Clase 181 — Corrección de comparaciones múltiples (Bonferroni, FDR)

Parte: 3 — Estadística Inferencial y Causal · Fuente: ISLP, cap. 13 Multiple Testing + Benjamini & Hochberg (1995). Duración estimada: 70 min.

Clase 181 — Corrección de comparaciones múltiples (Bonferroni, FDR)

> *Parte: 3 — Estadística Inferencial y Causal · Fuente: ISLP, cap. 13 Multiple Testing + Benjamini & Hochberg (1995). Duración estimada: 70 min.*

Objetivo

Entender por qué hacer 100 tests al α=0.05 produce ≈ 5 falsos positivos esperados aunque todas las H sean verdaderas, y aplicar las dos familias de corrección: family-wise error rate (FWER) con Bonferroni y Holm, y false discovery rate (FDR) con Benjamini-Hochberg (BH). Saber elegir entre ambas según el contexto (medicina/seguridad → FWER; screening exploratorio → FDR).

Resultados de aprendizaje

Al finalizar, el estudiante podrá:

- Cuantificar la inflación de α al hacer k tests independientes: 1 - (1-α)^k.
- Aplicar Bonferroni: α_corregido = α / m. Conservador pero simple.
- Aplicar Holm-Bonferroni (statsmodels.stats.multitest.multipletests(..., method='holm')) — uniformemente más poderoso que Bonferroni.
- Aplicar Benjamini-Hochberg (BH/FDR) y entender que controla la proporción esperada de falsos positivos entre los rechazos, no el FWER.
- Distinguir FWER (P[al menos 1 falso positivo] ≤ α) de FDR (E[V/R] ≤ q, donde V son falsos positivos y R rechazos totales).
- Reportar p-values ajustados (q-values) y umbrales claros.

Temas

- El problema: si m=20 tests independientes con H verdadera y α=0.05, P(al menos uno rechaza) = 1 - 0.95^{20} ≈ 64 %.
- FWER: probabilidad de al menos 1 falso positivo en toda la familia.
- FDR: proporción esperada de falsos positivos entre los rechazos (no entre todos los tests).
- Bonferroni: rechazar si p_i ≤ α/m. Controla FWER exactamente.
- Holm: ordenar p-values y comparar p_(i) ≤ α/(m-i+1). Uniformemente más poderoso que Bonferroni.
- Benjamini-Hochberg (BH): ordenar p_(1) ≤ ... ≤ p_(m); rechazar todos los p_(i) tales que p_(i) ≤ (i/m)·q. Controla FDR a nivel q.
- Cuándo usar cada uno: FWER si un falso positivo es catastrófico (drug approval, security). FDR si esperás muchos descubrimientos verdaderos y querés tolerar algunos falsos (genómica, A/B testing masivo).

Definiciones y características

- Family-Wise Error Rate (FWER): P(rechazar al menos una H verdadera). Sin corrección, crece con m. Bonferroni y Holm lo acotan.
- False Discovery Rate (FDR): E[V / max(R, 1)] donde V = falsos positivos, R = total de rechazos.

Concepto introducido por Benjamini & Hochberg (1995). Mucho menos restrictivo que FWER.

- Bonferroni: divide α por m. Si m=100 y α=0.05, cada test usa α_local=0.0005. Conservador, pierde poder con m grande.
- Šidák: α_corregido = 1 - (1-α)^(1/m). Marginalmente menos conservador que Bonferroni; asume independencia.
- Holm-Bonferroni (1979): step-down. Ordena p-values, compara el más chico contra α/m, el siguiente contra α/(m-1), etc. Uniformemente mejor que Bonferroni.
- BH (Benjamini-Hochberg): step-up. Ordena, encuentra el mayor i tal que p_(i) ≤ (i/m)·q, rechaza todos hasta ese. Controla FDR si los tests son independientes o tienen positive regression dependence (BY si la dependencia es arbitraria).
- q-value: p-value ajustado bajo FDR. Interpretación: "si rechazo todo con q ≤ 0.05, espero que ≤ 5 % de mis rechazos sean falsos".

Dataset / recursos

- Genómica sintética: simular m=1000 tests, m=950 nulos verdaderos y m=50 alternativos. Generar p-values y mostrar comportamiento de cada método.
- A/B testing real: 20 métricas testeadas a la vez → controlar familia.
- Librerías: statsmodels.stats.multitest, pingouin.multicomp, scipy.stats.

