SUBSECRETARÍA DE POSGRADO · FHyCS · UNJu · JUNIO 2026

UNIDAD 2 · MATERIAL COMPLEMENTARIO DE LECTURA

Diseño y Estructura de Bases de Datos

para Proyectos de Investigación

Dr. Javier Altamirano · Mg. Ana María Chalabe

Facultad de Humanidades y Ciencias Sociales – UNJu

Introducción

Este material de lectura acompaña el segundo encuentro del curso y tiene como propósito desarrollar en profundidad los conceptos trabajados durante la clase sincrónica. Su lectura previa o posterior a la sesión en vivo le permitirá afianzar los fundamentos del diseño y la estructura de bases de datos y contar con una referencia escrita para las actividades prácticas de la unidad.

Los temas que se desarrollan aquí son la base sobre la que se construirán las unidades siguientes: no es posible construir ni gestionar una base de datos de calidad sin haber diseñado previamente su estructura con criterio metodológico. Una base de datos bien diseñada es el resultado de responder correctamente un conjunto de preguntas antes de recolectar o cargar el primer dato.

El material está organizado en seis secciones: tablas dinámicas como herramienta de exploración; clasificación de variables; diseño conceptual de la base de datos; normalización; diccionario de datos y libro de códigos; y fuentes de datos primarias y secundarias.

U N I D A D 0 2

Diseño y Estructura de Bases de Datos

para proyectos de investigación científica

1. Tablas dinámicas como herramienta de exploración

Una tabla dinámica es un resumen interactivo de una base de datos que permite reorganizar, agrupar y calcular valores estadísticos —suma, promedio, conteo, máximo, mínimo— sin modificar en ningún momento los datos originales. Es probablemente la herramienta más poderosa que ofrece Excel para el análisis exploratorio de datos, y su dominio marca una diferencia significativa en la capacidad analítica de un investigador.

La principal ventaja de las tablas dinámicas sobre las funciones convencionales como CONTAR.SI o PROMEDIO es que no requieren escribir fórmulas: el investigador arrastra variables hacia diferentes áreas de análisis y Excel recalcula los resultados de forma inmediata. Esto permite explorar los datos con una agilidad que sería imposible con funciones individuales, y facilita la formulación y el testeo de hipótesis durante el propio proceso de análisis.

1.1 Requisitos para crear una tabla dinámica

Para que Excel pueda generar una tabla dinámica, la base de datos debe cumplir con los principios de organización trabajados en la Unidad 1:

- Una sola fila de encabezados en la primera fila, con nombres únicos para cada columna.
- Sin filas ni columnas vacías dentro del rango de datos.
- Cada columna con un solo tipo de dato — no mezclar texto y números en la misma columna.
- Sin celdas fusionadas ni filas de totales intercaladas en la tabla.

> *Una base de datos mal estructurada generará tablas dinámicas incorrectas o difíciles de interpretar. El paso previo de limpieza y normalización no es optativo: es el requisito mínimo para que el análisis sea válido.*

1.2 Las cuatro áreas de una tabla dinámica

Una tabla dinámica en Excel organiza las variables en cuatro áreas que determinan cómo se presenta el análisis:

Filas	Variable cuyos valores forman las filas de la tabla. Generalmente la variable principal de análisis. Ejemplo: tipo de accidente.
Columnas	Variable cuyos valores se distribuyen en columnas. Útil para comparar categorías entre sí. Ejemplo: día de la semana.
Valores	Lo que se calcula: CONTAR para variables categóricas, SUMA o PROMEDIO para cuantitativas. Ejemplo: cantidad de casos (Contar).
Filtros	Variable que permite filtrar toda la tabla dinámica mostrando solo una categoría. Ejemplo: zona (urbana / rural).

1.3 ¿Qué función usar en el área de Valores?

La elección de la función de resumen en el área de Valores no es arbitraria: depende del tipo de variable que se está analizando.

