

MAT-269: Sesión 1, Análisis Estadístico Multivariado

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM



Horario:

Clases: Lunes y Miércoles, bloque 1-2 (08:00-09:30 hrs.)¹

Ayudantía: Por definir.

Contacto:

E-mail: felipe.osorios@usm.cl.

Web: <http://fosorios.mat.utfsm.cl/teaching.html>

Evaluación:

Se realizará **2 Certámenes** y **Tareas**.

Ponderaciones:

Sea \bar{C} y \bar{T} el promedio de certámenes y tareas, respectivamente. De este modo, la nota de presentación (NP) es dada por:

$$NP = 0.8\bar{C} + 0.2\bar{T}.$$




¹Volviendo a la modalidad presencial, se realizará en [Sala F-265](#).



- ▶ Inferencia en análisis multivariado.
 - ▶ Estimación y test de hipótesis para una muestra aleatoria desde $N_p(\mu, \Sigma)$.
- ▶ Técnicas multivariadas.
 - ▶ Regresión multivariada y GMANOVA.
 - ▶ Análisis de componentes principales.
 - ▶ Análisis factorial.
 - ▶ Métodos de clasificación y agrupamiento.
- ▶ Tópicos adicionales.*



Referencias bibliográficas

-  Anderson, T.W. (2003).
An Introduction to Multivariate Statistical Analysis (3rd Ed.).
Wiley, New York.
-  Härdle, W.K., Simar, L. (2012).
Applied Multivariate Statistical Analysis (3rd Ed.).
Springer, New York.
-  Seber, G.A.F. (2004).
Multivariate Observations.
Wiley, New York.



- ▶ ¿Existe competencia en el [mercado de AFPs](#) chileno?
- ▶ El desempeño de los estudiantes chilenos en el [SIMCE](#).
- ▶ Un problema de clasificación clásico, o por qué nos presta dinero el banco.
- ▶ Recordatorio: el [esquema de modelación](#).



Esto NO es una crítica al sistema de AFP...



Aplicación:

El sistema de AFP (o de capitalización individual) chileno está en vigor desde 1980.

Ahorros de los contribuyentes son administrados en un sistema de multifondos.

Existe 5 tipos de fondos (A, B, C, D y E) divididos por la proporción del portafolio que es invertido en títulos de renta variable.

El fondo A tiene la mayor proporción de inversión en renta variable, la que disminuye progresivamente para los fondos B, C, D y E.

Conjunto de datos:

Rentabilidades mensuales de AFPs: Cuprum, Habitat, PlanVital y ProVida en el periodo de agosto/2005 a diciembre/2013.

Datos fueron obtenidos desde el sitio web de la superintendencia de pensiones (www.spensiones.cl)

Conjunto de datos con 101 observaciones y 4 variables (solamente datos del Fondo D).

Obs. 28, 30, 35, 38, 39, 42, 66, 73 y 75 son identificadas como outliers.

QQ-plot de distancias transformadas revelan la presencia de colas pesadas.



Administradoras de Fondos de Pensiones (AFP) de Chile

Aplicación:

El sistema de AFP (o de capitalización individual) chileno está en vigor desde 1980.

Ahorros de los contribuyentes son administrados en un sistema de multifondos.

Existe 5 tipos de fondos (A, B, C, D y E) divididos por la proporción del portafolio que es invertido en títulos de renta variable.

El fondo A tiene la mayor proporción de inversión en renta variable, la que disminuye progresivamente para los fondos B, C, D y E.

Conjunto de datos:

Rentabilidades mensuales de AFPs: Cuprum, Habitat, PlanVital y ProVida en el periodo de agosto/2005 a diciembre/2013.

Datos fueron obtenidos desde el sitio web de la superintendencia de pensiones (www.spensiones.cl)

Conjunto de datos con 101 observaciones y 4 variables (solamente datos del Fondo D).

Obs. 28, 30, 35, 38, 39, 42, 66, 73 y 75 son identificadas como outliers.

