Module 2: Introduction to Decision Theory

Rebecca C. Steorts

Agenda

- What is decision theory?
- Simple set up for discrete case
- General setup
- Loss functions
- Frequentist Risk
- Bayesian risk
- Bayes rule under squared error loss
- Lab/Homework: Application to resource allocation

What will you learn?

The basics of decision theory, which include:

- notation/terminology
- discrete example
- loss function(s)
- frequentist risk and posterior risk
- practice working with these concepts via derivations, lab, and via homework

Background/review: Convex function
https://en.wikipedia.org/wiki/Convex_function Result: "A twice-differentiable function of a single variable is convex if and only if its second derivative is non-negative on its entire domain."
-Wikipedia

Principles/Notation

Decision theory consists of three basic principles:

1. Possible states of nature Θ (the parameter space).
2. A set of possible actions $\mathcal{A}$
3. A loss function $\ell(\theta, a)$, where $\theta \in \Theta$ and $a \in \mathcal{A}$.

If the actions are data dependent, the literature denotes this as a decision rule $\delta(x)$.

$(\Theta, \mathcal{A}, \ell)$ define a game, where typically the goal is to minimize the loss function.

Note: We may or may not have observed data $x \in \mathcal{X}$.

Rules of "the game"

1. Nature selects a point $\theta \in \Theta$.

Example: Nature decides it will rain today.

2. You (the statistician) chose an action $a \in \mathcal{A}$ **where you are not informed of the state of nature.**

Example: Your **action** is to NOT carry an umbrella.

Based upon 1 and 2, you will lose an amount $\ell(\theta, a)$.

Motivating Example

$$\Theta = \{\text{Rain}, \text{No Rain}\}$$

$$\mathcal{A} = \{\text{Umbrella}, \text{No Umbrella}\}$$

		$\mathcal{A}$	
		Umbrella	No Umbrella
Θ	Rain	0	10
	No Rain	1	0

The 2 by 2 matrix illustrates a loss function $\ell(\theta, a)$ for all combinations of a, θ.

What do you learn from this figure?

Risk

We will define two types of risk that are commonly used, the *frequentist risk* and the *posterior risk*.

Our goal is to minimize our risk (we want to avoid losing, such as losing money, bad outcomes, or getting wet!)

Frequentist risk

Suppose that $\theta \in \Theta$. Suppose that X has distribution $p(x|\theta)$.

The ***risk*** (or ***frequentist risk***) associated with an action or decision procedure $\delta(X)$ is

$$R(\theta, \delta(X)) = \mathbb{E}(\ell(\theta, \delta(X)) \mid \theta) = \int \ell(\theta, \delta(x))\, p(x|\theta)\, dx,$$

if X is continuous, while the integral is replaced with a sum if X is discrete.

Posterior risk (or Bayes' risk)

The **posterior risk** is defined to be

$$\rho(\theta, \delta(X)) = \mathbb{E}(\ell(\theta, \delta(X)) \mid X = x) = \int \ell(\theta, \delta(X))\ p(\theta \mid x)\ d\theta,$$

if Θ is continuous, while the integral is replaced with a sum if Θ is discrete.

Exercise

Let's return to the umbrella problem. Notice that there is no data defined. (This is called a no data problem).

1. What is the frequentist risk?

Solution:

$$R(\theta, a) = \mathbb{E}_X(\ell(\theta, a) \mid \theta) = \ell(\theta, a)$$

because there is no data X.

Observe that the frequentist risk and the loss are the same. This is *always* the case when there is no data.

Exercise (continued)

2. Suppose that the state of nature θ is rain all the time. Find the action a that minimizes the frequentist risk.

Solution: Assuming that it rains all the time, we know that

$$R(\text{Rain}, a) = \begin{cases} 0 & \text{if } a = \text{Umbrella} \\ 10 & \text{if } a = \text{No Umbrella} \end{cases}$$

The action a that minimizes the frequentist risk above is $a = \text{Umbrella}$.

Exercise (continued)

3. **Consider the prior below on the state of nature θ. Define the posterior risk and simplify it.**

$$\theta = \begin{cases} \text{Rain} & \text{with probability } p \\ \text{No Rain} & \text{with probability } 1-p \end{cases}$$

The posterior risk is

$$\rho(\theta, a) = \mathbb{E}_{\Theta}\left(\ell(\theta, \delta(X)) \mid X = x\right) = \sum_{\Theta} \ell(\theta, \delta(X)) p(\theta).$$

Exercise (continued)

In the situation, where one carries the umbrella

$$\rho(\theta, a) = p \times 0 + (1 - p) \times 1 = 1 - p.$$

In the situation, where one decides to not carry the umbrella

$$\rho(\theta, a) = p \times 10 + (1 - p) \times 0 = 10p.$$

Exercise (continued)

4. **Find the action(s) that minimizes the posterior risk.**

To minimize the posterior risk, we consider the following:

If $1 - p < 10p$, the optimal decision is to carry the umbrella.

