STAT 480 Statistical Computing Applications

Unit 5. Resampling Methods

Lecture 1. Bootstrap

Department of Statistics, Iowa State University
Spring 2019

Impact of Revolution in Computing

- The continuing revolution in computing is having a dramatic influence on statistics.
- Exploratory analysis of data becomes easier as graphs and calculations are automated.
- Statistical study of very large and very complex data sets becomes feasible.
- Computationally intensive methods
  - Bootstrap
  - Randomization (Permutation) Tests
  - Cross Validation

Motivation

- It is often relatively easy to devise an estimator $\hat{\theta}$ of a parameter θ of interest, but it is difficult or impossible to determine the distribution or variance (sampling variability) of that estimator. Variance helps in assessing the accuracy of the estimators.
- One might fit a parametric model to the dataset, yet not be able to assign confidence intervals to see how accurately the parameters are determined.

Motivation: Example 1

- Let $X_1, \ldots, X_n$ be i.i.d from an unknown distribution F with mean μ and variance σ^2.
- We can estimate μ and σ^2 by

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

- Interest: Estimate the noise ratio

$$\tau = \frac{\mu}{\sigma}.$$

- A natural estimator for τ is $\hat{\tau} = \hat{\mu}/\hat{\sigma}$.
- What is the $Var(\hat{\tau})$?

Motivation: Example 2

- Let $X_1, \ldots, X_n$ be i.i.d from an unknown distribution F with mean μ and variance σ^2.
- Interest: Construct a confidence interval for μ.
- $$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$
- We can estimate μ and σ^2 by

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

- Obtain $100(1-\alpha)\%$ confidence interval

$$\left[\bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right]$$

Motivation: Example 3

- Consider two datasets $x = (x_1, \cdots, x_n)$, $y = (y_1, \cdots, y_n)$.
- You want to know whether there is any relationship between them.
- One approach is to calculate the correlation coefficient

$$\begin{aligned} \rho &= Corr(x, y) = \frac{Cov(x, y)}{\sqrt{Var(x)Var(y)}} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}. \end{aligned}$$

- Interest: Construct confidence interval for ρ.
- Confidence intervals for ρ are not easy to calculate because both ends of the interval have to lie between -1 and 1.

Motivation: Example 4

- Data: Mouse data
  - Survival times to 16 mice after a test surgery
  - 7 mice in treatment group (new medical treatment)
  - 9 mice in control group (no treatment)

Group	Survival time (in days)									Mean
Treatment	94	197	16	38	99	144	23			86.86
Control	52	104	146	10	51	30	40	27	46	56.22

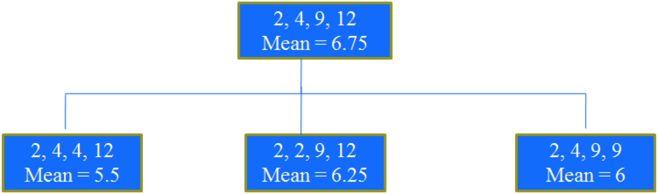
- Research Question: Did treatment prolong survival?
- Consider a two sample T test.
- Problem: samples show high fluctuation $\rightarrow$ need to assess accuracy of estimates.

Alternative Approach: Bootstrap

- Statistical inference is based on the sampling distributions of sample statistics.
- In absence of any other information about the distribution, the observed sample contains all the available information about the underlying distribution.
- Resampling the sample might be the best guide to what can be expected from resampling from the distribution.
- The method of drawing samples from the sample with replacement is called bootstrap.

The Resampling Idea

- Interest: estimate the mean of a population and sampling variability.
- Suppose we have the sample 2, 4, 9, 12.
- To study the average, we calculate all possible averages from numbers selected from 2, 4, 9, 12 with replacement.



The Resampling Idea

- We are interested in learning the average of a population, we have the sample 2, 4, 9, 12.
- To study the average, we calculate all possible averages from numbers selected from 2, 4, 9, 12 with replacement.
- If we enumerate all possible samples of size 4, taken with replacement from 2, 4, 9, 12, and for each calculate the mean, we again obtain an empirical distribution of the average.

