Lecture 10 - Monte Carlo methods in inference

Björn Andersson (w/ Xijia Liu)

Department of Statistics, Uppsala University

March 5, 2014

Table of Contents

Monte Carlo experiment

- In statistics, we can investigate the properties of estimators, tests, model selection methods etc by simulation
- A Monte Carlo experiment simulates data from a probability model repeatedly
- The data retrieved from the simulation procedure is then used to evaluate the estimator, test or method of interest, in terms of MSE, confidence level, type I error, asymptotic properties and so on
- A Monte Carlo simulation can investigate statistical properties which are not possible or feasible to investigate analytically
  - In many applications, statistical properties are very difficult or impossible to derive or are derived under ideal conditions which are not realistic

Simulation

Suppose we want to evaluate an estimator T under a certain parametric model for the data but we are unable to derive the properties of the estimator analytically. Then there is a need for simulating data under the parametric model.
Example: we want to investigate the properties of an estimator T which is a function of a series of Bernoulli trials where the success probability is

$$P(X = 1) = c + \frac{1 - c}{1 + \exp[-a(1 - b)]}.$$

We then simulate data from our parametric model and evaluate the properties of T under this model.

Simulation

Other possible simulations:

- Investigate the properties of the estimator under different assumptions about the underlying distribution of a nuisance parameter (robustness study)
- Choose parameter values that are suspected to make the model more difficult to estimate and simulate under such a model

How many replications?

To get absolute certainty we need to repeat the simulation in infinity. How many replications are needed in practice? We can investigate the variance of the simulation estimators. Assume a sample from the $N(\mu, \sigma^2)$ distribution.
The bias estimator $B_R = \sum_{i=1}^{R} \bar{Y}_i^* - \bar{Y}$:

$$\mathrm{Var}(B_R) = \frac{\sigma^2}{n} \times \frac{n-1}{nR}.$$

The variance estimator $V_R = \sum_{i=1}^{R} (\bar{Y}_i^* - \bar{Y}^*)^2/(R-1)$ $(\bar{Y}^* = \sum_{i=1}^{R} \bar{Y}^*/R)$:

$$\mathrm{Var}(V_R) \approx \frac{2\sigma^4}{n^3} + \frac{2}{R}\left(\frac{2\sigma^4}{n^3} + \frac{\sigma^4}{n^2}\right).$$

The first term is due to the variability in the data and the second term is due to the simulation variability. To have simulation variability account for 10% of total variability select $R = 10n$.

Table of Contents

Notation

Let $x_1, \ldots, x_n$ be a random sample from the distribution of a random variable X. An estimator for a parameter θ is an n-variate function of the random sample:

$$\hat{\theta} = f(x_1, \ldots, x_n).$$

The standard error of the estimator $\hat{\theta}$ is

$$\mathrm{se}(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})}.$$

Monte Carlo estimation and standard error

- With the Monte Carlo method, samples $x_1^{(j)}, \ldots x_n^{(j)}$ are repeatedly generated from the distribution of X. Let the total number of samples be m.
- For the estimator $\hat{\theta}$ we then retrieve m estimates $\hat{\theta}^{(j)} = f(x_1^{(j)}, \ldots x_n^{(j)})$, $j \in \{1, \ldots, m\}$ from the Monte Carlo samples.
- The sample variance from the Monte Carlo estimates $\{\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(m)}\}$ can then be used to estimate the standard error of the estimator $\hat{\theta}$:

$$\hat{\mathrm{se}}(\hat{\theta}) = \sqrt{\frac{1}{m-1} \sum_{j=1}^{m} \left(\hat{\theta}^{(j)} - \frac{\sum_{j=1}^{m} \hat{\theta}^{(j)}}{m} \right)^2}.$$

Monte Carlo estimation and standard error

In item response theory (IRT), the responses to a number of multiple-choice questions can be modelled. The probability of answering item k correctly is modelled by

$$P(\theta; a, b) = \frac{1}{1 + \exp[-a(\theta - b)]}.$$

Under an assumption of the distribution of the abilities θ in the population, the parameters can be estimated using a variant of the EM algorithm. The asymptotic standard errors with this estimator can be estimated with the Delta method. We want to investigate the small sample performance of the standard error estimates through a Monte Carlo simulation.

