This Week: Dr. Osonde Osoba

- Confidence Intervals (C-I)
  - Pollster's C-I
  - Hypothesis Testing
- Estimation
  - Properties: Bias, MSE, Consistency
  - Method of moments
  - Maximum Likelihood Estimation (MLE)
  - Maximum A Posteriori (MAP)
    - Expectation-Maximization (EM)
  - Maximum Entropy (Max Ent)
  - Minimum Mean-Squared Error (MMSE)

H/W:

L-G: 8.24 (ab), 8.28 (ab), 8.29 (a), 8.30 (a)

4.147 – 4.149

8.39 – 8.41

G: 6.27, 6.28

Last week, we covered CLT:

$\{X_i\}_{i=1}^n$: iid $\quad \sigma_X^2 < \infty$

$$Z_n = \text{std}(\bar{X}_n)$$

$$\Rightarrow \quad Z_n \xrightarrow{d} Z \sim N(0,1)$$

Qu 1: How good is the CLT CDF approximation for a given sample size, n?

i.e characterize $\sup_z |F_n(z) - \Phi(z)|$ as $n \to \infty$

Ans: Berry-Essen Theorem

$\exists\, C > 0 \quad \forall\, z \quad \forall n$ when $E[|X|^3] < \infty$

such that

$$|F_n(z) - \Phi(z)| \leq \frac{C}{\sqrt{n}} \cdot \frac{E[|X-\mu|^3]}{\sigma^3}$$

Convergence rate $\propto 1/\sqrt{n}$

where $F_n(z)$ is the cdf of $Z_n = \text{std}(\bar{X}_n)$.

$C \in [0.4097, 0.71]$

[$\Phi(z)$ is the standard normal CDF]

Qu 2: Can we use the convergence of $z_n = std(\bar{X}_n)$ to infer a probable range of values for μ?

i.e find $I = [a, b]$ such that $P(\mu \in I) = (1-\alpha)$.

Ans:

Consider the statement

$$P(z_n \in I) = 1 - \alpha$$

$n \to \infty \Rightarrow$ By CLT we can use the CDF $\Phi(z)$ to estimate this probability. Φ is symmetric around 0. So we expect $I = [-z_{\alpha/2}, z_{\alpha/2}]$ where the value of $z_{\alpha/2}$ depends on α and comes from analyzing $\Phi(z)$

$$\begin{aligned}
1-\alpha &= P(z_n \in [-z_{\alpha/2}, z_{\alpha/2}]) \\
&= P(-z_{\alpha/2} \le z_n \le z_{\alpha/2}) \\
&= P\left(-z_{\alpha/2} \le \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) \\
&= P(-z_{\alpha/2} \cdot \sigma/\sqrt{n} \le \bar{X}_n - \mu \le z_{\alpha/2} \cdot \sigma/\sqrt{n}) \\
&= P(\bar{X}_n - z_{\alpha/2} \cdot \sigma/\sqrt{n} \le \mu \le \bar{X}_n + z_{\alpha/2} \cdot \sigma/\sqrt{n})
\end{aligned}$$

$$1-\alpha = P\left(\mu \in \left[\bar{X}_n - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \bar{X}_n + \frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}}\right]\right)$$

So our range $\boxed{I = \left[\bar{X}_n - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \bar{X}_n + \frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}}\right]}$

(C-I)

I is called the $(1-\alpha)\cdot 100\%$ Confidence Interval for the mean μ.

μ is a constant and $\bar{X}_n$ is a r.v. So the statement $P(\mu \in I) = 1-\alpha$ does not have the typical "measure of preimage" meaning since μ is not a measurable function. The randomness is in I.

Confidence Intervals are an example of a random set.

i.e $I : \Omega \longrightarrow S \subset 2^{\mathbb{R}}$

In general, the C-I of an estimate $\hat{\theta}_n$ for θ is:

$$\hat{\theta}_n \pm \varepsilon$$

where ε is the margin of error (M.O.E).

e.g $\bar{X}_n$ estimating μ:

$$\bar{X}_n \pm \varepsilon$$

$$\boxed{\varepsilon = \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}} \quad (**)$$

$\sqrt{\mathrm{Var}(\bar{X}_n)}$ = standard error

$\Rightarrow$ To get a C-I for μ at the $(1-\alpha)100\%$ level with ε M.O.E. we need

$$\boxed{n = \frac{(z_{\alpha/2})^2 \cdot \sigma_x^2}{\varepsilon^2}}$$

Pollster's C-I:

e.g. "63% of Americans believe ___."

