Statistical Inference
Lecture Four

https://people.bath.ac.uk/masss/APTS/2022-23/LectureFour.pdf

Simon Shaw

University of Bath

APTS, 13-16 December 2022

Overview of Lecture Four

Last time, Bayesian statistical decision problem, $[\Theta, \mathcal{D}, \pi(\theta), L(\theta, d)]$.

- The risk of decision $d \in \mathcal{D}$ under the distribution $\pi(\theta)$ is $\rho(\pi(\theta), d) = \int_\theta L(\theta, d)\pi(\theta)\, d\theta$.
- A decision $d^* \in \mathcal{D}$ for which $\rho(\pi, d^*) = \rho^*(\pi)$ is a Bayes rule.
- The Bayes rule for the posterior decision respects the SLP.

Today, we'll look at decision theory from a classical perspective.

- The classical risk for the model $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \,|\, \theta)\}$ is

$$R(\theta, \delta) \;=\; \int_X L(\theta, \delta(x)) f_X(x \,|\, \theta)\, dx.$$

- A decision rule δ_0 is admissible if there is no decision rule δ_1 which dominates it.
- Wald's Complete Class Theorem, CCT: a decision rule is admissible if and only if it is a Bayes rule for some prior distribution.
- Admissible decision rules respect the SLP.
- Loss functions for point estimation, set estimation and hypothesis testing.

Example

Let $X = (X_1, \ldots, X_n)$ where $X_i \sim N(\theta, \sigma^2)$ and σ^2 is known. Suppose that $L(\theta, d) = (\theta - d)^2$ and consider a conjugate prior $\theta \sim N(\mu_0, \sigma_0^2)$. Possible decision functions include:

1. $\delta_1(x) = \overline{x}$, the sample mean.
2. $\delta_2(x) = \text{med}\{x_1, \ldots, x_n\} = \tilde{x}$, the sample median.
3. $\delta_3(x) = \mu_0$, the prior mean.
4. $\delta_4(x) = \mu_n$, the posterior mean where

$$\mu_n = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\overline{x}}{\sigma^2}\right),$$

the weighted average of the prior and sample mean accorded to their respective precisions.

Example - continued

The respective classical risks are

1. $R(\theta, \delta_1) = \frac{\sigma^2}{n}$, a constant for θ, since $\overline{X} \sim N(\theta, \sigma^2/n)$.
2. $R(\theta, \delta_2) = \frac{\pi\sigma^2}{2n}$, a constant for θ, since $\tilde{X} \sim N(\theta, \pi\sigma^2/2n)$ (approximately).
3. $R(\theta, \delta_3) = (\theta - \mu_0)^2 = \sigma_0^2 \left(\frac{\theta - \mu_0}{\sigma_0}\right)^2$.
4. $R(\theta, \delta_4) = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-2} \left\{\frac{1}{\sigma_0^2}\left(\frac{\theta - \mu_0}{\sigma_0}\right)^2 + \frac{n}{\sigma^2}\right\}$.

Which decision do we choose? We observe that $R(\theta, \delta_1) < R(\theta, \delta_2)$ for all $\theta \in \Theta$ but other comparisons depend upon θ.

- The accepted approach for classical statisticians is to narrow the set of possible decision rules by ruling out those that are obviously bad.

Definition (Admissible decision rule)

A decision rule δ_0 is inadmissible if there exists a decision rule δ_1 which dominates it, that is

$$R(\theta, \delta_1) \;\leq\; R(\theta, \delta_0)$$

for all $\theta \in \Theta$ with $R(\theta, \delta_1) < R(\theta, \delta_0)$ for at least one value $\theta_0 \in \Theta$. If no such δ_1 exists then δ_0 is admissible.

- If δ_0 is dominated by δ_1 then the classical risk of δ_0 is never smaller than that of δ_1 and δ_1 has a smaller risk for θ_0.
- Thus, you would never want to use δ_0.[1]
- The accepted approach is to reduce the set of possible decision rules under consideration by only using admissible rules.

[1] Here I am assuming that all other considerations are the same in the two cases: e.g. for all $x \in \mathcal{X}$, $\delta_1(x)$ and $\delta_0(x)$ take about the same amount of resource to compute.

