STA 35C: Statistical Data Science III

Lecture 19: Multiple Hypotheses Testing (cont'd) + Review for Midterm 2

Dogyoon Song

Spring 2025, UC Davis

Announcement

Midterm 2 on Fri, May 16 (12:10 pm–1:00 pm in class)

- See Canvas announcement (or Lec. 17/18 slides) for allowed materials, etc.
- Coverage: Lectures 12–19
- A practice midterm and answer key are available on the course webpage
- Office hours this week:
  - Instructor: Wed, 4–6pm (extended); no OH Thu
  - TA: Thu 1–2pm

Remote lecture (Zoom) on Mon, May 19

- Zoom link will be emailed via Canvas

Today's topics

- **Recap: Multiple hypotheses testing**
  - Goals to control false positives

- **Brief review for midterm 2**
  - Cross-validation
  - Bootstrap
  - Subset selection
  - Regularization
  - Multiple hypotheses testing

Recap: Multiple testing

Single-hypothesis test:

- Typically set up H_0, and gather data to reject it if there is significant evidence
- Type I error $=$ false positive; Type II error $=$ false negative
- Each test has Type I error at most α (e.g. 0.05)

Modern data analysis: multiple tests simultaneously

- E.g. Testing thousands of predictors or biomarkers to discovery significant ones
- If m is large, false rejections can occur easily by chance
- On average, $\alpha \times m$ false positives if each is tested at level α

Key challenge: How to address inflated false positives as m grows

Hypothesis testing as classification

A single hypothesis test classifies H_0 as "true or not":

- **Goal:** Discover "real phenomenon" (H_1) or conclude non-existence (H_0)
  - H_0 is true $\iff$ no real effect
  - H_0 is false $\iff$ there is a real effect (H_1)
  - We "discover" an effect by rejecting H_0
- **Test as classification:** Depending on evidence gathered from data,
  - Reject $H_0 \iff$ classify $\hat{H} = 1$
  - Fail to reject $H_0 \iff$ classify $\hat{H} = 0$

	H_0 is true (**"H=0"**)	H_0 is not true (**"H=1"**)
Reject H_0 (**"$\widehat{\mathbf{H}}$=1"**)	FP (Type I)	TP
Not reject H_0 (**"$\widehat{\mathbf{H}}$=0"**)	TN	FN (Type II)

Hypothesis test at level α

Consider the probabilities of each outcome for hypothesis test

	H_0 is true (**"H=0"**)	H_0 is not true (**"H=1"**)
Reject H_0 (**"$\widehat{\mathbf{H}}$=1"**)	p_{FP}	p_{TP}
Not reject H_0 (**"$\widehat{\mathbf{H}}$=0"**)	p_{TN}	p_{FN}

Hypothesis test at level α:

- $\Pr(\text{reject } H_0 \mid H_0 \text{ true}) \;\leq\; \alpha$
- That is, the chance of a false positive is at most α

$$\Pr(\hat{H} = 1 \mid H = 0) = \frac{\Pr(\hat{H} = 1 \;\&\; H = 0)}{\Pr(H = 0)} = \frac{p_{\mathrm{FP}}}{p_{\mathrm{FP}} + p_{\mathrm{TN}}} \leq \alpha$$

Testing multiple hypotheses at level α

Suppose we test m hypotheses $H_{0,1}, \ldots, H_{0,m}$, all at level α, obtaining confusion matrix:

	H_0 is true	H_0 is not true
Reject H_0	N_{FP}	N_{TP}
Not reject H_0	N_{TN}	N_{FN}

- $N_{\mathrm{FP}}, N_{\mathrm{TP}}, N_{\mathrm{TN}}, N_{\mathrm{FN}}$ are random variables that sum to m
- Roughly, we expect $N_{\mathrm{FP}} \approx m \cdot p_{\mathrm{FP}}$; when all m nulls are true, $N_{\mathrm{FP}} \approx m \cdot \alpha$

If these m tests are independent,

- Probability of *at least one* false positive $\approx 1 - (1 - \alpha)^m$
- For $m = 20, \alpha = 0.05$, that probability is $\approx 64\%$

Family-wise error rate (FWER)

	H_0 is true	H_0 is not true
Reject H_0	N_{FP}	N_{TP}
Not reject H_0	N_{TN}	N_{FN}

Goal: Ensure $N_{\text{FP}} < 1$ with high probability

$$\mathrm{FWER} = \Pr(N_{\text{FP}} \geq 1)$$

- Bonferroni correction sets each test at α/m to keep $\mathrm{FWER} \leq \alpha$ (union bound)
- Holm's step-down procedure refines this by adapting thresholds step by step

Interpretation: Controlling $\mathrm{FWER} \leq \alpha$ ensures we have *no* Type I errors with probability at least $1 - \alpha$

False discovery rate (FDR)

	H_0 is true	H_0 is not true
Reject H_0	N_{FP}	N_{TP}
Not reject H_0	N_{TN}	N_{FN}

FDR Strategy: Increase N_{TP} at the cost of tolerating a moderate N_{FP}

- Strict FWER control often yields many Type II errors (missing real signals)
- FDR-based approach lets us accept some false positives but aims for higher power (detecting more TP)
    - N_{FP}: "false discoveries"
    - N_{TP}: "true discoveries"

