Recap:


Methods of finding estimators

(i) Method of moments
(ii) Method of maximum likelihood estimation
(iii) Bayes and minimax estimators

Method of moments

Sometimes we do not know how to construct estimators for a bunch of parameters and we need to depend on intuitions. In that case, creating a sample analogue of the population parameters might work as an estimator.

In general we create the method of moment estimator as following.

Let \( m_j = \frac{1}{n} \sum_{i=1}^{n} x_i^j \),

also assume \( m_j = E(x_j) \)

Generally \( m_j \)'s are functions of unknown parameters \( \theta_1, \ldots, \theta_k \). We find the method of moment estimators (MOM) of \( \theta_1, \ldots, \theta_k \).
by solving $k$ equations

$$m_j = \mu_j(\theta_1, \ldots, \theta_k), \quad j = 1, \ldots, k.$$ 

**Example:** $x_1, \ldots, x_n \overset{iid}{\sim} N(\mu, \sigma^2)$

we will use, $m_1 = E[x]$

$$m_2 = E[x^2]$$

$$\Rightarrow \quad \frac{1}{n} \sum_{i=1}^{n} x_i = E[x] \quad \cdots \quad (1)$$

$$\Rightarrow \quad \frac{1}{n} \sum_{i=1}^{n} x_i^2 = E[x^2] \quad \cdots \quad (2)$$

$$\Rightarrow \quad \frac{1}{n} \sum_{i=1}^{n} x_i = \mu$$

$$\frac{1}{n} \sum_{i=1}^{n} x_i^2 = \mu^2 + \sigma^2 \quad \Rightarrow \quad \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} x_i\right)^2$$

In this case, MOM estimators are functions of complete, sufficient statistics. However, it is not the case in many distributions.

**e.g.:** $x_1, \ldots, x_n \overset{iid}{\sim} U(\mu-2, \mu+2)$

$$\frac{1}{n} \sum_{i=1}^{n} x_i = E[x] = \mu$$

$$\Rightarrow \text{MOM estimator of } \mu \text{ in } \bar{x}.$$

Here minimal sufficient statistic in a function of $X(n), X(n)$... Hence MOM estimator is not
Another problem of the MOM estimator is that it might not provide estimators in the support of the parameter space.

Example: \( X_1, \ldots, X_n \sim \text{Bin}(k, p) \), \( k \) and \( p \) are both unknown.

\[
\frac{1}{n} \sum_{i=1}^{n} x_i = E[X] \quad \Rightarrow \quad \frac{1}{n} \sum_{i=1}^{n} x_i = kp
\]

\[
\frac{1}{n} \sum_{i=1}^{n} x_i^2 = E[X^2] \quad \Rightarrow \quad \frac{1}{n} \sum_{i=1}^{n} x_i^2 = kp(1-p) + kp^2
\]

\( \Rightarrow \quad p = \frac{\bar{x}}{k} \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} x_i^2 = kp - kp^2 + kp^2
\]

\[
= \bar{x} \left( 1 + kp(\bar{x} - k) \right)
\]

\( \Rightarrow \quad \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x} = k \left( \frac{\bar{x}}{k} \right) \left( \bar{x} - k \right)
\]

\( \Rightarrow \quad \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x} = \frac{\bar{x}^2 \left( \bar{x} - k \right)}{k}
\]

\( \Rightarrow \quad k \left[ \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x} \right] = \bar{x} \left( \bar{x} - k \right)
\]

\( \Rightarrow \quad k = \frac{\bar{x}^2}{\bar{x} - \frac{1}{n} \sum_{i=1}^{n} x_i^2 + \bar{x}}
\]

The R.H.S. may not be an integer, while L.H.S. is an integer. Hence MOM estimator is not even a feasible estimator here.
Example: \( x_1, \ldots, x_n \sim \mathcal{N}(0, \theta^2) , \theta > 0 \)

\[
\frac{1}{n} \sum_{i=1}^{n} x_i = \theta
\]

and \( \frac{1}{n} \sum_{i=1}^{n} x_i^2 = E[x^2] = \theta^2 + \theta^2 = 2\theta^2 \)

\[\Rightarrow \theta = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{and} \quad \theta = \sqrt{\frac{1}{2n} \sum_{i=1}^{n} x_i^2}\]

