Máster en Planificación y Gestión de Procesos Empresariales

Excel para el análisis de datos

Álvaro Briz Redón

Departamento de Estadística e Investigación Operativa
Universitat de València

alvaro.briz@uv.es

27 de noviembre de 2025

VNIVERSITAT DE VALÈNCIA

Objetivos y advertencias

Objetivos:

- Conocer el potencial uso de Excel como herramienta de análisis estadístico de datos.
- En concreto, utilizar el paquete **"Herramientas para Análisis"** de Excel.
- Combinar Excel con R (RStudio) para elaborar documentos Excel de análisis de datos.

Advertencias:

- Realizar un análisis estadístico "avanzado" puede resultar muy complicado (o inviable) mediante Excel.
- Sí que podemos realizar análisis estadísticos sencillos con Excel de forma directa (con el paquete o manualmente).
- La combinación entre Excel y R puede ser de interés en **contextos donde trabajen o colaboren personas de diferente perfil**.

Mi experiencia

- Uso de Excel fundamentalmente como medio de lectura/escritura de bases de datos en investigación.
  - Carga de datos para su análisis en R.
  - Escritura de resultados en Excel para su posterior visualización (no tan frecuente).
- Uso combinado de Excel y R para generar informes estadísticos en la Oficina de Estadística del Ayuntamiento.
  - Ejemplo:
    Estudio sobre *Mortalidad en la ciudad de València 2011-2015* (link).
- Uso de macros de Excel (VBA) para explotar ficheros y generar bases de datos.
  - Ejemplo:
    Explotación de ficheros .txt del BORME para crear bases de datos sobre empresas constituidas y extinguidas en la Oficina de Estadística del Ayuntamiento.

Contenidos

Algunos contenidos a tratar

1. Tablas Dinámicas
2. Estadística Descriptiva
3. Contrastes de Hipótesis
4. Regresión Lineal
5. Series Temporales
6. **Conexión entre Excel y R**

Nuestro dataset: Online Retail

- Datos reales de transacciones de ventas minoristas en línea (Aula Virtual).

  Puede descargarse desde el repositorio Kaggle:
  `https://www.kaggle.com/datasets/kabilan45/online-retail-ii-dataset`
- **Para Tablas Dinámicas:** Columnas Categóricas (`Country`, Producto).
- **Para Estadística/Regresión:** Columnas Numéricas (`Quantity`, `Price`).
- **Para Series Temporales:** Columna de Fecha (`InvoiceDate`).

Crear columna de Total

$\texttt{Total} = \texttt{Quantity} \times \texttt{Price}$

Requisito: Activar “Herramientas para Análisis”

1. Ir a **Archivo** $\rightarrow$ **Opciones**.
2. Seleccionar **Complementos**.
3. En *Gestionar: Complementos de Excel*, hacer clic en **Ir...**.
4. Marcar la casilla **“Herramientas para análisis”**.
5. Aceptar.

Resultado

Aparece el grupo **Análisis** en la pestaña **Datos**.

Tablas dinámicas

- Herramienta clave para **transformar** una base de datos en una tabla cruzada.
- Permiten **agrupar, filtrar y agregar** (sumar, contar, promediar) según variables categóricas.

Ejemplo

¿Cuál es la **Suma de Total** por **País** y **Producto**?

Creación de una tabla dinámica

1. Selecciona cualquier celda dentro del conjunto de datos.
2. Pestaña **Insertar** $\rightarrow$ **Tabla Dinámica**.
3. Elegir **Nueva hoja de cálculo**.

Construcción

- Arrastrar `Country` y `Description` a **Filas**.
- Arrastrar `Total` a **Valores** (se suma por defecto).

De suma a promedio

- Por defecto, los valores se **suman**. Podemos cambiar la función.

Construcción

1. Clic derecho sobre el campo en **Valores** → **Configuración de campo de valor...**
2. Seleccionar **Promedio** → Aceptar.

Agrupación temporal

- Clave para obtener series temporales.

Pasos

1. Arrastrar `InvoiceDate` a **Filas**.
2. Clic derecho en cualquier celda con contenido temporal en la tabla → **Agrupar...**
3. Seleccionar **Meses** y **Años** → Aceptar.

Resultado: Total desagregadas por Mes y Año.

Filtrado por variable categórica

- Podemos filtrar según el valor de una de las variables categóricas disponibles.

Pasos

1. Añadir `Country` a filtro.
2. Elegir un país desde la propia tabla dinámica.

La tabla dinámica se actualizará al hacer clic en cualquier país.

Ejercicio

Ejercicio

Construye una Tabla Dinámica que muestre:

- **Filas:** País y producto.
- **Valores:** Recuento del número de pedidos.

¿Cuál es el producto más vendido en cantidad de pedidos en Reino Unido?

Importante

Para un uso avanzado de tablas dinámicas debemos controlar el apartado **Analizar tabla dinámica** que aparece en el menú superior cuando tenemos seleccionada una tabla.

Entre otras funcionalidades, nos permite:

- Cambiar el origen de los datos (los datos que se emplean para la tabla).
- Crear nuevos campos en la tabla.
- Generar gráficos dinámicos.