- We now show that admissible rules can be related to a Bayes rule δ^* for a prior distribution $\pi(\theta)$.

Theorem

If a prior distribution $\pi(\theta)$ is strictly positive for all Θ with finite Bayes risk and the classical risk, $R(\theta, \delta)$, is a continuous function of θ for all δ, then the Bayes rule δ^* is admissible.

Proof (Robert, 2007)

Letting $f(\theta, x) = f_X(x \,|\, \theta)\pi(\theta)$ we have

$$\begin{aligned}
\mathbb{E}\{L(\theta, \delta(X))\} &= \int_x \int_\theta L(\theta, \delta(x)) f(\theta, x)\, d\theta dx \\
&= \int_\theta \left\{ \int_x L(\theta, \delta(x)) f_X(x \,|\, \theta)\, dx \right\} \pi(\theta)\, d\theta \\
&= \int_\theta R(\theta, \delta)\pi(\theta)\, d\theta
\end{aligned}$$

Proof continued

- Suppose that the Bayes rule δ^* is inadmissible and dominated by δ_1.
- Thus, in an open set C of θ, $R(\theta, \delta_1) < R(\theta, \delta^*)$ with $R(\theta, \delta_1) \leq R(\theta, \delta^*)$ elsewhere.
- Consequently, $\mathbb{E}\{L(\theta, \delta_1(X))\} < \mathbb{E}\{L(\theta, \delta^*(X))\}$ which is a contradiction to δ^* being the Bayes rule. $\square$

- The relationship between a Bayes rule with prior $\pi(\theta)$ and an admissible decision rule is even stronger.
- The following result was derived by Abraham Wald (1902-1950)

Wald's Complete Class Theorem, CCT

In the case where the parameter space Θ and sample space $\mathcal{X}$ are finite, a decision rule δ is admissible if and only if it is a Bayes rule for some prior distribution $\pi(\theta)$ with strictly positive values.

- An illuminating blackboard proof of this result can be found in Cox and Hinkley (1974, Section 11.6).
- There are generalisations of this theorem to non-finite decision sets, parameter spaces, and sample spaces but the results are highly technical.
- We'll proceed assuming the more general result, which is that a decision rule is admissible if and only if it is a Bayes rule for some prior distribution $\pi(\theta)$, which holds for practical purposes.

So what does the CCT say?

1. Admissible decision rules respect the SLP. This follows from the fact that admissible rules are Bayes rules which respect the SLP. This provides support for using admissible decision rules.
2. If you select a Bayes rule according to some positive prior distribution $\pi(\theta)$ then you cannot ever choose an inadmissible decision rule.

Point estimation

- We now look at possible choices of loss functions for different types of inference.
- For point estimation the decision space is $\mathcal{D} = \Theta$, and the loss function $L(\theta, d)$ represents the (negative) consequence of choosing d as a point estimate of θ.
- It will not be often that an obvious loss function $L : \Theta \times \Theta \to \mathbb{R}$ presents itself. There is a need for a generic loss function which is acceptable over a wide range of applications.

Suppose that Θ is a convex subset of $\mathbb{R}^p$. A natural choice is a convex loss function,

$$L(\theta, d) \;=\; h(d - \theta)$$

where $h : \mathbb{R}^p \to \mathbb{R}$ is a smooth non-negative convex function with $h(0) = 0$.

- This type of loss function asserts that small errors are much more tolerable than large ones.
- One possible further restriction is that h is an even function, $h(d-\theta) = h(\theta - d)$.
- In this case, $L(\theta, \theta + \epsilon) = L(\theta, \theta - \epsilon)$ so that under-estimation incurs the same loss as over-estimation.
- We saw previously, that for quadratic loss $\Theta \subset \mathbb{R}$, $L(\theta, d) = (\theta - d)^2$, the Bayes rule was the expectation of $\pi(\theta)$. As we will see, this attractive feature can be extended to more dimensions.
- There are many situations where this is not appropriate and the loss function should be asymmetric and a generic loss function should be replaced by a more specific one.