- **Variables cualitativas (nominal u ordinal):** usar CONTAR o CONTAR DISTINTOS. No tiene sentido sumar o promediar categorías.
- **Variables cuantitativas discretas:** SUMA para totales (número de plantas, cantidad de eventos), PROMEDIO para tendencia central, CONTAR para frecuencias.
- **Variables cuantitativas continuas:** PROMEDIO, MÍN, MÁX para medidas descriptivas. También es posible calcular la desviación estándar directamente desde la tabla dinámica.

2. Clasificación de variables

Una variable es cualquier característica o atributo de la unidad de análisis que puede tomar diferentes valores entre los distintos casos registrados en la base de datos. La clasificación correcta de las variables que componen una base de datos tiene consecuencias directas sobre cómo se diseña la base, qué tipo de valores se almacenan en cada columna, cómo se codifican los datos y qué técnicas de análisis son apropiadas para cada variable.

Existen dos grandes grupos: las variables cualitativas y las variables cuantitativas. Cada grupo se subdivide a su vez en subtipos con características y tratamientos diferentes.

2.1 Variables cualitativas

Las variables cualitativas —también llamadas categóricas— registran atributos o categorías que no son intrínsecamente numéricas. Sus valores son nombres, etiquetas o clasificaciones. Se subdividen en:

Nominales

Sus categorías no tienen un orden entre sí: ninguna es "mayor" o "mejor" que otra. Son simplemente diferentes. Pueden ser binomiales —con exactamente dos categorías posibles— o multinomiales —con más de dos categorías.

Ejemplos: sexo biológico (MASCULINO / FEMENINO), tipo de vehículo (Automóvil / Bicicleta / Moto), zona (urbana / rural), tipo de accidente.

Consideraciones para el diseño: las categorías deben definirse antes de comenzar la recolección de datos y mantenerse consistentes en todos los registros. En el diccionario de datos deben listarse todos los valores posibles.

Ordinales

Sus categorías tienen un orden jerárquico o una escala de valores definida, pero las distancias entre categorías no son necesariamente iguales ni cuantificables. El orden importa, pero no la magnitud de las diferencias.

Ejemplos: nivel educativo (primario / secundario / universitario / posgrado), grado de severidad de una lesión (leve / moderada / grave), estadío de desarrollo de un insecto (huevo / larva / pupa / adulto).

Consideraciones para el diseño: debe establecerse explícitamente el orden de las categorías en el diccionario de datos. Si se asignan códigos numéricos (1, 2, 3...) para representar las categorías ordinales, es fundamental documentar qué número corresponde a qué categoría para evitar interpretaciones erróneas.

2.2 Variables cuantitativas

Las variables cuantitativas representan cantidades numéricas sobre las que tiene sentido realizar operaciones aritméticas: suma, resta, promedio. Se subdividen en:

Discretas

Toman valores enteros que provienen de conteos. Entre dos valores consecutivos no existen valores intermedios posibles: no puede haber 2,5 plantas en una parcela ni 1,3 hijos en una familia.

Ejemplos: número de plantas por parcela, cantidad de vainas por planta, número de eventos registrados, número de hijos, cantidad de publicaciones.

Consideraciones para el diseño: la columna debe formatearse como número entero en Excel. Si se permite el ingreso de valores decimales, es probable que exista un error de carga.

Continuas

Pueden tomar cualquier valor numérico dentro de un rango, incluyendo decimales. Provienen de mediciones y, en teoría, su precisión solo está limitada por el instrumento de medición.

Ejemplos: peso en kilogramos, altura en centímetros, temperatura en grados Celsius, rendimiento en kg/ha, concentración de una sustancia.

Consideraciones para el diseño: deben especificarse la unidad de medida y la cantidad de decimales que se van a registrar. Esto reduce la variabilidad en el ingreso de datos y facilita el análisis posterior.

> *La clasificación de variables no es un ejercicio académico: determina qué se puede hacer con los datos. No tiene sentido calcular el promedio de una variable nominal, ni representar con un gráfico de barras una variable continua sin agrupar primero sus valores en intervalos.*

3. Diseño conceptual de la base de datos

El diseño conceptual es el conjunto de decisiones que se toman antes de abrir cualquier software y de ingresar el primer dato. Es la etapa en la que se define qué se va a registrar, por qué, en qué forma y con qué criterios. Una base de datos bien diseñada conceptualmente es mucho más fácil de construir, mantener y analizar. Una base de datos mal diseñada genera problemas que a menudo son irreparables sin reconstruirla desde cero.

