Universidad
Rey Juan Carlos
Dpto. Economía Aplicada I

I. Métodos Estadísticos y Construcción de Modelos

DR. FRANCISCO RABADÁN PÉREZ

ESTADÍSTICA SUPERIOR

Índice

1. Definición AMV
2. Objetivos AMV.
3. Estructura
4. Software (Data Mining y lenguajes)
5. Conceptos previos:
    1. Valor teórico
    2. Principios de Validez y Fiabilidad.
    3. Error de medida
    4. Significación vs. Potencia
    5. P-valor
6. Selección de la técnica AMV (Hair)

Universidad Rey Juan Carlos
Dpto. Economía Aplicada I

I.1. El análisis de datos multivariante

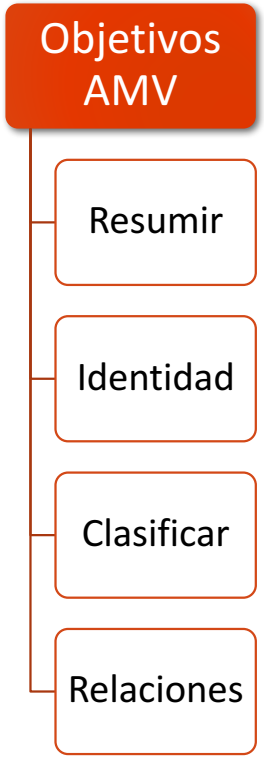
AMV es necesario para describir/comprender una realidad compleja.

Definición amplia: Todos los métodos estadísticos que analizan simultáneamente varias variables para cada individuo u objeto de investigación (Hair,p.5)

Definición estricta: las variables deben ser aleatorias y estar interrelacionadas de tal forma que sus diferentes efectos no puedan ser interpretados separadamente con algún sentido (Hair, p.5)

Universidad Rey Juan Carlos
Dpto. Economía Aplicada I

I.1. El análisis de datos multivariante



AMV: estudio estadístico de multitud de variables medidas en elementos de una población con los siguientes **objetivos** (Peña, p.2):

- **Resumir datos** mediante un pequeño conjunto de nuevas variables (Parsimonia).
- **Identidad**: Encontrar grupos homogéneos, si existen.
- **Clasificar** nuevas observaciones (inferencia)
- **"Causalidad"-interacción**: Relacionar conjuntos de variables

I.1. El análisis de datos multivariante

Objetivos AMV

- Resumir
- Identidad
- Clasificar
- relaciones

Se busca simplificar la descripción con pocas variables indicadoras.

Ventajas:

- Representación gráfica
- Comparar conjuntos de datos
- Comparar en distintos instantes de tiempo.
- Mejorar nuestro conocimiento/comprensión de la realidad.

I.1. El análisis de datos multivariante

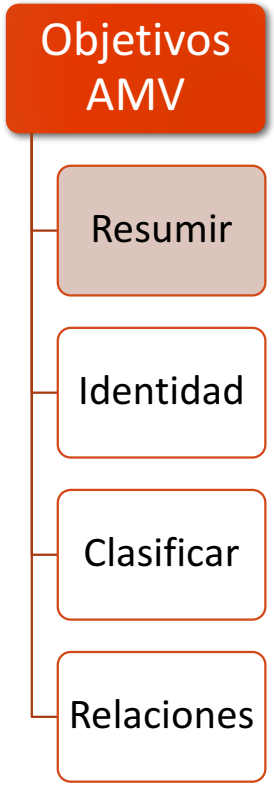
Objetivos AMV

- Resumir
- Identidad
- Clasificar
- Causalidad

Fases en el examen de datos

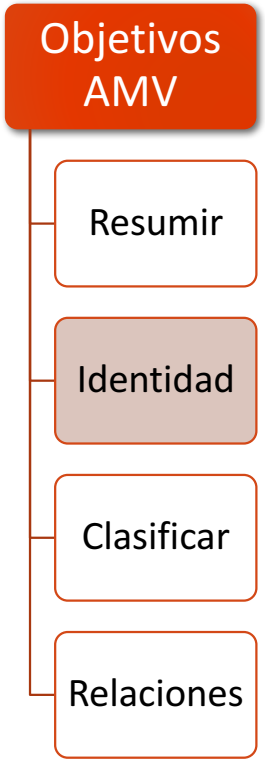
1. Análisis gráfico
2. Identificación y evaluación de **valores perdidos** (*missing values*).
3. Identificación y tratamiento de los **valores atípicos** (*outliers*).
4. Verificar si se cumplen las **hipótesis exigidas** por la técnica multivariante.
5. **Comprensión preliminar** primera descripción de los datos y del comportamiento de las variables observadas.

I.1. El análisis de datos multivariante



- Tipos de variables resumen:
  - **Variables latentes:** explican el comportamiento de un conjunto de variables observadas.
  - **De pertenencia**: asignan un caso a un grupo.
  - **Proxy**: Variable observada que representa a un conjunto de variables con un comportamiento similar.
  - **Índices**: Construidas "a priori" sobre una conjunto de variables para comparar / explicar su evolución.