Monte Carlo estimation and standard error

We proceed as follows:

1. Data is simulated in accordance with the IRT model
2. The model is estimated using the EM algorithm
3. The estimated parameters and their associated estimated standard errors are saved

The average of the analytical standard errors from the delta method are calculated and compared to the Monte Carlo estimates of the standard error. Of course we can investigate the bias and so forth, as well.

Mean Squared Error (MSE)

- The mean squared error (MSE) of an estimator $\hat{\theta}$ of θ is

$$\mathrm{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2.$$

- We have the following identity

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + [E(\hat{\theta} - \theta)]^2.$$

  Note that $E(\hat{\theta} - \theta)$ is the bias of $\hat{\theta}$ with respect to θ.
- We can estimate the MSE using the Monte Carlo samples:

$$\widehat{\mathrm{MSE}}(\hat{\theta}) = \frac{1}{m} \sum_{j=1}^{m} \left(\hat{\theta}^{(j)} - \theta_{MC} \right)^2,$$

  where θ_{MC} is the value of the parameter used in the simulation.

Estimation of the MSE of the trimmed mean

- Let $x_1, \ldots, x_n$ be a random sample from the distribution of the R.V. X. We wish to estimate μ, the mean of X.
- One estimator of μ is the k-th level trimmed sample mean is defined as

$$\bar{X}_{[-k]} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)},$$

  where $x_{(1)}, \ldots, x_{(n)}$ is the ordered sample. Hence $\bar{X}_{[-k]}$ is the sample mean of the sample where the k smallest and k largest observations have been removed.
- The small sample variance of $\bar{X}_{[-k]}$ is not easily derived.

Estimation of the MSE of the trimmed mean

- We wish to estimate the MSE of $\bar{X}_{[-k]}$ when the distribution of X is a mixture of independent R.V.s $Y \sim N(0,1)$ and $Z \sim N(0,100)$ such that $X = Y$ with probability p and $X = Z$ with probability $1-p$, where $p \in [0,1]$.
- Note that the regular sample mean $\bar{X}$ is unbiased but has a possibly large variance.
- The MSE of $\bar{X}_{[-k]}$ is then estimated by

$$\widehat{MSE}(\bar{X}_{[-k]}) = \frac{1}{m}\sum_{j=1}^{m}\left(\bar{X}_{[-k]}^{(j)}\right)^2,$$

since $E(X) = 0$.

Confidence level

- If (L, U) is the $100(1-\alpha)\%$ confidence interval estimator for an unknown parameter θ, then the lower and upper limits L and U are statistics with distributions that depend on the distribution of the random variable X which was sampled.
- The confidence level is the probability that the interval (L, U) covers the true value of the parameter θ, i.e.

$$P[\theta \in (L, U)].$$

- Obviously, a $100(1-\alpha)\%$ empirical confidence level is desired.

Confidence level estimation

The empirical confidence level of a confidence interval estimator can be estimated with Monte Carlo methods:

1. For each replicate $j \in \{1, \ldots, m\}$:
   1. Generate the j-th random sample $x_1^{(j)}, \ldots, x_n^{(n)}$
   2. Compute the $100(1-\alpha)\%$ confidence interval (L_j, U_j) for the j-th sample
   3. Check if θ is in (L_j, U_j), i.e. compute $y_j = \mathbf{1}(\theta \in (L_j, U_j))$
2. Compute the empirical confidence level

$$\bar{y} = \frac{\sum_{j=1}^{m} y_j}{m}.$$

Confidence level example

- Let $x_1, \ldots, x_n$ be a random sample from $N(\mu, \sigma^2)$, where $n \geq 2$. Let S^2 be the sample variance. Then

$$V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(df = n-1).$$

- A one-sided $100(1-\alpha)\%$ confidence interval is given by $(0, (n-1)S^2/\chi^2_{1-\alpha}(df = n-1))$.
- The empirical confidence level with this interval can be estimated using the Monte Carlo method, when, for example, the data is not normal distributed.

