# Introduction to Statistics at an Advanced Level

STAT 201B

Shizhe Zhang

Last updated: November 20, 2025

Online Resources:

https://bcourses.berkeley.edu/courses/1548317

https://edstem.org/us/courses/84592/discussion

## Contents

## ※ Lecture 1

### 1.1 Information

Instructor: Dr. Haiyan Huang Tu/Th 11:00am-12:29pm Lecture, 106 Stanley Office: 317 Evans

GSI: Karissa Huang (krhuang@berkeley.edu) W 12:00pm-1:59pm (101 Discussion Section), 334 Evans W 2:00pm-3:59pm (102 Discussion Section), 334 Evans

GSI: Drew Thanh Nguyen (drew.t.nguyen@berkeley.edu) W 4:00pm-5:59pm (103 Discussion Section), 344 Evans Online tools:

1. Bcourses
2. Ed discussion
3. Gradescope

Grade:

1. Homework: 30%

   Problem sets will be assigned roughly each Wednesday, for a total of 9 assignments. You should download the assignments from Bcourses. Each problem set is to be turned in on Friday a week later. No late assignments will be accepted. The homework with lowest score will not be included in the final homework grade. Some problems may not be graded, and you should review the solutions carefully for those problems. Students can discuss homework assignments. Each student must write up his/her own solutions individually. Any evidence of cheating will be subject to disciplinary action.

2. Midterm: 25%

   October 16, A double sided A4 page of handwritten notes is allowed.

3. Final: 45%

   Dec 17 8-11am, Two double sided A4 pages of handwritten notes are allowed.

Office hour: Thursday 1-2pm 317 Evans

### 1.2 Introduction to Inference

Different types of inference:

- Nonparametric
- Parametric: Frequentist; Bayesian

  Treats parameters as unknown fixed constants; Focuses on point estimation, confidence intervals, and hypothesis tests.

  Makes probability statements about parameters, reflecting beliefs. Bases all inference on the posterior distribution, which we can summarize in various ways.

  e.g. Assume $\sigma^2 \sim \chi^2(1)$ and use the data to modify it.

**Parametric models** can be described by a finite number of parameters. Generally we consider a family of distributions that are parameterized by a finite set of parameters. e.g. $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), \qquad i = 1, \dots, n$

Use $\theta$ to indicate an arbitrary parameter. Use $P_\theta(Y \in A)$ to emphasize the $F_Y$'s dependence on $\theta$.

**Nonparametric models** require an infinite number of parameters to describe the distribution. They are called distribution free to indicate that we make few restirctions on the family of distributions.

## 1.3 Point Estimation

A statistic is any function of the data. A point estimator $\hat{\theta}_n$ is a statistic that provides a single value as an estimate of an unknown parameter $\theta$.

We call $\hat{\theta}(X_1, \dots, X_n)$ the **RV** an **estimator**, while we call $\hat{\theta}(x_1, \dots, x_n)$ an **estimate**

Note that

$$\hat{X}_n \sim N(\mu, \frac{\sigma^2}{n})$$

Bias: $bias(\hat{\theta}) = E[\hat{\theta}] - \theta$

Standard error: $se(\hat{\theta}) = \sqrt{Var_\theta(\hat{\theta})}$

Standard deviation for the population $sd(Y) = \sigma$

Mean squared error:

$$MSE(\hat{\theta}_n) = E_n[(\hat{\theta}_n - \theta)^2] = Var_n(\hat{\theta}_n) + bias(\hat{\theta}_n)^2$$

Trick is $E[(\hat{\theta}_n - E(\hat{\theta}_n))(E(\hat{\theta}_n) - \theta)] = 0$

**Definition**

If $\hat{\theta}_n \xrightarrow{p} \theta$, then $\hat{\theta}_n$ is a weakly consistent estimator of $\theta$.

**Example**

For $X_1 \dots, X_n \sim N(\mu, \sigma^2)$, we have

$$\bar{X}_n, \hat{S}_n^2 \xrightarrow{p} \mu, \sigma^2$$

**Definition**

Asymptotic normality:

$$\frac{\hat{\theta}_n - \theta}{\sqrt{Var(\hat{\theta}_n)}} \xrightarrow{d} N(0,1)$$

Note Slutsky's Thm allow us to replace $se$ by some weakly consistent estimator $\hat{\sigma}_n$

# ※ Lecture 2

**Definition** ***Plug-in Estimator***

Let $X_1, ..., X_n \overset{i.i.d.}{\sim} F$, where F can be parametric or nonparametric. Assume that we are interested in estimating the quantities that are related to F , such as the mean, median, variance, quantiles, etc, by a nonparametric way.
No matter F is parametric or non-parametric, we can write the quantities of interest as a function of $F$ , $\theta(F)$. The substitution (plug-in) method is to estimate $\theta(F)$ with $\theta(\hat{F}_n)$, where $\hat{F}_n$ is the empirical distribution of F

Empirical distribution function:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) = \#\{X_i \leq x\}/n$$

$$p = P(Y_i = 1) = P(X_i = x) = F(x)$$

$$\begin{aligned}\mathrm{E}[\hat{F}_n(x)] &= F(x)\\ \mathrm{V}[\hat{F}_n(x)] &= \frac{F(x)[1-F(x)]}{n}\\ \mathrm{MSE}[\hat{F}_n(x)] &= \mathrm{V}[\hat{F}_n(x)] \longrightarrow 0\\ \hat{F}_n(x) &\xrightarrow{P} F(x)\end{aligned}$$

Plug in estimator:

$$\hat{\theta}_{\text{plug-in}}(F) \triangleq E_{\hat{F}_n}(X) = \sum_t t \cdot P_{\hat{F}_n}(X_i = t)$$

$$= \sum_t t \sum_{i=1}^{n} \frac{I(X_i = t)}{n} = \sum_{i=1}^{n}\sum_t t \cdot \frac{I(X_i = t)}{n} = \bar{X}_n$$

Now we are interested in $\theta(F) = Var_F(X)$

One possible estimator of $\theta(F)$ is $\hat{\theta}(F) = \theta(\hat{F}_n)$

$$\begin{aligned}\theta(\hat{F}_n) &= \mathrm{var}_{\hat{F}_n}(X) = E_{\hat{F}_n}(X^2) - \left(E_{\hat{F}_n}(X)\right)^2\\ &= \frac{\sum_{i=1}^{n} X_i^2}{n} - \left(\frac{\sum_{i=1}^{n} X_i}{n}\right)^2\end{aligned}$$

$$= \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

This is biased but consistent.

**Theorem** ***Glivenko-Cantelli Theorem***

$$\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{a.s.} 0$$

**Theorem**

Suppose the function $\theta(F)$ is continuous in the sup-norm:

$$\forall \epsilon > 0, \exists \delta > 0 \text{ such that "} \|G - F\|_\infty < \delta \text{ implies } |\theta(G) - \theta(F)| < \epsilon\text{".}$$

[That is for any $\epsilon$, if there is some $G$ close enough to $F$, then $\theta(G)$ is close to $\theta(F)$.]

Then,

$$\theta(\hat{F}_n) \xrightarrow{P} \theta(F).$$

**Definition** ***Linear statistics***

A statistic is a linear function of $F$ if it can be written as

$$T(F) = \int r(x)\, dF(x)$$

for some measurable function $r(x)$.

The mean is a linear functional, but the variance and quantile function are not.

The plug-in estimator of $T(F)$ is just $T(\hat{F}_n)$. When $T$ is a linear functional,

$$T(\hat{F}_n) = \int r(x) d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} r(X_i)$$

## ※ Lecture 3

**Theorem**

The Dvoretzky-Kiefer-Wolfowitz Inequality states that for i.i.d. random variables $X_1, \ldots, X_n$ with empirical distribution $\hat{F}_n$ and true distribution $F$, the following holds:

$$P(\sup_x |F(x) - \hat{F}_n(x)| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

Let the RHS be $1-\alpha \longrightarrow \epsilon = \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}$

Then we have

$$P(\hat{F}_n(x)-\epsilon \leq F(x) \leq \hat{F}_n(x)+\epsilon, \forall x) \geq 1-\alpha$$

Let $L(x) = \max\{\hat{F}_n(x)-\epsilon, 0\}$ and $U(x) = \min\{\hat{F}_n(x)+\epsilon, 1\}$

Then we have $P(L(x) \leq F(x) \leq U(x), \forall x) \geq 1-\alpha$

Often we have $T(\hat{F}_n) \approx N(T(F), \hat{se}^2)$, which allows us to form an approximate $1-\alpha$ confidence interval. We need to find an asymptotic distribution of $T(\hat{F}_n)$.

$\theta(F) = T(F)$ quantity of interest (often a single value instead of function like F)

We will have

$$P(|\frac{T(f)-T(\hat{F}_n)}{\hat{se}}| \leq z_{\alpha/2}) \approx 1-\alpha$$

And we focus on this interval:

$$T(\hat{F}_n) \pm z_{\alpha/2}\hat{se}$$

## 3.1 Bootstrap

### Monte Carlo

$$E(h(Y)) = \int h(y)\,dF_Y(y) \approx \frac{1}{n}\sum_{i=1}^{n} h(Y_i) \text{ where } Y_i \overset{\text{i.i.d.}}{\sim} F_Y$$

Note that if $E[h(Y)] < \infty$, then

$$RHS \overset{a.s.}{\longrightarrow} E[h(Y)] \text{ as } n \longrightarrow \infty$$

**Example**

Approx $\int_{-\infty}^{\infty} sin^2(x)e^{-x^2}\,dx$ using Monte Carlo with $n = 1000$ samples.

$$\sqrt{\pi}\int_{-\infty}^{\infty} sin^2(x)\frac{1}{\sqrt{\pi}}e^{-x^2}\,dx = E[sin^2(X)] \text{ where } X \sim N(0, 1/2)$$

```python
import numpy as np
n = 10000
X = np.random.normal(0, np.sqrt(1/2), n)
np.sqrt(np.pi) * np.mean(np.sin(X)**2)
```

Even though, the target density is $h$. More generally, we can use Monte Carlo for:

$$E_h[q(\theta)] = \int h(\theta)q(\theta)\,d\theta = \int q(\theta)\frac{h(\theta)g(\theta)}{g(\theta)}\,d\theta \approx \frac{1}{n}\sum_{i=1}^{n}\frac{h(\theta_i)q(\theta_i)}{g(\theta_i)} \text{ where } \theta_i \overset{\text{i.i.d.}}{\sim} g(\theta)$$

i.e. we can sample from a different distribution $g$ and use importance weights $\dfrac{q(\theta)}{g(\theta)}$ to adjust.

```
import numpy as np
n = 1000
X = np.random.normal(0, 1, n)
np.mean(X > 3)
# np.float64(0.002)
```

Now try to stimualate using importance sampling:

```
import numpy as np
n = 1000
X = np.random.normal(3, 1, n)
np.mean((X > 3) * np.exp(-X**2/2 + (X-3)**2/2))
# np.float64(0.0014236252168949273)
```

If we knew F , we could use MC integration to approximate Var$F(T_n)$. However, we don't in practice, so we make an initial approximation of F with the empirical CDF $\hat{F}_n$ and then use MC integration to approximate $V_{\hat{F}_n}[T_n]$.

$$V_F[T_n] \overset{ECDF}{\approx} V_{\hat{F}_n} \overset{MC}{\approx} \hat{V}_{\hat{F}_n}$$

## ❊ Lecture 4

We know $F$. The bootstrap procedure to estimate $V_F(T_n)$ is:

At the $j$-th iteration, for $j = 1, \dots, B$:

1. Sample $X_{1,j} \dots X_{n,j} \sim F$
2. Compute $T_{n,j} = g(X_{1,j}, \dots, X_{n,j})$
3. The bootstrap estimate of $V_F(T_n)$ is

$$\hat{V}_{\hat{F}_n} = \frac{1}{B}\sum_{j=1}^{B}(T^*_{n,j} - \bar{T}^*_n)^2, \quad \text{where } \bar{T}^*_n = \frac{1}{B}\sum_{j=1}^{B} T^*_{n,j}$$

## 4.1 Bootstrapping method for estimating bias

$X_1, \ldots, X_n \overset{i.i.d.}{\sim} F_0$. Let $F_1$ be the corresponding empirical distribution. (i.e. $\hat{F_n}$) Then $\theta(F_1)$ is an empirical Plug-in estimate of $\theta(F_0)$. How to estimate

$$t_0 = E_{F_0}(\theta(F_0) - \theta(F_1))$$

Answer: Draw $Y_1, \ldots, Y_n \sim F_1$ and derive the empirical distribution $F_2$ based on $Y_1, \ldots, Y_n$. Then $\theta(F_2)$ is an empirical Plug-in estimate of $\theta(F_1)$.

$$\hat{t}_0 = E_{F_1}(\theta(F_1) - \theta(F_2))$$

Mimicing the $F_0$ with $F_1$.

$$E_{F_1}(Y) = \sum_{i=1}^{n} X_i P(Y = X_i) = \sum_{i=1}^{n} X_i \frac{1}{n} = \bar{X}_n$$

$$Var_{F_1}(Y) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

**Example** ***Why this is good?***

$$T_n = median(X_1, \ldots, X_n)$$

$$C_n = T_n \pm z_{\alpha/2} \sqrt{\hat{V}_{F_1}(T_n)}$$

This only works well if the distribution of $T_n$ is close to Normal. Note that asymptotic normality does not always hold. For example, if $X_i \sim U(0, \theta)$, then $T_n = \max(X_1, \ldots, X_n)$ and the asymptotic distribution relies on $n$ instead of $B$.

**Example *Bias correction***

We want to estimate $\theta(F_0) = (E_{F_0}X)^2 = \mu^2$ where $X \sim F$ with mean $\mu$ and variance $\sigma^2$. The EPI is $\theta(F_1) = (\bar{X}_n)^2$. The bias is

$$t_0 = E_{F_0}(\theta(F_0) - \theta(F_1)) = E_{F_0}(\mu^2 - (\bar{X}_n)^2) = \mu^2 - Var_{F_0}(\bar{X}_n) - [E_{F_0}(\bar{X}_n)]^2 = -Var(X)/n$$

Now we consider

$$\tilde{\theta} = \theta(F_1) + \hat{t_0} = \theta(F_1) + E_{F_1}(\theta(F_1) - \theta(F_2)) = \theta(F_1) + \theta(F_1) - E_{F_1}(\theta(F_2))$$

$Z_1 \ldots Z_k \sim F_2$ and $E_{F_2}(Z) = \bar{Y}_m$ and $Var_{F_2}(Z) = \frac{1}{m}\sum_{i=1}^{m}(Y_i - \bar{Y}_m)^2$
By definition,

$$\theta(F_2) = (E_{F_1}Z)^2 = (\bar{Y})^2 = \bar{Y}_n^2 + Var_{F_1}(\bar{Y}_n) = \frac{1}{m}(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2) + (E_{F_1}(\bar{Y}))^2$$

$$\tilde{\theta} = 2(\bar{X})^2 - [(\bar{X})^2 + \frac{1}{m}(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2)] = (\bar{X})^2 - \frac{1}{mn}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$E_{F_0}(\tilde{\theta}) = Var_{F_0}(\bar{X}) + E_{F_0}(\mu^2 - \frac{1}{mn}\sum_{i=1}^{n}(X_i - \bar{X})^2) = \mu^2 - \frac{m-n+1}{mn}\sigma^2$$

If $m = n$, $E_{F_0}(\tilde{\theta}) = \mu^2 + \frac{1}{n}\sigma^2$ If $m = n-1$, $E_{F_0}(\tilde{\theta}) = \mu^2$ – unbiased!

## 4.2 Parametric Inference

$\mathcal{F} = \{F(x;\theta) : \theta \in \Theta\}$ where $\Theta \subseteq \mathbb{R}^k$ is the parameter space. Choose class of distributions $\mathcal{F}$ based on knowledge of the problem.

- Sufficient statistic: $T(X_1, \ldots, X_n)$ is sufficient for $\theta$ if the conditional distribution of $X_1, \ldots, X_n$ given $T = t$ does not depend on $\theta$. Keep the information about the parameters.

- Likelihood functions summarizes the information about $\theta$ contained in the data. Into a parameter-based function that drives inference.

**Definition** ***Sufficient Statistic***

$X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathcal{P} = P_\theta : \theta \in \Omega$ .
A statistic $T$ is **sufficient** for $\theta$ if , for every $t$ in the range of $\mathcal{T} of\ T$, the conditional distribution of $P_\theta(X|T(X) = t)$ is independent of $\theta$.

# ※ Lecture 5

## 5.1 Sufficiency

**Motivation** We hope to separate the information contained in the data into the information relevant for making inference about $\theta$ and the information irrelevant for these inferences. In other words, we would like to compress the data to, e.g. $T(X)$, without loss of information. (Actually, it often turns out that some part of the data carries no information about the unknown distribution that produces the data)

**Benefits** 1. increasing computational efficiency and decreasing storage requirements 2. involving irrelevant information may increase an estimator's risk (see Rao-Blackwell Theorem) 3. Improving the scientific interpretability of our data

**Example**

Let $X_i \overset{i.i.d.}{\sim} Ber(\theta)$. Show that $T(X) = \sum_{i=1}^{n} X_i$ is sufficient for $\theta$.

$$P_0(X_1 = x_1, \dots, X_n = x_n | T(X) = t) = \frac{P_0(X_1 = x_1, \dots, X_n = x_n, T(X) = t)}{P_0(T(X) = t)}$$

$$= \frac{P_0(X_1 = x_1, \dots, X_n = x_n | \sum_{i=1}^{n} X_i = t)}{P_0(\sum_{i=1}^{n} X_i = t)}$$

$$= \begin{cases} 0 & \text{when } t \neq \sum_{i=1}^{n} x_i \\ \frac{1}{\binom{n}{t}} & \text{when } t = \sum_{i=1}^{n} x_i \end{cases}$$

**Theorem** ***Neyman Factorization Theorem***

Suppose the family $\{P_\theta : \theta \in \Omega\}$ of distributions have joint mass functions or densities $\{p(x;\theta) : \theta \in \Omega\}$. Then a statistic $T$ is sufficient for $\theta$ if and only if there are functions $h$ and $g$ such that the density/mass function can be written

$$p(x;\theta) = h(x)\,g(T(x),\theta).$$

**Proof** $\Rightarrow$

If $T$ is sufficient for $\theta$, then

$$P_\theta(X = x) = \underbrace{P_\theta(X = x|T(X) = T(x))}_{h(x)} \cdot P_\theta(T(X) = T(x))$$

$$= h(x) \cdot g(T(x),\theta)$$

The first term is independent of $\theta$ according to the definition of Sufficient Statistics.

$\Leftarrow$ If $p(x;\theta) = h(x)g(T(x),\theta)$, then

$$P_\theta(X = x|T(X) = t) = \frac{P_\theta(X = x, T(X) = t)}{P_\theta(T(X) = t)} = \frac{h(x)g(T(x),\theta)}{\sum\limits_{y:T(y)=t} h(y)g(T(y),\theta)}.$$

Since $P(X = x, T(X) = T(x)) = P(X = x)$ and we need to run through all $y$ such that $T(y) = t$, the $g(T(y),\theta)$ term cancels out. So the conditional distribution does not depend on $\theta$.

$$= \frac{h(x)}{\sum\limits_{y:T(y)=t} h(y)}$$

According to the definition of Sufficient Statistics, $T$ is sufficient for $\theta$.

**Example**

Let $X_i \sim U(0,\theta)$. Show that $T(X) = \max(X_1,\dots,X_n)$ is sufficient for $\theta$.

$$\begin{aligned} p(x_1,\dots,x_n;\theta) &= \frac{1}{\theta^n}\cdot I(0<x_1,\dots,x_n<\theta) = \frac{1}{\theta^n} I(0<\max(x_1,\dots,x_n)<\theta) \\ &= \frac{1}{\theta^n} I(0<Y_{(1)})\cdot I(Y_{(n)}<\theta) \\ &= I(Y_{(1)}>0)\cdot\frac{1}{\theta^n}\cdot I(Y_{(n)}<\theta) \\ &= h(Y)\cdot g(T(Y),\theta) \end{aligned}$$

$$T(Y) = Y_{(n)}$$

**Theorem** *The Rao-Blackwell Theorem*

Suppose $X$ is distributed according to $P_\theta(x) \in \{P_\theta : \theta \in \Omega\}$ and a statistic $T(X)$ is sufficient for $\theta$. Given any estimator $\delta(X)$ of $\theta$, define

$$\eta(T) = \mathbb{E}_\theta\big[\delta(X) \mid T(X)\big].$$

If the loss function $\mathcal{L}(\theta,\delta(X))$ is convex and the risk function

$$R(\theta,\delta(X)) = \mathbb{E}\big[\mathcal{L}(\theta,\delta(X))\big] < \infty,$$

then

$$R(\theta,\eta) \le R(\theta,\delta).$$

If $\mathcal{L}$ is strictly convex, then the inequality is strict unless $\delta = \eta$.

Note that the loss function reflects the degree of wrongness of an estimate. The commonly used quadratic loss function is defined as

$$\mathcal{L}(\theta,\delta) = \big(\theta - \delta(X)\big)^2.$$

*Proof.* $\delta(x)$: an estimator of $\theta$.

$\eta(x) := \mathbb{E}_\theta\big[\delta(X) \mid T(X)\big] = \eta(T(X))$ a function of $T(X)$.

$$E_{\theta,x}[\eta(x)|T(x)] = E_{\theta,x|T(x)}[\eta(x)] = \int \eta(x) f(x|T(x))dx \text{ no theta}$$

$\mathbb{E}_\theta(\eta(x)) = \mathbb{E}_\theta\big[\mathbb{E}_\theta\big[\delta(X) \mid T(X)\big]\big] = \mathbb{E}_\theta\big[\delta(X)\big]$

$\mathcal{L}(\theta,\eta)$ loss function

$R(\theta,\delta) = \mathbb{E}_\theta(\mathcal{L}(\theta,\delta(X)))$

**Lemma** ***Jensen Inequality***

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $X : \Omega \to \mathbb{R}$ be an integrable random variable, i.e. $E[|X|] < \infty$.
Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function such that $\varphi(X)$ is integrable. Then

$$\varphi\big(E[X]\big) \;\le\; E[\varphi(X)].$$

Moreover, if $\varphi$ is strictly convex, then equality holds if and only if $X$ is almost surely constant.

*Proof.* Let $a = \mathbb{E}[X]$. By the definition of convexity, for any $x$,

$$\phi(x) \geq \phi(a) + \phi'(a)(x-a).$$

Taking expectation on both sides gives

$$\mathbb{E}[\phi(X)] \geq \phi(a) + \phi'(a)(\mathbb{E}[X] - a) = \phi(a).$$

□

$$\begin{aligned} R(\theta, \eta) &= \mathbb{E}_\theta(\mathcal{L}(\theta, \eta(X))) = \mathbb{E}_\theta(\mathcal{L}(\theta, \eta(T(X)))) \\ &= \mathbb{E}_{\theta,x}[L(\theta, E_{\theta,x}[\delta(X)|T(X)])] = \mathbb{E}_{\theta,x}[\mathcal{L}(\theta, E_{\theta,x|T(X)}[\delta(X)])] \\ &\leq \mathbb{E}_{\theta,x}[E_{\theta,x|T(X)}[\mathcal{L}(\theta, \delta(X))]] \qquad \text{Jensen Inequality} \\ &= \mathbb{E}_{\theta,x}[\mathcal{L}(\theta, \delta(X))] = R(\theta, \delta(x)) \qquad \text{Law of iterated expectation} \end{aligned}$$

□

# ※ Lecture 6

**Example**

Let $X_i \sim N(\theta, 1) i.i.d. i = 1, \dots, n$. Show that $T = \sum\limits_{i=1}^{n} X_i$ is a sufficient statistic for $\theta$.

