Statistical Decision Theory and Bayesian Analysis

CH 1 BASIC CONCEPTS

§1.1 Introduction

Example 1: A drugmaker cares about $\begin{cases} \theta_1: \text{ \% of people for whom the drug is effective} \\ \theta_2: \text{ \% of market.} \end{cases}$

Classical statistics use only sample information.

Non-sample information

- Loss (Statistitians are pessimistic)
- Prior

E.g. A lady claims to tell whether tea or milk is poured into the cup first

A music expert claims to distinguish Haydn score from Mozart score

A drunker claims to predict the outcome of a fair coin.

In 10 trials, they are all correct.

Frequentist's hypothesis test:

H_0: $\theta = 0.5$ (The person is guessing)

H_0 is rejected with one tailed significance level of 2^{-10}

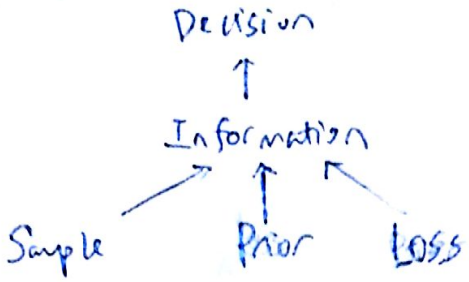
N.B. $\text{p-value}(\mathcal{D}) \triangleq P(f(\hat{\mathcal{D}}) \geq f(\mathcal{D}) \mid \hat{\mathcal{D}} \sim H_0)$

(See Machine Learning: A Probabilistic Perspective) P.213

f: test statistics

here: f is the summarization of the outcome

SDT/BA- CH1-2



Frequentist → Bayesian → Waldish

§1.2 Basic Elements

Definitions.

θ: state of nature / parameter that affects the decision process

Θ: parameter space

a: action (decision) ← Think of an estimator $\hat{\theta}$

$\mathcal{A}$: all actions

Loss function: $L(\theta, a): \Theta \times \mathcal{A} \to \mathbb{R}$

A particular action a_0 with true state of nature θ_0 results in loss $L(a_0, \theta_0)$.

Samples $X = (X_1, \cdots, X_n)$ X_i iid.

Sample space $\mathcal{X}$: possible outcomes of X_i

Reminders. A is not related to $\mathcal{A}$

- Let A be a set of events. $A \subseteq \mathcal{X}$

$$P_\theta(A) = \int_A f(x|\theta)\, dx$$

or

$$P_\theta(A) = \sum_{x \in A} f(x|\theta).$$

- $E_\theta[h(X)] = \int_{\mathcal{X}} h(x) \cdot f(x|\theta)\, dx$

or $E_\theta[h(X)] = \sum_{x \in \mathcal{X}} h(x) \cdot f(x|\theta).$

§1.3

Loss, Risk, etc

Definition. If $\pi^*(\theta)$ is the believed probability distribution of θ at the time of decision making, the Bayesian expected loss of an action a is

$$\rho(\pi^*, a) = E^{\pi^*} L(\theta, a) = \int_{\Theta} L(\theta, a)\, dF^{\pi^*}(\theta)$$

(a: your decision)

Definition A (nonrandomized) decision rule $\delta(x)$ is a function from $\mathcal{X}$ into $\mathcal{A}$. If the data sample is a, then the action is $\delta(x)$

(δ: your decision (function))

