Introduction to Statistics at an Advanced Level

STAT 201B

Shizhe Zhang

shizhe_zhang@berkeley.edu

Last updated: December 26, 2025

Online Resources:

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

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

Contents

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. $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), \qquad i = 1, \ldots, n$

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

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 θ.

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

Example

For $X_1 \ldots, X_n \sim N(\mu, \sigma^2)$, 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:

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

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

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

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:

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

[That is, for any ϵ, 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 $1-\alpha \to \epsilon = \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}$

Then we have

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

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

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

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

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

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

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

Example

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

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

```
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 $\frac{q(\theta)}{g(\theta)}$ 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 $j = 1, \ldots, B$:

1. Sample $X_{1,j} \ldots X_{n,j} \sim F$
2. Compute $T_{n,j} = g(X_{1,j}, \ldots, X_{n,j})$
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 μ and variance σ^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

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

$Z_1 \dots Z_k \sim F_2$ 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: $T(X_1, \dots, X_n)$ is sufficient for θ if the conditional distribution of $X_1, \dots, X_n$ given $T = t$ does not depend on θ. Keep the information about the parameters.
- Likelihood functions summarizes the information about θ contained in the data. Into a parameter-based function that drives inference.

Definition ***Sufficient Statistic***

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

❖ Lecture 5

5.1 Sufficiency

Motivation We hope to separate the information contained in the data into the information relevant for making inference about θ 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

- Increasing computational efficiency and decreasing storage requirements
- Involving irrelevant information may increase an estimator's risk (see Rao-Blackwell Theorem)
- 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 θ.

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

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

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

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 θ 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 θ, then

$$\begin{aligned} P_\theta(X = x) &= P_\theta(X = x | T(X) = T(x)) \cdot P_\theta(T(X) = T(x)) \\ &\underset{h(x)}{=} h(x) \cdot g(T(x),\theta) \end{aligned}$$

The first term is independent of θ according to the definition of Sufficient Statistics.

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

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

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 θ.

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

According to the definition of Sufficient Statistics, T is sufficient for θ. □

Example

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

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

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

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 θ. Given any estimator $\delta(X)$ of θ, 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

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

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 θ.

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

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

Taking expectation on both sides gives

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

□

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

□

❖ Lecture 6

Example

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

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

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

□

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 $\eta(\Theta) = \{\eta_1(\theta), \ldots, \eta_d(\theta)\}$ has non-empty interior in $\mathbb{R}^d$ and $T_1(x), \ldots, T_d(x)$ are linearly independent.

In a full rank exponential family, the natural sufficient statistic

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

is minimally sufficient.

Example

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

Solution.

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

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

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

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

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

■

6.2 Moments estimation

Definition

Suppose $\theta = (\theta_1, \ldots, \theta_k)$. For $j = 1, \ldots, k$, 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 θ such that

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

Example ***Normal*** μ, σ^2

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

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

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

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

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

Example

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

Solution.

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

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

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

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

■

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.

Solution.

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

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

The likelihood is decreasing in θ 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

Some properties of MLE:

- If $\hat{\theta}_n$ is MLE of θ, 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 θ.
4. The parameter space Θ contains an open set ω of which the true parameter value θ^* is an interior point.

5. The function $f(x;\theta)$ is differentiable with respect to θ in ω.

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

Note that

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

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

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

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

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

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

Thus

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

□

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* λ

For the exponential family, we have

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

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*** μ, σ^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

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

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

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

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

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 $X_1, \ldots, X_n$, 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 θ.
2. **Support stability:** The support $\{x : f(x;\theta) > 0\}$ does not depend on θ.
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 θ 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 θ,

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

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:**

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

- **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 θ.

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 θ.

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.

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

Theorem *CLT*

Let $X_1, X_2, \ldots, X_n$ be iid with mean μ 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 $X_n Y_n \xrightarrow{D} cX$.

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

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

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:

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

And

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

Thus by Slutsky's theorem,

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

□

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

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Exp(\theta)$, 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 θ is

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

9.2 Fisher Information Matrix

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

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

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 $\nabla g = \left(\frac{\partial}{\partial\theta_1} g(\theta), \ldots, \frac{\partial}{\partial\theta_k} g(\theta)\right)^\top$ be the gradient of g, and suppose that ∇g evaluated at $\hat{\theta}_n$ is not zero. Then

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

where

$$\hat{se}(\hat{\tau}_n) = \sqrt{(\nabla\hat{g})^\top J_n(\hat{\theta}_n)(\nabla\hat{g})}$$

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

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

❖ Lecture 10

10.1 Nonparametric Methods

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

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

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

❖ 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 Θ into two disjoint sets Θ_0 and Θ_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 $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$.

		Predicted condition		Sources: [4][5][6][7][8][9][10][11] view · talk · edit	
	Total population = P + N	**Predicted positive**	**Predicted negative**	Informedness, bookmaker informedness (BM) = TPR + TNR − 1	Prevalence threshold (PT) = $\frac{\sqrt{\text{TPR} \times \text{FPR}} - \text{FPR}}{\text{TPR} - \text{FPR}}$
Actual condition	**Real Positive (P)** [a]	**True positive** (TP), hit[b]	**False negative** (FN), miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power = $\frac{\text{TP}}{\text{P}} = 1 - \text{FNR}$	False negative rate (FNR), miss rate type II error [c] = $\frac{\text{FN}}{\text{P}} = 1 - \text{TPR}$
	Real Negative (N)[d]	**False positive** (FP), false alarm, overestimation	**True negative** (TN), correct rejection[e]	False positive rate (FPR), probability of false alarm, fall-out type I error [f] = $\frac{\text{FP}}{\text{N}} = 1 - \text{TNR}$	True negative rate (TNR), specificity (SPC), selectivity = $\frac{\text{TN}}{\text{N}} = 1 - \text{FPR}$
	Prevalence = $\frac{\text{P}}{\text{P} + \text{N}}$	Positive predictive value (PPV), precision = $\frac{\text{TP}}{\text{TP} + \text{FP}} = 1 - \text{FDR}$	False omission rate (FOR) = $\frac{\text{FN}}{\text{TN} + \text{FN}} = 1 - \text{NPV}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$
	Accuracy (ACC) = $\frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}$	False discovery rate (FDR) = $\frac{\text{FP}}{\text{TP} + \text{FP}} = 1 - \text{PPV}$	Negative predictive value (NPV) = $\frac{\text{TN}}{\text{TN} + \text{FN}} = 1 - \text{FOR}$	Markedness (MK), deltaP (Δp) = PPV + NPV − 1	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
	Balanced accuracy (BA) = $\frac{\text{TPR} + \text{TNR}}{2}$	F_1 score = $\frac{2\,\text{PPV} \times \text{TPR}}{\text{PPV} + \text{TPR}} = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FP} + \text{FN}}$	Fowlkes–Mallows index (FM) = $\sqrt{\text{PPV} \times \text{TPR}}$	*phi* or Matthews correlation coefficient (MCC) = $\sqrt{\text{TPR} \times \text{TNR} \times \text{PPV} \times \text{NPV}} - \sqrt{\text{FNR} \times \text{FPR} \times \text{FOR} \times \text{FDR}}$	Threat score (TS), critical success index (CSI), Jaccard index = $\frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}$

Figure 1: Positive and Negative Predictive Values vs Prevalence

Example ***Normal distribution***

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

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

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)
```