The bilinear loss function for $\Theta \subset \mathbb{R}$ is, for $\alpha, \beta > 0$,

$$L(\theta, d) = \begin{cases} \alpha(\theta - d) & \text{if } d \leq \theta, \\ \beta(d - \theta) & \text{if } d \geq \theta. \end{cases}$$

- The Bayes rule is a $\frac{\alpha}{\alpha+\beta}$-fractile of $\pi(\theta)$.
- If $\alpha = \beta = 1$ then $L(\theta, d) = |\theta - d|$, the absolute loss which gives a Bayes rule of the median of $\pi(\theta)$.
- $|\theta - d|$ is smaller that $(\theta - d)^2$ for $|\theta - d| > 1$ and so absolute loss is smaller than quadratic loss for large deviations. Thus, it takes less account of the tails of $\pi(\theta)$ leading to the choice of the median.
- If $\alpha > \beta$, so $\frac{\alpha}{\alpha+\beta} > 0.5$, then under-estimation is penalised more than over-estimation and so that Bayes rule is more likely to be an over-estimate.

Example

If $\Theta \in \mathbb{R}^p$, the Bayes rule δ^* associated with the distribution $\pi(\theta)$ and the quadratic loss

$$L(\theta, d) \;\; = \;\; (d-\theta)^T Q \, (d-\theta)$$

is the expectation $\mathbb{E}_{(\pi)}(\theta)$ for every positive-definite symmetric $p \times p$ matrix Q.

Example (Robert, 2007), $Q = \Sigma^{-1}$

Suppose $X \sim N_p(\theta, \Sigma)$ where the known variance matrix Σ is diagonal with elements σ_i^2 for each i. Then $\mathcal{D} = \mathbb{R}^p$. A possible loss function is

$$L(\theta, d) \;\; = \;\; \sum_{i=1}^{p} \left(\frac{d_i - \theta_i}{\sigma_i} \right)^2$$

so that the total loss is the sum of the squared component-wise errors.

- As the Bayes rule for $L(\theta, d) = (d - \theta)^T Q (d - \theta)$ does not depend upon Q, it is the same for an uncountably large class of loss functions.
- If we apply the Complete Class Theorem to this result we see that for quadratic loss, a point estimator for θ is admissible if and only if it is the conditional expectation with respect to some positive prior distribution $\pi(\theta)$.
- The value, and interpretability, of the quadratic loss can be further observed by noting that, from a Taylor series expansion, an even, differentiable and strictly convex loss function can be approximated by a quadratic loss function.

Stein's Example

- Let $X = (X_1, \ldots, X_p)^T$, $\theta = (\theta_1, \ldots, \theta_p)^T$ for $p \geq 3$.
- Suppose that $X \mid \theta \sim N_p(\theta, I_p)$ where I_p is the $p \times p$ identity matrix.
- Thus, given θ, the X_is are independent $N(\theta_i, 1)$.
- For a single observation $X = x$ the maximum likelihood estimate is $\delta^0(x) = x = (x_1, \ldots, x_p)^T$. This is unbiased.
- For quadratic loss $L(\theta, d) = (\theta - d)^T(\theta - d)$ the classical risk of δ^0 is

$$\begin{aligned} R(\theta, \delta^0) &= \mathbb{E}[L(\theta, \delta^0(X)) \mid \theta] \\ &= \sum_{i=1}^{p} \mathbb{E}[(\theta_i - X_i)^2 \mid \theta] \\ &= \sum_{i=1}^{p} Var(X_i \mid \theta) \; = \; p. \end{aligned}$$

- We'll show that δ^0 is inadmissible.

- Consider the set of James-Stein estimators

$$\delta^a(X) \;=\; \left(1 - \frac{a}{X^TX}\right)X$$

for $a \geq 0$ ($a = 0$ gives $\delta^0(X) = X$) which, for $a > 0$, are biased.

