Análisis Factorial Confirmatorio

Análisis Avanzado de Datos II

Gabriel Sotomayor, Basado en material preparado por Anais Herrera Leighton

2025-06-02

Medición en Ciencias Sociales

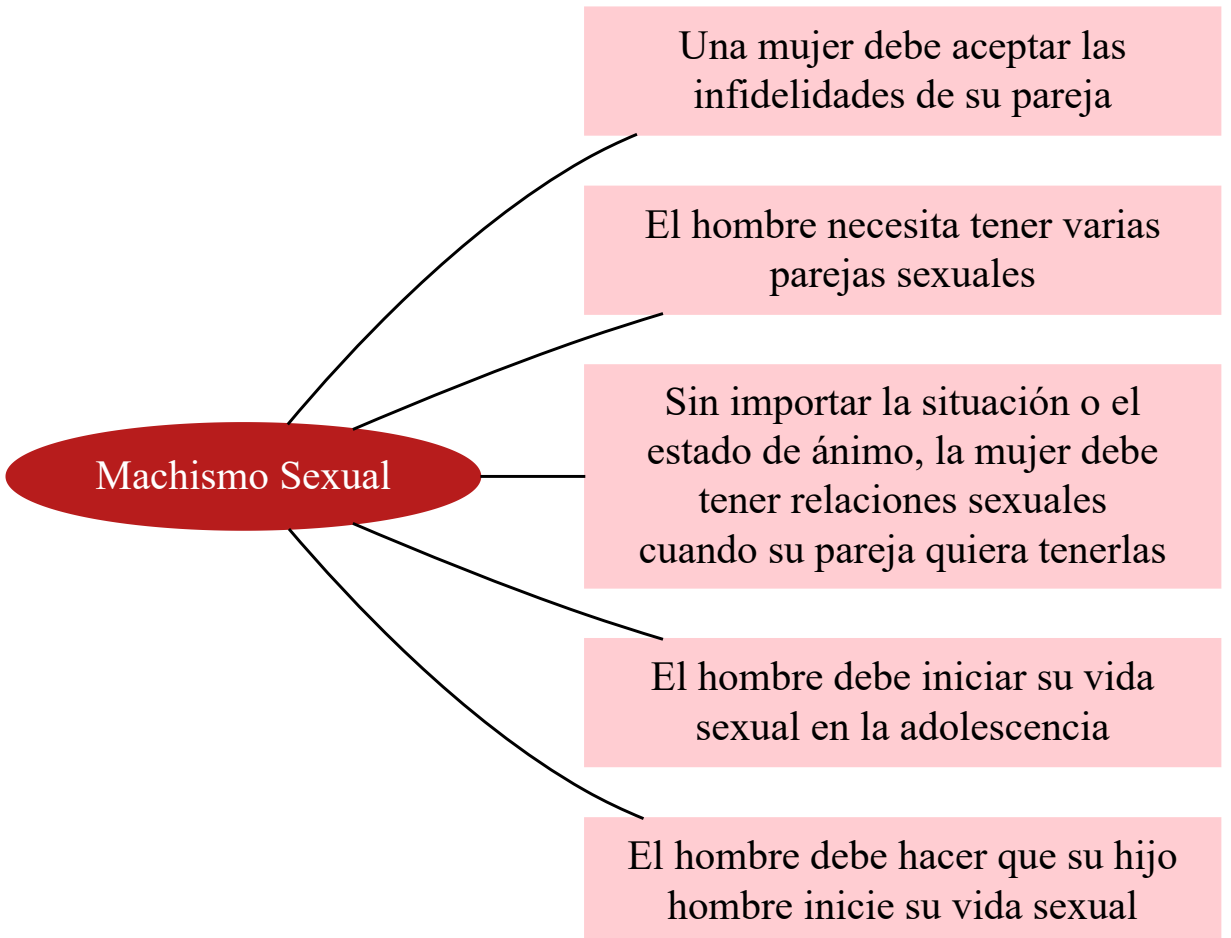
- En ocasiones un indicador puede ser suficiente para capturar el concepto que se quiere medir
  - Variables observadas (ej. edad, sexo)
- Pero algunos conceptos no pueden ser medidos directamente, para lo cual requerimos distintos indicadores
  - Variables latentes (ej. Actitudes machistas, clase social)

Medición en Ciencias Sociales

Ejemplo de medición de una variable latente (unidimensional): Escala de Machismo Sexual (Díaz, Rosas & González, 2010).

