Fundamentos de Análisis de Datos
Introducción al Análisis Multivariado

Dra. Andrea Alejandra Rey

Especialización en Ciencia de Datos - ITBA


Obtención de información
Análisis de datos
Acceso a gran cantidad de datos

Terminología

Observaciones

Son las unidades sobre las cuales se miden los datos.

Población

Es el conjunto de todas las unidades de estudio.

Muestra

Es un conjunto de observaciones de la población.

Variable

Es una característica particular de la población que se recolecta estadísticamente.

Ejemplo de base

País	Calidad de vida	Seguridad	Costo de vida	Polución*
Argentina	111.77	35.89	30.78	51.32
Brasil	105.13	33.36	33.77	53.68
México	127.36	45.55	36.99	58.52
Estados Unidos	173.79	50.97	71.60	36.06
Francia	154.59	44.73	67.55	42.83
Italia	141.49	53.14	60.39	54.73
India	118.31	55.36	22.15	73.00
Nueva Zelanda	177.61	54.16	71.60	24.19
Nepal	84.51	62.40	24.67	84.14
Finlandia	191.31	73.71	66.47	11.99

*Descargado desde `https://www.numbeo.com/cost-of-living/` (enero, 2023).

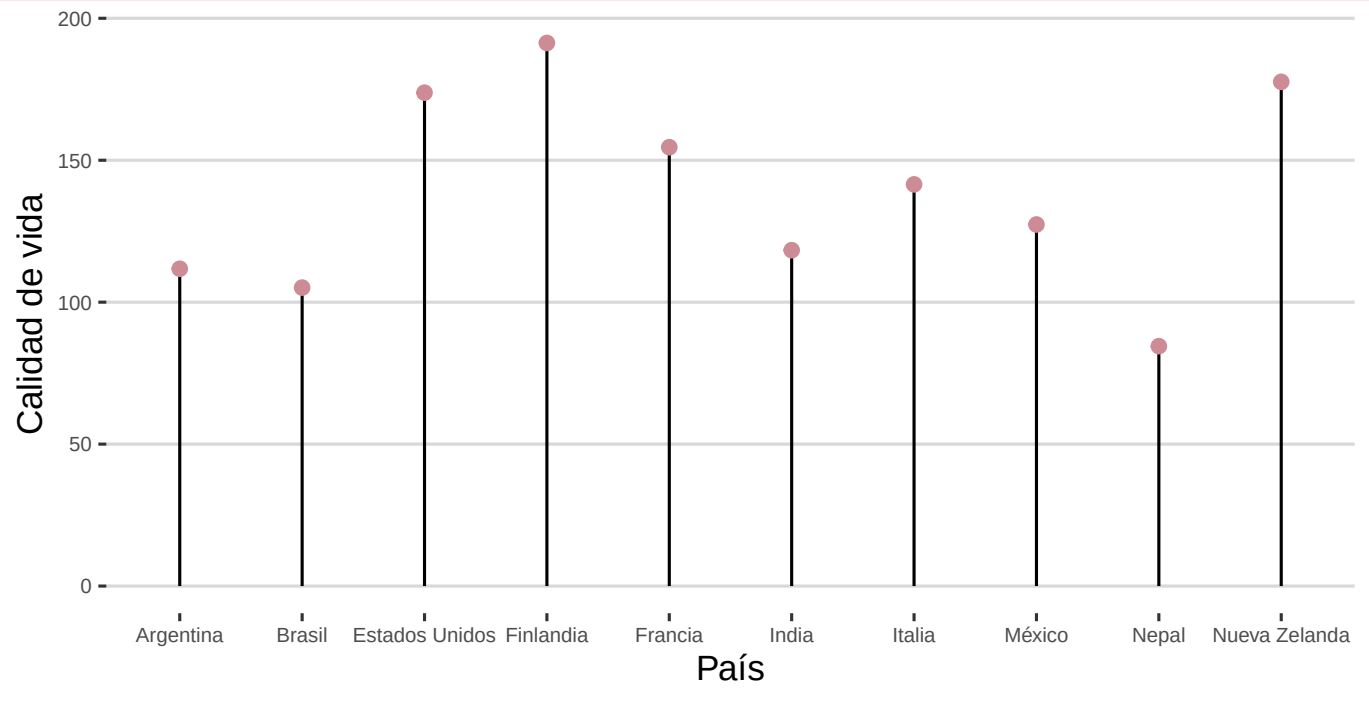
Análisis univariado

Existe un único tipo de variable.

