Statistical Computing

Michael Mayer

February 2023

Statistical Computing: What will we do?

Chapters

1. R in Action
2. Statistical Inference
3. Linear Models
4. Model Selection and Validation
5. Trees
6. Neural Nets

Remarks

- Chapters 3 to 6:
  Statistical ML in Action
- Two weeks per chapter
- Exercises at end of chapter notes

Statistical Inference

Outline

1. Statistical Inference
2. The Bootstrap
3. Permutation Tests

Aims

- Learn computer-intensive methods in statistical inference
- Apply programming techniques from last chapter

Statistical Inference

Use data to make statements about unknown population parameter θ:

- True proportion of patients that benefit from some novel treatment
- True average claim count per insured car year
- True correlation coefficient between the price of a diamond and its size

Main Tasks

1. Point estimation: Estimate θ by an estimator $\hat{\theta}(\text{data})$
2. Interval estimation: Provide confidence interval $I(\text{data})$ for θ
3. Testing: Use test statistic $T(\text{data})$ to measure statistical evidence against null hypothesis, e.g., $\theta = 0$. Reject if evidence is strong enough

Math stats solves tasks for different parameters θ, and under different circumstances

Classic Results for the Mean

- Distribution F with $\mu = \mathbb{E}(F)$ and $\sigma = \sqrt{\mathrm{Var}(F)} < \infty$
- $\boldsymbol{X} = (X_1, \ldots, X_n)$: Sequence of independent RVs, each with distribution F

Properties of sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ viewed as RV

1. $\hat{\mu}$ is an unbiased estimator of μ:

$$\mathbb{E}(\hat{\mu}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \mu$$

2. Law of large numbers: As $n \to \infty$, $\hat{\mu}$ converges in probability to μ
3. CLT: For $Z \sim N(0,1)$ and standard deviation $\sigma(\hat{\mu})$ of $\hat{\mu}$:

$$\frac{\hat{\mu} - \mu}{\sigma(\hat{\mu})} \xrightarrow[n\to\infty]{d} Z$$

Standard Deviation $\sigma(\hat{\mu})$ of the Sample Mean

$$\sigma(\hat{\mu}) = \sqrt{\mathrm{Var}(\hat{\mu})} = \sqrt{\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)} = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i)} = \sqrt{\frac{1}{n}\mathrm{Var}(X_i)} = \frac{\sigma}{\sqrt{n}}$$

Since $\sigma = \sigma(F) = \sigma(X_i)$ unknown, replace it by sample standard deviation
$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2} \rightarrow$ estimated standard deviation of mean: $\hat{\sigma}(\hat{\mu}) = \hat{\sigma}/\sqrt{n}$

Remarks and outlook

- Standard deviation $\sigma(\hat{\theta})$ of estimator called standard error
- Estimated standard error denoted by $\hat{\sigma}(\hat{\theta})$
- Accuracy of estimator $\rightarrow$ confidence intervals for θ
- Formulas are rare $\rightarrow$ Bootstrap

Computer Simulations

Illustration via repeated sampling from known(!) distribution

- Law of large numbers
- Central Limit Theorem

From the CLT to Confidence Intervals (CI)

Approximate $(1-\alpha)\cdot 100\%$-CI for μ:

$$[\hat{\mu} \pm z_{1-\alpha/2}\hat{\sigma}(\hat{\mu})],$$

where $\hat{\sigma}(\hat{\mu}) = \hat{\sigma}/\sqrt{n}$, and z_β is the β-quantile of $N(0,1)$ ($\rightarrow$ lecture notes)

Remarks

- "z-confidence interval" for the mean μ
- Usually more accurate: Student CI with $n-1$ degrees of freedom
- "Probability" or "Confidence"?

Examples

- z-CI for the mean
- Accuracy: Compare nominal coverage probability $1-\alpha$ with real coverage

Other Estimators

- Many estimators $\hat{\theta}$ are asymptotically normal
- We can use the same formula to calculate approximate $(1-\alpha)\cdot 100\%$ CI for θ:

$$I_{1-\alpha} = [\hat{\theta} \pm z_{1-\alpha/2}\hat{\sigma}(\hat{\theta})]$$

Limitation

Usually, no general formula for $\hat{\sigma}(\hat{\theta})$ available

Solution (Bradley Efron, 1979)

The Bootstrap offers a fully generic and automatic way of finding $\hat{\sigma}(\hat{\theta})$

The Bootstrap

- Observe sample: $\boldsymbol{x} = (x_1, \ldots, x_n)$
- Standard error $\hat{\sigma}(\hat{\theta})$ of estimator $\hat{\theta}(\boldsymbol{x})$?

