Theory of Statistical Inference
Topic 3. Confidence Set and Hypothesis Testing

Hung Chen
Division of Biostatistics
The Joseph L. Mailman School of Public Health
Columbia University
1/16/2001
email: chen@ams.sunysb.edu
Web page: www.ams.sunysb.edu/∼chen/

OUTLINE

1. Introduction
   - Examples
   - The Neyman-Pearson lemma
   - The duality between confidence intervals and tests
   - Bayesian analysis and elementary decision theory
2. Tests in Parametric Models
   - Likelihood ratio tests
   - Asymptotic tests based on likelihoods
   - χ^2- tests
   - Asymptotic Confidence Sets
3. Optimal Tests and Confidence Sets
   - UMP Tests
   - UMP Unbiased Tests
   - UMP Invariant Tests
   - Construction of Confidence Sets-Pivotal quantities and Test duality
   - Properties of Confidence Sets
     * Lengths of confidence intervals
     * UMA and UMAU confidence sets
4. References
   - Bickel and Doksum (1977) Chapters 5, 6, and 10.
   - Silvey (1999) Chapters 5, 6, 7, and 11.

Introduction

Up to now, we discuss how to find the most *plausible* parameter value after an observation $\mathbf{x}$ has been made. For example, the method of maximum likelihood tries to assign that value of the parameter which gives greatest probability to $\mathbf{x}$. However, we should be extremely reluctant to believe that an estimate coincided with the true parameter in all circumstances, and it is natural to ask how near the true parameter we might expect an estimate to be. The very use of the phrase *how near* implies that there is a metric on the parameter space.

We may take the point of view that, when an observation has been made, this observation divides the parameter set into two disjoint subsets: a plausible subset and an implausible subset; and that what we really want to do, rather than to fix attention on a particular parameter value as an estimate of the true parameter, is to determine this plausible subset of parameter values. Then our conclusion based on an observation would be, "The true parameter is in such-and-such a subset of the set of possible parameters." The formalization of this idea leads to the problem of *set estimation.*

Example 3.1 Suppose that we have available a random sample $\mathbf{x} = (x_1, \ldots, x_n)$ from a normal distribution with unknown mean μ and unknown variance σ^2 and we wish to determine a *plausible* set of values of μ.

- Let $\boldsymbol{\theta} = (\mu, \sigma^2)$ and let $t(\mathbf{x}, \mu) = \sqrt{n}(\bar{x} - \mu)/s$, where $\bar{x} = n^{-1}\sum_{i=1}^n x_i$ and $s^2 = (n-1)^{-1}\sum_{i=1}^n (x_i - \bar{x})^2$.
- Note that $t(\mathbf{x}, \mu)$ is distributed as Student's t with $n-1$ degrees of freedom which does not depend on unknown $\boldsymbol{\theta}$.
- Without knowing what $\boldsymbol{\theta}$ is, we can find a number $t_{\alpha/2}$ such that
$$P_\theta(-t_{\alpha/2} \le t(\mathbf{x}, \mu) \le t_{\alpha/2}) = 1 - \alpha,$$
where α is a small preassigned number between 0 and 1.
- The above can be rewritten in the following form:
$$P_\theta\left(\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1 - \alpha.$$
It reads as **Whatever the true value of μ may be, the probability that the random interval**
$$\left[\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right]$$
contains this true value is $1 - \alpha$.

- For each given $\mathbf{x}$, the interval

$$\left[\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}\bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

  may be regarded as a *plausible* set of values of μ, plausible in the sense that we are $100(1-\alpha)$ per cent confident that this set contains the true parameter value.

- The interval is called a *confidence interval* for μ with *confidence coefficient* $1-\alpha$.

- Do you note the difference between $\bar{X}$ and $\bar{x}$?

- Think of the meaning of $1-\alpha$.

Remarks:

- As α decreases, $t_{\alpha/2}$ increases, so that this $100(1-\alpha)$ per cent confidence interval, corresponding to any given $\mathbf{x}$, widens as α decreases.
  In other words. If we wish to have great confidence in a chosen plausible interval, we must choose a larger interval than is necessary if we are content to have less confidence in our chosen interval.

- For fixed α and any given $\mathbf{x}$, there is not a **unique** $100(1-\alpha)$ per cent confidence interval for μ.
  For example, we can choose two numbers $t_{1\alpha}$ and $t_{2\alpha}$ such that $t_{1\alpha} \neq -t_{2\alpha}$, and still have

$$P_\theta\{t_{1\alpha} \leq t(\mathbf{x}, \mu) \leq -t_{2\alpha}\} = 1-\alpha.$$

  If we do so, then we are led by exactly the same argument to the $100(1-\alpha)$ per cent confidence interval

$$\left[\bar{x} - t_{2\alpha}\frac{s}{\sqrt{n}}, \bar{x} - t_{1\alpha}\frac{s}{\sqrt{n}}\right].$$

In science, medicine, public policy, and indeed most human activities, we are quite often to get a yes or no answer to important questions. In one of Mendel's famous experiments,

- He crossed peas heterozygous for a trait with two alleles, one of which was dominant.

- The progeny exhibited approximately the expected ratio of one homozygous dominant to two heterozygous dominants (to one recessive).

- In a modern formulation, if there were n dominant offspring (seeds), the natural model is to assume, if the inheritance ratio can be arbitrary, that N_{AA}, the number of homozygous corresponds to $H_0 : p = 1/3$ with the alternative $H_1 : p \neq 1/3$.

As a further example, the graduate division of the University of California at Berkeley attempted to study the possibility that sex bias operated in graduate admissions in 1973 by examining admission data. In this case, what does the hypothesis of no sex bias corresponds to? It is natural to translate this into

$$P[Admit|Male] = P[Admit|Female].$$

Example 3.2 Suppose that a new drug is being considered with a view to curing a certain disease.

- The drug is given to n patients suffering from the disease and the number r of cures is noted.
- We wish to test the hypothesis that there is at least a $50 - 50$ chance of a cure by this drug based on the following data:

$$r \text{ cures among } n \text{ patients.}$$

- Put the problem in the following framework of statistical test:
    - The sample space $\mathcal{X}$ is simple-it is the set $\{0, 1, 2, \ldots, n\}$.
    - The family $\{P_\theta\}$ of possible distributions on $\mathcal{X}$ is (assuming independent patients) the family of binomial distributions, parametrized by the real parameter θ taking values in $[0, 1]$. θ is being interpreted as the probability of cure.
    - The stated hypothesis defines the subset $\Theta_0 = [1/2, 1]$ of the parameter space.
    - In this situation, only a small class of tests which seem worth considering on a purely intuitive basis.
      We will only consider those for which the set of x taken to be consistent with Θ_0 have the form $\{x : x \geq k\}$
    - **Question**: Does it make sense to consider that r cures out of n patients were consistent with Θ_0, while $r+1$ were not?
    - What is a **reasonable** test?

