(Copied from the lecture notes of last week)

Stein’s lemma. (see Stein (1981, Ann. Stat.) or Stein (1973))
Let \( Y \) be a \( N(0, 1) \) real random variable and \( g : \mathbb{R} \rightarrow \mathbb{R} \) be an indefinite integral of the Lebesgue measurable function \( g' \), essentially the derivative of \( g \). Suppose that \( E|g'(Y)| < \infty \). Then

\[
E(Y g(Y)) = Eg' (Y).
\]

Proof of the lemma: Write

\[
E g'(Y) = \int g'(y) \phi(y) \, dy
\]
\[
= \int_0^\infty g'(y) \int_y^\infty x \phi(x) \, dx \, dy - \int_0^y g'(y) \int_{-\infty}^y x \phi(x) \, dx 
\]
then apply Fubini’s theorem.

Stein’s unbiased estimate of the risk (SURE).
Let \( X \sim N(\theta, \Sigma_{n \times n}) \) with \( \Sigma \) positive definite and \( g(X) \) be absolutely continuous, then

\[
E[(X + g(X) - \theta)^T \Sigma^{-1}(X + g(X) - \theta)] = E[n + g(X)^T \Sigma^{-1}g(X) + 2\nabla \cdot g(X)]
\]
\[
E \|X + g(X) - \theta\|^2 = E[tr(\Sigma) + \|g(X)\|^2 + 2tr(\Sigma \cdot Dg(X))]
\]
where \( \nabla g(X) = \sum_{i=1}^p \partial g_i / \partial X_i \) and \( Dg(X) \) is a \( p \times p \) matrix with \( (Dg)_{ij} = \partial g_i / \partial X_j \).

Homework problem: prove the identities above.

James-Stein estimator.

\[
\delta_{J-S}(X) = \left( 1 - \frac{C\sigma^2}{\|X\|^2} \right) X, \ C > 0.
\]

Theorem. Let \( X \sim N(\theta, \sigma^2 I_n) \). Let \( 0 < C \leq 2(n-2) \) (hence \( n \geq 3 \)). Then

\[
R(\theta, \delta_{J-S}) = E \|\delta_{J-S}(X) - \theta\|^2 \leq n\sigma^2.
\]

Proof of the theorem:

\[
E \|\delta_{J-S}(X) - \theta\|^2 = n\sigma^2 - E \left[ \frac{\sigma^4}{\|X\|^2} C (2(n-2) - C) \right]
\]
Question: a similar result for other losses, e.g. $L(\theta, \delta) = \Sigma^n_{i=1} |\delta_i - \theta_i|$

**Domination of positive-part estimator.**

The following lemma is a generalization of Theorem 6.2 on page 302 of Lehmann (1983) and Theorem 5.4 on page 356 of Lehmann and Casella (1998). It shows that taking the positive part will improve the estimator componentwise. Specifically for an estimator $(\hat{\theta}_1(X), \ldots, \hat{\theta}_d(X))$ where

$$\hat{\theta}_i(X) = (1 - g_i(X))X_i,$$

the positive part estimator of $\hat{\theta}_i(X)$ is denoted as

$$\hat{\theta}_i^+(X) = (1 - g_i(X))_+ X_i.$$

**Lemma.** Assume that $g_i(X)$ is symmetric with respect to the $i$th coordinate, then

$$E(\theta_i - \hat{\theta}_i^+)^2 \leq E(\theta_i - \hat{\theta}_i)^2.$$ 

Furthermore, if

$$P_\theta(g_i(X) > 1) > 0,$$

then

$$E(\theta_i - \hat{\theta}_i^+)^2 < E(\theta_i - \hat{\theta}_i)^2.$$

**Proof of the lemma:** Simple calculation shows that

$$E(\theta_i - \hat{\theta}_i^+)^2 - E(\theta_i - \hat{\theta}_i)^2 = E((\hat{\theta}_i^+)^2 - \hat{\theta}_i^2) - 2\theta_iE(\hat{\theta}_i^+ - \hat{\theta}_i).$$

Let’s calculate the expectation by conditioning on $g_i(X)$. For $g_i(X) \leq 1$, $\hat{\theta}_i^+ = \hat{\theta}_i$. Hence it is sufficient to condition on $g_i(X) = b$ where $b > 1$ and show that

$$E((\hat{\theta}_i^+)^2 - \hat{\theta}_i^2 \mid g_i(X) = b) - 2\theta_iE(\hat{\theta}_i^+ - \hat{\theta}_i \mid g_i(X) = b) \leq 0,$$

or equivalently,

$$-E(\hat{\theta}_i^2 \mid g_i(X) = b) + 2\theta_i(1 - b)E(X_i \mid g_i(X) = b) \leq 0.$$