- Expresa en tu opinión tu grado de acuerdo o desacuerdo con las siguientes frases. Por favor responde honestamente utilizando estas opciones: (1) Totalmente en desacuerdo; (2) En desacuerdo; (3) Sin opinión; (4) De acuerdo; (5) Totalmente de acuerdo
  - Una mujer debe aceptar las infidelidades de su pareja
  - El hombre necesita tener varias parejas sexuales
  - Sin importar la situación o el estado de ánimo, la mujer debe tener relaciones sexuales cuando su pareja quiera tenerlas
  - El hombre debe iniciar su vida sexual en la adolescencia
  - El hombre debe hacer que su hijo hombre inicie su vida sexual

Medición en Ciencias Sociales



Medición en Ciencias Sociales

Ejemplos de variables latentes (con más de una dimensión)

- Actitudes meritocráticas
  - Percepciones meritocráticas
  - Preferencias meritocráticas
- Actitudes hacia la violencia de carabineros
  - Justificación uso de violencia para disolver marchas
  - Justificación uso de violencia al allanar comunidades mapuche
- Inteligencia
  - Habilidades verbales
  - Habilidades matemáticas

Medición en Ciencias Sociales

Ejemplos de variables latentes (con más de una dimensión): Actitudes meritocráticas (Castillo, Iturra, Meneses & Maldonado, 2021)

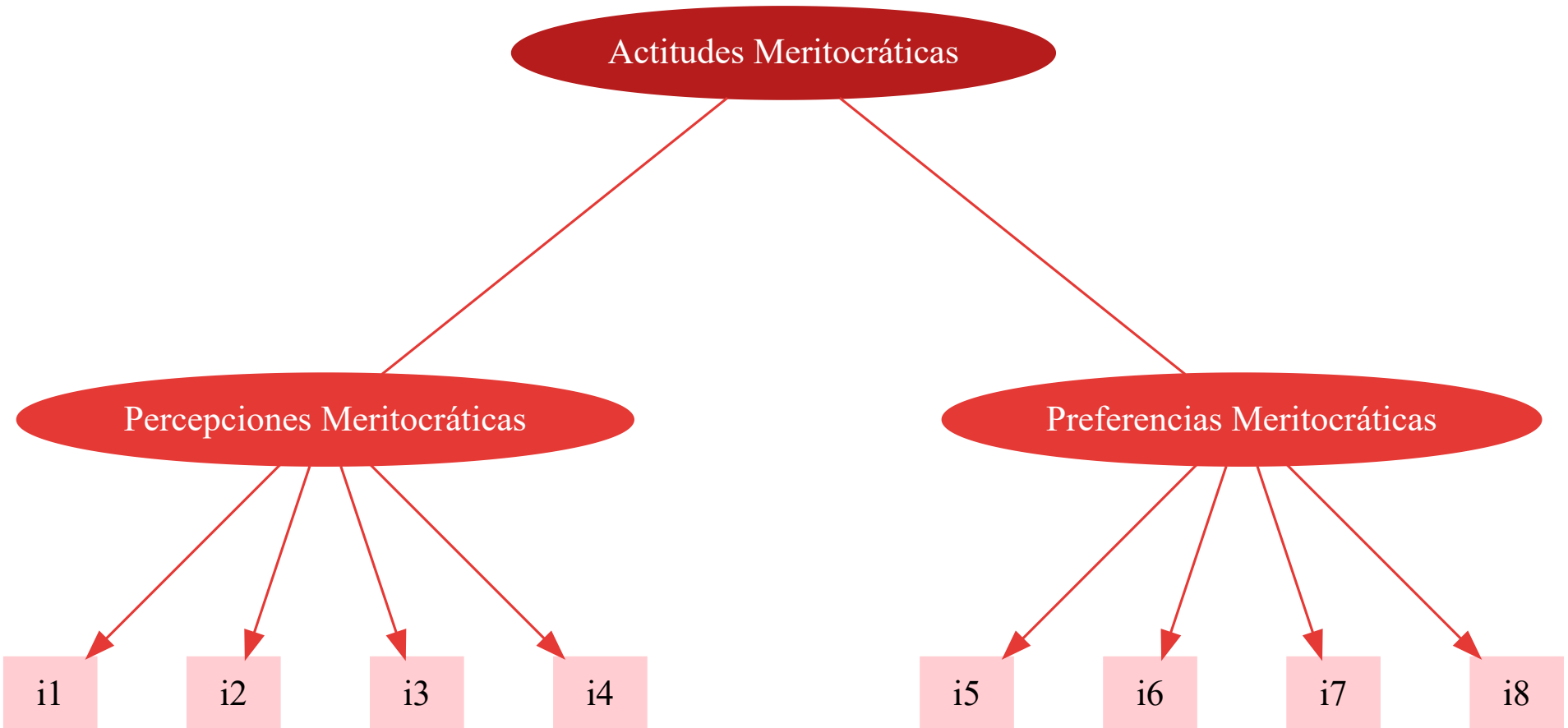
Dimensión	Subdimensión	Indicador
Percepción	Meritocrática	Quienes más se esfuerzan logran obtener mayores recompensas que quienes se esfuerzan menos.(i1)
		Quienes poseen más talento logran obtener mayores recompensas que quienes poseen menos talento. (i2)
	No meritocrática	Quienes tienen padres ricos logran salir adelante. (i3)
		Quienes tienen buenos contactos logran salir adelante. (i4)

