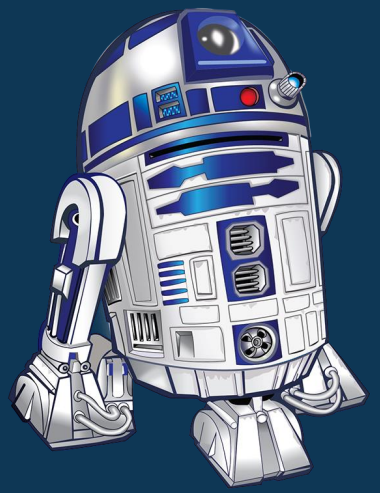
UNSAAC

Universidad Nacional de
San Antonio Abad del Cusco

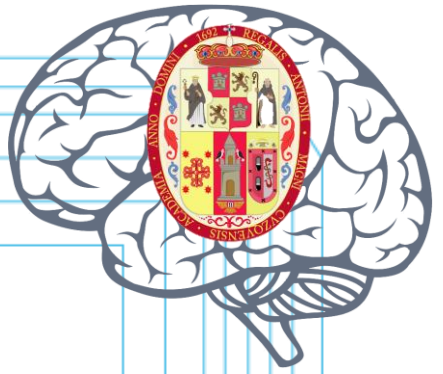
Técnicas multivariadas

- PROFESOR: ARTURO ZUÑIGA



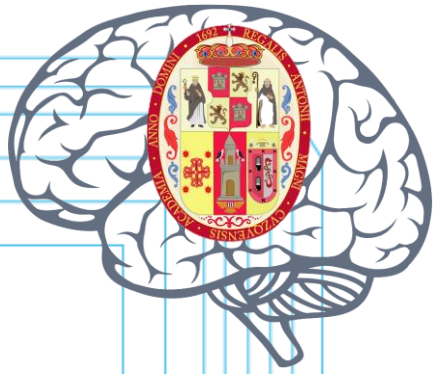
1.1. Entendimiento y preparación de datos

¿Por qué preparar los datos?



- Algún tipo de preparación de datos siempre es necesario para la mayoría de técnicas multivariadas y de minería de datos o machine learning.
- El propósito de la preparación es transformar los conjuntos de datos de tal forma que la información que contienen este mejor expuesta para la herramienta de minería de datos que se utilizará.
- Los errores de predicción deberían ser menores (o en el peor caso similares) luego de la preparación de datos, en comparación con la data inicial.
- La preparación de datos también prepara al analista para producir mejores modelos y de manera más rápida.
- Tener buenos datos es un prerrequisito para producir modelos efectivos de cualquier tipo.
- Los datos necesitan ser formateados para cada software en particular.

¿Por que preparar los datos? II



- Los datos necesitan ser adecuados para un método en particular
- Los datos en la vida real están “sucios":
  - **Incompletos:** Falta de valores en los atributos, carecen de algunos atributos de interés, sólo contienen datos agregados: ej., ocupación = “ ”
  - **anómalos:** errores y outliers
    ej., Salario = “-10"
  - **Inconsistentes:** contienen discrepancias en códigos y nombres
    ej., Edad = “42” , Cumpleaños = “03/07/1997"
    ej., Rating previo “1,2,3” Rating actual “A, B, C"
    ej., Discrepancia con registros duplicados

¿Por qué los datos están sucios?

- Los datos incompletos pueden venir de
  - Datos "No aplicables al momento de ser colectados.
  - Diferentes consideraciones de tiempo cuando fueron recolectados y cuando son analizados
  - Problemas Humanos/hardware/software
- Datos anómalos (valores incorrectos) pueden venir de
  - Instrumentos de recolección de datos defectuoso
  - Errores humanos o de computadora en la entrada de los datos
  - Errores en la transmisión de datos
- Datos inconsistentes pueden venir de
  - Diferentes fuentes de datos
  - Violación de dependencias funcionales (ej., modificación en algunos datos relacionados)
- Registros duplicados también necesitan ser limpiados

