UNSAAC

Universidad Nacional de San Antonio Abad del Cusco

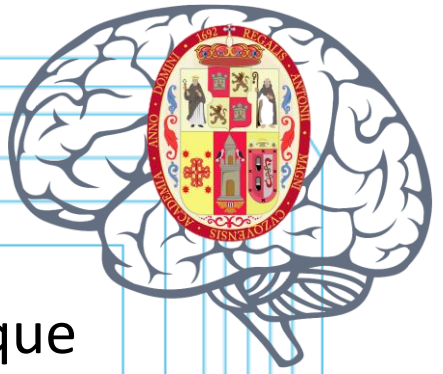
Técnicas multivariadas

- PROFESOR: ARTURO ZUÑIGA

Método del Análisis Factorial

1. *Introducción*
2. *Metodología Estadística*
   - *2.1 Pruebas estadísticas preliminares*
   - *2.2 El modelo del análisis factorial*
   - *2.3 Métodos de estimación*
   - *2.4 Rotación de factores*
   - *2.5 Interpretación de los factores*
   - *2.6 Puntuaciones o Scores.*
3. *Ejemplo de aplicación*

Introducción al Análisis Factorial



El análisis factorial tiene por objetivo explicar un conjunto de variables observadas por un pequeño número de variables latentes, o no observadas, que llamaremos factores.

Por ejemplo, supongamos que hemos tomado veinte medidas físicas del cuerpo de una persona: estatura, longitud del tronco y de las extremidades, anchura de hombros, peso, etc.

Es intuitivo que todas estas medidas no son independientes entre sí, y que conocidas algunas de ellas podemos prever con poco error las restantes porque las dimensiones del cuerpo humano dependen de ciertos factores, y si estos fuesen conocidos podríamos prever con poco error los valores de las variables observadas.

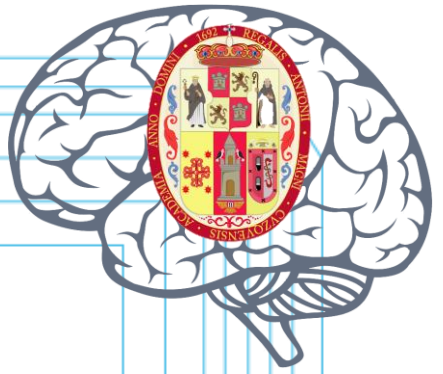
Análisis Factorial: economía

Otro ejemplo, supongamos que estamos interesados en estudiar el desarrollo humano en los países del mundo.

Disponemos de muchas variables económicas, sociales y demográficas, en general dependientes entre sí, que están relacionadas con el desarrollo.

Podemos preguntarnos si el desarrollo de un país depende de un pequeño número de factores tales que, conocidos sus valores, pudiéramos prever el conjunto de las variables económicas que teníamos de cada país.

Análisis Factorial: inteligencia



Como tercer ejemplo, supongamos que medimos con distintas pruebas la capacidad mental de un individuo para procesar información y resolver problemas.

Podemos preguntarnos si existen unos factores, no directamente observables, que expliquen el conjunto de resultados observados.

El conjunto de estos factores será lo que llamamos inteligencia y es importante conocer cuántas dimensiones distintas tiene este concepto y cómo caracterizarlas y medirlas.

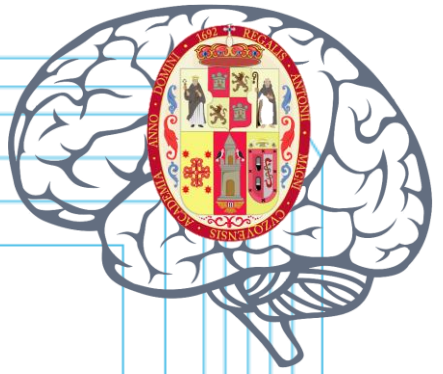
Análisis Factorial vs Componentes Principales

El análisis factorial surge impulsado por el interés de Karl Pearson y Charles Spearman en comprender las dimensiones de la inteligencia humana en los años 30, y muchos de sus avances se han producido en el área de la psicometría.

El análisis factorial esta relacionado con los componentes principales, pero existen ciertas diferencias.

En primer lugar, los componentes principales se construyen para explicar las varianzas, mientras que los factores se construyen para explicar las covarianzas o correlaciones entre las variables.

En segundo lugar, componentes principales es un herramienta descriptiva, mientras que el análisis factorial presupone un modelo estadístico formal de generación de la muestra dada.



El AF y el ACP son técnicas para examinar interdependencia de variables. El objetivo del ACP es explicar la mayor parte de la variabilidad total de un conjunto de variables con el menor número de Comp. posibles. En el AF, los Fact. son seleccionados para explicar las interrelaciones entre variables. El AF es un método que permite construir un modelo para explicar la correlación existente entre un conjunto de variables, en términos de otro conjunto de menor número de variables denominadas factores. En el AF las variables originales juegan el rol de variables dependientes, cuya interrelación se desea que sean explicadas con la selección de un conjunto de factores (comunes y únicos) que no son observables y que deberán ser hallados.

