UNIVERSIDADE DA CORUÑA

Aprendizaje automático I

Máster Universitario en Informática Industrial y Robótica

Tema 4

Metodología para el análisis de resultados

Óscar Fontenla Romero

Escuela Politécnica de Ingeniería de Ferrol

http://www.udc.es/epef

Índice

- Métodos de estimación del error
- Métodos de comparación de dos modelos
- Métodos de comparación de múltiples modelos

Introducción

- En el tema anterior hemos analizado distintas métricas para la evaluación de los modelos

- La cuestión ahora es: usando alguna de esas métricas ¿cómo podemos estimar su valor de la forma más realista posible?

- Nos interesa el error de generalización del modelo

Introducción

- Idealmente, el error de un modelo debería ser estimado sobre toda la población de la que proceden los datos
    - Sin embargo, sólo se dispone de una muestra limitada de datos
- Solución más simple: emplear todo el conjunto de datos para entrenar el modelo y para estimar el error
- Problemas:
    - El modelo obtenido probablemente sobreajustará los datos
    - El error obtenido será muy optimista

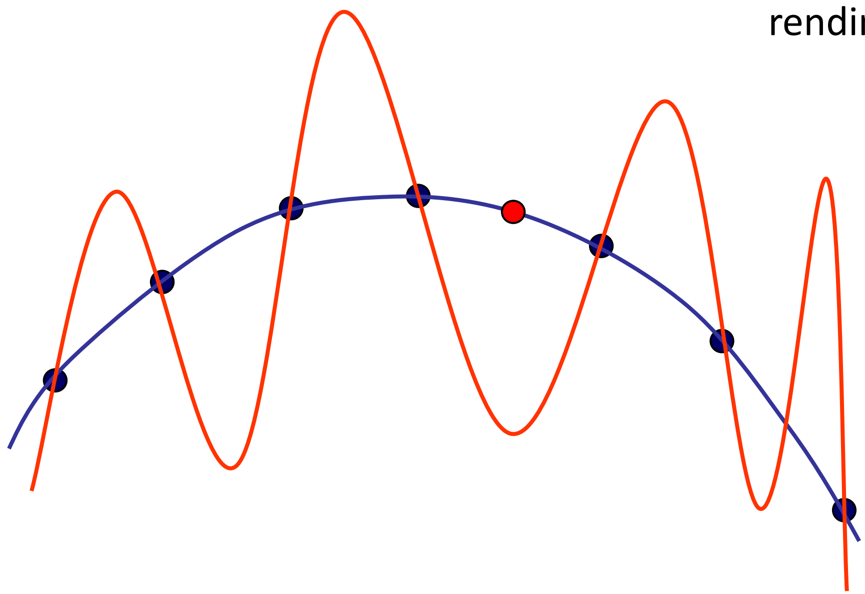
Introducción

- Ejemplo de estimación del error:

Error empírico (entrenamiento) = 0 ⟹ Error ***optimista***

Error real (prueba) > 0

No es válido para conocer el rendimiento ***real*** del sistema



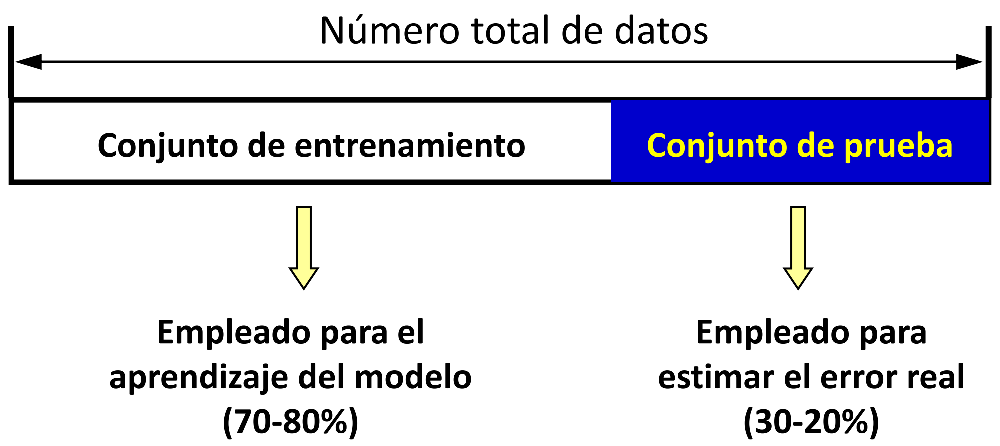
Métodos de estimación del error

- Partición simple del conjunto de datos (*holdout*)

- Validación cruzada:
  - Submuestreo aleatorio
  - *K*-fold cross-validation
  - Stratified *K*-fold cross-validation
  - Leaving one-out cross-validation

Partición simple (*holdout*)

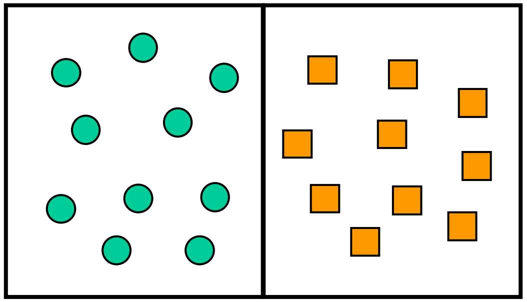
- Dividir el conjunto de datos en dos subconjuntos:
    - El primero de ellos se empleará para entrenar (conjunto de entrenamiento, *training set*)
    - El segundo se emplea para estimar el error (conjunto de prueba, *test set*)



Partición simple (*holdout*)

