Winter 2019 Math 106
Topics in Applied Mathematics
Data-driven Uncertainty Quantification

Yoonsang Lee (yoonsang.lee@dartmouth.edu)

Lecture 4: Parametric Inference
4.1 Statistical Inference

Statistical inference or learning is the process of using data to infer the distribution that generated the data.

Therefore, we can estimate statistical functionals of the unknown distribution

Note that any map of a distribution is called a statistical functional of the distribution

\[ F = F(P). \]

For example, for a distribution \( P(x) \) and its corresponding density \( p(x) \)

\[ E[X] = \int xp(x)dx \]
\[ \text{median} = P^{-1}(1/2) \]

For a sample of two random variables \( X \) and \( Y \) with a joint density \( p(x, y) \)

\[ E[Y|X = x] = \int yp(x, y)/p(x)dy \]
4.1 Statistical Inference

Example. Let $X_1, X_2, ..., X_n$ is a sample from a density $p(x)$. Infer $p(x)$ using the sample.

1. If we assume that $p(x)$ is a Gaussian, we need to estimate only the mean and variance using the sample mean and variance

$$\hat{m} = \frac{1}{n} \sum_i X_i$$

and

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_i (X_i - \hat{m})^2$$

2. Without assuming any form for $p(x)$, we estimate the $p(x)$ using a histogram
Example. Let $X_1, X_2, \ldots, X_n$ is a sample from a density $p(x)$. Infer $p(x)$ using the sample.

1. If we assume that $p(x)$ is a Gaussian, we need to estimate only the mean and variance using the sample mean and variance

$$\hat{m} = \frac{1}{n} \sum_{i} X_i$$

and

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i} (X_i - \hat{m})^2$$

2. Without assuming any form for $p(x)$, we estimate the $p(x)$ using a histogram

Example 1 is an example of *parametric inference* (where the unknown parameters are the mean and the variance). Example is an example of *nonparametric inference*. 
4.1 Statistical Inference

Broadly speaking, inferential problems fall into one of the three types

1. Point estimation
2. Confidence set (interval for 1D)
3. Hypothesis testing
4.1.1 Point Estimation

Let $F$ be a statistical functional of an unknown distribution $P$ and \{${X}_i$\} be a independent and identically distributed sample of $P$.

Point estimation provide a single best guess of $F$, often denoted by

$$\hat{F} = g(X_1, X_2, ..., X_n),$$

which is a function of the sample.
4.1.1 Point Estimation

Let \( F \) be a statistical functional of an unknown distribution \( P \) and \( \{X_i\} \) be a independent and identically distributed sample of \( P \).

Point estimation provide a single best guess of \( F \), often denoted by

\[
\hat{F} = g(X_1, X_2, ..., X_n),
\]

which is a function of the sample. This means that if we have a different sample \( \hat{F} \) changes. To be more precise, \( \hat{F} \) is a random variable.
4.1.1 Point Estimation

Let $F$ be a statistical functional of an unknown distribution $P$ and \( \{X_i\} \) be a independent and identically distributed sample of $P$.

Point estimation provide a single best guess of $F$, often denoted by

\[
\hat{F} = g(X_1, X_2, ..., X_n),
\]

which is a function of the sample.

The distribution of $\hat{F}$ is called the **sampling distribution** and its standard deviation is called the **standard error**, denoted by $se$.

\[
se = \sqrt{Var(\hat{F})}
\]
4.1.1 Point Estimation

Let $F$ be a statistical functional of an unknown distribution $P$ and 
\{X_i\} be a independent and identically distributed sample of $P$.

Point estimation provide a single best guess of $F$, often denoted by

$$\hat{F} = g(X_1, X_2, ..., X_n),$$

which is a function of the sample.

- If the expected value of the point estimator is equal to the 
  true value $F_{true}$, then the estimator is called \textbf{unbiased}.
- If the estimator converges in probability to the true value as 
  the sample size, $n$, increases, the estimator is called \textbf{consistent}.
- The estimator is asymptotically Normal if the estimator 
  converges in distribution to a normal as the sample size 
  increases.
4.1.1 Point Estimation

The **mean squared error (MSE)** defined as

\[ E[(\hat{\theta} - \theta)^2] \]

can be written as

\[ \text{MSE} = \text{bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta}). \]
4.1.1 Point Estimation

Example. Let $X_1, X_2, ..., X_n$ is a sample of a Bernoulli($p$). The estimator of $p$ is given by

$$\hat{p} = \frac{1}{n} \sum X_i.$$ 

▶ $\hat{p}$ is unbiased.
▶ From the law of large numbers, it is also consistent.
▶ From the central limit theorem, it is asymptotically normal.
▶ The standard error $se = \sqrt{Var(\hat{p})} = \sqrt{\frac{p(1-p)}{n}}$.
▶ The estimated $se$ uses the estimated $\hat{p}$ for the standard error

$$se = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}.$$
4.1.2 Confidence Sets

Let \( \{X_i\} \) be an independent, identically distributed sample. A \( 1 - \alpha \) **confidence set** is a set \( C \), which is a function of the sample, such that

\[
\mu(F \in C) = 1 - \alpha.
\]

That is, the probability that \( C \) traps the true value \( F \) is \( 1 - \alpha \).