False discovery rate control

	H_0 is true	H_0 is not true
Reject H_0	N_{FP}	N_{TP}
Not reject H_0	N_{TN}	N_{FN}

False discovery proportion: fraction of false discoveries among all "claimed" ($\hat{H} = 1$)

$$\text{FDP} = \frac{N_{\text{FP}}}{N_{\text{FP}} + N_{\text{TP}}}$$

False discovery rate (FDR): $\text{FDR} = \mathbb{E}[\text{FDP}]$

- Controlling FDR at q (e.g., 5% or 10%) means $\mathbb{E}[\text{FDP}] \leq q$
- Methods like Benjamini–Hochberg aim to maintain FDR $\leq q$ while rejecting more nulls than strict FWER approaches

Pop-up quiz: Comparing FDR vs. FWER

You have m hypotheses to test. The *False Discovery Rate* (FDR) is defined as $\mathbb{E}[\text{FDP}]$, where $\text{FDP} = \frac{\#\text{FP}}{\#\text{FP}+\#\text{TP}}$. **Which statement best captures differences between FDR and FWER**?

(A) FDR forces the probability of *zero* false positives to stay below α, whereas FWER allows a small fraction q.

(B) FDR aims to keep $\mathbb{E}[\text{fraction of false positives among rejections}] \leq q$, while FWER demands $\Pr(\text{at least one false positive}) \leq \alpha$.

(C) Under FDR control, no false positives are allowed once you discover enough true positives.

(D) FDR only works for independent tests, but FWER can handle correlated tests without adjustments.

Answer: (B).
FDR control (e.g., Benjamini–Hochberg) allows a certain fraction of false positives *on average*, whereas FWER control (e.g., Bonferroni/Holm) requires the chance of *any* false positive be controlled below α.

Review: Cross-validation

Goal: Estimate test performance from training data alone

Key ideas:

- Single split (validation set): random partition into train/test; simple but high variance
- LOOCV (leave-one-out): train on $n-1$ points, validate on 1 point, repeat for all points
- k-fold CV: partition data into k folds, systematically rotate which fold is the validation set

Trade-offs:

- Fewer folds (e.g. 5- or 10-fold) reduce computation but can have slightly higher variance
- LOOCV uses maximum training size ($n-1$) but is more expensive and can have higher correlation across folds

Usage:

- Model selection: pick model that yields lowest CV error
- Tuning parameters (e.g. λ in ridge/lasso)

Review: Bootstrap

Goal: Approximate the sampling distribution (e.g. standard errors) using just one dataset

Method:

- Sample n points *with replacement* from the original dataset of size n (a "bootstrap sample")
- Compute desired statistic (mean, regression coefficient, etc.) on the bootstrap sample
- Repeat B times, forming a distribution of the statistic estimates $\{\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*\}$

Bootstrap SE/CI:

- Standard error $\approx \mathrm{SD}(\hat{\theta}_b^*) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \bar{\theta}^*)^2}$
- Use percentiles or normal approximation to construct confidence intervals
- Interpreting the coverage of confidence intervals requires care

Key premise: The observed sample is representative of the population

Review: Subset selection

Goal: Identify a relevant subset of predictors among many

Best subset selection:

- Tries all 2^p subsets (exhaustive); picks the best model for each size k, then chooses among them by adjusted R^2, CV, etc.
- Feasible only if p is small or moderate (can be very expensive for large p)

Forward/backward stepwise:

- Greedy approximations: add/remove one predictor at a time
- Complexity $\mathcal{O}(p^2)$ vs. 2^p for best subset
- Might miss the absolute best subset but often works well in practice

Pros/Cons:

- Direct variable selection (some coefficients set to zero)
- Can be unstable for large p; small changes in data may change chosen subset

Review: Regularization

Motivation: Least squares can be unstable or undefined if $p \approx n$ or $p > n$; high variance or collinearity issues

Ridge regression:

- Add penalty $\lambda \sum_j \beta_j^2$
- Typically shrinks all coefficients; no exact zeros
- More stable under collinearity

Lasso:

- Add penalty $\lambda \sum_j |\beta_j|$
- Can zero out some coefficients, enabling variable selection
- Slightly less stable than ridge if predictors are highly correlated

Tuning λ: Usually chosen by cross-validation; neither ridge nor lasso always wins—depends on data and interpretability needs

Review: Multiple hypotheses testing

Problem: Testing many hypotheses inflates chance of false positives

- Probability(≥ 1 false positive) can be $1 - (1 - \alpha)^m$ if tests are independent
- p-hacking: repeatedly searching for small p-values leads to spurious "discoveries"

FWER (Family-Wise Error Rate):

- Probability of any (=at least 1) false positive
- Bonferroni, Holm's step-down keep FWER $\leq \alpha$
- Often conservative, can reduce power when m is large

FDR (False Discovery Rate):

- Expected fraction of false positives among rejections (=FP + TP)
- Benjamini–Hochberg procedure can control FDR
- Less conservative, typically yields more rejections, tolerating some false positives