Definition The frequentist's risk function of a decision rule $\delta(x)$ is

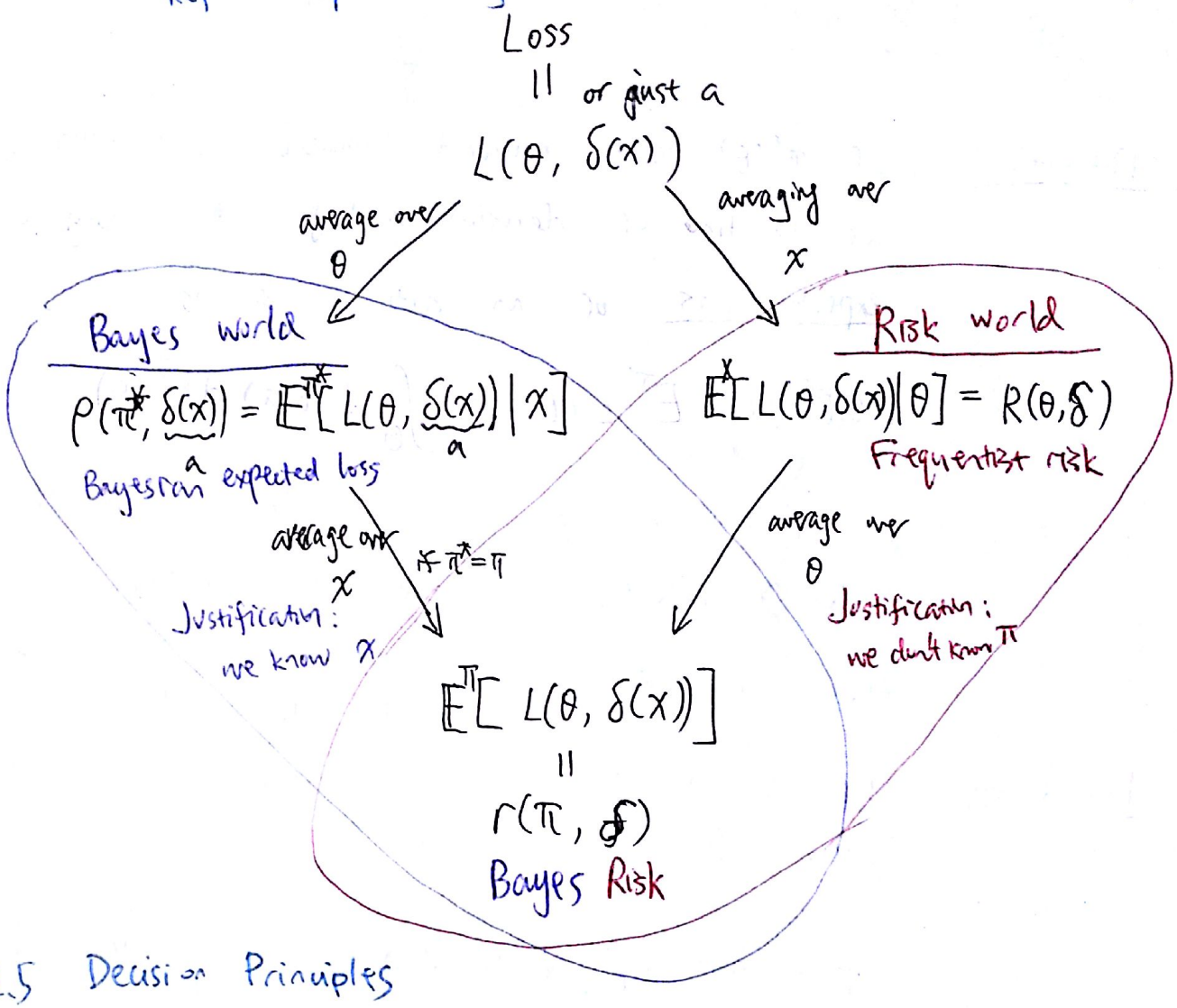
$$R(\theta, \delta) = E_\theta^X [L(\theta, \delta(X))] = \int_{\mathcal{X}} L(\theta, \delta(x))\, d F^X(x|\theta)$$

For no-data problem, $R(\theta, \delta) \equiv L(\theta, \delta)$

Definition The Bayes risk of a decision rule δ, with respect to a prior distribution π on Θ is

$$r(\pi, \delta) = E^{\pi} [R(\theta, \delta)]$$

Ref: http://www.youtube.com/user/mathematicalmonk

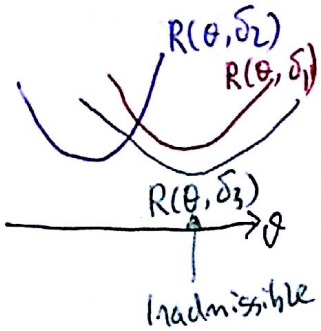


§ 1.5 Decision Principles

Which decision to use?