*Proof.* $f_\theta(x_1, \dots, x_n) = \prod_{i=1}^n f_\theta(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^n (x_i-\theta)^2}$

$$= \left[\frac{1}{(2\pi)^{n/2}} e^{\frac{\sum\limits_{i=1}^{n} x_i^2}{2}}\right] \cdot e^{-\frac{n\theta^2}{2} + \theta \sum\limits_{i=1}^{n} x_i} = h(x) \cdot g_\theta(T(x))$$

□

## 6.1 Minimal Sufficiency

**Definition** ***Minimal Sufficiency***

Suppose $T(X)$ is sufficient for $P = \{P_\theta : \theta \in \Omega\}$. For any other sufficient statistic $S(X)$, if we can always find a function $f$ such that $T = f(S)$, then $T$ is minimally sufficient.

$T = f(S)$ means
(i) the knowledge of $S$ implies the knowledge of $T$, and
(ii) $T$ provides a greater reduction of data unless $f$ is one-to-one.

A $d$-parameter exponential family has pdf in the following form

$$p(x, \theta) = h(x) \exp\left[\sum_{i=1}^{d} \eta_i(\theta) T_i(x) - A(\theta)\right],$$

which is of full rank if $\eta(\Theta) = \{\eta_1(\theta), \dots, \eta_d(\theta)\}$ has non-empty interior in $\mathbb{R}^d$ and $T_1(x), \dots, T_d(x)$ are linearly independent.

In a full rank exponential family, the natural sufficient statistic

$$T = (T_1, \dots, T_d)$$

is minimally sufficient.

**Example**

Let $X_i \sim N(\theta, \sigma^2) i.i.d. i = 1, \dots, n$.

$$f_{\mu,\sigma^2}(x_1, \dots, x_n) = \prod_{i=1}^{n} f_{\mu,\sigma^2}(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \cdot \exp\{\frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n\mu^2}{2\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\}$$

$$\eta_1(\theta) = \frac{\mu}{\sigma^2} \qquad T_1 = \sum_{i=1}^{n} x_i$$

$$\eta_2(\theta) = -\frac{1}{2\sigma^2} \qquad T_2 = \sum_{i=1}^{n} x_i^2$$

$$A(\theta) = \frac{n\mu^2}{2\sigma^2} \qquad h(x) = \frac{1}{(2\pi)^{n/2}\sigma^n}$$

## 6.2 Moments estimation

Suppose $\theta = (\theta_1, \dots, \theta_k)$. For $j = 1, \dots, k$, define the $j^{th}$ moment

$$\alpha_j \equiv \alpha_j(\theta) = E_\theta[X^j] = \int x^j \, dF_\theta(x)$$

and the $j^{th}$ sample moment

$$\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j.$$

The method of moments estimator $\hat{\theta}_n$ is defined to be the value of $\theta$ such that

$$\begin{aligned} \alpha_1(\hat{\theta}_n) &= \hat{\alpha}_1, \\ \alpha_2(\hat{\theta}_n) &= \hat{\alpha}_2, \\ &\vdots \\ \alpha_k(\hat{\theta}_n) &= \hat{\alpha}_k. \end{aligned}$$

**Example**

For normal distribution $N(\mu, \sigma^2)$, we have

$$\alpha_1(\theta) = E[X] = \mu, \qquad \alpha_2(\theta) = E[X^2] = \mu^2 + \sigma^2.$$

The method of moments estimators are

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \hat{\mu}^2.$$

**MOM generalization:** Instead of using $\alpha_j(\theta) = E_\theta[X^j]$, we can consider

$$\alpha_j(\theta) = E_\theta[g(X)^j]$$

and find $\hat{\theta}_n$ such that

$$\alpha_j(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} g(X_i)^j, \qquad j = 1, \ldots, n.$$

*Why do this?*

1. Flexibility: Sometimes raw moments don't exist (e.g., Cauchy distribution has no mean/variance), or are not convenient to solve.
2. Efficiency: Choosing $g_j$ cleverly can give better estimators (lower variance).
3. Connection to GMM: The generalized method of moments (GMM) in econometrics formalizes this idea—use more (possibly redundant) moment conditions than parameters, and solve them optimally.

## ※ Lecture 7

### 7.1 Maximum Likelihood Estimation

$$\mathcal{L}_n(\theta) = f_\theta(X_1, \ldots, X_n; \theta) = \prod_{i=1}^{n} f_\theta(X_i; \theta) \text{if the data are independent}$$

log-likelihood function

$$l_n(\theta) = \log \mathcal{L}_n(\theta) = \sum_{i=1}^{n} \log f_\theta(X_i; \theta)$$

If the log-likelihood function is differentiable, then the MLE $\hat{\theta}$ satisfies

$$\frac{\partial l_n(\theta)}{\partial \theta_j} = 0 \text{ for j = 1,...,p}$$

But still need to check the second order condition and boundaries where the likelihood is maximized.

**Example**

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$

$$\mathcal{L}_n(\theta) = f_\theta(X_1, \ldots, X_n; \theta) = \prod_{i=1}^{n} f_\theta(X_i; \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i-\theta)^2}{2}} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2}$$

$$l_n(\theta) = \log \mathcal{L}_n(\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(X_i - \theta)^2$$

$$\frac{\partial l_n(\theta)}{\partial \theta} = \sum_{i=1}^{n}(X_i - \theta) = 0 \implies \hat{\theta} = \bar{X}_n$$

$$\frac{\partial^2 l_n(\theta)}{\partial \theta^2} = -n < 0 \text{ (max)}$$

But if with the restriction $\theta \in [0, \infty)$, then

$$\hat{\theta} = \max(0, \bar{X}_n)$$

**Example**

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U[0, \theta]$. Find MLE and MOM.

$$\mathcal{L}_n(\theta) = f_\theta(X_1, \ldots, X_n; \theta) = \prod_{i=1}^{n} f_\theta(X_i; \theta) = \prod_{i=1}^{n} \frac{1}{\theta} I(X_i \in [0, \theta]) = \frac{1}{\theta^n} I(\max(X_i) \leq \theta)$$

$$l_n(\theta) = \log \mathcal{L}_n(\theta) = -n\log\theta + \log I(\max(X_i) \leq \theta)$$

The likelihood is decreasing in $\theta$ for $\theta \geq \max(X_i)$, so the MLE is

$$\hat{\theta}_{MLE} = \max(X_i)$$

The MOM estimator is

$$\alpha_1(\theta) = EX_1 = \frac{\theta}{2}, \qquad \hat{\alpha}_1 = \bar{X}_n$$

$$\hat{\theta}_{MOM} = 2\bar{X}_n$$

# ※ Lecture 8

## 8.1 MLE

- If $\hat{\theta}_n$ is MLE of $\theta$, then $g(\hat{\theta}_n)$ is MLE of $g(\theta)$.
- Under certain conditions $\hat{\theta}_n \xrightarrow{p} \theta$.

  *We assert:* The following conditions are sufficient for consistency of the MLE:

  1. $X_1, \ldots, X_n$ are *iid* with density $f(x;\theta)$.
  2. Identifiability, i.e. if $\theta \neq \theta'$, then $f(x;\theta) \neq f(x;\theta')$.
  3. The densities $f(x;\theta)$ have common support, i.e. $\{x : f(x;\theta) > 0\}$ is the same for all $\theta$.
  4. The parameter space $\Theta$ contains an open set $\omega$ of which the true parameter value $\theta^*$ is an interior point.
  5. The function $f(x;\theta)$ is differentiable with respect to $\theta$ in $\omega$.

These conditions ensure uniform convergence in probability of a normalized form of the log-likelihood to its expected value.

Note that

$$\ell_n(\theta) = \sum_{i=1}^{n} \log f(X_i;\theta) \propto \frac{1}{n}\sum_{i=1}^{n} \log f(X_i;\theta) \xrightarrow{P} \mathbb{E}_{\theta^*}\left[\log f(X_1;\theta)\right] \quad \text{for any fixed } \theta \text{ by WLLN.}$$

where $\theta^*$ denotes the true value of $\theta$. Showing consistency requires that the convergence is uniform in $\theta$. We also need to show that

$$\mathbb{E}_{\theta^*}\left[\log f(X_1;\theta)\right]$$

is maximized at $\theta = \theta^*$ since $\hat{\theta}_n$ maximizes $\ell_n(\theta)$.

*Proof.* By property of $iid$ and common support, we have

$$\mathbb{E}_{\theta^*}[\log f(X_1;\theta)] - \mathbb{E}_{\theta^*}[\log f(X_1;\theta^*)] = \int f(x;\theta^*) \log \frac{f(x;\theta)}{f(x;\theta^*)}\, dx$$

Since $\log$ is a concave function, by Jensen's inequality we have

$$\int f(x;\theta^*) \log \frac{f(x;\theta)}{f(x;\theta^*)}\, dx \leq \log \int f(x;\theta^*) \frac{f(x;\theta)}{f(x;\theta^*)}\, dx = 0$$

Given by the fact that $\int f(x;\theta)\,dx = 1$ for any $\theta$.

Thus

$$\mathbb{E}_{\theta^*}[\log f(X_1;\theta)] \leq \mathbb{E}_{\theta^*}[\log f(X_1;\theta^*)] \quad \text{for any } \theta$$

□

One class of distributions that satisfies the conditions is known as the **exponential family**. For $\Theta \subseteq \mathbb{R}$, these have densities that can be written as

$$f(x;\theta) = h(x)c(\theta)\exp\{\eta(\theta)T(x)\}.$$

**Example** ***Exponential*** $\lambda$

For the exponential family, we have

$$f(x;\lambda) = \lambda e^{-\lambda x} \text{ for } x \geq 0, \lambda > 0.$$

Here, $h(x) = 1_{[0,\infty)}(x)$, $c(\lambda) = \lambda$, $\eta(\lambda) = -\lambda$, and $T(x) = x$.

**Example** ***Binomial*** $n, p$

For the exponential family, we have

$$f(x;n,p) = \binom{n}{x}p^x(1-p)^{n-x} \text{ for } x = 0, 1, \ldots, n, n \in \mathbb{N}, p \in (0,1).$$

Here, $h(x) = \binom{n}{x}$, $c(p) = (1-p)^n$, $\eta(p) = \log\frac{p}{1-p}$, and $T(x) = x$.

**Example** ***Normal*** $\mu, \sigma^2$

For the exponential family, we have

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \text{ for } x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma^2 > 0.$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right)$$

Here, $h(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, $c(\mu,\sigma^2) = \frac{1}{\sqrt{\sigma^2}}e^{-\mu^2/(2\sigma^2)}$, $\eta^\top(\theta) = (-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2})$, and $T(x)^\top = (x, x^2)$.

**Definition** ***Fisher Information***

Define the score function $s(X;\theta) = \dfrac{\partial}{\partial\theta} \log f(X;\theta)$.
Then the **Fisher information** (based on $n$ observations) is

$$I_n(\theta) = V_\theta\left(\frac{\partial}{\partial\theta}\ell_n(\theta)\right) = V_\theta\left(\sum_{i=1}^{n} s(X_i;\theta)\right).$$

$$= \sum_{i=1}^{n} V_\theta(s(X_i;\theta)) \quad \text{(if } X_1,\dots,X_n \text{ are independent)}$$

$$= nV_\theta(s(X_1;\theta)) \quad \text{(if } X_1,\dots,X_n \text{ are identically distributed)}$$

$$= nI_1(\theta) \equiv nI(\theta).$$

where $V_\theta(\cdot)$ stands for variance.

