UNIVERSIDAD DE CÓRDOBA

ANÁLISIS DE ENCUESTAS

TÉCNICAS MULTIVARIANTES

1. Introducción
2. Clasificación de las técnicas
3. Etapas de análisis
4. Supuestos básicos
5. Valores perdidos y anómalos

Definición.

- Conjunto de **métodos** estadísticos cuya finalidad es **analizar simultáneamente** conjuntos de **datos multivariantes**: hay varias variables medidas para cada caso.
- Permiten un **mejor entendimiento del fenómeno** objeto de estudio, obteniendo información que los métodos univariantes y bivariantes son incapaces de conseguir.

Objetivos.

- Proporcionar **métodos para estudiar datos multivariantes** que el análisis estadístico uni y bidimensional es incapaz de conseguir.
- **Ayudar al investigador a tomar decisiones óptimas en el contexto en el que se encuentre** teniendo en cuenta la información disponible por el conjunto de datos analizado.

3 grupos:

- Métodos de **dependencia**
- Métodos de **interdependencia**
- Métodos **estructurales**

Métodos de dependencia:

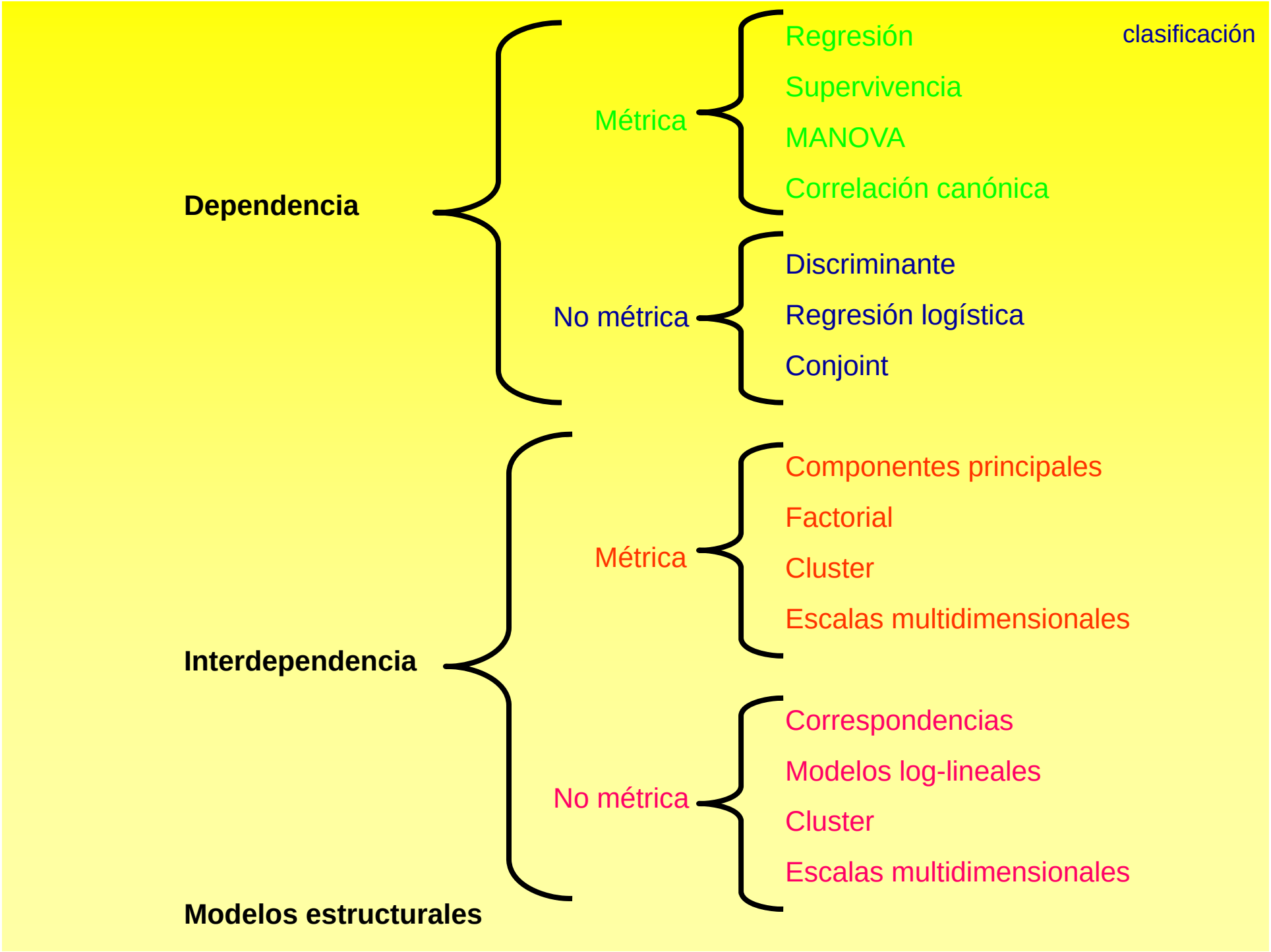
- Suponen que las variables analizadas están divididas en dos grupos: las variables **dependientes** y las variables **independientes**.
- El **objetivo** consiste en **determinar** si el conjunto de variables **independientes** **afecta** al conjunto de variables **dependientes** y de **qué forma**.