I.1. El análisis de datos multivariante



- **Identidad** : Esperamos que existan individuos/casos lo suficientemente parecidos como para formar grupos con características homogéneas que a su vez se diferencien del resto.
- Construiremos tipologías.

I.1. El análisis de datos multivariante

- Objetivos AMV
  - Resumir
  - Identidad
  - Clasificar
  - Relaciones

Cuando aumenta el espacio muestral y queremos juzgar a que grupos pre-existentes pertenecen los nuevos casos.

Ejemplo: riesgo de cobro con nuevos clientes, perfiles de comportamiento, nichos de mercado,…

I.1. El análisis de datos multivariante

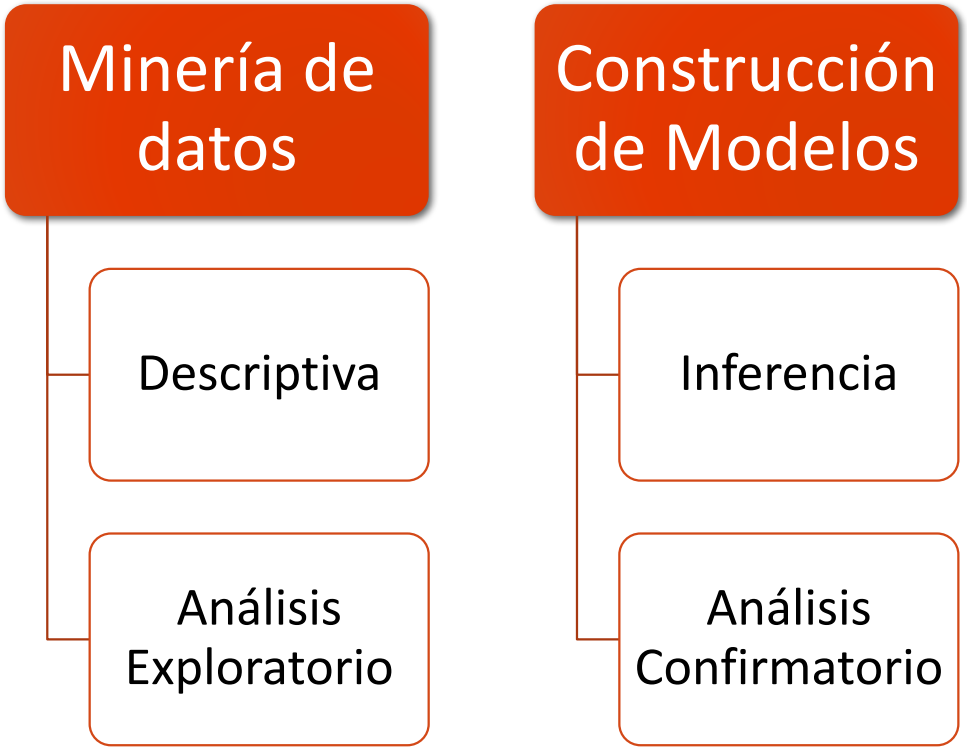
Objetivos AMV

- Resumir
- Identidad
- Clasificar
- Relaciones

Relacionar conjuntos de variables.

- Ej: capacidades intelectuales vs. resultados profesionales.
- Objetivo: descubrir nuevas dimensiones (latentes) en la relación de grupos de variables.
- También podemos comparar el mismo grupo de variables en distintos momentos del tiempo.

