Purdue-NCKU program

Lecture 2
Point and Interval Estimations

Dr. Qifan Song

Statistical Modeling

- Statistical inferences aim to learn the underlying distribution of data
- Make some mathematical assumptions on the distribution of the observations
- For random observations based on different subjects, usually we assume

  $$X_1, \ldots, X_n \sim f \text{ independently, and } f \in \mathcal{F},$$

  where $\mathcal{F}$ is a set of *candidate* distributions, any distribution outside $\mathcal{F}$ will not be considered
- Search a distribution in $\mathcal{F}$ as the plausible truth.
- Tradeoff between flexibility and solvability

Parametric vs Nonparametric Modelings

- Parametric Model: the cardinality of $\mathcal{F}$ is not bigger or equivalent to R^d for some fixed dimension d.
- Nonparametric Model: the cardinality of $\mathcal{F}$ is equivalent to R^d where dimension $d = \infty$ or $d \to \infty$.

- This course mostly focuses on parametric modeling
- d Dimensional Parametric Model: $\mathcal{F} = \{f_\theta; \theta \in \Theta \subset R^d\}$
- θ is called parameter, it may or may not have an actual meaning
- Identifiablity and Re-parametrization
- Finding the plausible truth distribution in $\mathcal{F}$ boils down to finding the plausible truth θ.

Point Estimation of the Parameter

Goal: We want to estimate the value of θ (or more generally, $\phi(\theta)$ for some known function ϕ).

Definition: A point estimation is any function of data $\widehat{\theta}_n := \widehat{\theta}(X_1, \ldots, X_n) \in R^d$

There are infinite possible estimators and the question is: can we find a *good* estimator?

We need certain measure to determine the quality of an estimator.

Accuracy Measure

Assume the statistical modeling is correct, and the true parameter is θ^*.(what if not?) This assumption will be used for the whole course unless otherwise stated

- Mean Squared Error (MSE): $E(\hat{\theta}_n - \theta^*)^2$
  $= \int (\hat{\theta}(x_1, \ldots, x_n) - \theta^*)^2 \prod_{i=1}^n f_{\theta^*}(x_i) dx_1 \ldots dx_n$
- Bias: $E\hat{\theta}_n - \theta^* = \int \hat{\theta}(x_1, \ldots, x_n) \prod_{i=1}^n f_{\theta^*}(x_i) dx_1 \ldots dx_n - \theta^*$
- Variance: $Var(\hat{\theta}(X_1, \ldots, X_n))$

- MSE = Bias*Bias+Variance
- An accurate estimator requires: small MSE, bias and variance

- MSE, bias and variance are functions over parameter space Θ
- Unbiased Estimator: if the bias is constant 0.
- Trade-off between bias and variance
- Example: $\widehat{\theta} = aY$ for estimating $\mu = E(Y)$
- It is common to find the least-variance estimator among all unbiased estimators

Correctness Measure

- Consistent estimator: as n increases to infinity, $\widehat{\theta}_n \to \theta^*$ for all possible value $\theta^* \in \Theta$
- The estimation procedure is correct: one can obtain the truth if inifnite data are given.
- If $\widehat{\theta}_n$ is consistent for θ, then $\phi(\widehat{\theta})$ is consistent for $\phi(\theta)$ for any continuous function ϕ
- If MSE converges to 0, then consistency holds.
- By LLN,
  - Sample mean $\bar{X}$ is consistent for population mean μ
  - Sample variance S^2 is consistent for population variance σ^2

LLN-based estimation

- $\mu^k := E(X^k)$ can be consistently estimated by $[\sum_{i=1}^n (X_i)^k]/n$ for all $k = 1, 2, \ldots..$
- If the parameter of interest can be formulated by μ^k's, i.e., $\theta = \phi(\mu^1, \mu^2, \ldots)$, then we estimate

$$\hat{\theta} = \phi(\sum_{i=1}^n (X_i)/n, \sum_{i=1}^n (X_i)^2/n, \ldots)$$

- Consistent is ensured
- Drawback 1: ϕ may not be unique
- Drawback 2: The estimator may be unreasonable.
  Example: $X_i \sim Unif[0, \theta]$ has an estimator $\hat{\theta} = 2\bar{X}$ which maybe smaller than max(X_i)

loss function-based estimation

We design a loss function $\mathcal{L}(X, \theta)$ which measures how good the distribution f_θ fits data X.