Bayesian
- 1° x known — Bayesian expected loss
- 2° x unknown (or no-data) — Bayes risk

Frequentist. Additional criterion needed.
- Admissibility
- Minmax
- Invariance
- (Resort to π). Bayes Risk.

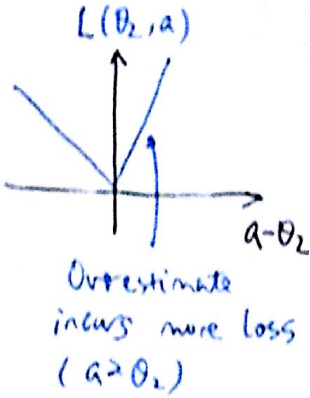


Example 1. Drugmaker wants to estimate the proportion of the market θ_2. Obviously, $\Theta_2 = \{\theta_2 : 0 \le \theta_2 \le 1\} = [0, 1]$

An action is an estimate of θ_2. $\mathcal{A} = [0, 1]$

The company determines that the loss function is

$$L(\theta_2, a) = \begin{cases} \theta_2 - a & \text{if } \theta_2 - a \ge 0 \\ 2(a - \theta_2) & \text{if } \theta_2 - a \le 0 \end{cases}$$



Experiment: a survey among n people, among whom X people would like to buy drug. ($X \sim \mathcal{B}(n, \theta_2)$ ← Binomial)

$$f(x|\theta_2) = \binom{n}{x} \theta_2^x (1 - \theta_2)^{n-x}$$

Prior: From previous experience, the drugmaker believes

$$\theta_2 \sim U[0.1, 0.2]$$

Thus,

$$\pi(\theta_2) = 10\, I_{(0.1, 0.2)}(\theta_2)$$

Bayesian expected loss Assume no experiment is conducted

i.e. $\pi^*(\theta_2) = \pi(\theta_2) = 10\, I_{(0.1, 0.2)}(\theta_2)$

$$\rho(\pi, a) = \int_0^1 L(\theta_2, a)\, \pi(\theta_2)\, d\theta_2$$

(a is what you decide to do when making decisions)

$$= \int_0^a 2(a - \theta_2) \cdot 10 \cdot I_{(0.1, 0.2)}(\theta_2)\, d\theta_2 + \int_a^1 (\theta_2 - a) \cdot 10\, I_{(0.1, 0.2)}(\theta_2)\, d\theta_2$$

$$= \begin{cases} 0.15 - a & \text{if } a \le 0.1 \\ 15a^2 - 4a + 0.3 & \text{if } 0.1 \le a \le 0.2 \\ 2a - 13 & \text{if } a \ge 0.2 \end{cases}$$

Conditional Bayes Principle

To minimize $\rho(\pi, a)$, $a^* = \frac{2}{15}$ and $\min \rho(\pi, a) = \frac{1}{30}$

Example 3. An investor do must decide whether to buy a rather risky ZZZ bonds.

Loss matrix:

	a_1 invest	a_2 not invest	
$\theta = p(\text{win})$	-500	-300	0.9
$p(\text{lose}) = 1-\theta$	1000	-300	0.1
	1000	-300	

Prior: The investor determines according to his/her experince that

$$\pi(\theta) = 0.9 \qquad \text{i.e. } p(\text{win}) = 0.9 \quad p(\text{lose}) = 0.1$$

Bayesian expected loss

$$\rho(\pi, a_1) = \mathbb{E}^{\pi}[L(\theta, a_1)] = L(\theta_1, a_1)\cdot\pi(\theta_1) + L(\theta_2, a_2)\cdot\pi(\theta_2)$$
$$= 0.9 \cdot (-500) + 0.1 \cdot 1000$$
$$= -350$$
$$\rho(\pi, a_2) = \mathbb{E}^{\pi}[L(\theta, a_2)] = L(\theta_1, a_2)\pi(\theta_2) + L(\theta_2, a_2)\pi(\theta_2)$$
$$= -300 \cdot 0.9 - 300 \cdot 0.1$$
$$= -300$$

Conditional Bayesian Decision Principle

To minimize $\rho(\pi, a)$, with respect to a.
we choose $a = a_1$, $\min_a \rho = -350$

Minimax Principle

$$\sup_\theta L(\theta, a_1) = \max\{-500, 1000\} = 1000$$
$$\sup_\theta L(\theta, a_2) = \max\{-300, -300\} = -300$$

To minimize $\sup_\theta L(\theta, a)$, we choose $a = a_2$

$$\inf_a \sup_\theta L(\theta, a) = -300.$$

Example 4: $X \sim N(\theta, 1)$

action: to estimate θ.

