas post-hoc son aquellas que controlan el error de tipo I (se produce en la comparación de cada uno de los pares de grupos por separado), algunas son más robustas que otras y se eligen en función del cumplimiento del requisito de homogeneidad de varianzas, número de grupos conformados y la cantidad de observaciones realizadas por grupo (Cuadro 5.13.).

Cuadro 5.13. Principales pruebas de comparaciones múltiples (post-hoc) aplicadas cuando se obtienen diferencias estadísticas significativas entre los tratamientos objeto de estudio.

| Requisito | Prueba de comparaciones múltiples | Características |
|---|---|---|
| Se asumen varianzas homogéneas entre los grupos | Diferencia Menos Significativa (DMS) | No controla el error de tipo I en el SPSS (no se utiliza porque se incrementa el error de tipo I). |
| | Bonferroni | Es más potente cuando el número de grupos a comparar es pequeño y los tamaños muestrales son diferentes, aunque debe formarse subconjuntos homogéneos a partir de la tabla de comparaciones múltiples que emite el SPSS. |

| Requisito | Prueba de comparaciones múltiples | Características |
|---|---|---|
| | Tukey | Es más potente cuando el número de grupos a comparar es grande y el tamaño de las muestras de los diferentes grupos es similar. |
| | Duncan | Es muy robusta, se utiliza cuando se presentan pocos grupos y el tamaño de las muestras son iguales. |
| | Gabriel | Es adecuada cuando el número de grupos a comparar es grande y los tamaños muéstrales son distintos en los diferentes grupos. |
| No se asumen varianzas homogéneas entre los grupos | Games-Howell | Aquella que se utiliza cuando no se asumen varianzas homogéneas y los tamaños muéstrales son distintos, aunque debe formarse subconjuntos homogéneos a partir de la tabla de comparaciones múltiples que emite el SPSS. |

Fuente: Adaptado de Serra (2017).

## Interpretación de los resultados

Debido a que previamente se obtuvo un *p-valor* de 0,002 en el estadístico de Levene no se asumen varianzas homogéneas, por lo que la prueba de comparaciones múltiples a desarrollar, para efectuar la interpretación, es la prueba de comparaciones múltiples de Games-Howel (anexo 9). Teniendo en cuenta que el asterisco significa diferencia significativa entre los sistemas productivos comparados, se elabora la tabla de subconjuntos homogéneos (Cuadro 5.14.) y se concluye que en el cultivo del banano es donde se presenta el mayor valor de pH del suelo (8,24), no diferente estadísticamente a los valores obtenidos en los cultivos de pasto (8,16) y maíz (7,76), aunque si lo hace con los valores alcanzados en bosque (7,42) y cacao (7,24), lo que evidencia que en los sistemas de producción se presentan condiciones de basicidad.

Cuadro 5.14. Tabla de subconjuntos homogéneos elaborada a partir de la prueba Games-Howell que emite las comparaciones múltiples entre los sistemas productivos en relación con el pH del suelo.

| Sistema de producción | N | Subconjunto para alfa = 0,05 | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| Cacao | 3 | 7,24 | | |
| Bosque | 3 | 7,42 | 7,42 | |
| Maíz | 3 | 7,76 | 7,76 | 7,76 |
| Pasto | 3 | | 8,16 | 8,16 |
| Banano | 3 | | | 8,24 |
| Sig. | | ,160 | 0,055 | 0,195 |

## Pruebas no paramétricas

Las pruebas no paramétricas se aplican para variables categóricas (nominales u ordinales) o cuando la variable numérica no cumple con al menos uno de los requisitos descritos en la sección "Distribución de la variable aleatoria".

## Prueba Chi-cuadrado

Chi-cuadrado de Pearson (denotada por $\chi^2$), es una prueba estadística utilizada para evaluar hipótesis acerca de la relación o asociación que se presenta entre dos variables categóricas (nominal u ordinal) ya sean dicotómicas o politómicas. Se aplica en el nivel relacional, el cual es bivariado y casi siempre la variable grupo es fija y la otra es aleatoria. Es la prueba no paramétrica más utilizada en la investigación científica, siempre parte del supuesto que indica que las dos variables en estudio no se encuentran relacionadas desde el punto de vista probabilístico y su finalidad es identificar diferencias entre los grupos participantes (pueden ser dos o más). Tiene numerosas aplicaciones, entre las que se encuentran:

– Prueba $\chi^2$ de bondad de ajuste (utilizada para comparar la frecuencia evaluada en un grupo con el parámetro de su población y la variable aleatoria es categórica).

– Prueba $\chi^2$ de independencia (utilizada cuando ambas variables son categóricas dicotómicas o politómicas, lo que

indica el estudio de la asociación entre las categorías de dichas variables).

– Prueba $\chi^2$ de homogeneidad (utilizada cuando la variable aleatoria es categórica politómica).

Ejemplo: se necesita conocer la posible influencia del lugar de tenencia de los animales (casa-terraza, terraza-patio, patio-calle) sobre el padecimiento a *B*. canis en caninos atendidos en una clínica veterinaria. Para ello, se crearon dos grupos de animales en función del padecimiento a *B*. canis (variable fija), un grupo con la enfermedad y el otro sin el padecimiento; y con los dueños de las mascotas se recolectó la información relacionada con el lugar de tenencia (variable aleatoria) durante una semana, posteriormente los datos fueron ordenados y tabulados en Microsoft Excel (base de datos del anexo 2).

**Procedimiento estadístico**

1. Planteamiento de hipótesis

$H_0$: El lugar de tenencia de los caninos no se encuentra relacionado con el padecimiento a *B. canis*.

$H_1$: El lugar de tenencia de los caninos se encuentra relacionado con el padecimiento a *B. canis*.

2. Nivel de significación de la prueba: $\alpha$=0,05.

3. Selección y desarrollo de la prueba estadística

Una vez organizada la matriz de datos en la vista de variables y vista de datos del SPSS se procede a aplicar el Test de Chi-cuadrado, la cual se realiza a continuación del procedimiento aplicado para obtención de la tabla de contingencia descrito en la sección “Tablas de contingencia”.

**Pasos en el SPSS**

Buscar la base de datos y abrirla con el SPSS>en la barra de menú del software buscar la opción analizar>estadísticos descriptivos>tablas cruzadas>pasar la variable fija (los dos grupos creados en función del padecimiento a *B. canis*)

para columnas y para la opción filas se traslada a la variable cuya frecuencia se necesita analizar (lugar de tenencia de los animales)>clic en casillas y dentro de la opción mostrar en casillas se selecciona dentro de porcentajes a la columna ya que lo que se compara son las columnas; lo cual se realiza a través de los porcentajes>continuar>estadísticos>seleccionar el test de Chi-cuadrado>continuar>aceptar.

En el visor de resultados del SPSS se muestra primeramente la TC analizada en la sección "Prueba Chi-cuadrado" en la cual se compara la frecuencia de las categorías de la variable lugar de tenencia dentro del grupo afectado con la patología y el no afectado. Sin embargo, las diferencias numéricas observadas no son suficientes para concluir que se presenta una asociación entre los lugares de tenencia y los casos positivos a *B. canis*, es por ello que se solicitó la prueba estadística chi-cuadrado que se muestra en la Cuadro 5.15.

Cuadro 5.15. Pruebas chi-cuadrado de homogeneidad.

| | Valor | df | Sig. asintótica (bilateral) |
|---|---|---|---|
| Chi-cuadrado de Pearson | 1,385a | 2 | ,500 |
| N de casos válidos | 40 | | |
| a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 6,50. | | | |

Lectura del *p-valor*

El valor de significación asintótica (bilateral) obtenido es de 0,500; el cual es mayor al nivel de significación establecido para realizar la prueba (0,05), por lo que no se presenta evidencia estadística para aceptar la hipótesis alternativa.

5. Tomar la decisión

A partir del Test de Chi-cuadrado y con una probabilidad de error del 50,0% se concluye que no existe asociación entre el lugar de tenencia de los caninos y el padecimiento a *B. canis*, o sea, son variables independientes; el padecimiento a la patología puede encontrarse condicionado a otras variables, pero no al lugar de tenencia de los caninos.

## Referencia Bibliográfica


Aguilera, A. M. (2005). *Análisis de tablas de contingencia bidimensionales*. Obtenido de http://www.ugr.es/~focana/dclasif/aaguilera.pdf

Aguilera, A. M. (2006). *Modelización de tablas de contingencia multidimensionales*. Editorial La Muralla. Obtenido de https://www.casadellibro.com/libro-modelizacion-de-tablas-de-contingencia-multidimensionales/9788471337603/1106248

Aguirre, C., & Vizcaino, M. (2010). *Aplicación de estimadores estadisticos y diseños experimentales en investigaciones forestales*. Ibarra, Ecuador: Editorial Universitaria. Universidad Técnica del Norte.

Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2008). *Estadística para administración y economía*. Santa Fe: CENGAGE LEARNING. 10a. edición. ISBN-13: 978-607-481-319-7.

Barnett, V. (1991). *Sample Survey Principles and Methods*. Londres: Edward Arnold.

Batanero, C. (2001). *Didáctica de la estadística*. Granada, España: Grupo de Educación Estadística de la Universidad de Granada (GEEUG). ISBN: 84-699-4295-6.

Canavos, G. C. (1988). *Probabilidad y estadística. Aplicaciones y métodos*. México: Editorial McGRAW HILL. Obtenido de https://estadisticaunicaes.files.wordpress.com/2012/05/george-c-canavos-probabilidad-y-estadc3adstica-aplicaciones-y-mc3a9todos.pdf

Casas, J. M. (1996). *Inferencia estadística para economía y administración de empresas*. Madrid, España: Editorial Universitaria. ISBN: 9788480041959. Obtenido de https://www.casadellibro.com/libro-inferencia-estadistica-para-economia-y-administracion-de-empresas/9788480041959/512409

Castañeda, M. B. (2010). *Procesamiento de datos y análisis estadísticos utilizando SPSS*. Río grande do sur: EDIPRUC.

Castro, A., Rodríguez, I., & Ramírez, I. (2018). Evaluación de la producción de leche en bovinos alimentados con cáscara de banano maduro. *Revista Científica Agroecosistemas, 6*(1), 108-114. Obtenido de https://aes.ucf.edu.cu/index.php/aes/article/view/171/206

Depool, R., & Monasterio, D. (2013). *Probabilidad y estadística. Aplicaciones a la ingeniería*. Barquisimeto, Venezuela: Universi-

dad Nacional Experimental Politécnica Antonio José de Sucre (Unexpo). Obtenido de http://www.bqto.unexpo.edu.ve/avisos/PROBABILIDADYESTADISTICA(2-7-13).pdf

Development Core Team. (2008). *A language and environment for statistical computing*. Vienna, Austria: Foundation for Statistical Computing. ISBN 3-900051-07-0. Obtenido de URL http://www.R-project.org.

EPPO. (2012). *Global Database*. European and Mediterranean Plant Protection Organization. Obtenido de Obtenido en: http://gel.eppo.int

FAOSTAT. (2017). *Área cosechada y producción mundial de cultivos por paises*. Rome, Italy: Food and Agriculture Organization of the United Nations. Obtenido de http://www.fao.org/faostat/es/#data/QC/visualize

Fisher, R. A. (1954). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd. 2th Ed., 1954.

Garriga, A. J., Lubin, P., Merino, J. M., Padilla, M., Recio, P., & Suárez, J. C. (2010). *Introducción al análisis de datos*. Madrid, España: Universidad Nacional de Educación a Distancia (UNED). Obtenido de https://es.scribd.com/doc/110899354/Introduccion-al-Analisis-de-Datos

Gorgas, J., Cardiel, N., & Zamorano, J. (2011). *Estadísticas básica para estudiantes de ciencias*. Madrid, España: Departamento de Astrofísica y Ciencias de la Atmósfera. Facultad de Ciencias Físicas. Universidad Complutense de Madrid. Obtenido de http://webs.ucm.es/info/Astrof/users/jaz/ESTADISTICA/libro_GCZ2009.pdf

Hernández, R., Fernández, C., & Baptista, P. (2014). *Metodología de la investigación* (Sexta ed.). D.F. México: McGraw-Hill.

Horra, J. D. (2003). *Estadística aplicada*. Editorial Díaz de Santos. Obtenido de https://estadisticaunicaes.files.wordpress.com/2012/05/uned-estadc3adstica-aplicada-julic3a1n-de-la-horra.pdf

IBM Corp. (2016). *SPSS Statistics versión 24.0.0.0 de prueba para Windows*. Barcelona: International Business Machines Corp.

INAMHI. (2015). *Anuario Meteorológico. No 251-2011*. Quito, Ecuador: Instituto Nacional de Meteorología e Hidrología, República del

Ecuador. Obtenido de http://www.serviciometeorologico.gob.ec/wp-content/uploads/anuarios/meteorologicos/Am%202012.pdf

INEC. (2012). *Estadísticas del Ecuador*. Quito - Ecuador : Instituto Nacional de Estadística y Censos. Obtenido de http://www.ecuadorencifras.gob.ec.

INEC. (24 de enero de 2016). *Instituto Nacional de Estadística y Censos*. Obtenido de Instituto Nacional de Estadística y Censos: http://www.ecuadorencifras.gob.ec/

Johnson, R., & Kuby, P. (2012). *Estadística elemental*. México: 11ª edición. CENGAGE Learning. ISBN. 978-0-538-73350-2.

Lind, D. A., Marchal, W. G., & Mason, R. D. (2004). *Estadística para administración y economía*. Bogotá, Colombia: Editorial Alfaomega. 11a Edición. Obtenido de https://es.scribd.com/doc/285320909/Estadistica-para-la-administracion-y-Economia-Lind-Marchal-11-edicion-pdf

Lind, D. A., Marchal, W. G., & Wathen, S. A. (2015). *Estadística aplicada a los negocios y la economía*. México, D. F.: McGraw-Hill Internacional.

Lizasoain, L., & Joaristi, L. (2003). *Gestión y análisis de datos con SPSS. Versión 11.* . Madrid: Thomson.

Maronna, R. A. (1995). *Probabilidad y Estadística Elementales para Estudiantes de Ciencias*. La Plata, Argentina: Facultad de Ciencias Exactas Universidad Nacional de La Plata. Obtenido de http://www.mate.unlp.edu.ar/~maron/MaronnaHome_archivos/Probabilidad%20y%20Estadistica%20Elementales.pdf

Milton, J. (1994). *Estadìstica para Biologìa y Ciencias de la Salud*. Madrid: McGraw-Hill.

Montero, I., & León, O. G. (2005). Sistema de clasificación del método en los informes de investigación en Psicología. *International Journal of Clinical and Health Psychology. Asociación Española de Psicología Conductual. Granada. España,* 5(1), 115-127.

Montgomery, D. C. (1991). *Diseño y Análisis de ExperimentoS*. México, D.F: Editorial Iberoamericana, primera edición. Obtenido de http://www.sidalc.net/cgi-bin/wxis.exe/?IsisScript=UCC.xis&method=post&formato=2&cantidad=1&expresion=mfn=028167

Morales, A. E. (2012). *Estadística y probabilidades*. Santiago de Chile.

Chile: Editorial Universidad Católica de la Santísima Concepción. Instituto Profesional Virgilio Gómez. Chile. Obtenido de http://www.x.edu.uy/inet/EstadisticayProbabilidad.pdf

Pardo, A., & Ruíz, M. A. (2005). *Análisis de datos con SPSS 13 base* (1ª edición. ISBN: 8448145364 ed.). España: McGraw-Hill Interamericana de España. Obtenido de https://www.agapea.com/libros/Analisis-de-datos-con-SPSS-13-Base-9788448145361-i.htm

Rodrigo, M. F., & Molina, J. G. (2011). *Estadística Inferencial en Psicología. Tema 3. Inferencia estadística: estimación de parámetros.* Valencia, España: Universidad de Valencia. Obtenido de http://ocw.uv.es/ciencias-de-la-salud/estadistica-estadistica-inferencial-en-psicologia/tema_3.pdf

Sáez, A. J. (2012). *Apuntes de estadística para ingenieros.* España: Dpto de Estadística e Investigación Operativa. Universidad de Jaén. Obtenido de http://www4.ujaen.es/~ajsaez/recursos/EstadisticaIngenieros.pdf

Salcedo, A. (2013). *Estadística en el investigación.* Caracas: ISBN: 978-980-00-2743-1.

Serra, P. (10 de abril de 2017). *http://statistics.blogs.uv.es/.* (U. d. Valencia, Editor, & D. d. Fisoterapia, Productor) Obtenido de http://statistics.blogs.uv.es/: http://statistics.blogs.uv.es/

Sokal, R. R., & Rohlf, P. J. (1994). *Biometry.* San Francisco. United States of America: 3rd ed. Freeman & Co.

Solarte, C., Garcia, H. A., & Imuez, M. A. (2009). *Bioestadìstica. Aplicaciones en producciòn y salud animal.* Nariño, Colombia: Editorial Universitaria. Universidad de Nariño.

Steel, R., & Torrie, H. (1985). *Bioestadìstica: Principios y procedimientos.* Bogotá: McGraw-Hill.

Supo, J. (10 de abril de 2017). *Bioestadìstico.com.* Obtenido de Bioestadìstico.com.

Tukey, J. W. (1977). *Exploratory data analysis.* Massachussets: Addison-Wesley.

# Anexos

Anexo 1. Base de datos del resultado de la categorización de una variable numérica, a partir de la cual se crean dos variables categóricas.

| No. | Peso (kg) (numérica de razón) | Clasificación de la raza por su peso (categórica ordinal politómica) | Presencia de *B. canis* (categórica nominal dicotómica) |
|---|---|---|---|
| 1 | 22,5 | Grande | Enfermo |
| 2 | 12,3 | Mediana | No enfermo |
| 3 | 3,5 | Pequeña | No enfermo |
| 4 | 4,8 | Pequeña | No enfermo |
| 5 | 35,7 | Grande | Enfermo |
| 6 | 3,9 | Pequeña | No enfermo |
| 7 | 15,6 | Mediana | No enfermo |
| 8 | 7,8 | Mediana | Enfermo |
| 9 | 23,6 | Grande | No enfermo |
| 10 | 15,7 | Mediana | No enfermo |
| 11 | 26,9 | Grande | No enfermo |
| 12 | 4,2 | Pequeña | Enfermo |
| 13 | 22,6 | Grande | No enfermo |
| 14 | 25,1 | Grande | No enfermo |
| 15 | 15,3 | Mediana | No enfermo |
| 16 | 4,6 | Pequeña | Enfermo |
| 17 | 14,8 | Mediana | Enfermo |
| 18 | 27,3 | Grande | No enfermo |
| 19 | 32,1 | Grande | No enfermo |
| 20 | 14,9 | Mediana | Enfermo |
| 21 | 3,7 | Pequeña | No enfermo |
| 22 | 61,0 | Grande | No enfermo |
| 23 | 34,9 | Grande | Enfermo |

| No. | Peso (kg) (numérica de razón) | Clasificación de la raza por su peso (categórica ordinal politómica) | Presencia de *B. canis* (categórica nominal dicotómica) |
|---|---|---|---|
| 24 | 4,9 | Pequeña | No enfermo |
| 25 | 25,4 | Grande | Enfermo |
| 26 | 23,2 | Grande | Enfermo |
| 27 | 27,3 | Grande | No enfermo |
| 28 | 22,9 | Grande | Enfermo |
| 29 | 4,9 | Pequeña | No enfermo |
| 30 | 13,8 | Mediana | Enfermo |

Anexo 2. Base de datos de la distribución de la variable fija y la variable aleatoria en el estudio de casos (con *Babesia canis*) y controles (sin *Babesia canis*).

| No. | Babesia canis | Lugar de tenencia |
|---|---|---|
| 1 | Con *Babesia canis* | Casa-terraza |
| 2 | Con *Babesia canis* | Terraza-patio |
| 3 | Con *Babesia canis* | Casa-terraza |
| 4 | Con *Babesia canis* | Patio-calle |
| 5 | Con *Babesia canis* | Terraza-patio |
| 6 | Con *Babesia canis* | Patio-calle |
| 7 | Con *Babesia canis* | Casa-terraza |
| 8 | Con *Babesia canis* | Terraza-patio |
| 9 | Con *Babesia canis* | Patio-calle |
| 10 | Con *Babesia canis* | Casa-terraza |
| 11 | Con *Babesia canis* | Patio-calle |
| 12 | Con *Babesia canis* | Terraza-patio |
| 13 | Con *Babesia canis* | Patio-calle |
| 14 | Con *Babesia canis* | Terraza-patio |
| 15 | Con *Babesia canis* | Casa-terraza |

| No. | Babesia canis | Lugar de tenencia |
|---|---|---|
| 16 | Con *Babesia canis* | Patio-calle |
| 17 | Con *Babesia canis* | Terraza-patio |
| 18 | Con *Babesia canis* | Patio-calle |
| 19 | Con *Babesia canis* | Terraza-patio |
| 20 | Con *Babesia canis* | Patio-calle |
| 21 | Sin *Babesia canis* | Patio-calle |
| 22 | Sin *Babesia canis* | Casa-terraza |
| 23 | Sin *Babesia canis* | Terraza-patio |
| 24 | Sin *Babesia canis* | Casa-terraza |
| 25 | Sin *Babesia canis* | Patio-calle |
| 26 | Sin *Babesia canis* | Patio-calle |
| 27 | Sin *Babesia canis* | Casa-terraza |
| 28 | Sin *Babesia canis* | Terraza-patio |
| 29 | Sin *Babesia canis* | Terraza-patio |
| 30 | Sin *Babesia canis* | Casa-terraza |
| 31 | Sin *Babesia canis* | Terraza-patio |
| 32 | Sin *Babesia canis* | Casa-terraza |
| 33 | Sin *Babesia canis* | Terraza-patio |
| 34 | Sin *Babesia canis* | Casa-terraza |
| 35 | Sin *Babesia canis* | Patio-calle |
| 36 | Sin *Babesia canis* | Patio-calle |
| 37 | Sin *Babesia canis* | Terraza-patio |
| 38 | Sin *Babesia canis* | Casa-terraza |
| 39 | Sin *Babesia canis* | Terraza-patio |
| 40 | Sin *Babesia canis* | Casa-terraza |

Anexo 3. Base de datos de precipitaciones ocurridas en el periodo junio-noviembre de 2017 en la granja Santa Inés de la Universidad Técnica de Machala.

| Días | Meses | | | | | |
|---|---|---|---|---|---|---|
| | Junio | Julio | Agosto | Septiembre | Octubre | Noviembre |
| 1 | 0 | 0,4 | 2,1 | 0,1 | 0 | 0,7 |
| 2 | 0 | 0 | 0,3 | 0 | 3,5 | 5,6 |
| 3 | 0 | 0,2 | 0,7 | 0,3 | 2 | 0,5 |
| 4 | 0 | 0 | 0 | 0 | 0,6 | 1,2 |
| 5 | 0 | 0 | 0 | 0,9 | 2,4 | 0 |
| 6 | 0,5 | 0,5 | 0 | 0,4 | 0,8 | 1,8 |
| 7 | 3,1 | 0 | 0 | 0,3 | 0 | 2 |
| 8 | 0,9 | 0 | 1,6 | 0,2 | 0 | 3,5 |
| 9 | 1,9 | 0 | 0 | 0 | 0 | 0,5 |
| 10 | 0 | 0 | 0 | 0 | 0,1 | 0 |
| 11 | 5,8 | 0 | 0,3 | 0,5 | 0,7 | 0 |
| 12 | 10,9 | 0 | 0,4 | 0,8 | 0,4 | 2 |
| 13 | 3 | 0 | 0 | 0,9 | 0 | 0 |
| 14 | 1,2 | 0 | 0 | 0 | 0 | 0 |
| 15 | 0,3 | 0 | 0,2 | 0,5 | 0 | 0,5 |
| 16 | 0 | 0 | 0 | 0,6 | 1,5 | 0,1 |
| 17 | 3,8 | 0 | 0,1 | 0,1 | 0,1 | 1,7 |
| 18 | 6,6 | 4,7 | 0,1 | 0,5 | 0 | 0 |
| 19 | 0 | 3,1 | 0 | 0,3 | 0 | 0 |
| 20 | 0,4 | 0 | 0 | 0 | 0 | 0 |
| 21 | 1 | 0,8 | 2,9 | 1,5 | 1 | 0 |
| 22 | 0,2 | 2,5 | 0,7 | 1,3 | 0 | 0 |
| 23 | 0,3 | 0 | 0,1 | 0 | 1,2 | 1,2 |
| 24 | 0,7 | 0 | 0 | 0 | 1 | 1 |
| 25 | 0,9 | 0,7 | 1 | 0,5 | 0 | 0,1 |
| 26 | 0 | 0 | 0 | 0 | 0 | 1,3 |

| Días | | | | Meses | | |
|---|---|---|---|---|---|---|
| 27 | 0 | 0,1 | 0 | 0 | 0 | 0 |
| 28 | 4,3 | 0,8 | 0 | 0,6 | 0,5 | 0 |
| 29 | 4,1 | 0,4 | 0 | 0 | 0 | 0 |
| 30 | 0 | 0,2 | 0 | 0 | 0,4 | 0 |
| 31 | 0 | 0,2 | 0 | 0 | 0,1 | 0 |
| Días con lluvia | 19 | 13 | 13 | 18 | 16 | 16 |

Anexo 4. Tabla de números aleatorios.