### 8.2 Fisher Information Identity

For a single observation $X \sim f(x;\theta)$, the score function is

$$s(X;\theta) = \frac{\partial}{\partial\theta} \log f(X;\theta).$$

The **Fisher information** is defined as

$$I(\theta) = \mathrm{Var}_\theta(s(X;\theta)).$$

For $n$ i.i.d. observations $X_1,\dots,X_n$, the Fisher information is

$$I_n(\theta) = \mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\ell_n(\theta)\right) = \mathrm{Var}_\theta\left(\sum_{i=1}^{n} s(X_i;\theta)\right),$$

where

$$\ell_n(\theta) = \sum_{i=1}^{n} \log f(X_i;\theta).$$

If $X_i$ are i.i.d., then

$$I_n(\theta) = nI(\theta).$$

—

**Proposition 8.1** (Sufficient Conditions for Fisher Information Identity)**.** Let $X \sim f(x;\theta)$ with pdf (or pmf) $f(x;\theta)$. If the following conditions hold:

1. **Differentiability:** $f(x;\theta)$ is twice differentiable with respect to $\theta$.
2. **Support stability:** The support $\{x : f(x;\theta) > 0\}$ does not depend on $\theta$.
3. **Interchange of differentiation and integration:** Differentiation under the integral sign is valid, i.e.

   $$\frac{\partial}{\partial\theta}\int f(x;\theta)\,dx = \int \frac{\partial}{\partial\theta} f(x;\theta)\,dx,$$

   and similarly for the second derivative. This is satisfied if there exists a function $g(x)$ such that

   $$\left|\frac{\partial}{\partial\theta} f(x;\theta)\right| \le g(x) \quad \text{and} \quad \left|\frac{\partial^2}{\partial\theta^2} f(x;\theta)\right| \le g(x)$$

   for all $\theta$ in an open interval containing the true parameter value, and

   $$\int g(x)\,dx < \infty$$

then the Fisher information admits the equivalent forms

$$I(\theta) = \mathbb{E}_\theta[s(X;\theta)^2] = -\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta} s(X;\theta)\right] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right],$$

where $s(X;\theta) = \frac{\partial}{\partial\theta}\log f(X;\theta)$ is the score function.

These are satisfied by exponential family distributions (e.g. Normal, Bernoulli, Poisson).

*Proof.* Start with the definition of the score:

$$s(X;\theta) = \frac{\partial}{\partial\theta}\log f(X;\theta).$$

Then

$$I(\theta) = \mathbb{E}_\theta\left[s(X;\theta)^2\right].$$

Note that

$$\mathbb{E}_\theta[s(X;\theta)] = \int \frac{\partial}{\partial\theta}\log f(x;\theta)\,dF(x;\theta).$$

Simplify:

$$\int \frac{1}{f(x;\theta)}\frac{\partial}{\partial\theta} f(x;\theta)\,f(x;\theta)\,dx = \int \frac{\partial}{\partial\theta} f(x;\theta)\,dx = \frac{\partial}{\partial\theta}\int f(x;\theta)\,dx = \frac{\partial}{\partial\theta}(1) = 0.$$

Thus the score has mean zero.

Now differentiate $s(X;\theta)$:

$$\frac{\partial}{\partial\theta}s(X;\theta) = \frac{\partial^2}{\partial\theta^2}\log f(X;\theta).$$

Taking expectation:

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}s(X;\theta)\right] = \mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right].$$

Note that

$$s(x;\theta) = \frac{f'(x;\theta)}{f(x;\theta)},$$

so

$$\frac{\partial^2}{\partial\theta^2}\log f(x;\theta) = \frac{f''(x;\theta)}{f(x;\theta)} - \left(\frac{f'(x;\theta)}{f(x;\theta)}\right)^2.$$

Multiply by $f(x;\theta)$:

$$\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\,f(x;\theta) = f''(x;\theta) - \frac{f'(x;\theta)^2}{f(x;\theta)}.$$

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}s(X;\theta)\right] = \int f''(x;\theta)\,dx - \int \frac{f'(x;\theta)^2}{f(x;\theta)}\,dx.$$

Since $\int f(x;\theta)\,dx = 1$ for all $\theta$,

$$\int f''(x;\theta)\,dx = \frac{\partial^2}{\partial\theta^2}\int f(x;\theta)\,dx = \frac{\partial^2}{\partial\theta^2}(1) = 0.$$

Thus

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}s(X;\theta)\right] = -\int \left(\frac{f'(x;\theta)}{f(x;\theta)}\right)^2 f(x;\theta)\,dx = -\mathbb{E}_\theta[s(X;\theta)^2].$$

Using integration by parts (or dominated convergence), one can show

$$I(\theta) = \mathbb{E}_\theta[s(X;\theta)^2] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right].$$

□

# ※ Lecture 9

## 9.1 MLE

Under two additional conditions (also satisfied by *iid* observations under exponential family models), we have

- **Asymptotic normality:**

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} N\left(0, \frac{1}{I(\theta)}\right)$$

- **Asymptotic efficiency:** If $\tilde{\theta}_n$ is some other estimator such that

$$\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{D} N(0, v(\theta)),$$

  then $v(\theta) \geq 1/I(\theta)$ for all $\theta$.

  Asymptotic normality still holds replacing $I(\theta)$ by $I(\hat{\theta})$, that is,

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{1/I(\hat{\theta}_n)}} \xrightarrow{D} N(0, 1)$$

  We can use this to construct approximate $1 - \alpha$ confidence intervals for $\theta$.

  Rmk.: In terms of exponential families, MLE has such nice properties because it is a solution to the likelihood equation, which involves the sufficient statistic.

$$f(x; \theta) = h(x)c(\theta)\exp\{\sum_{i=1}^{k} \eta_i(\theta)T_i(x)\}$$

i.e. the estimator is sufficient. The Rao-Blackwell theorem says that if we have an unbiased estimator, then conditioning on a sufficient statistic will give us a better (lower variance) unbiased estimator. MLE is already a function of the sufficient statistic, so it is already optimal in this sense.

*Proof.*

$$\frac{\partial \ell}{\partial \theta} |_{\theta^*} = 0$$

**Theorem** ***CLT***

Let $X_1, X_2, \ldots, X_n$ be iid with mean $\mu$ and variance $\sigma^2 < \infty$. Then

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{D} N(0, 1)$$

**Theorem** *Slutsky's theorem*

If $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} c$, then $X_nY_n \xrightarrow{D} cX$.

$$\frac{\partial}{\partial\theta}\ell_n(\hat{\theta}_n) - \frac{\partial}{\partial\theta}\ell_n(\theta^*) \overset{Taylor}{\approx} (\hat{\theta}_n - \theta^*)\frac{\partial^2}{\partial\theta^2}\ell_n(\theta^*)$$

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \approx -\frac{\sqrt{n}\frac{\partial}{\partial\theta}\ell_n(\theta^*)}{\frac{\partial^2}{\partial\theta^2}\ell_n(\theta^*)}$$

The expectation of the numerator:

$$\mathbb{E}\left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i;\theta^*)\right] = 0$$

The variance of the numerator:

$$\mathrm{Var}\left[\frac{\partial}{\partial\theta}\log f(X_i;\theta^*)\right] = I(\theta^*)$$

Rearrange the terms:

$$\frac{\frac{\sqrt{n}}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i;\theta^*)}{\sqrt{I(\theta^*)}} \quad \xrightarrow{D} \quad N(0,1)$$

And

$$\frac{n\sqrt{I(\theta^*)}}{-\frac{\partial^2}{\partial\theta^2}\ell_n(\theta^*)} \quad \xrightarrow{P} \quad \frac{1}{\sqrt{I(\theta^*)}}$$

Thus by Slutsky's theorem,

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \quad \xrightarrow{D} \quad N\left(0, \frac{1}{I(\theta^*)}\right)$$

□

**Example $X \overset{iid}{\sim} Exp(\theta)$**

Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} Exp(\theta)$, with pdf

$$f(x;\theta) = \theta e^{-\theta x}, \quad x > 0, \theta > 0$$

The log-likelihood is

$$\ell(\theta) = n \log \theta - \theta \sum_{i=1}^{n} x_i$$

The score function is

$$S(\theta) = \frac{\partial}{\partial \theta} \ell(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i$$

The MLE is

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{X}_n}$$

The Fisher information is

$$I_n(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \ell(\theta)\right] = \frac{n}{\theta^2}$$

Thus by asymptotic normality,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N\left(0, \theta^2\right)$$

So an approximate $1 - \alpha$ confidence interval for $\theta$ is

$$\hat{\theta} \pm z_{\alpha/2} \frac{\hat{\theta}}{\sqrt{n}}$$

## 9.2 Fisher Information Matrix

For a $p$-dimensional parameter $\theta = (\theta_1, \dots, \theta_p)$, the Fisher information matrix is

$$\begin{aligned} I(\theta)_n &= \mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \ell(\theta)\right)\left(\frac{\partial}{\partial \theta} \ell(\theta)\right)^T\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta \partial \theta^T} \ell(\theta)\right] \\ &= \begin{pmatrix} I_{1,1}(\theta) & I_{1,2}(\theta) & \cdots & I_{1,p}(\theta) \\ I_{2,1}(\theta) & I_{2,2}(\theta) & \cdots & I_{2,p}(\theta) \\ \vdots & \vdots & \ddots & \vdots \\ I_{p,1}(\theta) & I_{p,2}(\theta) & \cdots & I_{p,p}(\theta) \end{pmatrix} \end{aligned}$$

Let $\hat{\theta}_n$ be the (vector valued) MLE, and let $J_n(\theta) = I_n(\theta)^{-1}$. Then under appropriate regularity conditions and for large $n$,

$$\sqrt{n}(\hat{\theta}_n - \theta) \overset{D}{\approx} N(0, nJ_n(\theta))$$

We can use the marginal densities

$$\hat{\theta}_{n,i} \overset{D}{\approx} N\big(\theta_i, J_{n,ii}(\theta)\big)$$

to construct 95% confidence intervals for the individual parameters.

**Example** $X \sim N(\mu, \sigma^2)$

The log-likelihood is

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

The information matrix is

$$I(\mu, \sigma^2)_n = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

The inverse is

$$J(\mu, \sigma^2)_n = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}$$

Thus by asymptotic normality,

$$\sqrt{n}\begin{pmatrix} \hat{\mu} - \mu \\ \hat{\sigma}^2 - \sigma^2 \end{pmatrix} \xrightarrow{D} N\left(\mathbf{0}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}\right)$$

### 9.3 Multiparameter Delta method

Suppose $\tau = g(\theta_1, \ldots, \theta_k)$ is a differentiable function. Let $\nabla g = \left(\frac{\partial}{\partial\theta_1}g(\theta), \ldots, \frac{\partial}{\partial\theta_k}g(\theta)\right)'$ be the gradient of $g$, and suppose that $\nabla g$ evaluated at $\hat{\theta}_n$ is not zero. Then

$$\frac{\hat{\tau}_n - \tau}{\hat{s}e(\hat{\tau}_n)} \xrightarrow{D} N(0, 1)$$

where

$$\hat{s}e(\hat{\tau}_n) = \sqrt{(\nabla\hat{g})' J_n(\hat{\theta}_n)(\nabla\hat{g})}$$

and $\nabla\hat{g}$ is $\nabla g$ evaluated at $\hat{\theta}_n$.

**Example:** Continuing the example on page 19, let $\tau = g(\mu, \sigma) = \mu/\sigma$. Find the MLE for $\tau$ and its limiting normal distribution.

