

# Statistical Inference

## Lecture 11a

ANU - RSFAS

Last Updated: Mon May 14 05:19:39 2018

# Decision Theory

- The elements of a basic decision theory problem:
  1. A number of 'actions' are possible; we must decide which to take.
  2. A number of 'states of nature' are possible; we don't know in general which will occur.
  3. The relative desirability of the various actions for each state of nature can be quantified.
  4. Prior information may be available regarding the relative probabilities of the various states of nature.
  5. Data may be available which will add to our knowledge of the relative probabilities of the states of nature.

# Decision Theory


- Let  $\theta$  denote the true state of nature.
- Suppose we are able to observe some data . . . a draw from the random variable  $\mathbf{X}$ , whose distribution depends on  $\theta$ . Sometimes no data are available.
- A decision procedure  $\delta$ , specifies which action to take for each value of  $\mathbf{X}$ .
- If we observe  $\mathbf{X} = \mathbf{x}$ , then we adopt the procedure  $\delta(\mathbf{x})$ .
- Whether  $\delta(\mathbf{x})$  was a good choice depends on the loss function, which measures the loss from  $\delta(\mathbf{x})$  when  $\theta$  holds.

$$L_S(\theta, \delta(\mathbf{x}))$$

Note: The negative of a loss function is a utility function.

# Decision Theory

or maximize expected utility

- Frequentists want to minimize expected loss. 
- Bayesians want to minimize posterior expected loss.

# Decision Theory

**Eg.** (Lindley 1985) A doctor has the task of deciding whether or not to carry out a dangerous operation on a person suspected of suffering from a disease. If he has the disease and does operate, the chance of recovery is only 50%; without the operation the chance is only 1 in 20. On the other hand if he does not have the disease and the operation is performed there is 1 chance in 5 of his dying as a result of the operation, whereas there is no chance of death without the operation. Assume the patient has the following utilities for recovery and death  $u(R) = 1$  and  $u(\bar{R}) = 0$ . Advise the doctor (You may assume there are always only two possibilities, death or recovery.).

- There are two possible decision procedure:  $\delta_1 = \text{operate}$ ,  $\delta_2 = \text{not operate}$ .
- There are two states of nature:  $\theta_1 = \text{patient has the disease}$ ,  
 $\theta_2 = \text{patient does not have the disease}$ .
- There are two outcomes:  $z_1 = \text{full recovery}$ ,  $z_2 = \text{death}$ .

- If the doctor **operates**, we have the following expected utility:

$$\begin{aligned}
 \sum_{i=1}^2 u(z_i)P(z_i) &= \begin{array}{cc} \text{P(R and D)} & \text{P(R and not D)} \\ [P(R|D)P(D) + P(R|\bar{D})P(\bar{D})]u(R) + & \\ \text{P(not R and D)} & \text{P(not R and not D)} \\ [P(\bar{R}|D)P(D) + P(\bar{R}|\bar{D})P(\bar{D})]u(\bar{R}) & \end{array} \\
 &= [0.05P(D) + 0.8P(\bar{D})]u(R) + [0.5P(D) + 0.2P(\bar{D})]u(\bar{R})
 \end{aligned}$$

- If the doctor **does not operate**, we have the following expected utility:

$$\begin{aligned}
 \sum_{i=1}^2 u(z_i)P(z_i) &= [P(R|D)P(D) + P(R|\bar{D})P(\bar{D})]u(R) + \\
 &\quad [P(\bar{R}|D)P(D) + P(\bar{R}|\bar{D})P(\bar{D})]u(\bar{R}) \\
 &= [0.05P(D) + 1P(\bar{D})]u(R) + [0.95P(D) + 0P(\bar{D})]u(\bar{R})
 \end{aligned}$$

```
dec.lind <- function(u.death, u.rec){
```

```
p.dis <- seq(0,1, by=0.01)
```

probability of disease (state of nature, we don't know what it is)

```
u.op <- rep(0, length(p.dis))
```

```
u.d.op <- rep(0, length(p.dis))
```

```
for(i in 1:length(p.dis)){
```

```
u.op[i] <- (0.5*p.dis[i] + 0.8*(1-p.dis[i]))*u.rec +  
  (0.5*p.dis[i] + 0.2*(1-p.dis[i]))*u.death
```

```
u.d.op[i] <- (0.05*p.dis[i] + 1*(1-p.dis[i]))*u.rec +  
  (0.95*p.dis[i] + 0*(1-p.dis[i]))*u.death  
}
```