*Centro de Ciencias Básicas*
*Departamento de Estadística*

**Tabla de Número Aleatorios**

| Renglón | Columna | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 34600 | 19108 | 69812 | 93480 | 65191 | 57359 | 24408 | 36527 | 60414 | 94913 |
| 2 | 79151 | 13078 | 01872 | 84469 | 83906 | 06881 | 22936 | 49856 | 97607 | 04230 |
| 3 | 92494 | 97825 | 58734 | 08516 | 37704 | 20133 | 70505 | 06395 | 54808 | 57036 |
| 4 | 44852 | 06858 | 81140 | 89296 | 54813 | 56856 | 24316 | 70468 | 90027 | 08372 |
| 5 | 97467 | 69926 | 51148 | 73026 | 43306 | 89484 | 33330 | 19093 | 80101 | 48435 |
| 6 | 96207 | 18877 | 70523 | 29690 | 44458 | 99242 | 35456 | 39595 | 87653 | 32716 |
| 7 | 60337 | 14292 | 12704 | 08359 | 36120 | 29596 | 67888 | 93498 | 74984 | 72836 |
| 8 | 04812 | 88937 | 96641 | 22579 | 73721 | 31921 | 35923 | 14615 | 40883 | 03776 |
| 9 | 30697 | 44518 | 57792 | 97046 | 99380 | 17005 | 30846 | 55406 | 22689 | 88659 |
| 10 | 60331 | 18044 | 08728 | 03094 | 03465 | 49651 | 90558 | 38744 | 11275 | 83301 |
| 11 | 18237 | 87670 | 02435 | 72480 | 99308 | 66631 | 17864 | 56993 | 98537 | 72231 |
| 12 | 98035 | 63712 | 25899 | 61025 | 35983 | 46596 | 59199 | 36711 | 03279 | 15780 |
| 13 | 67961 | 65714 | 61082 | 75324 | 85711 | 68100 | 91197 | 62429 | 68027 | 21201 |
| 14 | 70218 | 24572 | 67326 | 26462 | 87248 | 17841 | 87067 | 78185 | 42740 | 57149 |
| 15 | 83363 | 17664 | 88351 | 55077 | 07062 | 17763 | 60613 | 60318 | 05146 | 02800 |
| 16 | 68761 | 46051 | 17313 | 89765 | 00076 | 37890 | 69373 | 83061 | 32370 | 43278 |
| 17 | 67671 | 08649 | 76236 | 27897 | 17142 | 49988 | 96564 | 96447 | 51142 | 19597 |
| 18 | 95378 | 01544 | 76192 | 69697 | 29253 | 70416 | 17232 | 38553 | 21685 | 22376 |
| 19 | 84149 | 79121 | 41425 | 91820 | 04102 | 66022 | 43084 | 52345 | 42530 | 13834 |
| 20 | 95722 | 26655 | 74689 | 06488 | 39904 | 89072 | 54856 | 41955 | 54177 | 23443 |
| 21 | 19752 | 28685 | 28588 | 43556 | 66010 | 50637 | 37566 | 74944 | 20588 | 98308 |
| 22 | 45683 | 63873 | 88430 | 66485 | 06903 | 21488 | 50694 | 63228 | 23797 | 55052 |
| 23 | 99371 | 57461 | 20036 | 00612 | 19257 | 63458 | 57497 | 08098 | 74158 | 72297 |
| 24 | 72580 | 53039 | 43441 | 98578 | 54184 | 45921 | 65127 | 01318 | 68949 | 48418 |
| 25 | 34315 | 22973 | 71948 | 22061 | 65262 | 45078 | 31623 | 68896 | 05562 | 69511 |
| 26 | 68713 | 40962 | 66760 | 59066 | 51208 | 26809 | 54870 | 28032 | 73369 | 85440 |
| 27 | 44238 | 95669 | 78727 | 90871 | 08582 | 59089 | 73503 | 97694 | 68497 | 94423 |
| 28 | 95424 | 98332 | 30624 | 05323 | 17194 | 75596 | 56225 | 48613 | 19599 | 79610 |
| 29 | 49124 | 66002 | 32001 | 79866 | 31301 | 48747 | 93177 | 34517 | 05604 | 90547 |
| 30 | 74466 | 36981 | 62140 | 54336 | 98307 | 84174 | 31450 | 67320 | 24019 | 93067 |
| 31 | 62705 | 87371 | 27786 | 60655 | 04768 | 28167 | 89910 | 73654 | 39125 | 08345 |
| 32 | 79052 | 64426 | 81519 | 48547 | 52989 | 10767 | 44335 | 37239 | 39975 | 01336 |
| 33 | 77042 | 31204 | 27133 | 25775 | 26464 | 74715 | 90253 | 64489 | 09105 | 40317 |
| 34 | 92857 | 14367 | 89222 | 76064 | 42903 | 39980 | 98344 | 59800 | 29486 | 71233 |
| 35 | 38651 | 89649 | 14561 | 27064 | 20533 | 91217 | 63644 | 53458 | 40351 | 58349 |
| 36 | 09378 | 90060 | 82564 | 32916 | 71102 | 81222 | 74533 | 11621 | 10693 | 96972 |
| 37 | 58472 | 38563 | 24415 | 34840 | 58615 | 33807 | 12195 | 90969 | 73059 | 29497 |
| 38 | 42007 | 40047 | 43323 | 68349 | 91582 | 39506 | 77927 | 89534 | 87133 | 02449 |
| 39 | 53059 | 66560 | 80929 | 47058 | 64544 | 75657 | 68385 | 92748 | 31013 | 97111 |
| 40 | 82725 | 70760 | 74764 | 97880 | 46162 | 58002 | 62728 | 78882 | 77898 | 23641 |

Anexo 5. Producción de leche promedio semanal por vaca (kg) durante el periodo evaluado en la investigación.

| No. | Grupo | Producción de leche promedio semanal por vaca (kg) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Semana 1 | Semana 2 | Semana 3 | Semana 4 | Semana 5 | Semana 6 | Semana 7 | Semana 8 |
| 1 | Casos | 7,26 | 7,69 | 6,91 | 6,89 | 6,06 | 6,17 | 5,80 | 6,80 |
| 2 | Casos | 10,46 | 10,91 | 10,83 | 10,23 | 10,57 | 10,54 | 9,57 | 10,00 |
| 3 | Casos | 8,06 | 8,20 | 7,80 | 7,66 | 7,89 | 7,03 | 7,34 | 7,34 |
| 4 | Casos | 5,91 | 5,40 | 5,34 | 4,83 | 5,66 | 5,03 | 4,91 | 4,91 |
| 5 | Casos | 11,23 | 11,11 | 10,77 | 10,51 | 10,23 | 8,43 | 9,46 | 9,63 |
| 6 | Casos | 7,94 | 7,97 | 7,20 | 7,09 | 6,91 | 5,86 | 6,23 | 6,03 |
| 7 | Casos | 7,91 | 7,17 | 7,03 | 7,17 | 7,06 | 6,46 | 5,51 | 6,46 |
| 8 | Casos | 5,89 | 5,51 | 4,11 | 3,09 | 6,37 | 6,60 | 6,37 | 6,80 |
| 9 | Casos | 7,09 | 7,77 | 7,69 | 6,91 | 7,74 | 7,03 | 7,26 | 8,00 |
| 10 | Casos | 8,74 | 8,60 | 7,97 | 7,17 | 7,80 | 7,27 | 7,03 | 7,40 |
| 11 | Casos | 7,37 | 7,40 | 7,69 | 7,77 | 6,97 | 6,94 | 6,06 | 6,37 |
| 12 | Casos | 8,20 | 8,51 | 7,51 | 7,06 | 8,03 | 8,26 | 7,80 | 8,60 |
| 13 | Casos | 8,83 | 8,97 | 7,67 | 8,29 | 6,69 | 5,86 | 6,14 | 5,51 |
| 14 | Casos | 7,63 | 6,83 | 6,77 | 6,91 | 8,23 | 9,14 | 8,23 | 8,51 |
| 15 | Casos | 9,57 | 9,86 | 9,63 | 9,57 | 10,49 | 9,43 | 10,26 | 11,80 |
| 16 | Casos | 10,57 | 12,03 | 12,14 | 11,51 | 9,23 | 7,00 | 7,03 | 7,09 |
| 17 | Casos | 8,97 | 8,34 | 7,63 | 7,77 | 7,69 | 7,60 | 7,26 | 7,83 |

| No. | Grupo | Producción de leche promedio semanal por vaca (kg) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 18 | Casos | 7,20 | 7,07 | 6,71 | 5,64 | 8,06 | 9,46 | 9,74 | 9,91 |
| 19 | Control | 6,96 | 6,69 | 6,44 | 5,49 | 7,66 | 7,17 | 7,23 | 7,46 |
| 20 | Control | 9,31 | 8,40 | 7,97 | 8,23 | 10,14 | 11,03 | 11,37 | 10,66 |
| 21 | Control | 5,66 | 5,54 | 4,63 | 5,29 | 7,03 | 6,83 | 6,91 | 6,77 |
| 22 | Control | 8,11 | 6,49 | 5,80 | 6,17 | 5,97 | 6,03 | 5,89 | 6,17 |
| 23 | Control | 7,26 | 8,09 | 8,23 | 7,31 | 6,94 | 7,74 | 7,31 | 7,17 |
| 24 | Control | 9,83 | 10,46 | 10,34 | 10,51 | 8,37 | 8,00 | 7,94 | 8,49 |
| 25 | Control | 4,49 | 4,77 | 4,37 | 4,59 | 6,86 | 6,83 | 6,40 | 6,49 |
| 26 | Control | 8,31 | 7,66 | 6,77 | 7,74 | 6,56 | 6,46 | 7,23 | 8,00 |
| 27 | Control | 12,86 | 12,91 | 12,46 | 12,86 | 6,54 | 6,71 | 6,80 | 7,09 |
| 28 | Control | 7,46 | 6,77 | 7,14 | 7,63 | 6,14 | 5,97 | 7,11 | 7,09 |
| 29 | Control | 7,91 | 7,17 | 6,80 | 7,17 | 8,69 | 8,74 | 8,46 | 8,26 |
| 30 | Control | 8,26 | 8,54 | 8,31 | 8,80 | 10,43 | 10,63 | 9,17 | 9,37 |
| 31 | Control | 9,31 | 10,09 | 9,43 | 9,89 | 8,14 | 8,26 | 9,09 | 8,86 |
| 32 | Control | 7,66 | 8,11 | 7,49 | 7,51 | 7,26 | 7,71 | 7,66 | 7,17 |
| 33 | Control | 8,71 | 8,43 | 7,60 | 7,40 | 4,80 | 5,86 | 5,31 | 5,46 |
| 34 | Control | 6,66 | 7,91 | 8,37 | 8,06 | 4,66 | 5,54 | 4,66 | 5,37 |
| 35 | Control | 8,17 | 6,71 | 7,29 | 7,06 | 7,23 | 6,11 | 6,71 | 7,14 |
| 36 | Control | 9,21 | 8,70 | 8,43 | 9,46 | 9,26 | 9,54 | 9,17 | 9,23 |

Anexo 6. Porcentaje de supervivencia de camarones obtenidos en cada laboratorio de procedencia de las larvas utilizadas para la siembra.

| Laboratorio de procedencia | Supervivencia de camarones (%) |
|---|---|
| Doble T | 65,0 |
| Doble T | 68,0 |
| Doble T | 70,0 |
| Doble T | 70,0 |
| Reydamar | 52,0 |
| Reydamar | 50,0 |
| Reydamar | 50,0 |
| Reydamar | 44,0 |
| Reydamar | 43,0 |
| Aqualab | 55,0 |
| Aqualab | 60,0 |
| Aqualab | 53,0 |
| Aqualab | 58,0 |
| Aqualab | 51,0 |
| Nutriagro | 64,0 |
| Nutriagro | 57,0 |
| Nutriagro | 59,0 |
| Nutriagro | 68,0 |
| Nutriagro | 66,0 |
| Génesis | 67,0 |
| Génesis | 68,0 |
| Génesis | 70,0 |
| Génesis | 68,0 |

Anexo 7. Densidad aparente del suelo (de 0-15 cm de profundidad) en sistemas productivos de la granja Santa Inés, pertenecientes a la Universidad Técnica de Machala.

| No. | Sistema productivo | Densidad aparente del suelo (g $cm^3$) |
|---|---|---|
| 1 | Pasto | 1,90 |
| 2 | Pasto | 1,68 |
| 3 | Pasto | 1,76 |
| 4 | Cacao | 1,65 |
| 5 | Cacao | 1,69 |
| 6 | Cacao | 1,69 |
| 7 | Banano | 1,80 |
| 8 | Banano | 1,74 |
| 9 | Banano | 1,74 |
| 10 | Maíz | 1,75 |
| 11 | Maíz | 1,63 |
| 12 | Maíz | 1,66 |
| 13 | Bosque | 1,75 |
| 14 | Bosque | 1,63 |
| 15 | Bosque | 1,66 |

Anexo 8. pH del suelo (entre 0-15 cm de profundidad) en sistemas productivos de la granja Santa Inés, pertenecientes a la Universidad Técnica de Machala.

| No. | Sistema productivo | pH del suelo (en $H_2O$) |
|---|---|---|
| 1 | Pasto | 1,90 |
| 2 | Pasto | 1,68 |
| 3 | Pasto | 1,76 |
| 4 | Cacao | 1,65 |
| 5 | Cacao | 1,69 |
| 6 | Cacao | 1,69 |
| 7 | Banano | 1,80 |
| 8 | Banano | 1,74 |
| 9 | Banano | 1,74 |
| 10 | Maíz | 1,75 |
| 11 | Maíz | 1,63 |
| 12 | Maíz | 1,66 |
| 13 | Bosque | 1,75 |
| 14 | Bosque | 1,63 |
| 15 | Bosque | 1,66 |

Anexo 9. Prueba de Games-Howell que muestra las comparaciones múltiples realizadas para la variable pH del suelo en función de los sistemas productivos estudiados.

| | (I) Sistema Productivo | (J) Sistema Productivo | Diferencia de medias (I-J) | Error estándar | Sig. | Intervalo de confianza al 95% | |
|---|---|---|---|---|---|---|---|
| | | | | | | Límite inferior | Límite superior |
| Games-Howell | Pasto | Cacao | ,91867* | ,12583 | ,041 | ,0824 | 1,7549 |
| | | Banano | -,08200 | ,19366 | ,989 | -1,4784 | 1,3144 |
| | | Maíz | ,39600 | ,29860 | ,711 | -1,8383 | 2,6303 |
| | | Bosque | ,74433 | ,36149 | ,457 | -1,9844 | 3,4731 |
| | Cacao | Pasto | -,91867* | ,12583 | ,041 | -1,7549 | -,0824 |
| | | Banano | -1,00067 | ,22615 | ,067 | -2,1055 | ,1042 |
| | | Maíz | -,52267 | ,32063 | ,575 | -2,3940 | 1,3486 |
| | | Bosque | -,17433 | ,37989 | ,986 | -2,5495 | 2,2008 |
| | Banano | Pasto | ,08200 | ,19366 | ,989 | -1,3144 | 1,4784 |
| | | Cacao | 1,00067 | ,22615 | ,067 | -,1042 | 2,1055 |
| | | Maíz | ,47800 | ,35281 | ,684 | -1,2389 | 2,1949 |
| | | Bosque | ,82633 | ,40741 | ,418 | -1,3136 | 2,9663 |
| | Maíz | Pasto | -,39600 | ,29860 | ,711 | -2,6303 | 1,8383 |
| | | Cacao | ,52267 | ,32063 | ,575 | -1,3486 | 2,3940 |
| | | Banano | -,47800 | ,35281 | ,684 | -2,1949 | 1,2389 |
| | | Bosque | ,34833 | ,46652 | ,934 | -1,7653 | 2,4620 |
| | Bosque | Pasto | -,74433 | ,36149 | ,457 | -3,4731 | 1,9844 |
| | | Cacao | ,17433 | ,37989 | ,986 | -2,2008 | 2,5495 |
| | | Banano | -,82633 | ,40741 | ,418 | -2,9663 | 1,3136 |
| | | Maíz | -,34833 | ,46652 | ,934 | -2,4620 | 1,7653 |

# 06 Capítulo Estadística predictiva con datos agropecuarios

Bill Serrano; Irán Rodríguez Delgado

Las características de la época actual hacen necesario poseer conocimientos sólidos en el uso de las herramientas estadísticas avanzadas para poder inferir resultados sobre una población desde el análisis de una o varias muestras -con datos históricos y actuales -, que mediante el uso de la estadística inferencial se modelizan procesos, sistemas, respuestas con fines predictivos permitiendo extraer patrones de comportamiento para identificar riesgos, oportunidades y con su ayuda cosechar resultados esperados. En el sector agropecuario estas técnicas se hacen pertinentes al existir la necesidad de mejorar la productividad, donde la estadística inferencial sirve como eje transversal para tomar las decisiones acertadas para dicho propósito.

**Bill Serrano:** Ingeniero Agrónomo e Ingeniero en Gestión Empresarial, Magister en Administración de Empresas y estudiante doctoral en Análisis Económico y Estrategia Empresarial en la Universidad A Coruña. Fue Gerente de Almacén y Jefe Comercial Zonal en ICESA, Gerente de producto en ICESA y COMPTECO. Actualmente Profesor Titular en la Universidad Técnica de Machala.

**Irán Rodríguez Delgado:** Ingeniero Agrónomo (1992) Universidad Central de Las Villas, Cuba Magister en Agricultura Sostenible (2009) Universidad de Cienfuegos, Cuba; Investigador Agregado (2009) Instituto de Investigaciones de la Caña de Azúcar, Cuba; Profesor Titular (2015) Universidad Técnica de Machala. Autor de cuatro libros y 17 artículos publicados.

En el presente capítulo se hace exposición relacionada con las correlaciones, el análisis de regresión y la causalidad. Además, se hace una presentación específica relacionada con el análisis de regresión simple, análisis de regresión y el análisis de varianza, análisis de regresión múltiple, análisis de regresión con variables dicotómicas, finalizando con análisis de regresión y ANCOVA. Todas las técnicas descritas se aplicarán a datos agropecuarios agregando varios ejemplos ilustrativos, gráficos y los cálculos correspondientes se obtendrán usando los software estadísticos como stata y SPSS.

## Análisis de regresión, correlación y causalidad

### Interpretación de la Regresión

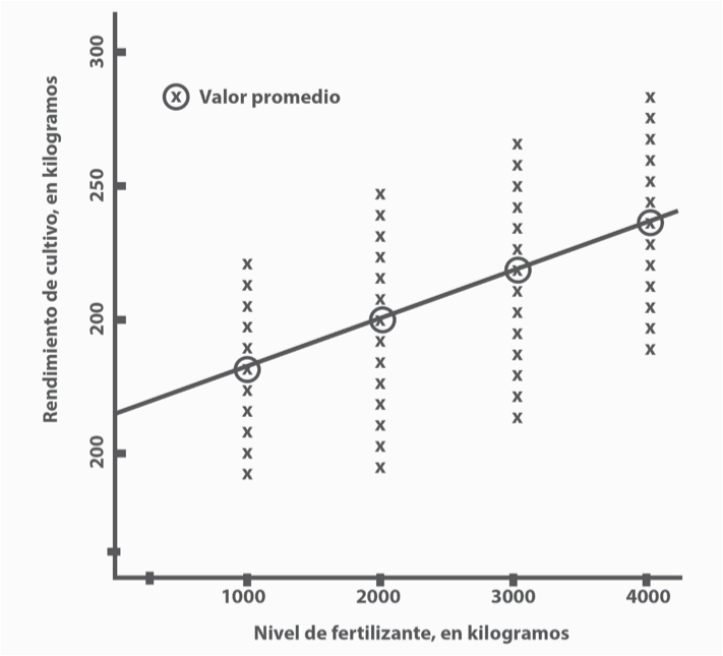
El análisis de regresión abarca el tratamiento de la dependencia de una variable (variable dependiente, variable explicada, predicha, regresada, variable de respuesta, endógena, resultado) respecto de una o varias variables (variables independientes, variables explicativas, predictora, regresora, estímulo, exógena, covariante, variable de control), con el propósito de estimar la media poblacional de la primera en términos de los valores conocidos de las segundas.

Cuando en el tratamiento de la dependencia de una variable se observa una sola variable independiente se denomina análisis de regresión simple y cuando existe más de una variable independiente toma el nombre de análisis de regresión múltiple.

Considere el siguiente ejemplo: A un agrónomo le interesa averiguar las razones de la estabilidad en la distribución del rendimiento de un cultivo dentro de una población. La regresión intenta averiguar cómo cambia el rendimiento promedio del cultivo dado la cantidad de fertilizante aplicado. Es decir, lo que intenta predecir es el rendimiento del cultivo a partir de la cantidad de fertilizante aplicado. Considere el Gráfico 6.1, correspondiente a un diagrama de dispersión. El gráfico muestra la distribución de los rendimientos del cultivo en una población hipotética, correspondientes

al conjunto de valores conocidos del fertilizante aplicado. Preste atención que, para cualquier cantidad de fertilizante aplicado, existe un rango (distribución) de rendimientos. Sin embargo, observe también que, a pesar de la variabilidad del rendimiento del cultivo conforme al valor del fertilizante aplicado, el rendimiento promedio del cultivo aumenta, por lo general, en la medida en que lo hace el fertilizante aplicado. Los cruces dentro de los círculos en el gráfico establecen que el rendimiento promedio del cultivo corresponde a una cantidad determinada de fertilizante aplicado. Estos promedios se conectan para obtener la línea recta del gráfico. Esta línea, se conoce como recta de regresión. Dicha recta muestra que el rendimiento del cultivo aumenta conforme crece la cantidad de fertilizante aplicado.

Gráfico 6.1. Diagrama de dispersión




Fuente: Modificado de Gujarati (2006)

En el análisis de regresión lo pertinente es lo que se denomina dependencia estadística entre variables. Esta es propia de variables aleatorias o estocásticas, es decir, variables con distribución de probabilidad. Por ejemplo, siguiendo el ejemplo del rendimiento del cultivo, éste depende del fertilizante aplicado, sin embargo, también lo hace de la lluvia, temperatura, sol, entre otros, y tal dependencia es de naturaleza aleatoria porque las variables explicativas, a pesar de ser importantes, no permiten al agrónomo predecir de forma exacta el rendimiento del cultivo, debido a los errores propios de la medición de estas variables. Por tal motivo, existirá alguna variabilidad intrínseca en la variable explicada.

## Regresión y causalidad

La causalidad en forma simple se dice que es el principio o el origen de algo. Este concepto es traído a la práctica para explicar la relación entre una causa y su efecto. En la estadística, este término explica la relación de necesidad de coocurrencia de dos variables.

A pesar de que la regresión establece la relación estadística que pueda existir entre la dependencia de una variable respecto a otras, y por más fuerte que esta sea, no implica que exista causalidad necesariamente. Para determinar la causalidad, es necesario acudir a consideraciones teóricas o a priori. Usando el ejemplo citado del rendimiento del cultivo, no existe motivo estadístico para suponer que el fertilizante depende del rendimiento del cultivo, sin embargo, la lógica indica que la relación es a la inversa, ya que no es posible controlar la cantidad de fertilizante aplicado mediante el rendimiento del cultivo.

## Regresión y correlación

El análisis de regresión y la correlación se vincula de manera estrecha. Por un lado, el propósito principal del análisis de correlación es determinar el grado de asociación lineal entre dos variables, por ejemplo, si se desea conocer la correlación entre la lluvia y el rendimiento de un cultivo; entre la cantidad

de fertilizante aplicado y el rendimiento de un cultivo; entre el balanceado utilizado y el crecimiento del camarón en una piscina; entre la salinidad del agua y la producción de tilapia. En cambio, en el análisis de regresión trata de estimar el valor promedio de una variable con base en los valores conocidos de otras. De tal manera, se desee estimar el promedio del rendimiento de una hectárea de camarón desde la cantidad de balanceado utilizado. La correlación al determinar el grado de asocial lineal entre las variables, me permite poner a consideración incluir o no las variables en el modelo de la regresión.

Hay que considerar que en un análisis de regresión existe una asimetría en el trato a las variables dependientes e independientes. También se supone que la variable dependiente tiene una distribución de probabilidad, es decir es estocástica, y se asume que las variables independientes tienen valores fijos o conocidos en muestreos repetitivos – las variables independientes pueden intrínsecamente ser estocásticas, pero para fines del análisis de regresión, se considera como supuesto que sus valores son fijos en el muestreo repetitivo, es decir que la variable explicativa toma los mismos valores en diferentes muestras. De esta forma, en el Gráfico 6.1 se supuso que la cantidad de fertilizante aplicado era fijo en los niveles dados y se obtuvieron rendimientos de los cultivos en esos niveles. En cambio, en el análisis de correlación, entre las variables que intervienen no existe distinción, de tal manera que, la correlación existente entre la cantidad de fertilizante aplicado y el rendimiento de cultivo, es la misma entre el rendimiento del cultivo y la cantidad de fertilizante aplicado.

Para interpretar la correlación, hay que conocer que toma valores en el intervalo de [-1,1], indicando el signo el sentido de la relación entre las variables. Por ejemplo: si el índice de correlación es de 1, se interpreta como una correlación positiva perfecta, es decir, existe una asociación lineal directa perfecta entre las variables: cuando una de ellas aumenta, la otra también lo hace en proporción constante. Es así que:

- Correlación =1, existe una asociación lineal positiva perfecta.
- Correlación mayor a 0 y menor a 1, existe una asociación lineal positiva.

- Correlación =0, no existe asociación lineal entre las variables.

- Correlación menor a 0 y mayor a -1, existe asociación lineal negativa.

- Correlación =-1, existe una asociación lineal negativa perfecta.

