Parameter Estimation

Fall 2018

Instructor:

Ajit Rajwade

Topic Overview

- Concept of maximum likelihood estimation
- Maximum likelihood estimates of the parameters of various distributions
- Concept of point estimate and confidence intervals
- Estimator bias, variance and mean-squared error

What is parameter estimation?

- In many applications, we have samples of a random variable which we know belongs to a distribution of so and so family (eg: Poisson, normal, etc.)

- However, we may **not** know *some* or *any* of the parameters of the distribution.

- We can however **estimate** the parameters given some *assumptions* about the samples. This is called **parameter estimation**.

Parameter estimation

- Let X be a continuous random variable with probability density $f_X(x;\theta)$.

- Consider a sample of X having value say x_1. Then we say that the **likelihood** of this sample is $f_X(x_1;\theta)$ assuming parameter θ.

- If we observe some n samples (rather n samples, one from each of the n iid random variables having density $f_X(x;\theta)$), then their **joint likelihood** is given as

  $f_{X1,X2,X3,\ldots,Xn}(x_1, x_2,\ldots, x_n;\theta)$ or in shorthand $f(x_1, x_2,\ldots, x_n;\theta)$.

Parameter estimation

- Now consider the case where we have the *n* sample values but we do not know θ.
- A reasonable way to estimate θ is to treat the joint likelihood as a function of θ and maximize it with respect to θ.
- Intuitively, such an estimate $\hat{\theta}$ is the one which maximizes the probability density of the observed values. It is called as the **maximum likelihood estimate** of θ.

Parameter estimation

- Maximum likelihood estimation is **equally applicable for discrete random variables** in which case the likelihood is defined in terms of discrete probabilities instead of probability density functions.

- In maximum likelihood estimation, it is often easier (in terms of calculations) to maximize $\log f(x_1, x_2, \ldots, x_n; \theta)$ instead of $f(x_1, x_2, \ldots, x_n; \theta)$.

- Doing so does not alter the estimate in any way for many densities that are not zero-valued except at +/- infinity.

Parameter estimation

- But how does one get a functional form for the joint likelihood $f(x_1, x_2, \ldots, x_n; \theta)$?

- One typically *assumes* that the samples are statistically *independent*.

- This is a reasonable assumption we shall use, though there are situations where it is violated (we will not see those situations in this course).

Examples (Derivations on the board and in the book)

- Maximum likelihood estimation of p parameter of a Bernoulli distribution
- Of the mean/variance λ of a Poisson distribution
- Of the mean and standard deviation of a Gaussian distribution
- Of the range $[\alpha,\theta]$ of a uniform distribution.
- ML with a twist: linear regression.

https://www.cse.iitb.ac.in/~ajitvr/CS215_Fall2018/MLE/

ML estimates are random variables!