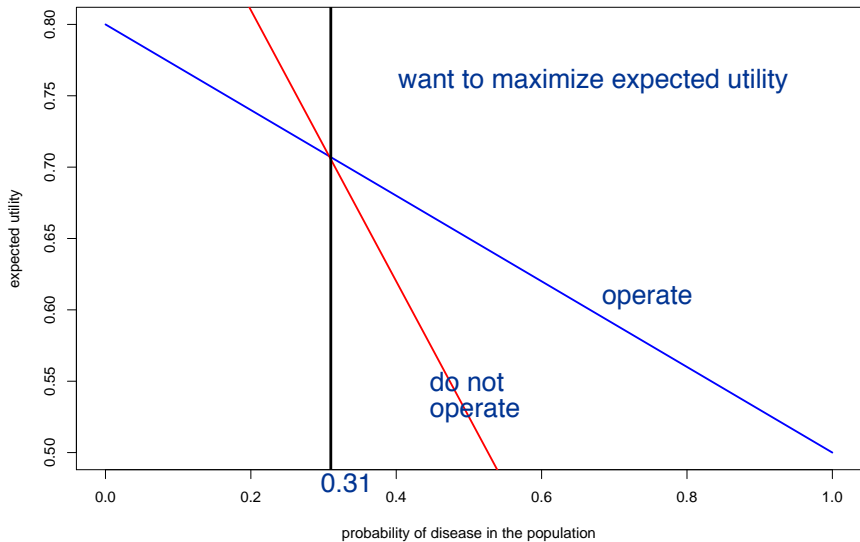
```
opt <- p.dis[u.op>u.d.op][1]
```

```
return(list(p.dis, u.op, u.d.op, opt))  
}
```



```
out <- dec.lind(u.death=0, u.rec=1)
p.dis <- out[[1]]
u.op <- out[[2]]
u.d.op <- out[[3]]
opt <- out[[4]]

plot(p.dis, u.op, type="l", lwd=2, ylab="expected utility",
      xlab="probability of disease in the population", col="blue")
lines(p.dis, u.d.op, col="red", lwd=2)
abline(v=opt, lwd=3)
```



The only information we don't have is the probability of disease in the population. This will be made based on the physician's judgment. The blue line in the figure represents the expected utility for the operation against the the probability of disease in the population. The red line is the expected utility for not operating against the disease in the population. The lines cross at  $P(D) = 0.31$ . Since we want to maximize expected utility, the the physician should not operate if  $P(D) < 0.31$  and should operate if  $P(D) > 0.31$ .

# Decision Theory

**Def 6.1:** The **risk function**  $R(\theta, \delta(\mathbf{x}))$  is defined as

$$R(\theta, \delta(\mathbf{x})) = \int \underbrace{L_S(\theta, \delta(\mathbf{x})) L(\theta; \mathbf{x})}_{=E(L, (\theta, \delta))} d\mathbf{x}.$$

This is the **expected loss** with respect to the joint distribution (i.e. likelihood).

# Decision Theory

**Def 6.1:** A procedure  $\delta_1$  is **inadmissible** if there exists another procedure  $\delta_2$  such that

$$R(\theta, \delta_1) \geq R(\theta, \delta_2) \quad \forall \theta,$$

with strict inequality for some  $\theta$ .

# Decision Theory

**Def 6.3:** The **minimax procedure** is such that

$$\max_{\theta} R(\theta, \delta)$$

is minimized.

assume the worst case

- Here we take a pessimistic view of the world by assuming **nature** is malevolent.

# Decision Theory

**Def 6.3:** A **Bayes** procedure is such that the **Bayes risk**

$$\int R(\theta, \delta) p(\theta) d\theta$$

is minimized. Expected prior loss is minimized.