¿Por qué los datos están sucios? II

- No hay calidad en los datos, no hay calidad en los resultados!
  - Decisiones de calidad deben de basarse en datos de calidad

    ej., datos duplicados o perdidos pueden producir estadísticas engañosas o incorrectas.
  - Data warehouse necesita una integración consistente de datos de calidad.
  - La selección de datos, la limpieza y la transformación comprende la mayor parte del trabajo de construir una data warehouse

Principales tareas en la preparación de datos I

- Limpieza de datos
  - Completa valores faltantes, suavizar datos ruidosos, identificar o remover outliers y resolver inconsistencias.
- Integración de datos
  - Integración de múltiples bases de datos, cubos de datos, archivos.
- Transformación de datos
  - Normalización y agregación (totalización)
- Reducción de datos
  - Se obtiene una representación más reducida en volumen pero que produce los mismos o similares resultados analíticos.
- Discretización de datos
  - Parte de la reducción de datos pero con particular importancia, especialmente para datos numéricos

A. Datos perdidos

Datos perdidos I

Los datos no siempre están disponibles.

La falta de valores se puede deber a:

- Mal funcionamiento de equipos.
- Inconsistencia con otros datos registrados y por lo tanto eliminados.
- Datos no ingresados debido a equivocaciones.
- Algunos datos pudieron no considerarse importantes al momento de ingresar datos.
- No se registro historial o cambios en los datos.

Puede ser necesario estimar estos valores faltantes.

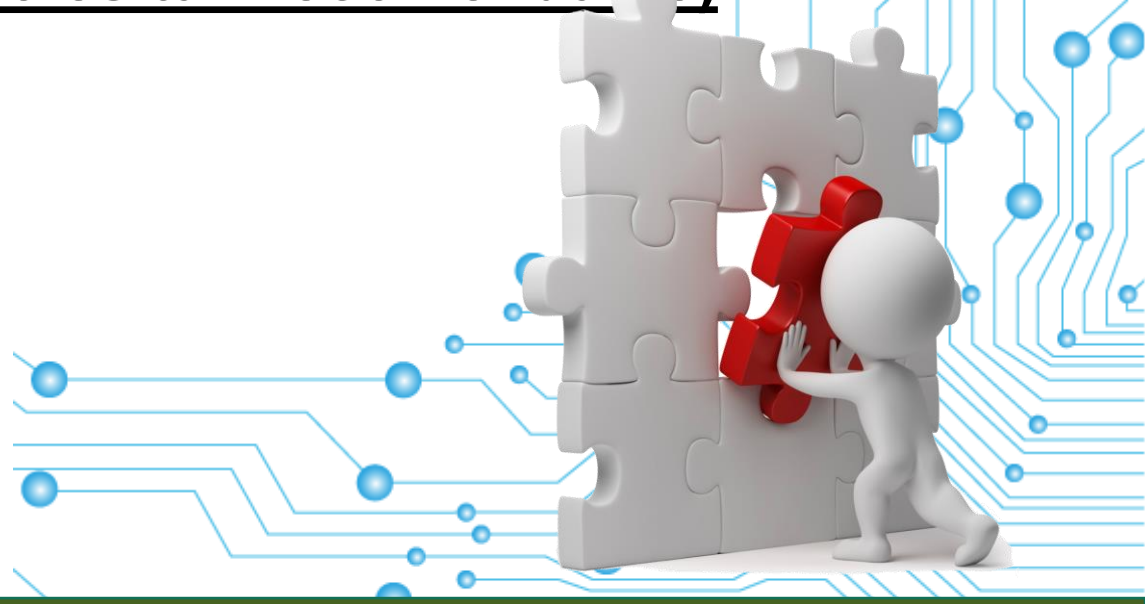
Los valores faltantes son un problema común en análisis estadístico.

Se ha propuesto muchos métodos para el tratamiento de valores faltantes. Muchos de estos métodos fueron desarrollados para el tratamiento de valores faltantes en encuestas por muestreo.

