Decision Theory, Chapter 12

Elizabeth Purdom

This document has last been compiled on Oct 14, 2024.

Contents

1 Decision Theory

Statistical decision theory is concerned with making decisions under uncertainty. We express our uncertainties around the problem and what information we know in terms of a probability distribution F that comes from a class of possible distributions $\mathcal{F}$. For example, we could make a decision based on data or an estimate of an important quantity, and F describes the uncertainty of that data or estimate.

The particular decision made is also referred to as an "action," and we'll denote it by a, with the collection of all possible actions denoted by $\mathcal{A}$. We assume that we have to choose a not completely knowing the true F.

To be able to decide which action a to take, we use a **loss** function to quantify the (negative) consequences of an action.

Definition 1.1 (Loss Function)**.** A loss function

$$L(F, a) : (\mathcal{F} \times \mathcal{A}) \to [0, \infty)$$

quantifies the consequences of taking an action a when the true state of nature is F. Larger values indicate more unfavorable outcomes.

The interest of decision theory and loss functions is that if the loss function is well selected, it implies what decision should be made – the decision that results in the minimum loss.

Decision theory can be formulated for general actions and losses, but we are going to focus on decision theory for estimation problems. In this case the actions are based on observed data $X \sim F$ and our actions are our estimates of some parameter θ of F,

$$a = \hat{\theta}(X),$$

so that $\mathcal{A}$ is some subset of R or R^d; or if we are estimating a discrete outcome $\mathcal{A}$ could be in Z, i.e. integer-valued.

Example: A drug company has developed a new pain reliever. They are trying to determine how much of the drug to produce, but they are uncertain about the proportion of the market the drug will capture (θ).

Suppose that we think that the entire market is size K. If we knew θ we would want to make $K\theta$ of the drug. So we want to estimate θ, and then that will tell us how much to make.

$$L(\theta,\hat{\theta}) = \begin{cases} K(\theta-\hat{\theta}) & \hat{\theta}-\theta < 0 \\ 2K(\hat{\theta}-\theta) & \hat{\theta}-\theta \geq 0 \end{cases}$$

This loss function implies that an overestimate of demand (leading to overproduction of the drug) is considered twice as costly as an underestimate. The loss is also taken to be linear in the difference of our estimate and the truth, which may be reasonable if the total cost is proportional to the number of units produced.

Standard Loss Functions Many decision theory problems for estimation are based on the following "standard" loss functions. These are expressed in generic "units of utility."

- Squared error loss: $L(\theta,\hat{\theta}) = (\theta-\hat{\theta})^2$
- Linear loss: $L(\theta,\hat{\theta}) = \begin{cases} K_1(\theta-\hat{\theta}) & \hat{\theta}-\theta < 0 \\ K_2(\hat{\theta}-\theta) & \hat{\theta}-\theta \geq 0 \end{cases}$
- Absolute error loss: $L(\theta,\hat{\theta}) = |\theta-\hat{\theta}|$ (linear loss with $K_1 = K_2$)
- L^p loss: $L(\theta,\hat{\theta}) = |\theta-\hat{\theta}|^p$
- Zero-one loss: $L(\theta,\hat{\theta}) = \begin{cases} 0 & \hat{\theta} = \theta \\ 1 & \hat{\theta} \neq \theta \end{cases}$

Note that Zero-one loss only makes sense for a predicting a discrete outcome, like predicting what category an object should be in.

2 Comparing Estimators with Risk

2.1 Risk

The actual loss depends on the unknown θ, so and the random data that creates the estimator $\hat{\theta}$. Instead, we may consider an "expected loss" and then choose an "optimal" decision with respect to this. This "expected loss" is known as risk. However, there are several ways of thinking about the expectation; hence, several different risks.[1]

[1] In the definitions that follow, we assume that θ is continuous when we define distributions on θ, but if we are estimating θ which takes on discrete values, the definitions of the expectations simply change to sums with the appropriate p.m.f.

1. **Risk** The frequentist risk (or sometimes just "risk")

$$R(\theta, \hat{\theta}) = E_\theta\left(L(\theta, \hat{\theta})\right) = \int L(\theta, \hat{\theta}(x)) f(x|\theta) dx$$

averages over different possible realizations x of the random variable, given that the true "state of nature" is θ.

It is a function of the unknown θ, as well as the particular choice of $\hat{\theta}$.

From a Bayesian Perspective, you can also write the frequentist risk as the expected loss condition on θ,[2]

$$E(L(\theta, \hat{\theta})|\theta)$$

2. **Posterior risk** As the name implies, this is a Bayesian concept, and is the expected loss conditional on the data,

$$r(\hat{\theta}|X) = E(L(\theta, \hat{\theta})|X) = \int L(\theta, \hat{\theta}(X)) f(\theta|X) d\theta$$

averages over uncertainty in θ after conditioning on observations X.

Note that conditional on X, $\hat{\theta}$ is not random, and the expectation is based on the distribution $\theta|X$, so this risk is a function of X (and the choice of $\hat{\theta}$), but not of θ.

3. **Bayes risk**: This is the risk considering the distribution of both X and θ. As the name implies, it is also a Bayesian definition, assuming θ has a distribution $f(\theta)$

$$\begin{aligned} r(f, \hat{\theta}) &= E(L(\theta, \hat{\theta})) \\ &= E(E(L(\theta, \hat{\theta})|\theta)) = E(R(\theta, \hat{\theta})) \\ &= E(E(L(\theta, \hat{\theta})|X)) = E(r(\hat{\theta}|X)) \end{aligned}$$

It depends on the particular form of $\hat{\theta}$ (but is no longer dependent on the unknown θ).