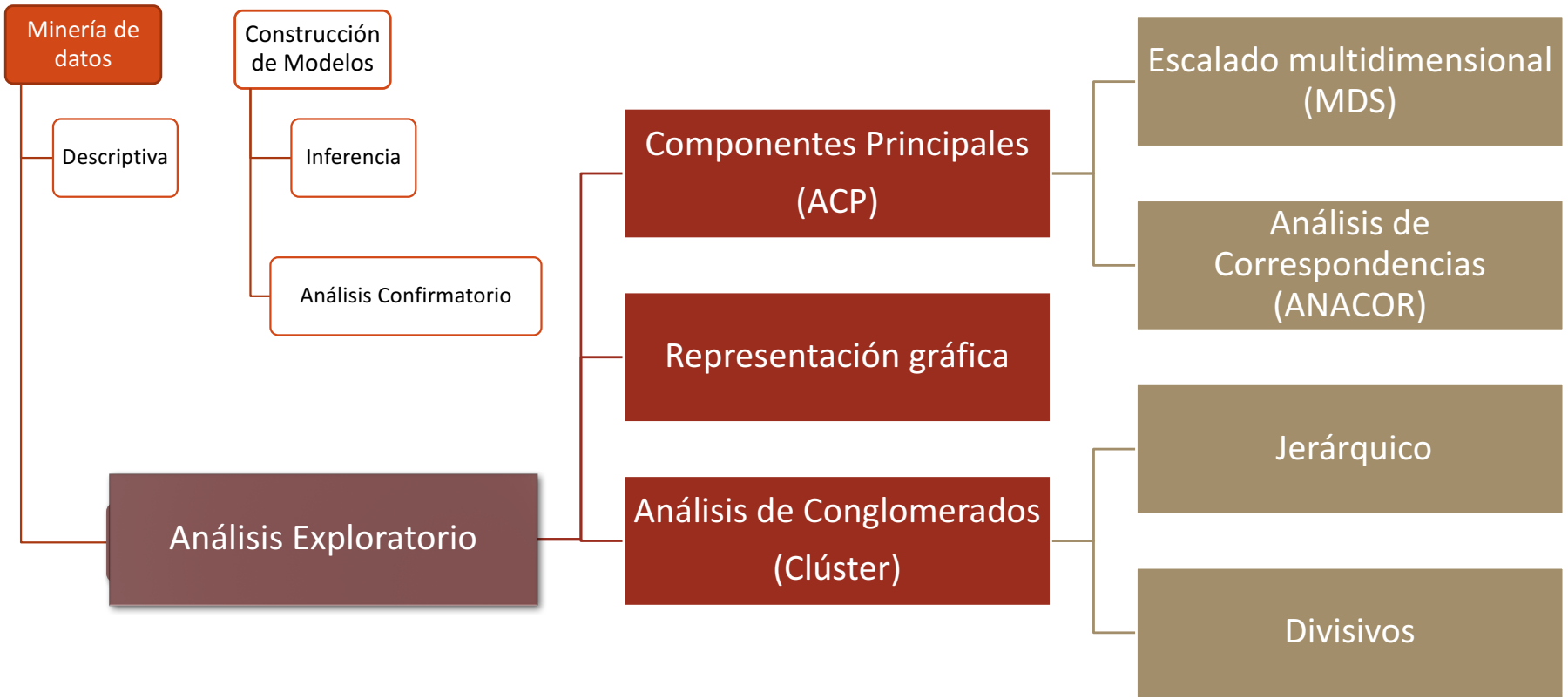
I.2. Estructura



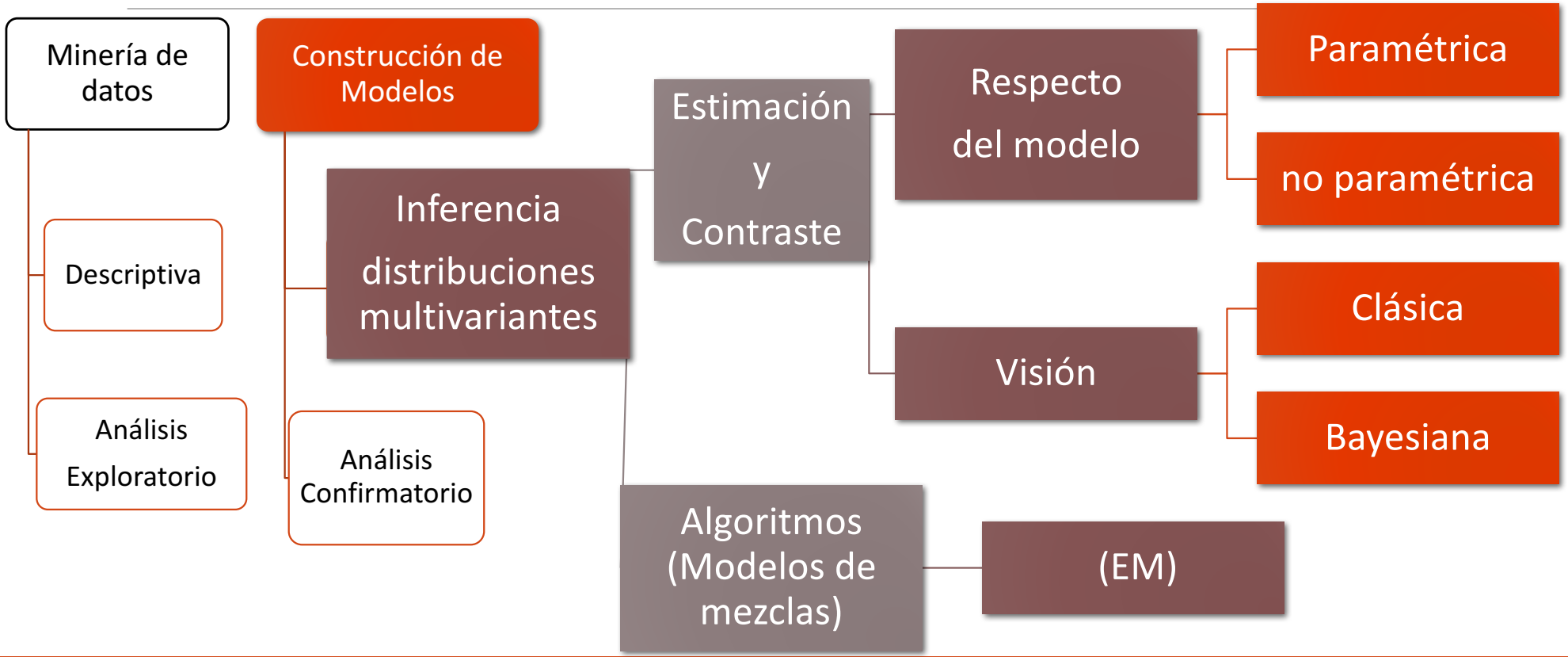
De la tabla multidimensional al algebra matricial:

- proyección ortogonal/oblicua,
- MCO,
- MCP…

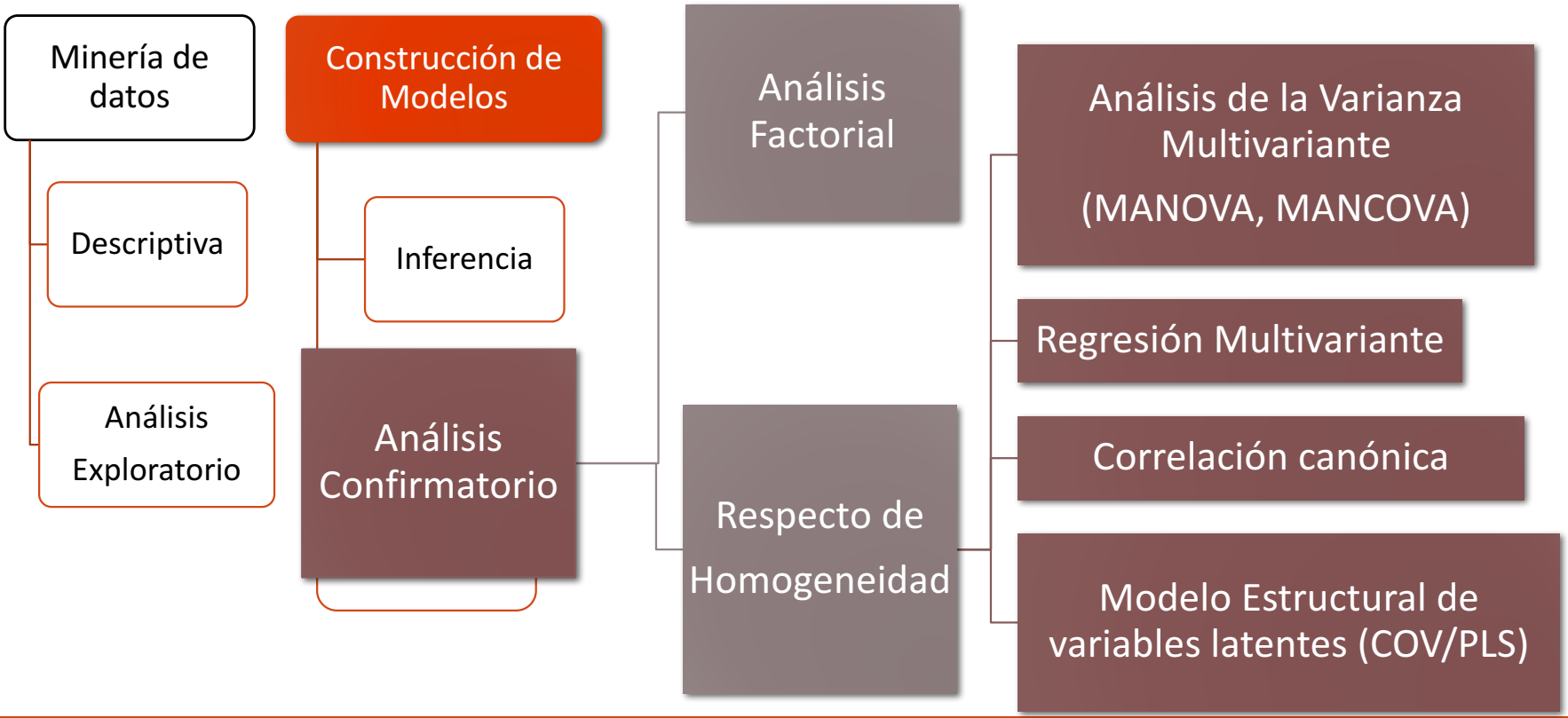
I.2. Estructura



I.2. Estructura



I.2. Estructura



I.2. Estructura

Objetivos	Descriptiva	Inferencia
Resumir datos	Descripción de datos	Construcción de Modelos
Obtener indicadores	• Componentes Principales (ACP) • Escalado Multidimensional • Análisis de Correspondencias (ANACOR)	Análisis Factorial (AF)
Clasificar	Análisis de conglomerados (Clúster)	Análisis Discriminante
Agrupar	Análisis de conglomerados	Clasificación con mezclas
Relacionar variables	• Correlación, Contingencia • Regresión Múltiple	• Correlación Canónica • Ecuaciones Estructurales (COV/PLS)