Usually at 95% confidence level and $\varepsilon = \pm 3$ percentage pt

This is really an attempt to estimate p in

$X_k \sim \text{Bernoulli}(p)$ where X_k encodes the binary behavior. (POP^n)

$\Rightarrow$ The sample mean is:

$$\hat{p}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$

and for large n $\hat{p}_n \overset{d}{\approx} N(\mu_x, \sigma_x^2/n)$ by CLT.

$\mu_x = p$; $\sigma_x^2 = p(1-p)$

[We have assume $\{X_k\}$ is iid and "representative"]

$\Rightarrow$ $(1-\alpha)100\%$ C-I for p

$$\hat{p}_n \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Now $p \in [0,1] \Rightarrow 0 \le p(1-p) \le 1/4$

MOE $= \pm 3$ pct pts

$\alpha = 5\% \Rightarrow \boxed{z_{\alpha/2} = 1.96}$	$\alpha = 1\% \Rightarrow \boxed{z_{\alpha/2} = 2.58}$
$n = \left(\frac{1.96}{(0.03)4}\right)^2$	$n = \left(\frac{2.58}{(0.03)4}\right)^2$
$\Rightarrow n \ge 1068$ samples	$\Rightarrow n \ge 1849$ samples

Confidence Intervals are the basis for frequentist hypothesis testing (HT). The result of a HT depend on whether a test statistic falls inside or outside a $(1-\alpha)100\%$ C-I.

e.g. [a basic Hypothesis Test]: (for μ_x)

① Formulate competing hypothesis about μ_x

$H_0: \mu_x = \mu_0$ [null hypothesis]

$H_{alt}: \mu_x \neq \mu_0$ [alternate hypothesis]

② Specify $(1-\alpha)100\%$ C-I for $\bar{X} - \mu_0$:

$(1-\alpha)100\%$ C-I : $\left[-z_{\alpha/2}\cdot \sigma/\sqrt{n},\ z_{\alpha/2}\cdot \sigma/\sqrt{n}\right]$

③ Calculate test statistic $\bar{x}$ for observed data.

④ Reject H_0 if $\bar{x} \notin (1-\alpha)100\%$ C-I

"Do not reject" H_0 / accept H_{alt} if

$\bar{x} \in (1-\alpha)100\%$ C-I

C-Is for ρ, σ^2, σ_1^2/σ_2^2 determine other types of Hypothesis tests for other parameters.

Tests of Hypotheses

Hypotheses	Assumptions	Critical Region
$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$	$N(\mu, \sigma^2)$ or n large, σ^2 known	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_\alpha$
$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$	$N(\mu, \sigma^2)$ σ^2 unknown	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} \geq t_\alpha(n-1)$
$H_0: \mu_X - \mu_Y = 0$ $H_1: \mu_X - \mu_Y > 0$	$N(\mu_X, \sigma_X^2)$ $N(\mu_Y, \sigma_Y^2)$ σ_X^2, σ_Y^2 known	$z = \dfrac{\bar{x} - \bar{y} - 0}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} \geq z_\alpha$
$H_0: \mu_X - \mu_Y = 0$ $H_1: \mu_X - \mu_Y > 0$	Variances unknown, large samples	$z = \dfrac{\bar{x} - \bar{y} - 0}{\sqrt{s_x^2/n + s_y^2/m}} \geq z_\alpha$
$H_0: \mu_X - \mu_Y = 0$ $H_1: \mu_X - \mu_Y > 0$	$N(\mu_X, \sigma_X^2)$ $N(\mu_Y, \sigma_Y^2)$ $\sigma_X^2 = \sigma_Y^2$, unknown	$t = \dfrac{\bar{x} - \bar{y} - 0}{s_p\sqrt{1/n + 1/m}} \geq t_\alpha(n+m-2)$ $s_p = \sqrt{\dfrac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}$
$H_0: \mu_D = \mu_X - \mu_Y = 0$ $H_1: \mu_D = \mu_X - \mu_Y > 0$	X and Y normal, but dependent	$t = \dfrac{\bar{d} - 0}{s_d/\sqrt{n}} \geq t_\alpha(n-1)$
$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$	$N(\mu, \sigma^2)$	$\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2} \geq \chi_\alpha^2(n-1)$
$H_0: \sigma_X^2/\sigma_Y^2 = 1$ $H_1: \sigma_X^2/\sigma_Y^2 > 1$	$N(\mu_X, \sigma_X^2)$ $N(\mu_Y, \sigma_Y^2)$	$F = \dfrac{s_x^2}{s_y^2} \geq F_\alpha(n-1, m-1)$
$H_0: p = p_0$ $H_1: p > p_0$	$b(n, p)$ n is large	$z = \dfrac{y/n - p_0}{\sqrt{p_0(1-p_0)/n}} \geq z_\alpha$
$H_0: p_1 - p_2 = 0$ $H_1: p_1 - p_2 > 0$	$b(n_1, p_1)$ $b(n_2, p_2)$	$z = \dfrac{y_1/n_1 - y_2/n_2 - 0}{\sqrt{\left(\dfrac{y_1 + y_2}{n_1 + n_2}\right)\left(1 - \dfrac{y_1 + y_2}{n_1 + n_2}\right)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \geq z_\alpha$