Notice the Bayes risk averages over both the other two types of risk above:

- averages over $R(\theta, \hat{\theta})$ over the likely values of θ (based on the prior $f(\theta)$)
- averages the posterior risk $r(\hat{\theta}|X)$ over likely values of X (based on the marginal distribution $f(x)$ of X)

[2]Really, the notation $E_\theta(X)$ is a frequentist way of indicating the dependence of the expectation on θ without writing $E(X|\theta)$, which for a frequentist would make no sense since θ is not random.

Example Suppose $\bar{X}_n \sim N(\theta, 1/n)$ and we are estimating θ under squared error loss. Consider $\hat{\theta}_c(X) = c\bar{X}_n$, where c is chosen ahead of time (non-random)

- Then our risk $R(\theta, \hat{\theta})$ is

$$\begin{aligned} R(\theta, \hat{\theta}) &= E_\theta(\theta - \hat{\theta})^2 \\ &= E_\theta(\theta - c\bar{X}_n)^2 \\ &= E_\theta(\theta^2 - 2c\theta\bar{X} + c^2\bar{X}^2) \\ &= \frac{c^2}{n} + (1 - 2c + c^2)\theta^2 \end{aligned}$$

- Consider the prior distribution $\theta \sim N(0, b^2)$. Then we have that

$$\theta|\bar{X}_n \sim N(\frac{b^2\bar{X}_n}{b^2 + 1/n}, \frac{b^2/n}{b^2 + 1/n})$$

  Our posterior risk $r(\hat{\theta}|X)$ is

$$\begin{aligned} r(\hat{\theta}|\bar{X}_n) &= E((\theta - \hat{\theta})^2|\bar{X}_n) \\ &= E((\theta - c\bar{X}_n)^2|\bar{X}_n) \\ &= var(\theta|X) + (E(\theta|\bar{X}))^2 - 2c\bar{X}E(\theta|\bar{X}) + c^2\bar{X}^2 \\ &= \frac{b^2}{b^2 + \frac{1}{n}}\left[\frac{1}{n} + \left(\frac{b^2}{b^2 + 1/n} - 2c + \frac{b^2 + 1/n}{b^2}c^2\right)\bar{X}^2\right] \end{aligned}$$

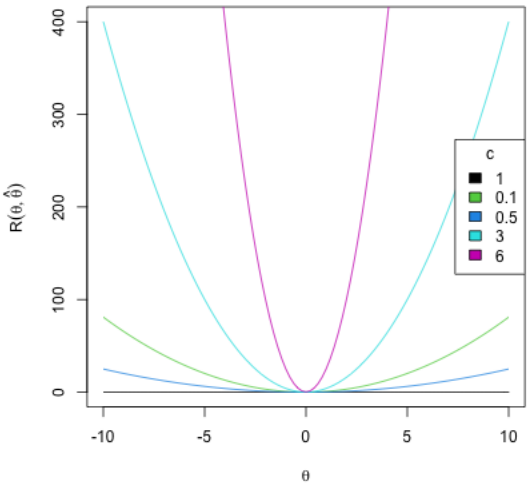
- Consider calculating the Bayes risk for our normal example under the same prior distribution $\theta \sim N(0, b^2)$:

$$\begin{aligned} r(\hat{\theta}) &= E(R(\theta, \hat{\theta})) \\ &= E\left(\frac{c^2}{n} + (1 - 2c + c^2)\theta^2\right) \\ &= \frac{c^2}{n} + (1 - 2c + c^2)b^2 \end{aligned}$$

Bayesian or Frequentist? Notice that we are defining ideas with respect to both Bayesian and Frequentist probability. Both Bayesian and Frequentists use decision-theoretic ideas (i.e. loss and risk) in considering the best estimation procedure to use. For Bayesians, the range of possible estimators will be those based on the posterior distribution, but there is still a question of what single estimator from the posterior distribution to use (mean, mode, median,...). But you will probably see far more treatment of the risk of estimators in Frequentist theory, where there are potentially unlimited numbers of estimators; decision theoretic ideas allow for comparison of the vast world of frequentist possibilities. For this reason, many of the definitions and treatments are from a frequentist point of view, even when using Bayesian definitions of risk.

2.2 Admissible estimators

Let's go back to the question of choosing a $\hat{\theta}$ based on our risk in of our example above. In this case, it's a question of the best c. If we look at the frequentist risk, $R(\theta, \hat{\theta})$, we have a function of θ for each choice of $\hat{\theta}$:



Which of these estimators (i.e. values of c) would be best?

Definition 2.1 (Inadmissible)**.** An estimator $\hat{\theta}$ is **inadmissible** if there exists another rule $\hat{\theta}'$ such that

$$\begin{aligned} R(\theta, \hat{\theta}') &\leq R(\theta, \hat{\theta}) \text{ for all } \theta \\ R(\theta, \hat{\theta}') &< R(\theta, \hat{\theta}) \text{ for at least one } \theta \end{aligned}$$

Otherwise $\hat{\theta}$ is called admissible.

Notice that admissibility of estimators is a frequentist definition and depends on the frequentist risk.

Back to our Example We can see that in our above example, $\hat{\theta} = c\bar{X}$, this estimator is not admissible for any of the values of $c \neq 1$, because $c = 1$ is always better. Does this mean that for $c = 1$ this estimator is admissible?