Ejemplo

El índice de calidad de vida de ciertos países:

$$111.77, 105.13, 127.36, 173.79, 154.59, 141.49, 118.31, 177.61, 84.51, 191.31.$$

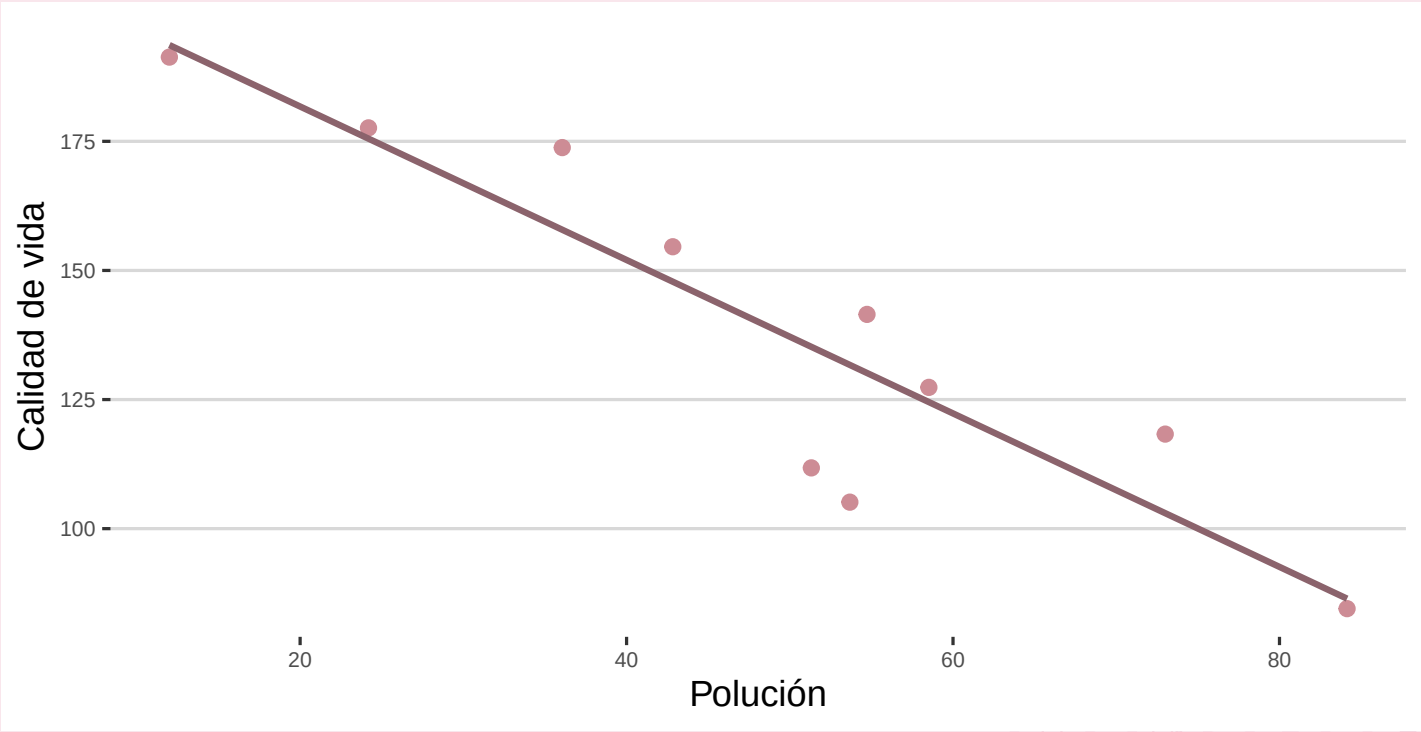


Análisis bivariado

Existen dos variable de las cuales se estudian relaciones simples de causalidad o asociación.

Pregunta

¿Influye el índice de polución en la calidad de vida de un país?



Análisis bivariado

Primer paso	Observar la naturaleza de la posible relación de las variables.
Segundo paso	Identificar los niveles de medición de los datos: valores nominales, ordinales o de ratios.
Tercer paso	Aplicar el rigor de la significancia estadística para validad resultados.

Análisis multivariado

Existen más de dos variable, dependientes o independientes, que producen una sola salida o *output*, en inglés.

> Todo lo que ocurre en el mundo ocurre por múltiples razones.

- Al momento de realizar predicciones o tomar decisiones investigadores, gerentes y consumidores utilizan una variedad de indicadores o factores junto con métricas asociadas para analizar su impacto en ciertas situaciones.
- Surge así la necesidad de técnicas confiables cada vez más sofisticadas para el análisis y la comprensión de datos.

Análisis multivariado

Características

- Contribuye a la reducción y simplificación de los datos sin perder detalles importantes.
- Las variables se agrupan y ordenan en base a sus características.
- Es importante verificar los datos recolectados y analizar su estado.
- Es de suma importancia comprender la relación entre todas las variables para poder predecir el comportamiento de las mismas en nuevas observaciones.
- Se debe testear la hipótesis estadística formulada a partir de los datos.