## ※ Lecture 10

### 10.1 Nonparametric Methods

We have $X_1, \dots, X_n \overset{\text{iid}}{\sim} F$, which we have no information about.

Bootstrap: Resample with replacement from the data $X_1, \dots, X_n$ to get $X_1^*, \dots, X_n^*$. Then compute the statistic of interest $T_n^* = T(X_1^*, \dots, X_n^*)$. Repeat this many times to get an empirical distribution of $T_n^*$, which approximates the sampling distribution of $T_n = T(X_1, \dots, X_n)$.

Say we have done $B$ bootstrap samples, and we have $T_{n,1}^*, \dots, T_{n,B}^*$. Then we have a vector $(T_{n,1}^*, \dots, T_{n,B}^*)$.

## ※ Lecture 11

### 11.1 Hypothesis Testing

A **statistical hypothesis** is a statement about a parameter (or a statistical functional in nonparametric models).

A hypothesis test partitions the parameter space $\Theta$ into two disjoint sets $\Theta_0$ and $\Theta_1$, and produces a decision rule for choosing between

$$H_0 : \theta \in \Theta_0 \quad \text{and} \quad H_1 : \theta \in \Theta_1$$

$H_0$ is called the *null hypothesis* and $H_1$ is the *alternative hypothesis*. The possible choices are:

- Reject $H_0$
- Fail to reject $H_0$

We evaluate a test using its *power function*, defined as

$$\beta(\theta) = P_\theta(X \in R)$$

## 11.2 Reject Rule

The decision of whether to reject $H_0$ is determined by whether the sample $X = (X_1, \ldots, X_n)$ falls into a predefined rejection region $R$.

Usually, the rejection region has the form

$$R = \{(x_1, \ldots, x_n) : T(x_1, \ldots, x_n) > c\}$$

where $T$ is called a *test statistic* and $c$ is the *critical value*.

The idea is to construct $R$ so that the probability of the data falling into it when $H_0$ is true is small.

And the test size would be $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$.

### Example *Normal distribution*

Suppose $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, and let $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ be the MLEs. If $H_0 : \mu = 0$, one test statistic we might consider is $T = |\hat{\mu}_n / \hat{\sigma}_n|$, reasoning that if $H_0$ is true, $T$ will tend to be small.

Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, with $\sigma^2$ known.
Test $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$ using rejection region

$$R = \{(x_1, \dots, x_n) : |\bar{X}_n| > c\}$$

Find and plot $\beta(\mu)$.

*Solution.*

$$\beta(\mu) = P_\mu(||(\bar{X}_n) > c|) = P_\mu\left(\bar{X}_n > c\right) + P_\mu\left(\bar{X}_n < -c\right)$$