Fuente: Peña

Universidad Rey Juan Carlos
Dpto. Economía Aplicada I

I.3. Software

Objetivos	Público Objetivo	Licencia	Interface/ Programación
SPSS	Ciencias Sociales	Privativa	Interface
STATA	Empresa	Privativa	Interface
SAS	Empresa	Privativa/gratuita	Programación
R	General e Investigación	Open Source	Programación
Rstudio	Empresas/general	Privativa / Open Source	Programación
RKWard	General	Open Source	Interface
Smart PLS	Investigación/Empresa Ecuaciones Estructurales	Privativa	Interface
Matlab	Científicos e investigadores	Privativa	Lenguaje de alto nivel
Mathematica Wolfram	Matemáticos e investigadores	Privativa	Lenguaje de alto nivel
Orange 3 / Python	General y Data Mining	Free / comercial	Gui/python

I.4. AMV: Valor teórico

Combinación lineal de variables con ponderaciones determinadas empíricamente (Hair)

$$VT = \sum_{i=1}^{n} w_i x_i = \hat{y}_j$$

- X : Variables observadas
- w_i: ponderaciones AMV
- Y: Variables dependientes

- **Significado**: valor único que mejor se adapta al objeto de nuestra investigación.
  - ***Regresión Múltiple***: Y guarda la mejor correlación con X
  - ***Discriminante***: Y busca maximizar las diferencias entre grupos.
  - ***Factorial***: Y representa mejor la dimensionalidad

1.4. Error de medida

Validez

- Grado en que la medida representa con precisión lo que se supone que representa
- = "Precisión en la definición de la variable" + "pregunta correcta en cuestionario"

Fiabilidad

- CN: Validez
- Grado en que la variable observada mide el valor verdadero y está libre de error
- Estadísticos consistentes: Δ nº experimentos → Δ regularidad estadística

1.4. Error de medida

Grado en que los valores observados no son representativos de los valores verdaderos.

Hipótesis: todas las variables tienen observaciones con errores de medida

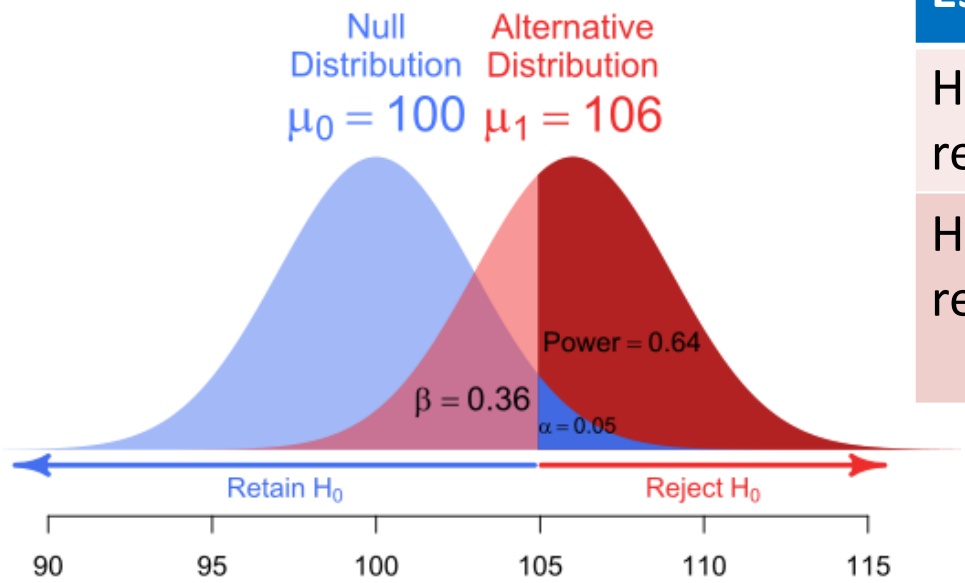
Causas:

- Imprecisión en el proceso de medida
- Mal diseño cuestionario
- Incapacidad del encuestado para contestar adecuadamente

Efecto: debilita las relaciones entre las variables

I.4. Significación estadística vs. Potencia del contraste

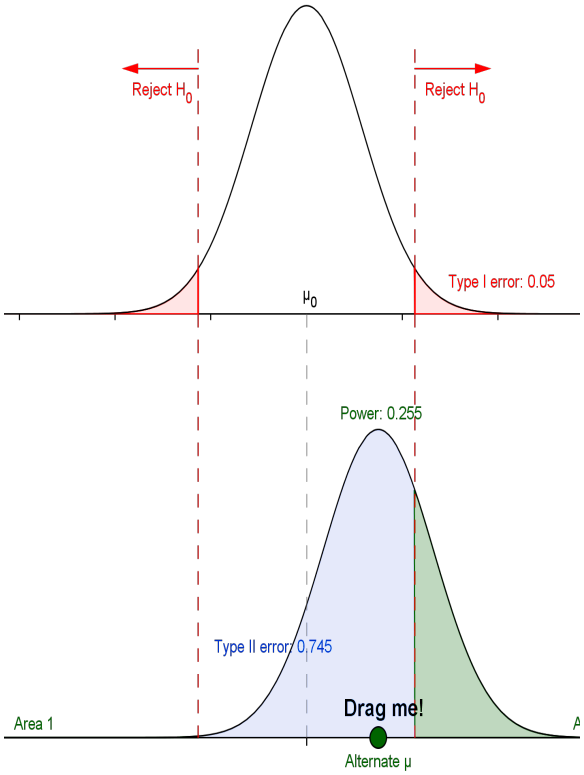
Universidad Rey Juan Carlos
Dpto. Economía Aplicada I



