# 10-705/36-705 Intermediate Statistics

Larry Wasserman
http://www.stat.cmu.edu/~larry/=stat705/

Fall 2011

| Week | Class I | Class II | Day III | Class IV |
|---|---|---|---|---|
| | **Syllabus** | | | |
| August 29 | Review | Review, Inequalities | | Inequalities |
| September 5 | No Class | O_P | HW 1 [soln] | VC Theory |
| September 12 | Convergence | Convergence | HW 2 [soln] | Test I |
| September 19 | Convergence Addendum | Sufficiency | HW 3 [soln] | Sufficiency |
| September 26 | Likelihood | Point Estimation | HW 4 [soln] | Minimax Theory |
| October 3 | Minimax Summary | Asymptotics | HW 5 [soln] | Asymptotics |
| October 10 | Asymptotics | Review | | Test II |
| October 17 | Testing | Testing | HW 6 [soln] | Mid-semester Break |
| October 24 | Testing | Confidence Intervals | HW 7 [soln] | Confidence Intervals |
| October 31 | Nonparametric | Nonparametric | | Review |
| November 7 | Test III | No Class | HW 8 [soln] | The Bootstrap |
| November 14 | The Bootstrap | Bayesian Inference | HW 9 [soln] | Bayesian Inference |
| November 21 | No Class | No Class | | No Class |
| November 28 | Prediction | Prediction | HW 10 [soln] | Model Selection |
| December 5 | Multiple Testing | Causation | | Individual Sequences |
| | **Practice Final** | | | |

# 10-705/36-705: Intermediate Statistics, Fall 2010


| | |
|---|---|
| Professor | Larry Wasserman |
| Office | Baker Hall 228 A |
| Email | larry@stat.cmu.edu |
| Phone | 268-8727 |
| Office hours | Mondays, 1:30-2:30 |
| Class Time | Mon-Wed-Fri 12:30 - 1:20 |
| Location | GHC 4307 |
| TAs | Wanjie Wang and Xiaolin Yang |


**Website** http://www.stat.cmu.edu/∼larry/=stat705

## Objective

This course will cover the fundamentals of theoretical statistics. Topics include: point and interval estimation, hypothesis testing, data reduction, convergence concepts, Bayesian inference, nonparametric statistics and bootstrap resampling. We will cover Chapters 5 – 10 from Casella and Berger plus some supplementary material. This course is excellent preparation for advanced work in Statistics and Machine Learning.

## Textbook

Casella, G. and Berger, R. L. (2002). *Statistical Inference, 2nd ed.*

## Background

I assume that you are familiar with the material in Chapters 1 - 4 of Casella and Berger.

## Other Recommended Texts

Wasserman, L. (2004). *All of Statistics: A concise course in statistical inference.*
Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics.*
Rice, J. A. (1977). *Mathematical Statistics and Data Analysis, Second Edition.*

## Grading

10% : Test I (Sept. 16) on the material of Chapters 1–4
20% : Test II (October 14)
20% : Test III (November 7)
30% : Final Exam (Date set by the University)
20% : Homework

**Exams**
All exams are closed book. **Do NOT buy a plane ticket until the final exam has been scheduled.**

**Homework**
Homework assigments will be posted on the web. Hand in homework to Mari Alice Mcshane, 228 Baker Hall by **3 pm Thursday**. **No late homework**.

**Reading and Class Notes**
Class notes will be posted on the web regularly. **Bring a copy to class.** The notes are not meant to be a substitute for the book and hence are generally quite terse. Read both the notes and the text before lecture. Sometimes I will cover topics from other sources.

**Group Work**
You are encouraged to work with others on the homework. But write-up your final solutions on your own.

## Course Outline

1. Quick Review of Chapters 1-4
2. Inequalities
3. Vapnik-Chervonenkis Theory
4. Convergence
5. Sufficiency
6. Likelihood
7. Point Estimation
8. Minimax Theory
9. Asymptotics
10. Robustness
11. Hypothesis Testing
12. Confidence Intervals
13. Nonparametric Inference
14. Prediction and Classification
15. The Bootstrap
16. Bayesian Inference
17. Markov Chain Monte Carlo
18. Model Selection

# Lecture Notes 1
# Quick Review of Basic Probability
# (Casella and Berger Chapters 1-4)

## 1 Probability Review

Chapters 1-4 are a review. **I will assume you have read and understood Chapters 1-4.** Let us recall some of the key ideas.

### 1.1 Random Variables

A *random variable* is a map $X$ from a probability space $\Omega$ to $\mathbb{R}$. We write

$$P(X \in A) = P(\{\omega \in \Omega : \ X(\omega) \in A\})$$

and we write $X \sim P$ to mean that $X$ has distribution $P$. The *cumulative distribution function (cdf)* of $X$ is

$$F_X(x) = F(x) = P(X \leq x).$$

If $X$ is discrete, its *probability mass function (pmf)* is

$$p_X(x) = p(x) = P(X = x).$$

If $X$ is continuous, then its *probability density function function (pdf)* satisfies

$$P(X \in A) = \int_A p_X(x)dx = \int_A p(x)dx$$

and $p_X(x) = p(x) = F'(x)$. The following are all equivalent:

$$X \sim P, \quad X \sim F, \quad X \sim p.$$

Suppose that $X \sim P$ and $Y \sim Q$. We say that $X$ and $Y$ have the same distribution if

$$P(X \in A) = Q(Y \in A)$$

for all $A$. In other words, $P = Q$. **In that case we say that $X$ and $Y$ are *equal in distribution* and we write $X \stackrel{d}{=} Y$. It can be shown that $X \stackrel{d}{=} Y$ if and only if $F_X(t) = F_Y(t)$ for all $t$.**

### 1.2 Expected Values

The *mean* or expected value of $g(X)$ is

$$\mathbb{E}\left(g(X)\right) = \int g(x)dF(x) = \int g(x)dP(x) = \begin{cases} \int_{-\infty}^{\infty} g(x)p(x)dx & \text{if } X \text{ is continuous} \\ \sum_j g(x_j)p(x_j) & \text{if } X \text{ is discrete.} \end{cases}$$

Recall that:

1. $\mathbb{E}(\sum_{j=1}^{k} c_j g_j(X)) = \sum_{j=1}^{k} c_j \mathbb{E}(g_j(X))$.
2. If $X_1, \ldots, X_n$ are independent then

$$\mathbb{E}\left(\prod_{i=1}^{n} X_i\right) = \prod_i \mathbb{E}\left(X_i\right).$$

3. We often write $\mu = \mathbb{E}(X)$.
4. $\sigma^2 = \mathsf{Var}\left(X\right) = \mathbb{E}\left((X-\mu)^2\right)$ is the **Variance**.
5. $\mathsf{Var}\left(X\right) = \mathbb{E}\left(X^2\right) - \mu^2$.
6. If $X_1, \ldots, X_n$ are independent then

$$\mathsf{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_i a_i^2 \mathsf{Var}\left(X_i\right).$$

7. The covariance is

$$\mathsf{Cov}(X,Y) = \mathbb{E}((X-\mu_x)(Y-\mu_y)) = \mathbb{E}(XY) - \mu_X \mu_Y$$

and the correlation is $\rho(X,Y) = \mathsf{Cov}(X,Y)/\sigma_x\sigma_y$. Recall that $-1 \leq \rho(X,Y) \leq 1$.

The conditional expectation of $Y$ given $X$ is the random variable $\mathbb{E}(Y|X)$ whose value, when $X = x$ is

$$\mathbb{E}(Y|X=x) = \int y\ p(y|x)dy$$

where $p(y|x) = p(x,y)/p(x)$. The *Law of Total Expectation* or *Law of Iterated Expectation*:

$$\mathbb{E}(Y) = \mathbb{E}\big[\mathbb{E}(Y|X)\big] = \int \mathbb{E}(Y|X=x)p_X(x)dx.$$

The *Law of Total Variance* is

$$\mathsf{Var}(Y) = \mathsf{Var}\big[\mathbb{E}(Y|X)\big] + \mathbb{E}\big[\mathsf{Var}(Y|X)\big].$$

The $n^{\text{th}}$ moment is $\mathbb{E}\left(X^n\right)$ and the $n^{\text{th}}$ central moment is $\mathbb{E}\left((X-\mu)^n\right)$. The *moment generating function (mgf)* is

$$M_X(t) = \mathbb{E}\left(e^{tX}\right).$$

Then, $M_X^{(n)}(t)|_{t=0} = \mathbb{E}\left(X^n\right)$.

**If $M_X(t) = M_Y(t)$ for all $t$ in an interval around 0 then $X \stackrel{d}{=} Y$.**

## 1.3 Exponential Families

A family of distributions $\{p(x;\theta):\ \theta \in \Theta\}$ is called an *exponential family* if

$$p(x;\theta) = h(x)c(\theta)\exp\left\{\sum_{i=1}^{k} w_i(\theta)t_i(x)\right\}.$$

**Example 1** *$X \sim \text{Poisson}(\lambda)$ is exponential family since*

$$p(x) = P(X=x) \quad = \quad \frac{e^{-\lambda}\lambda^x}{x!} = \frac{1}{x!}e^{-\lambda}\exp\{\log\lambda \cdot x\}.$$

**Example 2** *$X \sim U(0,\theta)$ is not an exponential family. The density is*

$$p_X(x) = \frac{1}{\theta}I_{(0,\theta)}(x)$$

*where $I_A(x) = 1$ if $x \in A$ and 0 otherwise.*

We can rewrite an exponential family in terms of a *natural parameterization.* For $k=1$ we have

$$p(x;\eta) = h(x)\exp\{\eta t(x) - A(\eta)\}$$

where

$$A(\eta) = \log\int h(x)\exp\{\eta t(x)\}dx.$$

For example a Poisson can be written as

$$p(x;\eta) = \exp\{\eta x - e^{\eta}\}/x!$$

where the natural parameter is $\eta = \log\lambda$.

Let $X$ have an exponential family distribution. Then

$$\mathbb{E}(t(X)) = A'(\eta), \quad \mathsf{Var}(t(X)) = A''(\eta).$$

**Practice Problem: Prove the above result.**

## 1.4 Transformations

Let $Y = g(X)$. Then

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(g(X) \le y) = \int_{A(y)} p_X(x)dx$$

where

$$A_y = \{x:\ g(x) \le y\}.$$

Then $p_Y(y) = F_Y'(y)$.

If $g$ is monotonic, then

$$p_Y(y) = p_X(h(y))\left|\frac{dh(y)}{dy}\right|$$

where $h = g^{-1}$.

**Example 3** *Let $p_X(x) = e^{-x}$ for $x > 0$. Hence $F_X(x) = 1 - e^{-x}$. Let $Y = g(X) = \log X$. Then*

$$\begin{aligned} F_Y(y) = P(Y \leq y) &= P(\log(X) \leq y) \\ &= P(X \leq e^y) = F_X(e^y) = 1 - e^{-e^y} \end{aligned}$$

*and $p_Y(y) = e^y e^{-e^y}$ for $y \in \mathbb{R}$.*

**Example 4 Practice problem.** *Let $X$ be uniform on $(-1, 2)$ and let $Y = X^2$. Find the density of $Y$.*

Let $Z = g(X, Y)$. For exampe, $Z = X + Y$ or $Z = X/Y$. Then we find the pdf of $Z$ as follows:

1. For each $z$, find the set $A_z = \{(x, y) : g(x, y) \leq z\}$.
2. Find the CDF

$$F_Z(z) = P(Z \leq z) = P(g(X, Y) \leq z) = P(\{(x, y) : g(x, y) \leq z\}) = \int\int_{A_z} p_{X,Y}(x, y)dxdy.$$

3. The pdf is $p_Z(z) = F'_Z(z)$.

**Example 5 Practice problem.** *Let $(X, Y)$ be uniform on the unit square. Let $Z = X/Y$. Find the density of $Z$.*

## 1.5 Independence

Recall that $X$ and $Y$ are *independent* if and only if

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

for all $A$ and $B$.

**Theorem 6** *Let $(X, Y)$ be a bivariate random vector with $p_{X,Y}(x, y)$. $X$ and $Y$ are independent iff $p_{X,Y}(x, y) = p_X(x)p_Y(y)$.*

$X_1, \ldots, X_n$ are independent if and only if

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i).$$

Thus, $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^n p_{X_i}(x_i)$.

If $X_1, \ldots, X_n$ are independent and identically distributed we say they are iid (or that they are a random sample) and we write

$$X_1, \ldots, X_n \sim P \quad \text{or} \quad X_1, \ldots, X_n \sim F \quad \text{or} \quad X_1, \ldots, X_n \sim p.$$

## 1.6 Important Distributions

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

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}.$$

If $X \in \mathbb{R}^d$ then $X \sim N(\mu, \Sigma)$ if

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right).$$

$X \sim \chi^2_p$ if $X = \sum_{j=1}^p Z_j^2$ where $Z_1, \ldots, Z_p \sim N(0,1)$.

$X \sim \text{Bernoulli}(\theta)$ if $\mathbb{P}(X=1) = \theta$ and $\mathbb{P}(X=0) = 1-\theta$ and hence

$$p(x) = \theta^x (1-\theta)^{1-x} \qquad x = 0, 1.$$

$X \sim \text{Binomial}(\theta)$ if

$$p(x) = \mathbb{P}(X=x) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \qquad x \in \{0, \ldots, n\}.$$

$X \sim \text{Uniform}(0, \theta)$ if $p(x) = I(0 \leq x \leq \theta)/\theta$.

## 1.7 Sample Mean and Variance

The sample mean is

$$\overline{X} = \frac{1}{n} \sum_i X_i,$$

and the sample variance is

$$S^2 = \frac{1}{n-1} \sum_i (X_i - \overline{X})^2.$$

Let $X_1, \ldots, X_n$ be iid with $\mu = \mathbb{E}(X_i) = \mu$ and $\sigma^2 = \mathsf{Var}(X_i) = \sigma^2$. Then

$$\mathbb{E}(\overline{X}) = \mu, \quad \mathsf{Var}(\overline{X}) = \frac{\sigma^2}{n}, \quad \mathbb{E}(S^2) = \sigma^2.$$

**Theorem 7** *If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ then*

*(a)* $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

*(b)* $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

*(c)* $\overline{X}$ *and* $S^2$ *are independent*

## 1.8 Delta Method

If $X \sim N(\mu, \sigma^2)$, $Y = g(X)$ and $\sigma^2$ is small then

$$Y \approx N(g(\mu), \sigma^2(g'(\mu))^2).$$

To see this, note that

$$Y = g(X) = g(\mu) + (X - \mu)g'(\mu) + \frac{(X-\mu)^2}{2}g''(\xi)$$

for some $\xi$. Now $\mathbb{E}((X-\mu)^2) = \sigma^2$ which we are assuming is small and so

$$Y = g(X) \approx g(\mu) + (X - \mu)g'(\mu).$$

Thus

$$\mathbb{E}(Y) \approx g(\mu), \quad \mathsf{Var}(Y) \approx (g'(\mu))^2\sigma^2.$$

Hence,

$$g(X) \approx N\left(g(\mu),\ (g'(\mu))^2\sigma^2\right).$$

---

# Appendix: Useful Facts

## Facts about sums

- $\sum_{i=1}^n i = \frac{n(n+1)}{2}$.
- $\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$.
- Geometric series: $a + ar + ar^2 + \ldots = \frac{a}{1-r}$, for $0 < r < 1$.
- Partial Geometric series $a + ar + ar^2 + \ldots + ar^{n-1} = \frac{a(1-r^n)}{1-r}$.
- Binomial Theorem

$$\sum_{x=0}^{n} \binom{n}{x} a^x = (1+a)^n, \quad \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x} = (a+b)^n.$$

- Hypergeometric identity

$$\sum_{x=0}^{\infty} \binom{a}{x}\binom{b}{n-x} = \binom{a+b}{n}.$$

# Common Distributions

## Discrete

### Uniform

- $X \sim U(1, \ldots, N)$
- $X$ takes values $x = 1, 2, \ldots, N$
- $P(X = x) = 1/N$
- $\mathbb{E}(X) = \sum_x xP(X = x) = \sum_x x\frac{1}{N} = \frac{1}{N}\frac{N(N+1)}{2} = \frac{(N+1)}{2}$
- $\mathbb{E}(X^2) = \sum_x x^2P(X = x) = \sum_x x^2\frac{1}{N} = \frac{1}{N}\frac{N(N+1)(2N+1)}{6}$

### Binomial

- $X \sim \text{Bin}(n, p)$
- $X$ takes values $x = 0, 1, \ldots, n$
- $P(X = x) = \binom{n}{x}p^x(1-p)^{n-x}$

### Hypergeometric

- $X \sim \text{Hypergeometric}(N, M, K)$
- $P(X = x) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}$

### Geometric

- $X \sim \text{Geom}(p)$
- $P(X = x) = (1-p)^{x-1}p$, $x = 1, 2, \ldots$
- $\mathbb{E}(X) = \sum_x x(1-p)^{x-1} = p\sum_x \frac{d}{dp}\left(-(1-p)^x\right) = p\frac{p}{p^2} = \frac{1}{p}$.

### Poisson

- $X \sim \text{Poisson}(\lambda)$
- $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$ $x = 0, 1, 2, \ldots$
- $\mathbb{E}(X) = \mathsf{Var}(X) = \lambda$
- $M_X(t) = \sum_{x=0}^{\infty} e^{tx}\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{\left(\lambda e^t\right)^x}{x!} = e^{-\lambda}e^{\lambda e^t} = e^{\lambda(e^t-1)}$.

- $\mathbb{E}(X) = \lambda e^t e^{\lambda(e^t-1)}|_{t=0} = \lambda$.
- Use mgf to show: if $X_1 \sim \text{Poisson}(\lambda_1)$, $X_2 \sim \text{Poisson}(\lambda_2)$, independent then $Y = X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

## Continuous Distributions

### Normal

- $X \sim N(\mu, \sigma^2)$
- $p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{\frac{-1}{2\sigma^2}(x-\mu)^2\}$, $x \in \mathcal{R}$
- mgf $M_X(t) = \exp\{\mu t + \sigma^2 t^2/2\}$.
- $E(X) = \mu$
- $\mathsf{Var}(X) = \sigma^2$.
- e.g., If $Z \sim N(0,1)$ and $X = \mu + \sigma Z$, then $X \sim N(\mu, \sigma^2)$. Show this...

**Proof.**

$$\begin{aligned} M_X(t) &= E\left(e^{tX}\right) = E\left(e^{t(\mu+\sigma Z)}\right) = e^{t\mu} E\left(e^{t\sigma Z}\right) \\ &= e^{t\mu} M_Z(t\sigma) = e^{t\mu} e^{(t\sigma)^2/2} = e^{t\mu + t^2\sigma^2/2} \end{aligned}$$

which is the mgf of a $N(\mu, \sigma^2)$.

Alternative proof:

$$\begin{aligned} F_X(x) &= P(X \le x) = P(\mu + \sigma Z \le x) = P\left(Z \le \frac{x-\mu}{\sigma}\right) \\ &= F_Z\left(\frac{x-\mu}{\sigma}\right) \\ p_X(x) &= F_X'(x) = p_Z\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}\frac{1}{\sigma} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \end{aligned}$$

which is the pdf of a $N(\mu, \sigma^2)$. □

**Gamma**

- $X \sim \Gamma(\alpha, \beta)$.
- $p_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$, $x$ a positive real.
- $\Gamma(\alpha) = \int_0^\infty \frac{1}{\beta^\alpha} x^{\alpha-1} e^{-x/\beta} dx$.
- Important statistical distribution: $\chi^2_p = \Gamma(\frac{p}{2}, 2)$.
- $\chi^2_p = \sum_{i=1}^p X_i^2$, where $X_i \sim N(0,1)$, iid.

**Exponential**

- $X \sim exponen(\beta)$
- $p_X(x) = \frac{1}{\beta} e^{-x/\beta}$, $x$ a positive real.
- $exponen(\beta) = \Gamma(1, \beta)$.
- e.g., Used to model waiting time of a Poisson Process. Suppose $N$ is the number of phone calls in 1 hour and $N \sim Poisson(\lambda)$. Let $T$ be the time between consecutive phone calls, then $T \sim exponen(1/\lambda)$ and $E(T) = (1/\lambda)$.
- If $X_1, \ldots, X_n$ are iid $exponen(\beta)$, then $\sum_i X_i \sim \Gamma(n, \beta)$.
- Memoryless Property: If $X \sim exponen(\beta)$, then

$$P(X > t + s | X > t) = P(X > s).$$

## Linear Regression

Model the response ($Y$) as a linear function of the parameters and covariates ($x$) plus random error ($\epsilon$).

$$Y_i = \theta(x, \beta) + \epsilon_i$$

where

$$\theta(x, \beta) = X\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k.$$

## Generalized Linear Model

Model the natural parameters as linear functions of the the covariates.

**Example: Logistic Regression**.

$$P(Y = 1|X = x) = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}}.$$

In other words, $Y|X = x \sim Bin(n, p(x))$ and

$$\eta(x) = \beta^T x$$

where

$$\eta(x) = \log\left(\frac{p(x)}{1 - p(x)}\right).$$

Logistic Regression consists of modelling the natural parameter, which is called the log odds ratio, as a linear function of covariates.

# Location and Scale Families, CB 3.5

Let $p(x)$ be a pdf.

$$\text{Location family}: \ \{p(x|\mu) = p(x - \mu): \ \mu \in \mathbb{R}\}$$

$$\text{Scale family}: \ \left\{p(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right): \ \sigma > 0\right\}$$

$$\text{Location} - \text{Scale family}: \ \left\{p(x|\mu, \sigma) = \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right): \ \mu \in \mathbb{R}, \sigma > 0\right\}$$

**(1) Location family**. Shifts the pdf.

e.g., Uniform with $p(x) = 1$ on $(0, 1)$ and $p(x - \theta) = 1$ on $(\theta, \theta + 1)$.
e.g., Normal with standard pdf the density of a $N(0, 1)$ and location family pdf $N(\theta, 1)$.
**(2) Scale family**. Stretches the pdf.

e.g., Normal with standard pdf the density of a $N(0, 1)$ and scale family pdf $N(0, \sigma^2)$.
**(3) Location-Scale family**. Stretches and shifts the pdf.

e.g., Normal with standard pdf the density of a $N(0, 1)$ and location-scale family pdf $N(\theta, \sigma^2)$, i.e., $\frac{1}{\sigma} p(\frac{x-\mu}{\sigma})$ .

# Multinomial Distribution

The multivariate version of a Binomial is called a Multinomial. Consider drawing a ball from an urn with has balls with $k$ different colors labeled "color 1, color 2, ..., color $k$." Let $p = (p_1, p_2, \ldots, p_k)$ where $\sum_j p_j = 1$ and $p_j$ is the probability of drawing color $j$. Draw $n$ balls from the urn (independently and with replacement) and let $X = (X_1, X_2, \ldots, X_k)$ be the count of the number of balls of each color drawn. We say that $X$ has a Multinomial $(n, p)$ distribution. The pdf is

$$p(x) = \binom{n}{x_1, \ldots, x_k} p_1^{x_1} \ldots p_k^{x_k}.$$

# Multivariate Normal Distribution

We now define the multivariate normal distribution and derive its basic properties. We want to allow the possibility of multivariate normal distributions whose covariance matrix is not necessarily positive definite. Therefore, we cannot define the distribution by its density function. Instead we define the distribution by its moment generating function. (The reader may wonder how a random vector can have a moment generating function if it has no density function. However, the moment generating function can be defined using more general types of integration. In this book, we assume that such a definition is possible but find the moment generating function by elementary means.) We find the density function for the case of positive definite covariance matrix in Theorem 5.

**Lemma 8** *(a). Let* $\mathbf{X} = \mathbf{AY} + \mathbf{b}$ *Then*

$$M_X(t) = \exp(\mathbf{b}'\mathbf{t}) M_Y(\mathbf{A}'\mathbf{t}).$$

*(b). Let $c$ be a constant. Let* $\mathbf{Z} = c\mathbf{Y}$*. Then*

$$M_Z(\mathbf{t}) = M_Y(c\mathbf{t}).$$

*(c). Let*

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} \mathbf{t}_1 \\ \mathbf{t}_2 \end{pmatrix}$$

*Then*

$$M_{\mathbf{Y}_1}(\mathbf{t}_1) = M_{\mathbf{Y}} \begin{pmatrix} \mathbf{t}_1 \\ 0 \end{pmatrix}.$$

*(d).* $\mathbf{Y}_1$ *and* $\mathbf{Y}_2$ *are independent if and only if*

$$M_{\mathbf{Y}} \begin{pmatrix} \mathbf{t}_1 \\ \mathbf{t}_2 \end{pmatrix} = M_{\mathbf{Y}} \begin{pmatrix} \mathbf{t}_1 \\ 0 \end{pmatrix} M_{\mathbf{Y}} \begin{pmatrix} 0 \\ \mathbf{t}_2 \end{pmatrix}.$$

We start with $Z_1, \ldots, Z_n$ independent random variables such that $Z_i \sim N_1(0,1)$. Let $\mathbf{Z} = (Z_1, \ldots, Z_n)'$. Then

$$E(\mathbf{Z}) = 0, \quad \text{cov}(\mathbf{Z}) = \mathbf{I}, \quad M_{\mathbf{Z}}(\mathbf{t}) = \prod \exp \frac{t_i^2}{2} = \exp \frac{\mathbf{t}'\mathbf{t}}{2}. \tag{1}$$

Let $\mu$ be an $n \times 1$ vector and $\mathbf{A}$ an $n \times n$ matrix. Let $\mathbf{Y} = \mathbf{AZ} + \mu$. Then

$$E(\mathbf{Y}) = \mu \quad \text{cov}(\mathbf{Y}) = \mathbf{AA}'. \tag{2}$$

Let $\Sigma = \mathbf{AA}'$. We now show that the distribution of $\mathbf{Y}$ depends only on $\mu$ and $\Sigma$. The moment generating function $M_{\mathbf{Y}}(\mathbf{t})$ is given by

$$M_{\mathbf{Y}}(\mathbf{t}) = \exp(\mu'\mathbf{t}) M_{\mathbf{Z}}(\mathbf{A}'\mathbf{t}) = \exp\left(\mu'\mathbf{t} + \frac{\mathbf{t}'(\mathbf{A}'\mathbf{A})\mathbf{t}}{2}\right) = \exp\left(\mu'\mathbf{t} + \frac{\mathbf{t}'\Sigma\mathbf{t}}{2}\right).$$

With this motivation in mind, let $\mu$ be an $n \times 1$ vector, and let $\Sigma$ be a nonnegative definite $n \times n$ matrix. Then we say that the $n$-dimensional random vector $\mathbf{Y}$ has an *n-dimensional normal distribution* with mean vector $\mu$, and covariance matrix $\Sigma$, if $\mathbf{Y}$ has moment generating function

$$M_{\mathbf{Y}}(\mathbf{t}) = \exp\left(\mu'\mathbf{t} + \frac{\mathbf{t}'\Sigma\mathbf{t}}{2}\right). \tag{3}$$

We write $\mathbf{Y} \sim N_n(\mu, \Sigma)$. The following theorem summarizes some elementary facts about multivariate normal distributions.

**Theorem 9** *(a). If $\mathbf{Y} \sim N_n(\mu, \Sigma)$, then $E(\mathbf{Y}) = \mu$, cov$(\mathbf{Y}) = \Sigma$.*

*(b). If $\mathbf{Y} \sim N_n(\mu, \Sigma)$, $c$ is a scalar, then $c\mathbf{Y} \sim N_n(c\mu, c^2\Sigma)$.*

*(c). Let $\mathbf{Y} \sim N_n(\mu, \Sigma)$. If $\mathbf{A}$ is $p \times n$, $\mathbf{b}$ is $p \times 1$, then $A\mathbf{Y} + \mathbf{b} \sim N_p(\mathbf{A}\mu + \mathbf{b}, \mathbf{A}\Sigma\mathbf{A}')$.*

*(d). Let $\mu$ be any $n \times 1$ vector, and let $\Sigma$ be any $n \times n$ nonnegative definite matrix. Then there exists $\mathbf{Y}$ such that $\mathbf{Y} \sim N_n(\mu, \Sigma)$.*

**Proof.** (a). This follows directly from (2) above.

(b) and (c). Homework.

(d). Let $Z_1, \ldots, Z_n$ be independent, $Z_i \sim N(0,1)$. Let $\mathbf{Z} = (Z_1, \ldots, Z_n)'$. It is easily verified that $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbf{I})$. Let $\mathbf{Y} = \Sigma^{1/2}\mathbf{Z} + \mu$. By part b, above,

$$\mathbf{Y} \sim N_n(\Sigma^{1/2}0 + \mu, \Sigma).$$

□

We have now shown that the family of normal distributions is preserved under linear operations on the random vectors. We now show that it is preserved under taking marginal and conditional distributions.

**Theorem 10** *Suppose that* $\mathbf{Y} \sim N_n(\mu, \Sigma)$. *Let*

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11}\, \Sigma_{12} \\ \Sigma_{21}\, \Sigma_{22} \end{pmatrix}.$$

*where* $\mathbf{Y}_1$ *and* $\mu_1$ *are* $p \times 1$, *and* $\Sigma_{11}$ *is* $p \times p$.

*(a).* $\mathbf{Y}_1 \sim N_p(\mu_1, \Sigma_{11}), \mathbf{Y}_2 \sim N_{n-p}(\mu_2, \Sigma_{22})$.

*(b).* $\mathbf{Y}_1$ *and* $\mathbf{Y}_2$ *are independent if and only if* $\Sigma_{12} = 0$.

*(c). If* $\Sigma_{22} > 0$, *then the condition distribution of* $\mathbf{Y}_1$ *given* $\mathbf{Y}_2$ *is*

$$\mathbf{Y}_1 | \mathbf{Y}_2 \sim N_p(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{Y}_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

**Proof.** (a). Let $\mathbf{t}' = (\mathbf{t}_1', \mathbf{t}_2')$ where $\mathbf{t}_1$ is $p \times 1$. The joint moment generating function of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ is

$$M_{\mathbf{Y}}(\mathbf{t}) = \exp(\mu_1'\mathbf{t}_1 + \mu_2'\mathbf{t}_2 + \frac{1}{2}(\mathbf{t}_1'\Sigma_{11}\mathbf{t}_1 + \mathbf{t}_1'\Sigma_{12}\mathbf{t}_2 + \mathbf{t}_2'\Sigma_{21}\mathbf{t}_1 + \mathbf{t}_2'\Sigma_{22}\mathbf{t}_2)).$$

Therefore,

$$M_{\mathbf{Y}}\begin{pmatrix} \mathbf{t}_1 \\ 0 \end{pmatrix} = \exp(\mu_1'\mathbf{t}_1 + \frac{1}{2}\mathbf{t}_1'\Sigma_{11}\mathbf{t}_1), \quad M_{\mathbf{Y}}\begin{pmatrix} 0 \\ \mathbf{t}_2 \end{pmatrix} = \exp(\mu_2'\mathbf{t}_2 + \frac{1}{2}\mathbf{t}_2'\Sigma_{22}\mathbf{t}_2).$$

By Lemma 1c, we see that $\mathbf{Y}_1 \sim N_p(\mu_1, \Sigma_{11}), \mathbf{Y}_2 \sim N_{n-p}(\mu_2, \Sigma_{22})$.

(b). We note that

$$M_{\mathbf{Y}}(\mathbf{t}) = M_{\mathbf{Y}}\begin{pmatrix} \mathbf{t}_1 \\ 0 \end{pmatrix} M_{\mathbf{Y}}\begin{pmatrix} 0 \\ \mathbf{t}_2 \end{pmatrix}$$

if and only if

$$\mathbf{t}_1'\Sigma_{12}\mathbf{t}_2 + \mathbf{t}_2'\Sigma_{21}\mathbf{t}_1 = 0.$$

Since $\Sigma$ is symmetric and $\mathbf{t}_2'\Sigma_{21}\mathbf{t}_1$ is a scalar, we see that $\mathbf{t}_2'\Sigma_{21}\mathbf{t}_1 = \mathbf{t}_1'\Sigma_{12}\mathbf{t}_2$.
Finally, $\mathbf{t}'\Sigma_{12}\mathbf{t}_2 = 0$ for all $\mathbf{t}_1 \in R^p, \mathbf{t}_2 \in R^{n-p}$ if and only if $\Sigma_{12} = 0$, and the result follows from Lemma 1d.

(c). We first find the joint distribution of

$$\mathbf{X} = \mathbf{Y}_1 - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{Y}_2 \text{ and } \mathbf{Y}_2.$$

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{I} & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix}$$

Therefore, by Theorem 2c, the joint distribution of $\mathbf{X}$ and $\mathbf{Y}_2$ is

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y}_2 \end{pmatrix} \sim N_n\left(\begin{pmatrix} \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}\right)$$

and hence $\mathbf{X}$ and $\mathbf{Y}_2$ are independent. Therefore, the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}_2$ is the same as the marginal distribution of $\mathbf{X}$,

$$\mathbf{X} | \mathbf{Y}_2 \sim N_p(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

Since $\mathbf{Y}_2$ is just a constant in the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}_2$ we have, by Theorem 2c, that the conditional distribution of $\mathbf{Y}_1 = \mathbf{X} + \Sigma_{12}\Sigma_{22}^{-1}\mathbf{Y}_2$ given $\mathbf{Y}_2$ is

$$\mathbf{Y}_1|\mathbf{Y}_2 \sim N_p(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 + \Sigma_{12}\Sigma_{22}^{-1}\mathbf{Y}_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

Note that we need $\Sigma_{22} > 0$ in part c so that $\Sigma_{22}^{-1}$ exists. □

**Lemma 11** *Let $\mathbf{Y} \sim N_n(\mu, \sigma^2\mathbf{I})$, where $\mathbf{Y}' = (Y_1, \ldots, Y_n), \mu' = (\mu_1, \ldots, \mu_n)$ and $\sigma^2 > 0$ is a scalar. Then the $Y_i$ are independent, $Y_i \sim N_1(\mu, \sigma^2)$ and*

$$\frac{||\mathbf{Y}||^2}{\sigma^2} = \frac{\mathbf{Y}'\mathbf{Y}}{\sigma^2} \sim \chi_n^2\left(\frac{\mu'\mu}{\sigma^2}\right).$$

**Proof.** Let $Y_i$ be independent, $Y_i \sim N_1(\mu_i, \sigma^2)$. The joint moment generating function of the $Y_i$ is

$$M_{\mathbf{Y}}(t) = \prod_{i=1}^{n}(\exp(\mu_i t_i + \frac{1}{2}\sigma^2 t_i^2)) = \exp(\mu'\mathbf{t} + \frac{1}{2}\sigma^2\mathbf{t}'\mathbf{t})$$

which is the moment generating function of a random vector that is normally distributed with mean vector $\mu$ and covariance matrix $\sigma^2\mathbf{I}$. Finally, $\mathbf{Y}'\mathbf{Y} = \Sigma Y_i^2, \mu'\mu = \Sigma\mu_i^2$ and $Y_i/\sigma \sim N_1(\mu_i/\sigma, 1)$. Therefore $\mathbf{Y}'\mathbf{Y}/\sigma^2 \sim \chi_n^2(\mu'\mu/\sigma^2)$ by the definition of the noncentral $\chi^2$ distribution. □

We are now ready to derive the nonsingular normal density function.

**Theorem 12** *Let $\mathbf{Y} \sim N_n(\mu, \Sigma)$, with $\Sigma > 0$. Then $\mathbf{Y}$ has density function*

$$p_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mu)'\Sigma^{-1}(\mathbf{y} - \mu)\right).$$

**Proof.** We could derive this by finding the moment generating function of this density and showing that it satisfied (3). We would also have to show that this function is a density function. We can avoid all that by starting with a random vector whose distribution we know. Let

$$\mathbf{Z} \sim N_n(0, \mathbf{I}). \quad \mathbf{Z} = (Z_1, \ldots, Z_n)'.$$

Then the $Z_i$ are independent and $Z_i \sim N_1(0, 1)$, by Lemma 4. Therefore, the joint density of the $Z_i$ is

$$p_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2}z_i^2\right) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\mathbf{z}'\mathbf{z}\right).$$

Let $\mathbf{Y} = \Sigma^{1/2}\mathbf{Z} + \mu$. By Theorem 2c, $\mathbf{Y} \sim N_n(\mu, \Sigma)$. Also $\mathbf{Z} = \Sigma^{-1/2}(\mathbf{Y} - \mu)$, and the transformation from $\mathbf{Z}$ to $\mathbf{Y}$ is therefore invertible. Furthermore, the Jacobian of this inverse transformation is just $|\Sigma^{-1/2}| = |\Sigma|^{-1/2}$. Hence the density of $\mathbf{Y}$ is

$$\begin{aligned} p_{\mathbf{Y}}(\mathbf{y}) &= p_Z(\Sigma^{-1/2}(\mathbf{y} - \mu))\frac{1}{|\Sigma|^{1/2}} \\ &= \frac{1}{|\Sigma|^{1/2}(2\pi)^{n/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mu)'\Sigma^{-1}(\mathbf{y} - \mu)\right). \end{aligned}$$

□

We now prove a result that is useful later in the book and is also the basis for Pearson's $\chi^2$ tests.

**Theorem 13** *Let* $\mathbf{Y} \sim N_n(\mu, \Sigma), \Sigma > 0$. *Then*
*(a).* $\mathbf{Y}'\Sigma^{-1}\mathbf{Y} \sim \chi^2_n(\mu'\Sigma^{-1}\mu)$.
*(b).* $(\mathbf{Y}-\mu)'\Sigma^{-1}(\mathbf{Y}-\mu) \sim \chi^2_n(0)$.

**Proof.** (a). Let $\mathbf{Z} = \Sigma^{-1/2}\mathbf{Y} \sim N_n(\Sigma^{-1/2}\mu, \mathbf{I})$. By Lemma 4, we see that

$$\mathbf{Z}'\mathbf{Z} = \mathbf{Y}'\Sigma^{-1}\mathbf{Y} \sim \chi^2_n(\mu'\Sigma^{-1}\mu).$$

(b). Follows fairly directly. □

# The Spherical Normal

For the first part of this book, the most important class of multivariate normal distribution is the class in which

$$\mathbf{Y} \sim N_n(\mu, \sigma^2\mathbf{I}).$$

We now show that this distribution is spherically symmetric about $\mu$. A rotation about $\mu$ is given by $\mathbf{X} = \Gamma(\mathbf{Y}-\mu)+\mu$, where $\Gamma$ is an orthogonal matrix (i.e., $\Gamma\Gamma' = \mathbf{I}$). By Theorem 2, $\mathbf{X} \sim N_n(\mu, \sigma^2\mathbf{I})$, so that the distribution is unchanged under rotations about $\mu$. We therefore call this normal distribution the *spherical normal distribution*. If $\sigma^2 = 0$, then $P(\mathbf{Y} = \mu) = 1$. Otherwise its density function (by Theorem 4) is

$$p_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}||\mathbf{y}-\mu||^2\right).$$

By Lemma 4, we note that the components of $Y$ are independently normally distributed with common variance $\sigma^2$. In fact, the spherical normal distribution is the only multivariate distribution with independent components that is spherically symmetric.

# Lecture Notes 2

## 1 Probability Inequalities

Inequalities are useful for bounding quantities that might otherwise be hard to compute. They will also be used in the theory of convergence.

**Theorem 1 (The Gaussian Tail Inequality)** *Let $X \sim N(0,1)$. Then*

$$\mathbb{P}(|X| > \epsilon) \leq \frac{2e^{-\epsilon^2/2}}{\epsilon}.$$

*If $X_1, \ldots, X_n \sim N(0,1)$ then*

$$\mathbb{P}(|\overline{X}_n| > \epsilon) \leq \frac{1}{\sqrt{n}\epsilon} e^{-n\epsilon^2/2}.$$

**Proof.** The density of $X$ is $\phi(x) = (2\pi)^{-1/2} e^{-x^2/2}$. Hence,

$$\begin{aligned} \mathbb{P}(X > \epsilon) &= \int_\epsilon^\infty \phi(s)ds \leq \frac{1}{\epsilon}\int_\epsilon^\infty s\,\phi(s)ds \\ &= -\frac{1}{\epsilon}\int_\epsilon^\infty \phi'(s)ds = \frac{\phi(\epsilon)}{\epsilon} \leq \frac{e^{-\epsilon^2/2}}{\epsilon}. \end{aligned}$$

By symmetry,

$$\mathbb{P}(|X| > \epsilon) \leq \frac{2e^{-\epsilon^2/2}}{\epsilon}.$$

Now let $X_1, \ldots, X_n \sim N(0,1)$. Then $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i \sim N(0, 1/n)$. Thus, $\overline{X}_n \stackrel{d}{=} n^{-1/2} Z$ where $Z \sim N(0,1)$ and

$$\mathbb{P}(|\overline{X}_n| > \epsilon) = \mathbb{P}(n^{-1/2}|Z| > \epsilon) = \mathbb{P}(|Z| > \sqrt{n}\,\epsilon) \leq \frac{1}{\sqrt{n}\epsilon} e^{-n\epsilon^2/2}.$$

□

**Theorem 2 (Markov's inequality)** *Let $X$ be a non-negative random variable and suppose that $\mathbb{E}(X)$ exists. For any $t > 0$,*

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(X)}{t}. \tag{1}$$

**Proof.** Since $X > 0$,

$$\begin{aligned} \mathbb{E}(X) &= \int_0^\infty x\, p(x)dx = \int_0^t x\, p(x)dx + \int_t^\infty xp(x)dx \\ &\geq \int_t^\infty x\, p(x)dx \geq t \int_t^\infty p(x)dx = t\, \mathbb{P}(X > t). \end{aligned}$$

□

**Theorem 3 (Chebyshev's inequality)** *Let $\mu = \mathbb{E}(X)$ and $\sigma^2 = \mathsf{Var}(X)$. Then,*

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \quad \text{and} \quad \mathbb{P}(|Z| \geq k) \leq \frac{1}{k^2} \tag{2}$$

*where $Z = (X - \mu)/\sigma$. In particular, $\mathbb{P}(|Z| > 2) \leq 1/4$ and $\mathbb{P}(|Z| > 3) \leq 1/9$.*

**Proof.** We use Markov's inequality to conclude that

$$\mathbb{P}(|X - \mu| \geq t) = \mathbb{P}(|X - \mu|^2 \geq t^2) \leq \frac{\mathbb{E}(X - \mu)^2}{t^2} = \frac{\sigma^2}{t^2}.$$

The second part follows by setting $t = k\sigma$. □

If $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ then and $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$ Then, $\mathsf{Var}(\overline{X}_n) = \mathsf{Var}(X_1)/n = p(1-p)/n$ and

$$\mathbb{P}(|\overline{X}_n - p| > \epsilon) \leq \frac{\mathsf{Var}(\overline{X}_n)}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2}$$

since $p(1-p) \leq \frac{1}{4}$ for all $p$.

# 2 Hoeffding's Inequality

Hoeffding's inequality is similar in spirit to Markov's inequality but it is a sharper inequality. We begin with the following important result.

**Lemma 4** *Supppose that $\mathbb{E}(X) = 0$ and that $a \leq X \leq b$. Then*

$$\mathbb{E}(e^{tX}) \leq e^{t^2(b-a)^2/8}.$$

Recall that a function $g$ is **convex** if for each $x, y$ and each $\alpha \in [0,1]$,

$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y).$$

**Proof.** Since $a \leq X \leq b$, we can write $X$ as a convex combination of $a$ and $b$, namely, $X = \alpha b + (1-\alpha)a$ where $\alpha = (X-a)/(b-a)$. By the convexity of the function $y \to e^{ty}$ we have

$$e^{tX} \leq \alpha e^{tb} + (1-\alpha)e^{ta} = \frac{X-a}{b-a}e^{tb} + \frac{b-X}{b-a}e^{ta}.$$

Take expectations of both sides and use the fact that $\mathbb{E}(X) = 0$ to get

$$\mathbb{E}e^{tX} \leq -\frac{a}{b-a}e^{tb} + \frac{b}{b-a}e^{ta} = e^{g(u)} \tag{3}$$

where $u = t(b-a)$, $g(u) = -\gamma u + \log(1-\gamma+\gamma e^u)$ and $\gamma = -a/(b-a)$. Note that $g(0) = g'(0) = 0$. Also, $g''(u) \leq 1/4$ for all $u > 0$. By Taylor's theorem, there is a $\xi \in (0,u)$ such that

$$g(u) \;=\; g(0) + ug'(0) + \frac{u^2}{2}g''(\xi) = \frac{u^2}{2}g''(\xi) \leq \frac{u^2}{8} = \frac{t^2(b-a)^2}{8}.$$

Hence, $\mathbb{E}e^{tX} \leq e^{g(u)} \leq e^{t^2(b-a)^2/8}$. □

Next, we need to use *Chernoff's method*.

