Notes for STA238H1: Probability, Statistics and Data Analysis II

Jaehyeon Park

February 28, 2024

Contents

1 Preface

About

This document is my organized notes that provides basic concepts for STA238H1 that I took during Winter 2024 semseter at the University of Toronto. Throughout the semester, I began to recognize my lack of knowledge and familiarity in LaTeX. Also, some of my friends asked for my notes, so I just created one for all. The purpose of this document is to assist students, not to skip lectures.

Please note that this does not contain any material related to R code. Further, materials covered in future semesters may vary. Unfortunately, I will not be able to adjust the content every semester.

Each section consists of material covered in approximatly one week (4 hours). All sections begin with a summarized material in a blue box. In case when reference sheets are permitted for tests, such part will be highly beneficial. Some subsections contain examples. Existence of examples does not imply heavier emphasis on such content. I was simply lazy.

How can I succeed?

Some asked me how to be successful in this course. My answer is simple: attend ALL lectures AND tutorials, complete ALL practice AND extra problems, and go to office hours. I had tutorials where only 3 of 50 students attended. Before you start complaining how ”hard” the term test was, ask yourself if you've put enough effort into these courses.

Updates

If you catch a typo or incorrect information, please contact me. I will update accordingly.

2 Introduction to data analysis

Readings: *Dekking et al., Chapter 15*

2.1 Histogram

Shapes of histogram

Constructing histogram

2.2 Kernel Density Estimate (KDE)

Properties of Kernels

Constructing KDE

2.3 Scatterplot

2.4 Empirical cumulative distribution function (eCDF)

2.5 Center and variabiltiy of the data

2.6 Five-number summary and boxplot

3 Statistical Modelling

Readings: *Dekking et al., Chapter 13 & 17*

3.1 General principle

3.2 Estimating features of the model distribution

4 Estimators and their distributions

Readings: *Dekking et al., Chapter 14 & 19*

4.1 General principles

4.2 Sampling distribution

4.3 Distributions related to Normal distribution

Standard Normal distribution

$\chi^2(n)$ distribution

$t(n)$ distribution

$F(m, n)$ distribution

5 Evaluating estimators

Readings: *Dekking et al., Chapter 20*

5.1 Unbiased estimator

5.2 Jensen's inequality

5.3 Efficiency of estimators

5.4 Cramer-Rao inequality

5.5 Mean squared error

6 Maximum likelihood estimation

Readings: *Dekking et al., Chapter 21*

Definition

Maximum likelihood estimator is an estimator based on highest likelihood of possible parameter values given a dataset.

Constructon of MLE

1. Construct a likelihood function of parameter θ, where $p_\theta(x_i)$ is the probability mass function associated with θ

$$L(\theta) = \prod_{i=1}^{n} p_\theta(x_i)$$

2. Construct a log-likelihood function based on likelihood function

$$l(\theta) = \sum_{i=1}^{n} \log p_\theta(x_i)$$

3. Take the partial derivative with respect to θ and set it equal to 0

$$\frac{\partial}{\partial\theta} l(\theta) = 0$$

4. (optional) Take the second partial derivative to ensure it is the maximum

$$\frac{\partial^2}{\partial^2\theta} l(\theta) < 0$$

note.

The above description is based on θ associated with discrete distribution. If dealing with continuous distribution, use the probability desnsity function $f_\theta(x_i)$ instead of $p_\theta(x_i)$.

6.1 General principle

In previous sections, some methods like methods of moments to estimate θ produced unrealistic estimate. So, we need a better method that can reliably produce a sensible estimate. Maximum likelihood principle is associated with constructing a sensible estimator.