Assume $L(\theta, a) = (\theta - a)^2$

Consider the decision rule

$$\delta_c(x) = cx$$

Frequentist risk

$$R(\theta, \delta_c) = \mathbb{E}_\theta^X [L(\theta, \delta_c(X))]$$

$$= \mathbb{E}_\theta^X [(\theta - cX)^2]$$

$$= \mathbb{E}_\theta^X [(c(\theta - X) + (1-c)\theta)^2]$$

$$= \mathbb{E}_\theta^X [c^2(\theta - X)^2] + 2 \cdot \mathbb{E}_\theta^X [c(1-c)\cdot\theta\cdot(\theta - X)] + \mathbb{E}_\theta^X [(1-c)^2\theta^2]$$

($(\theta - X)^2 = 1$; $(\theta - X) = \theta - \theta = 0$)

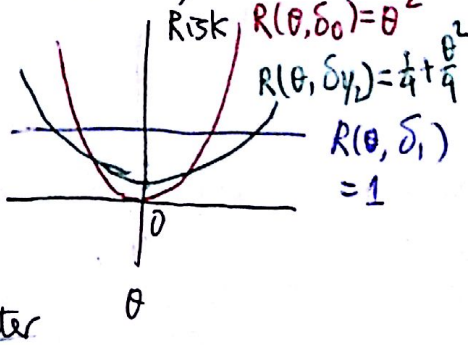
$$= c^2 + (1-c)^2\theta^2$$

For $c > 1$

$$R(\theta, \delta_c) > c^2 + (1-c)^2\theta^2 > c^2 = R(\theta, \delta_1)$$

Thus δ_c $(c > 1)$ inadmissible

For $0 \le c \le 1$: noncomparable



Definition A decision rule δ_1 is R-better than a decision rule δ_2 if $R(\theta, \delta_1) \le R(\theta, \delta_2), \forall \theta \in \Theta$

Definition A decision rule is admissible if there exists no R-better decision rule. A decision rule is inadmissible if there does exist an R-better decision rule.

Bayes risk Suppose $\pi(\theta) = N(\theta|\ 0, \tau^2)$

$$r(\pi, \delta_c) = E^{\pi}[R(\theta, \delta_c)]$$

$$= E^{\pi}[c^2 + (1-c)^2\theta^2]$$

$$= c^2 + (1-c)^2\tau^2$$

The Bayes Risk Principle A decision rule δ_1 is preferred to a rule δ_2 if

$$r(\pi, \delta_1) < r(\pi, \delta_2)$$

A decision rule minimizes $r(\pi, \delta)$ is optimal; it is called a Bayes rule, denoted at δ^{π}.

$r(\pi) = r(\pi, \delta^{\pi})$ is called Bayes risk for π

In the aforementioned example

~~$r(\pi) = r(\pi$~~,

$$r(\pi, \delta_c) = c^2 + (1-c)^2\tau^2$$

$$\frac{\partial}{\partial c} r(\pi, \delta_c) = 2c + 2(1-c)\cdot(-1)\tau^2 \triangleq 0$$

$$\Rightarrow \quad c^* = \frac{\tau^2}{1+\tau^2}$$

Thus $$r(\pi) = r(\pi, \delta_{c^*}) = \left(\frac{\tau^2}{1+\tau^2}\right)^2 + \left(\frac{1}{1+\tau^2}\right)^2 \cdot \tau^2 = \frac{\tau^2}{1+\tau^2}$$

§ 1.6. Foundations

§ 1.6.1. Misuse of Classical Inference Procedures

- p-value is harmful!