Another Example (Wasserman) Consider estimating p from $X_i \sim Bernoulli(p)$ and two different estimators

$$\hat{p}_1 = \bar{X}, \hat{p}_2 = \frac{\sum_i X_i + a}{a + b + n}$$

for some constants a and b. $\hat{p}_2$ is the estimator if we were bayesian and set the prior to be $Beta(a, b)$, while $\hat{p}_1$ is the standard frequentist estimate. $\hat{p}_2$ actually depends on the constants, a, b, it's actually a whole class of estimates. To be concrete, I will set $\hat{p}_2$ with $a = b = \sqrt{n/4}$ in what follows.

Then with squared error loss we have that the risk of an estimator is its mean squared error,

$$R(\theta, \hat{\theta}) = var(\hat{\theta}) + (bias(\hat{\theta}))^2$$

For our two Bernoulli estimates, this gives us,

$$R(p, \hat{p}_1) = var(\bar{X}) = \frac{p(1-p)}{n}$$

and

$$\begin{aligned} R(p, \hat{p}_2) &= var(\hat{p}_2) + (bias(\hat{p}_2))^2 \\ &= \frac{np(1-p)}{(a+b+n)^2} + \left(\frac{np + a}{a + b + n} - p\right)^2 \\ &= \frac{1}{(a+b+n)^2}\left\{np(1-p) + [(1-p)a - pb]^2\right\} \end{aligned}$$

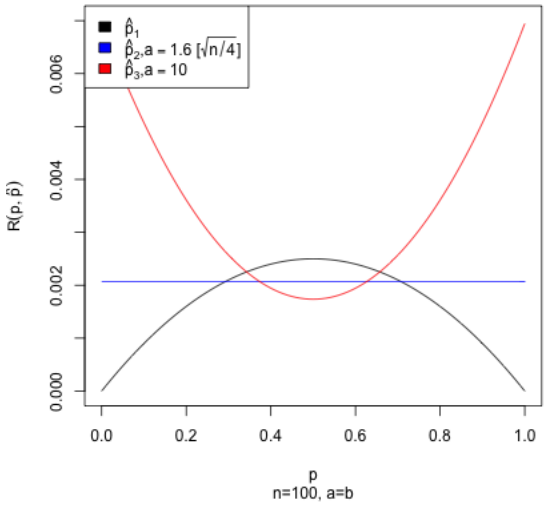
Notice with the choice of $a = b = \sqrt{n/4}$ we have

$$\hat{p}_2 = \frac{\sum_i X_i + \sqrt{n/4}}{\sqrt{n} + n}$$

which gives a constant risk for $\hat{p}_2$,

$$R(p, \hat{p}_2) = \frac{n}{4(n + \sqrt{n})^2}$$

We can compare the risk of these estimators. I also the risk of a third estimator, $\hat{p}_3$, based on choosing $a = b = 10$.



These estimators perform better under different values of p (neither one dominates the other). Which means that neither estimator shows anything about the inadmissibility of the other – they could be both admissible, for example, if there is no estimator that dominates either of them; or one or the other or both could be inadmissible because you can find estimators that dominate them.

Notice that inadmissiblity can be shown by example, in the sense that you simply need to construct a single example of an estimator that dominates an estimator and that shows it is inadmissible. Proving admissibility means showing there is no other estimate that dominates it; this can be much trickier.

Admissibility is a minimal standard – don't be uniformally worse that something else. But it can be suprising what standard estimators are not admissible, particularly in high dimensions, as we will see with the Stein Paradox below.

3 Comparing Estimators with Risk

Admissibility isn't enough to pick among estimators. We can have multiple admissible estimators with different performances for different θ. To decide on the "best" estimator, we need a one-number summary of our risk, not just a function, to compare different estimators and pick the best one. There are two different summaries we will consider:

- **Bayes Risk** as defined above, which averages the risk, relative to the probability of the parameter.

In our above example, consider the uniform prior $f(\theta) = 1$. [3] Then we have,

$$r(\hat{p}_1, f) = \frac{1}{n}\int_0^1 p(1-p)dp = 1/6n$$
$$r(\hat{p}_2, f) = \frac{n}{4(n+\sqrt{n})^2}\int_0^1 1dp = \frac{n}{4(n+\sqrt{(}n))^2}$$

Then for $n > 20$, $r(\hat{p}_1, f) < r(\hat{p}_2, f)$ so we would choose $\hat{p}_1 = \bar{X}$.

- **Maximum Risk** $\bar{R}(\hat{\theta}) = \sup_\theta R(\theta, \hat{\theta})$ This compares each estimator's "worst-case" scenario. In our previous example, with $a = b = \sqrt{n/4}$,

$$\bar{R}(\hat{p}_1) = \max_{0\leq p\leq 1}\frac{p(1-p)}{n} = \frac{1}{4n}$$
$$\bar{R}(\hat{p}_2) = \max_{0\leq p\leq 1}\frac{n}{4(n+\sqrt{(}n))^2} = \frac{n}{4(n+\sqrt{(}n))^2}$$

Therefore, under maximum risk critera, $\bar{R}(\hat{p}_2) < \bar{R}(\hat{p}_1)$ so we would pick $\hat{p}_2$ over $\hat{p}_1$.

Note that while the worst case for $\hat{p}_2$ is better than that of $\hat{p}_1$, $\hat{p}_1$ is better for a wider range of values (when $n = 100$) than p_2.

We can see that how we evaluate our estimator will change which is "better".

A Best Estimator Above we used a simple comparison of just two estimators. But could be other estimators that are better than both $\hat{p}_1$ or $\hat{p}_2$. For example, $\hat{p}_1$ is NOT the estimator that minimizes the Bayes risk for a uniform prior. So the real question is to look at *all* estimators and find a $\hat{p}$ that minimizes the Bayes risk across all estimators, or minimizes the maximum risk across all estimators (and there might be more than one, since it might not be a unique estimator that minimizes these risks).