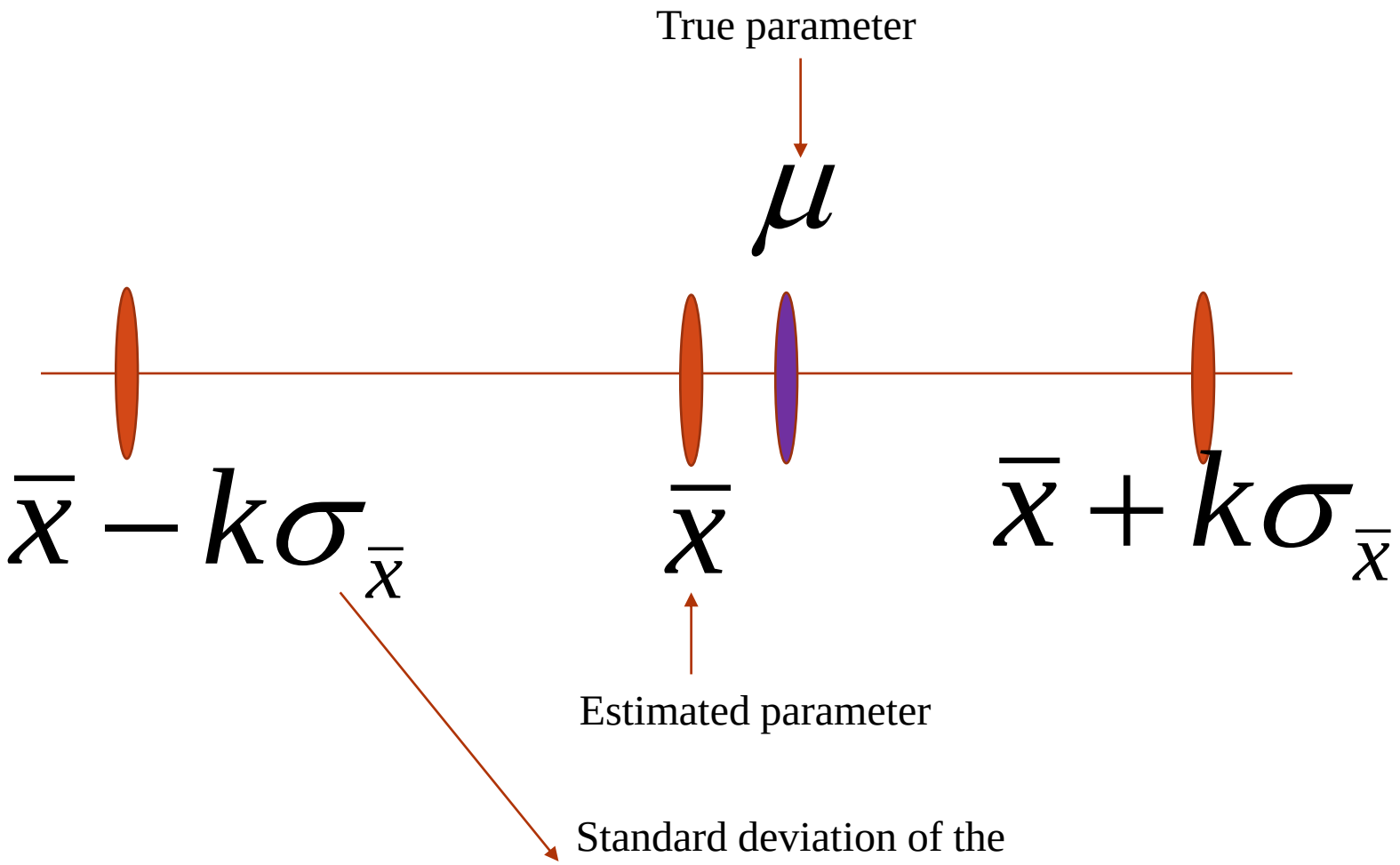
- The ML estimate is a random variable!
- Why? Because its value is a function of the samples from some underlying distribution.
- The ML estimate has its **own** probability density function, and its own mean, variance, etc.
- When you compute the value of an ML estimate, it is the value of a *sample* of that random variable.

Confidence intervals

- The value of the ML estimate of a parameter is a single scalar which *may differ* from the true parameter value. Such an estimate is called a **point estimate**.

- We do expect/wish the point estimate to be *close* to the true value – especially when the number of samples is *high*.

- One way to see this closeness, is to construct an interval around the ML estimate and show that the true value will lie inside this interval with high probability. This is called a **confidence interval**.


True parameter
μ
$\bar{x} - k\sigma_{\bar{x}}$
$\bar{x}$
$\bar{x} + k\sigma_{\bar{x}}$
Estimated parameter
Standard deviation of the estimated parameter

Confidence interval: empirical mean of a Gaussian (known variance)

- You know the following fact:

See clarification on next two slides

$$\sqrt{n}\left(\frac{\overline{X}-\mu}{\sigma}\right) \sim N(0,1)$$

Empirical mean of Gaussian samples

Known σ

- Hence

$$P\left[-2.5 < \sqrt{n}\left(\frac{\overline{X}-\mu}{\sigma}\right) < +2.5\right] \approx 0.99$$

$$\therefore P\left[\frac{-2.5\sigma}{\sqrt{n}} < \overline{X}-\mu < \frac{+2.5\sigma}{\sqrt{n}}\right] \approx 0.99$$

$$\therefore P\left[\frac{-2.5\sigma}{\sqrt{n}} < \mu-\overline{X} < \frac{+2.5\sigma}{\sqrt{n}}\right] \approx 0.99$$

$$\therefore P\left[\overline{X}-\frac{2.5\sigma}{\sqrt{n}} < \mu < \overline{X}+\frac{2.5\sigma}{\sqrt{n}}\right] \approx 0.99$$

Two-sided 99% confidence interval:

$$\left[\overline{X}-\frac{2.5\sigma}{\sqrt{n}}, \overline{X}+\frac{2.5\sigma}{\sqrt{n}}\right]$$

Clarification

- For Gaussian samples, the empirical mean given by $\overline{X}$ is always Gaussian distributed. Why?