The Neyman-Pearson Theory

Now we provide a theoretical discussion on testing of statistical hypotheses to resolve the difficulty stated in Example 3.2. Neyman and Pearson (1933) presented Neyman-Pearson Fundamental Lemma which unfolded the various complex problems in testing statistical hypotheses. The basic mathematical framework is: a sample space $\mathcal{X}$ and a family $\{P_\theta : \theta \in \Theta\}$ of probability distributions on $\mathcal{X}$, labeled by a parameter θ which ranges over a parameter space Θ.

- A hypothesis is a statement which implies that the true probability distribution describing the inherent variability in an observational situation belongs to a proper subset of a family of possible probability distribution.
  Alternatively we may say that a hypothesis implies that the true parameter θ belongs to a proper subset of the parameter space Θ; and it is convenient to identify the hypothesis with the subset, to talk about the hypothesis Θ_0, where $\Theta_0 \subset \Theta$.
- The theory of hypothesis testing is concerned with the problem: Is a given observation consistent with some stated hypothesis or is it not?
- A statistical test of a hypothesis is a rule which assigns each possible observation to one of two exclusive categories: *consistent with the hypothesis under consideration* and *not consistent with the hypothesis.*

Set-Up

- $\mathbf{X}$: a sample from a population P in $\mathcal{P}$
  $\mathcal{P}$: a family of population
- Test a given hypothesis $H_0 : P \in \mathcal{P}_0$ versus $H_1 : P \in \mathcal{P}_1$ where $\mathcal{P}_0$ and $\mathcal{P}_1$ are two disjoint subsets of $\mathcal{P}$ and $\mathcal{P}_0 \cup \mathcal{P}_1 = \mathcal{P}$
- H_0 is called the *null hypothesis* and H_1 is called the *alternative hypothesis.*
  The names of these two hypotheses suggest that in some sense they are **not** on an equal footing, a point to which we shall return.
- The action space for this problem contains only two elements, i.e., $\mathcal{A} = \{0, 1\}$, where 0 is the action of accepting H_0 and 1 is the action of rejecting H_0.
- A decision rule is called a *test.*
  It partitions the sample space into two subsets, a set of points

each of which is consistent with H_0 (a region of acceptance of H_0); and its complement, consisting of points not consistent with H_0 (the critical region of the test).

- Our object in constructing a **good** test may then be interpreted as choosing a critical region which is optimum relative to some criterion.

- There are only two types of statistical errors we may commit:

  - Reject H_0 when H_0 is true. It is called *the type I error.*
  - Accept H_0 when H_0 is wrong. It is called *the type II error.*
  - In statistical inference, a test T, which is a statistic from $\mathcal{X}$ to $\{0,1\}$, is assessed by the probabilities of making two types of errors:

    $$\alpha_T(P) = P(T(\mathbf{X}) = 1), \quad P \in \mathcal{P}_0$$

    and

    $$1 - \alpha_T(P) = P(T(\mathbf{X}) = 0), \quad P \in \mathcal{P}_1,$$

    were denoted by $\alpha_T(\theta)$ and $1-\alpha_T(\theta)$ if P is in a parametric family indexed by θ.
  - For a given test $T(\mathbf{X})$, its power function is defined to be

    $$\beta_T(P) = E[T(\mathbf{X})], \quad P \in \mathcal{P},$$

    which is the type I error probability of $T(\mathbf{X})$ when $P \in \mathcal{P}_0$ and one minus the type II error probability of $T(\mathbf{X})$ when $P \in \mathcal{P}_1$.
  - Here we only consider $T(\mathbf{X}) = 1$ or 0.
    This kind of test is called a nonrandomized test.
  - Later on, we may consider randomized tests in which $T(\mathbf{X})$ take values in $[0, 1]$.

- Neyman-Pearson framework:
  An **optimal** test is to assign a small bound α to one of the error probabilities, say $\alpha_T(P)$, $P \in \mathcal{P}_0$, and then to attempt to minimize the other error probability $1-\alpha_T(P)$, $P \in \mathcal{P}_1$ subject to

  $$\sup_{P\in\mathcal{P}_0} \alpha_T(P) \le \alpha.$$

  The bound α is called the level of significance. The left-hand side of the above is called the size of the test T.

Why Neyman-Pearson framework is being accepted?

- A test whose error probabilities are as small as possible is clearly desirable.
  However, we cannot choose the critical region in such a way that $\alpha(\theta)$ and $\beta(\theta)$ are simultaneously uniformly minimized.
  By taking the critical region as the empty set, we can make $\alpha(\theta) = 0$ and by taking the critical region as the sample space, we can make $\beta(\theta) = 0$. Hence a test which uniformly minimized both error-probability functions would require to have zero error probabilities, and usually no such test exists.

- The modification suggested by Neyman and Pearson is based on the fact that in most circumstances our attitudes to the hypotheses Θ_0 and $\Theta - \Theta_0$ are different- we are often asking if there is sufficient evidence to reject the hypothesis Θ_0.
  In terms of the two possible errors this may be translated into the statement that often the Type I error is more serious than the Type II error.

- We should control the probability of the Type I error at some pre-assigned small value α, and then, subject to this control, look for a test which uniformly minimizes the function describing the probabilities of Type II error.

- Is this asymmetry on (H_0, H_1) reasonable?
  - Suppose we use this testing technique in searching for regions of the genome that resemble other regions that are known to have significant biological activity.
  - One way of doing this is to align the known and unknown regions and compute statistics based on the number of matches.
  - To determine significant values of these statistics a (more complicated) version of the following is done.
    Thresholds (critical values) are set so that if the matches occur at random and the probability of a match is 1/2, then the probability of exceeding the threshold (type I) error is smaller than α.
  - No one really believes that H_0 is true and possible types of alternatives are vaguely known at best, but computation under H_0 is easy.

Now we use the following example to motivate Neyman-Pearson lemma. We start from the simplest possible situation, that where Θ has only two elements θ_0 and θ_1, say, and where $\Theta_0 = \{\theta_0\}$, $\Theta - \Theta_0 = \{\theta_1\}$. Note that a hypothesis which specifies a set in the parameter space containing only one element is called a *simple* hypothesis.