Decisión Estadística	Realidad	
	H_0: cierta	H_0: falsa
H_0: No rechazar	$\gamma = 1-\alpha$ nivel de confianza	β (error tipo II)
H_0: rechazar	α (error tipo I)	$1-\beta$ (potencia)

1-β: dicta la probabilidad de éxito en la búsqueda de las diferencias, si es que existen realmente

I.4. Factores que influyen en la potencia



Efecto Tamaño

- Diferencia entre hipótesis alternativas
- Se basa en la correlación efectiva

Nivel de Significación

- $\nabla\alpha \rightarrow \nabla(1-\beta)$
- Se reduce la posibilidad de encontrar un efecto negativo significativo

Tamaño muestral (n)

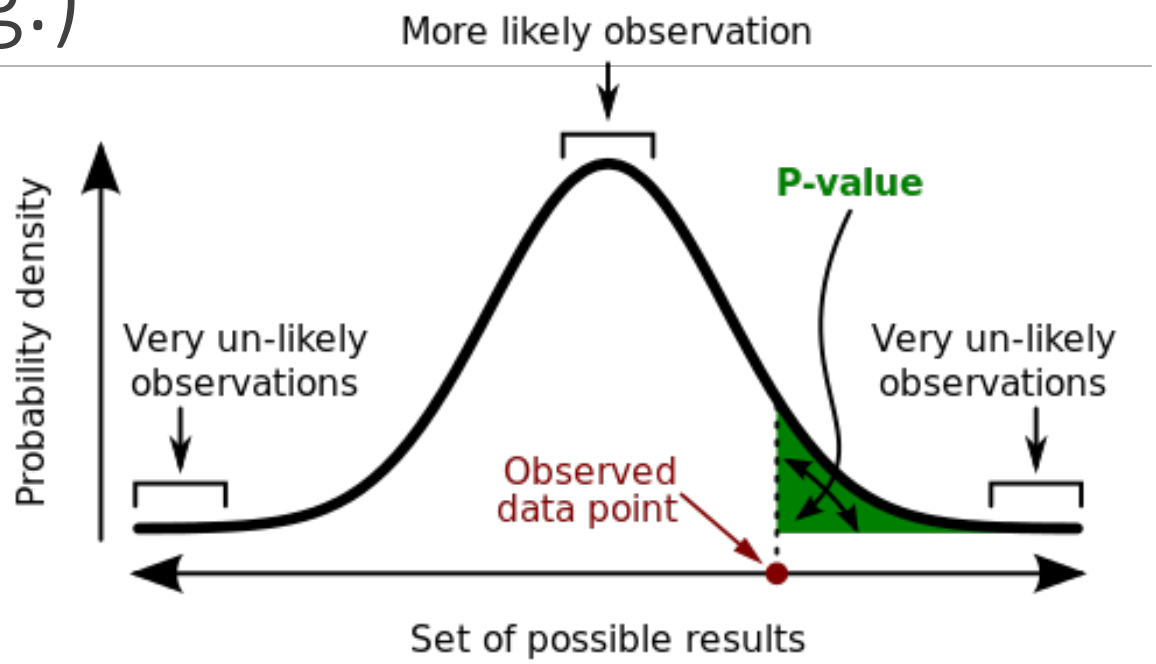
- Δn--> puede producir demasiada potencia (los efectos pequeños serán cada vez más significativos).

1.4. p-valor (sig.)

La falacia de la condición transpuesta

$$P\left(\hat{y}/_{H_0}\right) \neq P\left(H_0/_{\hat{y}}\right)$$

Usar el p-valor como un resultado cierto a nivel poblacional es un error atroz desde el punto de vista lógico



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

1.5. Selección técnica AMV

ES - MÉTODOS ESTADÍSTICOS Y CONSTRUCCIÓN DE MODELOS

Regresión múltiple (explica la relación lineal).

Sirve para explicar el comportamiento de una única variable dependiente en función del comportamiento de otras.

Sólo sirve para predecir si las variables independientes son indicadores adelantados (ARIMA, alisado exponencial…).

Análisis discriminante múltiple (clasifica).

Sirve para clasificar en relación a una variable dependiente categórica (no métrica) construida a partir de un conjunto de variables independientes cuantitativas y continuas.

Predice la verosimilitud de una entidad respecto de un conjunto de identidades estadísticas (categorías) expresadas en funciones discriminantes.

1.5. Selección técnica AMV

ES - MÉTODOS ESTADÍSTICOS Y CONSTRUCCIÓN DE MODELOS

Regresión logística (clasifica).

Cumple la misma función que el discriminante, pero con distinta metodología.

MANOVA (Análisis Multivariante de la varianza) y **MANCOVA** (Análisis Multivariante de la la Varianza que incluye la Covarianza).