QQ-plot de distancias transformadas revelan la presencia de colas pesadas.

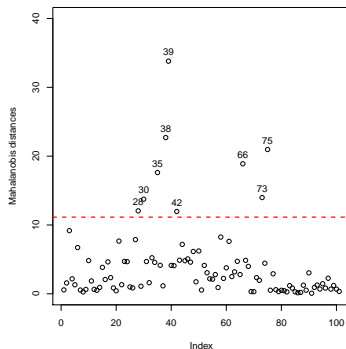


Identificando observaciones atípicas

En mercados emergentes como el chileno suele ocurrir periodos con **alta volatilidad**.

Existe una batería de procedimientos para detectar observaciones que presentan un comportamiento es **aberrante/atípico**.

Este tipo de observaciones puede tener un **efecto nefasto** sobre la inferencia estadística.²



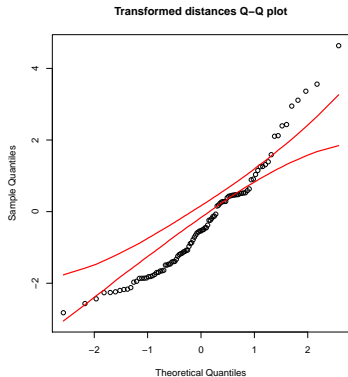
²Es decir, Ud. puede llegar a **conclusiones erróneas!**

Evaluando los supuestos distribucionales

El **supuesto de normalidad** es habitual en este tipo de problemas.

Es decir, suponga x_1, \dots, x_n una **muestra aleatoria** desde $N_p(\mu, \Sigma)$.

Usando **test de hipótesis** y **técnicas gráficas** se concluye que el supuesto de normalidad **no es soportado por los datos**.



Análisis multivariado usando la distribución t de Student

- ▶ Osorio, F., Galea, M., Arellano, R. (2020+).
Using the multivariate t distribution for robust modelling in multivariate analysis.
Working paper.

Características del problema:

- ▶ AFPs invierten esencialmente en la **misma cartera de inversiones**.
- ▶ Mercados emergentes suelen presentar **alta volatilidad**.
- ▶ Los datos son **bien modelados** usando la distribución t multivariada.

Conclusiones:

- ▶ Aparentemente, **no existe competencia** en el mercado de AFP.
- ▶ Cálculo óptimo de los **porcentajes de inversión** en los distintos fondos.
- ▶ Test para evaluar la **igualdad entre razones de Sharpe**.³

³Trabajo en desarrollo junto al prof. Manuel Galea (PUC).



Desde 1988 el SIMCE evalúa los **resultados de aprendizaje** de los estudiantes del sistema de educación chileno.

Objetivos:

- ▶ Describir el **comportamiento del aprendizaje** de los estudiantes.
- ▶ Determinar si existe diferencias significativas entre el **tipo de dependencia** (municipal, subvencionado, particular).

Características del problema:

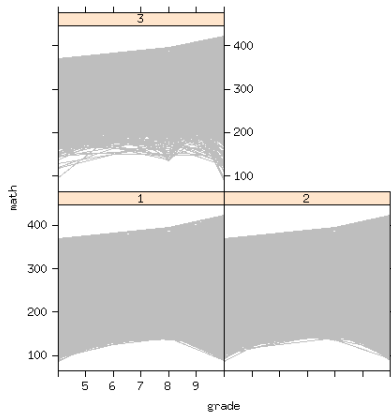
- ▶ Mediciones de un mismo individuo (estudiante) **a través del tiempo** (4° y 8° básico, 2° medio).⁴
- ▶ Datos disponibles para los años 2007, 2011 y 2013, pruebas de Lenguaje y Matemáticas.

⁴Conocido como: **datos con estructura longitudinal**.



Datos del SIMCE

Perfiles individuales de los puntajes del SIMCE en matemáticas, organizados por tipo de dependencia.