On the other hand, if $1 - p > 10p$, the optimal decision is to not carry the umbrella.

Exercise (continued)

5. **Find the action that minimizes the posterior risk when p = 0.2.**

Then the posterior risk from carrying the umbrella is $1 - p = 0.8$ and the posterior risk from not carrying the umbrella is $10p = 2$.

In this situation and fixed value of p, the posterior risk is minimized by carrying the umbrella.

See written details here: https://github.com/resteorts/modern-bayes/blob/master/reading/babybayes-master.pdf pages 99– 100.

General setup

Assume an unknown state of nature Θ ($\theta \in \Theta$).

Also,

- we observe data $x = x_{1:n} \in \mathcal{X}$,
- we take an action a or decision procedure $\delta(X)$,
- we incur a real-valued loss function $\ell(\theta, \delta(X))$.

Bayes rule

A ***Bayes rule*** is an optimal decision procedure $\hat{\delta}(X)$ that **minimizes the posterior risk** for all values of $x \in \mathcal{X}$.[1]

[1] Sometimes the loss is restricted to be nonnegative, to avoid certain pathologies.

Assumptions

1. Assume observed data $x = x_{1:n}$.
2. Assume that $\hat{\delta}(X)$ is the **optimal decision rule** $\delta(X)$.
3. Assume squared error loss

$$\ell(\theta, \delta(X)) = (\theta - \delta(X))^2.$$

Theorem

Show that the posterior mean $\hat{\delta}(x) = E[\theta \mid x_{1:n}]$ minimizes the posterior risk. (It's important to verify the solution is unique).

Proof

By definition, we want to **minimize the posterior risk.**

The **posterior risk** can be written as

$$\begin{aligned}\rho(\theta, \delta(x)) &= \mathbb{E}(\ell(\theta, \delta(x))|x_{1:n}) = \mathbb{E}((\theta - \delta(x)))^2|x_{1:n}) \\ &= \mathbb{E}(\theta^2 - 2\theta\ \delta(x)) + \delta^2(x))|x_{1:n}) \\ &= \mathbb{E}(\theta^2|x_{1:n}) - 2\delta(x))\mathbb{E}(\theta|x_{1:n}) + \delta^2(x)).\end{aligned}$$

Proof (continued)

Let's now **minimize the posterior risk.**

Recall that

$$\rho(\theta, \delta(x)) = \mathbb{E}(\theta^2 | x_{1:n}) - 2\delta(x))\mathbb{E}(\theta | x_{1:n}) + \delta^2(x))$$

$$\begin{aligned}\frac{\partial \rho(\theta, \delta(x))}{\partial \delta(x))} &= \frac{\partial \{\mathbb{E}(\theta^2 | x_{1:n}) - 2\delta(x))\mathbb{E}(\theta | x_{1:n}) + \delta^2(x))\}}{\partial \delta x} \\ &= -2\mathbb{E}(\theta | x_{1:n}) + 2\theta\end{aligned}$$

Proof (continued)

Now, let

$$0 - 2\mathbb{E}(\theta|x_{1:n}) + 2\delta(x) =: 0,$$

which implies that

$$\delta(x) = \mathbb{E}(\theta|x_{1:n}).$$

Because the loss function is convex, $\delta(x) = \mathbb{E}(\theta|x_{1:n})$ is the unique solution.[2]

[2] Alternatively, the second partial derivative is 0, which is non-negative implying the function is convex and the solution is unique.

Summary of Theorem

To summarize, we just showed that the Bayes rule under squared error loss is

$$\hat{\delta}(x) = \mathbb{E}(\theta | x_{1:n}).$$

That is, the **Bayes rule is the posterior mean**!

Resource allocation for disease prediction

This material corresponds with lab 3 and homework 3.

Suppose public health officials in a small city need to decide how much resources to devote toward prevention and treatment of a certain disease, but the fraction θ of infected individuals in the city is unknown.

Resource allocation for disease prediction (continued)

Suppose they allocate enough resources to accomodate a fraction c of the population. Recall that θ is the fraction of the infected individuals in the city.

- If c is too large, there will be wasted resources, while if it is too small, preventable cases may occur and some individuals may go untreated.
- After deliberation, they adopt the following loss function:

$$\ell(\theta, c) = \begin{cases} |\theta - c| & \text{if } c \geq \theta \\ 10|\theta - c| & \text{if } c < \theta. \end{cases}$$

This applied example corresponds to lab 3 and homework 3.