$$= 1 - \Phi(\sqrt{n}(c - \mu)) + \Phi(\sqrt{n}(-c - \mu))$$

■

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm
# parameters
n = 10
sigma = 1
c = 1
x = np.linspace(-5, 5, 400)
f = 1 - norm.cdf(np.sqrt(n) * (c-x) / sigma) + norm.cdf(-np.sqrt(n) * (c+x) / sigma)
```

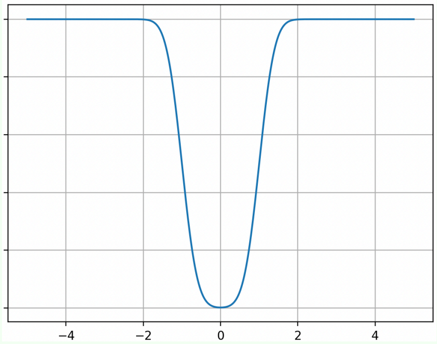


*Figure 1*

**Example**

Let $X \sim \text{Bin}(5, p)$. Test $H_0 : p \le \frac{1}{2}$ vs $H_1 : p > \frac{1}{2}$ with rejection regions:

$$R_1 = \{x : x = 5\}, \quad R_2 = \{x : x \ge 3\}$$

Plot and compare $\beta_1(p)$ and $\beta_2(p)$.

*Solution.* For a rejection region $R$, the power function is

$$\beta(p) = P_p(X \in R).$$

For $R_1 = \{x = 5\}$,

$$\beta_1(p) = P_p(X = 5) = \binom{5}{5} p^5 (1-p)^0 = p^5.$$

For $R_2 = \{x \ge 3\}$,

$$\begin{aligned} \beta_2(p) = P_p(X \ge 3) &= \sum_{x=3}^{5} \binom{5}{x} p^x (1-p)^{5-x} \\ &= 10p^3(1-p)^2 + 5p^4(1-p) + p^5. \end{aligned}$$

At $p = \frac{1}{2}$, the test sizes are

$$\alpha_1 = \beta_1(0.5) = (0.5)^5 = 0.03125, \quad \alpha_2 = \beta_2(0.5) = P_{0.5}(X \ge 3) = 0.5.$$

Hence $R_2$ gives higher power but also much larger size. ■

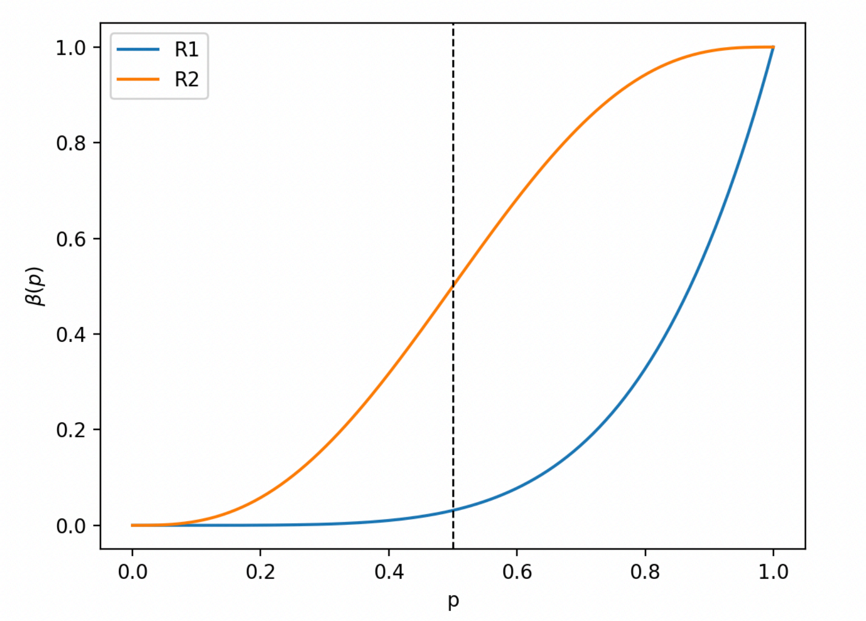


*Figure 2*

### 11.3 Size and Level of a Test

A test has *level* $\alpha$ if its size $\leq \alpha$, where

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$$

That is, $\alpha$ is the largest probability of rejecting $H_0$ when $H_0$ is true (Type I error).

| | Fail to reject $H_0$ | Reject $H_0$ |
|---|---|---|
| $H_0$ true | Correct | Type I error |
| $H_1$ true | Type II error | Correct |

$$P_{H_0 \text{ True}}(\text{Type I error}) = P_{H_0}(X \in R) \leq \sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$$

$$P_{H_1 \text{ True}}(\text{Type II error}) = P_{H_1}(X \notin R) = 1 - P_{H_1}(X \in R) \leq 1 - \inf_{\theta \in \Theta_1} \beta(\theta)$$

## ※ Lecture 12

### 12.1 Hypothesis Testing

$$H_0 : g(\theta) = g(\theta_0)$$

If the distribution is from *an exponential family*, and $g(\theta)$ is a linear function of the natural parameter, then

$$\frac{g(\hat{\theta}_n) - g(\theta_0)}{\hat{se}(g(\hat{\theta}_n))} \xrightarrow{D} N(0,1)$$

where $T(F) = \mathbb{E}_F r(x)$ for any $r$

**Example**

Suppose that $X \sim \text{Bin}(m, p_1)$ and $Y \sim \text{Bin}(n, p_2)$. Construct a size $\alpha$ Wald test for $H_0 : p_1 = p_2$.

$$H_0 : p_1 - p_2 = 0 \qquad H_1 : p_1 - p_2 \neq 0$$

where $\hat{p_1} - \hat{p_2} = X/m - Y/n$ is the MLE of $p_1 - p_2$.

**Example**

Let $F(u, v)$ be the joint distribution of two random variables $U$ and $V$. Let $\theta = T(F) = \rho(U, V)$, where $\rho$ denotes the correlation. Describe how to construct a size $\alpha$ Wald test for $H_0 : \rho = 0$ using the plug-in estimator and the bootstrap.

*Solution.*

$$\rho(U, V) = \frac{\mathbb{E}[(U - \mu_U)(V - \mu_V)]}{\sigma_U \sigma_V} = \frac{\mathbb{E}[UV] - \mu_U \mu_V}{\sigma_U \sigma_V} = \frac{\frac{1}{n}\sum_{i=1}^{n} U_i V_i - \bar{U}\bar{V}}{\hat{se}(U)\hat{se}(V)}$$

where $\hat{se}(U)$ and $\hat{se}(V)$ are the sample standard deviations of $U$ and $V$.

$$\hat{\rho} = \frac{\frac{1}{n}\sum_{i=1}^{n} U_i V_i - \bar{U}\bar{V}}{\hat{se}(U)\hat{se}(V)}$$

$$\hat{se}(\hat{\rho}) = \text{bootstrap estimate of standard error of } \hat{\rho}$$

The Wald test rejects $H_0$ when

$$\left|\frac{\hat{\rho} - 0}{\hat{se}(\hat{\rho})}\right| > z_{\alpha/2}$$

■

## 12.2 Likelihood Ratio Test (LRT)

Another broadly applicable class of tests is the **likelihood ratio test (LRT)**. Let

$$T(X) = \frac{\sup_{\theta \in \Theta} L_n(\theta)}{\sup_{\theta \in \Theta_0} L_n(\theta)}.$$

If $T(X)$ is large, it means there are values of $\theta$ in $\Theta_1$ that yield larger likelihood than any in $\Theta_0$. The likelihood ratio test rejects $H_0$ when

$$R = \{x : T(x) > c\}.$$

If $\hat{\theta}_n$ is the MLE and $\hat{\theta}_{n,0}$ is the MLE under the constraint $\theta \in \Theta_0$, then

$$T(X) = \frac{L_n(\hat{\theta}_n)}{L_n(\hat{\theta}_{n,0})}.$$

**Remark 12.1.** This LRT is always greater than or equal to 1, since the numerator is the unconstrained MLE and the denominator is the constrained MLE.

**Example**

Suppose $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, 1)$. Test $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$. Find $T(X)$ and simplify the rejection region. Use this to find the size $\alpha$ LRT.

*Solution.*

$$T(X) = \frac{L_n(\hat{\theta}_n)}{L_n(\hat{\theta}_{n,0})} = \frac{\exp\left(-\frac{1}{2}\sum_{i=1}^n (X_i - \bar{X})^2\right)}{\exp\left(-\frac{1}{2}\sum_{i=1}^n (X_i - \theta_0)^2\right)} = \exp\left(-\frac{1}{2}\left[\sum_{i=1}^n (X_i - \bar{X})^2 - \sum_{i=1}^n (X_i - \theta_0)^2\right]\right)$$

$$= \exp\left(-\frac{1}{2}\left[n(\bar{X} - \theta_0)^2 - 2(\bar{X} - \theta_0)\sum_{i=1}^n (X_i - \bar{X})\right]\right) = \exp\left(-\frac{n}{2}(\bar{X} - \theta_0)^2\right)$$

The rejection region is

$$R = \{x : T(x) > c\} = \{x : \exp(-\frac{n}{2}(\bar{X}-\theta_0)^2) > c\} = \left\{x : |\bar{X} - \theta_0| > \sqrt{-\frac{2}{n}\log c}\right\}$$

Power function:

$$\beta(\theta) = P_\theta(X \in R) = P_\theta\left(|\bar{X} - \theta_0| > \sqrt{-\frac{2}{n}\log c}\right) = 2 \cdot P\left(\bar{X} - \theta_1 > \sqrt{-\frac{2}{n}\log c} + \theta_0 - \theta_1\right)$$

The $\bar{X} - \theta_1 \sim N(0, \frac{1}{n})$.

■

When the exact power function cannot be computed, and $\Theta_0$ consists of fixing certain elements of $\theta$, we can use

$$\lambda(X) = 2\log T(X) \xrightarrow{D} \chi^2_{r-q},$$

where $r = \dim(\Theta)$ and $q = \dim(\Theta_0)$.

**Example**

Suppose $X_i \overset{iid}{\sim}$ Poisson$(\theta)$, and let $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be the MLE. For testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$,

$$\lambda = 2\log\frac{L(\hat{\theta}_n)}{L(\theta_0)} = 2n[(\theta_0 - \hat{\theta}_n) - \hat{\theta}_n \log(\theta_0/\hat{\theta}_n)].$$

Since $\lambda \xrightarrow{D} \chi_1^2$, reject $H_0$ if $\lambda > \chi^2_{1,\alpha}$.
Notice that

$$\hat{\theta}_n[\log(\theta_0)-\log(\hat{\theta}_n)] = \hat{\theta}_n[\frac{1}{\hat{\theta}_n}(\theta_0-\hat{\theta}_n)-\frac{1}{2\hat{\theta}_n^2}(\theta_0-\hat{\theta}_n)^2] = (\theta_0-\hat{\theta}_n)-\frac{1}{2\hat{\theta}_n}(\theta_0-\hat{\theta}_n)^2$$

Thus,

$$\lambda = 2n[(\theta_0-\hat{\theta}_n)-(\theta_0-\hat{\theta}_n)+\frac{1}{2\hat{\theta}_n}(\theta_0-\hat{\theta}_n)^2] = n\frac{(\theta_0-\hat{\theta}_n)^2}{\hat{\theta}_n} = \left(\frac{\theta_0-\hat{\theta}_n}{\sqrt{\frac{\hat{\theta}_n}{n}}}\right)^2 \sim \chi_1^2$$

# ❊ Lecture 13

## 13.1 Pearson's Test

**Example** ***Poisson***

$$H_0 : X_1, X_2 \ldots, X_n \sim Poisson(\lambda) H_1 : notnull$$

Construct $K$ categories, where category $i$ corresponds to observing $i-1$ events for $i = 1, 2, \ldots, K-1$ and category $K$ corresponds to observing at least $K-1$ events. Let $O_i$ be the observed counts in category $i$ and let $E_i$ be the expected counts in category $i$ under $H_0$. Then the test statistic is:

$$X^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$$

In practice, usually we use at least 5 categories.

$$\begin{aligned} \{0\} &:= i = 1 \\ \{1\} &:= i = 2 \\ &\vdots \\ \{K-2\} &:= i = K-1 \\ \{K-1, K, K+1, \ldots\} &:= i = K \end{aligned}$$

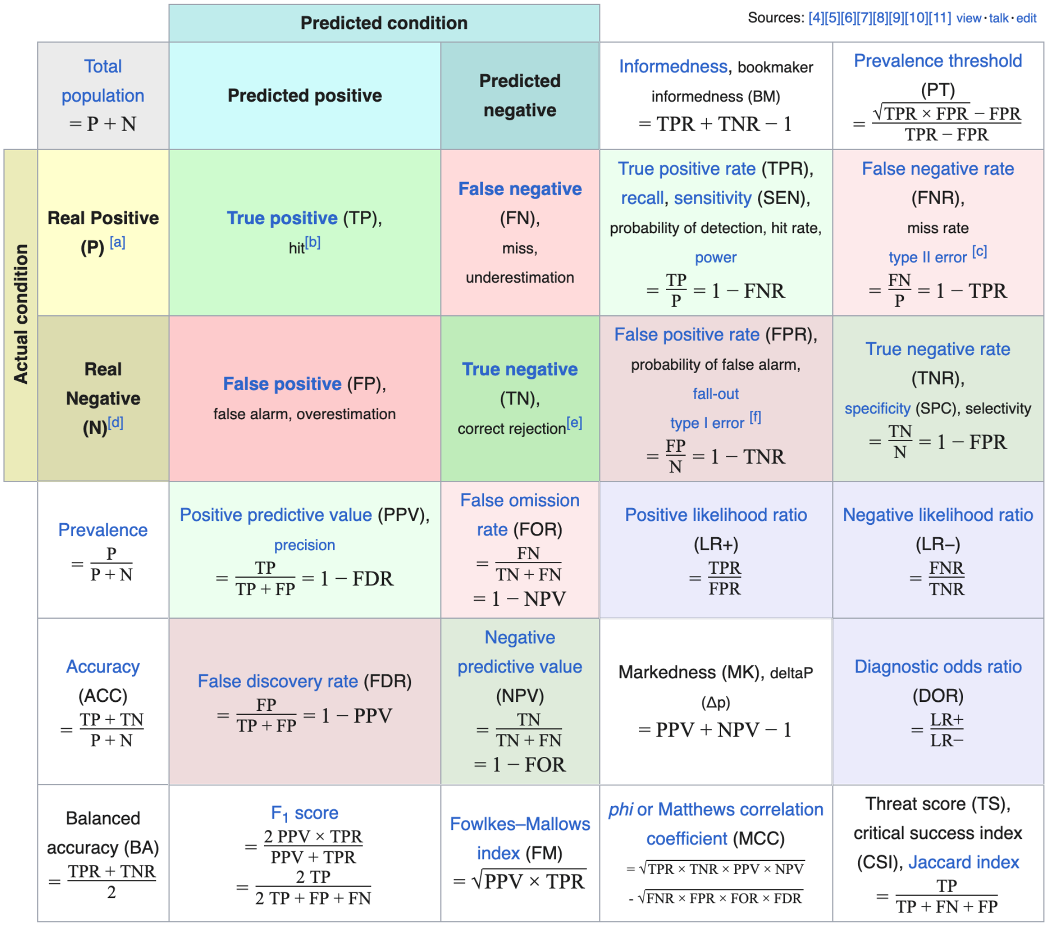
$$Y_j = \#\{x_i | x_i = j-1\} \text{ for } j = 1, 2, \ldots, K-1$$
$$Y_K = \#\{x_i | x_i \geq K-1\}$$
$$p_j(\lambda) = \begin{cases} e^{-\lambda} \frac{\lambda^{j-1}}{(j-1)!} & j = 1, 2, \ldots, K-1 \\ 1 - \sum_{i=0}^{K-2} e^{-\lambda} \frac{\lambda^i}{i!} & j = K \end{cases}$$

If $\lambda$ is known, then under $H_0$,

$$T(X) = \sum_{j=1}^{K} \frac{(Y_j - np_j(\lambda))^2}{np_j(\lambda)} \xrightarrow{D} \chi^2_{K-1}$$

If $\lambda$ is unknown, then

$$T(X) = \sum_{j=1}^{K} \frac{(Y_j - np_j(\lambda))^2}{np_j(\lambda)} \xrightarrow{D} \chi^2_{K-1-1}$$



| | | Predicted condition | | | Sources: [4][5][6][7][8][9][10][11] view · talk · edit |
|---|---|---|---|---|---|
| | Total population<br>$= \mathrm{P} + \mathrm{N}$ | Predicted positive | Predicted negative | Informedness, bookmaker informedness (BM)<br>$= \mathrm{TPR} + \mathrm{TNR} - 1$ | Prevalence threshold (PT)<br>$= \frac{\sqrt{\mathrm{TPR} \times \mathrm{FPR}} - \mathrm{FPR}}{\mathrm{TPR} - \mathrm{FPR}}$ |
| Actual condition | Real Positive (P) [a] | True positive (TP),<br>hit[b] | False negative (FN),<br>miss,<br>underestimation | True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power<br>$= \frac{\mathrm{TP}}{\mathrm{P}} = 1 - \mathrm{FNR}$ | False negative rate (FNR), miss rate<br>type II error [c]<br>$= \frac{\mathrm{FN}}{\mathrm{P}} = 1 - \mathrm{TPR}$ |
| | Real Negative (N)[d] | False positive (FP),<br>false alarm, overestimation | True negative (TN),<br>correct rejection[e] | False positive rate (FPR), probability of false alarm, fall-out<br>type I error [f]<br>$= \frac{\mathrm{FP}}{\mathrm{N}} = 1 - \mathrm{TNR}$ | True negative rate (TNR), specificity (SPC), selectivity<br>$= \frac{\mathrm{TN}}{\mathrm{N}} = 1 - \mathrm{FPR}$ |
