# Module 2: Introduction to Decision Theory

Rebecca C. Steorts

# Agenda

- What is decision theory?
- General setup
- Example of general set up
- Bayesian risk
- Frequentist Risk
- Integrated Risk
- An Application to Resource Allocation

# General setup

Assume an unknown state $S$ (a.k.a. the state of nature). Assume

- we receive an observation $x$,
- we take an action $a$, and
- we incur a real-valued loss $\ell(S, a)$.

| | |
|---|---|
| $S$ | state (unknown) |
| $x$ | observation (known) |
| $a$ | action |
| $\ell(s, a)$ | loss |

# Example

1. State: $S = \boldsymbol{\theta}$
2. Observation: $x = x_{1:n}$
3. Action: $a = \hat{\theta}$
4. Loss: $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ (quadratic loss, a.k.a. square loss)

Question: Why do we consider quadratic loss over other loss functions?

# Bayesian approach

- $S$ is a random variable,
- the distribution of $x$ depends on $S$,
- and the optimal decision is to choose an action $a$ that minimizes the ***posterior expected loss*** or ***posterior risk***,

$$\rho(a, x) = \mathbb{E}(\ell(S, a)|x).$$

In other words, $\rho(a, x) = \sum_s \ell(s, a)p(s|x)$ if $S$ is a discrete random variable, while if $S$ is continuous then the sum is replaced by an integral.

# Bayesian approach (continued)

1. A ***decision procedure*** $\delta$ is a systematic way of choosing actions $a$ based on observations $x$. Typically, this is a deterministic function $a = \delta(x)$ (but sometimes introducing some randomness into $a$ can be useful).
2. A ***Bayes procedure*** is a decision procedure that chooses an $a$ minimizing the posterior expected loss $\rho(a, x)$, for each $x$.
3. Note: Sometimes the loss is restricted to be nonnegative, to avoid certain pathologies.

# What is the optimal decision rule?

- Goal: Minimize the posterior risk
- First note that

$$\ell(\theta, \hat{\theta}) = \theta^2 - 2\theta\hat{\theta} + \hat{\theta}^2$$

- It then follows that the **posterior loss** is

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

which is a convex function of $\hat{\theta}$.

# What is the optimal decision rule?

We just showed that

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

Setting the derivative with respect to $\hat{\theta}$ equal to 0, and solving, we find that the minimum occurs at $\hat{\theta} = \mathbb{E}(\boldsymbol{\theta}|x_{1:n})$, **the posterior mean**.

Let's walk through this derivation together.

# What is the optimal decision rule?

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

Now, let

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

which implies that

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

Why is the solution unique?

# Resource allocation for disease prediction

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 $\theta$ 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 $\theta$ 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:

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

# 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 $b = 1$.[1]
- 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., $\sum_{i=1}^{n} x_i = 1$.

[1] 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 $x = x_{1:n}$.