- Inconvenientes de este método:
  - Si se dispone de pocos datos es un "lujo" disponer de una parte importante como conjunto de prueba
  - Puesto que sólo se realiza un único experimento con un conjunto de entrenamiento → resultado engañoso si la partición no es adecuada

Conjunto de entrenamiento (no adecuado)



Conjunto de prueba (no adecuado)

- Por tanto, en general no es un buen método …

Submuestreo aleatorio

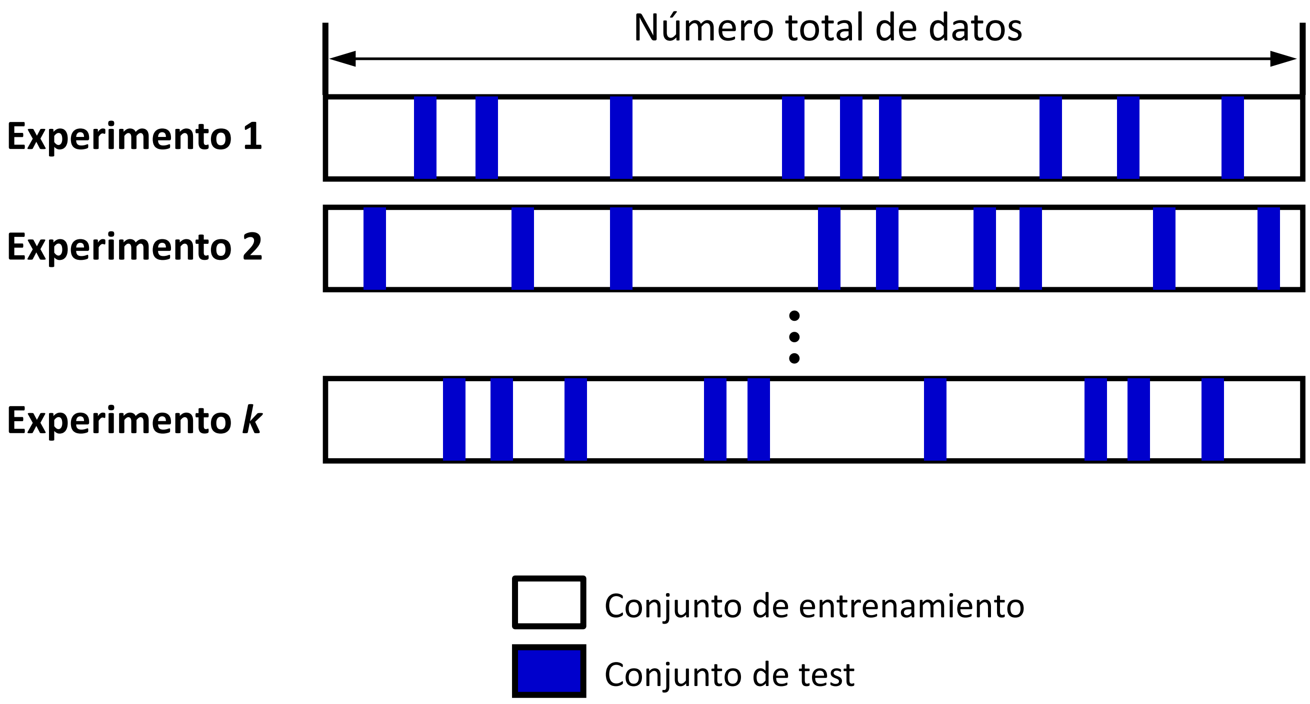
- Realiza *K* experimentos empleando como conjunto de prueba diferentes subconjuntos del conjunto de datos:
  - Cada subconjunto de prueba se escoge aleatoriamente del total de muestras (sin reemplazamiento)
  - El resto de datos se emplean para entrenar

- El estimador del error real se obtiene como la media de los errores obtenidos en los *K* experimentos (entrenamientos):

$$\mathrm{E} = \frac{1}{K}\sum_{i=1}^{K} \mathrm{E}_i$$

Submuestreo aleatorio

- Ejemplo:



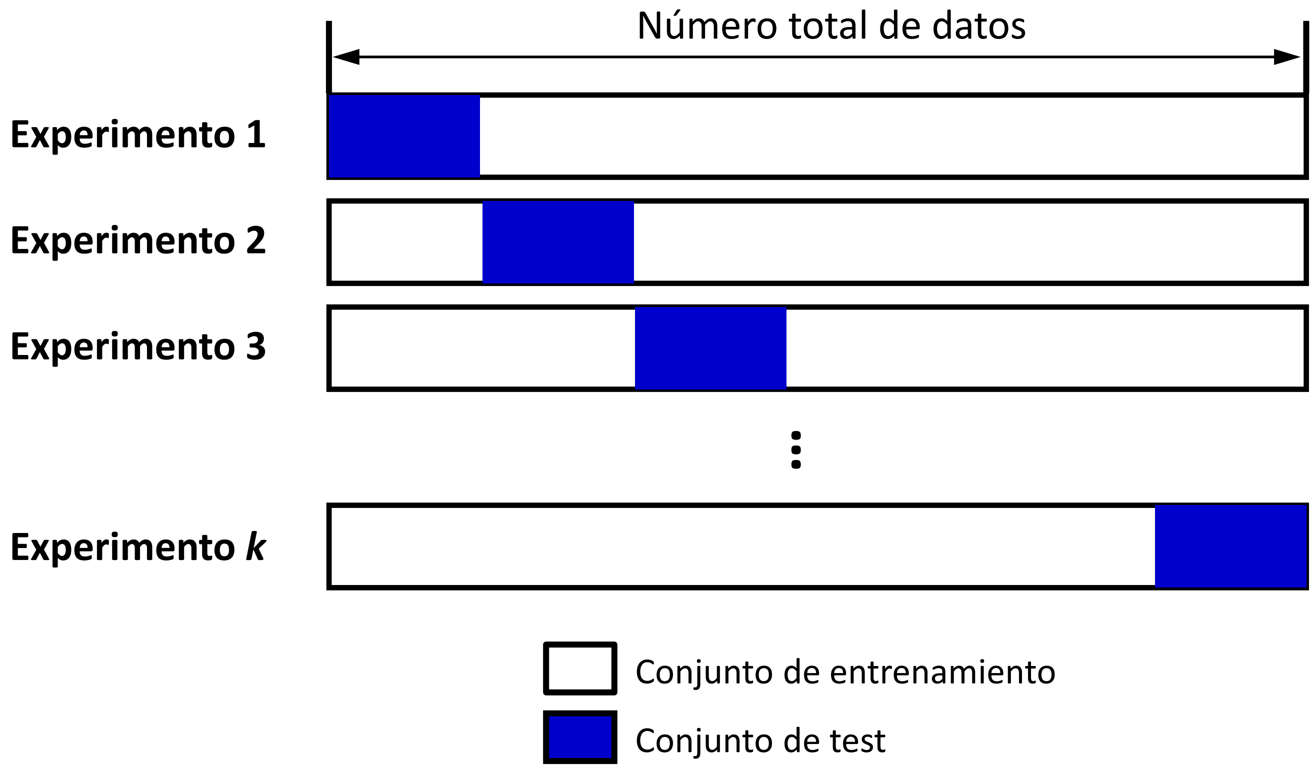
Validación cruzada *K*-fold

- Dividir el conjunto de datos en *K* subconjuntos disjuntos de aproximadamente igual tamaño
- Para $i = 1, \dots, K$ hacer:
  - Para el subconjunto *i* entrenar el sistema con los *i*-1 subconjuntos restantes (*conjunto de entrenamiento*)
  - Estimar el error sobre el conjunto *i* (*conjunto de prueba/test*): E_i
- El error de validación se calcula como la media de los errores anteriores:

$$\mathrm{E} = \frac{1}{K}\sum_{i=1}^{K} \mathrm{E}_i$$

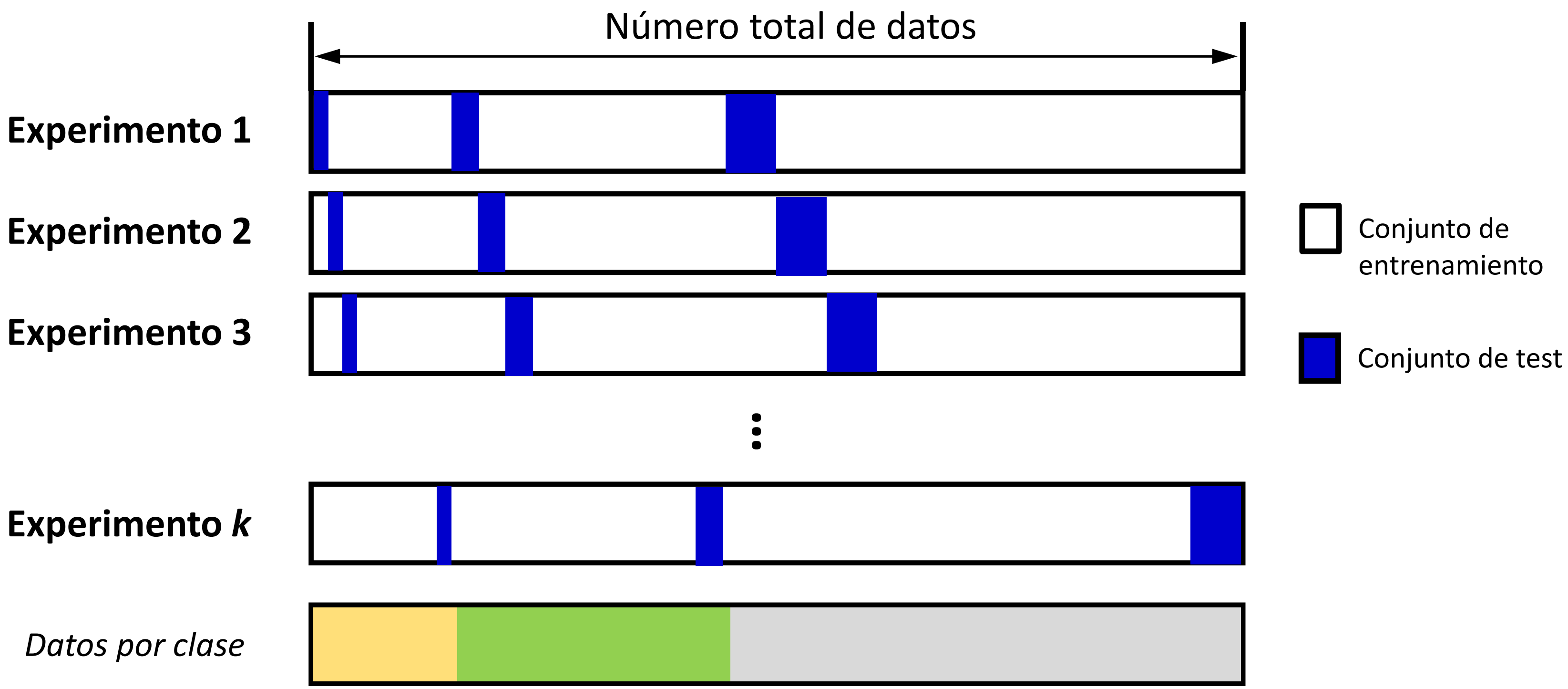
Validación cruzada *K*-fold

- Similar al submuestreo aleatorio pero tiene la ventaja de que todas las muestras del conjunto de datos se usan alguna vez para entrenar o como parte del conjunto de prueba