El diseño conceptual no requiere conocimientos técnicos avanzados: requiere claridad metodológica sobre el proyecto de investigación. Las cuatro preguntas fundamentales que estructuran este diseño son:

3.1 Las cuatro preguntas del diseño conceptual

¿Cuál es el objetivo? Define qué información es relevante y cuál es prescindible. El objetivo determina las variables que se necesitan, el nivel de detalle requerido y el tipo de análisis que se realizará. Si el objetivo no está claro antes de diseñar la base, la base resultante suele tener o demasiadas variables irrelevantes o carecer de variables fundamentales.

¿Cuál es la unidad de observación? Es lo que representa cada fila de la base de datos. Puede ser una persona, un animal, una parcela, un evento, una institución, un momento de medición, un documento. La unidad de observación debe ser consistente en toda la tabla: cada fila debe representar exactamente el mismo tipo de cosa.

¿Cuáles son las variables respuesta? Son las variables que se miden o registran como resultado del fenómeno que se estudia —las variables dependientes en el lenguaje del diseño experimental. Son lo que el investigador quiere explicar o describir.

¿Cuáles son las variables explicatorias? Son los factores, condiciones o características que podrían explicar los valores de las variables respuesta —las variables independientes. Incluyen los tratamientos experimentales, las características demográficas, las condiciones ambientales o cualquier otra variable que el investigador considere relevante para el análisis.

> *Responder estas cuatro preguntas antes de recolectar los datos ahorra enormes problemas posteriores. Es mucho más costoso rediseñar una base de datos con cientos de registros cargados que planificar bien desde el inicio.*

3.2 Ejemplo de diseño conceptual: ensayo de cultivo de soja

Para ilustrar cómo se aplican estas cuatro preguntas, tomemos el caso del ensayo de cultivo de soja que se trabaja en la práctica de esta unidad:

Objetivo	Comparar el efecto de 6 tratamientos de fertilización sobre el rendimiento de soja en diferentes parcelas.
Unidad de observación	Cada planta individual muestreada dentro de una parcela y tratamiento específico.
Variables respuesta	Número de vainas por planta, número de nudos, rendimiento en kg/parcela, peso de 1000 semillas (g).
Variables explicatorias	Tratamiento (T1–T6: tipos de fertilizante), Parcela (A, B, C), Repetición (1, 2, 3).

Nótese que este diseño conceptual determina exactamente qué columnas va a tener la base de datos y qué valor va en cada celda antes de registrar un solo dato. Esto es el diseño conceptual en acción.

4. Normalización: evitar redundancia y asegurar consistencia

El concepto de normalización proviene del campo de las bases de datos relacionales, pero sus principios fundamentales aplican igualmente a las bases de datos tabulares que se trabajan en investigación con Excel o similares. Normalizar una base de datos significa organizarla de manera que cada hecho se registre una sola vez, en el lugar correcto, con un formato consistente.

El estadístico Hadley Wickham formalizó estos principios bajo el nombre de Tidy Data ("datos ordenados"), que se ha convertido en el estándar de referencia en la comunidad de análisis de datos. Los tres principios del Tidy Data —cada variable en una columna, cada observación en una fila, cada valor en una celda— son equivalentes a los principios básicos de normalización aplicados a las hojas de cálculo.

4.1 Problemas que la normalización resuelve

Redundancia. Cuando el mismo dato aparece en múltiples lugares de la base, cualquier corrección debe hacerse en todos esos lugares simultáneamente. Si se olvida alguno, la base queda en estado inconsistente. La normalización elimina la redundancia registrando cada dato exactamente una vez.

