ISyE 6412: Thereotical Statistics

- **Lectures: TR 8:00-9:15am, IC 209**
- **Instructor: Dr. Yajun Mei**
  <ymei@isye.gatech.edu>
  Instructor Office Hours: 9:15-9:45am TR, Groseclose #343
- **TA: TBD**
- **Course Homepage: see Canvas**
- **HW#1 is due at Canvas at 9:30am on Thur, Aug 29.**

My academic pathway

- **Undergraduate: Math, Peking Univ., BS in 1996**
- **Work as a computer programmer in a Chinese bank, 1996-1998**
- **Graduate: PhD in Math with a minor in EE, Caltech, 1998-2003 (advisor: Dr. Gary Lorden)**
- **Post Doc in biostatistics: FHCRC, Seattle, 2003-Sep 2005 (supervisor: Dr. Sarah Holte)**
- **New Research Fellow: SAMSI & Duke Univ., Fall 2005**
- **Joined ISyE of GT since Jan 2006, and received tenure in 2011. Currently an associate professor.**
- **Current Research Interests: Streaming Data Analysis, High-dim change-point detection, IE, Biostatistics, etc.**

Lecture 1

Agenda

- Course organization
- Introduction to Statistical Decision Theory

Organization of the Course

- **Course homepage: Please check Canvas for handouts, HWs and solution sets.**
- **Text book: "Statistical Inference" (2nd edition), by Casella and Berger. Notes/slides provided.**
- **Pre-requisites: Ch. 1 – 5 of the textbook. (Solid math skills, esp. calculus).**
  This is a required core graduate course for PhD students in Statistics. In the past, half of students are PhD, and half are MS.
- **Grading:**
  - **Homework: 20%**
  - **Midterms: 20%+20% (Thur, Sept 26 & Thur, Nov 07)**
  - **Final: 40% (Dec. 12, Thur, 8:00-10:50am)**

Catalog Description

- **Rigorous introduction to theory of statistical inference. Estimation and testing. Construction and assessment of estimators and tests. Fundamentals of decision theory, minimax, and Bayes Paradigms.**

Main Topics of the Course

- **Statistical Decision Theory**
  - **Risk function**
  - **Bayesian, Minimax, Admissible**
- **Sufficient Statistics**
  - **Sufficient Statistics: Factorization theorem**
  - **Minimal sufficient**
  - **Complete statistics (Basu' theorem)**
- **Point Estimator:**
  - **Method of Moments**
  - **Maximum Likelihood Estimator: asymptotic properties**
  - **The best unbiased estimator: how to find?**
    - **Complete sufficient statistics**
    - **Information Inequality**
- **Testing Hypothesis: Type I & Type II errors, power, p-value, Likelihood Ratio Test, Neyman-Pearson Lemma.**
- **Confidence Interval (if we have time)**

This Lecture

- **Introduction to Statistical Decision Theory**

Modern Statistics/Statisticians

- Modern Statistics relay more on computers (any not?)
- Statisticians not only use computers, but must use brain to think: how to explain the results from computers? Why to get certain outcome? Can we improve it? How to improve it?
- You might know how to use R, SAS, Python, or cloud computing, but if you do not use your brain to understand and interpret the outcomes, then you are just a programmer, not a statistician.
- If you use statistical methods appropriately, and work hard to genuinely think and understand the data sources, the outcomes and interpretations, then what you are doing is exactly the Statistics, no matter whether you receive PhD in statistics or not.
- In any field full with dogmatic practices, jobs/human will be replaced by machines/robots.

Probability versus Statistics

A typical problem in probability theory:

- **specify the sample space and its underlying probability law, and**
- **compute the probability of a given chance event.**

For example, if $X_1, \cdots, X_n$ are iid Bernoulli RV with $P(X_i=1) = p$ and $P(X_i = 0) = 1-p$, what is $P(X_1+\cdots+X_n=r)$, where r is an integer in [0,n].

- **What is a typical problem in Statistics?**

A typical problem in Statistics

A typical problem in Statistics:

- We observe the outcome from some studies or experiments
- We specify a class of probability laws --- we know that the true underlying probability law is a member of this class, but we do not know which one it is.
- The objective might be to determine a "good" way of guessing on the basis of the outcome of experiment, which of the possible underlying probability law is the one which actually governs the experiment whose outcome we are to observe.

Example 1

- We are given a coin about which we know nothing. We are allowed to perform 10 independent flips with the coin, on each of which the probability of getting a heads is p. We do not know p, but only know that $0 < p < 1$.

- Objective: on the basis of what we observe on the 10 flips, guess the value of p.

  (Of course, the outcome of this experiment cannot tell us with complete sureness the exact value of p. But we can try to make an accurate guess).

Two ways to form a guess

1. We might compute
   # of heads obtained in 10 tosses / 10
   and use this as our guess;

2. Or we might observe how many flips it took to get the first head, and guess
   p=0 if no heads come up in 10 tosses, and guess p to be
   p=1/#of tosses on which first heads appears if at least one heads occurs.

   (can you think other ways to make a guess?)

Two more ways to guess

3. We can guess p to be $\pi/8$ or 2/3, depending on whether the number of heads in 10 tosses is odd or even;

4. We can also ignore the experiment entirely and always guess p =1/2.

(Are they valid guesses? Are they reasonable?)