Maximum likelihood estimator

Suppose \( f_{\theta_1}, \ldots, f_{\theta_K}(x) \) be the p.d.f. of \( x \). Then the likelihood of \( x = (x_1, \ldots, x_n) \) is given by

\[L(\theta_1, \ldots, \theta_K) = \prod_{i=1}^{n} f_{\theta_i}(x_i)\]

Maximum likelihood estimator (MLE) is the value of \( \hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_K) \) that maximizes the likelihood \( L(\theta_1, \ldots, \theta_K) \). MLE is found by solving

\[\frac{\partial L(\theta)}{\partial \theta_i} = 0 , \quad i = 1, \ldots, K\]

Sometimes you use

\[\frac{\partial \log L(\theta)}{\partial \theta_i} = 0 , \quad i = 1, \ldots, K\]

to find the MLE.

Note that \( \frac{\partial \log L(\theta)}{\partial \theta_i} = 0 \Rightarrow \frac{1}{L(\theta)} \frac{\partial L(\theta)}{\partial \theta_i} = 0 \)

\[\Rightarrow \frac{\partial L(\theta)}{\partial \theta_i} = 0 \quad \text{. Thus they are equivalent.}\]

\[\frac{\partial}{\partial \theta_i} \log L(\theta) = 0 \Rightarrow \frac{\partial}{\partial \theta_i} \left[ \frac{1}{n} \sum_{j=1}^{n} \log f_{\theta_1, \ldots, \theta_K}(x_j) \right] = 0 \]

\[\Rightarrow \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta_i} \log f_{\theta_1, \ldots, \theta_K}(x_j) = 0 , \quad i = 1, \ldots, K.\]
Note that \( \frac{\partial \log f_{\theta_i}(x)}{\partial \theta_i} \) for \( i = 1, \ldots, k \) are components of the score vector.

In the one-dimensional case where \( k = 1 \), the equations boil down to

\[
\sum_{j=1}^{n} \frac{\partial \log f_\theta(x_j)}{\partial \theta} = 0 \quad \Rightarrow \quad \sum_{j=1}^{n} u_\theta(x_j) = 0
\]

Let \( \hat{\theta}_1, \ldots, \hat{\theta}_k \) be the solutions of the system of equations to find the MLE. The MLE should also satisfy that \( \frac{\partial^2 L(\theta)}{\partial \theta^2} \bigg|_{\theta = \hat{\theta}} < 0 \)

MLE sometimes lies in the boundary of the parameter space.

**Example:** \( X_1, \ldots, X_n \sim N(\theta, 1), \theta > 0 \)

Find the MLE of \( \theta \).

When \( \theta > 0 \) restriction is not there, MLE of \( \theta \) is given by \( \bar{X} \). However, when we have this additional restriction we cannot use \( \bar{X} \) as the MLE.

If \( \bar{X} > 0 \), then obviously \( \bar{X} \) is going to be the MLE.
If $\bar{x} \leq 0$ then $\bar{x}$ is not the MLE.

$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(x_i - \bar{x})^2}{2} \right\}

= \left( \frac{1}{\sqrt{2\pi}} \right)^n \exp \left\{ -\frac{1}{2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right\}

= \left( \frac{1}{\sqrt{2\pi}} \right)^n \exp \left\{ -\frac{1}{2} \left[ \sum_{i=1}^{n} (x_i - \bar{x})^2 + (\bar{x} - \theta)^2 + 2(\bar{x} - \bar{z})(\bar{z} - \theta) \right] \right\}

= \left( \frac{1}{\sqrt{2\pi}} \right)^n \exp \left\{ -\frac{1}{2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right\} \exp \left\{ -\frac{n}{2} (\bar{x} - \theta)^2 \right\}

Note that $\bar{x} \leq 0$ so, let $\bar{x} = -z$, $z \geq 0$

$L(\theta) \propto \exp \left\{ -\frac{n}{2} (z - \theta)^2 \right\}

= \exp \left\{ -\frac{n}{2} (z + \theta)^2 \right\}

Note that the range of $\theta$ is $\theta > 0$.

As $\theta$ increases the function $L(\theta)$ monotonically decreases.

the maximum is achieved when $\theta = 0$

which is outside the range of $\theta$.

thus no MLE exists in the case where

$\bar{x} \leq 0$. 