Métodos de interdependencia:

- No distinguen entre variables dependientes e independientes y su **objetivo** consiste en **identificar qué variables están relacionadas, cómo lo están y por qué**.

Métodos estructurales:

- Suponen que las variables están divididas en dos grupos: el de las variables **dependientes** y el de las **independientes**.
- El objetivo es analizar **como las variables independientes afectan a las variables dependientes y las relaciones de las variables de los dos grupos entre sí**.


clasificación
Dependencia
Métrica
Regresión
Supervivencia
MANOVA
Correlación canónica
No métrica
Discriminante
Regresión logística
Conjoint
Interdependencia
Métrica
Componentes principales
Factorial
Cluster
Escalas multidimensionales
No métrica
Correspondencias
Modelos log-lineales
Cluster
Escalas multidimensionales
Modelos estructurales

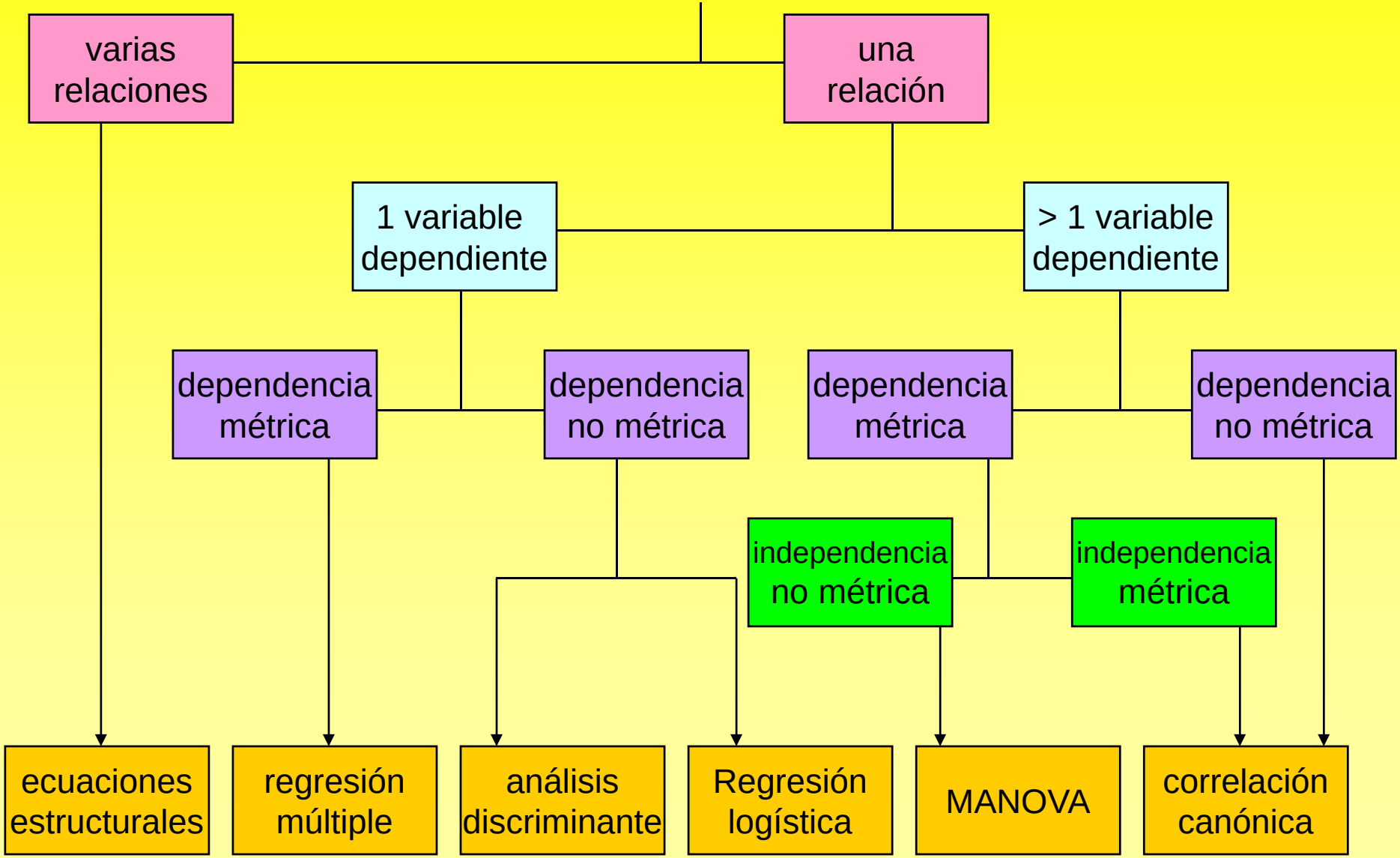
¿La investigación responde a un problema de dependencia entre variables o de interdependencia de las mismas?