- For quadratic loss the classical risk of δ^a is

$$\begin{aligned}
R(\theta, \delta^a) &= \mathbb{E}[(\theta - \delta^a(X))^T(\theta - \delta^a(X)) \,|\, \theta] \\
&= \mathbb{E}\left[\left((\theta - X) + \frac{aX}{X^TX}\right)^T\left((\theta - X) + \frac{aX}{X^TX}\right) \,\middle|\, \theta\right] \\
&= \mathbb{E}[(\theta - X)^T(\theta - X) \,|\, \theta] + a^2\mathbb{E}\left[\frac{1}{X^TX} \,\middle|\, \theta\right] \\
&\quad -2a\mathbb{E}\left[\frac{X^T(X - \theta)}{X^TX} \,\middle|\, \theta\right] \\
&= R(\theta, \delta^0) + a^2\mathbb{E}\left[\frac{1}{X^TX} \,\middle|\, \theta\right] - 2a\sum_{i=1}^{p}\mathbb{E}\left[\frac{X_i(X_i - \theta_i)}{X^TX} \,\middle|\, \theta\right]
\end{aligned}$$

- Stein's Lemma states that for $X \mid \theta \sim N_p(\theta, I_p)$ and $g(X)$ a suitably behaved real valued function

$$\mathbb{E}(g(X)(X_i - \theta_i) \mid \theta) = \mathbb{E}\left[\frac{\partial g(X)}{\partial X_i} \,\middle|\, \theta\right].$$

- Using this result we can show that

$$\sum_{i=1}^{p} \mathbb{E}\left[\frac{X_i}{X^T X}(X_i - \theta_i) \,\middle|\, \theta\right] = (p-2)\mathbb{E}\left[\frac{1}{X^T X} \,\middle|\, \theta\right]$$

so that

$$R(\theta, \delta^a) = R(\theta, \delta^0) + (a^2 - 2a(p-2))\mathbb{E}\left[\frac{1}{X^T X} \,\middle|\, \theta\right].$$

- Now, $X^T X \geq 0$ so that $\mathbb{E}[1/X^T X \mid \theta] \geq 0$ (actually positive) and thus if $a^2 - 2a(p-2) < 0$ then $R(\theta, \delta^a) < R(\theta, \delta^0)$.
- Hence, if $0 < a < 2(p-2)$ (exists as $p \geq 3$) then δ^0 is inadmissible.

- Note that $a = p - 2$ minimises $R(\theta, \delta^a)$
- The ith term of $\delta^a(X) = \left(1 - \frac{a}{X^T X}\right) X$ is $\left(1 - \frac{a}{X^T X}\right) X_i$ and so depends on all $X_1, \ldots, X_p$ even though the X_is are independent.
- This outcome, often called Stein's Paradox, can be shown to occur in many situations when comparing three or more populations.
- It occurs because the loss function is dealing with simultaneous estimation of all parameters and so is an on average property.
- Note that δ^a shrinks some of the estimates towards 0 and this idea - using shrinkage to reduce variance (at the expense of introducing bias) - is widely used in statistics.
- The inadmissible δ^0 means that I can't find a proper prior for which δ^0 is the Bayes rule (in this case, it's essentially the Bayes rule of an improper uniform).

Set estimation

- For set estimation the decision space is a set of subsets of Θ so that each $d \subset \Theta$.
- There are two contradictory requirements for set estimators of Θ.
  1. We want the sets to be small.
  2. We also want them to contain θ.
- A simple way to represent these two requirements is to consider the loss function

$$L(\theta, d) \;=\; |d| + \kappa(1 - \mathbb{1}_{\theta \in d})$$

  for some $\kappa > 0$ where $|d|$ is the volume of d.
- The value of κ controls the trade-off between the two requirements.
  - If $\kappa \downarrow 0$ then minimising the expected loss will always produce the empty set.
  - If $\kappa \uparrow \infty$ then minimising the expected loss will always produce Θ.

- For loss functions of the form $L(\theta, d) = |d| + \kappa(1 - \mathbb{1}_{\theta \in d})$ we'll show there is a a simple necessary condition for a rule to be a Bayes rule.

Definition (Level set)

A set $d \subset \Theta$ is a level set of the posterior distribution exactly when $d = \{\theta : \pi(\theta \,|\, x) \geq k\}$ for some k.

Theorem (Level set property, LSP)

If δ^* is a Bayes rule for $L(\theta, d) = |d| + \kappa(1 - \mathbb{1}_{\theta \in d})$ then it is a level set of the posterior distribution.