**Lemma 5** *Let $X$ be a random variable. Then*

$$\mathbb{P}(X > \epsilon) \leq \inf_{t \geq 0} e^{-t\epsilon}\,\mathbb{E}(e^{tX}).$$

**Proof.** For any $t > 0$,

$$\mathbb{P}(X > \epsilon) = \mathbb{P}(e^X > e^\epsilon) = \mathbb{P}(e^{tX} > e^{t\epsilon}) \leq e^{-t\epsilon}\mathbb{E}(e^{tX}).$$

Since this is true for every $t \geq 0$, the result follows. □

**Theorem 6 (Hoeffding's Inequality)** *Let $Y_1, \ldots, Y_n$ be iid observations such that $\mathbb{E}(Y_i) = \mu$ and $a \leq Y_i \leq b$ where $a < 0 < b$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}\left(|\overline{Y}_n - \mu| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2/(b-a)^2}. \tag{4}$$

**Proof.** Without los of generality, we asume that $\mu = 0$. First we have

$$\begin{aligned}\mathbb{P}(|\overline{Y}_n| \geq \epsilon) &= \mathbb{P}(\overline{Y}_n \geq \epsilon) + \mathbb{P}(\overline{Y}_n \leq -\epsilon) \\ &= \mathbb{P}(\overline{Y}_n \geq \epsilon) + \mathbb{P}(-\overline{Y}_n \geq \epsilon).\end{aligned}$$

Next we use Chernoff's method. For any $t > 0$, we have, from Markov's inequality, that

$$\begin{aligned}
\mathbb{P}(\overline{Y}_n \geq \epsilon) &= \mathbb{P}\left(\sum_{i=1}^n Y_i \geq n\epsilon\right) = \mathbb{P}\left(e^{\sum_{i=1}^n Y_i} \geq e^{n\epsilon}\right) \\
&= \mathbb{P}\left(e^{t\sum_{i=1}^n Y_i} \geq e^{tn\epsilon}\right) \leq e^{-tn\epsilon}\mathbb{E}\left(e^{t\sum_{i=1}^n Y_i}\right) \\
&= e^{-tn\epsilon}\prod_i \mathbb{E}(e^{tY_i}) = e^{-tn\epsilon}(\mathbb{E}(e^{tY_i}))^n.
\end{aligned}$$

From Lemma 4, $\mathbb{E}(e^{tY_i}) \leq e^{t^2(b-a)^2/8}$. So

$$\mathbb{P}(\overline{Y}_n \geq \epsilon) \leq e^{-tn\epsilon}e^{t^2 n(b-a)^2/8}.$$

This is minimized by setting $t = 4\epsilon/(b-a)^2$ giving

$$\mathbb{P}(\overline{Y}_n \geq \epsilon) \leq e^{-2n\epsilon^2/(b-a)^2}.$$

Applying the same argument to $\mathbb{P}(-\overline{Y}_n \geq \epsilon)$ yields the result. □

**Example 7** *Let $X_1, \ldots, X_n \sim$ Bernoulli$(p)$. Chebyshev's inequality yields*

$$\mathbb{P}(|\overline{X}_n - p| > \epsilon) \leq \frac{1}{4n\epsilon^2}.$$

*According to Hoeffding's inequality,*

$$\mathbb{P}(|\overline{X}_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

*which decreases much faster.*

**Corollary 8** *If $X_1, X_2, \ldots, X_n$ are independent with $\mathbb{P}(a \leq X_i \leq b) = 1$ and common mean $\mu$, then, with probability at least $1 - \delta$,*

$$|\overline{X}_n - \mu| \leq \sqrt{\frac{c}{2n}\log\left(\frac{2}{\delta}\right)} \tag{5}$$

*where $c = (b-a)^2$.*

# 3 The Bounded Difference Inequality

So far we have focused on sums of random variables. The following result extends Hoeffding's inequality to more general functions $g(x_1, \ldots, x_n)$. Here we consider McDiarmid's inequality, also known as the Bounded Difference inequality.

**Theorem 9 (McDiarmid)** *Let $X_1, \ldots, X_n$ be independent random variables. Suppose that*

$$\sup_{x_1,\ldots,x_n,x_i'} \left| g(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n) - g(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n) \right| \leq c_i \tag{6}$$

*for $i = 1, \ldots, n$. Then*

$$\mathbb{P}\left( g(X_1,\ldots,X_n) - \mathbb{E}(g(X_1,\ldots,X_n)) \geq \epsilon \right) \leq \exp\left\{ -\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2} \right\}. \tag{7}$$

**Proof.** Let $V_i = \mathbb{E}(g|X_1,\ldots,X_i) - \mathbb{E}(g|X_1,\ldots,X_{i-1})$. Then $g(X_1,\ldots,X_n) - \mathbb{E}(g(X_1,\ldots,X_n)) = \sum_{i=1}^n V_i$ and $\mathbb{E}(V_i|X_1,\ldots,X_{i-1}) = 0$. Using a similar argument as in Hoeffding's Lemma we have,

$$\mathbb{E}(e^{tV_i}|X_1,\ldots,X_{i-1}) \leq e^{t^2 c_i^2/8}. \tag{8}$$

Now, for any $t > 0$,

$$\begin{aligned}
\mathbb{P}\left(g(X_1,\ldots,X_n) - \mathbb{E}(g(X_1,\ldots,X_n)) \geq \epsilon\right) &= \mathbb{P}\left(\sum_{i=1}^n V_i \geq \epsilon\right) \\
&= \mathbb{P}\left(e^{t\sum_{i=1}^n V_i} \geq e^{t\epsilon}\right) \leq e^{-t\epsilon}\mathbb{E}\left(e^{t\sum_{i=1}^n V_i}\right) \\
&= e^{-t\epsilon}\mathbb{E}\left(e^{t\sum_{i=1}^{n-1} V_i}\mathbb{E}\left(e^{tV_n} \;\middle|\; X_1,\ldots,X_{n-1}\right)\right) \\
&\leq e^{-t\epsilon}e^{t^2c_n^2/8}\mathbb{E}\left(e^{t\sum_{i=1}^{n-1} V_i}\right) \\
&\vdots \\
&\leq e^{-t\epsilon}e^{t^2\sum_{i=1}^n c_i^2}.
\end{aligned}$$

The result follows by taking $t = 4\epsilon/\sum_{i=1}^n c_i^2$. □

**Example 10** *If we take $g(x_1,\ldots,x_n) = n^{-1}\sum_{i=1}^n x_i$ then we get back Hoeffding's inequality.*

**Example 11** *Suppose we throw $m$ balls into $n$ bins. What fraction of bins are empty? Let $Z$ be the number of empty bins and let $F = Z/n$ be the fraction of empty bins. We can write $Z = \sum_{i=1}^n Z_i$ where $Z_i = 1$ of bin $i$ is empty and $Z_i = 0$ otherwise. Then*

$$\mu = \mathbb{E}(Z) = \sum_{i=1}^n \mathbb{E}(Z_i) = n(1-1/n)^m = ne^{m\log(1-1/n)} \approx ne^{-m/n}$$

*and $\theta = \mathbb{E}(F) = \mu/n \approx e^{-m/n}$. How close is $Z$ to $\mu$? Note that the $Z_i$'s are not independent so we cannot just apply Hoeffding. Instead, we proceed as follows.*

*Define variables $X_1, \ldots, X_m$ where $X_s = i$ if ball $s$ falls into bin $i$. Then $Z = g(X_1, \ldots, X_m)$. If we move one ball into a different bin, then $Z$ can change by at most 1. Hence, (6) holds with $c_i = 1$ and so*

$$\mathbb{P}(|Z - \mu| > t) \leq 2e^{-2t^2/m}.$$

*Recall that he fraction of empty bins is $F = Z/m$ with mean $\theta = \mu/n$. We have*

$$\mathbb{P}(|F - \theta| > t) = \mathbb{P}(|Z - \mu| > nt) \leq 2e^{-2n^2t^2/m}.$$

# 4 Bounds on Expected Values

**Theorem 12 (Cauchy-Schwartz inequality)** *If $X$ and $Y$ have finite variances then*

$$\mathbb{E}\,|XY| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}. \tag{9}$$

The Cauchy-Schwarz inequality can be written as

$$\mathrm{Cov}^2(X, Y) \leq \sigma_X^2 \sigma_Y^2.$$

Recall that a function $g$ is **convex** if for each $x, y$ and each $\alpha \in [0, 1]$,

$$g(\alpha x + (1 - \alpha)y) \leq \alpha g(x) + (1 - \alpha)g(y).$$

If $g$ is twice differentiable and $g''(x) \geq 0$ for all $x$, then $g$ is convex. It can be shown that if $g$ is convex, then $g$ lies above any line that touches $g$ at some point, called a tangent line. A function $g$ is **concave** if $-g$ is convex. Examples of convex functions are $g(x) = x^2$ and $g(x) = e^x$. Examples of concave functions are $g(x) = -x^2$ and $g(x) = \log x$.

**Theorem 13 (Jensen's inequality)** *If $g$ is convex, then*

$$\mathbb{E}g(X) \geq g(\mathbb{E}X). \tag{10}$$

*If $g$ is concave, then*

$$\mathbb{E}g(X) \leq g(\mathbb{E}X). \tag{11}$$

**Proof.** Let $L(x) = a + bx$ be a line, tangent to $g(x)$ at the point $\mathbb{E}(X)$. Since $g$ is convex, it lies above the line $L(x)$. So,

$$\mathbb{E}g(X) \geq \mathbb{E}L(X) = \mathbb{E}(a + bX) = a + b\mathbb{E}(X) = L(\mathbb{E}(X)) = g(\mathbb{E}X).$$

□

**Example 14** *From Jensen's inequality we see that $\mathbb{E}(X^2) \geq (\mathbb{E}X)^2$.*

**Example 15 (Kullback Leibler Distance)** *Define the Kullback-Leibler distance between two densities $p$ and $q$ by*

$$D(p,q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx.$$

*Note that $D(p,p) = 0$. We will use Jensen to show that $D(p,q) \geq 0$. Let $X \sim f$. Then*

$$-D(p,q) = \mathbb{E}\log\left(\frac{q(X)}{p(X)}\right) \leq \log \mathbb{E}\left(\frac{q(X)}{p(X)}\right) = \log \int p(x)\frac{q(x)}{p(x)}dx = \log \int q(x)dx = \log(1) = 0.$$

*So, $-D(p,q) \leq 0$ and hence $D(p,q) \geq 0$.*

**Example 16** *It follows from Jensen's inequality that 3 types of means can be ordered. Assume that $a_1, \ldots, a_n$ are positive numbers and define the arithmetic, geometric and harmonic means as*

$$\begin{aligned} a_A &= \frac{1}{n}(a_1 + \ldots + a_n) \\ a_G &= (a_1 \times \ldots \times a_n)^{1/n} \\ a_H &= \frac{1}{\frac{1}{n}(\frac{1}{a_1} + \ldots + \frac{1}{a_n})}. \end{aligned}$$

*Then $a_H \leq a_G \leq a_A$.*

Suppose we have an exponential bound on $\mathbb{P}(X_n > \epsilon)$. In that case we can bound $\mathbb{E}(X_n)$ as follows.

**Theorem 17** *Suppose that $X_n \geq 0$ and that for every $\epsilon > 0$,*

$$\mathbb{P}(X_n > \epsilon) \leq c_1 e^{-c_2 n \epsilon^2} \tag{12}$$

*for some $c_2 > 0$ and $c_1 > 1/e$. Then,*

$$\mathbb{E}(X_n) \leq \sqrt{\frac{C}{n}}. \tag{13}$$

*where $C = (1 + \log(c_1))/c_2$.*

**Proof.** Recall that for any nonnegative random variable $Y$, $\mathbb{E}(Y) = \int_0^\infty \mathbb{P}(Y \geq t)dt$. Hence, for any $a > 0$,

$$\mathbb{E}(X_n^2) = \int_0^\infty \mathbb{P}(X_n^2 \geq t)dt = \int_0^a \mathbb{P}(X_n^2 \geq t)dt + \int_a^\infty \mathbb{P}(X_n^2 \geq t)dt \leq a + \int_a^\infty \mathbb{P}(X_n^2 \geq t)dt.$$

Equation (12) implies that $\mathbb{P}(X_n > \sqrt{t}) \leq c_1 e^{-c_2 n t}$. Hence,

$$\mathbb{E}(X_n^2) \leq a + \int_a^\infty \mathbb{P}(X_n^2 \geq t)dt = a + \int_a^\infty \mathbb{P}(X_n \geq \sqrt{t})dt \leq a + c_1 \int_a^\infty e^{-c_2 n t}dt = a + \frac{c_1 e^{-c_2 n a}}{c_2\, n}.$$

Set $a = \log(c_1)/(nc_2)$ and conclude that

$$\mathbb{E}(X_n^2) \leq \frac{\log(c_1)}{nc_2} + \frac{1}{nc_2} = \frac{1+\log(c_1)}{nc_2}.$$

Finally, we have

$$\mathbb{E}(X_n) \leq \sqrt{\mathbb{E}(X_n^2)} \leq \sqrt{\frac{1+\log(c_1)}{nc_2}}.$$

□

Now we consider bounding the maximum of a set of random variables.

**Theorem 18** *Let $X_1, \ldots, X_n$ be random variables. Suppose there exists $\sigma > 0$ such that $\mathbb{E}(e^{tX_i}) \leq e^{t\sigma^2/2}$ for all $t > 0$. Then*

$$\mathbb{E}\left(\max_{1\leq i\leq n} X_i\right) \leq \sigma\sqrt{2\log n}. \tag{14}$$

**Proof.** By Jensen's inequality,

$$\begin{aligned}
\exp\left\{t\mathbb{E}\left(\max_{1\leq i\leq n} X_i\right)\right\} &\leq \mathbb{E}\left(\exp\left\{t\max_{1\leq i\leq n} X_i\right\}\right) \\
&= \mathbb{E}\left(\max_{1\leq i\leq n} \exp\{tX_i\}\right) \leq \sum_{i=1}^{n} \mathbb{E}\left(\exp\{tX_i\}\right) \leq ne^{t^2\sigma^2/2}.
\end{aligned}$$

Thus,

$$\mathbb{E}\left(\max_{1\leq i\leq n} X_i\right) \leq \frac{\log n}{t} + \frac{t\sigma^2}{2}.$$

The result follows by setting $t = \sqrt{2\log n}/\sigma$. □

# 5 $O_P$ and $o_P$

In statisics, probability and machine learning, we make use of $o_P$ and $O_P$ notation.

Recall first, that $a_n = o(1)$ means that $a_n \to 0$ as $n \to \infty$. $a_n = o(b_n)$ means that $a_n/b_n = o(1)$.

$a_n = O(1)$ means that $a_n$ is eventually bounded, that is, for all large $n$, $|a_n| \leq C$ for some $C > 0$. $a_n = O(b_n)$ means that $a_n/b_n = O(1)$.

We write $a_n \sim b_n$ if both $a_n/b_n$ and $b_n/a_n$ are eventually bounded. In computer sicence this s written as $a_n = \Theta(b_n)$ but we prefer using $a_n \sim b_n$ since, in statistics, $\Theta$ often denotes a parameter space.

Now we move on to the probabilistic versions. Say that $Y_n = o_P(1)$ if, for every $\epsilon > 0$,

$$\mathbb{P}(|Y_n| > \epsilon) \to 0.$$

Say that $Y_n = o_P(a_n)$ if, $Y_n/a_n = o_P(1)$.

Say that $Y_n = O_P(1)$ if, for every $\epsilon > 0$, there is a $C > 0$ such that

$$\mathbb{P}(|Y_n| > C) \leq \epsilon.$$

Say that $Y_n = O_P(a_n)$ if $Y_n/a_n = O_P(1)$.

Let's use Hoeffding's inequality to show that sample proportions are $O_P(1/\sqrt{n})$ within the the true mean. Let $Y_1, \ldots, Y_n$ be coin flips i.e. $Y_i \in \{0, 1\}$. Let $p = \mathbb{P}(Y_i = 1)$. Let

$$\widehat{p}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

We will show that: $\widehat{p}_n - p = o_P(1)$ and $\widehat{p}_n - p = O_P(1/\sqrt{n})$.

We have that

$$\mathbb{P}(|\widehat{p}_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2} \to 0$$

and so $\widehat{p}_n - p = o_P(1)$. Also,

$$\begin{aligned}
\mathbb{P}(\sqrt{n}|\widehat{p}_n - p| > C) &= \mathbb{P}\left(|\widehat{p}_n - p| > \frac{C}{\sqrt{n}}\right) \\
&\leq 2e^{-2C^2} < \delta
\end{aligned}$$

if we pick $C$ large enough. Hence, $\sqrt{n}(\widehat{p}_n - p) = O_P(1)$ and so

$$\widehat{p}_n - p = O_P\left(\frac{1}{\sqrt{n}}\right).$$

Now consider $m$ coins with probabilities $p_1, \ldots, p_m$. Then

$$\begin{aligned}
\mathbb{P}(\max_j |\widehat{p}_j - p_j| > \epsilon) &\leq \sum_{j=1}^{m} \mathbb{P}(|\widehat{p}_j - p_j| > \epsilon) \qquad \text{union bound} \\
&\leq \sum_{j=1}^{m} 2e^{-2n\epsilon^2} \quad \text{Hoeffding} \\
&= 2me^{-2n\epsilon^2} = 2\exp\left\{-(2n\epsilon^2 - \log m)\right\}.
\end{aligned}$$

Supose that $m \leq e^{n^\gamma}$ where $0 \leq \gamma < 1$. Then

$$\mathbb{P}(\max_j |\widehat{p}_j - p_j| > \epsilon) \leq 2\exp\left\{-(2n\epsilon^2 - n^\gamma)\right\} \to 0.$$

Hence,

$$\max_j |\widehat{p}_j - p_j| = o_P(1).$$

# Lecture Notes 3

## 1 Uniform Bounds

Recall that, if $X_1, \ldots, X_n \sim$ Bernoulli$(p)$ and $\widehat{p}_n = n^{-1}\sum_{i=1}^n X_i$ then, from Hoeffding's inequality,

$$\mathbb{P}(|\widehat{p}_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

Sometimes we want to say more than this.

**Example 1** *Suppose that $X_1, \ldots, X_n$ have cdf $F$. Let*

$$F_n(t) = \frac{1}{n}\sum_{i=1}^n I(X_i \leq t).$$

*We call $F_n$ the* **empirical cdf**. *How close is $F_n$ to $F$? That is, how big is $|F_n(t) - F(t)|$? From Hoeffding's inequality,*

$$\mathbb{P}(|F_n(t) - F(t)| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

*But that is only for one point $t$. How big is $\sup_t |F_n(t) - F(t)|$? We would like a bound of the form*

$$\mathbb{P}\left(\sup_t |F_n(t) - F(t)| > \epsilon\right) \leq \text{ something small.}$$

**Example 2** *Suppose that $X_1, \ldots, X_n \sim P$. Let*

$$P_n(A) = \frac{1}{n}\sum_{i=1}^n I(X_i \in A).$$

*How close is $P_n(A)$ to $P(A)$? That is, how big is $|P_n(A) - P(A)|$? From Hoeffding's inequality,*

$$\mathbb{P}(|P_n(A) - P(A)| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

*But that is only for one set $A$. How big is $\sup_{A\in\mathcal{A}} |P_n(A) - P(A)|$ for a class of sets $\mathcal{A}$? We would like a bound of the form*

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \leq \text{ something small.}$$

**Example 3** *(Classification.) Suppose we observe data $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $Y_i \in \{0, 1\}$. Let $(X, Y)$ be a new pair. Suppose we observe $X$. Now we want to predict $Y$. A classifier $h$ is a function $h(x)$ which takes values in $\{0, 1\}$. When we observe $X$ we predict $Y$ with $h(X)$. The classification error, or risk, is the probability of an error:*

$$R(h) = \mathbb{P}(Y \neq h(X)).$$

*The training error is the fraction of errors on the observed data* $(X_1, Y_1), \ldots, (X_n, Y_n)$*:*

$$\widehat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \neq h(X_i)).$$

*By Hoeffding's inequality,*

$$\mathbb{P}(|\widehat{R}(h) - R(h)| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

*How do we choose a classifier? One way is to start with a set of classifiers* $\mathcal{H}$*. Then we define* $\widehat{h}$ *to be the member of* $\mathcal{H}$ *that minimizes the training error. Thus*

$$\widehat{h} = \text{argmin}_{h\in\mathcal{H}}\widehat{R}(h).$$

*An example is the set of linear classifiers. Suppose that* $x \in \mathbb{R}^d$*. A linear classifier has the form* $h(x) = 1$ *of* $\beta^T x \geq 0$ *and* $h(x) = 0$ *of* $\beta^T x < 0$ *where* $\beta = (\beta_1, \ldots, \beta_d)^T$ *is a set of parameters.*

*Although* $\widehat{h}$ *minimizes* $\widehat{R}(h)$*, it does not minimize* $R(h)$*. Let* $h_*$ *minimize the true error* $R(h)$*. A fundamental question is: how close is* $R(\widehat{h})$ *to* $R(h_*)$*? We will see later than* $R(\widehat{h})$ *is close to* $R(h_*)$ *if* $\sup_h |\widehat{R}(h) - R(h)|$ *is small. So we want*

$$\mathbb{P}\left(\sup_h |\widehat{R}(h) - R(h)| > \epsilon\right) \leq \quad \text{something small.}$$

---

More generally, we can state out goal as follows. For any function $f$ define

$$P(f) = \int f(x)dP(x), \quad P_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

Let $\mathcal{F}$ be a set of functions. In our first example, each $f$ was of the form $f_t(x) = I(x \leq t)$ and $\mathcal{F} = \{f_t : \ t \in \mathbb{R}\}$.

We want to bound

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| > \epsilon\right).$$

We will see that the bounds we obtain have the form

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| > \epsilon\right) \leq c_1\kappa(\mathcal{F})e^{-c_2 n\epsilon^2}$$

where $c_1$ and $c_2$ are positive constants and $\kappa(\mathcal{F})$ is a measure of the size (or complexity) of the class $\mathcal{F}$.

Similarly, if $\mathcal{A}$ is a class of sets then we want a bound of the form

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \leq c_1\kappa(\mathcal{A})e^{-c_2 n\epsilon^2}$$

where $P_n(A) = n^{-1}\sum_{i=1}^{n} I(X_i \in A)$.

Bounds like these are called *uniform bonds* since they hold uniformly over a class of functions or over a class of sets.

## 2 Finite Classes

Let $\mathcal{F} = \{f_1, \ldots, f_N\}$. Suppose that

$$\max_{1 \leq j \leq N} \sup_x |f_j(x)| \leq B.$$

We will make use of the *union bound*. Recall that

$$\mathbb{P}\left(A_1 \bigcup \cdots \bigcup A_N\right) \leq \sum_{j=1}^{N} \mathbb{P}(A_j).$$

Let $A_j$ be the event that $|P_n(f_j) - P(f)| > \epsilon$. From Hoeffding's inequality, $\mathbb{P}(A_j) \leq 2e^{-n\epsilon^2/(2B^2)}$. Then

$$\begin{aligned} \mathbb{P}\left(\sup_{f \in \mathcal{F}} |P_n(f) - P(f)| > \epsilon\right) &= \mathbb{P}(A_1 \bigcup \cdots \bigcup A_N) \\ &\leq \sum_{j=1}^{N} \mathbb{P}(A_j) \leq \sum_{j=1}^{N} 2e^{-n\epsilon^2/(2B^2)} = 2Ne^{-n\epsilon^2/(2B^2)}. \end{aligned}$$

Thus we have shown that

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |P_n(f) - P(f)| > \epsilon\right) \leq 2\kappa e^{-n\epsilon^2/(2B^2)}$$

where $\kappa = |\mathcal{F}|$.

The same idea applies to classes of sets. Let $\mathcal{A} = \{A_1, \ldots, A_N\}$ be a finite collection of sets. By the same reasoning we have

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \leq 2\kappa e^{-n\epsilon^2/(2B^2)}$$

where $\kappa = |\mathcal{F}|$ and $P_n(A) = n^{-1} \sum_{i=1}^{n} I(X_i \in A)$.

To extend these ideas to infinite classes like $\mathcal{F} = \{f_t : \ t \in \mathbb{R}\}$ we need to introduce a few more concepts.

## 3 Shattering

Let $\mathcal{A}$ be a class of sets. Some examples are:

1. $\mathcal{A} = \{(-\infty, t] : \ t \in \mathbb{R}\}$.
2. $\mathcal{A} = \{(a, b) : \ a \leq b\}$.
3. $\mathcal{A} = \{(a, b) \cup (c, d) : \ a \leq b \leq c \leq d\}$.

4. $\mathcal{A}$ = all discs in $\mathbb{R}^d$.

5. $\mathcal{A}$ = all rectangles in $\mathbb{R}^d$.

6. $\mathcal{A}$ = all half-spaces in $\mathbb{R}^d = \{x : \ \beta^T x \geq 0\}$.

7. $\mathcal{A}$ = all convex sets in $\mathbb{R}^d$.

Let $F = \{x_1, \ldots, x_n\}$ be a finite set. Let $G$ be a subset of $F$. Say that $\mathcal{A}$ **picks out** $G$ if

$$A \cap F = G$$

for some $A \in \mathcal{A}$. For example, let $\mathcal{A} = \{(a, b) : \ a \leq b\}$. Suppose that $F = \{1, 2, 7, 8, 9\}$ and $G = \{2, 7\}$. Then $\mathcal{A}$ picks out $G$ since $A \cap F = G$ if we choose $A = (1.5, 7.5)$ for example.

Let $S(\mathcal{A}, F)$ be the number of these subsets picked out by $\mathcal{A}$. Of course $S(\mathcal{A}, F) \leq 2^n$.

**Example 4** *Let* $\mathcal{A} = \{(a, b) : \ a \leq b\}$ *and* $F = \{1, 2, 3\}$. *Then* $\mathcal{A}$ *can pick out:*

$$\emptyset, \ \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}.$$

*So* $s(\mathcal{A}, F) = 7$. *Note that* $7 < 8 = 2^3$. *If* $F = \{1, 6\}$ *then* $\mathcal{A}$ *can pick out:*

$$\emptyset, \ \{1\}, \{6\}, \{1, 6\}.$$

*In this case* $s(\mathcal{A}, F) = 4 = 2^2$.

We say that $F$ is **shattered** if $s(\mathcal{A}, F) = 2^n$ where $n$ is the number of points in $F$.

Let $\mathcal{F}_n$ denote all finite sets with $n$ elements.

Define the **shatter coefficient**

$$s_n(\mathcal{A}) = \sup_{F \in \mathcal{F}_n} s(\mathcal{A}, F).$$

Note that $s_n(\mathcal{A}) \leq 2^n$.

The following theorem is due to Vapnik and Chervonenis. The proof is beyond the scope of the course. (If you take 10-702/36-702 you will learn the proof.)

| Class $\mathcal{A}$ | VC dimension $V_{\mathcal{A}}$ |
|---|---|
| $\mathcal{A} = \{A_1, \ldots, A_N\}$ | $\leq \log_2 N$ |
| Intervals $[a, b]$ on the real line | 2 |
| Discs in $\mathbb{R}^2$ | 3 |
| Closed balls in $\mathbb{R}^d$ | $\leq d + 2$ |
| Rectangles in $\mathbb{R}^d$ | $2d$ |
| Half-spaces in $\mathbb{R}^d$ | $d + 1$ |
| Convex polygons in $\mathcal{R}^2$ | $\infty$ |
| Convex polygons with $d$ vertices | $2d + 1$ |

Table 1: The VC dimension of some classes $\mathcal{A}$.

**Theorem 5** *Let $\mathcal{A}$ be a class of sets. Then*

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \leq 8\ s_n(\mathcal{A})\ e^{-n\epsilon^2/32}. \tag{1}$$

This partly solves one of our problems. But, how big can $s_n(\mathcal{A})$ be? Sometimes $s_n(\mathcal{A}) = 2^n$ for all $n$. For example, let $\mathcal{A}$ be all polygons in the plane. Then $s_n(\mathcal{A}) = 2^n$ for all $n$. But, in many cases, we will see that $s_n(\mathcal{A}) = 2^n$ for all $n$ up to some integer $d$ and then $s_n(\mathcal{A}) < 2^n$ for all $n > d$.

The **Vapnik-Chervonenkis (VC) dimension** is

$$d = d(\mathcal{A}) = \quad \text{largest n such that } s_n(\mathcal{A}) = 2^n.$$

In other words, $d$ is the size of the largest set that can be shattered.

Thus, $s_n(\mathcal{A}) = 2^n$ for all $n \leq d$ and $s_n(\mathcal{A}) < 2^n$ for all $n > d$. The VC dimensions of some common examples are summarized in Table 1. Now here is an interesting question: for $n > d$ how does $s_n(\mathcal{A})$ behave? It is less than $2^n$ but how much less?

**Theorem 6 (Sauer's Theorem)** *Suppose that $\mathcal{A}$ has finite VC dimension $d$. Then, for all $n \geq d$,*

$$s(\mathcal{A}, n) \leq (n + 1)^d. \tag{2}$$

We conclude that:

**Theorem 7** *Let $\mathcal{A}$ be a class of sets with VC dimension $d < \infty$. Then*

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \leq 8\ (n+1)^d\ e^{-n\epsilon^2/32}. \tag{3}$$

**Example 8** *Let's return to our first example. Suppose that $X_1, \ldots, X_n$ have cdf $F$. Let*

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq t).$$

*We would like to bound $\mathbb{P}(\sup_t |F_n(t) - F(t)| > \epsilon)$. Notice that $F_n(t) = P_n(A)$ where $A = (-\infty, t]$. Let $\mathcal{A} = \{(-\infty, t] :\ t \in \mathbb{R}\}$. This has VC dimension $d = 1$. So*

$$\mathbb{P}(\sup_t |F_n(t) - F(t)| > \epsilon) = \mathbb{P}\left(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \leq 8\ (n+1)\ e^{-n\epsilon^2/32}.$$

*In fact, there is a tighter bound in this case called the DKW (Dvoretsky-Kiefer-Wolfowitz) inequality:*

$$\mathbb{P}(\sup_t |F_n(t) - F(t)| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

# 4 Bounding Expectations

Eearlier we saw that we can use exponential bounds on probabilities to get bounds on expectations. Let us recall how that works.

Consider a finite collection $\mathcal{A} = \{A_1, \ldots, A_N\}$. Let

$$Z_n = \max_{1\leq j\leq N} |P_n(A_j) - P(A_j)|.$$

We know that

$$\mathbb{P}(Z_n > \epsilon) \leq 2me^{-2n\epsilon^2}. \tag{4}$$

But now we want to bound

$$\mathbb{E}(Z_n) = \left(\max_{1\leq j\leq N} |P_n(A_j) - P(A_j)|\right).$$

We can rewrite (4) as

$$\mathbb{P}(Z_n^2 > \epsilon^2) \leq 2Ne^{-2n\epsilon^2}.$$

or, in other words,

$$\mathbb{P}(Z_n^2 > t) \leq 2Ne^{-2nt}.$$

Recall that, in general, if $Y \geq 0$ then

$$\mathbb{E}(Y) = \int_0^\infty \mathbb{P}(Y > t)dt.$$

Hence, for any $s$,

$$\begin{aligned}
\mathbb{E}(Z_n^2) &= \int_0^\infty \mathbb{P}(Z_n^2 > t)dt \\
&= \int_0^s \mathbb{P}(Z_n^2 > t)dt + \int_s^\infty \mathbb{P}(Z_n^2 > t)dt \\
&\leq s + \int_s^\infty \mathbb{P}(Z_n^2 > t)dt \\
&\leq s + 2N \int_s^\infty e^{-2nt} dt \\
&= s + 2N \left( \frac{e^{-2ns}}{2n} \right) \\
&= s + \frac{N}{n} e^{-2ns}.
\end{aligned}$$

Let $s = \log(N)/(2n)$. Then

$$\mathbb{E}(Z_n^2) \leq s + \frac{N}{n} e^{-2ns} = \frac{\log N}{2n} + \frac{1}{n} = \frac{\log N + 2}{2n}.$$

Finally, we use Cauchy-Schwartz:

$$\mathbb{E}(Z_n) \leq \sqrt{\mathbb{E}(Z_n^2)} \leq \sqrt{\frac{\log N + 2}{2n}} = O\left( \sqrt{\frac{\log N}{n}} \right).$$

In summary:

$$\mathbb{E}\left( \max_{1 \leq j \leq N} |P_n(A_j) - P(A_j)| \right) = O\left( \sqrt{\frac{\log N}{n}} \right).$$

For a single set $A$ we would have $\mathbb{E}|P_n(A) - P(A)| \leq O(1/\sqrt{n})$. The bound only increases logarithmically with $N$.

# Lecture Notes 4

## 1 Random Samples

Let $X_1, \ldots, X_n \sim F$. A **statistic** is any function $T = g(X_1, \ldots, X_n)$. Recall that the sample mean is

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and sample variance is

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Let $\mu = \mathbb{E}(X_i)$ and $\sigma^2 = \mathsf{Var}(X_i)$. Recall that

$$\mathbb{E}(\overline{X}_n) = \mu, \quad \mathsf{Var}(\overline{X}_n) = \frac{\sigma^2}{n}, \quad \mathbb{E}(S_n^2) = \sigma^2.$$

**Theorem 1** *If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ then $\overline{X}_n \sim N(\mu, \sigma^2/n)$.*

**Proof.** We know that

$$M_{X_i}(s) = e^{\mu s + \sigma^2 s^2/2}.$$

So,

$$\begin{aligned} M_{\overline{X}_n}(t) &= \mathbb{E}(e^{t\overline{X}_n}) = \mathbb{E}(e^{\frac{t}{n}\sum_{i=1}^n X_i}) \\ &= (\mathbb{E}e^{tX_i/n})^n = (M_{X_i}(t/n))^n = \left(e^{(\mu t/n) + \sigma^2 t^2/(2n^2)}\right)^n \\ &= \exp\left\{\mu t + \frac{\frac{\sigma^2}{n}t^2}{2}\right\} \end{aligned}$$

which is the mgf of a $N(\mu, \sigma^2/n)$. ■

**Example 2** *(Example 5.2.10). Let $Z_1, \ldots, Z_n \sim \text{Cauchy}(0,1)$. Then $\overline{Z}_n \sim \text{Cauchy}(0,1)$.*

**Lemma 3** *If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ then*

$$T_n = \frac{\overline{X}_n - \mu}{S/\sqrt{n}} \sim t_{n-1} \approx N(0,1).$$

Let $X_{(1)}, \ldots, X_{(n)}$ denoted the ordered values:

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}.$$

Then $X_{(1)}, \ldots, X_{(n)}$ are called the *order statistics*.

# 2 Convergence

Let $X_1, X_2, \ldots$ be a sequence of random variables and let $X$ be another random variable. Let $F_n$ denote the cdf of $X_n$ and let $F$ denote the cdf of $X$.

1. $X_n$ **converges almost surely to** $X$, written $X_n \stackrel{a.s.}{\to} X$, if, for every $\epsilon > 0$,
$$\mathbb{P}(\lim_{n\to\infty} |X_n - X| < \epsilon) = 1. \tag{1}$$

2. $X_n$ **converges to** $X$ **in probability**, written $X_n \stackrel{P}{\to} X$, if, for every $\epsilon > 0$,
$$\mathbb{P}(|X_n - X| > \epsilon) \to 0 \tag{2}$$
as $n \to \infty$. In other words, $X_n - X = o_P(1)$.

3. $X_n$ **converges to** $X$ **in quadratic mean** (also called convergence in $L_2$), written $X_n \stackrel{qm}{\to} X$, if
$$\mathbb{E}(X_n - X)^2 \to 0 \tag{3}$$
as $n \to \infty$.

4. $X_n$ **converges to** $X$ **in distribution**, written $X_n \rightsquigarrow X$, if
$$\lim_{n\to\infty} F_n(t) = F(t) \tag{4}$$
at all $t$ for which $F$ is continuous.

*Convergence to a Constant.* A random variable $X$ has a **point mass distribution** if there exists a constant $c$ such that $\mathbb{P}(X = c) = 1$. The distribution for $X$ is denoted by $\delta_c$ and we write $X \sim \delta_c$. If $X_n \stackrel{P}{\to} \delta_c$ then we also write $X_n \stackrel{P}{\to} c$. Similarly for the other types of convergence.

**Theorem 4** *$X_n \stackrel{as}{\to} X$ if and only if, for every $\epsilon > 0$,*
$$\lim_{n\to\infty} \mathbb{P}(\sup_{m\geq n} |X_m - X| \leq \epsilon) = 1.$$

**Example 5** *(Example 5.5.8). This example shows that convergence in probability* **does not** *imply almost sure convergence. Let $S = [0, 1]$. Let $P$ be uniform on $[0, 1]$. We draw $S \sim P$. Let $X(s) = s$ and let*
$$X_1 = s + I_{[0,1]}(s), \quad X_2 = s + I_{[0,1/2]}(s), \quad X_3 = s + I_{[1/2,1]}(s)$$
$$X_4 = s + I_{[0,1/3]}(s), \quad X_5 = s + I_{[1/3,2/3]}(s), \quad X_6 = s + I_{[2/3,1]}(s)$$
*etc. Then $X_n \stackrel{P}{\to} X$. But, for each $s$, $X_n(s)$ does* **not** *converge to $X(s)$. Hence, $X_n$ does not converge almost surely to $X$.*

**Example 6** *Let $X_n \sim N(0, 1/n)$. Intuitively, $X_n$ is concentrating at 0 so we would like to say that $X_n$ converges to 0. Let's see if this is true. Let $F$ be the distribution function for a point mass at 0. Note that $\sqrt{n}X_n \sim N(0,1)$. Let $Z$ denote a standard normal random variable. For $t < 0$,*

$$F_n(t) = \mathbb{P}(X_n < t) = \mathbb{P}(\sqrt{n}X_n < \sqrt{n}t) = \mathbb{P}(Z < \sqrt{n}t) \to 0$$

*since $\sqrt{n}t \to -\infty$. For $t > 0$,*

$$F_n(t) = \mathbb{P}(X_n < t) = \mathbb{P}(\sqrt{n}X_n < \sqrt{n}t) = \mathbb{P}(Z < \sqrt{n}t) \to 1$$

*since $\sqrt{n}t \to \infty$. Hence, $F_n(t) \to F(t)$ for all $t \neq 0$ and so $X_n \rightsquigarrow 0$. Notice that $F_n(0) = 1/2 \neq F(1/2) = 1$ so convergence fails at $t = 0$. That doesn't matter because $t = 0$ is not a continuity point of $F$ and the definition of convergence in distribution only requires convergence at continuity points.*

*Now consider convergence in probability. For any $\epsilon > 0$, using Markov's inequality,*

$$\mathbb{P}(|X_n| > \epsilon) \;=\; \mathbb{P}(|X_n|^2 > \epsilon^2) \leq \frac{\mathbb{E}(X_n^2)}{\epsilon^2} = \frac{\frac{1}{n}}{\epsilon^2} \to 0$$

*as $n \to \infty$. Hence, $X_n \xrightarrow{P} 0$.*

The next theorem gives the relationship between the types of convergence.

**Theorem 7** *The following relationships hold:*
*(a) $X_n \xrightarrow{qm} X$ implies that $X_n \xrightarrow{P} X$.*
*(b) $X_n \xrightarrow{P} X$ implies that $X_n \rightsquigarrow X$.*
*(c) If $X_n \rightsquigarrow X$ and if $\mathbb{P}(X = c) = 1$ for some real number $c$, then $X_n \xrightarrow{P} X$.*
*(d) $X_n \xrightarrow{as} X$ implies $X_n \xrightarrow{P} X$.*
*In general, none of the reverse implications hold except the special case in (c).*

**Proof.** We start by proving (a). Suppose that $X_n \xrightarrow{qm} X$. Fix $\epsilon > 0$. Then, using Markov's inequality,

$$\mathbb{P}(|X_n - X| > \epsilon) = \mathbb{P}(|X_n - X|^2 > \epsilon^2) \leq \frac{\mathbb{E}|X_n - X|^2}{\epsilon^2} \to 0.$$

Proof of (b). Fix $\epsilon > 0$ and let $x$ be a continuity point of $F$. Then

$$\begin{aligned} F_n(x) &= \mathbb{P}(X_n \leq x) = \mathbb{P}(X_n \leq x, X \leq x + \epsilon) + \mathbb{P}(X_n \leq x, X > x + \epsilon) \\ &\leq \mathbb{P}(X \leq x + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon) \\ &= F(x + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon). \end{aligned}$$

Also,

$$\begin{aligned} F(x-\epsilon) &= \mathbb{P}(X \leq x-\epsilon) = \mathbb{P}(X \leq x-\epsilon, X_n \leq x) + \mathbb{P}(X \leq x-\epsilon, X_n > x) \\ &\leq F_n(x) + \mathbb{P}(|X_n - X| > \epsilon). \end{aligned}$$

Hence,

$$F(x-\epsilon) - \mathbb{P}(|X_n - X| > \epsilon) \leq F_n(x) \leq F(x+\epsilon) + \mathbb{P}(|X_n - X| > \epsilon).$$

Take the limit as $n \to \infty$ to conclude that

$$F(x-\epsilon) \leq \liminf_{n\to\infty} F_n(x) \leq \limsup_{n\to\infty} F_n(x) \leq F(x+\epsilon).$$

This holds for all $\epsilon > 0$. Take the limit as $\epsilon \to 0$ and use the fact that $F$ is continuous at $x$ and conclude that $\lim_n F_n(x) = F(x)$.

Proof of (c). Fix $\epsilon > 0$. Then,

$$\begin{aligned} \mathbb{P}(|X_n - c| > \epsilon) &= \mathbb{P}(X_n < c-\epsilon) + \mathbb{P}(X_n > c+\epsilon) \\ &\leq \mathbb{P}(X_n \leq c-\epsilon) + \mathbb{P}(X_n > c+\epsilon) \\ &= F_n(c-\epsilon) + 1 - F_n(c+\epsilon) \\ &\to F(c-\epsilon) + 1 - F(c+\epsilon) \\ &= 0 + 1 - 1 = 0. \end{aligned}$$

Proof of (d). This follows from Theorem 4.

Let us now show that the reverse implications do not hold.

*Convergence in probability does not imply convergence in quadratic mean.* Let $U \sim \text{Unif}(0,1)$ and let $X_n = \sqrt{n} I_{(0,1/n)}(U)$. Then $\mathbb{P}(|X_n| > \epsilon) = \mathbb{P}(\sqrt{n} I_{(0,1/n)}(U) > \epsilon) = \mathbb{P}(0 \leq U < 1/n) = 1/n \to 0$. Hence, $X_n \xrightarrow{P} 0$. But $\mathbb{E}(X_n^2) = n \int_0^{1/n} du = 1$ for all $n$ so $X_n$ does not converge in quadratic mean.

*Convergence in distribution does not imply convergence in probability.* Let $X \sim N(0,1)$. Let $X_n = -X$ for $n = 1, 2, 3, \ldots$; hence $X_n \sim N(0,1)$. $X_n$ has the same distribution function as $X$ for all $n$ so, trivially, $\lim_n F_n(x) = F(x)$ for all $x$. Therefore, $X_n \rightsquigarrow X$. But $\mathbb{P}(|X_n - X| > \epsilon) = \mathbb{P}(|2X| > \epsilon) = \mathbb{P}(|X| > \epsilon/2) \neq 0$. So $X_n$ does not converge to $X$ in probability. ■

The relationships between the types of convergence can be summarized as follows:

$$\begin{array}{ccccc} & & \text{q.m.} & & \\ & & \downarrow & & \\ \text{a.s.} & \to & \text{prob} & \to & \text{distribution} \end{array}$$

**Example 8** *One might conjecture that if* $X_n \xrightarrow{P} b$, *then* $\mathbb{E}(X_n) \to b$. *This is not true. Let* $X_n$ *be a random variable defined by* $\mathbb{P}(X_n = n^2) = 1/n$ *and* $\mathbb{P}(X_n = 0) = 1 - (1/n)$. *Now,* $\mathbb{P}(|X_n| < \epsilon) = \mathbb{P}(X_n = 0) = 1 - (1/n) \to 1$. *Hence,* $X_n \xrightarrow{P} 0$. *However,* $\mathbb{E}(X_n) = [n^2 \times (1/n)] + [0 \times (1 - (1/n))] = n$. *Thus,* $\mathbb{E}(X_n) \to \infty$.