3.1 Bayes Rule

A decision rule that minimizes the Bayes risk is called a Bayes rule.

Definition 3.1 (Bayes Rule)**.** The estimator $\hat{\theta}$ is a Bayes rule, or Bayes estimator (under a particular prior f and loss function) if

$$r(f, \hat{\theta}) = \inf_{\tilde{\theta}} r(f, \tilde{\theta})$$

[3] Notice above I said $\hat{p}_2$ was the estimator based on a Beta distribution, but I am considering it just as an estimator and can pick a different prior, even though that is incoherent for my Bayesian analysis.

Minimizing Posterior Risk is sufficient Notice that while we defined the Bayes rule based on minimizing the Bayesian Risk, it is equivalent to find the Bayes estimator using the posterior risk. We can see this as follows:

For each X, let $\hat{\theta}(X)$ be the value of $\hat{\theta}$ that minimizes the posterior risk $r(\hat{\theta}|X)$. (Recall that for each X, $r(\hat{\theta}|X)$ returns a single number for each $\hat{\theta}$.) The estimator defined in this way is the Bayes estimator. This is because

$$r(f, \hat{\theta}) = E(r(\hat{\theta}|X)) = \int r(\hat{\theta}|x) f(x)$$

If $\hat{\theta}$ minimizes $r(\hat{\theta}|X)$ being integrated *for each* X, then we've also minimized the integral of $r(\hat{\theta}|X)$, regardless of what $f(x)$ is.

Example In our above example, with $\bar{X}|\theta \sim N(\theta, 1/n)$, and $\theta \sim N(0, b^2)$, then we have that

$$\theta|\bar{X}_n \sim N(\frac{b^2\bar{X}_n}{b^2 + 1/n}, \frac{b^2/n}{b^2 + 1/n})$$

We want to find the estimator $\hat{\theta}$ that minimizes

$$r(f, \hat{\theta}) = E((\theta - \hat{\theta}(X))^2|X)$$

Well, generally if we have a random variable Y and want to find the c that minimizes

$$E(Y - c)^2$$

this is given by $E(Y)$. So we have that the Bayes rule is

$$\hat{\theta} = E(\theta|X)$$

which in this case is

$$\frac{b^2\bar{X}_n}{b^2 + 1/n}$$

Bayes rule for common loss functions Note that the above is a general phenomena, and didn't depend on the particulars of the problem above for squared error loss. By the same logic, we can calculate the Bayes rule explicitly for several standard loss functions:

- Squared error loss: posterior mean
- Absolute error loss: posterior median
- Zero-one loss: posterior mode

Back to Bernoulli Example Our estimator

$$\hat{p}_2 = \frac{\sum_i X_i + \sqrt{n/4}}{\sqrt{n} + n}$$

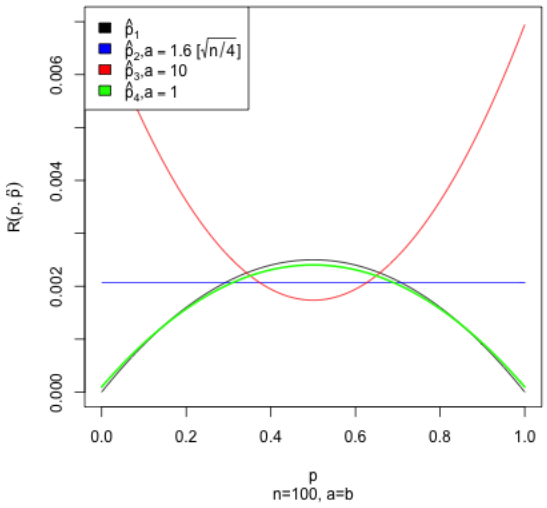
is the posterior mean when the prior is $Beta(\sqrt{n/4}, \sqrt{n/4})$. It is therefore the Bayes rule (i.e. minimizes the Bayes risk) for this prior.

That seems contrary to what I said above, where $\hat{p}_1 = \bar{X}_n$ had *lower* Bayes risk than $\hat{p}_2$. However, the Bayes risk calculation we did was based on a uniform prior for p (a uniform is equivalent to $Beta(1, 1)$). $\hat{p}_2$ is NOT the Bayes estimator for a uniform prior; it's the Bayes estimator for a prior with $Beta(\sqrt{n/4}, \sqrt{n/4})$.

Let's define yet another estimator $\hat{p}_4$ that is the Bayes estimator for a uniform prior ($a = b = 1$). It's risk is

$$R(p, \hat{p}_4) = \frac{1}{(2+n)^2}\left(np(1-p) + [1-2p]^2\right)$$

We can plot its risk and see that it is quite similar to that of $\bar{X}_n$:



We can calculate the Bayes risk of $\hat{p}_4$ under a uniform prior:

$$\begin{aligned} r(f, \hat{p}_4) &= \int_0^1 R(p, \hat{p}_4) f(p) dp \\ &= \frac{1}{(2+n)^2} \int_0^1 np(1-p) + [1-2p]^2 dp \\ &= \frac{1}{(2+n)^2} * \left(1 + \frac{1}{2}(n-4) + \frac{1}{3}(4-n)\right) \\ &= \frac{1}{6(2+n)} \end{aligned}$$

Recall, that the risk of $\hat{p}_1$ was $\frac{1}{6n} > \frac{1}{6(2+n)}$. So as claimed, the posterior mean *assuming the Uniform prior* has lower Bayes risk than $\hat{p}_1$.

Exercise Suppose $X_1, \ldots, X_n|\sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$, where θ is known. Let the prior distribution for σ^2 be inverse gamma with parameters a and b. The prior PDF is

$$f(\sigma^2; a, b) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-a-1}\exp\{-b/\sigma^2\}$$

- What is the posterior distribution of σ^2?