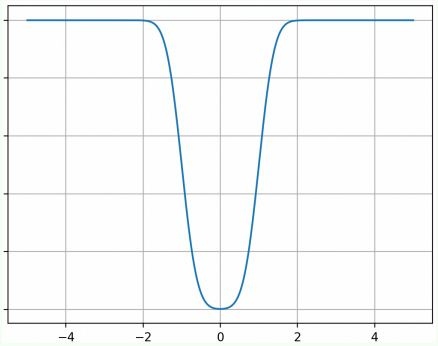

Figure 2

Example

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

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

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 $R_2 = \{x \geq 3\}$,

$$\beta_2(p) = P_p(X \geq 3) = \sum_{x=3}^{5} \binom{5}{x} p^x (1-p)^{5-x}$$

$$= 10p^3(1-p)^2 + 5p^4(1-p) + p^5.$$

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

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

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

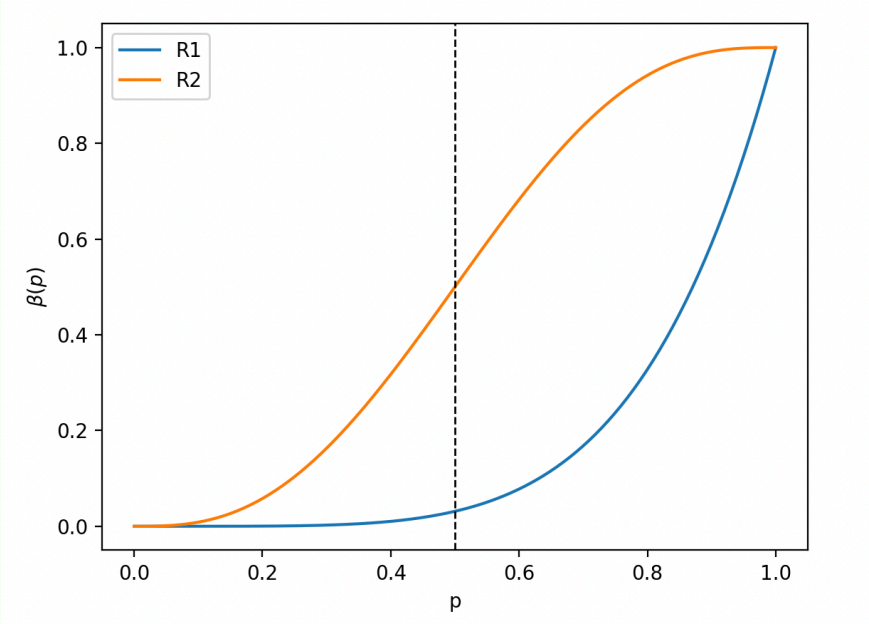

Figure 3

11.3 Size and Level of a Test

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

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

That is, α 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

Wald Test: Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Let $\hat{\theta}_n$ be an estimator such that

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

The size α Wald test rejects H_0 when $T > z_{\alpha/2}$, where

$$T = \left| \frac{\hat{\theta}_n - \theta_0}{\widehat{\text{se}}(\hat{\theta}_n)} \right|.$$

We can show that asymptotically, the Wald test has size α, and that it is obtained by inverting the approximate $1 - \alpha$ normal-based CI for θ.

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

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

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

Example

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

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

where $\hat{p}_1 - \hat{p}_2 = X/m - Y/n$ is the MLE of $p_1 - p_2$.

Solution. We have

$$\hat{p}_1 = \frac{X}{m}, \qquad \hat{p}_2 = \frac{Y}{n},$$

so the MLE of $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 = \frac{X}{m} - \frac{Y}{n}.$$

Since $X \sim \text{Bin}(m, p_1)$ and $Y \sim \text{Bin}(n, p_2)$ are independent,

$$\text{Var}(\hat{p}_1 - \hat{p}_2) = \text{Var}(\hat{p}_1) + \text{Var}(\hat{p}_2) = \frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}.$$

Under the null hypothesis $H_0 : p_1 - p_2 = 0$, we have $p_1 = p_2 = p$. We estimate p using the pooled estimator

$$\hat{p} = \frac{X + Y}{m + n}.$$

Thus, the estimated standard error of $\hat{p}_1 - \hat{p}_2$ is

$$\widehat{\text{se}}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}(1-\hat{p})}{m} + \frac{\hat{p}(1-\hat{p})}{n}}.$$

The Wald test statistic is

$$W = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\widehat{\text{se}}(\hat{p}_1 - \hat{p}_2)}.$$

Asymptotically, under H_0, $W \xrightarrow{D} N(0, 1)$. Therefore, the size α Wald test rejects H_0 if

$$|W| > z_{\alpha/2},$$

where $z_{\alpha/2}$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution. ■

Rmk. The $\hat{se}$ here is different from the one in Section handouts 5. The corresponding link in ED. Please correct me if I am wrong.

Example

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

Solution.

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

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

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

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

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 θ in Θ_1 that yield larger likelihood than any in Θ_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 $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$. 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 α 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 $\bar{X} - \theta_1 \sim N(0, \frac{1}{n})$, where θ_1 is the true value of θ. ■

When the exact power function cannot be computed, and Θ_0 consists of fixing certain elements of θ, 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 $X_i \overset{iid}{\sim} \text{Poisson}(\theta)$, 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 $\lambda \xrightarrow{D} \chi^2_1$, 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,

$$\lambda = 2n[(\theta_0 - \hat{\theta}_n) - (\theta_0 - \hat{\theta}_n) + \frac{1}{2\hat{\theta}_n}(\theta_0 - \hat{\theta}_n)^2] = n\frac{(\theta_0 - \hat{\theta}_n)^2}{\hat{\theta}_n} = \left(\frac{\theta_0 - \hat{\theta}_n}{\sqrt{\frac{\hat{\theta}_n}{n}}}\right)^2 \sim \chi^2_1$$

12.3 P-value

Suppose that for every $\alpha \in (0, 1)$ we have a size α test with rejection region R_α. When

$$R_\alpha = \{x : T(x) \geq c_\alpha\},$$
$$\text{p-value} = \sup_{\theta \in \Theta_0} P_\theta\big(T(X) \geq T(x)\big)$$

where x is the observed data.

Therefore, the p-value is the probability under H_0 of observing a value $T(X)$ the same as or more extreme than what was actually observed.

Equivalently,

$$\text{p-value} = \inf\{\alpha : T(x) \in R_\alpha\}.$$

That is, the p-value is the smallest level at which we can reject H_0 with x observed.