6.2 MLE with Discrete distribution

The above summary basically explains how to derive $L(\theta)$ and $l(\theta)$

Example

Consider *i.i.d* distributions $X_1, ..., X_n \sim Bin(5, \theta)$. We are interested in deriving MLE $\hat{\theta}$ for θ. First, we know that θ is associated with discrete distribution. Also, we know that probability mass fucntion for binomial distribution is:

$$p_\theta(x_i) = \binom{5}{x} \theta^{x_i} (1-\theta)^{5-x_i}$$

1. Construct a likelihood function $L(\theta)$:

$$L(\theta) = \prod_{i=1}^{n} \binom{5}{x} \theta^{x_i} (1-\theta)^{5-x_i}$$

2. Construct a log-likelihood function $l(\theta)$:

$$l(\theta) = \sum_{i=1}^{n} \ln \binom{5}{x} + \ln(\theta^{x_i}) + \ln((1-\theta)^{5-x_i}) = \sum_{i=1}^{n} \ln(\binom{5}{x}) + x_i \ln(\theta) + (5-x_i)\ln(1-\theta)$$

3. Take the partial derivative with respect to θ.

Note that $\binom{5}{x}$ is a constant, so after taking the partial derivative, it will be gone.

$$\frac{\partial}{\partial\theta}l(\theta) = \sum_{i=1}^{n}\frac{1}{\theta}x_i - \frac{1}{1-\theta}(5-x_i)$$

$$\frac{\partial}{\partial\theta}l(\theta) = \frac{1}{\theta}\sum_{i=1}^{n}x_i - \frac{1}{1-\theta}\sum_{i=1}^{n}(5-x_i)$$

4. Set $l(\theta) = 0$ and solve for θ

$$\frac{1}{\theta}\sum_{i=1}^{n}x_i = \frac{1}{1-\theta}\sum_{i=1}^{n}(5-x_i) \implies \theta = \frac{\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}5 - x_i + x_i} = \frac{1}{5n}\sum_{i=1}^{n}x_i$$

Thus, $\hat{\theta} = \frac{1}{5n}\sum_{i=1}^{n}x_i$ is the MLE for θ.

6.3 MLE with continuous distribution

Example

Consider *i.i.d* distributions $X_1, ..., X_n \sim Exp(\lambda)$. We are interested in deriving MLE $\hat{\lambda}$ for λ. First, we know that θ is associated with continuous distribution. Also, we know that probability density function for binomial distribution is:

$$f_\lambda(x) = \lambda e^{-\lambda x}, \text{ for } x \geq 0$$

1. Construct a likelihood function $L(\lambda)$

$$L(\lambda) = \prod_{i=1}^{n}\lambda e^{-\lambda x_i}$$

2. Construct a log-likelihood function $l(\lambda)$

$$l(\lambda) = \sum_{i=1}^{n}\ln(\lambda e^{-\lambda x_i}) = \sum_{i=1}^{n}(\ln\lambda - \lambda x_i)$$

3. Take the partial derivative with respect to λ

$$\frac{\partial}{\partial\lambda}l(\lambda) = \sum_{i=1}^{n}(\frac{1}{\lambda} - x_i) = \frac{n}{\lambda} - \sum_{i=1}^{n}x_i$$

4. Set $l(\lambda) = 0$ and solve for λ

$$\frac{n}{\lambda} = \sum_{i=1}^{n}x_i \implies \lambda = \frac{1}{n}\sum_{i=1}^{n}x_i$$

Thus, $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n}x_i$ is the MLE for λ

7 Bootstrap principles

Readings: *Dekking et al., Chapter 18*

General definition

Bootstrapping approximates sampling distribution of sample statistic by resampling dataset B times.

Empirical bootsrapping method procedure

Repeat step 1,2,3 for each $b = 1, 2, ..., B$.

step 1. Sample bootsrap sample of size n with replacement from the given dataset.

step 2. Compute the bootsrap statistic $\hat{\theta}_b^*$ from the bootsrap sample obtained in step 1.

step 3. Compute the centered boostrap statistic from the bootstrap dataset:

$$(\hat{\theta} - \theta)_b^* =$$

Parametric bootsrapping method procedure First make assumption about the shape of the sampling distribution.