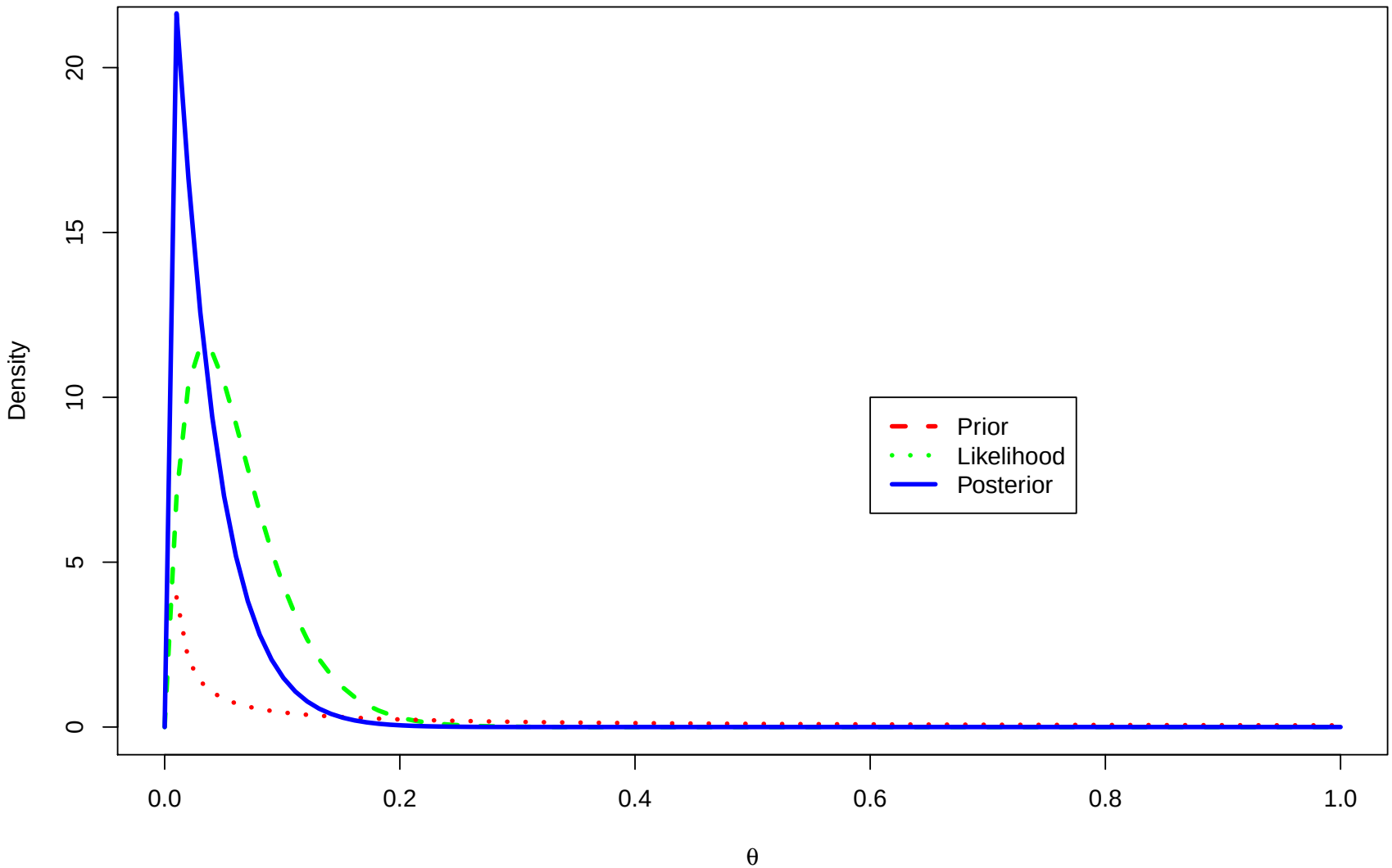
1. We know $p(\theta|x)$ as an updated Beta, so we can numerically compute this integral for each $c$.
2. Figure 1 shows $\rho(c, x)$ for our example.
3. The minimum occurs at $c \approx 0.08$, so under the assumptions above, this is the optimal amount of resources to allocate.
4. How would one perform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