1. For Wald test with statistic

$$\text{p-value} = P_{\theta_0}(|W| > |w|) \approx P(|Z| > |w|) = 2(1 - \Phi(|w|)),$$

2. For LRT with statistic

$$\lambda(X) = 2\log T(X) \xrightarrow{D} \chi^2_{r-q}, \quad \text{p-value} = P\big(\chi^2_{r-q} \geq \lambda(x)\big).$$

Theorem

Theorem 12.2 (Neyman–Pearson lemma, simple vs. simple)**.** Let $X = (X_1, \dots, X_n)$ have likelihood $L_n(\theta) = f(x_1, \dots, x_n \mid \theta)$. Consider testing the simple hypotheses

$$H_0 : \theta = \theta_0 \qquad \text{vs.} \qquad H_1 : \theta = \theta_1.$$

Among all tests of size α (i.e. tests φ with $\mathbb{E}_{\theta_0}[\varphi(X)] \le \alpha$), the most powerful test has rejection region of the form

$$\left\{ x : \frac{L_n(\theta_1)}{L_n(\theta_0)} > k \right\}$$

for some constant k chosen so that the test has size exactly α.

Proof. Let $f_0(x)$ and $f_1(x)$ denote the joint densities of X under θ_0 and θ_1, respectively, so $L_n(\theta_0) = f_0(X)$ and $L_n(\theta_1) = f_1(X)$. A (non-randomized) test is determined by its rejection region $C \subseteq \mathcal{X}$: we reject H_0 if $X \in C$, and we do not reject otherwise.

The size and power of a test with region C are

$$\alpha(C) = \mathbb{P}_{\theta_0}(X \in C) = \int_C f_0(x)\,dx, \qquad \beta(C) = \mathbb{P}_{\theta_1}(X \in C) = \int_C f_1(x)\,dx.$$

Define the likelihood ratio

$$\Lambda(x) = \frac{f_1(x)}{f_0(x)}.$$

Fix $\alpha \in (0, 1)$. Consider the test which rejects H_0 when $\Lambda(x) > k$, where $k > 0$ is chosen so that

$$\alpha^* := \int_{\{\Lambda(x) > k\}} f_0(x)\,dx = \alpha.$$

(We ignore the set $\{\Lambda(x) = k\}$ or assume it has probability zero under f_0, so the equality can be achieved without randomization.) Call this rejection region

$$C^* = \{x : \Lambda(x) > k\}.$$

We must show that for any other test region C with $\alpha(C) \le \alpha$, we have $\beta(C) \le \beta(C^*)$.

Consider the difference in powers:

$$\beta(C^*) - \beta(C) = \int_{C^*} f_1(x)\,dx - \int_C f_1(x)\,dx = \int (I_{C^*}(x) - I_C(x)) f_1(x)\,dx,$$

where I_A is the indicator of set A.

Rewrite this using $\Lambda(x) = f_1(x)/f_0(x)$:

$$\beta(C^*) - \beta(C) = \int (I_{C^*}(x) - I_C(x))\Lambda(x) f_0(x)\, dx.$$

Now add and subtract k inside the integrand:

$$\beta(C^*) - \beta(C) = \int (I_{C^*}(x) - I_C(x))(\Lambda(x) - k) f_0(x)\, dx + k \int (I_{C^*}(x) - I_C(x)) f_0(x)\, dx.$$

By our choice of k,

$$\int I_{C^*}(x) f_0(x)\, dx = \alpha,$$

and since $\alpha(C) \le \alpha$,

$$\int I_C(x) f_0(x)\, dx \le \alpha.$$

Hence

$$\int (I_{C^*}(x) - I_C(x)) f_0(x)\, dx = \alpha(C^*) - \alpha(C) \ge 0,$$

so the second term is nonnegative:

$$k \int (I_{C^*}(x) - I_C(x)) f_0(x)\, dx \ge 0.$$

For the first term, note that by definition of C^*:

$$\Lambda(x) - k > 0 \quad \text{on } C^*, \qquad \Lambda(x) - k < 0 \quad \text{on } C^{*c}.$$

Also, $I_{C^*}(x) - I_C(x)$ can only take the values $-1, 0, 1$ and

$$I_{C^*}(x) - I_C(x) = \begin{cases} 1, & x \in C^* \setminus C, \\ -1, & x \in C \setminus C^*, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, when $I_{C^*}(x) - I_C(x) = 1$ we have $x \in C^*$ and $\Lambda(x) - k > 0$, so the product $(I_{C^*}(x) - I_C(x))(\Lambda(x) - k)$ is nonnegative; when $I_{C^*}(x) - I_C(x) = -1$ we have $x \in C \setminus C^* \subseteq C^{*c}$, so $\Lambda(x) - k < 0$, and the product is again nonnegative. Therefore

$$(I_{C^*}(x) - I_C(x))(\Lambda(x) - k) \ge 0 \quad \text{for all } x,$$

and hence

$$\int (I_{C^*}(x) - I_C(x))(\Lambda(x) - k) f_0(x)\, dx \ge 0.$$

Combining the two parts, we obtain

$$\beta(C^*) - \beta(C) \geq 0,$$

i.e. $\beta(C^*) \geq \beta(C)$ for every test C with $\alpha(C) \leq \alpha$.

Thus the likelihood-ratio test with rejection region $C^* = \{x : \Lambda(x) > k\}$ is most powerful of size α, which completes the proof. □

❖ Lecture 13

13.1 Pearson's Test

Review connection See LRT practice for more.

Multinomial model:

$$X = (X_1, X_2, \ldots, X_k) \sim \text{Multinomial}(n, p_1, p_2, \ldots, p_k)$$

where $\sum_{i=1}^k p_i = 1$ and $p_i \geq 0$ for all i.

The pdf is:

$$f(x_1, x_2, \ldots, x_k | p_1, p_2, \ldots, p_k) = \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

where $x_i \geq 0$ for all i and $\sum_{i=1}^k x_i = n$.

MLE:

$$\hat{p}_i = \frac{X_i}{n}, \quad i = 1, 2, \ldots, k$$

Definition ***Pearson's Chi-Squared Test***

We want to test:

$$H_0 : p = p^0 = (p_1^0, p_2^0, \ldots, p_k^0) \qquad H_1 : p \neq p^0$$

The test statistic is:

$$T = \sum_{i=1}^k \frac{(X_i - np_i^0)^2}{np_i^0} = \sum_{i=1}^k \frac{(X_i - E_i)^2}{E_i}$$

Under H_0, when n is large,

$$T \xrightarrow{D} \chi^2_{k-1}$$

The test rejects H_0 when $T > \chi^2_{k-1,\alpha}$.

Example *Poisson*

$$H_0 : X_1, X_2 \ldots, X_n \sim Poisson(\lambda) \qquad H_1 : \text{not null}$$

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 λ 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 λ is unknown, then
we use the MLE $\hat{\lambda} = \bar{X}_n$ to estimate λ. Then under H_0,

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

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 θ.
2. $f(x^n|\theta)$ is the likelihood function.
3. $f(\theta|x^n)$ is the posterior distribution of θ 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 μ 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.

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

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

■

Example *Poisson-Gamma Model*

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

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

Solution.

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

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

■

Example

For $X_1, \ldots, X_n|\theta \overset{iid}{\sim} N(\theta, \sigma^2)$, where σ^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 $n \to \infty$, the weight for the prior $\to 0$.