- If X and Y are independent random variables, then $Z = X+Y$ has a pdf given by:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx$$

- If X and Y are Gaussian distributed, then it can be shown that Z is also Gaussian distributed (easy using MGFs, a bit tedious using basic integral calculus within the above formula).

Clarification

- This result extends to the sum of more than two Gaussian random variables as well, i.e. the sum(and hence average) of $n > 2$ Gaussian random variables is also a Gaussian random variable.

- Note: the empirical mean of any n i.i.d. random variables has an **approximate** (for finite n) Gaussian distribution by Central Limit Theorem.

- But if the original n random variables are themselves Gaussian, the empirical mean **exactly** has a Gaussian distribution.

Confidence interval: empirical mean of a Gaussian (known variance)

- You know that the following fact:

$$\sqrt{n}\left(\frac{\overline{X}-\mu}{\sigma}\right) \sim N(0,1)$$

Empirical mean of Gaussian samples

Known σ

- Hence

$$\therefore P\left[\overline{X}-\frac{2.5\sigma}{\sqrt{n}} < \mu < \overline{X}+\frac{2.5\sigma}{\sqrt{n}}\right] \approx 0.99$$

(two-sided) 99% confidence interval:

$$\left[\overline{X}-\frac{2.5\sigma}{\sqrt{n}}, \overline{X}+\frac{2.5\sigma}{\sqrt{n}}\right]$$

Note that this analysis and hence this confidence interval is **not** applicable in the case when the σ is **unknown** and hence needs to be estimated. In fact, the following random variable does **not** have a normal distribution but a student-t distribution instead – which we have not covered in class so far:

$$\sqrt{n}\left(\frac{\overline{X}-\mu}{\hat{\sigma}}\right) \text{ where } \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}}$$

Confidence interval: empirical mean of a Gaussian (known variance)

- Sometimes we may be interested in whether the true parameter *exceeds* its estimate with q% confidence.

- Hence

$$P\left[\sqrt{n}\left(\frac{\overline{X}-\mu}{\sigma}\right) < +2.35\right] \approx 0.99$$

$$P\left[\overline{X} - 2.35\left(\frac{\sigma}{\sqrt{n}}\right) < \mu\right] \approx 0.99$$

Why is this 2.35 instead of 2.5?

(upper one-sided) 99% confidence interval

Confidence interval: a clarification

- With 99% probability, our estimate (which **is** a random variable) will be such that the true parameter happens to lie inside a confidence interval around it (the estimated value).

Confidence interval: variance of a Gaussian

- Given samples from a normal distribution with unknown parameters, we know that

$$(n-1)\left(\frac{S^2}{\sigma^2}\right) \sim \chi^2_{n-1}, S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Definition :

$P(X \geq \chi^2_{\alpha,n-1}) = \alpha$

There are readily available tables for the chi-square distribution which give you $\chi^2_{\alpha,n-1}$ given any α, n.

$$P\left((n-1)\left(\frac{S^2}{\sigma^2}\right) \geq \chi^2_{\alpha/2,n-1}\right) = \frac{\alpha}{2}$$

$$P\left((n-1)\left(\frac{S^2}{\sigma^2}\right) \geq \chi^2_{1-\alpha/2,n-1}\right) = 1-\frac{\alpha}{2}$$

$$\therefore P\left(\chi^2_{1-\alpha/2,n-1} \leq (n-1)\left(\frac{S^2}{\sigma^2}\right) \leq \chi^2_{\alpha/2,n-1}\right) = 1-\alpha$$

$$\therefore P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right) = 1-\alpha$$

(approximate) Confidence interval: Mean of a Bernoulli Random variable

- Let X be the number of successes in a sequence of n Bernoulli trials with success probability p.