**Example 9** *Let* $X_1, \ldots, X_n \sim \text{Uniform}(0, 1)$. *Let* $X_{(n)} = \max_i X_i$. *First we claim that* $X_{(n)} \xrightarrow{P} 1$. *This follows since*

$$\mathbb{P}(|X_{(n)} - 1| > \epsilon) = \mathbb{P}(X_{(n)} \leq 1 - \epsilon) = \prod_i \mathbb{P}(X_i \leq 1 - \epsilon) = (1 - \epsilon)^n \to 0.$$

*Also*

$$\mathbb{P}(n(1 - X_{(n)}) \leq t) = \mathbb{P}(X_{(n)} \leq 1 - (t/n)) = (1 - t/n)^n \to e^{-t}.$$

*So* $n(1 - X_{(n)}) \rightsquigarrow \text{Exp}(1)$.

Some convergence properties are preserved under transformations.

**Theorem 10** *Let* $X_n, X, Y_n, Y$ *be random variables. Let* $g$ *be a continuous function.*

*(a) If* $X_n \xrightarrow{P} X$ *and* $Y_n \xrightarrow{P} Y$, *then* $X_n + Y_n \xrightarrow{P} X + Y$.
*(b) If* $X_n \xrightarrow{qm} X$ *and* $Y_n \xrightarrow{qm} Y$, *then* $X_n + Y_n \xrightarrow{qm} X + Y$.
*(c) If* $X_n \rightsquigarrow X$ *and* $Y_n \rightsquigarrow c$, *then* $X_n + Y_n \rightsquigarrow X + c$.
*(d) If* $X_n \xrightarrow{P} X$ *and* $Y_n \xrightarrow{P} Y$, *then* $X_n Y_n \xrightarrow{P} XY$.
*(e) If* $X_n \rightsquigarrow X$ *and* $Y_n \rightsquigarrow c$, *then* $X_n Y_n \rightsquigarrow cX$.
*(f) If* $X_n \xrightarrow{P} X$, *then* $g(X_n) \xrightarrow{P} g(X)$.
*(g) If* $X_n \rightsquigarrow X$, *then* $g(X_n) \rightsquigarrow g(X)$.

- Parts (c) and (e) are know as **Slutzky's theorem**
- Parts (f) and (g) are known as **The Continuous Mapping Theorem**.
- It is worth noting that $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y$ does not in general imply that $X_n + Y_n \rightsquigarrow X + Y$.

# 3 The Law of Large Numbers

The law of large numbers (LLN) says that the mean of a large sample is close to the mean of the distribution. For example, the proportion of heads of a large number of tosses of a fair coin is expected to be close to 1/2. We now make this more precise.

Let $X_1, X_2, \ldots$ be an iid sample, let $\mu = \mathbb{E}(X_1)$ and $\sigma^2 = \mathsf{Var}(X_1)$. Recall that the sample mean is defined as $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$ and that $\mathbb{E}(\overline{X}_n) = \mu$ and $\mathsf{Var}(\overline{X}_n) = \sigma^2/n$.

**Theorem 11 (The Weak Law of Large Numbers (WLLN))**
*If $X_1, \ldots, X_n$ are iid, then $\overline{X}_n \xrightarrow{P} \mu$. Thus, $\overline{X}_n - \mu = o_P(1)$.*

> Interpretation of the WLLN: The distribution of $\overline{X}_n$ becomes more concentrated around $\mu$ as $n$ gets large.

**Proof.** Assume that $\sigma < \infty$. This is not necessary but it simplifies the proof. Using Chebyshev's inequality,

$$\mathbb{P}\left(|\overline{X}_n - \mu| > \epsilon\right) \leq \frac{\mathsf{Var}(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

which tends to 0 as $n \to \infty$. ■

**Theorem 12 The Strong Law of Large Numbers.** *Let $X_1, \ldots, X_n$ be iid with mean $\mu$. Then $\overline{X}_n \xrightarrow{as} \mu$.*

The proof is beyond the scope of this course.

# 4 The Central Limit Theorem

The law of large numbers says that the distribution of $\overline{X}_n$ piles up near $\mu$. This isn't enough to help us approximate probability statements about $\overline{X}_n$. For this we need the central limit theorem.

Suppose that $X_1, \ldots, X_n$ are iid with mean $\mu$ and variance $\sigma^2$. The central limit theorem (CLT) says that $\overline{X}_n = n^{-1}\sum_i X_i$ has a distribution which is approximately Normal with mean $\mu$ and variance $\sigma^2/n$. This is remarkable since nothing is assumed about the distribution of $X_i$, except the existence of the mean and variance.

**Theorem 13 (The Central Limit Theorem (CLT))** *Let $X_1, \ldots, X_n$ be iid with mean $\mu$ and variance $\sigma^2$. Let $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$. Then*

$$Z_n \equiv \frac{\overline{X}_n - \mu}{\sqrt{\mathsf{Var}(\overline{X}_n)}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \rightsquigarrow Z$$

*where $Z \sim N(0,1)$. In other words,*

$$\lim_{n\to\infty} \mathbb{P}(Z_n \leq z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

> Interpretation: Probability statements about $\overline{X}_n$ can be approximated using a Normal distribution. It's the probability statements that we are approximating, not the random variable itself.

A consequence of the CLT is that

$$\overline{X}_n - \mu = O_P\left(\sqrt{\frac{1}{n}}\right).$$

In addition to $Z_n \rightsquigarrow N(0,1)$, there are several forms of notation to denote the fact that the distribution of $Z_n$ is converging to a Normal. They all mean the same thing. Here they are:

$$\begin{aligned}
Z_n &\approx N(0,1) \\
\overline{X}_n &\approx N\left(\mu, \frac{\sigma^2}{n}\right) \\
\overline{X}_n - \mu &\approx N\left(0, \frac{\sigma^2}{n}\right) \\
\sqrt{n}(\overline{X}_n - \mu) &\approx N\left(0, \sigma^2\right) \\
\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} &\approx N(0,1).
\end{aligned}$$

Recall that if $X$ is a random variable, its moment generating function (mgf) is $\psi_X(t) = \mathbb{E}e^{tX}$. Assume in what follows that the mgf is finite in a neighborhood around $t = 0$.

**Lemma 14** *Let $Z_1, Z_2, \ldots$ be a sequence of random variables. Let $\psi_n$ be the mgf of $Z_n$. Let $Z$ be another random variable and denote its mgf by $\psi$. If $\psi_n(t) \to \psi(t)$ for all $t$ in some open interval around 0, then $Z_n \rightsquigarrow Z$.*

**Proof of the central limit theorem.** Let $Y_i = (X_i - \mu)/\sigma$. Then, $Z_n = n^{-1/2}\sum_i Y_i$. Let $\psi(t)$ be the mgf of $Y_i$. The mgf of $\sum_i Y_i$ is $(\psi(t))^n$ and mgf of $Z_n$ is $[\psi(t/\sqrt{n})]^n \equiv \xi_n(t)$. Now $\psi'(0) = \mathbb{E}(Y_1) = 0$, $\psi''(0) = \mathbb{E}(Y_1^2) = \mathsf{Var}(Y_1) = 1$. So,

$$\begin{aligned}
\psi(t) &= \psi(0) + t\psi'(0) + \frac{t^2}{2!}\psi''(0) + \frac{t^3}{3!}\psi'''(0) + \cdots \\
&= 1 + 0 + \frac{t^2}{2} + \frac{t^3}{3!}\psi'''(0) + \cdots \\
&= 1 + \frac{t^2}{2} + \frac{t^3}{3!}\psi'''(0) + \cdots
\end{aligned}$$

Now,

$$\begin{aligned}
\xi_n(t) &= \left[\psi\left(\frac{t}{\sqrt{n}}\right)\right]^n \\
&= \left[1 + \frac{t^2}{2n} + \frac{t^3}{3!n^{3/2}}\psi'''(0) + \cdots\right]^n \\
&= \left[1 + \frac{\frac{t^2}{2} + \frac{t^3}{3!n^{1/2}}\psi'''(0) + \cdots}{n}\right]^n \\
&\to e^{t^2/2}
\end{aligned}$$

which is the mgf of a N(0,1). The result follows from Lemma 14. In the last step we used the fact that if $a_n \to a$ then

$$\left(1 + \frac{a_n}{n}\right)^n \to e^a. \quad \blacksquare$$

---

The central limit theorem tells us that $Z_n = \sqrt{n}(\overline{X}_n - \mu)/\sigma$ is approximately N(0,1). However, we rarely know $\sigma$. We can estimate $\sigma^2$ from $X_1, \ldots, X_n$ by

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2.$$

This raises the following question: if we replace $\sigma$ with $S_n$, is the central limit theorem still true? The answer is yes.

**Theorem 15** *Assume the same conditions as the CLT. Then,*

$$T_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \rightsquigarrow N(0,1).$$

**Proof.** We have that

$$T_n = Z_n W_n$$

where

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$$

and

$$W_n = \frac{\sigma}{S_n}.$$

Now $Z_n \rightsquigarrow N(0,1)$ and $W_n \xrightarrow{P} 1$. The result follows from Slutzky's theorem. $\blacksquare$

---

There is also a multivariate version of the central limit theorem. Recall that $X = (X_1, \ldots, X_k)^T$ has a multivariate Normal distribution with mean vector $\mu$ and covariance matrix $\Sigma$ if

$$f(x) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$

In this case we write $X \sim N(\mu, \Sigma)$.

**Theorem 16 (Multivariate central limit theorem)** *Let $X_1, \ldots, X_n$ be iid random vectors where $X_i = (X_{1i}, \ldots, X_{ki})^T$ with mean $\mu = (\mu_1, \ldots, \mu_k)^T$ and covariance matrix $\Sigma$. Let $\overline{X} = (\overline{X}_1, \ldots, \overline{X}_k)^T$ where $\overline{X}_j = n^{-1}\sum_{i=1}^n X_{ji}$. Then,*

$$\sqrt{n}(\overline{X} - \mu) \rightsquigarrow N(0, \Sigma).$$

# 5 The Delta Method

If $Y_n$ has a limiting Normal distribution then the delta method allows us to find the limiting distribution of $g(Y_n)$ where $g$ is any smooth function.

**Theorem 17 (The Delta Method)** *Suppose that*

$$\frac{\sqrt{n}(Y_n - \mu)}{\sigma} \rightsquigarrow N(0,1)$$

*and that $g$ is a differentiable function such that $g'(\mu) \neq 0$. Then*

$$\frac{\sqrt{n}(g(Y_n) - g(\mu))}{|g'(\mu)|\sigma} \rightsquigarrow N(0,1).$$

*In other words,*

$$Y_n \approx N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{implies that} \quad g(Y_n) \approx N\left(g(\mu), (g'(\mu))^2 \frac{\sigma^2}{n}\right).$$

**Example 18** *Let $X_1, \ldots, X_n$ be iid with finite mean $\mu$ and finite variance $\sigma^2$. By the central limit theorem, $\sqrt{n}(\overline{X}_n - \mu)/\sigma \rightsquigarrow N(0,1)$. Let $W_n = e^{\overline{X}_n}$. Thus, $W_n = g(\overline{X}_n)$ where $g(s) = e^s$. Since $g'(s) = e^s$, the delta method implies that $W_n \approx N(e^\mu, e^{2\mu}\sigma^2/n)$.*

There is also a multivariate version of the delta method.

**Theorem 19 (The Multivariate Delta Method)** *Suppose that $Y_n = (Y_{n1}, \ldots, Y_{nk})$ is a sequence of random vectors such that*

$$\sqrt{n}(Y_n - \mu) \rightsquigarrow N(0, \Sigma).$$

*Let $g : \mathbb{R}^k \to \mathbb{R}$ and let*

$$\nabla g(y) = \begin{pmatrix} \frac{\partial g}{\partial y_1} \\ \vdots \\ \frac{\partial g}{\partial y_k} \end{pmatrix}.$$

*Let $\nabla_\mu$ denote $\nabla g(y)$ evaluated at $y = \mu$ and assume that the elements of $\nabla_\mu$ are nonzero. Then*

$$\sqrt{n}(g(Y_n) - g(\mu)) \rightsquigarrow N\left(0, \nabla_\mu^T \Sigma \nabla_\mu\right).$$

**Example 20** *Let*

$$\begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix}, \begin{pmatrix} X_{12} \\ X_{22} \end{pmatrix}, \ \ldots, \ \begin{pmatrix} X_{1n} \\ X_{2n} \end{pmatrix}$$

*be iid random vectors with mean $\mu = (\mu_1, \mu_2)^T$ and variance $\Sigma$. Let*

$$\overline{X}_1 = \frac{1}{n}\sum_{i=1}^n X_{1i}, \qquad \overline{X}_2 = \frac{1}{n}\sum_{i=1}^n X_{2i}$$

*and define* $Y_n = \overline{X}_1\overline{X}_2$. *Thus,* $Y_n = g(\overline{X}_1, \overline{X}_2)$ *where* $g(s_1, s_2) = s_1 s_2$. *By the central limit theorem,*

$$\sqrt{n}\begin{pmatrix} \overline{X}_1 - \mu_1 \\ \overline{X}_2 - \mu_2 \end{pmatrix} \rightsquigarrow N(0, \Sigma).$$

*Now*

$$\nabla g(s) = \begin{pmatrix} \frac{\partial g}{\partial s_1} \\ \frac{\partial g}{\partial s_2} \end{pmatrix} = \begin{pmatrix} s_2 \\ s_1 \end{pmatrix}$$

*and so*

$$\nabla_\mu^T \Sigma \nabla_\mu = (\mu_2 \quad \mu_1)\begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}\begin{pmatrix} \mu_2 \\ \mu_1 \end{pmatrix} = \mu_2^2\sigma_{11} + 2\mu_1\mu_2\sigma_{12} + \mu_1^2\sigma_{22}.$$

*Therefore,*

$$\sqrt{n}(\overline{X}_1\overline{X}_2 - \mu_1\mu_2) \rightsquigarrow N\Big(0, \mu_2^2\sigma_{11} + 2\mu_1\mu_2\sigma_{12} + \mu_1^2\sigma_{22}\Big). \quad \blacksquare$$

# Addendum to Lecture Notes 4

Here is the proof that

$$T_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \rightsquigarrow N(0,1)$$

where

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

**Step 1.** We first show that $R_n^2 \xrightarrow{P} \sigma^2$ where

$$R_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Note that

$$R_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2.$$

Define $Y_i = X_i^2$. Then, using the LLN (law of large numbers)

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{P} \mathbb{E}(Y_i) = \mathbb{E}(X_i^2) = \mu^2 + \sigma^2.$$

Next, by the LLN,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu.$$

Since $g(t) = t^2$ is continuous, the continuous mapping theorem implies that

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \xrightarrow{P} \mu^2.$$

Thus

$$R_n^2 \xrightarrow{P} (\mu^2 + \sigma^2) - \mu^2 = \sigma^2.$$

**Step 2.** Note that

$$S_n^2 = \left(\frac{n}{n-1}\right) R_n^2.$$

Since, $R_n^2 \xrightarrow{P} \sigma^2$ and $n/(n-1) \to 1$, we have that $S_n^2 \xrightarrow{P} \sigma^2$.

**Step 3.** Since $g(t) = \sqrt{t}$ is continuous, (for $t \geq 0$) the continuous mapping theorem implies that $S_n \xrightarrow{P} \sigma$.

**Step 4.** Since $g(t) = t/\sigma$ is continuous, the continuous mapping theorem implies that $S_n/\sigma \xrightarrow{P} 1$.

**Step 5.** Since $g(t) = 1/t$ is continuous (for $t > 0$) the continuous mapping theorem implies that $\sigma/S_n \xrightarrow{P} 1$. Since convergence in probability implies convergence in distribution, $\sigma/S_n \rightsquigarrow 1$.

**Step 5.** Note that

$$T_n = \left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}\right)\left(\frac{\sigma}{S_n}\right) \equiv V_n W_n.$$

Now $V_n \rightsquigarrow Z$ where $Z \sim N(0,1)$ by the CLT. And we showed that $W_n \rightsquigarrow 1$. By Slutzky's theorem, $T_n = V_n W_n \rightsquigarrow Z \times 1 = Z$.

# Lecture Notes 5

## 1 Statistical Models

A statistical model $\mathcal{P}$ is a collection of probability distributions (or a collection of densities). An example of a **nonparametric model** is

$$\mathcal{P} = \left\{ p : \int (p''(x))^2 dx < \infty \right\}.$$

A **parametric model** has the form

$$\mathcal{P} = \left\{ p(x;\theta) : \ \theta \in \Theta \right\}$$

where $\Theta \subset \mathbb{R}^d$. An example is the set of Normal densities $\{p(x;\theta) = (2\pi)^{-1/2} e^{-(x-\theta)^2/2}\}$. For now, we focus on parametric models.

The model comes from **assumptions**. Some examples:

- Time until something fails is often modeled by an exponential distribution.
- Number of rare events is often modeled by a Poisson distribution.
- Lengths and weights are often modeled by a Normal distribution.

**These models are not correct.** But they might be useful. Later we consider nonparametric methods that do not assume a parametric model

## 2 Statistics

Let $X_1, \ldots, X_n \sim p(x;\theta)$. Let $X^n \equiv (X_1, \ldots, X_n)$. Any function $T = T(X_1, \ldots, X_n)$ is itself a random variable which we will call a *statistic*.

Some examples are:

- order statistics, $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$

- sample mean: $\overline{X} = \frac{1}{n}\sum_i X_i$,
- sample variance: $S^2 = \frac{1}{n-1}\sum_i (X_i - \overline{x})^2$,
- sample median: middle value of ordered statistics,
- sample minimum: $X_{(1)}$
- sample maximum: $X_{(n)}$
- sample range: $X_{(n)} - X_{(1)}$
- sample interquartile range: $X_{(.75n)} - X_{(.25n)}$

**Example 1** *If* $X_1, \ldots, X_n \sim \Gamma(\alpha, \beta)$, *then* $\overline{X} \sim \Gamma(n\alpha, \beta/n)$.

*Proof:*

$$\begin{aligned} M_{\overline{X}} &= E[e^{t\overline{x}}] = E[e^{\sum X_i t/n}] = \prod_i E[e^{X_i(t/n)}] \\ &= [M_X(t/n)]^n = \left[\left(\frac{1}{1-\beta t/n}\right)^{\alpha}\right]^n = \left[\frac{1}{1-\beta/nt}\right]^{n\alpha}. \end{aligned}$$

*This is the mgf of* $\Gamma(n\alpha, \beta/n)$.

**Example 2** *If* $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ *then* $\overline{X} \sim N(\mu, \sigma^2/n)$.

**Example 3** *If* $X_1, \ldots, X_n$ *iid Cauchy(0,1),*

$$p(x) = \frac{1}{\pi(1+x^2)}$$

*for* $x \in \mathbb{R}$, *then* $\overline{X} \sim$ *Cauchy(0,1).*

**Example 4** *If* $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ *then*

$$\frac{(n-1)}{\sigma^2} S^2 \sim \chi^2_{(n-1)}.$$

*The proof is based on the mgf.*

**Example 5** *Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the order statistics, which means that the sample $X_1, X_2, \ldots, X_n$ has been ordered from smallest to largest:*

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}.$$

*Now,*

$$\begin{aligned} F_{X_{(k)}}(x) &= P(X_{(k)} \leq x) \\ &= P(\textit{at least } k \textit{ of the } X_1, \ldots, X_n \leq x) \\ &= \sum_{j=k}^{n} P(\textit{exactly } j \textit{ of the } X_1, \ldots, X_n \leq x) \\ &= \sum_{j=k}^{n} \binom{n}{j} [F_X(x)]^j [1 - F_X(x)]^{n-j} \end{aligned}$$

*Differentiate to find the pdf (See CB p. 229):*

$$p_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} [F_X(x)]^{k-1} \; p(x) [1 - F_X(x)]^{n-k}.$$

# 3 Sufficiency

(Ch 6 CB) We continue with **parametric inference**. In this section we discuss **data reduction** as a formal concept.

- Sample $X^n = X_1, \cdots, X_n \sim F$.
- Assume $F$ belongs to a family of distributions, (e.g. $F$ is Normal), indexed by some parameter $\theta$.
- We want to learn about $\theta$ and try to summarize the data without throwing any information about $\theta$ away.
- If a statistic $T(X_1, \cdots, X_n)$ contains all the information about $\theta$ in the sample we say $T$ is **sufficient**.

## 3.1 Sufficient Statistics

**Definition**: $T$ is sufficient for $\theta$ if the conditional distribution of $X^n|T$ does not depend on $\theta$. Thus, $f(x_1, \ldots, x_n|t; \theta) = f(x_1, \ldots, x_n|t)$.

**Example 6** *$X_1, \cdots, X_n \sim \text{Poisson}(\theta)$. Let $T = \sum_{i=1}^n X_i$. Then,*

$$p_{X^n|T}(x^n|t) = \mathbb{P}(X^n = x^n|T(X^n) = t) = \frac{P(X^n = x^n \text{ and } T = t)}{P(T = t)}.$$

*But*

$$P(X^n = x^n \text{ and } T = t) = \begin{cases} 0 & \text{if} \quad T(x^n) \neq t \\ P(X^n = x^n) & \text{if} \quad T(X^n) = t \end{cases}$$

*Hence,*

$$P(X^n = x^n) = \prod_{i=1}^n \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\sum x_i}}{\prod(x_i!)} = \frac{e^{-n\theta}\theta^t}{\prod(x_i!)}.$$

*Now, $T(x^n) = \sum x_i = t$ and so*

$$P(T = t) = \frac{e^{-n\theta}(n\theta)^t}{t!} \qquad \text{since } T \sim \text{Poisson}(n\theta).$$

*Thus,*

$$\frac{P(X^n = x^n)}{P(T = t)} = \frac{t!}{(\prod x_i)! n^t}$$

*which does not depend on $\theta$. So $T = \sum_i X_i$ is a sufficient statistic for $\theta$. Other sufficient statistics are: $T = 3.7 \sum_i X_i$, $T = (\sum_i X_i, X_4)$, and $T(X_1, \ldots, X_n) = (X_1, \ldots, X_n)$.*

## 3.2 Sufficient Partitions

It is better to describe sufficiency in terms of partitions of the sample space.

**Example 7** *Let $X_1, X_2, X_3 \sim \text{Bernoulli}(\theta)$. Let $T = \sum X_i$.*

| $x^n$ | | $t$ | $p(x\|t)$ |
|---|---|---|---|
| *(0, 0, 0)* | $\to$ | $t = 0$ | *1* |
| *(0, 0, 1)* | $\to$ | $t = 1$ | *1/3* |
| *(0, 1, 0)* | $\to$ | $t = 1$ | *1/3* |
| *(1, 0, 0)* | $\to$ | $t = 1$ | *1/3* |
| *(0, 1, 1)* | $\to$ | $t = 2$ | *1/3* |
| *(1, 0, 1)* | $\to$ | $t = 2$ | *1/3* |
| *(1, 1, 0)* | $\to$ | $t = 2$ | *1/3* |
| *(1, 1, 1)* | $\to$ | $t = 3$ | *1* |
| *8 elements* | $\to$ | *4 elements* | |

1. A partition $B_1, \ldots, B_k$ is sufficient if $f(x|X \in B)$ does not depend on $\theta$.

2. A statistic $T$ induces a partition. For each $t$, $\{x : \ T(x) = t\}$ is one element of the partition. $T$ is sufficient if and only if the partition is sufficient.

3. Two statistics can generate the same partition: example: $\sum_i X_i$ and $3\sum_i X_i$.

4. If we split any element $B_i$ of a sufficient partition into smaller pieces, we get another sufficient partition.

**Example 8** *Let* $X_1, X_2, X_3 \sim$ Bernoulli$(\theta)$. *Then* $T = X_1$ *is* **not** *sufficient. Look at its partition:*

| $x^n$ | | $t$ | $p(x\|t)$ |
|---|---|---|---|
| *(0, 0, 0)* | $\to$ | $t=0$ | $(1-\theta)^2$ |
| *(0, 0, 1)* | $\to$ | $t=0$ | $\theta(1-\theta)$ |
| *(0, 1, 0)* | $\to$ | $t=0$ | $\theta(1-\theta)$ |
| *(0, 1, 1)* | $\to$ | $t=0$ | $\theta^2$ |
| *(1, 0, 0)* | $\to$ | $t=1$ | $(1-\theta)^2$ |
| *(1, 0, 1)* | $\to$ | $t=1$ | $\theta(1-\theta)$ |
| *(1, 1, 0)* | $\to$ | $t=1$ | $\theta(1-\theta)$ |
| *(1, 1, 1)* | $\to$ | $t=1$ | $\theta^2$ |
| *8 elements* | $\to$ | *2 elements* | |

## 3.3 The Factorization Theorem

**Theorem 9** *$T(X^n)$ is sufficient for $\theta$ if the joint pdf/pmf of $X^n$ can be factored as*

$$p(x^n;\theta) = h(x^n) \times g(t;\theta).$$

**Example 10** *Let $X_1, \cdots, X_n \sim$ Poisson. Then*

$$p(x^n;\theta) = \frac{e^{-n\theta}\theta^{\sum X_i}}{\prod(x_i!)} = \frac{1}{\prod(x_i!)} \times e^{-n\theta}\theta^{\sum_i X_i}.$$

**Example 11** *$X_1, \cdots, X_n \sim N(\mu, \sigma^2)$. Then*

$$p(x^n;\mu,\sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{\sum(x_i-\overline{x})^2 + n(\overline{x}-\mu)^2}{2\sigma^2}\right\}.$$

*(a) If $\sigma$ known:*

$$p(x^n;\mu) = \underbrace{\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{\frac{-\sum(x_i-\overline{x})^2}{2\sigma^2}\right\}}_{h(x^n)} \quad \underbrace{exp\ \left\{\frac{-n(\overline{x}-\mu)^2}{2\sigma^2}\right\}}_{g(T(x^n)|\mu)}.$$

*Thus, $\overline{X}$ is sufficient for $\mu$.*

*(b) If $(\mu, \sigma^2)$ unknown then $T = (\overline{X}, S^2)$ is sufficient. So is $T = (\sum X_i, \sum X_i^2)$.*

## 3.4 Minimal Sufficient Statistics (MSS)

We want the greatest reduction in dimension.

**Example 12** *$X_1, \cdots, X_n \sim N(0, \sigma^2)$. Some sufficient statistics are:*

$$\begin{aligned} T(X_1, \cdots, X_n) &= (X_1, \cdots, X_n) \\ T(X_1, \cdots, X_n) &= (X_1^2, \cdots, X_n^2) \\ T(X_1, \cdots, X_n) &= \left( \sum_{i=1}^{m} X_i^2, \sum_{i=m+1}^{n} X_i^2 \right) \\ T(X_1, \cdots, X_n) &= \sum X_i^2. \end{aligned}$$

**Definition:** $T$ is a **Minimal Sufficient Statistic** if the following two statements are true:

1. $T$ is sufficient and
2. If $U$ is any other sufficient statistic then $T = g(U)$ for some function $g$.

In other words, $T$ generates the **coarsest sufficient partition**.

Suppose $U$ is sufficient. Suppose $T = H(U)$ is also sufficient. $T$ provides greater reduction than $U$ unless $H$ is a $1-1$ transformation, in which case $T$ and $U$ are equivalent.

**Example 13** *$X \sim N(0, \sigma^2)$. $X$ is sufficient. $|X|$ is sufficient. $|X|$ is MSS. So are $X^2, X^4, e^{X^2}$.*

**Example 14** *Let $X_1, X_2, X_3 \sim$ Bernoulli$(\theta)$. Let $T = \sum X_i$.*

| $x^n$ | | $t$ | $p(x\|t)$ | $u$ | $p(x\|u)$ |
|---|---|---|---|---|---|
| *(0, 0, 0)* | $\rightarrow$ | $t=0$ | *1* | $u=0$ | *1* |
| *(0, 0, 1)* | $\rightarrow$ | $t=1$ | *1/3* | $u=1$ | *1/3* |
| *(0, 1, 0)* | $\rightarrow$ | $t=1$ | *1/3* | $u=1$ | *1/3* |
| *(1, 0, 0)* | $\rightarrow$ | $t=1$ | *1/3* | $u=1$ | *1/3* |
| *(0, 1, 1)* | $\rightarrow$ | $t=2$ | *1/3* | $u=73$ | *1/2* |
| *(1, 0, 1)* | $\rightarrow$ | $t=2$ | *1/3* | $u=73$ | *1/2* |
| *(1, 1, 0)* | $\rightarrow$ | $t=2$ | *1/3* | $u=91$ | *1* |
| *(1, 1, 1)* | $\rightarrow$ | $t=3$ | *1* | $u=103$ | *1* |

*Note that $U$ and $T$ are both sufficient but $U$ is not minimal.*

## 3.5 How to find a Minimal Sufficient Statistic

**Theorem 15** *Define*

$$R(x^n, y^n; \theta) = \frac{p(y^n;\theta)}{p(x^n;\theta)}.$$

*Suppose that $T$ has the following property:*

| $R(x^n, y^n; \theta)$ **does not depend on** $\theta$ **if and only if** $T(y^n) = T(x^n)$**.** |
|---|

*Then $T$ is a MSS.*

**Example 16** *$Y_1, \cdots, Y_n$ iid Poisson $(\theta)$.*

$$p(y^n;\theta) = \frac{e^{-n\theta}\theta^{\sum y_i}}{\prod y_i}, \quad \frac{p(y^n;\theta)}{p(x^n;\theta)} = \frac{\theta^{\sum y_i - \sum x_i}}{\prod y_i! / \prod x_i!}$$

*which is independent of $\theta$ iff $\sum y_i = \sum x_i$. This implies that $T(Y^n) = \sum Y_i$ is a minimal sufficient statistic for $\theta$.*

The minimal sufficient statistic is not unique. But, the minimal sufficient partition is unique.

**Example 17** *Cauchy.*

$$p(x;\theta) = \frac{1}{\pi(1+(x-\theta)^2)}.$$

*Then*

$$\frac{p(y^n;\theta)}{p(x^n;\theta)} = \frac{\prod_{i=1}^{n}\{1+(x_i-\theta)^2\}}{\prod_{j=1}^{n}\{1+(y_j-\theta)^2\}}.$$

*The ratio is a constant function of* $\theta$ *if*

$$T(Y^n) = (Y_{(1)}, \cdots, Y_{(n)}).$$

*It is technically harder to show that this is true only if* $T$ *is the order statistics, but it could be done using theorems about polynomials. Having shown this, one can conclude that the order statistics are the minimal sufficient statistics for* $\theta$*.*

**Note: Ignore the material on completeness and ancillary statistics.**

# Lecture Notes 6

## 1 The Likelihood Function

**Definition.** Let $X^n = (X_1, \cdots, X_n)$ have joint density $p(x^n; \theta) = p(x_1, \ldots, x_n; \theta)$ where $\theta \in \Theta$. The **likelihood function** $L : \Theta \to [0, \infty)$ is defined by

$$L(\theta) \equiv L(\theta; x^n) = p(x^n; \theta)$$

where $x^n$ is fixed and $\theta$ varies in $\Theta$.

1. The likelihood function is a function of $\theta$.
2. The likelihood function is **not** a probability density function.
3. If the data are iid then the likelihood is

$$L(\theta) = \prod_{i=1}^{n} p(x_i; \theta) \qquad \text{iid case only.}$$

4. The likelihood is only defined up to a constant of proportionality.
5. The likelihood function is used (i) to generate estimators (the maximum likelihood estimator) and (ii) as a key ingredient in Bayesian inference.

**Example 1** *These 2 samples have the same likelihood function:*

$$(X_1, X_2, X_3) \sim \textit{Multinomial}\ (n = 6, \theta, \theta, 1 - 2\theta)$$

$$X = (1, 3, 2) \quad \Longrightarrow \quad L(\theta) = \frac{6!}{1!3!2!}\theta^1\theta^3(1 - 2\theta)^2 \propto \theta^4(1 - 2\theta)^2$$

$$X = (2, 2, 2) \quad \Longrightarrow \quad L(\theta) = \frac{6!}{2!2!2!}\theta^2\theta^2(1 - 2\theta)^2 \propto \theta^4(1 - 2\theta)^2$$

**Example 2** $X_1, \cdots, X_n \sim N(\mu, 1)$. *Then,*

$$L(\mu) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \propto \exp\left\{-\frac{n}{2}(\overline{x} - \mu)^2\right\}.$$

**Example 3** *Let* $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. *Then*

$$L(p) \propto p^X (1-p)^{n-X}$$

*for* $p \in [0,1]$ *where* $X = \sum_i X_i$.

**Theorem 4** *Write* $x^n \sim y^n$ *if* $L(\theta|x^n) \propto L(\theta|y^n)$. *The partition induced by* $\sim$ *is the minimal sufficient partition.*

**Example 5** *A non iid example. An AR(1) time series auto regressive model. The model is:* $X_1 \sim N(0, \sigma^2)$ *and*

$$X_{i+1} = \theta X_i + e_{i+1} \qquad e_i \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

*It can be show that we have the Markov property:* $o(x_{n+1}|x_n, x_{n-1}, \cdots, x_1) = p(x_{n+1}|x_n)$. *The likelihood function is*

$$\begin{aligned} L(\theta) &= p(x^n; \theta) \\ &= p(x_1;\theta)p(x_2|x_1;\theta)\cdots p(x_n|x_1,\ldots,x_{n-1};\theta) \\ &= p(x_n|x_{n-1};\theta)p(x_{n-1}|x_{n-2};\theta)\cdots p(x_2|x_1;\theta)p(x_1;\theta) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\theta} \exp\left(\frac{-1}{2\theta^2}(x_{n+i-1} - \theta x_{n-i})^2\right). \end{aligned}$$

# 2 Likelihood, Sufficiency and the Likelihood Principle

The likelihood function is a minimal sufficient statistic. That is, if we define the equivalence relation: $x^n \sim y^n$ when $L(\theta; x^n) \propto L(\theta; y^n)$ then the resulting partition is minimal sufficient.

Does this mean that the likelihood function contains all the relevant information? Some people say yes it does. This is sometimes called *the likelihood principle.* That is, the likelihood principle says that the likelihood function contains all the infomation in the data.

This is FALSE. Here is a simple example to illustrate why. Let $\mathcal{C} = \{c_1, \ldots, c_N\}$ be a finite set of constants. For simplicity, asssume that $c_j \in \{0, 1\}$ (although this is not important). Let $\theta = N^{-1} \sum_{j=1}^{N} c_j$. Suppose we want to estimate $\theta$. We proceed as follows.

Let $S_1, \ldots, S_n \sim \text{Bernoulli}(\pi)$ where $\pi$ is known. If $S_i = 1$ you get to see $c_i$. Otherwise, you do not. (This is an example of survey sampling.) The likelihood function is

$$\prod_i \pi^{S_i}(1-\pi)^{1-S_i}.$$

The unknown parameter does not appear in the likelihood. In fact, there are no unknown parameters in the likelihood! The likelihood function contains no information at all.

But we can estimate $\theta$. Let

$$\widehat{\theta} = \frac{1}{N\pi}\sum_{j=1}^{N} c_j S_j.$$

Then $\mathbb{E}(\widehat{\theta}) = \theta$. Hoeffding's inequality implies that

$$\mathbb{P}(|\widehat{\theta} - \theta| > \epsilon) \leq 2e^{-2n\epsilon^2\pi^2}.$$

Hence, $\widehat{\theta}$ is close to $\theta$ with high probability.

Summary: the minimal sufficient statistic has all the information you need to compute the likelihood. But that does not mean that all the information is in the likelihood.

# Lecture Notes 7

## 1 Parametric Point Estimation

$X_1, \ldots, X_n \sim p(x;\theta)$. Want to estimate $\theta = (\theta_1, \ldots, \theta_k)$. An *estimator*

$$\widehat{\theta} = \widehat{\theta}_n = w(X_1, \ldots, X_n)$$

is a function of the data.

**Methods**:

1. Method of Moments (MOM)
2. Maximum likelihood (MLE)
3. Bayesian estimators

**Evaluating Estimators**:

1. Bias and Variance
2. Mean squared error (MSE)
3. Minimax Theory
4. Large sample theory (later).

## 2 Some Terminology

- $\mathbb{E}_\theta(\widehat{\theta}) = \int \cdots \int \widehat{\theta}(x_1, \ldots, x_n) p(x_1;\theta) \cdots p(x_n;\theta) dx_1 \cdots dx_n$
- Bias: $\mathbb{E}_\theta(\widehat{\theta}) - \theta$
- the distribution of $\widehat{\theta}_n$ is called its *sampling distribution*
- the standard deviation of $\widehat{\theta}_n$ is called the *standard error* denoted by $\text{se}(\widehat{\theta}_n)$
- $\widehat{\theta}_n$ is *consistent* if $\widehat{\theta}_n \xrightarrow{\text{P}} \theta$
- later we will see that if bias $\to 0$ and $\text{Var}(\widehat{\theta}_n) \to 0$ as $n \to \infty$ then $\widehat{\theta}_n$ is consistent
- an estimator is *robust* if it is not strongly affected by perturbations in the data (more later)

# 3 Method of Moments

Define

$$\begin{aligned} m_1 &= \frac{1}{n}\sum_{i=1}^n X_i, & \mu_1(\theta) &= \mathbb{E}(X_i) \\ m_2 &= \frac{1}{n}\sum_{i=1}^n X_i^2, & \mu_2(\theta) &= \mathbb{E}(X_i^2) \\ & \vdots & \vdots & \\ m_k &= \frac{1}{n}\sum_{i=1}^n X_i^k, & \mu_k(\theta) &= \mathbb{E}(X_i^k). \end{aligned}$$

Let $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ solve:

$$m_j = \mu_j(\widehat{\theta}), \quad j = 1, \ldots, k.$$

**Example 1** *$N(\beta, \sigma^2)$ with $\theta = (\beta, \sigma^2)$. Then $\mu_1 = \beta$ and $\mu_2 = \sigma^2 + \beta^2$. Equate:*

$$\frac{1}{n}\sum_{i=1}^n X_i = \widehat{\beta}, \quad \frac{1}{n}\sum_{i=1}^n X_i^2 = \widehat{\sigma}^2 + \widehat{\beta}^2$$

*to get*

$$\widehat{\beta} = \overline{X}, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2.$$

**Example 2** *Suppose*

$$X_1, \ldots, X_n \sim \text{Binomial}(k, p)$$

*where both $k$ and $p$ are unknown. We get*

$$kp = \overline{X}_n, \quad \frac{1}{n}\sum_{i=1}^n X_i^2 = kp(1-p) + k^2p^2$$

*giving*

$$\widehat{p} = \frac{\overline{X}}{k}, \quad \widehat{k} = \frac{\overline{X}^2}{\overline{X} - \frac{1}{n}\sum_i (X_i - \overline{X})^2}.$$

# 4 Maximum Likelihood

Let $\widehat{\theta}$ maximize

$$L(\theta) = p(X_1, \ldots, X_n; \theta).$$

Same as maximizing

$$\ell(\theta) = \log L(\theta).$$

Often it suffices to solve

$$\frac{\partial \ell(\theta)}{\partial \theta_j} = 0, \quad j = 1, \ldots, k.$$

**Example 3** *Binomial.* $L(p) = \prod_i p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}$ *where* $S = \sum_i X_i$. *So*

$$\ell(p) = S \log p + (n - S) \log(1 - p)$$

*and* $\widehat{p} = \overline{X}$.

**Example 4** $X_1, \ldots, X_n \sim N(\mu, 1)$.

$$L(\mu) \propto \prod_i e^{-(X_i-\mu)^2/2} \propto e^{-n(\overline{X}-\mu)^2}, \quad \ell(\mu) = -\frac{n}{2}(\overline{X} - \mu)^2$$

*and* $\widehat{\mu} = \overline{X}$. *For* $N(\mu, \sigma^2)$ *we have*

$$L(\mu, \sigma^2) \propto \prod_i \frac{1}{\sigma} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \right\}$$

*and*

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

*Set*

$$\frac{\partial \ell}{\partial \mu} = 0, \quad \frac{\partial \ell}{\partial \sigma^2} = 0$$

*to get*

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i, \quad \widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X})^2.$$

**Example 5** *Let* $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$. *Then*

$$L(\theta) = \frac{1}{\theta^n} I(\theta > X_{(n)})$$

*and so* $\widehat{\theta} = X_{(n)}$.

The mle is *equivariant*. if $\eta = g(\theta)$ then $\widehat{\eta} = g(\widehat{\theta})$. Suppose $g$ is invertible so $\eta = g(\theta)$ and $\theta = g^{-1}(\eta)$. Define $L^*(\eta) = L(\theta)$ where $\theta = g^{-1}(\eta)$. So, for any $\eta$,

$$L^*(\widehat{\eta}) = L(\widehat{\theta}) \geq L(\theta) = L^*(\eta)$$

and hence $\widehat{\eta} = g(\widehat{\theta})$ maximizes $L^*(\eta)$. For non invertible functions this is still true if we define

$$L^*(\eta) = \sup_{\theta : \tau(\theta) = \eta} L(\theta).$$

**Example 6** *Binomial. The mle is* $\widehat{p} = \overline{X}$. *Let* $\psi = \log(p/(1-p))$. *Then* $\widehat{\psi} = \log(\widehat{p}/(1-\widehat{p}))$.

Later, we will see that maximum likelihood estimators have certain optimality properties.

# 5 Bayes Estimator

Regard $\theta$ as random. Start with *prior distribition* $\pi(\theta)$. Note that $f(x|\theta)\pi(\theta) = f(x, \theta)$. Now Compute the *posterior distribition* by Bayes' theorem:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)}$$

where

$$m(x) = \int f(x|\theta)\pi(\theta) d\theta.$$

This can be written as

$$\pi(\theta|x) \propto L(\theta)\pi(\theta).$$

Now compute a point estimator from the posterior. For example:

$$\widehat{\theta} = \mathbb{E}(\theta|x) = \int \theta\pi(\theta|x)d\theta = \frac{\int \theta f(x|\theta)\pi(\theta)d\theta}{\int f(x|\theta)\pi(\theta)d\theta}.$$

This approach is controversial. We will discuss the controversey and the meaning of the prior later in the course. For now, we just think of this as a way to define an estimator.

**Example 7** *Let* $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. *Let the prior be* $p \sim \text{Beta}(\alpha, \beta)$. *Hence*

$$\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

*and*

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt.$$

*Set* $Y = \sum_i X_i$. *Then*

$$\pi(p|X) \propto \underbrace{p^Y 1-p^{n-Y}}_{\text{likelihood}} \times \underbrace{p^{\alpha-1}1-p^{\beta-1}}_{\text{prior}} \propto p^{Y+\alpha-1}1-p^{n-Y+\beta-1}.$$

*Therefore,* $p|X \sim \text{Beta}(Y+\alpha, n-Y+\beta)$. *(See page 325 for more details.) The Bayes estimator is*

$$\widetilde{p} = \frac{Y+\alpha}{(Y+\alpha)+(n-Y+\beta)} = \frac{Y+\alpha}{\alpha+\beta+n} = (1-\lambda)\widehat{p}_{mle} + \lambda\ \overline{p}$$

*where*

$$\overline{p} = \frac{\alpha}{\alpha+\beta}, \quad \lambda = \frac{\alpha+\beta}{\alpha+\beta+n}.$$

*This is an example of a* conjugate prior.

**Example 8** *Let* $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ *with* $\sigma^2$ *known. Let* $\mu \sim N(m, \tau^2)$. *Then*

$$\mathbb{E}(\mu|X) = \frac{\tau^2}{\tau^2+\frac{\sigma^2}{n}}\overline{X} + \frac{\frac{\sigma^2}{n}}{\tau^2+\frac{\sigma^2}{n}}m$$

*and*

$$\text{Var}(\mu|X) = \frac{\sigma^2\tau^2/n}{\tau^2+\frac{\sigma^2}{n}}.$$

# 6 MSE

The mean squared error (MSE) is

$$\mathbb{E}_\theta(\widehat{\theta}-\theta)^2 = \int \cdots \int (\widehat{\theta}(x_1,\ldots,x_n)-\theta)^2 f(x_1;\theta)\cdots f(x_n;\theta)dx_1\ldots dx_n.$$

The bias is

$$B = \mathbb{E}_\theta(\widehat{\theta}) - \theta$$

and the variance is

$$V = \mathrm{Var}_\theta(\widehat{\theta}).$$

**Theorem 9** *We have*

$$MSE = B^2 + V.$$

**Proof.** Let $m = \mathbb{E}_\theta(\widehat{\theta})$. Then

$$\begin{aligned} MSE &= \mathbb{E}_\theta(\widehat{\theta}-\theta)^2 = \mathbb{E}_\theta(\widehat{\theta}-m+m-\theta)^2 \\ &= \mathbb{E}_\theta(\widehat{\theta}-m)^2 + (m-\theta)^2 + 2\mathbb{E}_\theta(\widehat{\theta}-m)(m-\theta) \\ &= \mathbb{E}_\theta(\widehat{\theta}-m)^2 + (m-\theta)^2 = V + B^2. \end{aligned}$$

■

An estimator is *unbiased* if the bias is 0. In that case, the MSE = Variance. There is often a tradeoff between bias and variance. So low bias can imply high variance and vice versa.