Confidence level example

Let $y_1, \ldots, y_n$ be a sample from the $t(3)$ distribution. We want to estimate the variance of Y. The true value is $\mathrm{Var}(Y) = \frac{v}{v-2} = 3$. What is the empirical confidence level of $(0, (n-1)S^2/\chi^2_{1-\alpha}(df = n-1))$? Choose $\alpha = 0.05$.

```
R> level <- numeric(1)
R> for(i in 1:1000){
+     tty <- rt(100, df = 3)
+     ci.t <- c(0, 99 * var(tty) / qchisq(0.95, df = 99))
+     if(ci.t[2] < 3) level <- level + 1
+ }
R> level / 1000
```

```
[1] 0.854
```

Confidence level example

Let $z_1, \ldots, z_n$ be a sample from the $N(0,1)$ distribution. We want to verify the validity of the confidence interval $(0, (n-1)S^2/\chi^2_{1-\alpha}(df = n-1))$. Let $\alpha = 0.05$.

```
R> level <- numeric(100)
R> for(j in 1:100){
+     for(i in 1:1000){
+         tty <- rnorm(100)
+         ci.t <- c(0, 99 * var(tty) / qchisq(0.95, df = 99))
+         if(ci.t[2] < 1) level[j] <- level[j] + 1
+     }
+  }
R> mean(level / 1000)

[1] 0.94995

R> var(level / 1000)

[1] 5.130051e-05
```

Estimation of the false positive rate in graphical models

A graphical model is a graphical representation of the relationships between a set of random variables. If each variable is connected by an edge in the graph, then the variables are dependent conditional on all the other variables in the graph. The edges can be inferred from the data using different algorithms which utilize partial correlations to derive the graph structure. The properties of different algorithms can be estimated using Monte Carlo simulation. For example, we could be interested in the false positive rate:

$$\text{FPR} = \frac{\#\text{ of incorrect edges in the estimated graph}}{\text{total }\#\text{ of missing edges in the true graph}}.$$

Estimation of the false positive rate in graphical models

We generate a set of variables with a given graph structure and investigate the properties of the algorithm to recover the correct underlying graph through simulation.

1. Generate data with a certain underlying graph structure
2. Estimate the graph from the algorithm
3. Calculate the false positive rate for the iteration by comparing the true graph to the estimated graph

Take the mean of the false positive rate for each iteration as the estimated false positive rate of the algorithm under the particular graph structure.

Table of Contents

Hypothesis testing

- We want to test a hypothesis regarding a parameter $\theta \in \Theta$, where Θ denotes the parameter space. The hypothesis is of the form:

$$\mathrm{H}_0 : \theta \in \Theta_0$$
$$\mathrm{H}_\mathrm{A} : \theta \in \Theta_A,$$

  where Θ_0 and Θ_A partition Θ.
- A type I error occurs when H_0 is rejected when $\theta \in \Theta_0$.
- A type II error occurs when H_0 is not rejected when $\theta \notin \Theta_0$.

Parametric Monte Carlo tests

- Monte Carlo methods are not useful just for evaluating estimators using simulation, it is also possible to conduct tests of a statistic generated from a given data set.
- In a parametric Monte Carlo test, data is simulated according to the statistical model under the stated null hypothesis and then the values of the statistic from each generated data set is calculated and compared to the statistic calculated from the observed data.
- Hence, using simulation we test if the statistic from the observed data is consistent with the statistics calculated from data generated under the null hypothesis.

Parametric Monte Carlo tests

Consider a one-sided hypothesis test. Let the test statistic be denoted T. Let the observed statistic be t. Let the simulated values of the statistic be $t_1^*, \ldots, t_R^*$. Under the null hypothesis, all values $t, t_1^*, \ldots, t_R^*$ are equally likely. Hence we have, for a continuous T,

$$P(T < T_{(r)}^* | H_0) = \frac{r}{R+1},$$

where $T_{(r)}^*$ denoted the r-th ordered value of T. If exactly k of the simulated values $\{t_i^*\}$ are larger then the observed statistic t, then

$$P(T \geq t | H_0) = p_{\mathrm{mc}} = \frac{k+1}{R+1},$$

which is called the Monte Carlo p-value.

Parametric Monte Carlo test example

Consider a regression model

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \mathbf{e},$$

where $\mathbf{x}_1$ is the known explanatory variable and the entries in $\mathbf{e}$ are iid $N(0, 1)$. We estimate β_1 with OLS. Let's say we want to test

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 > 0$$

The statistic of interest is then $\hat{\beta}_1$. To obtain the test, we draw a random permutation $\mathbf{y}^i$ from the observed values $\mathbf{y}$ and calculate $\hat{\beta}_1^{(i)}$ for each replication i. The p-value of the test is then

$$p_{\text{mc}} = \frac{\sum_{i=1}^{R} \mathbf{1}(\hat{\beta}_1^{(i)} > \hat{\beta}_1) + 1}{R + 1},$$

and we reject H_0 at $\alpha = 0.05$ if $p_{\text{mc}}^{\text{obs}} \leq 0.05$.