- X is a binomial random variable, and can be approximated as a normal random variable with mean np and variance $np(1-p)$. Hence we have:

$$\frac{X-np}{\sqrt{np(1-p)}} \sim N(0,1)$$

$$\therefore \forall \alpha, P\left(-z_{\alpha/2} < \frac{X-np}{\sqrt{np(1-p)}} < z_{\alpha/2}\right) \approx 1-\alpha, P(Z \geq z_{\alpha}) = \alpha$$

(approximate) Confidence interval: Mean of a Bernoulli Random variable

$$\frac{X-np}{\sqrt{np(1-p)}} \sim N(0,1)$$

$$\therefore \forall \alpha \in (0,1), P\left(-z_{\alpha/2} < \frac{X-np}{\sqrt{np(1-p)}} < z_{\alpha/2}\right) \approx 1-\alpha, P(Z \geq z_\alpha) = \alpha$$

$$\hat{p} = X/n$$

$$\therefore \forall \alpha \in (0,1), P\left(-z_{\alpha/2} < \frac{X-np}{\sqrt{n\hat{p}(1-\hat{p})}} < z_{\alpha/2}\right) \approx 1-\alpha, P(Z \geq z_\alpha) = \alpha$$

(this is an approximation)

$$\therefore P\left(-z_{\alpha/2}\sqrt{n\hat{p}(1-\hat{p})} < np - X < z_{\alpha/2}\sqrt{n\hat{p}(1-\hat{p})}\right) \approx 1-\alpha$$

$$\therefore P\left(-z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < p - X/n < z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right) \approx 1-\alpha$$

(approximate) Confidence interval: Mean of a Bernoulli Random variable

$$\therefore P\left(-z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < p - X/n < z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right) \approx 1-\alpha$$

$$\therefore P\left(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < p < \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right) \approx 1-\alpha$$

Estimator bias, variance and mean squared error

- Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a set of n i.i.d. random variables from a given distribution with parameter θ.
- Suppose θ' is an estimate of θ (obtained through some estimator).
- Note that the ML estimator is one particular estimator of θ, but there exist many other types of estimators.
- How does one conclude whether or not a given estimator is a good estimator?

Estimator bias, variance and mean squared error

- One metric for this is to evaluate $(\theta'-\theta)^2$.
- But this is a random variable! So we can consider its expected value, i.e. $E[(\theta'-\theta)^2]$.
- This is called the **mean squared error of the estimator**.
- We desire estimators with low mean-squared error.

Estimator bias, variance and mean squared error

- In some estimators, the expected value of the estimate, i.e. $E(\theta')$ may not be equal to the true parameter θ. Such estimators are called **biased**.

- The following quantity is called as the **bias of the estimator**: $E(\theta') - \theta$.

- Example of **unbiased estimator** (bias = 0): ML estimator for the mean of a Gaussian, Bernoulli, etc.

- Example of biased estimator: ML estimator of the variance of a Gaussian when mean is unknown, interval of a uniform distribution.

Estimator bias, variance and mean squared error

- Merely knowing the estimator bias does not give us a complete idea of the quality of the estimator.
- What if different samples of **X** gave wildly different estimates of θ for the same estimator?
- Such estimators are said to have a high **variance**.
- More formally, the **variance of an estimator** is defined as : $E[(\theta' - E(\theta'))^2]$.

Estimator bias, variance and mean squared error

- It can be proved that (proof on board and in book):

$$MSE(\theta') = E[(\theta' - E(\theta'))^2] + (E(\theta') - \theta)^2$$

variance

Squared bias

- Point to note: a biased estimator may have lower MSE than an unbiased estimator (because the latter may have higher variance).

Estimator bias, variance and mean squared error

- Examples of estimator bias, variance and MSE – on the board!
- Consider n samples X_1, X_2, ..., X_n from some distribution with parameter θ.
- An estimator of the form $\theta' = 1$ has very high bias (unless of course the true value of θ was 1) but very low variance (why?).
- Suppose θ represented the expected value of the samples. Then an estimator of the form $\theta'' = X_i$ will have no bias (why?) but variance equal to that of X_i. In particular note, that the variance of θ'' does not decrease with the number of samples n – which is not desirable.

Estimator consistency

- Let θ be the parameter of a distribution. Let an estimator of this parameter produce value $\hat{\theta}$.

- We say that the estimator is (asymptotically) **consistent** if

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0.$$

- The estimates θ' and θ'' on the previous slide are not consistent estimators.

- Note an estimator may be biased but still consistent (eg: ML estimator for variance of a Gaussian with unknown mean). And an unbiased estimator may be inconsistent (eg: θ'').

Motivation for MLE

- The MLE is a consistent estimator.
- No consistent estimators can achieve a lower asymptotic MSE than the MLE.
- We state these properties without proof.