Thus we are now considering testing a simple null-hypothesis against a simple alternative. In this case, the power function of any test reduces to a single number, and we examine the question of the existence of a most-powerful test of given significance level α.

Example 3.3 Consider the problem that r cures out of n patients when $n = 5$. We wish to test

$$H_0 : p = 0.5 \quad \text{versus} \quad H_1 : p = 0.3.$$

- The probability distribution of r is

r	0	1	2	3	4	5
$p = 0.5$	0.031	0.156	0.313	0.313	0.156	0.031
$p = 0.3$	0.168	0.360	0.309	0.132	0.028	0.003
$f_1(r)/f_0(r)$	5.419	2.308	0.987	0.422	0.179	0.097

- Think of the meaning of likelihood ratio $f_1(r)/f_0(r)$.
- We consider all possible nonrandomized tests of significance level 0.2.

critical region	$\beta_{0.5}$	$\beta_{0.3}$	critical region	$\beta_{0.5}$	$\beta_{0.3}$
$\{0\}$	0.031	0.168	$\{0, 1\}$	0.187	0.528
$\{1\}$	0.156	0.360	$\{0, 4\}$	0.187	0.196
$\{4\}$	0.156	0.028	$\{1, 5\}$	0.187	0.363
$\{5\}$	0.031	0.003	$\{4, 5\}$	0.187	0.031
$\{0, 5\}$	0.062	0.171			

- The best test is the one with critical region $\{0, 1\}$. Can you give a reason for that? Or, can you find a rule?
  Try to think in terms of likelihood ratio by noting

$$f_1(r) = \frac{f_1(r)}{f_0(r)} \cdot f_0(r).$$

  As a hint, compare the two tests $\{0, 1\}$ and $\{0, 4\}$ with the same α. Observe that their power are

$$\begin{aligned} \beta_{\{0,1\}} &= [P_{\{p=0.3\}}(r = 0)] + P_{\{p=0.3\}}(r = 1) \\ \beta_{\{0,4\}} &= [P_{\{p=0.3\}}(r = 0)] + P_{\{p=0.3\}}(r = 4). \end{aligned}$$

  Compare $P_{\{p=0.3\}}(r = 4)$ to $P_{\{p=0.3\}}(r = 1)$.

- Conclusion: Refer to Remark 2 of Theorem 1.

Definition A test T_* of size α is a *uniformly most powerful* (UMP) test if and only if $\beta_{T_*}(P) \geq \beta_T(P)$ for all $P \in \mathcal{P}_1$ and T of size α.

If $U(\mathbf{X})$ is a sufficient statistic for $P \in \mathcal{P}$, then for any test $T(\mathbf{X})$, $E(T|U)$ has the same power function as T and, therefore, to find a UMP test we may consider tests that are functions of U only.

We now state and prove the Neyman-Pearson fundamental lemma. Suppose that the probability distributions P_0 and P_1 on the sample space $\mathcal{X}$ and defined by density functions f_0 and f_1 respectively with respect to some fixed measure on $\mathcal{X}$. (There is no loss of generality in this assumption since the fixed measure may be taken, for instance, to be $P_0 + P_1$.) The following result gives a recipe to find the UMP tests when both H_0 and H_1 are simple.

Theorem 1 *(The Neyman-Pearson lemma). Suppose that $\Theta_0 = \{P_0\}$ and $\Theta - \Theta_0 = \{P_1\}$. Let f_j be the pdf of P_j with respect to a σ-finite measure ν.*
(i) (Existence of a UMP test). For every α, there exists a UMP test of size α, which is equal to

$$T_*(X) = \begin{cases} 1 & f_1(X) > cf_0(X) \\ \gamma & f_1(X) = cf_0(X) \\ 0 & f_1(X) < cf_0(X) \end{cases},$$

where $\gamma \in (0,1)$ and $c \geq 0$ are some constants chosen so that $E[T_*(X)] = \alpha$ when $P = P_0$ ($c = \infty$ is allowed).
(ii) (Uniqueness). If T_{**} is a UMP test of size α,then

$$T_{**}(X) = \begin{cases} 1 & f_1(X) > cf_0(X) \\ 0 & f_1(X) < cf_0(X) \end{cases} \quad a.s.\ P.$$

Proof. Assume now that $0 < \alpha < 1$. We first prove (i).

- Show that there exist γ and c such that $E_0[T_*(X)] = \alpha$, where E_j is the expectation wrt P_j.
  - Let $\gamma(t) = P_0(f_1(X) > tf_0(X))$. Then $\gamma(t)$ is nonincreasing, $\gamma(-\infty) = 1$, and $\gamma(\infty) = 0$.
    To get a better understanding, you can think in terms of Example 3.3 with $\alpha = 0.2$. Here t refers to the likelihood ratio
  - Thus, there exists a $c \in (0, \infty)$ such that $\gamma(c) \leq \alpha \leq \gamma(c-)$. Set

    $$\gamma = \begin{cases} \frac{\alpha - \gamma(c)}{\gamma(c-) - \gamma(c)} & \gamma(c-) \neq \gamma(c) \\ 0 & \gamma(c-) = \gamma(c) \end{cases}$$

    Note that $\gamma(c-) - \gamma(c) = P(f_1(X) = cf_0(X))$.
  - Then

    $$E_0[T_*(X)] = P_0(f_1(X) > cf_0(X)) + \gamma P_0(f_1(X) = cf_0(X)) = \alpha.$$

- Next, we show that T_* is a UMP test.

- Suppose that $T(X)$ is a test satisfying $E_0[T(X)] \leq \alpha$.
  If $T_*(x) - T(x) > 0$, then $T_*(x) > 0$ and, therefore, $f_1(x) \geq cf_0(x)$.
  If $T_*(x) - T(x) < 0$, then $T_*(x) < 1$ and, therefore, $f_1(x) \leq cf_0(x)$.
- Since $[T_*(x) - T(x)][f_1(x) - cf_0(x)] \geq 0$,

$$\begin{aligned} \int [T_*(x) - T(x)][f_1(x) - cf_0(x)] d\nu &\geq 0, \\ \int [T_*(x) - T(x)] f_1(x) d\nu &\geq c \int [T_*(x) - T(x)] f_0(x) d\nu. \end{aligned}$$

- The above inequality leads to

$$E_1[T_*(X)] - E_1[T(X)] \geq c\{E_0[T_*(X)] - E_0[T(X)]\} \geq 0.$$

We now prove (ii). Let $T_{**}(X)$ be a UMP test of size α. Define

$$A = \{x : T_*(X) \neq T_{**}(X), f_1(x) \neq cf_0(X)\}.$$

Then $[T_*(x) - T_{**}(x)][f_1(x) - cf_0(x)] > 0$ when $x \in A$ and $= 0$ when $x \in A^c$, and

$$\int [T_*(x) - T_{**}(x)][f_1(x) - cf_0(x)] d\nu = 0,$$

since both T_* and T_{**} are UMP tests of size α. This implies that $\nu(A) = 0$.