- ▶ Maturana, P., and Osorio, F., (2020+).
An approach for robust estimation in growth curve models.
Working paper.

Características del problema:

- ▶ Aproximadamente **132K estudiantes** para ser analizados (base de datos de **mediano porte**).
- ▶ **Crecimiento lineal** (cuadrático?) a través del tiempo.
- ▶ Igual número de mediciones por individuo (**datos balanceados**).

Alternativas para análisis:

- ▶ Modelos con efectos-mixtos.
- ▶ Modelos multi-nivel.
- ▶ Modelo de curvas de crecimiento (GMANOVA).



Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

Iris setosa



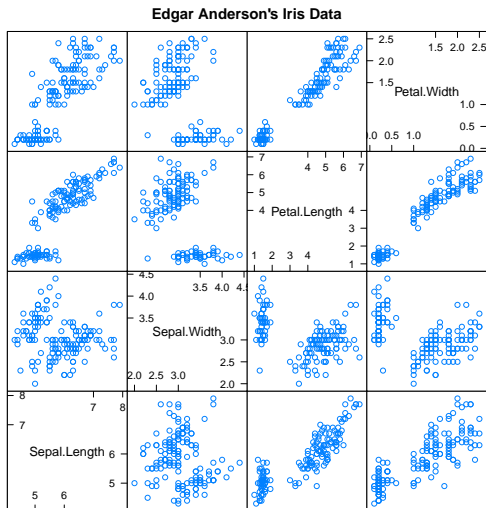
Iris versicolor



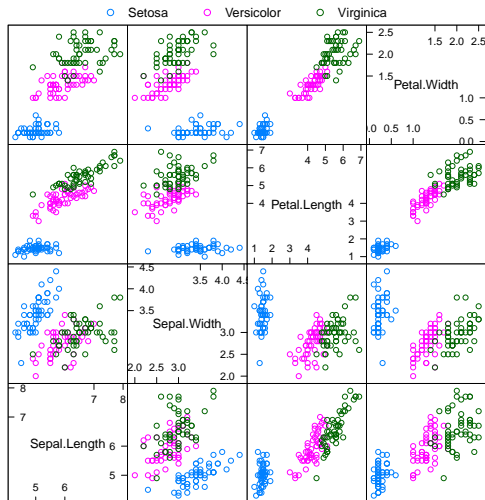
Iris virginica



Datos de Flores Iris (Anderson, 1935; Fisher, 1936)



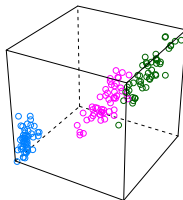
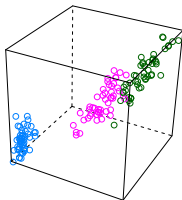
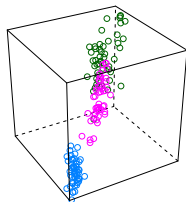
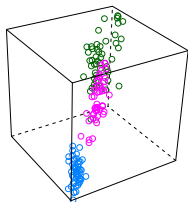
Datos de Flores Iris (Anderson, 1935; Fisher, 1936)



Scatter Plot Matrix



Datos de Flores Iris (Anderson, 1935; Fisher, 1936)



Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

Datos observados:

Mediciones (cm) del largo y ancho de los **sépalos** y el largo y ancho de **pétalos** para 50 flores desde 3 especies de **Iris** (setosa, virginica y versicolor).

Objetivo:

- ▶ Obtener una función que permita **discriminar** entre especies.
- ▶ Usando las medidas de una flor, **clasificarla** apropiadamente.

Características del problema:

- ▶ El análisis exploratorio revela una separación evidente en **2 grupos**.
- ▶ Técnicas más refinadas permiten identificar las 3 especies, p.ej.:
 - ▶ Análisis discriminante,
 - ▶ Técnicas de clasificación (Reconocimiento de patrones),
 - ▶ Aprendizaje de máquina (Máquinas de soporte vectorial, Data mining).



Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

Datos observados:

Mediciones (cm) del largo y ancho de los sépalos y el largo y ancho de pétalos para 50 flores desde 3 especies de Iris (setosa, virginica y versicolor).

Objetivo:

- ▶ Obtener una función que permita discriminar entre especies.
- ▶ Usando las medidas de una flor, clasificarla apropiadamente.

Características del problema:

- ▶ El análisis exploratorio revela una separación evidente en 2 grupos.
- ▶ Técnicas más refinadas permiten identificar las 3 especies, p.ej.:
 - ▶ Análisis discriminante,
 - ▶ Técnicas de clasificación (Reconocimiento de patrones),
 - ▶ Aprendizaje de máquina (Máquinas de soporte vectorial, Data mining).



Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

Datos observados:

Mediciones (cm) del largo y ancho de los **sépalos** y el largo y ancho de **pétalos** para 50 flores desde 3 especies de **iris** (setosa, virginica y versicolor).

Objetivo:

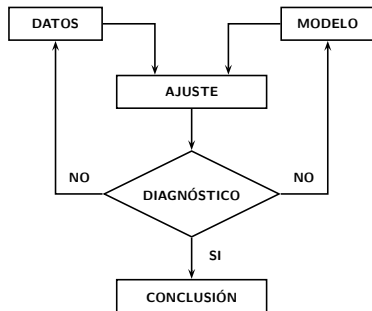
- ▶ Obtener una función que permita **discriminar** entre especies.
- ▶ Usando las medidas de una flor, **clasificarla** apropiadamente.

Características del problema:

- ▶ El análisis exploratorio revela una separación evidente en **2 grupos**.
- ▶ Técnicas más refinadas permiten identificar las 3 especies, p.ej.:
 - ▶ **Análisis discriminante**,
 - ▶ **Técnicas de clasificación** (Reconocimiento de patrones),
 - ▶ **Aprendizaje de máquina** (Máquinas de soporte vectorial, Data mining).



Esquema de Modelación Estadística



Recolección de datos: **Muestreo**.

Análisis exploratorio de datos.

Análisis Multivariado.

Técnicas de Regresión.

Series de Tiempo, entre (muchas) otras.

Inferencia Estadística.

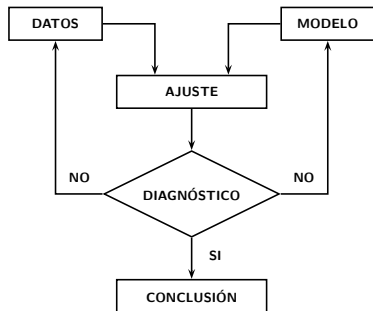
Bondad de ajuste, técnicas gráficas.

Análisis de Sensibilidad.

Comuniquen sus resultados!



Esquema de Modelación Estadística



Recolección de datos: **Muestreo**.

Análisis exploratorio de datos.

Análisis Multivariado.

Técnicas de Regresión.

Series de Tiempo, entre (muchas) otras.

Inferencia Estadística.

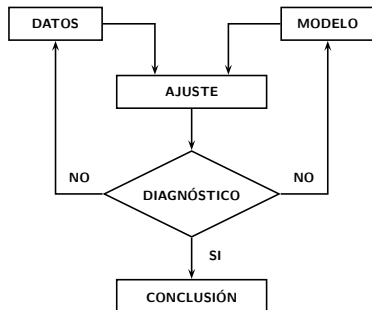
Bondad de ajuste, técnicas gráficas.

Análisis de Sensibilidad.

Comuniquen sus resultados!



Esquema de Modelación Estadística



Recolección de datos: **Muestreo**.

Análisis exploratorio de datos.

Análisis Multivariado.

Técnicas de Regresión.

Series de Tiempo, entre (muchas) otras.

Inferencia Estadística.