El AF tiene como objetivo explicar la estructura causal que origina las relaciones entre un conjunto de variables, así como la variación específica de cada una de ellas.

El AF, presupone la existencia de un conjunto de variables **subyacentes** (factores latentes) desconociendo cuántas y cuáles son, pero que deberán ser halladas en base de la estructura de correlación o variabilidad y de la explicación que puede aportar cada una de ellas sobre las variables originales.

El ACP es una técnica de reducción de datos que se sitúa en el campo de la Estadística Descriptiva, mientras que el AF implica la elaboración de un modelo que requiere la formulación de hipótesis estadísticas y la aplicación de métodos de inferencia.

2.1 Pruebas Estadísticas Preliminares

1) Prueba de los coeficientes de correlación

Examinando la matriz de correlaciones se puede evidenciar la existencia de correlaciones significativas. Para corroborar se realiza la prueba estadística de significación de cada uno de los coeficientes de correlación:

Formulación de hipótesis:

Hp: $\rho = 0$

Ha: $\rho \neq 0$

Cálculo de la prueba estadística: $t_C = \dfrac{r - \rho}{S_r}$ donde: $S_r = \sqrt{\dfrac{1 - r^2}{n - 2}}$

Decisión estadística: Se acepta Hp, si $t_{(\frac{\alpha}{2},\, n-2)} \leq t_C \leq t_{(1-\frac{\alpha}{2},\, n-2)}$

2.1 Pruebas Estadísticas Preliminares

2) Prueba de esfericidad de Barlett

Permite probar si existe una intercorrelación significativa entre las variables originales.

Así, para p variables se puede usar la matriz de correlación R_p cuyos elementos de la diagonal son 1s y los que están fuera de la diagonal son coeficientes que miden la intercorrelación entre cada par de variables. Si todos estos coeficientes son nulos (no existe correlaciones entre las p variables), entonces la matriz R_p será igual a la identidad, con lo que su determinante será igual a la unidad.

Formulación de las hipótesis:

$$\text{Hp} : |R_p| = 1$$

$$\text{Ha} : |R_p| \neq 1$$

$$\left|\begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix}\right| = 1$$

Estadística de Barlett:

$$\chi_c^2 = -\left[n - 1 - \frac{1}{6}(2p+5)\right]\ln|R|$$

Decisión estadística. Se rechaza Hp, si $\chi_c^2 \geq \chi^2_{(1-\alpha)(p^2-p)}$. Si se acepta Hp, entonces las variables no están correlacionadas y por lo tanto no tiene sentido aplicar el análisis de componentes principales.

2.1 Pruebas Estadísticas Preliminares

3. Con el Software R se hará la prueba de normalidad p-variada de Shapiro.

H0: Las p variables tienen distribución normal p-variada.

NOTA: Desde un punto de vista estadístico, se pueden obviar los supuestos de normalidad, homocedasticidad y linealidad siendo conscientes que su incumplimiento produce una disminución en las correlaciones observadas.. En realidad sólo es necesaria la normalidad cuando se aplica una prueba estadística a la significación de los factores; sin embargo raramente se utilizan estas pruebas. Es deseable que haya cierto grado de multicolinealidad, porque uno de los objetivos es identificar series de variables interrelacionadas.

2.1 Pruebas Estadísticas Preliminares

3) Ajuste del Modelo

Para evaluar la adecuación el conjunto de datos de la muestra al análisis factorial, se utiliza el estadístico propuesto por Kaiser-Meyer-Olkin (KMO), definido como:

$$KMO = \frac{\sum \sum_{h \neq j} r_{jh}^2}{\sum \sum_{h \neq j} r_{jh}^2 + \sum \sum_{h \neq j} a_{jh}^2}$$

donde r_{jh} son los coeficiente de correlación simple entre las variables
a_{jh} son los coeficiente de correlación parcial entre las variables

En el caso que existiera una adecuación de los datos, los coeficientes de correlación parcial serán pequeños y por consiguiente el KMO estará próximo a 1. El KMO varía entre 0 y 1. Para valores de KMO menores a 0.5 se considerará no aceptable la aplicación del análisis factorial al conjunto de datos.

2.1 Pruebas Estadísticas Preliminares

4) Bondad de ajuste de cada variable

También se propone una medida de adecuación para cada una de las variables. Los índices de de adecuación es:

$$MSA_j = \frac{\sum_{h \neq j} r_{jh}^2}{\sum_{h \neq j} r_{jh}^2 + \sum_{h \neq j} a_{jh}^2}$$

Si MSA_j se próximo a 1, indicaría que la variable j es adecuada para el análisis factorial. Por el contrario, para valores de MSA_j menores a 0.5 se considera a la variable no aceptable para el AF y puede ser eliminada para el análisis.