Empirical type I rate with Monte Carlo

We can also evaluate the properties of a particular test using simulation. The empirical type I rate of a test can be estimated using Monte Carlo simulation:

1. For each replication $j \in \{1, \ldots, m\}$
   1. Generate the j-th random sample $x_1^{(j)}, \ldots, x_n^{(n)}$ from the null distribution, i.e. for $\theta \in \Theta_0$.
   2. Compute the test statistic T_j from the j-th sample
   3. Let $y_j = \mathbf{1}(\mathrm{H}_0 \text{ is rejected at significance level } \alpha)$
2. Calculate the sum of the y_j, $\bar{y} = \sum_{j=1}^{m} y_j / m$ as the estimate of the type I rate.

Empirical type I rate example

Let $y_1, \ldots, y_n$ be a sample from density $\sim N(10, 10)$. Consider the hypothesis test defined by

$$H_0 : \mu = 10$$
$$H_A : \mu \neq 10$$

The estimator is $\hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{n}$ and the test statistic is $t = \frac{\hat{\mu} - 10}{\sqrt{\hat{\text{Var}}(\hat{\mu})/n}}$. Under H_0, $t \sim t(df = n - 1)$. At $\alpha = 0.05$ we reject H_0 if $t_{\text{obs}} < t_{\alpha/2}(df = n - 1)$ or $t_{\text{obs}} > t_{1-\alpha/2}(df = n - 1)$

Empirical type I rate example

Test when data is $N(10, 10)$.

```
R> reject <- numeric(1)
R> for(i in 1:5000){
+     yN <- rnorm(20, 10, sqrt(10))
+     Sy <- sqrt(var(yN) / 20)
+     muY <- sum(yN) / length(yN)
+     tobs <- (muY - 10) / Sy
+     if(abs(tobs) >  qt(0.975, df = 19))
+         reject <- 1 + reject
+ }
R> reject / 5000
```

```
[1] 0.0528
```

Empirical type I rate example

Test when data is $t(df = 3)$.

```
R> reject <- numeric(1)
R> for(i in 1:5000){
+     yN <- rt(20, 3)
+     Sy <- sqrt(var(yN) / 20)
+     muY <- sum(yN) / length(yN)
+     tobs <- (muY) / Sy
+     if(abs(tobs) >  qt(0.975, df = 19))
+         reject <- 1 + reject
+ }
R> reject / 5000
```

```
[1] 0.0436
```

Power of a test

- The power of a hypothesis test, with test statistic T and rejection region R, is defined as

  $$1-\text{Type II error rate} = 1-P(T \notin R|\theta \notin \Theta_0) = P(T \in R|\theta \notin \Theta_0),$$

  i.e. the probability of rejecting a false null hypothesis.
- The empirical power of a test against a fixed alternative can be estimated using a Monte Carlo experiment.

Power of a test

The empirical power of a test can be estimated with Monte Carlo methods:

1. For each replication $j \in \{1, \ldots, m\}$
   1. Generate the j-th random sample $x_1^{(j)}, \ldots, x_n^{(n)}$ from the alternative distribution, i.e. for $\theta \in \Theta_A$.
   2. Compute the test statistic T_j from the j-th sample
   3. Let $y_j = \mathbf{1}(\mathrm{H}_0 \text{ is rejected at significance level } \alpha)$
2. Calculate $\bar{y} = \sum_{j=1}^{m} y_j / m$ as the estimate of the power of the test.

Power of a test: example

For the hypothesis test of

$$H_0 : \mu = 10$$
$$H_A : \mu = 12$$

```
R> power <- numeric(1)
R> for(i in 1:10000){
+     yN <- rnorm(20, 12, 5)
+     Sy <- sqrt(var(yN) / 20)
+     muY <- sum(yN) / length(yN)
+     tobs <- (muY-10) / Sy
+     if(abs(tobs) >  qt(0.975, df=19)) power <- 1 + power
+  }
R> power / 10000
[1] 0.3959
```