Resource allocation for disease prediction (continued)

- By considering data from other similar cities, they determine a prior $p(\theta)$. For simplicity, suppose $\boldsymbol{\theta} \sim \text{Beta}(a, b)$ (i.e., $p(\theta) = \text{Beta}(\theta|a, b)$), with $a = 0.05$ and $b = 1$.[3]
- They conduct a survey assessing the disease status of $n = 30$ individuals, $x_1, \ldots, x_n$.

This is modeled as $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$, which is reasonable if the individuals are uniformly sampled and the population is large. Suppose all but one are disease-free, i.e., $\sum_{i=1}^{n} x_i = 1$.

[3] We could certainly consider other choices of a, b but we consider these choices for simplicity. You'll look at other choices in lab/homework.

The Bayes procedure

The **Bayes procedure** is to minimize the posterior expected loss

$$\rho(c, x) = \mathbb{E}(\ell(\boldsymbol{\theta}, c)|x) = \int \ell(\theta, c)p(\theta|x)d\theta$$

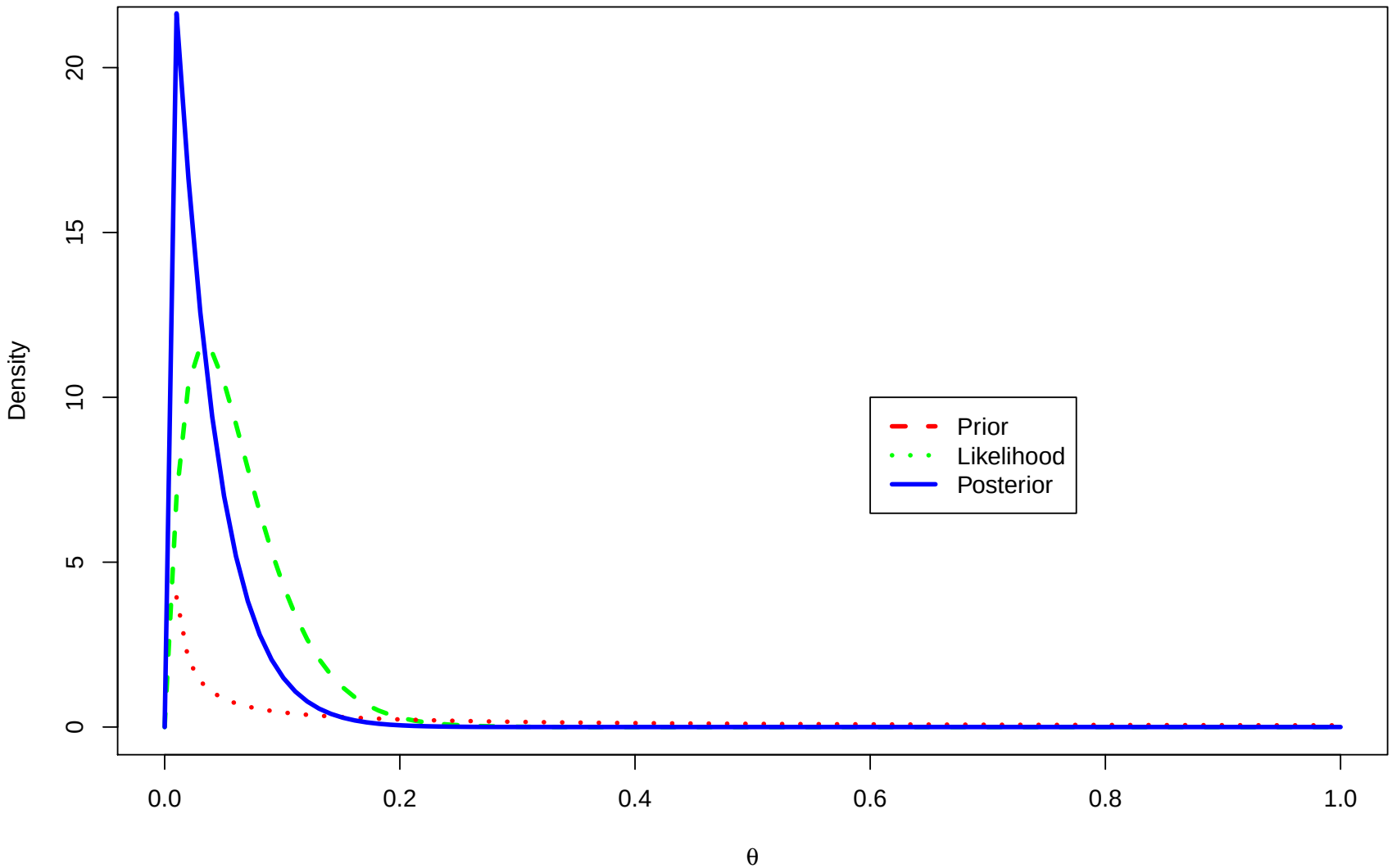
where $x = x_{1:n}$.