**Example 10** *Let* $X_1,\ldots,X_n \sim N(\mu,\sigma^2)$. *Then*

$$\mathbb{E}(\overline{X}) = \mu, \quad \overline{E}(S^2) = \sigma^2.$$

*The MSE's are*

$$\mathbb{E}(\overline{X}-\mu)^2 = \frac{\sigma^2}{n}, \quad \mathbb{E}(S^2-\sigma^2)^2 = \frac{2\sigma^4}{n-1}.$$

*See p 331 for calculations.*

# 7 Best Unbiased Estimators

What is the smallest variance of an unbiased estimator? This was once considered an important question. Today we consider it not so important. There is no reason to require an estmator to be unbiased. Having small MSE is more important. However, for completeness, we will briefly consider the question.

An estimator $W$ is UMVUE (Uniform Minimum Variance Unbiased Estimator) for $\tau(\theta)$ if (i) $E_\theta(W) = \tau(\theta)$ for all $\theta$ and (ii) if $E_\theta(W') = \tau(\theta)$ for all $\theta$ then $\mathrm{Var}_\theta(W) \leq \mathrm{Var}_\theta(W')$.

The Cramer-Rao inequality gives a lower bound on the variance of any unbaised estimator. The bound is:

$$\mathrm{Var}_\theta(W) \geq \frac{\left(\frac{d}{d\theta} E_\theta W\right)^2}{E_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right)} = \frac{(\tau'(\theta))^2}{I_n(\theta)}.$$

There is also a link with sufficiency.

**Theorem 11 The Rao-Blackwell Theorem.** *Let $W$ be an unbiased estimator of $\tau(\theta)$ and let $T$ be a sufficient statistic. Define $W' = \phi(T) = E(W|T)$. Then $W'$ is unbiased and $\mathrm{Var}_\theta(W') \leq \mathrm{Var}_\theta(W)$ for all $\theta$.*

Note that $\phi$ is a well-defined estimator since, by sufficiency, it does not depend on $\theta$.

**Proof.** We have

$$E_\theta(W') = E_\theta(E(W|T)) = E_\theta(W) = \tau(\theta)$$

so $W'$ is unbiased. Also,

$$\begin{aligned}
\mathrm{Var}_\theta(W) &= \mathrm{Var}_\theta(E(W|T)) + E_\theta(\mathrm{Var}(W|T)) \\
&= \mathrm{Var}_\theta(W') + E_\theta(\mathrm{Var}(W|T)) \\
&\geq \mathrm{Var}_\theta(W').
\end{aligned}$$

■

**Ignore the material on completeness.**

# Lecture Notes 8

# 1 Minimax Theory

Suppose we want to estimate a parameter $\theta$ using data $X^n = (X_1, \ldots, X_n)$. What is the best possible estimator $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ of $\theta$? Minimax theory provides a framework for answering this question.

## 1.1 Introduction

Let $\widehat{\theta} = \widehat{\theta}(X^n)$ be an estimator for the parameter $\theta \in \Theta$. We start with a **loss function** $L(\theta, \widehat{\theta})$ that measures how good the estimator is. For example:

| | |
|---|---|
| $L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$ | squared error loss, |
| $L(\theta, \widehat{\theta}) = \lvert\theta - \widehat{\theta}\rvert$ | absolute error loss, |
| $L(\theta, \widehat{\theta}) = \lvert\theta - \widehat{\theta}\rvert^p$ | $L_p$ loss, |
| $L(\theta, \widehat{\theta}) = 0$ if $\theta = \widehat{\theta}$ or 1 if $\theta \neq \widehat{\theta}$ | zero–one loss, |
| $L(\theta, \widehat{\theta}) = I(\lvert\widehat{\theta} - \theta\rvert > c)$ | large deviation loss, |
| $L(\theta, \widehat{\theta}) = \int \log\left(\frac{p(x;\theta)}{p(x;\widehat{\theta})}\right) p(x;\, \theta)dx$ | Kullback–Leibler loss. |

If $\theta = (\theta_1, \ldots, \theta_k)$ is a vector then some common loss functions are

$$L(\theta, \widehat{\theta}) = ||\theta - \widehat{\theta}||^2 = \sum_{j=1}^{k} (\widehat{\theta}_j - \theta_j)^2,$$

$$L(\theta, \widehat{\theta}) = ||\theta - \widehat{\theta}||_p = \left( \sum_{j=1}^{k} |\widehat{\theta}_j - \theta_j|^p \right)^{1/p}.$$

When the problem is to predict a $Y \in \{0, 1\}$ based on some classifier $h(x)$ a commonly used loss is

$$L(Y, h(X)) = I(Y \neq h(X)).$$

For real valued prediction a common loss function is

$$L(Y, \widehat{Y}) = (Y - \widehat{Y})^2.$$

The **risk** of an estimator $\widehat{\theta}$ is

$$R(\theta, \widehat{\theta}) = \mathbb{E}_\theta\Big( L(\theta, \widehat{\theta}) \Big) = \int L(\theta, \widehat{\theta}(x_1, \ldots, x_n)) p(x_1, \ldots, x_n; \theta) dx. \tag{1}$$

When the loss function is squared error, the risk is just the MSE (mean squared error):

$$R(\theta, \widehat{\theta}) = \mathbb{E}_\theta(\widehat{\theta} - \theta)^2 = \mathsf{Var}_\theta(\widehat{\theta}) + \text{bias}^2. \tag{2}$$

If we do not state what loss function we are using, assume the loss function is squared error.

The **minimax risk** is

$$R_n = \inf_{\widehat{\theta}} \sup_{\theta} R(\theta, \widehat{\theta})$$

where the infimum is over all estimators. An estimator $\widehat{\theta}$ is **a minimax estimator** if

$$\sup_{\theta} R(\theta, \widehat{\theta}) = \inf_{\widehat{\theta}} \sup_{\theta} R(\theta, \widehat{\theta}).$$

**Example 1** *Let $X_1, \ldots, X_n \sim N(\theta, 1)$. We will see that $\overline{X}_n$ is minimax with respect to many different loss functions. The risk is $1/n$.*

**Example 2** *Let $X_1, \ldots, X_n$ be a sample from a density $f$. Let $\mathcal{F}$ be the class of smooth densities (defined more precisely later). We will see (later in the course) that the minimax risk for estimating $f$ is $Cn^{-4/5}$.*

## 1.2 Comparing Risk Functions

To compare two estimators, we compare their risk functions. However, this does not provide a clear answer as to which estimator is better. Consider the following examples.

**Example 3** *Let $X \sim N(\theta, 1)$ and assume we are using squared error loss. Consider two estimators: $\widehat{\theta}_1 = X$ and $\widehat{\theta}_2 = 3$. The risk functions are $R(\theta, \widehat{\theta}_1) = \mathbb{E}_\theta(X - \theta)^2 = 1$ and $R(\theta, \widehat{\theta}_2) = \mathbb{E}_\theta(3 - \theta)^2 = (3 - \theta)^2$. If $2 < \theta < 4$ then $R(\theta, \widehat{\theta}_2) < R(\theta, \widehat{\theta}_1)$, otherwise, $R(\theta, \widehat{\theta}_1) < R(\theta, \widehat{\theta}_2)$. Neither estimator uniformly dominates the other; see Figure 1.*

**Example 4** *Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Consider squared error loss and let $\widehat{p}_1 = \overline{X}$. Since this has zero bias, we have that*

$$R(p, \widehat{p}_1) = \mathsf{Var}(\overline{X}) = \frac{p(1-p)}{n}.$$

*Another estimator is*

$$\widehat{p}_2 = \frac{Y + \alpha}{\alpha + \beta + n}$$

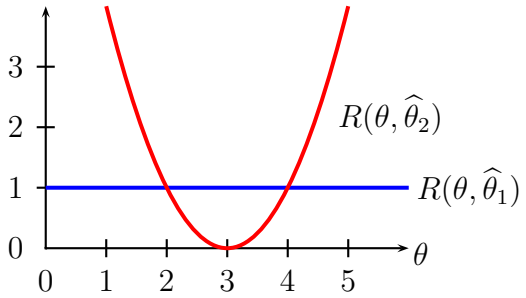



Figure 1: Comparing two risk functions. Neither risk function dominates the other at all values of $\theta$.

*where* $Y = \sum_{i=1}^n X_i$ *and* $\alpha$ *and* $\beta$ *are positive constants.*[1] *Now,*

$$\begin{aligned} R(p, \widehat{p}_2) &= \mathsf{Var}_p(\widehat{p}_2) + (\mathrm{bias}_p(\widehat{p}_2))^2 \\ &= \mathsf{Var}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) + \left(\mathbb{E}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) - p\right)^2 \\ &= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left(\frac{np+\alpha}{\alpha+\beta+n} - p\right)^2. \end{aligned}$$

*Let* $\alpha = \beta = \sqrt{n/4}$. *The resulting estimator is*

$$\widehat{p}_2 = \frac{Y + \sqrt{n/4}}{n + \sqrt{n}}$$

*and the risk function is*

$$R(p, \widehat{p}_2) = \frac{n}{4(n+\sqrt{n})^2}.$$

*The risk functions are plotted in figure 2. As we can see, neither estimator uniformly dominates the other.*

These examples highlight the need to be able to compare risk functions. To do so, we need a one-number summary of the risk function. Two such summaries are the maximum risk and the Bayes risk.

The **maximum risk** is

$$\overline{R}(\widehat{\theta}) = \sup_{\theta \in \Theta} R(\theta, \widehat{\theta}) \tag{3}$$

[1] This is the posterior mean using a Beta $(\alpha, \beta)$ prior.

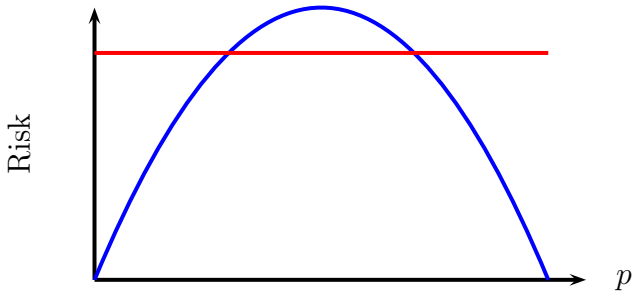



Figure 2: Risk functions for $\widehat{p}_1$ and $\widehat{p}_2$ in Example 4. The solid curve is $R(\widehat{p}_1)$. The dotted line is $R(\widehat{p}_2)$.

and the **Bayes risk** under prior $\pi$ is

$$B_\pi(\widehat{\theta}) = \int R(\theta, \widehat{\theta})\pi(\theta)d\theta. \tag{4}$$

**Example 5** *Consider again the two estimators in Example 4. We have*

$$\overline{R}(\widehat{p}_1) = \max_{0\leq p\leq 1} \frac{p(1-p)}{n} = \frac{1}{4n}$$

*and*

$$\overline{R}(\widehat{p}_2) = \max_p \frac{n}{4(n+\sqrt{n})^2} = \frac{n}{4(n+\sqrt{n})^2}.$$

*Based on maximum risk, $\widehat{p}_2$ is a better estimator since $\overline{R}(\widehat{p}_2) < \overline{R}(\widehat{p}_1)$. However, when $n$ is large, $\overline{R}(\widehat{p}_1)$ has smaller risk except for a small region in the parameter space near $p = 1/2$. Thus, many people prefer $\widehat{p}_1$ to $\widehat{p}_2$. This illustrates that one-number summaries like maximum risk are imperfect.*

These two summaries of the risk function suggest two different methods for devising estimators: choosing $\widehat{\theta}$ to minimize the maximum risk leads to minimax estimators; choosing $\widehat{\theta}$ to minimize the Bayes risk leads to Bayes estimators.

An estimator $\widehat{\theta}$ that minimizes the Bayes risk is called a **Bayes estimator**. That is,

$$B_\pi(\widehat{\theta}) = \inf_{\widetilde{\theta}} B_\pi(\widetilde{\theta}) \tag{5}$$

where the infimum is over all estimators $\widetilde{\theta}$. An estimator that minimizes the maximum risk is called a **minimax estimator**. That is,

$$\sup_{\theta} R(\theta, \widehat{\theta}) = \inf_{\widetilde{\theta}} \sup_{\theta} R(\theta, \widetilde{\theta}) \tag{6}$$

where the infimum is over all estimators $\widetilde{\theta}$. We call the right hand side of (6), namely,

$$R_n \equiv R_n(\Theta) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} R(\theta, \widehat{\theta}), \tag{7}$$

the **minimax risk**. Statistical decision theory has two goals: determine the minimax risk $R_n$ and find an estimator that achieves this risk.

Once we have found the minimax risk $R_n$ we want to find the minimax estimator that achieves this risk:

$$\sup_{\theta \in \Theta} R(\theta, \widehat{\theta}) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} R(\theta, \widehat{\theta}). \tag{8}$$

Sometimes we settle for an asymptotically minimax estimator

$$\sup_{\theta \in \Theta} R(\theta, \widehat{\theta}) \sim \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} R(\theta, \widehat{\theta}) \quad n \to \infty \tag{9}$$

where $a_n \sim b_n$ means that $a_n/b_n \to 1$. Even that can prove too difficult and we might settle for an estimator that achieves the minimax rate,

$$\sup_{\theta \in \Theta} R(\theta, \widehat{\theta}) \asymp \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} R(\theta, \widehat{\theta}) \quad n \to \infty \tag{10}$$

where $a_n \asymp b_n$ means that both $a_n/b_n$ and $b_n/a_n$ are both bounded as $n \to \infty$.

## 1.3 Bayes Estimators

Let $\pi$ be a prior distribution. After observing $X^n = (X_1, \ldots, X_n)$, the posterior distribution is, according to Bayes' theorem,

$$\mathbb{P}(\theta \in A | X^n) = \frac{\int_A p(X_1, \ldots, X_n|\theta)\pi(\theta)d\theta}{\int_\Theta p(X_1, \ldots, X_n|\theta)\pi(\theta)d\theta} = \frac{\int_A \mathcal{L}(\theta)\pi(\theta)d\theta}{\int_\Theta \mathcal{L}(\theta)\pi(\theta)d\theta} \tag{11}$$

where $\mathcal{L}(\theta) = p(x^n; \theta)$ is the likelihood function. The posterior has density

$$\pi(\theta|x^n) = \frac{p(x^n|\theta)\pi(\theta)}{m(x^n)} \tag{12}$$

where $m(x^n) = \int p(x^n|\theta)\pi(\theta)d\theta$ is the **marginal distribution** of $X^n$. Define the **posterior risk** of an estimator $\widehat{\theta}(x^n)$ by

$$r(\widehat{\theta}|x^n) = \int L(\theta, \widehat{\theta}(x^n))\pi(\theta|x^n)d\theta. \tag{13}$$

**Theorem 6** *The Bayes risk $B_\pi(\widehat{\theta})$ satisfies*

$$B_\pi(\widehat{\theta}) = \int r(\widehat{\theta}|x^n)m(x^n)\,dx^n. \tag{14}$$

*Let $\widehat{\theta}(x^n)$ be the value of $\theta$ that minimizes $r(\widehat{\theta}|x^n)$. Then $\widehat{\theta}$ is the Bayes estimator.*

**Proof.**Let $p(x,\theta) = p(x|\theta)\pi(\theta)$ denote the joint density of $X$ and $\theta$. We can rewrite the Bayes risk as follows:

$$\begin{aligned} B_\pi(\widehat{\theta}) &= \int R(\theta,\widehat{\theta})\pi(\theta)d\theta = \int\left(\int L(\theta,\widehat{\theta}(x^n))p(x|\theta)dx^n\right)\pi(\theta)d\theta \\ &= \int\int L(\theta,\widehat{\theta}(x^n))p(x,\theta)dx^n d\theta = \int\int L(\theta,\widehat{\theta}(x^n))\pi(\theta|x^n)m(x^n)dx^n d\theta \\ &= \int\left(\int L(\theta,\widehat{\theta}(x^n))\pi(\theta|x^n)d\theta\right)m(x^n)\,dx^n = \int r(\widehat{\theta}|x^n)m(x^n)\,dx^n. \end{aligned}$$

If we choose $\widehat{\theta}(x^n)$ to be the value of $\theta$ that minimizes $r(\widehat{\theta}|x^n)$ then we will minimize the integrand at every $x$ and thus minimize the integral $\int r(\widehat{\theta}|x^n)m(x^n)dx^n$.

Now we can find an explicit formula for the Bayes estimator for some specific loss functions.

**Theorem 7** *If $L(\theta,\widehat{\theta}) = (\theta-\widehat{\theta})^2$ then the Bayes estimator is*

$$\widehat{\theta}(x^n) = \int \theta\pi(\theta|x^n)d\theta = \mathbb{E}(\theta|X=x^n). \tag{15}$$

*If $L(\theta,\widehat{\theta}) = |\theta-\widehat{\theta}|$ then the Bayes estimator is the median of the posterior $\pi(\theta|x^n)$. If $L(\theta,\widehat{\theta})$ is zero–one loss, then the Bayes estimator is the mode of the posterior $\pi(\theta|x^n)$.*

**Proof.**We will prove the theorem for squared error loss. The Bayes estimator $\widehat{\theta}(x^n)$ minimizes $r(\widehat{\theta}|x^n) = \int(\theta-\widehat{\theta}(x^n))^2\pi(\theta|x^n)d\theta$. Taking the derivative of $r(\widehat{\theta}|x^n)$ with respect to $\widehat{\theta}(x^n)$ and setting it equal to zero yields the equation $2\int(\theta-\widehat{\theta}(x^n))\pi(\theta|x^n)d\theta = 0$. Solving for $\widehat{\theta}(x^n)$ we get 15.

**Example 8** *Let $X_1,\ldots,X_n \sim N(\mu,\sigma^2)$ where $\sigma^2$ is known. Suppose we use a $N(a,b^2)$ prior for $\mu$. The Bayes estimator with respect to squared error loss is the posterior mean, which is*

$$\widehat{\theta}(X_1,\ldots,X_n) = \frac{b^2}{b^2+\frac{\sigma^2}{n}}\overline{X} + \frac{\frac{\sigma^2}{n}}{b^2+\frac{\sigma^2}{n}}a. \quad \blacksquare \tag{16}$$

## 1.4 Minimax Estimators

Finding minimax estimators is complicated and we cannot attempt a complete coverage of that theory here but we will mention a few key results. The main message to take away from this section is: **Bayes estimators with a constant risk function are minimax**.

**Theorem 9** *Let $\widehat{\theta}$ be the Bayes estimator for some prior $\pi$. If*

$$R(\theta, \widehat{\theta}) \leq B_\pi(\widehat{\theta}) \text{ for all } \theta \tag{17}$$

*then $\widehat{\theta}$ is minimax and $\pi$ is called a* **least favorable prior**.

**Proof.**Suppose that $\widehat{\theta}$ is not minimax. Then there is another estimator $\widehat{\theta}_0$ such that $\sup_\theta R(\theta, \widehat{\theta}_0) < \sup_\theta R(\theta, \widehat{\theta})$. Since the average of a function is always less than or equal to its maximum, we have that $B_\pi(\widehat{\theta}_0) \leq \sup_\theta R(\theta, \widehat{\theta}_0)$. Hence,

$$B_\pi(\widehat{\theta}_0) \leq \sup_\theta R(\theta, \widehat{\theta}_0) < \sup_\theta R(\theta, \widehat{\theta}) \leq B_\pi(\widehat{\theta}) \tag{18}$$

which is a contradiction.

**Theorem 10** *Suppose that $\widehat{\theta}$ is the Bayes estimator with respect to some prior $\pi$. If the risk is constant then $\widehat{\theta}$ is minimax.*

**Proof.**The Bayes risk is $B_\pi(\widehat{\theta}) = \int R(\theta, \widehat{\theta})\pi(\theta)d\theta = c$ and hence $R(\theta, \widehat{\theta}) \leq B_\pi(\widehat{\theta})$ for all $\theta$. Now apply the previous theorem.

**Example 11** *Consider the Bernoulli model with squared error loss. In example 4 we showed that the estimator*

$$\widehat{p}(X^n) = \frac{\sum_{i=1}^n X_i + \sqrt{n/4}}{n + \sqrt{n}}$$

*has a constant risk function. This estimator is the posterior mean, and hence the Bayes estimator, for the prior* $\text{Beta}(\alpha, \beta)$ *with $\alpha = \beta = \sqrt{n/4}$. Hence, by the previous theorem, this estimator is minimax.*

**Example 12** *Consider again the Bernoulli but with loss function*

$$L(p, \widehat{p}) = \frac{(p - \widehat{p})^2}{p(1-p)}.$$

*Let $\widehat{p}(X^n) = \widehat{p} = \sum_{i=1}^n X_i/n$. The risk is*

$$R(p, \widehat{p}) = E\left(\frac{(\widehat{p} - p)^2}{p(1-p)}\right) = \frac{1}{p(1-p)}\left(\frac{p(1-p)}{n}\right) = \frac{1}{n}$$

*which, as a function of $p$, is constant. It can be shown that, for this loss function, $\widehat{p}(X^n)$ is the Bayes estimator under the prior $\pi(p) = 1$. Hence, $\widehat{p}$ is minimax.*

What is the minimax estimator for a Normal model? To answer this question in generality we first need a definition. A function $\ell$ is **bowl-shaped** if the sets $\{x : \ell(x) \leq c\}$ are convex and symmetric about the origin. A loss function $L$ is bowl-shaped if $L(\theta, \widehat{\theta}) = \ell(\theta - \widehat{\theta})$ for some bowl-shaped function $\ell$.

**Theorem 13** *Suppose that the random vector $X$ has a Normal distribution with mean vector $\theta$ and covariance matrix $\Sigma$. If the loss function is bowl-shaped then $X$ is the unique (up to sets of measure zero) minimax estimator of $\theta$.*

If the parameter space is restricted, then the theorem above does not apply as the next example shows.

**Example 14** *Suppose that $X \sim N(\theta, 1)$ and that $\theta$ is known to lie in the interval $[-m, m]$ where $0 < m < 1$. The unique, minimax estimator under squared error loss is*

$$\widehat{\theta}(X) = m\left(\frac{e^{mX} - e^{-mX}}{e^{mX} + e^{-mX}}\right).$$

*This is the Bayes estimator with respect to the prior that puts mass 1/2 at $m$ and mass 1/2 at $-m$. The risk is not constant but it does satisfy $R(\theta, \widehat{\theta}) \leq B_\pi(\widehat{\theta})$ for all $\theta$; see Figure 3. Hence, Theorem 9 implies that $\widehat{\theta}$ is minimax. This might seem like a toy example but it is not. The essence of modern minimax theory is that the minimax risk depends crucially on how the space is restricted. The bounded interval case is the tip of the iceberg.*

**Proof That $\overline{X}_n$ is Minimax Under Squared Error Loss.** Now we will explain why $\overline{X}_n$ is justified by minimax theory. Let $X \sim N_p(\theta, I)$ be multivariate Normal with mean vector $\theta = (\theta_1, \ldots, \theta_p)$. We will prove that $\widehat{\theta} = X$ is minimax when $L(\theta, \widehat{\theta}) = ||\widehat{\theta} - \theta||^2$. Assign the prior $\pi = N(0, c^2 I)$. Then the posterior is

$$\Theta | X = x \ \sim N\left(\frac{c^2 x}{1 + c^2}, \frac{c^2}{1 + c^2} I\right). \tag{19}$$

The Bayes risk for an estimator $\widehat{\theta}$ is $R_\pi(\widehat{\theta}) = \int R(\theta, \widehat{\theta})\pi(\theta)d\theta$ which is minimized by the posterior mean $\widetilde{\theta} = c^2 X/(1 + c^2)$. Direct computation shows that $R_\pi(\widetilde{\theta}) = pc^2/(1 + c^2)$. Hence, if $\theta^*$ is any estimator, then

$$\begin{aligned} \frac{pc^2}{1 + c^2} &= R_\pi(\widetilde{\theta}) \leq R_\pi(\theta^*) && (20) \\ &= \int R(\theta^*, \theta)d\pi(\theta) \leq \sup_\theta R(\theta^*, \theta). && (21) \end{aligned}$$

We have now proved that $R(\Theta) \geq pc^2/(1 + c^2)$ for every $c > 0$ and hence

$$R(\Theta) \geq p. \tag{22}$$

But the risk of $\widehat{\theta} = X$ is $p$. So, $\widehat{\theta} = X$ is minimax.

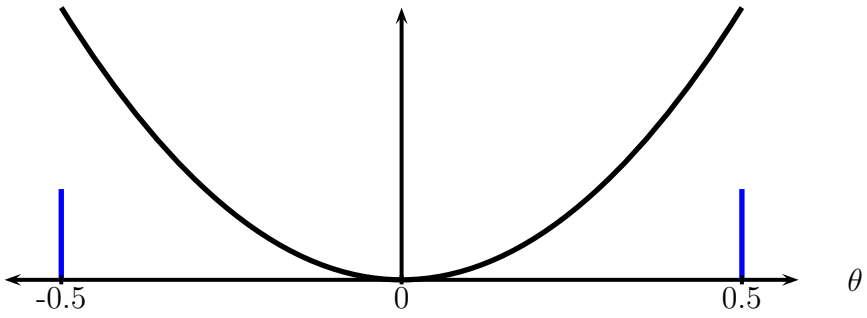



Figure 3: Risk function for constrained Normal with m=.5. The two short dashed lines show the least favorable prior which puts its mass at two points.

## 1.5 Maximum Likelihood

For parametric models that satisfy weak regularity conditions, the maximum likelihood estimator is approximately minimax. Consider squared error loss which is squared bias plus variance. In parametric models with large samples, it can be shown that the variance term dominates the bias so the risk of the mle $\widehat{\theta}$ roughly equals the variance:[2]

$$R(\theta, \widehat{\theta}) = \mathsf{Var}_\theta(\widehat{\theta}) + \text{bias}^2 \approx \mathsf{Var}_\theta(\widehat{\theta}). \tag{23}$$

The variance of the mle is approximately $\mathsf{Var}(\widehat{\theta}) \approx \frac{1}{nI(\theta)}$ where $I(\theta)$ is the Fisher information. Hence,

$$nR(\theta, \widehat{\theta}) \approx \frac{1}{I(\theta)}. \tag{24}$$

For any other estimator $\theta'$, it can be shown that for large $n$, $R(\theta, \theta') \geq R(\theta, \widehat{\theta})$. So **the maximum likelihood estimator is approximately minimax. This assumes that the dimension of $\theta$ is fixed and $n$ is increasing.**

## 1.6 The Hodges Example

Here is an interesting example about the subtleties of optimal estimators. Let $X_1, \ldots, X_n \sim N(\theta, 1)$. The mle is $\widehat{\theta}_n = \overline{X}_n = n^{-1} \sum_{i=1}^n X_i$. But consider the following estimator due to

[2]Typically, the squared bias is order $O(n^{-2})$ while the variance is of order $O(n^{-1})$.

Hodges. Let

$$J_n = \left[ -\frac{1}{n^{1/4}}, \ \frac{1}{n^{1/4}} \right] \tag{25}$$

and define

$$\widetilde{\theta}_n = \begin{cases} \overline{X}_n & \text{if } \overline{X}_n \notin J_n \\ 0 & \text{if } \overline{X}_n \in J_n. \end{cases} \tag{26}$$

Suppose that $\theta \neq 0$. Choose a small $\epsilon$ so that 0 is not contained in $I = (\theta - \epsilon, \theta + \epsilon)$. By the law of large numbers, $\mathbb{P}(\overline{X}_n \in I) \to 1$. In the meantime $J_n$ is shrinking. See Figure 4. Thus, for $n$ large, $\widetilde{\theta}_n = \overline{X}_n$ with high probability. We conclude that, for any $\theta \neq 0$, $\widetilde{\theta}_n$ behaves like $\overline{X}_n$.

When $\theta = 0$,

$$\begin{aligned} \mathbb{P}(\overline{X}_n \in J_n) &= \mathbb{P}(|\overline{X}_n| \leq n^{-1/4}) && (27) \\ &= \mathbb{P}(\sqrt{n}|\overline{X}_n| \leq n^{1/4}) = \mathbb{P}(|N(0,1)| \leq n^{1/4}) \to 1. && (28) \end{aligned}$$

Thus, for $n$ large, $\widetilde{\theta}_n = 0 = \theta$ with high probability. This is a much better estimator of $\theta$ than $\overline{X}_n$.

We conclude that Hodges estimator is like $\overline{X}_n$ when $\theta \neq 0$ and is better than $\overline{X}_n$ when $\theta = 0$. So $\overline{X}_n$ is not the best estimator. $\widetilde{\theta}_n$ is better.

Or is it? Figure 5 shows the mean squared error, or **risk**, $R_n(\theta) = \mathbb{E}(\widetilde{\theta}_n - \theta)^2$ as a function of $\theta$ (for $n = 1000$). The horizontal line is the risk of $\overline{X}_n$. The risk of $\widetilde{\theta}_n$ is good at $\theta = 0$. At any $\theta$, it will eventually behave like the risk of $\overline{X}_n$. But the maximum risk of $\widetilde{\theta}_n$ is terrible. We pay for the improvement at $\theta = 0$ by an increase in risk elsewhere.

There are two lessons here. First, we need to pay attention to the maximum risk. Second, it is better to look at uniform asymptotics $\lim_{n\to\infty} \sup_\theta R_n(\theta)$ rather than pointwise asymptotics $\sup_\theta \lim_{n\to\infty} R_n(\theta)$.

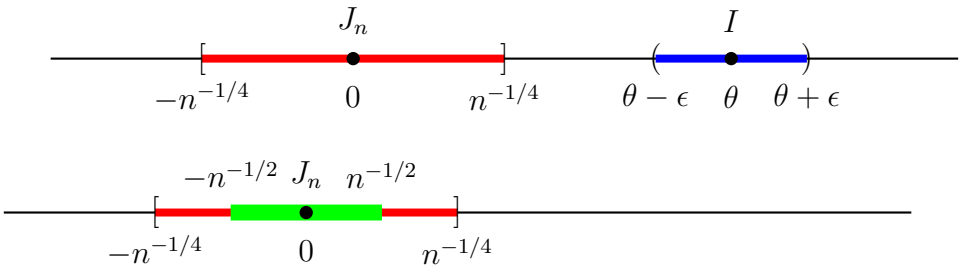



Figure 4: Top: when $\theta \neq 0$, $\overline{X}_n$ will eventually be in $I$ and will miss the interval $J_n$. Bottom: when $\theta = 0$, $\overline{X}_n$ is about $n^{-1/2}$ away from 0 and so is eventually in $J_n$.

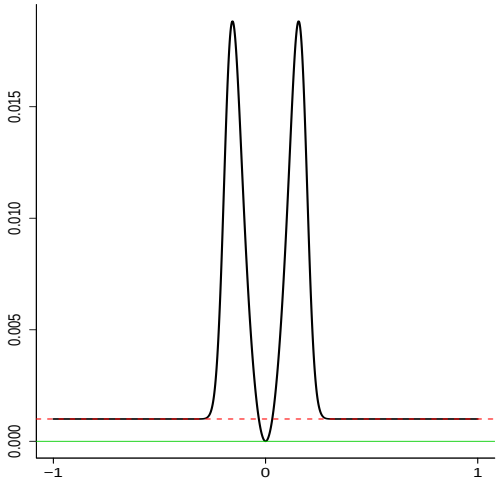



Figure 5: The risk of the Hodges estimator for $n = 1000$ as a function of $\theta$. The horizontal line is the risk of the sample mean.

**36-705/10-705**
**Summary of Minimax Theory**
**Larry Wasserman**
October 5, 2011

1. Loss $L(\theta, \widehat{\theta})$ where $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$. Remember that $\widehat{\theta}$ is a function of $X_1, \ldots, X_n$.
2. Risk

$$R(\theta, \widehat{\theta}) = E_\theta[L(\theta, \widehat{\theta})] = \int \cdots \int L(\theta, \widehat{\theta}(x_1, \ldots, x_n)) p(x_1, \ldots, x_n; \theta) dx_1 \cdots dx_n.$$

3. If $L(\theta, \widehat{\theta}) = (\widehat{\theta} - \theta)^2$ then

$$R(\theta, \widehat{\theta}) = E_\theta(\widehat{\theta} - \theta)^2 = \text{MSE} = \text{bias}^2 + \text{variance}.$$

4. Maximum risk: we define how good an estimator is by its maximum risk

$$\sup_\theta R(\theta, \widehat{\theta}).$$

5. Minimax risk:

$$R_n = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} R(\theta, \widehat{\theta}).$$

6. An estimator $\widehat{\theta}$ is minimax if

$$\sup_{\theta \in \Theta} R(\theta, \widehat{\theta}) = R_n.$$

7. The Bayes risk for an estimator $\widehat{\theta}$, with respect to a prior $\pi$ is

$$B_\pi(\widehat{\theta}) = \int R(\theta, \widehat{\theta}) \pi(\theta) d\theta.$$

8. An estimator $\widehat{\theta}_\pi$ is the Bayes estimator with respect to a prior $\pi$ if

$$B_\pi(\widehat{\theta}_\pi) = \inf_{\widehat{\theta}} B_\pi(\widehat{\theta}).$$

   In other words, $\widehat{\theta}_\pi$ minimizes $B_\pi(\widehat{\theta})$ over all estimators.
9. The Bayes risk can we re-written as

$$B_\pi(\widehat{\theta}) = \int r(\widehat{\theta})\, m(x_1, \ldots, x_n) dx_1 \cdots dx_n$$

   where $m(x_1, \ldots, x_n) = \int p(x_1, \ldots, x_n; \theta) \pi(\theta) d\theta$ and $r(\widehat{\theta}) = \int L(\theta, \widehat{\theta}) p(\theta | x_1, \ldots, x_n) d\theta$. Hence, to minimize $B_\pi(\widehat{\theta}_\pi)$ is suffices to minimize $r(\widehat{\theta})$.
10. **Key Theorem:** Suppose that (i) $\widehat{\theta}$ is the Bayes estimator with respect to some prior $\pi$ and (ii) $R(\theta, \widehat{\theta})$ is constant. Then $\widehat{\theta}$ is minimax.

11. **Bounds.** Sometimes it is hard to find $R_n$ so it is useful to find a lower bound and an upper bound on the minimax risk. The following result is helpful:
**Theorem:** Let $\widehat{\theta}_\pi$ by a Bayes estimator with respect to some prior $\pi$. Let $\widehat{\theta}^*$ be any estimator. Then:

$$B_\pi(\widehat{\theta}_\pi) \leq R_n \leq \sup_\theta R(\theta, \widehat{\theta}^*). \tag{1}$$

**Proof of the Lower Bound.** Let $\widehat{\theta}_\pi$ be the Bayes estimator for some prior $\pi$. Let $\widehat{\theta}$ be any other estimator. Then,

$$B_\pi(\widehat{\theta}_\pi) \leq B_\pi(\widehat{\theta}) = \int R(\theta, \widehat{\theta})\pi(\theta)d\theta \leq \sup_\theta R(\theta, \widehat{\theta}).$$

Take the inf over all $\widehat{\theta}$ and conclude that

$$B_\pi(\widehat{\theta}_\pi) \leq \inf_{\widehat{\theta}_n} \sup_\theta R(\theta, \widehat{\theta}) = R_n.$$

Hence, $R_n \geq B_\pi(\widehat{\theta}_\pi)$.
**Proof of the Upper bound.** Choose any estimator $\widehat{\theta}^*$. Then

$$R_n = \inf_{\widehat{\theta}} \sup_\theta R(\theta, \widehat{\theta}) \leq \sup_\theta R(\theta, \widehat{\theta}^*).$$

12. How to prove that $\overline{X}_n$ is minimax for the Normal model. Let $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ where $\sigma^2$ is known. Let $L(\theta, \widehat{\theta}) = (\widehat{\theta} - \theta)^2$.
    (a) First we show that $R_n = \sigma^2/n$. We do this by getting a lower bound and an upper bound on $R_n$.
    (b) **Lower Bound.** Let $\pi = N(0, c^2)$. The posterior $p(\theta|X_1, \ldots, X_n)$ is $N(a, b^2)$ where

$$a = \frac{n\overline{X}/\sigma^2}{\frac{1}{c^2} + \frac{n}{\sigma^2}} \quad \text{and} \quad b^2 = \frac{1}{\frac{1}{c^2} + \frac{n}{\sigma^2}}.$$

The Bayes estimator minimizes $r(\widehat{\theta}) = \int (\widehat{\theta} - \theta)^2 p(\theta|x_1, \ldots, x_n)d\theta$. This is minimized by $\widehat{\theta}_\pi = \int \theta p(\theta|x_1, \ldots, x_n)d\theta = E(\theta|X_1, \ldots, X_n)$. But $E(\theta|X_1, \ldots, X_n) = a$. So the Bayes esimator is

$$\widehat{\theta}_\pi = \frac{n\overline{X}/\sigma^2}{\frac{1}{c^2} + \frac{n}{\sigma^2}}.$$

Next we compute $R(\widehat{\theta}_\pi, \theta)$. This means we need to compute the MSE of $\widehat{\theta}_\pi$. The bias $\widehat{\theta}_\pi$ is $-\theta\sigma^2/(\sigma^2 + nc^2)$. The variance of $\widehat{\theta}_\pi$ is $nc^4\sigma^2/(\sigma^2 + nc^2)^2$. So

$$R(\theta, \widehat{\theta}_\pi) = \text{bias}^2 + \text{variance} = \frac{\theta^2\sigma^4}{(\sigma^2 + nc^2)^2} + \frac{nc^4\sigma^2}{(\sigma^2 + nc^2)^2} = \frac{\theta^2\sigma^4 + nc^4\sigma^2}{(\sigma^2 + nc^2)^2}.$$

Let us now compute the Bayes risk of this estimator. It is

$$\begin{aligned} B_\pi(\widehat{\theta}_\pi) &= \int R(\theta, \widehat{\theta}_\pi)\pi(\theta)d\theta = \frac{\sigma^4 \int \theta^2 \pi(\theta)d\theta + nc^4\sigma^2}{(\sigma^2 + nc^2)^2} \\ &= \frac{\sigma^4 c^2 + nc^4\sigma^2}{(\sigma^2 + nc^2)^2} = \frac{\sigma^2}{\frac{\sigma^2}{c^2} + n}. \end{aligned}$$

By (1), this proves that

$$R_n \geq \frac{\sigma^2}{\frac{\sigma^2}{c^2} + n}.$$

(c) **Upper Bound.** Choose $\widehat{\theta} = \overline{X}_n$. Then $R(\theta, \widehat{\theta}) = \sigma^2/n$. By (1),

$$R_n \leq \sup_\theta R(\theta, \widehat{\theta}) = \frac{\sigma^2}{n}.$$

(d) Combining the lower and upper bound we see that

$$\frac{\sigma^2}{\frac{\sigma^2}{c^2} + n} \leq R_n \leq \frac{\sigma^2}{n}.$$

This bound is true for all $c > 0$. If take the limit as $c \to \infty$ then we get that $R_n = \frac{\sigma^2}{n}$. We have succeeded in finding the minimax risk $R_n$.

(e) The last step is to find a minimax estimator. We have to find an estimator whose maximum risk is $R_n$. But we already saw that $\overline{X}$ has maximum risk equal to $R_n$. Hence $\overline{X}_n$ is minimax.

# Lecture Notes 9

# Asymptotic (Large Sample) Theory

## 1 Review of $o$, $O$, etc.

1. $a_n = o(1)$ mean $a_n \to 0$ as $n \to \infty$.
2. A random sequence $A_n$ is $o_p(1)$ if $A_n \xrightarrow{\text{P}} 0$ as $n \to \infty$.
3. A random sequence $A_n$ is $o_p(b_n)$ if $A_n/b_n \xrightarrow{\text{P}} 0$ as $n \to \infty$.
4. $n^p\, o_p(1) = o_p(n^p)$, so $\sqrt{n}\, o_p(1/\sqrt{n}) = o_p(1) \xrightarrow{\text{P}} 0$.
5. $o_p(1) \times o_p(1) = o_p(1)$.

1. $a_n = O(1)$ if $|a_n|$ is bounded by a constant as $n \to \infty$.
2. A random sequence $Y_n$ is $O_p(1)$ if for every $\epsilon > 0$ there exists a constant $M$ such that $\lim_{n\to\infty} P(|Y_n| > M) < \epsilon$ as $n \to \infty$.
3. A random sequence $Y_n$ is $O_p(b_n)$ if $Y_n/b_n$ is $O_p(1)$.
4. If $Y_n \rightsquigarrow Y$, then $Y_n$ is $O_p(1)$.
5. If $\sqrt{n}(Y_n - c) \rightsquigarrow Y$ then $Y_n = O_P(1/\sqrt{n})$. (potential test qustion: prove this)
6. $O_p(1) \times O_p(1) = O_p(1)$.
7. $o_p(1) \times O_p(1) = o_p(1)$.

## 2 Distances Between Probability Distributions

Let $P$ and $Q$ be distributions with densities $p$ and $q$. We will use the following distances between $P$ and $Q$.

1. Total variation distance $V(P, Q) = \sup_A |P(A) - Q(A)|$.

2. $L_1$ distance $d_1(P,Q) = \int |p - q|$.

3. Hellinger distance $h(P,Q) = \sqrt{\int (\sqrt{p} - \sqrt{q})^2}$.

4. Kullback-Leibler distance $K(P,Q) = \int p \log(p/q)$.

5. $L_2$ distance $d_2(P,Q) = \int (p - q)^2$.

Here are some properties of these distances:

1. $V(P,Q) = \frac{1}{2} d_1(P,Q)$. (prove this!)

2. $h^2(P,Q) = 2(1 - \int \sqrt{pq})$.

3. $V(P,Q) \leq h(P,Q) \leq \sqrt{2V(P,Q)}$.

4. $h^2(P,Q) \leq K(P,Q)$.

5. $V(P,Q) \leq h(P,Q) \leq \sqrt{K(P,Q)}$.

6. $V(P,Q) \leq \sqrt{K(P,Q)/2}$.

# 3 Consistency

$\widehat{\theta}_n = T(X^n)$ is *consistent* for $\theta$ if

$$\widehat{\theta}_n \xrightarrow{\text{P}} \theta$$

as $n \to \infty$. In other words, $\widehat{\theta}_n - \theta = o_p(1)$. Here are two common ways to prove that $\widehat{\theta}_n$ consistent.

**Method 1:** Show that, for all $\varepsilon > 0$,

$$\mathbb{P}(|\widehat{\theta}_n - \theta| \geq \varepsilon) \longrightarrow 0.$$

**Method 2.** Prove convergence in quadratic mean:

$$\text{MSE}(\widehat{\theta}_n) = \text{Bias}^2(\widehat{\theta}_n) + \text{Var}(\widehat{\theta}_n) \longrightarrow 0.$$

If bias $\to 0$ and var $\to 0$ then $\widehat{\theta}_n \xrightarrow{qm} \theta$ which implies that $\widehat{\theta}_n \xrightarrow{p} \theta$.

**Example 1** *Bernoulli(p). The mle $\widehat{p}$ has bias 0 and variance $p(1-p)/n \to 0$. So $\widehat{p} \xrightarrow{\text{P}} p$ and is consistent. Now let $\psi = \log(p/(1-p))$. Then $\widehat{\psi} = \log(\widehat{p}/(1-\widehat{p}))$. Now $\widehat{\psi} = g(\widehat{p})$ where $g(p) = \log(p/(1-p))$. By the continuous mapping theorem, $\widehat{\psi} \xrightarrow{\text{P}} \psi$ so this is consistent. Now consider*

$$\widehat{p} = \frac{X+1}{n+1}.$$

*Then*

$$\text{bias} = E(\widehat{p}) - p = -\frac{p-1}{n(1+n)} \to 0$$

*and*

$$\text{var} = \frac{p(1-p)}{n} \to 0.$$

*So this is consistent.*

**Example 2** $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$. *Let $\widehat{\theta}_n = X_{(n)}$. By direct proof (we did it earlier) we have $\widehat{\theta}_n \xrightarrow{\text{P}} \theta$.*

Method of moments estimators are typically consistent. Consider one parameter. Recall that $\mu(\widehat{\theta}) = m$ where $m = n^{-1}\sum_{i=1}^n X_i$. Assume that $\mu^{-1}$ exists and is continuous. So $\widehat{\theta} = \mu^{-1}(m)$. By the WLLN $m \xrightarrow{\text{P}} \mu(\theta)$. So, by the continuous mapping Theorem,

$$\widehat{\theta}_n = \mu^{-1}(m) \xrightarrow{\text{P}} \mu^{-1}(\mu(\theta)) = \theta.$$

# 4 Consistency of the MLE

Under regularity conditions (see page 516), the mle is consistent. Let us prove this in a special case. This will also reveal a connection between the mle and Hellinger distance.