| | Prevalence<br>$= \frac{\mathrm{P}}{\mathrm{P} + \mathrm{N}}$ | Positive predictive value (PPV), precision<br>$= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} = 1 - \mathrm{FDR}$ | False omission rate (FOR)<br>$= \frac{\mathrm{FN}}{\mathrm{TN} + \mathrm{FN}}$<br>$= 1 - \mathrm{NPV}$ | Positive likelihood ratio (LR+)<br>$= \frac{\mathrm{TPR}}{\mathrm{FPR}}$ | Negative likelihood ratio (LR−)<br>$= \frac{\mathrm{FNR}}{\mathrm{TNR}}$ |
| | Accuracy (ACC)<br>$= \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}}$ | False discovery rate (FDR)<br>$= \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}} = 1 - \mathrm{PPV}$ | Negative predictive value (NPV)<br>$= \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}$<br>$= 1 - \mathrm{FOR}$ | Markedness (MK), deltaP (Δp)<br>$= \mathrm{PPV} + \mathrm{NPV} - 1$ | Diagnostic odds ratio (DOR)<br>$= \frac{\mathrm{LR+}}{\mathrm{LR-}}$ |
| | Balanced accuracy (BA)<br>$= \frac{\mathrm{TPR} + \mathrm{TNR}}{2}$ | $\mathrm{F}_1$ score<br>$= \frac{2\,\mathrm{PPV} \times \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}}$<br>$= \frac{2\,\mathrm{TP}}{2\,\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$ | Fowlkes–Mallows index (FM)<br>$= \sqrt{\mathrm{PPV} \times \mathrm{TPR}}$ | phi or Matthews correlation coefficient (MCC)<br>$= \sqrt{\mathrm{TPR} \times \mathrm{TNR} \times \mathrm{PPV} \times \mathrm{NPV}}$<br>$- \sqrt{\mathrm{FNR} \times \mathrm{FPR} \times \mathrm{FOR} \times \mathrm{FDR}}$ | Threat score (TS), critical success index (CSI), Jaccard index<br>$= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}}$ |

*Figure 3: Positive and Negative Predictive Values vs Prevalence*

## 13.2 Bayesian Statistics

$$f(\theta|x^n) = \frac{f(x^n|\theta)(f(\theta))}{f(x^n)}$$

1. $f(\theta)$ is the prior distribution of $\theta$.
2. $f(x^n|\theta)$ is the likelihood function.
3. $f(\theta|x^n)$ is the posterior distribution of $\theta$ given data $x^n$.
4. $f(x^n)$ is the marginal likelihood of the data, can be hard to compute. Serve as the normalizing constant.

Good news: We often do not need to compute $f(x^n)$, since the family of prior and posterior distributions are often the same (conjugate prior).

**Example** ***Normal Model***

Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, and the prior distribution of $\mu$ is $N(\mu_0, \sigma_0^2)$, i.e.,

$$f(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}}$$

*Solution.*

$$f(\mu|x_1, \ldots, x_n) \propto f(x_1, \ldots, x_n|\mu) f(\mu) \propto \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}}$$

$$\propto e^{-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left(\mu-\frac{\frac{n\bar{x}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}\right)^2}$$

■

**Example** ***Poisson-Gamma Model***

Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Poisson(\theta)$, and the prior distribution of $\theta$ is $Gamma(\alpha, \beta)$, i.e.,

$$f(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}, \theta > 0$$

*Solution.*

$$f(\theta|x_1, \ldots, x_n) \propto f(x_1, \ldots, x_n|\theta) f(\theta) \propto \left(\prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!}\right) \theta^{\alpha-1} e^{-\beta\theta} \propto \theta^{(\sum_{i=1}^n x_i)+\alpha-1} e^{-(n+\beta)\theta}$$

$$\mathbb{E}(\theta|X_1, \ldots, X_n) = \frac{\sum\limits_{i=1}^{n} X_i + \alpha}{n+\beta} = \frac{n}{n+\beta}\bar{X}_n + \frac{\beta}{n+\beta}\frac{\alpha}{\beta}$$

■

**Example**

For $X_1, \dots, X_n|\theta \overset{iid}{\sim} N(\theta, \sigma^2)$ , where $\sigma^2$ is known, $\theta \sim N(a, b^2)$ The posterior distribution is

$$\theta|x^n \sim N\left(\frac{\frac{n\bar{x}}{\sigma^2}+\frac{a}{b^2}}{\frac{n}{\sigma^2}+\frac{1}{b^2}}, \frac{1}{\frac{n}{\sigma^2}+\frac{1}{b^2}}\right)$$

The mean could be written as a weighted average of the prior mean and the sample mean:

$$\frac{\frac{n\bar{x}}{\sigma^2}+\frac{a}{b^2}}{\frac{n}{\sigma^2}+\frac{1}{b^2}} = \left(\frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2}+\frac{1}{b^2}}\right)\bar{x} + \left(\frac{\frac{1}{b^2}}{\frac{n}{\sigma^2}+\frac{1}{b^2}}\right)a$$

When $n \longrightarrow \infty$, the weight for the prior $\rightarrow 0$.

# ※ Lecture 14

## 14.1 Bayesian

For rejection sampling, the $\theta_i^F$ := the $i$-th item from final list of $B$ size.
It follows some distribution $q$

$$q(\theta) = \frac{f(\theta)f(X_1, \dots, X_n|\theta)}{f(X_1, \dots, X_n)}$$

$$q(\theta) \propto f(\theta)\frac{L_n(\theta)}{L_n(\hat{\theta}_n)} \propto f(\theta)f(X_1, \dots, X_n|\theta) \propto f(\theta|X_1, \dots, X_n)$$

## 14.2 Importance Sampling

$$\begin{aligned} E_h[q(\theta)] &= \int q(\theta)h(\theta)\,d\theta \\ &= \int q(\theta)\frac{h(\theta)}{g(\theta)}g(\theta)\,d\theta \\ &\approx \frac{1}{B}\sum_{i=1}^{B} q(\theta_i)\frac{h(\theta_i)}{g(\theta_i)}. \end{aligned}$$

We choose $h(\theta)$ to be the posterior distribution $f(\theta|X_1, \dots, X_n)$, and $g(\theta)$ to be the prior distribution $f(\theta)$.

Then we have

$$E_{f(\theta|X_1,\dots,X_n)}[q(\theta)] \approx \frac{1}{B}\sum_{i=1}^{B} q(\theta_i)\frac{f(\theta_i|X_1,\dots,X_n)}{f(\theta_i)}$$

How to get $f(\theta_i|X_1,\dots,X_n)$?

Denote $L_n(\theta) = f(X_1,\dots,X_n|\theta)$, then

Denote the fraction to be :

$$w_i := \frac{f(\theta_i|X_1,\dots,X_n)}{B\cdot f(\theta_i)}$$

The numerator

$$f(\theta_i|X_1,\dots,X_n) = \frac{f(X_1,\dots,X_n|\theta_i)f(\theta_i)}{f(X_1,\dots,X_n)} = \frac{L_n(\theta_i)f(\theta_i)}{f(X_1,\dots,X_n)}$$

Thus, the weight is :

$$= \frac{L_n(\theta_i)}{B\cdot\int f(X_1,\dots,X_n|\theta)f(\theta)d\theta} \approx \frac{L_n(\theta_i)}{\sum_{j=1}^{B} L_n(\theta_j)}$$

(by approximating the denominator via Monte Carlo integration)

$$f(x_1,\dots,x_n) = \int f(x_1,\dots,x_n,\theta)d\theta = \int f(x_1,\dots,x_n|\theta)\frac{f(\theta)}{g(\theta)}g(\theta)d\theta \approx \frac{1}{B}\sum_{j=1}^{B} L_n(\theta_j)\frac{f(\theta_j)}{g(\theta_j)}$$

We can choose the second MC to use the same as the first MC, i.e., $g(\theta) = f(\theta)$

Finally, we have

$$\mathbb{E}[q(\theta)|x^n] \approx \sum_{i=1}^{B} w_i q(\theta_i)$$

# ❊ Lecture 15

## 15.1 Hypothesis Testing using Posterior Odds

**Example**

Albert Pujols (St. Louis Cardinals) and Ichiro Suzuki (Seattle Mariners) had very similar batting averages over 2001–2010. Their career totals in that span were:

Pujols: $n = 5146$ at-bats, $x = 1717$ hits    Suzuki: $m = 6099$ at-bats, $y = 2030$ hits.

Let $X \mid p_1 \sim \mathrm{Bin}(n, p_1)$ be Pujols' hits and $Y \mid p_2 \sim \mathrm{Bin}(m, p_2)$ be Suzuki's hits. We wish to assess evidence for/against the hypothesis $p_{\text{Pujols}} = p_{\text{Suzuki}}$.
Under $H_1 : \quad p_1 \neq p_2$, assign independent priors $p_1 \sim \mathrm{Unif}(0, 1)$ and $p_2 \sim \mathrm{Unif}(0, 1)$. Compute the marginal likelihood

$$f(x, y \mid H_1) = \int_0^1 \int_0^1 f(x \mid p_1)\, f(y \mid p_2)\, dp_1\, dp_2.$$

*Solution.* With $X \mid p_1 \sim \mathrm{Bin}(n, p_1)$ and $Y \mid p_2 \sim \mathrm{Bin}(m, p_2)$,

$$f(x \mid p_1, H_1) = \binom{n}{x} p_1^x (1 - p_1)^{n-x}, \qquad f(y \mid p_2, H_1) = \binom{m}{y} p_2^y (1 - p_2)^{m-y}.$$

Using independence and the Uniform$(0, 1) = \mathrm{Beta}(1, 1)$ priors,

$$\begin{aligned}
f(x, y \mid H_1) &= \iint f(x_1 \mid p_1, H_1)\, f(p_1\underset{1}{|}H_1)\, f(y \mid p_2, H_1) f(p_2\underset{1}{|}H_1)\, dp_1\, dp_2 \\
&= \binom{n}{x}\binom{m}{y} \left( \int_0^1 p_1^x (1 - p_1)^{n-x}\, dp_1 \right) \left( \int_0^1 p_2^y (1 - p_2)^{m-y}\, dp_2 \right) \\
&= \binom{n}{x}\binom{m}{y} B(x + 1, n - x + 1)\, B(y + 1, m - y + 1),
\end{aligned}$$

where $B(a, b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}$ is the Beta function. Since

$$\binom{n}{x} B(x + 1, n - x + 1) = \frac{n!}{x!(n - x)!} \cdot \frac{x!(n - x)!}{(n + 1)!} = \frac{1}{n + 1},$$

and similarly for $m, y$, the marginal likelihood simplifies to

$$\boxed{f(x, y \mid H_1) = \frac{1}{(n + 1)(m + 1)}}.$$

**Numerical value for these data.** With $n = 5146$ and $m = 6099$,

$$f(x, y \mid H_1) = \frac{1}{(5146+1)(6099+1)} = \frac{1}{31{,}396{,}700} \approx 3.19 \times 10^{-8}.$$

And for the null hypothesis $H_0 : \ p_1 = p_2 = p$, with prior $p \sim \text{Unif}(0, 1)$,

$$\begin{aligned} f(x, y \mid H_0) &= \int_0^1 f(x, y \mid p, H_0)\, dp \\ &= \int_0^1 f(x \mid p, H_0) f(y \mid p, H_0)\, dp \\ &= \binom{n}{x}\binom{m}{y} \int_0^1 p^{x+y}(1-p)^{(n-x)+(m-y)}\, dp \\ &= \binom{n}{x}\binom{m}{y} B(x + y + 1, n + m - (x + y) + 1). \end{aligned}$$

■

Let $p = P(H_1)$, so $P(H_0) = 1 - p$, and let $p^* = P(H_1 \mid \text{Data})$. By Bayes' rule,

$$p^* = \frac{f(\text{Data} \mid H_1)P(H_1)}{f(\text{Data} \mid H_1)P(H_1) + f(\text{Data} \mid H_0)P(H_0)}.$$

Divide numerator and denominator by $f(\text{Data} \mid H_0)P(H_0)$:

$$p^* = \frac{\dfrac{f(\text{Data} \mid H_1)}{f(\text{Data} \mid H_0)} \cdot \dfrac{p}{1-p}}{1 + \dfrac{f(\text{Data} \mid H_1)}{f(\text{Data} \mid H_0)} \cdot \dfrac{p}{1-p}}.$$

Define the Bayes factor $BF_{10} = \dfrac{f(\text{Data} \mid H_1)}{f(\text{Data} \mid H_0)}$ to obtain