# Decision Theory

- Consider:

$$\begin{aligned}\int R(\theta, \delta) p(\theta) d\theta &= \int_{\Theta} \left[ \int_{\mathcal{X}} L_S(\theta, \delta) L(\theta; \mathbf{x}) d\mathbf{x} \right] p(\theta) d\theta \\&= \int_{\Theta} \int_{\mathcal{X}} L_S(\theta, \delta) \frac{L(\theta; \mathbf{x}) p(\theta)}{p(\mathbf{x})} p(\mathbf{x}) d\mathbf{x} d\theta \\&= \int_{\Theta} \int_{\mathcal{X}} L_S(\theta, \delta) p(\theta | \mathbf{x}) p(\mathbf{x}) d\mathbf{x} d\theta \\&= \int_{\mathcal{X}} p(\mathbf{x}) \left[ \int_{\Theta} L_S(\theta, \delta) p(\theta | \mathbf{x}) d\theta \right] d\mathbf{x}\end{aligned}$$

frequentists worry  
about unseen  $\mathbf{x}$ 's  
sampled.

bayesians worry  
about the posterior  
expected loss



- For any value of  $\mathbf{x}$  we should minimize  $\int_{\Theta} L_S(\theta, \delta) p(\theta | \mathbf{x}) d\theta$ , which is **posterior expected loss**.



# Decision Theory

- Under certain conditions:

1. A Bayes procedure is necessarily admissible.
  2. Every admissible procedure is a Bayes procedure for some prior distribution.     you don't know what that prior is.
- For example, if  $\theta$  is discrete and only takes a number of finite values then (2) holds. Additionally if  $P(\theta) > 0$  for all  $\theta$  then (1) holds.

# Decision Theory

- The link between Bayes and minimax procedures:
  1. A Bayes procedure with constant risk for  $\theta$  is minimax.
  2. A minimax procedure is generally a Bayes procedure for some prior distribution. In particular, the so called **least favourable prior distribution**.

# Decision Theory

**Eg.** (Berger 2006) An insurance company is faced with taking one of the following 3 actions (decision procedures):  $\delta_1$ : increase sales force by 10%;  $\delta_2$ : maintain present sales force;  $\delta_3$ : decrease sales by 10%. Depending on whether the economy is good ( $\theta_1$ ), mediocre ( $\theta_2$ ), or bad ( $\theta_3$ ). The company would expect to lose the following amounts of money in each case:

|            | $\theta_1$ | $\theta_2$ | $\theta_3$ |
|------------|------------|------------|------------|
| $\delta_1$ | -10        | -5         | -3         |
| $\delta_2$ | -5         | -5         | -1         |
| $\delta_3$ | 1          | 0          | -1         |

loss table

1. Determine if each action is inadmissible.
2. The company believes that  $\theta$  has the probability distribution  $\pi(\theta_1) = 0.2, \pi(\theta_2) = 0.3, \pi(\theta_3) = 0.5$ . Order the actions according to their Bayesian expected loss (equivalent to the Bayes risk here since this is a no data problem) and state the Bayes action.
3. Order the actions according to the minimax principle and find the minimax action.

- We have no data in this case, which means we have no likelihood, which means the Risk is just the Loss.

$$R(\theta, \delta) = E(L_S(\theta, \delta))$$

$$\cancel{E[R(\theta, \delta)]} = L_S(\theta, \delta)$$

- So we just want to see if a  $\delta$  is dominated by other  $\delta$ s. Note:

$$L_S(\theta, \delta_3) \geq L_S(\theta, \delta_j)$$

for  $j = 2, 3$  and for all  $\theta$ . So  $\delta_3$  is inadmissible.

- To find the Bayes decision procedure, we want to minimize posterior expected loss. Here we don't have data, so we just minimize the Bayes risk:

$$\min_{\delta} \sum_{i=1}^3 L_s(\theta_i, \delta) p(\theta_i)$$

$$\sum_{i=1}^3 L_s(\theta_i, \delta_1) p(\theta_i) = 0.2(-10) + 0.3(-5) + 0.5(-3) = -5$$

$$\sum_{i=1}^3 L_s(\theta_i, \delta_2) p(\theta_i) = 0.2(-5) + 0.3(-5) + 0.5(-1) = -3$$

$$\sum_{i=1}^3 L_s(\theta_i, \delta_3) p(\theta_i) = 0.2(1) + 0.3(0) + 0.5(-1) = -0.3$$

So  $\delta_* = \delta_1$ .

- To find the minimax procedure (no data):

$$\min_{\delta} \max_{\theta} R(\theta, \delta) = \min_{\delta} \max_{\theta} L_S(\theta, \delta)$$

pick the largest value in loss table

$$\max_{\theta} L_S(\theta, \delta) = \begin{cases} \delta_1 : & -3 \\ \delta_2 : & -1 \\ \delta_3 : & 1 \end{cases}$$

- Now take the minimum:  $\delta^* = \delta_1$ .