A Comparison of these 4 guesses

- **Can you "prove that method 1 is better than method 3"?**

- **Many statistical textbooks will tell you to use method 1, but probably you will not find a reference to any other methods;**
- **You will find some mention of a "Principle of unbiased estimation" or "Principle of Maximum Likelihood" to justify the use of method 1, but you may not find any real justification of the use of these "principles" in obtaining guessing methods.**
- **In fact, you will not even find any satisfactory discussions of what is meant by a "good" method, let it alone a proof that method 1 really is good.**

Statistical Inference

- Surely, it should be possible to make precise what is meant to be a "good" guessing method and to determine which methods are good.
- Such a rational approach to the subject to "statistical inference" (that is, to the subject to obtaining good guessing methods) came into being with the work of Neyman in 1930s and of Wald in 1940s.
- This course discusses the important ideas of statistical inference.

  First, we must describe the way in which a statistical problem is specified.

Specification of a Statistical Problem

Suppose that the statisticians can observe the outcome X(=$(X_1, \ldots, X_n)$) of an experiment. We need to specify:

- **S: the sample space** (i.e., possible value of X);
- **Ω: the set of all possible distribution functions** (or probability laws) of X. Often referred as "possible state of nature". In the parametric approach, it is convenient to think of df's as being labeled by θ;
- **D: the decision space** (or collection of possible actions at the conclusion of the experiment);
- **L: the Loss function** L(θ, d).

Example 1

In the previous example of flipping a coin 10 times, and want to estimate the problem of getting a heads.

- **What is S (the sample space)?**
- **What is Ω (the set of all possible distribution functions)?**

Example 1 (Cont.)

We can let $X = (X_1, X_2, \ldots, X_{10})$, where X_i is 1 or 0 according to whether the i-th toss is a head or a tail. Then

- S (sample space): consists of 2^{10} possible value of $X = (X_1, X_2, \ldots, X_{10})$.
- Ω: the class of possible df's consists of all df's for which the X_i's are iid with $P(X_i = 1) = 1 - P(X_i = 0) = \theta$, where $0 \leq \theta \leq 1$. It is convenient to think of these df's as being labeled by θ.
  - WLOG, we will simply say that "$\theta = 1/3$" meaning that "the X_i's are iid with $P(X_i = 1) = 1 - P(X_i = 0) = 1/3$." So $\Omega = \{0 \leq \theta \leq 1\}$.

Remark #1

- It is implied in our discussion of specifying a statistical problem that by concluding an experiment the statistician can obtain some information about the "actual state of nature" θ in Ω.
- Often by the law of large numbers that if one could repeat the experiment an infinite number of times, then one could discover θ exactly.
- In reality, the statistician performs only a finite number of experiments. The more experiments made, the better is the approximation to perform an infinite # of experiments.

Remark #2

It is frequently the case that statisticians or experimenters proceed as follows.

- We decide on an experiment and specify the statistical problem (as described later).
- We then decide whether the results that would be obtained would be accurate enough
- If not, we decide on a new experiment (often working only more replications) and work out the specification of a new statistical problem.

This behavior of a statistician is often called "design of an experiments" (ISyE 6413).

Notation

Returning to our general considerations, we introduce some notation:

- $\mathbf{P}_{\theta_0}(A)$ **(or** $\mathbf{P}_{F_0}(A)$**) to mean "the probability of the chance event A, computed when the underlying probability law of X is given by** $\theta = \theta_0$**.**
- **We shall carry this notation over to expectations** $\mathbf{E}_{\theta_0}(g(X))$ **and** $\mathbf{Var}_{\theta_0}(g(X))$**.**
- **In our example 1,**

$$\mathbf{P}_\theta\{\sum_{i=1}^{10} X_i = 4\} = \binom{10}{4}\theta^4(1-\theta)^6.$$

$$\mathbf{E}_\theta(\textstyle\sum_{i=1}^{10} X_i) = 10\theta, \qquad \mathbf{Var}_\theta(\textstyle\sum_{i=1}^{10} X_i) = 10\theta(1-\theta).$$

Decision Space

- We now describe the next aspect of a statistical problem: the collection of possible actions which the statistician can take, or of possible statements which can be made, at the conclusion of the experiment.
- We denote this collection by **D, the decision space,** its elements being called decisions.
- At the conclusion of the experiment, the statistician actually only choose <u>one</u> decision out of the possible choices in D.
  And the statistician must make such a decision.

Example 1(a)

Continue Example 1 (flipping a coin):

- **The statistician required to guess the value of θ. In this case, we can think of D as the set of real numbers d satisfying $0 < d < 1$:**

$$D = \{d : 0 \leq d \leq 1\} = [0, 1].$$

Thus, the decision "0.35" stands for the statement "my guess is that θ is 0.35."

(Point Estimation)

Example 1(b)

- Suppose a gambler does NOT want to have a numerical guess as to the value of θ, but only wants to know whether the coin is fair, is biased towards to heads or is biased towards tails.
  In this case $D = \{d_1, d_2, d_3\}$, where
  d1 stands for "the coin is fair"
  d2 stands for "the coin is biased towards heads"
  d3 stands for "the coin is biased towards tails."

Note that in both examples, D can be viewed as the collection of possible answers to a question.

(Multi-Hypothesis testing)

Example 1(c)

- An even simpler question would be "Is the coin fair?" We would have $D = \{d_1, d_2\}$, where
  d1 means "Yes, the coin is fair"
  d2 means "No, the coin is biased".

- Application: Suppose the US mint attempted to discourage gamblers from unfairly using its coin by issuing only coins which were judged to be fair. The Mint's statisticians would either throw it into a barrel to be shipped to a bank (d1) or throw it back into the melting pot (d2).