Para ilustrar el análisis de correlación, considere los datos del Cuadro 6.1, que detalla el rendimiento de un cultivo y el fertilizante aplicado. El coeficiente de correlación resultante es de 1, es decir, que existe una asociación lineal positiva perfecta. Cuando el fertilizante aplicado por hectárea aumenta, el rendimiento por hectárea también lo hace, sin embargo, que exista correlación entre las variables no quiere decir necesariamente que exista causalidad. La causalidad solo puede aceptarse cuando hay suficientes razones claras. La correlación no implica causalidad.

Cuadro 6.1. Rendimiento de un cultivo y el fertilizante aplicado

| Obs | Rendimiento/Kg x Ha | Fertilizante /kg x Ha |
|---|---|---|
| 1 | 200 | 100 |
| 2 | 250 | 125 |
| 3 | 300 | 150 |
| 4 | 350 | 175 |
| 5 | 400 | 200 |
| 6 | 450 | 225 |
| 7 | 500 | 250 |
| 8 | 550 | 275 |
| 9 | 600 | 300 |
| 10 | 650 | 325 |
| 11 | 700 | 350 |
| 12 | 750 | 375 |
| 13 | 800 | 400 |
| 14 | 850 | 425 |

## Análisis de regresión simple

El análisis de regresión simple también conocido como análisis de regresión con dos variables, tiene como finalidad establecer el valor promedio de la variable dependiente, con base en los valores conocidos de una sola variable independiente. Para entender esto -mediante un ejemplo hipotético- consideremos los datos de la Cuadro 6.2. Estos datos se refieren a 80 piscinas de camarón, donde cada una tiene un tamaño de 1 hectárea, así como el nivel balanceado aplicado (X), en libras y la producción obtenida por corrida (110 días), en libras. Las 80 piscinas se dividen en 8 grupos por la cantidad de balanceado (de 2000 libras a 3200); de igual manera, se reflejan las producciones por corrida de cada piscina de los distintos grupos. Por efecto, existen 8 valores fijos de X y los correspondientes valores de Y (producción por corrida) para cada valor de X.

Cuadro 6.2. Rendimiento de un cultivo y el fertilizante aplicado

| X | 2000 | 2200 | 2400 | 2600 | 2800 | 3000 | 3200 | 3400 |
|---|---|---|---|---|---|---|---|---|
| Y | | | | | | | | |
| Producción por piscina por corrida Y, libras | 1800 | 2000 | 2100 | 2300 | 2350 | 2420 | 2590 | 2790 |
| | 1850 | 2100 | 2150 | 2200 | 2360 | 2460 | 2610 | 2840 |
| | 2000 | 2050 | 2200 | 2000 | 2450 | 2500 | 2650 | 2880 |
| | 2050 | 2130 | 2000 | 2100 | 2500 | 2580 | 2800 | 3100 |
| | 1800 | 2120 | 1960 | 2050 | 2300 | 2420 | 2570 | 2800 |
| | 1900 | 2200 | 2200 | 2200 | 2330 | 2400 | 2550 | 2780 |
| | 2020 | 1980 | 2090 | 2340 | 2220 | 2340 | 2400 | 2640 |
| | 1890 | 1800 | 2100 | 2200 | 2280 | 2400 | 2550 | 2760 |
| | 1940 | 2100 | 2080 | 2150 | 2300 | 2400 | 2530 | 2760 |
| | 2000 | 2010 | 2120 | 2300 | 2230 | 2320 | 2470 | 2690 |
| Total | 19250 | 20490 | 21000 | 21840 | 23320 | 24240 | 25720 | 28040 |
| Media condicional de Y | 1925 | 2049 | 2100 | 2184 | 2332 | 2424 | 2572 | 2804 |

Se observa un cambio considerable entre la producción por corrida de cada grupo de balanceado. Es decir, que a pesar de la variabilidad de la producción en cada rango de balanceado considerado, en promedio, la producción se incrementa a medida que aumenta las libras de balanceado aplicado. En el Cuadro 6.2 se muestra la media de la producción por corrida que corresponde a cada uno de los 8 niveles de balanceado aplicado. De tal forma, el nivel de balanceado de 2000 libras por hectárea y por corrida le corresponde una media de producción de 1925 libras de camarón, sin embargo, al nivel de 3000, la media correspondiente es de 2424. Existen en total 8 medias, las que son denominadas como valores esperados condicionales, en consideración que dependen de los valores de la variable X (variable condicional).

Se hace necesario diferenciar el valor esperado incondicional y los valores esperados condicionales de la producción por corrida/ha. Si sumamos las 80 producciones que forman la población y la dividimos para 80, obtendremos la cantidad de 2298.75 (183900/80), que es el valor de la media incondicional, del nivel de balanceado aplicado; como no consideramos los niveles de balanceado aplicado es considerada incondicional. Como es de suponer, los distintos valores esperados condicionales de Y del Cuadro 6.2 difieren respecto del valor esperado incondicional de Y (2298.75). Siguiendo con ejemplo, cuando se desea conocer el valor esperado de la producción de una piscina por corrida, la respuesta es de 2298.75 (la media incondicional), sin embargo, si se desea saber cuál es el valor esperado de la producción de una piscina por corrida cuando se aplica 2400 libras de balanceado, la respuesta es 2100 libras (la media condicional). Por tal razón, conocer el nivel de balanceado permite predecir con mayor exactitud el valor medio de la producción. Esta es la esencia de la regresión.

Del anterior ejemplo expuesto, es evidente que cada media condicional está en función de X (lineal). Esta función explica como la media de Y varía con respecto a X. La función adopta una relación lineal, ya que un Biólogo o un

productor se puede plantear que la producción de camarón tiene una relación lineal con el balanceado aplicado. Por tanto, inicialmente se puede aproximar que, Y es una función lineal de X, del tipo: $Y = \beta 0 + \beta 1X$, donde $\beta 0$ y $\beta 1$ son parámetros no conocidos pero fijos que se denominan coeficientes de regresión, también se conocen como coeficientes de intersección y de pendiente. Este conjunto de parámetros y variables se conoce como la ecuación de regresión, pero también se la identifica como regresión. En el análisis de regresión el objetivo es estimar los parámetros $\beta 0$ y $\beta 1$ con base en las observaciones de X y Y.

## Método de los mínimos cuadrados

El método de los mínimos cuadrados se ha convertido en uno de los métodos más utilizados para el análisis de regresión por su eficacia y por la calidad de sus propiedades estadísticas, por tal motivo para realizar los análisis de regresión a forma da ejemplificar se utilizará el método de los mínimos cuadrados. En este sentido, los estimadores obtenidos se conocen como estimadores de mínimos cuadrados (MC), los que cumplen las siguientes propiedades:

- Se enuncian en términos de las cantidades observables, es decir de los X y Y que se encuentran en la muestra.
- Cada estimador proporciona un solo valor puntual del parámetro poblacional pertinente.
- Una vez generados los estimadores de MC, se obtiene la línea de la regresión muestral. La regresión obtenida tiene la propiedad de que el valor de la Y estimada (muestra) representa a la Y real (población).

### Ejemplo Ilustrativo:

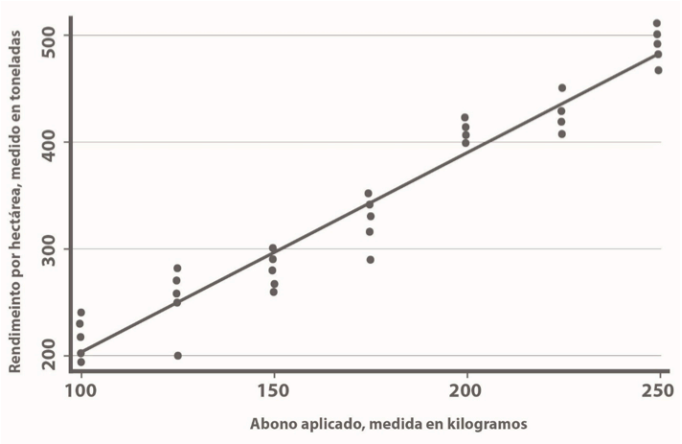
El Cuadro 6.3 proporciona datos sobre el rendimiento de un cultivo por hectárea (medido en tonelada) y el abono aplicado (medido en kilogramos). Mediante el software stata encontramos la ecuación de la regresión:

Y = 17.56429 + 1.856857X (es decir, β0=17.56429 y β1 = 1.856857).

$r^2$ = 0.9460

La línea de la regresión se muestra en el Gráfico 6.2

Gráfico 6.2. Distribución hipotética de rendimiento por hectárea correspondiente a la cantidad de abono aplicado




El valor de β1 =1.856857, que mide la pendiente, indica que, entre el intervalo muestral de X entre 100 y 250 kg de abono por hectárea, a medida que el valor X aumenta 1 kg, el incremento estimado en el rendimiento promedio por hectárea es 1.856857 toneladas. Es decir, que, por cada kg de abono adicional, en promedio, produce aumento en el rendimiento por hectárea de 1.856857 toneladas.

El valor de β0=17.56429, que pertenece al intercepto, enseña el nivel promedio del rendimiento del cultivo cuando la cantidad de abono aplicado es cero.

El valor de $r^2$ de 0.946 se interpreta que el nivel de abono aplicado explica el 94.6% de la variación del rendimiento por hectárea.

Antes de finalizar este ejemplo, considere que el modelo ejemplificado es muy sencillo. La teoría y la experiencia indican que, el rendimiento de un cultivo, a parte del abono,

existen otros factores importantes en la determinación del rendimiento. Al ingresar más variables X al modelo, esta toma el nombre de regresión múltiple.

Cuadro 6.3. Rendimiento de un cultivo y el fertilizante aplicado

| Abono /kg x Ha | Rendimiento/tn x Ha | | | | |
|---|---|---|---|---|---|
| 100 | 200 | 230 | 195 | 240 | 217 |
| 125 | 250 | 270 | 200 | 280 | 257 |
| 150 | 300 | 280 | 260 | 290 | 267 |
| 175 | 350 | 330 | 290 | 340 | 317 |
| 200 | 400 | 412 | 408 | 422 | 399 |
| 225 | 450 | 420 | 430 | 430 | 407 |
| 250 | 500 | 480 | 510 | 490 | 467 |

## Análisis de regresión y análisis de varianza

En el análisis de la regresión al hacer el análisis de varianza (ANOVA) permite descomponer la variabilidad de la variable dependiente en variabilidad explicada por el modelo más la variabilidad no explicada, esto permitirá contrastar si el modelo es significativo o no. Es decir SCT=SCE + SCR, que fragmenta la suma de los cuadrados totales (SCT) en dos componentes: la suma de cuadrados explicada (SCE) y la suma de cuadrados de residuos (SCR). El estudio de estos componentes de SCT se conoce como análisis de varianza desde el punto de vista de la regresión.

Adicional debe considerar que para crear un modelo de regresión lineal es necesario que cumpla con los siguientes supuestos:

Supuesto 1. Linealidad: El modelo de regresión debe ser lineal en los parámetros, no es necesario que lo sea en las variables. Es decir, el modelo de regresión debe ser de la forma $Y = \beta 0 + \beta 1X + u$ (regresión lineal simple); $Y = \beta 0 + \beta 1X + \beta nXn + u$ (regresión lineal múltiple).

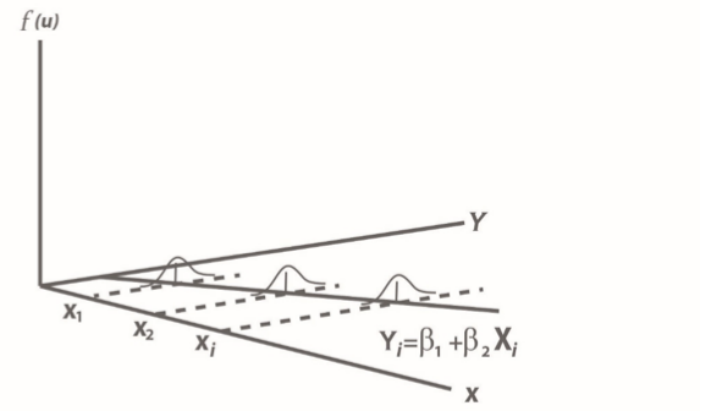
Supuesto 2. Valores fijos de X, o valores de X independientes del término error: Los valores que toma la variable independiente X pueden ser considerados como fijos en el caso de la variable independiente fija, o haber sido muestreada con la variable dependiente Y en el caso de la regresora estocástica. Así mismo que la(s) variable(s) X y el término error son independientes.

Las razones para suponer que los valores de la variable independiente son no estocásticos en ciertas condiciones, es porque en situaciones experimentales asociadas al sector agropecuario se necesite fijar los valores de la variable X. Por ejemplo, un Biólogo puede aplicar distintas cantidades de balanceado en distintas piscinas de camarón para ver el efecto en la producción. De tal manera, puede fijar una cantidad específica de balanceado aplicado en distintas piscinas con el propósito de obtener la producción promedio.

Supuesto 3. El valor de la perturbación (u) tenga una esperanza matemática igual a 0: este supuesto establece que el valor de la media de la perturbación, que depende las variables independientes dadas, es cero. Lo que mantiene el supuesto es que los factores que no se incluyen explícitamente en un modelo dado, y, por consiguiente, pertenecen a u, no afectan el valor de la media de Y. De tal manera, el supuesto 3 manifiesta que X y u no están correlacionadas (ambas ejercen influencia independientes y aditivas en Y). Si fuera lo contrario, no se podría analizar el efecto de cada una (X y u) sobre la variable dependiente, ya que, X aumentaría cuando u aumente y viceversa (correlación positiva) o X se incrementaría cuando u disminuye, y se reduciría cuando u aumenta (correlación negativa). Por tal razón, de forma general el supuesto 3 expresa que no existe error de especificación en el modelo de regresión elegido.

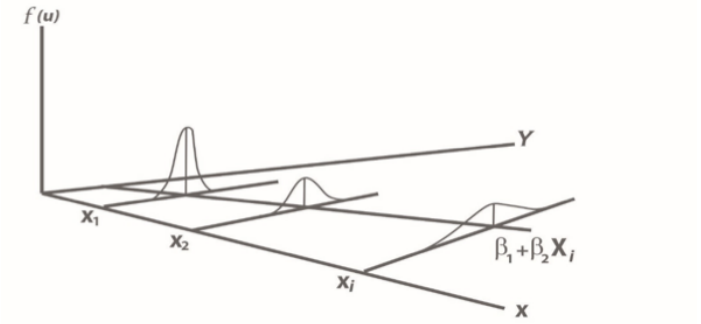
Supuesto 4. Homocedasticidad: La varianza de las perturbaciones es la misma sin importar el valor de X. Es decir, la variación alrededor de la línea de la regresión formada por elpromedio entre X y Y, es la misma para todos los valores de X (en el Gráfico 6.3 se aprecia la situación), lo contrario se conoce como heteroscedasticidad (ver Gráfico 6.4).

Gráfico 6.3. Homocedasticidad.




Fuente: Gujarati (2006)

Gráfico 6.4.Heterocedasticidad

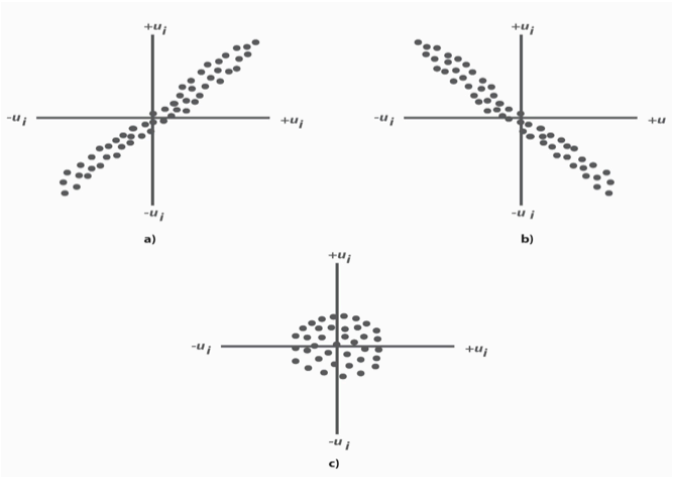



Fuente: Gujarati (2006)

Para entenderlo claramente, supongamos que Y es la producción por hectárea de un cultivo y X el fertilizante aplicado. Los gráficos 6.3 y 6.4 reflejan que aumenta la producción a medida que la cantidad de fertilizante aplicado también aumenta. Sin embargo, en el gráfico 6.3 la varianza de la producción permanece igual para todos los niveles de fertilizante aplicado, mientras que en el 6.4 aumenta la variabilidad en la producción a medida que va aumentando la dosis del fertilizante.

Supuesto 5. Incorrelación: La correlación entre las perturbaciones es igual a cero, es decir, dado X, las desviaciones de dos valores cualesquiera de Y en relación a su valor promedio muestran patrones como en el gráfico 6.5, donde se nota que las u están correlacionadas, pues a una u positiva la sigue una u positiva, y lo que requiere el supuesto 4 es que las correlaciones estén ausentes.

Gráfico 6.5. Patrones de correlación entre las perturbaciones a) correlación serial positiva; b) correlación serial negativa; c) correlación cero.




Supuesto 6. No colinealidad (para regresiones múltiples): Es decir la inexistencia de colinealidad perfecta – si una de las variables independientes tiene una relación lineal con otras variables independientes – y colinealidad parcial – si entre las variables independientes existen altas correlaciones -. Al existir colinealidad los predictores se encontrarían en combinación lineal, y la influencia de cada uno de ellos en la variable dependiente no puede distinguirse al quedar solapados unos con otros.

Supuesto 7. La naturaleza de la variables X: Los valores de la variable X no deben ser atípicos, es decir, valores muy dispersos en relación con el resto de las observaciones, con el fin de que los resultados de las regresiones estén subyugadas por tales valores atípicos.  Así mismo, todos los valores de X

en una muestra específica no deben ser iguales. Si los valores de X son idénticos se imposibilita la estimación de los Beta, la variación que existe tanto en los valores de X como de Y es necesario para poder utilizar la regresión como herramienta.

Informe de Resultados del análisis de regresión

Existen distintas formas de dar a conocer los resultados de un análisis de regresión; sin embargo, en este texto utilizaremos el que creemos es el más conveniente por su facilidad de análisis, como ilustración utilizamos el ejemplo del Cuadro 6.2 producción por piscina por corrida en libras y balanceado aplicado en libras:

Y = 709.6071 + 0.5885714X1

ee = (72.12439) (0.0263361)

t = (9.84) (22.35)

p = (0.001)(0.001)

F = 499.45

p = 0.001

r2 = 0.8649

Los primeros valores representan a la ecuación de la regresión; el primer conjunto de paréntesis son los errores estándar de los beta de la regresión; los valores del segundo conjunto son los estadísticos calculados (valores de t estimados), es decir, t (22.35) = 0.5885714 / 0.0263361; y los valores del tercer grupo son valores de p estimados. De tal manera, la probabilidad de obtener un valor de t igual o mayor de 22.35 es de 0.001. o prácticamente 0. En otras palabras, se rechaza la hipótesis nula que establece que el verdadero valor poblacional de cada coeficiente de regresión individual es cero.

Mientras menor sea el valor de p, más significativo es el modelo, – se considera significativo cuando es el valor de p es menor o igual a 0.05, sin embargo, es preferible dejar que el investigador decida si debe rechazar la hipótesis nula con el valor de p dado – es decir, que cuanto menor sea el valor de p, menor será la probabilidad de cometer un error

si se rechaza la hipótesis nula. Por ejemplo, si un análisis del rendimiento de un cultivo, el valor de p de un estadístico de prueba resulta ser 0.13 o 13% y el investigador establecer rechazar la hipótesis en este nivel de significancia, que así sea, está asumiendo el riesgo de equivocarse el 13% de las veces si se rechaza la hipótesis nula verdadera.

Siguiendo con el ejemplo de la regresión, se presenta el valor de significancia exacto de cada valor de t estimado. Así, en relación a la hipótesis nula de que el verdadero valor poblacional de cada coeficiente individual es cero (que la cantidad de balanceado aplicado en las piscinas de camarón no produce ningún efecto en el nivel de producción), la probabilidad de que sea cierto es prácticamente cero. Por lo tanto, se rechaza la hipótesis nula y se acepta la alternativa de que el verdadero valor de la pendiente poblacional es diferente de cero (que la cantidad de balanceado aplicado en las piscinas de camarón si produce efecto en el nivel de producción). Como el signo del $\beta 1$ es positivo el efecto de la variable independiente sobre la dependiente es directa, para los casos donde el signo es negativo, se interpreta como una relación inversa.

## Análisis de regresión múltiple

El análisis de regresión simple, estudiando anteriormente, suele ser no tan útil en la práctica. Es el caso del cuadro 6.2, donde se estableció que solo el nivel de balanceado X se relaciona con la producción del camarón por hectárea Y. Definitivamente rara vez es tan simple, pues, algunas otras variables afectan la producción. Para citar otro ejemplo, es altamente probable que la producción de una bananera no solo dependa de la fertilización aplicada sino también de otros insumos y recursos, tales como horas sol, agua, etc. Por tal motivo, es necesario ampliar el modelo de regresión simple y considerar modelos con más de dos variables. Al momento de adherir más variables toma el nombre de análisis de regresión múltiple, donde interviene una variable dependiente y al menos dos variables independientes.

La ecuación de la regresión - para el caso de dos variables independientes – se la representa de la siguiente forma:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + u$$

El β0 es el término del intercepto, este se conoce como el valor promedio sobre Y de todas las variables no incluidas en el modelo, aunque para mejor interpretación se conocen como el valor promedio de Y cuando las demás variables independientes se igualan a cero. Los coeficientes β1 y β2 son conocidos como coeficientes de regresión parcial o pendientes. El coeficiente β1 estima el cambio en el valor de la media de Y por unidad de cambio en X1, manteniendo X2 constante. Es decir, es el efecto que tiene X1 por cada cambio unitario sobre el valor medio de Y, neto del efecto que X2 ejerza en la media Y. De la misma manera β2 estima el cambio en el valor de la media de Y por unidad de cambio en X2, manteniendo X1 constante. Es decir, es el efecto que tiene una unidad de cambio en X2 sobre el valor medio de Y, neto del efecto que X1 ejerza en la media Y.

El modelo de regresión lineal múltiple debe ser lineal en los parámetros, no es necesario que lo sea en las variables. Es decir, el modelo de regresión debe ser de la forma $Y = \beta 0 + \beta 1X + \beta nXn + u$.

Como ilustración utilizamos el ejemplo del Cuadro6.4 que muestra el rendimiento de un cultivo en kg/ha como variable dependiente y al abono aplicado y los mm de agua recibidos como variables independientes:

Cuadro 6.4. Rendimiento por hectárea considerando el abono aplicado y los milímetros de agua recibidos.

| Rendimiento/kg x Ha | Riego (mm) | Abono / kg x Ha | Rendimiento/kg x Ha | Riego (mm) | Abono / kg x Ha |
|---|---|---|---|---|---|
| 200 | 1500 | 100 | 195 | 1640 | 100 |
| 250 | 1400 | 125 | 200 | 1540 | 125 |
| 300 | 1500 | 150 | 260 | 1640 | 150 |
| 350 | 2300 | 175 | 290 | 2440 | 175 |

| Rendimiento/kg x Ha | Riego (mm) | Abono / kg x Ha | Rendimiento/kg x Ha | Riego (mm) | Abono / kg x Ha |
|---|---|---|---|---|---|
| 400 | 2300 | 200 | 408 | 2440 | 200 |
| 450 | 2300 | 225 | 430 | 2440 | 225 |
| 500 | 2600 | 250 | 510 | 2740 | 250 |
| 230 | 2000 | 100 | 240 | 1900 | 100 |
| 270 | 1900 | 125 | 280 | 1800 | 125 |
| 280 | 2000 | 150 | 290 | 1900 | 150 |
| 330 | 3200 | 175 | 340 | 3100 | 175 |
| 412 | 3200 | 200 | 422 | 3100 | 200 |
| 420 | 3200 | 225 | 430 | 3100 | 225 |
| 480 | 3500 | 250 | 490 | 3400 | 250 |
| 217 | 1500 | 100 | 317 | 3100 | 317 |
| 257 | 1400 | 125 | 399 | 2900 | 399 |
| 267 | 1500 | 150 | 407 | 3100 | 407 |

Así, con el paquete estadístico stata obtenemos la siguiente regresión:

Y = 83.66359 + 0.0775272X2 + 0.3918139X3

ee= (35.84106) (0.0210883) (0.1868419)

t = (2.33) (3.68) (2.10)

p = (0.026) (0.044) (0.001)

F = 29.74

p = 0.001

$r^2$ = 0.6574

Interpretemos los coeficientes de regresión: 0.0775272 es el coeficiente de regresión parcial del riego, este indica que, si se mantiene constante la influencia del abono aplicado, a medida que los mm de agua aplicado en la planta (riego) se incrementa, por ejemplo, en 1 mm en promedio, el ren-

dimiento del cultivo aumenta en 0.0775272 kg/ha. Interpretando esto desde el punto de vista agronómico, si el productor incrementa en el riego 100 mm de agua, en promedio, el rendimiento del cultivo por ha aumenta en 7.75272 kilogramos. El coeficiente 0.3918139 establece que, si la influencia del riego se mantiene constante, el rendimiento del cultivo aumentará, en promedio 0.3918139kg/ha, si el abono aplicado subiera 1kg/ha. El valor del intercepto de 83.66359, interpretado de forma mecánica, expresa que, si los valores del riego y el abono fuesen cero, el rendimiento del cultivo en promedio sería de 83.66359 kg/Ha. El valor de $R^2$ es de 0.6574, lo que significa que el 65.74% de la variación en el rendimiento del cultivo por hectárea se explica mediante el abono y el nivel de riego aplicado. Sin embargo, es necesario aclarar, que a pesar de que el $R^2$ es una medida de bondad de ajuste del modelo, que permite determinar que tan bueno es el modelo para predecir, este desempeña un papel modesto en el análisis de regresión, ya que en un análisis de regresión el objetivo no es obtener una $R^2$, sino obtener estimadores confiables de los verdaderos coeficientes de regresión poblacional que permita hacer inferencia estadística sobre ellos. Por lo expresado, la pertinente es establecer correctamente las variables explicativas para la variable dependiente y por su significancia estadística.