# Resource allocation for disease prediction in R

```r
# set seed
set.seed(123)

# data
sum_x = 1
n = 30
# prior parameters
a = 0.05; b = 1
# posterior parameters
an = a + sum_x
bn = b + n - sum_x
th = seq(0,1,length.out = 100)
# writing the likelihood as a beta
# a trick from module 1
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)
  risk = 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

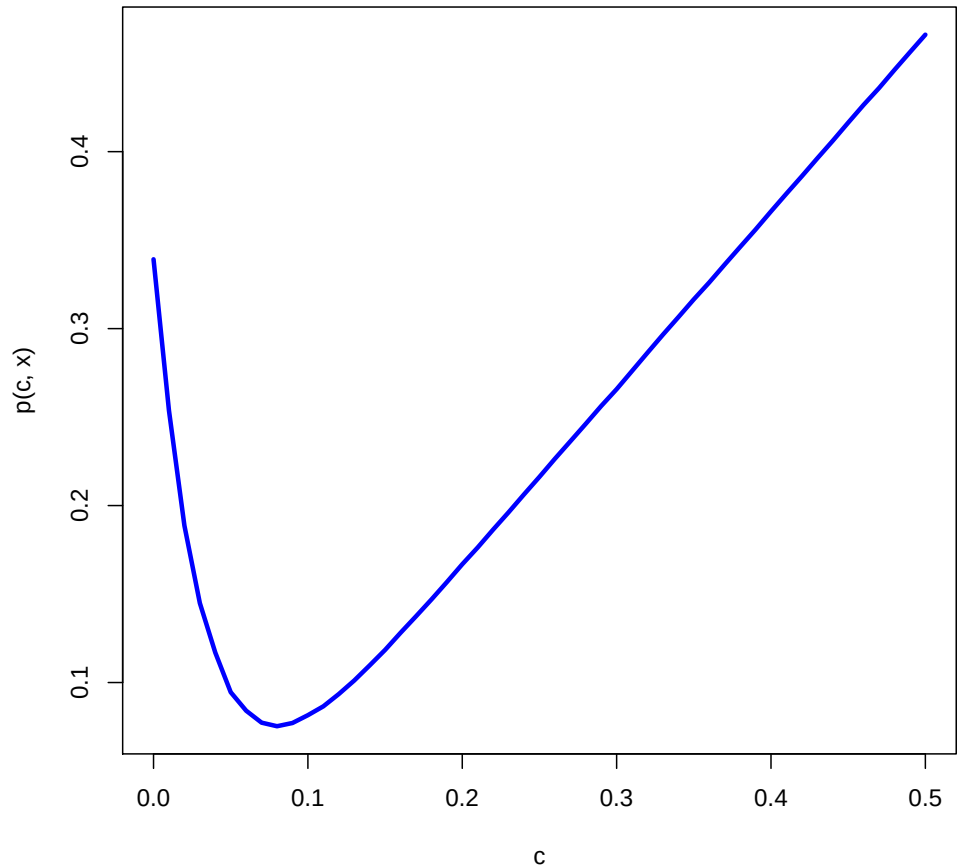



Figure 1:

# 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
  risk = mean(loss)
}
```

# Posterior Risk Function (More Advanced)

```
# 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
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 the more advanced function
  # of the 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 = T)
plot(c, post_risk[2,], type = 'l',
     col='red', lty = 2, yaxt = "n", ylab = "")
par(new = T)
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
```

