Statistical Inference

> *"Statistics is the grammar of science."* – Karl Pearson

Inference v.s. Prediction

- Prediction: finding the best model that generalizes
- Inference: confirming a hypothetical (causal) relationship from data

Bayesian v.s. Frequentist

- Bayesian: parameters are random but data are fixed
  Uncertainty directly available from the posterior distribution
- Frequentist: true parameters fixed but unknown
  Uncertainty as a result from having finite samples

Sampling distributions

The distribution of a statistics under repeated sampling from the same population using the same sampling mechanism

- Acknowledging uncertainty of evidence from data
- Foundation of statistical reasoning
- Key to propert statistical inference

Example: Central Limit Theorem and Sampling Distribution

Statistical Reasoning

Given that the observations are finite samples from a true distribution...

- Confidence Interval
  A set of estimators that guarantee chosen coverage probability.
- Hypothesis Testing
  Assuming the null hypothesis, is the dataset an extreme case?

- Multiple Testing
  If one tests infinite number of hypotheses, one of them might be positive by chance

Other Common Tests

- Model-based Tests
- Rank-based Tests
- Simulation-based Tests

Rank-based Tests

One way to avoid parametric assumptions is to consider the order/ranking instead of the actual numeric values.

Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ denote the ordered sample. The **rank** of an observation x_i is

$$R_i \;=\; 1 + \#\{\, x_j : x_j < x_i \, j \neq i\}.$$

Bottom-line: For similar hypotheses, a parametric test is more efficient than a nonparametric test, while the nonparametric offers more robustness towards violation of assumptions.

Model-based Tests

These tests are based on the comparison between two *nested* models, where the full model contain all parameters in the reduced model.

- F-test/ANOVA
- Likelihood ratio test
- ...

Model-based Tests: F-test/ANOVA

Consider a linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0},\, \sigma^2\mathbf{I}\right)$$

Want to test q linear constraints $H_0:\; \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$

Test statistic

$$F = \frac{(\mathrm{RSS_r} - \mathrm{RSS_f})/q}{\mathrm{RSS_f}/(n-p)}.$$

where $\mathrm{RSS_r}$ is the residual sum of squares from reduced model and $\mathrm{RSS_f}$ is the residual sum of squares from the full model.

Then under the null hypothesis $F \ \sim \ F_{q,\, n-p}$ q = 1 here.

Likelihood Ratio test (LRT)

Using the logistic regression as an example

$$\Pr(Y = 1 \mid \mathbf{x}) \ = \ \mathrm{logit}^{-1}(\beta_0 + \beta_1 \mathrm{age} + \beta_2 \mathrm{smoke}).$$

To test $H_0 : \beta_2 = 0$ ("smoking not associated"):

$$\Lambda = -2\left[\log \mathcal{L}_\mathrm{r} - \log \mathcal{L}_\mathrm{f}\right],$$

where $\mathcal{L}_\mathrm{r}$ is the likelihood of the reduced model and $\mathcal{L}_\mathrm{f}$ is the likelihood of the full model.

Under H_0, $\Lambda \rightarrow \ \chi^2_{df=1}$ as sample size approaches infinity.

Assumptions

Correctly specified likelihood (independence, link function); sample size large enough.

> Note: For linear regression with normal errors, F-test is equivalent to LRT as sample size approaches infinity.

Simulation-based Tests

Hypothesis testing hinges on the null distribution of a chosen test statistic.

When the closed-form null distribution is unavailable, we can approximate it.

- Simulation
  - Using Monte-Carlo methods to obtain the null distribution
  - Need to know the *true distribution*
  - Uncommon in practice but require no calculous
- Permutation
  - Shuffling data to break possible relationships
  - Work for specific null hypotheses
  - Limited applications but assumption-less
- Bootstrap

- Approximate sampling distribution via resampling
- Require large sample size
- Computationally expansive but universal