(Binary Hypothesis testing)

Example 1(d)

- Suppose that our gambler does NOT merely want a guess as to the value of θ which governs the coin at hand, but rather a statement of an interval of values which is thought to include θ.
- In this case, we can think of each element d of D as being an interval of real numbers from d′ to d″ inclusive, where $0 \leq d' \leq d'' \leq 1$, and D is the set of all such intervals.
  d=[0.35, 0.40] → "I guess $0.35 \leq \theta \leq 0.40$"

  (Confidence Interval. Driving distance & time)

Loss Function

- In order to specify a statistical problem completely, we must state precisely how right or wrong the various possible decisions are for each possible underlying probability law of X.

- In many cases it is useful to think that making a correct decision causes us to incur no loss, whereas any incorrect decision may cause us to incur some positive loss.

Example 1(c) cont.

- In example 1(c) "is the coin fair?" it may be that, if the true value θ is fairly close to 0.50, say, $0.495 < \theta < 0.505$, then we deem it correct to make decision d1 ("the coin is fair") and incorrect to make decision d2, since the coin is judged to be close enough to fair for us to call it such.
  On the other hand, if $|\theta - 0.50| \geq 0.005$, we may feel that d2 is to be thought of as correct and d1 as incorrect.
- One possible loss function is:

$$L(\theta, d_1) = \begin{cases} 0, & \text{if } |\theta - 0.5| < 0.005 \\ L, & \text{otherwise.} \end{cases}$$

$$L(\theta, d_2) = \begin{cases} L, & \text{if } |\theta - 0.5| < 0.005 \\ 0, & \text{otherwise.} \end{cases}$$

Loss Function

- In any statistical problem, we define $L(\theta, d)$ to be the loss incurred if the true distribution of X is given by θ and the statistician makes decision d.
- Important: the loss $L(\theta, d)$ must be stated precisely for every possible θ and every possible decision d in D. Why?
  - The choice of a good statistical procedure (guessing rule) depends on L;
  - A procedure which is good for one loss function L may be very poor for another loss function which might be appropriate in other circumstances.

Loss Function (cont)

How do we obtain the loss function?

- **In practice, it is often difficult to judge the relative seriousness of the various possible incorrect decisions;**
- **Sometimes L can be written down in units of dollars on the basis of economic considerations, but in other settings, particularly in basic scientific research, the losses are difficult to make precise (better measurements → winning Nobel prize?)**
- **Practically speaking, there are many important practice setting where we can find "one" statistical procedure which is fairly good for any of a variety of Ls whose form are somewhat similar.**

Example 2

- Suppose the astronomer's measurements are iid N(μ, σ^2), where mean μ is the actual distance to the star. If this distance is known to be at least h light-years but nothing else is known. Then
- $\Omega = \{\theta = (\mu, \sigma^2) : \mu \geq h, \sigma > 0\}.$
- The decision is a real number: $D = \{d : d \geq h\}.$ and "d= 100" means "I guess the distance to be 100 light years."
- The loss function would reflect the fact that the larger the amount by which the astronomer misguess, the larger the loss. For example, $L((\mu, \sigma^2), d) = |\mu - d|$ or $(\mu - d)^2.$
- In practice, it is often suffice to make a rough guess as to the form of L.

Example 2 (Cont)

- Although the same statistical procedure may be good for many L's in similar form, such a procedure may be a poor one if we are faced with a L which is quite different.
- For example, suppose that under-guesses of the value of μ are more serious than over-guesses (they may result in giving the dog too little fuel). This may be reflected in a loss function such as

$$L((\mu, \sigma^2), d) = \begin{cases} 2(\mu - d), & \text{if } d \leq \mu; \\ d - \mu, & \text{if } d \geq \mu. \end{cases}$$

Example 3

We shall now give an example to illustrate how the L might be determined on economic ground.

- Suppose a middleman buys some products from a manufactory at $15 per unit. He can have two selling strategies:
  - d1: He can sell them to customers at $25 per unit with a guarantee that a barrel contains at most 2% defective items (otherwise the customer receives his money back and keep the item)
  - d2: he can sell the items without guarantee at $18 per unit.
- Let θ = the (unknown) proportion of defectives in a given unit.
  What is the loss function?

Example 3 (Cont)

Profit	Proportion of Defective	
	$\theta \leq 0.02$	$\theta > 0.02$
d1 (with guarantee)	\$25-\$15 = \$10	0 - \$15 = -\$15
d2 (without guarantee)	\$18-\$15 = \$3	\$18-\$15 = \$3

Loss = - Profit:

$$L(\theta, d1) = \begin{cases} -10 & if\ \theta \leq 0.02 \\ 15 & if\ \theta > 0.02 \end{cases}$$ and

$$L(\theta, d2) = \begin{cases} -3 & if\ \theta \leq 0.02 \\ -3 & if\ \theta > 0.02 \end{cases}$$

Loss function

- **Sometimes statisticians work not with the absolute loss, but rather with a quantity called the "regret":**

$$L^*(\theta, d) = L(\theta, d) - \min_{d \in D} L(\theta, d).$$

That is, for each θ, the excess of the loss L(θ, d) over the minimum possible loss for this θ.