(6)
However if we begin with the distribution
\( X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1), \theta \geq 0 \)
Then when \( \bar{X} \geq 0 \), \( \theta = \bar{X} \) maximizes the likelihood.
When \( \bar{X} < 0 \), the likelihood is maximized at \( \theta = 0 \).

MLE of \( \theta \) in this problem is \( \max \{ 0, \bar{X} \} \)

Remark: MLE does not need to be a function of a minimal sufficient statistic.

\( X_1, \ldots, X_n \sim \mathcal{U}(\theta-1, \theta+1) \)

\[ L(\theta) = \frac{1}{2^n} I[ X(n) - 1 < \theta < X(1) + 1 ] \]

Any value of \( \theta \in (X(n) - 1, X(1) + 1) \) maximizes the likelihood. Hence MLE is not unique.

Also,
\[ \hat{\theta} = \left( \frac{\bar{X}}{\bar{X} + 1} \right) (X(n) - 1) + \frac{1}{(\bar{X} + 1)} (X(1) + 1) \]

\( \hat{\theta} \in (X(n) - 1, X(1) + 1) \) being a weighted average of \( X(n) - 1 \) and \( X(1) + 1 \).

Thus \( \hat{\theta} \) is an MLE, but it is not a function of a minimal sufficient statistic.
Invariance property of MLE

If \( \hat{\theta} \) is an MLE of \( \theta \), \( \hat{\tau}(\hat{\theta}) \) is an MLE of \( \tau(\theta) \), for any function \( \tau(\theta) \).
Recall:
1. Method of moments estimator
2. Maximum likelihood estimator

Invariance property of MLE:
If $\hat{\theta}$ in the MLE of $\theta$ then $\mathbb{L}(\hat{\theta})$ in the MLE of $\mathbb{L}(\theta)$.

Asymptotic distribution of MLE:
In smooth regular problems MLE $\hat{\theta}$ has the property that $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \frac{1}{I(\theta)})$

If $g$ is any function of $\theta$ s.t. $g'(\theta) \neq 0$.

$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, \frac{[g'(\theta)]^2}{I(\theta)})$ (By Delta method)

when $n$ is very large.

the distribution of $\sqrt{n}(g(\hat{\theta}) - g(\theta))$ can be approximated by $\mathcal{N}(0, \frac{[g'(\theta)]^2}{I(\theta)})$

$E[\sqrt{n}(g(\hat{\theta}) - g(\theta))] = 0 \Rightarrow E[g(\hat{\theta})] = g(\theta)$

and $\text{Var}(\sqrt{n}(g(\hat{\theta}) - g(\theta))) \approx \frac{[g'(\theta)]^2}{I(\theta)}$

$\Rightarrow \text{Var}(g(\hat{\theta})) \approx \frac{[g'(\theta)]^2}{n I(\theta)}$

thus $g(\hat{\theta})$ is an asymptotically unbiased estimator of $g(\theta)$ whose variance asymptotically achieves the Cramer-Rao lower bound.
Asymptotic distribution of MLE is Normal when the regularity conditions hold.

It is not always true.

Example: \( X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{U}(0, \theta), \theta > 0 \)

\[
\theta \left( x_{(n)} \right) = \frac{1}{\theta^n} \mathbf{1}(x_{(n)} \leq \theta)
\]

If \( \theta > x_{(n)} \) then \( \frac{1}{\theta} \mathbf{1} < \frac{1}{X_{(n)}} \)

Here the MLE is \( X_{(n)} \).

Now, \( \sqrt{n} \left( X_{(n)} - \theta \right) \overset{d}{\rightarrow} \mathcal{N}(0, \frac{1}{\theta^2}) \)

\( \sqrt{n} \left( X_{(n)} - \theta \right) \overset{d}{\rightarrow} \mathcal{E}(\theta) \)

Bayes and Minimax estimator

\( X_1, \ldots, X_n \overset{iid}{\sim} f(x|\theta) \)

Assume a prior distribution \( \pi(\theta) \),

Suppose the prior dist. is \( \pi(\theta) \). There are objective and subjective choices of priors.

