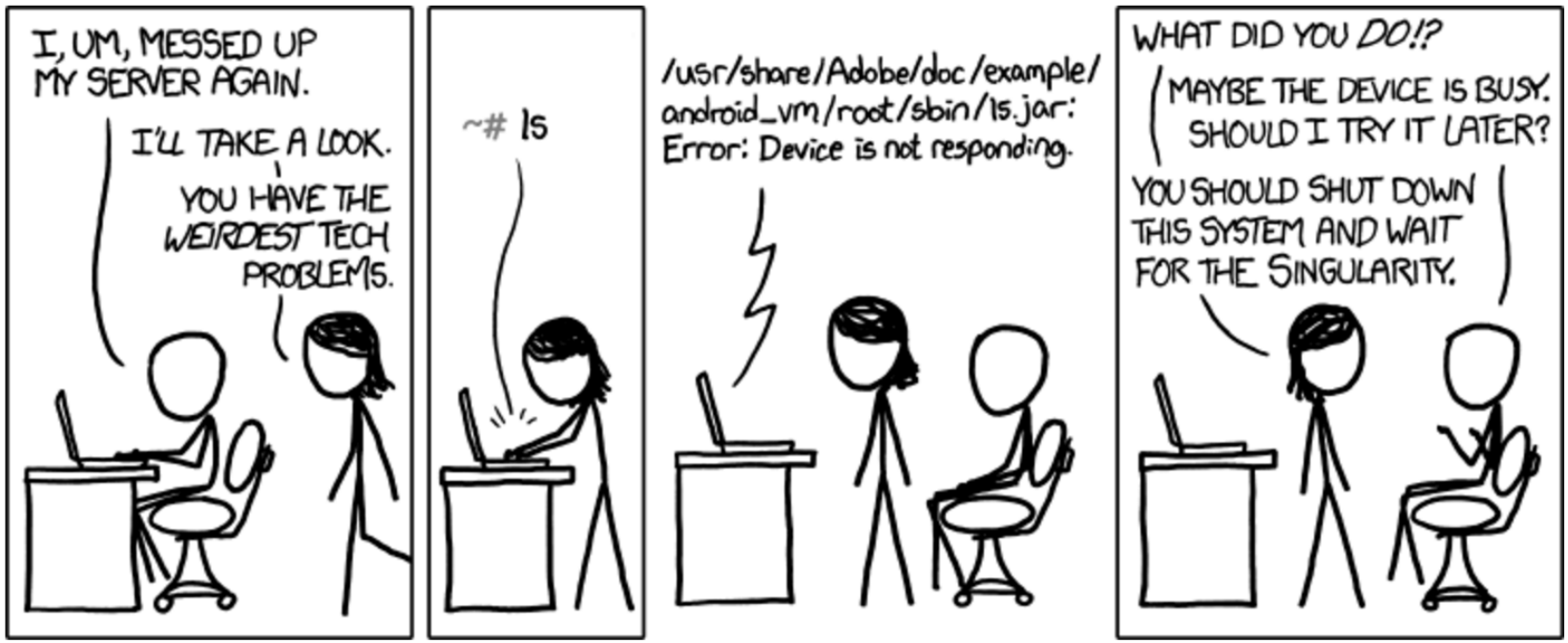
Intro to Decision Theory

Rebecca C. Steorts

Bayesian Methods and Modern Statistics: STA 360/601

Lecture 3

Please be patient with the Windows machine....



Topics

- Loss function
- Risk
- Posterior Risk
- Bayes risk
- Bayes estimator
- Minimax estimators
- An Example

A Nice Relaxation

1. Derivations on the homework, you can write it up and scan it.
2. Any computational piece must be done in R/Markdown and be reproducible (this includes the writing here).

What about file submissions. If you choose to do it as homework 1, submit as before.