Suppose that the model consists of finitely many distinct densities $\{p_0, p_1, \ldots, p_N\}$. The likelihood function is

$$L(p_j) = \prod_{i=1}^{n} p_j(X_i).$$

The mle $\widehat{p}$ is the density $p_j$ that maximizes $L(p_j)$. Without loss of generality, assume that the true density is $p_0$.

**Theorem 3**

$$\mathbb{P}(\widehat{p} \neq p_0) \to 0$$

*as* $n \to \infty$.

**Proof.** Let us begin by first proving an inequality. Let $\epsilon_j = h(p_0, p_j)$. Then, for $j \neq 0$,

$$\begin{aligned}
\mathbb{P}\left(\frac{L(p_j)}{L(p_0)} > e^{-n\epsilon_j^2/2}\right) &= \mathbb{P}\left(\prod_{i=1}^{n} \frac{p_j(X_i)}{p_0(X_i)} > e^{-n\epsilon_j^2/2}\right) = \mathbb{P}\left(\prod_{i=1}^{n} \sqrt{\frac{p_j(X_i)}{p_0(X_i)}} > e^{-n\epsilon_j^4/2}\right) \\
&\leq e^{n\epsilon_j^2/4}\mathbb{E}\left(\prod_{i=1}^{n} \sqrt{\frac{p_j(X_i)}{p_0(X_i)}}\right) = e^{n\epsilon_j^2/4}\prod_{i=1}^{n}\mathbb{E}\left(\sqrt{\frac{p_j(X_i)}{p_0(X_i)}}\right) \\
&= e^{n\epsilon_j^2/4}\left(\int \sqrt{p_j\, p_0}\right)^n = e^{n\epsilon_j^2/4}\left(1 - \frac{h^2(p_0,p_j)}{2}\right)^n = e^{n\epsilon_j^2/4}\left(1 - \frac{\epsilon_j^2}{2}\right)^n \\
&= e^{n\epsilon_j^2/4}\exp\left\{n\log\left(1 - \frac{\epsilon_j^2}{2}\right)\right\} \leq e^{n\epsilon_j^2/4}e^{-n\epsilon_j^2/2} = e^{-n\epsilon_j^2/2}.
\end{aligned}$$

We used the fact that $h^2(p_0, p_j) = 2 - 2\int \sqrt{p_0\, p_j}$ and also that $\log(1-x) \leq -x$ for $x > 0$. Let $\epsilon = \min\{\epsilon_1, \ldots, \epsilon_N\}$. Then

$$\begin{aligned}
\mathbb{P}(\widehat{p} \neq p_0) &\leq \mathbb{P}\left(\frac{L(p_j)}{L(p_0)} > e^{-n\epsilon_j^2/2} \text{ for some } j\right) \\
&\leq \sum_{j=1}^{N} \mathbb{P}\left(\frac{L(p_j)}{L(p_0)} > e^{-n\epsilon_j^2/2}\right) \\
&\leq \sum_{j=1}^{N} e^{-n\epsilon_j^2/2} \leq N e^{-n\epsilon^2/2} \to 0.
\end{aligned}$$

■

We can prove a similar result using Kullback-Leibler distance as follows. Let $X_1, X_2, \ldots$ be iid $F_\theta$. Let $\theta_0$ be the true value of $\theta$ and let $\theta$ be some other value. We will show that

$L(\theta_0)/L(\theta) > 1$ with probability tending to 1. We assume that the model is **identifiable**; this means that $\theta_1 \neq \theta_2$ implies that $K(\theta_1, \theta_2) > 0$ where $K$ is the Kullback-Leibler distance.

**Theorem 4** *Suppose the model is identifiable. Let $\theta_0$ be the true value of the parameter. For any $\theta \neq \theta_0$*

$$\mathbb{P}\left(\frac{L(\theta_0)}{L(\theta)} > 1\right) \to 1$$

*as $n \to \infty$.*

**Proof.** We have

$$\begin{aligned}
\frac{1}{n}(\ell(\theta_0) - \ell(\theta)) &= \frac{1}{n}\sum_{i=1}^{n} \log p(X_i;\theta_0) - \frac{1}{n}\sum_{i=1}^{n} \log p(X_i;\theta) \\
&\xrightarrow{p} E(\log p(X;\theta_0)) - E(\log p(X;\theta)) \\
&= \int (\log p(x;\theta_0)) p(x;\theta_0) dx - \int (\log p(x;\theta)) p(x;\theta_0) dx \\
&= \int \left(\log \frac{p(x;\theta_0)}{p(x;\theta)}\right) p(x;\theta_0) dx \\
&= K(\theta_0, \theta) > 0.
\end{aligned}$$

So

$$\begin{aligned}
\mathbb{P}\left(\frac{L(\theta_0)}{L(\theta)} > 1\right) &= \mathbb{P}\left(\ell(\theta_0) - \ell(\theta) > 0\right) \\
&= \mathbb{P}\left(\frac{1}{n}(\ell(\theta_0) - \ell(\theta)) > 0\right) \to 1. \quad \square
\end{aligned}$$

■

This is not quite enough to show that $\widehat{\theta}_n \to \theta_0$.

**Example 5** *Inconsistency of an mle. In all examples so far $n \to \infty$, but the number of parameters is fixed. What if the number of parameters also goes to $\infty$? Let*

$$\begin{aligned}
Y_{11}, Y_{12} &\sim N(\mu_1, \sigma^2) \\
Y_{21}, Y_{22} &\sim N(\mu_2, \sigma^2) \\
\vdots &\sim \vdots \\
Y_{n1}, Y_{n2} &\sim N(\mu_n, \sigma^2).
\end{aligned}$$

*Some calculations show that*

$$\widehat{\sigma}^2 = \sum_{i=1}^{n}\sum_{j=1}^{2}\frac{(Y_{ij}-\overline{Y}_i)^2}{2n}.$$

*It is easy to show (good test question) that*

$$\widehat{\sigma}^2 \xrightarrow{p} \frac{\sigma^2}{2}.$$

*Note that the modified estimator $2\widehat{\sigma}^2$ is consistent.*

The reason why consistency fails is because the dimension of the parameter space is increasing with $n$.

**Theorem 6** *Under regularity conditions on the model $\{p(x;\theta) : \theta \in \Theta\}$, the mle is consistent.*

# 5 Score and Fisher Information

The score and Fisher information are the key quantities in many aspects of statistical inference. (See Section 7.3.2 of CB.) Suppose for now that $\theta \in \mathbb{R}$.

- $L(\theta) = p(x^n;\theta)$
- $\ell(\theta) = \log\ L(\theta)$
- $S(\theta) = \frac{\partial}{\partial\theta}\,\ell(\theta) \leftarrow$ **score function.**

Recall that the value $\widehat{\theta}$ that maximizes $L(\theta)$ is the **maximum likelihood estimator (mle)**. Equivalently, $\widehat{\theta}$ maximizes $\ell(\theta)$. Note that $\widehat{\theta} = T(X_1,\ldots,X_n)$ is a function of the data. Often, we get $\widehat{\theta}$ by differentiation. In that case $\widehat{\theta}$ solves

$$S(\widehat{\theta}) = 0.$$

We'll discuss the mle in detail later.

**Some Notation:** Recall that

$$\mathbb{E}_\theta(g(X)) \equiv \int g(x)p(x;\theta)dx.$$

**Theorem 7** *Under regularity conditions,*

$$\mathbb{E}_\theta[S(\theta)] = 0.$$

*In other words,*

$$\int \cdots \int \left( \frac{\partial \log p(x_1, \ldots, x_n; \theta)}{\partial \theta} \right) p(x_1, \ldots, x_n; \theta) dx_1 \ldots dx_n = 0.$$

That is, if the expected value is taken at the same $\theta$ as we evaluate $S$, then the expectation is 0. This does not hold when the $\theta$'s mismatch: $\mathbb{E}_{\theta_0}[S(\theta_1)] \neq 0$.

**Proof.**

$$\begin{aligned}
\mathbb{E}_\theta[S(\theta)] &= \int \cdots \int \frac{\partial \ \log \ p(x^n;\theta)}{\partial \theta} \, p(x^n;\theta) \, dx_1 \cdots dx_n \\
&= \int \cdots \int \frac{\frac{\partial}{\partial \theta} \, p(x^n;\theta)}{p(x^n;\theta)} p(x^n;\theta) \, dx_1 \cdots dx_n \\
&= \frac{\partial}{\partial \theta} \underbrace{\int \cdots \int p(x^n;\theta) \, dx_1 \cdots dx_n}_{1} \\
&= 0.
\end{aligned}$$

■

**Example 8** *Let* $X_1, \ldots, X_n \sim N(\theta, 1)$. *Then*

$$S(\theta) = \sum_{i=1}^{n} (X_i - \theta).$$

**Warning:** If the support of $f$ depends on $\theta$, then $\int \cdots \int$ and $\frac{\partial}{\partial \theta}$ cannot be switched.

The next quantity of interest is the **Fisher Information or Expected Information**. The information is used to calculate the variance of quantities that arise in inference problems

such as the mle $\widehat{\theta}$. It is called *information* because it tells how much information is in the likelihood about $\theta$. The definition is:

$$\begin{aligned} I(\theta) &= \mathbb{E}_\theta[S(\theta)^2] \\ &= \mathbb{E}_\theta[S(\theta)^2] - \big(\mathbb{E}_\theta[S(\theta)]\big)^2 \\ &= \mathrm{Var}_\theta(S(\theta)) \qquad\qquad \text{since } \mathbb{E}_\theta[S(\theta)] = 0 \\ &= \mathbb{E}_\theta\left[-\frac{\partial^2}{\partial\theta^2}\,\ell(\theta)\right] \leftarrow \text{ easiest way to calculate} \end{aligned}$$

We will prove the final equality under regularity conditions shortly. $I(\theta)$ grows linearly in $n$, so for an iid sample, a more careful notation would be $I_n(\theta)$

$$\begin{aligned} I_n(\theta) &= E\left[-\frac{\partial^2}{\partial\theta^2}\,l(\theta)\right] = E\left[-\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\,\log p(X_i;\theta)\right] \\ &= -nE\left[\frac{\partial^2}{\partial\theta^2}\,\log p(X_1;\theta)\right] = nI_1(\theta). \end{aligned}$$

Note that the Fisher information is a function of $\theta$ in two places:

- The derivate is w.r.t. $\theta$ and the information is evaluated at a particular value of $\theta$.
- The expectation is w.r.t. $\theta$ also. The notation only allows for a single value of $\theta$ because the two quantities should match.

A related quantity of interest is the **observed information**, defined as

$$\widehat{I}_n(\theta) = -\frac{\partial^2}{\partial\theta^2}\,\ell(\theta) = -\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\,\log p(X_i;\theta).$$

By the LLN $\frac{1}{n}\widehat{I}_n(\theta) \xrightarrow{\mathrm{P}} I_1(\theta)$. So observed information can be used as a good approximation to the Fisher information.

Let us prove the identity: $\mathbb{E}_\theta[S(\theta)^2] = \mathbb{E}_\theta\left[\frac{-\partial^2}{\partial\theta^2}\,\ell(\theta)\right]$. For simplicity take $n = 1$. First note that

$$\int p = 1 \quad\Rightarrow\quad \int p' = 0 \quad\Rightarrow\quad \int p'' = 0 \quad\Rightarrow\quad \int \frac{p''}{p} p = 0 \quad\Rightarrow\quad \mathbb{E}\left(\frac{p''}{p}\right) = 0.$$

Let $\ell = \log p$ and $S = \ell' = p'/p$. Then $\ell'' = (p''/p) - (p'/p)^2$ and

$$\begin{aligned} V(S) &= \mathbb{E}(S^2) - (\mathbb{E}(S))^2 = \mathbb{E}(S^2) = \mathbb{E}\left(\frac{p'}{p}\right)^2 \\ &= \mathbb{E}\left(\frac{p'}{p}\right)^2 - \mathbb{E}\left(\frac{p''}{p}\right) \\ &= -\mathbb{E}\left(\left(\frac{p''}{p}\right) - \left(\frac{p'}{p}\right)^2\right) \\ &= -\mathbb{E}(\ell''). \quad \square \end{aligned}$$

Why is $I(\theta)$ called "Information"? Later we will see that $\mathrm{Var}(\widehat{\theta}) \approx 1/I_n(\theta)$.

**The Vector Case.** Let $\theta = (\theta_1, \cdots, \theta_K)$. $L(\theta)$ and $\ell(\theta)$ are defined as before.

- $S(\theta) = \left[\frac{\partial \ell(\theta)}{\partial \theta_i}\right]_{i=1,\cdots,K}$ a vector of dimension $K$
- Information $I(\theta) = \mathrm{Var}[S(\theta)]$ is the variance-covariance matrix of

  $$S(\theta) = [I_{ij}]_{ij=1,\cdots,k}$$

  where

  $$I_{ij} = -\mathbb{E}_\theta\left[\frac{\partial^2 \ell(\theta)}{\partial \theta_i \, \partial \theta_j}\right].$$

- $I(\theta)^{-1}$ is the asymptotic variance of $\widehat{\theta}$. (This is the inverse of the matrix, evaluated at the proper component of the matrix.)

**Example 9**

$$\begin{aligned}
X_1, \cdots, X_n &\sim N(\mu, \gamma) \\
L(\mu, \gamma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\gamma}} \exp\left\{\frac{-1}{2\gamma}(x_i - \mu)^2\right\} \propto \gamma^{\frac{-n}{2}} \exp\left\{\frac{-1}{2\gamma}\Sigma(x_i - \mu)^2\right\} \\
\ell(\mu, \gamma) &= K - \frac{n}{2}\ \log\ \gamma - \frac{1}{2\gamma}\Sigma(x_i - \mu)^2 \\
S(\mu, \gamma) &= \begin{bmatrix} \frac{1}{\gamma}\Sigma(x_i - \mu) \\ -\frac{n}{2\gamma} + \frac{1}{2\gamma^2}\Sigma(x_i - \mu)^2 \end{bmatrix} \\
I(\mu, \gamma) &= -E\begin{bmatrix} \frac{-n}{\gamma} & \frac{-1}{\gamma^2}\Sigma(x_i - \mu) \\ \frac{-1}{\gamma^2}\Sigma(x_i - \mu) & \frac{n}{2\gamma^2} - \frac{1}{\gamma^3}\Sigma(x_i - \mu)^2 \end{bmatrix} \\
&= \begin{bmatrix} \frac{n}{\gamma} & 0 \\ 0 & \frac{n}{2\gamma^2} \end{bmatrix}
\end{aligned}$$

*You can check that* $\mathbb{E}_\theta(S) = (0, 0)^T$.

# 6 Efficiency and Asymptotic Normality

If $\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow N(0, v^2)$ then we call $v^2$ the *asymptotic variance* of $\widehat{\theta}_n$. This is not the same as the limit of the variance which is $\lim_{n\to\infty} n\text{Var}(\widehat{\theta}_n)$.

Consider $\overline{X}_n$. In this case, the asymptotic variance is $\sigma^2$. We also have that $\lim_{n\to\infty} n\text{Var}(\overline{X}_n) = \sigma^2$. In this case, they are the same. In general, the latter may be larger (or even infinite).

**Example 10** *(Example 10.1.10) Suppose we observe* $Y_n \sim N(0, 1)$ *with probability* $p_n$ *and* $Y_n \sim N(0, \sigma_n^2)$ *with probability* $1 - p_n$. *We can write this as a* **hierachical model***:*

$$\begin{aligned}
W_n &\sim \text{Bernoulli}(p_n) \\
Y_n | W_n &\sim N(0, W_n + (1 - W_n)\sigma_n^2).
\end{aligned}$$

*Now,*

$$\begin{aligned}
\text{Var}(Y_n) &= \text{Var}E(Y_n|W_n) + E\text{Var}(Y_n|W_n) \\
&= \text{Var}(0) + E(W_n + (1 - W_n)\sigma_n^2) = p_n + (1 - p_n)\sigma_n^2.
\end{aligned}$$

*Suppose that* $p_n \to 1$, $\sigma_n \to \infty$ *and that* $(1-p_n)\sigma_n^2 \to \infty$. *Then* $\mathrm{Var}(Y_n) \to \infty$. *Then*

$$\mathbb{P}(Y_n \leq a) \;\; = \;\; p_n \mathbb{P}(Z \leq a) + (1-p_n)\mathbb{P}(Z \leq a/\sigma_n) \to \mathbb{P}(Z \leq a)$$

*and so* $Y_n \rightsquigarrow N(0,1)$. *So the asymptotic variance is 1.*

Suppose we want to estimate $\tau(\theta)$. Let

$$v(\theta) = \frac{|\tau'(\theta)|^2}{I(\theta)}$$

where

$$I(\theta) = \mathrm{Var}\left(\frac{\partial}{\partial\theta}\log p(X;\theta)\right) = -\mathbb{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\log p(X;\theta)\right).$$

We call $v(\theta)$ the **Cramer-Rao lower bound.** Generally, any well-behaved estimator will have a limiting variance bigger than or equal to $v(\theta)$. We say that $W_n$ is **efficient** if $\sqrt{n}(W_n - \tau(\theta)) \rightsquigarrow N(0, v(\theta))$.

**Theorem 11** *Let* $X_1, X_2, \ldots,$ *be iid. Assume that the model satisfies the regularity conditions in 10.6.2. Let* $\widehat{\theta}$ *be the mle. Then*

$$\sqrt{n}(\tau(\widehat{\theta}) - \tau(\theta)) \rightsquigarrow N(0, v(\theta)).$$

*So* $\tau(\widehat{\theta})$ *is consistent and efficient.*

We will now prove the asymptotic normality of the mle.

**Theorem 12**

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow N\left(0, \frac{1}{I(\theta)}\right).$$

*Hence,*

$$\widehat{\theta}_n = \theta + O_P\left(\frac{1}{\sqrt{n}}\right).$$

**Proof.** By Taylor's theorem

$$0 = \ell'(\widehat{\theta}) = \ell'(\theta) + (\widehat{\theta} - \theta)\ell''(\theta) + \cdots.$$

Hence

$$\sqrt{n}(\widehat{\theta}-\theta) \approx \frac{\frac{1}{\sqrt{n}}\ell'(\theta)}{-\frac{1}{n}\ell''(\theta)} = \frac{A}{B}.$$

Now

$$A = \frac{1}{\sqrt{n}}\ell'(\theta) = \sqrt{n} \times \frac{1}{n}\sum_{i=1}^{n} S(\theta, X_i) = \sqrt{n}(\overline{S} - 0)$$

where $S(\theta, X_i)$ is the score function based on $X_i$. Recall that $E(S(\theta, X_i)) = 0$ and $\text{Var}(S(\theta, X_i)) = I(\theta)$. By the central limit theorem, $A \rightsquigarrow N(0, I(\theta)) = \sqrt{I(\theta)}Z$ where $Z \sim N(0,1)$. By the WLLN,

$$B \xrightarrow{\text{P}} -E(\ell'') = I(\theta).$$

By Slutsky's theorem

$$\frac{A}{B} \rightsquigarrow \frac{\sqrt{I(\theta)}Z}{I(\theta)} = \frac{Z}{\sqrt{I(\theta)}} = N\left(0, \frac{1}{I(\theta)}\right).$$

So

$$\sqrt{n}(\widehat{\theta}-\theta) \rightsquigarrow N\left(0, \frac{1}{I(\theta)}\right). \quad \square$$

■

Theorem 11 follows by the delta method:

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow N(0, 1/I(\theta))$$

implies that

$$\sqrt{n}(\tau(\widehat{\theta}_n) - \tau(\theta)) \rightsquigarrow N(0, (\tau'(\theta))^2/I(\theta)).$$

The standard error of $\widehat{\theta}$ is

$$se = \sqrt{\frac{1}{nI(\theta)}} = \sqrt{\frac{1}{I_n(\theta)}}.$$

The estimated standard error is

$$\widehat{se} = \sqrt{\frac{1}{I_n(\widehat{\theta})}}.$$

The standard error of $\widehat{\tau} = \tau(\widehat{\theta})$ is

$$se = \sqrt{\frac{|\tau'(\theta)|}{nI(\theta)}} = \sqrt{\frac{|\tau'(\theta)|}{I_n(\theta)}}.$$

The estimated standard error is

$$\widehat{se} = \sqrt{\frac{|\tau'(\widehat{\theta})|}{I_n(\widehat{\theta})}}.$$

**Example 13** *$X_1, \cdots, X_n$ iid Exponential $(\theta)$. Let $t = \overline{x}$. So: $p(z;\theta) = \theta e^{-\theta x}$, $L(\theta) = e^{-n\theta t + n \ln \theta}$, $l(\theta) = -n\theta t + n \ \ln \ \theta$, $S(\theta) = \frac{n}{\theta} - nt \Rightarrow \widehat{\theta} = \frac{1}{t} = \frac{1}{\overline{X}}$, $l''(\theta) = \frac{-n}{\theta^2}$, $I(\theta) = E[-l''(\theta)] = \frac{n}{\theta^2}$, $\widehat{\theta} \approx N\left(\theta, \frac{\theta^2}{n}\right)$.*

**Example 14** *$X_1, \ldots, X_n \sim$ Bernoulli$(p)$. The mle is $\widehat{p} = \overline{X}$. The Fisher information for $n = 1$ is*

$$I(p) = \frac{1}{p(1-p)}.$$

*So*

$$\sqrt{n}(\widehat{p} - p) \rightsquigarrow N(0, p(1-p)).$$

*Informally,*

$$\widehat{p} \approx N\left(p, \ \frac{p(1-p)}{n}\right).$$

*The asymptotic variance is $p(1-p)/n$. This can be estimated by $\widehat{p}(1-\widehat{p})/n$. That is, the estimated standard error of the mle is*

$$\widehat{se} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

*Now suppose we want to estimate $\tau = p/(1-p)$. The mle is $\widehat{\tau} = \widehat{p}/(1-\widehat{p})$. Now*

$$\frac{\partial}{\partial p}\frac{p}{1-p} = \frac{1}{(1-p)^2}$$

*The estimated standard error is*

$$\widehat{se}(\widehat{\tau}) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \times \frac{1}{(1-\widehat{p})^2} = \sqrt{\frac{\widehat{p}}{n(1-\widehat{p})^3}}.$$

# 7 Relative Efficiency

If

$$\begin{aligned} \sqrt{n}(W_n - \tau(\theta)) &\rightsquigarrow N(0, \sigma_W^2) \\ \sqrt{n}(V_n - \tau(\theta)) &\rightsquigarrow N(0, \sigma_V^2) \end{aligned}$$

then the *asymptotic relative efficiency (ARE)* is

$$\text{ARE}(V_n, W_n) = \frac{\sigma_W^2}{\sigma_V^2}.$$

**Example 15** *(10.1.17). Let* $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$. *The mle of* $\lambda$ *is* $\overline{X}$. *Let*

$$\tau = \mathbb{P}(X_i = 0).$$

*So* $\tau = e^{-\lambda}$. *Define* $Y_i = I(X_i = 0)$. *This suggests the estimator*

$$W_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

*Another estimator is the mle*

$$V_n = e^{-\widehat{\lambda}}.$$

*The delta method gives*

$$\text{Var}(V_n) \approx \frac{\lambda e^{-2\lambda}}{n}.$$

*We have*

$$\begin{aligned} \sqrt{n}(W_n - \tau) &\rightsquigarrow N(0, e^{-\lambda}(1 - e^{-\lambda})) \\ \sqrt{n}(V_n - \tau) &\rightsquigarrow N(0, \lambda e^{-2\lambda}). \end{aligned}$$

*So*

$$\text{ARE}(W_n, V_n) = \frac{\lambda}{e^{\lambda} - 1} \leq 1. \quad \square$$

Since the mle is efficient, we know that, in general, $\text{ARE}(W_n, \text{mle}) \leq 1$.

# 8 Robustness

The mle is efficient only if the model is right. The mle can be bad if the model is wrong. That is why we should consider using nonparametric methods. One can also replace the mle with estimators that are more *robust*.

Suppose we assume that $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$. The mle is $\widehat{\theta}_n = \overline{X}_n$. Suppose, however that we have a perturbed model $X_i$ is $N(\theta, \sigma^2)$ with probability $1-\delta$ and $X_i$ is Cauchy with probability $\delta$. Then, $\text{Var}(\overline{X}_n) = \infty$.

Consider the median $M_n$. We will show that

$$ARE(\text{median}, \text{mle}) = .64.$$

But, under the perturbed model the median still performs well while the mle is terrible. In other words, we can trade efficiency for robustness. Let us now find the limiting distribution of $M_n$.

Let $Y_i = I(X_i \leq \mu + a/\sqrt{n})$. Then $Y_i \sim \text{Bernoulli}(p_n)$ where

$$p_n = P(\mu + a/\sqrt{n}) = P(\mu) + \frac{a}{\sqrt{n}}p(\mu) + o(n^{-1/2}) = \frac{1}{2} + \frac{a}{\sqrt{n}}p(\mu) + o(n^{-1/2}).$$

Also, $\sum_i Y_i$ has mean $np_n$ and standard deviation

$$\sigma_n = \sqrt{np_n(1-p_n)}.$$

Note that,

$$M_n \leq \mu + \frac{a}{\sqrt{n}} \quad \text{if and only if} \quad \sum_i Y_i \geq \frac{n+1}{2}.$$

Then,

$$\begin{aligned}
\mathbb{P}(\sqrt{n}(M_n - \mu) \leq a) &= \mathbb{P}\left(M_n \leq \mu + \frac{a}{\sqrt{n}}\right) = \mathbb{P}\left(\sum_i Y_i \geq \frac{n+1}{2}\right) \\
&= \mathbb{P}\left(\frac{\sum_i Y_i - np_n}{\sigma_n} \geq \frac{\frac{n+1}{2} - np_n}{\sigma_n}\right).
\end{aligned}$$

Now,

$$\frac{\frac{n+1}{2} - np_n}{\sigma_n} \to -2ap(\mu)$$

and hence

$$\mathbb{P}(\sqrt{n}(M_n - \mu) \leq a) \to \mathbb{P}(Z \geq -2ap(\mu)) = \mathbb{P}\left(-\frac{Z}{2p(\mu)} \leq a\right) = \mathbb{P}\left(\frac{Z}{2p(\mu)} \leq a\right)$$

so that

$$\sqrt{n}(M_n - \mu) \rightsquigarrow N\left(0, \frac{1}{(2p(\mu))^2}\right).$$

For a standard Normal, $(2p(0))^2 = .64$.

# Lecture Notes 10

# Hypothesis Testing

## 1 Introduction

(See Chapter 8 and Chapter 10.3.)

Null hypothesis: $H_0 : \theta \in \Theta_0$

Alternative hypothesis: $H_1 : \theta \in \Theta_1$

where $\Theta_0 \cap \Theta_1 = \emptyset$.

**Example 1** $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$.

$$H_0 : p = \frac{1}{2} \qquad H_1 : p \neq \frac{1}{2}. \quad \square$$

The question is not whether $H_0$ is true or false. The question is whether there is sufficient evidence to reject $H_0$, much like a court case.

Our possible actions are: reject $H_0$ or retain (don't reject) $H_0$.

| | Decision | |
|---|---|---|
| | Retain $H_0$ | Reject $H_0$ |
| $H_0$ true | $\surd$ | Type I error (false positive) |
| $H_1$ true | Type II error (false negative) | $\surd$ |

**Warning: Hypothesis testing should only be used when it is appropriate. Often times, people use hypothesis tetsing when it would be much more appropriate to use confidence intervals (which is the next topic).**

# 2 Constructing Tests

1. Choose a *test statistic* $W = W(X_1, \ldots, X_n)$.
2. Choose a rejection region $R$.
3. If $W \in R$ we reject $H_0$ otherwise we retain $H_0$.

**Example 2** $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$.

$$H_0 : p = \frac{1}{2} \qquad H_1 : p \neq \frac{1}{2}.$$

*Let* $W = n^{-1} \sum_{i=1}^n X_i$. *Let* $R = \{x^n : \ |w(x^n) - 1/2| > \delta\}$. *So we reject* $H_0$ *if* $|W - 1/2| > \delta$.

We need to choose $W$ and $R$ so that the test has good statistical properties. We will consider the following tests:

1. Neyman-Pearson Test
2. Wald test
3. Likelihood Ratio Test (LRT)
4. the permutation test
5. the score test (optional)

Before we discuss these methods, we first need to talk about how we evaluate tests.

# 3 Evaluating Tests

Suppose we reject $H_0$ when $X^n = (X_1, \ldots, X_n) \in R$. Define the *power function* by

$$\beta(\theta) = P_\theta(X^n \in R).$$

**We want $\beta(\theta)$ to be small when $\theta \in \Theta_0$ and we want $\beta(\theta)$ to be large when $\theta \in \Theta_1$.** The general strategy is:

1. Fix $\alpha \in [0, 1]$.
2. Now try to maximize $\beta(\theta)$ for $\theta \in \Theta_1$ subject to $\beta(\theta) \leq \alpha$ for $\theta \in \Theta_0$.

We need the following definitions. A test is *size* $\alpha$ if

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

A test is *level* $\alpha$ if

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

A size $\alpha$ test and a level $\alpha$ test are almost the same thing. The distinction is made bcause sometimes we want a size $\alpha$ test and we cannot construct a test with exact size $\alpha$ but we can construct one with a smaller error rate.

**Example 3** *$X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ with $\sigma^2$ known. Suppose*

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta > \theta_0.$$

*This is called a* **one-sided alternative**. *Suppose we reject $H_0$ if $W > c$ where*

$$W = \frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}}.$$

*Then*

$$\begin{aligned} \beta(\theta) &= P_\theta\left(\frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}} > c\right) \\ &= P_\theta\left(\frac{\overline{X}_n - \theta}{\sigma/\sqrt{n}} > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \\ &= 1 - \Phi\left(c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \end{aligned}$$

*where $\Phi$ is the cdf of a standard Normal. Now*

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \beta(\theta_0) = 1 - \Phi(c).$$

*To get a size $\alpha$ test, set $1 - \Phi(c) = \alpha$ so that*

$$c = z_\alpha$$

*where $z_\alpha = \Phi^{-1}(1-\alpha)$. Our test is: reject $H_0$ when*

$$W = \frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}} > z_\alpha.$$

**Example 4** *$X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ with $\sigma^2$ known. Suppose*

$$H_0 : \theta \leq \theta_0, \qquad H_1 : \theta \neq \theta_0.$$

*This is called a* **two-sided** *alternative. We will reject $H_0$ if $|W| > c$ where $W$ is defined as before. Now*

$$\begin{aligned}
\beta(\theta) &= P_\theta(W < -c) + P_\theta(W > c) \\
&= P_\theta\left(\frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}} < -c\right) + P_\theta\left(\frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}} > c\right) \\
&= P\left(Z < -c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) + P\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \\
&= \Phi\left(-c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) + 1 - \Phi\left(c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \\
&= \Phi\left(-c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) + \Phi\left(-c - \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

*since $\Phi(-x) = 1 - \Phi(x)$. The size is*

$$\beta(\theta_0) = 2\Phi(-c).$$

*To get a size $\alpha$ test we set $2\Phi(-c) = \alpha$ so that $c = -\Phi^{-1}(\alpha/2) = \Phi^{-1}(1-\alpha/2) = z_{\alpha/2}$. The test is: reject $H_0$ when*

$$|W| = \left|\frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}}\right| > z_{\alpha/2}.$$

# 4 The Neyman-Pearson Test

Let $\mathcal{C}_\alpha$ denote all level $\alpha$ tests. A test in $\mathcal{C}_\alpha$ with power function $\beta$ is **uniformly most powerful (UMP)** if the following holds: if $\beta'$ is the power function of any other test in $\mathcal{C}_\alpha$ then $\beta(\theta) \leq \beta'(\theta)$ for all $\theta \in \Theta_1$.

Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. (Simple null and simple alternative.)

**Theorem 5** *Suppose we set*

$$R = \left\{ x = (x_1, \ldots, x_n) : \ \frac{f(X_1, \ldots, X_n; \theta_1)}{f(X_1, \ldots, X_n; \theta_0)} > k \right\} = \left\{ x^n : \ \frac{L(\theta_1)}{L(\theta_0)} > k \right\}$$

*where $k$ is chosen so that*

$$P_{\theta_0}(X^n \in R) = \alpha.$$

*In other words, reject $H_0$ if*

$$\frac{L(\theta_1)}{L(\theta_0)} > k.$$

*This test is a UMP level $\alpha$ test.*

This is theorem 8.3.12 in the book. The proof is short; you should read the proof.

Notes:

1. Ignore the material on union-intersection tests and monotonote likelihood ratios (MLR).
2. In general it is hard to find UMP tests. Sometimes they don't even exist. Still, we can find tests with good properties.

# 5 The Wald Test

Let

$$W = \frac{\widehat{\theta}_n - \theta_0}{se}.$$

Under the uusal conditions we have that under $H_0$, $W \rightsquigarrow N(0,1)$. Hence, an asymptotic level $\alpha$ test is to reject when $|W| > z_{\alpha/2}$.

For example, with Bernoulli data, to test $H_0 : p = p_0$,

$$W = \frac{\widehat{p} - p_0}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}}.$$

You can also use

$$W = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

In other words, to compute the standard error, you can replace $\theta$ with an estimate $\widehat{\theta}$ or by the null value $\theta_0$.

# 6 The Likelihood Ratio Test (LRT)

This test is simple: reject $H_0$ if $\lambda(x^n) \leq c$ where

$$\lambda(x^n) = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)} = \frac{L(\widehat{\theta}_0)}{L(\widehat{\theta})}$$

where $\widehat{\theta}_0$ maximizes $L(\theta)$ subject to $\theta \in \Theta_0$.

**Example 6** *$X_1, \ldots, X_n \sim N(\theta, 1)$. Suppose*

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta \neq \theta_0.$$

*After some algebra (see page 376),*

$$\lambda = \exp\left\{-\frac{n}{2}(\overline{X}_n - \theta_0)^2\right\}.$$

*So*

$$R = \{x : \ \lambda \leq c\} = \{x : \ |\overline{X} - \theta_0| \geq c'\}$$

*where $c' = \sqrt{-2 \log c/n}$. Choosing $c'$ to make this level $\alpha$ gives: reject if $|W| > z_{\alpha/2}$ where $W = \sqrt{n}(\overline{X} - \theta_0)$ which is the test we constructed before.*

**Example 7** *$X_1, \ldots, X_n \sim N(\theta, \sigma^2)$. Suppose*

$$H_0 : \theta \leq \theta_0, \qquad H_1 : \theta \neq \theta_0.$$

*Then*

$$\lambda(x^n) = \frac{L(\theta_0, \widehat{\sigma}_0)}{L(\widehat{\theta}, \widehat{\sigma})}$$

*where $\widehat{\sigma}_0$ maximizes the likelihood subject to $\theta = \theta_0$. In the homework, you will prove that $\lambda(x^n) < c$ corresponds to rejecting when $|T_n| > k$ for some constant $k$ where*

$$T_n = \frac{\overline{X}_n - \theta_0}{S/\sqrt{n}}.$$

*Under $H_0$, $T_n$ has a t-distribution with $n-1$ degrees of freedom. So the final test is: reject $H_0$ if*

$$|T_n| > t_{n-1,\alpha/2}.$$

*This is called Student's t-test. It was invented by William Gosset working at Guiness Breweries and writing under the pseudonym Srudent.*

**Theorem 8** *Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ where $\theta \in \mathbb{R}$. Under $H_0$,*

$$-2\log\lambda(X^n) \rightsquigarrow \chi^2_1.$$

*Hence, if we let $W_n = -2\log\lambda(X^n)$ then*

$$P_{\theta_0}(W > \chi^2_{1,\alpha}) \to \alpha$$

*as $n \to \infty$.*

**Proof.** Using a Taylor expansion:

$$\ell(\theta) \approx \ell(\widehat{\theta}) + \ell'(\widehat{\theta})(\theta - \widehat{\theta}) + \ell''(\widehat{\theta})\frac{(\theta-\widehat{\theta})^2}{2} = \ell(\widehat{\theta}) + \ell''(\widehat{\theta})\frac{(\theta-\widehat{\theta})^2}{2}$$

and so

$$\begin{aligned} -2\log\lambda(x^n) &= 2\ell(\widehat{\theta}) - 2\ell(\theta_0) \\ &\approx 2\ell(\widehat{\theta}) - 2\ell(\widehat{\theta}) - \ell''(\widehat{\theta})(\theta-\widehat{\theta})^2 = -\ell''(\widehat{\theta})(\theta-\widehat{\theta})^2 \\ &= \frac{-\ell''(\widehat{\theta})}{I_n(\theta_0)} I_n(\theta_0)(\sqrt{n}(\widehat{\theta}-\theta_0))^2 = A_n \times B_n. \end{aligned}$$

Now $A_n \xrightarrow{\mathrm{P}} 1$ by the WLLN and $\sqrt{B_n} \rightsquigarrow N(0,1)$. The result follows by Slutsky's theorem. ■

**Example 9** *$X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$. We want to test $H_0 : \lambda = \lambda_0$ versus $H_1 : \lambda \neq \lambda_0$. Then*

$$-2\log\lambda(x^n) = 2n[(\lambda_0 - \widehat{\lambda}) - \widehat{\lambda}\log(\lambda_0/\widehat{\lambda})].$$

*We reject $H_0$ when $-2\log\lambda(x^n) > \chi^2_{1,\alpha}$.*

Now suppose that $\theta = (\theta_1, \ldots, \theta_k)$. Suppose that $H_0$ fixes some of the parameters. Then

$$-2 \log \lambda(X^n) \rightsquigarrow \chi^2_\nu$$

where

$$\nu = \dim(\Theta) - \dim(\Theta_0).$$

**Example 10** *Consider a multinomial with $\theta = (p_1, \ldots, p_5)$. So*

$$L(\theta) = p_1^{y_1} \cdots p_5^{y_5}.$$

*Suppose we want to test*

$$H_0 : p_1 = p_2 = p_3 \quad \text{and} \quad p_4 = p_5$$

*versus the alternative that $H_0$ is false. In this case*

$$\nu = 4 - 1 = 3.$$

*The LRT test statistic is*

$$\lambda(x^n) = \frac{\prod_{i=1}^5 \widehat{p}_{0j}^{Y_j}}{\prod_{i=1}^5 \widehat{p}_j^{Y_j}}$$

*where $\widehat{p}_j = Y_j/n$, $\widehat{p}_{10} = \widehat{p}_{20} = \widehat{p}_{30} = (Y_1 + Y_2 + Y_3)/n$, $\widehat{p}_{40} = \widehat{p}_{50} = (1 - 3\widehat{p}_{10})/2$. These calculations are on p 491. Make sure you understand them. Now we reject $H_0$ if $-2\lambda(X^n) > \chi^2_{3,\alpha}$.* □

# 7 p-values

When we test at a given level $\alpha$ we will reject or not reject. It is useful to summarize what levels we would reject at and what levels we woud not reject at.

**The p-value is the smallest $\alpha$ at which we would reject $H_0$.**

In other words, we reject at all $\alpha \geq p$. So, if the pvalue is 0.03, then we would reject at $\alpha = 0.05$ but not at $\alpha = 0.01$.

Hence, to test at level $\alpha$ when $p < \alpha$.

**Theorem 11** *Suppose we have a test of the form: reject when $W(X^n) > c$. Then the p-value when $X^n = x^n$ is*

$$p(x^n) = \sup_{\theta \in \Theta_0} P_\theta(W(X^n) \geq W(x^n)).$$

**Example 12** $X_1, \ldots, X_n \sim N(\theta, 1)$. *Test that $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. We reject when $|W|$ is large, where $W = \sqrt{n}(\overline{X}_n - \theta_0)$. So*

$$p = P_{\theta_0}\left(|\sqrt{n}(\overline{X}_n - \theta_0)| > w\right) = P(|Z| > w) = 2\Phi(-|w|).$$

**Theorem 13** *Under $H_0$, $p \sim \text{Unif}(0, 1)$.*

**Important.** Note that $p$ is NOT equal to $P(H_0|X_1, \ldots, X_n)$. The latter is a Bayesian quantity which we will discuss later.

# 8 The Permutation Test

This is a very cool test. It is distribution free and it does not involve any asymptotic approximations.

Suppose we have data

$$X_1, \ldots, X_n \sim F$$

and

$$Y_1, \ldots, Y_m \sim G.$$

We want to test:

$$H_0 : F = G \quad \text{versus} \quad H_1 : F \neq G.$$

Let

$$Z = (X_1, \ldots, X_n, Y_1, \ldots, Y_m).$$

Create labels

$$L = (\underbrace{1, \ldots, 1}_{n \text{ values}}, \underbrace{2, \ldots, 2}_{m \text{ values}}).$$

A test statistic can be written as a function of $Z$ and $L$. For example, if

$$W = |\overline{X}_n - \overline{Y}_n|$$

then we can write

$$W = \left| \frac{\sum_{i=1}^N Z_i I(L_i = 1)}{\sum_{i=1}^N I(L_i = 1)} - \frac{\sum_{i=1}^N Z_i I(L_i = 2)}{\sum_{i=1}^N I(L_i = 2)} \right|$$

where $N = n + m$. So we write $W = g(L, Z)$.

Define

$$p = \frac{1}{N!} \sum_{\pi} I(g(L_\pi, Z) > g(L, Z))$$

where $L_\pi$ is a permutation of the labels and the sum is over all permutations. Under $H_0$, permuting the labels does not change the distribution. In other words, $g(L, Z)$ has an equal chance of having any rank among all the permuted values. That is, under $H_0$, $\approx \text{Unif}(0, 1)$ and if we reject when $p < \alpha$, then we have a level $\alpha$ test.

Summing over all permutations is infeasible. But it suffices to use a random sample of permutations. So we do this:

1. Compute a random permutation of the labels and compute $W$. Do this $K$ times giving values $W_1, \ldots, W_K$.

2. Compute the p-value

$$\frac{1}{K} \sum_{j=1}^{K} I(W_j > W).$$

# 9 The Score Test (Optional)

Recall that the score statistic is

$$S(\theta) = \frac{\partial}{\partial \theta} log f(X_1, \ldots, X_n; \theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} log f(X_i \theta).$$

Recall that $E_\theta S(\theta) = 0$ and $V_\theta S(\theta) = I_n(\theta)$. By the CLT,

$$Z = \frac{S(\theta_0)}{\sqrt{I_n(\theta_0)}} \rightsquigarrow N(0, 1)$$

under $H_0$. So we reject if $|Z| > z_{\alpha/2}$. The advantage of the score test is that it does not require maximizing the likelihood function.

**Example 14** *For the Binomial,*

$$S(p) = \frac{n(\widehat{p}_n - p)}{p(1-p)}, \quad I_n(p) = \frac{n}{p(1-p)}$$

*and so*

$$Z = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

*This is the same as the Wald test in this case.*

# Lecture Notes 11
# Interval Estimation (Confidence Intervals)

Chapter 9 and Chapter 10.4

## 1 Introduction

Find $C_n = [L(X_1, \ldots, X_n), U(X_1, \ldots, X_n)]$ so that

$$P_\theta\Big( L(X_1, \ldots, X_n) \leq \theta \leq U(X_1, \ldots, X_n) \Big) \geq 1 - \alpha \quad \text{for all } \theta \in \Theta.$$

In other words:

$$\inf_{\theta \in \Theta} P_\theta\Big( L(X_1, \ldots, X_n) \leq \theta \leq U(X_1, \ldots, X_n) \Big) \geq 1 - \alpha.$$

We say that $C_n$ has **coverage** $1-\alpha$ or that $C_n$ is a $1-\alpha$ **confidence interval. Note that $C_n$ is random and $\theta$ is fixed (but unknown).**

More generally, a $1-\alpha$ *confidence set* $C_n$ is a (random) set $C_n \subset \Theta$ such that

$$\inf_{\theta \in \Theta} P_\theta\Big( \theta \in C_n(X_1, \ldots, X_n) \Big) \geq 1 - \alpha.$$

Again, $C_n$ is random, $\theta$ is not.