My informal note on the risk function of frequentist hypothesis test:

$R(\theta, \delta) = \begin{cases} \alpha \text{ if } \theta \in \\ \text{type II err. if } \theta \in \end{cases}$

Example: $X_1 \cdots X_n \overset{iid}{\sim} N(\theta, 1)$

Conduct a size $\alpha = 0.05$ test

The solution is not unique

$H_0: \theta = 0 \iff H_1: \theta \neq 0$

The classical test: reject H_0 if $\sqrt{n}\,|\bar{x}| > 1.96$

What about $n = 10^{24}$?

Another Example: Kepler's data

Example: $X \sim N(\theta, 9)$.

$H_0: \theta \leq 0 \iff H_1: \theta > 0.$

A sample of 9 observation resulting in $\bar{x} = 1.$ has $\alpha = 0.16$.

Moderately convincing.

Collecting more data ?

§ 1.6.2. The Frequentist Perspective (Frequentists' Justification)

The original idea was from J. Neyman and E. Pearson.

A frequentists intends to

1. find out a method independent of θ or its prior knowledge, i.e., get a value $\bar{R}$ by $\delta(x)$ and $L(\theta, \delta(x))$, which satisfies that
2. the expectation of L is $\bar{R}$ while Repeating $\delta(x)$.

Example. For dealing with standard univariate normal theory problems consider the usual 95% confidence rule for the unknown mean θ.

$$\delta(x) = (\bar{x} - ts,\ \bar{x} + ts)$$

where $\bar{x}$ is the sample mean and s is the sample std.
t is the appropriate percentile from relevant t distribution.

Suppose we measure the performance of δ by loss

$$L(\theta, \delta(x)) = 1 - I_{\delta(x)}(\theta) = \begin{cases} 0, & \text{if } \theta \in \delta(x) \\ 1. & \text{if } \theta \notin \delta(x) \end{cases}$$

Then the frequentist risk is,

$$R((\theta,\sigma),\delta) = \mathbb{E}^{X}_{\theta,\sigma}[L(\theta,\delta(X)]$$

$$= P_{\theta,\sigma}\{\delta(X) \text{ does not contain } \theta\}$$

$$= 0.05 \qquad (\forall\, \theta, \sigma).$$

The universal measure of performance is, if we have different problems θ_i, σ_i. with data $X^{(i)}$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} L(\theta_i, \delta(X^{(i)})) = 0.05 \equiv \bar{R}$$

GREAT!

However, in general $R(\theta,\delta)$ 1° has not eliminate dependence of θ, 2° is repeated over X rather than use of δ.

Bound of R for certain θ is NOT satisfactory.

Example: Consider a simple null hypothesis test

$$H_0: \theta = \theta_0 \quad \longleftrightarrow \quad H_1: \theta = \theta_1$$

Let us define loss to be "0-1" loss

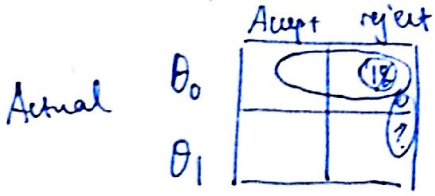
Then the risk function

$$R(\theta_0, \delta) = \alpha_0 \quad \text{(Type I error, FP)}$$

$$R(\theta_1, \delta) = \alpha_1 \quad \text{(Type II error, FN)}$$

We claim only 1% of true null hypotheses will be rejected.

However, we can say nothing ~~that the null hypothe~~ about how often one errs when rejecting.



Let $\alpha_1 = .99$ and $p(\theta_0) = p(\theta_1) = 0.5$

Then $P(\text{Error} \mid \text{Reject}) = 0.5$

⇓

"real" error, because we do not know θ

§ 1.6.3 The Conditional Perspective.

Example. Let $X = (X_1, X_2)$. X_i iid.

$$P_\theta(X_i = \theta - 1) = P_\theta(X_i = \theta + 1) = \frac{1}{2}$$

Frequentist: $\delta(X) = \begin{cases} \frac{1}{2}(X_1 + X_2) & \text{if } X_1 \neq X_2 \\ X_1 - 1 & \text{if } X_1 = X_2 \end{cases}$

Loss: "0-1" loss

half chance $X_1 = X_2$

$$R(\theta, \delta) = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

half chance $X_1 \neq X_2$

half chance $\delta(X) = \theta$ if $X_1 = X_2$

Nonsensible to report $\frac{1}{4}$ risk even though X is provided.