Posterior dist., denoted by \( \pi(\theta|x_1, \ldots, x_n) \),

\[
\pi(\theta|x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n|\theta) \pi(\theta)}{\int f(x_1, \ldots, x_n|\theta) \pi(\theta) \, d\theta}
\]
Example: \( X_1, \ldots, X_n \overset{iid}{\sim} \text{Beta}(p) \), \( p \sim \text{Beta}(\alpha, \beta) \)

\[
\pi(p | x_1, \ldots, x_n) \propto \left[ \frac{n}{\pi} \ p^{x_i} (1-p)^{1-x_i} \right] \ p^{\alpha-1} (1-p)^{\beta-1} \\
\propto p^{\sum_{i=1}^n x_i + \alpha - 1} (1-p)^{n - \sum_{i=1}^n x_i + \beta - 1} \\
\text{Thus } p | x_1, \ldots, x_n \sim \text{Beta} \left( \sum_{i=1}^n x_i + \alpha, \ n - \frac{n}{\pi} \sum_{i=1}^n x_i + \beta \right)
\]

Example: \( X_1, \ldots, X_n \overset{iid}{\sim} \text{Poi}(\lambda) \), \( \lambda \sim \text{Gamma}(\alpha, \beta) \)

\[
\pi(\lambda | x_1, \ldots, x_n) \propto \left[ \frac{n}{\lambda} \ e^{-\lambda} \lambda^{x_i} \right] \ e^{-\lambda/\beta} \ \lambda^{\alpha-1} \\
\propto e^{-\lambda(n+1)} \ \lambda^{\sum_{i=1}^n x_i + \alpha - 1} \\
\lambda | x_1, \ldots, x_n \sim \text{Gamma} \left( \sum_{i=1}^n x_i + \alpha, \ \frac{1}{n+1+\beta} \right)
\]

Definition (Conjugate Family)

Let \( \mathcal{F} \) denote the class of p.r.d.f.s on p.r.f.s \( f(x | \theta) \). A class \( \mathcal{F} \) of prior distributions is a conjugate family for \( F \) if the posterior distribution is in the class \( \mathcal{F} \) for all \( f \in \mathcal{F} \) and all priors in \( \mathcal{F} \).

Now we will learn how to construct good estimators from Bayesian statistics. We will be using the "risk function" notion in Bayesian statistics.
Let $S(x)$ be an estimator of $\theta$. The loss in estimating $\theta$ by $S(x)$ is represented by a function known as the loss function. Mathematically, we write it as $L(\theta, S)$. Define $R(\theta, S) = E_{x} \{L(\theta, S(x))\}$ as the risk function of $S$.

When $L(\theta, S(x)) = (S(x) - \theta)^2$, then $R(\theta, S(x)) = E_{x} \{ (S(x) - \theta)^2 \}$ is the MSE of $S(x)$.

Given two estimators $S_1, S_2$, we say $S_1$ is better than $S_2$ if $R(\theta, S_1) \leq R(\theta, S_2)$ for all $\theta$ and the inequality is strict for at least one $\theta$.

\[ \text{risk} \quad \frac{S_2}{\quad S_1} \quad \text{risk} \]

\[ \frac{\quad S_2}{\text{risk}} \quad \frac{S_1}{\text{risk}} \quad \frac{\quad S_1}{\text{risk}} \]

When two risks intersect each other, we do not know which estimator to choose. At that time, we have to propose a summary measure from the entire risk function.

(4)
Different summary measures have been proposed. We will discuss the following two.

1. Average risk = $E_\theta \exp\left[ l(\theta, s(x)) \right]$. Minimizing this average risk will lead to the Bayes estimator.

2. Supremum risk = $\sup E_{\theta} \exp\left[ l(\theta, s(x)) \right]$. Minimizing the supremum risk will give rise to the Minimax estimator.

Bayes estimator

\[ E_\theta \exp\left[ l(\theta, s(x)) \right] = E_x s(x) \exp\left[ l(\theta, s(x)) \right] = E_x E_\theta | x \left[ l(\theta, s(x)) \right] \]

we have to minimize this average risk w.r.t. $s(x)$. We will find $s(x)$ that minimizes

\[ E_\theta | x \left[ l(\theta, s(x)) \right] \]

for all $x$. There are a few loss functions where it can be done easily.

