STA261: Probability and Statistics II

Shahriar Shams

Week 2 (Point estimation and some properties)

UNIVERSITY OF TORONTO

Winter 2020

Recap

- Probability expressed as Expectation: $P[A] = E[I_A]$
- **LLN:** $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} E[X_i]$
- **CLT:** $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} N(0,1)$
- Linear combination of Normal variables follows Normal
- t, χ^2 and F distributions and how they are realted to Normal distribution

Learning goals for this week

- Basic idea of Population vs. Sample, Parameter vs. Statistic
- Types of inferences.
- Method of Moments Estimator
- Likelihood function and the idea behind it.
- Maximum Likelihood Estimator (MLE).
- Measuring quality of an estimator.
- Unbiasedness (one of the properties of an estimator)

These are selected topics from Evans and Rosenthal Chap 6.1 – 6.3 and John A. Rice Chap 8

Population vs. Sample

- **Population:** A collection of ALL the subjects that have something in common.
  - Example: all UofT students, all registered voters in Canada/Ontario, all vehicles manufactured in Canada.
  - Often we want to know some features (e.g. average/variance) of our population of interest.

- **Sample:** A subset of the population.
  - We use the sample to make inference about the unknown characteristics of our population.
  - The goal should be to make sure the sample is a representative one.

Parameter vs. Statistic (some confusing letters!)

- **Parameter:** A characteristic (or a summary) of the population
  - Example: Mean (μ), Standard deviation (σ) etc...
  - We use the the Greek letter θ to represent the parameter(s) of our population.
  - For example, when we say $X \sim Poisson(\lambda)$; θ stands for λ.
  - On the other hand when we say $X \sim Normal(\mu, \sigma)$; θ stands for both μ and σ

- **Statistic:** Any summary of the sample.
  - Example: sample total ($\sum X_i$), sample mean ($\bar{X}$), sample standard deviation (S) etc...
  - When a statistic is used to estimate a parameter it's called an estimator.
  - $\bar{X}$ is an estimator of μ or S is an estimator of σ.
  - Since a function of sample observations, often the sign $T(X)$ is used to represent a statistic/estimator.
  - For example, if we are dealing with sample mean then $T(X) = \bar{X}$ and for the sample standard deviation then $T(X) = S$

Estimate from an Estimator

- When we have observed a sample and calculate the value of an estimator, then that numerical value is called the estimate.
- Typically lower case letters are used to represent an estimate.

Parameter (θ)	Estimator (T)	Estimate (t)
μ	$\bar{X}$	$\bar{x}$
σ	S	s
Unknown Constant*	Random variable	Known Constant

- Notation of "estimator" used in Textbooks:
  - $T(s)$ in Evans & Rosenthal, where "s" stands for sample
  - $T(X_1, X_2, ...X_n)$ in John A Rice.
  - In some places just the letter T has been used in both the books.

*Note: In Bayesian school of thought parameters are not assumed to be fixed rather treated as random variables. We will talk about it briefly in couple of weeks.

Types of inferences

- **Estimation:**
  - Point estimation: Based on the sample observations, calculating a *particular value* as an estimate of the parameter θ
  - Interval estimation: Calculating a *range of values* that is likely to contain the parameter θ
- **Hypothesis testing:**
  - Based on the sample, assess whether a hypothetical value θ_0 is a plausible value of the parameter θ or not.

Different types of estimation

- Method of Moments Estimation.
- Maximum Likelihood Estimation.

Method of Moments Estimation [Rice-P260]

- Let $X_1, X_2, ..., X_n$ are independently and identically distributed (i.i.d.) random variables.
- Let the k^{th} population moment be

$$\mu_k = E[X^k]$$

- k^{th} sample moment based on sample

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

- We use $\hat{\mu}_k$ as an estimator of μ_k
- In other words, we use the sample moments as estimators of the population moments.

Examples

- Example-1: $X_1, X_2, ..., X_n \overset{iid}{\sim} Poisson(\lambda)$. Find the method of moments estimator of λ. [Rice, page-261]

- Example-2: $X_1, X_2, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Find the method of moments estimators of μ and σ^2. [Rice, page-263]

Summary of Method of Moments Estimator

- Express the lower order population moment(s) in terms of the parameter(s).
- Invert the expression(s) to express the parameter(s) in terms of the population moment(s)
- Replace the population moments using the sample moments.

Intuition behind likelihood inference [E&R-P297]

- Assume we have two distributions: $P1$ and $P2$, both discrete uniforms.
- Under $P1$, $X \sim Unif\{1, 2, ..., 10^3\}$
- Under $P2$, $X \sim Unif\{1, 2, ..., 10^6\}$
- We observe one sample value of X, say 10.
- Which distribution did it come from?
- Which distribution is more *likely* to have produced this random number?