Inconsistencia. Cuando el mismo valor se registra de maneras diferentes en distintos registros —por ejemplo, "Masculino", "M", "masc." y "MASCULINO" para la misma categoría— el software de análisis los trata como cuatro categorías distintas, produciendo resultados incorrectos. La normalización requiere que los valores sean idénticos para el mismo concepto.

Valores calculados incrustados. Incluir totales, promedios o resúmenes dentro de la tabla de datos genera múltiples problemas: se confunden con observaciones reales, distorsionan los análisis automáticos y hacen imposible la creación de tablas dinámicas correctas. Los resúmenes deben calcularse fuera de la tabla, con funciones o tablas dinámicas.

Formato ancho vs. formato largo. Una fuente de datos muy frecuente en investigación es el formato ancho: varias columnas para lo que en realidad es la misma variable medida en distintas condiciones. Por ejemplo, una columna por tratamiento con los valores de rendimiento. El formato normalizado ("largo") usa una sola columna para el rendimiento y otra para el tratamiento. El formato largo es el único compatible con tablas dinámicas y con la mayoría de los software de análisis estadístico.

> *La base de datos de cabras que se muestra como ejemplo en clase ilustra exactamente estos problemas: formato ancho con grupos en bloques de columnas, totales y promedios incrustados, y fechas sin formato estandarizado. Reconocer estos problemas en fuentes secundarias es una habilidad fundamental para cualquier investigador.*

5. Diccionario de datos y libro de códigos

Un diccionario de datos es un documento que describe sistemáticamente cada variable de una base de datos: su nombre técnico, su tipo, la unidad de medida cuando corresponde, una descripción en lenguaje natural de qué mide y cuáles son los valores posibles o los códigos que puede tomar. Es, en términos simples, el "manual de lectura" de la base de datos.

El libro de códigos es un documento complementario que define con mayor detalle los códigos numéricos o alfanuméricos utilizados para representar categorías. Por ejemplo, si en la variable nivel_educativo se usan los códigos 1, 2, 3 y 4, el libro de códigos especifica que 1 = primario, 2 = secundario, 3 = universitario y 4 = posgrado.

Ambos documentos son fundamentales por la misma razón: una base de datos sin documentación solo puede ser interpretada correctamente por quien la construyó, y solo mientras recuerda los criterios que utilizó. Con el tiempo —o cuando otros investigadores acceden a los datos— la información sin documentar se convierte en información inaccesible.

5.1 Estructura del diccionario de datos

Un diccionario de datos bien estructurado incluye para cada variable al menos los siguientes campos:

Variable	Tipo	Descripción	Valores / Códigos	Dato faltante
`edad`	Cuantitativa discreta	Edad en años cumplidos del lesionado al momento del accidente	Entero entre 0 y 120	NA
`sexo`	Cualitativa nominal	Sexo biológico del lesionado	MASCULINO / FEMENINO	NA
`fecha_lesion`	Fecha	Fecha del accidente en formato DD/MM/AAAA	01/01/2010 a 31/12/2010	NA
`zona`	Cualitativa nominal	Tipo de zona geográfica donde ocurrió el accidente	urbana / rural	NA
`tipo_accidente`	Cualitativa nominal	Clasificación del tipo de accidente según protocolo	Caída del vehículo / Vuelco / Colisión vehículos / Colisión con objeto fijo	NA

Este diccionario se construye una vez, antes o durante la recolección de datos, y se actualiza cada vez que se toma una decisión de codificación nueva. Es un documento vivo, no un trámite que se hace al final.

5.2 Por qué el diccionario es una herramienta de investigación

Más allá de su función documental, el diccionario de datos tiene una función metodológica activa: construirlo obliga al investigador a tomar decisiones explícitas sobre la codificación antes de comenzar a cargar datos. Estas decisiones incluyen:

- ¿Qué categorías son posibles para cada variable categórica? ¿Son mutuamente excluyentes?
- ¿Qué rango de valores es razonable para cada variable cuantitativa? ¿Qué se hace con valores atípicos?
- ¿Cómo se codifican los datos faltantes? ¿Con NA, con 999, con -9?
- ¿Qué unidad de medida se usa para cada variable cuantitativa? ¿Gramos o kilogramos? ¿Centímetros o metros?