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \mathcal{L}(X_i, \theta)$$

- MLE: choose $\mathcal{L}$ to be the likelihood function, $\mathcal{L}(x, \theta) = -\log f_\theta(x)$
- Optimization problem
- Other choices are used, for the sake of better robustness or faster optimization.
- Quick exercise

Maximum Likelihood Estimation

Why likelihood function is a good choice?

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \mathcal{L}(X_i, \theta)$$

$$\Leftrightarrow 0 = \frac{1}{n} \sum \frac{\partial \mathcal{L}(X_i, \widehat{\theta})}{\partial \theta} \approx E \frac{\partial \mathcal{L}(X, \widehat{\theta})}{\partial \theta}$$

Furthermore,

$$E \frac{\partial \mathcal{L}(X, \theta^*)}{\partial \theta} = - \int \frac{\partial \log f_{\theta^*}(x)}{\partial \theta} f_{\theta^*}(x) dx$$

$$= - \int \frac{1}{f_{\theta^*}(x)} \frac{\partial f_{\theta^*}(x)}{\partial \theta} f_{\theta^*}(x) dx = - \int \frac{\partial f_{\theta^*}(x)}{\partial \theta} dx$$

$$= - \frac{\partial \int f_{\theta^*}(x) dx}{\partial \theta} = 0$$

Maximum Likelihood Estimation

Under proper conditions

- MLE is consistent
- MLE is almost the most efficient (sort of smallest MSE)
- MLE has a CLT type results

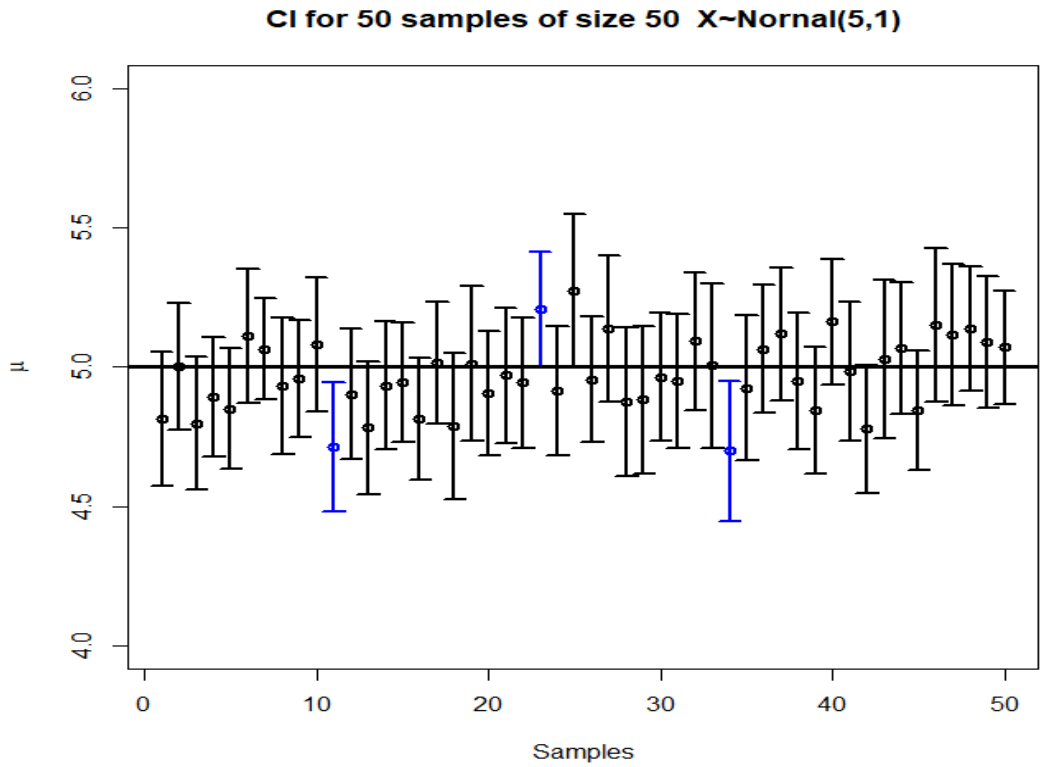
$$\sqrt{n}(\hat{\theta} - \theta^*) \approx N(0, \tau^2)$$

for some τ^2.