13.3 Posterior Inference

In Bayesian statistics, all inference is based on the posterior distribution. We can use the posterior to calculate quantities similar to those under frequentist statistics (point estimates and intervals), or we can examine the posterior probability of any event of interest.

The posterior mean is a commonly used point estimator:

$$\mathbb{E}[\theta \mid X_1, \ldots, X_n] = \int \theta\, f(\theta \mid X_1, \ldots, X_n)\, d\theta.$$

It can often be written as a weighted average of the prior mean and the MLE.

For example, in Example on the previous page,

$$\mathbb{E}[\theta \mid X_1, \ldots, X_n] = \frac{b^2 \sum_{i=1}^n X_i + a\sigma^2}{nb^2 + \sigma^2} = \frac{nb^2}{nb^2 + \sigma^2} \bar{X}_n + \frac{\sigma^2}{nb^2 + \sigma^2} a.$$

13.4 Credible Intervals

A $1 - \alpha$ credible interval for θ (also called a posterior interval) is an interval C_n satisfying

$$P(\theta \in C_n \mid X_1, \ldots, X_n) = 1 - \alpha.$$

Note a few differences compared to a confidence interval:

- The probability statement is about θ, not C_n. The interval C_n is a function of $X_1, \ldots, X_n$, which we are conditioning on in the probability statement.
- The statement is an equality. This is different from a frequentist interval, which puts a lower bound on the probability of coverage.
- The intervals constructed this way may or may not have good frequentist coverage rates.

13.5 Types of Credible Intervals

Note that C_n is not uniquely defined. There are several popular methods for finding such intervals.

A $1 - \alpha$ equal-tail credible interval is an interval (a, b) such that

$$\int_{-\infty}^{a} f(\theta \mid x^n)\, d\theta = \int_{b}^{\infty} f(\theta \mid x^n)\, d\theta = \frac{\alpha}{2}.$$

A $1 - \alpha$ highest posterior density (HPD) region R_n is defined such that:

1. $P(\theta \in R_n \mid x^n) = 1 - \alpha$,
2. $R_n = \{\theta : f(\theta \mid x^n) > k\}$ for some constant k.

When $f(\theta \mid x^n)$ is unimodal, R_n is an interval.

Often it is more informative to plot $f(\theta \mid x^n)$ than it is to report an interval.

❖ Lecture 14

14.1 MCMC

If the posterior distribution is complicated, we may not be able to get closed-form expressions for posterior quantities of interest, e.g. posterior mean, variance, quantiles, etc.

In such cases, we can use Monte Carlo methods to approximate these quantities.

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 $f(\theta|X_1,\ldots,X_n)$, and $g(\theta)$ to be the prior distribution $f(\theta)$.

Then we have

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

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

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

Denote the fraction to be :

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

The numerator

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

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

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

Thus, the weight is :

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

Finally, we have

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

14.3 Rejection Sampling

Goal: get s sample for $r(x)$

Proposal distribution: $g(x)$, which is easy to sample from.

Rejection procedure: find M s.t. $\forall x, \frac{r(x)}{M \cdot g(x)} < 1$

1. Sample $X_{cand} \sim g(x)$
2. Sample $U \sim \text{Uniform}(0, 1)$
3. If $U \leq \frac{r(X_{cand})}{M \cdot g(X_{cand})}$, accept X_{cand}; else reject X_{cand} and go to step 1.

Following the above procedure, the sample probability density function is:

$$f_{X_{acc}}(x) \propto g(x) \cdot \frac{r(x)}{M \cdot g(x)} \propto r(x)$$

Example

For posterior distribution sampling, we can set

$$r(\theta) = f(x|\theta), \quad g(\theta) = f(\theta)$$

We can multiply by some function that does not depend on θ to make the sampling easier:

$$\theta_1, \ldots, \theta_B \sim r(\theta) \propto r(\theta) f(x) = f(\theta, x) \propto f(x \mid \theta) g(\theta)$$

We need to find M such that

$$\frac{f(x|\theta)g(\theta)}{M \cdot g(\theta)} = \frac{f(x|\theta)}{M} < 1 \quad \forall \theta$$

$$\Rightarrow M = \sup_{\theta} f(x|\theta) = f(x|\hat{\theta}_{MLE})$$

❖ Lecture 15

15.1 Bayesian Hypothesis Testing

In a Bayesian analysis, hypotheses, like parameters, can be described using probability distributions.

The simplest case is when the hypotheses describe regions into which θ can fall, and these all have positive prior probability. If $H_0 : \theta \in \Theta_0$, then

- **Prior probability:**
$$P(H_0) = \int_{\Theta_0} f(\theta)\, d\theta$$
- **Posterior probability:**
$$P(H_0 \mid x^n) = \int_{\Theta_0} f(\theta \mid x^n)\, d\theta$$

Suppose $H_0, \ldots, H_{K-1}$ are K hypotheses under consideration (typically $K = 2$, but in theory we can have more). Suppose that under hypothesis H_k,

$$\theta \sim f(\theta \mid H_k).$$

Note that θ may mean different things under the various hypotheses.

Then,

$$P(H_k \mid x^n) = \frac{f(x^n \mid H_k)P(H_k)}{\sum_{k=1}^{K} f(x^n \mid H_k)P(H_k)}.$$

Therefore, the posterior odds of H_i relative to H_j equals

$$\frac{P(H_i \mid x^n)}{P(H_j \mid x^n)} = \frac{f(x^n \mid H_i)}{f(x^n \mid H_j)} \times \frac{P(H_i)}{P(H_j)}.$$

The term $\frac{f(x^n|H_i)}{f(x^n|H_j)}$ is called the **Bayes Factor** for comparing H_i to H_j, and is denoted by BF_{ij}.

When H_i and H_j represent regions of the parameter space, it is often easier to calculate the prior and posterior odds directly, and from these compute the Bayes Factor.

If

$$H_i : \theta = \theta_i \quad \text{and} \quad H_j : \theta = \theta_j,$$

then the Bayes Factor is simply the ratio of likelihoods under the two values:

$$\mathrm{BF}_{ij} = \frac{f(x^n \mid \theta_i)}{f(x^n \mid \theta_j)}.$$

More generally,

$$f(x^n \mid H_i) = \int_{\Theta} f(x^n \mid \theta, H_i)\, f(\theta \mid H_i)\, d\theta,$$

which is called the **marginal likelihood**.

If $f(\theta \mid H_i)$ is conjugate, this integral can often be calculated in closed form. Otherwise, we use sampling methods to approximate it. For example, we could use Monte Carlo integration by sampling from $f(\theta \mid H_i)$.