Validación cruzada *K*-fold estratificada

- Es una variante de la *K*-fold donde cada conjunto contiene aproximadamente el mismo porcentaje de muestras de cada clase que el conjunto completo



Leaving one-out

- Es el caso extremo de la validación cruzada *K*-fold tomando *K* como el número de muestras (*N*):

  - Para un conjunto de *N* muestras se realizan *N* experimentos

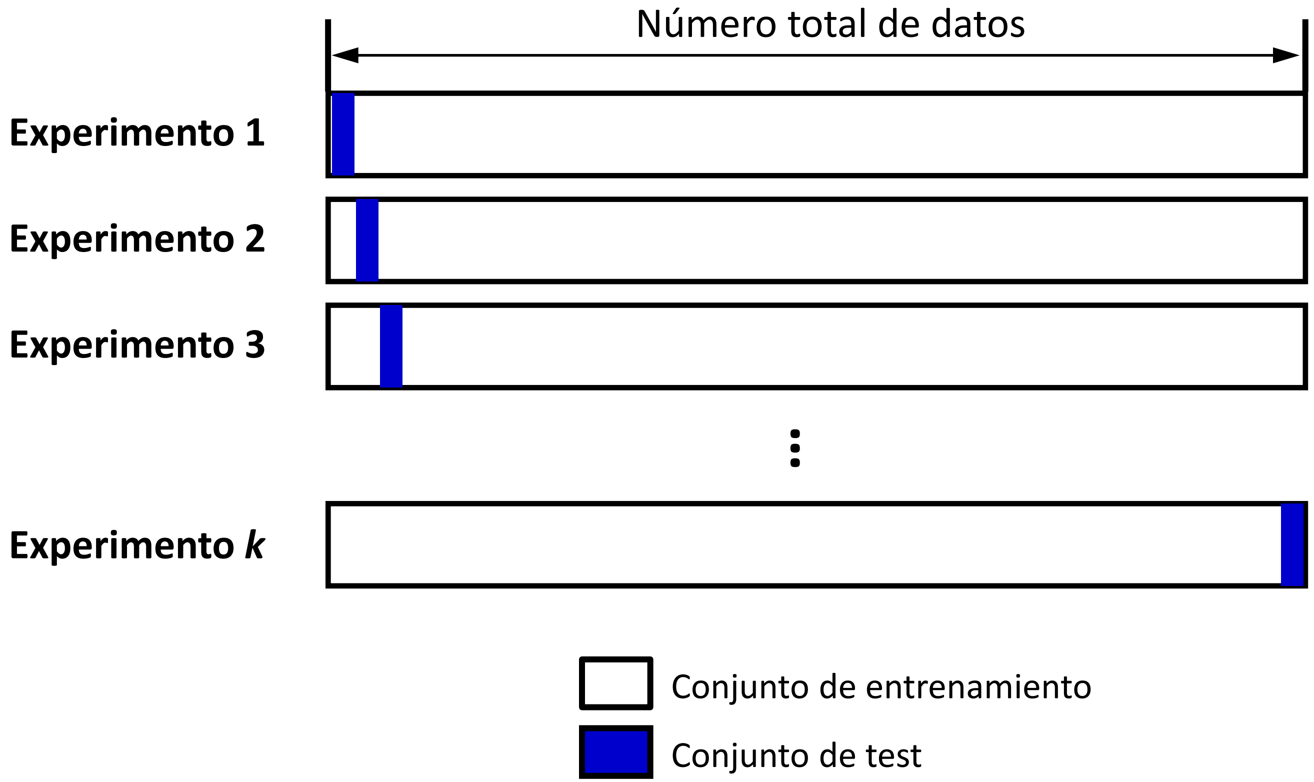
  - En cada experimento se emplean *N-1* datos para entrenar y el dato restante para prueba

- Como en los casos anteriores el error real se estima como la media de todos los conjuntos de prueba:

$$\mathrm{E} = \frac{1}{K}\sum_{i=1}^{K}\mathrm{E}_i$$

Leaving one-out

- Ejemplo:



Validación cruzada

- ¿Cuántos subconjuntos y experimentos realizar?:
  - Si se elige un gran número de subconjuntos
    - \+ El error estimado será muy preciso (sesgo pequeño respecto al error real)
    - – La varianza del error real será elevada
    - – Tiempo computacional elevado (muchos experimentos)
  - Si se eligen pocos subconjuntos
    - \+ Tiempo computacional reducido (pocos experimentos)
    - \+ La varianza del estimador será pequeña
    - – El error estimado será menos preciso (sesgo mayor respecto al error real)

Validación cruzada

- En la práctica la elección del número de subconjuntos (paquetes) depende del tamaño del conjunto de datos:

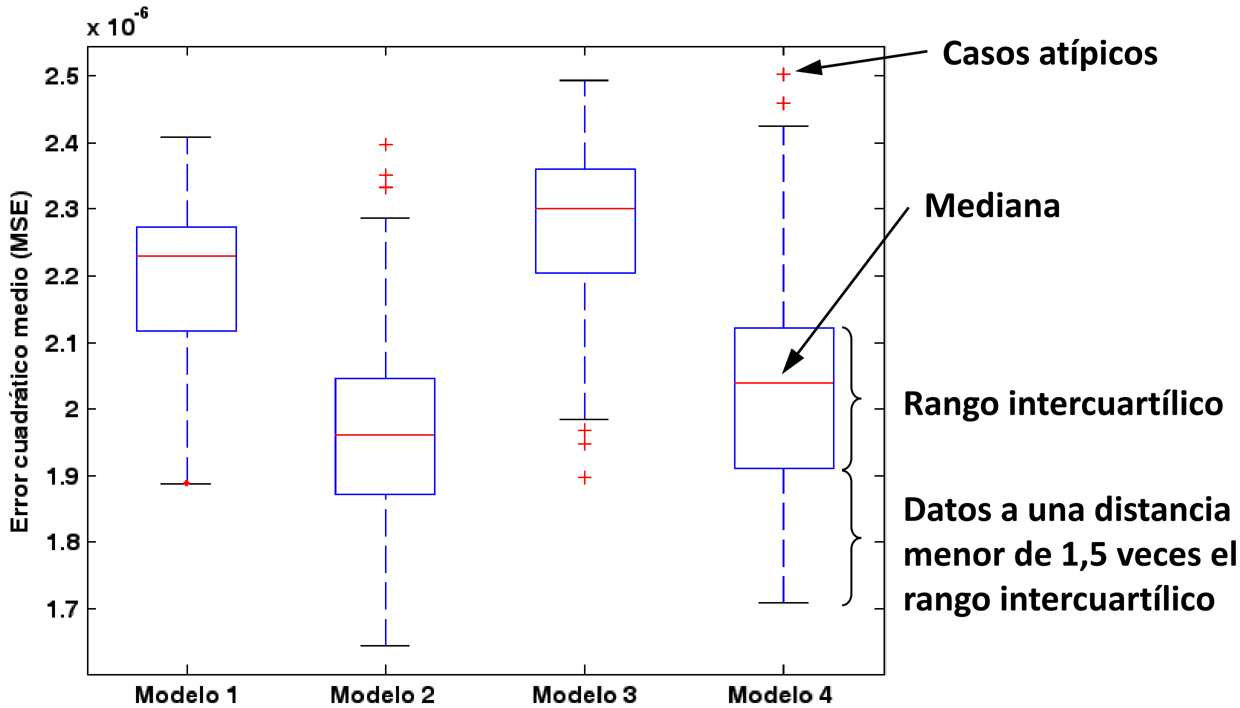
  - Para conjuntos de datos de gran tamaño incluso una validación cruzada 3-fold será bastante precisa

  - Para conjuntos de datos pequeños, se puede emplear la *leaving one-out* para tener en el conjunto de entrenamiento tantos datos como sea posible

- Una elección habitual de la *K*-fold es *K*=10

Diagrama de caja (*boxplot*)

- Herramienta interesante para mostrar gráficamente los resultados de varias simulaciones del modelo:



Comparación de modelos

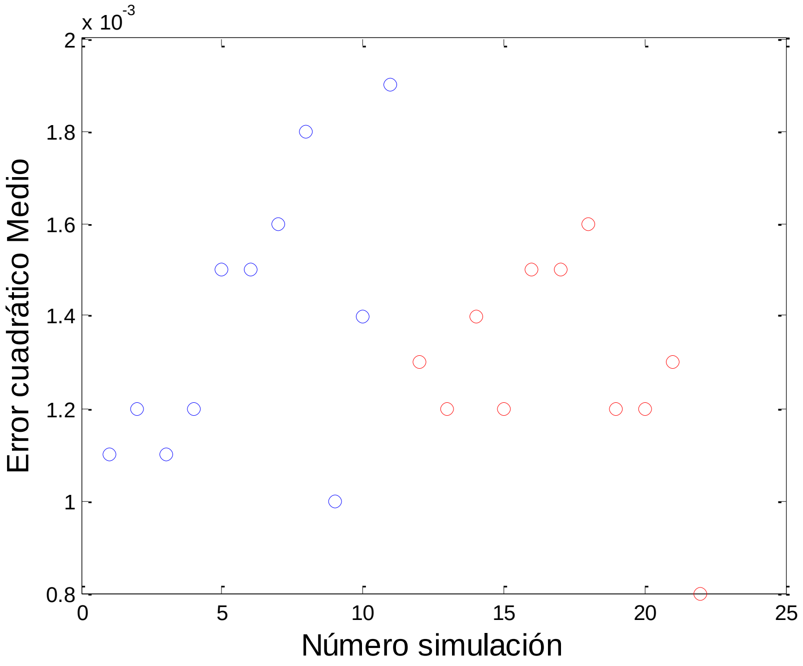
- Objetivos:

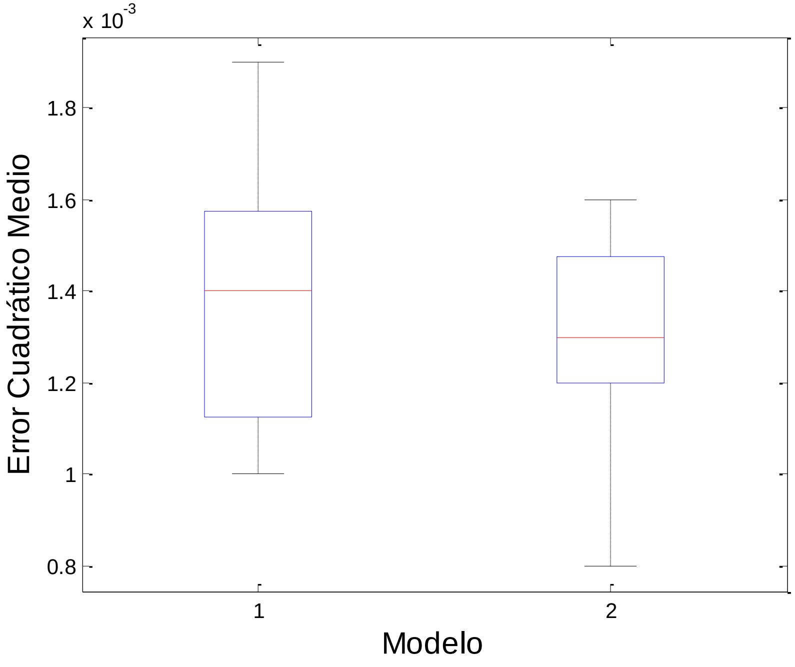
  - Comparar varios modelos para determinar el mejor: el que proporcionará mejor rendimiento en el futuro (con nuevos datos)

  - Determinar si las diferencias de error observadas entre los modelos son estadísticamente significativas

Comparación de modelos

- Ejemplo: ¿Cuál es el mejor de estos modelos en términos de error?





- Las diferencias entre los métodos, ¿son estadísticamente significativas?

Comparación de dos modelos

- Premisas:
  - Dos modelos a comparar sobre un conjunto de datos
  - Un conjunto *k* de errores cometidos por el modelo 1
  - Un conjunto *k* de errores cometidos por el modelo 2
- Preguntas:
  - ¿El rendimiento de ambos modelos es el mismo?
  - ¿Hay diferencias estadísticamente significativas entre ambos?
- Solución a las preguntas anteriores: contraste (test) de hipótesis
  - Método estadístico para comprobar la validez o no de una hipótesis (hipótesis nula)

Comparación de dos modelos

- **Etapas** del contraste de hipótesis:

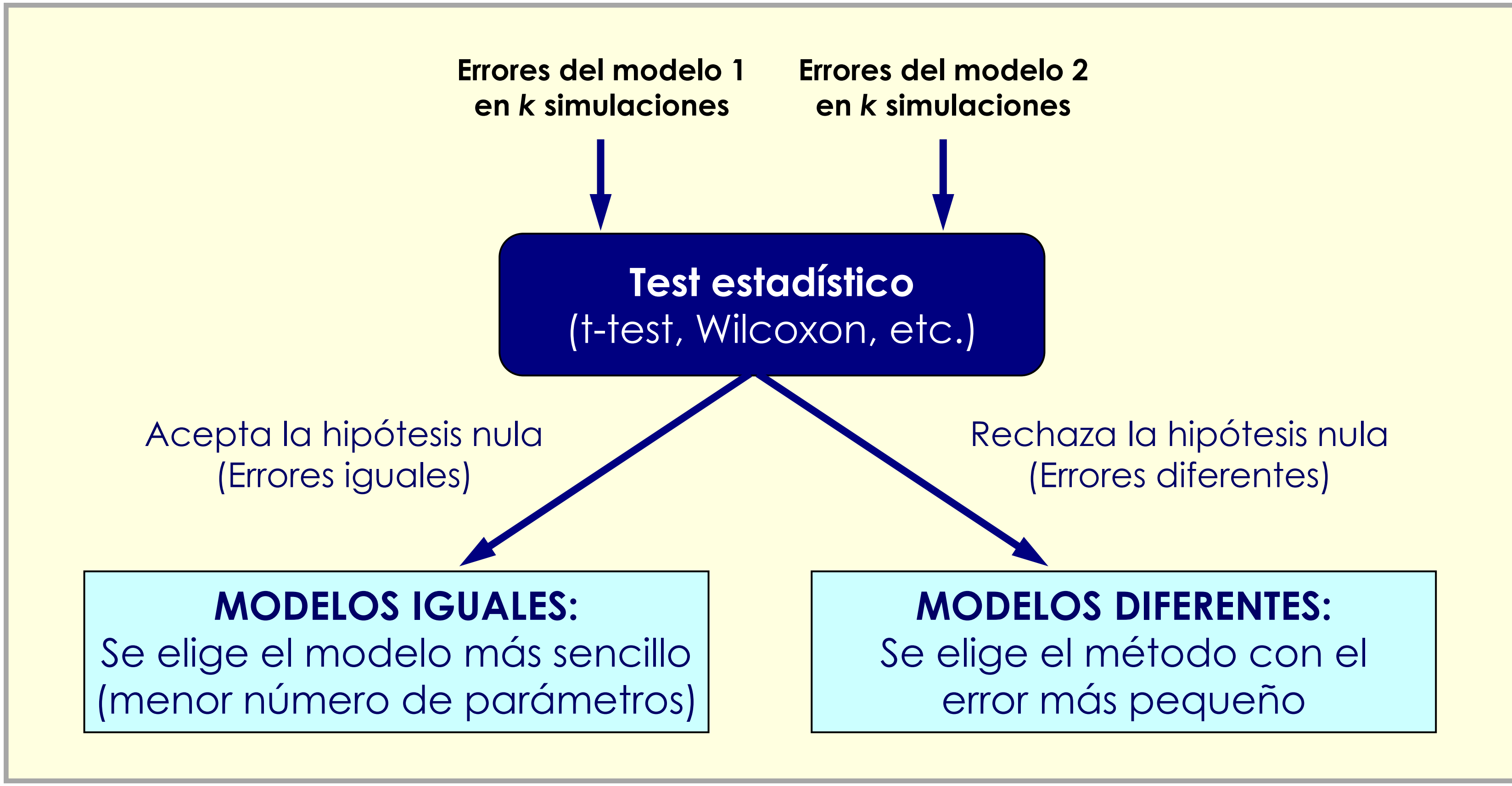
1. Definir la hipótesis nula (H_0)
2. Seleccionar un test estadístico (*estadístico del contraste*) que pueda emplearse para evaluar la validez de H_0
3. Elegir el nivel de significación (α) del test: probabilidad de rechazar H_0 siendo cierta
4. Calcular el *p*-valor (probabilidad de obtener una discrepancia mayor de la observada siendo H_0 cierta)
5. Comparar el *p*-valor obtenido con nivel de significación:
   - Si $p \leq \alpha$ → Rechazar H_0
   - Si $p > \alpha$ → Aceptar H_0