Conditionalist : if $x_1 \neq x_2$ $\quad a = \frac{1}{2}(x_1 + x_2)$

$$\rho(\pi^*, a) = 0$$

if $x_1 = x_2$ $\quad \pi^*(\theta = x_1 | \cdot) = \frac{1}{2}$

$\pi^*(\theta = x_1 + 1 | \cdot) = \frac{1}{2}$

Take action $a = x_1 - 1$

$$\rho(\pi^*, a) = 0.5$$

A Conditionalist knows whether one is in the 100% or 50% case.

Example.

	reject H_0		accept H_0
	1	2	3
$f(x\|\theta=0)$	0.005	0.005	0.99
$f(x\|\theta=1)$	0.00051	0.5849	0.01

Let $\theta = 0, 1$.
$x = 1, 2$ or 3
$f(x|\theta)$ as left.

The classical most powerful test of $H_0: \theta = 0 \Leftrightarrow H_1: \theta = 1$.
at level $\alpha = 0.01$:

Reject H_0 (conclude H_1) if $x = 1$ or 2
Accept H_0 if $x = 3$

Type I error $\alpha_0 = 0.01$

Type II error $\alpha_1 = 0.01$

Standard frequentist: upon observing $x=1$, reports the decision is H_1 with error probability

Conditionalist: when observing $x=1$, likelihood ration

$$\frac{f(x|\theta=0)}{f(x|\theta=1)} = \frac{0.005}{0.00051} \approx 1$$

Example. Cox 1958.

Suppose a substance to be analyzed can be sent either to a laboratory in New York or a laboratory in California. The two labs seem equally good, so a fair coin is flipped to choose between them, with "heads" denoting that the lab in New York will be chosen. The coin is flipped and comes up tails, so the California lab is used. After a while, the experimental result come back and a conclusion and report must be developed. Should this conclusion take into account the fact that the coin could have been heads, and hence the experiment in New York might

have been performed instead? Common sense (and the conditional viewpoint) vies to, that only the experiment actually performed is relevant, but frequentist reasoning would call for averaging over all possible data, even the possible New York data.

§ 1.6.4 The Likelihood Principle

Definition For observed data, x, the function $l(\theta) = f(x|\theta)$, considered as a function of θ, is called the **likelihood function**.

The Likelihood Principle In making inferences or decisions about θ after x is observed, all relevant experimental information is contained in the likelihood function for observed x. Furthermore two likelihood functions contain the same information about θ if they are proportional to each other (as functions of θ)

Example: p(Head) of a coin is θ

9 heads and 3 tails

Let X = # heads

How is the experiment conducted?

case #1: The experiment consisted of a predetermined 12 flips

$$l_1(\theta) = f_1(x|\theta) = \binom{n}{x}\theta^{x=9}(1-\theta)^{n-x}_{12} = 220\cdot\theta^9(1-\theta)^3$$

Notes

1° By "proportion", we impose constraint on the same parameter.

2° Likihood principle says about experimental information

case #2: The experiment consisted of flipping the coin until 3 tails were observed

$$l_2(\theta) = f_2(x|\theta) = \binom{n+x-1}{x}\theta^{x=9}(1-\theta)^{n=3} = 55\cdot\theta^9(1-\theta)^3$$

Classical hypothesis test $H_0: \theta = \frac{1}{2} \leftrightarrow H_1: \theta > \frac{1}{2}$

$$\alpha_1 = P_{1/2}(X \geq 9) = f_1(9|\theta=\tfrac{1}{2}) + f_1(10|\theta=\tfrac{1}{2}) + f_1(11|\theta=\tfrac{1}{2}) + f_1(12|\theta=\tfrac{1}{2})$$
$$= 0.075$$

$$\alpha_2 = P_{1/2}(X \geq 9) = f_2(9|\theta=\tfrac{1}{2}) + f_2(10|\theta=\tfrac{1}{2}) + \cdots$$
$$= 0.0325$$

Example. Let $\mathcal{X} = \{1,2,3\}$. $\Theta = \{0,1\}$. Two experiments E_1, E_2 with observing data X_1 and X_2, respectively.