Datos perdido: Impacto

Impacto de los valores faltantes:

- 1 % datos faltantes - trivial. **(eliminar –MCAR eliminar)**
- 1-5 % - manejable **(imputar medidas de tendencia central)**
- 5-15 % - requiere métodos sofisticados **(modelos de regresión o KNN)**
- Más del 15 % - interpretación perjudicial **(no es tan recomendable)**



Mecanismo de datos perdidos I

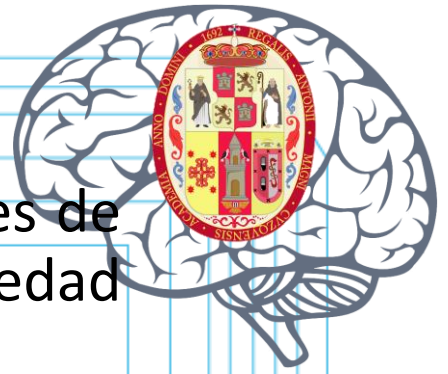
- **Valores faltantes completamente al azar (MCAR):**La probabilidad que una instancia tenga un valor faltante para un atributo es la misma para todas las instancias. Es decir, esta probabilidad no depende ni de los valores observados ni de los valores faltantes. La mayoría de los valores faltantes no son MCAR.

  Por ejemplo en el caso de tener en un estudio las variables ingreso y edad. Estaremos bajo un modelo MCAR cuando al analizar conjuntamente edad e ingresos, suponemos que la falta de respuesta en el campo ingresos es independiente del verdadero valor de los ingresos y la edad.

  Este mecanismo es mas adecuado para datos a ser usados en clasificación no supervisada.

- **Valores faltantes al azar (MAR):** La probabilidad que una instancia tenga un valor faltante en un atributo depende de los valores observados, como por ejemplo la clase a la cual pertenece la instancia, pero no depende de los valores faltantes. Este mecanismo es mas adecuado para datos usados en clasificación supervisada.

Mecanismo de datos perdidos II



En el ejemplo anterior si suponemos que los ingresos son independientes de los ingresos del miembro del hogar pero puede depender de la edad estaremos bajo un modelo MAR.

Este mecanismo es mas adecuado para datos usados en clasificación supervisada.

Valores faltantes no al azar o no ignorables (NMAR): La probabilidad de que una instancia tenga un valor faltante en un atributo depende de los valores faltantes en el conjunto de datos. Ocurre cuando las personas entrevistadas no quieren revelar algo muy personal acerca de ellas. El patrón de valores faltantes no es aleatorio. Este tipo de valores faltantes es el mas difícil de tratar y es el que **ocurre mas frecuentemente**.

En el ejemplo anterior, se obtiene que la función respuesta de la variable ingresos depende del propio valor de la variable ingresos, además de poder depender de otros factores.

"Missing at Random Data: MAR"

"Missing Not at Random Data: MNAR"

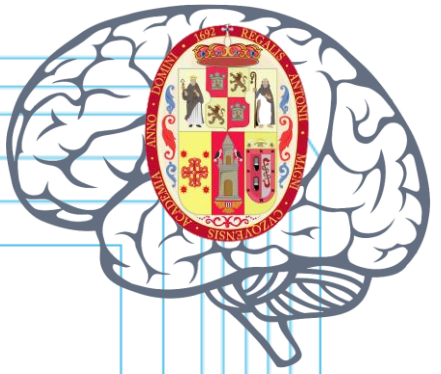
"Missing Completely at Random Data: MCAR"

Consideraciones practicas

- Para conjuntos de datos con un bajo porcentaje de valores faltantes el mecanismo se puede considerar MCAR.
- Para conjuntos de datos con un alto porcentaje de valores faltantes el mecanismo se puede considerar NMAR.
- En muchas aplicaciones lo prudente será considerar distintos modelos plausibles para el mecanismo de no respuesta y realizar un análisis de sensibilidad de las estimaciones.