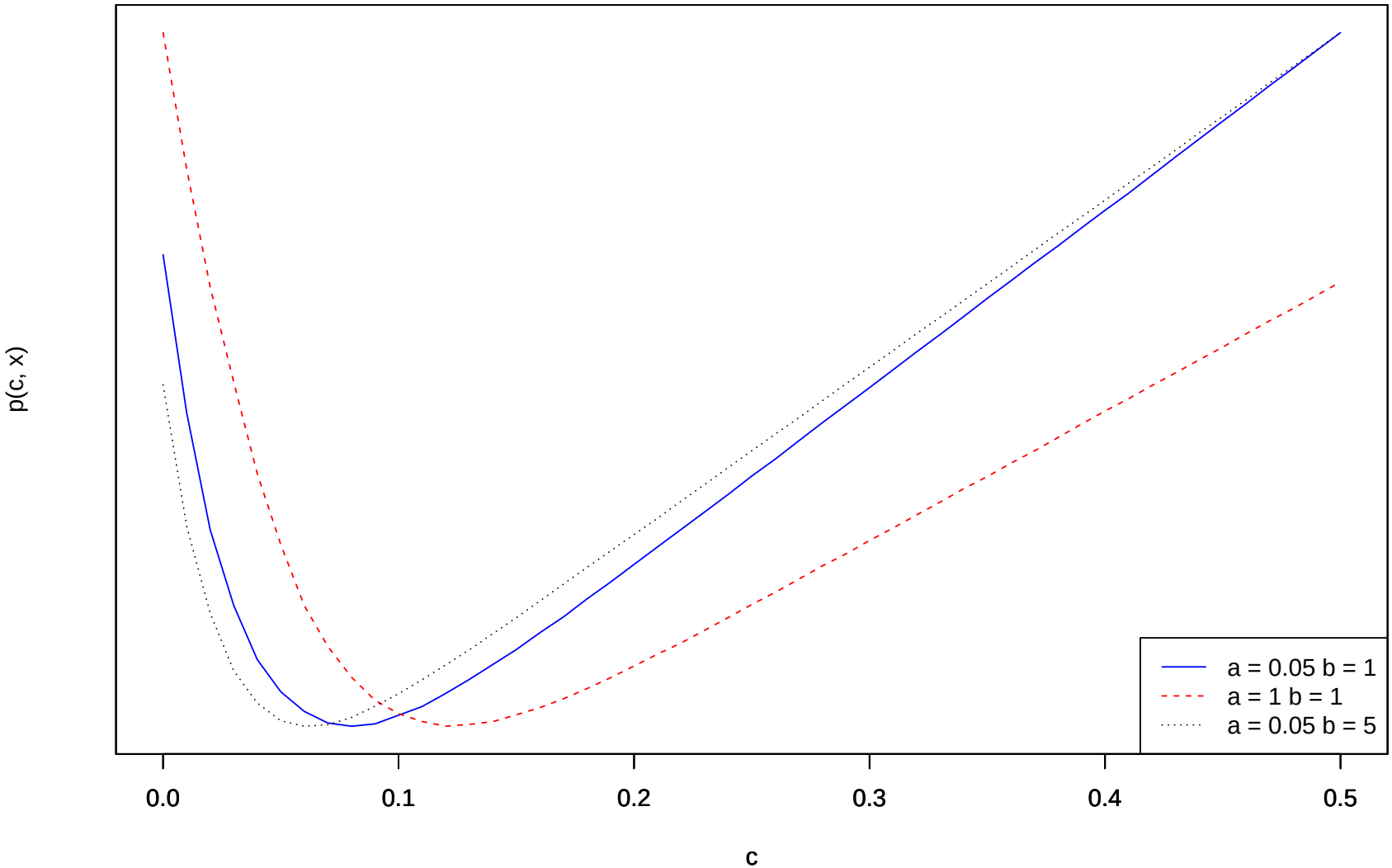
```
(c[which.min(post_risk[2,])])
```

```
## [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 $\delta$ 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.

# The integrated risk

The ***integrated risk*** associated with $\delta(X)$ is

$$r(\delta) = \mathbb{E}(\ell(\boldsymbol{\theta}, \delta(X)) = \int R(\theta, \delta)\, p(\theta)\, d\theta \tag{1}$$

$$= \int \int \ell(\theta, \delta(x))\, p(x|\theta) p(\theta)\, dx\, d\theta \tag{2}$$

# 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).[2]
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.)

[2] 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 $\bar{x}$.

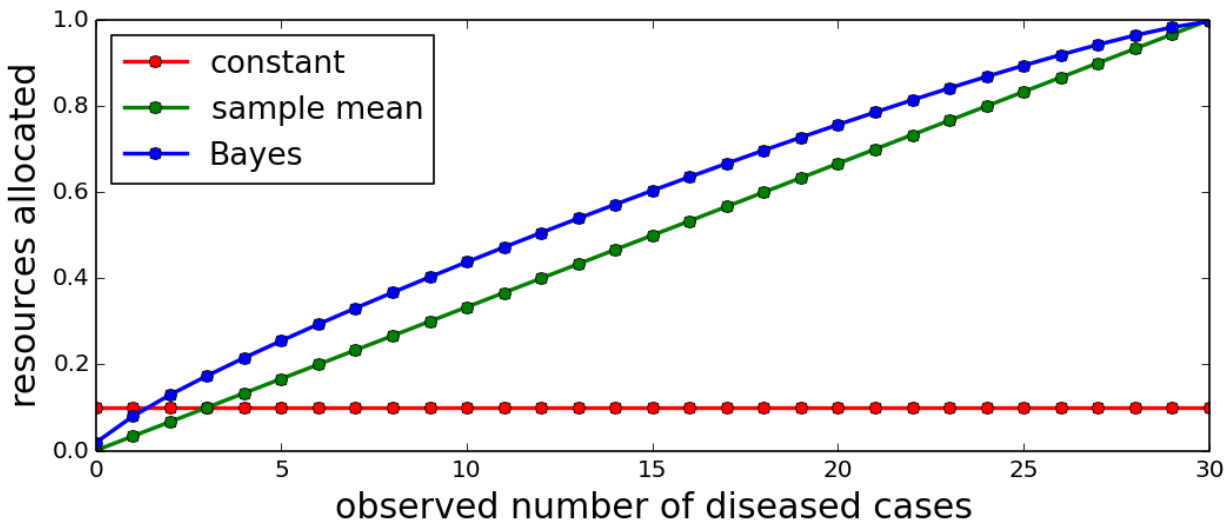



Figure 2: Resources allocated $c$, as a function of the number of diseased individuals observed, $\sum x_i$, for the three different procedures.