Example 3 (Cont)

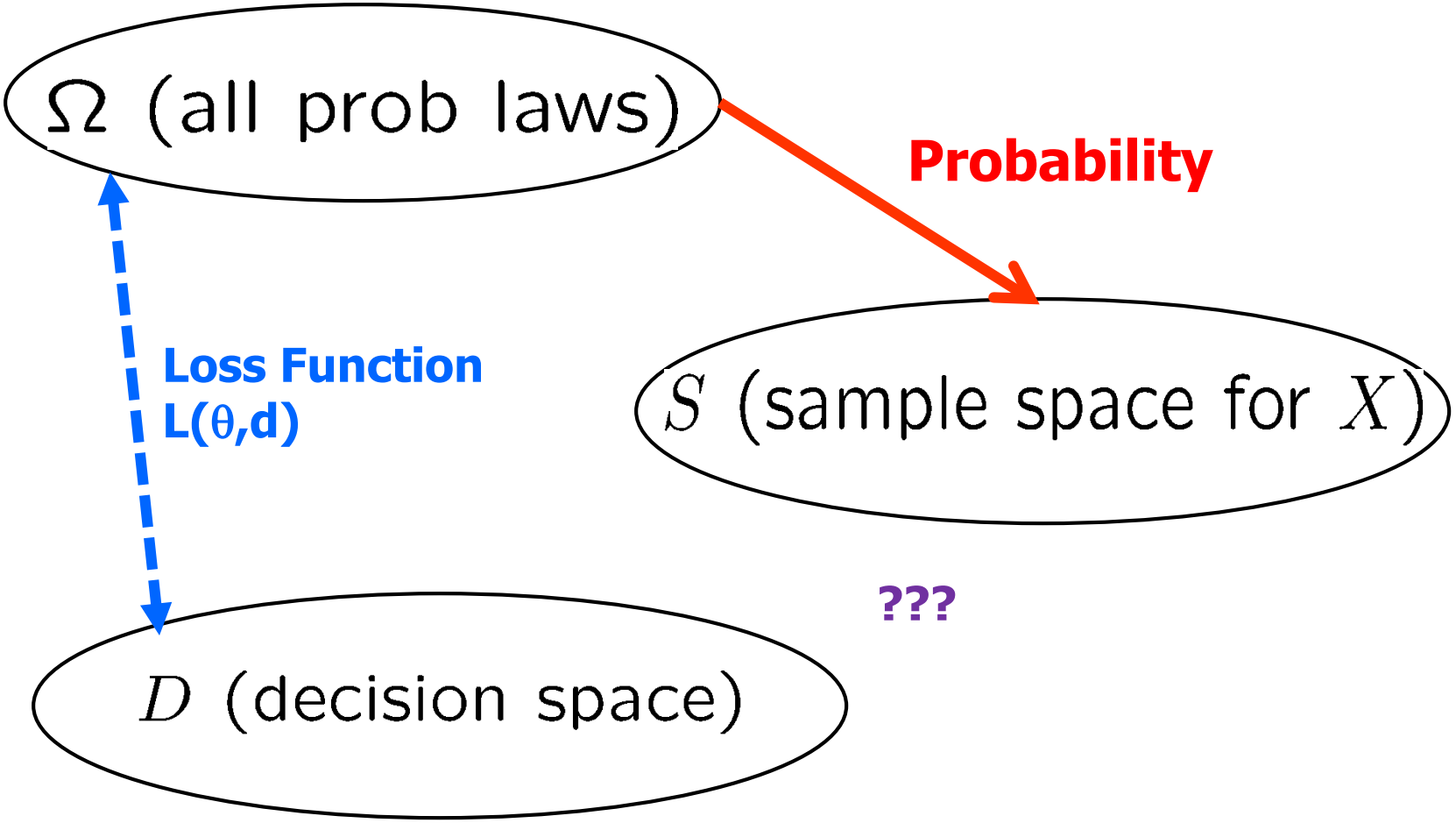
Loss	Proportion of Defective	
("regret" function)	$\theta \leq 0.02$	$\theta > 0.02$
d1 (with guarantee)	-10 (0)	15 (18)
d2 (without guarantee)	-3 (7)	-3 (0)

