Statistical Analysis of Simulated Data

Chapter 7. Statistical Analysis of Simulated Data

Estimate of the Population Mean

$\theta = EX = ?$

Recall: 1. If $X_1, \ldots, X_n$ are i.i.d. with mean θ and variance σ^2.

Then $E\bar{X} = \theta$ and $E(\bar{X} - \theta)^2 = Var\bar{X} = \sigma^2/n$.

2. *Chebyshev's inequality*:

$$P(|\bar{X} - \theta| > k\sigma/\sqrt{n}) \leq 1/k^2$$

guarantees that the sample mean is unlikely to be too far away from the population mean.

Hence, take $X_1, \ldots, X_n$, use $\bar{X} = \sum_{i=1}^{n} X_i/n \to \theta$ *a.s.* as $n \to \infty$.

Q: How large is n?

Answer 1: Specify a tolerated error bound (ϵ) with a desired probability ($1-\alpha$). Then apply

(i) *Chebyshev's Inequality*:

$$P(|\bar{X}-\theta| \le k\sigma/\sqrt{n}) \ge 1 - 1/k^2. \quad \therefore \mathbf{n} \ge \frac{\sigma^2}{\alpha\epsilon^2}.$$

(ii) *Central Limit Theorem*: As n **large**,

$$P(|\bar{X}-\theta| \le z_{\alpha/2}\sigma/\sqrt{n}) \approx 1-\alpha. \quad \therefore \mathbf{n} \ge \frac{\sigma^2 \mathbf{z}_{\alpha/2}}{\epsilon^2}.$$

Problem: But σ is usually **unknown**. $\Longrightarrow$ Give a prior *guess* or give an *estimate*.

Recall: The sample variance $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ is unbiased for σ^2. $n = ?$

Solution: Start with a '**pilot**' sample of size $n_0 \geq 30$, compute its S_0^2. Then replace σ^2 by S_0^2 in (i) or (ii) to get n. Select additional $n - n_0$ random variates and take

$$\bar{X} = \sum_{i=1}^{n} X_i/n.$$

Answer 2: **Sequentially** generate X_i until the 'estimated' s.d. of $\bar{X}$ is less than the s.e. that one can tolerate. i.e. $\hat{\sigma}_n/\sqrt{n} < d$ (a pre-specified bound).

Algorithm:
1. Specify d and generate $X_1, \ldots, X_n$, $n \geq 30$.
2. Compute $\bar{X}_n$ and $S_n^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$.
3. If $S/\sqrt{n} > d$, generate another X_{n+1} and replace n by $n+1$, return to 2; otherwise, set $\hat{\theta} = \bar{X}$.

Note: $\bar{X}_n$ and S_n^2 can be computed **recursively**.

Facts: 1. $\bar{X}_{n+1} = \bar{X}_n + \frac{1}{n+1}(X_{n+1} - \bar{X}_n)$.

2. $$S_{n+1}^2 = (1 - \tfrac{1}{n})S_n^2 + \tfrac{1}{n+1}(X_{n+1} - \bar{X}_n)^2$$
$$= (1 - \tfrac{1}{n})S_n^2 + (n+1)(\bar{X}_{n+1} - \bar{X}_n)^2.$$

Proof: Exercise.

Interval Estimate of θ.

1. If $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\theta, \sigma^2)$,

$P(\bar{\mathbf{X}} - \mathbf{t}_{\alpha/2;n-1}\mathbf{S}/\sqrt{\mathbf{n}} < \theta < \bar{\mathbf{X}} + \mathbf{t}_{\alpha/2;n-1}\mathbf{S}/\sqrt{\mathbf{n}}) = 1 - \alpha$.

2. If $X_1, \ldots, X_n \overset{i.i.d.}{\sim} EX = \theta$, $Var(X) = \sigma^2$,

$P(\bar{\mathbf{X}} - \mathbf{z}_{\alpha/2}\mathbf{S}/\sqrt{\mathbf{n}} < \theta < \bar{\mathbf{X}} + \mathbf{z}_{\alpha/2}\mathbf{S}/\sqrt{\mathbf{n}}) \approx 1 - \alpha$, for large n.