Interval Estimation

In order to incorporate accuracy information into the estimation. Wide intervals mean loose estimations, while narrow intervals mean accurate estimations

- A data dependent interval $[l(X_1, \ldots, X_n), u(X_1, \ldots, X_n)]$
- Confidence Interval (C.I.) $P(\theta^* \in [l(X_1, \ldots, X_n), u(X_1, \ldots, X_n)]) \geq 1 - \alpha$
- $1 - \alpha$ is the confidence level
- larger α leads to wider interval, in general.
- Long-run frequency justification


CI for 50 samples of size 50 X~Nornal(5,1)
μ
6.0
5.5
5.0
4.5
4.0
0
10
20
30
40
50
Samples

Designing C.I. for $\phi(\theta)$

- pivotal quantity for $\phi(\theta)$: A function g such that $g(X_1, \ldots, X_n, \phi(\theta^*))$ has a fixed distribution that doesn't depend on θ^*
- No matter what is the true parameter, g follows a distribution $\mathcal{G}_n$
- For this distribution, there always exist some constants g_1 and g_2 such that

$$Pr(g_1 \leq \mathcal{G}_n \leq g_2) \geq 1 - \alpha$$

- Equivalently

$$Pr(g_1 \leq g(X_1, \ldots, X_n, \phi(\theta^*)) \leq g_2) \geq 1 - \alpha$$

- $g_1 \leq g(X_1, \ldots, X_n, \phi(\theta^*)) \leq g_2$ means $\phi(\theta^*) \in \{\phi : g_1 \leq g(X_1, \ldots, X_n, \phi) \leq g_2\}$.

Designing C.I. for $\phi(\theta)$

Combining all above together, we have that

$$Pr(\phi(\theta^*) \in \{\phi : g_1 \leq g(X_1, \ldots, X_n, \phi) \leq g_2\}) \geq 1 - \alpha,$$

i.e., $\{\phi : g_1 \leq g(X_1, \ldots, X_n, \phi) \leq g_2\}$ is the $1 - \alpha$ C.I. for $\phi(\theta)$

Example z-C.I.

Observe $X_1, \ldots, X_n$. Assume they come from a $norm(\mu, \sigma_0^2)$ with known σ_0^2 and unknown μ

- pivotal quantity: $\sqrt{n}(\bar{X} - \mu^*)/\sigma_0$
- pivotal distribution: standard normal distribution
- choices of g_1, $g_2 = \pm z_{1-\alpha/2}$ (not unique choices)
- $\{\mu : |\sqrt{n}(\bar{X} - \mu)/\sigma_0| \leq z_{1-\alpha/2}\} = \bar{X} \pm z_{1-\alpha/2}\sigma_0/\sqrt{n}$

Example t-C.I.

Observe $X_1, \ldots, X_n$. Assume they come from a $norm(\mu, \sigma^2)$ with unknown σ_0^2 and unknown μ. Want a C.I. for μ

- pivotal quantity: $\sqrt{n}(\bar{X} - \mu^*)/S$, where $S = \sqrt{S^2}$ is the sample standard deviation
- pivotal distribution: t_{n-1} distribution
- choices of g_1, $g_2 = \pm t_{n-1,1-\alpha/2}$ (not unique choices)
- $\{\mu : |\sqrt{n}(\bar{X} - \mu)/S| \leq t_{n-1,1-\alpha/2}\} = \bar{X} \pm t_{n-1,1-\alpha/2} S/\sqrt{n}$

Example χ^2-C.I.

Observe $X_1, \ldots, X_n$. Assume they come from a $norm(\mu, \sigma^2)$ with unknown σ_0^2 and unknown μ. Want a C.I. for σ^2

- pivotal quantity: $(n-1)S^2/\sigma^{*2}$
- pivotal distribution: χ^2_{n-1} distribution
- choices of g_1, $g_2 = \chi^2_{n-1,\alpha/2}, \chi^2_{n-1,1-\alpha/2}$ (not unique choices)
- $\{\sigma^2 : \chi^2_{n-1,\alpha/2} \leq (n-1)S^2/\sigma^2 \leq \chi^2_{n-1,1-\alpha/2}\}$
  $= [(n-1)S^2/\chi^2_{n-1,1-\alpha/2}, (n-1)S^2/\chi^2_{n-1,\alpha/2}]$

Pivotal quantity for local/scale parameter

- θ is a location parameter if $X \sim f_\theta$ implies that $X - \theta \sim f_0$
- X can be represented as $Z + \theta$ where Z follow some reference distribution

- θ is a scale parameter if $X \sim f_\theta$ implies that $X/\theta \sim f_0$
- X can be represented as $Z\theta$ where Z follow some reference distribution