If you choose to submit with handwritten, R/Markdown, then you will need:

1. A .pdf file with clearly written problems (note anything that we can't read, won't be graded).
2. Any computational piece must be done in R/Markdown and be reproducible (this includes the writing here).
3. You must attach all .pdf, .tex, .Rmd files above that are needed to grade your homework. If you're unsure, come and ask in office hours.

Other stuff that comes up (come talk to me)....

Motivating Example: Skiing Anyone?

Motivating Example: Skiing Anyone?

Risk and Skiing

What types of things could go wrong?

One way Bayesian methods are often used are in making optimal decisions.

In statistical decision theory, we formalize good and bad results with a loss function.

What is loss?

- Loss function $L(\theta, \delta(x))$ is a function of unknown parameter $\theta \in \Theta$.
- $\delta(x)$ is a decision based on the data $x \in X$.

What are some examples of $\delta(x)$?

- Does Duke win or lose a given basketball game (0-1 loss).
- Two player game based on set of non-binary rules (point system).[1]
- Sample average of the data.

Back to our skiing example: θ: probability that you tear your ACL. δ: estimator of θ.

[1] (Discrete loss, for this example see Ch 5 of Baby Bayes notes).

Loss

- The loss function determines the penalty for deciding how well $\delta(x)$ estimates θ.

One discrete loss (0-1):

$$L(\theta, \delta(x)) = \begin{cases} 0 & \text{if } \delta(x) = \theta, \\ 1 & \text{if } \delta(x) \neq \theta, \end{cases}$$

Another loss is squared error: $L(\theta, \delta(x)) =$

Why do we use such choices?

What is risk?

We understand loss. But how do we understand a concept of "risk" in a formal setting?

Example: "I hate getting wet in the rain. How can I minimize the risk that I will never get wet?"

Intuition: Just always carry an umbrella.

We will come back to this example later. Now, some defintitions.

Frequentist Risk (Risk)

The frequentist risk (risk) is

$$R(\theta, \delta(x)) = E_\theta[L(\theta, \delta(x))] = \int_X L(\theta, \delta) f(x|\theta)\, dx.$$

where θ is held fixed and the expectation is taken over X.

Risk measures the long-term average loss resulting from using δ.