Medición en Ciencias Sociales

Ejemplos de variables latentes (con más de una dimensión): Actitudes meritocráticas (Castillo, Iturra, Meneses & Maldonado, 2021)

Dimensión	Subdimensión	Indicador
Preferencia	Meritocrática	Quienes más se esfuerzan deberían obtener mayores recompensas que quienes se esfuerzan menos. (i5)
		Quienes poseen más talento deberían obtener mayores recompensas que quienes poseen menos talento. (i6)
	No meritocrática	Está bien que quienes tienen padres ricos salgan adelante. (i7)
		Está bien que quienes tienen buenos contactos salgan adelante (i8)

Medición en Ciencias Sociales



Medición en Ciencias Sociales

- Los conceptos complejos no se pueden medir directamente, por lo que se realiza una operacionalización
  - Variable latente o factor
- Generamos mediciones que se aproximan a medir lo que representa el concepto
  - Variables observadas, indicadores o ítems
- No observamos directamente la variable latente, si no que esta es deducida a partir de las correlaciones entre las variables observadas

Medición de variables latentes

- Podemos utilizar modelos estadísticos para entender constructos sociales y responder preguntas cómo
  - ¿Cómo se relacionan entre sí distintos indicadores de un mismo concepto?
  - ¿Son las variables adecuadas para capturar un determinado concepto?
  - ¿Cuántas dimensiones tiene un concepto?
- Análisis factorial (variables continuas)
  - Permite estudiar la interrelación (o interdependencia) de variables observadas
  - Se agrupan las variables en un factor o en un número reducido de factores
  - Cada indicador tiene una varianza común que es explicada por el factor latente y una varianza única