- (θ_1, θ_2) is location and scale parameters X can be represented as $\theta_2 Z + \theta_1$ where Z follow some reference distribution

- Example: μ and σ^2 of the normal family, $1/\lambda$ of the exponential family, $1/\beta$ for Gamma distribution family

Pivotal quantity for local/scale parameter

- θ is a location parameter, then (any statistic about data center) $-\theta$ is a pivotal quantity, such as $\bar{X}-\theta$ or (median of data)$-\theta$

- θ is a scale, then (any statistic about data dispersion) $/\theta$ is a pivotal quantity, such as S/θ or (IQR of data)$/\theta$ or (Range of data)$/\theta$

- (θ_1, θ_2) is location and scale parameters, then
  - ((any statistic about data center) - θ)/(any statistic about data dispersion) is a pivotal quantity for θ_1
  - (any statistic about data dispersion) / θ_2 is a pivotal quantity for θ_2

Example C.I. for exponential distribution

Observe $X_1, \ldots, X_n$. Assume they come from a exponential distribution $f_\lambda(x) = \lambda \exp -\lambda x$

- pivotal quantity: $S/(1/\lambda)$
- what is the pivotal distribution? and the corresponding g_1, g_2?
- Use simulation: when $n = 10$, $\alpha = 5\%$, $g_1, g_2 \approx 0.37, 1.51$
- $\{\lambda : 0.37 \leq S/(1/\lambda) \leq 1.51\} = [0.37/S, 1.51/S]$

Designing Approximate C.I.

C.I. requires a probability statement while CLT gives an approximation for probability. So we aim to use CLT to construct C.I.

Example: Observe $X_1, \ldots, X_n$. Assume they come from a Bernoulli distribution with unknown p.

- Consider $\sqrt{n}(\bar{X} - p^*)/\sqrt{p^*(1-p^*)}$
- Is it a pivotal quantity? Ans: No, even its range depends on p^*
- By CLT: $\sqrt{n}(\bar{X} - p^*)/\sqrt{p^*(1-p^*)} \approx N(0,1)$, this implies that $Pr(z_{\alpha/2} \leq \sqrt{n}(\bar{X} - p^*)/\sqrt{p^*(1-p^*)} \leq z_{1-\alpha/2}) \approx 1 - \alpha$
- Approximate C.I. $\{p : |\sqrt{n}(\bar{X} - p)/\sqrt{p(1-p)}| \leq z_{1-\alpha/2}\}$

- Hate to solve a quadratic function? Add an additional approximation
- if we have a consistent estimator $\widehat{p}$ such as $\bar{X}$, i.e., $\sqrt{p^*(1-p^*)}/\sqrt{\widehat{p}(1-\widehat{p})} \to 1$
- Combine the above with CLT: $\sqrt{n}(\bar{X}-p^*)/\sqrt{\widehat{p}(1-\widehat{p})} \approx N(0,1)$.
- Repeat the previous procedure, we get an approximate C.I. $\{p : |\sqrt{n}(\bar{X}-p)/\sqrt{\widehat{p}(1-\widehat{p})}| \leq z_{1-\alpha/2}\} = \bar{X} \pm z_{1-\alpha/2}\sqrt{\widehat{p}(1-\widehat{p})/n}$

General Result of C.I. of mean

Observe $X_1, \ldots, X_n$. Assume they come from a distribution f_θ with unknown θ. We need a C.I. for the mean of this distribution

- Denote true mean μ^* and the true standard deviation is $\sigma(\theta^*)$
- we have a consistent estimator $\widehat{\theta}$
- By CLT: $\sqrt{n}(\bar{X} - \mu^*)/\sigma(\widehat{\theta}) \approx \sqrt{n}(\bar{X} - \mu^*)/\sigma(\theta^*) \approx N(0, 1)$, this implies that $Pr(z_{\alpha/2} \leq \sqrt{n}(\bar{X} - \mu^*)/\sigma(\widehat{\theta}) \leq z_{1-\alpha/2}) \approx 1 - \alpha$
- Approximate C.I. $\{\mu : |\sqrt{n}(\bar{X} - \mu)/\sigma(\widehat{\theta})| \leq z_{1-\alpha/2}\} = \bar{X} \pm z_{1-\alpha/2}\sigma(\widehat{\theta})/\sqrt{n}$
- Exercise: C.I. for mean of Poisson or exponential modelings

Wald C.I.

Observe $X_1, \ldots, X_n$. Assume they come from a distribution f_θ with unknown θ.