In our example, the loss (or "regret") function is

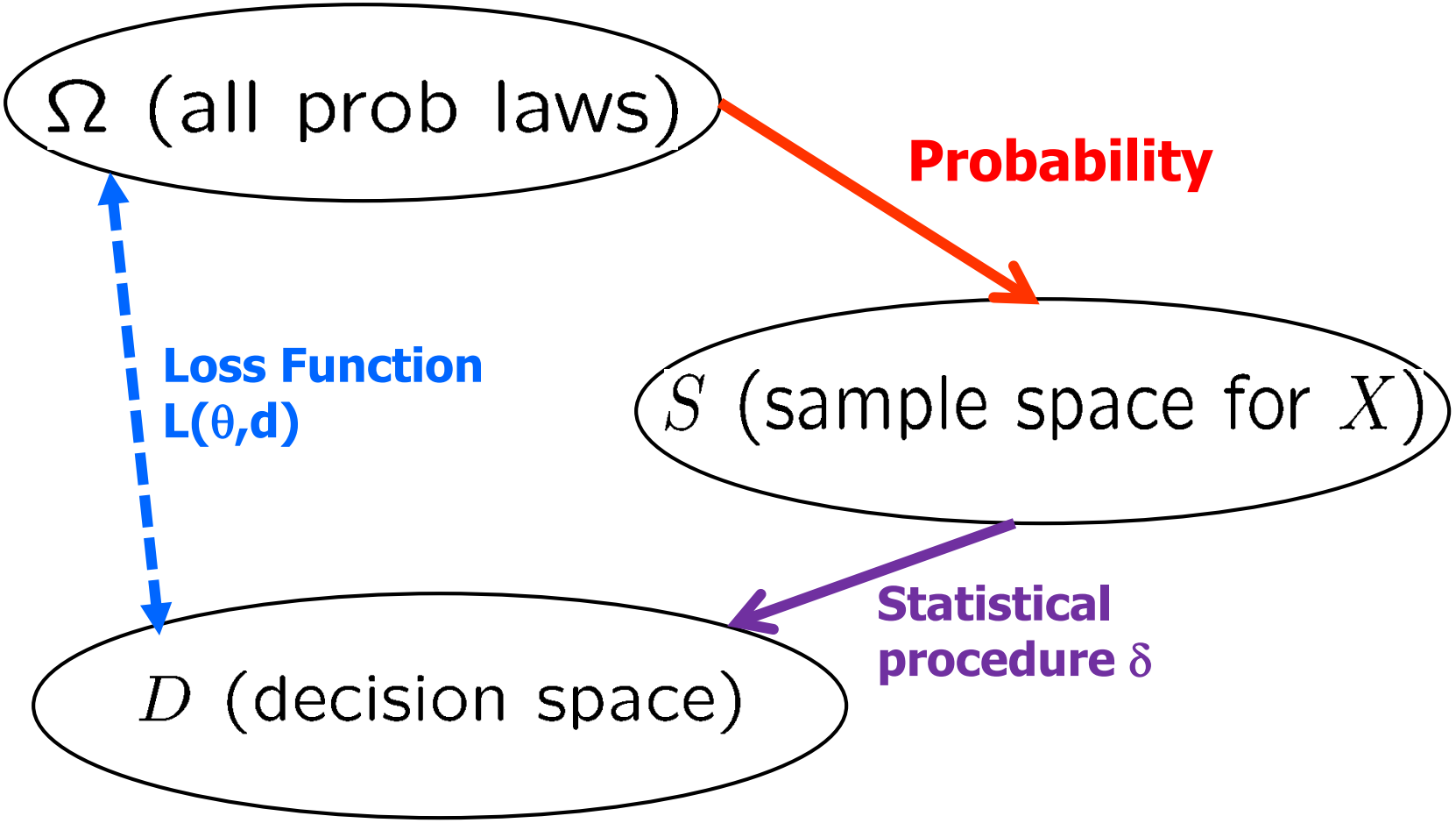
$$L^*(\theta, d1) = \begin{cases} 0 & if\ \theta \leq 0.02 \\ 18 & if\ \theta > 0.02 \end{cases} \quad \text{and}$$

$$L^*(\theta, d2) = \begin{cases} 7 & if\ \theta \leq 0.02 \\ 0 & if\ \theta > 0.02 \end{cases}$$

Statistical Problems



Statistical Problems



A Statistical procedure

- **A statistical procedure δ is a function from S into D (which we earlier called a "guessing" method and which is sometimes also called a "decision function")**
- **The experiment is conducted, the chance variable X being observed to take on a value x0, an element of S. The function δ then assigns an element d0 of D, given by d0 $=\delta$(x0). The statistician makes decision d0.**
- **There are many statistical procedures in any problem.**

Example 2(a)

When estimating θ=P(heads) based on 10 flips $X=(X_1, \cdots, X_{10})$,

- **The class of all possible df's:** $\Omega = \{\theta : 0 \leq \theta \leq 1\}.$
- **The decision space:** $D = \{d : 0 \leq d \leq 1\}.$
- **Four statistical procedures:**

$$\delta_1(X_1,\ldots,X_{10}) = \sum_{i=1}^{10} X_i/10;$$

$$\delta_2(X_1,\ldots,X_{10}) = \begin{cases} 1/j, & \text{if } X_1 = \ldots = X_{j-1} = 0 \text{ and } X_j = 1, \\ 0, & \text{if } X_1 = \ldots = X_{10} = 0; \end{cases}$$

$$\delta_3(X_1,\ldots,X_{10}) = \begin{cases} \pi/8, & \text{if } \sum_{i=1}^{10} X_i \text{ is odd}, \\ 2/3, & \text{if } \sum_{i=1}^{10} X_i \text{ is even}; \end{cases}$$

$$\delta_4(X_1,\ldots,X_{10}) = 1/2.$$

Example 2(b)

- **In the problem when D={d1,d2,d3} with d1 (fair), d2 (biased heads) and d3 (biased tails), two possible statistical procedures are**

$$\delta'(X_1,\ldots,X_{10}) = \begin{cases} d_1, & \text{if } \sum_{i=1}^{10} X_i = 5 \\ d_2, & \text{if } \sum_{i=1}^{10} X_i > 5 \\ d_3, & \text{if } \sum_{i=1}^{10} X_i < 5 \end{cases}$$

$$\delta''(X_1,\ldots,X_{10}) = d_1.$$

(δ" ignores the data and states that the coin is fair)

What is a good procedure?

- In any practical problems, any statistical procedures we use can possibly lead to an unfortunate decision.
- The essential point: such incorrect decisions, although possible, will not necessarily be very probable.
- Intuitively, a good statistical procedure is one for which the probability is large that a favorable decision will be made, whatever the true df θ or F may be.
  Expected loss or risk function!

Risk Function

- **Given S (sample space), Ω (all possible prob. Laws), D (decision space), and the loss function L(θ, d),**

  A statistical procedure δ is a function from S into D.