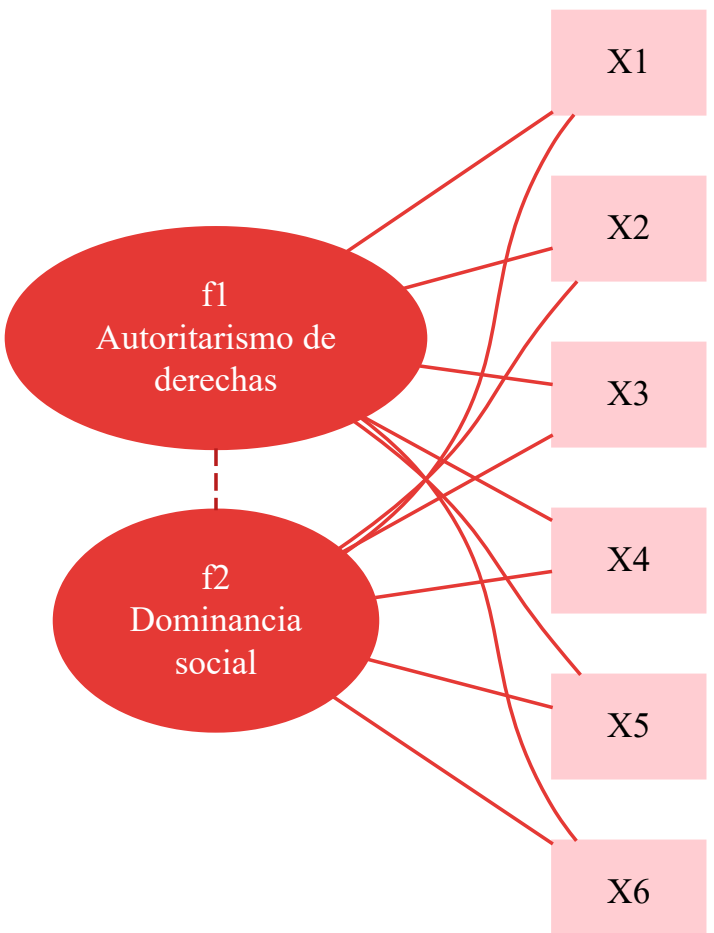
Análisis factorial

Análisis Factorial Exploratorio (AFE o EFA)	Análisis Factorial Confirmatorio (AFC o CFA)
Cuando no hay un modelo teórico que sustenta la manera en que las variables observadas se relacionan entre sí	Un modelo teórico especifica qué variables observadas se relacionan con qué variable(s) latente(s)
Busca identificar las variables latentes que subyacen a un conjunto de indicadores correlacionados entre sí	Generalmente, cada variable observada se relaciona solamente con una variable latente
El modelo plantea que cada variable observada se relaciona con todas las variables latentes	Podemos evaluar si el modelo planteado se ajusta a los datos observados

Análisis factorial: Exploratorio

¿Corresponden autoritarismo y dominancia a dos dimensiones distintas del conservadurismo?

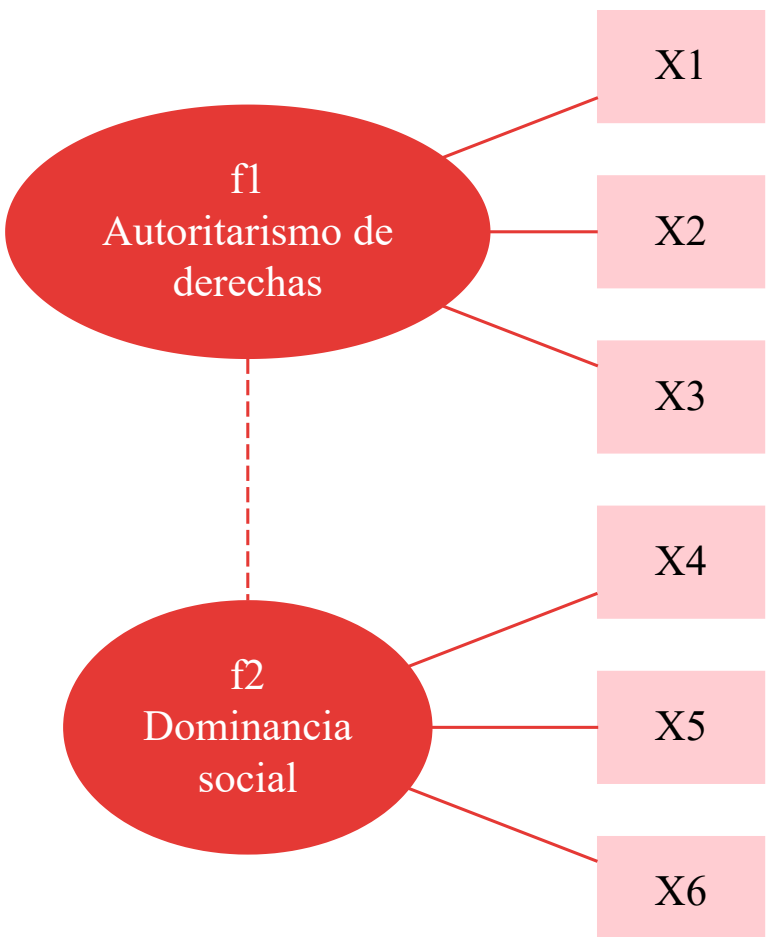
Enfoque Exploratorio (AFE):



Análisis factorial: Confirmatorio

¿Corresponden autoritarismo y dominancia a dos dimensiones distintas del conservadurismo?