1. We know $p(\theta|x)$ as an updated Beta, so we can numerically compute this integral for each c.
2. Figure 1 shows $\rho(c, x)$ for our example.
3. The minimum occurs at $c \approx 0.08$, so under the assumptions above, this is the optimal amount of resources to allocate.
4. How would one perform a sensitivity analysis of the prior assumptions?

Resource allocation for disease prediction in R

```r
## set seed
set.seed(123)
## data and number of total obs.
sum_x <- 1
n <- 30
# prior parameters
a <- 0.05
b <- 1
# posterior parameters
an <- a + sum_x
bn <- b + n - sum_x
# seq of theta values
th <- seq(0, 1, length.out = 100)
## likelihood, prior, posterior
like <- dbeta(th, sum_x + 1, n - sum_x + 1)
prior <- dbeta(th, a, b)
post <- dbeta(th, sum_x + a, n - sum_x + b)
```

Likelihood, Prior, and Posterior



The loss function

```
## Compute the loss given theta and c.
loss_function <- function(theta, c) {
  if (c < theta) {
    return(10 * abs(theta - c))
  }else {
    return(abs(theta - c))
  }
}
```

Posterior risk

```r
# Compute the posterior risk given c.
# S is the number of random draws.
posterior_risk <- function(c, s = 30000) {
  # Randow draws from posterior distribution,
  # which is a beta with params an and bn.
  theta <- rbeta(s, an, bn)
  # Calculating values of the loss times the posterior.
  loss <- apply(as.matrix(theta), 1, loss_function, c)
  # average values from the loss function (integral)
  mean(loss)
}
```

Posterior Risk (continued)

```r
# a sequence of c in [0, 0.5]
c <- seq(0, 0.5, by = 0.01)
post_risk <- apply(as.matrix(c), 1, posterior_risk)
head(post_risk)
```

```
## [1] 0.33917940 0.25367603 0.18868962 0.14489894 0.11693
```

Posterior expected loss/posterior risk for disease prevelance

```
# Plot posterior risk against c.
pdf(file = "posterior-risk.pdf")
plot(c, post_risk, type = "l", col = "blue",
    lwd = 3, ylab = "p(c, x)")
dev.off()
```

```
## pdf
##   2
```

```
# minimum of posterior risk occurs at c = 0.08
(c[which.min(post_risk)])
```

```
## [1] 0.08
```