Si se desea conocer qué pasaría con el rendimiento del cultivo si el abono y el nivel de riego se incrementaran simultáneamente, solo se tendría que multiplicar ambos coeficientes por los cambios generados y sumar los resultados. Como ejemplo, supongamos que el producto se pregunta ¿Cuál sería el efecto simultáneo en una unidad de cambio en el abono (1kg) y una de riego (1 mm) en el rendimiento del cultivo?, esto da:

$$0.0775272(1) + 0.3918139(1) = 0.4693411$$

Es decir, como resultado de este cambio simultaneo en el abono y en el riego, el rendimiento del cultivo por ha aumentará 0.4693411 kg.

## Análisis de regresión con variables dicotómicas

Los modelos de regresión analizados anteriormente fueron en esencia con variables de tipo escala de razón, sin embargo, debemos considerar que los modelos de regresión también trabajan con otro tipo de variables (escala ordinal, escala de intervalo y escala nominal). En este apartado se consideran modelos que no solo tengan variables en escala de razón, sino también variables en escala nominal. Estas últimas también son denominadas variables categóricas, variables cualitativas o variables indicadoras.

En el análisis de regresión la variable dependiente frecuentemente no solo es influenciada por variables en escala de razón (por ejemplo: producción, cantidad de fertilizante aplicado) sino también por variables de naturaleza cualitativa (por ejemplo: variedad, tipo de suelo). Como ejemplo, con los demás factores constantes, se ha visto que la producción de un cultivo es superior en suelos francos que en suelos arenosos. Por tal motivo, las variables cualitativas, como el tipo de suelo y la variedad, si influyen en la variable dependiente deben ser incluidas en el modelo como parte de las independientes o explicativas.

Como la característica de estas variables es la de indicar la presencia o ausencia de un atributo como suelo franco o no, variedad Cavendish o no, son variables en escala nominal principalmente. Si la variable tiene más de dos categorías, conviene dicotomizarla mediante la introducción de variables dummy que toman valores de 0 y 1, estos atributos se pueden cuantificar, donde 1 indica la presencia del atributo y 0 su ausencia. Por ejemplo, 1 puede indicar que el suelo de tipo arcilloso y 0 de un suelo de distinto al arcilloso; 0 1 puede indicar que el banano es de la variedad Cavendish y 0 que es de otra variedad. Las variables que toman los valores de 0 y 1 se denominan variables dicotómicas. Dicotomizar

Las variables dicotómicas pueden tratarse de igual forma en los modelos de regresión como las variables cuantitativas. Incluso, un modelo de regresión puede estar conformado únicamente por variables independientes dicotómicas. Los

modelos con la composición mencionada se conocen como modelos de análisis de varianza (ANOVA), descrito en el capítulo 5 como ANOVA de un factor.

Para ilustrar el análisis de varianza con variables dummy se usarán los datos que se presentan en el cuadro 6.5, éstos se refieren a la producción del cultivo de arroz (en toneladas) de 60 productores en distintas provincias del Ecuador – Las provincias es una variable cualitativa nominal con k=3 categorías, por tal motivo, para incluirla en el modelo de regresión ésta debe dicotomizarse; es decir, crear k-1=2 variables dummy (D1 y D2) –. Las provincias son: 1) El Oro (19 productores en total); 2) Los Ríos (25 productores en total), y 3) Guayas (16 productores en total).

Se desea conocer si la producción promedio de los productores de arroz difiere en relación a las 3 provincias del Ecuador. Si se toma el promedio aritmético simple de los productores de las tres provincias, obtenemos los siguientes promedios para las tres provincias: 4.57 (El Oro), 3.64 (Los ríos) y 4.18 (Guayas). Esos números difieren entre sí, pero, ¿son estadísticamente diferentes? Para compara dos o más valores medios por lo general se usa la técnica conocida como análisis de varianza, pero se logra lo mismo por medio de una regresión. En relación a lo mencionado, considere el siguiente modelo para continuar con el ejemplo:

$$Y = \beta 0 + \beta 1 D1 + \beta 2 D2 + u$$

Donde:

Y = producción (promedio) de los productores de arroz

D1 = 1 si el productor pertenece a la provincia de El Oro

= 0 para otra provincia del País

D2 = 1 si el productor pertenece a la provincia de El Guayas

= 0 para otra provincia del País

Observe que el modelo es como los otros modelos presentados en este capítulo, a diferencia que, en vez de tener variables independientes cuantitativas, se tienen solo variables dicotómicas o cualitativas.

Cuadro 6.5. Producción hipotética promedio del cultivo de arroz en las provincias de El Oro, Guayas y Los Ríos.

| Productor | Producción | D2 | D3 | Productor | Producción | D2 | D3 |
|---|---|---|---|---|---|---|---|
| 1 | 3.8 | 0 | 0 | 31 | 4.5 | 0 | 1 |
| 2 | 3.6 | 0 | 0 | 32 | 4.3 | 0 | 1 |
| 3 | 3.9 | 0 | 0 | 33 | 4 | 0 | 1 |
| 4 | 3.2 | 0 | 0 | 34 | 4 | 0 | 1 |
| 5 | 3.6 | 0 | 0 | 35 | 3.9 | 0 | 1 |
| 6 | 3.7 | 0 | 0 | 36 | 4.4 | 0 | 1 |
| 7 | 3.7 | 0 | 0 | 37 | 4.3 | 0 | 1 |
| 8 | 3.7 | 0 | 0 | 38 | 4.2 | 0 | 1 |
| 9 | 4 | 0 | 0 | 39 | 4.4 | 0 | 1 |
| 10 | 3.6 | 0 | 0 | 40 | 4.3 | 0 | 1 |
| 11 | 3.5 | 0 | 0 | 41 | 3.8 | 0 | 1 |
| 12 | 3.9 | 0 | 0 | 42 | 4 | 1 | 0 |
| 13 | 3.5 | 0 | 0 | 43 | 4.4 | 1 | 0 |
| 14 | 3.3 | 0 | 0 | 44 | 4.7 | 1 | 0 |
| 15 | 3.4 | 0 | 0 | 45 | 4.8 | 1 | 0 |
| 16 | 3.7 | 0 | 0 | 46 | 4.7 | 1 | 0 |
| 17 | 3.8 | 0 | 0 | 47 | 4.9 | 1 | 0 |
| 18 | 3.8 | 0 | 0 | 48 | 4.5 | 1 | 0 |
| 19 | 3.9 | 0 | 0 | 49 | 4.7 | 1 | 0 |
| 20 | 3.7 | 0 | 0 | 50 | 4.6 | 1 | 0 |
| 21 | 3.4 | 0 | 0 | 51 | 4.7 | 1 | 0 |
| 22 | 3.5 | 0 | 0 | 52 | 4.8 | 1 | 0 |
| 23 | 3.6 | 0 | 0 | 53 | 4.3 | 1 | 0 |
| 24 | 3.6 | 0 | 0 | 54 | 4.4 | 1 | 0 |
| 25 | 3.5 | 0 | 0 | 55 | 4.6 | 1 | 0 |

| Productor | Producción | D2 | D3 | Productor | Producción | D2 | D3 |
|---|---|---|---|---|---|---|---|
| 26 | 4.2 | 0 | 1 | 56 | 4.5 | 1 | 0 |
| 27 | 3.9 | 0 | 1 | 57 | 4.4 | 1 | 0 |
| 28 | 4 | 0 | 1 | 58 | 4.8 | 1 | 0 |
| 29 | 4.4 | 0 | 1 | 59 | 4.4 | 1 | 0 |
| 30 | 4.3 | 0 | 1 | 60 | 4.7 | 1 | 0 |

El modelo expresa, que la producción promedio de los productores de Los Ríos está dado por el intercepto β0, en la regresión múltiple; los coeficientes β1 y β2 indican la producción promedio de los productores de El Oro, así como los del Guayas, difieren respecto de la producción promedio de los productores de Los Ríos. Sin embargo, ¿cómo determinar si las diferencias encontradas son significativas estadísticamente? Con los datos del Cuadro 6.5 se obtienen los siguientes resultados.

Y = 3.636+ 0.9376842D1 + 0.54525D2

ee = (0.0420427) (0.0639794) (0.0673012)

t = (86.48) (14.66) (8.10)

p = (0.001) (0.001) (0.001)

F = 110.01

p = 0.001

$r^2$ = 0.7942

Como reflejan los resultados de la regresión, la producción promedio de los productores de Los Ríos es de 3.64 toneladas, el de los productores de la provincia de El Oro es mayor al de los Ríos por cerca de 0.9377 toneladas, así como los de Guayaquil, es mayor por alrededor de 0.545 toneladas. Las producciones medias reales de las últimas provincias se obtienen sumando estos promedios diferenciales a la producción promedio de los productores de los Ríos. Al realizar esto, tendremos que la producción promedio de las últimas dos provincias son próximos a 4.57 y 4.18 respectivamente.

Pero, para saber si estas producciones promedio son estadísticamente diferentes de la producción promedio de Los Ríos, que es la provincia con la que se compara, hay que averiguar si cada coeficiente β1 y β2 son significativos. Como se observa en la regresión, el coeficiente estimado de la pendiente para la provincia de El Oro es estadísticamente significativo, porque el valor de p es menor a 0.05; también el del Guayas es estadísticamente significativo, ya que el valor de p también es menor a 0.05. Por tal manera, se concluye, que estadísticamente, la producción de arroz de los productores de las provincias de Los Ríos, El Oro y Guayas son distintas entre sí.

Debe considerar que las variables dicotómicas solo establecen las diferencias, si es que existen, pero no explica las razones por las que se generan. A lo mejor, el nivel de fertilizante, la nutrición, el tipo de suelo, los mm de precipitación puedan ejercen influencia en los resultados.

Es necesario aclarar que: 1) la categoría a la que no se la asigna variable dicotómica es conocida como categoría omitida, de referencia, de comparación, de control u categoría base; 2) el valor del intercepto β0 representa el valor medio de la variable de referencia, en el ejemplo trabajado dicha categoría es la correspondiente a la provincia de Los Ríos; 3) los coeficientes asociados a las variables dicotómicas se conocen como coeficientes de intercepto diferencial, indicando la medida en que la categoría que toma el valor de 1 difiere del coeficiente del intercepto correspondiente a la categoría de referencia o comparación y 4) queda a criterio del analista o investigador establecer la categoría de comparación.

## Modelo de regresión y ANCOVA

Los modelos de regresión que contienen como variables independientes una mezcla de variables cualitativas y cuantitativas se denominan modelos de análisis de covarianza (ANCOVA). Para ilustrar el modelo utilizaremos los datos del Cuadro 6.6.

Cuadro 6.6.Rendimiento promedio de un cultivo considerando el riego, la cantidad de abono aplicado y el tipo de suelo.

| Rendimiento/ kg x Ha | Riego (mm) | Abono /kg x Ha | D3 | D4 | Rendimiento/ kg x Ha | Riego (mm) | Abono / kg x Ha | D3 | D4 |
|---|---|---|---|---|---|---|---|---|---|
| 200 | 1500 | 100 | 0 | 0 | 195 | 1640 | 100 | 0 | 0 |
| 250 | 1400 | 125 | 0 | 0 | 200 | 1540 | 125 | 1 | 0 |
| 300 | 1500 | 150 | 0 | 0 | 260 | 1640 | 150 | 1 | 0 |
| 350 | 2300 | 175 | 0 | 0 | 290 | 2440 | 175 | 0 | 0 |
| 400 | 2300 | 200 | 0 | 0 | 408 | 2440 | 200 | 1 | 0 |
| 450 | 2300 | 225 | 1 | 0 | 430 | 2440 | 225 | 1 | 0 |
| 500 | 2600 | 250 | 1 | 0 | 510 | 2740 | 250 | 0 | 1 |
| 230 | 2000 | 100 | 0 | 0 | 240 | 1900 | 100 | 0 | 0 |
| 270 | 1900 | 125 | 0 | 0 | 280 | 1800 | 125 | 0 | 0 |
| 280 | 2000 | 150 | 0 | 1 | 290 | 1900 | 150 | 0 | 0 |
| 330 | 3200 | 175 | 0 | 0 | 340 | 3100 | 175 | 0 | 0 |
| 412 | 3200 | 200 | 1 | 0 | 422 | 3100 | 200 | 1 | 0 |
| 420 | 3200 | 225 | 1 | 0 | 430 | 3100 | 225 | 0 | 1 |
| 480 | 3500 | 250 | 1 | 0 | 490 | 3400 | 250 | 1 | 0 |
| 217 | 1500 | 100 | 0 | 0 | 317 | 3100 | 317 | 0 | 1 |
| 257 | 1400 | 125 | 0 | 0 | 399 | 2900 | 399 | 0 | 1 |
| 267 | 1500 | 150 | 0 | 0 | 407 | 3100 | 407 | 0 | 1 |

Con los datos en mención (Cuadro 6.6) se desarrolla el siguiente modelo:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3D3 + \beta 4D4$$

Donde:

Y = rendimiento (promedio) de un cultivo

D3 = 1 si el suelo es franco

= 0 para otro tipo de suelo

D4 = 1 si el suelo es franco arcilloso

= 0 para otro tipo de suelo

X1 = mm de riego

X2 = Abono aplicado

Tenga presente que los datos proporcionados en el Cuadro 6.6 considera al suelo franco arenoso como la categoría de comparación. También, note que, además de las dos variables independientes o regresoras cualitativas, se tienen dos variables cuantitativas, X1 y X2, que en los modelos de ANCOVA se denominan covariante. Ya que, estos modelos representan una extensión de los modelos ANOVA en el sentido de que establecen un método para controlar estadísticamente los efectos de las variables independientes cuantitativas (nombradas variables concominantes o covariables).

De los datos del Cuadro 6.6, los resultados obtenidos del modelo (4) son los siguientes:

Y = 92.73969 + 0.60828X1 + 0.470243X2 + 53.1269D3 - 10.97473D4

ee = (35.6568) (0.203315) (0.223505) (24.84383) (37.64764)

t = (2.60) (2.99) (2.10) (2.14) (-0.29)

p = (0.014) (0.006) (0.044) (0.041) (0.773)

F = 19.48

p = 0.001

r2 = 0.7287

Como los resultados reflejan, cuando todo lo demás se mantiene constante: mientras el riego aumenta 1mm, en promedio, el rendimiento por hectárea se incrementa alrededor de 0.60 kilogramos. Así mismo, mientras el abono aplicado aumenta en 1 kg/ha, el rendimiento se incrementa más o menos en 0.47 kg/ha. Si controlamos los mm de riego y el abono, se observa que el coeficiente del intercepto diferencial es significativo para el suelo franco, pero no lo es para el suelo franco arcilloso.

## Referencia bibliográfica


Berndt, Ernst R. (1991). The Practice of Econometrics, Classic and Contemporary, Addison-Wesley.

Cameron, A. Colin y Pravin K. Trivedi. (2005). Microeconomics: Methods and Applications, Cambridge University Press, Nueva York.

Depool, R., & Monasterio, D. (2013). *Probabilidad y estadística. Aplicaciones a la ingeniería.* Barquisimeto, Venezuela: Universidad Nacional Experimental Politécnica Antonio José de Sucre (Unexpo). Obtenido de http://www.bqto.unexpo.edu.ve/avisos/PROBABILIDADYESTADISTICA(2-7-13).pdf

Garriga, A. J., Lubin, P., Merino, J. M., Padilla, M., Recio, P., & Suárez, J. C. (2010). *Introducción al análisis de datos.* Madrid, España: Universidad Nacional de Educación a Distancia (UNED).

Goldberger, Arthur S. (1998). Introductory Econometrics, Harvard University Press.

Goldberger, A. S. (1968). Topics in Regression Analysis, Macmillan, Nueva York.

Greene, William H. (2000). Econometric Analysis, 4a. ed., Prentice Hall, Englewood Cliffs.

Gujarati, Damodar N. (2006). Essentials of Econometrics, 3a. ed., McGraw-Hill, Nueva York.

Gujarati y Porter. (2010). *Econometría,* 5a. ed., McGraw-Hill, Nueva York.

Hayashi, Fumio. (2000). Econometrics, Princeton University Press, Princeton, N. J.

Hernández, R., Fernández, C., & Baptista, P. (2014). *Metodología de la investigación* (Sexta ed.). D.F. México: McGraw-Hill.

IBM Corp. (2016). *SPSS Statistics versión 24.0.0.0* de prueba para Windows. Barcelona: International Business Machines Corp.

Johnston, J. (1984). Econometric Methods, 3a. ed., McGraw-Hill, Nueva York.

Kennedy, Peter. (1998). A Guide to Econometrics, 4a. ed., MIT Press, Cambridge, Mass.

Klein, Lawrence R. (1962). An Introduction to Econometrics, Prentice Hall, Englewood Cliffs, N.J.

Lind, D. A., Marchal, W. G., & Wathen, S. A. (2015). *Estadística aplicada a los negocios y la economía*. México, D. F.: McGraw-Hill Internacional.

Patterson, Kerry. (2000). An Introduction to Applied Econometrics, A Time Series Approach, St. Martin's Press, Nueva York.

Peracchi, Franco. (2001). Econometrics, John Wiley & Sons, Nueva York.

Sáez, A. J. (2012). *Apuntes de estadística para ingenieros*. España: Dpto de Estadística e Investigación Operativa. Universidad de Jaén. Obtenido de http://www4.ujaen.es/~ajsaez/recursos/EstadisticaIngenieros.pdf

Salcedo, A. (2013). *Estadística en el investigación*. Caracas: ISBN: 978-980-00-2743-1.

Walters, A. A. (1968). An Introduction to Econometrics, Macmillan, Londres.

Wooldridge, Jeffrey M.. (2000). Introductory Econometrics, 3a. ed., South-Western College Publishing.

# 07 Capítulo Inteligencia de negocios en el sector agropecuario

Bertha Mazon-Olivo; Alberto Pan; Raquel Tinoco-Egas

La Inteligencia de Negocios (BI) se encarga de obtener el conocimiento a partir del procesamiento de datos crudos, con la finalidad de apoyar la toma de decisiones en una organización en sus niveles tácticos y estratégicos. Los temas que se abordan son: fundamentos de sistemas de información e inteligencia de negocios, el diseño e implementación de almacenes de datos (Data warehouse y Data mart), el proceso de extracción, transformación y carga de datos (ETL) y las técnicas de visualización de datos OLAP y tableros de control (dashboards); y, se concluye con el análisis de un caso de estudio y su desarrollo práctico de una solución BI.

**Bertha Mazon-Olivo:** Ingeniera en Sistemas y Magíster en Informática Aplicada. Profesora Titular en la Universidad Técnica de Machala. Líneas de investigación: Internet de las Cosas, integración, procesamiento y análisis de datos. Es estudiante del programa doctoral en Tecnologías de la Información y las Comunicaciones en Universidade da Coruña, España. Sus líneas de investigación son: Internet de las cosas, Ciencia de datos y Desarrollo de Aplicaciones Informaticas. Cuenta con varias publicaciones indexadas.

**Alberto Pan:** Es Director Técnico de Denodo y Profesor Asociado de la Universidad de A Coruña. Reibió una Licenciatura en Ciencias de la Computación en la Universidad de A Coruña en 1996 y un Ph.D. en Informática por la misma universidad en 2002. Sus intereses de investigación estan relacionados con la extracción e integración de datos y la automatizaación de la web. Alberto ha dirigido varios proyectos a nivel nacional y reegional en el campo de la integraciónde datos y accesi a la Web oculta. También es autor y coautor de numerosas publicaciones en revistas científicas y actas de congresos.

**Raquel Tinoco-Egas:** Ingeniera en Gestión Empresarial Internacional. Máster en Desarrollo de Negocios Internacionales por la Universidad de Neuchâtel de Suiza. Estudiante doctoral en Análisis Económico y Estrategia Empresarial en la Universidade da Coruña. Profesora Titular en la Universidad Técnica de Machala. Investiga temas como la innovación y tecnología para el desarrollo empresarial. Cuenta con varias publicaciones.

## Necesidades de información en una empresa agropecuaria

Una organización, dependiendo de su tamaño, objeto de actuación o negocio, genera cientos, miles o quizá millones de transacciones diarias que se traducen en datos; el sector agropecuario no es la excepción; el seguimiento de procesos productivos agrícolas o ganaderos, la gestión de recursos: financieros, humanos, materias primas, maquinaria, etc., generan muchos datos; si estos datos no son procesados, terminarán por ser olvidados y desaprovechados.

En la actualidad, con la Ciencia de Datos, es posible obtener el máximo provecho de la materia prima denominada "datos crudos" o "datos en bruto"; mediante el procesamiento de estos datos se genera información útil para la toma de decisiones en beneficio de la propia organización. Las disciplinas relacionadas con esta ciencia son: la Estadística, Inteligencia de Negocios, Minería de Datos, Inteligencia Artificial, Aprendizaje Automático, Bases de Datos y otras.

Para comprender la necesidad de la información en una organización del sector agropecuario, es necesario revisar el siguiente concepto:

Explotación Agropecuaria, comprende la producción agrícola y/o ganadera, que puede ser llevada a cabo mediante una gestión única, por una persona, un hogar, grupos familiares, asociaciones, cooperativas o empresas. La Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO, 2016), la define como:. "*... la unidad económica de producción agropecuaria bajo gestión única, que comprende todo el ganado mantenido en ella y toda la tierra dedicada total o parcialmente a fines agrícolas, independientemente del título, forma jurídica o tamaño. ...Las tierras de la explotación agropecuaria pueden constar de una o más parcelas, situadas en una o más áreas separadas en una o más divisiones territoriales o administrativas, siempre que todas las parcelas compartan los mismos recursos*".

El proceso de toma de decisiones en una organización de producción agrícola (unidad de explotación agropecuaria) no se diferencia de otras empresas o instituciones. La Imagen 7.1, está basada en el libro de Laudon & Laudon (2012), ilustra los niveles de una organización, los tipos de sistemas de información, el proceso de conversión de datos en una decisión y finalmente la valoración de los datos según los criterio de volumen y valor. A continuación se describen en detalle cada componente del gráfico.

Imagen 7.1. Relación entre los sistemas de información y los niveles de una organización


Fuente: Elaboración propia

## Niveles de una organización

En una organización, los tipos de sistemas de información se distribuyen en tres niveles:

- Nivel Operativo. En este nivel se ubican los sistemas de información que gestionan las transacciones diarias de una organización y son los generadores de "*datos crudos*" o en bruto. Las tareas, recursos y metas están predefinidos y bien estructurados. Los Sistemas de Procesamiento Transaccional (TPS) se encargan del control de

recursos y procesos, por ejemplo: producción, compras, ventas, inventarios, etc.

- Nivel Táctico. En el nivel táctico se ubican sistemas de información de mandos medios como son los Sistemas de Soporte de Decisiones (DSS) y los Sistema de Información Gerencial (MIS). Los usuarios de este nivel planifican, dirigen y controlan las acciones del nivel operativo, toman decisiones a mediano plazo con afectación según su área o departamento; generan reportes con *"información"* consolidada a partir de los datos crudos de acuerdo a su ámbito de acción, para mantener informados a los ejecutivos de mandos estratégicos.
- Nivel Estratégico. En este nivel se encuentra los sistemas de Información para Ejecutivos (EIS) que apoyan a la alta gerencia en la toma de decisiones a largo plazo y que afectan a toda la organización. La junta de accionistas, el gerente general o propietario, aprovechan la información de los niveles inferiores para su *"conocimiento"* y en base a su experiencia o *"sabiduría"*, son capaces de tomar la *"decisión"* más acertada.

## Sistema de información (SI)

- Comprende un conjunto de personas, procedimientos, datos y tecnologías que apoyan las actividades de una organización. Según Laudon & Laudon (2012), lo definen como: *"...conjunto de componentes interrelacionados que recolectan (o recuperan), procesan, almacenan y distribuyen información para apoyar los procesos de toma de decisiones y de control en una organización"*. *Además, "...ayudan a los gerentes y trabajadores del conocimiento a analizar problemas, visualizar temas complejos y crear nuevos productos"*.

## Tipos de sistemas de información

Transaction Processing System (TPS). El sistema de Procesamiento Transaccional se encargan de recolectar, almacenar, procesar, calcular, ordenar, modificar y recuperar los datos

obtenidos de las transacciones diarias de una empresa.

Decision Support System (DSS). El sistema de Soporte de Decisiones, genera información a partir de los datos que provee el TPS y apoya las decisiones semi-estructuradas de los usuarios de mandos medios y estratégicos. Los DSS presentan información estadística, dinámica, multidimensional y consolidada acorde a los Indicadores Clave de Rendimiento (KPI's) de cada área de la organización, que regularmente se estructura en cubos OLAP (On-Line Analytical Processing).

Management Information System (MIS). El sistema de Información Gerencial comprende una colección de sistemas de información, que relacionados entre sí, apoyan las actividades de una organización en todos sus niveles. Las funciones principales de esto sistema son la planeación, organización, dirección y control de las actividades del nivel operativo, apoyo a las decisiones de nivel táctico y a su vez, informar los avances y resultados de actividades, al nivel estratégico.

Executive Information System (EIS). El sistema de Información Ejecutiva, está orientado a los usuarios de alta gerencia, tiene la función de presentar de forma sencilla, el estado actual de los KPI's relevantes de la organización, basándose en otros sistemas como los DSS, cubos OLAP y varias fuentes de datos empresariales. La interfaz típica de estos sistemas son los cuadros de mando o tableros de control (dashboards) que presentan información estadística dinámica descriptiva, diagnóstica o predictiva.

