# Foundations of Statistical Inference

Julien Berestycki

Department of Statistics
University of Oxford

MT 2016



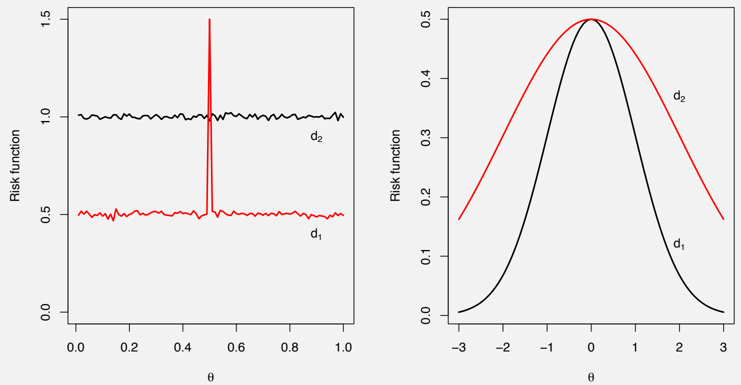
Lecture 11 : Finding minimax rules. Hypothesis testing with loss functions.



## Minimax rules

**Definition** A rule $\delta$ is a minimax rule if $\max_\theta R(\theta, \delta) \leq \max_\theta R(\theta, \delta')$ for any other rule $\delta'$. It minimizes the maximum risk.

Sometimes this doesn't produce a sensible choice of decision rule.






## Finding minimax rules

**Theorem 1** If $\delta$ is a Bayes rule for prior $\pi$, with $r(\pi, \delta) = C$, and $\delta_0$ is a rule for which $\max_\theta R(\theta, \delta_0) = C$, then $\delta_0$ is minimax.

**Proof** If for some other rule $\delta'$, $\max_\theta R(\theta, \delta') = C - \epsilon$ for some $\epsilon > 0$ (so $\delta_0$ is not minimax), then

$$\begin{aligned} r(\pi, \delta') &= \int R(\theta, \delta')\pi(\theta)d\theta \\ &\leq \int (C - \epsilon)\pi(\theta)d\theta \\ &= (C - \epsilon) \\ &< r(\pi, \delta) \end{aligned}$$

so $\delta$ is not the Bayes rule for $\pi$, a contradiction. [*This is an informal treatment which assumes the min and max exist - see Y&S Ch 2 Sec 2.6*]

## Finding minimax rules

**Theorem 2** If $\delta$ is a Bayes rule for prior $\pi$ with the property that $R(\theta,\delta)$ does not depend on $\theta$, then $\delta$ is minimax.

**Proof** (Y&S Ch 2) Let $R(\theta,\delta) = C$ (no $\theta$ dependence). This implies

$$r(\pi,\delta) = \int R(\theta,\delta)\pi(\theta)d\theta = C.$$

If for some other rule $\delta'$, $\max_\theta R(\theta,\delta') = C - \epsilon$ for some $\epsilon > 0$ (so $\delta_0$ is not minimax), then we have $r(\pi,\delta') \le C - \epsilon$. But $r(\pi,\delta) = C$ so $\delta$ is not the Bayes rule for $\pi$, a contradiction.



## Finding minimax rules

Theorem 2 tells us that

**The Bayes estimator with constant risk is minimax.**

This result is useful, as it gives an approach to finding minimax rules.

Bayes rules are sometimes easy to compute, so if we find a prior that yields a Bayes rule with constant risk for all $\theta$ we have the minimax rule.



## Example : minimax estimator for quadratic loss

$X$ is Binomial $(n,\theta)$, and the prior $\pi(\theta)$ is a Beta $(\alpha,\beta)$ distribution. For a quadratic loss function, the Bayes estimator is $(\alpha+X)/(\alpha+\beta+n)$

The risk function is

$$\begin{aligned}
\mathbb{E}_{X|\theta}[(\widehat{\theta}-\theta)^2] &= \text{MSE}(\widehat{\theta}) = [\text{Bias}(\widehat{\theta})]^2 + \text{Var}[\widehat{\theta}] \\
&= \left[\theta - \mathbb{E}\left(\frac{\alpha+X}{\alpha+\beta+n}\right)\right]^2 + \text{Var}\left[\frac{\alpha+X}{\alpha+\beta+n}\right] \\
&= \left[\theta - \left(\frac{\alpha+n\theta}{\alpha+\beta+n}\right)\right]^2 + \frac{n\theta(1-\theta)}{(\alpha+\beta+n)^2} \\
&= \frac{[\theta(\alpha+\beta)-\alpha]^2 + n\theta(1-\theta)}{[\alpha+\beta+n]^2}
\end{aligned}$$

The Bayes estimator with constant risk is minimax. This occurs when $\alpha = \beta = \sqrt{n}/2$, so the minimax estimator using quadratic loss is $(\alpha+x)/(\alpha+\beta+n) = (x+\sqrt{n}/2)/(n+\sqrt{n})$.