Remarks

1. When both H_0 and H_1 are simple, there exists a UMP test that can be determined by Theorem 1(ii) uniquely except on the set $B = \{x : f_1(x) = cf_0(x)\}$.
2. The critical region determined by $\{x : f_1(x)/f_0(x) \geq c\}$ is quite intuitive. Suppose that we set out to order points in the sample space according to the amount of evidence they provide for P_1 rather than P_0. We should naturally order them according to the value of the ratio $f_1(x)/f_0(x)$; any x for which this ratio is large provides evidence than P_1 rather than P_0 is the true underlying probability distribution. The Neyman-Pearson analysis gives us a basis for choosing c so that

$$P_1 \left\{ x : \frac{f_1(x)}{f_0(x)} \geq c \right\} = \alpha.$$

3. If $\nu(B) = 0$, then we have a unique nonrandomized UMP test; otherwise UMP tests are randomized on the set B and the randomization is necessary for UMP tests to have the given size α.

4. To overcome the difficulty caused by possible discreteness of the probability distributions involved is to allow randomized tests, according to which, having observed an x in the sample space, with probability γ we decide that H_1 is true and with probability $1-\gamma$ we decide that H_0 is true.

Now we use the Neyman-Pearson lemma to derive UMP test in the following two examples.

Example 3.4 Suppose that X is a sample of size 1. We wish to test whether it comes from $N(0,1)$ or the double exponential distribution $DE(0,2)$ with the pdf $4^{-1}\exp(-|x|/2)$.

- Since $P(f_1(x) = cf_0(x)) = 0$, there is a unique nonrandomized UMP test.
- The UMP test $T_*(x) = 1$ if and only if

$$\frac{\pi}{8}\exp(x^2 - |x|) > c^2$$

  for some $c > 0$, which is equivalent to $|x| > t$ or $|x| < 1-t$ for some $t > 1/2$.
- Suppose that $\alpha < 1/4$. We use

$$\alpha = E_0[T_*(X)] = P_0(|X| > t) = 0.3374 > \alpha.$$

  Hence t should be greater than 1 and

$$\alpha = \Phi(-t) + 1 - \Phi(t).$$

  Thus, $t = \Phi^{-1}(1-\alpha/2)$ and $T_*(X) = I_{(t,\infty)}(|X|)$.
- Why the UMP test rejects H_0 when $|X|$ is large?
- The power of T_* under H_1 is

$$E_1[T_*(X)] = P_1(|X| > t) = 1 - \frac{1}{4}\int_{-t}^{t} e^{-|x|/2}dx = e^{-t/2}.$$

Example 3.5 Let $X_1, \ldots, X_n$ be iid binary random variables with $p = P(X_1 = 1)$. Suppose that we wish to test $H_0 : p = p_0$ versus $H_1 : p = p_1$, where $0 < p_0 < p_1 < 1$.

- Since $P(f_1(x) = cf_0(x)) \neq 0$, we may need to consider randomized UMP test.
- A UMP test of size α is

$$T_*(Y) = \begin{cases} 1 & \lambda(Y) > c \\ \gamma & \lambda(Y) = c \\ 0 & \lambda(Y) < c, \end{cases}$$

where $Y = \sum_{i=1}^n X_i$ and

$$\lambda(Y) = \left(\frac{p_1}{p_0}\right)^Y \left(\frac{1-p_1}{1-p_0}\right)^{n-Y}.$$

- Since $\lambda(Y)$ is increasing in Y, there is an integer $m > 0$ such that
$$T_*(Y) = \begin{cases} 1 & Y > m \\ \gamma & Y = m \\ 0 & Y < m, \end{cases}$$
where m and γ satisfy
$$\alpha = E_0[T_*(Y)] = P_0(Y > m) + \gamma P_0(Y = m).$$

- Since Y has the binomial distribution $Bin(n, p)$, we can determine m and γ from
$$\alpha = \sum_{j=m+1}^n \left(\begin{array}{c} n \\ j \end{array}\right) p_0^j (1-p_0)^{n-j} + \gamma \left(\begin{array}{c} n \\ m \end{array}\right) p_0^m (1-p_0)^{n-m}.$$

- Unless
$$\alpha = \sum_{j=m+1}^n \left(\begin{array}{c} n \\ j \end{array}\right) p_0^j (1-p_0)^{n-j}$$
for some integer m, the UMP test is a randomized test.

- Do you notice that the UMP test T_* does not depend on p_1?
  - Neyman-Pearson lemma tells us that we should put those x into rejection region according to its likelihood ratio until the level of test achieves α.
  - Think of two hypothesis testing problems: The first one is $H_0 : p = p_0$ versus $H_1 : p = p_1$ and the second one is $H_0 : p = p_0$ versus $H_1 : p = p_2$ where $p_1 > p_0$ and $p_2 > p_0$.
  - For the above two testing problems, both their likelihood ratios increase as y increases.
  - T_* is in fact a UMP test for testing $H_0 : p = p_0$ versus $H_1 : p > p_0$.

- Suppose that there is a test T_* of size α such that for every $P_1 \in \mathcal{P}$, T_* is UMP for testing H_0 versus the hypothesis $P = P_1$. Then T_* is UMP for testing H_0 versus H_1.

Before we move to next topic, we discuss an example in which the two kinds of probability might be equally important.

Example 3.6 (Fisher's Discriminant Function)
Suppose $\mathbf{X} \sim N(\boldsymbol{\mu}_j, \Sigma_j)$, $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \Sigma_j)$, $j = 0, 1$.

- In a classification context, $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ correspond to two known populations and we desire to classify a new observation $\mathbf{X}$ as belonging to one or the other.
  - Learning sample, training sample and etc can be used to estimate $\boldsymbol{\theta}_j$, $j = 0, 1$.
  - What is the optimal classifier?
  - Denote unknown parameter associated with the new observation $\mathbf{x}$ as $\boldsymbol{\theta}$. If we put it as $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$, the probability of type I error and the probability of type II error will be its misclassification error probability.
  - Which misclassification error probability is more serious?
- Use N-P lemma, we reject $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ if

  $$Q = (\mathbf{X} - \boldsymbol{\mu}_0)^T \Sigma_1^{-0} (\mathbf{X} - \boldsymbol{\mu}_0) - (\mathbf{X} - \boldsymbol{\mu}_1)^T \Sigma_1^{-1} (\mathbf{X} - \boldsymbol{\mu}_1)$$

  is large.
- In the case $\Sigma_0 = \Sigma_1$, when Q is large is equivalent to

  $$F = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \Sigma_0^{-1} \mathbf{X}$$

  is large.
- How do we determine the optimal cutting value of Q if our goal is to minimize the sum of misclassification error probabilities?
- F is known as the Fisher discriminant function.