Enterprise Resource Planning (ERP). El sistema de Planificación de Recursos Empresariales, es un sistema integral, que gestionan recursos de toda la empresa como por ejemplo: recursos (naturales) para la producción, inventario de bienes y productos (físicos), logística, distribución, compras a proveedores y ventas a clientes (mercados), contable-financiero, recursos humanos y sociales. Los ERP se relacionan con otros sistemas como los CRM, SCM y MIS, DSS y EIS; estos sistemas pueden ser desarrollados a medida como sistemas independientes o como módulos de un ERP.

Customer Relationship Management (CRM). El sistema

de Gestión de la Relación con los Consumidores, se enfoca en servicios y estrategias de marketing para lograr la fidelización de los clientes que ya son parte de la empresa, así como para captar nuevos clientes. Permite llevar un control de reuniones, el registro del historial de acuerdos en procesos de negociación, contratos, convenios, etc. Por ejemplo las empresas inmobiliarias manejan este tipo de sistemas.

Supply Chain Management (SCM). El sistema de Gestión de la Cadena de Suministros, involucra a toda la logística que va desde el contacto con proveedores que suministran la materia prima necesaria para la producción de nuevos artículos; y, la organización de la cadena de distribuidores y vendedores para llegar al cliente o consumidor final.

Existen otros tipos de sistemas de información que contribuyen en las actividades de una organización, por ejemplo: los sistemas de ofimática, de mensajería o comunicación, de gestión documental, sistemas de comercio electrónico, de trabajo colaborativo o en grupo, etc.

## Etapas para la obtención del conocimiento y toma de decisión

Como se observa en parte derecha de la Imagen 7.1, los datos que se obtienen de los sistemas de procesamiento transaccional, son procesados para generar información, la misma que al ser presentada de manera oportuna a la persona adecuada y con la experiencia necesaria, ésta es capaz de analizarla, convertirla en conocimiento y actuar con sabiduría tomando la decisión más acertada en beneficio de su organización.

## Volumen y valor de los datos

En el nivel más bajo de la pirámide de la Imagen 7.1, los datos se obtienen en grandes cantidades de las transacciones diarias de la organización; éstos son considerados de poco valor porque no están procesados o no tiene una estructura enfocada en los Indicadores Clave de Rendimiento del negocio (KPI's). Los mismos datos al pasar por los niveles

táctico y estratégico se van procesando, resumiendo y consolidando de tal forma que su volumen disminuye pero su valor aumenta. Por ejemplo: supóngase que en promedio se generan 100 facturas de venta diariamente; al año serían un estimado de 36 000 facturas. Al gerente no le interesa las copias de todas las facturas, sino un reporte consolidado de ventas según el criterio específico; por ejemplo, puede ser el total de ventas por mes, trimestre, semestre o año; dicho reporte no pasará de una o dos páginas.

## Los objetivos y usuarios de la información

En la Imagen 7.2, se aprecia los usuarios de la información categorizados por niveles:

Usuarios operativos. En el nivel operativo se ubican algunos trabajadores administrativos, de servicio y producción, por ejemplo: bodeguero, vendedores, auxiliares contables, operarios, etc. Este tipo de usuarios tienen como objetivos: realizar o ejecutar acciones encomendadas por su jefe inmediato superior, planificar acciones a corto plazo (diarias, semanales o mensuales) e informar periódicamente el resultado de sus actividades a su jefe.

Usuarios de mandos medios. En el nivel táctico se encuentran los directivos, gerentes o jefes de sucursales o departamentos; por ejemplo, administrador de finca o hacienda, el director financiero, el director de producción, el gerente de zona, etc. Los usuarios de este nivel se encargan de: realizar planes operativos a mediano plazo (máximo 1 año), organizar y coordinar actividades, recursos y el personal; dirigir, controlar y supervisar al personal y los resultados que generan; y finalmente, tomar decisiones por área o departamento.

Usuarios de alta gerencia. En el nivel estratégico se ubican la junta de accionistas, gerente general o propietario. Los objetivos a este nivel son: tomar decisiones que afectan a toda la organización, preparar planes (estratégicos) a largo plazo (3 o más años), dirigir, controlar y supervisar la implementación de estrategias enfocadas en incrementar la productividad, con una mejora de la calidad de sus productos o

Imagen 7.2. Objetivos y usuarios de la información en una organización

| | Usuarios de la información | Objetivos de los usuarios respecto al manejo de la información |
|---|---|---|
| **ESTRATÉGICO** alta dirección | **Usuarios de alta gerencia:**<br>- Junta de accionistas/socios<br>- Gerente o director general<br>- Propietario | **Con alcance global:**<br>- Tomar decisiones que afectan a toda la organización<br>- Planear a largo plazo (plan estratégico a 3 años o más)<br>- Dirigir, controlar y supervisar la implementación de estrategias enfocadas a incrementar la productividad, mejorando la calidad con el menor costo posible |
| **TÁCTICO** de mandos medios | **Usuarios de mandos medios.**<br>Director, jefe, gerente de sucursal o de departamento:<br>- producción,<br>- contabilidad y finanzas,<br>- recursos humanos,<br>- comercial y mercadeo, | **Por área, departamento o zona:**<br>- Tomar decisiones con afectación en su área<br>- Elaborar planes operativos (máximo 1 año)<br>- Organizar y coordinar actividades, recursos y el personal<br>- Dirigir, controlar y supervisar al personal y los resultados obtenidos (productos o servicios) |
| **OPERATIVO** transaccional | **Usuarios operativos:**<br>algunos trabajadores administrativos, de servicio y producción:<br>- bodeguero, vendedores<br>- auxiliar contable,<br>- operarios de maquinaria, etc. | **Según el cargo:**<br>- Realizar o ejecutar acciones encomendadas por su jefe inmediato superior<br>- Planificar actividades a corto plazo (diarias, semanal o mensual)<br>- Informar resultados de su actividad o función |

Fuente: Elaboración propia

## Los problemas con los datos en una organización

Obtener un valor agregado de los datos para apoyar la toma de decisiones, es una tarea que requiere de conocimiento en el manejo de técnicas de Ciencia de Datos y herramientas tecnológicas. Algunas empresas desaprovechan sus datos y no analizan sus patrones de comportamiento en beneficio propio, debido al poco o inexistente conocimiento en el manejo de herramientas tecnológicas. A continuación se describen algunos problemas que tienen las empresas con sus datos:

- La falta de automatización e integración de sus departamentos y procesos hace que se ocupe el tiempo en tareas triviales de gestión manual de documentos.
- El sistema de procesamiento transaccional de la empresa puede ser limitado o tener algunos defectos que provoca la desconfianza de los usuarios en los informes o resultados de búsquedas.
- La alta gerencia y mandos medios pueden dar poca importancia a la información de un sistema informático, debido al inadecuado tratamiento y presentación.

- Los datos de las transacciones diarias sirven para realizar el proceso del momento y pronto éstos son olvidados en archivos impresos o en bases de datos electrónicas.
- Falta de conocimiento de la tecnología adecuada para el tratamiento de la información debido al presupuesto de la empresa.
- Los datos electrónicos pueden estar expuestos a perderse o ser robados por falta de seguridad.
- Falta de políticas de gobierno de datos.

## Fundamentos de inteligencia de negocios

La Inteligencia de Negocios es la tecnología que permite extraer, transformar y analizar los datos para generar escenarios, informes y pronósticos que apoyen a la toma de decisiones, lo que se traduce en una ventaja competitiva. La información adecuada en el lugar y momento oportuno incrementa la efectividad de cualquier empresa y las del sector agropecuario no son la excepción, los datos crudos se generan de distintas áreas como: producción, mercadeo, ventas, finanzas, recurso humano, etc. Una solución BI involucra la creación de nuevos almacenes de datos (data warehouse), que son alimentados mediante un proceso de ETL (Extraction Transformation and Loading), de distintas fuentes de datos estructurados y no estructurados (por ejemplo: bases de datos relacionales, hojas de cálculo, archivos planos, etc.), para proporcionar la información oportuna a las aplicaciones BI (análisis multidimensional OLAP, consultas y reportes, tableros de control, minería de datos y otras aplicaciones personalizadas) y a los usuarios que toman decisiones.

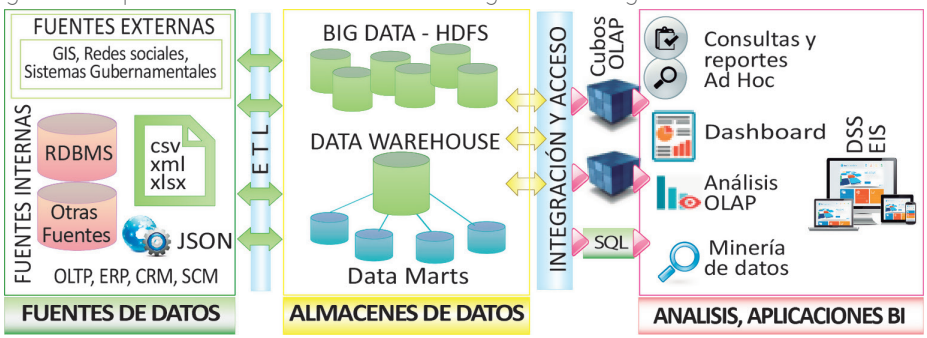
Según Mazon-Olivo et al. (2017), BI comprende un conjunto de estrategias y componentes que permiten transformar los datos operacionales en información y ésta en conocimiento útil para la toma de decisiones; es decir, facilita el monitoreo del cumplimento de los objetivos organizacionales y admite el análisis de la información histórica, contribuyendo

a la creación de estrategias comerciales que generan ventajas competitivas en el mercado. BI tiene múltiples campos de aplicación, entre los más comunes están los sectores: comercial, empresarial, industrial, educativo, banca, turismo, y otros que requieren del análisis de sus datos para identificar tendencias o patrones que a su vez orientan la toma de decisiones.

## Arquitectura de una solución de inteligencia de negocios

Para el desarrollo de una solución de inteligencia de negocios es necesario comprender su arquitectura, en la Imagen 7.3 es posible apreciar 3 capas: la capa de fuentes de datos, la capa de almacenes de datos y la capa de análisis y aplicaciones BI. Entre la primera y segunda capas se requiere el proceso de extracción transformación y carga (ETL) que suministre los datos crudos o en bruto (raw data) al almacén de datos (data warehouse). Para el análisis de datos se requieren aplicaciones BI que accedan al data warehouse mediante cubos OLAP y procesos de integración de datos.

Imagen 7.3. Arquitectura de una solución de Inteligencia de Negocios




Fuente: (Mazon-Olivo et al., 2017)

### La capa de datos fuente

Los datos en bruto existen en gran cantidad y diversidad de formatos y pueden clasificarse según:

**El origen:**

- Internos. Son de la propia empresa, pueden presentarse en bases de datos, archivos o reportes de sistemas informáticos (ERP, CRM, SCM)
- Externos. Se obtienen de otras organizaciones (redes sociales, sistemas de información geográfica, sistemas gubernamentales) mediante web services, descarga de sitios web, por email, reportes de sistemas informáticos.

**El formato:**

- Estructurados. Se obtienen de bases de datos relacionales, otras data warehouse
- Semi-estructurados. Por lo general son archivos en formatos: CSV, JSON, XML, HTML, etc.
- No estructurados. datos de archivos como PDF, imagen, sonido, video, etc.

**El tamaño:**

- Volúmenes de datos normales. Cientos o miles de registros medidos en KB o MB.
- Grandes volúmenes de datos (big data). Millones de registros medidos en GB, TB o PB (Peta Byte), generalmente se encuentran en sistemas clusterizados con mecanismos de procesamiento y almacenamiento distribuido.

### La capa de almacenes de datos

Comprende los grandes repositorios de metadatos, en este ámbito se encuentran los Data Warehouse y los Big Data.

**Data Warehouse (DW).** Es considerado un almacén o bodega de datos estructurados que contiene información temática, histórica e integrada según los indicadores clave de desempeño que se hayan previsto en una organización (Rosado G. & Rico B., 2010). Un DW es temático debido a que está conformado por áreas o grupo de datos de una organización. Cada tema del DW es representado por un Data mart.

**Big Data.** Representa un conjunto de recursos de información de gran volumen, que se obtienen a altas velocidades y con una variedad de formatos, que demandan un almacenamiento escalable y eficiente, formas innovadoras de procesamiento de información para mejorar el análisis, la comprensión y toma de decisiones (NIST, 2015). Una Big Data es una oportunidad para que las organizaciones obtengan ventajas competitivas en el mundo actual digitalizado y globalizado (De Mauro, Greco, & Grimaldim, 2015). Un término asociado a Big Data es Data Lake, que comprende grandes conjuntos de datos (big data sets), también llamados lagos de datos, que se etiquetan para realizar consultas o buscar patrones. Los tipos de Big Data en función de sus prestaciones pueden ser: 1) de alto rendimiento y 2) distribuidos, por lo general de bajo rendimiento. Ejemplos de sistemas de alto rendimiento son: Teradata, HP Vertica, IBM Netezza, Oracle Exadata (Moniruzzaman & Hossain, 2013; Ţăranu, 2015). Un ejemplo de sistema distribuido es Hadoop[1] (HDFS es Hadoop Distributed File System), considerado comos la plataforma más utilizadas para el procesamiento y almacenamiento distribuido de datos (Sawant & Shah, 2013; Ţăranu, 2015; White, 2015); ejemplos de versiones de plataformas que integran Hadoop son Hortonworks[2] , Cloudera[3] , AWS[4] , Microsoft Azure[5] , etc.

Las V's de la Big Data según (NIST, 2015):

- Volumen. Característica enfocada al tamaño del conjunto de datos (data set). Para esto se requiere tecnologías más eficientes enfocadas en la recolección, almacenamiento y procesamiento de datos. Ejemplos de generadores de grandes volúmenes de datos son: aplicaciones de la Web 2.0 como las redes sociales, el Internet

[1] Apache Hadoop: http://hadoop.apache.org/

[2] Hortonworks: https://es.hortonworks.com/

[3] Cloudera: https://www.cloudera.com/

[4] AWS Amazon Web Services: https://aws.amazon.com/es/

[5] Microsoft Azure: https://azure.microsoft.com/es-es/

de las Cosas (Internet of Things), los objetos inteligentes (Smart objects), entre otros.

- Velocidad. Se refiere a la tasa de flujo de datos. La rapidez con la que se producen los datos se debe en gran medida a la concurrencia de generadores de datos (personas y objetos), muchos de esos datos se transmiten en tiempo real.
- Variedad. Se refiere a la diversidad de formatos de datos provenientes de múltiples repositorios y dominios; por ejemplo: datos estructurados (Bases de datos relacionales, Data Warehouse), semi-estructurados (archivos de datos en formatos JSON, XML, HTML, xls etc.) y no estructurados (archivos de imágenes, videos, audios, etc.).
- Variabilidad. Representa el cambio de características de los datos. Los requisitos de seguridad y privacidad pueden cambiar dependiente de la naturaleza y el tiempo que toma llevar a cabo funciones como: recogida, procesamiento, agregación y almacenamiento de datos. La gobernabilidad también puede cambiar a medida que las organizaciones responsables se fusionan o incluso desaparecen.
- Veracidad. Los datos recopilados deben ser confiables, es decir, se refiere a que se debe hacer un control de calidad de los datos antes de realizar alguna operación con ellos.
- Otras v's o características también atribuidas a big data son: valor, viabilidad y visualización.

**Data Mart.** Comprende un subconjunto de datos enfocados en el análisis de un tema, área o ámbito específico en una organización (Mosquera & Hallo, 2014). El conjunto de data marts comprende un data warehouse. Algunos ejemplos de data mart en una empresa agropecuaria aplicados al proceso productivo puede ser: *prod_parcelas*, *prod_cultivos*, *prod_siembra*, *prod_cosecha*, *prod_ingresos*, *prod_costos*, *inventario*, *recursos-humanos y otros*.

Componentes de un Data Mart:

- Tabla de hechos (table fact). Es la tabla central del data mart, que contiene los datos o medidas (indicadores claves de desempeño del negocio KPI's) que se utilizan para análisis y las claves de las tablas de dimensiones. Es común nombrar las tablas de hechos en función del área/tema que representa el data mart; por ejemplo, si el data mart representa los *cultivos agrícolas* que se han producido o que están en proceso de producción, la tabla de hechos puede ser: *th_produccion_cultivos*.
- Medida o KPI (Key Performance Indicator). Representa un valor numérico que contribuye al análisis de un hecho. Por ejemplo: *densidad de siembra (número de plantas por hectárea), edad del cultivo (días o años), rendimiento de un cultivo (toneladas métricas por hectárea TM/ha), costos de producción por hectárea, rentabilidad de un cultivo (ingresos – egresos), etc.*
- Tablas de Dimensiones: Corresponden a las perspectivas o vistas a través de las cuales es posible analizar las medidas del negocio o KPI's. Por ejemplo: para la medida *densidad de siembra (número de plantas por hectárea)*, las posibles dimensiones pueden ser: *especie de cultivo (banano, cacao, café, etc.), clase de cultivo (frutales, hortalizas, cereales, etc.), tipo de cultivo según duración (ciclo corto, perennes), tiempo de la siembra (año, semestre, trimestre y mes),* tipo de suelo, etc.

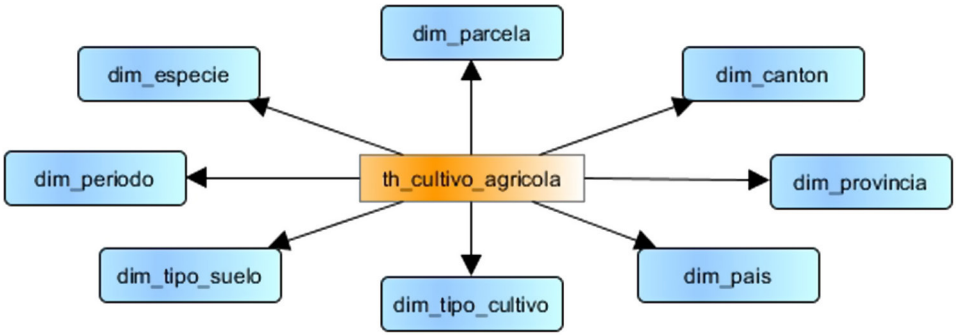
**Tipos de esquemas de diseño de un data marts y data warehouse.**

Existen 2 tipos de esquemas de diseños para crear modelos de data marts, estos son: estrella y copo de nieve; la integración de varios esquemas estrella o copo de nieve dan lugar a un esquema en forma de constelación que representa el data warehouse. A continuación se explican con mayor detalle:

Esquema en Estrella. Contiene una sola tabla de hechos con los datos de análisis y las claves de todas las tablas de

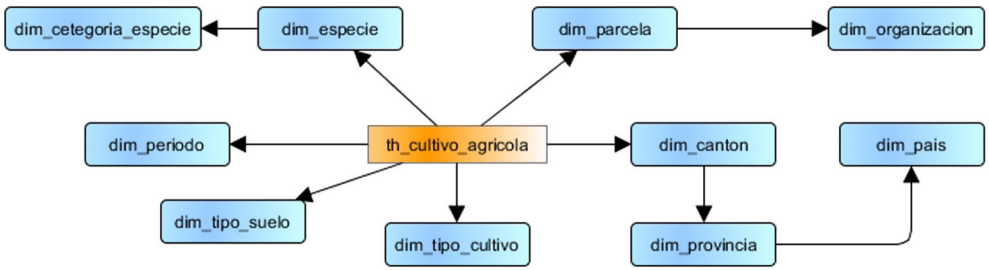
dimensiones. No existen relaciones entre las dimensiones. Ver Imagen 7.4.

Imagen 7.4. Esquema de un data mart en estrella




Esquema en Copo de Nieve. Similar al esquema en estrella con la diferencia de que las dimensiones pueden relacionarse creando jerarquías. Ver Imagen 7.5.

Imagen 7.5. Esquema de un data mart en copo de nieve




Esquema en Constelación. Integración de varios esquemas estrella o copos de nieve que representan un data warehouse. Su diseño puede llegar a ser complejo debido a la cantidad de relaciones entre las tablas de hechos y dimensiones, más si las dimensiones se comparten entre varios data marts. Es aconsejable manejar por separado el diseño de los data marts; sin embargo, en la implementación física del data warehouse es necesario optimizar y no repetir dimensiones.

## El proceso de Extracción Transformación y Carga (ETL)

ETL de sus siglas en inglés Extraction, Transformation and Load, es el proceso que permite seleccionar datos desde múltiples fuentes, utilizando una herramienta de integración se pre-procesan y se cargan en un nuevo almacén de datos (data warehouse).

Las actividades ETL según Cornejo et al. (2014), se describen a continuación:

- Extracción. Consiste en la identificación de los datos fuentes, verificación de su calidad, lectura de datos crudos, obtención de agregados y establecimiento de la estructura de la metadata del data warehouse.
- Transformación. Comprende la aplicación de una serie de reglas de negocio sobre los datos extraídos para convertirlos en datos con el formato del data warehouse. Algunas de estas actividades son: limpieza, cambio de formato, generación de datos calculados, creación de nuevos datos o claves, filtrado, ordenación, asociaciones y agregaciones.
- Carga. Las actividades que se realizan son: integración de datos, pruebas de carga, carga (escritura) de datos en el data warehouse, gestión de errores y mantenimiento de metadata. La carga de datos normalmente es mediante procesos batch, sea por lotes, por registros, por totales, u otra forma.

## Integración y virtualización de datos

La virtualización de datos consiste en el acceso a datos procedentes de distintos repositorios, ocultando la complejidad interna a las aplicaciones consumidoras. Las herramientas de integración y virtualización de datos, realizan dicha función, creando un único punto de acceso a los datos mediante una base datos lógica; y de esta manera facilitando el gobierno

de los datos (Van Der Lans, 2012). Ejemplos de este tipo de herramientas son: Denodo[6], Informatica[7], Cisco Data Virtualization[8], etc.

Cubos OLAP. OLAP significa procesamiento analítico en línea. Un cubo OLAP representa el esquema o definición de la estructura multidimensional para el análisis de datos empresariales que se encuentra agregados y organizados en un data warehouse. Un cubo se construye definiendo una tabla de hechos, las medidas y las dimensiones. Las consultas a un cubo OLAP se realizan mediante el lenguaje MDX (MultiDimensional eXpressions o expresiones multidimensionales). Si un data warehouse se implementa en un Sistema Gestor de Base de Datos (DBMS) relacional, se considera un sistema ROLAP, si se lo implementa en un sistema de almacenamiento multidimensional es MOLAP y si se implementa en los dos tipos entonces es HOLAP (Morales, Cuevas, & Martínez, 2016; Rosado & Rico, 2010).

### La capa analítica o de aplicaciones de inteligencia de negocios

En esta capa se ubican las aplicaciones y herramientas para el análisis descriptivo y predictivo de los datos, las más destacadas son: visores OLAP, tableros de control (dashboards), reportes y consultas Ad Hoc, Minería de datos y otras. Este tipo de aplicaciones se clasifican en dos tipos de sistemas que ya fueron mencionados en las secciones anteriores: los sistemas de soporte de decisiones (DSS) y los sistemas de información para ejecutivos (EIS). (Ghosh, Halder, & Sen, 2015; Gounder, Iyer, Professor-ccis, Mazyad, & Prof, 2016; Marinheiro & Bernardino, 2015; Vassell, Apperson, Calyam, Gillis, & Ahmad, 2016), como se puede observar en la Imagen 7.6. A continuación se explican con más detalle:

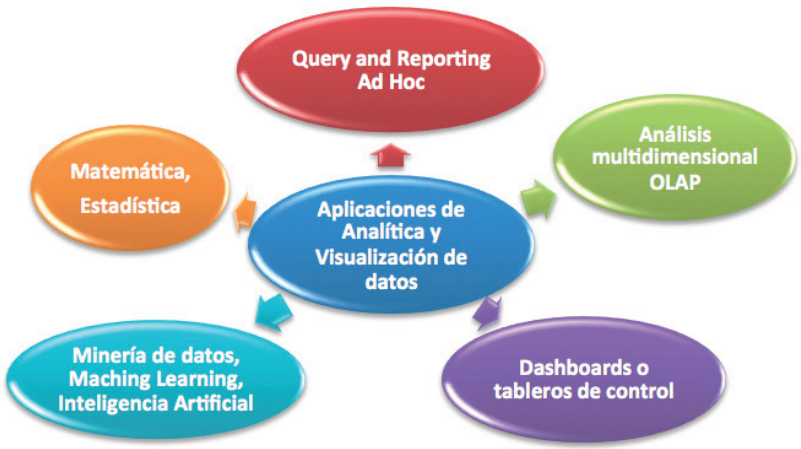
[1] https://www.denodo.com/en
[2] https://www.informatica.com
[3] http://www.compositesw.com/products-services/data-virtualization-platform/

Consultas o Reportes Ad Hoc. Consiste en el diseño de consultas o reportes personalizados para resolver un problema específico sin posibilidad de generalizar.

Análisis multidimensional OLAP. Mediante visores de cubos OLAP, es posible generar consultas dinámicas de tipo MDX. Es decir, el usuario puede elegir los KPI's o medidas y dimensiones a visualizar. Las operaciones OLAP que se pueden realizar son:

- Roll-up (Agregación). Eliminación de un criterio de agrupación para el análisis.
- Drill-down (Disgregación). Introducción de un nuevo criterio de agrupación, disgregando los grupos actuales.
- Slice and dice. Consiste en seleccionar y proyectar datos en una consulta.
- Pivote. Rotar la visualización de los datos, transformando filas en columnas.