Ejemplo

Calificaciones de performance en el trabajo con valores faltantes MCAR, MAR y MNAR				
	Calificaciones de performance en el trabajo			
IQ	Completo	MCAR	MAR	MNAR
78	9	-	-	9
84	13	13	-	13
84	10	-	-	10
85	8	8	-	-
87	7	7	-	-
91	7	7	7	-
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	-	7	-
99	7	7	7	-
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	-	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	-	12	12



Estos datos ejemplifican un escenario en donde se seleccionan candidatos a un empleo y los candidatos completan un test de IQ, durante su entrevista de trabajo posteriormente un supervisor evalúa su performance luego de un periodo de prueba de 6 meses.

Caso: Tao

El **proyecto del Océano de Atmósfera Tropical (TAO)** es un esfuerzo internacional importante que instrumentó a todo el Océano Pacífico tropical con aproximadamente 70 corrientes oceánicos profundos . El desarrollo de la matriz TAO en 1985 fue motivado por el evento El Niño de 1982-1983 y, en última instancia, diseñado para el estudio de las variaciones climáticas anuales relacionadas con El Niño y la Oscilación del Sur (ENSO) . Dirigido por la Oficina de Proyectos TAO del Laboratorio Ambiental Marino del Pacífico (PMEL) , la gama completa de las 70 corrientes se completó en 1994.

Variables

Un marco de datos con 736 observaciones sobre las siguientes 8 variables.

- Age: Año (numérico)
- Latitude: Latitud (numérico)
- Longitude: Longitud (numérico)
- Sea.Surface.Temp: temperatura en la superficie del mar (numérico)
- Air.Temp: temperatura del mar (numérico)
- Humidity: Humedad (numérico)
- Uwind: viento (numérico)
- Vwind: viento (numérico)

Tratamiento de la no respuesta I

Eliminar: Es la opción mas sencilla y consiste en eliminar las observaciones o variables que tengan los datos perdidos. Solamente debe realizarse si es poco el porcentaje de observaciones a eliminar y si es posible asumir que los valores faltantes provienen de un proceso MCAR.

Reemplazar (imputar): Reemplazar el valor perdido con un valor conocidos. Variedad de métodos, desde opciones sencillas (reemplazar por la media o mediana) hasta otras más complejas (modelos de regresión).

Mantener: No realizar imputación. A veces es factible analizar la información por separado. Por ejemplo, en algunas situaciones los procedimientos de Máxima Verosimilitud que usan variantes del algoritmo EM (Expectation-Maximization) pueden manejar la estimación de parámetros en presencia de valores faltantes.

Datos faltantes: Eliminación de casos

Utilizaremos la función na.omit

tao.cl = na.omit(tao)

Datos faltantes

Imputación: Los valores faltantes son reemplazados con valores estimados basados en la información disponible.

- Imputación por la media
- Imputación por la mediana
- Imputación por la moda

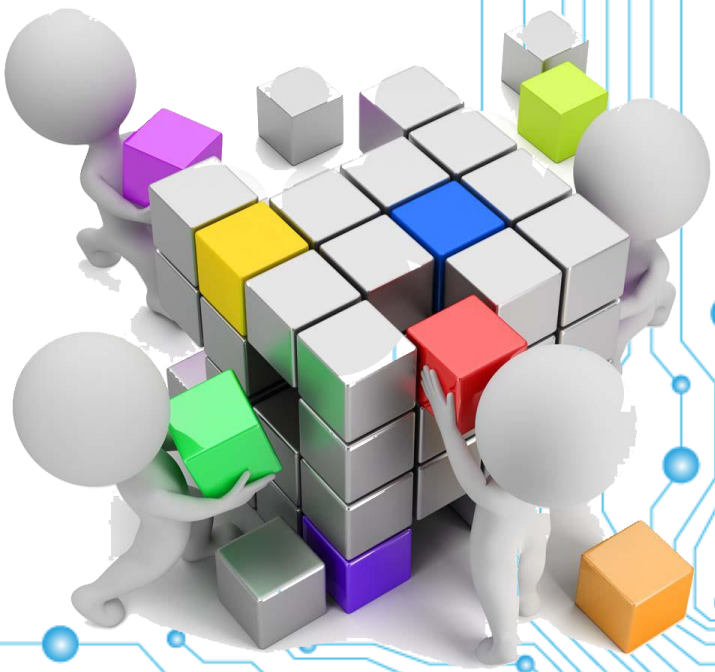
La librería ***DMwR*** tiene la función *centralImputation* que reemplaza lo valores faltantes de la siguiente manera:

- Si la variable es numérica (numeric o integer en R) reemplaza los valores faltantes con la **mediana.**
- Si la variable es categórica (factor en R) reemplaza los valores faltantes con la **moda.**

Datos faltantes

La librería ***VIM*** tiene la función *initialise* que reemplaza los valores faltantes de la siguiente manera:

- Si la variable es numérica continua (numeric en R) reemplaza los valores faltantes con la **media.**
- Si la variable es numérica discreta (integer en R) reemplaza los valores faltantes con la **mediana**.
- Si la variable es categórica (factor en R) reemplaza los valores faltantes con la **moda.**



Datos faltantes: K Vecinos mas cercanos KNN

El método consiste en que para cada valor faltante se encuentran las k-observaciones o instancias que están más cercanas considerando las otras variables.

Luego se reemplaza el valor faltante de la siguiente manera:

- Si la variable es categórica se reemplaza por la moda de las k-observaciones más cercanas.
- Si la variable es numérica se reemplaza por la media de las k observaciones más cercanas.

Datos faltantes

La librería **DMwR** tiene la función **knnImputation** que reemplaza los valores faltantes mediante el método de k-valores mas cercanos:

knnImputation(data, k = 10, scale = T, meth = "weighAvg", distData = NULL)

data: conjunto de datos

k: número de vecinos mas cercanos.

scale: indica si para calcular las distancias primero se estandarizan las variables.

meth: método para reemplazar el valor faltante para variables numéricas.

Opciones: 'median' (mediana) or 'weighAvg' (media ponderada por la distancia). En variables categoricas se usa la moda

Tarea: Data Census o Adult

Descripción: Con los datos del censo puedes verificar si existen valores perdidos y si los hubiera trata de reemplazarlos (imputar datos)

La extracción fue realizada por Barry Becker de la base de datos del Censo de 1994.

Se extrajo un conjunto de registros razonablemente limpios utilizando las siguientes condiciones: ((AAGE> 16) && (AGI> 100) && (AFNLWGT> 1) && (HRSWK> 0))

La tarea de predicción era determinar si una persona gana más de 50 mil al año.

Contiene:

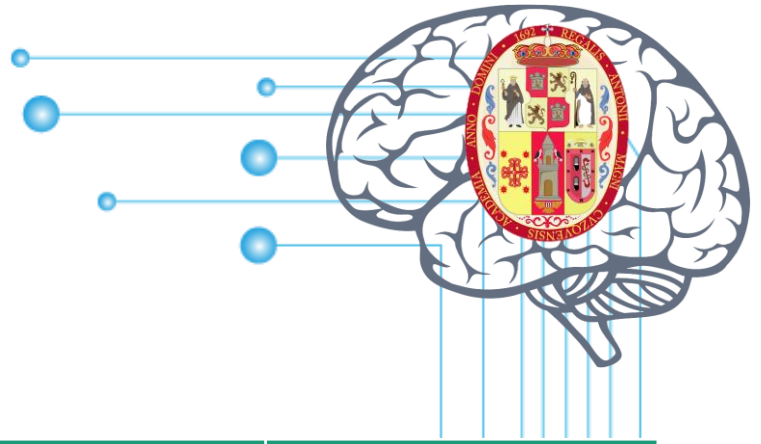
48842 instancias, contiene variables continuas y discretas (entrenamiento=32561, prueba=16281).