**Example.** Let \( F \) is a scalar value. If an estimator \( \hat{F} \) is asymptotically normal and the sample size \( n \) is large, the \( 1 - \alpha \) confidence interval \( C_n \) is given by

\[
(\hat{F} - z_{\alpha/2} \hat{se}, \hat{F} + z_{\alpha/2} \hat{se})
\]

where \( z = \Phi^{-1}(1 - (\alpha/2)) \) for the standard normal distribution \( \Phi \).
4.1.2 Confidence Sets

Let \( \{X_i\} \) be an independent, identically distributed sample. A \( 1 - \alpha \) confidence set is a set \( C \), which is a function of the sample, such that

\[
\mu(F \in C) = 1 - \alpha.
\]

That is, the probability that \( C \) traps the true value \( F \) is \( 1 - \alpha \).

A frequently asked question for a data scientist position. The interpretation, ”the probability of the true value \( F \) is in the set \( C \) is \( 1 - \alpha \)” is an incorrect statement.

When we construct a confidence set \( C \) using a sample \( \{X_i\} \), \( C \) is a random variable while the true value \( F \) is fixed. Thus, the definition of the confidence set

\[
\mu(F \in C) = 1 - \alpha.
\]

is about a probability of the random variable \( C \), not \( F \).
4.1.3 Hypothesis Testing

Hypothesis testing starts with a null hypothesis and check if the sample provide sufficient evidence to reject the theory. Check one of your favorite statistics books for details.
4.2 Parameteric Inference

Let \( \{X_i\} \) be an IID sample of a distribution \( P \). In the parametric inference, we assume that the form of the unknown distribution is parameterized by a set of parameters \( \theta = (\theta_1, \ldots, \theta_m) \)

\[
P(x) = P(x; \theta).
\]

If we have an estimate of the parameter, say \( \hat{\theta} \), the estimator provides an estimate of the distribution \( P(x; \hat{\theta}) \).

**Example.**

- If we assume that the sample is from a Gaussian distribution with a mean \( m \) and a variance \( \sigma^2 \), the parameter is a pair \( (m, \sigma^2) \).
- If we assume that the sample is from a Bernoulli\((p)\), the parameter is the mean \( p \).
4.2 Parameteric Inference

We will consider two methods for parametric inference

- Method of Moments
- Max Likelihood Estimator (MLE)
4.2.1 Method of Moments

For a sample $X_1, X_2, \ldots, X_n$, the $j$-th moment is

$$\alpha_j(\theta) = E[X^j] = \int x^j p(x; \theta) \, dx,$$

i.e., a function of $\theta$,

where $p(x; \theta)$ is the parametrized density of the parametrized distribution $P(x; \theta)$. The $j$-th sample moment, $\hat{\alpha}_j$, is

$$\hat{\alpha}_j = \frac{1}{n} \sum_{i} X_i^j$$

If the size of the parameter $\theta$ is $m$, the method of moments estimator $\hat{\theta}$ is defined to be the value $\theta$ such that

$$\alpha_j(\hat{\theta}) = \hat{\alpha}_j, \quad j = 1, 2, \ldots, k.$$
4.2.1 Method of Moments

Example. Let $X_1, X_2, ..., X_n$ be an IID sample of Bernoulli($p$).

- The size of parameter $\theta = p$ is 1.
- The first moment $\alpha_1(\theta) = \alpha_1(p) = p$ and the first sample moment $\hat{\alpha}_1$ is

$$\hat{\alpha}_1 = \frac{1}{n} \sum X_i.$$

- By setting $\alpha_1(\theta) = \hat{\alpha}_1$, we have

$$\hat{\theta} = \hat{p} = \frac{1}{n} \sum X_i.$$
4.2.1 Method of Moments

Example. Let $X_1, X_2, \ldots, X_n$ be an IID sample of Normal($m, \sigma^2$).