# Decision Theory - Point Estimation

1. The possible actions are possible estimators.
2. Possible states of nature correspond to the true value of  $\theta$ .
3. Loss is some function of the estimator  $\hat{\theta}$  and  $\theta$ :  $L_S(\theta, \delta = \hat{\theta})$ .
4. Prior information is quantified by means of a prior distribution  $p(\theta)$ .
5. The available data, are our draws from  $f(x|\theta)$ .

# Decision Theory - Point Estimation

- Some simple loss functions:

## 1. Zero-one loss:

$$L_S(\theta, \delta = \hat{\theta}) = \begin{cases} 0, & |\hat{\theta} - \theta| < b, \\ a, & |\hat{\theta} - \theta| \geq b \end{cases}$$

where  $a, b > 0$ .

## 2. Absolute error loss:

$$L_S(\theta, \delta = \hat{\theta}) = a|\hat{\theta} - \theta|$$

## 3. Squared Error Loss (quadratic loss):

$$L_S(\theta, \delta = \hat{\theta}) = a(\hat{\theta} - \theta)^2$$



# Decision Theory - Point Estimation

- For squared-error loss we, we want to choose  $\hat{\theta}$  to minimize:  
or bales risk

$$\begin{aligned}\int (\hat{\theta} - \theta)^2 p(\theta|\mathbf{x}) d\theta &= E_{\theta|\mathbf{x}} [(\hat{\theta} - \theta)^2] \\&= E_{\theta|\mathbf{x}} [(\hat{\theta} - \bar{\theta} + \bar{\theta} - \theta)^2] \\&= E[(\hat{\theta} - \bar{\theta}) - (\theta - \bar{\theta})]^2 \\&= E[(\hat{\theta} - \bar{\theta})^2 - 2(\hat{\theta} - \bar{\theta})(\theta - \bar{\theta}) + (\theta - \bar{\theta})^2] \\&= E[(\hat{\theta} - \bar{\theta})^2] - 2(\hat{\theta} - \bar{\theta})E[(\theta - \bar{\theta})] + E[(\theta - \bar{\theta})^2] \\&= E[(\hat{\theta} - \bar{\theta})^2] - 2(\hat{\theta} - \bar{\theta})E[(\theta - \bar{\theta})] + E[(\theta - \bar{\theta})^2] \\&= (\hat{\theta} - \bar{\theta})^2 + E[(\theta - \bar{\theta})^2] \\&= [\text{Bias}(\theta)]^2 + V[\theta]\end{aligned}$$

only theta is  
random

theta.hat,  
theta.bar  
are not random

Where  $\bar{\theta}$  is the posterior mean.

- To minimize, we set  $\hat{\theta} = \bar{\theta}$ . Estimator is our posterior mean

# Decision Theory - Point Estimation

- Absolute error loss:  $E[|\theta - \hat{\theta}|] = \int |\theta - \hat{\theta}| p(\theta|\mathbf{x}) d\theta$ .
- This is minimized when  $\hat{\theta} = \text{median}$ .

**Proof:** Let  $m$  be the median of  $p(\theta|\mathbf{x})$ , and let  $a > m$  be another decision procedure. Note that:

$$L_S(\theta, m) - L_S(\theta, a) = \begin{cases} m - a & \text{if } \theta \leq m \\ 2\theta - (m + a) & \text{if } m < \theta < a \\ a - m & \text{if } \theta \geq a \end{cases}$$

Handwritten notes and derivations:

- $\hat{\theta} = m$  (red)
- $\theta \leq m < a$  (blue)
- $(m - \theta) - (a - \theta) = m - a$  (blue)
- $(m - \theta) - (a - \theta) = m - a$  (blue)
- $-(\theta - m) = m - \theta$  (blue)
- $| \theta - m |$  (blue)
- $-(\theta - m)$  (blue)
- $= (m - \theta)$  (blue)
- $\theta \leq m < a$  (blue)