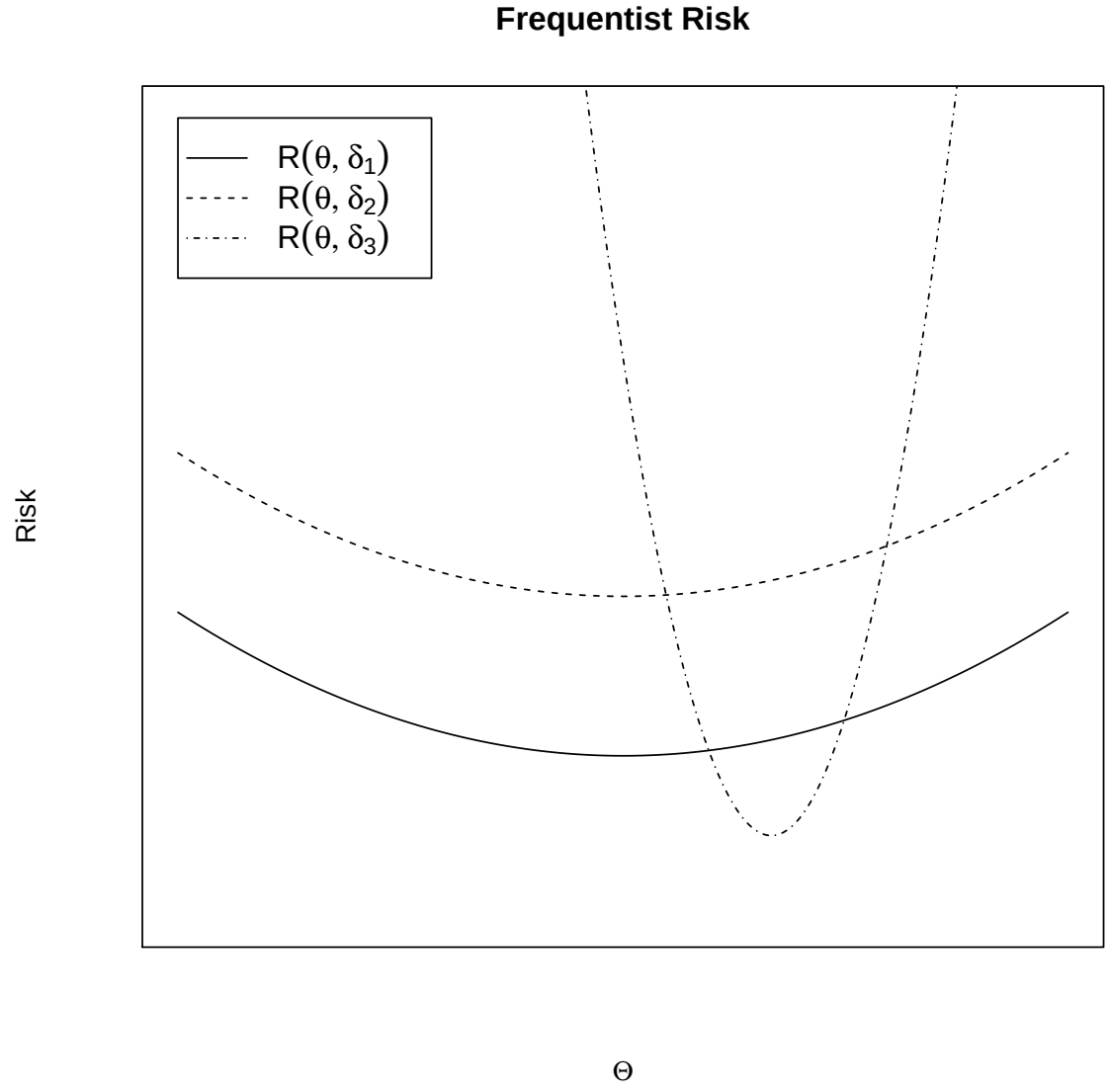
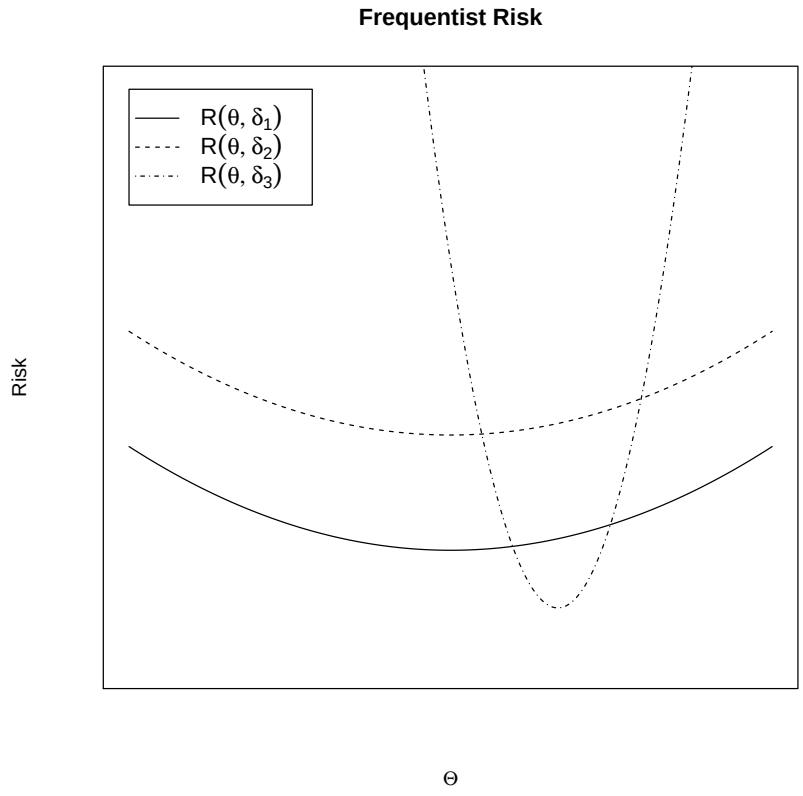


Figure 1: Shows the risk of three different decisions as a function of $\theta \in \Theta$



Do any of the estimators dominate the other (over θ)?