Enfoque Confirmatorio (AFC):



Términos relevantes

- **Factores:** variables latentes a la base de las correlaciones entre los indicadores
- **Cargas factoriales:** medida estandarizada de asociación entre el indicador y la variable latente
- **Comunalidad:** proporción del indicador que se asocia a factor(es) comun(es)

Supuestos Análisis Factorial Confirmatorio

- Un número importante de variables observadas de nivel de medición intervalo/razón (mínimo 5 categorías de respuesta en escala).
- Base de datos suficientemente grande. Esto depende de la calidad de los datos, de las correlaciones y del número de ítems. Existen distintos criterios. En general, utilizar bases de al menos 200 casos.
- El modelo está correctamente especificado
- Relación lineal entre variables
- Ausencia de multicolinealidad
- Normalidad multivariante en las variables (al usar el método de estimación de máxima verosimilitud)

Pasos AFC

1. Identificación modelo
2. Evaluación ajuste del modelo
3. Cargas factoriales y comunalidades
4. Comparación modelos
5. Reespecificar modelo (si es necesario)

Identificación del modelo

- Un requisito para poder realizar el AFC es que el modelo se encuentre “identificado”, es decir, que exista una solución única para todos los parámetros del modelo, es decir, contamos con información suficiente para estimarlo.
- Esto depende de la relación entre la cantidad de variables que utilizamos y la cantidad de parámetros a estimar.
- Los grados de libertad del modelo deben ser mayores a 0.
- En caso de que no se cumpla con esto, el software nos los hará saber, debemos cambiar la especificación del modelo.

Identificación del modelo

- En caso de que el modelo no se encuentre identificado debemos chequear ciertas condiciones necesarias, pero no suficientes (es decir, que pueden cumplirse y aun así el modelo no se encontrará identificado).
- Regla t: condición necesaria pero no suficiente

$$t \leq \frac{1}{2} \cdot p(p+1)$$

- Donde: t = número total de parámetros a ser estimados (coeficientes de regresión, varianzas de los errores y correlaciones) y p = número de variables observadas
- La escala de las variables latentes debe estar fijada (ya sea fijando una de las cargas factoriales para cada variable latente en 1 o asignando una varianza de 1 a la variable latente).

Identificación del modelo

- Deben haber suficientes indicadores para cada variable latente:
  - Al menos 2 indicadores por variable latente cuando el modelo tiene al menos 2 variables latentes
  - Al menos 3 indicadores por variable latente cuando el modelo tiene solamente 1 variable latente

Especificación y estimación del modelo

- Utilizaremos el paquete `{lavaan}` (Rosseel, 2012).
- Comenzamos especificando el modelo en un objeto:

```
mod_conf <- 'autoritarismo =~ aut1 + aut2 + aut3
             dominancia =~ dom1 + dom2 + dom3'
```

- El símbolo `=~` significa que la variable latente (izquierda) es medida por los indicadores (derecha).
- La función `cfa()` se usa para estimar el modelo:

```
mod_conf_cfa <- cfa(mod_conf,   # modelo especificado
                    data = datos) # base de datos con ítems
```

- `mod_conf_cfa` es un objeto con los resultados del modelo ajustado.

Evaluación ajuste del modelo

Prueba de Chi-Cuadrado

- Evalúa si existen discrepancias significativas entre la matriz de covarianza observada y la que es estimada por el modelo
- La hipótesis nula plantea que no son significativamente distintas
  - En este caso, p > 0,05 indica un buen ajuste del modelo con 95% de confianza
  - Buscamos mantener la hipótesis nula
- Sin embargo, es sensible al tamaño de la muestra
  - Con muestras grande (n > 400) es muy difícil encontrar un buen ajuste
- Criterio menos estricto: Chi2/ grados de libertad < 2
  - Aunque hay autores que proponen que si el valor es igual o menor a 4 el ajuste es adecuado

Evaluación ajuste del modelo

RMSEA (Root Mean Square Error of Approximation)

- Evalúa el ajuste del modelo considerando el tamaño de la muestra y la complejidad del modelo.
- Valores menores a 0,05 indican un buen ajuste
- Valores entre 0,05 y 0,08 indican un ajuste razonable
- Valores sobre 0,10 indican un mal ajuste

Evaluación ajuste del modelo

CFI (Comparative Fit Index)

- Índice que compara el ajuste del modelo con un modelo base o nulo (modelo sin covarianzas entre las variables)
- Da cuenta del incremento en el ajuste al pasar de un modelo base al modelo propuesto
- Es menos sensible al tamaño de la muestra
- Obtiene valores entre 0 y 1. Mientras más cercano a 1, mejor es el ajuste
- En general, se considera que un modelo tiene un ajuste razonable si CFI > 0,90 y que tiene un ajuste bueno si CFI > 0,95.