**Example 1** *Let $X_1, \ldots, X_n \sim N(\theta, \sigma)$. Suppose that $\sigma$ is known. Let $L = L(X_1, \ldots, X_n) = \overline{X} - c$ and $U = U(X_1, \ldots, X_n) = \overline{X} + c$. Then*

$$\begin{aligned} P_\theta(L \leq \theta \leq U) &= P_\theta(\overline{X} - c \leq \theta \leq \overline{X} + c) \\ &= P_\theta(-c < \overline{X} - \theta < c) = P_\theta\left( -\frac{c\sqrt{n}}{\sigma} < \frac{\sqrt{n}(\overline{X} - \theta)}{\sigma} < \frac{c\sqrt{n}}{\sigma} \right) \\ &= P\left( -\frac{c\sqrt{n}}{\sigma} < Z < \frac{c\sqrt{n}}{\sigma} \right) = \Phi(c\sqrt{n}/\sigma) - \Phi(-c\sqrt{n}/\sigma) \\ &= 1 - 2\Phi(-c\sqrt{n}/\sigma) = 1 - \alpha \end{aligned}$$

*if we choose $c = \sigma z_{\alpha/2}/\sqrt{n}$. So, if we define $C_n = \overline{X}_n \pm \sigma z_{\alpha/2}\sqrt{n}$ then*

$$P_\theta(\theta \in C_n) = 1 - \alpha$$

*for all $\theta$.*

**Example 2** *$X_i \sim N(\theta_i, 1)$ for $i = 1, \ldots, n$. Let*

$$C_n = \{\theta \in \mathbb{R}^n : \ ||X - \theta||^2 \leq \chi^2_{n,\alpha}\}.$$

*Then*

$$P_\theta(\theta \notin C_n) = P_\theta(||X - \theta||^2 > \chi^2_{n,\alpha}) = P(\chi^2_n > \chi^2_{n,\alpha}) = \alpha.$$

Four methods:

1. Probability Inequalities
2. Inverting a test
3. Pivots
4. Large Sample Approximations

Optimal confidence intervals are confidence intervals that are as short as possible but we will not discuss optimality.

# 2 Using Probability Inequalities

Intervals that are valid for finite samples can be obtained by probability inequalities.

**Example 3** *Let $X_1, \ldots, X_n \sim$ Bernoulli$(p)$. By Hoeffding's inequality:*

$$\mathbb{P}(|\widehat{p} - p| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

*Let*

$$\epsilon_n = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}.$$

*Then*

$$\mathbb{P}\left(|\widehat{p} - p| > \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}\right) \leq \alpha.$$

*Hence, $\mathbb{P}(p \in C) \geq 1 - \alpha$ where $C = (\widehat{p} - \epsilon_n, \widehat{p} + \epsilon_n)$.*

**Example 4** *Let $X_1, \ldots, X_n \sim F$. Suppose we want a* **confidence band** *for $F$. We can use VC theory. Remember that*

$$\mathbb{P}\left(\sup_x |F_n(x) - F(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

*Let*

$$\epsilon_n = \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}.$$

*Then*

$$\mathbb{P}\left(\sup_x |F_n(x) - F(x)| > \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}\right) \leq \alpha.$$

*Hence,*

$$P_F(L(t) \leq F(t) \leq U(t) \ \text{ for all } t) \geq 1 - \alpha$$

*for all $F$, where*

$$L(t) = \widehat{F}_n(t) - \epsilon_n, \quad U(t) = \widehat{F}_n(t) + \epsilon_n.$$

*We can improve this by taking*

$$L(t) = \max\left\{\widehat{F}_n(t) - \epsilon_n,\ 0\right\}, \quad U(t) = \min\left\{\widehat{F}_n(t) + \epsilon_n,\ 1\right\}.$$

# 3 Inverting a Test

For each $\theta_0$, construct a level $\alpha$ test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Define $\phi_{\theta_0}(x^n) = 1$ if we reject and $\phi_{\theta_0}(x^n) = 0$ if we don't reject. Let $A(\theta_0)$ be the acceptance region, that is, $A(\theta_0) = \{x^n :\ \phi_{\theta_0}(x^n) = 0\}$. Let

$$C(x^n) = \{\theta :\ x^n \in A(\theta)\} = \{\theta :\ \phi_\theta(x^n) = 0\}.$$

**Theorem 5** *For each $\theta$,*

$$P_\theta(\theta \in C(x^n)) = 1 - \alpha.$$

**Proof.** $1 - P_\theta(\theta \in C(x^n))$ is the probability of rejecting $\theta$ when $\theta$ is true which is $\alpha$. □

■

The converse is also true: if $C(x^n)$ is a $1-\alpha$ confidence interval then the test:

$$\text{reject } H_0 \text{ if } \theta_0 \notin C(x^n)$$

is a level $\alpha$ test.

**Example 6** *Suppose we use the LRT. We reject $H_0$ when*

$$\frac{L(\theta_0)}{L(\widehat{\theta})} \leq c.$$

*So*

$$C = \left\{ \theta : \ \frac{L(\theta)}{L(\widehat{\theta})} \geq c \right\}.$$

*See Example 9.2.3 for a detailed example involving the exponential distribution.*

**Example 7** *Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with $\sigma^2$ known. The LRT of $H_0 : \mu = \mu_0$ rejects when*

$$|\overline{X} - \mu_0| \geq \frac{\sigma}{\sqrt{n}} z_{\alpha/2}.$$

*So*

$$A(\mu) = \left\{ x^n : \ |\overline{X} - \mu_0| < \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right\}$$

*and so $\mu \in C(X^n)$ if and only if*

$$|\overline{X} - \mu| \leq \frac{\sigma}{\sqrt{n}} z_{\alpha/2}.$$

*In other words,*

$$C = \overline{X} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}.$$

*If $\sigma$ is unknown, then this becomes*

$$C = \overline{X} \pm \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}.$$

*(Good practice question.)*

# 4 Pivots

A function $Q(X_1, \ldots, X_n, \theta)$ is a *pivot* if the distribution of $Q$ does not depend on $\theta$.

For example, if $X_1, \ldots, X_n \sim N(\theta, 1)$ then

$$\overline{X}_n - \theta \sim N(0, 1/n)$$

so $Q = \overline{X}_n - \theta$ is a pivot.

Let $a$ and $b$ be such that

$$P_\theta(a \leq Q(X, \theta) \leq b) \geq 1 - \alpha$$

for all $\theta$. We can find such an $a$ and $b$ because $Q$ is a pivot. It follows immediately that

$$C(x) = \{\theta : \ a \leq Q(x, \theta) \leq b\}$$

has coverage $1 - \alpha$.

**Example 8** *Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. ($\sigma$ known.) Then*

$$Z = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0, 1).$$

*We know that*

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

*and so*

$$P\left(-z_{\alpha/2} \leq \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

*Thus*

$$C = \overline{X} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}.$$

*If $\sigma$ is unknown, then this becomes*

$$C = \overline{X} \pm \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}$$

*because*

$$T = \frac{\sqrt{n}(\overline{X} - \mu)}{S} \sim t_{n-1}.$$

**Example 9** *Let* $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$. *Let* $Q = X_{(n)}/\theta$. *Then*

$$\mathbb{P}(Q \leq t) = \prod_i \mathbb{P}(X_i \leq t\theta) = t^n$$

*so* $Q$ *is a pivot. Let* $c_n = \alpha^{1/n}$. *Then*

$$\mathbb{P}(Q \leq c_n) = \alpha.$$

*Also,* $\mathbb{P}(Q \leq 1) = 1$. *Therefore,*

$$\begin{aligned} 1 - \alpha &= \mathbb{P}(c \leq Q \leq 1) = \mathbb{P}\left(c \leq \frac{X_{(n)}}{\theta} \leq 1\right) \\ &= \mathbb{P}\left(\frac{1}{c} \geq \frac{\theta}{X_{(n)}} \geq 1\right) \\ &= \mathbb{P}\left(X_{(n)} \leq \theta \leq \frac{X_{(n)}}{c}\right) \end{aligned}$$

*so a* $1 - \alpha$ *confidence interval is*

$$\left(X_{(n)},\ \frac{X_{(n)}}{\alpha^{1/n}}\right).$$

# 5 Large Sample Confidence Intervals

We know that, under regularity conditions,

$$\frac{\widehat{\theta}_n - \theta}{se} \rightsquigarrow N(0, 1)$$

where $\widehat{\theta}_n$ is the mle and $se = 1/\sqrt{I_n(\widehat{\theta})}$. So this is an asymptotic pivot and an approximate confidence interval is

$$\widehat{\theta}_n \pm z_{\alpha/2} se.$$

By the delta method, a confidence interval for $\tau(\theta)$ is

$$\tau(\widehat{\theta}_n) \pm z_{\alpha/2} se(\widehat{\theta}) |\tau'(\widehat{\theta}_n)|.$$

By inverting the LRT and using the $\chi^2$ limiting distribution we get the LRT large sample confidence set:

$$C = \left\{\theta :\ -2 \log\left(\frac{L(\theta)}{L(\widehat{\theta})}\right) \leq \chi^2_{k,\alpha}\right\}.$$

Then

$$P_\theta(\theta \in C) \to 1 - \alpha$$

for each $\theta$.

**Example 10** *Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Using the Wald statistic*

$$\frac{\widehat{p} - p}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}} \rightsquigarrow N(0,1)$$

*so an approximate confidence interval is*

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

*Using the LRT we get*

$$C = \left\{ p : \ -2\log\left(\frac{p^Y(1-p)^{n-Y}}{\widehat{p}^Y(1-\widehat{p})^{n-Y}}\right) \leq \chi^2_{1,\alpha} \right\}.$$

*These intervals are different but, for large $n$, they are nearly the same. A finite sample interval can be constructed by inverting a test.*

# 6 A Pivot For the cdf

Let $X_1, \ldots, X_n \sim F$. We want to construct two functions $L(t) \equiv L(t, X)$ and $U(t) \equiv U(t, X)$ such that

$$P_F(L(t) \leq F(t) \leq U(t) \ \text{ for all } t) \geq 1 - \alpha$$

for all $F$.

Let

$$K_n = \sup_x |F_n(x) - F(x)|$$

where

$$F_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \leq x) = \frac{\#\{X_i \leq x\}}{n}$$

is the empirical distribiton function. We claim that $K_n$ is a pivot. To see this, let $U_i = F(X_i)$. Then $U_1, \ldots, U_n \sim \text{Uniform}(0,1)$. So

$$\begin{aligned} K_n &= \sup_x |F_n(x) - F(x)| \\ &= \sup_x \left| \frac{1}{n} \sum_{i=1}^n I(X_i \leq x) - F(x) \right| \\ &= \sup_x \left| \frac{1}{n} \sum_{i=1}^n I(F(X_i) \leq F(x)) - F(x) \right| \\ &= \sup_x \left| \frac{1}{n} \sum_{i=1}^n I(U_i \leq F(x)) - F(x) \right| \\ &= \sup_{0 \leq t \leq 1} \left| \frac{1}{n} \sum_{i=1}^n I(U_i \leq t) - t \right| \end{aligned}$$

and the latter has a distribution depending only on $U_1, \ldots, U_n$. We could find, by simulation, a number $c$ such that

$$\mathbb{P}\left( \sup_{0 \leq t \leq 1} \left| \frac{1}{n} \sum_{i=1}^n I(U_i \leq t) - t \right| > c \right) = \alpha.$$

A confidence set is then

$$C = \{F : \sup_x |F_n(x) - F(x)| < c\}.$$

# Lecture Notes 12
# Nonparametric Inference

**This is not in the text.**

Suppose we want to estimate something without assuming a parametric model. Some examples are:

1. Estimate the cdf $F$.
2. Estimate a density function $p(x)$.
3. Estimate a regression function $m(x) = \mathbb{E}(Y|X = x)$.
4. Estimate a functional $T(P)$ of a distribution $P$ for example $T(P) = \mathbb{E}(X) = \int x\, p(x)dx$.

## 1 The cdf and the Empirical Probability

We already solved this problem when we did VC theory. Given $X_1, \ldots, X_n \sim F$ where $X_i \in \mathbb{R}$ we use,

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x).$$

We saw that

$$\mathbb{P}\Big(\sup_x |\widehat{F}(x) - F(x)| > \epsilon\Big) \leq 2e^{-2n\epsilon^2}.$$

Hence,

$$\sup_x |\widehat{F}(x) - F(x)| \xrightarrow{\text{P}} 0$$

and

$$\sup_x |\widehat{F}(x) - F(x)| = O_P\left(\sqrt{\frac{1}{n}}\right).$$

It can be shown that this is the minimax rate of convergence. In other words,

More generally, for $X_i \in \mathbb{R}^d$, we set

$$P_n(A) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \in A).$$

We saw that, for any class $\mathcal{A}$ with VC dimension $v$,

$$\mathbb{P}\Big(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \epsilon\Big) \leq c_1 n^v e^{-c_2 n\epsilon^2}.$$

# 2 Density Estimation

$X_1, \ldots, X_n$ are iid with density $p$. For simplicity assume that $X_i \in \mathbb{R}$. What happens if we try to do maximum likelihood? The likelihood is

$$L(p) = \prod_{i=1}^{n} p(X_i).$$

We can make this as large as we want by making $p$ highly peaked at each $X_i$. So $\sup_p L(p) = \infty$ and the mle is the density that puts infinite spikes at each $X_i$.

We will need to put some restriction on $p$. For example

$$p \in \mathcal{P} = \left\{ p :\ p \geq 0,\ \int p = 1,\ \int |p''(x)|^2 dx \leq C \right\}.$$

The most commonly used nonparametric density estimator is probably the histogram. Another common estimator is the *kernel density estimator*. A *kernel* $K$ is a symmetric density function with mean 0. The estimator is

$$\widehat{p}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

where $h > 0$ is called the *bandwidth*.

The bandwidth controls the smoothness of the estimator. Larger $h$ makes $\widehat{f}_n$ smoother. As a loss function we will use

$$\mathcal{L}(p, \widehat{p}) = \int (p(x) - \widehat{p}(x))^2 dx.$$

The risk is

$$R = \mathbb{E}\left(\mathcal{L}(p, \widehat{p})\right) = \int \mathbb{E}(p(x) - \widehat{p}(x))^2 dx = \int (b^2(x) + v(x)) dx$$

where

$$b(x) = \mathbb{E}(\widehat{p}(x)) - p(x)$$

is the bias and

$$v(x) = \mathrm{Var}(\widehat{p}(x)).$$

Let

$$Y_i = \frac{1}{h} K\left(\frac{x - X_i}{h}\right).$$

Then $\widehat{p}_n(x) = n^{-1} \sum_{i=1}^n Y_i$ and

$$\begin{aligned}
\mathbb{E}(\widehat{p}(x)) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n Y_i\right) = \mathbb{E}(Y_i) \\
&= \mathbb{E}\left(\frac{1}{h} K\left(\frac{X_i - x}{h}\right)\right) \\
&= \int \frac{1}{h} K\left(\frac{u - x}{h}\right) p(u) du \\
&= \int K(t) p(x + ht) dt \qquad \text{where } u = x + ht \\
&= \int K(t)\left(p(x) + htp'(x) + \frac{h^2 t^2}{2} p''(x) + o(h^2)\right) dt \\
&= p(x)\int K(t)dt + hp'(x)\int tK(t)dt + \frac{h^2}{2} p''(x)\int t^2 K(t)dt + o(h^2)dt \\
&= (p(x) \times 1) + (hp'(x) \times 0) + \frac{h^2}{2} p''(x)\kappa + o(h^2)
\end{aligned}$$

where $\kappa = \int t^2 K(t)dt$. So

$$\mathbb{E}(\widehat{p}(x)) \approx p(x) + \frac{h^2}{2} p''(x)\kappa$$

and

$$b(x) \approx \frac{h^2}{2} p''(x)\kappa.$$

Thus

$$\int b^2(x)dx = \frac{h^4}{4}\kappa^2 \int (p''(x))^2 dx.$$

Now we compute the variance. We have

$$v(x) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n Y_i\right) = \frac{\mathrm{Var} Y_i}{n} = \frac{\mathbb{E}(Y_i^2) - (\mathbb{E}(Y_i))^2}{n}.$$

Now

$$\begin{aligned}
\mathbb{E}(Y_i^2) &= \mathbb{E}\left(\frac{1}{h^2}K^2\left(\frac{X_i - x}{h}\right)\right) \\
&= \int \frac{1}{h^2}K^2\left(\frac{u-x}{h}\right)p(u)du \\
&= \frac{1}{h}\int K^2(t)p(x+ht)dt \quad \mathrm{u = x + ht} \\
&\approx \frac{p(x)}{h}\int K^2(t)dt = \frac{p(x)\xi}{h}
\end{aligned}$$

where $\xi = \int K^2(t)dt$. Now

$$(\mathbb{E}(Y_i))^2 \approx \left(p(x) + \frac{h^2}{2}p''(x)\kappa\right)^2 = f^2(x) + O(h^2) \approx f^2(x).$$

So

$$v(x) = \frac{\mathbb{E}(Y_i^2)}{n} - \frac{(\mathbb{E}(Y_i))^2}{n} \approx \frac{p(x)}{nh} + f^2(x) = \frac{p(x)\xi}{nh} + o\left(\frac{1}{nh}\right) \approx \frac{p(x)\xi}{nh}$$

and

$$\int v(x)dx \approx \frac{\xi}{nh}.$$

Finally,

$$R \approx \frac{h^4}{4}\kappa^2\int (p''(x))^2dx + \frac{\xi}{nh} = Ch^4 + \frac{\xi}{nh}.$$

Note that

$$h \uparrow \quad \longrightarrow \quad \text{bias} \uparrow,\ \text{variance} \downarrow$$

$$h \downarrow \quad \longrightarrow \quad \text{bias} \downarrow,\ \text{variance} \uparrow .$$

If we choose $h = h_n$ to satisfy

$$h_n \to 0, \quad nh_n \to \infty$$

then we see that $\widehat{p}_n(x) \xrightarrow{\mathrm{P}} p(x)$.

If we minimize over $h$ we get

$$h = \left(\frac{\xi}{4nC}\right)^{1/5} = O\left(\frac{1}{n}\right)^{1/5}.$$

This gives

$$R = \frac{C_1}{n^{4/5}}$$

for some constant $C_1$.

Can we do better? The answer, based on minimax theory, is no.

**Theorem 1** *There is a constant $a$ such that*

$$\inf_{\widehat{p}} \sup_{f \in \mathcal{F}} R(f, \widehat{p}) \geq \frac{a}{n^{4/5}}.$$

So the kernel estimator achieves the minimax rate of convergence. The histogram converges at the sub-optimal rate of $n^{-2/3}$. Proving these facts is beyond the scope of the course.

There are many practical questions such as: how to choose $h$ in practice, how to extend to higher dimensions etc. These are discussed in 10-702 as well as other courses.

# 3 Regression

We observe $(X_1, Y_1), \ldots, (X_n, Y_n)$. Given a new $X$ we want to predict $Y$. If our prediction is $m(X)$ then the predictive loss os $(Y - m(X))^2$. Later in the course we will discuss prediction in detail and we will see that the optimal predictor is the regression function

$$m(x) = \mathbb{E}(Y|X = x) = \int y p(y|x) dy.$$

The kernel estimator is

$$\widehat{m}_n(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)}.$$

The properties are similar to kernel density estimation. Again, you will study this in more detail in some other classes.

# 4 Functionals

Let $X_1, \ldots, X_n \sim F$. Let $\mathcal{F}$ be all distributions. A map $T : \mathcal{F} \to \mathbb{R}$ is called a statistical functional.

**Notation.** Let $F$ be a distribution function. Let $f$ denote the probability mass function if $F$ is discrete and the probability density function if $F$ is continuous. The integral $\int g(x)dF(x)$ is interpreted as follows:

$$\int g(x)dF(x) = \begin{cases} \sum_j g(x_j)p(x_j) & \text{if } F \text{ is discrete} \\ \int g(x)p(x)dx & \text{if } F \text{ is continuous.} \end{cases}$$

A **statistical functional** $T(F)$ is any function of of the cdf $F$. Examples include the mean $\mu = \int x\, dF(x)$, the variance $\sigma^2 = \int (x-\mu)^2 dF(x)$, the median $m = F^{-1}(1/2)$, and the largest eigenvalue of the covariance matrix $\Sigma$.

The **plug-in estimator** of $\theta = T(F)$ is defined by

$$\widehat{\theta}_n = T(\widehat{F}_n).$$

A functional of the form $\int a(x)dF(x)$ is called a **linear functional.** The empirical cdf $\widehat{F}_n(x)$ is discrete, putting mass $1/n$ at each $X_i$. Hence, if $T(F) = \int a(x)dF(x)$ is a linear functional then the plug-in estimator for linear functional $T(F) = \int a(x)dF(x)$ is:

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

Let $\widehat{\mathsf{se}}$ be an estimate of the standard error of $T(\widehat{F}_n)$. In many cases, it turns out that

$$\widehat{\theta}_n = T(\widehat{F}_n) \approx N(T(F), \widehat{\mathsf{se}}^2).$$

In that case, an approximate $1-\alpha$ confidence interval for $T(F)$ is then

$$\widehat{\theta}_n \pm z_{\alpha/2}\, \widehat{\mathsf{se}}.$$

We can use the Wald statistic

$$W = \frac{\widehat{\theta}_n - \theta_0}{\text{se}}$$

to do a hypothesis test.

**Example 2 (The mean)** *Let* $\mu = T(F) = \int x\, dF(x)$. *The plug-in estimator is* $\widehat{\mu} = \int x\, d\widehat{F}_n(x) = \overline{X}_n$. *The standard error is* $\mathsf{se} = \sqrt{\mathsf{Var}(\overline{X}_n)} = \sigma/\sqrt{n}$. *If* $\widehat{\sigma}$ *denotes an estimate of* $\sigma$, *then the estimated standard error is* $\widehat{\mathsf{se}} = \widehat{\sigma}/\sqrt{n}$. *A Normal-based confidence interval for* $\mu$ *is* $\overline{X}_n \pm z_{\alpha/2}\, \widehat{\sigma}/\sqrt{n}$.

**Example 3 (The variance)** *Let* $\sigma^2 = \mathsf{Var}(X) = \int x^2\, dF(x) - \left(\int x\, dF(x)\right)^2$. *The plug-in estimator is*

$$\widehat{\sigma}^2 = \int x^2 d\widehat{F}_n(x) - \left(\int x d\widehat{F}_n(x)\right)^2 \tag{1}$$

$$= \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2 \tag{2}$$

$$= \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2. \tag{3}$$

**Example 4 (Quantiles)** *Let* $F$ *be strictly increasing with density* $f$. *Let* $T(F) = F^{-1}(p)$ *be the* $p^{\text{th}}$ *quantile. The estimate of* $T(F)$ *is* $\widehat{F}_n^{-1}(p)$. *We have to be a bit careful since* $\widehat{F}_n$ *is not invertible. To avoid ambiguity we define* $\widehat{F}_n^{-1}(p) = \inf\{x :\ \widehat{F}_n(x) \geq p\}$. *We call* $\widehat{F}_n^{-1}(p)$ *the* $p^{th}$ **sample quantile**.

How do we estimate the standard error? There are two approaches. One is based on something called *the influence function* which is a nonparametric version of the score function. We won't cover that in this course. The second approach is to use the *bootstrap* which we will discuss in an upcoming lecture.

# 5 Optional: The Influence Function

If you are curious what the influence is, I will describe it here. This section is optional and you can skip it if you prefer.

The **influence function** is defined by

$$L_F(x) = \lim_{\epsilon \to 0} \frac{T\big((1-\epsilon)F + \epsilon\delta_x\big) - T(F)}{\epsilon}$$

where $\delta_x$ denote a point mass distribution at $x$: $\delta_x(y) = 0$ if $y < x$ and $\delta_x(y) = 1$ if $y \geq x$. The **empirical influence function** is defined by

$$\widehat{L}(x) = \lim_{\epsilon \to 0} \frac{T\big((1-\epsilon)\widehat{F}_n + \epsilon\delta_x\big) - T(\widehat{F}_n)}{\epsilon}.$$

The influence function is the nonparametric version of the score function. More precisely, it behaves like the score divided by the Fisher information, $L = \text{score/information} = S/I$.

**Theorem 5** *If $T$ is Hadamard differentiable*[1] *with respect to $d(F,G) = \sup_x |F(x) - G(x)|$ then*

$$\sqrt{n}(T(\widehat{F}_n) - T(F)) \rightsquigarrow N(0, \tau^2)$$

*where $\tau^2 = \int L_F^2(x) dF(x)$. Also,*

$$\frac{(T(\widehat{F}_n) - T(F))}{\widehat{\mathsf{se}}} \rightsquigarrow N(0,1)$$

*where $\widehat{\mathsf{se}} = \widehat{\tau}/\sqrt{n}$ and*

$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \widehat{L}^2(X_i).$$

We call the approximation $(T(\widehat{F}_n) - T(F))/\widehat{\mathsf{se}} \approx N(0,1)$ the **functional delta method** or the **nonparametric delta method**.

From the normal approximation, a large sample confidence interval is:

$$T(\widehat{F}_n) \pm z_{\alpha/2}\, \widehat{\mathsf{se}}.$$

**Example 6 (The mean)** *Let $\theta = T(F) = \int x\, dF(x)$. The plug-in estimator is $\widehat{\theta} = \int x\, d\widehat{F}_n(x) = \overline{X}_n$. Also, $T((1-\epsilon)F + \epsilon\delta_x) = (1-\epsilon)\theta + \epsilon x$. Thus, $L(x) = x - \theta$, $\widehat{L}(x) = x - \overline{X}_n$ and $\widehat{\mathsf{se}}^2 = \widehat{\sigma}^2/n$ where $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. A pointwise asymptotic nonparametric 95 percent confidence interval for $\theta$ is $\overline{X}_n \pm 2\,\widehat{\mathsf{se}}$.*

[1] Hadamard differentiability is a smoothness condition on $T$.

**Example 7 (Quantiles)** *Let $F$ be strictly increasing with positive density $f$, and let $T(F) = F^{-1}(p)$ be the $p^{\text{th}}$ quantile. The influence function is*

$$L(x) = \begin{cases} \frac{p-1}{p(\theta)}, & x \leq \theta \\ \frac{p}{p(\theta)}, & x > \theta. \end{cases}$$

*The asymptotic variance of $T(\widehat{F}_n)$ is*

$$\frac{\tau^2}{n} = \frac{1}{n}\int L^2(x)dF(x) = \frac{p(1-p)}{nf^2(\theta)}.$$

# Lecture Notes 13
# The Bootstrap

**This is mostly not in the text.**

## 1 Introduction

Can we estimate the mean of a distribution without using a parametric model? Yes. The key idea is to first estimate the distribution function nonparametrically. Then we can get an estimate of the mean (and many other parameters) from the distribution function.

How can we get the standard error of that estimator? The answer is: the bootstrap. The bootstrap is a nonparametric method for finding standard errors and confidence intervals.

**Notation.** Let $F$ be a distribution function. Let $p$ denote the probability mass function if $F$ is discrete and the probability density function if $F$ is continuous. The integral $\int g(x)dF(x)$ is interpreted as follows:

$$\int g(x)dF(x) = \begin{cases} \sum_j g(x_j)p(x_j) & \text{if } F \text{ is discrete} \\ \int g(x)p(x)dx & \text{if } F \text{ is continuous.} \end{cases} \tag{1}$$

For $0 < \alpha < 1$ define $z_\alpha$ by $\mathbb{P}(Z > z_\alpha) = \alpha$ where $Z \sim N(0,1)$. Thus $z_\alpha = \Phi^{-1}(1-\alpha) = -\Phi^{-1}(\alpha)$.

## 2 Review of The Empirical Distribution Function

The bootstrap uses the empirical distribution function. Let $X_1, \ldots, X_n \sim F$ where $F(x) = \mathbb{P}(X \leq x)$ is a distribution function on the real line. We can estimate $F$ with the **empirical distribution function** $\widehat{F}_n$, the cdf that puts mass $1/n$ at each data point $X_i$.

Recall that the empirical distribution function $\widehat{F}_n$ is defined by

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) \tag{2}$$

where

$$I(X_i \leq x) = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{if } X_i > x. \end{cases} \tag{3}$$

From (1) it follows that $\int g(x) d\widehat{F}_n(x) = n^{-1} \sum_{i=1}^n g(X_i)$. According to the **Glivenko–Cantelli Theorem**,

$$\sup_x |\widehat{F}_n(x) - F(x)| \xrightarrow{\text{as}} 0. \tag{4}$$

Hence, $\widehat{F}_n$ is a consistent estimator of $F$. In fact, the convergence is fast. According to the **Dvoretzky–Kiefer–Wolfowitz (DKW)** inequality, for any $\epsilon > 0$,

$$\mathbb{P}\left( \sup_x |F(x) - \widehat{F}_n(x)| > \epsilon \right) \leq 2e^{-2n\epsilon^2}. \tag{5}$$

If $\epsilon_n = c_n/\sqrt{n}$ where $c_n \to \infty$, then $\mathbb{P}(\sup_x |F(x) - \widehat{F}_n(x)| > \epsilon_n) \to 0$. Hence, $\sup_x |F(x) - \widehat{F}_n(x)| = O_P(n^{-1/2})$.

# 3 Statistical Functionals

Recall that a **statistical functional** $T(F)$ is any function of of the cdf $F$. Examples include the mean $\mu = \int x \, dF(x)$, the variance $\sigma^2 = \int (x - \mu)^2 dF(x)$, $m = F^{-1}(1/2)$, and the largest eigenvalue of the covariance matrix $\Sigma$.

The **plug-in estimator** of $\theta = T(F)$ is defined by

$$\widehat{\theta}_n = T(\widehat{F}_n). \tag{6}$$

Let $\widehat{\mathsf{se}}$ be an estimate of the standard error of $T(\widehat{F}_n)$. (We will see how to get this later.) In many cases, it turns out that

$$T(\widehat{F}_n) \approx N(T(F), \widehat{\mathsf{se}}^2). \tag{7}$$

In that case, an approximate $1 - \alpha$ confidence interval for $T(F)$ is then

$$T(\widehat{F}_n) \pm z_{\alpha/2} \, \widehat{\mathsf{se}}. \tag{8}$$

**Example 1 (The mean)** *Let* $\mu = T(F) = \int x\, dF(x)$. *The plug-in estimator is* $\widehat{\mu} = \int x\, d\widehat{F}_n(x) = \overline{X}_n$. *The standard error is* $\mathsf{se} = \sqrt{\mathsf{Var}(\overline{X}_n)} = \sigma/\sqrt{n}$. *If* $\widehat{\sigma}$ *denotes an estimate of* $\sigma$, *then the estimated standard error is* $\widehat{\mathsf{se}} = \widehat{\sigma}/\sqrt{n}$. *A Normal-based confidence interval for* $\mu$ *is* $\overline{X}_n \pm z_{\alpha/2}\, \widehat{\sigma}/\sqrt{n}$.

**Example 2** *A functional of the form* $\int a(x)dF(x)$ *is called a* **linear functional.** *(Recall that* $\int a(x)dF(x)$ *is defined to be* $\int a(x)p(x)dx$ *in the continuous case and* $\sum_j a(x_j)p(x_j)$ *in the discrete case.) The empirical* cdf $\widehat{F}_n(x)$ *is discrete, putting mass* $1/n$ *at each* $X_i$. *Hence, if* $T(F) = \int a(x)dF(x)$ *is a linear functional then the plug-in estimator for linear functional* $T(F) = \int a(x)dF(x)$ *is:*

$$T(\widehat{F}_n) = \int a(x)d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} a(X_i). \tag{9}$$

**Example 3 (The variance)** *Let* $\sigma^2 = \mathsf{Var}(X) = \int x^2\, dF(x) - \left(\int x\, dF(x)\right)^2$. *The plug-in estimator is*

$$\begin{aligned}
\widehat{\sigma}^2 &= \int x^2 d\widehat{F}_n(x) - \left(\int x d\widehat{F}_n(x)\right)^2 && (10)\\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 && (11)\\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2. && (12)
\end{aligned}$$

**Example 4 (The skewness)** *Let* $\mu$ *and* $\sigma^2$ *denote the mean and variance of a random variable* $X$. *The skewness — which measures the lack of symmetry of a distribution — is defined to be*

$$\kappa = \frac{\mathbb{E}(X-\mu)^3}{\sigma^3} = \frac{\int (x-\mu)^3 dF(x)}{\left\{\int (x-\mu)^2 dF(x)\right\}^{3/2}}. \tag{13}$$

*To find the plug-in estimate, first recall that* $\widehat{\mu} = n^{-1}\sum_{i=1}^{n} X_i$ *and* $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2$. *The plug-in estimate of* $\kappa$ *is*

$$\widehat{\kappa} = \frac{\int (x-\mu)^3 d\widehat{F}_n(x)}{\left\{\int (x-\mu)^2 d\widehat{F}_n(x)\right\}^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})^3}{\widehat{\sigma}^3}. \tag{14}$$

**Example 5 (Correlation)** *Let $Z = (X, Y)$ and let $\rho = T(F) = \mathbb{E}(X-\mu_X)(Y-\mu_Y)/(\sigma_x\sigma_y)$ denote the correlation between $X$ and $Y$, where $F(x, y)$ is bivariate. We can write $T(F) = a(T_1(F), T_2(F), T_3(F), T_4(F), T_5(F))$ where*

$$\begin{aligned} T_1(F) &= \int x\, dF(z) \quad T_2(F) = \int y\, dF(z) \quad T_3(F) = \int xy\, dF(z) \\ T_4(F) &= \int x^2\, dF(z) \quad T_5(F) = \int y^2\, dF(z) \end{aligned} \tag{15}$$

*and*

$$a(t_1, \ldots, t_5) = \frac{t_3 - t_1 t_2}{\sqrt{(t_4 - t_1^2)(t_5 - t_2^2)}}. \tag{16}$$

*Replace $F$ with $\widehat{F}_n$ in $T_1(F)$, ..., $T_5(F)$, and take*

$$\widehat{\rho} = a(T_1(\widehat{F}_n), T_2(\widehat{F}_n), T_3(\widehat{F}_n), T_4(\widehat{F}_n), T_5(\widehat{F}_n)). \tag{17}$$

*We get*

$$\widehat{\rho} = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \overline{X}_n)^2}\sqrt{\sum_{i=1}^n (Y_i - \overline{Y}_n)^2}} \tag{18}$$

*which is called the* **sample correlation**.

**Example 6 (Quantiles)** *Let $F$ be strictly increasing with density $f$. Let $T(F) = F^{-1}(p)$ be the $p^{\text{th}}$ quantile. The estimate of $T(F)$ is $\widehat{F}_n^{-1}(p)$. We have to be a bit careful since $\widehat{F}_n$ is not invertible. To avoid ambiguity we define $\widehat{F}_n^{-1}(p) = \inf\{x :\ \widehat{F}_n(x) \geq p\}$. We call $\widehat{F}_n^{-1}(p)$ the $p^{th}$* **sample quantile**.

# 4 The Bootstrap

Let $T_n = g(X_1, \ldots, X_n)$ be a statistic and let $\mathsf{Var}_F(T_n)$ denote the variance of $T_n$. We have added the subscript $F$ to emphasize that the variance is itself a function of $F$. In other words

$$\mathsf{Var}_F(T_n) = \int\int \cdots \int (g(X_1, \ldots, X_n) - \mu)^2 dF(x_1) dF(x_2) \cdots dF(x_n)$$

where

$$\mu = \mathbb{E}(T_n) = \int\int \cdots \int g(X_1, \ldots, X_n) dF(x_1) dF(x_2) \cdots dF(x_n).$$

If we knew $F$ we could, at least in principle, compute the variance. For example, if $T_n = n^{-1}\sum_{i=1}^n X_i$, then

$$\mathsf{Var}_F(T_n) = \frac{\sigma^2}{n} = \frac{\int x^2 dF(x) - \left(\int x dF(x)\right)^2}{n}. \tag{19}$$

In other words, the variance of $\widehat{\theta} = T(F_n)$ is itself a function of $F$. We can write

$$\mathsf{Var}_F(T_n) = U(F)$$

for some $U$. Therefore, to estimate $\mathsf{Var}_F(T_n)$ we can use

$$\widehat{\mathsf{Var}_F(T_n)} = U(\widehat{F}_n).$$

This is the bootstrap estimate of the standard error. To repeat: we estimate $U(F) = \mathsf{Var}_F(T_n)$ with $U(\widehat{F}_n) = \mathsf{Var}_{\widehat{F}_n}(T_n)$. In other words, we use a plug-in estimator of the variance.

But how can we compute $\mathsf{Var}_{\widehat{F}_n}(T_n)$? We approximate it with a simulation estimate denoted by $v_{\text{boot}}$. Specifically, we do the following steps:

---

**Bootstrap Variance Estimation**

1. Draw $X_1^*, \ldots, X_n^* \sim \widehat{F}_n$.
2. Compute $T_n^* = g(X_1^*, \ldots, X_n^*)$.
3. Repeat steps 1 and 2, $B$ times to get $T_{n,1}^*, \ldots, T_{n,B}^*$.
4. Let

$$v_{\text{boot}} = \frac{1}{B}\sum_{b=1}^B \left(T_{n,b}^* - \frac{1}{B}\sum_{r=1}^B T_{n,r}^*\right)^2. \tag{20}$$

---

By the law of large numbers, $v_{\text{boot}} \xrightarrow{\text{as}} \mathsf{Var}_{\widehat{F}_n}(T_n)$ as $B \to \infty$. The estimated standard error of $T_n$ is $\widehat{\mathsf{se}}_{\text{boot}} = \sqrt{v_{\text{boot}}}$. The following diagram illustrates the bootstrap idea:

$$\begin{array}{rccccc} \text{Real world:} & F & \Longrightarrow & X_1, \ldots, X_n & \Longrightarrow & T_n = g(X_1, \ldots, X_n) \\ \text{Bootstrap world:} & \widehat{F}_n & \Longrightarrow & X_1^*, \ldots, X_n^* & \Longrightarrow & T_n^* = g(X_1^*, \ldots, X_n^*) \end{array}$$

**Bootstrap for the Median**

```
Given data X = (X(1), ..., X(n)):


T     = median(X)
Tboot = vector of length B
for(i in 1:N){
    Xstar    = sample of size n from X (with replacement)
    Tboot[i] = median(Xstar)
    }
se = sqrt(variance(Tboot))
```

Figure 1: Pseudo-code for bootstrapping the median.

$$\mathsf{Var}_F(T_n) \overbrace{\approx}^{O(1/\sqrt{n})} \mathsf{Var}_{\widehat{F}_n}(T_n) \overbrace{\approx}^{O(1/\sqrt{B})} v_{boot}. \tag{21}$$

How do we simulate from $\widehat{F}_n$? Since $\widehat{F}_n$ gives probability $1/n$ to each data point, drawing $n$ points at random from $\widehat{F}_n$ is the same as drawing a sample of size $n$ with replacement from the original data. Therefore step 1 can be replaced by:

1. **Draw $X_1^*, \ldots, X_n^*$ with replacement from $X_1, \ldots, X_n$.**

**Example 7** *Figure 1 shows pseudo-code for using the bootstrap to estimate the standard error of the median.*

# 5 The Parametric Bootstrap

So far, we have estimated $F$ nonparametrically. There is also a **parametric bootstrap.** If $F_\theta$ depends on a parameter $\theta$ and $\widehat{\theta}$ is an estimate of $\theta$, then we simply sample from $F_{\widehat{\theta}}$

instead of $\widehat{F}_n$. This is just as accurate, but much simpler than, the delta method. Here is more detail.

Suppose that $X_1, \ldots, X_n \sim p(x;\theta)$. Let $\widehat{\theta}$ be the mle. Let $\tau = g(\theta)$. Then $\widehat{\tau} = g(\widehat{\theta})$. To get the standard error of $\widehat{\tau}$ we need to compute the Fisher information and then do the delta method. The bootstrap allows us to avoid both steps. We just do the following:

1. Compute the estimate $\widehat{\theta}$ from the data $X_1, \ldots, X_n$.
2. Draw a sample $X_1^*, \ldots, X_n^* \sim f(x;\widehat{\theta})$. Compute $\widehat{\theta}_1^*$ and $\widehat{\tau}_1^* = g(\widehat{\theta}_1^*)$ from the new data. Repeat $B$ times to get $\widehat{\tau}_1^*, \ldots, \widehat{\tau}_B^*$.
3. Compute the standard deviation

$$\widehat{\text{se}} = \frac{1}{B}\sum_{b=1}^{B}(\widehat{\tau}_j^* - \overline{\tau})^2 \qquad \text{where} \qquad \overline{\tau} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\tau}_j^*. \tag{22}$$

No need to get the Fisher information or do the delta method.

# 6 Bootstrap Confidence Intervals

There are several ways to construct bootstrap confidence intervals. They vary in ease of calculation and accuracy.

**Normal Interval.** The simplest is the Normal interval

$$\widehat{\theta}_n \pm z_{\alpha/2}\, \widehat{\mathsf{se}}_{\text{boot}} \tag{23}$$

where $\widehat{\mathsf{se}}_{\text{boot}}$ is the bootstrap estimate of the standard error.

**Pivotal Intervals.** Let $\theta = T(F)$ and $\widehat{\theta}_n = T(\widehat{F}_n)$. We can also construct an approximate confidence interval for $\theta$ using the (approximate) pivot $\sqrt{n}(\widehat{\theta}^* - \widehat{\theta})$ as follows:

$$C = \left( \widehat{\theta}_n - \frac{\breve{H}^{-1}\left(1 - \frac{\alpha}{2}\right)}{\sqrt{n}},\ \widehat{\theta}_n - \frac{\breve{H}^{-1}\left(\frac{\alpha}{2}\right)}{\sqrt{n}} \right) \tag{24}$$

where

$$\breve{H}(r) = \frac{1}{B}\sum_{j=1}^{B} I\big(\sqrt{n}(\widehat{\theta}_j^* - \widehat{\theta}) \leq r\big) \tag{25}$$

where

$$H(r) = P\big(\sqrt{n}(\widehat{\theta}_n - \theta) \leq r\big), \qquad \widehat{H}(r) = P_n\big(I\big(\sqrt{n}(\widehat{\theta}_j^* - \widehat{\theta}) \leq r\big). \tag{26}$$

**Theorem 8** *Under appropriate conditions on $T$, $\sup_u |H(u) - \widehat{H}(u)| \xrightarrow{P} 0$ as $n \to \infty$ and $\sup_u |\widehat{H}(u) - \breve{H}(u)| \xrightarrow{P} 0$ as $B \to \infty$.*

Now we can show that the confidence interval has coverage that is approximately equal to $1-\alpha$. Applying Theorem 8 we have

$$\begin{aligned}
\mathbb{P}(\theta \in C) &= \mathbb{P}\left(\widehat{\theta}_n - \frac{\breve{H}^{-1}\left(1-\frac{\alpha}{2}\right)}{\sqrt{n}} \leq \theta \leq \widehat{\theta}_n - \frac{\breve{H}^{-1}\left(\frac{\alpha}{2}\right)}{\sqrt{n}}\right) \\
&= \mathbb{P}\left(\breve{H}^{-1}\left(\frac{\alpha}{2}\right) \leq \sqrt{n}(\widehat{\theta}_n - \theta) \leq \breve{H}^{-1}\left(1-\frac{\alpha}{2}\right)\right) \\
&= H\left(\breve{H}^{-1}\left(1-\frac{\alpha}{2}\right)\right) - H\left(\breve{H}^{-1}\left(\frac{\alpha}{2}\right)\right) \\
&\approx H\left(\widehat{H}^{-1}\left(1-\frac{\alpha}{2}\right)\right) - H\left(\widehat{H}^{-1}\left(\frac{\alpha}{2}\right)\right) \\
&\approx H\left(H^{-1}\left(1-\frac{\alpha}{2}\right)\right) - H\left(H^{-1}\left(\frac{\alpha}{2}\right)\right) \\
&= \left(1-\frac{\alpha}{2}\right) - \frac{\alpha}{2} = 1-\alpha.
\end{aligned}$$

# 7 Remarks About The Bootstrap

1. The bootstrap is nonparametric but it does require some assumptions. You can't assume it is always valid. (See the appendix.)
2. The bootstrap is an asymptotic method. Thus the coverage of the confidence interval is $1-\alpha+r_n$ where the remainder $r_n \to 0$ as $n \to \infty$.
3. There is a related method called the jackknife where the standard error is estimated by leaving out one observation at a time. However, the bootstrap is valid under weaker conditions than the jackknife. See Shao and Tu (1995).

4. Another way to construct a bootstrap confidence interval is to set $C = [a, b]$ where $a$ is the $\alpha/2$ quantile of $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$ and $b$ is the $1-\alpha/2$ quantile. This is called the percentile interval. This interval seems very intuitive but does not have the theoretical support of the interval in (24). However, in practice, the percentile interval and the interval in (24) are often quite similar.
5. There are many cases where the bootstrap is not formally justified. This is especially true with discrete structures like trees and graphs. Nonethless, the bootstrap can be used in an informal way to get some intuition of the variability of the procedure. But keep in mind that the formal guarantees may not apply in these cases. For example, see Holmes (2003) for a discussion of the bootstrap applied to phylogenetic tres.
6. There is an improvement on the bootstrap called subsampling. In this case, we draw samples of size $m < n$ without replacement. Subsampling produces valid confidence intervals under weaker conditions than the bootstrap. See Politis, Romano and Wolf (1999).
7. There are many modifications of the bootstrap that lead to more accurate confidence intervals; see Efron (1996).