To get an interval estimate of θ based on simulation, one may generate random variates **sequentially** until the interval length, $2z_{\alpha/2}S_n/\sqrt{n} \leq l$, a specified bound, then the resulting $\bar{X}_n \pm z_{\alpha/2}S_n/\sqrt{n}$ is the estimate.

The Bootstrapping Technique

Ex: 1. $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\theta, 2.19)$. $\theta = EX$. $\hat{\theta} = \bar{X}$ and

$Var(\hat{\theta}) = 2.19/n$; Interval estimate: $\bar{X} \pm 1.96\sqrt{2.19}/\sqrt{n}$. $\square$

Ex: 2. $X_1, \ldots, X_n \overset{i.i.d.}{\sim} C(\theta, 1)$. θ is the **median** of X_i, i.e. θ is

such that $E[\mathbf{1}_{(-\infty,\theta)}(X)] = 1/2$.

Consider $\hat{\theta} =$ **sample median** $= X_{(n+1)/2}$, n odd.

Q: *Sampling distribution* of $\hat{\theta}$? *Interval estimate* of θ?

Mean square error of $\hat{\theta}$?

Idea: In general, the parameter of interest, say θ, is a function of the distribution, i.e. $\theta = \theta(F)$.

Now F is unknown $\Longrightarrow$ Take a **sample** $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$, **estimate** F by $\hat{F}_e$, the **empirical distribution** of $X_1, \ldots, X_n$, and estimate θ by $\theta(\hat{F}_e)$.

Recall: 1. $\mathbf{\hat{F}_e(x)} = \frac{\# \text{ of } X_i's \leq x}{n}, \quad i = 1, \ldots, n.$

(or $P(X^\star = X_i) = 1/n, \quad i = 1, \ldots, n$; or $X^\star \sim \hat{F}_e$.)

2. As $n \to \infty$, $\hat{F}_e(x) \to F(x), \forall$ continuity point x.

$\therefore \hat{\theta} = \theta(\hat{F}_e) \to \theta(F) = \theta$ $a.s.$ as $n \to \infty$ under general conditions.

Ex: 1. $\theta = \theta(F) = EX = \int x dF(x) = ?$ Let $X_1, \ldots, X_n \sim F$.

If $X^\star \sim \hat{F}_e, \quad P(X^\star = X_i) = 1/n, \;\; i = 1, \ldots, n.$

$$\therefore \hat{\theta} = \theta(\hat{\mathbf{F}}_{\mathbf{e}}) = E\mathbf{X}^\star = \sum_{i=1}^{n} X_i P(X^\star = X_i) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}.$$

$\square$

Ex

: 2. $\theta =$ median of $F = \theta(F)$=?

Note that $\int_{-\infty}^{\theta} dF(x) \geq 1/2$ and $\int_{\theta}^{\infty} dF(x) \geq 1/2$.

Thus, $\hat{\theta}$ is such that

$$\int_{-\infty}^{\hat{\theta}} d\hat{F}_e(x) \geq 1/2 \quad \text{and} \quad \int_{\hat{\theta}}^{\infty} d\hat{F}_e(x) \geq 1/2.$$

That is, given $X_1, \ldots, X_n$,

$$\sum_{x \le \hat{\theta}} P(X^\star = x) \ge 1/2 \iff \sum_{i:X_i \le \hat{\theta}} P(X^\star = X_i) \ge 1/2$$

$$\iff \sum_{\mathbf{i:X_i} \le \hat{\theta}} \frac{\mathbf{1}}{\mathbf{n}} \ge 1/2.$$

i.e. $\sum_{i:X_i \le \hat{\theta}} 1 \ \ge n/2$ and $\sum_{i:X_i \ge \hat{\theta}} 1 \ \ge n/2$.