Estadística descriptiva: ¿Cómo son nuestros datos?

- Permite resumir las características de una única variable.
- **Tendencia central:** Media, Mediana.
- **Dispersión:** Desviación Estándar, Varianza, Rango.

Ejercicio

¿Cuál es la cantidad (`Quantity`) promedio de artículos por compra y cómo de variable es?

Funciones básicas en Excel

Algunas funciones básicas de estadística en Excel que podemos escribir como fórmulas desde cualquier celda:

- **Media:** `=PROMEDIO(rango)`
- **Mediana:** `=MEDIANA(rango)`
- **Desv. Estándar:** `=DESVEST.M(rango)`
- **Asimetría:** `=ASIMETRIA(rango)` (Para ver la forma de la distribución).

Herramienta "Estadística Descriptiva"

1. Pestaña **Datos** $\rightarrow$ **Análisis de Datos**.
2. Seleccionar **Estadística Descriptiva**.
3. **Rango de Entrada:** Columna `Quantity`.
4. Seleccionar **Resumen de estadísticas**.

Ventaja

Obtiene múltiples métricas con un solo click, incluyendo el error típico y la curtosis.

Interpretando la descriptiva

- **Media y mediana:** Si son muy diferentes, la distribución está sesgada (asimétrica).
- **Desviación estándar (s):** Mide la dispersión. Si s es muy grande en comparación con la media, los datos son muy dispersos.
- **Asimetría:** Si es muy positiva, hay una cola larga hacia la derecha (muchos valores bajos y algunos muy altos). Si es negativa, al revés (muchos valores altos y algunos muy bajos).

Contrastes de hipótesis: ¿Hay diferencias?

- Nos permite comparar dos grupos y determinar si la diferencia observada es **estadísticamente significativa**.

Problema

¿Es el precio promedio de pedido en **Reino Unido** significativamente diferente al de **Francia**?

- **Hipótesis Nula (H_0):** $\mu_{UK} = \mu_{Francia}$ (No hay diferencias).
- **Hipótesis Alternativa (H_a):** $\mu_{UK} \neq \mu_{Francia}$ (Sí hay diferencias).

El p-valor y el nivel de significatividad

- **Nivel de significatividad (α):** La tasa de error que estamos dispuestos a asumir (típicamente $\alpha = 0.05$ o 5 %).
- p-**valor:** La probabilidad de observar nuestros datos (o más extremos) si H_0 fuera cierta.

Regla de decisión

Si $p < 0.05$**:** Rechazamos H_0, la diferencia es estadísticamente significativa.
Si $p \geq 0.05$**:** No rechazamos H_0, la diferencia no es estadísticamente significativa.

Prueba t en Excel (Dos muestras)

1. **Preparación:** Filtrar la columna `Price` para **Reino Unido** y **Francia** y llevar los datos a otra hoja.
2. Pestaña **Datos** $\rightarrow$ **Análisis de Datos**.
3. Seleccionar **Prueba t para dos muestras suponiendo varianzas desiguales**.
4. **Rango Variable 1:** Datos de Francia.
5. **Rango Variable 2:** Datos de Reino Unido.

Regresión lineal: Estimación y predicción

- Permite **modelar** la relación lineal entre dos o más variables.
- Usada para **predecir** el valor de una variable (Y) basándose en el valor de otra (X).

Pregunta

¿Podemos predecir las `Total` (Y) basándonos en `Quantity` (X)?

El modelo y los coeficientes

- **Ecuación:** $Y = \beta_0 + \beta_1 X + \epsilon$
- Y **(Dependiente):** `Total` (Lo que queremos predecir).
- X **(Independiente):** `Quantity` (El predictor).
- β_0 **(Incercepto):** Valor de Y cuando $X = 0$.
- β_1 **(Pendiente):** Cuánto cambia Y por cada unidad de cambio en X.

Regresión lineal en Excel

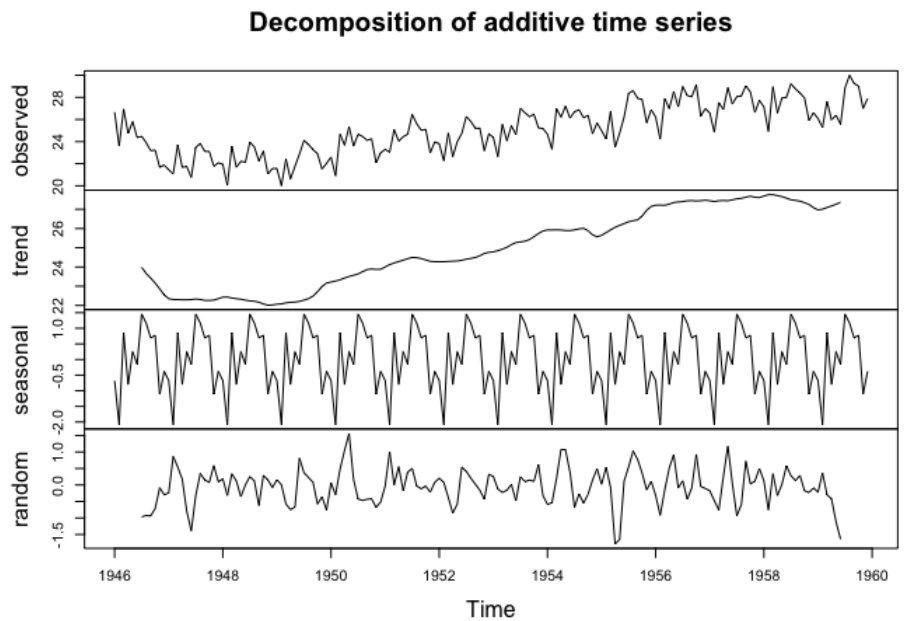
1. Pestaña **Datos** → **Análisis de Datos**.
2. Seleccionar **Regresión**.
3. **Rango Y de entrada:** Columna `Total`.
4. **Rango X de entrada:** Columna `Quantity`.