Posterior expected loss/posterior risk for disease prevelance

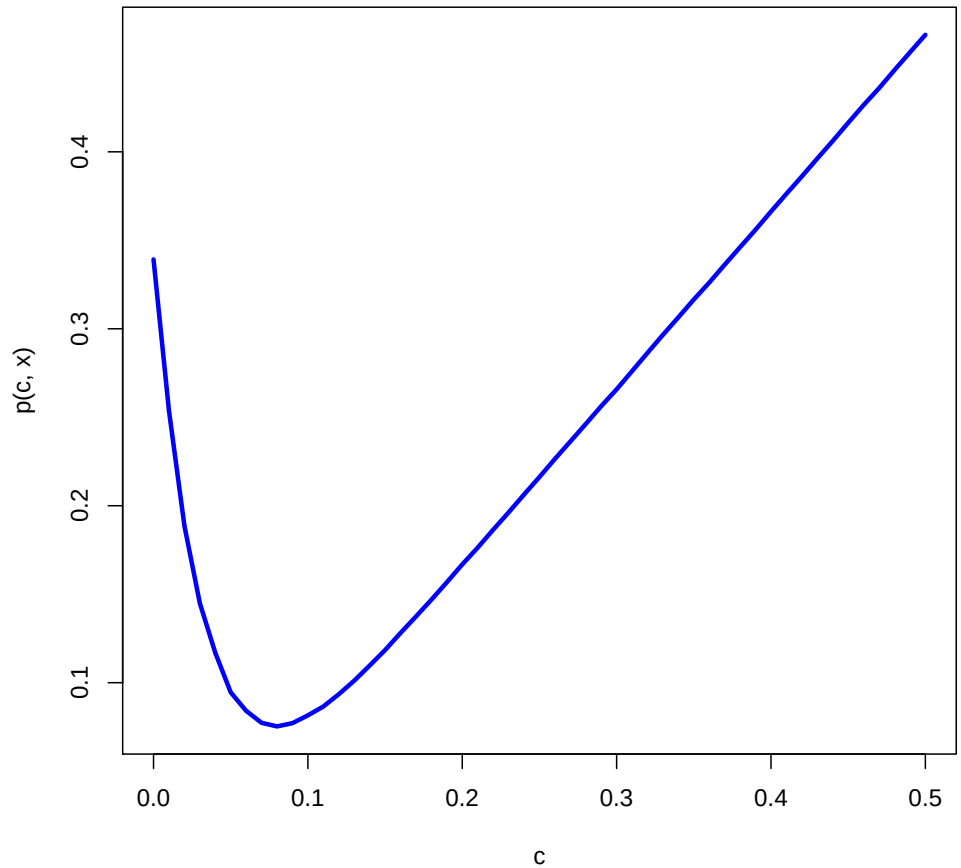


Figure 1:

Sensitivity Analysis

Suppose now that $a = 0.05, 1, 0.05$ and $b = 1, 2, 10$.

If we have different prior, the posterior risk is minimized at different c values. The optimal c depends on not only the data, but also the prior setting.

Posterior Risk Function (More Advanced)

```
# Compute the posterior risk given c.
# S is the number of random draws.
posterior_risk <- function(c, a_prior, b_prior,
                           sum_x, n, s = 30000) {
  # Randow draws from beta distribution.
  a_post <- a_prior + sum_x
  b_post <- b_prior + n - sum_x
  theta <- rbeta(s, a_post, b_post)
  loss <- apply(as.matrix(theta), 1, loss_function, c)
  # average values from the loss function
  mean(loss)
}
```

Posterior Risk Function (More Advanced)

```r
# a sequence of c in [0, 0.5]
c <- seq(0, 0.5, by = 0.01)
post_risk <- apply(as.matrix(c), 1, posterior_risk,
                   a, b, sum_x, n)
head(post_risk)
```

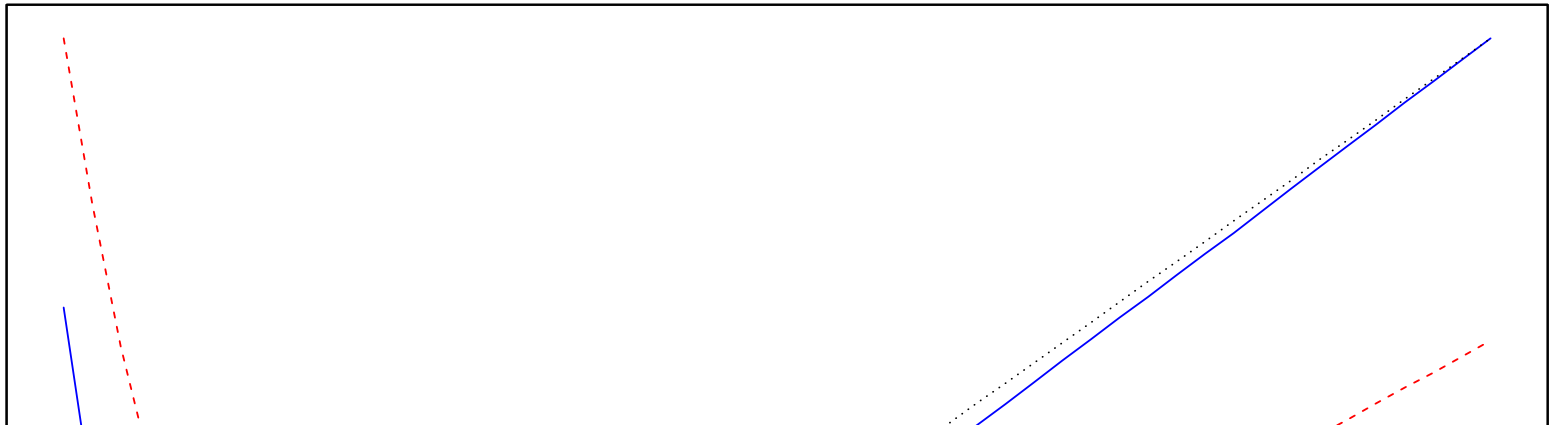
```
## [1] 0.33742709 0.25432988 0.19124960 0.14450410 0.11565
```

Sensitivity Analysis

```
# set prior values
as <- c(0.05, 1, 0.05)
bs <- c(1, 1, 10)
post_risk <- matrix(NA, 3, length(c))
# for each pair of a and b, compute the posterior risks
for (i in 1:3) {
  a_prior <- as[i]
  b_prior <- bs[i]
# Using advanced function of posterior risk.
post_risk[i,  ] <- apply(as.matrix(c),  1,
                         posterior_risk,  a_prior,
                         b_prior,  sum_x,  n)
}
```