Estimation:

The basic idea here is to use random samples (i.e iid samples) to infer parameters/information about the underlying population. So we have

$\{X_i\}_{i=1}^n$: iid samples

and we define a statistic (a function of the data)

$$\hat{\theta}_n = g(X_1, X_2, \cdots, X_n).$$

Suppose θ is an parameter of the population distribution. The goal is to specify statistics $g(\cdot)$ that yield "optimal" estimators $\hat{\theta}_n$ for the estimand θ.

There are 3 basic tools for judging estimator "quality":

(a) Bias:

$$B(\hat{\theta}_n; \theta) = |E[\hat{\theta}_n] - \theta|$$

if $B(\hat{\theta}_n; \theta) = 0$ we say $\hat{\theta}_n$ is an unbiased estimate for θ.

If $B(\hat{\theta}_n; \theta) \neq 0$ but $B(\hat{\theta}_n; \theta) \to 0$ as $n \to \infty$ then $\hat{\theta}_n$ is asymptotically unbiased.

Else $\hat{\theta}_n$ is biased.

(b) Mean Squared-Error (MSE($\hat{\theta}_n, \theta$))

$$MSE(\hat{\theta}_n; \theta) = E[(\hat{\theta}_n - \theta)^2]$$

$$= Var(\hat{\theta}_n) + Bias^2(\hat{\theta}_n; \theta)$$

[Variance - Bias decomposition for MSE]

(c) Consistency:

$\hat{\theta}_n$ is a consistent estimate for θ

$\iff$ $\hat{\theta}_n \xrightarrow{P} \theta$

$\iff$ $\forall \varepsilon > 0 \quad \lim_{n \to \infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0$

Maximum Likelihood Estimation (MLE)

Assuming iid $\{X_i\}_{i=1}^n$ with common pdf $f(x;\theta)$ parametrized by an unknown variable θ. The goal of MLE is to use $\{X_i\}_{i=1}^n$ to estimate θ. The MLE approach is to find the value of θ that maximizes the probability of observing the sample $\vec{x} = \{x_i\}_{i=1}^n$. This requires the idea of a Likelihood $L(\theta) = f_{\vec{X}}(\vec{x}; \theta)$. This is the pdf of $\vec{X}$ considered as a function of θ instead of $\vec{X}$.

$$\hat{\theta}^{MLE} = \underset{\theta}{\operatorname{argmax}} \{ f_{\vec{x}}(\vec{x};\theta) \}$$

$\vec{X} = \{X_i\}_{i=1}^{n}$ iid

$$\Leftrightarrow \quad \hat{\theta}^{MLE} = \underset{\theta}{\operatorname{argmax}} \left\{ \prod_{i=1}^{n} f_x(x_i;\theta) \right\}$$

$$\hat{\theta}^{MLE} = \underset{\theta}{\operatorname{argmax}} \{ \prod_{i=1}^{n} L(\theta) \}$$