Cuando se eliminan las instancias con valores faltantes quedan 45222 (entrenamiento=30162 faltantes quedan 45222 (entrenamiento=30162, prueba=15060).

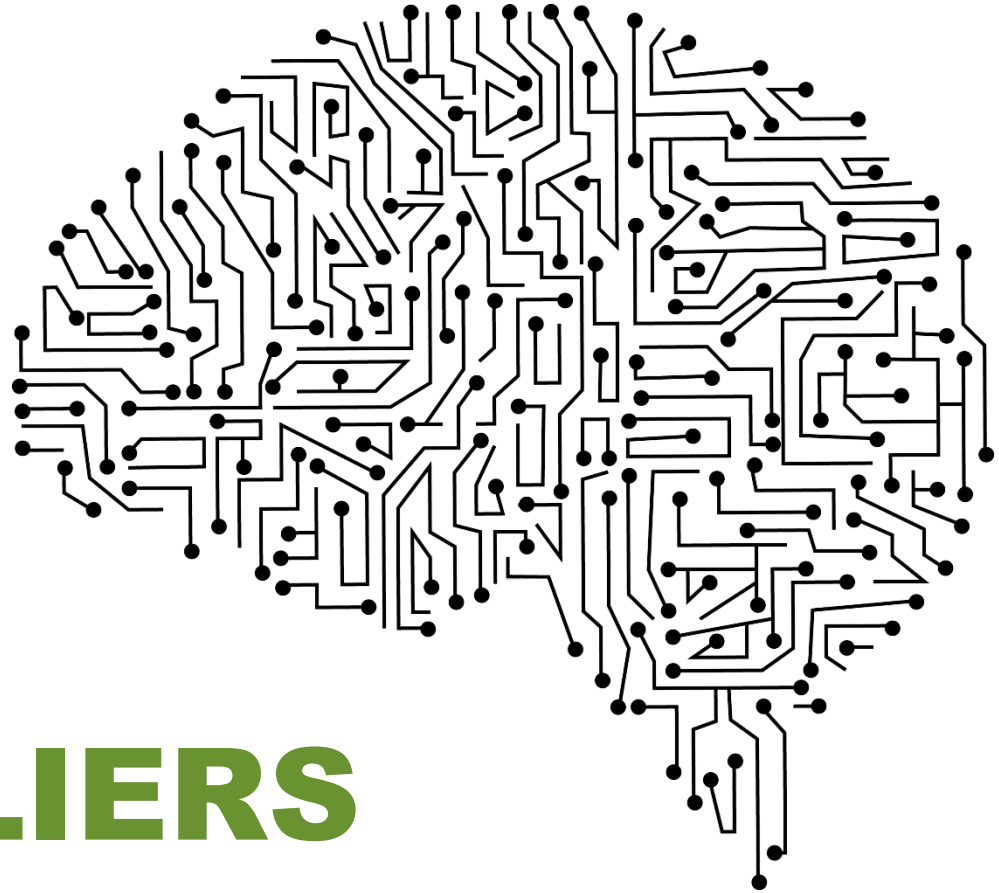
Disponible en: https://archive.ics.uci.edu/ml/datasets/Adult

Donantes: Ronny Kohavi y Barry Becker (1996)

Variables de la data



Cod	Variable	Tipo	Cod	Variable	Tipo
V1	Edad	Continua	V8	Raza	Cualitativa
V2	Clase de trabajo	Cualitativa	V9	Sexo	Cualitativa
V3	Impuestos anuales	Continua	V10	Capital ganado	Continua
V4	Educación	Cualitativa	V11	Capital perdido	Continua
V5	Estado civil	Cualitativa	V12	Horas por semana	Continua
V6	Ocupación	Cualitativa	V13	Ciudad nativa	Cualitativa
V7	Relación de trabajo	Cualitativa	V14	Salario	Cualitativa