- The size of parameter $\theta = (m, \sigma^2)$ is 2.
- The first and the second moments are
  \[
  \alpha_1(m, \sigma^2) = \mu, \quad \alpha_2(m, \sigma^2) = m^2 + \sigma^2
  \]
- The sample first and the sample second moments are
  \[
  \hat{\alpha}_1 = \frac{1}{n} \sum X_i, \quad \hat{\alpha}_2 = \frac{1}{n} \sum X_i^2
  \]
- Solving the system of equations gives
  \[
  \hat{\mu} = \frac{1}{n} \sum X_i, \quad \hat{\sigma}^2 = \frac{1}{n} \sum (X_i - \hat{\mu})^2.
  \]

Note that $\sigma^2$ is biased (but consistent).
4.2.2 Maximum Likelihood Estimator

Let $X_1, X_2, \ldots, X_n$ be IID with a density $p(x; \theta)$. The joint distribution of the sample $p(x_1, x_2, \ldots, x_n; \theta)$ is

$$p(x_1, x_2, \ldots, x_n; \theta) = \prod_i^n p(x_i; \theta) = p(x_1; \theta)p(x_2; \theta) \cdots p(x_n; \theta)$$

This joint density as a function of $\theta$ is called the likelihood function

$$\mathcal{L}_n(\theta) = \prod_i^n p(x_i; \theta).$$

The likelihood is the probability (density) of the sample under the assumption of the parametric model. Note that $n$ is the sample size.

**Warning.** The likelihood function is not a density of $\theta$. 
4.2.2 Maximum Likelihood Estimator

**Definition.** The maximum likelihood estimator (MLE) $\hat{\theta}$ is the value $\theta$ that maximizes the likelihood function $\mathcal{L}_n(\theta)$. 
4.2.2 Maximum Likelihood Estimator

**Definition.** The maximum likelihood estimator (MLE) $\hat{\theta}$ is the value $\theta$ that maximizes the likelihood function $L_n(\theta)$.

**Example.** Let $X_1, X_2, ..., X_n$ is IID Bernoulli($p$). The likelihood function is

$$L_n(p) = \prod_{i}^{n} p^{X_i} (1 - p)^{1-X_i} = p^S (1 - P)^{n-S}$$

where $S = \sum X_i$.

Hence,

$$\ln L(p) = S \ln p + (n - S) \ln(1 - p).$$

Take the derivative and set it equal to zero gives

$$\hat{p} = \frac{S}{n}.$$
4.2.2 Maximum Likelihood Estimator

**Definition.** The **maximum likelihood estimator** (MLE) \( \hat{\theta} \) is the value \( \theta \) that maximizes the likelihood function \( L_n(\theta) \).

**Example.** Let \( X_1, X_2, ..., X_n \) is IID Normal \((m, \sigma^2)\). The likelihood function after a scaling is

\[
L(m, \sigma) = \prod \frac{1}{\sigma} \exp \left( -\frac{1}{2\sigma^2} (X_i - m)^2 \right) = \sigma^{-n} \exp \left( -\frac{1}{2\sigma^2} \sum_i (X_i - m)^2 \right)
\]

\[
= \sigma^{-n} \exp \left( -\frac{nS^2}{2\sigma^2} \right) \exp \left( -\frac{n(\overline{X} - m)^2}{2\sigma^2} \right)
\]

where \( \overline{X} = \frac{1}{n} \sum X_i \) and \( S^2 = \frac{1}{n} \sum (X_i - m)^2 \). The log-likelihood is

\[
l(m, \sigma) = -n \ln \sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\overline{X} - m)^2}{2\sigma^2}.
\]

Solving the gradient of \( l(m, \sigma) \) equal to zero gives

\[
\hat{m} = \overline{X} \quad \text{and} \quad \hat{\sigma} = S.
\]
4.2.2 Maximum Likelihood Estimator

**Exercise.** Let $X_1, X_2, ..., X_n$ is IID Uniform$(0, \theta)$. Find the MLE of $\theta$. 
4.2.3 Properties of MLE

Under certain conditions on the model, the MLE has the following properties

1. It is **consistent**. That is, $\hat{\theta}_n \rightarrow \theta_{true}$ in probability.

2. It is **equivalent**. If $\hat{\theta}_n$ is the MLE of $\theta$, then $g(\hat{\theta})$ is the MLE of $g(\theta)$.

3. It is **asymptotically normal**. $\hat{\theta}_n - \theta_{true}$ converges in distribution to $N(0, se^2)$.

4. It is **asymptotically optimal**. That is, roughly speaking, among all well-behaved estimators, the MLE has the smallest variance, at least for large samples.