Do any not?

Frequentists have a few answers for deciding which is better:

1. **Restricted classes of procedure**.
   - We could force $E_\theta[\hat{\theta}] = \theta$ for all θ (unbiased).
     - Suppose we only look at unbiased estimators.
     - Then we can often reduce the situation to only risk curves like δ_1 and δ_2 in Figure 1, eliminating overlapping curves like δ_3.
   - Existence of an optimal unbiased procedure is a nice frequentist theory, but many good procedures are biased—for example Bayesian procedures are typically biased.
   - Food for thought: will an unbiased estimator always exist?!?!?
2. **Minimax**. Here, we look at $\sup_\Theta R(\theta, \delta(x))$, where

$$R(\theta, \delta(x)) = E_\theta[L(\theta, \delta(x))].$$

   - In Figure 2, δ_2 would be chosen over δ_1 because its maximum worst-case risk (the grey dotted line) is lower.
     - Specifically, find the global max of δ_1 and δ_2. Now choose the minimum. This gives you the minimimax estimator which is δ_2 here.
     - First, maximize over all possible $\theta \in \Theta$. Then take the minimum.
   - sup = supremum. See the definition given in class.

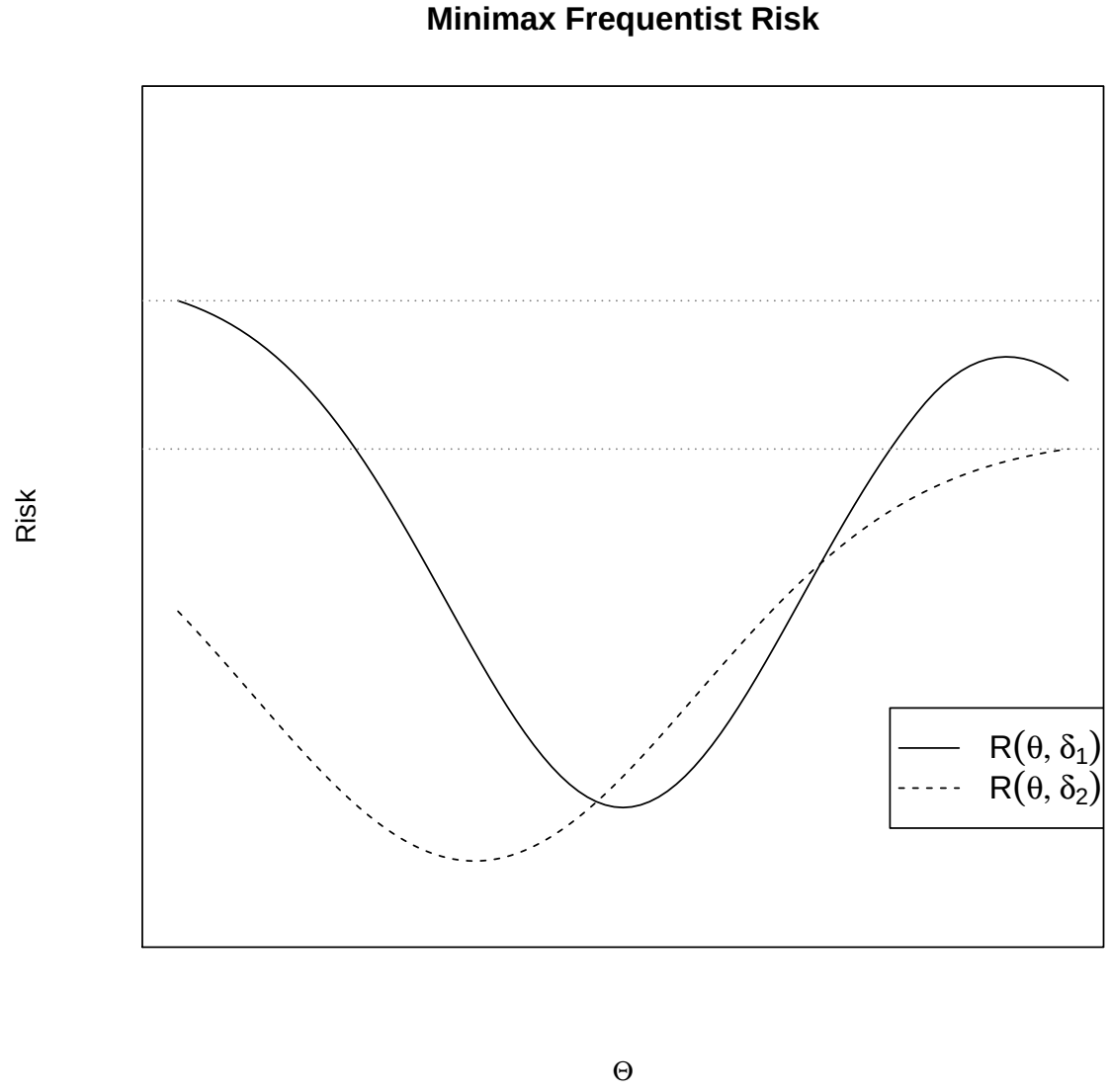


Figure 2: Minimax frequentist Risk

Bayesian Decision Theory

Define the posterior risk as

$$\rho(\pi, \delta(x)) = \int_{\Theta} L(\theta, \delta(x))\pi(\theta|x)\, d\theta.$$

The Bayes action $\delta^*(x)$ for any fixed x is the decision $\delta(x)$ that minimizes the posterior risk.

If the problem at hand is to estimate some unknown parameter θ, then we typically call this the Bayes estimator instead.

Theorem: Under squared error loss, the Bayes estimator ($\hat{\theta}_B$) minimizes the posterior risk.

$\hat{\theta}_B$ happens to be the posterior mean $E(\theta \mid \boldsymbol{X})$!