Variance of the Sample Mean

- Interest: Estimate the variance of the sample mean $\bar{X}$.
- Let $\bar{x}_k^*$ be the sample for the kth resample x_k^*, $k = 1, \ldots, 256$.
- We can estimate $Var(\bar{X})$ by

$$Var(\bar{X}) \approx \frac{1}{256}\sum_{k=1}^{256}(\bar{x}_k^* - \bar{x}^*)^2, \quad \bar{x}^* = \frac{1}{256}\sum_{k=1}^{256}\bar{x}_k^*.$$

- In the original sample, $\bar{x} = 6.75$,
  $\hat{\sigma}^2 = \frac{1}{4}\sum_{i=1}^{4}(x_i - \bar{x})^2 = 15.6875$, so $\hat{\sigma}^2/4 = 3.92$.
- If we calculate all 256 samples, $\bar{x}^* = 6.75$, $Var(\bar{X}) \approx 3.94$.

Bootstrap

- There is a problem! In the example with $n = 4$, there were $n^n = 256$ different bootstrap resamples, so we could get them all.
- In more typical sample sizes, n^n grows so large as to be incomputable, so we just select B resamples.

Bootstrap Variance

- Let $\hat{\theta}$ be an estimator of θ based on $x = (x_1, \cdots, x_n)$.
- Calculate the variance by repeating the following steps $k = 1, \ldots, B$
  1. Create pseudo data x^* by sampling n onservations from $(x_1, \ldots, x_n)$ with replacement.
  2. Calculate $\hat{\theta}^*$ of the pseudo data x^*.
- Now you have $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$, the bootstrap variance is

$$Var(\hat{\theta}) \approx \frac{1}{B}\sum_{k=1}^{B}(\hat{\theta}_k^* - \bar{\theta}^*)^2,$$

where $\bar{\theta}^* = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_k^*$.

How Many Bootstrap Samples?

- Choice of B depends on
  - Computer availability
  - Type of the problem: standard errors, confidence intervals
  - Complexity of the problem
- My Recipe
  - Choose a large but tolerable number of replications. Obtain the bootstrap estimates.
  - Change the random-number seed. Obtain the bootstrap estimates again, using the same number of replications.
  - Do the results change meaningfully? If so, the first number you chose was too small. Try a larger number. If results are similar enough, you probably have a large enough number.

Note: To be sure, you should probably perform step 2 a few more times, but I seldom do.

Bootstrap for Correlation

- Law School data: average LSAT and GPA scores for the 1973entering classes of 15 American law schools.

LSAR (x)	576	635	558	578	666	580	555	661
GPA (y)	3.39	3.30	2.81	3.03	3.44	3.07	3.00	3.43
LSAR (x)	651	605	653	575	545	572	594	
GPA (y)	3.36	3.13	3.12	2.74	2.76	2.88	2.96	

- Confidence intervals for the correlation.
  - One common approach is to transform the correlation coefficient, i.e, find a function of the correlation that is approximately normally distributed.
  - Both ends of the interval have to lie between -1 and 1.
  - With bootstrap, we can avoid this.

Bootstrap for Correlation (Cont.)

- An alternative approach is to calculate confidence intervals for ρ using a resampling bootstrap.
- Sampling n elements from (x, y) with replacement, so you obtain a pseudo data set with n elements.
- The correlation of these pseudo data can then be calculated, which will be different to the correlation in the real data.
- Repeating this data re-sampling process a large number of times gives a large number of correlation coefficients.

Bootstrap for Correlation (Cont.)

- Estimate the correlation coefficient $\hat{\rho}$ for the real data.
- Calculate a confidence interval by repeating the following steps $k = 1, \ldots, B$.
  1. Create pseudo data (x^*, y^*) by sampling n pairs (x_i, y_i) from (x, y) with replacement.
  2. Calculate the correlation coefficient of the pseudo data (x^*, y^*).
- The 95% confidence intervals for ρ are simply the 2.5% and 97.5% percentiles from the set of estimated correlation coefficients from the pseudo data.