$$\boxed{p^* = \frac{\dfrac{p}{1-p} BF_{10}}{1 + \dfrac{p}{1-p} BF_{10}}}$$

which is equivalent to the odds form

$$\frac{p^*}{1-p^*} = \frac{p}{1-p} BF_{10}.$$

# ❈ Lecture 16

## 16.1 Decision theory basics

**Definition**

Define a loss function:

$$\mathcal{L}(\theta,\hat{\theta}) : (\Theta \times \mathcal{A}) \longrightarrow [0,\infty)$$

e.g. squared error loss: $\mathcal{L}(\theta,\hat{\theta}) = (\theta - \hat{\theta})^2$

**Risk** (Average loss over all possible data) a.k.a. frequentist risk:

$$R(\theta,\hat{\theta}) = \mathbb{E}_{X|\theta}\left[\mathcal{L}(\theta,\hat{\theta}(X))\right] = \int \mathcal{L}(\theta,\hat{\theta}(x))f(x|\theta)dx$$

**Posterior risk** (Average loss over posterior distribution of $\theta$):

$$r(\hat{\theta}|\mathbf{x}) = E_{\theta|\mathbf{x}}\left[\mathcal{L}(\theta,\hat{\theta})\right] = \int \mathcal{L}(\theta,\hat{\theta})f(\theta|\mathbf{x})\mathbf{d}\theta$$

where $\mathbf{x} \triangleq (x_1,\ldots,x_n)$

**Example**

Prior: $X_1,\ldots,X_n \overset{\text{iid}}{\sim} N(\theta,1), \qquad f(\theta) \sim N(0,\tau^2)$

$$f(\theta|\mathbf{x}) \propto \exp\{-\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2}\}\exp\{-\frac{\theta^2}{2\tau^2}\} \propto \exp\{-\frac{1}{2}\left((n+\frac{1}{\tau^2})\theta^2 - 2n\bar{x}\theta\right)\}$$

$$\theta|x \sim N(\frac{n\tau^2}{n\tau^2+1}\bar{X}_n, \frac{\tau^2}{n\tau^2+1})$$

$$\begin{aligned} r(\hat{\theta}|x) &= E_{\theta|x}(\theta-\hat{\theta})^2 = E_{\theta|x}(\theta^2) - 2\hat{\theta}\cdot E_{\theta|x}(\theta) + \hat{\theta}^2 \\ &= -(\hat{\theta} - \frac{n\tau^2}{n\tau^2+1}\bar{X}_n)^2 + \frac{\tau^2}{n\tau^2+1} \end{aligned}$$

Posterior risk is minimized at $\hat{\theta} = E[\theta|x] = \dfrac{n\tau^2}{n\tau^2+1}\bar{X}_n$

**Example** ***Frequentist risk***

$$\mathcal{L}(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$

$$\begin{aligned} R(\theta, \hat{\theta}) &= E_{x|\theta}(\hat{\theta})^2 - 2\theta \cdot E_{x|\theta}(\hat{\theta}) + \theta^2 \\ &= Var_{x|\theta}(\hat{\theta}) + (E_{x|\theta}(\hat{\theta}) - \theta)^2 = MSE(\hat{\theta}) \end{aligned}$$

Consider two estimators, $\hat{\theta}$ and $\hat{\theta}'$. We say $\hat{\theta}'$ **dominates** $\hat{\theta}$ if

$$R(\theta, \hat{\theta}') \leq R(\theta, \hat{\theta}) \quad \text{for all } \theta,$$

and

$$R(\theta, \hat{\theta}') < R(\theta, \hat{\theta}) \quad \text{for at least one } \theta.$$

The estimator $\hat{\theta}$ is called **inadmissible** if there is at least one other estimator $\hat{\theta}'$ that dominates it. Otherwise it is called **admissible**.

**Definition** ***Bayes risk***

The Bayes risk is defined as:

$$\begin{aligned} r(f, \hat{\theta}) &= \int \underbrace{R(\theta, \hat{\theta})}_{\text{frequentist risk}} f(\theta)\, d\theta \\ &= \int \left[\int L(\theta, \hat{\theta}(x)) f(x \mid \theta)\, dx\right] f(\theta)\, d\theta \\ &= \int\int L(\theta, \hat{\theta}(x)) f(x, \theta)\, dx\, d\theta \\ &= \int \left[\int L(\theta, \hat{\theta}(x)) f(\theta \mid x)\, d\theta\right] f(x)\, dx \\ &= \int \underbrace{r(\hat{\theta} \mid x)}_{\text{posterior risk}} f(x)\, dx \end{aligned}$$

This expression averages over both $\theta$ and $X$. It depends on the particular form of $\hat{\theta}$, and on the probability models for the data $f(x \mid \theta)$ and the parameter $\theta$ ($f(\theta)$).

**Example**

$$X \sim N(\theta, 1), \hat{\theta}_c(x) = cx, \theta \sim N(0, \tau^2), \mathcal{L}(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2 = (\theta - cX)^2$$

**Frequentist -> Integral over the parameter**

$$R(\theta, \hat{\theta}_c) = E_{x|\theta}(\theta - cx)^2 = (c-1)^2\theta^2 + c^2$$

Bayes risk:

$$r(f, \hat{\theta}_c) = \int R(\theta, \hat{\theta}_c) f(\theta) d\theta = (c-1)^2\tau^2 + c^2$$

F.O.C. w.r.t. $c$:

$$\implies c = \frac{\tau^2}{\tau^2 + 1}$$

Thus, the Bayes rule is $\hat{\theta}_{\text{Bayes}} = \frac{\tau^2}{\tau^2+1} X$.

**Bayesian -> Integral over the data**

$$f(\theta \mid x) \propto f(x|\theta) f(\theta) \propto \exp\{-\frac{(x-\theta)^2}{2}\} \exp\{-\frac{\theta^2}{2\tau^2}\} \propto \exp\{-\frac{1}{2}\left((1 + \frac{1}{\tau^2})\theta^2 - 2x\theta\right)\}$$

Posterior distribution:

$$\theta | x \sim N(\frac{\tau^2}{\tau^2+1} x, \frac{\tau^2}{\tau^2+1})$$

Posterior risk:

$$r(\hat{\theta}_c | x) = E_{\theta|x}(\theta - cx)^2 = (\frac{\tau^2}{\tau^2+1} - c)^2 x^2 + \frac{\tau^2}{\tau^2+1}$$

Posterior rule is $\hat{\theta}_c = \frac{\tau^2}{\tau^2+1} x$, and the posterior risk is $\frac{\tau^2}{\tau^2+1}$ at this value. Thus, the posterior risk is invariant to $x$ at the posterior minimizing value. The Bayes risk is:

$$r(f, \hat{\theta}_c) = \int r(\hat{\theta}_c | x) f(x) dx = \frac{\tau^2}{\tau^2+1}$$

# ❊ Lecture 17

## 17.1 Decision Theory

**Example**

An investor is deciding whether or not to purchase \$1000 of risky ZZZ bonds. If the investor buys the bonds, they can be redeemed at maturity for a net gain of \$500. There could, however, be a default on the bonds, in which case the original \$1000 investment would be lost. If the investor doesn't buy the bonds, she will put her money in a "safe" investment, for which she will be guaranteed a net gain of \$300 over the same time period. She estimates the probability of a default to be 0.1.

*Solution.*

$$\mathcal{A} = \{a_1, a_2\} = \{\text{buy ZZZ bonds}, \text{don't buy ZZZ bonds}\}$$

$$\Theta = \{\theta_1, \theta_2\} = \{\text{default}, \text{no default}\}$$

$$R(\theta, a_1(x)) = \mathbb{E}_{x|\theta} L(\theta, a_1(x)) = \int L(\theta, a_1(x)) f(x \mid \theta)\, dx = L(\theta, a_1) \int f(x \mid \theta)\, dx = L(\theta, a_1)$$

$$r(f, a_1) = \mathbb{E}_\theta R(\theta, a_1) = 150 \qquad r(f, a_2) = \mathbb{E}_\theta R(\theta, a_2) = 300$$

■

## 17.2 Minimax

We want to choose an action that minimizes the worst-case risk. The maximum Risk:

$$\bar{R}(a) = \sup_\theta R(\theta, a) = \max\{R(\theta_1, a_1), R(\theta_2, a_2)\}$$

**Example *Cont'd***

$$\bar{R}(a_1) = \sup_\theta R(\theta, a_1) = \max\{R(\theta_1.a_1), R(\theta_2, a_1)\} = 1500$$

$$\bar{R}(a_2) = \sup_\theta R(\theta, a_2) = \max\{R(\theta_1, a_2), R(\theta_2, a_2)\} = 300$$

$$a_2 := \text{minimax}$$

In the estimation context, our possible actions are estimators $\hat{\theta}$. Then the maximum risk is

$$\bar{R}(\hat{\theta}) = \sup_{\theta} R(\theta, \hat{\theta})$$

**Example**

$$X \sim N(\theta, 1) \qquad R(\theta, \hat{\theta}) = \mathbb{E}_{x|\theta}(\theta - \hat{\theta})^2 \qquad \hat{\theta}_c(x) = cx$$

1. Find $R(\theta, \hat{\theta}_c)$.
2. Find minimax rule $\hat{\theta}_{c^*}$.
3. Let prior $\theta \sim N(a, b)$. Determine Bayes rule $\theta$.

*Solution.*

$$\bar{R}(\hat{\theta}_c) = \sup_{\theta} R(\theta, \hat{\theta}_c) = \sup_{\theta} \mathbb{E}_{x|\theta}(\theta - cx)^2 = \theta^2(c^2 - 2c + 1) + c^2 = \begin{cases} +\infty & c \neq 1 \\ 1 & c = 1 \end{cases}$$

For minimax rule, choose $c = 1$.

$$f(\theta \mid x) \propto \exp\{-\frac{(x-\theta)^2}{2}\} \cdot \exp\{-\frac{(\theta - a)^2}{2b}\} \propto \exp\{-\frac{1+b}{2b}\left(\theta - \frac{bx + a}{1+b}\right)^2\}$$

Bayes rule:

$$\implies \mathbb{E}_{\theta|x}[\theta] = \frac{b}{1+b}x + \frac{a}{1+b} \neq \hat{\theta}_{c^*} \text{ which is of the form } cx$$

But the above is not in the form $cx$. So we compute the Bayes risk:

$$r(f, \hat{\theta}_c) = \mathbb{E}_{\theta} R(\theta, \hat{\theta}_c) = \mathbb{E}_{\theta}[(c-1)^2\theta^2 + c^2] = (a^2+b+1)(c - \frac{a^2+b}{a^2+b+1})^2 + a^2 + b - \frac{(a^2+b)^2}{a^2+b+1}$$

$$c = (a^2+b)/(a^2+b+1), \qquad \hat{\theta}_c \text{ minimizes } r(f, \hat{\theta}_c)$$

■

## 17.3 Geometry of Bayes and Minimax Rules for Finite $\Omega$

Given a finite parameter space $\Omega = \{\theta_1, \cdots, \theta_k\}$, we define the risk set as $S \subseteq \mathbb{R}^k$ such that

$$S = \{(y_1, \cdots, y_k) : y_i = R(\theta_i, \delta) \text{ for } \delta \in \mathcal{A}\}.$$

We can visualize $S$ in $\mathbb{R}^k$. Each decision rule $\delta$ corresponds to a point in $S$. The goal of decision theory is to find optimal points in $S$.

And by allowing randomized estimators, we can form convex combinations of points in $S$.

**Lemma.** The risk set $S$ is always convex when $\mathscr{A}$ has randomized estimators.

In this setting, a prior of $\theta$ can be considered as a finite vector

$$\lambda(\theta) = (\lambda_1, \cdots, \lambda_k) = (\lambda(\theta_1), \cdots, \lambda(\theta_k)),$$

with $\sum_{i=1}^k \lambda_i = 1$ and $\lambda \geq 0$. The Bayes risk is

$$r(\lambda, \delta) = \sum_{i=1}^{k} \lambda_i R(\theta_i, \delta) = (\lambda_1, \cdots, \lambda_k) \begin{pmatrix} y_1 \\ \vdots \\ y_k \end{pmatrix}.$$

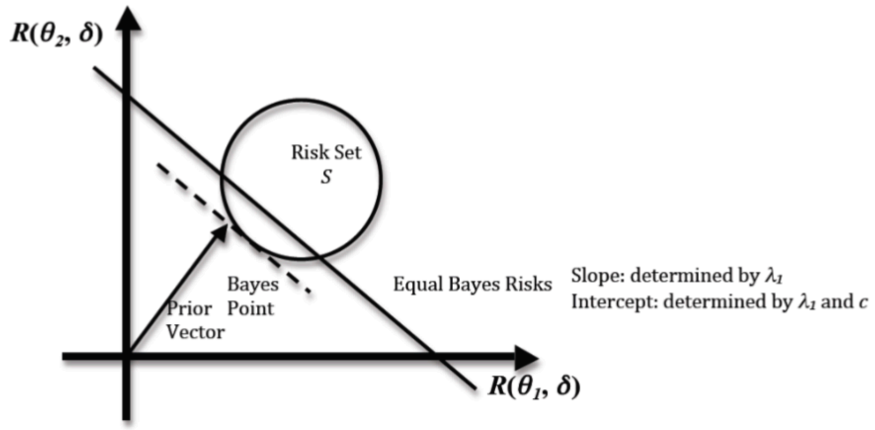


*Figure 4: Geometry of a Bayes Point for $k = 2$*

The tangent line with slope $-\lambda_1/\lambda_2$ corresponds to the Bayes rule with prior $\lambda = (\lambda_1, \lambda_2)$.

$$\lambda_1 \cdot R(\theta_1, \delta) + \lambda_2 \cdot R(\theta_2, \delta) = c$$

## References

[1] Larry Wasserman *All of Statistics*. Section 2 & 3

[2] Morris H. DeGroot *Probability and Statistics*.