- What is the Bayes estimator under squared error loss?
- What is the Bayes estimator under absolute error loss?
- What is the Bayes estimator under zero-one loss?

You may use the fact the mean of an $InverseGamma(a, b)$ distribution is $b/(a-1)$ when $a > 1$, the mode is $b/(a+1)$, and the median is not available in closed form.

3.1.1 Admissibility of Bayes Rules

One final note about Bayes rules: under weak conditions, they are admissible. The intuition for this is that if there existed a rule that had lower risk, it would also have lower Bayes risk. We have to be a little bit careful, because a Bayes rule depends on a particular choice of prior and the loss function, and you could potentially choose very strange priors.

One simple condition is that if our Bayes rule is unique (i.e. there is a unique Bayes estimator that minimizes the Bayes risk), it is admissible:

Theorem 1 (Unique Bayes Rules are Admissible)**.** *Any unique Bayes rule (with respect to a prior f) is admissible.*

Proof. By contradition: Suppose their exists $\tilde{\theta}$ that dominates $\hat{\theta}^f$, i.e. $R(\theta, \tilde{\theta}) \leq R(\theta, \hat{\theta})$ for all θ and strictly so for one θ. Then this implies that

$$\int R(\theta, \tilde{\theta}) f(\theta) d\theta \leq \int R(\theta, \hat{\theta}) f(\theta) d\theta$$

However, by the definition, $\hat{\theta}$ *uniquely* minimizes the Bayes risk (i.e. the above integral), so this is a contradition. □

Non-unique Bayes rules Non-unique Bayes rules can also be admissible. Here is one set of conditions, given by Wasserman, for when they are admissible (but there are others).

Theorem 2 (Bayes Rules are Admissible)**.** *Suppose that* $\Theta \subseteq \mathbb{R}$ *and that* $R(\theta, \hat{\theta})$ *is a continuous function of* θ *for every* $\hat{\theta}$*. Let* f *be a prior density that assigns positive probability to any open subset of* Θ*. Let* $\hat{\theta}^f$ *be a Bayes rule, with finite Bayes risk. Then* $\hat{\theta}^f$ *is admissible.*

Implications A couple of points:

- This is a frequentist criteria (admissibility) used to evaluate Bayesian estimators.
- All posterior means are admissible estimators for squared error loss of the parameter.
- The above criteria are for *proper* priors
- The admissibility of Bayes estimators means that frequentist theory frequently makes use Bayesian estimators for showing admissibility of estimators

Harder Example: $\bar{X}$ For example, if $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$, then you can show that $\bar{X}$ is admissible. This is proved through relating $\bar{X}$ to Bayesian estimators. $\bar{X}_n$ is not itself a Bayes estimate for a proper prior. But we've seen that it is the *limit* of Bayes estimates which do have proper priors. We've seen that if the prior of θ is a $N(0, b^2)$, then the posterior mean is

$$\hat{\theta}_{Bayes} = E(\theta|X) = \frac{b^2 \bar{X}_n}{b^2 + \sigma^2/n}.$$

This is the unique Bayes estimator under squared error loss, and therefore is admissible for θ. As $b^2 \to \infty$ the Bayesian estimator approaches $\bar{X}_n$, so $\bar{X}_n$ is the limit of admissible estimators.

This turns out to be usually enough to show estimators are admissible; indeed under certain conditions, you can show that admissible estimators must be limits of Bayesian estimators (if and only if statement):

Theorem 3 (Stein's necessary and Sufficient Condition for Admissibility[4])**.** *A decision rule* $\hat{\theta}$ *is admissibile if and only if there exists a sequence* f_n *of prior distributions with support on a finite set such that*

$$\hat{\theta}^{f_n} \to \hat{\theta}$$

[4]I am not giving a precise statement of the theorem. See Berger *Statistical Decision Theory and Bayesian Analysis* p. 546 or Lehmann *Theory of Point Estimation* p. 382

where $\hat{\theta}^{f_n}$ is the Bayes rule for prior f_n.

3.2 Minimax Rules

We'll now consider the other strategy for choosing an action, based on the minimizing the maximum risk, called a **minimax rule**.

Definition 3.2 (Minimax Rule)**.** An estimator $\hat{\theta}$ that minimizes the maximum risk is called a **minimax rule**:

$$\sup_{\theta} R(\theta, \hat{\theta}) = \inf_{\tilde{\theta}} \sup_{\theta} R(\theta, \tilde{\theta})$$

This is a "worst-case scenario" type of estimator.

Example Suppose $\bar{X} \sim N(\theta, 1/n)$ and we are estimating θ under squared error loss. Consider again $\hat{\theta}_c = c\bar{X}$. It's risk is

$$\frac{c^2}{n} + (c-1)^2\theta^2$$

So unless $c = 1$, we have

$$\sup_{\theta} R(\theta, \hat{\theta}_c) = \infty$$

But if $c = 1$ ($\hat{\theta}_n = \bar{X}$), then the risk is the same for all θ and we have,

$$\sup_{\theta} R(\theta, \hat{\theta}_c) = \frac{1}{n},$$

and the maximum risk is bounded.

Note this isn't over all estimators $\hat{\theta}$, but just of the form $\hat{\theta}_c$. So this doesn't show that $\bar{X}$ is a minimax rule.

3.2.1 Connection of Admissibility and minimax

There is are some relationships between admissibility and minimax estimators.