# 8 Examples

**Example 9 (The Median)** *The top left plot of Figure 2 shows the density for a $\chi^2$ distribution with 4 degrees of freedom. The top right plot shows a histogam of $n = 50$ draws from this distribution. Let $\theta = T(P)$ be the median. The true value is $\theta = 3.36$. The samlpe median turns out to be $\widehat{\theta}_n = 3.22$. We computed $B = 1000$ bootstrap values $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$ shown in the histogram (bottom left plot). The estimated standard error is 0.35. This is smaller than the true standard error which is 0.44.*

*Next we conducted a small simulation. We drew a sample of size $n$ and computed the 95 percent bootstrap confidence interval. We repeated this process $N = 100$ times. The bottom right plot shows the 100 intervals. The vertical line is the true value of $\theta$. The percentage*

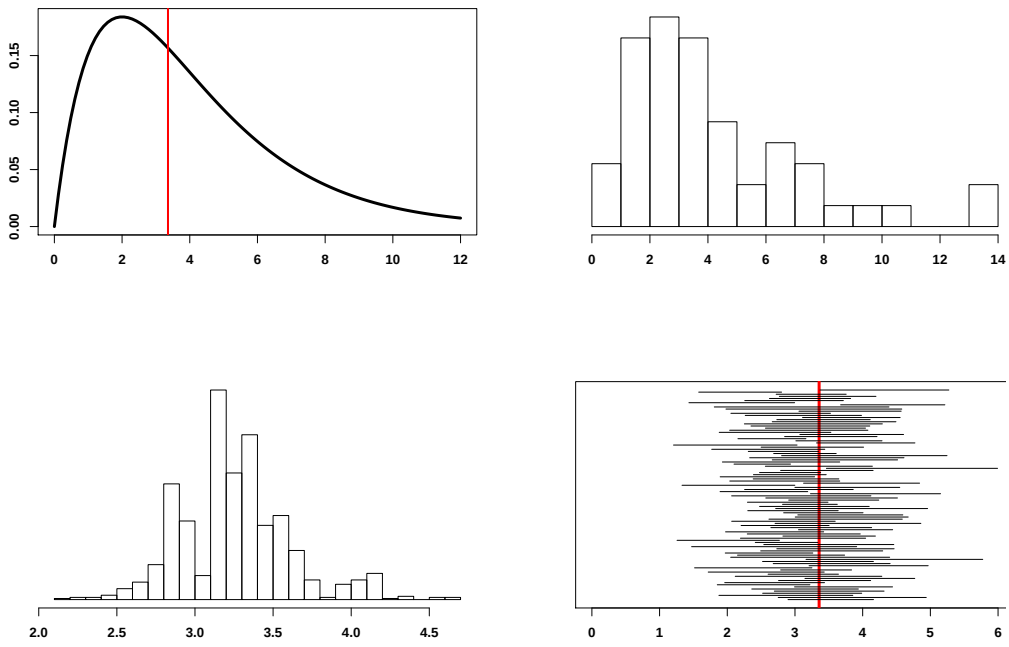



Figure 2: Top left: density of a $\chi^2$ with 4 degrees of freedom. The vertical line shows the median. Top right: $n = 50$ draw from the distribution. Bottom left: $B = 1000$ boostrap values $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$. Bottom right: Bootstrap confidence intervals from 100 experiments.

*of intervals that cover $\theta$ is 0.83 which shows that the bootstrap interval undercovers in this case.*

**Example 10 (Nonparametric Regression)** *The bootstrap is often used informally to get a sense of the variability of a procedure. Consider the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ in the top left plot of Figure 3. To estimate the regression function $m(x) = \mathbb{E}(Y|X = x)$ we use a kernel regression estimator given by $\widehat{m}(x) = \sum_{i=1}^n Y_i\, w_i(x)$ where $w_i(x) = K((x - X_i)/h)/\sum_j K((x - X_j)/h)$ and $K(x) = e^{-x^2/2}$ is a Gaussian kernel. The estimated curve is shown in the top right plot. We now create $B = 1,000$ boostrap replications resulting in curves $\widehat{m}_1^*, \ldots, \widehat{m}_B^*$ in the bottom left plot. At each $x$, we find the .025 and .975 quantile of*

*the bootstrap replications. This reults in the upper and lower band in the bottom right plot. The bootstrap reveals greater variability in the estimated curve around $x = -0.5$. The reason why we call this an informal use of the bootstrap is that the bands shown in the lower right plot are not rigorous confidence bands. There are several reasons for this. First, we used a percentile interval (described in the earlier list of remarks) rather than the interval defined by (24). Second, we have not adjusted for the fact that we are making simultaneous bands over all $x$. Finally, the theory of the bootstrap does not directly apply to nonparametric smoothing. Roughly speaking, we are really creating approximate confidence intervals for $\mathbb{E}(\widehat{m}(x))$ rather than for $m(x)$. Despite these shortcomings, the bootstrap is still regarded as a useful tool here but we must keep in mind that it is being used in an informal way. Some authors refer to the bands as variability bands rather than confidence bands for this reason.*

**Example 11 (Estimating Eigenvalues)** *Let $X_1, \ldots, X_n$ be random vectors where $X_i \in \mathbb{R}^p$ and let $\Sigma$ be the covariance matrix of $X_i$. A common dimension reduction technique is principal components which involves finding the spectral decomposition $\Sigma = E\Lambda E^T$ where the columns of $E$ are the eigenvectors of $\Sigma$ and $\Lambda$ is a diagonal matrix whose diagonal elements are the ordered eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$. The data dimension can be reduced to $q < p$ by projecting each data point onto the first $q$ eigenvalues. We choose $q$ such that $\sum_{j=q+1}^{p} \lambda_j^2$ is small. Of course, we need to estimate the eigenvectors and eigenvalues. For now, let us focus on estimating the largest eigenvalue and denote this by $\theta$. An estimate of $\theta$ is the largest principal component $\widehat{\theta}$ of the sample covariance matrix*

$$S = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^T. \tag{27}$$

*It is not at all obvious how can can estimate the standard error of $\widehat{\theta}$ or how to find a confidence interval for $\theta$. In this example, the bootstrap works as follows. Draw a sample of size $n$ with replacement from $X_1, \ldots, X_n$. The new sample is denoted by $X_1^*, \ldots, X_n^*$. Compute the sample covariance matrix $S^*$ of the new data and let $\widehat{\theta}^*$ denote the largest eigenvector of $S^*$. Repeat this process $B$ times where $B$ is typically about 10,000. This yields bootstrap values*

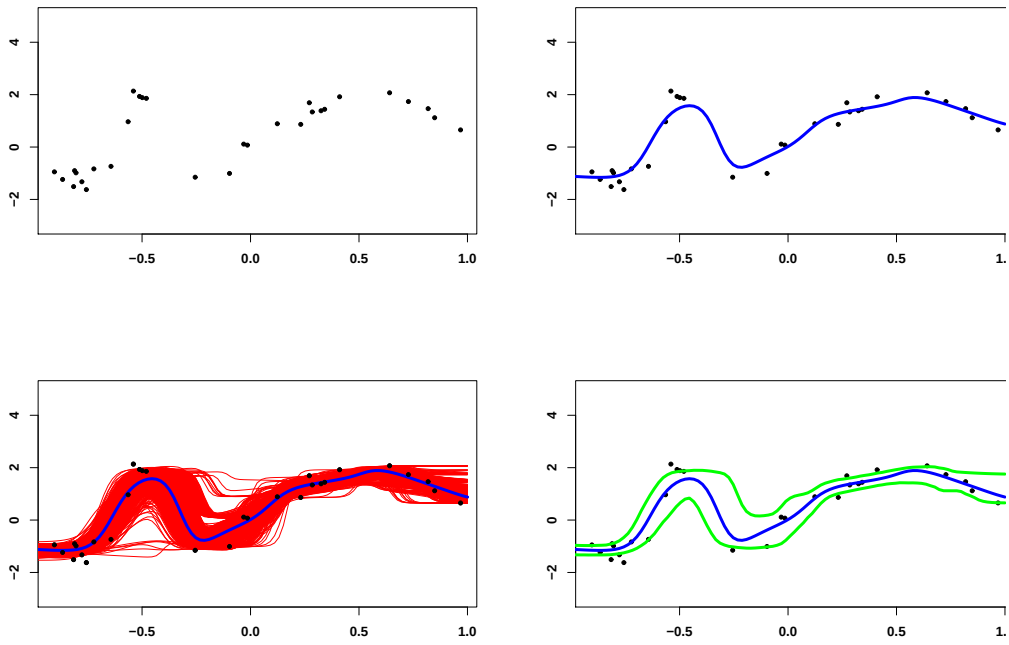



Figure 3: Top left: the data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Top right: kernel regression estimator. Bottom left: 1,000 bootstrap replications. Bottom right: 95 percent variability bands.

.

$\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$. *The standard deviation of* $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$ *is an estimate of the standard error of the original estimator* $\widehat{\theta}$.

*Figure 4 shows a PCA analysis of US arrest data. The last plot shows bootstrap replications of the first principal component.*

**Example 12 (Median Regression)** *Consider the linear regression model*

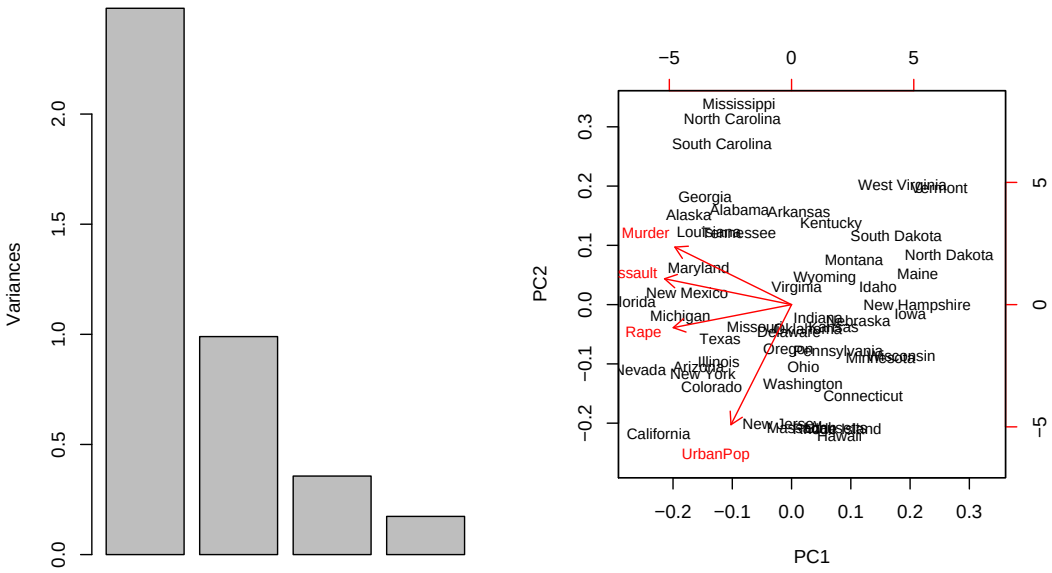
$$Y_i = X_i^T \beta + \epsilon_i. \tag{28}$$

*Instead of using least squares to estimate* $\beta$, *define* $\widehat{\beta}$ *to minimize*

$$\text{median}|Y_i - X_i^T \beta|. \tag{29}$$

*The resulting estimator* $\widehat{\beta}$ *is more resistant to outliers than the least squares estimator. But how can we find the standard error of* $\widehat{\beta}$*? Using the bootstrap approach, we resample the pairs of data to get the bootstrap sample* $(X_1^*, Y_1^*), \ldots, (X_n^* Y_n^*)$ *and then we get the corresponding bootstrap estimate* $\widehat{\beta}^*$. *We can repeat this many times and use the standard deviation of the bootstrap estimates to estimate the standard error of* $\widehat{\beta}$. *Figure 5 shows bootstrap replications of fits from regression and robust regression (minimizing* $L_1$ *error instead of squared error) in a dataset with an outlier.*

**Warning!** The bootstrap is not magic. Its validity requires some conditions to hold. When the conditions don't hold, the bootstrap, like any method, can give misleading answers.



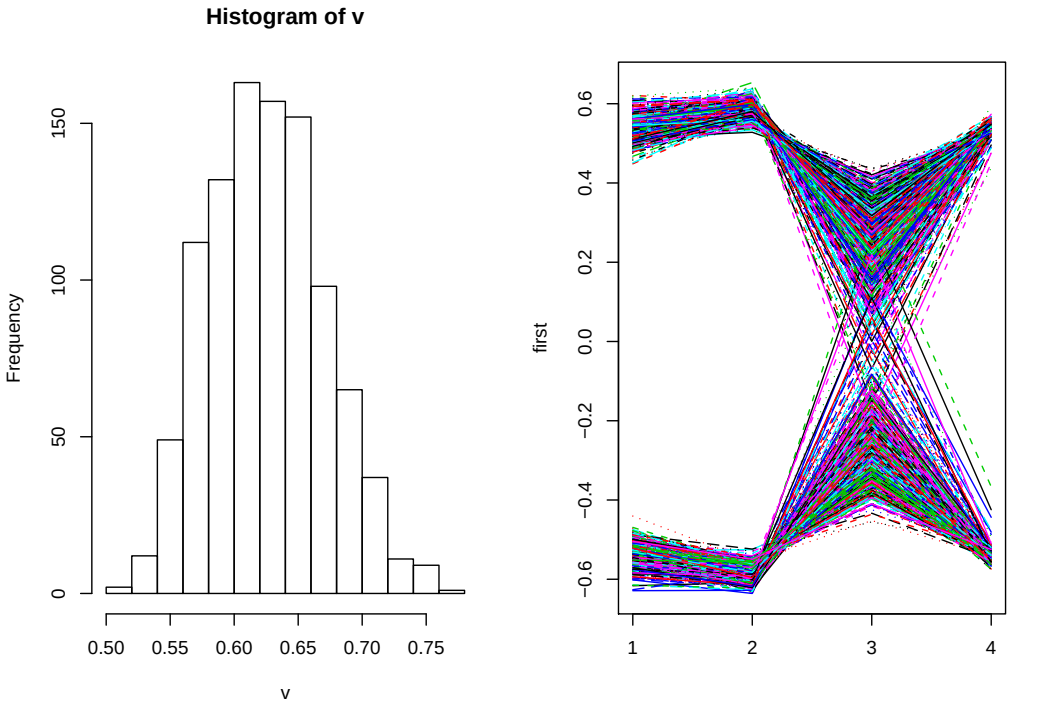




Figure 4: US Arrest Data

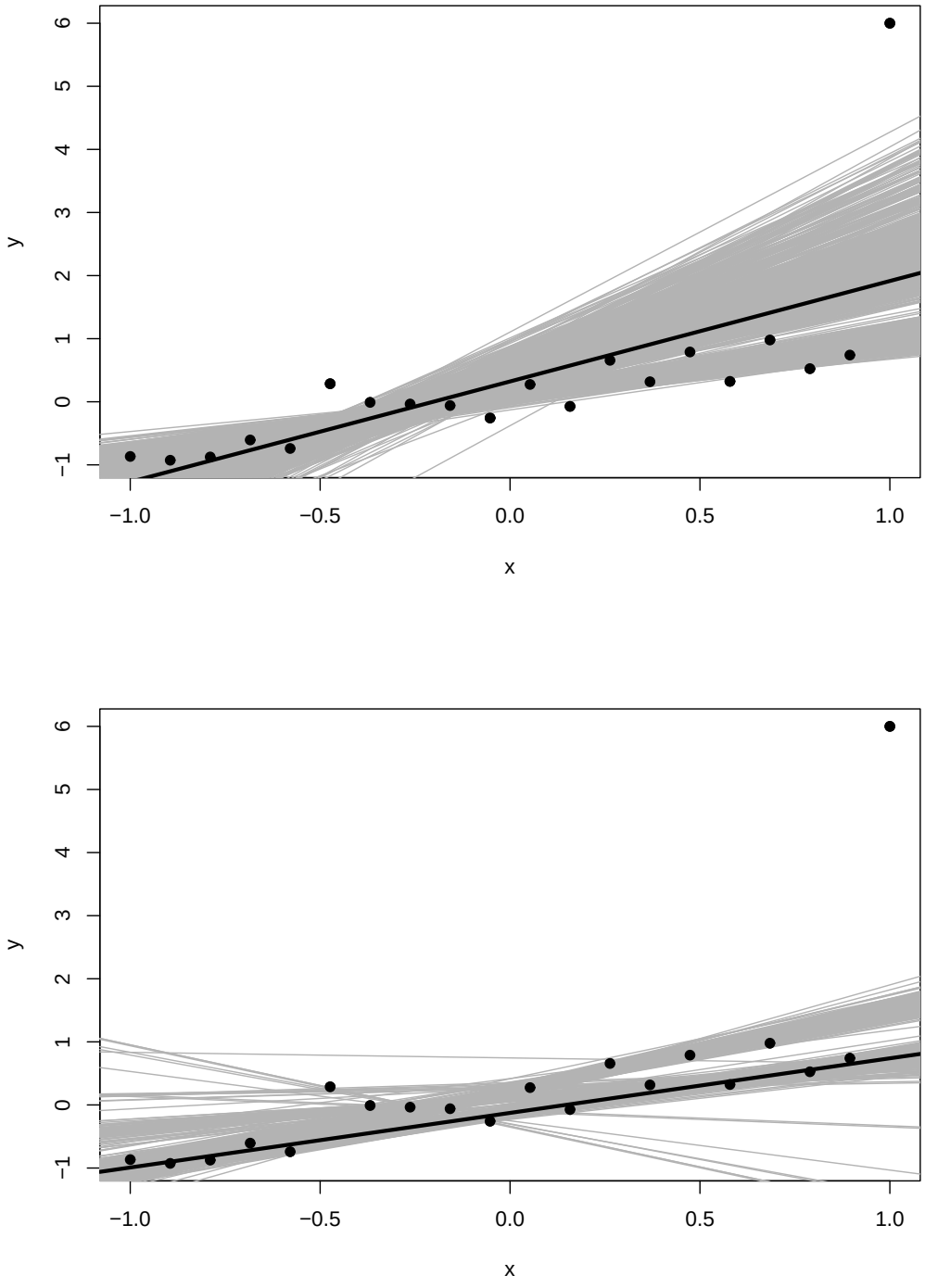



Figure 5: Robust Regression

# Lecture Notes 14
# Bayesian Inference

Relevant material is scattered throughout the book: see sections 7.2.3, 8.2.2, 9.2.4 and 9.3.3. We will also cover some material that is not in the book.

## 1 Introduction

So far we have been using *frequentist (or classical) methods.* In the frequentist approach, probability is interpreted as long run frequencies. The goal of frequentist inference is to create procedures with long run guarantees. Indeed, a better name for frequentist inference might be *procedural inference*. Moreover, the guarantees should be uniform over $\theta$ if possible. For example, a confidence interval traps the true value of $\theta$ with probability $1-\alpha$, no matter what the true value of $\theta$ is. **In frequentist inference, procedures are random while parameters are fixed, unknown quantities.**

In the *Bayesian approach*, probability is regarded as a measure of subjective degree of belief. In this framework, everything, including parameters, is regarded as random. There are no long run frequency guarantees. Bayesian inference is quite controversial.

Note that when we used Bayes estimators in minimax theory, we were not doing Bayesian inference. We were simply using Bayesian estimators as a method to derive minimax estimators.

## 2 The Mechanics of Bayes

Let $X_1, \ldots, X_n \sim p(x|\theta)$. In Bayes we also include a prior $\pi(\theta)$. It follows from Bayes' theorem that the posterior distribution of $\theta$ given the data is

$$\pi(\theta|X_1, \ldots, X_n) = \frac{p(X_1, \ldots, X_n|\theta)\pi(\theta)}{m(X_1, \ldots, X_n)}$$

where

$$m(X_1, \ldots, X_n) = \int p(X_1, \ldots, X_n|\theta)\pi(\theta)d\theta.$$

Hence,

$$\pi(\theta|X_1,\ldots,X_n) \propto L(\theta)\pi(\theta)$$

where $L(\theta) = p(X_1,\ldots,X_n|\theta)$ is the likelihood function. The interpretation is that $\pi(\theta|X_1,\ldots,X_n)$ represents your subjective beliefs about $\theta$ after observing $X_1,\ldots,X_n$.

A commonly used point estimator is the posterior mean

$$\overline{\theta} = \mathbb{E}(\theta|X_1,\ldots,X_n) = \int \theta\pi(\theta|X_1,\ldots,X_n)d\theta = \frac{\int \theta L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)}.$$

For interval estimation we use $C = (a,b)$ where $a$ and $b$ are chosen so that

$$\int_a^b \pi(\theta|X_1,\ldots,X_n) = 1-\alpha.$$

This interpretation is that

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

This does **not** mean that $C$ traps $\theta$ with probability $1-\alpha$. We will discuss the distinction in detail later.

**Example 1** *Let $X_1,\ldots,X_n \sim$ Bernoulli$(p)$. Let the prior be $p \sim$ Beta$(\alpha,\beta)$. Hence*

$$\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

*and*

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt.$$

*Set $Y = \sum_i X_i$. Then*

$$\pi(p|X) \propto \underbrace{p^Y 1-p^{n-Y}}_{\text{likelihood}} \times \underbrace{p^{\alpha-1}1-p^{\beta-1}}_{\text{prior}} \propto p^{Y+\alpha-1}1-p^{n-Y+\beta-1}.$$

*Therefore, $p|X \sim$ Beta$(Y+\alpha, n-Y+\beta)$. (See page 325 for more details.) The Bayes estimator is*

$$\widetilde{p} = \frac{Y+\alpha}{(Y+\alpha)+(n-Y+\beta)} = \frac{Y+\alpha}{\alpha+\beta+n} = (1-\lambda)\widehat{p}_{mle} + \lambda\ \overline{p}$$

*where*

$$\overline{p} = \frac{\alpha}{\alpha+\beta}, \quad \lambda = \frac{\alpha+\beta}{\alpha+\beta+n}.$$

*This is an example of a* conjugate prior.

**Example 2** *Let* $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ *with* $\sigma^2$ *known. Let* $\mu \sim N(m, \tau^2)$. *Then*

$$\mathbb{E}(\mu|X) = \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n}}\overline{X} + \frac{\frac{\sigma^2}{n}}{\tau^2 + \frac{\sigma^2}{n}}m$$

*and*

$$\text{Var}(\mu|X) = \frac{\sigma^2\tau^2/n}{\tau^2 + \frac{\sigma^2}{n}}.$$

# 3 Where Does the Prior Come From?

This is the million dollar question. In principle, the Bayesian is supposed to choose a prior $\pi$ that represents their prior information. This will be challenging in high dimensional cases to say the least. Also, critics will say that someone's prior opinions should not be included in a data analysis because this is not scientific.

There has been some effort to define "noninformative priors" but this has not worked out so well. An example is the *Jeffreys prior* which is defined to be

$$\pi(\theta) \propto \sqrt{I(\theta)}.$$

You can use a flat prior but be aware that this prior doesn't retain its flatness under transformations. In high dimensional cases, the prior ends up being highly influential. The result is that Bayesian methds tend to have poor frequentist behavior. We'll return to this point soon.

It is common to use flat priors even if they don't integrate to 1. This is posible since the posterior might still integrate to 1 even if the prior doesn't.

# 4 Large Sample Theory

There is a Bayesian central limit theorem. In nice models, with large $n$,

$$\pi(\theta|X_1, \ldots, X_n) \approx N\left(\widehat{\theta}, \frac{1}{I_n(\widehat{\theta})}\right) \tag{1}$$

where $\widehat{\theta}_n$ is the mle and $I$ is the Fisher information. In these cases, the $1-\alpha$ Bayesian intervals will be approximately the same as the frequentist confidence intervals. That is, an approximate $1-\alpha$ posterior interval is

$$C = \widehat{\theta} \pm \frac{z_{\alpha/2}}{\sqrt{I_n(\widehat{\theta})}}$$

which is the Wald confidence interval. However, this is only true if $n$ is large and the dimension of the model is fixed.

Here is a rough derivation of (1). Note that

$$\log \pi(\theta|X_1,\ldots,X_n) = \sum_{i=1}^{n} \log p(X_i|\theta) + \log \pi(\theta) - \log C$$

where $C$ is the normalizing constant. Now the sum has $n$ terms which grows with sample size. The last two terms are O(1). So the sum dominates, that is,

$$\log \pi(\theta|X_1,\ldots,X_n) \approx \sum_{i=1}^{n} \log p(X_i|\theta) = \ell(\theta).$$

Next, we note that

$$\ell(\theta) \approx \ell(\widehat{\theta}) + (\theta-\widehat{\theta})\ell'(\widehat{\theta}) + \frac{(\theta-\widehat{\theta})^2\ell''(\widehat{\theta})}{2}.$$

Now $\ell'(\widehat{\theta}) = 0$ so

$$\ell(\theta) \approx \ell(\widehat{\theta}) + \frac{(\theta-\widehat{\theta})^2\ell''(\widehat{\theta})}{2}.$$

Thus, approximately,

$$\pi(\theta|X_1,\ldots,X_n) \propto \exp\left(-\frac{(\theta-\widehat{\theta})^2}{2\sigma^2}\right)$$

where

$$\sigma^2 = -\frac{1}{\ell''(\widehat{\theta})}.$$

Let $\ell_i = \log p(X_i|\widehat{\theta}_0)$ where $\theta_0$ is the true value. Since $\widehat{\theta} \approx \theta_0$,

$$\ell''(\widehat{\theta}) \approx \ell''(\theta_0) = \sum_i \ell_i'' = n\frac{1}{n}\sum_i \ell_i'' \approx -nI_1(\theta_0) \approx -nI_1(\widehat{\theta}) = -I_n(\widehat{\theta})$$

and therefore, $\sigma^2 \approx 1/I_n(\widehat{\theta})$.

# 5 Bayes Versus Frequentist

In general, Bayesian and frequentist inferences can be quite different. If $C$ is a $1-\alpha$ Bayesian interval then

$$P(\theta \in C|X) = 1 - \alpha.$$

This does **not imply** that

$$\text{frequentist coverage} = \inf_{\theta} P_\theta(\theta \in C) = 1 - \alpha..$$

Typically, a $1 - \alpha$ Bayesian interval has coverage lower than $1 - \alpha$. Suppose you wake up everyday and produce a Bayesian 95 percent interval for some parameter. (A different parameter everyday.) The fraction of times your interval contains the true parameter will not be 95 percent. Here are some examples to make this clear.

**Example 3 Normal means.** *Let $X_i \sim N(\mu_i, 1)$, $i = 1, \ldots, n$. Suppose we use the flat prior $\pi(\mu_1, \ldots, \mu_n) \propto 1$. Then, with $\mu = (\mu_1, \ldots, \mu_n)$, the posterior for $\mu$ is multivariate Normal with mean $X = (X_1, \ldots, X_n)$ and covariance matrix equal to the identity matrix. Let $\theta = \sum_{i=1}^n \mu_i^2$. Let $C_n = [c_n, \infty)$ where $c_n$ is chosen so that $\mathbb{P}(\theta \in C_n|X_1, \ldots, X_n) = .95$. How often, in the frequentist sense, does $C_n$ trap $\theta$? Stein (1959) showed that*

$$\mathbb{P}_\mu(\theta \in C_n) \to 0, \quad \text{as } n \to \infty.$$

*Thus, $\mathbb{P}_\mu(\theta \in C_n) \approx 0$ even though $\mathbb{P}(\theta \in C_n|X_1, \ldots, X_n) = .95$.*

**Example 4 Sampling to a Foregone Conclusion.** *Let $X_1, X_2, \ldots \sim N(\theta, 1)$. Suppose we continue sampling until $T > k$ where $T = \sqrt{n}|\overline{X}_n|$ and $k$ is a fixed number, say, $k = 20$. The sample size $N$ is now a random variable. It can be shown that $\mathbb{P}(N < \infty) = 1$. It can also be shown that the posterior $\pi(\theta|X_1, \ldots, X_N)$ is the same as if $N$ had been fixed in advance. That is, the randomness in $N$ does not affect the posterior. Now if the prior $\pi(\theta)$ is smooth then the posterior is approximately $\theta|X_1, \ldots, X_N \sim N(\overline{X}_n, 1/n)$. Hence, if $C_n = \overline{X}_n \pm 1.96/\sqrt{n}$ then $\mathbb{P}(\theta \in C_n|X_1, \ldots, X_N) \approx .95$. Notice that 0 is never in $C_n$ since, when we stop sampling, $T > 20$, and therefore*

$$\overline{X}_n - \frac{1.96}{\sqrt{n}} > \frac{20}{\sqrt{n}} - \frac{1.96}{\sqrt{n}} > 0. \tag{2}$$

*Hence, when $\theta = 0$, $\mathbb{P}_\theta(\theta \in C_n) = 0$. Thus, the coverage is*

$$\text{Coverage} = \inf_\theta \mathbb{P}_\theta(\theta \in C_n) = 0.$$

*This is called sampling to a foregone conclusion and is a real issue in sequential clinical trials.*

**Example 5** *Here is an example we discussed earlier. Let $\mathcal{C} = \{c_1, \ldots, c_N\}$ be a finite set of constants. For simplicity, asssume that $c_j \in \{0, 1\}$ (although this is not important). Let $\theta = N^{-1} \sum_{j=1}^N c_j$. Suppose we want to estimate $\theta$. We proceed as follows. Let $S_1, \ldots, S_n \sim$* Bernoulli($\pi$) *where $\pi$ is known. If $S_i = 1$ you get to see $c_i$. Otherwise, you do not. (This is an example of survey sampling.) The likelihood function is*

$$\prod_i \pi^{S_i}(1-\pi)^{1-S_i}.$$

*The unknown parameter does not appear in the likelihood. In fact, there are no unknown parameters in the likelihood! The likelihood function contains no information at all. The posterior is the same as the prior.*

*But we can estimate $\theta$. Let*

$$\widehat{\theta} = \frac{1}{N\pi} \sum_{j=1}^N c_j S_j.$$

*Then $\mathbb{E}(\widehat{\theta}) = \theta$. Hoeffding's inequality implies that*

$$\mathbb{P}(|\widehat{\theta} - \theta| > \epsilon) \leq 2e^{-2n\epsilon^2\pi^2}.$$

*Hence, $\widehat{\theta}$ is close to $\theta$ with high probability. In particular, a $1-\alpha$ confidence interval is $\widehat{\theta} \pm \sqrt{\log(2/\alpha)/(2n\pi^2)}$.*

# 6 Bayesian Computing

If $\theta = (\theta_1, \ldots, \theta_p)$ is a vector then the posterior $\pi(\theta|X_1, \ldots, X_n)$ is a multivariate distribution. If you are interested in one parameter, $\theta_1$ for example, then you need to find the marginal posterior:

$$\pi(\theta_1|X_1, \ldots, X_n) = \int \pi(\theta_1, \ldots, \theta_p|X_1, \ldots, X_n) d\theta_2 \ \cdots \ d\theta_p.$$

Usually, this integral is intractable. In practice, we resort to Monte Carlo methods. These are discussed in 36/10-702.

# 7 Bayesian Hypothesis Testing

Bayesian hypothesis testing can be done as follows. Suppose that $\theta \in \mathbb{R}$ and we want to test

$$H_0 : \theta = \theta_0 \quad \text{and} \quad H_1 : \theta \neq \theta_0.$$

If we really believe that there is a positive prior probability that $H_0$ is true then we can use a prior of the form

$$a\delta_{\theta_0} + (1-a)g(\theta)$$

where $0 < a < 1$ is the prior probability that $H_0$ is true and $g$ is a smooth prior density over $\theta$ which represents our prior beliefs about $\theta$ when $H_0$ is false. It follows from Bayes' theorem that

$$P(\theta = \theta_0 | X_1, \ldots, X_n) = \frac{ap(X_1, \ldots, X_n|\theta_0)}{ap(X_1, \ldots, X_n|\theta_0) + (1-a)\int p(X_1, \ldots, X_n|\theta)g(\theta)d\theta} = \frac{aL(\theta_0)}{aL(\theta_0) + (1-a)m}$$

where $m = \int L(\theta)g(\theta)d\theta$. It can be shown that $P(\theta = \theta_0|X_1, \ldots, X_n)$ is very sensitive to the choice of $g$.

Sometimes, people like to summarize the test by using the *Bayes factor* $B$ which is defined to be the posterior odds divided by the prior odds:

$$B = \frac{\text{posterior odds}}{\text{prior odds}}$$

where

$$\begin{aligned} \text{posterior odds} &= \frac{P(\theta = \theta_0|X_1, \ldots, X_n)}{1 - P(\theta = \theta_0|X_1, \ldots, X_n)} \\ &= \frac{\frac{aL(\theta_0)}{aL(\theta_0)+(1-a)m}}{\frac{(1-a)m}{aL(\theta_0)+(1-a)m}} \\ &= \frac{aL(\theta_0)}{(1-a)m} \end{aligned}$$

and

$$\text{prior odds} = \frac{P(\theta = \theta_0)}{P(\theta \neq \theta_0)} = \frac{a}{1-a}$$

and hence

$$B = \frac{L(\theta_0)}{m}.$$

**Example 6** *Let $X_1, ldots, X_n \sim N(\theta, 1)$. Let's test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. Suppose we take $g(\theta)$ to be $N(0, 1)$. Thus,*

$$g(\theta) = \frac{1}{\sqrt{2\pi}} e^{-\theta^2/2}.$$

*Let us further take $a = 1/2$. Then, after some tedious integration to compute $m(X_1, \ldots, X_n)$ we get*

$$\begin{aligned} P(\theta = \theta_0 | X_1, \ldots, X_n) &= \frac{L(0)}{L(0) + m} \\ &= \frac{e^{-n\overline{X}^2/2}}{e^{-n\overline{X}^2/2} + \sqrt{\frac{n}{n+1}} e^{-n\overline{X}^2/(2(n+1))}}. \end{aligned}$$

*On the other hand, the p-value for the usual test is $p = 2\Phi(-\sqrt{n}|\overline{X}|)$. Figure 1 shows the posterior of $H_0$ and the p-value as a function of $\overline{X}$ when $n = 100$. Note that they are very different. Unlike in estimation, in testing there is little agreement between Bayes and frequentist methods.*

# 8 Conclusion

Bayesian and frequentist inference are answering two different questions. Frequentist inference answers the question:

How do I construct a procedure that has frequency guarantees?

Bayesian inference answers the question:

How do I update my subjective beliefs after I observe some data?

In parametric models, if $n$ is large and the dimension of the model is fixed, Bayes and frequentist procedures will be similar. Otherwise, they can be quite different.

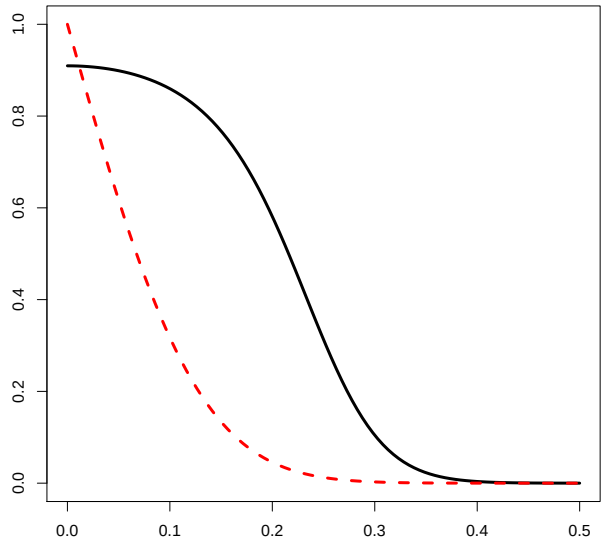



Figure 1: Solid line: $P(\theta = 0|X_1, \ldots, X_n)$ versus $\overline{X}$. Dashed line: p-value versus $\overline{X}$.

# Lecture Notes 15
# Prediction

**This is mostly not in the text. Some relevant material is in Chapters 11 and 12.**

## 1 Introduction

We observe *training data* $(X_1, Y_1), \ldots, (X_n, Y_n)$. Given a new pair $(X, Y)$ we want to predict $Y$ from $X$. There are two common versions:

1. $Y \in \{0, 1\}$. This is called *classification*, or discrimination, or pattern recognition. (More generally, $Y$ can be discrete.)

2. $Y \in \mathbb{R}$. This is called *regression*.

For classification we will use the following loss function. Let $h(x)$ be or prediction of $Y$ when $X = x$. Thus $h(x) \in \{0, 1\}$. The function $h$ is called a **classifier**. The classification loss is $I(Y \neq h(X))$ and the the **classification risk** is

$$R(h) = \mathbb{P}(Y \neq h(X)) = \mathbb{E}(I(Y \neq h(X))).$$

For regression, suppose our prediction of $Y$ when $X = x$ is $g(x)$. We will use the squared error prediction loss $(Y - g(X))^2$ and the risk is

$$R(g) = \mathbb{E}(Y - g(X))^2.$$

## 2 Regression

**Theorem 1** *$R(g)$ is minimized by*

$$m(x) = \mathbb{E}(Y|X = x) = \int y\, p(y|x) dy.$$

**Proof.** Let $g(x)$ be any function of $x$. Then

$$\begin{aligned}
R(g) &= \mathbb{E}(Y-g(X))^2 = \mathbb{E}(Y-m(X)+m(X)-g(X))^2 \\
&= \mathbb{E}(Y-m(X))^2 + \mathbb{E}(m(X)-g(X))^2 + 2\mathbb{E}((Y-m(X))(m(X)-g(X))) \\
&\geq \mathbb{E}(Y-m(X))^2 + 2\mathbb{E}((Y-m(X))(m(X)-g(X))) \\
&= \mathbb{E}(Y-m(X))^2 + 2\mathbb{E}\mathbb{E}\left( (Y-m(X))(m(X)-g(X)) \;\middle|\; X \right) \\
&= \mathbb{E}(Y-m(X))^2 + 2\mathbb{E}\left( (\mathbb{E}(Y|X)-m(X))(m(X)-g(X)) \right) \\
&= \mathbb{E}(Y-m(X))^2 + 2\mathbb{E}\left( (m(X)-m(X))(m(X)-g(X)) \right) \\
&= \mathbb{E}(Y-m(X))^2 = R(m).
\end{aligned}$$

■

Hence, to do make predictions, we need to estimate $m(x) = \mathbb{E}(Y|X=x)$. The simplest apprach is to use a parametric model. In particular, the *linear regression model* assumes that $m(x)$ is a linear function of $x$. (More precisely, we seek the best linear predictor.)

Suppose that $X_i \in \mathbb{R}^p$ so that

$$X_i = (X_{i1}, \ldots, X_{ip})^T.$$

Then the linear regression model is

$$m(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j.$$

We can write

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i, \quad i = 1, \ldots, n$$

where $\epsilon_1, \ldots, \epsilon_n$ are iid with mean 0.

If we use the convention that $X_{i1} = 1$ then we can write the model more simply as

$$Y_i = \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i = \beta^T X_i + \epsilon_i, \quad i = 1, \ldots, n \tag{1}$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$ and $X_i = (X_{i1}, \ldots, X_{ip})^T$.

Let us define $Y = (Y_1, \ldots, Y_n)^T$, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ and let $X$ be the $n \times p$ matrix with $X(i,j) = X_{ij}$. Then we can write (1) as

$$Y = X\beta + \epsilon.$$

The least squares estimator $\widehat{\beta}$ is the $\beta$ that minimizes

$$\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 = ||Y - X\beta||^2.$$

**Theorem 2** *Suppose that $X^TX$ is invertible. Then the least squares estimator is*

$$\widehat{\beta} = (X^TX)^{-1}X^TY.$$

The *fitted values* or *predicted values* are $\widehat{Y} = (\widehat{Y}_1, \ldots, \widehat{Y}_n)^T$ where

$$\widehat{Y}_i = X_i^T\widehat{\beta}.$$

Hence,

$$\widehat{Y} = X\widehat{\beta} = HY$$

where

$$H = X(X^TX)^{-1}X^T$$

is called the *hat matrix*.

**Theorem 3** *The matrix $H$ is symmetric and idempotent: $H^2 = H$. Moreover, $HY$ is the projection of $Y$ onto the column space of $X$.*

This is discussed in more detail in 36-707 and 10/36-702.

**Theorem 4** *Suppose that the linear model is correct.*[1] *Also, suppose that* $\mathrm{Var}(\epsilon_i) = \sigma^2$. *Then,*

$$\sqrt{n}(\widehat{\beta} - \beta) \rightsquigarrow N(0, \sigma^2 X^TX).$$

[1]This model is virtually never correct, so view this result with caution.

Under the (questionable) assumption that the linear model is correct, we can also say the following. A consistent estimator of $\sigma^2$ is

$$\widehat{\sigma}^2 = \frac{\text{RSS}}{n-p}$$

and

$$\frac{\sqrt{n}(\widehat{\beta}_j - \beta_j)}{s_j} \rightsquigarrow N(0,1)$$

where the standard error $s_j$ is the $j^{\text{th}}$ diagonal element of $\widehat{\sigma}^2 X^T X$. To test $H_0 : \widehat{\beta}_j = 0$ versus $H_1 : \widehat{\beta}_j \neq 0$ we reject if $|\widehat{\beta}_j|/s)j > z_{\alpha/2}$. An approximate $1-\alpha$ confidence interval for $\beta_j$ is

$$\widehat{\beta}_j \pm z_{\alpha/2} s_j.$$

**Theorem 5** *Suppose that the linear model is correct and that $\epsilon_1, \ldots, \epsilon_n \sim N(0, \sigma^2)$. Then the least squares estimator is the maximum likelihood estimator.*

# 3 Linear Prediction When the Model is Wrong

When the model is wrong (and it always is) the least squares estimator still has the following good property. Let $\beta_*$ minimize

$$R(\beta) = \mathbb{E}(Y - X^T\beta)^2.$$

We call $\ell_*(x) = x^T\beta_*$ the best linear predictor.

**Theorem 6** *Under weak conditions,*

$$R(\widehat{\beta}) - R(\beta_*) \xrightarrow{P} 0.$$

Hence, the least squares estimator approximates the best linear predictor. Let's prove this in the case with one covariate. Then

$$R(\beta) = \mathbb{E}(Y - X\beta)^2 = \mathbb{E}(Y)^2 - 2\beta\mathbb{E}(XY) + \beta^2\mathbb{E}(X^2).$$

Minimizing with respect to $\beta$ we get

$$\beta_* = \frac{\mathbb{E}(XY)}{\mathbb{E}(X^2)}$$

assuming that $0 < \mathbb{E}(X^2) < \infty$ and $\mathbb{E}(XY) < \infty$. Now

$$\widehat{\beta} = \frac{\sum_i X_i Y_i}{\sum_i X_i^2} = \frac{\frac{1}{n}\sum_i X_i Y_i}{\frac{1}{n}\sum_i X_i^2}.$$

By the law of large numbers and the continuous mapping theorem:

$$\widehat{\beta} \xrightarrow{P} \beta_*.$$

Since $R(\beta)$ is a continuous function of $\beta$, it follows from the continuous mapping theorem that

$$R(\widehat{\beta}) \xrightarrow{P} R(\beta_*).$$

In fact,

$$\widehat{\beta} = \frac{\frac{1}{n}\sum_i X_i Y_i}{\frac{1}{n}\sum_i X_i^2} = \frac{\mathbb{E}(XY) + O_P(1/\sqrt{n})}{\mathbb{E}(X^2) + O_P(1/\sqrt{n})} = \beta_* + O_P(1/\sqrt{n})$$

and

$$R(\widehat{\beta}) = R(\beta_*) + (\widehat{\beta} - \beta_*)R'(\beta_*) + o(\widehat{\beta} - \beta_*)$$

and so

$$R(\widehat{\beta}) - R(\beta_*) + O_P(1/\sqrt{n}).$$

The message here is that least squares estimates the best linear predictor: we don't have to assume that the truth is linear.