Duality between confidence sets and tests

- Confidence regions are random subsets of the parameter space that contain the true parameter with probability at least $1-\alpha$.
- Acceptance regions of statistical tests are, for a given hypothesis H_0, subsets of the sample space with probability of accepting H_0 at least $1-\alpha$ when H_0 is true.
- To illustrate the duality, we consider the example of two-sided tests for the mean of a normal distribution.
  - An established theory postulates the value θ_0 for a certain physical constant.
  - A scientist has reasons to believe that the theory is incorrect and measures the constant n times obtaining measurements $X_1, \ldots, X_n$.
  - Knowledge of his instruments leads him to assume that $X_1, \ldots, X_n$ are iid $N(\theta, \sigma^2)$.
  - Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.
  - A size α test can be obtained by the level $(1-\alpha)$ confidence interval

    $$[\bar{x} - st_{n-1}(1-\alpha/2)/\sqrt{n}, \bar{x} + st_{n-1}(1-\alpha/2)/\sqrt{n}].$$

    Namely, we *accept* H_0, if and only if, the postulated value θ_0 is a member of the level $(1-\alpha)$ confidence interval

    $$[\bar{x} - st_{n-1}(1-\alpha/2)/\sqrt{n}, \bar{x} + st_{n-1}(1-\alpha/2)/\sqrt{n}].$$

  - Set $T(\theta) = \sqrt{n}(\bar{X} - \theta)/s$. Because the same interval is used for every θ_0, it generates a family of level α tests. Reject H_0 if

    $$T_*(\mathbf{x}, \theta_0) = \begin{cases} 1 & \text{if } |T(\theta_0)| \geq t_{n-1}(1-\alpha/2) \\ 0 & \text{otherwise.} \end{cases}$$

    It leads to a two-sided test with size α.
  - On the other hand, by starting with the test T_*, we obtain the above confidence interval by finding the set of θ where $T_*(\mathbf{x}, \theta) = 0$.
    Note that

    $$P_{\theta_0}(T_*(\mathbf{x}, \theta_0) = 0) = P_{\theta_0}\left[\sqrt{n}\left|\frac{\bar{X} - \theta_0}{s}\right| \geq t_{n-1}(1-\alpha/2)\right] = 1-\alpha.$$

Now we consider the duality theorem.

Theorem 2 *For each* $\boldsymbol{\theta}_0 \in \Theta$*, let* T_{θ_0} *be a test for* $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ *(versus some* H_1*) with significance level* α *and acceptance region* $A(\boldsymbol{\theta}_0)$*. For eaxh* $\mathbf{x}$ *in the range of* X*, define*

$$C(\mathbf{x}) = \{\boldsymbol{\theta} : \mathbf{x} \in A(\boldsymbol{\theta})\}.$$

(i) $C(\mathbf{X})$ *is a level* $1-\alpha$ *confidence set for* $\boldsymbol{\theta}$*.*
(ii) If $T_{\boldsymbol{\theta}_0}$ *is a nonrandomized and has size* α *for every* $\boldsymbol{\theta}_0$*, then* $C(\mathbf{X})$ *has confidence coefficient* $1-\alpha$*.*
(iii) Let $C(\mathbf{X})$ *be a confidence set for* $\boldsymbol{\theta}$ *with significance level (or confidence coefficient)* $1-\alpha$*. For any* $\boldsymbol{\theta}_0 \in \Theta$*, define a region* $A(\boldsymbol{\theta}_0) = \{\mathbf{x} : \boldsymbol{\theta}_0 \in C(\mathbf{x})\}$*. Then the test* $T_*(\mathbf{X}) = 1 - I_{A(\boldsymbol{\theta}_0)}(\mathbf{X})$ *has significance level* α *for testing* $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ *versus some* H_1*.*

Proof. We only prove the first assertion. The proofs for the second and third assertions are similar.

- Under the given condition,

  $$\sup_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} P(\mathbf{X} \notin A(\boldsymbol{\theta}_0)) = \sup_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} P(T_{\boldsymbol{\theta}_0} = 1) \leq \alpha,$$

  which is the same as

  $$1 - \alpha \leq \inf_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} P(X \in A(\boldsymbol{\theta}_0)) = \inf_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} P(\boldsymbol{\theta}_0 \in C(\mathbf{X})).$$

- The above holds for all $\boldsymbol{\theta}_0$, the result follows from

  $$\inf_{P\in\mathcal{P}} P(\boldsymbol{\theta} \in C(\mathbf{X})) = \inf_{\boldsymbol{\theta}_0\in\Theta} \inf_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} P(\boldsymbol{\theta}_0 \in C(\mathbf{X})) \geq 1 - \alpha.$$

Remarks.

- $C(\mathbf{X})$ in Theorem 2 can be determined numerically, if it does not have an explicit form. Note that we just try to solve an equation as in the above example.
- Theorem 2 can be best illustrated in the case when $\boldsymbol{\theta}$ is real-valued and $A(\boldsymbol{\theta}) = \{Y : a(\boldsymbol{\theta}) \leq Y \leq b(\boldsymbol{\theta})\}$ for a real-valued statistic $Y(\mathbf{X})$ and some nondecreasing functions $a(\boldsymbol{\theta})$ and $b(\boldsymbol{\theta})$.
  - When we observe $Y = y$, $C(\mathbf{X})$ is an interval with limits $\underline{\boldsymbol{\theta}}$ and $\overline{\boldsymbol{\theta}}$, which are the $\boldsymbol{\theta}$-values at which the horizontal line $Y = y$ intersects the curve $Y = b(\boldsymbol{\theta})$ and $Y = a(\boldsymbol{\theta})$, respectively.
  - If $y = b(\boldsymbol{\theta})$ (or $y = a(\boldsymbol{\theta})$) has no solution or more than one solution, $\underline{\boldsymbol{\theta}} = \inf\{\boldsymbol{\theta} : y \leq b(\boldsymbol{\theta})\}$ (or $\overline{\boldsymbol{\theta}} = \sup\{\boldsymbol{\theta} : a(\boldsymbol{\theta}) \leq y\}$

Example 3.7 Assume $X_1, \ldots, X_n$ are iid binary random variables with $p = P(X_i = 1)$. Suppose that we need a lower confidence bound for p.