# Example: Resource allocation, revisited

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

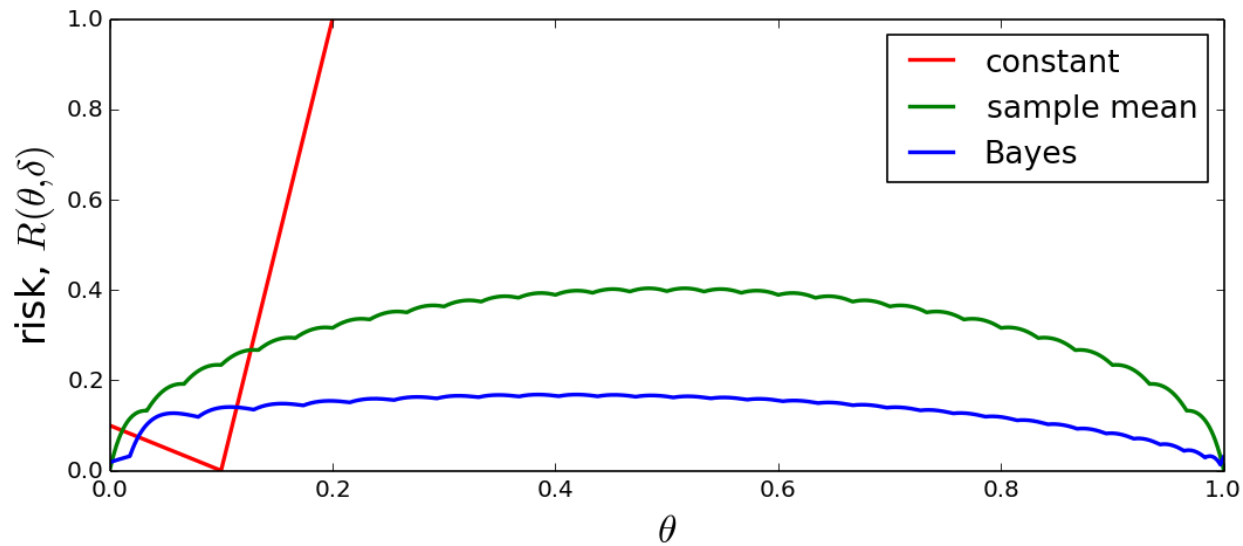



Figure 3: Risk functions for the three different procedures.

# Example: Resource allocation, revisited

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

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

for all $\theta$ and $R(\theta, \delta') < R(\theta, \delta)$ for at least one $\theta$. 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 ($\theta$) 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 three that we considered. We considered minimizing the posterior risk (Bayes rule), the risk (frequentist risk), or the integrated risk.
- Finally, we defined admissible/inadmissible rules.