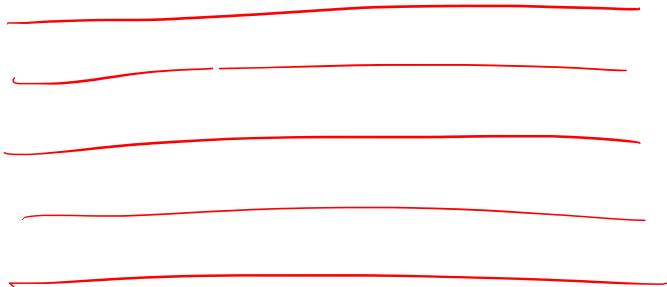
So:

$$L_S(\theta, m) - L_S(\theta, a) \leq (m - a) I_{(-\infty, m)}(\theta) + (a - m) I_{(m, \infty)}(\theta)$$

indicator function      indicator func

$$\theta \leq m < a$$

$$L_s(\theta, m)$$



$$E(L_S(\theta, m)) \leq E(L_S(\theta, a))$$

$$\begin{aligned} E[L_S(\theta, m) - L_S(\theta, a)] &\leq (m - a)E[l_{(-\infty, m)}(\theta)] + (a - m)E[l_{(-\infty, m)}(\theta)] \\ &\leq (m - a)P(\theta \leq m|\mathbf{x}) + (a - m)P(\theta > m|\mathbf{x}) \\ &\leq (m - a)(1/2) + (a - m)(1/2) = 0 \end{aligned}$$

- So the expected loss for  $m$  is smaller or equal  $a$ . We can make the same argument for  $a < m$ . Thus the median minimizes  $E[|\theta - \hat{\theta}|]$ .

# Decision Theory - Point Estimation (Eg. from L02b)

**Eg.:**  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\theta)$ .

- Consider the following prior,  $\theta \sim \text{beta}(a, b)$ .

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}$$

- From this, we can determine the posterior:

$$[\theta|\mathbf{x}] \sim \text{beta}(a^*, b^*)$$

$$a^* = a + y; \quad b^* = b + n - y$$

Where  $Y = \sum_{i=1}^n X_i$ .

# Decision Theory - Point Estimation

- Let's consider the Bayesian estimator under squared error loss (this is the mean of the posterior):

$$\hat{\theta} = \frac{y + a}{a + b + n}$$

- If we want the Bayes estimator under absolute error loss, it would be the median of the posterior.

$$\int_0^c p(\theta/x) d\theta = \frac{1}{2}$$

# Decision Theory - Point Estimation

- What is the minimax estimator under squared error loss? A Bayes procedure with constant risk for  $\theta$  is minimax.
- The risk function is (be careful . . . we are leaving Bayesian land for the moment):

Risk  
function

$$R(\theta, \delta = \hat{\theta}) = \int L_S(\theta, \delta(x)) L(\theta; x) dx$$

$$E_{x|\theta} = [\text{Bias}(\hat{\theta})]^2 + \text{Var}[\hat{\theta}]$$

(R) expected value of risk

$$\left\{ \begin{array}{l} E \left[ \frac{Y + a}{a + b + n} \right] = \frac{E[Y] + a}{a + b + n} = \frac{n\theta + a}{a + b + n} \\ V \left[ \frac{Y + a}{a + b + n} \right] = \left[ \frac{1}{a + b + n} \right]^2 V(Y) = \left[ \frac{1}{a + b + n} \right]^2 n\theta(1 - \theta) \end{array} \right.$$

# Decision Theory - Point Estimation

$$\begin{aligned} R(\theta, \delta(\mathbf{x})) &= \overset{\text{Var}}{\left[ \frac{n\theta(1-\theta)}{(a+b+n)^2} \right]} + \overset{\text{Bias}^2}{\left[ \frac{n\theta+a}{a+b+n} - \theta \right]^2} \\ &= \frac{[\theta(a+b)-a]^2 + n\theta(1-\theta)}{[a+b+n]^2} \end{aligned}$$