Under squared error loss, $\delta(\hat{x})$ minimizing the posterior risk is $\hat{\theta}_B$.

Bayes risk

The Bayes risk is denoted by $r(\pi, \delta(x))$. While the Bayes risk is a frequentist concept since it averages over X, the expression can also be interpreted differently. Consider

$$r(\pi, \delta(x)) = \int \int L(\theta, \delta(x)) \; f(x|\theta) \;\; \pi(\theta) \; dx \; d\theta \tag{1}$$

$$r(\pi, \delta(x)) = \int \int L(\theta, \delta(x)) \; \pi(\theta|x) \;\; \pi(x) \; dx \; d\theta \tag{2}$$

$$r(\pi, \delta(x)) = \int \rho(\pi, \delta(x)) \; \pi(x) \; dx. \tag{3}$$

Note that the last equation is the posterior risk averaged over the marginal distribution of x.

Connection with frequentist theory includes that finding a Bayes rule against the "worst possible prior" gives you a minimax estimator.

The Umbrella and the Statistician

Suppose a statistician doesn't particularly like getting wet, so he's interested in knowing how often he should carry his umbrella given the probability that it will rain on a given day.

		Statistician	
		Takes Umbrella (T)	Doesn't Take (D)
Nature	Rain (R)	0	10
	No Rain (N)	1	0

Table 1: Loss Function

Consider what our loss function is in terms of the table above.

$$L(\theta, d) = \begin{cases} 0, & \text{for } L(R,T), \\ 10, & \text{for } L(R,D), \\ 1, & \text{for } L(N,T), \\ 0, & \text{for } L(N,D). \end{cases}$$

- Notice that in this example, there is no data X.
- Thus, $R(\theta, d) = E_\theta[L(\theta, d(X))]$ is really just $E_\theta[L(\theta, d)] = L(\theta, d)$, so the risk and the loss are the same. This is *always* the case in no data problems.
- Suppose we can predict "Rain" with 100% accuracy.
- Let's now find the value of d that minimizes $R(\theta, d)$.