Theorem 4. *Suppose that $\hat{\theta}$ has constant risk and is admissible. Then $\hat{\theta}$ is minimax.*

Proof. If $\hat{\theta}$ were not minimax, then there exists a rule $\tilde{\theta}$ such that

$$R(\theta, \tilde{\theta}) \leq \sup_{\theta} R(\theta, \tilde{\theta}) < \sup_{\theta} R(\theta, \hat{\theta}) = c$$

Therefore the risk of $\tilde{\theta}$ would always be less than that of $\hat{\theta}$ which is a contradition of the fact that $\hat{\theta}$ is admissible. □

Therefore, returning to our example with $\bar{X} \sim N(\theta, 1/n)$, we know $\bar{X}_n$ is admissible for squared error loss. Because it has constant risk (i.e. it's risk doesn't depend on θ), this is enough to show that it is minimax for squared error loss.[5]

Generally, minimax estimators do not have to be admissibile. However, if an estimator is the *unique* minimax estimator for θ, then it is admissible.

To sum up the relationship between admissibility and Minimax:

- Admissible Estimators + Constant Risk ⇒ Minimax
- Unique Minimax ⇒ Admissible

Exercise An investor is deciding whether or not to purchase \$1000 of risky ZZZ bonds. If the investor buys the bonds, they can be redeemed at maturity for a net gain of \$500. There could, however, be a default on the bonds, in which case the original \$1000 investment would be lost. If the investor doesn't buy the bonds, she will put her money in a "safe" investment, for which she will be guaranteed a net gain of \$300 over the same time period. She estimates the probability of a default to be 0.1.

- Describe the parameter space Θ and the space of possible actions $\mathcal{A}$.
- What is the prior distribution?
- For each possible $\theta \in \Theta$ and $a \in \mathcal{A}$, compute the loss.
- Is any action inadmissible?
- Explain why the Bayes rule is for the investor to buy the bonds
- What is the minimax strategy?

[5] In fact it is minimax for any "reasonable" loss function (the only estimator with this property)

3.2.2 Connection of Bayes rule and minimax

In general it can be difficult to find minimax rules when the parameter space is infinite. Above, we have a link between admissibility, but as we've seen the most reliable way to find admissibility is to show it is a Bayes Rule.

This suggests looking at the relationship between Bayes rules and minimax.

- A Bayes estimator $\hat{\theta}^f$ which has constant risk is minimax.

  Theorem 5 (Wasserman 12.11)**.** *Suppose that $\hat{\theta}^f$ is the Bayes rule with respect to some prior f. Suppose further that $\hat{\theta}^f$ has constant risk: $R(\theta, \hat{\theta}^f) = c$ for some c. Then $\hat{\theta}^f$ is minimax.*

  We know Bayes estimators are generally admissibile, so this is similar to our statement above.

- Another way we can think about this is to start with the class of all Bayes rules, meaning considering the Bayes rules across *all* priors, $f(\theta)$. Bayes rules are generally pretty good – they are basically all admissible, for example, and we know they minimize the average risk over some choice of "weighting" the possible θ.

  Indeed, for any $\hat{\theta}$ the maximum risk is greater than the Bayes risk,

  $$r(f, \hat{\theta}) = E_f(R(\theta, \hat{\theta})|\theta) \leq E_f(\sup_\theta R(\theta, \hat{\theta})|\theta) = \sup_\theta R(\theta, \hat{\theta})$$

  So if we take the infinimum over all estimators, we get that the Bayes risk of a Bayes rule is always less than the minimax risk:

  $$inf_{\hat{\theta}} r(f, \hat{\theta}) = r(f, \hat{\theta}^f) \leq inf_{\hat{\theta}} \sup_\theta R(\theta, \hat{\theta}) = \text{Minimax Risk}$$

  We'd like this to be equal; in otherwords, we would like to find a Bayes rule whose Bayes risk is equal to the minimax risk. Then it would *also* be a minimax rule.

  So this suggests that we look across all possible priors to find the prior f that has the *largest* Bayes risk among all priors. We call such a distribution a **least favorable prior**.[6]

  How can we then go the final step and say that the least favorable prior actually achieves the minimax risk? The following will give it to us:

[6] Lehman, p. 310. This is a slightly different description of the definition in Wasserman, but more intuitive.

Theorem 6 (Wasserman 12.10)**.** *Suppose that $\hat{\theta}^f$ is the Bayes rule with respect to some prior f, i.e.*

$$r(f, \hat{\theta}^f) = \inf_{\tilde{\theta}} r(f, \tilde{\theta}).$$

Suppose that

$$R(\theta, \hat{\theta}^f) \leq r(f, \hat{\theta}^f) \quad \forall \theta$$

Then $\hat{\theta}^f$ is minimax.

Namely the frequentist risk of our Bayes estimator $\hat{\theta}^f$ is less than or equal to its Bayes risk.

You might ask, how is this possible? But recall our prior can put more mass on some θ than others to get the Bayes risk. We get to pick our prior, including point masses on particular values of θ, so it is clearly putting mass the θ with largest risk so that its average (the Bayes risk) is equal to its maximum risk.

Notice that the theorem didn't specify that f is least favorable prior, but if it satisfies the conditions of the above theorem, it will be least favorable, in terms of having lowest Bayesian risk across all possible priors.

- A *unique* bayes estimator $\hat{\theta}^f$ which has Bayes risk equal to $\sup_\theta R(\theta, \hat{\theta}^f)$ is the *unique* minimax procedure and the prior f is least favorable

Example Previously we considered $X|p \sim Bin(n, p)$. One of our estimators, was

$$p_2 = \frac{X + a}{a + b + n}.$$

This is the posterior mean with prior $Beta(a, b)$. This means that $\hat{p}_2$ is the Bayes estimator for squared error loss for this prior. Then the risk we found to be

$$R(p, \hat{p}_2) = \frac{np(1-p)}{(a+b+n)^2} + \left(\frac{np + a}{a + b + n} - p\right)^2$$

If we set $a = b = \sqrt{n/4}$, then we have that $R(p, \hat{p}_2)$ is a constant value. So $\hat{p}_2$ is a minimax estimator.