El p-valor informa sobre cuál sería el nivel de significación más pequeño que nos permitiría rechazar la hipótesis nula

Comparación de dos modelos

- Tipos de test estadísticos empleados en selección entre dos modelos:
  - T-test
    - Evalúa las diferencias entre las medias (errores medios) de dos modelos. Hipótesis nula: $\mu_1 - \mu_2 = 0$
    - Suposiciones de este test estadístico: ambas distribuciones siguen una distribución *Normal* con idénticas varianzas
  - Test de Wilcoxon
    - Evalúa las diferencias entre las mediana de dos modelos. Hipótesis nula: $m_1 - m_2 = 0$
    - Suposiciones de este test estadístico: ninguna

- Metodología:

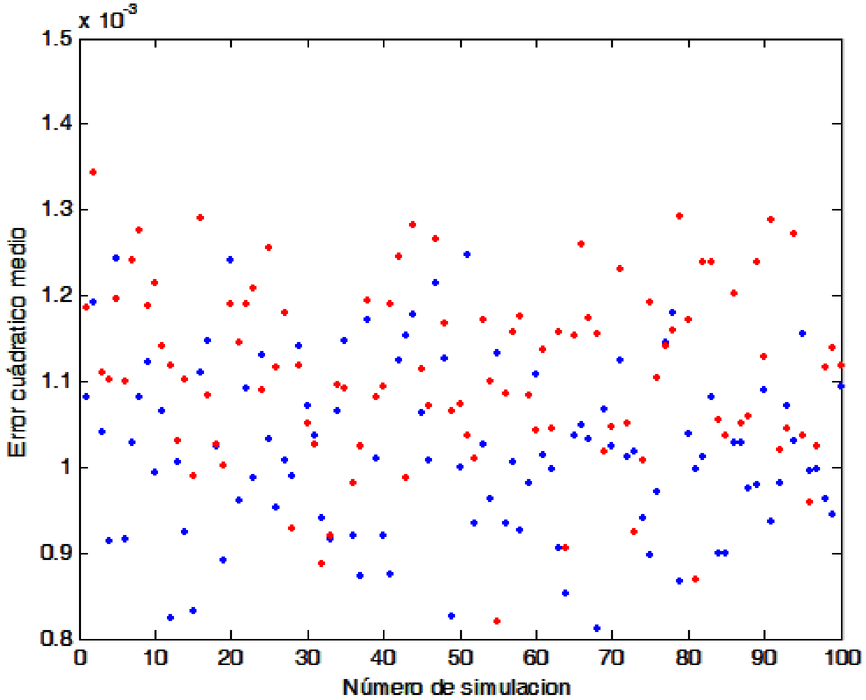


Comparación de dos modelos

- ¿Cuál de los dos métodos de contrastes de hipótesis es más adecuado?:
  - Si se cumplen las suposiciones del t-test, este método es más *potente* (mayor probabilidad de rechazar H_0 cuando es falsa)
  - Cuando no se cumplen las suposiciones del t-test, el del Wilcoxon es más potente y más *fiable* (no asume ninguna distribución)
  - El test de Wilcoxon es más robusto frente a casos atípicos (*outliers*)
- Conclusión general: si no se conoce la distribución de los errores de cada método → emplear test de Wilcoxon

Comparación de dos modelos

- **Ejemplo**: Dados dos modelos diferentes con los siguientes errores en el conjunto de prueba de 100 simulaciones



¿Hay diferencias *estadísticamente significativas* entre ambos modelos?

Comparación de dos modelos

- **Ejemplo**: Se realiza un t-test para comprobarlo
  - Hipótesis nula (H_0): $\mu_1 - \mu_2 = 0$ (ambas medias son iguales)
  - Nivel de significación (α): 0,01
  - Resultado obtenido:
    - *p*-valor del test: $1{,}1934 \times 10^{-9}$
    - Puesto que $p \leq \alpha$ → Rechazamos H_0 con un nivel de confianza del 99%

Comparación de múltiples modelos

- Si sólo se dispone de dos grupos de observaciones (dos modelos)
  - Se puede comparar el error medio empleando un t-test o un test de Wilcoxon
- Sin embargo, si existen más de dos grupos (tres o más modelos):
  - No es apropiado simplemente comparar cada par de modelos empleando por ejemplo un t-test ya que la posibilidad de detectar incorrectamente una diferencia significativa aumenta con el número de comparaciones
- ¿Cuál es la metodología correcta en este caso?