Squared error loss:

Let $l(\theta, s(x)) = (s(x) - \theta)^2$

\[ E_\theta | x \left[ (s(x) - \theta)^2 \right] = E_\theta | x \left[ (s(x) - E[\theta | x] + E[\theta | x] - \theta)^2 \right] \]

\[ = E_\theta | x \left[ (s(x) - E[\theta | x])^2 \right] + \underbrace{E_\theta | x \left[ (\theta - E[\theta | x])^2 \right]}_{\text{constant with respect to } s(x)} \]

5
\[ E_0 \mid x \left[ (s(x) - \theta)^2 \right] \geq E_0 \mid x \left[ (\infty \cdot E[\theta \mid x] - \theta)^2 \right] \]

Thus \( s(x) = E[\theta \mid x] \) is going to be the Bayes estimator.

\[ E_0 \mid x \left[ L(\theta, s(x)) \right] \] you want to minimize w.r.t. \( s(x) \)

\[ \frac{d}{ds} E_0 \mid x \left[ L(\theta, s) \right] = 0 \]

\[ \frac{d}{ds} E_0 \mid x \left[ (\theta - s)^2 \right] = 0 \]

\[ \Rightarrow E_0 \mid x \left[ 2(s - \theta) \right] = 0 \]

\[ \Rightarrow E_0 \mid x \left[ s \right] = E_0 \mid x \left[ \theta \right] \]

\[ \Rightarrow s(x) = E_0 \mid x \left[ \theta \right] \]

**Weighted squared error loss**

\[ L(\theta, s(x)) = w(\theta) (\theta - s(x))^2 \quad w(\theta) > 0 \]

\[ \frac{d}{ds} E_0 \mid x \left[ w(\theta) (\theta - s)^2 \right] = 0 \]

\[ \Rightarrow E_0 \mid x \left[ 2w(\theta) (s - \theta) \right] = 0 \]

\[ \Rightarrow E_0 \mid x \left[ w(\theta) s \right] = E_0 \mid x \left[ \theta \cdot w(\theta) \right] \]

\[ \Rightarrow s(x) = \frac{E_0 \mid x \left[ \theta \cdot w(\theta) \right]}{E_0 \mid x \left[ \theta \cdot w(\theta) \right]} \]
Example: \( x_1, \ldots, x_n \overset{iid}{\sim} \text{Ber}(p), \quad p \overset{\sim}{\sim} \text{Beta}(\alpha, \beta) \)

\[ p | x_1, \ldots, x_n \sim \text{Beta}(\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i) \]

Under squared error loss \( L(\theta, S(x)) = (\theta - S(x))^2 \),

What is the Bayes estimator?

Bayes estimator: \( \mathbb{E}[p | x_1, \ldots, x_n] = \frac{\alpha + \sum_{i=1}^{n} x_i}{\alpha + \sum_{i=1}^{n} x_i + \beta + n} \)

\[ = \frac{\alpha + \sum_{i=1}^{n} x_i}{\alpha + \beta + n} \]

\( \mathbb{E}[\mathbb{E}[p | x_1, \ldots, x_n]] = \frac{\alpha + np}{\alpha + \beta + n} \neq p \)

This is a biased estimator of \( p \).

Theorem: No unbiased estimator \( S(x) \) of \( \theta \)

can be a Bayes estimator under squared error loss unless \( E_\theta E_x [ (\theta - S(x))^2 ] = 0 \)

If \( S(x) \) is an unbiased estimator
which is also a Bayes estimator,

\( E[S(x) | \theta] = \theta \) for all \( \theta \)

and \( E[\theta | x] = S(x) \)

Now,

\( E_\theta E_x [ S(x) | \theta ] = E_\theta [ \theta E_x [ S(x) ] ] = E_\theta [ \theta ] \)

\[ \mathbb{E} \mathbb{E} [ S(x) | \theta ] = \mathbb{E} [ S(x) \mathbb{E} [ S(x) | \theta ] ] = \mathbb{E} [ S(x) ] \]
\[ E_\theta E_{X|\theta}[s(x)^{\gamma}] = E_\theta E_{X|\theta}[s(x) \theta] = E_\theta E_{X|\theta}[^\theta \gamma] \]

Hence,

\[ E_\theta, X[(\theta - s(x))^\gamma] = E_\theta, X[\theta^\gamma - 2s(x)\theta + s(x)^\gamma] \]

By (\ref{eq:example}), \( \theta = 0 \)

Sometimes you can show that the Bayes estimator is a convex combination of the prior mean and data mean.