Plot

```
plot(c,  post_risk[1,  ], type = "l",
     col = "blue",  lty = 1,  yaxt = "n",  ylab = "p(c, x)"
par(new = TRUE)
plot(c, post_risk[2, ], type = "l",
     col = "red", lty = 2, yaxt = "n", ylab = "")
par(new = TRUE)
plot(c, post_risk[3, ], type = "l",
     lty = 3, yaxt = "n", ylab = "")
legend("bottomright", lty = c(1, 2, 3),
       col = c("blue", "red", "black"),
       legend = c("a = 0.05 b = 1",
                  "a = 1 b = 1", "a = 0.05 b = 5"))
```



Optimal resources (a,b vary)

For $a = 0.05, 1, 0.05$ and $b = 1, 2, 10$ respectively, the optimal value for c is:

```
(c[which.min(post_risk[1, ])])
```

```
## [1] 0.08
```

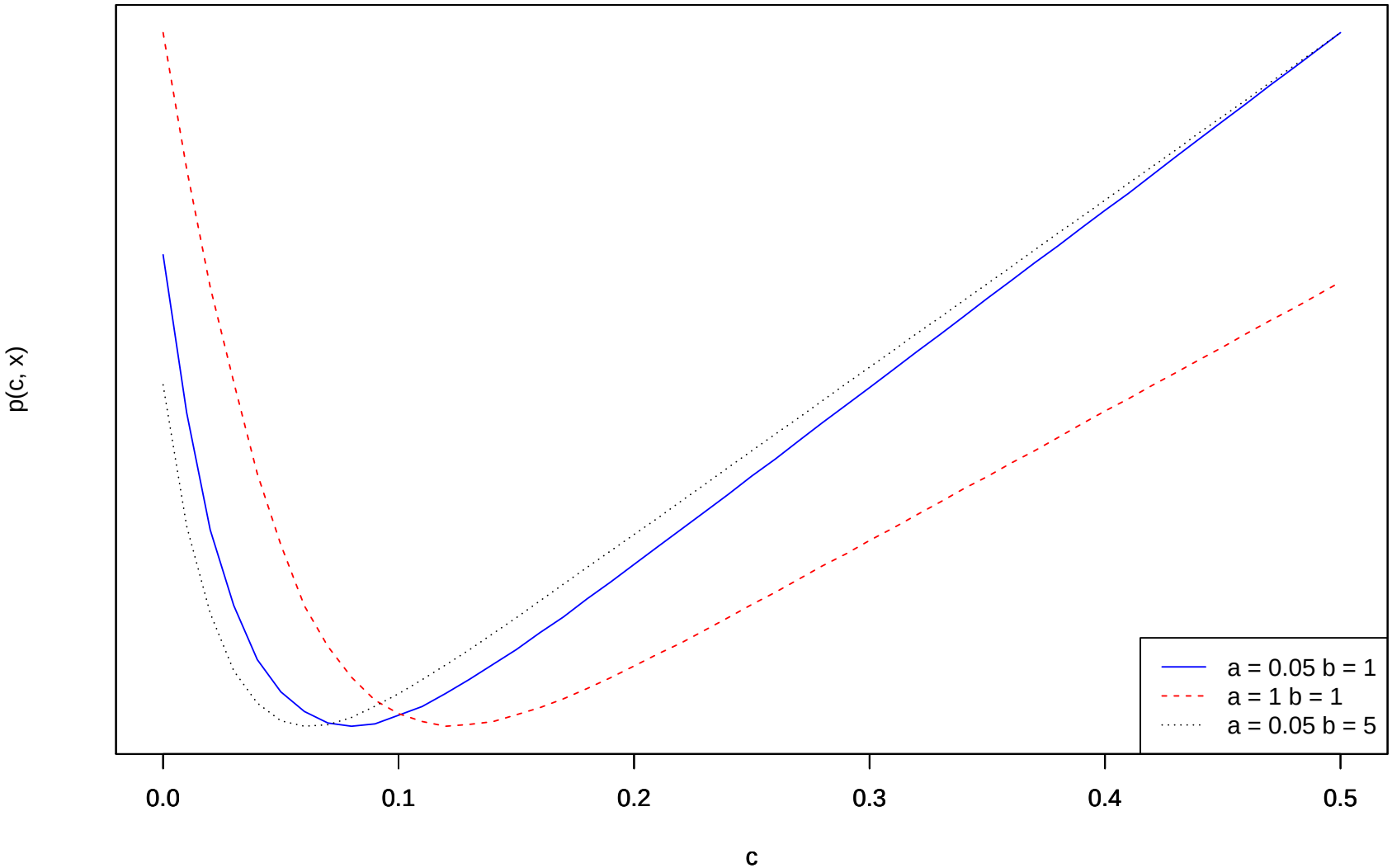
```
(c[which.min(post_risk[2, ])])
```

```
## [1] 0.12
```

```
(c[which.min(post_risk[3, ])])
```

```
## [1] 0.06
```