Definition of Likelihood Func.

- Suppose $X_1, X_2, ..., X_n$ has a joint density or mass function $f(x_1, x_2, ..., x_n|\theta)$
- We observe sample, $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$
- Given the sample, the likelihood function of θ, noted as $L(\theta|x_1, x_2, ..., x_n)$, is defined as

$$L(\theta|x_1, x_2, ..., x_n) = f(x_1, x_2, ..., x_n|\theta)$$

- $L(\theta|x_1, x_2, ..., x_n)$, or often written as $L(\theta)$, is a function of θ
- **If** X follows a discrete distribution, it gives the probability of observing the sample as a function of the parameter θ.

Definition of Likelihood Func. (cont...)

- If $X_1, X_2, ..., X_n$ are $i.i.d.$ then their joint density is the product of marginal densities, $f_\theta(x)$.
- Hence, in $i.i.d$ case we write

$$L(\theta) = f_\theta(x_1) * f_\theta(x_2) * ... * f_\theta(x_n) = \prod_{i=1}^{n} f_\theta(x_i)$$

$X \sim Bernoulli(\theta)$, $0 < \theta < 1$

- we observe $i.i.d.$ sample $(1, 0, 1, 1, 0)$
- $L(\theta|X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 0) = \theta^3(1-\theta)^2$
- In general, for n samples,

$$L(\theta|x_1, x_2, ...x_n) = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}$$

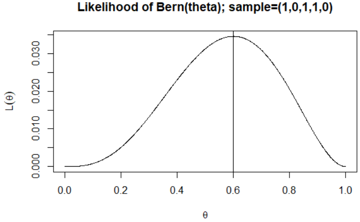
Comments on Likelihood Function

- $L(\theta)$ is NOT a pdf or pmf of θ.
- Likelihood introduces a belief ordering on parameter space, Ω
- For $\theta_1, \theta_2 \in \Omega$, we believe in θ_1 as the true value of θ over θ_2 whenever $L(\theta_1) > L(\theta_2)$
- Which means, the data is more likely to come from f_{θ_1} than f_{θ_2}
- The value $L(\theta)$ is very small for every value of θ (in the discrete case, as it's a multiplication of bunch of probabilities).
- So often, we are interested in the the *likelihood ratios*:

$$\frac{L(\theta_1)}{L(\theta_2)}$$

Maximum likelihood estimation [E&R-p308]

- Let's say we are interested in a point estimate of θ.
- A sensible choice will be to pick $\hat{\theta}$ that maximizes $L(\theta)$
- So, $\hat{\theta}$ satisfies $L(\hat{\theta}) \geq L(\theta)$ for all $\theta \in \Omega$
- $\hat{\theta}$ is called the *maximum likelihood estimate (MLE)* of θ



- For the numerical example used in slide 14, MLE of θ, $\hat{\theta} = 0.6$

Computation of the MLE

- Define, *log-likelihood function*, $l(\theta) = ln(L(\theta))$
- $ln(x)$ is a 1-1 increasing function of $x > 0 \implies L(\hat{\theta}) \geq L(\theta)$ for all $\theta \in \Omega$ if and only if $l(\hat{\theta}) \geq l(\theta)$
- In other words, if $L(\theta)$ is maximized at $\hat{\theta}$ then $l(\theta)$ will also be maximized at $\hat{\theta}$
- Therefore,

$$l(\theta) = ln \prod_{i=1}^{n} f_\theta(x_i) = \sum_{i=1}^{n} ln f_\theta(x_i)$$

- The obvious benifit $\implies$ It's much much easier to differentiate a sum than a product.
- Also $l(\theta)$ has some great properties which we will learn in couple of weeks.

Computation of the MLE (cont...)

- Solve the equation, $\frac{\partial l(\theta)}{\partial \theta} = 0$ for θ
- Say, $\hat{\theta}$ is the solution. But it's still not the MLE
- Need to check whether or not

$$\left.\frac{\partial^2 l(\theta)}{\partial \theta^2}\right|_{\theta=\hat{\theta}} < 0$$

- Example 1: $X_1, X_2, ..., X_n \overset{iid}{\sim} Poisson(\lambda)$. Find the MLE of λ.
- Example 2: $X_1, X_2, ..., X_n \overset{iid}{\sim} N(\mu, \sigma_0^2)$ where σ_0^2 is known. Find the MLE μ.