## Hypothesis testing with loss functions

Suppose $X_i \sim f(x;\theta)$ iid for $i = 1,2,...,n$ and we wish to test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$.

The decision rule is $\delta_C$ when we use a critical region $C$ so $\delta_C(x) = H_1$ if $x \in C$ and otherwise $\delta_C(x) = H_0$.

Loss function:

$$L_S(\theta,\delta_C(x)) = \begin{cases} a & \theta=\theta_0, x\in C \\ 0 & \theta=\theta_0, x\notin C \\ b & \theta=\theta_1, x\notin C \\ 0 & \theta=\theta_1, x\in C \end{cases}$$

so the loss for Type I error is $a$ ($H_0$ holds and we accept $H_1$) and the loss for Type II error is $b$ ($H_1$ holds and we accept $H_0$).

## Hypothesis testing with loss functions

The risk function for the rule $\delta_C$ is

$$\begin{aligned} R(\theta_0;\delta_C) &= \int L_S(\theta_0,\delta_C(x))f(x;\theta_0)dx \\ &= \int a\mathbb{I}(x\in C)f(x;\theta_0)dx \\ &= a\alpha \end{aligned}$$

as $\alpha = P(X \in C|H_0)$ is the probability for Type I error.

Similarly

$$R(\theta_1;\delta_C) = b\beta,$$

as $\beta = P(X \notin C|H_1)$ is the probability for a Type II error.

**Note** $\alpha$ and $\beta$ are functions of the critical region $C$.



## Hypothesis testing with loss functions

To calculate the Bayes risk we need a prior. Let $\pi(\theta_0) = p_0$ and $\pi(\theta_1) = p_1$ be the prior probabilities that $H_0$ and $H_1$ hold.

The Bayes risk is

$$\begin{aligned} r(\pi,\delta_C) &= \sum_{\theta\in\{\theta_0,\theta_1\}} R(\theta;\delta_C)\pi(\theta) \\ &= p_0 a\alpha(C) + p_1 b\beta(C). \end{aligned}$$

The **Bayes test** chooses the critical region $C$ to minimize the Bayes risk.

**Question** How do we do this?



## Hypothesis testing with loss functions

To answer this question, let's revise what we know about Frequentist hypothesis testing.

**Neyman-Pearson lemma** The best test of size $\alpha$ of $H_0$ vs $H_1$ is a likelihood ratio test with critical region

$$C' = \left\{x; \frac{L(\theta_1;x)}{L(\theta_0;x)} \geq A\right\}$$

for some constant $A > 0$ chosen so that $P(X \in C'|H_0) = \alpha$.

By 'best test' we mean the test with the highest power, where Power = 1 - Type II error.

In frequentist statistics we fix the Type I error ($\alpha$) and this determines the value of $A$, thus defining the critical region for the test.



## Bayes tests

The following theorem tells us how to choose the critical region to minimize the Bayes risk, in the same way that the Neyman-Pearson lemma tells us how to maximize the power at fixed size.

**Theorem** The critical region for the Bayes test is the critical region for a LR test with

$$A = \frac{p_0 a}{p_1 b}$$

Every LR test is a Bayes test for some $p_0, p_1$.

## Proof

The Bayes test minimizes

$$\begin{aligned} p_0 a\alpha + p_1 b\beta &= p_0 a P(X \in C|H_0) + p_1 b P(X \in C'|H_1) \\ &= p_0 a \int_C L(\theta_0; x)dx + p_1 b \int_{C'} L(\theta_1; x)dx \\ &= p_0 a \int_C L(\theta_0; x)dx + p_1 b \Big[1 - \int_C L(\theta_1; x)dx\Big] \\ &= p_1 b + \int_C \Big[p_0 a L(\theta_0; x) - p_1 b L(\theta_1; x)\Big] dx \end{aligned}$$