Análisis multivariado

Ventajas ☺

- La dependencia o independencia de las variables permiten la búsqueda de factores que ayudan a establecer conclusiones certeras.
- Las conclusiones obtenidas son cercanas a situaciones de la vida real debido al testeo del análisis.
- La exploración de múltiples variables proporciona un conocimiento más profundo de la realidad.

Análisis multivariado

Desventajas ☹

- Se requiere de cálculos complejos, tarea que puede resultar muy laboriosa.
- La recolección y procesamiento de un gran volumen de datos conlleva un largo tiempo de trabajo.

Dependencia versus independencia

¿Qué tipo de relación existe entre los datos?

Métodos de dependencia

La dependencia entre variables mide la relación causa-efecto, estudiando si los valores de ciertas variables independientes pueden ser usados para explicar, describir o predecir el valor de otra variable dependiente.
Las técnicas de dependencia se usan para construir modelos predictivos.

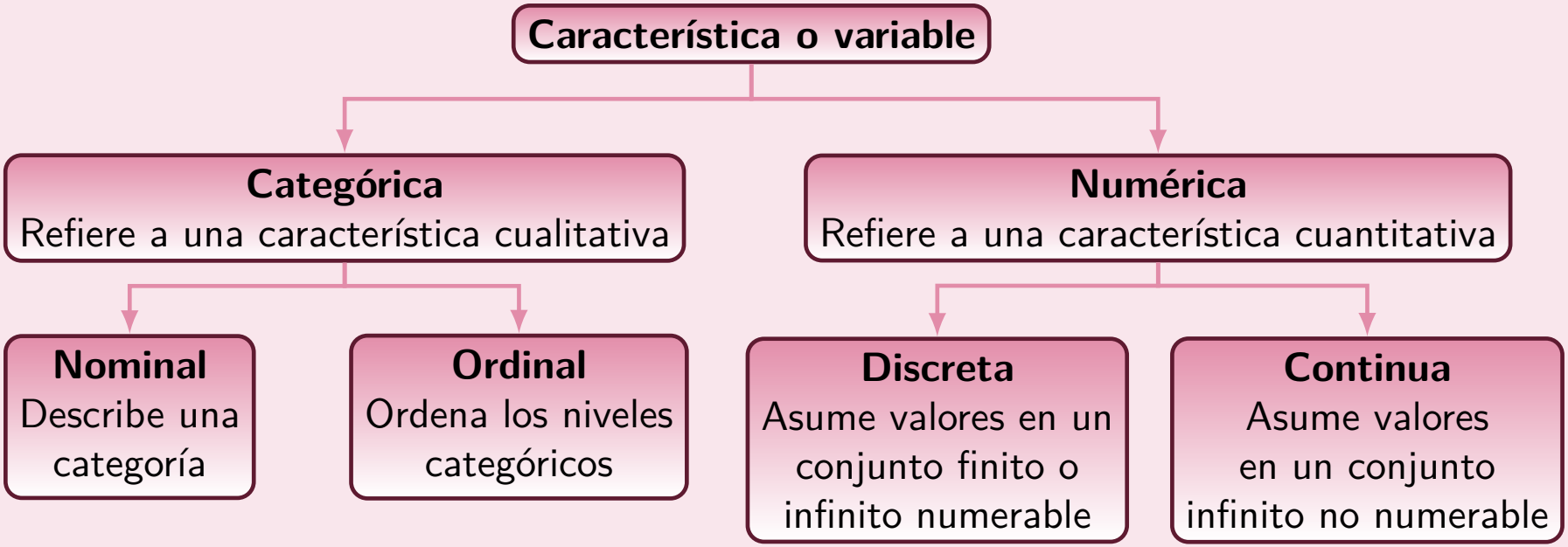
Métodos de independencia

La independencia entre variables se utiliza para comprender el significado estructural y las características subyacentes dentro de un conjunto de datos.
Las técnicas de independencia se usan para agrupar datos de manera significativa.

Técnicas de análisis multivariado

- Medidas estadísticas descriptivas.
- Representaciones gráficas.
- Análisis de varianza.
- Regresión lineal simple.
- Regresión lineal múltiple.
- Regresión logística.
- Análisis de componentes principales.