2.2. El Modelo del Análisis Factorial

1. Formulación del Modelo

Considerando el vector aleatorio $\underline{X}$ (px1) de variables observables con vector de media $\underline{\mu}$ y matriz de variancia-covariancia Σ. Entonces el modelo de factores en forma matricial se define:

$$\underline{X} = L\underline{f} + \underline{e} + \underline{\mu}$$

Donde:

L Matriz (pxm) de constantes. Cuyos elementos l_{jh} son coeficientes denominados **cargas factoriales**, que representan el peso que tiene factor h sobre la variables j.

$\underline{f}$ Vector aleatorio (mx1). Los elementos F_j son los llamados **factores comunes**.

$\underline{e}$ Vector aleatorio (px1). Los elementos e_j son llamados **factores únicos o específicos**.

μ Vector (px1) cuyos elementos μ_j son las medias poblacionales de la variable j.

2.2. El Modelo del Análisis Factorial

El modelo del AF muestra que cada una de las p variables observables X_j, se expresan como una combinación lineal de m factores comunes (m<p) y un factor único. Esto es, todas las variables originales están influenciadas por todos los factores comunes y sólo por un factor único. Las ecuaciones que representan estas combinaciones lineales se expresan como:

$$X_1 = l_{11} F_1 + l_{12} F_2 + \ldots + l_{1m} F_m + e_1 + \mu_1$$
$$X_2 = l_{21} F_1 + l_{22} F_2 + \ldots + l_{2m} F_m + e_2 + \mu_2$$
$$\vdots$$
$$X_p = l_{p1} F_1 + l_{p2} F_2 + \ldots + l_{pm} F_m + e_p + \mu_p$$

2.2. El Modelo del Análisis Factorial

2.. Supuestos del Modelo

Para el proceso de inferencia estadísticas se formulan suposiciones acerca del vector de medias y la matriz de covariancias de los factores comunes y únicos.

- $E(\underline{f}) = 0\,, \qquad E(\underline{f}\,\underline{f}') = I$
- $E(\underline{e}) = 0\,, \qquad E(\underline{e}\,\underline{e}') = \Omega$
- $E(\underline{f}\,\underline{e}') = 0$
- I es una matriz identidad y Ω es una matriz diagonal

- Los factores comunes son variables normalizadas con media 0 y variancia1 y están incorrelacionadas entre sí.
- Los factores únicos son variables con media 0 y variancias diferentes y están incorrelacionadas entre sí.
- Los factores comunes y únicos están incorrelacionados entre sí.

2.2. El Modelo del Análisis Factorial

3.. Propiedades del Modelo

El problema matemático del análisis factorial es descomponer la variabilidad (matriz de variancia-covarinacia Σ) de las variables observadas (vector $\underline{X}$), en dos partes: una variabilidad común con todas las variables (matriz $L'L$) y otra propia o específica (matriz Ω). Entonces, las matrices de variancia-covariancias y correlación poblacionales para el vector aleatorio $\underline{X}$ se pueden expresar por:

$$Cov(\underline{X}) = \Sigma = L'L + \Omega \qquad \text{y} \qquad R = L'L + \Omega \qquad \text{... (2)}$$

donde $L'L$ y Ω son las matrices de variancia-covariancias correspondiente a los factores comunes y factores únicos respectivamente.

Así mismo, se tiene que la variancia para la variable x_j queda expresada por:

$$Var(x_j) = l_{j1}^2 + l_{j2}^2 + \ldots + l_{jm}^2 + \omega_j^2$$

2.2. El Modelo del Análisis Factorial

La suma de las m cargas para una variables ($l_{j1}^2 + l_{j2}^2 + \ldots + l_{jm}^2$) determina su comunalidad.

Haciendo $h_j^2 = l_{j1}^2 + l_{j2}^2 + \ldots + l_{jm}^2$ entonces se nota que la variancia de x_j se expresa (es explicada) en términos de los elementos de la diagonal de las matrices de variancia-covariancia de los factores comunes y únicos, esto es:

$$Var(x_j) = \sigma_{jj} = h_j^2 + \omega_j^2 \quad \text{y}$$

$$1 = h_j^2 + \omega_j^2 \qquad \text{(si las variables están estandarizadas)}$$

donde:

h_j Es la **Comunalidad**. Parte de la variancia de x_j que es explicada por los factores comunes.

ω_j^2 Es la **Especificidad**. Parte de la vaiancia de x_j explicada por los factores únicos.

l_{jh} Son coeficientes llamadas **cargas factoriales**. Son los pesos de los distintos factores asociada a la variable j.