15.2 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 $X \mid p_1 \sim \text{Bin}(n, p_1)$ be Pujols' hits and $Y \mid p_2 \sim \text{Bin}(m, p_2)$ be Suzuki's hits. We wish to assess evidence for/against the hypothesis $p_{\text{Pujols}} = p_{\text{Suzuki}}$.
Under $H_0 : p_1 = p_2$, $H_1 : p_1 \neq p_2$, assign independent priors $p_1 \sim \text{Unif}(0, 1)$ and $p_2 \sim \text{Unif}(0, 1)$. 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 $X \mid p_1 \sim \text{Bin}(n, p_1)$ and $Y \mid p_2 \sim \text{Bin}(m, p_2)$,

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

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

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

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:

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

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 θ):

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

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

Example

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

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

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

$$r(\hat{\theta}|x) = E_{\theta|x}(\theta - \hat{\theta})^2 = E_{\theta|x}(\theta^2) - 2\hat{\theta} \cdot E_{\theta|x}(\theta) + \hat{\theta}^2$$

$$= -(\hat{\theta} - \frac{n\tau^2}{n\tau^2 + 1}\bar{X}_n)^2 + \frac{\tau^2}{n\tau^2 + 1}$$

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

Example *Frequentist risk*

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

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

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 & Bayes rule*

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 θ 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 θ ($f(\theta)$).
And the Bayes rule is defined as the decision rule $\hat{\theta}_{\text{Bayes}}$ that minimizes the Bayes risk:

$$\hat{\theta}_{\text{Bayes}} = \arg\min_{\hat{\theta}} r(f, \hat{\theta})$$

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:

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

Takeaway:

1. Posterior risk:

$$r(\hat{\theta} \mid x) = \mathbb{E}_{\theta|X}\left[\, L(\theta, \hat{\theta}(x)) \,\right].$$

2. Frequentist risk:

$$R(\theta, \hat{\theta}) = \mathbb{E}_{X|\theta}\left[L(\theta, \hat{\theta}(X)) \right].$$

3. Bayes risk:

$$r(f, \hat{\theta}) = \mathbb{E}_{\theta,X}\left[L(\theta, \hat{\theta}(X)) \right].$$

By iterated expectation, we also have that

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

and

$$r(f, \hat{\theta}) = \mathbb{E}_{X}\left[\mathbb{E}_{\theta|X}\left[L(\theta, \hat{\theta}(X))\right]\right] = \mathbb{E}_{X}\left[r(\hat{\theta} \mid X)\right].$$

Example

Suppose $X_1, \ldots, X_n \mid \sigma^2 \overset{\text{iid}}{\sim} N(\theta, \sigma^2)$ where θ is known. Let the prior for σ^2 be an inverse–gamma distribution $\text{Inv-Gamma}(a, b)$ with density

$$p(\sigma^2; a, b) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-a-1} \exp\{-b/\sigma^2\}, \qquad \sigma^2 > 0.$$

Solution:
The likelihood (as a function of σ^2) is

$$L(\sigma^2) \propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \theta)^2\right\}.$$

Write $S = \sum_{i=1}^{n}(X_i - \theta)^2$.
Multiplying prior by likelihood gives the posterior kernel

$$p(\sigma^2 \mid X) \propto (\sigma^2)^{-\left(a+1+\frac{n}{2}\right)} \exp\left\{-\frac{b + \frac{1}{2}S}{\sigma^2}\right\}.$$

Hence the posterior is again inverse–gamma:

$$\sigma^2 \mid X \sim \text{Inv-Gamma}\left(a + \frac{n}{2},\ b + \frac{1}{2}S\right).$$

Example *cont'd*

Define

$$a' = a + \frac{n}{2}, \qquad b' = b + \frac{1}{2}S.$$

The Bayes estimator under squared error is the posterior mean (provided it exists). For an $\text{Inv-Gamma}(a', b')$ distribution the mean is $b'/(a'-1)$ whenever $a' > 1$. Thus

$$\hat{\sigma}^2_{(\text{MSE})} = \mathbb{E}[\sigma^2 \mid X] = \frac{b + \frac{1}{2}S}{a + \frac{n}{2} - 1}, \qquad (\text{exists if } a + \tfrac{n}{2} > 1).$$

—

The Bayes estimator under absolute error loss is the posterior median:

$$\hat{\sigma}^2_{(\text{MAE})} = \text{median}\big(\sigma^2 \mid X\big).$$

The median of an inverse–gamma distribution does not have a simple closed form; it must be found numerically from

$$\int_0^{\hat{\sigma}^2_{(\text{MAE})}} p(t \mid X)\, dt = \tfrac{1}{2}.$$

—

A Bayes estimator under (point) 0–1 loss is any maximizer of the posterior density (the MAP estimator). For $\text{Inv-Gamma}(a', b')$ the mode is

$$\text{mode}(\sigma^2 \mid X) = \frac{b'}{a'+1} = \frac{b + \frac{1}{2}S}{a + \frac{n}{2} + 1},$$

provided $a' > 1$ (mode exists for $a' > 0$ but this formula holds for usual parameter ranges). Thus one convenient 0–1 Bayes estimate is

$$\hat{\sigma}^2_{(\text{MAP})} = \frac{b + \frac{1}{2}S}{a + \frac{n}{2} + 1}.$$

—

$$\sigma^2 \mid X \sim \text{Inv-Gamma}\Big(a + \tfrac{n}{2},\ b + \tfrac{1}{2}\sum_{i=1}^{n}(X_i - \theta)^2\Big),$$

$$\hat{\sigma}^2_{(\text{MSE})} = \frac{b + \frac{1}{2}S}{a + \frac{n}{2} - 1}, \qquad \hat{\sigma}^2_{(\text{MAE})} = \text{posterior median}\,, \qquad \hat{\sigma}^2_{(\text{MAP})} = \frac{b + \frac{1}{2}S}{a + \frac{n}{2} + 1}.$$

❖ 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.

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

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

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

$$r(f, a_1) = \mathbb{E}_\theta R(\theta, a_1) = -350 \qquad r(f, a_2) = \mathbb{E}_\theta R(\theta, a_2) = -300$$

■

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*

$$\bar{R}(a_1) = \sup_\theta R(\theta, a_1) = \max\{R(\theta_1.a_1), R(\theta_2, a_1)\} = 1000$$

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

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

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

maximum risk is

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

Definition *Minimax Rule*

A decision rule $\hat{\theta}_{MM}$ is minimax if it minimizes the maximum risk:

$$\hat{\theta}_{MM} = \arg\min_{\hat{\theta} \in \mathcal{A}} \bar{R}(\hat{\theta}) = \arg\min_{\hat{\theta} \in \mathcal{A}} \sup_{\theta} R(\theta, \hat{\theta})$$

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 θ.

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:

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

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 Ω

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 δ 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 $\mathcal{A}$ has randomized estimators.