- **The risk function of δ is**

$$
\begin{aligned}
R_\delta(\theta) &= R(\theta,\delta) = \mathbf{E}_\theta(L(\theta,\delta(\mathbf{X})) \\
&= \sum_{d\in D} L(\theta,d)\mathbf{P}_\theta(\delta(\mathbf{X})=d) \quad \text{(if D discrete)} \\
&= \begin{cases} \sum_{\mathbf{x}} L(\theta,\delta(\mathbf{x}))\mathbf{P}_\theta(\mathbf{X}=\mathbf{x}), & \text{if S discrete;} \\ \int L(\theta,\delta(\mathbf{x}))\, f_\theta(\mathbf{x})d\mathbf{x}, & \text{if S continuous.} \end{cases}
\end{aligned}
$$

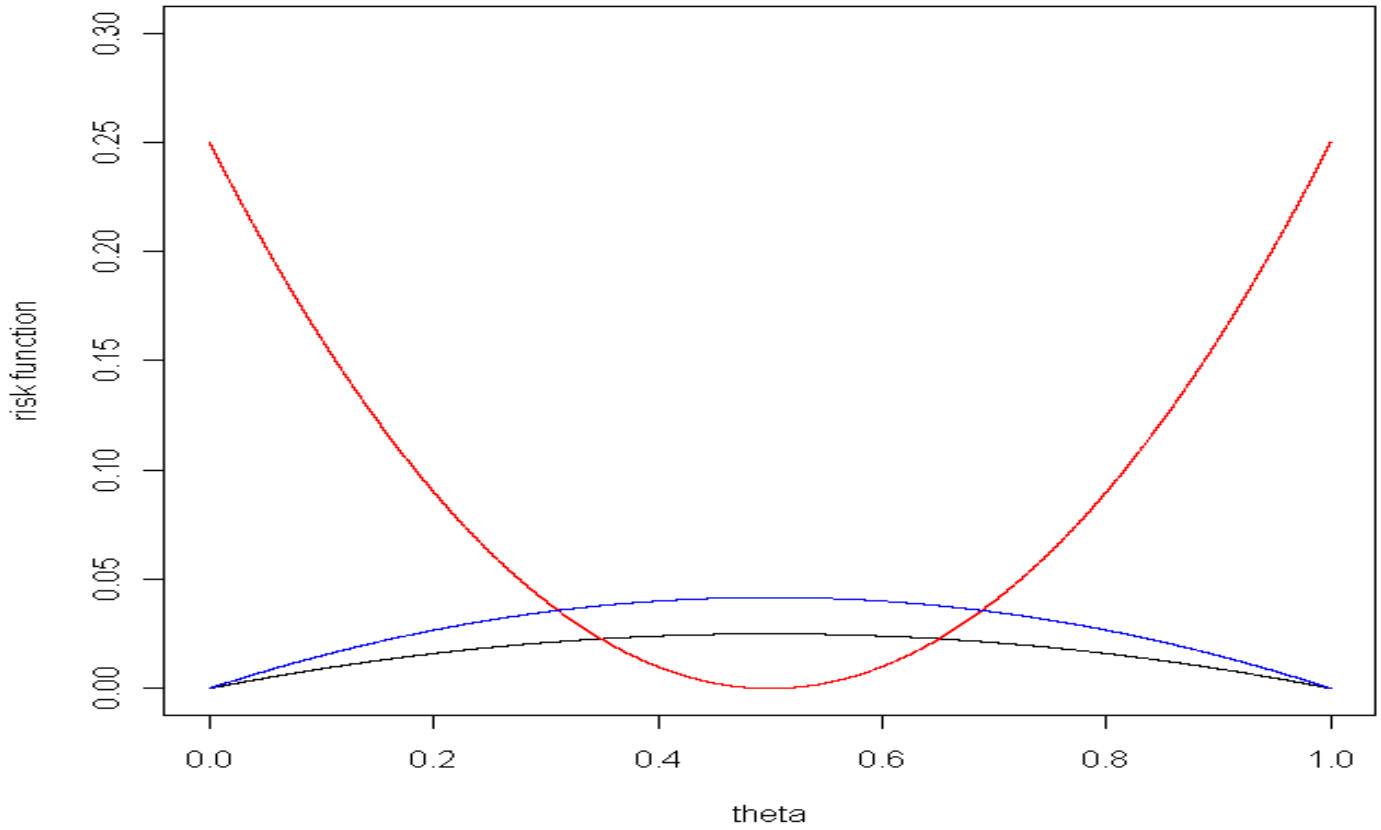
Example

In the problem of independently flipping a coin 10 times with θ = prob of getting a heads each time , suppose that we estimate θ under the loss function $L(\theta, d) = (\theta-d)^2$. Consider the following procedures:

- $\delta_1(X_1,\ldots, X_{10})= (X_1+\ldots+X_{10})/10$
- $\delta_4(X_1,\ldots, X_{10})= ½$
- $\delta_5(X_1,\ldots, X_{10})= (X_1+\ldots+X_6)/6$

$$R_{\delta_1}(\theta) = \frac{\theta(1-\theta)}{10}, \quad R_{\delta_4}(\theta) = (\frac{1}{2}-\theta)^2, \quad R_{\delta_5}(\theta) = \frac{\theta(1-\theta)}{6}$$

Plot of three risk functions



The smaller the risk, the better is the performance!