So choose $C$ such that $x \in C$ iff $p_0 a L(\theta_0; x) - p_1 b L(\theta_1; x) \leq 0$, i.e. when

$$\frac{L(\theta_1; x)}{L(\theta_0; x)} \geq \frac{p_0 a}{p_1 b}$$



## Example

Let $X_1, \dots, X_n$ be $\mathrm{N}(\mu, \sigma^2)$ with $\sigma^2$ known, and we want to test $H_0 : \mu = \mu_0$ vs $H_1 : \mu = \mu_1$, with $\mu_1 > \mu_0$. The critical region for the LR test

$$\frac{L(\theta_1; x)}{L(\theta_0; x)} \geq A$$

becomes

$$\bar{x} \geq \frac{\sigma^2 \log(A)}{n(\mu_1 - \mu_0)} + \frac{1}{2}(\mu_0 + \mu_1) = B(\text{ say})$$

In the classical case we ignore the exact value of $B$, but in a Bayes test $A = p_0 a/(p_1 b)$ and we substitute into $B$. As an example take

$$\mu_0 = 0, \mu_1 = 1, \sigma^2 = 1, n = 4, a = 2, b = 1, p_0 = 1/4, p_1 = 3/4$$



Then the Bayes test has critical region

$$\bar{x} \geq \frac{1}{4}\log(\frac{1}{3} \times \frac{2}{1}) + \frac{1}{2} = \frac{1}{4}\log(\frac{2}{3}) + \frac{1}{2} = 0.399$$

For this test (using the fact that $\bar{X}$ is $N(\mu, 1/4)$)

$$\alpha = P(\bar{X} \geq 0.3999 \mid \mu = 0, \sigma^2/n = 1/4) = 0.212$$

and

$$\beta = P(\bar{X} < 0.3999 \mid \mu = 1, \sigma^2/n = 1/4) = 0.115$$

In a classical approach fixing $\alpha = 0.05$, $B = 1.645\sqrt{1/4} = 0.822$, so

$$\beta = P(\bar{X} < 0.822 \mid \mu = 1, \sigma^2 = 1/4) = 0.363$$

In the Bayes test $\alpha$ has been increased and $\beta$ decreased.



## Example

Let $X_1, \dots, X_{10}$ be i.i.d. Bernoulli$(p)$. $H_0 :\ p = 0.5$ and $H_1 : p = 0.8$ with respective prior probabilities $\pi_0$ and $\pi_1$. As usual

$$\alpha = P(\text{reject } H_0 | H_0) \text{ and } \beta = P(\text{acept } H_0 | H_1).$$

The loss function is specified by

| | $H_0$ holds | $H_1$ holds |
|---|---|---|
| Accept $H_0$ | 0 | $b$ |
| Accept $H_1$ | $a$ | 0 |

Let $\delta_B$ be the test with critical region $S = \sum X_i \geq B$. The **Bayes risk** is

$$p_0 a\alpha + p_1 b\beta$$

and the Bayes test minimizes this expression.

## Example

Suppose $a = 2, b = 1$.

| $B$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $a\alpha$ | 2.000 | 1.980 | 1.978 | 1.890 | 1.656 | 1.246 |
| $b\beta$ | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.006 |
| $B$ | 6 | 7 | 8 | 9 | 10 | 11 |
| $a\alpha$ | 0.754 | 0.344 | 0.110 | 0.022 | 0.002 | 0.000 |
| $b\beta$ | 0.033 | 0.121 | 0.322 | 0.624 | 0.893 | 1.000 |

Let $\delta_B$ be the test with critical region $S = \sum X_i \geq B$. What is $\delta_{11}$ and can it be the Bayes test? Always accept $H_0$ and yes if $p_0$ is very close to 1 (in fact as soon as $p_0 \geq 0.982$).
Similarly, $\delta_7$ is optimal when $p_0 \in (0.177, 0.462)$. Graphical interpretation?What is the minimax test? (first only for non random rules and then in general)



## Sequential procedure

With sequential procedure, sample items are drawn **one at a time**. After each sample there is a choice of stopping and making decision/estimation or to continue. In general a **cost** is associated to sampling.
.
**Example:** Suppose the $X_i$ are iid Bernoulli($\theta$) and we want to test $H_0 : \theta = 0.10$ against $H_1 : \theta = 0.20$. If we choose in advance the sample size $n$, it might become clear in hindsight that the sample was unnecessarily large or small.

- $n = 100$, $S_n = \sum_{i=1}^{n} X_i = 25$. The sample was probably too large. The evidence was clear before the 100th draw.
- $n = 100, S_n = 15$. We need a larger sample.