Ejercicios

1. Inflación de α: simulá 10 000 experimentos. En cada uno, hacé 20 tests con H verdadera (scipy.stats.ttest_ind entre dos grupos N(0,1), n=30). Contá en qué % al menos 1 da p<0.05. Verificá que ≈ 64 %.
2. Bonferroni: con un vector pvals de 20 p-values, calculá pvals_adj = np.minimum(pvals * 20, 1) y compará contra multipletests(pvals, method='bonferroni').
3. Holm: multipletests(pvals, alpha=0.05, method='holm'). Comparar cuántos rechaza vs Bonferroni con el mismo vector.
4. BH/FDR: genera 1000 p-values, 950 de Uniform(0,1) y 50 de Beta(0.5, 5) (concentrados cerca de 0 — alternativos). Aplicá multipletests(pvals, alpha=0.05, method='fdr_bh'). Contá cuántos rechaza y estimá el FDR empírico (rechazos del primer grupo / total rechazos).
5. Comparación: mismo vector del ej. 4, aplicar Bonferroni, Holm y BH. Tabla con: # rechazos, % de los 50 verdaderos descubiertos (recall), FDR empírico. Verificá que BH descubre mucho más con FDR controlado.

Homework verificable

Sobre un dataset con 30 features y un target binario (ej.: load_breast_cancer):

1. Para cada feature, t-test entre la clase 0 y la clase 1.
2. Aplicar tres correcciones: Bonferroni, Holm, BH (q=0.05).
3. Tabla comparativa: # features significativas según cada método.
4. Justificar en 3 líneas qué método elegirías si: (a) vas a publicar los hallazgos en un paper médico, (b) usás esto como screening para una etapa siguiente.

Criterio de aceptación: BH debe descubrir más features que Holm, que descubre ≥ que Bonferroni. La justificación debe mencionar que (a) → FWER (Bonferroni/Holm), (b) → FDR (BH).

Errores comunes

Síntoma / mensaje	Causa y cómo arreglar	
Hago 50 t-tests sin corrección y reporto l	Cherry-picking estadístico. Fix: BH como m	
Aplico Bonferroni con m=1000 y no rechaza	Demasiado conservador para screening. Fix:	
Aplico BH a tests no independientes (ej.:	BH clásico asume independencia o PRDS. Fi	
Reporto p-value ajustado como "probabilida	El p-value ajustado sigue siendo un p-valu	datos)`. Fix: usar lenguaje preciso ("cont
Decido cuál corrección usar después de ver	También es p-hacking. Fix: pre-especificar	

Preguntas frecuentes

¿Cuándo Bonferroni y cuándo BH?

Regla simple: Bonferroni cuando el costo de un falso positivo es alto (medicina, seguridad, decisiones binarias). BH cuando hacés screening y aceptás cierta proporción de FP entre los hallazgos para no perder verdaderos (genómica, A/B testing masivo, feature selection).

¿Por qué Holm es siempre mejor que Bonferroni?

Porque rechaza al menos los mismos tests y a veces más, sin perder control del FWER. La única razón para usar Bonferroni es simplicidad pedagógica.

¿BH controla FDR exactamente al 5 %?

Controla FDR ≤ $q \cdot (m/m)$, donde m es el número de H verdaderas. En la práctica, $m \leq m$, así que FDR ≤ q. Si esperás muchos nulos, BH es un poco más conservador de lo que parece.

¿Storey's q-value es lo mismo que BH?

Es una versión adaptativa: estima $\pi = m/m$ de los datos y corrige menos cuando hay muchos rechazos esperados. En genómica es estándar; en data science general, BH alcanza.

¿Tengo que corregir si los tests son sobre datasets distintos?

Si forman parte del mismo objetivo de inferencia (la misma "familia"), sí. Si son análisis independientes con conclusiones separadas, no. El límite es subjetivo; pre-registrar la familia ayuda.

Referencias

- ISLP, cap. 13 — Multiple Testing.
- Benjamini, Y. & Hochberg, Y. (1995), Controlling the False Discovery Rate, JRSS Series B.
- Holm, S. (1979), A Simple Sequentially Rejective Multiple Test Procedure, Scandinavian Journal of Statistics.
- statsmodels.stats.multitest.multipletests — todos los métodos en una sola función.
- Efron, B. (2010), Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction.

Siguiente clase

Clase 182 — Intervalos de confianza

Apéndice: notebook (primer bloque)

Primera celda ejecutable del notebook de la clase.

```
# Imports y configuración inicial
```

Archivos complementarios

- notebook.ipynb