7.1 General principle

7.2 Empirical bootstrapping method

7.3 Parametric bootstrapping method

8 Confidence intervals

Readings: *Dekking et al., Chapter 23 & 24*

General definition
With $(1-\alpha)\%$ confidence, where α is called *significance level*,
θ is in *confidence interval* of:

$$(\hat{\theta}-c, \hat{\theta}+c)$$

Chebyshev's inequality
Given *i.i.d.* distributions X_n with known n, σ^2, and random estimate $\hat{\mu}$ for μ:
With at least$(1-\alpha)\%$ confidence, μ is in *confidence interval* of :

$$(-\frac{\sigma}{\sqrt{\alpha n}}+\hat{\mu}, \frac{\sigma}{\sqrt{\alpha n}}+\hat{\mu})$$

Confidence interval of mean for Normal distribution
Given $X_n \sim Normal(\mu, \sigma^2)$ and $(1-\alpha)\%$ confidence:
If σ^2 is given, μ is in *confidence interval* of:

$$(\bar{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

If σ^2 is unknown, μ is in *confidence interval* of:

$$(\bar{x}_n - \frac{t_{n-1,\alpha/2}(S_n)}{\sqrt{n}}, \bar{x}_n + \frac{t_{n-1,\alpha/2}(S_n)}{\sqrt{n}})$$

Confidence interval of mean for unknown distribution
With $(1-\alpha)\%$ confidence,
When n is considered large, apply CLT to approximate sampling distribution to be standard Normal. μ is in approximate *confidence interval* of:

$$(\bar{x}_n - z_{\alpha/2}\frac{s_n}{\sqrt{n}}, \bar{x}_n + z_{1-\alpha/2}\frac{s_n}{\sqrt{n}})$$

When n is not cosidered large, use empirical bootstrap method to estimate the distribution. μ is in approximate *confidence interval* of

$$(\bar{x}_n - c_u^*\frac{s_n}{\sqrt{n}}, \bar{x}_n + c_u^*\frac{s_n}{\sqrt{n}})$$

8.1 General principles

Previously, *sample statistics* were used as estimators for distribution parameters. Specifically, a single estimate was used for each parameter, but we know there are underlying uncertainties associated. We can quantify the uncertainty (or how "confident" we are about the estimate) by using *confidence intervals.*

Confidence interval
Consider sample size of n from a distribution X with parameter θ with finite expectation and variance ($\mu_X, \sigma_X < \infty$). We want to determine the possible values for θ, given the data.
Suppose $\hat{\theta}$ is an estimator for θ. Then using sampling distribution of $\hat{\theta}$, we can study

$$P(\hat{\theta}-c < \theta < \hat{\theta}+c) = 1-\alpha$$

Here, the interval $(\hat{\theta}-c, \hat{\theta}+c)$, or (L_n, U_n) is *confidence interval*, which contains θ with the probability of $1-\alpha$. In other words, given a dataset, we are $(1-\alpha)\%$ *confident* that true θ is

between $\hat{\theta} - c$ and $\hat{\theta} + c$, where $1 - \alpha$ is *confidence* level and α is *significance* level.
note. L_n, U_n are random, while θ, α are constant.

The following subsections will introduce how to construct c under different conditions.

8.2 Chebyshev's inequality

Consider *i.i.d.* distributions $X_1, ..., X_n$ with parameter with $\mu_X, \sigma_X < \infty$.
If we expand the absolute value from the inequality, we can construct the confidence interval.

Suppose $\hat{\mu}$ is an estimator for μ. Then $\mathbb{E}[\hat{\mu}] = \mu$ and $Var(X) = \frac{\sigma^2}{n}$. Now, we can construct the Chebyshev's inequality.

$$P(\|\hat{\mu} - \mathbb{E}[\hat{\mu}]\| \leq \epsilon) \geq 1 - \frac{Var(X)}{\epsilon^2}$$

$$P(\|\hat{\mu} - \mu\| \leq \epsilon) \geq 1 - \frac{\frac{\sigma^2}{n}}{\epsilon^2} = 1 - \frac{\sigma^2}{n\epsilon^2}$$