More about risk function

- In any given experiment, once a decision has been made, there is no question of chance: the decision is good or bad, and a definite loss has been incurred.
- The risk tells us the expected loss in advance of performing the change experiments.

A uniformly best procedure?

- If the were a procedure δ^* whose risk function were, for all theta, no greater tan that of any other procedure (that is, $R_{\delta*}(\theta) \leq R_{\delta}(\theta)$ for all θ and all procedure δ), that would clearly be the procedure to use.
- Unfortunately, there are no practical statistical problems for which such a "uniformly best" (or "an optimal") procedure δ^* exists.
- As an example, in the previous example, for one true θ (e.g., =1/2), one procedure (δ_4) is best, whereas for another true θ (e.g., =1), another procedure (δ_1) is better.

What to do in practice?

- Of course, we do not know the value of θ which governs our experiment, so we cannot know whether we would be better off using δ_1 or δ_4 with experiment at hand!
- How do we choose between δ_1 and δ_4? This will require quite a bit of additional discussion.

However, one thing is clear already: we would certainly never use $\delta_5 = (X_1+\ldots+X_6)/6$, since $R_{\delta_1}(\theta) \leq R_{\delta_5}(\theta)$ for all θ. In other words, among $\delta_1, \delta_4, \delta_5$, we could rule out δ_5 from consideration.

Definitions

a) A procedure δ' is **as good as** another procedure δ'' in a given statistical problem if
$$R_{\delta'}(\theta) \leq R_{\delta''}(\theta) \quad \text{for all } \theta;$$

b) and is **better than** δ'' if the above strict inequality holds for some θ.

c) If $R_{\delta'}(\theta) = R_{\delta''}(\theta)$ for all θ, we say that δ' and δ'' are **equivalent**. (It is perfectly possible for two different procedures to have identical risk functions)

d) If neither of two procedure is better than the other and they are not equivalent, we say that they are **incomparable**.

Definitions (cont.)

e) If, for a given procedure δ, there is another procedure δ' which is better than it, we say that δ is **inadmissible**. Otherwise we say that δ is **admissible**.

- In general, there is no need to consider inadmissible procedures.
- If two procedures are equivalent, either one can be used.

Prove admissible!

Example: In the problem of independently flipping a coin 10 times with θ = prob of getting a heads each time, suppose that we estimate θ under the loss function $L(\theta, d) = (\theta - d)^2$.

Prove that $\delta_4(X_1, \ldots, X_{10}) = 1/2$ is admissible.

Statistical Inference

- One general aim in statistical inference is to make a list of the admissible procedures for each statistical problem;
- But this will not usually be our chief concern, as practical users want a single procedure to use to make a decision.
- Thus we will be more concerned with the selection of the procedure to be used, than with obtaining a list of all admissible procedures.

How to select a procedure?

How are we to select one procedure put of the class of all admissible procedures?

- **Mathematicians and statisticians agreed to only use admissible procedures, but many disagree which admissible procedure should be used.**
- **Need some extra-math criterion!**
- **Sometimes statisticians restrict considerations to procedures in some specified class D^* rather than D= all procedures.**
  "δ^* is admissible relative to D^* "
  "δ^* is D^*-optimal" (uniformly better than any other procedures in D^*)

Admissible in some specified class $D^* \subset D$

- "δ^* is admissible relative to D^*"=
  "δ^* is in D^* and no δ' in D^* is better than δ^*"
- Therefore,
  - if δ^* is admissible relative to D^*, there may not may not exist a δ'' in D (all procedures) that is better than δ^*
  - If δ^* is inadmissible relative to D^*, then δ^* is also inadmissible in D.