Comparación de múltiples modelos

- Metodología para múltiples modelos:

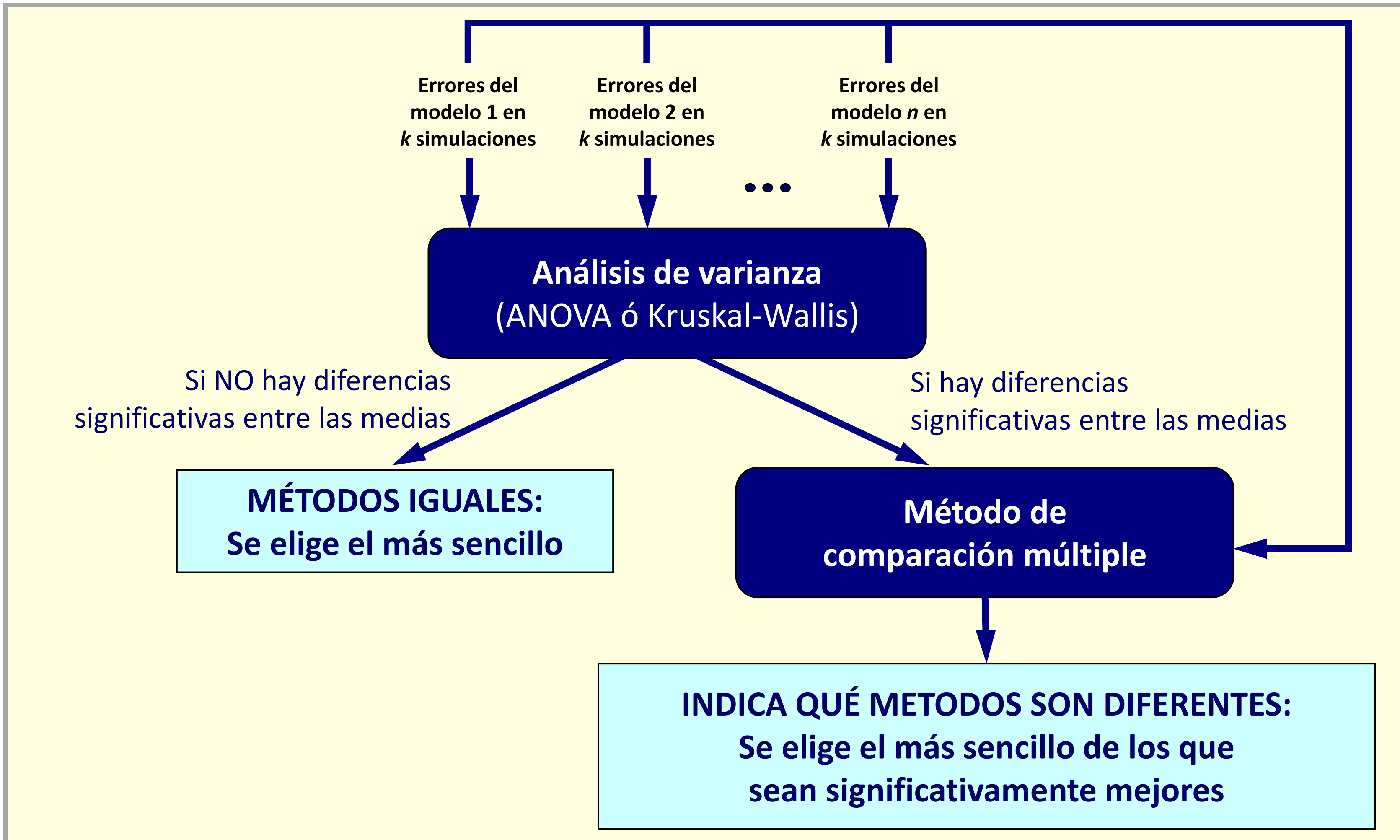
  1. Emplear un análisis de varianza ANOVA o un test de Kruskal-Wallis para identificar si hay una diferencia significativa entre todas las medias

  2. Si el test de varianza concluye que sí hay diferencias
     - Hay que investigar cuáles son diferentes empleando un método de comparación múltiple

  3. Si el test de varianza concluye que no hay diferencias
     - Todas las medias iguales → todos los modelos iguales (se elige el más sencillo)

Comparación de múltiples modelos

- Metodología:



Comparación de múltiples modelos

- Análisis de varianza (ANOVA):
  - Test paramétrico que compara las medias de modelos
  - Hipótesis nula: todas las medias son iguales (provienen de la misma población o de diferentes poblaciones pero con la misma media)
  - Suposiciones del test:
    - Todas las muestras de las diferentes poblaciones están normalmente distribuidas
    - Todas de las muestras de las diferentes poblaciones tienen la misma varianza
    - Todas las observaciones son mutuamente independientes
  - El test sigue siendo robusto para observaciones que no cumplan "ligeramente" las dos primeras suposiciones

Comparación de múltiples modelos

- Análisis de varianza (Kruskal-Wallis):
  - Test no paramétrico que compara las medias de diversos modelos
  - Hipótesis nula: todas las medias son iguales (provienen de la misma población o de diferentes poblaciones pero con la misma media)
  - Suposiciones del test:
    - Todas las observaciones provienen de una población continua
    - Todas las observaciones son mutuamente independientes

Comparación de múltiples modelos

- Métodos de comparación múltiple: comparan las diferencias entre cada par de medias con ajustes apropiados a la comparación múltiple:
  - Método de Tukey
  - Método de Holm-Bonferroni
  - Método de Scheffé