Then, we know $1 - \frac{\sigma^2}{n\epsilon^2}$ is the confidence level. Now, given the significance level, we are able to construct general confidence interval in terms of $\hat{\mu}, n, \sigma^2, \alpha$:

$$P(-\frac{\sigma}{\sqrt{\alpha n}} + \hat{\mu} \leq \mu \leq \frac{\sigma}{\sqrt{\alpha n}} + \hat{\mu}) \geq 1 - \alpha$$

Interpretation
μ is in the interval $(-\frac{\sigma}{\sqrt{\alpha n}} + \hat{\mu}, \frac{\sigma}{\sqrt{\alpha n}} + \hat{\mu})$ with confidence of *at least* $(1 - \alpha)\%$

note.
Remember that Chebyshev's inequality provided above is associated with the *minimum* value.
If we were to use the inequality related to maximum value, it would provide the *maximum* value for confidence level.

Example
Suppose we are looking for confidence interval for μ, with *at least* 70% confidence level for distribution X with $\sigma^2 = 25$, $n = 100$, $\hat{\mu} = 10$.
Referring back to how we constructed confidence interval using Chebyshev's inequality:

$$P(\|\hat{\mu} - \mu\| \leq \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2} = 0.7$$

Substiuting the variables with given values, we can solve for ϵ:

$$1 - \frac{25}{100\epsilon^2} = 0.7$$

$$\epsilon = \frac{5}{\sqrt{0.3(100)}}$$

Now we can compute the confidence interval:

$$P(\|10 - \mu\| \leq \frac{5}{\sqrt{0.3(100)}}) \geq 0.7$$

$$P(-\frac{5}{\sqrt{0.3(100)}} - 10 \leq -\mu \leq \frac{5}{\sqrt{0.3(100)}} + 10) \geq 0.7$$

$$P(-\frac{5}{\sqrt{0.3(100)}} + 10 \leq \mu \leq \frac{5}{\sqrt{0.3(100)}} + 10) \geq 0.7$$

$$P(9.087 \leq \mu \leq 10.913) \geq 0.7$$

Interpretation
μ is in the interval $(9.087, 10.913)$ with confidence of least 70%.

8.3 Normal distribution for mean

In some conditions, we are given that the dataset follows Normal distribution, which means we don't have to approximate the distribution of the dataset. Then, there are two methods to construct confidence interval for μ, depending on whether the other parameter, variance, is known. Unlike Chebyshev's inequality, this method will provide exact confidence level for constructed interval.

Consider *i.i.d* distributions $X_1, ...X_n \sim Normal(\mu, \sigma^2)$. Since we are looking for confidence interval for μ, we know it will be unknown, but what about σ^2? We can construct confidence intervals regardless by estimating σ^2 or using the given σ^2.

Known variance
If variance is known, the construction of confidence interval is fairly simple when using standard Normal:
Going back to general definition of confidence interval:

$$P(L_n < \mu < U_n) = 1 - \alpha$$

We can convert this with *critical values* of standard Normal distribution:

$$P(z_{1-\alpha/2} < \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) = 1 - \alpha$$

$$P(z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X}_n - \mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

$$P(\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

Interpretation
μ is in the interval $(\bar{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}})$ with confidence of $(1-\alpha)\%$

Unknown variance
Since the variance is unknown, the best alternative would be to estimate σ^2 with sample variance S_n^2. Here, since true variance was not used, we cannot directly use standard Normal distribution. Instead, we can use t-distribution as an alternative:
Going back to general definition of confidence interval and *studentized mean* $\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$:

$$P(L_n < \mu < U_n) = 1 - \alpha$$

We can convert this with $t_{m,p}$, which is the $(1-p)^{th}$ quantile of the $t(m)$ distribution:

$$P(t_{n-1,1-\alpha/2} < \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} < t_{n-1,\alpha/2}) = 1 - \alpha$$

$$P(t_{n-1,1-\alpha/2}\frac{S_n}{\sqrt{n}} < \bar{X}_n - \mu < t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}) = 1 - \alpha$$

$$P(\bar{X}_n - t_{n-1,1-\alpha/2}\frac{S_n}{\sqrt{n}} < \mu < \bar{X}_n + t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}) = 1 - \alpha$$