¿Cómo están medidas las variables: métricas o no métricas?

Si es un problema de dependencias,

¿cuántas variables dependientes existen?

Análisis de dependencias



Regresión lineal múltiple:

- Relación entre **1 variable dependiente métrica** y **varias variables independientes métricas o no métricas**.

$$Y_1 \leftarrow (X_1, X_2, X_3, ... X_m)$$

- Por ejemplo: Determinar si existe o no relación entre el **resultado neto** y la **superficie**, **dimensión** e **inversión** inicial.

¿Y si el resultado neto está codificado en Pérdidas = 0, y Ganancias = 2?

- **Análisis discriminante.** Proporciona **reglas de clasificación** óptimas de nuevas observaciones de las que se desconoce su grupo de procedencia basándose en la **información proporcionada los valores que en ella toman las variables independientes**.

- **Modelos de regresión logística.** Se utilizan como una alternativa al **análisis discriminante cuando no hay normalidad**.

Análisis de correlación canónica:

clasificación

$Y_1 \leftarrow (X_1, X_2, X_3, ... X_m)$ **regresión, discriminante y logistica**

$(Y_1, Y_2, Y_3, ... Y_m) \leftarrow (X_1, X_2, X_3, ... X_m)$ **correlación canónica**

- Asociación lineal entre un conjunto de variables dependientes y otro de variables independientes.
  - Si la dependencia es no métrica
  - Si la dependencia es métrica → sólo si la independencia también lo es

- Por ejemplo:
  - Determinar si existe o no relación entre el **resultado neto** y la **producción de contaminantes** de una explotación con la **superficie**, **dimensión** e **inversión** inicial. **Corr. canónica**
  - Determinar la misma relación pero con el **género del ganadero** y el **tipo de explotación** (intensivo, extensivo) **MANOVA**

Ecuaciones estructurales:

- Varias relaciones: estructuras de la covarianza y análisis factorial confirmatorio