## Sequential procedure

A **sequential decision rule** involves

1. A stopping rule (when to stop)
2. and a terminal decision rule (upon stopping, what decision do we take).

Notation:

- $X_1, X_2, \ldots$ stream of observations, supposed iid $\sim f(x, \theta)$, and write $\underline{x}_N = (x_1, \ldots, x_n)$ for the vector of the first $n$ observations.
- The stopping rule $S$ is a stopping time in the filtration of the $X_i$. i.e. the event $S = n$ depend only on $\underline{x}_n$.
- The terminal decision rule is a family of decisions rule indexed by $n$: $d_n(\underline{x}_n)$
- $c_i =$ **cost** of observation $i$
- the total loss function is

$$L(\theta, d_S(\underline{x}_S) + \sum_{i=1}^{S} c_i.$$



## Sequential procedure

The Risk of a sequential procedure is thus

$$\begin{aligned} R(\theta, S, d_S) &= E_{\bar{x}|\theta}\left[L(\theta, d_S(\underline{x}_S) + \sum_{i=1}^{S} c_i\right] \\ &= E\left[\sum_n 1_{\{S=n\}}\left(L(\theta, d_n(\underline{x}_n) + \sum_{i=1}^{n} c_i\right)\right] \end{aligned}$$

The Bayes risk integrates this quantity against the prior $\pi(\theta)$

$$\begin{aligned} B(S, d_S) &= E_\theta\left[E_{\underline{X}|\theta}\left[L(\theta, d_S(\underline{x}_S) + \sum_{i=1}^{S} c_i\right]\right] \\ &= E_{\underline{X}}\left[E_{\theta|\underline{x}_S}\left[L(\theta, d_S(\underline{x}_S) + \sum_{i=1}^{S} c_i\right]\right] \\ &= E_{\underline{X}}\left[\sum_n 1_{\{S=n\}} E_{\theta|\underline{x}_n}\left[L(\theta, d_n(\underline{x}_n) + \sum_{i=1}^{n} c_i\right]\right] \end{aligned}$$

# Sequential procedure

$$B(S, d_S) = E_{\underline{X}}\left[\sum_n 1_{\{S=n\}} E_{\theta|\underline{x}_n}\left[L(\theta, d_n(\underline{x}_n) + \sum_{i=1}^n c_i\right]\right]$$

Therefore for each fixed $n$ we need to minimize $E_{\theta|\underline{x}_n}[L(\theta, d_n(\underline{x}_n)]$ and thus **for each $n$ fixed $d_n$ must be the Bayes rule**.
The hard part is to decide when to stop. In general: **stop** iff current Bayes risk of making the decision now is less that expected Bayes risk of taking one more observation.
Generally hard, except when:

1. For each $n$, the Bayes risk does not depend on the observation $\underline{x}_n$
2. When there is a fixed upper bound for the number of observation (using backward induction).



# Example: Backward induction

Max # of obs.= 3. $X_i \sim Ber(\theta)$ we want to test $H_0 : \theta = 1/4$ against $H_1 : \theta = 1/2$ with uniform prior $\pi_0 = \pi_1 = 1/2$. The loss function is specified by $a = b = 30$ and $c_i = 1$. Then

$$\begin{aligned} P(H_0|S_n) &= \frac{P(S_n|H_0)P(H_0)}{P(S_n|H_0)P(H_0) + P(S_n|H_1)P(H_1)} \\ &= \frac{\binom{n}{s}4^{-s}(3/4)^{n-s}}{\binom{n}{s}4^{-s}(3/4)^{n-s} + \binom{n}{s}2^{-n}} \\ &= \frac{3^{n-s}}{3^{n-s} + 2^n} \end{aligned}$$

and $P(H_0|S_n) = \frac{2^n}{3^{n-s}+2^n}$. The Bayes rule choses $H_0$ when

$$3^{n-s} < 2^n.$$

Expected risk from terminal decision is

$$30(3^{n-s}, 2^n)/(3^{n-s} + 2^n)$$



# Example: Backward induction

After $n = 3$ obs.

| $S$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| *Bayesrisk* | 9.857 | 17.118 | 11.182 | 6.333 |

Suppose now that $n = 2, S_2 = 1$. Bayes risk of decision at this stage is **12.857**, so **14.857** with costs .