Exercise: Suppose $X|p \sim Bin(n, p)$ and the loss is squared error.

- Show $\hat{p} = X/n$ is not minimax. *Hint: Consider the randomized estimator*

$$\tilde{p} = \left\{ \begin{array}{ll} X/n & \text{with probability } 1 - \frac{1}{n+1} \\ 1/2 & \text{with probability } \frac{1}{n+1} \end{array} \right\}$$

Exercise: Consider a decision problem with possible states of nature θ_1 and θ_2 . Let X be a random variable with probability function $p(x|\theta)$:
$P(X = 0|\theta_1) = 0.2, P(X = 1|\theta_1) = 0.8;$
$P(X = 0|\theta_2) = 0.4, P(X = 1|\theta_2) = 0.6.$

Two non-randomized actions a_1 and a_2 are considered with the following loss function:
$L(\theta_1, a_1(0)) = 1, L(\theta_1, a_1(1)) = 2, L(\theta_1, a_2(0)) = 4, L(\theta_1, a_2(1)) = 0;$
$L(\theta_2, a_1(0)) = 3, L(\theta_2, a_1(1)) = 1, L(\theta_2, a_2(0)) = 1, L(\theta_2, a_2(1)) = 4.$

1. Give and plot the risk set $S = \{(r_1, r_2) : r_1 = \lambda R(\theta_1, a_1) + (1-\lambda)R(\theta_1, a_2), r_2 = \lambda R(\theta_2, a_1) + (1-\lambda)R(\theta_2, a_2), \lambda \in [0, 1]\}$.
2. Suppose θ has the prior distribution $\Lambda(\theta)$ defined by $P(\theta = \theta_1) = 0.9, P(\theta = \theta_2) = 0.1$. What is the Bayes rule with respect to $\Lambda(\theta)$?
3. Find the minimax rule(s).

Writing the frequentist risk of action a as $R(\theta, a)$, the maximum risk

$$\bar{R}(a) = \sup_{\theta} R(\theta, a)$$

Which action in the example minimizes $\bar{R}(a)$?

4 MLE, Shrinkage estimators and Stein's Paradox

If we return to our standard MLE estimator $\hat{\theta}$, we could ask whether MLE estimators are admissible? Are they minimax? Are they approximately Bayes? We saw that for the example of normal data that the answer to these questions is "Yes" for $\bar{X}$. More generally, for estimating a univariate parameter θ (based on a parametric model), the MLE will be approximately minimax and Bayes.

However, when we look at multivariate parameters, we will see that MLE performs poorly.

But first let's be clear about loss functions for multivariate. Again, you can choose many different functions, but the most common is squared error loss, which we generalize to

$$L(\theta, \hat{\theta}) = \sum_{j=1}^{p} (\theta_j - \hat{\theta}_j)^2 = ||\theta - \hat{\theta}||^2$$

Notice that we can write the frequentist loss as a bias-variance trade off, just as for a univariate θ. Let $\mu = E(\hat{\theta}) \in R^p$ and $\Sigma = var(\hat{\theta}) \in R^{p \times p}$ be the variance-covariance matrix of $\hat{\theta}$. We have the same type of bias-variance trade off of (frequentist) risk as before:

$$\begin{aligned} E_\theta L(\theta, \hat{\theta}) &= E||\theta - \hat{\theta}||^2 \\ &= E||\theta - \mu + \mu - \hat{\theta}||^2 \\ &= E||\theta - \mu||^2 + E||\mu - \hat{\theta}||^2 - \underbrace{2E(\theta - \mu)'(\mu - \hat{\theta})}_{=0} \\ &= E||\theta - \mu||^2 + tr(\Sigma) \end{aligned}$$

The Simplest Setup The simplest case we have always consider can be boiled down to this: Suppose we observe single data value $X \sim N(\theta, \sigma^2)$, with σ^2 known. Usually we think of multiple data $X_1, \ldots, X_n$ i.i.d $N(\theta, \sigma^2)$, but you can think of replacing the single value X above $\bar{X}$ and σ^2 with σ^2/n, and you will see that it doesn't change anything meaningful about the problem. So we will stick with a single value $X \sim N(\theta, \sigma^2)$.

How do we estimate θ? In this case, our MLE is $\hat{\theta}^{MLE} = X$.

How can we expand this simple example? We instead assume that the value we observe X is a vector in R^p. For simplicity, we are going to again present it as a single observation X_j from each mean θ_j, but if $X_j = \bar{Y}_j$ with for each j $Y_{ij} \sim N(\theta_j, \tau^2)$, then everything holds for $\sigma^2 = \tau^2/n$. Or X_j can be used for any other setting where we are trying to estimate the different means of normal variables (like regression, though there we don't have independence).

What is our model for X? The simplest explanation is that each element X_j is completely independent of each other, $X_j \overset{indep}{\sim} N(\theta_j, \sigma^2)$. We could write this more succinctly in terms of our vector of data X as a multivariate normal,

$$X \sim N_p(\theta, \sigma^2 I_p)$$

The MLE for θ is

$$\hat{\theta}_j^{MLE} = X_j.$$

$\hat{\theta}_j^{MLE}$ is unbiased and is also minimax for squared error loss..