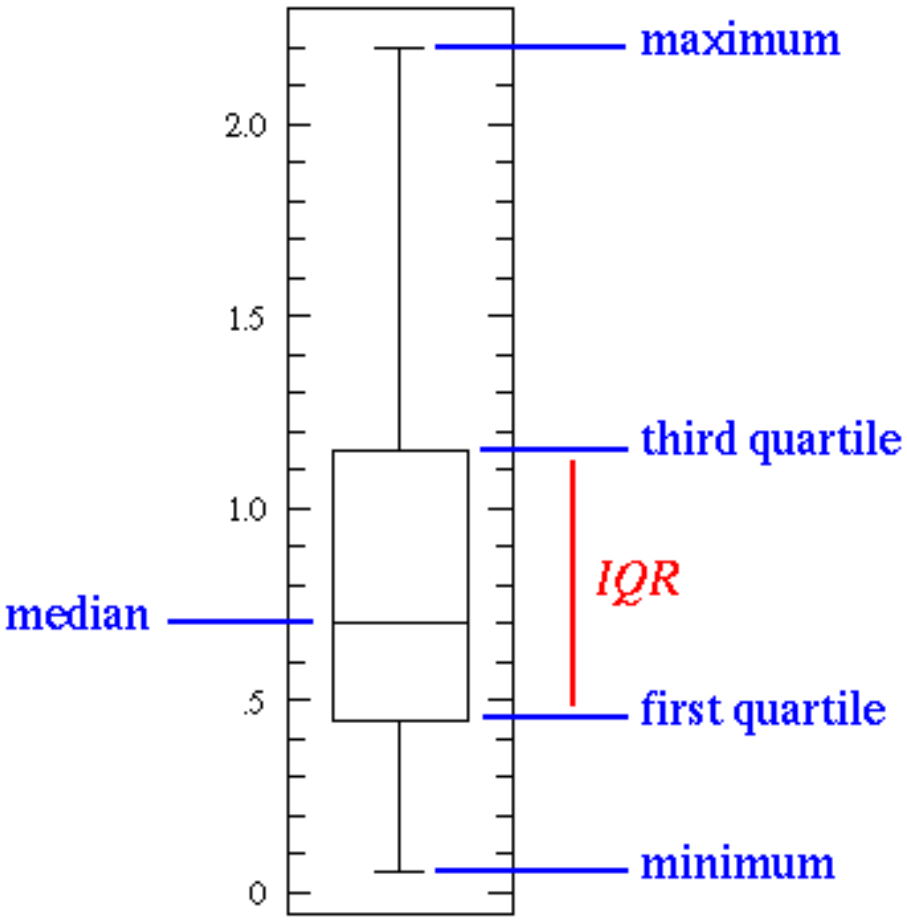
B. OUTLIERS

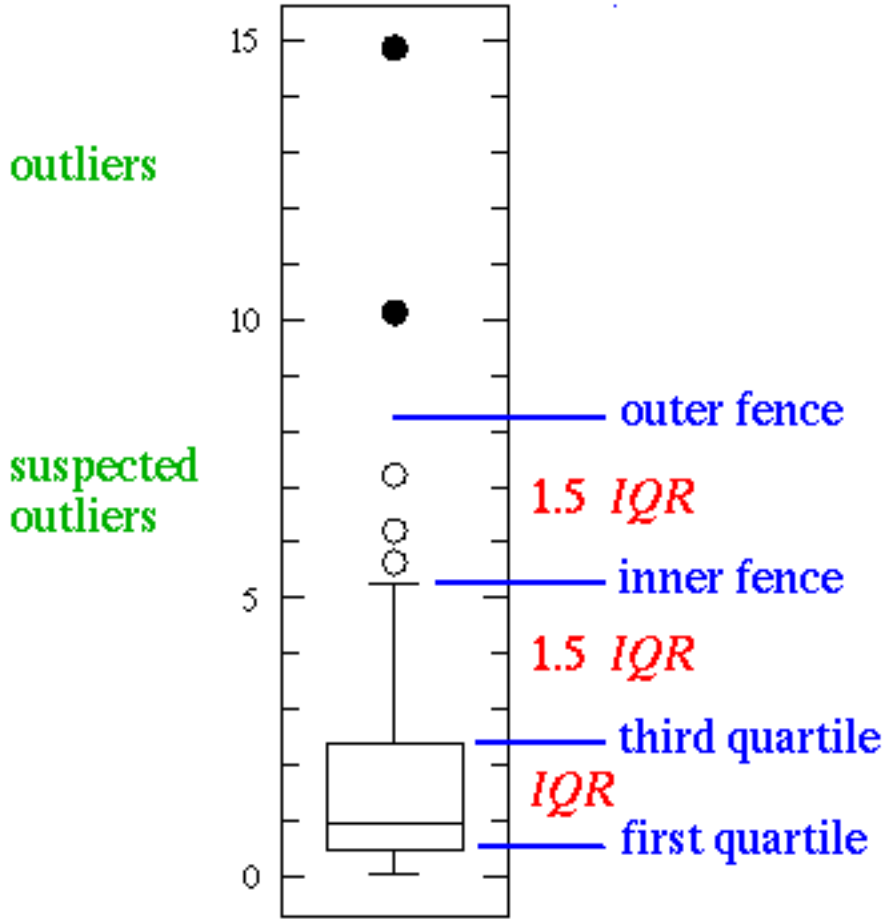
Valor outlier

Un “outlier" es una observación que se desvía tanto de las otras observaciones como para crear la sospecha de que fue generado por un mecanismo diferente.

- Considerar outliers valores que $\frac{|x-\bar{x}|}{s} > k$
- donde k es 2 ó 3 si consideramos normalidad.
- Considerando el Boxplot (Tukey, 1977), se considera outlier a los valores que caen fuera de este intervalo. $(Q1 - 3\,x\,IQR;\ Q3 + 3x\,IQR)$

Outliers Univariados: Boxplots





Outliers multivariado

Si tiene dos dimensiones el grafico de dispersión es una buena forma de detectar outliers (en tu curso de regresión lineal aprendiste outliers) si tiene mas de dos dimensiones se recomienda un estimador robusto de la distancia de mahalonobis.

Sea x una observación de un conjunto de datos multivariado consistente de n observaciones y p variables.

Sea $\bar{x}$ el centroide del conjunto de datos, el cual es un vector p-dimensional que tiene como componentes la media de cada variable.

Sea $\tilde{x}$ la matriz del conjunto de datos original con columnas centradas por sus medias.

Luego la matriz $S = \frac{1}{n-1}\tilde{x}'\,\tilde{x}$ de orden pxp representa la matriz de covarianza de p variables