Risk\_squared loss  
= MSE

- It turns out that if we set  $a = b = \sqrt{n/4}$  then we get (which is constant for  $\theta$ ):

$$R(\theta, \delta = \hat{\theta}) = \frac{1}{(4(1 + \sqrt{n})^2)}$$

$$\hat{\theta} = \frac{Y + \sqrt{n/4}}{n + \sqrt{n}}$$

minimax  
estimation



# James-Stein Estimator

- Stein was the first person to realize that  $\bar{X}$  for a multivariate normal distribution with  $p \geq 3$  is **inadmissible**!
- See Stein 1995
- This discussion is taken from *Decision Theory* by Parmigiani and Inou. For full details see the textbook.
- Consider data  $X = (x_1, \dots, x_p)' \sim \text{multivariate normal}(\theta, I_p)$ .
- We can show, but won't (see Parmigiani and Inoue):

$$E_{x|\theta} \left[ \frac{\sum x_i \theta_i}{\sum x_i^2} \right] = E_{y|\theta} \left[ \frac{2y}{p-2+2y} \right]$$

$$E_{x|\theta} \left[ \frac{1}{\sum x_i^2} \right] = E_{y|\theta} \left[ \frac{1}{p-2+2y} \right]$$

Where  $y \sim \text{Poisson}(\sum \theta_i^2/2)$ .

$$x \sim \text{MVN}(\theta, \Sigma = I_\theta)$$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \sim \text{MVN} \left( \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{pmatrix}, \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} \right)$$

$$\hat{\theta}_1 = x_1, \hat{\theta}_2 = x_2, \dots, \hat{\theta}_n = x_n$$

$$x \sim N(\mu, I) \Rightarrow \hat{\mu} = x$$

- For a single multivariate data point  $x$ , the MLE is  $\hat{\theta} = x$ .
- Consider the following loss function:

$$L_S(\theta, \delta) = \sum_{i=1}^p (\theta_i - \delta_i)^2$$

- Consider the decision procedure:

*specifically.*  $\delta_i = x_i \left[ 1 - \frac{p-2}{\sum_{i=1}^p x_i^2} \right]$

*James Stein estimator*  $\rightarrow \delta_1 = x \left[ 1 - \frac{p-2}{\sum_{i=1}^p x_i^2} \right]$

$\delta_1$  dominates  $\delta = x$ , which is the MLE!

**Proof:** The risk function for  $\delta$  is:

$$\begin{aligned}
 R(\theta, \delta) &= \int L_S(\theta, \delta) p(x|\theta) dx && \text{likelihood} \\
 &= \int \left( \sum_{i=1}^p (\theta_i - \delta_i)^2 \right) p(x_i|\theta_i) dx_i && \begin{array}{l} \text{loss function} \\ \text{likelihood} \end{array} \\
 &= \sum_{i=1}^p \underbrace{V(x_i|\theta_i)}_1 = p && \begin{array}{l} \text{variance of each is 1} \\ \text{sum them up to } p. \end{array}
 \end{aligned}$$

*Handwritten notes:*  
 $\hat{\theta}_i = x_i, E(\hat{\theta}_i) = \theta$


- The risk function for  $\delta_1$  is:

*moving & grouping things around.*

$$\begin{aligned}
 R(\theta, \delta) &= E_{\mathbf{x}|\theta} \left\{ \sum \left[ \theta_i - x_i \left[ 1 - \frac{p-2}{\sum_{i=1}^p x_i^2} \right] \right]^2 \right\} \\
 &= E_{\mathbf{x}|\theta} \left\{ \sum \left[ \theta_i - x_i + \frac{p-2}{\sum_{i=1}^p x_i^2} x_i \right]^2 \right\} \\
 &= E_{\mathbf{x}|\theta} \left\{ \sum \left[ x_i - \theta_i - \frac{p-2}{\sum_{i=1}^p x_i^2} x_i \right]^2 \right\} \\
 &= E_{\mathbf{x}|\theta} \left\{ \sum \left[ (x_i - \theta_i)^2 - 2(x_i - \theta_i) \frac{p-2}{\sum_{i=1}^p x_i^2} x_i + \left( \frac{p-2}{\sum_{i=1}^p x_i^2} x_i \right)^2 \right] \right\} \\
 &= E_{\mathbf{x}|\theta} \left\{ \sum (x_i - \theta_i)^2 - 2(p-2) \sum \frac{(x_i - \theta_i)x_i}{\sum_{i=1}^p x_i^2} + (p-2)^2 \sum \left( \frac{x_i}{\sum_{i=1}^p x_i^2} \right)^2 \right\}
 \end{aligned}$$