Proof

Note that

$$\begin{aligned}
\mathbb{E}\{L(\theta, d) \,|\, X\} &= |d| + \kappa(1 - \mathbb{E}(\mathbb{1}_{\theta \in d} \,|\, X)) \\
&= |d| + \kappa\mathbb{P}(\theta \notin d \,|\, X).
\end{aligned}$$

Proof continued

- For fixed x, we show that if d is not a level set of the posterior distribution then there is a $d' \neq d$ which has a smaller expected loss so that $\delta^*(x) \neq d$.
- Suppose that d is not a level set of $\pi(\theta \,|\, x)$. Then there is a $\theta \in d$ and $\theta' \notin d$ for which $\pi(\theta' \,|\, x) > \pi(\theta \,|\, x)$.
- Let $d' = d \cup d\theta' \setminus d\theta$ where $d\theta$ is the tiny region of Θ around θ and $d\theta'$ is the tiny region of Θ around θ' for which $|d\theta| = |d\theta'|$.
- Then $|d'| = |d|$ but

$$\mathbb{P}(\theta \notin d' \,|\, X) < \mathbb{P}(\theta \notin d \,|\, X)$$

Thus, $\mathbb{E}\{L(\theta, d') \,|\, X\} < \mathbb{E}\{L(\theta, d) \,|\, X\}$ showing that $\delta^*(x) \neq d$. $\square$

- The Level Set Property Theorem states that δ having the level set property is necessary for δ to be a Bayes rule for loss functions of the form $L(\theta, d) = |d| + \kappa(1 - \mathbb{1}_{\theta \in d})$.
- The Complete Class Theorem states that being a Bayes rule is a necessary condition for δ to be admissible.
- Being a level set of a posterior distribution for some prior distribution $\pi(\theta)$ is a necessary condition for being admissible for loss functions of this form.
- Bayesian HPD regions satisfy the necessary condition for being a set estimator.
- Classical set estimators achieve a similar outcome if they are level sets of the likelihood function, because the posterior is proportional to the likelihood under a uniform prior distribution.[2]

[2] In the case where Θ is unbounded, this prior distribution may have to be truncated to be proper.

Hypothesis tests

- For hypothesis tests, the decision space is a partition of Θ, denoted

$$\mathcal{H} \;:=\; \{H_0, H_1, \ldots, H_d\}.$$

- Each element of $\mathcal{H}$ is termed a hypothesis.
- The loss function $L(\theta, H_i)$ represents the (negative) consequences of choosing element H_i, when the true value of the parameter is θ.
- It would be usual for the loss function to satisfy

$$\theta \in H_i \implies L(\theta, H_i) \;=\; \min_j L(\theta, H_j)$$

on the grounds that an incorrect choice of element should never incur a smaller loss than the correct choice.

- Consider the test of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ where $\Theta_1 = \Theta \setminus \Theta_0$. Let $\mathcal{D} = \{d_0, d_1\}$ where d_i corresponds to accepting H_i. A generic loss function is the 0-1 ('zero-one') loss function

$$L(\theta, d_i) = \begin{cases} 0 & \text{if } \theta \in \Theta_i, \\ 1 & \text{if } \theta \notin \Theta_i. \end{cases}$$

- The classical risk is the probability of making a wrong decision,

$$R(\theta, \delta) = \begin{cases} \mathbb{P}(\delta(X) = d_1 \,|\, \theta) & \text{if } \theta \in \Theta_0, \\ \mathbb{P}(\delta(X) = d_0 \,|\, \theta) & \text{if } \theta \in \Theta_1, \end{cases}$$

which correspond to the familiar Type I and Type II errors.

- The Bayes rule is to choose H_0 if $\mathbb{P}_\pi(\theta \in \Theta_0) > \mathbb{P}_\pi(\theta \in \Theta_1)$ and H_1 otherwise, where $\mathbb{P}_\pi(\cdot)$ is the probability when $\theta \sim \pi(\theta)$.
- Hence, if $\pi(\theta) = f(\theta \,|\, x)$, the Bayes rule is to choose the hypothesis with the largest posterior probability.