Interpretation
μ is in the interval $(\bar{x}_n - t_{(n-1,1-\alpha/2)}\frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{(n-1,\alpha/2)}\frac{s_n}{\sqrt{n}})$ with confidence of $(1-\alpha)\%$

note.
Remember that when dealing with S_n^2 for t-distribution, you need to first find S_n^2 with $\frac{Var(X_i)}{n-1}$ before you take the root of it.

8.4 Unknown distribution for mean

When we are not given with the distribution, often times we would have to estimate the distribution as well. If we have large enough sample size, we can apply central limit theorem. If not, we can apply bootstrpping principle to estimate the distribution.

Large sample

Consider *i.i.d* unknown distributions $X_1, ...X_n$. If we conclude that n is sufficiently large enough, we can estimate that $X_i \sim Normal(\mu_x, \frac{\sigma_x^2}{n})$. Since σ_x^2 is also unknown, we would have to approximate it with sample variance S_n^2. Now, similar to how we constructed the confidence interval for Normal distribution:

$$P(z_{1-\alpha/2} < \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} < z_{\alpha/2}) = 1 - \alpha$$

$$P(z_{1-\alpha/2}\frac{S_n}{\sqrt{n}} < \bar{X}_n - \mu < z_{\alpha/2}\frac{S_n}{\sqrt{n}}) = 1 - \alpha$$

$$P(\bar{X}_n - z_{\alpha/2}\frac{S_n}{\sqrt{n}} < \mu < \bar{X}_n + z_{1-\alpha/2}\frac{S_n}{\sqrt{n}}) = 1 - \alpha$$

Interpretation

μ is in the interval $(\bar{x}_n - z_{\alpha/2}\frac{s_n}{\sqrt{n}}, \bar{x}_n + z_{1-\alpha/2}\frac{s_n}{\sqrt{n}})$ with confidence of $(1-\alpha)\%$

Small sample

Consider *i.i.d* unknown distributions $X_1, ...X_n$. If we conclude that n is not large enough, we cannot apply central limit theorem to approximate the distribution. So, we use empirical bootstrapping method as an alternative.

bootstrapping

Repeat step 1,2 for each $b = 1, 2, ..., B$.

step 1. Sample bootsrap sample of size n with replacement from the given dataset.

step 2. Compute the bootsrap studentized mean $t_b^* = \frac{\bar{x}_n^* - \bar{x}_n}{s_n^*/\sqrt{n}}$ from step 1.

Now, we have B copies of t_b^*, which approximates the distribution of studentized mean. Then, we approximate critical values c_l^*, c_u^* using quantiles of bootstrap sample and construct the confidence interval:

$$P(c_l^* < \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} < c_u^*) = 1 - \alpha$$

$$P(c_u^*\frac{S_n}{\sqrt{n}} < \bar{X}_n - \mu < c_l^*\frac{S_n}{\sqrt{n}}) = 1 - \alpha$$

$$P(\bar{X}_n - c_l^*\frac{S_n}{\sqrt{n}} < \mu < \bar{X}_n + c_u^*\frac{S_n}{\sqrt{n}}) = 1 - \alpha$$

Interpretation

μ is in the interval $(\bar{x}_n - c_l^*\frac{s_n}{\sqrt{n}}, \bar{x}_n + c_u^*\frac{s_n}{\sqrt{n}})$ with confidence of $(1-\alpha)\%$

note.

As mentioned in previous section, make sure to practice bootsrapping with small values of B instead of relying on computing powers.

9 Statistical testing

15

10 Goodness of fit

11 Introduction to Bayesian inference

12 Estimation in Bayesian inference

13 Predictive modelling