$$Y_1 \leftarrow (X_{11}, X_{12}, X_{13}, ... X_{1m})$$

$$Y_2 \leftarrow (X_{21}, X_{22}, X_{23}, ... X_{2m})$$

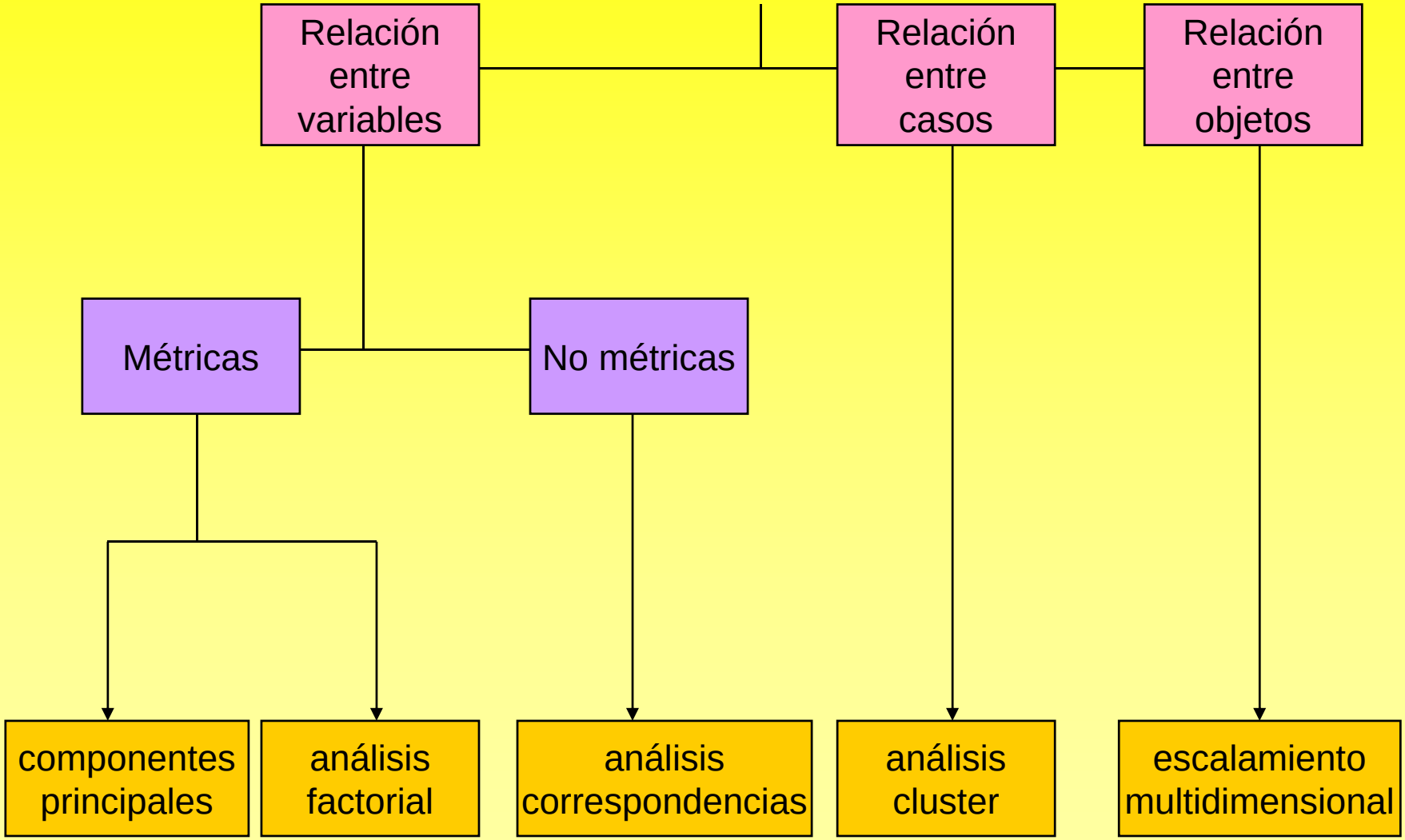
$$Y_3 \leftarrow (X_{31}, X_{32}, X_{33}, ... X_{3m})$$

Análisis de interdependencia.

- Las variables no se pueden separar en dependientes e independientes.

- **Objetivo**: determinar cómo y por qué las variables están correlacionadas.

Análisis de interdependencias



Análisis de componentes principales.

- Técnica de **reducción de datos**.
- **Objetivo**: construir combinaciones lineales de las variables iniciales que expliquen la mayor parte de la información contenida en esas variables.
- Esas combinaciones se denominan **Componentes Principales**, están **incorrelacionados** y cada componente sucesivo **explica menos varianza**.
- Por ejemplo: para comparar 10 explotaciones, es mejor utilizar 5 Componentes Principales que 150 variables

Análisis factorial.

- Técnica de **reducción de datos**.
- **Objetivo**: establecer qué causas latentes (factores) causan la correlación entre las variables observadas.

- Por ejemplo: el **desarrollo** de un animal no se puede medir directamente, pero sí es posible medir algunos de sus indicadores:
  - El peso y su incremento
  - La alzada y sus incrementos (cruz, palomillas, etc.)
  - Las dimensiones de algunas regiones corporales y su relación respecto a otras
  - **El análisis factorial establecería que el factor "desarrollo" explica todas estas variables y cómo se relaciona cada variable con el factor**

Análisis de correspondencias.

- Permite visualizar gráficamente **tablas de contingencia**.
- Por ejemplo: Si existe relación entre la formación del ganadero y el tipo de gestión
  - Formación: sin formación, primaria, bachillerato, universidad, formación profesional, módulos, escuela de capataces o sus combinaciones
  - Tipo de gestión: ninguna, asesor fiscal, veterinario clínico, agrónomo, veterinario asesor o sus combinaciones

Análisis de escalamiento multidimensional.