- Consider $H_0 : p = p_0$ versus $H_1 : p > p_0$.
- The acceptance region of a UMP test of size $\alpha \in (0, 1)$ is $A(p_0) = \{y : y \leq m(p_0)\}$, where $y = \sum_{i=1}^n x_i$ and $m(p_0)$ is an integer between 0 and n such that

$$\sum_{j=m(p_0)+1}^{n} C(n,j)p_0^j(1-p_0)^{n-j} \leq \alpha < \sum_{j=m(p_0)}^{n} C(n,j)p_0^j(1-p_0)^{n-j}.$$

- $m(p)$ is an integer-valued, nondecreasing step-function of p.
- Define

$$\underline{p} = \inf\{p : m(p) \geq y\} = \inf\left\{\sum_{j=y}^{n} C(n,j)p^j(1-p)^{n-j} \geq \alpha\right\}.$$

  $(\underline{p}, 1)$ is a level $1 - \alpha$ confidence interval.

Now we give an example in which neither estimation or testing is appropriate.

Example 3.8 A hazardous toxic waste site requires clean-up when the true chemical concentration θ in the contaminated soil is higher than a given level $\theta_0 \geq 0$.

- Because of the limitation in resources, we would like to spend our money and efforts more in those areas that pose high risk to public health.
- In a particular area where soil samples are obtained, we would like to take one of these three actions: a complete clean-up (a_1), a partial clean-up (a_2), and no clean-up (a_3).
- Suppose that the cost for a complete clean-up is c_1 and for a partial clean-up is $c_2 < c_1$; the risk to public health is $c_3(\theta - \theta_0)$ if $\theta > \theta_0$ and 0 if $\theta \leq \theta_0$; a complete clean-up can reduce the toxic concentration to an amount $\leq \theta_0$, whereas a partial clean-up can only reduce a fixed amount of the toxic concentration, i.e., the chemical concentration becomes $\theta - t$ after a partial clean-up, where t is a known constant.
- The loss in making a decision for the above can be written as

$L(\theta, a)$	a_1	a_2	a_3
$\theta \leq \theta_0$	c_1	c_2	0
$\theta_0 < \theta \leq \theta_0 + t$	c_1	c_2	$c_3(\theta - \theta_0)$
$\theta > \theta_0 + t$	c_1	$c_2 + c_3(\theta - \theta_0 - t)$	$c_3(\theta - \theta_0)$

- The true chemical concentration θ is unknown, we now use the Bayesian approach to tackle this question. θ is viewed as a realization of a random vector $\boldsymbol{\theta}$ whose prior distribution is Π. Π is based on past experience, past data, or statistician's belief and, thus, can be very subjective.
  Π is a probability distribution on a class of measurable sets in Θ.
- A sample X is drawn from $P_\theta = P_{x|\theta}$, which is viewed as the conditional distribution of X given $\boldsymbol{\theta} = \theta$. The sample $X = x$ is then used to obtain an updated prior distribution, which is called the *posterior* distribution.
- If both X and $\boldsymbol{\theta}$ are discrete, the Bayes formula appeared in elementary probability leads to

$$P(\boldsymbol{\theta} = \theta | X = x) = \frac{P(X = x|\boldsymbol{\theta} = \theta)P(\boldsymbol{\theta} = \theta)}{\sum_{\theta \in \Theta} P(X = x|\boldsymbol{\theta} = \theta)P(\boldsymbol{\theta} = \theta)}.$$

Assume that $\{P_{x|\theta} : \theta \in \Theta\}$ is dominated by a σ-finite measure ν and $f_\theta(x) = dP_{x|\theta}(x)/d\nu$. Suppose that $m(x) = \int_\Theta f_\theta(x) d\Pi > 0$ and $d\Pi/d\lambda = \pi(\theta)$ for a σ-finite measure λ. Then

$$\frac{dP_{\theta|x}}{d\lambda} = \pi(\theta)\frac{f_\theta(x)}{m(x)},$$

where $m(x) = \int_\Theta \pi(\theta) f_\theta(x) d\theta$.

- An observed result changes our degrees of belief in different parameter values by changing a prior distribution into a posterior distribution.
- In the Bayesian approach, the posterior distribution $P_{\theta|x}$ contains all the information we have about θ and, therefore, statistical decisions and inference should be made based on $P_{\theta|x}$, conditional on the observed $X = x$.

Choosing a prior distribution

- If we have accepted that degrees of belief can properly be described by probability distributions, it remains to establish a method of determining the appropriate prior distribution for each problem we encounter.
- In practice, prior knowledge is often rather vague and there is a whole class of prior distributions, each one of which is adequate for describing an individual's degrees of belief.
- What is the consequence of choosing a wrong prior?
- Robustness: It is claimed that the choice of prior distribution is not crucial as long as it is from a *good enough* class.
- The above claim can be justified via the following setting of asymptotic analysis.
  - Suppose that $\boldsymbol{\theta} = \theta$ (in frequentist terms).
  - Does the Bayes posterior distribution concentrate all mass more and more tightly around θ as $n \to \infty$?
  - Assume $\Theta = \{\theta_1, \ldots, \theta_k\}$ (finite). Let $\pi = P(\boldsymbol{\theta} = \theta_j)$, $j = 1, \ldots, k$ denote the prior distribution of $\boldsymbol{\theta}$.
  - Use Bayes theorem, we have
$$P(\boldsymbol{\theta} = \theta_j | X_1, \ldots, X_n) = \frac{\pi_j \prod_{i=1}^n p(X_i|\theta_j)}{\sum_{a=1}^k \pi_a \prod_{i=1}^n p(X_i|\theta_a)}.$$
  - Use the above fact, $P(\boldsymbol{\theta} = \theta_j | X_1, \ldots, X_n) = 0$ if $\pi_j = 0$. It means that no amount of data can convince a Bayesian who has decided a prior that θ_j is impossible.