The log-likelihood ($L(\theta)$) is often easier to optimize

$$L(\theta;x) = \ln f_x(x;\theta) = \ln L(\theta)$$

i.e $$\hat{\theta}^{MLE} = \underset{\theta}{\operatorname{argmax}} \{ L(\theta;\vec{x}) \}$$

ex: $\{X_i\}_{i=1}^{n}$: iid ; $f_x(x;a) = 2ax \cdot \exp(-ax^2)$; $x \in \mathbb{R}^+$

$$\Rightarrow \quad L(a;\vec{x}) = \log\left(\prod_{i}^{n} f_x(x_i;a) \right) = \sum_{i=1}^{n} \ln\left(2ax_i \cdot \exp(-ax_i^2) \right)$$

$$L(a;\vec{x}) = \sum_{i}^{n} [\ln 2 + \ln a + \ln x_i - ax_i^2]$$

$$\frac{\partial L(a;\vec{x})}{\partial a} = \sum_{i=1}^{n} \frac{1}{a} - x_i^2 = \frac{n}{a} - \sum_{i=1}^{n} x_i^2$$

$$L'(a;\vec{x}) = 0$$

$$\Rightarrow \quad n / \hat{a}^{MLE} = \sum_{i=1}^{n} x_i^2$$

$$\therefore \quad \hat{a}^{MLE} = \frac{n}{\sum_{i=1}^{n} x_i^2}$$

The general MLE procedure for data $\{X_i\}_{i=1}^n$:

i/ Formulate $L(\theta; \vec{X})$:

$$L(\theta; \vec{X}) = \ln f_{\vec{X}}(\vec{X}; \theta)$$

ii/ Solve for critical point of $L(\theta; \vec{X})$:

find $\hat{\theta}^{MLE}$ such that $\partial L / \partial \theta \Big|_{\theta = \hat{\theta}^{MLE}} = 0$

iii/ Make sure $\hat{\theta}^{MLE}$ is a maximum crit. pt. for $L(\theta; \vec{X})$

i.e. $\partial^2 L / \partial \theta^2 \Big|_{\theta = \hat{\theta}^{MLE}} < 0$

ML Estimates have 3 properties:

i/ $\hat{\theta}^{MLE}$ is <u>consistent</u>:

$$\hat{\theta}^{MLE} \xrightarrow{p} \theta$$

ii/ $\hat{\theta}^{MLE}$ obey the <u>Invariance Principle</u> (IP):

$$\boxed{g(\theta)^{MLE} = g(\hat{\theta}^{MLE})}$$

iii/ $\hat{\theta}^{MLE}$ is <u>asymptotically normal</u>:

$$\boxed{\hat{\theta}^{MLE} \xrightarrow{d} N\left(\theta, \frac{1}{J_n(\theta)}\right)}$$

$J_n(\theta)$: Fisher Information for θ.

Ex: $f(x;\theta) = \theta x^{\theta-1}$ $\quad x \in (0,1)$ $\quad \theta \in \mathbb{R}^+$

$\{X_i\}_{i=1}^n$: iid $\quad$ (a) $\hat{\theta}^{MLE}$ $\quad$ (b) $E[\hat{\theta}^{MLE}]$

$$\Rightarrow L(\theta;\vec{x}) = \ln\left(\prod_{i=1}^{n} \theta x_i^{\theta-1}\right)$$

$$= \sum_{i=1}^{n} \left[\ln\theta + (\theta-1)\ln x_i\right]$$

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

$$\partial L/\partial\theta = n/\theta + \sum_{i}^{n} \ln x_i$$

$$\partial L/\partial\theta\Big|_{\theta=\hat{\theta}} = 0 \Rightarrow n/\hat{\theta} = -\sum_{i=1}^{n} \ln x_i$$

$$\Rightarrow \boxed{\hat{\theta} = \frac{-n}{\sum_{i=1}^{n} \ln x_i}}$$

In more general case:

$\vec{\theta} = \{\theta_k\}_{k=1}^{p}$

i.e $\nabla_{\vec{\theta}} L(\vec{\theta})\Big|_{\vec{\theta}=\hat{\vec{\theta}}} = \vec{0}$

$$\hat{\vec{\theta}}^{MLE} = \arg\max_{\vec{\theta}} \left\{L(\vec{\theta};\vec{x})\right\}$$

e.g. $X_i \sim N(\mu, \sigma^2)$

$\vec{\theta} = (\mu, \sigma^2)^T$

$\Rightarrow \hat{\vec{\theta}}_n^{MLE} = \begin{pmatrix} \bar{X}_n \\ \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n) \end{pmatrix}$

We can view the log-likelihood as a statistic:

$$L(\theta; \vec{X}) = \sum_{i=1}^{n} \ln f_X(X_i; \theta)$$

How much does each extra sample of X improve the average precision of estimators for θ?