2.3. Métodos de Estimación

La estimación de los parámetros del modelo factorial implica el uso de métodos que conlleven a lo que se denomina la extracción de factores, determinando así el número de factores comunes a ser seleccionados en el análisis factorial. Existen varios métodos:

1) Método de los componentes principales *(método exploratorio),* cuando se busca la existencia y número de los factores

2) Método de máxima verosimilitud *(método confirmatorio)*
3) Método de mínimos cuadrados no ponderados
4) Método de mínimos cuadrados generalizados

2.4. Métodos de Rotación de Factores

La rotación de factores tiene como finalidad determinar fácilmente los factores comunes y la interpretación de sus interrelaciones con cada una de las variables originales. Los factores rotados obtenidos a partir de una solución inicial, presentarán una correlación alta (próxima a 1) con uno o grupo de variables originales y correlaciones bajas (próxima a 0) con el resto de variables. Así, se puede identificar rasgos o características comunes en un grupo de variables asociadas a un factor y dar una denominación a estas interrelaciones encontradas entre las variables del grupo.

1) Rotación ortogonal
2) Rotación oblicua.

2.5. Interpretación de los factores

Correlación entre las componentes principales y las variables originales

La correlación entre la j-ésima variable original y la k-ésima componente principal, representa el grado de asociación entre ellas y su valor cuantificará la proporción de la variación total de la variable original j que es explicada por la componente k. Así se tiene:

$$r_{jh} = \frac{w_{hj}\sqrt{\lambda_h}}{\sqrt{Var(X_j)}}$$

Si la variable X_j esta tipificada: $r_{jh} = w_{hj}\sqrt{\lambda_h}$

donde los w_{hj} son los coeficientes de los autovectores o cargas factoriales.

Estas correlaciones también representan la parte de variancia de cada variable que es explicada cada factor. Se cumple que la suma horizontal de los cuadrados de las cargas factoriales de una variable es igual a uno.

3. Ejemplo de aplicación

La base a datos que se utilizó corresponde a alpacas adultas en actividad reproductiva. Se consideran alpacas adultas aquellas de más de dos años; edad en la cual son capaces de producir crías.

Los datos que se procesaron fueron recolectados durante la esquila del año 2017 en alpacas del plantel de reproductores de la SAIS Pachacutec en Junín.

Las variables se describen a continuación:

Descripción de las variables

Variable	Descripción
Peso vivo=Peso	Medido en unidades de Kilos (Kg.). El registro se realiza momentos previos a la esquila con una balanza de plataforma.
Longitud de mecha =Mecha	Medido en Cm. El registro se realiza momentos posteriores a la esquila. Se utiliza una regla de 30cm, midiendo la longitud de la mecha de fibra en posición perpendicular al cuerpo del animal.
Diámetro promedio de fibra (DF) = Diametro	Medido en micras. Corresponde al promedio del diámetro medido de 1000 fibras. Se midieron 1000 fibras por animal.
Desviación estándar del DF = DesvE	Medido en micras. Corresponde a la desviación estándar de las 1000 fibras medidas por animal.
Coeficiente de variación del (DF) =CV	Medido en %. Corresponde al coeficiente de variación de las 1000 fibras medidas por animal.
Factor picazón = Picazon	Medido en %. Corresponde al porcentaje de fibras que presentan un diámetro de fibra mayor a 30.5 micras; de las 1000 fibras analizadas.

4. Ejemplo de aplicación

Use el archivo **Distritos Peruanos** para hacer el Análisis Factorial.

Las variables son:

4. Ejemplo de aplicación

NOASIST1: Porcentaje de población en el hogar de 6 a 11 años que no asiste a un centro educativo.

NOASIST2: Porcentaje de población en el hogar de 2 a 17 años que no asiste a un centro educativo.

ANALFT: Porcentaje de población en el hogar de 15 años a más que no saben leer ni escribir..

EDUPRIM1: Porcentaje de población en el hogar de 15 años a más que tiene educación primaria completa.

EDUSEC1: Porcentaje de población en el hogar de 18 años a más que tiene educación secundaria completa.

EDUSUP1: Porcentaje de población en el hogar de 18 años a más que tiene educación superior no universitaria completa.

ICASTELL: Porcentaje de población con idioma castellano como lengua aprendida desde la niñez.

INATIVA: Porcentaje de población con idioma quechua, aymara u otra lengua nativa como lengua aprendida desde la niñez

EDUYEARS: Nº de años promedio de educación de la población en el hogar.

EDU1564: Nº de años promedio de educación de la población en el hogar de 15 y 64 años.

EDU1599: Nº de años promedio de educación de la población en el hogar de 15 años y más.

EDUJEFE: Nº de años promedio de educación del jefe del hogar.

EDUCONY: Nº de años promedio de educación del cónyuge del jefe del hogar

Gracias