Obviously, the last inequality is satisfied if we can show

$$\theta_iE(X_i \mid g_i(X) = b) \geq 0.$$

We may further condition on $X_j = X_j$ for $j \neq i$ and it suffices to establish

$$\theta_iE(X_i \mid g_i(X) = b, X_j = x_j, j \neq i) \geq 0.$$ 

Given that $X_i = x_j, j \neq i$, consider only the case where $g_i(X) = b$ has solutions. Due to symmetry of $g_i(X)$, these solutions are in pairs. Let $\pm y_k, k \in K$, denote the solutions. Hence the left hand side of the equation above equals

$$\theta_iE(X_i \mid \pm y_k, k \in K)$$

$$= \sum_{k \in K} \theta_iE(X_i \mid X_i = \pm y_k)P_\theta(X_i = \pm y_k \mid X_i = \pm y_k, k \in K).$$
Note that
\[ \theta_i E(X_i \mid X_i = \pm y_k) = \frac{\theta_i y_k e^{y_k \theta_i} - \theta_i y_k e^{-y_k \theta_i}}{e^{y_k \theta_i} + e^{-y_k \theta_i}}, \]
which is symmetric in \( \theta_i y_k \) and is increasing for \( \theta_i y_k > 0 \). Hence it is bounded below by zero, a bound obtained by substituting \( \theta_i y_k = 0 \).

The strict inequality of the theorem can be established by noting that
\[ E[(\hat{\theta}_i^+) - \hat{\theta}_i^2] \] is strictly negative.

**Question (a long standing problem):** find a Bayes estimator to dominate the positive part of JS estimator.
Lecture 3. The Canonical normal means estimation problem (cont.).

Shrink toward a common mean.

Theorem. Let \( X \sim N (\theta, \sigma^2 I_n) \). Let \( 0 < C \leq 2(n - 3) \) (hence \( n \geq 4 \)).

Define

\[
\delta_+ (X) = \overline{X} + \left( 1 - \frac{C \sigma^2}{||X - \overline{X}||^2} \right) (X - \overline{X}),
\]

Then

\[
R (\theta, \delta_+) < R (\theta, \delta) \leq n \sigma^2.
\]

Homework problem: prove the theorem above (cf. Lindley and Smith (1972, JRSSB)).

An example from Efron and Morris.

Efron and Morris (1975,1977) looked at the batting averages of a sample of 18 baseball players for the 1970 season (Batting averages are the proportion of "base hits" for a player out of his total "at bats").

Explanation for the table:

| \( Y_i \) | \( \frac{1}{45} \text{Bin}(45, \gamma_i) \) : batting average for the first 45 at bats |
| \( P_i \) | \( \frac{1}{n_i} \text{Bin}(n_i, \gamma_i) \) : batting average for the rest of the season |
| \( n_i \) | number of at bats for the rest of the season |
| \( X_i \) | \( X_i = \text{arcsin} \left( 2Y_i - 1 \right) \approx N \left( \mu_i, \frac{1}{45} \right), \mu_i = \text{arcsin} \left( 2\gamma_i - 1 \right) \) |
| \( \overline{R} \) | \( R_i = \text{arcsin} \left( 2P_i - 1 \right) \) |

Variance stabilizing transformations

Let \( X_i, i = 1, 2, \ldots, K \), be a sequence of i.i.d.r.v.'s with distribution in the exponential family \( P \) with parameter set \( \Theta \),

\[
P_\theta (dx) = \exp \{ \theta U (x) - V (\theta) \} \mu (dx)
\]

where \( U \) is a measurable map, and \( V (\theta) \) is the cumulant generating function associated with the exponential family. Let \( Y_i = U (X_i) \), then \( \sum_{i=1}^K Y_i \) is a sufficient statistics for the i.i.d. model. Set for brevity, \( S_K = \sum_{i=1}^K Y_i \) and \( \mu (\theta) = V (\theta) = EY_1 \), \( I (\theta) = V' (\theta) = Var \left( Y_1 \right) \).