[assume $\hat{\theta}_n$ is unbiased]

The "Fisher Information", quantifies this idea:

$$J_n(\theta) = E_{X|\theta}\left[\left(\frac{\partial \ln f(\vec{X}|\theta)}{\partial \theta}\right)^2\right] = E_{X|\theta}\left[\frac{-\partial^2 \ln f(\vec{X}|\theta)}{\partial \theta^2}\right]$$

under "regularity conditions"

[Precision of estimator $\hat{\theta}_n \propto 1/\mathrm{Var}(\hat{\theta}_n)$]

The Cramer-Rao Lower Bound:

$$\mathrm{Var}(\hat{\theta}_n) \geq \frac{1}{J_n(\theta)}$$

establishes a lower bound for the variance of any unbiased estimator $\hat{\theta}_n$.

$J_n(\theta)$ quantifies how much "information" X has about θ on average. Any unbiased $\hat{\theta}_n$ cannot "squeeze out" $> J_n(\theta)$ worth of info out of X on average.

Maximum (MAP) A Posteriori Estimation:

Make the "Bayesian Assumption" that the unknown θ is also a rv with pdf $h(\theta)$ (prior). Then the best estimate for θ is one that maximizes the posterior pdf $f(\theta|x)$. This is the MAP Estimate $\hat{\theta}^{MAP}$.

$$\boxed{\hat{\theta}^{MAP} = \arg\max_{\theta} \left(f(\theta|x)\right)}$$

$$\Leftrightarrow \boxed{\hat{\theta}^{MAP} = \arg\max_{\theta} \left(g(x|\theta) \cdot h(\theta)\right)}$$

[By Bayes theorem]

i/ Recall that $\hat{\theta}^{MAP}$ is a Bayes-optimal estimate for the loss function $L(d;\theta) = 1 - \delta(\theta - d)$.

ii/ $\hat{\theta}^{MAP} = \hat{\theta}^{MLE}$ when $h(\theta) = \text{Constant } \forall \theta$.

iii/ The prior pdf $h(\theta)$ encodes prior/subjective info about θ.

MLE/MAP rely heavily on the parametric form of $g(x|\theta)$. In practice, $g(x|\theta)$ might be difficult or too complicated to analyze. This often happens when X is a complicated/corrupted version of a simple rv Z.

ie $X = T(Z)$

where $T(\cdot)$ is a corruption process.

The Expectation-Maximization (EM) algorithm is a method for handling such complicated MLE problems. The approach is to replace the complicated $\ln g(x|\theta)$ with a simpler surrogate:

$$Q(\theta|\theta_t) = E\left[\ln f(z|\theta) \,|\, X, \theta_t\right]$$

This is intended as a simpler $L(\theta)$ for X.

EM is traditionally done in 2 steps:

repeat until $\{\theta_t\}$ converges

E-step: $Q(\theta|\theta_t) = E_{z|x,\theta_t}\left[\ln f(z|\theta) | X, \theta_t\right]$

M-step: $\theta_{t+1} = \operatorname{argmax}\left[Q(\theta|\theta_t)\right]$

The procedure is guarateed to converge to locally optimal estimates for θ.

All estimation methods so far have assumed a pdf family $f(x|\theta)$ for the observed data. Suppose we want to match observed data $\{x_i\}_{i=1}^n$ to a pdf but we have no reason to pick any one pdf family over another. Then choose a pdf family with the maximum Entropy or highest disorder.

Entropy:

The entropy of X, $H(X)$, is the average information associated with X, $I(X) = \log\left(\frac{1}{P(X=x)}\right)$

$$H(X) = E_X\left[\log \frac{1}{P(X=x)}\right] \quad [X \text{ discrete}]$$

$$= \sum_k -\log(P(X=k)) \cdot P(X=k) \quad [\text{discrete}]$$

$$= -\int \log f_X(x) \cdot f_X(x)\, dx \quad [\text{continuous}]$$

e.g 1 $X \sim$ Bernoulli(p)