Cargas factoriales y comunalidades

- Estimamos el modelo y obtenemos cargas para cada variable en su factor correspondiente
- Obtenemos soluciones estandarizadas y no estandarizadas
- Un buen modelo tiene cargas factoriales estandarizadas sobre 0,7 (o al menos sobre 0,5)
- **Recordatorio:** las cargas factoriales son una medida estandarizada de asociación entre el indicador y la variable latente
- Las comunalidades corresponden al R-cuadrado, esto es, el porcentaje de varianza de una variable que es explicado por la variable latente (carga estandarizada al cuadrado).
- Las cargas (y comunalidades) más altas implican que las variables son más relevantes a la hora de definir un factor (mediciones más “puras” de este)
- La información de las cargas sirve para determinar eventualmente que ítem sería mejor eliminar (si es que es necesario eliminar algún ítem), pudiéndose descartar la(s) variable(s) con la(s) carga(s) más baja(s).

Cargas factoriales y comunalidades

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|) Std.lv Std.all
  autoritarismo =~
    aut1              1.000                             1.646   0.907
    aut2              0.921    0.060   15.457    0.000   1.517   0.887
    aut3              0.929    0.069   13.502    0.000   1.530   0.803
  dominancia =~
    dom1              1.000                             1.201   0.919
    dom2              1.096    0.080   13.729    0.000   1.316   0.882
    dom3              0.861    0.084   10.298    0.000   1.034   0.685

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|) Std.lv Std.all
  autoritarismo ~~
    dominancia        0.678    0.174    3.891    0.000   0.343   0.343
```

- **Estimate:** cargas no estandarizadas
- `aut1` y `dom1` igual a 1
  - Fijado así para dar escala de los factores
  - `lavaan` asume que el primer indicador de un factor se fija en 1
- **Std.lv:** solución con factores estandarizados
- **Std.all:** solución con factores e indicadores estandarizados

Cargas factoriales y comunalidades

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|) Std.lv Std.all
  autoritarismo =~
    aut1              1.000                             1.646   0.907
    aut2              0.921    0.060   15.457    0.000   1.517   0.887
    aut3              0.929    0.069   13.502    0.000   1.530   0.803
  dominancia =~
    dom1              1.000                             1.201   0.919
    dom2              1.096    0.080   13.729    0.000   1.316   0.882
    dom3              0.861    0.084   10.298    0.000   1.034   0.685

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|) Std.lv Std.all
  autoritarismo ~~
    dominancia        0.678    0.174    3.891    0.000   0.343   0.343
```

- Todos los indicadores p<0,05
  - Su relación con los factores es significativa al 95% de confianza
- `aut1` y `dom1` no tienen valor p porque fueron fijados
- Los coeficientes estandarizados (Std.all) > 0,5
- `aut1` y `dom1` son los indicadores más puros de sus factores

Cargas factoriales y comunalidades

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|) Std.lv Std.all
  autoritarismo =~
    aut1              1.000                             1.646   0.907
    aut2              0.921    0.060   15.457    0.000   1.517   0.887
    aut3              0.929    0.069   13.502    0.000   1.530   0.803
  dominancia =~
    dom1              1.000                             1.201   0.919
    dom2              1.096    0.080   13.729    0.000   1.316   0.882
    dom3              0.861    0.084   10.298    0.000   1.034   0.685