$$P(X_3 = 1|S_2 = 1) = \frac{1}{4}P(H_0|S_2 = 1) + \frac{1}{2}P(H_1|S_2 = 1) = \frac{1}{4}\frac{3}{7} + \frac{1}{2}\frac{4}{7} = 0.3929.$$

So expected Bayes risk of taking one more observation is

$$0.3929 \times 11.182 + (1 - 0.3929) \times 17.118 = 14.786$$

So should take the third observation. The Bayes risk after $n = 2$ obs. as a function of $s$ is

| $S$ | 0 | 1 | 2 |
|---|---|---|---|
| *Bayesrisk* | 11.23 | 14.786 | 8.0 |



# Sequential probability ratio test

Can be viewed both as frequentist or bayesian.
$X_1, X_2, \dots$ stream of iid observations $\sim f(x, \theta)$.

| Hypothesis | $H_0$ | $H_1$ |
|---|---|---|
| | $\theta = \theta_0$ | $\theta = \theta_1$ |

**Frequentist approach**

$$LR(n) = \frac{L(\theta_1, \underline{x}_n)}{L(\theta_0, \underline{x}_n)} = \frac{\prod_{i=1}^n f(\theta_1, x_i)}{\prod_{i=1}^n f(\theta_0, x_i)}$$

Test procedure: take $0 < K_1 < K_2 < \infty$.

- If $LR(n) \le K_1$ stop and accept $H_0$,
- If $LR(n) \ge K_2$ stop and accept $H_1$,
- If $LR(n) \in (K_1, K_2)$ take one more observation.

# Sequential probability ratio test

**Bayesian approach** Each obs. cost $c$.

| Hypothesis | $H_0$ | $H_1$ |
|---|---|---|
| | $\theta = \theta_0$ | $\theta = \theta_1$ |
| Prior probability | $\pi_0$ | $\pi_1$ |
| incorrect rejection cost | $a$ | $b$ |

Then

**Lemma**

*For a good choice of* $0 < K_1 < K_2 < \infty$.

- *If* $LR(n) \le K_1$ *stop and accept* $H_0$,
- *If* $LR(n) \ge K_2$ *stop and accept* $H_1$,
- *If* $LR(n) \in (K_1, K_2)$ *take one more observation.*

*is the Bayes procedure.*

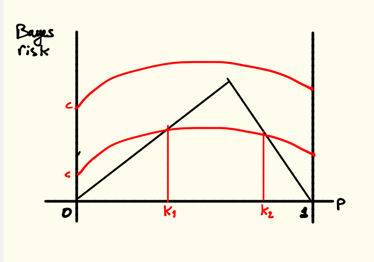


# Sequential probability ratio test

At stage $n$ of sequential sampling, decision (stop and decide or continue) depends on

1. cost $c$ of future obs.
2. Loss of decision $a, b$
3. The amount of information about $\theta$ that future obs. can provide
4. The current info about $\theta$ given by $p = P(\theta = \theta_0 | \underline{x}_n)$

If we stop, expected losses are $b(1-p)$ if we chose $H_0$ and $ap$ if we chose $H_1$. So decision is $H_0 \Leftrightarrow p > b/(a+b)$.




$B^*(p) = \min_\delta B(\delta, p)$ where $\delta$ any rule that take more obs. Is Bayes risk.
$B^*(p) \ge c, B^*(0) = B^*(1) = c$
$p \mapsto B^*(p)$ is concave.



# Sequential probability ratio test

The rule is then

- If $LR(n) \le K_1$ stop and accept $H_0$,
- If $LR(n) \ge K_2$ stop and accept $H_1$,
- If $LR(n) \in (K_1, K_2)$ take one more observation.

Observe that

$$p = \frac{P(H_0)}{P(H_0)} + LR(n)P(H_1)$$

so we continue sampling if

$$\begin{aligned} k_1 < p < k_2 \quad &\equiv \quad \frac{1}{k_2} < \frac{1}{p} < \frac{1}{k_1} \\ &\equiv \quad \frac{1}{k_2} < 1 + LR(n)P(H_1)/P(H_0) < \frac{1}{k_1} \\ &\equiv \quad \frac{1-k_2}{k_2}\frac{P(H_0)}{P(H_1)} < LR(n) < \frac{1-k_1}{k_1}\frac{P(H_0)}{P(H_1)} \end{aligned}$$

Agree with SPRT WITH $K_1 = \frac{1-k_1}{k_1}\frac{P(H_0)}{P(H_1)}, K_2 = \frac{1-k_2}{k_2}\frac{P(H_0)}{P(H_1)}$