- If all π_j are positive,

  $$\log \frac{P(\theta_a|X_1,\ldots,X_n)}{P(\theta_j|X_1,\ldots,X_n)} = n\left(\frac{1}{n}\log\frac{\pi_a}{\pi_j} + \frac{1}{n}\sum_{i=1}^n \log\frac{p(X_i|\theta_a)}{p(X_i|\theta_j)}\right).$$

  By the weak law of large numbers, under P_{θ_j},

  $$\frac{1}{n}\sum_{i=1}^n \log\frac{p(X_i|\theta_a)}{p(X_i|\theta_j)} \to E_{\theta_j}\left(\log\frac{p(X_1|\theta_a)}{p(X_1|\theta_j)}\right)$$

  in probability. But

  $$E_{\theta_j}\left(\log\frac{p(X_1|\theta_a)}{p(X_1|\theta_j)}\right) < 0$$

  if $\theta_a \neq \theta_j$. Therefore,

  $$\log\frac{P(\theta_a|X_1,\ldots,X_n)}{P(\theta_j|X_1,\ldots,X_n)} \to -\infty.$$

- This proof depends critically on the consistency of the MLE. In fact, we can also derive asymptotic normality of Bayes estimator under regularity conditions.

- Another major issue that arises is computation. There are two major methods: Laplace's method and Markov chain Monte Carlo.

For Example 3.8, it concerns about the cost. Under Bayesian analysis, it can be dealt as follows:

- The average loss for the action $a(X)$, which is called the risk of using a, is defined to be

  $$R_a(P) = E[L(P, a(X))] = \int_{\mathcal{X}} L(P, a(x))dP(x).$$

  $R_a(P)$ is also denoted by $R_a(\theta)$ if P is a parametric family indexed by θ.

- Since $\theta \sim \Pi$, we use the Bayesian approach by considering an average of $R_a(\theta)$ over θ:

  $$r_a(\Pi) = \int_\Theta R_a(\theta)d\Pi(\theta),$$

  which is called the *Bayes risk* of a with respect to Π.

- We now find a good action with the smallest Bayes's risk. Observe that

  $$\begin{aligned} r_a(\Pi) &= \int_\Theta \pi(\theta)d\lambda \int_{\mathcal{X}} L(\theta, a(x))f_\theta(x)d\nu \\ &= \int_{\mathcal{X}} m(x)d\nu \int_\Theta L(\theta, a(x))\frac{f_\theta(x)\pi(\theta)}{m(x)}d\lambda, \end{aligned}$$

since

$$\pi(\theta)f_\theta(x) = \frac{dP_{x|\theta}}{d\nu}(x)\pi(\theta)d\lambda = m(x)\frac{f_\theta(x)\pi(\theta)}{m(x)}.$$

Here $m(x)$ is the marginal density function and $f_\theta(x)\pi(\theta)/m(x)$ is the posterior density function of θ after observing x.
Hence, we choose a to minimize the expected posterior loss for fixed x since all terms in the above double integral are non-negative.

- For this problem, a Bayes action can be obtained by comparing

$$E_{\theta|x}[L(\theta, a_j)] = \begin{cases} c_1 & j = 1 \\ c_2 + c_3 E_{\theta|x}[\psi(\boldsymbol{\theta}, t)] & j = 2 \\ c_3 E_{\theta|x}[\psi(\boldsymbol{\theta}, 0)] & j = 3, \end{cases}$$

where $\psi(\theta, t) = (\theta - \theta_0 - t)I_{(\theta_0+t,\infty)}(\theta)$.

Revisits of Decision Theory

When we talk about the estimation, we discuss how to measure the performance of an estimator $T(\mathbf{X})$ of a parameter $q(\boldsymbol{\theta})$? When we do hypothesis testing, we discuss why to adopt the Neyman-Pearson framework and how to finds a UMP test and etc. These lead to

- clarify the objectives of a study,
- point to what the different possible actions are,
- provide assessments of risk, accuracy, and reliability of statistical procedures,
- provide guidance in the choice of procedures for analyzing outcomes of experiments.

This leads to the consideration of decision theoretic framework. We begin with a statistical model with an observation vector $\mathbf{x}$ whose distribution P ranges over a set $\mathcal{P}$. We usually take $\mathcal{P}$ to be parametrized, $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$.

- **Action space** A: It is the space of actions or decisions or claims that we can contemplate making.
  - For point estimation of $q(\boldsymbol{\theta})$, a is $T(\mathbf{x}) \in R^d$.
  - For testing, only two actions are contemplated: accepting or rejecting the *specialness* of P (i.e. $H_0 : P \in \mathcal{P}_0$).

- **Loss function** $\ell(\boldsymbol{\theta}, a)$: It is defined as a function

  $$\ell : \mathcal{P} \times \mathcal{A} \to R^+.$$

  It is the loss incurred if we take action a when $\boldsymbol{\theta}$ is the true state of nature.
  Think of mean squared error in estimation and $0 - 1$ loss in testing (Type I error and Type II error).

- **Decision procedures** $\delta(\mathbf{x})$:

- **Risk function (expected or average loss)**:

  $$R(\boldsymbol{\theta}, \delta) = E[\ell(\boldsymbol{\theta}, \delta(\mathbf{X}))].$$

  Consider the $0 - 1$ loss in testing (α and β).

Comparison of decision procedures

How do we find the *optimum* procedure?

- Two decision rules which have the same risk function are considered equivalent from the point of view of decision theory.

- Avoid *bad* procedure approach: Avoid the estimate that can be improved by others.
  A procedure δ *improves* a procedure δ^* if, and only if,

  $$R(\boldsymbol{\theta}, \delta) \leq R(\boldsymbol{\theta}, \delta^*)$$

  for all $\boldsymbol{\theta}$ with strict inequality for some $\boldsymbol{\theta}$.
  If the above hold, δ^* is then called *inadmissible.*
  Typically, there is no rule δ that improves all others. (You can think in terms of estimation.)

- Minimax approach: Look at the worst possible risk.

  $$\sup_{\boldsymbol{\theta}} R(\boldsymbol{\theta}, \delta^*) = \inf_{\delta} \sup_{\boldsymbol{\theta}} R(\boldsymbol{\theta}, \delta).$$

- Bayesian approach: Consider an average of $R(\boldsymbol{\theta}, \delta)$ over $\boldsymbol{\theta}$.

  $$r(\delta) = E[R(\boldsymbol{\theta}, \delta)] = \int R(\boldsymbol{\theta}, \delta) d\pi(\boldsymbol{\theta}).$$

  - π is called a prior density in Bayesian analysis. (We can think of π as a weight function in general.)
  - $r(\delta)$ is called the *Bayes risk* of δ.
  - In the Bayesian analysis, $\boldsymbol{\theta}$ is viewed as a realization of a random vector whose prior distribution is π.

- The prior distribution is based on past experience, past data, or statistician's belief and, thus, can be very subjective.
- A sample $\mathbf{X}$ is drawn from $P_{\mathbf{x}|\boldsymbol{\theta}}$, which is viewed as the conditional distribution of $\mathbf{X}$ given $\boldsymbol{\theta}$.
- The sample $\mathbf{x}$ is then used to obtain an updated prior distribution, which is called the *posterior* distribution.