> *El diccionario de datos es un entregable obligatorio de la tarea asincrónica de esta unidad. Deberán construirlo para su propio proyecto de investigación: es el primer paso concreto hacia la base de datos final.*

6. Fuentes de datos: primarias y secundarias

Una de las primeras decisiones metodológicas en cualquier proyecto de investigación es determinar de dónde provienen los datos: ¿los generará el propio investigador, o utilizará datos ya existentes generados por otros? Esta distinción tiene consecuencias profundas sobre el diseño de la base de datos, el nivel de control sobre la calidad de los datos y los desafíos que deberán enfrentarse durante el análisis.

6.1 Fuentes primarias

Una fuente primaria es aquella generada por el propio investigador para un objetivo específico. El investigador diseña el instrumento de recolección (encuesta, planilla de campo, protocolo de observación), define las variables, establece los criterios de codificación y controla el proceso de recolección.

Las ventajas son significativas: las variables se diseñan exactamente para el objetivo de la investigación, el investigador tiene control sobre la calidad del registro, y las decisiones de codificación son propias. La desventaja principal es el costo en tiempo y recursos que implica la recolección original de datos.

El ejemplo del ensayo de cultivo de soja es un caso paradigmático de fuente primaria: el diseño experimental, los instrumentos de medición y los criterios de registro fueron definidos por el investigador antes de comenzar la recolección.

6.2 Fuentes secundarias

Una fuente secundaria es aquella generada por otra persona u organización, adoptada por el investigador para un objetivo distinto al original. Los datos ya existen: el investigador los reutiliza, adapta o combina con otros datos para responder sus propias preguntas.

Las ventajas incluyen el acceso inmediato a grandes volúmenes de datos históricos o de alcance masivo que sería imposible recolectar individualmente —como datos censales, encuestas nacionales o registros administrativos—, así como el bajo costo de recolección. La desventaja principal es que las variables disponibles pueden no ajustarse exactamente al objetivo, y el investigador debe comprender y respetar la codificación original del generador de los datos.

6.3 Desafíos específicos de las fuentes secundarias

Trabajar con fuentes secundarias implica un conjunto de desafíos metodológicos que no existen cuando los datos son propios:

- **Comprender el diccionario ajeno:** las bases de datos secundarias frecuentemente usan códigos numéricos para representar categorías. Sin el diccionario de datos del generador original, esos códigos son ininterpretables.
- **Evaluar la calidad del registro:** el investigador no controló la recolección y no puede asumir que los estándares de calidad propios se aplicaron.
- **Restructurar el formato:** muchas fuentes secundarias —especialmente registros administrativos— tienen formatos no tabulares, con múltiples hojas, encabezados anidados o datos en formato ancho que deben transformarse antes de poder analizarse.
- **Respetar las limitaciones originales:** los datos secundarios fueron generados para un objetivo diferente. Algunas variables del objetivo original pueden no estar disponibles, o pueden estarlo en una granularidad diferente a la necesaria.

> *La base de datos de cabras de Lozano es un ejemplo real de fuente secundaria con problemas de estructura. Tiene datos valiosos, pero requiere una transformación significativa antes de poder analizarse. En la Unidad 3 aprenderemos a realizar esa transformación usando Power Query.*

Referencias bibliográficas

Mayo, F. (2024). *Excel para el análisis de datos: Técnicas y trucos para simplificar el trabajo*. Madrid: Anaya Multimedia. 400 pp. ISBN: 978-84-415-5035-3. (Capítulos 1-3, pp. 21-180)

Romero-Carazas, R., Mayta-Huiza, D., Ancaya-Martínez, M., Tasayco-Barrios, S. y Berrio-Quispe, M. (2024). *Método de investigación científica: Diseño de proyectos y elaboración de protocolos en las Ciencias Sociales*. Lima: Editorial Idicap Pacífico. 198 pp. https://doi.org/10.53595/eip.012.2024 (Capítulo 3, pp. 65-98)

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, 59(10), 1-23. https://doi.org/10.18637/jss.v059.i10