Recap: we're minimizing the risk first.

Recall the loss.

$$L(\theta, d) = \begin{cases} 0, & \text{for } L(R,T), \\ 10, & \text{for } L(R,D), \\ 1, & \text{for } L(N,T), \\ 0, & \text{for } L(N,D). \end{cases}$$

Solution: Note, I'm using R for two different things (risk and Rain)!

Let's look at $\theta = (Rain)$

$$R(R, d) = \begin{cases} 0, & \text{if } d = T, \\ 10, & \text{if } d = D. \end{cases}$$

- The d that minimizes the risk above is $d = T$, meaning the statistician would take his umbrella.
- What happens when $\theta = N$ with 100% accuracy?

The cases above are unreasonable, so let's consider the situation where we know

$$\theta = \begin{cases} R, & \text{with probability } p, \\ N, & \text{with probability } 1-p. \end{cases}$$

This is a prior $p(\theta)$ on the values of θ.

Now we would like to minimize the Bayes risk, $r(d) = E_X[R(\theta, d)]$. Recall the loss.

$$L(\theta, d) = \begin{cases} 0, & \text{for } L(R,T), \\ 10, & \text{for } L(R,D), \\ 1, & \text{for } L(N,T), \\ 0, & \text{for } L(N,D). \end{cases}$$

Solution: If the statistician *takes the umbrella (T)*, then

$$r(d) = E_X[R(\theta, T)] = p \cdot 0 + (1-p) \cdot 1 = 1 - p.$$

If the statistician decides to *does not take the umbrealla (N)* then

$$r(d) = E_X[R(\theta, D)] = p \cdot 10 + (1-p) \cdot 0 = 10p.$$

Recall from the previous slide: If the statistician *takes the umbrella*, then

$$r(d) = E_X[R(\theta, T)] = p \cdot 0 + (1 - p) \cdot 1 = 1 - p.$$

If the statistician decides to *leave his umbrella* at home then

$$r(d) = E_X[R(\theta, D)] = p \cdot 10 + (1 - p) \cdot 0 = 10p.$$

Then

- If $1 - p < 10p$, then the statistician should take his umbrella.
- On the other hand, if $1 - p > 10p$, the statistician should leave his umbrella at home. Note that we have minimized the Bayes risk.

At home: think about what is the value of p when the statistician takes and doesn't take his umbrella that results in the two situations having equal Bayes risk?

Back to Minimaxity

An estimator is minimax if it minimizes the maximum risk. For solving problems:

1. maximize the risk over all possible parameter values, e.g. θ.
2. Then we find the estimator that minimizes this maximum risk.

Often applying the definition can be hard.

More on Minimaxity

A more useful way of showing minimaxity!

If the Bayes estimate, $\hat{\theta}_B$, has constant (frequentist) risk under the given prior, then $\hat{\theta}_B$ is considered to be minimax.

Recall our "umbrella example":

$$L(\theta, d) = \begin{cases} 0, & \text{for } L(R, T), \\ 10, & \text{for } L(R, D), \\ 1, & \text{for } L(N, T), \\ 0, & \text{for } L(N, D). \end{cases}$$

- This is a no data problem, so $R(\theta, d) = L(\theta, d)$.
- Then to find the minimax estimator d, we first maximize over all possible values of θ for each estimator d, i.e., we maximize over rain and not rain.
- The maximum risks for D and T are $R(R, D) = 10$ and $R(N, T) = 1$. Then minimizing the risk functions over the estimators, we find $R(N, T) = 1$.
- So, the minimax estimator is $d = T$, or rather for the statistician to *always* carry his umbrella.
- How would this change with a prior on θ? Think about this at home.