- Permite aflorar los **criterios** que utilizan los individuos para **considerar que distintos objetos son parecidos o distintos**.
- Por ejemplo: Estudios de preferencia del jamón curado

Análisis de conglomerados (cluster).

- A diferencia del factorial que agrupa variables, pretende **agrupar observaciones**.
- De tal modo que las observaciones **dentro de los grupos** sean parecidas respecto a las variables utilizadas para agrupar.
- Y que las observaciones **entre los grupos** sean lo más diferentes posibles respecto a las mismas variables.

- Por ejemplo, para hacer grupos de animales en un programa de mejora genética, o de explotaciones de cara a optimizar su gestión.

Proceso de aplicación de la técnica multivariante.

1. Definir el problema que se está investigando (**modelo conceptual**)
   - Análisis conceptual de su objeto de estudio
   - Identificar las **relaciones** fundamentales **que se van a abordar**
   - Elección de la **técnica** a aplicar

Proceso de aplicación de la técnica multivariante.

Por ejemplo:

Analizar la gestión de los sectores ecológicos

- Relaciones entre las variables y los casos:
  - AF/ACP + ANOVA
  - Correlación canónica
- Dependencia de una variable y las demás:
  - Regresión logística
  - AF/ACP + Discriminante

2. Desarrollo del **plan de análisis**
    - Tamaño **muestral** mínimo para la técnica concreta
    - Las **escalas** de las variables a analizar son correctas

3. Condiciones de **aplicabilidad de la técnica** elegida

4. **Desarrollo de la técnica**, incorporando o eliminando variables según la bondad de ajuste

5. **Interpretación** de los resultados
    - Interpretar el modelo global
    - Analizar las variables individuales: cargas factoriales, coeficientes, varianzas, etc.
    - La interpretación retroalimenta al paso 4

6. **Validación del modelo.** Técnicas de diagnóstico que permitan generalizar los resultados a la población.

Condiciones de aplicabilidad.

- Normalidad
- Homocedasticidad
- Linealidad
- Independencia

Análisis multivariante de la normalidad.

- Existen **pocos contrastes** (Mardia-curtosis y Mardia-apuntalamiento) y **no se conoce bien su distribución**
- También el **gráfico chi-cuadrado**:
  - Se calculan las distancias de Mahanalobis (**D**)
  - Su cuadrado se ordenan de menor a mayor (**D^2**)
  - En cada distancia se calcula su percentil **(j-0,5)/n**
  - Se calculan los valores **X^2** de los percentiles de una distribución X^2 con p grados de libertad (**p=número de variables estudiadas**)
  - Se representan **D^2 y X^2**
- Con Statgraphics se utiliza "**Multivariate Control Chart**"
- **La variable o variables problemáticas se pueden transformar o eliminar**

Homoscedasticidad (univariante):

- **Contraste de Levene** (hipótesis nula: la varianza de la variable X es igual en todos los niveles que forma la variable Z)

Homoscedasticidad (multivariante):

- **Contraste M de Box**
  - Es **muy sensible** (se recomienda que p<0,001)
  - **Es necesaria normalidad multivariante** para el contraste

Por ejemplo:

- Estudiar si los **ganaderos son conscientes** de que la **producción intensiva perjudica el medio ambiente**
- **O por el contrario**, **los ganaderos intensivos lo son porque no son conscientes de esto**
- Si esto es así, los **ganaderos intensivos estarían significativamente más en desacuerdo con la afirmación que los extensivos**

Por ejemplo:

- Esto es un problema de **análisis discriminante**:
  - **Una variable dependiente no métrica** (intensivo o extensivo)
  - **Varias variables independientes métricas**:
    - Y1: Opinión (1 a 5): la **g.intensiva perjudica** el m. ambiente
    - Y2: Opinión (1 a 5): **no permitir g.intensiva** en espacios protegidos y naturales
    - Y3: Opinión (1 a 5): **reducir ayudas** a g.intensiva U.E.
    - Y4: Opinión (1 a 5): **debe informarse más** sobre los efectos de la g. Intensiva a la opinión pública