Por tanto $D^2(x, \bar{x}) = (\mathrm{x} - \bar{x})'\mathrm{S}^{-1}(\mathrm{x} - \bar{x}) > \mathrm{k}$

Donde D^2 es llamada la distancia de Mahalanobis cuadrada estiamda desde x al centroide de los datos

Outliers multivariados: Distancia de Mahalanobis II

Una observación con una distancia de Mahalanobis grande puede ser considerada como un outlier.

Si se asume que los datos vienen de una distribución normal multivariada (p dimensiones):

- Entonces la distancia de Mahalanobis cuadrada de las observaciones siguen una distribución Chi-cuadrado con p grados de libertad.
- Es posible realizar una grafica QQ de la distribución Chi-cuadrado para detectar a los outliers.
- En R: qqplot()

Consideraciones practicas:

- La distribución de chi-cuadrado sigue siendo razonablemente buena para la distancia de Mahalanobis estimada.
- Si los datos no siguen una distribución normal multivariada, los puntos con una distancia de Mahalanobis grande son todavía potenciales outliers.

Aplicación: Tasas

Se obtuvo las estadísticas sobre la Estimación y Análisis de la Fecundidad y mortalidad infantil de 51 ciudades de Latinoamérica.

- Variables
- Mortalidad infantil (numérico)
- Fertilidad (numérico)
- No.Catolicos (numérico)

Gracias