According to the central limit theorem, the sequence \( \sqrt{K} \left( \frac{S_K}{K} - \mu (\theta) \right) \) converges weakly to the normal r.v. with zero mean and variance \( I (\theta) \). Define a function \( F : R \rightarrow R : F (\lambda) = I \left( \mu^{-1} (\lambda) \right)^{-1/2} \) such that \( F' (\mu (\theta)) = I (\theta)^{-1/2} \). The so called delta method gives

\[
\sqrt{K} \left\{ F \left( \frac{S_K}{K} - \mu (\theta) \right) \right\} \overset{d}{\rightarrow} N (0, 1).
\]
We then call $F$ a variance stabilization transformation. For two finite constants $a$ and $c$, it is also true that

$$\sqrt{K} \left\{ F \left( \frac{S_K + a}{K + c} \right) - F ( \mu (\theta) ) \right\} \xrightarrow{d} N (0, 1).$$

This suggests we have freedom to choose $a$ and $c$ in practice.

### 1970 Batting Averages for 18 Major League Players and Transformed Values $X$, $\theta$

<table>
<thead>
<tr>
<th>i</th>
<th>Player</th>
<th>$Y_i$, batting average for first 45 at bats</th>
<th>$p_i$, batting average for remainder of season</th>
<th>At bats for remainder of season</th>
<th>$X$, $\theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Clemente (Pits, NL)</td>
<td>.400</td>
<td>.346</td>
<td>367</td>
<td>-1.35</td>
</tr>
<tr>
<td>2</td>
<td>F. Robinson (Balt, AL)</td>
<td>.378</td>
<td>.268</td>
<td>456</td>
<td>-1.65</td>
</tr>
<tr>
<td>3</td>
<td>F. Howard (Wasl, AL)</td>
<td>.359</td>
<td>.276</td>
<td>521</td>
<td>-1.97</td>
</tr>
<tr>
<td>4</td>
<td>Johnstone (Cal, AL)</td>
<td>.333</td>
<td>.222</td>
<td>275</td>
<td>-2.29</td>
</tr>
<tr>
<td>5</td>
<td>Berry (Chi, AL)</td>
<td>.311</td>
<td>.273</td>
<td>418</td>
<td>-2.60</td>
</tr>
<tr>
<td>6</td>
<td>Spence (Cal, AL)</td>
<td>.311</td>
<td>.270</td>
<td>496</td>
<td>-2.99</td>
</tr>
<tr>
<td>7</td>
<td>Kessinger (Chi, NL)</td>
<td>.293</td>
<td>.263</td>
<td>586</td>
<td>-2.92</td>
</tr>
<tr>
<td>8</td>
<td>L. Alfonso (Bos, AL)</td>
<td>.267</td>
<td>.210</td>
<td>138</td>
<td>-3.25</td>
</tr>
<tr>
<td>9</td>
<td>Santo (Chi, NL)</td>
<td>.244</td>
<td>.220</td>
<td>110</td>
<td>-3.60</td>
</tr>
<tr>
<td>10</td>
<td>Swoboda (NY, NL)</td>
<td>.244</td>
<td>.230</td>
<td>200</td>
<td>-3.93</td>
</tr>
<tr>
<td>11</td>
<td>Unser (Wash, AL)</td>
<td>.222</td>
<td>.224</td>
<td>180</td>
<td>-3.95</td>
</tr>
<tr>
<td>12</td>
<td>Williams (Chi, AL)</td>
<td>.222</td>
<td>.226</td>
<td>210</td>
<td>-3.95</td>
</tr>
<tr>
<td>13</td>
<td>Scott (Bos, AL)</td>
<td>.222</td>
<td>.303</td>
<td>415</td>
<td>-3.95</td>
</tr>
<tr>
<td>14</td>
<td>Petrocelli (Bos, AL)</td>
<td>.222</td>
<td>.264</td>
<td>538</td>
<td>-3.95</td>
</tr>
<tr>
<td>15</td>
<td>E. Rodriguez (KC, AL)</td>
<td>.222</td>
<td>.226</td>
<td>186</td>
<td>-3.95</td>
</tr>
<tr>
<td>16</td>
<td>Campanellis (Oak, AL)</td>
<td>.200</td>
<td>.285</td>
<td>550</td>
<td>-4.32</td>
</tr>
<tr>
<td>17</td>
<td>Munson (NY, AL)</td>
<td>.178</td>
<td>.316</td>
<td>448</td>
<td>-4.70</td>
</tr>
<tr>
<td>18</td>
<td>Alvis (Mil, NL)</td>
<td>.156</td>
<td>.200</td>
<td>70</td>
<td>-5.10</td>
</tr>
</tbody>
</table>
Lecture 4. Bayes estimation, minimaxity and Admissibility.