Hence, $X_{([n/2])} \le \hat{\theta} \le X_{([n/2]+1)}$, $\hat{\theta} =$ **sample median**. $\square$

Q: Now $\hat{\theta} = \theta(\hat{F}_e) = g(X_1, \ldots, X_n)$, $X_i \sim F$, how 'good' is $\hat{\theta}$?

e.g. $E(\hat{\theta})=$? $Var(\hat{\theta})=$? or $MSE(\hat{\theta})=$?

If F is unknown, then the 'behavior' of $\hat{\theta}$ is unknown. For example

$$
\begin{aligned}
MSE(F) &= E^F(\hat{\theta} - \theta)^2 = E^F(\theta(\hat{F}_e) - \theta(F))^2 \\
&= E^F(g(X_1, \ldots, X_n) - \theta(F))^2, \quad X_i \sim F \\
&= \int (g(x_1, \ldots, x_n) - \theta(F))^2 dF(x_1, \ldots, x_n),
\end{aligned}
$$

function of F, **unknown**! Again, estimate it by $MSE(\mathbf{\hat{F}_e})$.

For given $X_1, \ldots, X_n \sim F$,

$$\begin{aligned} MSE(\hat{F}_e) &= E^{\hat{F}_e}(g(X_1^\star, \ldots, X_n^\star) - \theta(\hat{F}_e))^2 \\ &= \int (g(x_1^\star, \ldots, x_n^\star) - \hat{\theta})^2 d\hat{F}_e(x_1^\star, \ldots, x_n^\star) \\ &= \sum_{x_1^\star} \cdots \sum_{x_n^\star} (g(x_1^\star, \ldots, x_n^\star) - g(X_1, \ldots, X_n))^2 \\ &\qquad \cdot P(X_1^\star = x_1^\star, \ldots, X_n^\star = x_n^\star). \end{aligned}$$

Here, $x_i^\star \in \{X_1, \ldots, X_n\}$,

$P(X_1^\star = x_1^\star, \ldots, X_n^\star = x_n^\star) = \prod_{i=1}^n P(X_i^\star = x_i^\star) = \prod_{i=1}^n \frac{1}{n} = \left(\frac{\mathbf{1}}{\mathbf{n}}\right)^{\mathbf{n}}$.

Note: The summation is over *all* possible combinations of

$(x_1^\star, \ldots, x_n^\star)$ and $x_i^\star \in \{X_1, \ldots, X_n\}$. *$n^n$*! *Big*!

Monte-Carlo approximation.

Given $X_1, \dots, X_n \sim F$,

$$\begin{aligned} \widehat{MSE}_B &= \widehat{MSE}(\hat{F}_e) \\ &= \frac{1}{N}\sum_{l=1}^{N}(g(X^\star_{1l}, \dots, X^\star_{nl}) - (g(X_1, \dots, X_n))^2 \\ &\to MSE(\hat{F}_e), N \text{ large }, \end{aligned}$$

where $X^\star_{1l}, \dots, X^\star_{nl} \sim X^\star_i \sim \hat{F}_e$,the **uniform distribution** on $\{X_1, \dots, X_n\}$.

Ex: $\theta = \theta(\mathbf{F}) = EX$, $\hat{\theta} = \theta(\hat{\mathbf{F}}_{\mathbf{e}}) = \bar{X} = g(X_1, \ldots, X_n)$.

$$MSE(\mathbf{F}) = E(\hat{\theta} - \theta)^2 = ?.$$

Note that $\hat{\theta}^\star = g(X_1^\star, \ldots, X_n^\star) = \bar{X}^\star$,

$X_i^\star \sim \hat{F}_e \sim$ uniform on $\{X_1, \ldots, X_n\}$.

$$\therefore \quad \widehat{MSE}(\hat{\mathbf{F}}_{\mathbf{e}}) = \frac{1}{N} \sum_{l=1}^{N} [\bar{\mathbf{X}}_{\mathbf{l}}^\star - \bar{\mathbf{X}}]^2,$$

where $\bar{X}_l^\star = \sum_{i=1}^n X_{il}^\star / n$, $l = 1, \ldots, N$; $X_{il}^\star \sim X_i^\star \sim \hat{\mathbf{F}}_{\mathbf{e}}$.