**Example:** \( X_1, \ldots, X_n \text{iid } \text{Ber}(\theta), \ \theta \sim \text{Beta}(\alpha, \beta) \)

\[ \Rightarrow \quad E[\theta | X_1, \ldots, X_n] = \frac{\sum_i x_i + \alpha}{\alpha + \beta + n} \]

\[ = \frac{\sum_i x_i}{n} \cdot \frac{n}{\text{data mean}} + \frac{\alpha}{\alpha + \beta + n} \cdot \frac{(\alpha + \beta + n)}{\text{prior mean}} \]

\( \omega_1 + \omega_2 = 1 \)

---

**Midterm**

1. \( X_1, \ldots, X_{2n} \text{iid } \text{Ber}(\theta), \ \ S_n = \sum_{i=1}^{n} X_{2i} X_{2i-1} \)

\( \sqrt{n} \left( \frac{S_n}{n} - \theta \right) \quad E\left[ \frac{S_n}{n} \right] = \theta \)

\( \text{Var}\left( \Theta X_{2i} X_{2i-1} \right) = E\left[ X_{2i} X_{2i-1} \right] - \left\{ E\left[ X_{2i} \right] E\left[ X_{2i-1} \right] \right\}^\gamma \]

\[ = \left\{ \left\{ \Theta \right\}^\gamma \text{Beta}(\theta) \left[ p(1-p) + \theta^\gamma \cdot \frac{e^{-\alpha}}{\alpha^\gamma} \right] \right\}^\gamma \]
\[ \text{Var}(X_2; X_{2i-1}) = p^r - p^u = p^r(1-p^r) \]

CLT tells us \( \sqrt{n} \left( \frac{\sum X_i}{n} - p^r \right) \overset{d}{\to} N(0, p^r(1-p^r)) \)

2. \( X = \begin{cases} 
-1 & \text{w.p. } \theta/2 \\
0 & \text{w.p. } 1-\theta \\
1 & \text{w.p. } \theta/2 
\end{cases} \)

\( |X_i| = 1 \) w.p. \( \theta \)

\( = 0 \) w.p. \( 1-\theta \)

\( \Rightarrow \{X_1, \ldots, X_n\} \overset{iid}{\sim} \text{Ber}(\theta). \)

Now the rest follows from classes.

\( \hat{\theta} \) that was \( \frac{\sum |X_i| (\frac{\sum |X_i| - 1}{n(n-1)})}{n(n-1)} \)

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

\( (\bar{X}, \widehat{\sigma^2}) \) in \( \varnothing \) a complete sufficient stat.

\( E[\bar{X}] = \mu \) thus by Lehmann-Scheffé theorem.

\( \bar{X} \) in the UMPU for \( \mu \).

\( \left( \frac{n-1}{\sigma^2} \right) \overset{\sim}{\in} \chi^2_{n-1} \)

\( \Rightarrow \frac{\bar{X}^2}{(n-1)\sigma^2} \overset{\sim}{\in} \text{Inverse-Gamma} \left( \frac{n-1}{2}, 2 \right) \)

\( \Rightarrow E \left[ \frac{\bar{X}^2}{(n-1)\sigma^2} \right] = \frac{1}{2(n-1)} = \frac{1}{n-3} \)

\( \Rightarrow E \left[ \frac{n-3}{(n-1)\sigma^2} \right] = \frac{1}{\bar{X}^2} \text{ this is a function of } \sigma \)
Recap:

1. Risk function: \( E_{\theta} \left[ L(\theta, s(x)) \right] \) - defines an estimator \( s(x) \).
2. \( E_{\theta} E_{x} \left[ L(\theta, s(x)) \right] \rightarrow \) an estimator \( d(x) \) that minimizes this average risk function called the Bayes estimator.
3. \( \sup_{\theta} E_{x} \left[ L(\theta, s(x)) \right] \). The minimax estimator minimizes \( \sup_{\theta} E_{x} \left[ L(\theta, s(x)) \right] \). Thus minimax estimators are going to protect us from the worst case scenarios.

Computing finding minimax estimators is difficult. However, we can use other tricks to compute minimax estimators. If the prior distribution of \( \theta \) gives high probability to those values which will increase the risk, then intuitively a minimax estimator is a Bayes estimator.