Bayes estimator

Example: Observe a normally distributed $n$-dimensional random variable $X$,

$$X \sim N(\theta, \Sigma)$$

where $\theta$ and $\Sigma$ are parameters.

We assume that $\theta$ has a prior distribution $G$. A Bayes estimator, denoted by $\delta_G$, solves the following minimization problem:

$$\int R(\theta, \delta_G) G(d\theta) = \inf_{\delta} \left\{ \int R(\theta, \delta) G(d\theta) \right\}.$$  

When $G$ has a density w.r.t. Lebesgue measure, the conditional density of $\theta$ given $X = x$ is

$$f(\theta|x) = \frac{f_\theta(x) g(\theta)}{\int f_\theta(x) g(\theta) d\theta}.$$  

When the loss is squared error, $L(\theta, \delta) = (\theta - \delta)^T M (\theta - \delta)$ with $M$ positive definite, then the posterior mean is the Bayes estimator,

$$\delta_G = \frac{\int \theta f_\theta(x) g(\theta) d\theta}{g^*(x)}$$

with $g^*(x) = \int f_\theta(x) g(\theta) d\theta$.

If $G$ is a general (non-negative) measure, it is typical not true that $\inf_{\delta} \left\{ \int R(\theta, \delta) G(d\theta) \right\} < \infty$. We call $\delta_G$ a Bayes estimator if

$$\inf_{\delta} \left\{ \int [R(\theta, \delta) - R(\theta, \delta_G)] G(d\theta) \right\} \geq 0,$$

and call the posterior mean "formal" or "generalized" Bayes estimator.

Example: Let $X \sim N(\theta, 1)$ and $g(\theta) = e^{\alpha \theta}$. The posterior mean is $X + a$, but it is not Bayes estimator, since

$$\int [R(\theta, X) - R(\theta, X + a)] G(d\theta) < 0.$$  

Alternate form for Bayes estimators (for normal location problem)

Define

$$\nabla_2 h = (\partial^2 h / \partial x_i \partial x_j)_{n \times n}.$$  

Theorem. Let $X \sim N(\theta, \Sigma)$ with $\Sigma$ known. Let $G$ be any prior such that $g^*(x) < \infty$ for all $x$. Then

$$\theta | X = x \sim N(x + \Sigma \nabla (\log (g^*(x))), \Sigma + \Sigma \nabla_2 (\log (g^*(x)))).$$  

Proof of the theorem:

\[
E(\theta | X) = x + \frac{\int (\theta - x) \varphi_{\Sigma} (x - \theta) g(\theta) \, d\theta}{g^*(x)} = x + \Sigma \nabla g^* (x) \frac{g^*(x)}{g^*(x)}
\]

and

\[
Cov(\theta | X) = E(\theta - E(\theta | X)) (\theta - E(\theta | X))^T = E[(\theta - x) (\theta - x)^T | x] - (E(\theta | x) - x) (E(\theta | x) - x)^T
\]

where

\[
E[(\theta - x) (\theta - x)^T] = \Sigma (\nabla_2 (g^* (x))) \Sigma + g^* (x) \Sigma
\]

by integration by parts.

Example. Consider a normal prior \( \theta \sim N(\mu, \Gamma) \). Then the posterior distribution of \( \theta \) is

\[
\theta | X \sim N(\mu + \Gamma (\Sigma + \Gamma)^{-1} (X - \mu), \Gamma (\Sigma + \Gamma)^{-1} \Sigma)
\]

Minimality of \( \delta_0 = X \)

Lemma. For a given procedure \( \delta' \) suppose there is a sequence of prior distributions \( \{G_i\} \) such that

\[
\lim_{i \to \infty} \int R(\theta, \delta_{G_i}) G_i (d\theta) = \sup \theta R(\theta, \delta')
\]

Then \( \delta' \) is minimax.

The squared error loss:

\[
L(\theta, \delta) = (\theta - \delta)^T M (\theta - \delta)
\]

Theorem. For the normal location problem, \( \delta_0 = X \) is a minimax estimator of \( \theta \) under the squared error loss.

Proof of the theorem: Let \( G_i = N(0, i^2 I) \). Then

\[
\lim_{i \to \infty} \int R(\theta, \delta_{G_i}) G_i (d\theta) = \sup \theta R(\theta, \delta') = Tr \left( \Sigma M \right)
\]

Admissibility of \( \delta_0 = X \) for \( n \leq 2 \). We will show that next week.