Bootstrap estimate of standard error

1. From $\boldsymbol{x}$, draw with replacement a Bootstrap sample $\boldsymbol{x}^*$ of size n
2. Calculate Bootstrap replication $\hat{\theta}(\boldsymbol{x}^*)$ of $\hat{\theta}(\boldsymbol{x})$
3. Repeat B times to get B Bootstrap replications $\hat{\theta}(\boldsymbol{x}^{*1}), \ldots, \hat{\theta}(\boldsymbol{x}^{*B})$
4. Calculate sample standard deviation of the B replications

Examples

- Mean
- Median

- Why "Bootstrap"?
- Bootstrap sample is to original sample what original sample is to population

Bootstrap Confidence Intervals

Standard normal Bootstrap confidence interval

- Take any asymptotically normal $\hat{\theta}$
- Approximate $(1-\alpha)\cdot 100\%$-confidence interval for θ:

$$[\hat{\theta} \pm z_{1-\alpha/2}\hat{\sigma}(\hat{\theta})]$$

- $\hat{\sigma}(\hat{\theta})$: Bootstrap estimate of standard error
- Sample size not too small

Example

Alternative: Percentile Bootstrap Confidence Interval

- Consider B Bootstrap replications $\hat{\theta}(\boldsymbol{x}^{*1}), \ldots, \hat{\theta}(\boldsymbol{x}^{*B})$
- Use empirical $(\alpha/2)$ and $(1-\alpha/2)$ quantiles as confidence limits
- No asymptotic normality of $\hat{\theta}$ required
- Transformation respecting
- Range-preserving
- Since extreme quantiles involved, use large B, e.g., 9999

Examples

- Median
- Simulation study on accuracy
- Better confidence intervals

Multiple Samples/Groups

Examples

- Mean difference between two groups
- Median difference between two groups
- R-squared of a one-way ANOVA between multiple groups

Two ways of drawing Bootstrap sample

1. Resample within group to keep group sizes fixed (usually recommended)
2. Resample rows in data with one column representing the group (more generic)

Example

Multivariate Situations

Examples

- Pearson correlation
- Kendall's rank correlation
- R-squared of a linear regression
- Mean difference of two groups (how?)

$\rightarrow$ Study associations between variables

How to create Bootstrap sample?

Sample whole rows of dataset (with replacement)

Example

Permutation Tests / Resampling tests

First described by R.A. Fisher in 1935!

Hypotheses tests in general

- Want to show interesting alternative hypothesis H_1 about θ, e.g. $\theta \neq 0$
- Measure evidence against contrary H_0, e.g. $\theta = 0$ by test statistic T
- If evidence is strong enough, reject H_0 in favor of H_1

p value normalised test statistic

Probability of observing at least as much evidence against H_0 as in the specific sample when H_0 holds $\rightarrow$ reject H_0 if p value ≤ 0.05 (or some other prespecified value)

$$\text{p value} = P_{H_0}\{T \geq t\}$$

observed is larger than in test statistic

Example

Tests for Association

Example showed: two-sample t-test is test of association between:

- $\boldsymbol{Y}$: Numeric variable representing the stacked values of both groups
- $\boldsymbol{X}$: Binary variable representing the group ("A" or "B")

Association is measured in terms of location shift in the grouped means

Many additional tests are tests of association between two variables. Examples?

Can be tackled by

- computer-intensive,
- fully automatic

technique called permutation test

Permutation Tests

T: Measures strength of association between $\boldsymbol{X}$ and $\boldsymbol{Y}$ with observations $\boldsymbol{x}$ and $\boldsymbol{y}$

How to find distribution of T under null?

1. $\boldsymbol{x}^*$: permutation of $\boldsymbol{x}$
   $\rightarrow$ destroys dependency between $\boldsymbol{x}$ and $\boldsymbol{y}$

   no need to shuffle both

2. Calculate $t^* = T(\boldsymbol{x}^*, \boldsymbol{y})$
3. Repeat above steps B times to get permutation replications $t^*(1), \ldots, t^*(B)$
   $\rightarrow$ empirical null hypothesis distribution of T
4. Permutation p value: $\frac{1}{B}\sum_{i=1}^{B} 1\{t^*(i) \geq T(\boldsymbol{x}, \boldsymbol{y})\}$

Remarks and Examples

Remarks

- B large, e.g. 10000
- Why not all permutations? $\rightarrow$ approximate or Monte-Carlo permutation tests
- "coin" package in R
- Permutation replications versus Bootstrap replications? resample with replacement
  change of data distribution
- Not completely assumption-free (iid. under H_0 is sufficient)

shuffling
under Null hypothesis

Examples

- Two-sample t-test
- Simulation: Type 1 error and power?

More Examples

- Wilcoxon's rank sum test
- Test for Pearson correlation
- Paired t-test

replace in the two variable + test the numerical var with their rank (permutation test without computers)

in coin package

person id etc

Y	x	block
5	A	1
6	B	1
7	A	2
⋮		⋮

just shuffle in blocks

Many more in the "coin" package

very small sample size (3-5)
permutation test are best for all other cases
there are no disadvantages