Additional Remarks

Additional Remarks on risk function

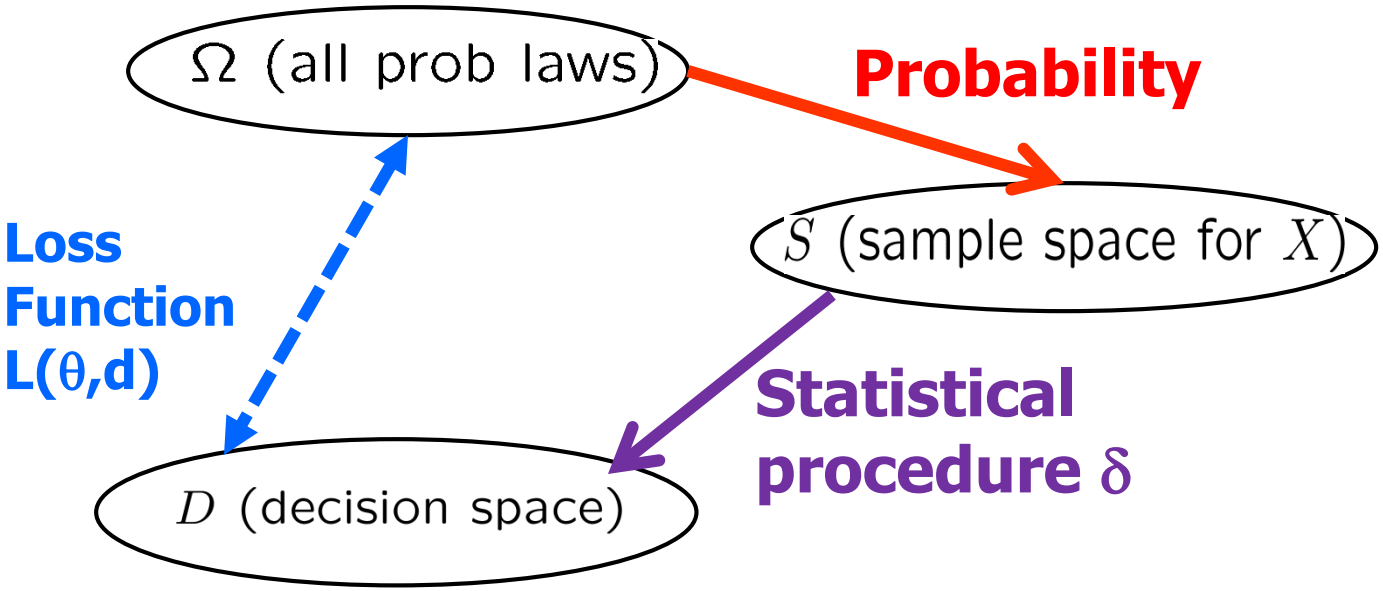
- **Scale of loss function L(θ, d):**

$$L^*(\theta, d) = 100L(\theta, a) \quad \text{or} \quad L^*(\theta, d) = L(\theta, d) + C$$

- **Simple loss function: L(θ, d) can only take on the values 0 and 1, depending on whether a decision d is "correct" or "incorrect" for this θ. In this case,**

$$\begin{aligned} R_\delta(\theta) &= E_\theta L(\theta, \delta(X)) = P_\theta(L(\theta, \delta(X)) = 1) \\ &= P_\theta(\delta \text{ reaches an incorrect detecion}). \end{aligned}$$

Summary: Statistical problems



- **Risk function:** $R_\delta(\theta) = R(\theta, \delta) = \mathbf{E}_\theta(L(\theta, \delta(X))$
- **Admissible / inadmissible.**

Not easy to prove admissible!

Example: In the problem of independently flipping a coin n times with θ = prob of getting a heads each time. Suppose that we estimate θ under the loss function $L(\theta, d) = (\theta - d)^2$.

- **Is the procedure $\delta(X_1, \ldots, X_n) = 0$ admissible?**

  Yes, we will prove it in class for $n = 1$.

ISyE 6412: Thereotical Statistics

A. Classification of Statistical Problems

B. Some criteria for Choosing a procedure

(These two topics are intended as an informal outline of some topics in statistics and it is OK if you may not understand all aspects)

A. Classification of Statistical Problems

- **Statistical Problems have been and are classified on any of several bases, as for example the structure of Ω or of D.**

- **Accordingly, we give several separate classification**

I. The structure of Ω

a) **Parametric model**: the class of all df's F in Ω can be represented in terms of a vector θ consisting of a finite number of real components in a natural way.

b) **Non-parametric model (distribution-free)**: the model structure is not specified a priori, but is instead determined from data

c) **Semi-parametric model**: it has both a finite dimensional component and an infinite dimensional component.

Example

Assume $X_1, \cdots, X_n$ are iid with unknown mean μ and known variance 1.

a) **Parametric model:** when the X_n's have a normal df, or an exponential df, or an uniform.

b) **Non-parametric model:** when Ω=all one-dim df's with mean between a and b and with finite variances (often use ranks of data)

c) **Semi-parametric model:** Cox model in survival analysis: The df is of the form

$$F(t) = 1 - \exp(-\int_0^t \lambda_0(u) e^{\beta' \mathbf{X}(u)} du)$$

where X(u) = known function of covariate and parameter $\theta = (\beta, \lambda_0(u))$ unknown.

II. The Structure of D

a) Point estimation
b) Interval estimation and region estimation (confidence interval/regions)
c) Testing hypothesis
d) Regression problems (overlapped with Design Of experiments)
e) Multiple decision problems (is the coin fair, biased towards heads or tails?)
f) Ranking problems (three items, A, B , C, which one has the largest mean?)
g) Other topics: one wants to find the best item together with a numerical guess of how much better it is than the next best one?

III. Other Topics

a) Sampling Methods
   i. Fixed sample size procedures
   ii. Two-stage procedures (Stein, 40s)
   iii. Sequential procedures (Wald, 50s) ---Clinical trials (offline versus online)

b) Cost considerations (experimental & computation)

c) Design of experiments

d) Prediction Problems: Having watched GT football practices in the summer, will they win ACC?

e) Mathematical Ideas of Importance:
   i. Sufficiency
   ii. Asymptotic theory
   iii. Randomization

B. Some Criteria for choosing a procedure

- Having specified the statistical problem or model associated with an experiment with which we are concerned, we must select a single procedure to be used with the data from that experiment

- For most statistical problems, there are no "uniformly best" procedures, some additional criterion is (are) necessary.

  The purpose of such a criterion is to specified uniquely a statistical procedure which will be used.

Some Criteria

1. Bayes criterion
2. Minimax criterion
3. Sufficiency criterion
4. Unbiased criterion
5. Maximum Likelihood & Likelihood Ratio
6. Invariance criterion
7. Robustness criterion
8. Method of Moments
9. Criteria of Asymptotic Efficiency
10. Computational tractability/Efficiency

Some Criteria

	Often Unique Solution?
1. Bayes criterion	Y
2. Minimax criterion	N
3. Sufficiency criterion	N
4. Unbiased criterion	N
5. Maximum Likelihood & Likelihood Ratio	Y
6. Invariance criterion	N
7. Robustness criterion	N
8. Method of Moments	Y
9. Criteria of Asymptotic Efficiency	N
10. Computational tractability/Efficiency	N

- It is possible that no procedures are admissible (#4,5,6)