Algorithm:

STEP 1: Compute $\bar{X} = \sum_{i=1}^{n} X_i/n$ for **given** data $\{X_1, \ldots, X_n\}$. $l = 1$.

STEP 2: Generate $\mathbf{X}_l^\star = (X_1^\star, \ldots, X_n^\star)$, where $X_j^\star \sim$ **uniformly** on $\{X_1, \ldots, X_n\}$.

STEP 3: Compute $\bar{X}_l^\star = \sum_{j=1}^{n} X_j^\star/n$.

STEP 4: Repeat STEP 2 to STEP 3, for $l = 1, \ldots, N$ times.

STEP 5: $\widehat{MSE}_B = \frac{1}{N} \sum_{l=1}^{N} [\bar{X}_l^\star - \bar{X}]^2$. $\square$

Note: 1. $\widehat{MSE}_B \rightarrow MSE(\hat{F}_e) \rightarrow$ **MSE(F)**.

2. $\hat{\theta}_l^\star$, $l = 1, \ldots, N$ can be used to approximate the distribution of $\hat{\theta}$.

$$\hat{F}_{\hat{\theta}}(x) = \{\# \text{ of } \hat{\theta}_l^\star \leq x\}/N, \quad -- \text{ **bootstrap distribution** of } \hat{\theta},$$

$$\hat{F}_{\hat{\theta}}(x) = \{\# \text{ of } \hat{\theta}_l^\star - \hat{\theta} \leq x\}/N, \quad -- \text{ **bootstrap distribution** of } \hat{\theta} - \theta.$$

3. One may use $\hat{\theta}_l^\star$, $l = 1, \ldots, N$ to construct an interval estimate of θ.

Ex: $\theta =$ Median of X, i.e. $P(X \leq \theta) \geq 1/2$ and $P(X \geq \theta) \geq 1/2$.

$\therefore \hat{\theta} = \theta(\hat{F}_e) =$ the middle of $\{X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}\} = X_{(\frac{n+1}{2})}$.

Thus, $MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = MSE(F)$ and

$MSE(\hat{F}_e) = E^{\hat{F}_e}(X^{\star}_{(\frac{n+1}{2})} - X_{(\frac{n+1}{2})})^2$ based on $X_1, \ldots, X_n$.

Therefore,

$$\widehat{MSE}_B = \frac{1}{N} \sum_{l=1}^{N} \left(X^{\star l}_{(\frac{n+1}{2})} - X_{(\frac{n+1}{2})} \right)^2,$$

where $X^{\star l}_{(\frac{n+1}{2})}$ is the median based on $\{X_1^{\star l}, \ldots, X_n^{\star l}\}$ and each $X_j^{\star l} \overset{i.i.d.}{\sim}$ uniform on $\{X_1, \ldots, X_n\}$. $\square$

Recall: Generation of $X^\star \sim$ uniform on $\{X_1, \ldots, X_n\}$:

STEP 1: Generate $U \sim U(0,1)$.

STEP 2: Set $I = \text{Int}(nU) + 1$.

STEP 3: Set $X^\star = X_I$.

Parametric Bootstrap

Use $\hat{F}(x) = F(x|\hat{\theta})$ where $\hat{\theta}$ is an estimate of θ.

e.g. $MSE(\hat{\theta})$=?

Algorithm:

1. Compute $\hat{\theta} = g(X_1, \ldots, X_n)$ based on $X_1, \ldots, X_n$.
2. Generate $X_1^\star, \ldots, X_n^\star$ from $F(\cdot|\hat{\theta})$.
3. Compute $\hat{\theta}^\star = g(X_1^\star, \ldots, X_n^\star)$.
4. Repeat 2-3 N times to get $\hat{\theta}_i^\star$, $i = 1, \ldots, N$.
5. $\widehat{MSE}_B(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^N \left(\hat{\theta}_i^\star - \hat{\theta}\right)^2$.