Series temporales

Definición (serie temporal)

Conjunto de observaciones indexado temporalmente: y_t, $t = 1, ..., T$.

- **Componentes clave de una serie temporal:**
  1. **Tendencia:** Crecimiento o decrecimiento a largo plazo.
  2. **Estacionalidad:** Patrones que se repiten periódicamente.
  3. **Ruido:** Fluctuaciones aleatorias.



Predicciones en Excel: Hoja de Previsión

- Una herramienta visual y rápida para predecir a futuro, basada en el método Error-Trend-Seasonal (ETS).

Pasos

1. Seleccionar el rango de fechas y el rango de valores.
2. Pestaña **Datos** $\rightarrow$ **Hoja de Previsión**.
3. Definir el **Final de la previsión**.

Suavizado Exponencial

- Un método clásico de suavizado de series temporales que pondera más las observaciones recientes.
- Útil para "eliminar" el ruido y ver la tendencia subyacente.

Pasos

1. Pestaña **Datos** $\rightarrow$ **Análisis de Datos**.
2. Seleccionar **Suavización Exponencial**.
3. **Factor de amortiguación:** Un valor entre 0 y 1. Cuanto más cerca esté de 1, mayor será el suavizado.

Excel + R: El paquete `openxlsx`

El paquete `openxlsx` nos permite cargar y generar archivos Excel con R.

Entre otras funcionalidades incluye:

- Cargar en R una base de datos almacenada en un Excel.
- Exportar a Excel una base de datos que hemos generado o trabajado desde R.
- **Crear un Excel desde 0, definiendo su formato y contenido.**
- **Abrir una plantilla de Excel con formato preestablecido para realizar modificaciones sobre el contenido de sus datos.**

Ver documento *Excel y R.pdf*.

Estudio de caso Excel + R: Conjunto de datos *stackloss*

Datos de operación de una planta de oxidación de amoníaco ($n = 21$).

Variable Respuesta (Y):

- `stack.loss`: 10 veces el porcentaje de amoníaco que escapa sin absorberse (ineficiencia).

Variables Predictoras (X):

- `Air.Flow`: Flujo de aire a la planta.
- `Water.Temp`: Temperatura del agua de enfriamiento.
- `Acid.Conc.`: Concentración de ácido nítrico circulante.

Modelización y objetivo

En R, cargamos los datos y fácilmente ajustamos un modelo de regresión lineal múltiple para explicar la respuesta con la función `lm()`.

Objetivo final:

Generar un Excel que permita evaluar visualmente la influencia de cada observación en el ajuste del modelo. Entre otras cuestiones, para:

- Identificar posibles observaciones anómalas.
- Plantearse reajustar el modelo sin dichas observaciones anómalas.

Existen muchas métricas para ello[1], pero nos centramos en dos: la distancia de Mahalanobis y la distancia de Cook.

[1] En el siguiente link se describen otras métricas (paquete `performance` de R).

Detección de outliers y valores influyentes: Distancia de Mahalanobis

Utilidad: Detectar *outliers* en el espacio de las variables predictoras (X), asumiendo que la dependencia entre estas es normal multivariante. Es decir, puntos con alto *leverage*.

Definición (distancia de Mahalanobis)

$$DM_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$$

Donde:

- $\mathbf{x}_i$: Observación multivariante i.
- $\bar{\mathbf{x}}$: Media de las observaciones.
- $\mathbf{S}$: Matriz observada de varianzas-covarianzas.

- Mide la distancia entre la observación y el centroide de los datos.
- Se compara con una distribución χ^2 con p grados de libertad ($p = 3$ predictores).

Detección de outliers y valores influyentes: Distancia de Cook

Utilidad: Identificar observaciones **influyentes** en el ajuste del modelo: una observación es **influyente** si, al eliminarla, el ajuste dado por el modelo cambia considerablemente.

Definición (distancia de Cook)

$$DC_i = \frac{1}{p \cdot \text{MSE}} \sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2$$

Donde:

- p: Número de predictores.
- MSE: Error cuadrático medio del modelo, $\text{MSE} = \frac{1}{n} \sum_{j=1}^{n} (y_j - \hat{y}_j)^2$.
- $\hat{y}_{j(i)}$: Valor ajustado sobre la observación j, si quitamos la observación i del modelo.