X_1

	1	2	3
$f_1(x_1\|0)$	0.90	0.05	0.05
$f_1(x_1\|1)$	0.09	0.055	0.855

X_2

	1	2	3
$f_2(x_2\|0)$	0.26	0.73	0.01
$f_2(x_2\|1)$	0.026	0.803	0.171

Likelihood Priciple says E_1 and E_2 are equivalent given x because $l_1(\theta) \propto l_2(\theta)$

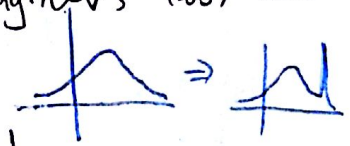
Frequentists: accept H_0 if $X_1 = 1$ (or $X_2 = 1$)
reject H_0 if $X_1 = 2, 3$ (or $X_2 = 2, 3$)

Type I error Experiment I: 0.1 Experiment II: 0.74

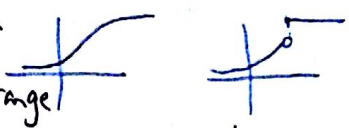
Type II error Experiment I: 0.09 Experiment II: 0.026

Example:

- Engineer draws a random sample of electron tubes and measure the plate voltages under certain conditions with a very accurate voltmeter
- A statistician examines the samples which seem to be Gaussian distributed, varying from 75 to 99, mean = 87, std = 4
- Later the statistician visited the engineer's lab, and found the voltmeter reads 100 at most



Hence a new analysis is needed



- The engineer says he has a high-range voltmeter and he would have use it if the voltage > 100
- The statistician is happy
- The engineer says accuddly the high-range voltmeter is broken
- The statistician says a new analysis is needed.

The Weak Conditionality Principle

Suppose one can perform either of two experiments E_1 or E_2, both pertaining to θ, and that the actual experiment conducted is the mixed experiment of first choosing $J=1$ or 2 with probability $\frac{1}{2}$ each (independent of θ), and then performing experiment E_J. Then the actual information about θ obtained from the overall mixed experiment should depend only on the experiment E_J that is actually performed.

Sufficiency + Weak Conditionality Principle $\Rightarrow$ Likelihood Principle

Frequentists and Conditionalists agree on sufficiency. Then frequentists are in trouble.

§ 1.7 Sufficient Statistics.

Definition

Let X be a random variable whose distribution depends on the unknown parameter θ, but is otherwise unknown. A function T of X is said to be sufficient statistics for θ if the $f(X|T=t)$ is independent of θ.

Fisher's factorization Theorem.

$T(X)$ is sufficient for $\theta \Leftrightarrow f_\theta(x) = h(x)\, g_\theta[T(x)]$.

Definition Partition induced by T

$$\mathcal{X}_t = \{x \in \mathcal{X}: T(x)=t\}$$

Argue: a sufficient statistics contain all information concerning θ from samples.

Give $T(X)=t$, denote $f_t(x)$ as a distribution over $\mathcal{X}$.

The density $f_t(x)$ does not depend on θ because $f(X|T=t) \perp\!\!\!\perp \theta$

$\Rightarrow$ $f_t(x)$ is explicitly calculable from $f_\theta(x|\theta)$ for whatever θ?

Random variable X can be thought of as

1° first arising from the random generation of T.

2° followed by the random choice of x from $\mathcal{X}_t$.

We would now argue sufficiency is sensible from a frequentist's perspective.

Theorem: Assume that T is a sufficient statistic for θ, let $\delta_0^*(X,\cdot)$ be any randomized rule in $\mathcal{D}^*$. Then (subject to measurability conditions) there exists a randomized rule $\delta_1^*(t,\cdot)$ depending only on $T(x)$, which is R-equivalent to δ_0^*

($\delta^*: \mathcal{X} \times \mathcal{A} \to [0,1]$

$\delta^*(x,a)$ is the probability that we take action a given data x.)