$\Rightarrow H(X) = -p\log p - (1-p)\log(1-p)$ //

e.g 2 $X \sim$ Uniform$(1, \cdots, n)$

$\Rightarrow H(X) = \sum_k -\frac{1}{n}\log(\frac{1}{n}) = -\frac{n}{n}\cdot\log(\frac{1}{n}) = -\log(\frac{1}{n}) = \log(n)$ //

$\log(y)$ concave $\Rightarrow E_X[\log(1/P(X))] \leq \log[E_X[1/P(X)]]$ (Jensen's)

$\therefore -H(X) \geq -\log(n) \iff \boxed{H(X) \leq \log(n)}$

i.e $\boxed{H(X) \leq H(U)}$ if $X, U \in \{1, \cdots, n\}$, $U \sim$ Uniform

Also:

i/ Conditional Entropy:

$$H(Y|X) = E_{Y|X}[-\ln P(Y|X)]$$

[Thm: $H(Y) \geq H(Y|X)$]

ii/ Mutual Information:

$$I(X,Y) = \sum_X \sum_Y P(x,y) \log \frac{P(x,y)}{P(x)\cdot P(y)}$$

$$I(X,Y) = E_{XY}\left[\log\left(\frac{P(X,Y)}{P(X)\cdot P(Y)}\right)\right]$$

$[I(X,Y) = 0] \iff [X, Y \text{ independent}]$

iii/ Joint Entropy:

$$H(X,Y) = E_{XY}[-\log P(X,Y)]$$

$$I(X,Y) = H(X) - H(X|Y)$$

$$H(X,Y) = H(X) + H(Y|X)$$

(MaxEnt)

The maximum entropy principle says match the data to the ~~pdf~~ with the maximum entropy, subject to constraints on moments. i.e. find pdf of $\{x_i\}_{i=1}^n$ such that:

i/ $\left\{ E[X^k] = \frac{1}{n} \sum_{i=1}^{n} x_i^k \right\}_{k=0}^{m}$ ← match m moments.

ii/ $f_X(x)$ has the max entropy for X.

We already saw that $U \sim U(1, \cdots, n)$ has the MaxEnt on $\{1, \cdots, n\}$ without any moment constraints ($m=0$)

$m=1 \Rightarrow \left\{ \begin{matrix} X \sim \exp \\ \text{or } X \sim \text{geometric} \end{matrix} \right\}$; $m=2 \Rightarrow X \sim N(\cdot, \cdot)$

Dropping the MaxEnt constraint (ii/) reduces this to the historical <u>method of moments</u>:

e.g. : $\{X_i\}$: iid $\sim \gamma(\alpha, \theta)$

find (α, θ) such that :

$$E[X] = \alpha\theta = \frac{1}{n} \sum X_i$$

$$E[X^2] = \alpha\theta^2 + \alpha^2\theta^2 = \frac{1}{n} \sum X_i^2$$

Minimum Mean-Squared Error Estimation:

A different kind of Estimation problem:

Find an estimate for X using Y : $g(Y)$

such that $g(Y)$ minimizes the average mean-square estimation error $E[(X - g(Y))^2]$

$$\Rightarrow \quad \boxed{\hat{X} = g(Y) = E[X|Y]} \qquad [\text{MMSE}]$$

This means that the estimation error is orthogonal to Y [no further "information" of value in Y about X]

$$E[(g(Y) - X) \cdot Y] = 0$$

If we constrain $g(\cdot)$ to be (also) linear we get the Linear MMSE (LMMSE)

$$(\text{MMSE} = \text{LMMSE}) \Leftrightarrow (X, Y \text{ are jointly Gaussian})$$

$\vec{X}, \vec{Y}$ JG

$$\Rightarrow \boxed{E[\vec{X}|\vec{Y}] = \vec{\mu}_X + \underline{K}_{XY}\,\underline{K}_{YY}^{-1}(\vec{Y} - \vec{\mu}_Y)} \qquad [\text{multi-dim}]$$

$$\boxed{\mu_X + \rho_{XY}\,\sigma_X/\sigma_Y\,(Y - \mu_Y)} \qquad [\underline{1}\text{-dim}]$$

$$\underline{K}_{X|Y} = \underline{K}_{XX} - \underline{K}_{XY}\,\underline{K}_{YY}^{-1}\,\underline{K}_{YX} \; ; \quad \sigma_{X|Y}^2 = \sigma_X^2(1 - \rho_{XY}^2)$$