Imagen 7.6. Esquema de las aplicaciones BI




Dashboard (Tablero de control). Son interfaces visuales que resumen la información del negocio mediante los indicadores clave de desempeño (KPI) utilizando gráficos estadísticos, valores escalares, medidores, semáforos, etc.

Minería de datos (DM: Data Mining). Es el proceso de exploración mediante técnicas descriptivas y predictivas que permiten descubrir un conocimiento oculto (patrones) a partir un conjunto de datos (data sets) o bases de datos (KDD: Knowledge Discovery in Databases). DM aplica métodos y técnicas de estadística, inteligencia artificial, y aprendizaje automático.

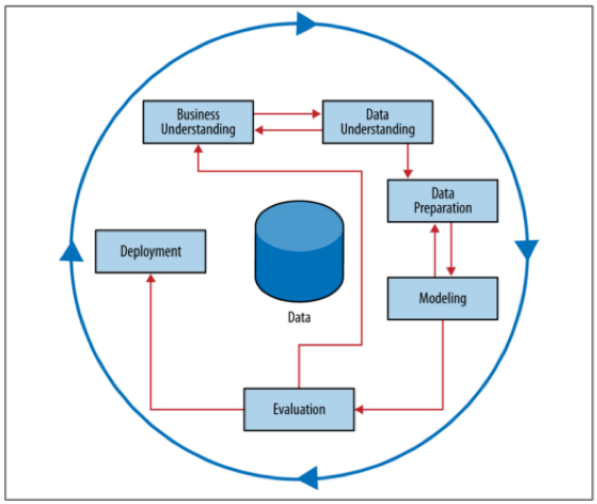
## Metodologías para el desarrollo de sistemas BI y DM

Existen varias metodologías que proponen un ciclo de vida para crear soluciones BI y/o DM; entre las más destacadas se encuentran: Imon, Kimball, HEFESTO, CRISP-DM.

### Metodología CRISP-DM (Cross Industry Standard Process for Data Mining)

Esta metodología es un estándar de la industria de la minería de datos y aplicaciones BI que describe un proceso organizado por fases para llevar a cabo un proyecto de minería de datos. Las actividades de esta metodología se observan en la Imagen 7.7 y se describen a continuación:

Imagen 7.7. Fases de la Metodología CRISP-DM




Fuente: (Provost & Fawcett, 2013)

- Comprensión del negocio. Consiste en realizar actividades como: identificar los requerimientos empresariales, el planteamiento del problema, los objetivos del proyecto de análisis de datos y el establecimiento de un plan de trabajo.
- Comprensión de Datos. En esta fase, se realizan actividades de identificación de los indicadores clave de desempeño, las dimensiones de análisis, familiarización con las estructuras de datos fuentes y verificación de su calidad.
- Preparación de datos. Comprende actividades de limpieza, transformación, formateo, selección e integración de datos. Estas actividades se ejecutan en reiteradas ocasiones hasta cumplir los requerimientos del negocio.
- Modelado. En esta fase se crean los modelos (diseños) de los conjuntos de datos o almacenes de datos en base a los indicadores de desempeño del negocio (KPI's) o requerimientos de análisis de datos. Con el propósito de refinar el diseño, se revisa y se corrige el (los) modelo(s) y si es necesario se vuelve a la fase de Preparación de datos.
- Evaluación. Cada uno de los modelos de datos con perspectiva de análisis, son evaluados según criterios de calidad.
- Despliegue. En esta actividad, se realizan tareas que pueden ser simples o complejas, como la generación de reportes (gráficos o tablas estadísticas) o la implantación de una plataforma de explotación de información que proporcione acceso controlado a los usuarios tomadores de decisiones en toda la organización.

# Caso de estudio: diseño e implementación de un sistema de inteligencia de negocios aplicado al sector camaronero

## Enunciado del caso de estudio

El consorcio de empresas ABC maneja N empresas que producen camarón; cada empresa tiene a cargo una camaronera que está ubicada en un país, provincia y cantón; Cada camaronera está organizada por piscinas. Cada piscina tiene un número de hectáreas de producción. El proceso de producción de camarón se denomina "corrida", cada corrida tiene un número secuencial, fecha de inicio y fin, estado (concluido o en proceso) y una o más piscinas donde se siembra el camarón. Por cada piscina que es parte de una corrida, se registra el costo de inversión, el número de larvas de camarón sembradas, la cantidad en Kilogramos de camarón cosechados clasificados por talla (Grande, Mediano y Pequeño) y el total de ingresos por la venta de la cosecha ($ por venta total y por talla). La empresa puede tener una o más corridas activas, pero una piscina no puede estar en más de una corrida activa en el mismo periodo de tiempo.

Aplicando un proceso metodológico de inteligencia de negocios, realizar el diseño de un data warehouse que satisfaga los requerimientos planteados; luego, el proceso ETL con datos simulados desde Excel; a continuación, la creación de Cubos OLAP; y finalmente, el diseño e implementación de un dashboard BI como aplicación EIS. Aplicar el proceso metodológico de CRISP-DM. En este caso se omite la actividad de Evaluación.

## Comprensión del negocio

### Requerimientos del negocio (preguntas del negocio)

1. ¿Cuántos Kilogramos de camarón se han producido por talla y empresa en el último año?

2. ¿Cuáles son los gastos (inversión), los ingresos y la utilidad por empresa en un año determinado?

3. ¿Cuántas hectáreas de producción de camarón tiene el consorcio por empresa, corridas activa y piscina en el último año?

4. ¿Cuántas hectáreas de producción de camarón dispone el consorcio por empresa, piscina y país?

5. ¿Cuál es el número promedio de larvas de camarón por hectárea que se siembran por empresa y camaronera?

6. ¿Cuál es la empresa que ha generado más utilidades por año?

7. ¿Cuál es la empresa que ha producido más camarón por talla y año?

Las preguntas se han obtenido en base a entrevistas aplicadas a algunos productores de camarón. Se ha resaltado el texto en color azul a aquellos posibles indicadores de desempeño o medidas del negocio, se ha subrayado las potenciales dimensiones, el color verde denota posible filtro aplicar a una dimensión y el color naranja denota sentido de orden.

### Planteamiento del problema

Necesidad de diseñar una solución de inteligencia de negocios para el proceso de producción en un consorcio camaronero.

### Objetivo

Diseñar una solución de inteligencia de negocios (Dashboard EIS) para el proceso de producción en un consorcio camaronero, empleando técnicas de análisis multidimensional OLAP y herramientas que permiten el diseño de un data warehouse, la ejecución del proceso ETL y la implementación de un dashboard BI (panel de control de los principales KPI's), con el propósito de mantener oportunamente informados a los ejecutivos que toman decisiones.

### Plan de trabajo

El plan de trabajo para este caso de estudio, se divide en actividades, tareas, recursos, responsables y distribución del tiempo y costos según la metodología CRISP-DM. Por cuestión de limitaciones de espacio no se lo incluye en este texto.

## Comprensión de los datos

En el Cuadro 7.1 se identifican los principales indicadores clave de desempeño (KPI's) con sus respectivas dimensiones.

Cuadro 7.1: Identificación de indicadores clave de rendimiento (KPI) y dimensiones

| Data Mart | KPI | Código KPI | Dimensión |
|---|---|---|---|
| Estadística de piscinas | Número de hectáreas de producción de camarón | pis_hectareas | estado, empresa, país, periodo (año, semestre), piscina |
| Estadística de corridas, piscinas y siembra | Número de larvas de camarón sembradas por hectárea | cp_num_larvas | empresa, camaronera, país |
| Estadística de cosechas | Número de kilogramos de camarón cosechado | cose_num_kilos | talla, empresa, periodo (año, semestre) |
| | Gastos de producción (inversión) | cose_inversion | |
| | Total de ingresos por ventas | cose_total | |
| | Utilidad | cose_utilidad | |

Fuente: Elaboración propia

En el Cuadro 7.2 se preparan los data marts con sus tablas de hechos, KPI's y funciones de agregación o fórmulas que se aplicarán a los datos. Las posibles funciones de agregación que se pueden aplicar son: sumarización (SUM), conteo (COUNT), promedio (AVG), mínimo (MIN), máximo (MAX). También es posible aplicar fórmulas matemáticas que combinan operadores y/o funciones de agregación.

Cuadro 7.2: Funciones de agregación por indicador clave de desempeño

| Data Mart | Tabla de hecho | Función de agregación o fórmula por KPI |
|---|---|---|
| Estadística de piscinas | th_piscina | SUM (pis_hectareas) |
| Estadística de corridas, piscinas y siembra | th_corrida_piscina | AVG(cp_num_larvas) |
| Estadística de cosechas | th_cosecha | SUM(cose_num_kilos)<br>SUM(cose_inversion)<br>SUM(cose_total)<br>SUM (cose_utilidad) = [ SUM(cose_total) - SUM(cose_inversion)] |

Fuente: Elaboración propia

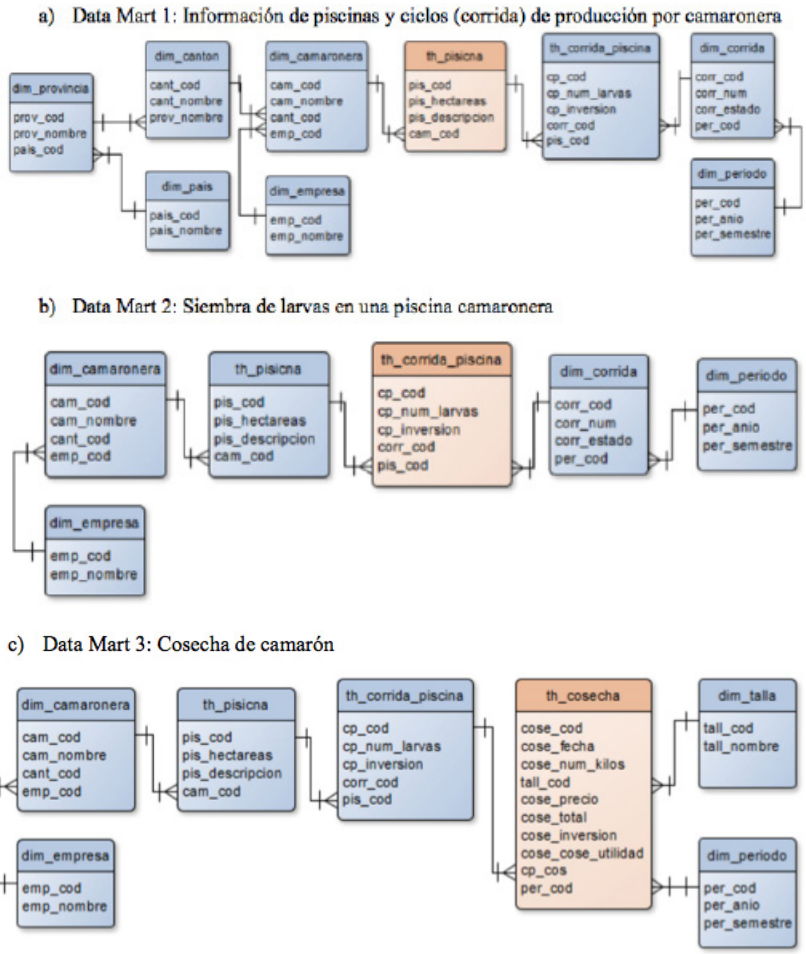
## Preparación de datos

En esta fase lo ideal es tener acceso a datos reales de una empresa como una base de datos de un sistema transaccional. Debido a que se está tratando con un caso de estudio simulado, se trabajó con datos en una hoja de cálculo.

## Modelado

En la Imagen 7.8, se muestran los modelos lógicos de los data marts que conforman la data warehouse según el caso de estudio de control de producción del consorcio camaronero.

Imagen 7.8. Modelos lógicos de data marts: a) Información de piscinas y proceso productivo (corrida) por camaronera, b) Siembra de larvas en una piscina y c) Cosecha de camarón.




Fuente: Elaboración propia

## Despliegue. Implementación de un dashboard en Power BI

### Descarga e instalación de Power BI Desktop

Un dashboard o tablero de control es una herramienta que representa de manera gráfica los indicadores claves de desempeño (KPI) de una organización. En este caso se utilizó la herramienta Microsoft Power BI Desktop versión freware, debido a que ofrece un conjunto de herramientas de inteligencia de negocios para el análisis de datos destinada a usuarios empresariales, profesionales de TI y desarrolladores.

Para su descarga, acceder al link en la página oficial del instalador PBIDesktop_x64.msi, https://go.microsoft.com/fwlink/?LinkId=521662&clcid=0x40a.

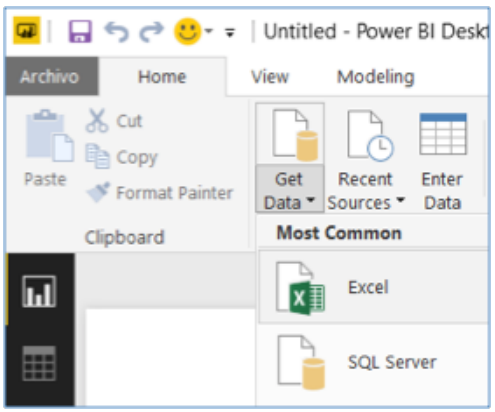
**Datos fuente.** Los datos fuente utilizados para este caso de estudio, se pueden descargar del siguiente enlace:

https://drive.google.com/file/d/1HYEuTuc_yU42g4y-ZpH7Ms1j3Ghw9o2fH/view?usp=sharing

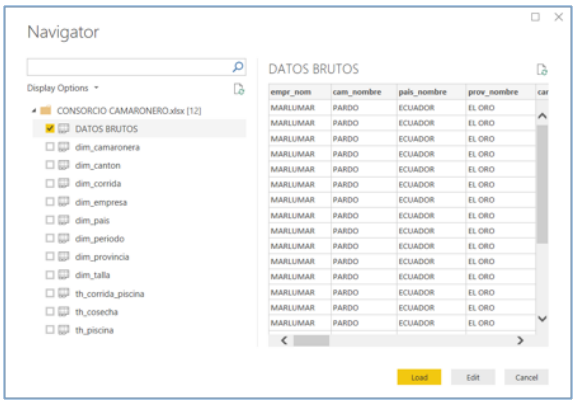
### Creación de cubos OLAP y gráficos

Primero se ingresar al entorno de trabajo de Power BI Desktop y se seguir los pasos del Cuadro 7.3.

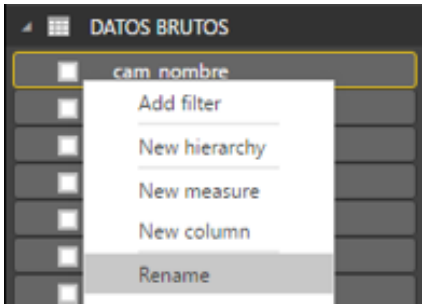
Cuadro 7.3: Pasos para la creación de gráficos estadísticos

| | |
|---|---|
| 1. Dar clic en Get Data, a continuación en Excel y seleccionar el archivo con los datos fuentes. |  |

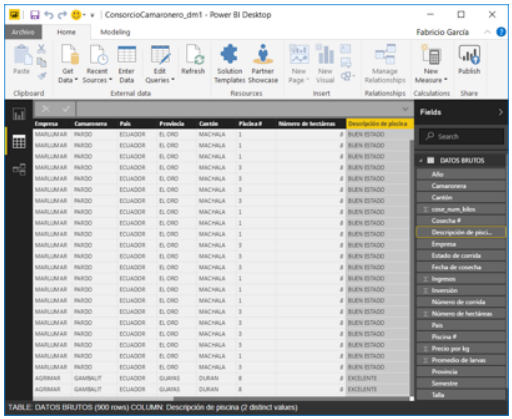
2. Seleccionar las medidas y dimensiones a utilizar, en este caso se asume que se dispone de datos condensados en una sola hoja de cálculo denominada "DATOS BRUTOS".



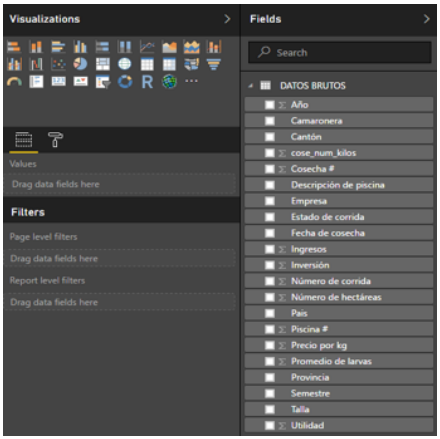

3. En el lado derecho de la pantalla se cargan todas las dimensiones y medidas, las cuales se pueden renombrar dando clic derecho.



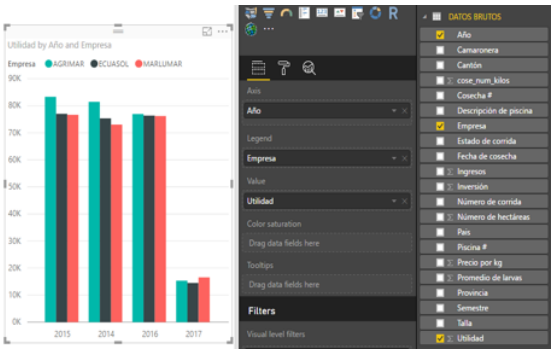

4. En la parte izquierda se pueden observar los datos seleccionando la opción Data. En la opción de Relaciones se pueden observar las relaciones de las tablas en caso de haber importado de una base de datos.



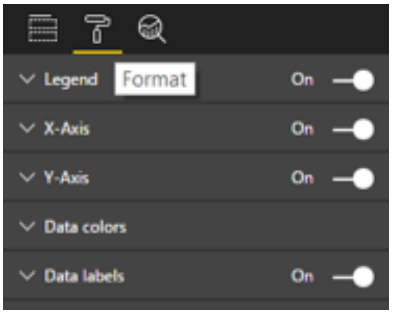
5. Para diseñar los gráficos estadísticos de acuerdo a las preguntas de negocio de cada data mart identificado, seleccionar el tipo de gráfico y luego las medidas y dimensiones que se deseen en la parte derecha de la ventana.



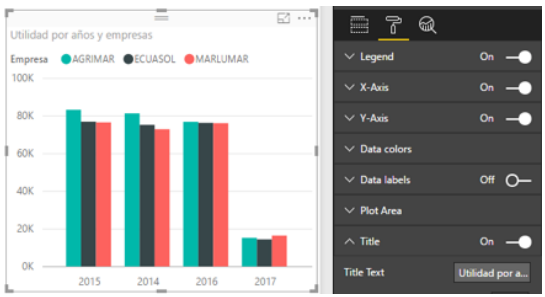

6. Una vez seleccionadas las medidas y dimensiones, el gráfico se va visualizando automáticamente, las medidas se añadirán a las sección de Valores, mientras que las dimensiones pueden ubicarse en Ejes o Leyenda, esto dependerá de la forma de visualización del gráfico.



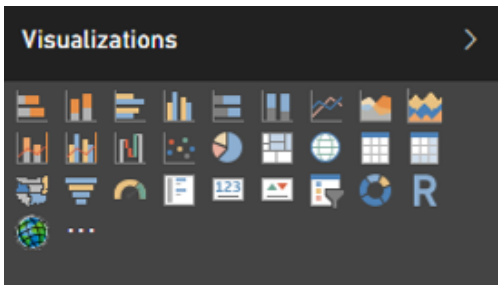

7. Se pueden mostrar los valores en las barras, dirigirse a la pestaña Formato y habilitar opción Etiquetas de datos.



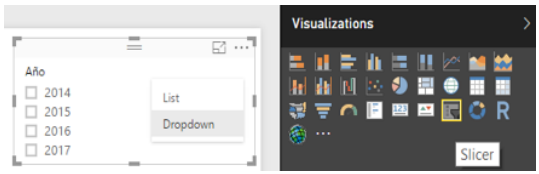
8. Para cambiar el título al gráfico ubicarse en Título de la pestaña Formato, se modifica el texto del título y listo.




9. Se puede cambiar el estilo del gráfico seleccionando otro diseño en Visualizaciones.




10. Para ubicar filtros se debe arrastrar la dimensión que se desee hacia un espacio en blanco del área de trabajo y seleccionar el ícono Slicer.



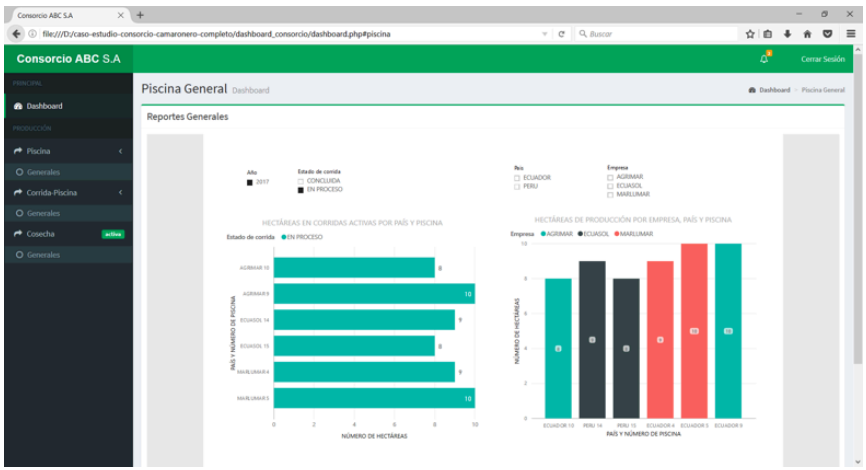

## Dashboards BI

Un dashboard es un tablero de control que permite el manejo integrado de la información importante de la empresa que es útil para la toma de decisiones. En este caso de uso, primero se diseñaron los gráficos estadísticos y luego se publicaron en internet en la cloud de Power BI mediante una cuenta que se puede crear gratuitamente; sin embargo, esta cuenta es limitada y, si se requieren de más servicios o prestaciones se debe cancelar un valor mensual o anual. Para darle un estilo personalizado de aplicación web, se creó

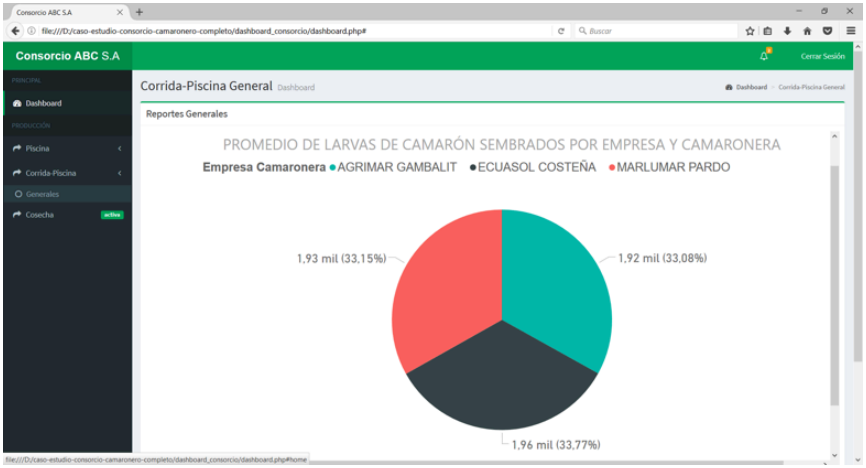
el menú de opciones mediante código HTML y se embebieron los gráficos de Power BI. En la Imagen 7.9, se visualizan tres capturas de pantalla de la aplicación web correspondiente a los dashboard BI creados.

Imagen 7.9. Capturas de pantalla de la aplicación web (dashboar BI) con estadísticas de producción de camarón de un consorcio camaronero. En a) estadística de piscinas, en b) estadísticas de siembra y en b) estadísticas de cosecha.

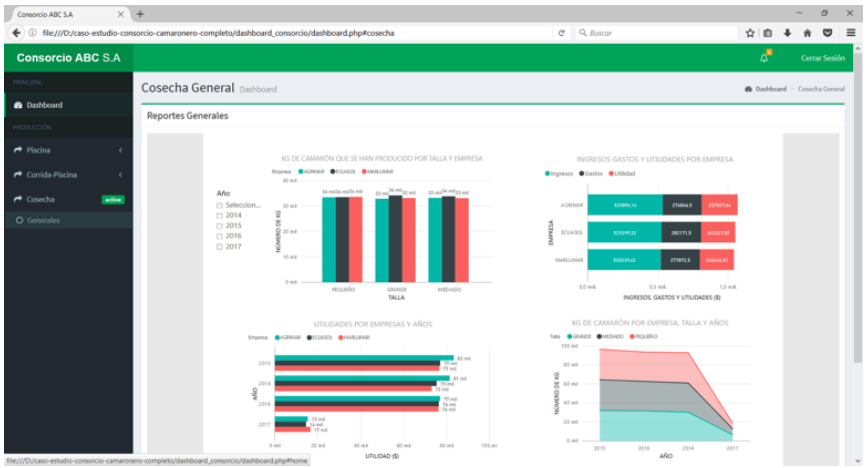
a) Estadística de piscinas




b) Estadística de siembra



c) Estadística de cosecha




En esta capítulo se presentó un panorama de lo que se puede hacer con la Inteligencia de Negocios en el sector agropecuario. Primero se explican los fundamentos teóricos de BI y mediante un caso de estudio enfocado en la producción de camarón, se evidencia una un ejemplo sencillo de dashboard BI.