How to find Bayes estimate?

- Consider the Bayes risk

$$r(\pi, \delta) = E[R(\boldsymbol{\theta}, \delta)] = E[\ell(\boldsymbol{\theta}, \delta(\mathbf{X}))],$$

  where $(\boldsymbol{\theta}, \mathbf{X})$ are random vector.

- Write $r(\pi, \delta)$ as

$$E\{E[\ell(\boldsymbol{\theta}, \delta(\mathbf{X}))|\mathbf{X}]\}.$$

- Recall that $\ell(\boldsymbol{\theta}, \delta(\mathbf{X}))$ is nonnegative. A possible strategy is to find the minimizer of $E[\ell(\boldsymbol{\theta}, \delta(\mathbf{X}))|\mathbf{X}]$.
  $E[\ell(\boldsymbol{\theta}, \delta(\mathbf{X}))|\mathbf{X}]$ is called the posterior risk.
  - For the squared error loss, the minimizer is

$$\delta^*(\mathbf{x}) = E[\boldsymbol{\theta}|\mathbf{x}].$$

  - For the absolute error loss, the minimizer is the median of posterior distribution.

- The function $\delta^*(\mathbf{x})$ with

$$E[\ell(\boldsymbol{\theta}, \delta^*(\mathbf{X}))|\mathbf{x}] = \inf_a E[\ell(\boldsymbol{\theta}, a(\mathbf{X}))|\mathbf{x}]$$

  is called a Bayes estimate.

Example 3.9 Suppose $X_1, \ldots, X_n$ is a $N(\theta, \sigma_0^2)$ sample, where σ_0^2 is known and θ is unknown. The prior distribution of θ is $N(\eta_0, \tau_0^2)$.

- If we observe $\sum_i X_i = s$, the posterior distribution $\pi(\theta|\mathbf{x})$ is a normal density with mean

$$\left(\frac{\sigma_0^2}{\tau_0^2} + n\right)^{-1} \left(s + \frac{\eta_0 \sigma_0^2}{\tau_0^2}\right)$$

  and variance

$$\left(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}\right)^{-1}.$$

- For the squared error loss, the Bayes estimate is

$$\delta^*(\bar{X}) = \eta_0^2 \frac{1/\tau_0^2}{n/\sigma_0^2 + 1/\tau_0^2} + \bar{X}\frac{n/\sigma_0^2}{n/\sigma_0^2 + 1/\tau_0^2}.$$

- The Bayes risk is

$$\frac{1}{n/\sigma_0^2 + 1/\tau_0^2}.$$

How to find minimax procedures?

- A worst-case analysis: (The true $\boldsymbol{\theta}$ is one that is as hard as possible.)
  δ_1 is better than δ_2 from a minimiax point of view if

$$\sup_{\boldsymbol{\theta}} R(\boldsymbol{\theta}, \delta_1) < \sup_{\boldsymbol{\theta}} R(\boldsymbol{\theta}, \delta_2).$$

- Suppose that we want to estimate the mean θ of a normal distribution with known variance σ_0^2 and use the squared error loss as a criterion.
  - Note that $E(\bar{X}_n - \theta)^2 = \sigma_0^2/n$.
  - The risk of using $\bar{X}$ as estimate of θ does not depend on the value of unknown parameter θ.
  - In Example 3.9, we just describe a Bayes estimate with Bayes risk is $(n/\sigma_0^2 + 1/\tau_0^2)^{-1}$.
    When n is large, the Bayes risk tends to σ_0^2/n.
  - Observe that

$$\sup_{\theta} R(\theta, \delta) \geq E_\pi(R(\theta, \delta)) \geq E_\pi(R(\theta, \delta^*)),$$

    where δ^* is Bayes estimate of θ under prior π.
  - For the above set-up,

$$E_\pi(R(\theta, \delta^*)) = (n/\sigma_0^2 + 1/\tau_0^2)^{-1} = \sup_{\theta} R(\theta, \bar{X}_n) - \frac{\sigma_0^2}{n}\frac{1}{\sigma_0^2/n + \tau_0^2}.$$

    When $\tau_0^2 \to \infty$,

$$\frac{E_\pi(R(\theta, \delta^*))}{\sup_\theta R(\theta, \bar{X}_n)} \to 1.$$

    We conclude that $\bar{X}_n$ is minimax.

Theorem 3 *Let δ^* be a rule such that $\sup_\theta R(\theta, \delta^*) = r < \infty$, let $\{\pi_k\}$ denote a sequence of prior distributions such that $\pi_k\{\theta : R(\theta, \delta^*) = r\} = 1$, and let $r_k = \inf_\delta r(\pi_k, \delta)$, where $r(\pi_k, \delta)$ denotes the Bayes risk wrt π_k. If*

$$r_k \to r \quad \text{as } k \to \infty,$$

then δ^* is minimax.

If δ^* is a Bayes rule whose risk is constant on $\boldsymbol{\theta}$, then δ^* is minimax.

How to find admissible procedures?

- A method to eliminate bad procedures.
- Any procedure which is strictly dominated by another is said to be *inadmissible.*
  δ strictly dominates δ^* if $R(\boldsymbol{\theta}, \delta) \leq R(\boldsymbol{\theta}, \delta^*)$, for all $\boldsymbol{\theta}$, and this inequality is strict for some $\boldsymbol{\theta}$.
- When Θ is finite and δ^* is Bayes with respect to a prior frequency function π such that $\pi(\boldsymbol{\theta}) > 0$ for every $\boldsymbol{\theta} \in \Theta$, then δ^* is admissible.
- When Θ is an interval and δ^* is Bayes with respect to a prior density π such that $\pi > 0$ on Θ and $R(\boldsymbol{\theta}, \delta)$ is a continuous function of $\boldsymbol{\theta}$ for all δ, then δ^* is admissible.
  If
  $$\frac{r_k(\delta^*) - r_k(\delta)}{\int_a^b \pi_k(\theta) d\theta} \to 0$$
  as $k \to \infty$ for every fixed $a < b$, then δ^* is admissible.

Example 3.10 Under the setting of Example 3.9, $\bar{X}$ is also admissible.

- Set π_k to be $N(\mu, k)$.
- Observe that
  $$\frac{r_k(\bar{X}) - r_k(\delta)}{\int_a^b \pi_k(\theta) d\theta} = \frac{(\sigma^2/n)(1/(1+k))}{\int_a^b \pi_k(\theta) d\theta} \to 0$$
  as $k \to \infty$.