If the prior only has mass at \( \theta^* \), then the minimax estimator is:

\[
E_{\theta} E_{x} \left[ L(\theta, s(x)) \right] = \sup_{\theta} E_{x} \left[ L(\theta, s(x)) \right]
\]
Definition (Least Favorable Distribution)

A prior distribution \( \pi(\theta) \) is known to be a least favorable prior if

\[
E_\theta E_{X|\theta} \left[ L(\theta, S_\pi(x)) \right] \\
\geq E_\theta E_{X|\theta} \left[ L(\theta, S_{\pi'}(x)) \right] \quad \text{for all \( \pi' \) prior distributions,} \\
\text{with} \quad S_\pi \quad \text{and} \quad S_{\pi'} \quad \text{are Bayes estimators w.r.t. priors} \ \pi \quad \text{and} \quad \pi' \quad \text{respectively.}
\]

So, \( \pi(\theta) \) is such that if increases the Bayes risk for the Bayes estimator w.r.t. any other prior.

Result: If \( \pi(\theta) \) be a prior distribution for which

\[
\int_{\Theta} E_{X|\theta} \left[ L(\theta, S_\pi) \right] \frac{d\theta}{\pi(\theta)} = \sup_{\theta} E_{X|\theta} \left[ L(\theta, S_\pi) \right]
\]

then when \( S_\pi \) is the Bayes estimator. Then

(a) \( S_\pi \) is minimax estimator

(b) \( \pi \) is least favorable distribution

Proof: a) For any \( g(x) \),

\[
\sup_{\theta} E_{X|\theta} \left[ L(\theta, g(x)) \right] \geq E_\theta E_{X|\theta} \left[ L(\theta, S_\pi(x)) \right] \\
\geq E_\theta E_{X|\theta} \left[ L(\theta, S_{\pi'}(x)) \right] = \sup_{\theta} E_{X|\theta} \left[ L(\theta, S_{\pi'}) \right]
\]

by assumption.

Thus \( S_\pi(x) \) is a minimax estimator.
(b) Note that, by assumption

\[ E_\theta E_\theta^1 \Omega \left[ L(\theta, s_{n\pi}) \right] = \sup_{\theta} E_\theta^1 \Omega \left[ L(\theta, s_{n\pi}) \right] \]

\[ \geq \int E_\theta^1 \Omega \left[ L(\theta, s_{n\pi}) \right] \pi'(\theta) d\theta \] for any other prior dist. \( \pi'(\theta) \) on \( \Theta \)

\[ E_\theta^1 \Omega \left[ L(\theta, s_{n\pi}) \right] \leq \sup_{\theta} E_\theta^1 \Omega \left[ L(\theta, s_{n\pi}) \right] \]

\[ \int E_\theta^1 \Omega \left[ L(\theta, s_{n\pi}) \right] \pi'(\theta) d\theta \leq \sup_{\theta} E_\theta^1 \Omega \left[ L(\theta, s_{n\pi}) \right] \int \pi'(\theta) d\theta \]

\[ \geq \int E_\theta^1 \Omega \left[ L(\theta, s_{n\pi}) \right] \pi'(\theta) d\theta \]

\[ = E_\theta E_\theta^1 \Omega \left[ L(\theta, s_{n\pi}) \right] \]

Thus, \( E_\theta E_\theta^1 \Omega \left[ L(\theta, s_{n\pi}) \right] \geq E_\theta E_\theta^1 \Omega \left[ L(\theta, s_{n\pi}) \right] \)

which implies that \( \pi^*(\theta) \) in the least favourable distribution.

To use this result, we need to have a prior dist. for which

\[ \int E_\theta^1 \Omega \left[ L(\theta, s_{n\pi}) \right] \pi(\theta) d\theta = \sup_{\theta} E_\theta^1 \Omega \left[ L(\theta, s_{n\pi}) \right] \]

then we can find the Bayes estimator w.r.t. that prior. Our result says that this Bayes estimator will be the minimax estimator.
If the risk fn. $E_x \left[ L(\theta, \delta_n) \right]$ is constant, then average is going to be the supremum.

One algorithm:

Step 1. Take a prior dist of your choice.

Step 2. Compute the risk fn. w.r.t. that prior.

Step 3. Try to find prior hyperparameters in such a way that the risk $r_x$ of the Bayes estimator becomes constant.