5. It is approximately the **Bayes estimator**.
4.2.3 Properties of MLE

Idea of the proof for the consistency.

- Maximizing $\mathcal{L}_n(\theta)$ is equivalent to maximizing

$$M_n(\theta) = \frac{1}{n} \sum \ln \frac{p(X_i; \theta)}{p(X_i; \theta_{true})}.$$ 

- From the law of large numbers, $M_n$ converges to the expected value

$$E \left( \ln \frac{p(X; \theta)}{p(X; \theta_{true})} \right) = \int \ln \frac{p(x; \theta)}{p(x; \theta_{true})} p(x; \theta_{true}) dx$$

$$= -D(p(x; \theta_{true}), p(x; \theta)) \leq 0$$

with equality when $\theta = \theta_{true}$. 
### 4.2.3 Properties of MLE

**Idea of the proof for the asymptotically normal property.**

For $I_n(\theta) = \log \mathcal{L}_n(\theta)$

$$0 = I'_n(\hat{\theta}) \approx I'_n(\theta) + (\hat{\theta} - \theta)I''_n(\theta)$$

which yields

$$\hat{\theta} - \theta = -\frac{I'_n(\theta)}{I''_n(\theta)}$$

From the central limit theorem, $I'_n(\theta)/\sqrt{n}$ converges in distribution to $N(0, I(\theta))$ where $I(\theta)$ is the variance of $\frac{\partial}{\partial x} \ln p(x; \theta)$.

Also, from the law of large numbers, $I''_n(\theta)/n$ converges in probability to the mean of $\frac{\partial^2}{\partial x^2} \ln p(x; \theta)$, which is $I(\theta)$.

**Exercise.** Show that the mean of $\frac{\partial}{\partial x} \ln p(x; \theta)$ is 0.

**Exercise.** Show that the mean of $\frac{\partial^2}{\partial x^2} \ln p(x; \theta)$ is the variance of $\frac{\partial}{\partial x} \ln p(x; \theta)$, that is $I(\theta)$. 
4.2.3 Properties of MLE

- The **score function** is the first derivative of the parametrized density
  \[
  s(X; \theta) = \frac{\partial}{\partial x} \ln p(x; \theta).
  \]

- The variance of the sum of the score functions is called **Fisher information**
  \[
  I_n(\theta) = \text{Var}(\sum_{i}^n s(X_i; \theta)).
  \]

That is, the Fisher information is $nl(\theta)$ where $l(\theta)$ is the variance of the score function.
4.2.4 The Expectation-Maximization (EM) Algorithm

**Goal:** Find a $\theta$ that maximize $\mathcal{L}_n(\theta)$, i.e., the MLE estimator.

**Algorithm:**

1. Pick an initial value $\theta^0$. For $j = 1, 2, \ldots$, repeat steps 1 and 2.
2. (The E-step): Calculate
   \[
   J(\theta|\theta^j) = E \left( \ln \frac{\prod p(x_i, y_i; \theta)}{\prod p(x_i, y_i; \theta^j)} | x \right)
   \]
   This expectation is over the missing variable $\{y_i\}$ treating $\theta^j$ and $\{x_i\}$ are fixed.
3. Find $\theta^{j+1}$ maximizing $J(\theta|\theta^j)$. 

4.2.4 The Expectation-Maximization (EM) Algorithm

**Idea of the proof.** We want to show that the procedure increases the likelihood, that is, $\mathcal{L}(\theta^{j+1}) \geq \mathcal{L}(\theta^j)$.

From

$$J(\theta^{j+1}|\theta^j) = E \left( \ln \frac{\prod p(x_i, y_i; \theta)}{\prod p(x_i, y_i; \theta^j)} | \{x_i\} \right)$$

we have

$$\ln \frac{\mathcal{L}(\theta^{j+1})}{\mathcal{L}(\theta^j)} = J(\theta^{j+1}|\theta^j) - E \left( \ln \frac{\prod p(y_i|x_i; \theta)}{\prod p(y_i|x_i; \theta^j)} | \{x_i\} \right)$$

$$= J(\theta^{j+1}|\theta^j) + D(f_j, f_j + 1) \geq 0$$

where $f_j = \prod p(y_i|x_i; \theta)$. 
Example. Let $X_1, X_2, \ldots, X_n$ be a sample from a parametrized density
\[
p(x) = \frac{1}{2} \phi(x; \mu_1, 1) + \frac{1}{2} \phi(x; \mu_0, 1)
\]
where $\phi(x; \mu_i, 1)$ is a Gaussian density with a mean $\mu_i$ and a variance 1. Find the MLE.