In this setting, a prior of θ 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

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

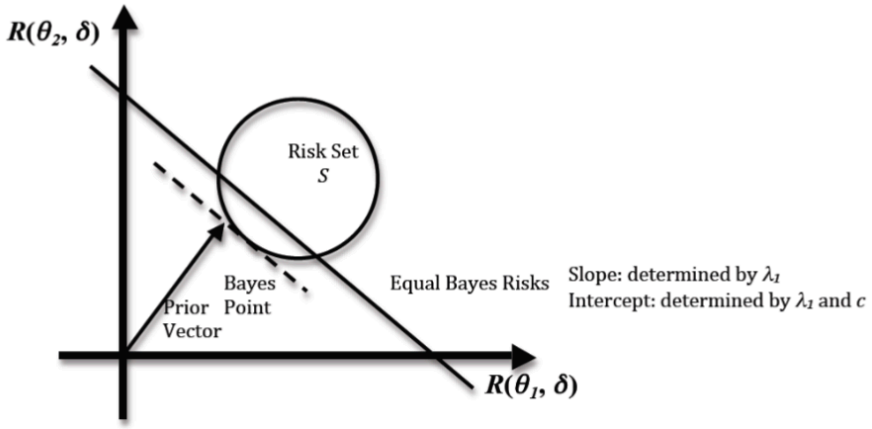

Figure 4: Geometry of a Bayes Point for $k = 2$

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

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

❖ Lecture 18

18.1 Minimax

In general it can be difficult to find minimax rules when the parameter space is infinite. One connection to Bayes rules is:

Theorem 18.1 (Bayes–Minimax Connection)**.** Suppose there exists a prior density $f(\theta)$ such that the Bayes rule $\hat{\theta}_f$ has constant risk; that is,

$$R(\theta, \hat{\theta}_f) = c \qquad \text{for all } \theta \in \Theta.$$

Then $\hat{\theta}_f$ is minimax with minimax risk c.

Proof. Let δ be any decision rule. Its Bayes risk under prior f is

$$r_f(\delta) = \int_\Theta R(\theta, \delta)\, f(\theta)\, d\theta.$$

Since $\hat{\theta}_f$ is the Bayes rule under f, we have

$$r_f(\hat{\theta}_f) \le r_f(\delta) \qquad \text{for all rules } \delta.$$

Because the risk of $\hat{\theta}_f$ is constant equal to c, its Bayes risk is

$$r_f(\hat{\theta}_f) = \int_\Theta R(\theta, \hat{\theta}_f) f(\theta)\, d\theta = \int_\Theta c\, f(\theta)\, d\theta = c.$$

Now take any rule δ.

$$r_f(\delta) = \int_\Theta R(\theta, \delta) f(\theta)\, d\theta \le \sup_{\theta' \in \Theta} R(\theta', \delta).$$

Combining with $c = r_f(\hat{\theta}_f) \le r_f(\delta)$ yields

$$c \le r_f(\delta) \quad \Rightarrow \quad c \le \sup_{\theta \in \Theta} R(\theta, \delta).$$

while for $\hat{\theta}_f$,

$$\sup_{\theta \in \Theta} R(\theta, \hat{\theta}_f) = c.$$

Hence $\hat{\theta}_f$ achieves the smallest possible maximum risk. Therefore, $\hat{\theta}_f$ is minimax with minimax risk c. □

Example

Let $X \mid p \sim \mathrm{Bin}(n, p)$ and consider squared–error loss.

1. Risk of $\hat{p} = X/n$:

$$R(p, \hat{p}) = \mathbb{E}_p\left[(X/n - p)^2\right] = \frac{1}{n^2}\,\mathrm{Var}_p(X) = \frac{1}{n^2}\,np(1-p) = \frac{p(1-p)}{n}.$$

This depends on p; its supremum over $p \in [0, 1]$ is attained at $p = 1/2$ and equals

$$\sup_p R(p, \hat{p}) = \frac{1}{4n}.$$

2. Randomized estimator showing $\hat{p}$ is not minimax Define the randomized estimator $\tilde{p}$ by

$$\tilde{p} = \begin{cases} X/n & \text{with probability } 1 - \dfrac{1}{n+1}, \\ 1/2 & \text{with probability } \dfrac{1}{n+1}, \end{cases}$$

(independent of X). Its risk equals the mixture of risks:

$$\begin{aligned} R(p, \tilde{p}) &= \left(1 - \frac{1}{n+1}\right) R(p, \hat{p}) + \frac{1}{n+1}\,\mathbb{E}_p\left[(1/2 - p)^2\right] \\ &= \frac{n}{n+1} \cdot \frac{p(1-p)}{n} + \frac{1}{n+1}(p - 1/2)^2 \\ &= \frac{1}{n+1}\Big(p(1-p) + (p - 1/2)^2\Big). \end{aligned}$$

But

$$p(1-p) + (p - 1/2)^2 = (p - p^2) + (p^2 - p + 1/4) = \tfrac{1}{4},$$

so

$$R(p, \tilde{p}) = \frac{1}{n+1} \cdot \frac{1}{4} = \frac{1}{4(n+1)} \quad \text{for all } p.$$

Hence $\tilde{p}$ has constant (and strictly smaller) maximum risk than $\hat{p}$:

$$\sup_p R(p, \tilde{p}) = \frac{1}{4(n+1)} < \frac{1}{4n} = \sup_p R(p, \hat{p}).$$

Therefore $\hat{p} = X/n$ is not minimax.

Example *cont'd*

3. Bayes estimator under $\mathrm{Beta}(a,b)$ prior and choice of (a,b) making risk constant

Let the prior be $p \sim \mathrm{Beta}(a,b)$. Given X, the posterior is

$$p \mid X \sim \mathrm{Beta}(a+X,\ b+n-X),$$

so the Bayes estimator under squared–error loss (posterior mean) is

$$\delta(X) = \mathbb{E}[p \mid X] = \frac{a+X}{a+b+n}.$$

Write $A = a+b$. Then we can write

$$\delta(X) = \frac{a+X}{A+n}.$$

Compute its risk $R(p) = \mathbb{E}_p\left[(\delta(X)-p)^2\right]$. Since δ is affine in X,

$$\mathbb{E}_p[\delta(X)] = \frac{a+np}{A+n},$$
$$\mathrm{Var}_p(\delta(X)) = \frac{1}{(A+n)^2}\,\mathrm{Var}_p(X) = \frac{np(1-p)}{(A+n)^2}.$$

Thus

$$\begin{aligned} R(p) &= \mathrm{Var}_p(\delta(X)) + \big(\mathbb{E}_p[\delta(X)] - p\big)^2 \\ &= \frac{np(1-p)}{(A+n)^2} + \frac{(a-pA)^2}{(A+n)^2} \\ &= \frac{1}{(A+n)^2}\Big(np(1-p) + (a-pA)^2\Big). \end{aligned}$$

$$np(1-p) + (a-pA)^2 = p^2(A^2-n) + p\,(n-2aA) + a^2.$$

For $R(p)$ to be constant (independent of p)

$$\begin{cases} A^2 - n = 0, \\ n - 2aA = 0. \end{cases}$$

$$\Longrightarrow\ a = b = \frac{\sqrt{n}}{2} \qquad (\text{i.e. } p \sim \mathrm{Beta}(\tfrac{\sqrt{n}}{2}, \tfrac{\sqrt{n}}{2}))$$

makes the frequentist risk of the Bayes estimator constant in p.

$$np(1-p) + (a-pA)^2 = n\Big(p(1-p) + (p-1/2)^2\Big) = \frac{n}{4},$$
$$R(p) = \frac{n/4}{(A+n)^2} = \frac{1}{4}\cdot\frac{n}{(n+\sqrt{n})^2} = \frac{1}{4\big(1+n^{-1/2}\big)^2}.$$

Because the Bayes estimator has constant frequentist risk, it is minimax.