Example: $x_1, \ldots, x_n \stackrel{iid}{\sim} \text{Ber}(p), \quad p \sim \text{Beta}(\alpha, \beta)$.

$$S_n(x) = \frac{\sum_{i=1}^{n} x_i + \alpha}{n + \alpha + \beta}$$

$$E_x[p \left( (S_n(x) - p)^2 \right)] = E_x[p \left( \left( \frac{\sum_{i=1}^{n} x_i + \alpha}{n + \alpha + \beta} - p \right)^2 \right)]$$

$$= \text{Var}_x[p \left( \frac{\sum_{i=1}^{n} x_i + \alpha}{n + \alpha + \beta} \right) + \left[ E_x[p \left( \frac{\sum_{i=1}^{n} x_i + \alpha}{n + \alpha + \beta} \right) - p \right]^2]$$

$$= \frac{np(1-p)}{(n+\alpha+\beta)^2} + \left[ \frac{np + \alpha}{n + \alpha + \beta} - p \right]^2$$

$$= \frac{np(1-p)}{(n+\alpha+\beta)^2} + \left[ \frac{\alpha - p\alpha - p\beta}{n + \alpha + \beta} \right]^2$$
\[
= \frac{1}{(n+\alpha+\beta)^2} \left[ \alpha^2 + \left( n - 2\alpha (\alpha+\beta) \right) \beta + \left( (\alpha+\beta) - n \right) \beta^2 \right]
\]

This to be a const. (not a function of \( p \)) we must have

\[
n = 2\alpha (\alpha+\beta) \quad \cdots \quad 1
\]

\[
\sqrt{n} = (\alpha+\beta) \quad \cdots \quad 2
\]

Solving them we obtain \( \alpha = \frac{\sqrt{n}}{2}, \beta = \frac{\sqrt{n}}{2} \).

Thus by our earlier thus: \( \text{Beta} (\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}) \)

\[
\Sigma_{i=1}^{n} x_i + \frac{\sqrt{n}}{2}
\]

is a minimax estimator.

---

**Testing of Hypothesis**

**Statistical hypothesis testing is all about**

1. **Beginning** with a tentative idea about the unknown parameter.

2. *Want to test* the validity of this tentative idea based on sample. Existing tentative idea, status quo: \( H_0 \) (null hypothesis), new idea: \( H_1 \) (alternative hypothesis).

3. We begin by assuming that the null hypothesis is true. Only when there is an overwhelming evidence contradicting the null do we reject it in favor of the alternative.
<table>
<thead>
<tr>
<th>$H_0$ true</th>
<th>$H_0$ false</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do not reject $H_0$</td>
<td>Incorrect</td>
</tr>
<tr>
<td>Reject $H_0$</td>
<td>Correct</td>
</tr>
</tbody>
</table>

Type 1 error = $P(\text{reject } H_0 \mid H_0 \text{ true})$

Type 2 error = $P(\text{do not reject } H_0 \mid H_0 \text{ false})$

Type 1 error is also known as the level (size) of the test.

On the other hand, $1 - \text{Type 2 error} = P(\text{Rejecting } H_0 \mid H_0 \text{ false}) = \text{power of the test}$

$0 \leq \text{level} \leq 1$, $0 \leq \text{power} \leq 1$

Ideally, we would like to minimize both type 1 and type 2 error.

But it turns out that it is not possible to simultaneously minimize both. So, we fix level at a pre-specified value and find a test that maximizes power.

**Parametric tests:** $X_1, \ldots, X_n \overset{iid}{\sim} f(x \mid \theta)$. We test $H_0: \theta \in \Theta_0$ vs. $H_1: \theta \in \Theta_1$. If $\Theta_0$ is singleton we will call it a simple null hypothesis. Otherwise we will call it a composite null hypothesis.
Let \( R = \{ x \in \mathcal{X} | \text{the null hypothesis is rejected for } x \} \).

\( R \) is called the rejection region or critical region of the test.

\[ \phi(x) = \text{prob. of rejecting } H_0 \text{ when } x \text{ is observed} \]

Power function of a test is given by

\[ \beta(\theta) = \int \phi(x) f(x | \theta) \, dx \]

\[ = \int P(\text{rejecting } H_0 \mid x \text{ is observed} \mid \theta) f(x | \theta) \, dx \]

Intuitively

\[ = P(\text{reject } H_0 \mid \theta \text{ in the truth}) \]

If level of a test in \( \alpha \) and the power function of that test is given by \( \beta(\theta) \) and \( \beta(\theta) \) is the hypothesis in \( H_0: \theta \in \Theta_0 \), vs. \( H_1: \theta \in \Theta_1 \),

then \( \beta(\theta) \leq \alpha \), \( \Theta \in \Theta_0 \).