Plot



Frequentist Risk

1. Consider a decision problem in which $S = \boldsymbol{\theta}$.
2. The ***risk*** (or ***frequentist risk***) associated with a decision procedure δ is

$$R(\theta, \delta) = \mathbb{E}\big(\ell(\boldsymbol{\theta}, \delta(X)) \mid \boldsymbol{\theta} = \theta\big),$$

where X has distribution $p(x|\boldsymbol{\theta})$. In other words,

$$R(\theta, \delta) = \int \ell(\theta, \delta(x))\, p(x|\theta)\, dx$$

if X is continuous, while the integral is replaced with a sum if X is discrete.

Example: Resource allocation, revisited

1. The frequentist risk provides a useful way to compare decision procedures in a prior-free way.
2. In addition to the Bayes procedure or Bayes rule that we have considered earlier in the lecture, consider two other potential optimal decision rules: choosing $c = \bar{x}$ (sample mean) or $c = 0.1$ (constant).[4]
3. Remark: both the frequentist rules are looking an optimal estimator in a prior free way. (There are many other examples, but we'll just look at two simple cases.)

[4] Recall: The Bayes rule minimizes the posterior risk with respect to the parameter of interest.

Example: Resource allocation, revisited

3. Figure 2 shows each procedure as a function of $\sum x_i$, the observed number of diseased cases. For the prior we have chosen, the Bayes procedure always picks c to be a little bigger than $\bar{x}$.

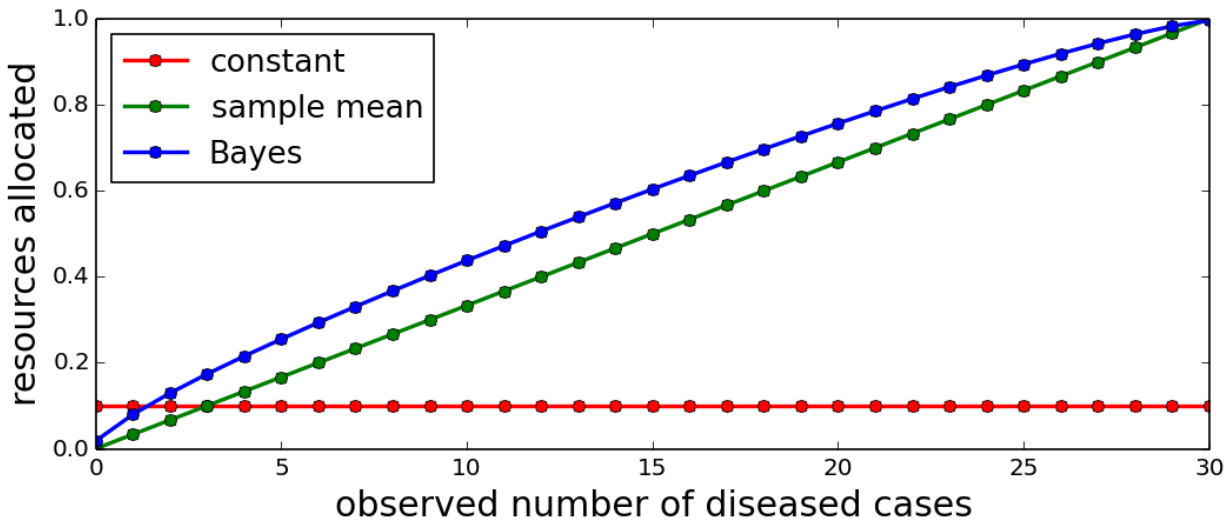


Figure 2: Resources allocated c, as a function of the number of diseased individuals observed, $\sum x_i$, for the three different procedures.

Example: Resource allocation, revisited

4. Figure 3 shows the risk $R(\theta, \delta)$ as a function of θ for each procedure. Smaller risk is better. (Recall that for each θ, the risk is the expected loss, averaging over all possible data sets. The observed data doesn't factor into it at all.)