Stein Paradox In 1956 Charles Stein shocked everyone by showing that if $p \geq 3$, the vector $X = \hat{\theta}_j^{MLE}$ is inadmissible for squared error loss.[7] He gave an estimator, the **James-Stein estimator**, which dominates X, thus making X inadmissible. The estimator is given as:

$$\hat{\theta}^{JS} = \left(1 - (p-2)\frac{\sigma^2}{\sum X_i^2}\right)^+ X$$

where $(z)^+ = max\{z, 0\}$ [8] In fact, the James-Stein estimator is *also* not admissible.

Notice that this estimator is equivalently,

$$\hat{\theta}^{MLE} - (p-2)\frac{\sigma^2}{||X||^2}X,$$

or 0 if this quantity is negative. So that the estimator of $\hat{\theta}_j$ is "shrunken" or made closer to 0. And the amount they will be closer is an amount that depends on all of the data via $\sum X_i^2 = ||X||^2$.

This is called a **shrinkage estimator**, where the original estimator X is shrunk toward the origin (i.e. each component of X is made smaller). Notice also that this shrinkage estimator is a biased estimator.

Why is this a "paradox"? Notice that the individual components of X are *independent* – they have nothing to do with each other with entirely different means (though importantly we are assuming they share the same variance). Yet to get a better estimation of each individual mean, we will combine this data . To put this in context, the individual X_i could be measuring entirely different things that have nothing to do with each other (cars and batting averages, to use the example of Efron), yet somehow we do better in estimating their individual means when we use all of the data.

It's also important to understand that our notion of doing "better" is based on the cumulative squared error loss of all the dimensions of θ. It doesn't mean for each individual element of θ we are doing better.

[7] For $p = 1$ or $p = 2$, then X is the *unique* minimax estimator, and is thus guaranteed to be admissible. But $p > 2$, X is not a unique minimax estimator.

[8] This is a modification of the original J-S estimator. The original James-Stein estimator was

$$\hat{\theta}^{JS} = \left(1 - (p-2)\frac{\sigma^2}{\sum X_j^2}\right) X$$

and did not take the positive part, and hence could change the sign of the estimate. The modified JS estimator given here has lower risk than the original JS estimator and is more interpretable.

The Bayes Estimator If instead we created a Bayesian model, we would put a prior on the vector θ. A logical prior is

$$\theta \sim N_p(0, b^2 I_p)$$

Then we would have

$$\theta|X \sim N\left(\frac{b^2}{b^2+\sigma^2}X, \left(\frac{1}{b^2}+\frac{1}{\sigma^2}\right)^{-1}\right)$$

The Bayes rule that minimizes squared error loss is the posterior mean,

$$\hat{\theta}_{Bayes} = \frac{b^2}{b^2+\sigma^2}X = (1 - \frac{\sigma^2}{b^2+\sigma^2})X$$

So we see that the (admissible) Bayes estimator will multiply each component of X by a factor, again creating a shrinkage estimator. The Bayes estimator shrinks X toward it's prior mean, in this case the origin 0, like the $J-S$ estimator. It doesn't use the data to decide how much to shrink, it uses the prior/known parameters, b and σ^2.

This is a common phenomena we've seen where Bayesian estimators adjust or correct MLE estimators toward the mean.

Empirical Bayes There is a link between the James-Stein Estimator and the Bayes estimator. Under the prior $\theta \sim N_p(0, b^2 I_p)$, the marginal distribution of X can be calculated. Notice that $f(x)$, the marginal distribution, depends on the prior parameter b. It can be shown that

$$\frac{b^2+\sigma^2}{||X||^2}$$

is distributed as an inverse chi-squared distribution with p df so that

$$E\left(1-(p-2)\frac{\sigma^2}{\sum X_i^2}\right) = \frac{b^2}{b^2+\sigma^2}$$

where the expectation is taken over the marginal distribution of X. In other words, the shrinkage factor in the James-Stein estimator is an unbiased estimator of $\frac{b^2}{b^2+\sigma^2}$, based on the marginal distribution of X of the Bayes shrinkage factor.

This is an example of what is known as **Empirical Bayes** estimation. Empirical Bayes estimation consists of using a Bayesian estimator relying on a specific prior, but rather than assuming the prior parameter is known, estimating the prior parameter

from the marginal distribution of X using frequentist estimation techniques (MLE, MOM, etc). Specifically

$$X|\theta \sim f(x|\theta), \theta \sim f(\theta;\tau)$$
$$\hat{\theta}_{Bayes} = E_\tau(\theta|X)$$

Then estimate $\hat{\tau}$ from the marginal distribution of the data $f(x;\tau)$ to get

$$\hat{\theta}_{EB} = E_{\hat{\tau}}(\theta|X)$$

Empirical Bayes estimators are frequently shrikage estimators, just like Bayesian estimators.

Moral of the Story Stein first demonstrated in the absolute simplest case what is now a widely understood as a general phenomena regarding MLE estimators. Specifically, when estimating a multivariate parameter of dimension p, the MLE does not do well in high dimensions. This has significant ramifications in a wide variety of problems, and modern statistical methods frequently involve different types of shrinkage or *regularization* to improve the performance of the estimator. Maximizing the likelihood is still a fundamental building block of modern statistics, but the maximization is generally not solely of the likelihood, but done with additional constraints or modifications of the likelihood.

Furthermore, this is also an example of the limits of considering unbiased estimators. Statistical theory traditionally focused a great deal on unbiased estimators (and $\bar{X}$ is the best *unbiased* estimator in terms of risk). Ultimately, it is now widely understood that in more complicated problems, shrinkage estimators greatly reduce the variance of the estimator at the cost of introducing a small amounts of bias and as a result greatly improve the risk.