- Let $\hat{\theta}$ be the MLE estimation
- $\hat{\theta}$ is consistent
- $\sqrt{n}(\hat{\theta} - \theta^*) \approx N(0, \tau^2(\theta^*))$
- Thus $\sqrt{n}(\hat{\theta} - \theta^*)/\tau(\hat{\theta}) \approx \sqrt{n}(\hat{\theta} - \theta^*)/\tau(\theta^*) \approx N(0,1)$, this implies that $Pr(z_{\alpha/2} \leq \sqrt{n}(\hat{\theta} - \theta^*)/\tau(\hat{\theta}) \leq z_{1-\alpha/2}) \approx 1 - \alpha$
- Approximate C.I. $\hat{\theta} \pm z_{1-\alpha/2}\tau(\hat{\theta})/\sqrt{n}$
- Most used C.I. for a general statistical modeling

Designing Bootstrapping C.I.

Observe $X_1, \ldots, X_n$. Assume they come from a distribution f_θ with unknown θ. Assume that we have a good consistent estimator $\widehat{\theta}$ for θ, based on which we want to construct a C.I. But we have difficulties to derive the (approximate) distribution of $\widehat{\theta}$

- To find a valide confidence interval $[\widehat{\theta} - a, \widehat{\theta} + b]$ for some constants a, b, i.e., $Pr(\theta^* \in [\widehat{\theta} - a, \widehat{\theta} + b]) = 1 - \alpha$.
- Eqivalently

$$Pr(\widehat{\theta} \in [\theta^* - b, \theta^* + a]) = 1 - \alpha$$

- If we know the distribution of $\widehat{\theta}$ (which changes w.r.t. θ^*), then one can determine a and b (*which also depends on* θ^*)

Use simulation to determine the distribution of $\widehat{\theta}$

- Sample $X_1^{(b)}, \ldots, X_n^{(b)} \sim f_{\theta^*}$ for $b = 1, \ldots, B$ with very large B (Bootstrapping samples)
- $\widehat{\theta}^{(b)} = \widehat{\theta}(X_1^{(b)}, \ldots, X_n^{(b)})$ for $b = 1, \ldots, B$ (Bootstrapping estimates)
- $\widehat{\theta}^{(b)}$ are iid samples follow the same distribution about which we want to know.
- Let $[\theta^* - b, \theta^* + a]$ be the lower/upper sample quantiles of $\widehat{\theta}^{(b)}$'s.

Unfortunately, the above procedure requires that we know θ^*

Parametric Bootstrapping

Repeat the above procedure with some other $\tilde{\theta}$, which is close to θ^*. The a, b values obtained under $\tilde{\theta}$ should be close to the a, b values under θ^*.

An immediate choice $\tilde{\theta} = \hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$, since it is consistent.

Bootstrapping Procedure:

- Sample $X_1^{(b)}, \ldots, X_n^{(b)} \sim f_{\hat{\theta}}$ for $b = 1, \ldots, B$ with very large B (Bootstrapping samples)
- $\hat{\theta}^{(b)} = \hat{\theta}(X_1^{(b)}, \ldots, X_n^{(b)})$ for $b = 1, \ldots, B$ (Bootstrapping estimates)
- Let $[\hat{\theta} - b, \hat{\theta} + a]$ be the lower/upper sample quantiles of $\hat{\theta}^{(b)}$'s.
- That is, we first sort all $\hat{\theta}^b$, then $\hat{\theta} - b = \hat{\theta}^{(\lfloor B*\alpha/2 \rfloor)}$ and $\hat{\theta} + a = \hat{\theta}^{(\lceil B*(1-\alpha/2) \rceil)}$
- So, the final C.I. is $[2\hat{\theta} - \hat{\theta}^{(\lceil B*(1-\alpha/2) \rceil)}, 2\hat{\theta} - \hat{\theta}^{(\lfloor B*\alpha/2 \rfloor)}]$

Last Thinking

For the same statistical problem, there may exist several different valid C.I. formulas. To obtain an accurate C.I. (i.e., narrow interval), can we compute all possible intervals and then pick the narrowest one?

For example, both z and t C.I.s can be applied to normal with a known variance problem. Given a data set, can we compute both z and t C.I.s and then use the narrower one for statistical inferences?

Chapter Review

- Statistical Modeling
- Point estimations and its evaluation
- Two popular point estimations
- Confidence Interval
- Three popular confidence intervals