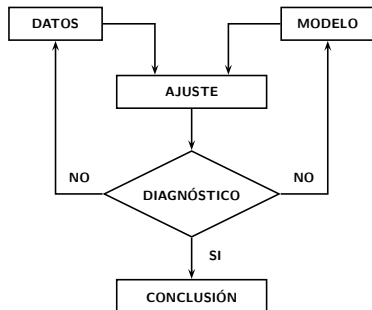
Bondad de ajuste, técnicas gráficas.

Análisis de **Sensibilidad**.

Comuniquen sus resultados!



Esquema de Modelación Estadística



Recolección de datos: **Muestreo**.

Análisis exploratorio de datos.

Análisis Multivariado.

Técnicas de Regresión.

Series de Tiempo, entre (muchas) otras.

Inferencia Estadística.

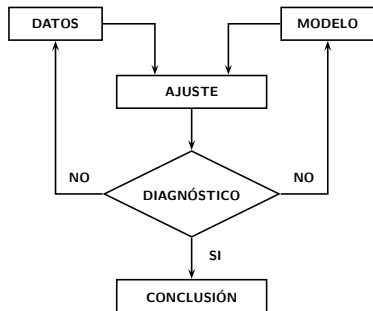
Bondad de ajuste, técnicas gráficas.

Análisis de Sensibilidad.

Comuniqué sus resultados!



Esquema de Modelación Estadística



Recolección de datos: **Muestreo**.

Análisis exploratorio de datos.

Análisis Multivariado.

Técnicas de Regresión.

Series de Tiempo, entre (muchas) otras.

Inferencia Estadística.

Bondad de ajuste, técnicas gráficas.

Análisis de **Sensibilidad**.

Comuniquen sus resultados!



Criterio para rendir global:

$$45 \leq NP \leq 54$$

Nota final:

Sea NG la nota obtenida en el global. De este modo, la **nota final** (NF) del curso se calcula como:

$$NF = 0.7 NP + 0.3 NG$$



Clases:

Lunes y Miércoles, bloques 1-2 (08:00-09:30 hrs.) via [Zoom](#)⁵

Ayudantías:

Horario y ayudante por definir.

Atención alumnos:

Para [agendar una reunión](#) enviar un e-mail a: felipe.osorios@usm.cl. Mencionar el código de la asignatura en el asunto MAT-269.

Disponibilidad horaria:

Un fichero PDF detallando la [disponibilidad horaria del profesor](#) será publicado en la [página web de la asignatura](#). Ud. también puede consultar por email.

⁵Disponible en: <https://zoom.us/>



- ▶ Se llevará un **control de asistencia**.
- ▶ Se puede realizar **preguntas** sobre la materia en **cualquier momento**.
- ▶ Conversaciones sobre asuntos ajenos a la clase no serán tolerados. Otros estudiantes tiene derecho a **asistir clases en silencio**.
- ▶ Al enviar algún **e-mail al profesor**, identificar el código de la asignatura en el asunto (**MAT269**).
- ▶ **E-mail** será el canal de **comunicación oficial** entre el profesor y los estudiantes.



- ▶ Evaluaciones serán coordinadas mediante email. El alumno deberá enviar **ficheros en PDF**, puede ser un documento escaneado pero con **letra legible**.
- ▶ Pedidos de corrección **deben ser argumentados por escrito**.
- ▶ **Cualquier tipo de fraude** en prueba (copia, WhatsApp, suplantación, etc.) implicará la **reprobación de los involucrados**.⁶

⁶Puede implicar la apertura de un **proceso disciplinario**.



Orientaciones de estudio

- ▶ **Mantener la frecuencia de estudio** de inicio a final del semestre. El ideal es estudiar el contenido luego de **cada** clase.
- ▶ Estudiar primeramente el contenido dado en clases, **buscando apoyo en las referencias bibliográficas**.
- ▶ Las **referencias son fuentes de ejemplos y ejercicios**. Resuelva una buena cantidad de ejercicios. **No deje esto para la víspera de la prueba**.
- ▶ **Buscar las referencias bibliográficas** al inicio del semestre, dando preferencia a las principales y complementarias.