Tipo de variables

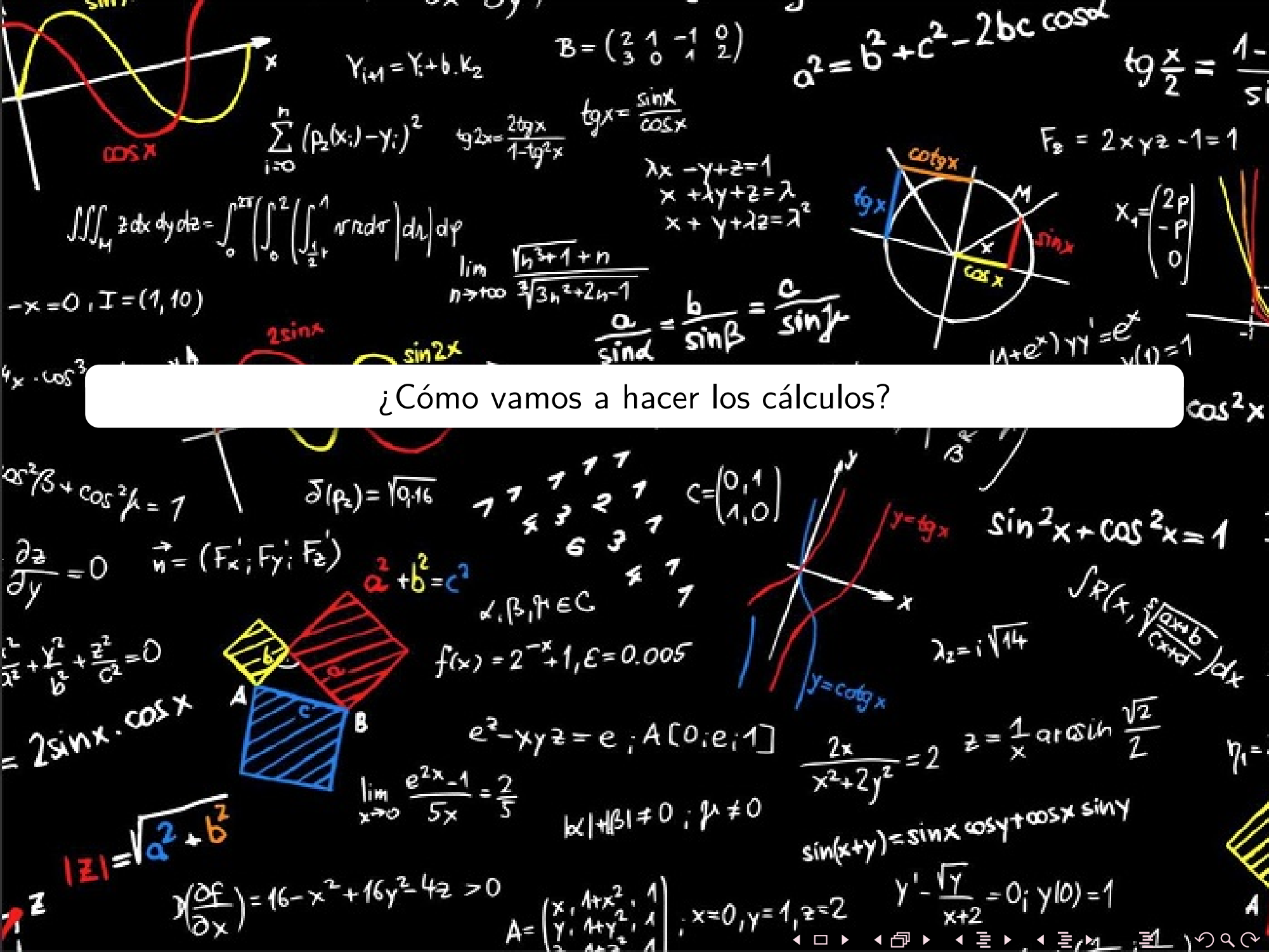


- Tipo de sangre
- Color de auto
- Género

- Respuestas en un rango desde la satisfacción a la insatisfacción
- Estadío de una enfermedad

- Población mundial
- Cantidad de hijos

- Edad precisa
- Temperatura precisa
- Tiempo de ejecución


¿Cómo vamos a hacer los cálculos?

R

The R Project for Statistical Computing

- Es un entorno de software libre utilizado para computación estadística y visualización gráfica.
- Compila y se ejecuta en una amplia variedad de plataformas.
- Permite un efectivo manejo y almacenamiento de datos.
- Es un lenguaje de programación bien desarrollado, simple y efectivo que incluye condicionales, bucles, funciones recursivas definidas por el usuario y facilidades de entrada y salida.

RStudio

RStudio IDE

- Es un entorno de desarrollo integrado basado en un conjunto de herramientas creadas con el fin de hacerlo más productivo con R y Python.
- Incluye una consola y un editor de sintaxis visualmente resaltado, que admiten la ejecución directa de código.
- Cuenta con varias herramientas de trazado, historial, depuración y administración del espacio de trabajo.

Primeros pasos en R