- This approach can be naturally extended to multiple hypotheses $\mathcal{H} = \{H_0, H_1, \ldots, H_d\}$ which partition Θ by taking

$$L(\theta, H_i) \;=\; 1 - \mathbb{1}_{\{\theta \in H_i\}}.$$

  i.e., zero if $\theta \in H_i$, and one if it is not.
- For the posterior decision, the Bayes rule is to select the hypothesis with the largest posterior probability.
- However, this loss function is hard to defend as being realistic.
- If we choose H_i and it turns out that $\theta \notin H_i$ then the inference is wrong and the loss is the same irrespective of where θ lies.
- An alternative approach is to co-opt the theory of set estimators.
- The statistician can use her set estimator δ to make at least some distinctions between the members of $\mathcal{H}$:
  - Accept H_i exactly when $\delta(x) \subset H_i$,
  - Reject H_i exactly when $\delta(x) \cap H_i = \emptyset$,
  - Undecided about H_i otherwise.

Confidence procedures and confidence sets

- We consider interval estimation, or more generally set estimation.
- Under the model $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$, for given data $X = x$, we wish to construct a set $C = C(x) \subset \Theta$ and the inference is the statement that $\theta \in C$.
- If $\theta \in \mathbb{R}$ then the set estimate is typically an interval.

Definition (Confidence procedure)

A random set $C(X)$ is a level-$(1-\alpha)$ confidence procedure exactly when

$$\mathbb{P}(\theta \in C(X) \mid \theta) \;\geq\; 1 - \alpha$$

for all $\theta \in \Theta$. C is an exact level-$(1-\alpha)$ confidence procedure if the probability equals $(1-\alpha)$ for all θ.

- The value $\mathbb{P}(\theta \in C(X) \,|\, \theta)$ is termed the coverage of C at θ.
- Exact is a special case: typically $\mathbb{P}(\theta \in C(X) \,|\, \theta)$ will depend upon θ.
- The procedure is thus conservative: for a given θ_0 the coverage may be much higher than $(1-\alpha)$.

Uniform example

- Let $X_1, \ldots, X_n$ be independent and identically distributed $\text{Unif}(0, \theta)$ random variables where $\theta > 0$. Let $Y = \max\{X_1, \ldots, X_n\}$.
- We consider two possible sets: (aY, bY) where $1 \leq a < b$ and $(Y + c, Y + d)$ where $0 \leq c < d$.
  1. $\mathbb{P}(\theta \in (aY, bY) \,|\, \theta) = \left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n$. Thus, the coverage probability of the interval does not depend upon θ.
  2. $\mathbb{P}(\theta \in (Y + c, Y + d) \,|\, \theta) = \left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n$. In this case, the coverage probability of the interval does depend upon θ.

- We distinguish between the confidence procedure C, which is a random interval and so a function for each possible x, and the result when C is evaluated at the observation x, which is a set in Θ.

Definition (Confidence set)

The observed $C(x)$ is a level-$(1-\alpha)$ confidence set exactly when the random $C(X)$ is a level-$(1-\alpha)$ confidence procedure.

- If $\Theta \subset \mathbb{R}$ and $C(x)$ is convex, i.e. an interval, then a confidence set (interval) is represented by a lower and upper value.
- The challenge with confidence procedures is to construct one with a specified level: to do this we start with the level and then construct a C guaranteed to have this level.

Definition (Family of confidence procedures)

- $C(X; \alpha)$ is a family of confidence procedures exactly when $C(X; \alpha)$ is a level-$(1 - \alpha)$ confidence procedure for every $\alpha \in [0, 1]$.
- C is a nesting family exactly when $\alpha < \alpha'$ implies that $C(x; \alpha') \subset C(x; \alpha)$.

- For $X_1, \ldots, X_n$ iid $\text{Unif}(0, \theta)$, $Y = \max\{X_1, \ldots, X_n\}$ then

$$C(Y; \alpha) = \left(\left(1 - \frac{\alpha}{2}\right)^{-1/n} Y, \left(\frac{\alpha}{2}\right)^{-1/n} Y \right)$$

  is a nesting family of exact confidence procedures.
- For example, if $n = 10$ then

$$C(y; 0.10) = (1.0051y, 1.3493y); \;\; C(y; 0.05) = (1.0025y, 1.4461y).$$

- If we start with a family of confidence procedures for a specified model, then we can compute a confidence set for any level we choose.