- ANOVA: compara la media de una variable de distintas poblaciones simultáneamente bajo el supuesto de igualdad de varianzas.
- MANOVA: compara varios valores medios de distintas variables en distintas muestras simultáneamente. Las variables independientes son categóricas.
- MANCOVA = MANOVA + COVARIABLES.

1.5. Selección técnica AMV

ES - MÉTODOS ESTADÍSTICOS Y CONSTRUCCIÓN DE MODELOS

Análisis conjunto.

Se busca ordenar las preferencias respecto a un conjunto de objetos partiendo de información múltiple de carácter cualitativo (atributos).

Correlación canónica (H. Hotteling).

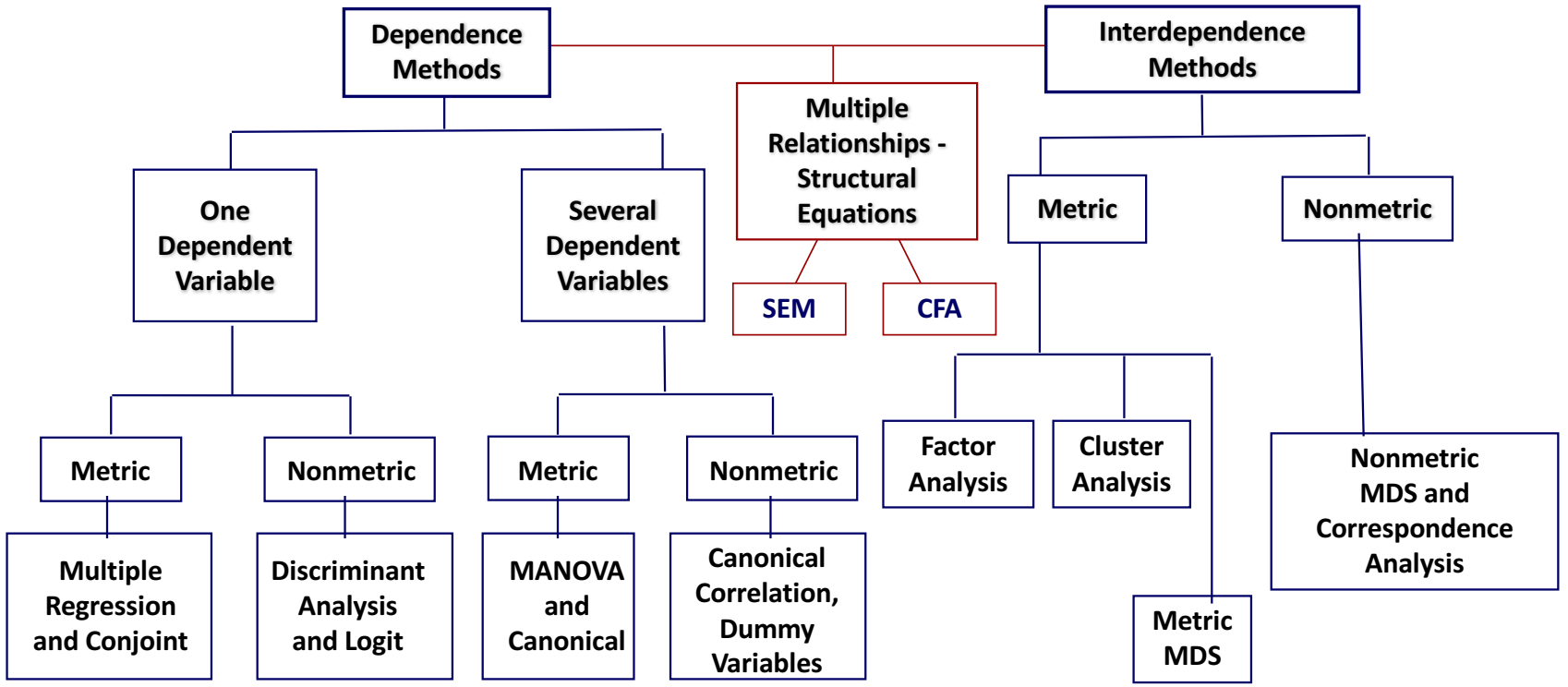
Busca relaciones entre dos grupos de variables y la validez de las mismas. Se diferencia de la Regresión múltiple en que hay muchas dependientes.

Modelos de ecuaciones estructurales (SEM).