Para el caso de estudio se utilizó datos en bruto simulados con el fin de explicar el proceso de construcción de una aplicación BI. El proceso metodológico se llevó a cabo mediante la metodología CRISP-DM. En la fase de Comprensión del Negocio se describió los requerimientos del negocio, mediante una serie de preguntas con una estructura particular (medida o KPI, dimensiones, filtros y sentido de orden). En la fase de Comprensión de los Datos, se determinó los principales data marts con sus respectivas medidas o indicadores clave de rendimiento (KPIs) y dimensiones. En la fase de Preparación de Datos, se creó el modelo de los data marts. Y en la fase de Despliegue, se construyó la aplicación BI mediante la herramienta Microsoft Power BI Desktop, que integra tres dashboards: a) estadística de piscinas, b) estadísticas de siembra b) estadísticas de cosecha.

## Referencia Bibliográfica


Cornejo, R., Navarrete, M., Valdivia, R., Aroca, P., & Aracena, S. (2014). Desarrollo de una base de datos integrada de Censo y encuesta mediante el uso de elementos de inteligencia de negocios y SIG. *Ingeniare. Revista Chilena de Ingeniería,* 22, 205–217. http://doi.org/10.4067/S0718-33052014000200007

De Mauro, A., Greco, M., & Grimaldim, M. (2015). What is Big Data ? A Consensual Definition and a Review of Key Research Topics. *In International Conference on Integrated Information (IC-ININFO 2014)* (Vol. 1644, pp. 97–104). http://doi.org/10.1063/1.4907823

FAO. (2016). *Programa mundial del censo agropecuario 2020. Volumen 1. Programa, definiciones y conceptos.* Retrieved from http://www.fao.org/3/a-i4913s.pdf

Ghosh, R., Halder, S., & Sen, S. (2015). An Integrated Approach to Deploy Data Warehouse in Business Intelligence Environment. *In Proceedings of the 2015 Third International Conference on Computer, Communication, Control and Information Technology* (C3IT) (p. 7). http://doi.org/10.1109/C3IT.2015.7060115

Gounder, M. S., Iyer, V. V., Professor-ccis, A., Mazyad, A. Al, & Prof, A. (2016). A Survey on Business Intelligence tools for University Dashboard development. *2016 3rd MEC International Conference on Big Data and Smart City.*

Laudon, K. C., & Laudon, J. P. (2012). *Sistemas De Información Gerencial.* (Pearson, Ed.) (12 Edición). México: Pearson Education.

Marinheiro, A., & Bernardino, J. (2015). Experimental Evaluation of Open Source Business Intelligence Suites using OpenBRR, 13(3), 810–817.

Mazon-Olivo, B., Rivas, W., Pinta, M., Mosquera, A., Astudillo, L., & Gallegos, H. (2017). Dashboard para el soporte de decisiones en una empresa del sector minero. *Conference Proceedings - Universidad Técnica de Machala, 1*, 1218–1229. Retrieved from http://investigacion.utmachala.edu.ec/proceedings/index.php/utmach/article/view/219/191

Moniruzzaman, A., & Hossain, S. (2013). Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. *International Journal of Database Theory and*

*Application, 6*(4), 1–14. Retrieved from http://www.sersc.org/journals/IJDTA/vol6_no4/1.pdf

Morales, A., Cuevas, R., & Martínez, J. (2016). Procesamiento Analítico con Minería de Datos. *Revista Iberoamericana de Las Ciencias Computacionales E Informática, 5*.

Mosquera, L., & Hallo, M. (2014). Data Mart Para El Sistema De Servicios Sociales Del Conadis. *Revista Politécnica,* 33(2).

NIST. (2015). NIST Special Publication 1500-1 NIST. Big Data Interoperability Framework : Volume 1 , Definitions. *National Institute of Standards and Technology, 1,* 32. http://doi.org/10.6028/NIST.SP.1500-1

Provost, F., & Fawcett, T. (2013). *Data Science for Business. What you need to know about Data Mining and Data-Analytic thinking*. O'Reilly Media.

Rosado, A., & Rico, D. (2010). Business Intelligence :State of the Art. *Scientia Et Technica, XVI*(44), 321–326. Retrieved from http://www.redalyc.org/articulo.oa?id=84917316060

Rosado G., A. A., & Rico B., D. W. (2010). Inteligencia de Negocios: Estado del Arte. *Scientia Et Technica, XVI,* 321–326. Retrieved from http://www.redalyc.org/articulo.oa?id=84917316060

Sawant, N., & Shah, H. (2013). *Big Data Application Architecture A&A. Apress*. http://doi.org/10.1007/978-1-4302-6293-0

Ţăranu, I. (2015). Big Data Analytics Platforms analyze from startups to traditional database players. *Database Systems Journal, VI(1),* 23–32. Retrieved from http://www.dbjournal.ro/archive/19/19_3.pdf

Van Der Lans, R. F. (2012). *Data Virtualization for Business Intelligence Systems. Revolutionizing Data Integration for Data Warehouses. Data Virtualization for Business Intelligence Systems*. Elsevier Inc. http://doi.org/10.1016/B978-0-12-394425-2.00014-9

Vassell, M., Apperson, O., Calyam, P., Gillis, J., & Ahmad, S. (2016). Intelligent Dashboard for Augmented Reality based Incident Command Response Co-ordination. *IEEE Consumer Communications and Networking Conference,* 0–3. http://doi.org/10.1109/CCNC.2016.7444921

White, T. (2015). *Hadoop: The Definitive Guide, 4th Edition*. O'Reilly (Vol. 54).

# 08 Capítulo

# Inteligencia Artificial aplicada a datos agropecuarios

Iván Ramírez-Morales, Eduardo Tusa; Daniel Rivero

La inteligencia artificial (IA), podría sonar aún como un término de ciencia ficción, sin embargo muchas personas no se percatan de que están utilizándose cada vez más en actividades de la vida cotidiana. Los asistentes personales como

**Iván Ramírez-Morales:** Doctor en Medicina Veterinaria y Zootecnia por la Universidad Agraria de la Habana, Máster en Desarrollo Comunitario por la Universidad Nacional de Loja y está finalizando su Doctorado en TIC por la Universidade A Coruña, ha realizado varios cursos en Brasil, Japón, Perú y Argentina. Fué Oficial de Territorio del Programa Marco ART/PNUD de la ONU, y Director de Planificación del Gobierno Provincial de El Oro. Actualmente es Profesor Titular en la Universidad Técnica de Machala, su área de investigación se centra en el uso de tecnologías para el mejoramiento de la productividad agropecuaria, Cuenta a la fecha más de 10 publicaciones indexadas, varias de ellas en revistas de alto impacto en los índices de JCR y SJR.

**Eduardo Tusa:** Ingeniero Electrónico (Magna Cum Laude) con una Subespecialización en Matemáticas de la Universidad San Francisco de Quito. Su cuarto año de formación de pregrado fue realizado en la Universidad de Illinois en Urbana - Champaign, USA. Máster en Visión, Imagen y Robótica (con distinción) de la Universidad de Borgoña (Francia), la Universidad de Girona (España) y la Universidad Heriot-Watt (Reino Unido). Es docente de la Unidad Académica de Ingeniería Civil de la Universidad Técnica de Machala, donde ha impartido las asignaturas de Programación en MATLAB, Informática, Nuevas Tecnologías de la Información y Comunicación, Cálculo Integral, Ecuaciones Diferenciales, Matemática Avanzada, Probabilidad y Estadístic

**Daniel Rivero:** Ingeniero en Informática por la Universidad de A Coruña, y Doctor en el área de conocimiento en Ciencias de la Información e Inteligencia Artificial. Trabaja como Profesor Contratado Doctor en el Departamento de Tecnologías de la Información y las Comunicaciones de la citada Universidad. Su área de investigación incluye las Redes de Neuronas Artificiales, Computación Evolutiva, y en general la aplicación de técnicas de Machine Learning en distintos entornos. Fruto de este trabajo de investigación, ha publicado una gran cantidad de artículos en distintas revistas indexadas con índice de impacto JCR, así como distintos libros, como autor, editor o autor de capítulos, y comunicaciones a congresos. Por otra parte, ha colaborado también en un gran número de proyectos de investigación financiados de forma competitiva.

Siri o Cortana, los vehículos autónomos, predicción de fraudes, o de condiciones idóneas de mercado, recomendaciones sobre tendencias, son entre otras aplicaciones de uso común.

Una rama de la inteligencia artificial conocida como aprendizaje automático o aprendizaje máquina (machine learning - ML), se refiere a los algoritmos computacionales que son capaces de realizar acciones complejas, sin que éstos hayan sido explícitamente programados para ello, si no que estos algoritmos son más bien, entrenados para realizar esta tarea.

En tareas muy complejas, esta particularidad es indispensable para que un computador sea capaz de realizar una tarea; por ejemplo, programar un juego como el ajedrez implica una complejidad que fue calculada por Shannon (1950) como $10^{120}$, como punto de comparación, se dice que todos los átomos del universo son $6^{79}$, en estos casos es mejor utilizar algoritmos que aprenden a partir de ejemplos.

Los algoritmos de aprendizaje máquina tienen la capacidad de generalizar un comportamiento de respuesta a partir de una información suministrada en forma de ejemplos, durante el proceso de entrenamiento.

Actualmente se está utilizando algoritmos de aprendizaje profundo que permite a las máquinas aprender de una manera muy similar a como lo hacen los humano.

Aunque estas tecnologías no están alejadas de la potencial aplicación en el sector agropecuario, existe todavía un rezago en cuanto a su uso por parte de los profesionales de este importante sector de la economía. En referencia al objetivo principal de este libro, en este capítulo se explora la aplicabilidad de algunas técnicas de IA enfocadas al análisis de información de ámbito agropecuario.

Para los profesionales que se desenvuelven en el ámbito agropecuario, es común utilizar estadísticas descriptivas para el procesamiento de sus datos. Estas técnicas permiten una adecuada comprensión de la información existente en el dato, sin embargo no siempre son capaces de extraer con

suficiente exhaustividad, la información relevante que apoye a la toma de decisiones por parte de los administradores y técnicos de producción.

En las fincas agropecuarias cada día se genera una gran cantidad de información, aunque por la naturaleza misma de los medios con los que se obtiene, esta información generalmente tiene mucho ruido, es decir, datos erróneos, o irregulares, que pueden enmascarar el conocimiento contenido en la relación de la información. Es por ello que se hace necesario utilizar nuevas técnicas de análisis bajo un enfoque de aprendizaje automático.

## Tipos de aprendizaje automático

### Aprendizaje no supervisado

El aprendizaje automático no supervisado, consiste en asignar una máquina la tarea de inferir una función que describa la estructura oculta de los datos, dado que éstos no han sido previamente etiquetados. En este caso no se cuenta con la posibilidad de evaluar fácilmente  la exactitud del resultado de la función inferida.

En este tipo de algoritmos, la salida se asocia con el grado de similitud entre las características de entrada, es decir que el aprendizaje se centra en las asociaciones que ocurren en un conjunto de datos tratando de encontrar cualquier tipo de regularidad en los datos.

Estas técnicas suelen ser utilizadas para agrupar datos según su criterio de similitud. Además son muy utilizadas para visualización de datos ya que permiten reducir a dos o tres dimensiones, datos multidimensionales. Precisamente por esta propiedad, los algoritmos no supervisados suelen ser utilizados para extracción de características previo al entrenamiento con alguno de los algoritmos de aprendizaje supervisado.

## Aprendizaje supervisado

El aprendizaje supervisado consiste en el descubrimiento de patrones válidos a partir de conjuntos de datos de entrenamiento que han sido previamente etiquetados. En el aprendizaje supervisado, cada ejemplo tiene un un objeto de entrada y un valor de salida deseada.

Un algoritmo de aprendizaje supervisado analiza los datos de entrenamiento y produce una función inferida, que puede ser utilizado para el mapeo de nuevos ejemplos. Una correcta selección de ejemplos, permitirá el algoritmo para determinar correctamente las etiquetas de clase para nuevas instancias. Esta capacidad de inferir la clase de datos nuevos, se conoce como generalización.

Para entrenar un algoritmo con técnicas de aprendizaje supervisado, es necesario en primer lugar identificar el conjunto de datos para el entrenamiento. Este tiene que ser representativo del universo de datos y debe haber sido etiquetado y revisado por expertos en el área.

La precisión va depender en gran medida de las características del vector de entrada, estas características deben contener suficiente información sobre el patrón de entrada para que sea capaz de predecir con precisión la salida deseada. Debido a un efecto que se denomina la “maldición de la multidimensionalidad”, el vector de entrada no debe tener demasiadas características.

Los algoritmos de aprendizaje automático suelen tener varios parámetros que deben ser ajustados durante el proceso de entrenamiento, estos paŕametros permiten modelar de mejor manera y elevan la precisión de la función aprendida.

Existe una gran variedad de algoritmos de aprendizaje supervisado. No existe uno que sea válido para todos los problemas, cada uno tiene sus particularidades. La selección del algoritmo idóneo se realiza habitualmente en un proceso que es empírico y requiere de muchas pruebas cuyo resultado final es la optimización del modelo.

Los problemas que se abordan con aprendizaje supervisado suelen ser generalmente de dos tipos: clasificación y regresión. Es importante diferenciar ambos tipos, ya que su comprensión es básica para entender el funcionamiento de las técnicas. La clasificación consiste en la asignación de una clase o tipo de acuerdo a sus características; por ejemplo de qué raza es un animal, o de qué variedad es una planta. La regresión por otro lado, busca predecir un valor cuantificable, por ejemplo, cuántos litros de leche producirá una vaca, o cuántos quintales por hectárea se obtendrán de una parcela. Como se puede apreciar, la clasificación tiene el objetivo de predecir valores discretos, mientras que la regresión predice valores continuos.

## Técnicas de ML más utilizadas

En este apartado se realizará una breve descripción de las técnicas más comunes utilizadas en aprendizaje automático. Además se presentan varios ejemplos de aplicaciones reales en las que los autores han implementado estas técnicas, así como otros ejemplos ilustrativos llevados a cabo por otros autores.

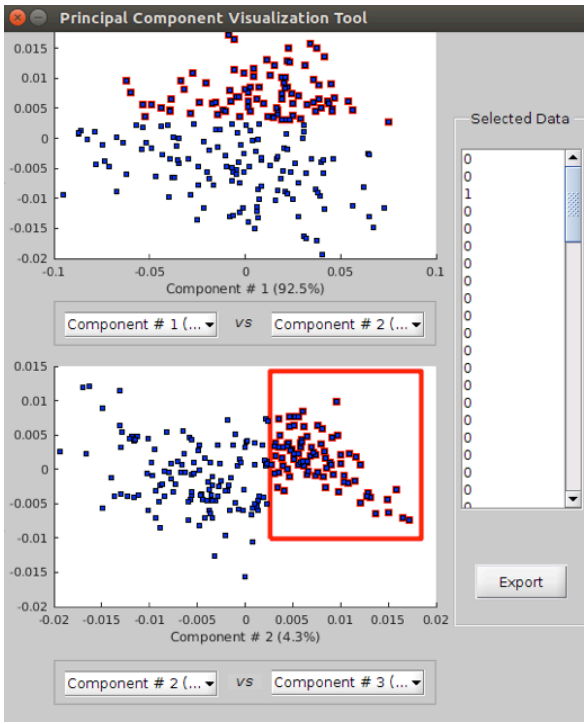
### Análisis de componentes principales

La técnica de análisis de componentes principales (Principal Component Analysis - PCA) es una técnica de aprendizaje no supervisado. Es comúnmente utilizada para reducir la dimensión de un conjunto de datos. La variabilidad de los datos se conserva y el número de variables o características se reduce.

Con esta técnica se busca la mejor representación de los mínimos cuadrados de los datos, de esta manera cuando en un patrón de datos existen algunas variables posiblemente correlacionadas, el algoritmo devuelve un nuevo conjunto de valores que no tienen una correlación lineal entre sí, a estos valores se les llama componentes principales.

El análisis de componentes principales, retiene los valores numéricos que explican en mayor proporción la varianza del conjunto de datos, e ignora aquellos que tienen menor influencia en la varianza. Comúnmente los componentes principales contienen lo más importante de la información original, sin embargo se pierde la representación original del dato.

Gráfico 8.1 Gráfico de un análisis de componentes principales para el diagnóstico de mastitis bovina utilizando espectrometría NIR.




En el Gráfico 8.1 se observa un gráfico del componente principal 1 versus el componente principal 2 y otro del componente principal 2 versus el componente principal 3. Se puede apreciar que los datos seleccionados pertenecen a la misma clase 0 (Infección Negativa) mientras que los datos no seleccionados pertenecen a la clase 1 (Infección Positiva). En la imagen también se observa que entre los seleccionados de la clase 0, aparece un dato mal clasificado.

En la agricultura de precisión se cuenta con grandes volúmenes de información georeferenciada, el análisis PCA clásico aplicado a este tipo de datos es capaz de evaluar las propiedades del suelo y el rendimiento del cultivo. De esta manera se detectan correlaciones entre variables que permiten la consolidación y homogeneización de zonas dentro de los lotes. Existen varias aplicaciones de esta técnica que serán abordadas más adelante, en general es una técnica que por su sencillez ha logrado una buena difusión entre la comunidad científica.

## K Vecinos Más Cercanos

También conocida como k-NN por su traducción en inglés (k Nearest Neighbors) Esta es una técnica muy versátil puesto que puede ser utilizada tanto para aprendizaje no supervisado, como para aprendizaje supervisado. En el caso no supervisado, se ha utilizado en clasificación y en regresión.

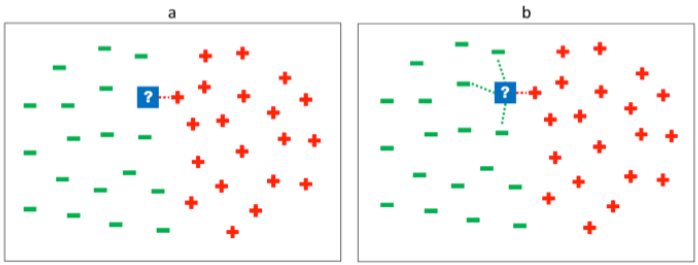
Su aplicación más habitual es en clasificación. Su funcionamiento consiste en asignar la clase a una patrón desconocido, según la clase que tengan los patrones conocidos más cercanos. El número de vecinos k normalmente es ajustado en un proceso de optimización con la finalidad de mejorar la precisión del clasificador, este ajuste especialmente importante cuando la muestra desconocida está rodeada de muestras conocidas que tienen diferentes clases.

Se genera una regla para la clasificación de acuerdo a la que la clase asignada será aquella que tenga la mayor parte de sus k vecinos más próximos.

El Gráfico 8.2 muestra la regla de decisión k-NN para k = 1 (a) y para k = 4 (b). El conjunto de datos ha sido etiquetado de tal manera que la clase negativa corresponde a la ausencia de una enfermedad y la clase positiva a la presencia de esta. En esta figura se ilustra cómo influye en la decisión el número k en la decisión de la asignación de una clase. En el primer caso que se puede observar en el Gráfico 8.2a, la muestra desconocida está fue clasificada con sólo un vecino más cercano, por lo tanto que de acuerdo a esta regla de

decisión, pertenecería a la clase positiva. En el segundo caso que se observa en el Gráfico 8.2b, se utilizan cuatro vecinos del conjunto de entrenamiento para clasificar la misma muestra desconocida, en esta ocasión, tres de los vecinos más cercanos pertenecen a la clase negativa.

Gráfico 8.2 Reglas de decisión en k-NN de acuerdo al número k



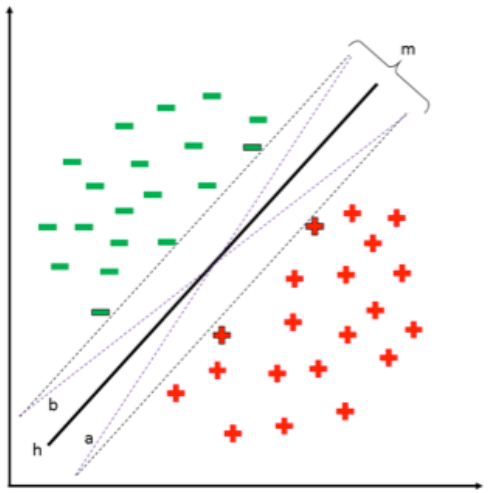

Esta técnica ha sido utilizada en clasificación y regresión debido en una amplia variedad de campos de la ciencia debido a su simplicidad y precisión. Se ha empleado en la predicción de enfermedades, pronóstico del clima, en la detección de deficiencias de nutrientes, entre otras aplicaciones.

## Máquinas de Soporte Vectorial

Las Máquinas de Soporte Vectorial (Support Vector Machine - SVM), están entre las técnicas más utilizadas en aprendizaje máquina. Se relacionan principalmente con la resolución de problemas de clasificación y regresión. Los principios de las SVM fueron desarrollados por Vapnik y colaboradores (1997). El enfoque original estaba dirigido a resolver problemas de clasificación binaria, sin embargo su aplicación se ha extendido a tareas de clasificación múltiple, aprendizaje no supervisado y regresión.

Las SVM tratan de obtener modelos que minimicen el riesgo estructural de cometer errores ante datos futuros. Su funcionamiento básico consiste en la separación del conjunto de datos en dos clases distintas gracias a un hiperplano definido en un espacio adecuado.

Gráfico 8.3 Representación del hiperplano y margen óptimos del modelo (m).

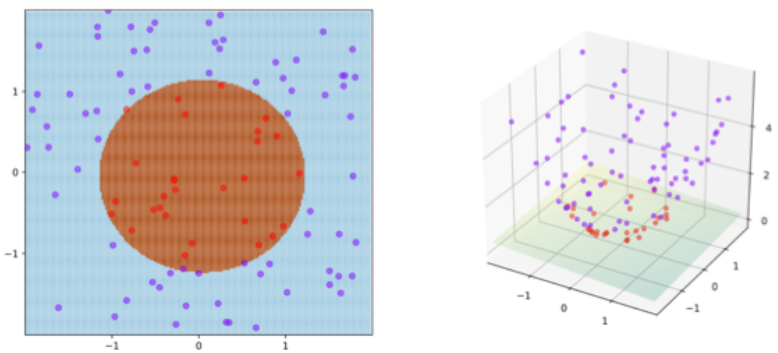



En el Gráfico 8.3 se ilustra los conceptos de hiperplano y margen óptimo del modelo. Como se puede observar en el espacio del margen, pueden existir un sinnúmero de hiperplanos alternativos, en el Gráfico se ilustran dos hiperplanos alternativos posibles (a y b). El hiperplano óptimo usado para separar las dos clases se define a partir de una pequeña cantidad de datos del conjunto de entrenamiento llamados vectores de soporte, que en el gráfico se encuentran sombreados. Estos vectores de soporte son los que determinan el margen del modelo. La elección del mejor hiperplano fue resuelta por Vapnik y Kotz (1982) con el planteamiento de que el hiperplano óptimo es definido como la función de decisión lineal con el máximo margen entre los vectores de soporte de las dos clases.

Sin embargo, en la mayoría de problemas del mundo real, los datos no son linealmente separables y por este motivo es necesario recurrir a estrategias como la identificación de otras dimensiones de separación. Las funciones kernel, son utilizadas para transformar el espacio original multidimensional, en otro espacio en el que las clases sean linealmente separables. En la práctica, las máquinas de soporte vectorial son entrenadas usando distintos kernels para seleccionar

aquel que tenga el mejor desempeño para el problema planteado. Entre los kernel más utilizados están el polinomial y el gaussiano (función de base radial), éste último cuenta con un parámetro sigma (σ) que ajusta el tamaño del kernel.

Gráfico 8.4 SVM con un kernel gaussiano φ ((a, b)) = (a, b, a2 + b2) (Shiyu, Nov, 13, 2016)



En el Gráfico 8.4 se observa cómo los datos de entrenamiento se trasladan a un nuevo espacio de 3 dimensiones en el que un hiperplano es capaz de separarlos linealmente con mayor facilidad.

La búsqueda de parámetros óptimos de una SVM es fundamental en la construcción de un modelo de predicción para que sea preciso y estable. Los parámetros del kernel son ajustables en las SVM para controlar la complejidad de la hipótesis resultante y evitar el sobreajuste del modelo.

Las SVM también pueden ser utilizadas en problemas regresión, esta versión fue propuesta por Vapnik, Golowich y Smola (1997). El método se llama Support Vector Regression (SVR). En este caso, el modelo depende únicamente de los vectores de soporte, ya que la función de pérdida para la construcción del modelo no considera los puntos que se encuentren fuera del margen, asimismo la función de pérdida ignora cualquier los datos que estén cerca al modelo de predicción, dentro de un umbral ε.

Las SVM se han aplicado en varios campos como series temporales, finanzas, aproximaciones de ingeniería, programación cuadrática convexa, clasificación binaria, regresión multivariada, entre otros.

## Redes de neuronas artificiales

Las Redes de Neuronas Artificiales (Artificial Neural Network - ANN) están inspiradas en el funcionamiento del sistema nervioso de los animales, cuyas redes de neuronas biológicas poseen bajas capacidades de procesamiento de forma individual, sin embargo su capacidad cognitiva se sustenta en la conectividad de éstas. De modo similar al mecanismo biológico, las ANN son capaces de realizar tareas complejas de clasificación, identificación, diagnóstico, optimización y predicción, gracias a la conectividad de unidades de procesamiento sencillas.