$$\begin{aligned}
R(\theta, \delta) &= E_{\mathbf{x}|\theta} \left\{ \sum (x_i - \theta_i)^2 - 2(p-2) \sum \frac{(x_i - \theta_i)x_i}{\sum_{i=1}^p x_i^2} + (p-2)^2 \sum \left( \frac{x_i}{\sum_{i=1}^p x_i^2} \right)^2 \right\} \\
&= E_{\mathbf{x}|\theta} \left\{ \sum (x_i - \theta_i)^2 - 2(p-2) \frac{\sum (x_i - \theta_i)x_i}{\sum_{i=1}^p x_i^2} + (p-2)^2 \left( \frac{1}{\sum_{i=1}^p x_i^2} \right) \right\} \\
&= p - 2(p-2) E_{\mathbf{x}|\theta} \left\{ \frac{\sum (x_i - \theta_i)x_i}{\sum_{i=1}^p x_i^2} \right\} + (p-2)^2 E_{\mathbf{x}|\theta} \left\{ \frac{1}{\sum_{i=1}^p x_i^2} \right\} \\
&= p - 2(p-2) E_{\mathbf{x}|\theta} \left\{ 1 - \frac{\sum x_i \theta_i}{\sum_{i=1}^p x_i^2} \right\} + (p-2)^2 E_{\mathbf{x}|\theta} \left\{ \frac{1}{\sum_{i=1}^p x_i^2} \right\} \\
&= p - 2(p-2) \left[ 1 - E_{\mathbf{x}|\theta} \left\{ \frac{\sum x_i \theta_i}{\sum_{i=1}^p x_i^2} \right\} \right] + (p-2)^2 E_{\mathbf{x}|\theta} \left\{ \frac{1}{\sum_{i=1}^p x_i^2} \right\} \\
&= p - 2(p-2) \left[ 1 - E_{\mathbf{y}|\theta} \left\{ \frac{2y}{p-2+2y} \right\} \right] + (p-2)^2 E_{\mathbf{y}|\theta} \left\{ \frac{1}{p-2+2y} \right\} \\
&= p - 2(p-2) \left[ E_{\mathbf{y}|\theta} \left\{ \frac{p-2+2y}{p-2+2y} \right\} - E_{\mathbf{y}|\theta} \left\{ \frac{2y}{p-2+2y} \right\} \right] + \\
&\quad (p-2)^2 E_{\mathbf{y}|\theta} \left\{ \frac{1}{p-2+2y} \right\}
\end{aligned}$$

$$\begin{aligned}
R(\theta, \delta) &= p - 2(p-2) \left[ E_{y|\theta} \left\{ \frac{p-2+2y}{p-2+2y} \right\} - E_{y|\theta} \left\{ \frac{2y}{p-2+2y} \right\} \right] + \\
&\quad (p-2)^2 E_{y|\theta} \left\{ \frac{1}{p-2+2y} \right\} \\
&= p - 2 \left[ E_{y|\theta} \left\{ \frac{(p-2)(p-2+2y)}{p-2+2y} \right\} - E_{y|\theta} \left\{ \frac{(p-2)2y}{p-2+2y} \right\} \right] + \\
&\quad (p-2)^2 E_{y|\theta} \left\{ \frac{1}{p-2+2y} \right\} \\
&= p - 2 \left[ E_{y|\theta} \left\{ \frac{(p-2)(p-2+2y) - (p-2)2y}{p-2+2y} \right\} \right] + E_{y|\theta} \left\{ \frac{(p-2)^2}{p-2+2y} \right\} \\
&= p - 2 \left[ E_{y|\theta} \left\{ \frac{(p-2)^2}{p-2+2y} \right\} \right] + E_{y|\theta} \left\{ \frac{(p-2)^2}{p-2+2y} \right\} \\
&= \underbrace{p}_{\text{blue underline}} - E_{y|\theta} \left\{ \frac{(p-2)^2}{p-2+2y} \right\} < \underbrace{p}_{\text{blue underline}}
\end{aligned}$$



• Recall  $p \geq 3$ .

- The James-Stein estimator is derived from a Bayesian approach. We note from Parmigiani and Inoue, that Robert (1994) states:

One of the major impacts of the Stein paradox is to signify the end of a “Golden Age” for classical Statistics, since it shows that the quest for the best estimator, i.e., the unique minimax admissible estimator, is hopeless, unless one restricts the class of estimators to be considered or incorporates some prior information. [ . . . ] its main consequences has been to reinforce the Bayesian–frequentist interface, by inducing frequentists to call for Bayesian techniques and Bayesians to robustify their estimators in terms of frequentist performances and prior uncertainty.