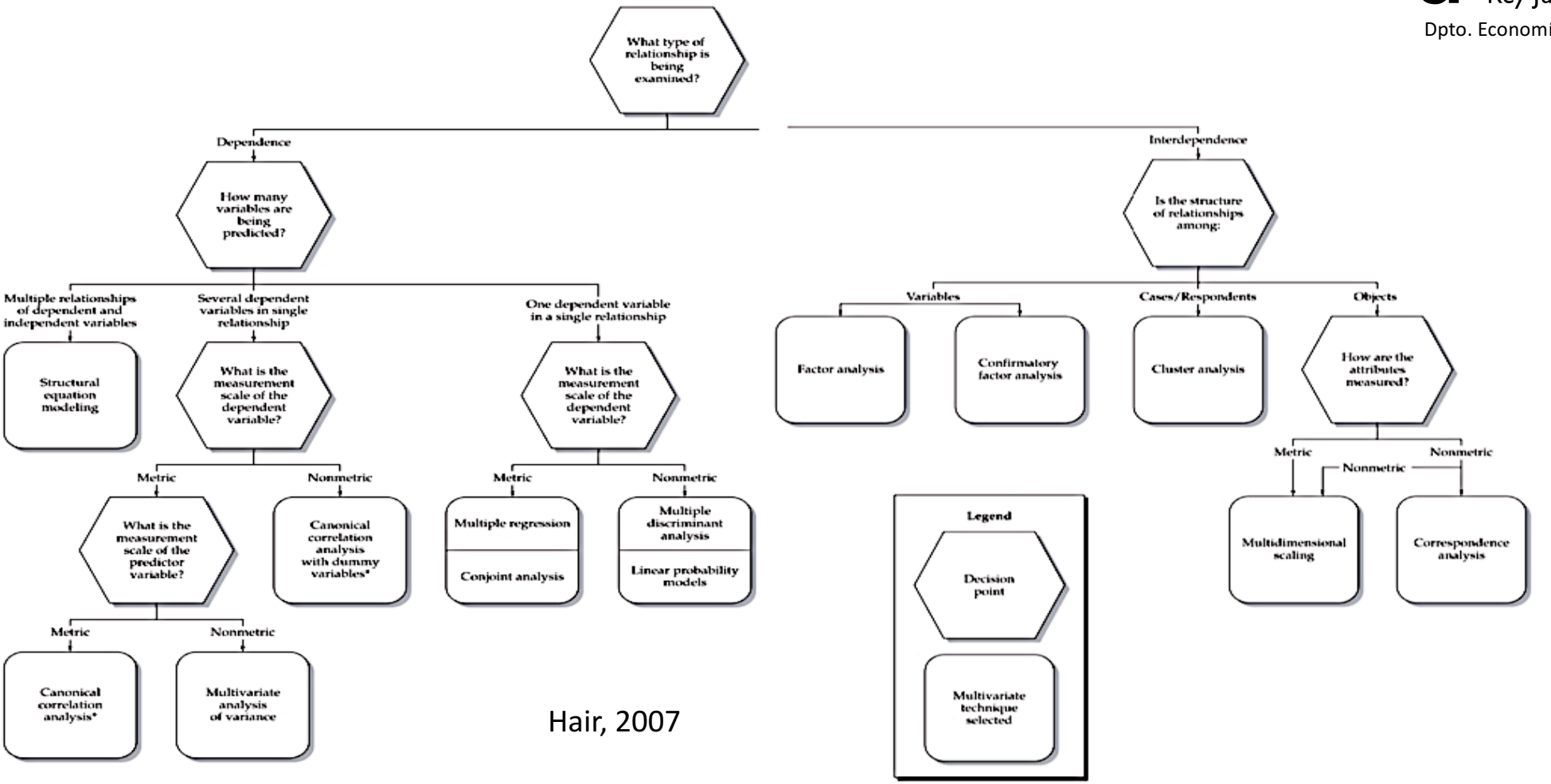
Estima múltiples relaciones de dependencia e interdependencia con los componentes: modelo de medida y modelo estructural. Las ecuaciones estructurales es el método confirmatorio por excelencia.

1.5. Selección técnica AMV

ES - MÉTODOS ESTADÍSTICOS Y CONSTRUCCIÓN DE MODELOS



Hair, 2007


What type of relationship is being examined?
Dependence
How many variables are being predicted?
Multiple relationships of dependent and independent variables
Structural equation modeling
Several dependent variables in single relationship
What is the measurement scale of the dependent variable?
Metric
What is the measurement scale of the predictor variable?
Metric
Canonical correlation analysis*
Nonmetric
Multivariate analysis of variance
Nonmetric
Canonical correlation analysis with dummy variables*
One dependent variable in a single relationship
What is the measurement scale of the dependent variable?
Metric
Multiple regression
Conjoint analysis
Nonmetric
Multiple discriminant analysis
Linear probability models
Interdependence
Is the structure of relationships among:
Variables
Factor analysis
Confirmatory factor analysis
Cases/Respondents
Cluster analysis
Objects
How are the attributes measured?
Metric
Multidimensional scaling
Nonmetric
Correspondence analysis
Nonmetric
Legend
Decision point
Multivariate technique selected

Hair, 2007

Fuentes

1. Daniel Peña (2002) Análisis de datos multivariantes, McGrawHill, Madrid, Tema 2, p.1-12.
2. Hair et al (2007) Análisis Multivariante, Pearson Prentice Hall, Madrid.p.1-28
3. my.ilstu.edu/~wjschne/138/Psychology138Lab14.html
4. https://upload.wikimedia.org/wikipedia/commons/3/3a/P-value_in_statistical_significance_testing.svg
5. http://www.melbapplets.ms.unimelb.edu.au/?portfolio=power-of-a-hypothesis-test