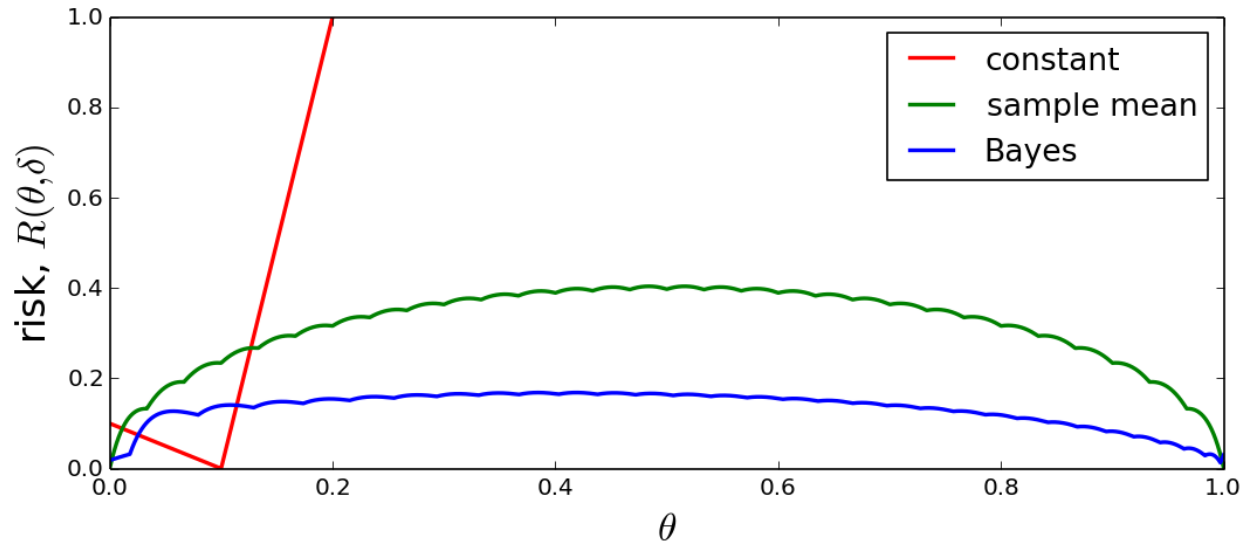


Figure 3: Risk functions for the three different procedures.

Example: Resource allocation, revisited

5. The constant procedure is fantastic when θ is near 0.1, but gets very bad very quickly for larger θ. The Bayes procedure is better than the sample mean for nearly all θ's. These curves reflect the usual situation—some procedures will work better for certain θ's and some will work better for others.
6. A decision procedure which is ***inadmissible*** is one that is dominated everywhere. That is, δ is ***admissible*** if there is no δ' such that

$$R(\theta, \delta') \leq R(\theta, \delta)$$

for all θ and $R(\theta, \delta') < R(\theta, \delta)$ for at least one θ. A decision rule is admissible so long as it is not being dominated everywhere.
7. Bayes procedures are admissible under very general conditions.
8. Admissibility is nice to have, but it doesn't mean a procedure is necessarily good. Silly procedures can still be admissible—e.g., in this example, the constant procedure $c = 0.1$ is admissible too!

Takeaways

- In understanding an optimal decision rule, we first must have a parameter of interest (θ) and define an optimal estimator ($\delta(X)$ or $\hat{\theta}$).
- There are many ways to define a loss function. A few that we talked about were the 0-1, quadratic, and absolute value loss.
- Next, we define several ways of finding an optimal decision rule. There were two that we considered. We considered minimizing the posterior risk (Bayes rule) and the risk (frequentist risk).
- Finally, we defined admissible/inadmissible rules.

Detailed Takeways for Exam

- $\hat{\theta}$ is an estimator of θ
- Loss function $L(\theta, \hat{\theta})$
- Examples of loss functions
- The difference between an action and an estimator
- Posterior risk
- Decision procedure
- Bayes rule (or Bayes estimator)
- How to derive the Bayes estimator
- What is the Bayes estimator under squared error loss?
- What is the Bayes estimator under weighted squared error loss?
- When you find the Bayes estimator, what condition do you always need to check?

Detailed Takeways for Exam (Continued)

- How to approach decision theory problems such as the resource allocation problem, where you're given the set up and then you must given the model and the loss function and back up the rational here.
- How to solve for the Bayes estimator for an applied problem such as the resource allocation problem, where the Bayes estimator is NOT in a closed form solution.
- How to conduct a sensitivity analysis for a posterior analysis and report your findings.
- The frequentist risk and why it's used.
- Admissibility and how to determine if an estimator is admissible or inadmissible.

Module 2 Derivations

Module 2 Derivations can be found below:

https://github.com/resteorts/modern-bayes/tree/master/lectures ModernBayes20/lecture-2/02-class-notes