❖ Review 1

19.1 Final Exam

Date: Tuesday, December 17, 2025, 8:00am – 11:00am

19.2 Questions

Review Question 1. Suppose we take a random sample of size n from a population of people. Let X_1 denote the number of individuals with a particular genotype AA, X_2 denote the number with Aa, and X_3 denote the number with aa. Assuming the gene frequencies are in equilibrium, the Hardy-Weinberg law says that the genotypes AA, Aa, and aa occur with probability:

$$p_1 = (1-\theta)^2, \quad p_2 = 2\theta(1-\theta), \quad p_3 = \theta^2$$

1. What is the likelihood function for θ, treating $X = (X_1, X_2, X_3)$ as a sample from the multinomial distribution with size n and $p = (p_1, p_2, p_3)$?
2. Find the maximum likelihood estimator (MLE) for θ under this model. Find the asymptotic distribution (after appropriate normalization) for the MLE.
3. Construct the likelihood ratio test statistic for the null hypothesis that $p_1 = (1-\theta)^2$, $p_2 = 2\theta(1-\theta)$, and $p_3 = \theta^2$. What is the asymptotic distribution of your test statistic under the null?

Solution.

1. The likelihood function is given by

$$L(\theta) = \frac{n!}{X_1!X_2!X_3!} p_1^{X_1} p_2^{X_2} p_3^{X_3} = \frac{n!}{X_1!X_2!X_3!} (1-\theta)^{2X_1} [2\theta(1-\theta)]^{X_2} (\theta^2)^{X_3}.$$

2. To find the MLE, we can maximize the log-likelihood:

$$\ell(\theta) = \log L(\theta) = \text{constant} + 2X_1 \log(1-\theta) + X_2 \log(2\theta(1-\theta)) + 2X_3 \log(\theta).$$

Taking the derivative with respect to θ and setting it to zero gives:

$$\frac{d\ell}{d\theta} = -\frac{2X_1}{1-\theta} + \frac{X_2}{\theta} - \frac{X_2}{1-\theta} + \frac{2X_3}{\theta} = 0.$$

Solving this equation yields the MLE $\hat{\theta}$. The asymptotic distribution of the MLE can be derived using the Fisher information.

$$\implies \hat{\theta} = \frac{X_2 + 2X_3}{2n}$$

And the second derivative is

$$\frac{d^2\ell}{d\theta^2} = -\frac{2X_1}{(1-\theta)^2} - \frac{X_2}{\theta^2} - \frac{X_2}{(1-\theta)^2} - \frac{2X_3}{\theta^2} < 0$$

3. The likelihood ratio test statistic is given by

$$\Lambda = 2\log\left(\frac{L(\hat{\theta})}{L(\hat{\theta}_0)}\right),$$

where $\hat{\theta}_0$ is the MLE under the null hypothesis. Under the null hypothesis, the asymptotic distribution of Λ follows a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters estimated under the null and alternative hypotheses.

The denominator has 1 free parameter (theta), while the numerator has 2 free parameters (p1, p2, p3 with one constraint that they sum to 1), so the degrees of freedom is 1.

Denominator:

$$L(\hat{\theta}_0) = \frac{n!}{X_1!X_2!X_3!}(1-\hat{\theta}_0)^{2X_1}[2\hat{\theta}_0(1-\hat{\theta}_0)]^{X_2}(\hat{\theta}_0^2)^{X_3}.$$

where

$$\hat{\theta}_0 = \frac{X_2 + 2X_3}{2n}$$

Numerator:

$$L(\hat{\theta}) = \frac{n!}{X_1!X_2!X_3!}\left(\frac{X_1}{n}\right)^{X_1}\left(\frac{X_2}{n}\right)^{X_2}\left(\frac{X_3}{n}\right)^{X_3}.$$

■

Review Question 2. Assume that we want to integrate $r(x)$ using a Monte Carlo Integration approach.

- How to select the density distribution $g(x)$ to sample $X_1, \ldots, X_B$?
- Can you assess the variance of your Monte Carlo integration result?

Remember that in Monte Carlo integration, we are finding the average of r/g for a number of sampled points. Hence if g is small for a given sample point, r/g will be arbitrarily large and can skew the sample mean from the true mean unless we generate a LARGE sample.

One way to avoid such cases is to use g that looks like r. The peaks and valleys of g should correspond to peaks and valleys of r.

Solution.

$$\int_\Omega r(x)\,dx = \int_\Omega \frac{r(x)}{g(x)} g(x)\,dx = \mathbb{E}_g\left[\frac{r(X)}{g(X)}\right] \approx \frac{1}{B}\sum_{i=1}^{B} \frac{r(X_i)}{g(X_i)}$$

Ideally, we want

1. $g(x)$ has the same support as $r(x)$, i.e., $g(x) = 0 \Leftrightarrow r(x) = 0$.
2. Approximate shape of $g(x)$ to be similar to $r(x)$.
3. $g(x)$ is easy to sample from and $\dfrac{r(x)}{g(x)}$ easy to compute.

$$\mathrm{Var}_{x\sim g}\left(\frac{1}{B}\sum_{i=1}^{B}\frac{r(x_i)}{g(x_i)}\right) = \frac{\mathrm{Var}_{x\sim g}\dfrac{r(x)}{g(x)}}{B}$$

which could be estimated by

$$\frac{1}{B^2}\sum_{i=1}^{B}\left(\frac{r(x_i)}{g(x_i)} - \frac{1}{B}\sum_{j=1}^{B}\frac{r(x_j)}{g(x_j)}\right)^2$$

■

Review Question 3. Suppose $X_1, \ldots, X_n$ are i.i.d. Poisson(λ).
Consider $S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$.

1. Is S^2 an unbiased estimator for λ? Justify your answer.
2. Does S^2 have the smallest variance among all unbiased estimators for λ? If not, please find the unbiased estimator (for λ) that has the smallest variance. Justify your answer.

Solution.

1. Yes. We have

$$\mathbb{E}[S^2] = \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right] = \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathbb{E}[X_i^2] - n\mathbb{E}[\overline{X}^2]\right).$$

   Since $X_i \sim \text{Poisson}(\lambda)$, we have $\mathbb{E}[X_i^2] = \lambda + \lambda^2$ and $\mathbb{E}[\overline{X}^2] = \frac{\lambda}{n} + \lambda^2$. Plugging these in, we get

$$\mathbb{E}[S^2] = \frac{1}{n-1}\left(n(\lambda + \lambda^2) - n\left(\frac{\lambda}{n} + \lambda^2\right)\right) = \lambda.$$

   Thus, S^2 is an unbiased estimator for λ.

2. No. Method 1(CR bound): The sample mean $\overline{X}$ is also an unbiased estimator for λ with variance $\frac{\lambda}{n}$. By the Cramér-Rao lower bound, the variance of any unbiased estimator for λ cannot be lower than that of $\overline{X}$. Since S^2 has a larger variance than $\overline{X}$, it does not have the smallest variance among all unbiased estimators for λ. Therefore, the unbiased estimator with the smallest variance is $\overline{X}$.