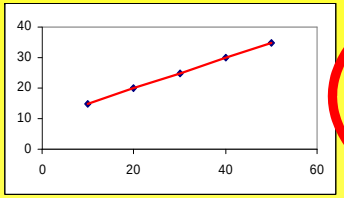
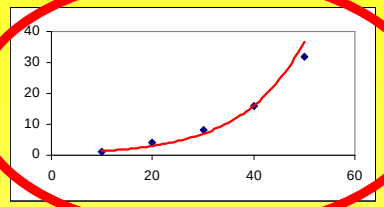
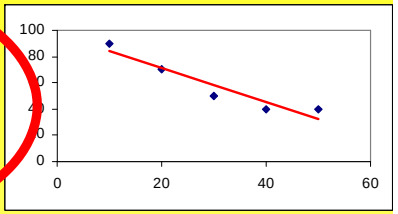
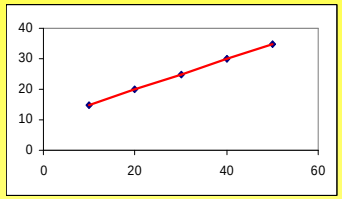
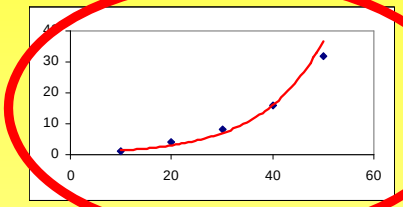
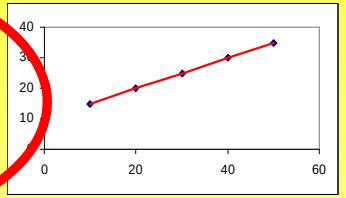
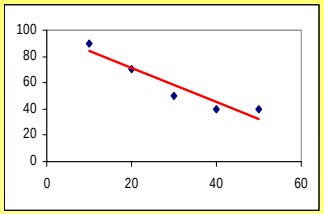
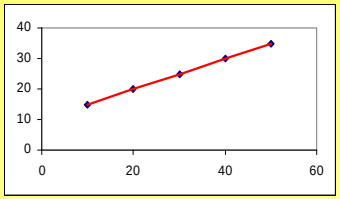
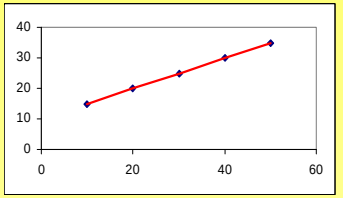
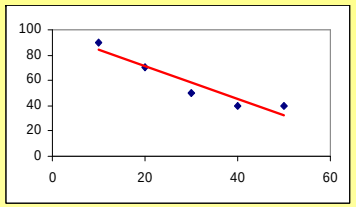
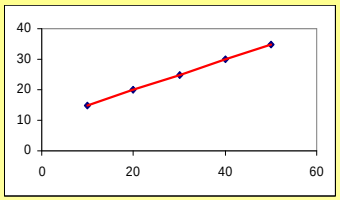
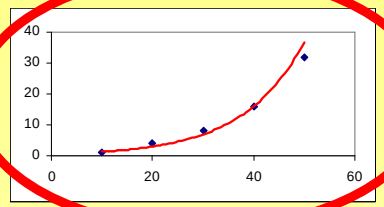
Por ejemplo:

- Debe comprobarse la hipótesis nula, que la matriz de varianzas-covarianzas de las variables Y es la misma para los niveles de X (intensivo-extensivo).
- **Contraste M de Box**.

Linealidad:

- Fundamental en todas las técnicas que se centren en el análisis de las **matrices de correlaciones o de covarianzas**
- Porque el coeficiente de **correlación de Pearson** sólo puede captar relaciones lineales
- Para la regresión lineal múltiple se analizan los residuos
- Para el resto de los casos: **gráficos de dispersión bivariante**
- **Por ejemplo:**

	consumo	inc. Peso	inc. Diám.	Digest
Consumo MS (kg/animal)	1	0,87	0,91	-0,66
Incremento de Peso	0,87	1	0,79	0,81
Incremento de Diámetro	0,91	0,79	1	0,92
Digestibilidad MS (%)	-0,66	0,81	0,92	1

	consumo	inc. Peso	inc. Diám.	Digestibilidad
Consumo				
Inc Peso				
Inc Diam				
Diges				

Independencia:

- Los valores que toman las variables en un caso no están influidos por los valores que toman en otro caso
- Si no se está seguro de esto, habría que **incrementar el nivel de significación de los contrastes 10 veces** (de $p<0,05$ a $p<0,005$)
- La independencia se asegura en el diseño experimental