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|) Std.lv Std.all
  autoritarismo ~~
    dominancia        0.678    0.174    3.891    0.000   0.343   0.343
```

- La covarianza entre ambos factores es significativa al 95% de confianza, p<0,05
- La correlación entre ambas variables es positiva (Std.all = 0,343)

Cargas factoriales y comunalidades

- Las comunalidades están expresadas como R cuadrado (R-Square)
- Estas se calculan como la carga factorial estandarizada al cuadrado
- En general, observamos comunalidades altas.
- El ítem `dom3` tiene la comunalidad más baja: el factor “dominancia” explica el 46,9% de la varianza de `dom3`.
  - Podríamos evaluar si sacar este ítem del modelo, en caso de querer reducir la escala

Reespecificar modelo

- Eventualmente, si el modelo no ajusta adecuadamente, se recomienda considerar índices de modificación
- En general, se proponen cambios relativos a la especificación de:
  - **Correlaciones entre los errores de indicadores**
    - Implica una asociación entre indicadores que no es explicada por la asociación con el factor, ej.: (a) Indicadores medidos en un mismo momento en el tiempo o (b) Indicadores medidos utilizando pruebas parecidas
  - **Cargas cruzadas:** un indicador carga en más de un factor
    - En caso de realizarse, debiera ser justificado teóricamente

Reespecificar modelo

- Observamos una lista de sugerencias de modificación bajo `Modification Indices`
- Mayores `mi` indican que este cambio significará un mayor cambio en la bondad de ajuste del modelo
- `=~`: indica agregar un efecto de una variable observada en una variable latente
- `~~`: indica agregar una correlación entre los errores de dos variables observadas
- En este caso, los índices de modificación más altos se relacionan con el `dom3`, se sugiere:
  - Agregar `dom3` como un indicador de autoritarismo (mi = 6,437)
  - Agregar una correlación entre los errores de `dom1` y `dom2` (mi = 6,437)
  - Agregar una correlación entre los errores de `aut2` y `dom3` (mi = 4,180)

Repaso de conceptos centrales de análisis factorial confirmatorio

- Evaluación formativa con el objetivo de hacer un repaso de los contenidos revisados en la sesión anterior sobre Análisis Factorial Confirmatorio a partir de los resultados de esta encuesta

https://forms.gle/hx3zXjGUu7iEJg886

Tratamiento de Ítems Ordinales

- En las diapositivas anteriores asumimos que los indicadores son continuos. Sin embargo, en ciencias sociales es muy común usar escalas tipo Likert (ej. 1 a 5), que son inherentemente **ordinales**.
- Cuando los ítems tienen menos de 5 categorías, o cuando se viola fuertemente el supuesto de normalidad, tratarlos como continuos puede sesgar los resultados.
- **Solución:** Tratar las variables como ordinales. Esto cambia dos cosas:
  - **Matriz de Correlación:** En lugar de una matriz de Pearson (para variables continuas), el modelo se basa en una **matriz de correlación policórica**, que estima la correlación entre las variables latentes continuas que se asumen subyacen a las respuestas ordinales observadas.
  - **Método de Estimación:** El estimador de Máxima Verosimilitud (ML) no es adecuado. Se utiliza un estimador robusto como el de **Mínimos Cuadrados Ponderados Diagonalmente (WLSMV)**, que está diseñado para datos categóricos.

Tratamiento de Ítems Ordinales en `lavaan`

- En lugar de modificar el dataframe previamente, `lavaan` permite declarar las variables como ordinales directamente dentro de la función `cfa()`.
- Este método es más directo y no altera el objeto de datos original. Se logra usando el argumento `ordered`.
- Al especificar variables ordinales, es crucial usar un estimador robusto como **WLSMV (Mínimos Cuadrados Ponderados Diagonalmente)**, el cual se especifica con el argumento `estimator`.

```
# Modelo sigue siendo el mismo
mod_conf <- 'autoritarismo =~ aut1 + aut2 + aut3
             dominancia =~ dom1 + dom2 + dom3'

# Vector con los nombres de los ítems a tratar como ordinales
items_ordinales <- c("aut1", "aut2", "aut3", "dom1", "dom2", "dom3")

# Ajustar el modelo especificando las variables ordinales y el estimador
mod_conf_ord <- cfa(mod_conf,
                    data = datos,
                    ordered = items_ordinales, # ¡Argumento clave!
                    estimator = "WLSMV")       # Estimador para datos categóricos
```

Inferencia y Comparación de Modelos

- El AFC es una herramienta de **testeo de hipótesis**. A menudo, no solo evaluamos un modelo, sino que comparamos modelos alternativos que representan teorías rivales.
- **Ejemplos de comparación:**
  - Modelo de 1 factor vs. Modelo de 2 factores.
  - Modelo donde dos factores correlacionan vs. Modelo donde no correlacionan.
  - Modelo original vs. Modelo reespecificado (ej. con una carga cruzada).
- Para comparar **modelos anidados** (un modelo es una versión más simple y restringida del otro), se utiliza la **Prueba de Diferencia de Chi-cuadrado**.
  - **H0:** El modelo más simple (con más restricciones) se ajusta igual de bien que el modelo más complejo.
  - Un p-valor significativo (< 0.05) sugiere que el modelo más complejo ofrece una mejora sustancial en el ajuste.

Inferencia y Comparación de Modelos

- En `lavaan`, se usa la función `anova()`:

```r
# Modelo 1: Un solo factor general
mod1 <- 'general =~ aut1+aut2+aut3+dom1+dom2+dom3'
fit1 <- cfa(mod1, data = datos)

# Modelo 2: Dos factores correlacionados (el original)
mod2 <- 'autoritarismo =~ aut1+aut2+aut3
         dominancia =~ dom1+dom2+dom3'
fit2 <- cfa(mod2, data = datos)

# Comparar los dos modelos anidados
anova(fit1, fit2)
```

Uso de Ponderadores (Pesos Muestrales)

- Cuando trabajamos con datos de encuestas, es común que las muestras no sean perfectamente representativas de la población.
- Los **ponderadores** o **pesos muestrales** (`weights`) se utilizan para corregir el desbalance en la muestra, permitiendo que las estimaciones de los parámetros (cargas, covarianzas) sean generalizables a la población de interés.
- `lavaan` **sí puede** incorporar pesos muestrales directamente para obtener estimaciones de parámetros correctas.
  - Esto se hace a través del argumento `sampling.weights` en la función `cfa()`.
- **Importante:** Usar solo los pesos **no es suficiente** si la encuesta tiene un diseño complejo (estratificación, conglomerados).

Más Allá de los Pesos: El Diseño Muestral Complejo

- Muchos diseños de encuesta son **complejos**: incluyen **estratificación** y **conglomerados (clusters)**.
- Ignorar esta estructura viola el supuesto de independencia de las observaciones.
- **Consecuencias:**
  - Los **errores estándar** de las estimaciones serán incorrectos (generalmente subestimados).
  - Los **p-valores** y las pruebas de hipótesis no serán confiables.
  - Los **índices de ajuste** (Chi-cuadrado, CFI, RMSEA) serán inválidos.
- **Solución:** Un proceso de dos pasos que combina `lavaan` y `lavaan.survey`.

Implementación Correcta en Dos Pasos

1. **Estimar el modelo en `lavaan` incluyendo los pesos muestrales.**
    - Se usa el argumento `sampling.weights = "nombre_del_ponderador"`.
    - Esto nos da las **estimaciones de los coeficientes (cargas, etc.) correctas**. Sin embargo, los errores estándar y los índices de ajuste aún son incorrectos si hay estratos/conglomerados.
2. **Corregir los resultados con `lavaan.survey`**.
    - Se crea un objeto de diseño de encuesta con el paquete `survey` que especifica los **estratos, conglomerados y pesos**.
    - Se usa la función `lavaan.survey()` para tomar el modelo del Paso 1 y el objeto de diseño para recalcular los **errores estándar** y los **índices de ajuste** correctamente.

Paso 1: Modelo CFA con Pesos en `lavaan`

- Primero, ajustamos el modelo con `cfa()`, pasando el nombre de la variable de ponderación al argumento `sampling.weights`.

```
# Ajustar el modelo CFA especificando la variable de ponderación
mod_conf_cfa_pesos <- cfa(mod_conf,
                          data = datos,
                          sampling.weights = "ponderador")

# AVISO: En este punto, los `Estimate` del summary son correctos y ponderados,
# pero los Std.Err, z-value, P(>|z|) y los índices de ajuste NO son
# confiables si el diseño es complejo.
```

Paso 2: Corrección del Diseño Complejo con `lavaan.survey`

- Ahora, usamos el resultado del paso anterior (`mod_conf_cfa_pesos`) y un objeto de diseño de encuesta para obtener los errores estándar e índices de ajuste correctos.

```r
# 1. Cargar paquetes e instalar si es necesario
# install.packages(c("survey", "lavaan.survey"))
library(survey)
library(lavaan.survey)

# 2. Crear el objeto de diseño de encuesta completo
diseno_encuesta <- svydesign(ids = ~conglomerado, # Variable de cluster
                             strata = ~estrato,     # Variable de estrato
                             weights = ~ponderador, # Variable de peso
                             data = datos)

# 3. Corregir el modelo lavaan usando el diseño de encuesta
fit_complejo <- lavaan.survey(lavaan.fit = mod_conf_cfa_pesos,
                              survey.design = diseno_encuesta)

# 4. Obtener el resumen con estimaciones, errores estándar e índices de ajuste corregidos
summary(fit_complejo)
```