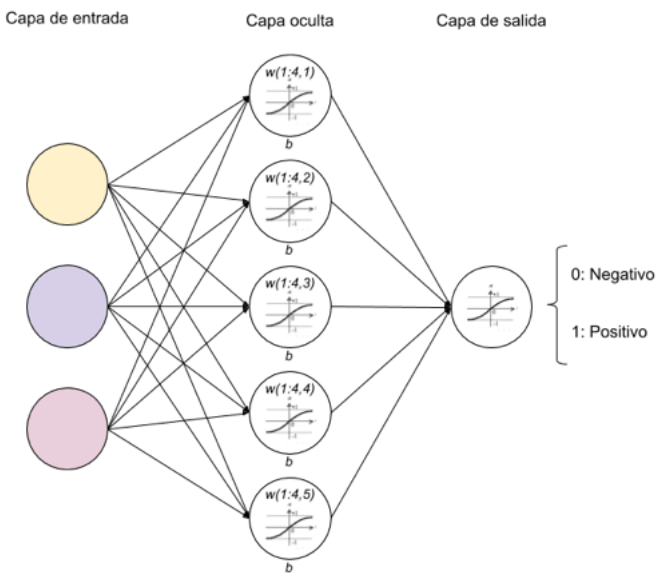
Las ANNs son algoritmos tanto de aprendizaje no supervisado, como de aprendizaje supervisado. Se pueden utilizar para agrupamiento, clasificación y regresión. La organización de las neuronas, permite aprender de los patrones y generalizar hacia nuevas entradas de datos.

Las redes de neuronas artificiales han atraído especial atención en los últimos años, sin embargo fueron McCulloch and Pitts, (1943) quienes presentaron el primer modelo de neurona artificial. Se plantea que las redes neuronales de múltiples capas ocultas, son capaces de aproximar cualquier función medible, por lo que se las considera aproximadores universales.

El perceptrón multicapa (Multilayer Perceptron - MLP) es un tipo de ANN cuyas neuronas están organizadas en capas. Las conexiones en esta red se realizan únicamente entre capas consecutivas. De manera general un MLP tiene una capa de entrada, una o múltiples capas ocultas y una capa de salida. La función de transferencia en las neuronas de la capa oculta y de la capa de salida usualmente es una sigmoidea, sin embargo, pueden estar presentes otras funciones como las lineales, las no lineales o las escalonadas.

En el Gráfico 8.5 se muestra una estructura característica del MLP, los patrones de entrada se proporcionan a la red a través de una capa que simplemente envía esta información a la siguiente capa. El procesamiento y la extracción de la información es realizado en las capas ocultas y en la capa de salida. Cada neurona recibe señales de salida de las neuronas en la capa anterior y envía su señal de salida a las neuronas de la capa siguiente. La capa de salida, recibe las entradas de las neuronas y de acuerdo a un cálculo probabilístico asigna la clase a la que pertenece el ejemplo desconocido. Los MLP pueden ser entrenados tanto para clasificación como para regresión.

Gráfico 8.5 Representación de un MLP con una capa oculta




Uno de los métodos más usados para optimizar el proceso de entrenamiento de un MLP busca localizar el error mínimo utilizando una técnica de gradiente descendiente. En primer lugar se inicializan con valores aleatorios los pesos y los bias de las neuronas, luego, se determina la dirección de la pendiente más pronunciada (gradiente descendiente), se

modifican los pesos, y se re-calcula el gradiente hasta llegar a un valor mínimo de la función.

Para mejorar el desempeño de una ANN es necesario seleccionar una arquitectura adecuada, esto consiste en determinar el número de capas ocultas, el número de neuronas y la forma como estarán interconectadas. La arquitectura de red va a depender del problema a resolver, y no existe una regla o método que permita decidir cuál es la mejor. Generalmente la selección de la mejor arquitectura, resulta de un proceso empírico, en el que es necesario probar distintas alternativas hasta que se encuentra una que proporcione buenos resultados.

El interés en el uso de las redes neuronales va en aumento gracias a su naturaleza paralela, lo que hace que puedan aumentar su velocidad de cálculo, por este motivo ha sido aplicada en una gran variedad de aplicaciones, entre las que destaca la predicción de precios futuros de productos exportables, estimación de la humedad del suelo, la predicción de los rendimientos de cultivos, la elaboración de mapas digitales de territorio, entre otras aplicaciones.

## Redes neuronales profundas

También conocidas como Deep Neural Networks (DNN), se distinguen de las redes neuronales comunes por su mayor profundidad, es decir, el número de capas ocultas a través de las cuales pasan los datos en un proceso de múltiples etapas de reconocimiento de patrones.

Las redes neuronales tradicionales se tienen hasta dos capas ocultas. Cuando se tiene más de tres capas, se considera DNN. En las DNN, cada capa se entrena con un conjunto distinto de características generadas como salidas de la capa anterior. Cuanto más se avanza hacia la red neuronal, la más compleja de las características de sus nodos pueden reconocer, ya que se agregan y se recombinan características de la capa anterior.

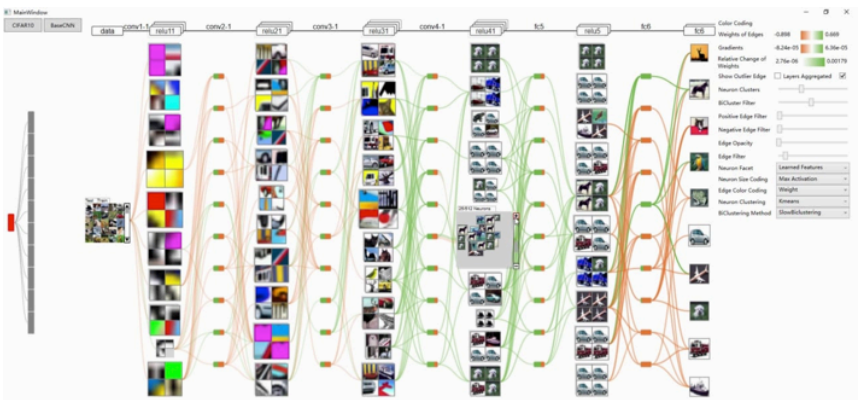
En el Gráfico 8.6 se puede observar una representación de una DNN, en la primera capa oculta se extraen caracterísi-

cas básicas como bordes, en las siguientes se tiene niveles mayores de complejidad como formas, en la siguiente se cuenta con representaciones más precisas del objeto a clasificar o detectar. Este avance creciente en la complejidad y la abstracción se conoce como jerarquía de características.

Las DNN son capaces de manejar conjuntos de datos muy grandes, de muy alta dimensión con miles de millones de parámetros que pasan a través de funciones no lineales.

Algo interesante es que las DNN son capaces de descubrir las estructuras latentes en datos no etiquetados. Un aspecto importante dado que la gran mayoría de los datos en el mundo no tienen etiqueta. Es decir que mediante redes profundas, es posible agrupar de acuerdo a su similitud conjuntos de datos de millones de imágenes, y de esta manera por ejemplo, contar automáticamente con un sistema que agrupe imágenes de plantas sin enfermedades, y distintos grupos según el tipo de enfermedad.

Gráfico 8.6 Representación de una Red Neuronal Profunda (Liu et al, 2017)



## Aplicaciones en el ámbito agropecuario

En el sector agropecuario, la inteligencia artificial tiene un gran potencial. Principalmente por su capacidad para el reconocimiento de patrones. Esta característica permite aplicaciones tales como la clasificación o estimación de paráme-

tros a partir de matrices numéricas, la detección temprana de problemas de producción utilizando series temporales, el análisis de imágenes para clasificación, el análisis de sonidos para detección de enfermedades o el análisis de videos para determinación de patrones de comportamiento.

La gama de posibles aplicaciones en el ámbito agropecuario es variada, en esta sección se realiza una descripción de algunas aplicaciones que han permitido optimizar algún proceso en el ámbito agropecuario con la consecuente mejora de los resultados económicos de las empresas.

## Análisis de señales

En el sector agropecuario es cada vez más común la generación de datos a partir de sensores, estos equipos generan señales que en ocasiones son muy complejas para su análisis manual. Es por esto que varios investigadores han recurrido al uso de las técnicas de aprendizaje automático.

Una experiencia que se desarrolló en la Universidad Técnica de Machala consiste en el desarrollo de un nuevo método para el análisis de mastitis subclínica en el ganado bovino. Este método se basa en el uso de un espectrómetro de reflectancia en el infrarrojo cercano (Near Infrared Reflectance - NIR), aplicado sobre muestras de leche cruda que fueron previamente etiquetadas con la metodología estándar de California Mastitis Test.

Se recogieron un total de 210 muestras de leche en receptores estériles etiquetados individuales. Se obtuvieron muestras de 67 vacas lecheras de raza mixta con 4,3 ± 1,8 años de edad, seleccionadas al azar de cinco granjas de la zona.

En el Gráfico 8.7 se observa las características de los espectrogramas NIR y sus ligeras diferencias que deberán ser analizadas utilizando técnicas de ML. El conjunto de datos estará disponible al público para su análisis una vez que el manuscrito sea publicado.

En el trabajo presentado, los modelos fueron desarrollados utilizando una técnica k-NN cuyo objetivo era detectar

el grado de mastitis. Los resultados de este trabajo muestran una gran potencialidad de la combinación de sensores de bajo costo con técnicas de ML. Los resultados no se detallan debido a que a la fecha de cierre de este libro, el manuscrito está en revisión por una importante revista científica del área.

Gráfico 8.7 Espectrogramas NIR de muestras de leche cruda etiquetadas de acuerdo al grado de mastitis bovina que presentan.

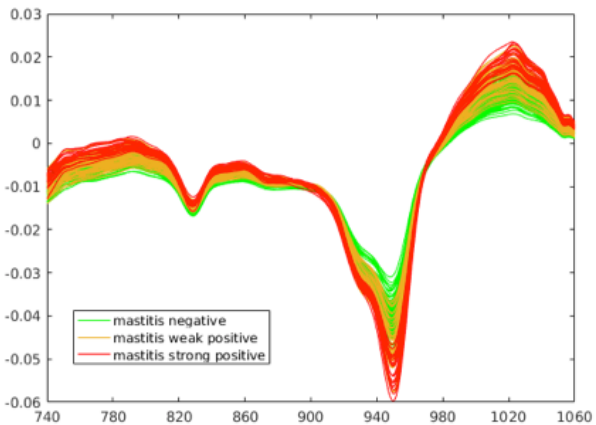
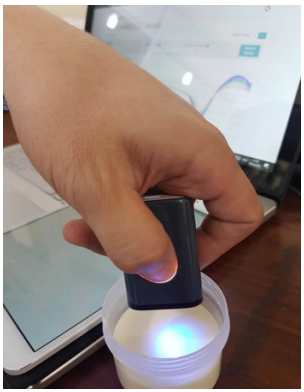



Imagen 8.1 Espectrómetro NIR portátil



El desarrollo de nuevos dispositivos portátiles abre un abanico de posibilidades para su aplicación en el campo agropecuario. En la Imagen 8.1 se puede observar la extracción de señales espectrales en una muestra de leche cruda uti-

lizando el dispositivo SCiO$^z$ desarrollado por la compañía Israelita Consumer Physics. Estos nuevos dispositivos son esencialmente una nueva fuente de señales que requieren ser analizadas utilizando técnicas precisas para la obtención de información relevante.

## Predicción en series temporales

En relación con el análisis de series temporales, su utilización en aplicaciones del sector agropecuario tiene que ver con la predicción de valores futuros y la alerta temprana de problemas.

Las series temporales, se analizan a partir de la reestructura de los patrones de entrada previo al entrenamiento de algoritmos de aprendizaje supervisado. Esto se hace, mediante la utilización de los datos previos como variables de entrada y utilizar un dato del siguiente día como la variable de salida (Kapoor & Bedi, 2013).

Este método se conoce como método de ventana deslizante, consiste en la creación de diferentes secuencias de puntos de datos consecutivos de la serie temporal. Existen dos parámetros en este método: tamaño de ventana y tamaño de paso en la ventana. El parámetro más importante es el tamaño de ventana, normalmente se experimenta con distintos valores hasta encontrar el valor óptimo, mientras que el tamaño de paso en la ventana se mantiene típicamente igual a 1.

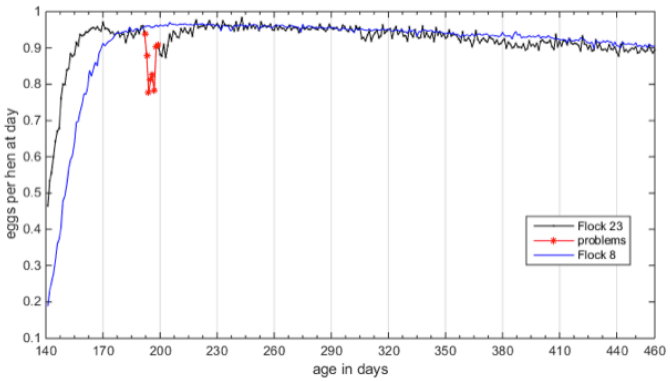
Dos trabajos publicados por Ramírez et al. (2016) y Ramírez et al. (2017) demuestra el uso de máquinas de soporte vectorial y de redes neuronales con una ventana deslizante para obtener un modelo de alerta temprana.

En este estudio, fueron registrados datos de campo de una granja de gallinas ponedoras alojadas en un sistema productivo de reemplazo denominado “todo dentro - todo fuera”, es decir que todas las aves de un mismo lote tienen la misma edad y son alojados juntos por grupos durante todo el tiempo de producción.

Los datos fueron recopilados diariamente desde enero 2008 a diciembre 2015. Debido a la organización y logística interna de la granja, los huevos fueron recogidos a distintas horas, por lo que el espaciado temporal de los datos no es de 24 horas. En algunos días el intervalo entre registros es de 20 horas y en otros de 28 horas. Los datos con variaciones en el espaciado temporal, representan un desafío para cualquier modelo (Jones, 1984), ya que debe ser capaz de de discriminar entre una anomalía en la curva producto de un problema real y las alteraciones relacionadas con el momento de la recolección.

En el gráfico 8.8, se muestra dos lotes representativos de la base de datos, el lote 8, que tiene una curva de producción característica, sin que se presenten problemas durante todo el tiempo de producción, y el lote 23, que a pesar de que inicia su producción con menos edad, muestra una fuerte caída entre los 191 días y los 199 días, este intervalo de tiempo fue etiquetado como anomalías en la curva, ya que a partir del día 199 las aves empiezan a recuperarse.

Gráfico 8.8 Producción por ave / día de dos lotes representativos de la base de datos




Los resultados de estos trabajos indican que es posible realizar un pronóstico automático de caídas de producción con una precisión, sensibilidad y especificidad superiores a 0.95. A nivel de finca, una pronostico con un día de antelación, podría resultar útil para la inspección diagnóstica en finca

en busca de síntomas clínicos, u otros hallazgos para la toma de medidas tendientes a la solución inmediata del problema. Esto mejora la capacidad preventiva en el sistema de producción avícola, brindando monitorización asistida de manera automática como complemento a la observación humana, lo que resulta especialmente útil, al manejar altas poblaciones de animales.

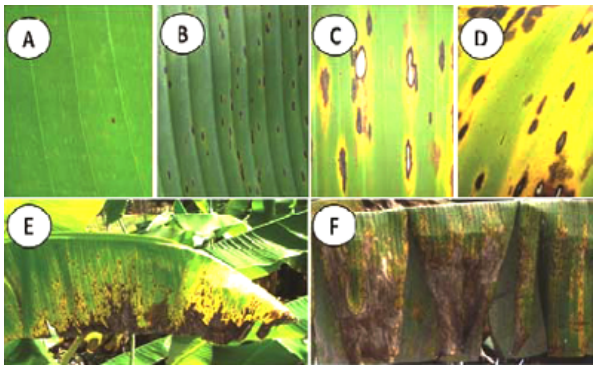
## Análisis de imágenes

Entre la comunidad del sector agropecuario es sabido que las enfermedades de las plantas y de los animales amenazan a la seguridad alimentaria, en esta área en particular, el uso de técnicas de inteligencia artificial tiene un papel fundamental para la identificación precisa y oportuna de enfermedades en los cultivos. Sin embargo esta tarea no es para nada trivial, y requiere de una gran cantidad de recursos para el entrenamiento y desarrollo de los algoritmos.

Actualmente se utilizan imágenes multiespectrales e hiperespectrales para el cálculo de índices de salud de la vegetación, sin embargo su utilización está muy limitada debido al alto costo de los equipos. Por otra parte, en los últimos 10 años se ha dado un fenómeno de universalización de la posesión de smartphones, al punto de que prácticamente en todas las unidades de producción agropecuaria hay al menos un dispositivo.

Esta particularidad ha hecho que el diagnóstico de enfermedades mediante smartphone sea una realidad cada vez más cercana. Existen bases de datos  tanto públicas como privadas que han recopilado y etiquetado decenas de miles de imágenes de plantas enfermas y sanas. En algunos casos estas imágenes han sido recolectadas en condiciones controladas, por lo que se infiere que su veracidad es alta. En el caso de bases de datos de animales sanos y enfermos, a criterio de los autores, no existen muchas fuentes de información, por lo que se recomienda iniciar una investigación en este sentido.

Imagen 8.2 Estadíos de afectación por Sigatoka Negra en hojas de banano (Vézina 2017).




En el Imagen 8.2 se puede observar los estadíos de afectación por el hongo que produce la enfermedad en el banano denominada Sigatoka Negra. En la Universidad Técnica de Machala, se ha propuesto para este año un proyecto que sea capaz de brindar una asistencia al diagnóstico en la evaluación del estado de afectación de las plantaciones de banano.

En la literatura científica se describen decenas de artículo científicos basados en ensayos experimentales que prueban la precisión de algoritmos de clasificación de imágenes, con resultados de más del 99% de exactitud, lo que pone en evidencia la viabilidad de este enfoque que más adelante será capaz de generar recomendaciones inteligentes asistidas por un smartphone a escala global.

## Análisis de sonidos

Uno de los signos para el diagnóstico de enfermedades en los animales de granja, está relacionado con el sonido que emiten los animales. Particularmente en las enfermedades respiratorias. Los médicos veterinarios consideran a la tos, como un mecanismo de defensa del cuerpo, contra la posible entrada de agentes extraños en el sistema respiratorio.

Las características de la tos son indicativos de posibles enfermedades respiratorias. Partiendo de esta premisa, varios investigadores han estudiado los sonidos durante un

cuadro de tos en los animales para monitorizar posibles problemas de salud con la ayuda de un sistema experto.

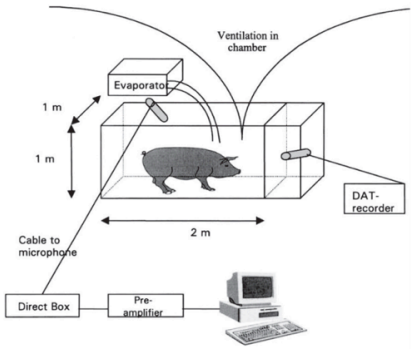
El uso de sistemas de soporte a la toma de decisiones en los sistemas agropecuarios tienen un alto potencial, debido a que en los sistemas de producción intensivos se manejan grandes cantidades de animales por lote, de tal manera que resulta un costo elevado tener sistemas de monitorización basados en observación humana.

El uso correcto de este tipo de sistemas es capaz de prevenir una zoonosis, o una epizootia, por este motivo, el desarrollo y aplicación de sensores y técnicas de detección para el diagnóstico automático es hoy un "hot topic" en la investigación y en la industria pecuaria.

Para el desarrollo de un sistema automático, se requiere que un experto etiquete una base de datos de sonidos de tos presencia de una enfermedad potencial, es decir, se utiliza técnicas de aprendizaje supervisado (Gráfico 8.9).

En el estudio de Chedad y colaboradores (2001) se utilizó redes de neuronas artificiales para predecir enfermedades respiratorias en cerdos. Para esto los autores construyeron una cámara de metal en la que cada cerdo es expuesto a variaciones de las condiciones ambientales tales como temperatura, el polvo,  concentración de NH3, y otras variables. El sistema logró un reconocimiento correcto de los sonidos con más del 90% de precisión.

Gráfico 8.9 Esquema del ensayo para la grabación de los sonidos emitidos por los cerdos en el estudio de Chedad y colaboradores (2001)



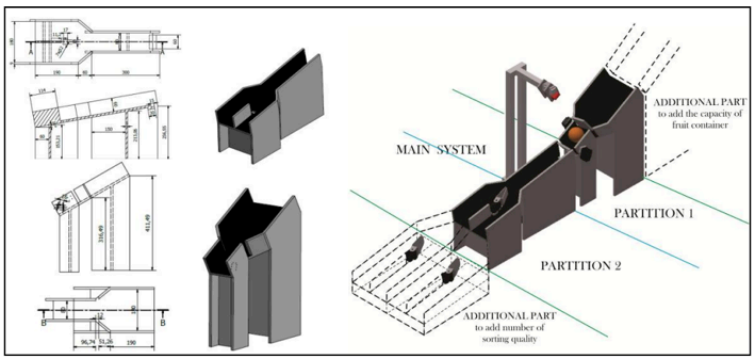
## Análisis de videos

La supervisión por video se utiliza comúnmente en aplicaciones de detección y clasificación en la industria agropecuaria, principalmente en las cadenas agroindustriales y procesos de postcosecha.

Una aplicación que resulta interesante debido a su potencialidad para automatizar las medianas y pequeñas fábricas agroindustriales tiene que ver con la utilización conjunta de técnicas de visión por computadora, técnicas de deep learning y algunos servo motores. En el trabajo de Afrisal et al (2013) se utilizó una webcam para obtener vídeos en una planta de procesamiento de frutas.

El algoritmo de visión por computadora transforma el RGB (rojo, verde y azul) en el espacio de color HSV (tono, saturación y valor) para facilitar los procesos de segmentación de color. Luego un algoritmo de agrupamiento separa las frutas de acuerdo con el nivel de madurez y tamaño. Finalmente, los servo motores se activan para mover la fruta a una bandeja de acuerdo con su grado de calidad.

Gráfico 8.10 Diseño del clasificador portátil desarrollado por Afrisal et al (2013)




En el Gráfico 8.10, se puede apreciar de mejor manera el diseño del clasificador. El sistema es capaz de realizar la tarea de forma precisa en menos de un segundo, y el operador tiene la posibilidad de ver en tiempo real los resultados y los datos en su computador.

## Referencia Bibliográfica


Afrisal, H., Faris, M., P., G. U., Grezelda, L., Soesanti, I., & F., M. A. (2013). Portable smart sorting and grading machine for fruits using computer vision. *In 2013 International Conference on Computer, Control, Informatics and Its Applications (IC3INA)* (pp. 71–75). ieeexplore.ieee.org.

Chedad, A., Moshou, D., Aerts, J. M., Van Hirtum, A., Ramon, H., & Berckmans, D. (2001). AP—Animal Production Technology: Recognition System for Pig Cough based on Probabilistic Neural Networks. *Journal of Agricultural Engineering Research*, 79(4), 449–457.

Jones, R. H. (1984). Fitting Multivariate Models to Unequally Spaced Data. In E. Parzen (Ed.), *Time Series Analysis of Irregularly Observed Data* (pp. 158–188). Springer New York.

Kapoor, P., & Bedi, S. S. (2013). Weather Forecasting Using Sliding Window Algorithm. *International Scholarly Research Notices, 2013*. https://doi.org/10.1155/2013/156540

Liu M, Shi J, Li Z, Li C, Zhu J, Liu S. Towards Better Analysis of Deep Convolutional Neural Networks. IEEE Trans Vis Comput Graph 2017;23:91–100.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics, 5*(4), 115–133.

Ramírez, I., Rivero Cebrián, D., Fernández Blanco, E., & Pazos Sierra, A. (2016). Early warning in egg production curves from commercial hens: A SVM approach. *Computers and Electronics in Agriculture, 121*, 169–179.

Ramírez-Morales, I., Fernández-Blanco, E., Rivero, D., & Pazos, A. (2017). Automated early detection of drops in commercial egg production using neural networks. *British Poultry Science*. https://doi.org/10.1080/00071668.2017.1379051

Shannon, C. E. (1950). XXII. Programming a computer for playing chess. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 41*(314), 256–275.

Shiyu, J. (Nov, 13, 2016). Kernel method in SVM. Retrieved from https://commons.wikimedia.org/w/index.php?curid=60458994

Vapnik, V., Golowich, S. E., & Smola, A. J. (1997). Support Vector Method for Function Approximation, Regression Estimation and Signal Processing. In M. I. Jordan & T. Petsche (Eds.), *Advances in Neural Information Processing Systems 9* (pp. 281–287). MIT Press.

Vapnik, V. N., & Kotz, S. (1982). *Estimation of dependences based on empirical data* (Vol. 41). Springer-Verlag New York.

Vézina A. Sigatoka leaf spot | The knowledge platform on the banana. The knowledge platform on the banana 2017. http://www.promusa.org/Sigatoka+leaf+spot (consultado el 11 de mayo de 2018).

*Análisis de Datos Agropecuarios*
Edición digital 2017- 2018.
www.utmachala.edu.ec

# Redes

Redes es la materialización del diálogo académico y propositivo entre investigadores de la UTMACH y de otras universidades iberoamericanas, que busca ofrecer respuestas glocalizadas a los requerimientos sociales y científicos. Los diversos textos de esta colección, tienen un espíritu crítico, constructivo y colaborativo. Ellos plasman alternativas novedosas para resignificar la pertinencia de nuestra investigación. Desde las ciencias experimentales hasta las artes y humanidades, Redes sintetiza policromías conceptuales que nos recuerdan, de forma empeñosa, la complejidad de los objetos construidos y la creatividad de sus autores para tratar temas de acalorada actualidad y de demanda creciente; por ello, cada interrogante y respuesta que se encierra en estas líneas, forman una trama que, sin lugar a dudas, inervará su sistema cognitivo, convirtiéndolo en un nodo de esta urdimbre de saberes.

UNIVERSIDAD TÉCNICA DE MACHALA
Editorial UTMACH
Km. 5 1/2 Vía Machala Pasaje
www.investigacion.utmachala.edu.ec / www.utmachala.edu.ec

ISBN: 978-9942-24-120-7