   Method 2(Sufficiency and Completeness): We define

$$\tilde{\lambda} = \mathbb{E}(S^2 \mid \sum_{i=1}^{n} X_i) \qquad \mathbb{E}(\tilde{\lambda}) = \mathbb{E}(S^2) = \lambda$$

   By Rao-Blackwell theorem, $\tilde{\lambda}$ has smaller variance than S^2. Since $\sum_{i=1}^{n} X_i$ is a complete sufficient statistic for λ (Poisson belongs to exponential family),

$$R(\lambda, \tilde{\lambda}) < R(\lambda, S^2) \implies \text{Var}(\tilde{\lambda}) < \text{Var}(S^2)$$

■

Review Question 4. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. with $X_i \sim N(0, 1)$ and $Y_i|(X_i = x) \sim N(x\theta, 1)$.

1. Is the above an exponential family?
2. Find the Fisher information $I(\theta)$.
3. Find an unbiased estimator of θ.

Solution.

1. Yes. The joint density function of (X_i, Y_i) is given by

$$f(\mathbf{x}, \mathbf{y}|\theta) = \prod_{i=1}^{n} f(y_i|x_i, \theta) f(x_i)$$

Substituting the given distributions $X_i \sim N(0,1)$ and $Y_i|X_i \sim N(x_i\theta, 1)$:

$$\begin{aligned} f(\mathbf{x}, \mathbf{y}|\theta) &= \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_i - x_i\theta)^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_i^2} \right) \\ &= (2\pi)^{-n} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} x_i^2 \right) \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (y_i^2 - 2x_i y_i \theta + x_i^2 \theta^2) \right) \\ &= \underbrace{(2\pi)^{-n} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (x_i^2 + y_i^2) \right)}_{h(\mathbf{x},\mathbf{y})} \cdot \exp\left(\theta \sum_{i=1}^{n} x_i y_i - \frac{\theta^2}{2} \sum_{i=1}^{n} x_i^2 \right) \end{aligned}$$

This takes the form of a (curved) exponential family with sufficient statistics $T_1 = \sum X_i Y_i, \eta_1 = \theta$ and $T_2 = \sum X_i^2, \eta_2 = -\frac{\theta^2}{2}$.

2. To find the Fisher information $I_n(\theta)$, we use the log-likelihood function derived from the joint density above:

$$\ell(\theta) = \log f(\mathbf{x}, \mathbf{y}|\theta) \propto -\frac{1}{2} \sum_{i=1}^{n} (y_i - x_i\theta)^2$$

The first derivative (score function) is:

$$\frac{\partial \ell}{\partial \theta} = -\frac{1}{2} \sum_{i=1}^{n} 2(y_i - x_i\theta)(-x_i) = \sum_{i=1}^{n} (x_i y_i - x_i^2 \theta)$$

The second derivative is:

$$\frac{\partial^2 \ell}{\partial \theta^2} = \sum_{i=1}^{n} (-x_i^2) = -\sum_{i=1}^{n} x_i^2$$

The Fisher Information is the negative expectation of the second derivative:

$$I_n(\theta) = -E\left[\frac{\partial^2 \ell}{\partial \theta^2} \right] = E\left[\sum_{i=1}^{n} X_i^2 \right] = \sum_{i=1}^{n} E[X_i^2]$$

Since $X_i \sim N(0,1)$, $E[X_i^2] = \text{Var}(X_i) + (E[X_i])^2 = 1$. Thus:

$$I_n(\theta) = n$$

3. We can use the Maximum Likelihood Estimator (MLE). Setting the score function to 0:

$$\sum_{i=1}^{n}(x_i y_i - x_i^2\hat{\theta}) = 0 \implies \hat{\theta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$

To check if it is unbiased, we use the law of iterated expectations, conditioning on X:

$$E[\hat{\theta}] = E_X\left[E_{Y|X}\left[\frac{\sum X_i Y_i}{\sum X_i^2}\right]\right] = E_X\left[\frac{1}{\sum X_i^2}\sum_{i=1}^{n} X_i E[Y_i|X_i]\right]$$

Since $E[Y_i|X_i] = X_i\theta$:

$$E[\hat{\theta}] = E_X\left[\frac{1}{\sum X_i^2}\sum_{i=1}^{n} X_i(X_i\theta)\right] = E_X\left[\frac{\theta\sum X_i^2}{\sum X_i^2}\right] = E_X[\theta] = \theta$$

Thus, $\hat{\theta} = \frac{\sum X_i Y_i}{\sum X_i^2}$ is an unbiased estimator of θ.

■

19.3 Application of Likelihood Ratio Test

Back to Peason's Test

1. $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$.

   For example, $X_1, \ldots, X_k \sim Multinomial(n, p_1, \ldots, p_k)$,
   $H_0 : p_1 = g_1(\theta), \ldots, p_k = g_k(\theta)$ Note that $\sum_{j=1}^{k} g_j(\theta) = 1$ H_1 : not null.

   Without parameter constraints, MLE is $\hat{p}_j = \dfrac{Y_j}{n}$, where Y_j is the observed count in category j.

2. Independent Test for multiple Populations

	Population 1	Population 2	...	Population m
Category 1	Y_{11}	Y_{12}	...	Y_{1m}
Category 2	Y_{21}	Y_{22}	...	Y_{2m}
...	...	...	...	...
Category k	Y_{k1}	Y_{k2}	...	Y_{km}

   Convert to probabilities table:

	Population 1	Population 2	...	Population m	
Category 1	p_{11}	p_{12}	...	p_{1m}	$p_{1\cdot}$
Category 2	p_{21}	p_{22}	...	p_{2m}	$p_{2\cdot}$
...	...	...	...	...	...
Category k	p_{k1}	p_{k2}	...	p_{km}	$p_{k\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	...	$p_{\cdot m}$	1

$$p_{ij} = P(\text{Category } i \text{ in Population } j)$$

$$H_0 : \text{independent} \quad i.e. p_{ij} = p_{i\cdot}p_{\cdot j}, \qquad H_a : \text{ not null}$$

Test statistic:

Under H_0, MLE is $\hat{p}_{ij} = p_{i\cdot}p_{\cdot j} = \frac{Y_{i\cdot}}{n}\frac{Y_{\cdot j}}{n}$, where $Y_{i\cdot} = \sum_{j=1}^{m} Y_{ij}$ and $Y_{\cdot j} = \sum_{i=1}^{k} Y_{ij}$.

Under H_a, MLE is $\hat{p}_{ij} = \frac{Y_{ij}}{n}$.

Thus,

$$T = 2\log\frac{L(\hat{p}_{ij})}{L(\hat{p}_{i\cdot}\hat{p}_{\cdot j})} = 2\sum_{i=1}^{k}\sum_{j=1}^{m} Y_{ij}\log\frac{Y_{ij}}{np_{i\cdot}p_{\cdot j}} \xrightarrow{D} \chi^2_{(k-1)(m-1)}$$

Rejection Sampling

Geometry Decision Theory

References

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

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