Proof: $\forall\ A \subseteq \mathcal{A}$ and $t \in \mathcal{T}$, we design δ_1^* base on δ_0^* as

$$\delta_1^*(t,A) = E^{X|t}[\delta_0^*(X,A)]$$

i.e. $\delta_1^*(t,\cdot)$ is formed by averaging δ_0^* over $\mathcal{X}_t$

$$L(\theta, \delta_1^*(t,\cdot)) = \mathbb{E}^{\delta_1^*(t,\cdot)}[L(\theta,a)]$$

$$= \mathbb{E}^{X|t}\,\mathbb{E}^{\delta_0^*(X,\cdot)}[L(\theta,a)]$$

[by substituting the design of δ_1^*, and then changing the order of expectation.]

Therefore

$$R(\theta,\delta_1^*) = \mathbb{E}^T[L(\theta,\delta_1^*(T,\cdot))]$$

$$= \mathbb{E}^T\mathbb{E}^{X|T}\mathbb{E}^{\delta_0^*(X,\cdot)}[L(\theta,a)]$$

$$= \mathbb{E}^X\mathbb{E}^{\delta_0^*(X,\cdot)}[L(\theta,a)] \qquad \mathbb{E}^X[h(X)] = \mathbb{E}^T\mathbb{E}^{X|T}[h(X)]$$

$$= \mathbb{E}^X[L(\theta,\delta_0^*(X,\cdot))]$$

$$= R(\theta,\delta_0^*) \qquad \#$$

The result implies that it is only necessary to consider rules based on a sufficient statistic.

The Likelihood Principle immediately implies sufficiency

$$l(\theta) = f(x|\theta) = h(x)\,g(T(x)|\theta) \propto_\theta g(T(x)|\theta)$$

§ 1.7 Convexity

Theorem: Assume that

1. $\mathcal{A} \subseteq \mathbb{R}^n$ convex
2. $\forall\, \theta \in \Theta$, $L(\theta,a)$ is convex w.r.t a
3. Let δ^* be a randomized in $\mathcal{D}^*$ for which $\mathbb{E}^{\delta^*(x,\cdot)}[|a|] < \infty \;\forall x \in \mathcal{X}$

Then. $\forall\, x$ and θ the nonrandomized decision rule

$$\delta(x) = \mathbb{E}^{\delta^*(x,\cdot)}[a]$$

has $L(\theta,\delta(x)) \le L(\theta,\delta^*(x,\cdot))$

Proof: 1° $\delta(x) \in \mathcal{A}$

2° $L(\theta,\delta(x)) = L(\theta, \mathbb{E}^{\delta^*(x,a)}[a]) \le \mathbb{E}^{\delta^*(x,a)}[L(\theta,a)] = L(\theta,\delta^*(x,a))$ #

Jenson's Inequality: $g(EX) \le E[g(X)]$ if g convex.

Theorem (Rao-Blackwell) Assume

1° $\mathcal{A} \subseteq \mathbb{R}^m$ convex

2° $\forall\, \theta \in \Theta$, $L(\theta, a)$ convex in a

3° Let $T(X)$ be a sufficient statistics, and $\delta_0 \in \mathcal{D}$

Then a nonrandomized decision rule δ_1, defined by

$$\delta_1(t) = E^{X|t}[\delta_0(X)]$$

is R-equivalent to or R-better than δ_0, provided that the expectation exists.

Proof: Jensen's Inequality

$$L(\theta, \delta_1(t)) = L(\theta, E^{X|t}[\delta_0(X)]) \le E^{X|t}[L(\theta, \delta_0(X))]$$

Then,

$$\begin{aligned} R(\theta, \delta_1) &= E^T[L(\theta, \delta_1(T))] \\ &\le E^T E^{X|T} L(\theta, \delta_0(X)) \\ &= E^X[L(\theta, \delta_0^*(X))] \\ &= R(\theta, \delta_0^*) \end{aligned}$$

Another Proof: Theorem p16

Define $\delta_1^*(t, A) = E^{X|t}[I_A(\delta_0^*(X))] \Leftrightarrow \delta_0^*$

Then $\delta_1(t) = E^{\delta_1^*(t,\cdot)}[a] = E^{X|t} E^{\mu_X}[a] = E^{X|t}[\delta_0(X)]$

Let $\mu_X(A) = I_A(\delta^0(X))$

is R-equivalent or R-better than δ^0

Theorem p17 #