Properties of MLE

- MLE is not unique.
- MLE may not exists.
- The likelihood may not always be differentiable.
  - Example: $X_1, X_2, ..., X_n \overset{iid}{\sim} Unif[0, \theta]$
  - In this case $\hat{\theta} = max(x_1, x_2, ...x_n)$
  - We have to be careful when range of X involves θ (as it is in this example)

Invariance Property of MLE

Let $\hat{\theta}$ be the MLE of θ and $\psi(\theta)$ be any 1-1 function of θ defined on Ω, then $\psi(\hat{\theta})$ is the MLE of $\psi(\theta)$

Sampling distribution of an Estimator

- Recall: An Estimator (T) is a random variable.
- If we repeat the sampling procedure and keep calculating T for each set of sample and finally draw a density histogram based on the T values we get the sampling distribution of T
- Example: Distribution of $\bar{X}$ approaches Normal (using CLT)
- **Standard error:** Standard deviation of an estimator is called the standard error (SE)

$E[\bar{X}]$ and $Var[\bar{X}]$

Assume $X_1, X_2, ...X_n$ is an $i.i.d.$ sequence of random variables each having finite mean μ and finite variance σ^2.

$$\begin{aligned}
E[\bar{X}] &= E[\frac{1}{n}X_1 + \frac{1}{n}X_2 + ... + \frac{1}{n}X_n] \\
&= \frac{1}{n}E[X_1] + \frac{1}{n}E[X_2] + ... + \frac{1}{n}E[X_n] \\
&= \frac{1}{n}\mu + \frac{1}{n}\mu + ... + \frac{1}{n}\mu = \frac{1}{n}n\mu = \mu
\end{aligned}$$

$$\begin{aligned}
Var[\bar{X}] &= V[\frac{1}{n}X_1 + \frac{1}{n}X_2 + ... + \frac{1}{n}X_n] \\
&= ....(\text{you do it}) \\
&= \frac{\sigma^2}{n} \\
\implies SE(\bar{X}) &= \frac{\sigma}{\sqrt{n}}
\end{aligned}$$

Some comments from the previous slides

- $\bar{X}$ is a linear combination of $X_1, X_2, ...X_n$
- $E[\bar{X}] = \mu$ and $var[\bar{X}] = \frac{\sigma^2}{n}$ regardless of the distribution of X.

$X_1, X_2, ...X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

Then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, doesn't matter whether n is small or large

$X_1, X_2, ...X_n \overset{iid}{\sim}$ _____$(mean = \mu, var = \sigma^2)$

$E[\bar{X}] = \mu$ and $var[\bar{X}] = \frac{\sigma^2}{n}$ (based on slide 21)

And CLT says, when $n \to \infty$, $\bar{X}_n \xrightarrow{D} N(\quad,\quad)$

Measuring quality of an estimator [E&R-page 322]

- Let $\psi(\theta)$ be any real valued function of θ
- Suppose, T is an estimator of $\psi(\theta)$
- The most commonly used measurement of accuracy of an estimator is *Mean Squared Error (MSE)*
- $MSE_\theta(T) = E_\theta[(T - \psi(\theta))^2]$
- The smaller the value of $MSE_\theta(T)$, the more concentrated the sampling distribution of T is about the value $\psi(\theta)$
- Since the true value of θ is unknown, often we evaluate the $MSE_\theta(T)$ at $\theta = \hat{\theta}$

More on MSE

$$MSE_\theta(T) = var_\theta(T) + (E_\theta(T) - \psi(\theta))^2$$

Proof...

Unbiasedness

- **Bias:** The bias of an estimator T of $\psi(\theta)$ is given by $E_\theta(T) - \psi(\theta)$

- **Unbiased estimator:** When the bias of an estimator is zero, it's called unbiased.
  - So T is unbiased estimator of $\psi(\theta)$ when $E_\theta(T) = \psi(\theta)$
  - In other words, T is unbiased if $\psi(\theta)$ is the mean of the sampling distribution of T.
  - Example: On slide 21, we have shown $E[\bar{X}] = \mu$. Therefore, sample mean is an unbiased estimator of the population mean.

Comments on MSE and Unbiasedness

- $MSE(T) = var(T) + (Bias(T))^2$
- For unbiased estimators, $MSE(T) = var(T)$
- If all the other properties (we haven't studied them yet) are similar, then an unbaised estimator is preferred over a biased estimator.
- In practice, often an biased estimator with lower variance is preferred over an unbiased estimator with really high variance $\implies$ we minimize MSE.

Assignment (Non-credit)

Evans and Rosenthal

6.2.1-6.2.10, 6.2.13, 6.2.17, 6.2.24

John A. Rice

Exercise 8: 4-7, 18-21, 57, 60