# 4 Nonparametric Regression

Suppose we want to estimate $m(x)$ where we only assume that $m$ is a smooth function. The kernel regression estimator is

$$\widehat{m}(x) = \sum_i Y_i \, w_i(x)$$

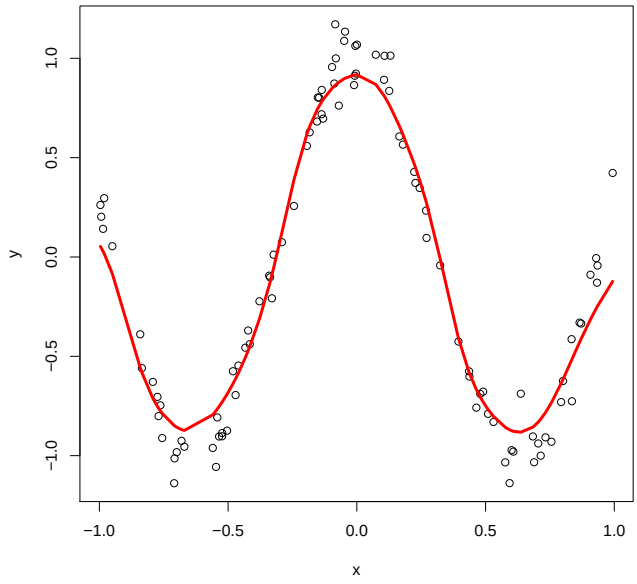



Figure 1: A kernel regression estimator.

where

$$w_i(x) = \frac{K\left(\frac{||x-X_i||}{h}\right)}{\sum_j K\left(\frac{||x-X_j||}{h}\right)}.$$

Here $K$ is a kernel and $h$ is a bandwidth. The properties are simialr to that of kernel density estimation. The properties of $\widehat{m}$ are discussed in more detail in the 36-707 and in 10-702. An example is shown in Figure 1.

# 5 Classification

The best classifier is the so-called *Bayes classifier* defined by:

$$h_B(x) = I(m(x) \geq 1/2)$$

where $m(x) = \mathbb{E}(Y|X = x)$.

**Theorem 7** *For any $h$, $R(h) \geq R(h_B)$.*

**Proof.** For any $h$,

$$\begin{aligned} R(h) - R(h_B) &= \mathbb{P}(Y \neq h(X)) - \mathbb{P}(Y \neq h_B(X)) \\ &= \int \mathbb{P}(Y \neq h(x)|X = x)p(x)dx - \int \mathbb{P}(Y \neq h_B(x)|X = x)p(x)dx \\ &= \int \left(\mathbb{P}(Y \neq h(x)|X = x) - \mathbb{P}(Y \neq h_B(x)|X = x)\right) p(x)dx. \end{aligned}$$

We will show that

$$\mathbb{P}(Y \neq h(x)|X = x) - \mathbb{P}(Y \neq h_B(x)|X = x) \geq 0$$

for all $x$. Now

$$\begin{aligned} \mathbb{P}(Y \neq h(x)|X = x) &- \mathbb{P}(Y \neq h_B(x)|X = x) \\ &= \left(h(x)\mathbb{P}(Y \neq 1|X = x) + (1 - h(x))\mathbb{P}(Y \neq 0|X = x)\right) \\ &\quad - \left(h_B(x)\mathbb{P}(Y \neq 1|X = x) + (1 - h_B(x))\mathbb{P}(Y \neq 0|X = x)\right) \\ &= (h(x)(1 - m(x)) + (1 - h(x))m(x)) \\ &\quad - (h_B(x)(1 - m(x)) + (1 - h_B(x))m(x)) \\ &= 2(m(x) - 1/2)(h_B(x) - h(x)) \geq 0 \end{aligned}$$

since $h_B(x) = 1$ if and only if $m(x) \geq 1/2$. □ ■

The most direct approach to classification is *empirical risk minimization* (ERM). We start with a set of classifiers $\mathcal{H}$. Each $h \in \mathcal{H}$ is a function $h : x \to \{0, 1\}$. The *training error* or *empirical risk* is

$$\widehat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} I(Y_i \neq h(X_i)).$$

We choose $\widehat{h}$ to minimize $\widehat{R}$:

$$\widehat{h} = \text{argmin}_{h \in \mathcal{H}} \widehat{R}(h).$$

For example, a linear classifier has the form $h_\beta(x) = I(\beta^T x \geq 0)$. The set of linear classifiers is $\mathcal{H} = \{h_\beta : \ \beta \in \mathbb{R}^p\}$.

**Theorem 8** *Suppose that $\mathcal{H}$ has VC dimension $d < \infty$. Let $\widehat{h}$ be the empirical risk minimizer and let*

$$h_* = \text{argmin}_{h \in \mathcal{H}} R(h)$$

*be the best classifier in $\mathcal{H}$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}(R(\widehat{h}) > R(h_*) + 2\epsilon) \leq c_2 n^d e^{-n c_2 \epsilon^2}$$

*for some constnts $c_1$ and $c_2$.*

**Proof.** Recall that

$$\mathbb{P}(\sup_{h \in \mathcal{H}} |\widehat{R}(h) - R(h)| > \epsilon) \leq c_2 n^d e^{-n c_2 \epsilon^2}.$$

But when $\sup_{h \in \mathcal{H}} |\widehat{R}(h) - R(h)| \leq \epsilon$ we have

$$R(\widehat{h}) \ \leq \ \widehat{R}(\widehat{h}) + \epsilon \leq \widehat{R}(h_*) + \epsilon \leq R(h_*) + 2\epsilon. \quad \square$$

■

Empirical risk minimization is difficult because $\widehat{R}(h)$ is not a smooth function. Thus, we often use other approaches. One idea is to use a *surrogate loss function.* To expain this idea, it will be convenient to relabel the $Y_i$'s as being +1 or -1. Many classifiers then take the form

$$h(x) = \text{sign}(f(x))$$

for some $f(x)$. For example, linear classifiers have $f(x) = x^T\beta$. Th classification loss is then

$$L(Y, f, X) = I(Yf(X) < 0)$$

since an error occurs if and only if $Y$ and $f(X)$ have different signs. An example of surrogate loss is the hinge function

$$(1 - Yf(X))_+.$$

Instead of minimizing classification loss, we minimize

$$\sum_i (1 - Y_i f(X_i))_+.$$

The resulting classifier is called a *support vector machine*.

Another approach to classification is *plug-in clasification*. We replace the Bayes rule $h_B = I(m(x) \geq 1/2)$ with

$$\widehat{h}(x) = I(\widehat{m}(x) \geq 1/2)$$

where $\widehat{m}$ is an estimate of the regression function. The estimate $\widehat{m}$ can be parametric or nonparametric.

A common parametric estimator is *logistic regression*. Here, we assume that

$$m(x; \beta) = \frac{e^{x^T\beta}}{1 + e^{x^T\beta}}.$$

Since $Y_i$ is Bernoulli, the likeihood is

$$L(\beta) = \prod_{i=1}^n m(X_i; \beta)^{Y_i}(1 - m(X_i; \beta))^{1-Y_i}.$$

We compute the mle $\widehat{\beta}$ numerically. See Section 12.3 of the text.

What is the relationship between classification and regression? Generally speaking, classification is easier. This follows from the next result.

**Theorem 9** *Let $m(x) = \mathbb{E}(Y|X = x)$ and let $h_m(x) = I(m(x) \geq 1/2)$ be the Bayes rule. Let $g$ be any function and let $h_g(x) = I(g(x) \geq 1/2)$. Then*

$$R(h_g) - R(h_m) \leq 2\sqrt{\int |g(x) - m(x)|^2 dP(x)}.$$

**Proof.** We showed earlier that

$$R(h_g) - R(h_m) = \int [\mathbb{P}(Y \neq h_g(x)|X = x) - \mathbb{P}(Y \neq h_m(x)|X = x)]\, dP(x)$$

and that

$$\mathbb{P}(Y \neq h_g(x)|X = x) - \mathbb{P}(Y \neq h_m(x)|X = x) = 2(m(x) - 1/2)(h_m(x) - h_g(x)).$$

Now

$$2(m(x) - 1/2)(h_m(x) - h_g(x)) = 2|m(x) - 1/2|\ I(h_m(x) \neq h_g(x)) \leq 2|m(x) - g(x)|$$

since $h_m(x) \neq h_g(x)$ implies that $|m(x) - 1/2| \leq |m(x) - g(x)|$. Hence,

$$\begin{aligned} R(h_g) - R(h_m) &= 2\int |m(x) - 1/2| I(h_m(x) \neq h_g(x)) dP(x) \\ &\leq 2\int |m(x) - g(x)| dP(x) \\ &\leq 2\sqrt{\int |g(x) - m(x)|^2 dP(x)} \end{aligned}$$

where the last setp follows from the Cauchy-Schwartz inequality. □ ■

Hence, if we have an estimator $\widehat{m}$ such that $\int |\widehat{m}(x) - m(x)|^2 dP(x)$ is small, then the excess classification risk is also small. But the reverse is not true.

# Lecture Notes 16

# Model Selection

**Not in the text.**

## 1 Introduction

Sometimes we have a set of possible models and we want to choose the best model. Model selection methods help us choose a good model. Here are some examples.

**Example 1** *Suppose you use a polynomial to model the regression function:*

$$m(x) = \mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p.$$

*You will need to choose the order of polynomial p. We can think of this as a sequence of models* $\mathcal{M}_1, \ldots, \mathcal{M}_p, \ldots$ *indexed by p.*

**Example 2** *Suppose you have data* $Y_1, \ldots, Y_n$ *on age at death for n people. You want to model the distribution of Y. Some popular models are:*

1. $\mathcal{M}_1$*: the exponential distribution:* $f(y;\theta) = \theta e^{-\theta y}$.
2. $\mathcal{M}_2$*: the gamma distribution:* $f(y;a,b) = (b^a/\Gamma(a))y^{a-1}e^{-by}$.
3. $\mathcal{M}_3$*: the log-normal distribution: we take* $\log Y \sim N(\mu, \sigma^2)$.

**Example 3** *Suppose you have time series data* $Y_1, Y_2, \ldots$. *A common model is the AR (autoregressive model):*

$$Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \cdots + a_k Y_{t-k} + \epsilon_t$$

*where* $\epsilon_t \sim N(0, \sigma^2)$. *The number k is called the order of the model. We need to choose k.*

**Example 4** *In a linear regression model, you need to choose which variables to include in the regression. This is called variable selection. This problem is discussed at length in 36-707 and 10-702.*

The most common model selections methods are:

1. AIC (and related methods like $C_p$).
2. Cross-validation.
3. BIC (and related methods like MDL, Bayesian model selection).

We need to distinguish between 2 goals:

1. Find the model that gives the best prediction (without assuming that any of the models are correct).
2. Assume one of the models is the true model and find the "true" model.

Generally speaking, AIC and cross-validation are used for goal 1 while BIC is used for goal 2.

# 2 AIC

Suppose we have models $\mathcal{M}_1, \ldots, \mathcal{M}_k$ where each model is a set of densities:

$$\mathcal{M}_j = \left\{ p(y; \theta_j) : \ \theta_j \in \Theta_j \right\}.$$

We have data $Y_1, \ldots, Y_n$ drawn from some density $f$. **We do not assume that $f$ is in any of the models.**

Let $\widehat{\theta}_j$ be the mle from model $j$. An estimate of $f$, based on model $j$ is $\widehat{f}_j(y) = p(y;\widehat{\theta}_j)$. The quality of $\widehat{f}_j(y)$ as an estimate of $f$ can be measured by the Kullback-Leibler distance:

$$\begin{aligned} K(f,\widehat{f}_j) &= \int p(y)\log\left(\frac{p(y)}{\widehat{f}_j(y)}\right)dy \\ &= \int p(y)\log p(y)dy - \int p(y)\log\widehat{f}_j(y)dy. \end{aligned}$$

The first term does not depend on $j$. So minimizing $K(f,\widehat{f}_j)$ over $j$ is the same as maximizing

$$K_j = \int p(y)\log p(y;\widehat{\theta}_j)dy.$$

We need to estimate $K_j$. Intuitively, you might think that a good estimate of $K_j$ is

$$\overline{K}_j = \frac{1}{n}\sum_{i=1}^{n}\log p(Y_i;\widehat{\theta}_j) = \frac{\ell_j(\widehat{\theta}_j)}{n}$$

where $\ell_j(\theta_j)$ is the log-likelihood funcion for model $j$. However, this estimate is very biased because the data are being used twice: first to get the mle and second to estimate the integral. Akaike showed that the bias is approximately $d_j/n$ where $d_j = \text{dimension}(\Theta_j)$. Therefore we use

$$\widehat{K}_j = \frac{\ell_j(\widehat{\theta}_j)}{n} - \frac{d_j}{n} = \overline{K}_j - \frac{d_j}{n}.$$

Now, define

$$\text{AIC}(j) = 2n\widehat{K}_j = \ell_j(\widehat{\theta}_j) - 2d_j.$$

Notice that maximizing $\widehat{K}_j$ is the same as maximizing $\text{AIC}(j)$ over $j$. Why do we multiply by $2n$? Just for historical reasons. We can multiply by any constant; it won't change which model we pick. In fact, different texts use different versions of AIC.

AIC stands for "Akaike Informaion Criterion." Akaike was a famous Japanese statistician who died recently (August 2009).

# 3 Theoretical Derivation of AIC

Let us now look closer to see where the formulas come from. Recall that

$$K_j = \int p(y)\log p(y;\widehat{\theta}_j)dy.$$

For simplicity, let us focus on one model and drop the subscript $j$. We want to estimate

$$K = \int p(y) \log p(y; \widehat{\theta}) dy.$$

Our goal is to show that

$$\overline{K} - \frac{d}{n} \approx K$$

where

$$\overline{K} = \frac{1}{n} \sum_{i=1}^{n} \log p(Y_i; \widehat{\theta})$$

and $d$ is the dimension of $\theta$.

**Some Notation and Background.** Let $\theta_0$ minimize $K(f, p(\cdot; \theta))$. So $p(y; \theta_0)$ is the closest density in the model to the true density. Let $\ell(y, \theta) = \log p(y; \theta)$ and

$$s(y, \theta) = \frac{\partial \log p(y; \theta)}{\partial \theta}$$

be the score and let $H(y, \theta)$ be the matrix of second derivatives.

Let $Z_n = \sqrt{n}(\widehat{\theta} - \theta_0)$ and recall that

$$Z_n \rightsquigarrow N(0, J^{-1} V J^{-1})$$

where $J = -\mathbb{E}[H(Y, \theta_0)]$ and

$$V = \text{Var}(s(Y, \theta_0)).$$

In class we proved that $V = J^{-1}$. But that proof assumed the model was correct. We are not assuming that. Let

$$S_n = \frac{1}{n} \sum_{i=1}^{n} s(Y_i, \theta_0).$$

By the CLT,

$$\sqrt{n} S_n \rightsquigarrow N(0, V)$$

Hence, in distribution

$$J Z_n \approx \sqrt{n} S_n. \tag{1}$$

Here we used the fact that $\mathrm{Var}(AX) = A(\mathrm{VarX})A^T$. Thus

$$\mathrm{Var}(JZ_n) = J(J^{-1}VJ^{-1})J^T = V.$$

We will need one other fact. Let $\epsilon$ be a random vector with mean $\mu$ and covariance $\Sigma$. Let

$$Q = \epsilon^T A \epsilon.$$

($Q$ is a called a quadratic form.) Then

$$\mathbb{E}(Q) = \mathrm{trace}(A\Sigma) + \mu^T A \mu.$$

**The details.** By using a Taylor series

$$\begin{aligned} K &\approx \int p(y)\left(\log p(y;\widehat{\theta}_0) + (\theta-\theta_0)^T s(y,\theta_0) + \frac{1}{2}(\widehat{\theta}-\theta_0)^T H(y,\theta_0)(\widehat{\theta}-\theta_0)\right)dy \\ &= K_0 - \frac{1}{2n} Z_n^T J Z_n \end{aligned}$$

where

$$K_0 = \int p(y)\log p(y;\theta_0)dy,$$

The second term dropped out because, like the score function, it has mean 0. Again we do a Taylor series to get

$$\begin{aligned} \overline{K} &\approx \frac{1}{n}\sum_{i=1}^n \left(\ell(Y_i,\theta_0) + (\widehat{\theta}-\theta_0)^T s(Y_i,\theta_0) + \frac{1}{2}(\widehat{\theta}-\theta_0)^T H(Y_i,\theta_0)(\widehat{\theta}-\theta_0)\right) \\ &= K_0 + A_n + (\widehat{\theta}-\theta_0)^T S_n - \frac{1}{2n} Z_n^T J_n Z_n \\ &\approx K_0 + A_n + \frac{Z_n^T S_n}{\sqrt{n}} - \frac{1}{2n} Z_n^T J Z_n \end{aligned}$$

where

$$J_n = -\frac{1}{n}\sum_{i=1}^n H(Y_i,\theta_0) - \xrightarrow{P} J,$$

and

$$A_n = \frac{1}{n}\sum_{i=1}^{n}(\ell(Y_i, \theta_0) - K_0).$$

Hence,

$$\overline{K} - K \approx A_n + \frac{\sqrt{n} Z_n^T S_n}{n} \approx A_n + \frac{Z_n^T J Z_n}{n}$$

where we used (1). We conclude that

$$\mathbb{E}(\overline{K} - K) \approx \mathbb{E}(A_n) + \mathbb{E}\left(\frac{Z_n^T J Z_n}{n}\right) = 0 + \frac{\text{trace}(J\ J^{-1} V J^{-1})}{n} = \frac{\text{trace}(J^{-1}V)}{n}.$$

Hence,

$$K \approx \overline{K} - \frac{\text{trace}(J^{-1}V)}{n}.$$

If the model is correct, then $J^{-1} = V$ so that $\text{trace}(J^{-1}V) = \text{trace}(I) = p$. Thus we would use

$$K \approx \overline{K} - \frac{p}{n}.$$

You can see that there are a lot of approximations and assumptions being used. So AIC is a very crude tool. Cross-validation is much more reliable.

# 4 Cross-Validation

There are various flavors of cross-validation. In general, the data are split into a training set and a test set. The models are fit on the training set and are used to predict the test set. Usually, many such splits are used and the result are averaged over splits. However, to keep things simple, we will use a single split.

Suppose again that we have models $\mathcal{M}_1, \ldots, \mathcal{M}_k$. Assume there are $2n$ data points. Split the data randomly into two halves that we will denote $D = (Y_1, \ldots, Y_n)$ and $T = (Y_1^*, \ldots, Y_n^*)$. Use $D$ to find the mle's $\widehat{\theta}_j$. Then define

$$\widehat{K}_j = \frac{1}{n}\sum_{i=1}^{n} \log p(Y_i^*; \widehat{\theta}_j).$$

Note that $\mathbb{E}(\widehat{K}_j) = K_j$; there is no bias because $\widehat{\theta}_j$ is independent of $Y_j^*$. We will assume that $|\log p(y;\theta)| \leq B < \infty$. By Hoeffding's inequality,

$$\mathbb{P}(\max_j |\widehat{K}_j - K_j| > \epsilon) \leq 2ke^{-2n\epsilon^2/(2B^2)}.$$

Let

$$\epsilon_n = \sqrt{\frac{2B^2 \log(2k/\alpha)}{n}}.$$

Then

$$\mathbb{P}(\max_j |\widehat{K}_j - K_j| > \epsilon_n) \leq \alpha.$$

If we choose $\widehat{j} = \text{argmax}_j \widehat{K}_j$, then, with probability at least $1 - \alpha$,

$$K_{\widehat{j}} \geq \max_j K_j - 2\sqrt{\frac{2B^2 \log(2k/\alpha)}{n}} = \max_j K_j - O\left(\frac{\log k}{n}\right).$$

So with high probability, you choose close to the best model. This argument can be improved and also applies to regression, classification etc. Of course, with regression, the loss function is $\mathbb{E}(Y - m(X))^2$ and the cross-validation score is then

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i^* - m(X_i^*))^2.$$

For classification we use

$$\frac{1}{n}\sum_{i=1}^{n} I(Y_i^* \neq h(X_i^*)).$$

We have made essentially no assumptions or approximations. (The bounded on $\log f$ can be relaxed.) The beauty of cross-validation is its simpicity and generality. It can be shown that AIC and cross-validation have very similar behavior. But, cross-validation works under weaker conditions.

# 5 BIC

BIC stands for *Bayesian Information Criterion*. It is also known as *the Schwarz Criterion* after Gideon Schwarz. It is virtually identical to the MDL (minimum description length) criterion.

We choose $j$ to maximize

$$\text{BIC}_j = \ell_j(\widehat{\theta}_j) - \frac{d_j}{2}\log n.$$

This is the same as AIC but the penalty is harsher. Thus, BIC tends to choose simpler models. Here is the derivation.

We put a prior $\pi_j(\theta_j)$ on the parameter $\theta_j$. We also put a prior probability $p_j$ that model $\mathcal{M}_j$ is the true model. By Bayes theorem

$$P(\mathcal{M}_j|Y_1,\ldots,Y_n) \propto p(Y_1,\ldots,Y_n|\mathcal{M}_j)p_j.$$

Furthermore,

$$p(Y_1,\ldots,Y_n|\mathcal{M}_j) = \int p(Y_1,\ldots,Y_n|\mathcal{M}_j,\theta_j)\pi_j(\theta_j)d\theta_j = \int L(\theta_j)\pi_j(\theta_j)d\theta_j.$$

We know choose $j$ to maximize $P(\mathcal{M}_j|Y_1,\ldots,Y_n)$. Equivalently, we choose $j$ to maximize

$$\log \int L(\theta_j)\pi_j(\theta_j)d\theta_j + \log p_j.$$

Some Taylor series approximations show that

$$\log \int L(\theta_j)\pi_j(\theta_j)d\theta_j + \log p_j \approx \ell_j(\widehat{\theta}_j) - \frac{d_j}{2}\log n = \text{BIC}_j.$$

What happened to the prior? It can be shown that the terms involving the prior are lower order than the term that appear in formula for $\text{BIC}_j$ so they have been dropped.

BIC behaves quite differently than AIC or cross-validation. It is also based on different assumptions. BIC assumes that one of the models is true and that you are trying to find the model most likely to be true in the Bayesian sense. AIC and cross-validation are trying to find the model that predict the best.

# 6 Model Averaging

**Bayesian Approach.** Suppose we want to predict a new observation $Y$. Let $D = \{Y_1,\ldots,Y_n\}$ be the observed data. Then

$$p(y|D) = \sum_j p(y|D,\mathcal{M}_j)\mathbb{P}(\mathcal{M}_j|D)$$

where

$$\mathbb{P}(\mathcal{M}_j|D) = \frac{\int L(\theta_j)\pi_j(\theta_j)d\theta_j}{\sum_s \int L(\theta_s)\pi_s(\theta_s)d\theta_s} \approx \frac{e^{\text{BIC}_j}}{\sum_s e^{\text{BIC}_s}}.$$

**Frequentist Approach.** There is a large and growing literaure on frequenist model averaging. It is discussed in 10-702.

# 7 Simple Normal Example

Let

$$Y_1, \ldots, Y_n \sim N(\mu, 1).$$

We want to compare two models:

$$M_0 : \ N(0,1), \quad \text{and} \quad M_1 : \ N(\mu, 1).$$

**Hypothesis Testing.** We test

$$H_0 : \mu = 0 \quad \text{versus} \quad \mu \neq 0.$$

The test statistic is

$$Z = \frac{\overline{Y} - 0}{\sqrt{\mathsf{Var}(\overline{Y})}} = \sqrt{n}\,\overline{Y}.$$

We reject $H_0$ if $|Z| > z_\alpha/2$. For $\alpha = 0.05$, we reject $H_0$ if $|Z| > 2$, i.e., if

$$|\overline{Y}| > \frac{2}{\sqrt{n}}.$$

**AIC.** The likelihood is proportional to

$$\mathcal{L}(\mu) = \prod_{i=1}^{n} e^{-(Y_i-\mu)^2/2} = e^{-n(\overline{Y}-\mu)^2/2} e^{-nS^2/2}$$

where $S^2 = \sum_i (Y_i - \overline{Y})^2$. Hence,

$$\ell(\mu) = -\frac{n(\overline{Y}-\mu)^2}{2} - \frac{nS^2}{2}.$$

Recall that $AIC = \ell_S - |S|$. The AIC scores are

$$AIC_0 = \ell(0) - 0 = -\frac{n\overline{Y}^2}{2} - \frac{nS^2}{2}$$

and

$$AIC_1 = \ell(\widehat{\mu}) - 1 = -\frac{nS^2}{2} - 1$$

since $\widehat{\mu} = \overline{Y}$. We choose model 1 if

$$AIC_1 > AIC_0$$

that is, if

$$-\frac{nS^2}{2} - 1 > -\frac{n\overline{Y}^2}{2} - \frac{nS^2}{2}$$

or

$$|\overline{Y}| > \frac{\sqrt{2}}{\sqrt{n}}.$$

Similar to but not the same as the hypothesis test.

**BIC.** The BIC scores are

$$BIC_0 = \ell(0) - \frac{0}{2}\log n = -\frac{n\overline{Y}^2}{2} - \frac{nS^2}{2}$$

and

$$BIC_1 = \ell(\widehat{\mu}) - \frac{1}{2}\log n = -\frac{nS^2}{2} - \frac{1}{2}\log n.$$

We choose model 1 if

$$BIC_1 > BIC_0$$

that is, if

$$|\overline{Y}| > \sqrt{\frac{\log n}{n}}.$$

| | |
|---|---|
| Hypothesis testing | controls type I errors |
| AIC/CV/$C_p$ | finds the most predictive model |
| BIC | finds the true model (with high probability) |

# Lecture Notes 17

## 1 Multiple Testing and Confidence Intervals

Suppose we need to test many null hypotheses

$$H_{0,1}, \ldots, H_{0,N}$$

where $N$ could be very large. We cannot simply test each hypotheses at level $\alpha$ because, if $N$ is large, we are sure to make lots of type I errors just by chance. We need to do some sort of *multiplicity adjustment.*

**Familywise Error Control.** Suppose we get a $p$-value $p_j$ for each null hypothesis. Let $I = \{i : \ H_{0,i} \text{ is true}\}$. If we reject $H_{0,i}$ for any $i \in I$ then we have made an error. Let $R = \{j : \text{ we reject } H_{0j}\}$ be the set of hypotheses we reject. We say that we have controlled the *familywise error rate* at level $\alpha$ if

$$\mathbb{P}(R \cap I \neq \emptyset) \leq \alpha.$$

The easiest way to control the familywise error rate is the *Bonferroni method.* The idea is to reject $H_{0,i}$ if and only if $p_i < \alpha/N$. Then

$$\begin{aligned}
\mathbb{P}(\text{making a false rejection}) &= \mathbb{P}\left(p_i < \frac{\alpha}{N} \quad \text{for some } i \in I\right) \\
&\leq \sum_{i \in I} \mathbb{P}\left(p_i < \frac{\alpha}{N}\right) \\
&= \sum_{i \in I} \frac{\alpha}{N} \quad \text{since } p_i \sim \text{Unif}(0,1) \text{ for } i \in I \\
&= \frac{\alpha \, |I|}{N} \leq \alpha.
\end{aligned}$$

So we have overall control of the type I error. However, it can have low power.

*The Normal Case.* Suppose that we have $N$ sample means $Y_1, \ldots, Y_N$ each based on $n$ Normal observations with variance 1. So $Y_j \sim N(\mu_j, 1/n)$. To test $H_{0,j} : \mu_j = 0$ we can use

the test statistic $T_j = \sqrt{n}Y_j$. The p-value is

$$p_j = 2\Phi(-|T_j|).$$

If we did uncorrected testing we rject when $p_j < \alpha$, which means, $|T_j| > z_{\alpha/2}$. A useful approximation is:

$$z_\alpha \approx \sqrt{2\log(1/\alpha)}.$$

So we reject when

$$|T_j| > \sqrt{2\log(2/\alpha)}.$$

Under the Bonferroni correction we reject when $p_j < \alpha/N$ which coresponds to

$$|T_j| > \sqrt{2\log(2N/\alpha)}.$$

Hence, the familywise rejection threshold grows like $\sqrt{\log N}$.

**False Discovery Control.** The Bonferroni adjustment is very strict. A weaker type of control is based on the *false discovery rate*.[1] Suppose we reject a set of hyptheses $R$. Define the *false discovery proportion*

$$\text{FDP} = \frac{|R \cap I|}{|R|}$$

where the ratio is defined to be 0 in case both the numerator and denominator are 0. Our goal is to find a method for choosing $R$ such that

$$\text{FDR} = \mathbb{E}(\text{FDP}) \leq \alpha.$$

The *Benjamini-Hochberg method* works as follows:

1. Find the ordered p-values $P_{(1)} < \cdots < P_{(N)}$.
2. Let $j = \max\{i : \ P_{(i)} < i\alpha/N\}$. Let $T = P_{(j)}$.
3. Let $R = \{i : \ P_i \leq T\}$.

[1]Reference: Benjamini and Hochberg (1995).

Let us see why this controls the FDR. Consider, in general, rejecting all hypothesis for which $P_i < t$. Let $W_i = 1$ if $H_{0,i}$ is true and $W_i = 0$ otherwise. Let $\widehat{G}$ be the empirical distribution of the p-values and let $G(t) = \mathbb{E}(\widehat{G}(t))$. In this case,

$$\text{FDP} = \frac{\sum_{i=1}^N W_i I(P_i < t)}{\sum_{i=1}^N I(P_i < t)} = \frac{\frac{1}{N}\sum_{i=1}^N W_i I(P_i < t)}{\frac{1}{N}\sum_{i=1}^N I(P_i < t)}.$$

Hence,

$$\begin{aligned}
\mathbb{E}(\text{FDP}) &\approx \frac{\mathbb{E}(\frac{1}{N}\sum_{i=1}^N W_i I(P_i < t))}{\frac{1}{N}\mathbb{E}(\sum_{i=1}^N I(P_i < t))} \\
&= \frac{\frac{1}{N}\sum_{i=1}^N W_i \mathbb{E}(I(P_i < t))}{\frac{1}{N}\sum_{i=1}^N \mathbb{E}(I(P_i < t))} \\
&= \frac{t|I|}{G(t)} \leq \frac{t}{G(t)} \approx \frac{t}{\widehat{G}(t)}.
\end{aligned}$$

Let $t = P_{(i)}$ for some $i$; then $\widehat{G}(t) = i/N$. Thus, FDR $\leq P_{(i)}N/i$. Setting this equal to $\alpha$ we get $P_{(i)} < i\alpha/N$ is the Benjamini-Hochberg rule.

FDR control typically has higher power than familywise control. But they are controlling different things. You have to decide, based on the context, which is appropriate.

**Example 1** *Figure 1 shows an example where $Y_j \sim N(\mu_j, 1)$ for $j = 1, \ldots, 1,000$. In this example, $\mu_j = 3$ for $1 \leq j \leq 50$ and $\mu_j = 0$ for $j > 50$. The figure shows the test statistics, the p-values, the sorted log p-values with the Bonferroni threshold and the sorted log p-values with the FDR threshold (using $\alpha = 0.05$). Bonferroni rejects 7 hypotheses while FDR rejects 22.*

**Multiple Confidence Intervals.** A similar problem occurs with confidence intervals. If we construct a confidence interval $C$ for one parameter $\theta$ then $\mathbb{P}(\theta \in C) \geq 1 - \alpha$. But if we construct confidence intervals $C_1, \ldots, C_N$ for $N$ parameters $\theta_1, \ldots, \theta_N$ then we want to ensure that

$$\mathbb{P}(\theta_j \in C_j, \quad \text{for all } j = 1, \ldots, N) \geq 1 - \alpha.$$

To do this, we construct each confidence interval $C_j$ at level $1 - \alpha/N$. Then

$$\mathbb{P}(\theta_j \notin C_j \text{ for some } j) \leq \sum_j \mathbb{P}(\theta_j \notin C_j) \leq \sum_j \frac{\alpha}{N} = \alpha.$$

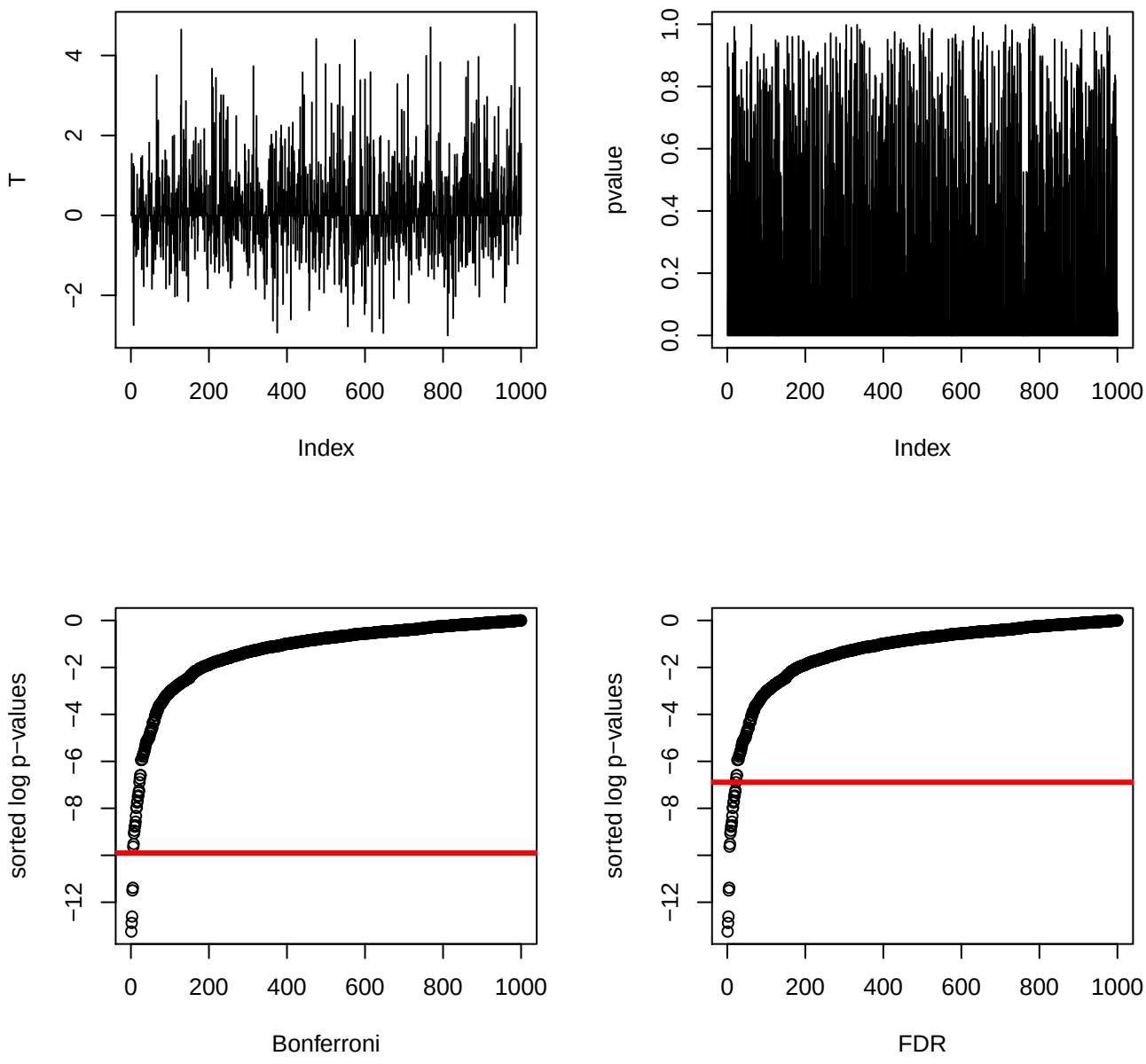



Figure 1: Top left: 1,000 test statistics. Top right: the p-values. Bottom left: sorted log p-values and Bonferroni threshold. Bottom right: sorted log p-values and FDR threshold.

# 2 Causation

Most of statistics and machine learning is concerned with prediction. A typical question is: what is a good prediction of $Y$ given that I **observe** that $X = x$? Causation is concerned with questions of the form: what is a good prediction of $Y$ given that I **set** $X = x$? The difference between passively observing $X = x$ and actively intervening and setting $X = x$ is significant and requires different techniques and, typically, much strnger assumptions.

Consider this story. A mother notices that tall kids have a higher reading level than short kids. (This is because the tall kids are older.) The mother puts her small child on a device and stretches the child until he is tall. She is dismayed to find out that his reading level has not changed.

Te mother is correct that height and reading skill are **associated**. Put another way, you can use height to predict reading skill. But that does not imply that height *causes* reading skill. This is what statisticians mean when they say:

**correlation is not causation.**

On the other hand, consider smoking and lung cancer. We know that smoking and lung cancer are associated. But we also believe that smoking causes lung cancer. In this case, we recognize that intervening and forcing someone to smoke does change his probability of getting lung cancer.

The difference between prediction (association/correlation) and causation is this: in prediction we are interested in

$$\mathbb{P}(Y \in A | X = x)$$

which means: the probability that $Y \in A$ given that we **observe** that $X$ is equal to $x$. For causation we are interested in

$$\mathbb{P}(Y \in A | \mathsf{set}\ X = x)$$

which means: the probability that $Y \in A$ given that we **set** $X$ equal to $x$. Prediction is about passive observation. Causation is about active intervention. Most of statistics and

machine learning concerns prediction. But sometimes causation is the primary focus. The phrase **correlation is not causation** can be written mathematically as

$$\mathbb{P}(Y \in A|X = x) \neq \mathbb{P}(Y \in A|\mathsf{set}\ X = x).$$

Despite the fact that causation and association are different, people mix them up all the time, even people trained in statistics and machine learning. On TV recently there was a report that good health is associated with getting seven hours of sleep. So far so good. Then the reporter goes on to say that, therefore, everyone should strive to sleep exactly seven hours so they will be healthy. Wrong. That's confusing causation and association. Another TV report pointed out a correlation between people who brush their teeth regularly and low rates of heart disease. An interesting correlation. Then the reporter (a doctor in this case) went on to urge people to brush their teeth to save their hearts. Wrong!

To avoid this confusion we need a way to discuss causation mathematically. That is, we need someway to make $\mathbb{P}(Y \in A|\mathsf{set}\ X = x)$ formal. There are two common ways to do this. One is to use **counterfactuals**. The other is to use **causal graphs**. These approaches are equivalent. There are two different languages for saying the same thing.

Causal inference is tricky and should be used with great caution. The main messages are:

1. Causal effects can be estimated consistently from randomized experiments.
2. It is difficult to estimate causal effects from observational (non-randomized) experiements.
3. All causal conclusions from observational studes should be regarded as very tentative.

Causal inference is a vast topic. We will only touch on the main ideas here.

**Counterfactuals.** Consider two variables $Y$ and $X$. Suppose that $X$ is a binary variable that represents some treatment. For example, $X = 1$ means the subject was treated and $X = 0$ means the subject was given placebo. The response variable $Y$ is real-valued.

We can address the problem of predicting $Y$ from $X$ by estimating $\mathbb{E}(Y|X = x)$. To address causal questions, we introduce *counterfactuals*. Let $Y_1$ denote the response we observe if the subject is treated, i.e. if we set $X = 1$. Let $Y_0$ denote the response we observe if the

subject is not treated, i.e. if we set $X = 0$. If we treat a subject, we observe $Y_1$ but we do not observe $Y_0$. Indeed, $Y_0$ is the value we would have observed if the subject had been treated. The unobserved variable is called a counterfactual.

We have enlarged our set of variables from $(X, Y)$ to $(X, Y, Y_0, Y_1)$. Note that

$$Y = XY_1 + (1 - X)Y_0. \tag{1}$$

A small dataset might look like this:

| $X$ | $Y$ | $Y_0$ | $Y_1$ |
|---|---|---|---|
| 1 | 1 | * | 1 |
| 1 | 1 | * | 1 |
| 1 | 0 | * | 0 |
| 1 | 1 | * | 1 |
| 0 | 1 | 1 | * |
| 0 | 0 | 0 | * |
| 0 | 1 | 1 | * |
| 0 | 1 | 1 | * |

The asterisks indicate unobserved variables. To answer causal questions, we are interested in the distribution $p(y_0, y_1)$. We can interpret $p(y_1)$ as $p(y|\mathsf{set}\ X = 1)$ and we can interpret $p(y_0)$ as $p(y|\mathsf{set}\ X = 0)$. In particular, we might want to estimate the *mean treatment effect* or *mean causal effect*

$$\theta = \mathbb{E}(Y_1) - \mathbb{E}(Y_0) = \mathbb{E}(Y|\mathsf{set}\ X = 1) - \mathbb{E}(Y|\mathsf{set}\ X = 0).$$

The parameter $\theta$ has the following intepretation: $\theta$ is the mean response if we forced everyone to take the treatment minus mean response if we forced everyone not to take the treatment.

Suppose now that we observe a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Can we estmate $\theta$? No. In general, there is no consistnt estimator of $\theta$. We can estimate $\alpha = \mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0)$ but $\alpha$ is not equal to $\theta$.

However, suppose that we did a randomized experiment where we randomly assigned each person to treatment of placebo by the flip of a coin. In this case, $X$ will be independent of $(Y_0, Y_1)$. In symbols:

$$\text{random treatment assignment implies : } \ (Y_0, Y_1) \amalg X.$$

Hence, in this case,

$$\begin{aligned} \alpha &= \mathbb{E}(Y|X=1) - \mathbb{E}(Y|X=0) \\ &= \mathbb{E}(Y_1|X=1) - \mathbb{E}(Y_0|X=0) \quad \text{since } Y = XY_1 + (1-X)Y_0 \\ &= \mathbb{E}(Y_1) - \mathbb{E}(Y_0) = \theta \quad \text{since } (Y_0, Y_1) \amalg X. \end{aligned}$$

Hence, random assignment makes $\theta$ equal to $\alpha$ and $\alpha$ can be consistently estimated. **If $X$ is randomly assigned then correlation = causation.** This is why people spend millions of dollars doing randomized experiements.

In some cases it is not feasible to do a randomized experiment. Smoking and lung cancer is an example. Can we estimate causal parameters from observational (non-randomized) studies? The answer is: sort of.

In an observational stsudy, the treated and untreated groups will not be comparable. Maybe the healthy people chose to take the treatment and the unhealthy people didn't. In other words, $X$ is not independent of $(Y_0, Y_1)$. The treatment may have no effect but we would still see a strong association between $Y$ and $X$. In other words, $\alpha$ might be large even though $\theta = 0$.

To account for the differences in the groups, we might measure confounding variables. These are the variables that affect both $X$ and $Y$. By definition, there are no such variables in a randomized experiment. The hope is that if we measure enough confounding variables $Z = (Z_1, \ldots, Z_k)$, then, perhaps the treated and untreated groups will be comparable, conditional on $Z$. Formally, we hope that $X$ is indpendent of $(Y_0, Y_1)$ conditional on $Z$. If this is true,

we can estimate $\theta$ since

$$\begin{aligned}
\theta &= \mathbb{E}(Y_1) - \mathbb{E}(Y_0) \\
&= \int \mathbb{E}(Y_1|Z=z)p(z)dz - \int \mathbb{E}(Y_0|Z=z)p(z)dz \\
&= \int \mathbb{E}(Y_1|X=1,Z=z)p(z)dz - \int \mathbb{E}(Y_0|X=0,Z=z)p(z)dz \\
&= \int \mathbb{E}(Y|X=1,Z=z)p(z)dz - \int \mathbb{E}(Y|X=0,Z=z)p(z)dz \qquad (2)
\end{aligned}$$

where we used the fact that $X$ is indpendent of $(Y_0, Y_1)$ conditional on $Z$ in the third line and the fact that $Y = (1-X)Y_1 + XY_0$ in the fourth line. The latter quantity can be estimated by

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(1, Z_i) - \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(0, Z_i)$$

where $\widehat{m}(x,z)$ is an estimate of the regression function $m(x,z) = \mathbb{E}(Y|X=x, Z=z)$. This is known as *adjusting for confounders* and $\widehat{\theta}$ is called the *adjusted treatment effect*.

It is instructive to compare the casual effect

$$\begin{aligned}
\theta &= \mathbb{E}(Y_1) - \mathbb{E}(Y_0) \\
&= \int \mathbb{E}(Y|X=1,Z=z)p(z)dz - \int \mathbb{E}(Y|X=0,Z=z)p(z)dz
\end{aligned}$$

with the predictive quantity

$$\begin{aligned}
\alpha &= \mathbb{E}(Y|X=1) - \mathbb{E}(Y|X=0) \\
&= \int \mathbb{E}(Y|X=1,Z=z)p(z|X=1)dz - \int \mathbb{E}(Y|X=0,Z=z)p(z|X=0)dz
\end{aligned}$$

which are mathematically (and conceptually) quite different.

We need to treat $\widehat{\theta}$ cautiously. It is very unlikely that we have successfully measured all the relevant confounding variables so $\widehat{\theta}$ should be regarded as a crude approximation to $\theta$ at best.

**Causal Graphs.** Another way to capture the difference between $P(Y \in A|X=x)$ and $P(Y \in A|\mathsf{set}\ X=x)$ is to represent the distribution using a directed graph and then we capture the second statement by performing certain operations on the graph.

A Directed Acyclic Graph (DAG) is a graph for a set of variables with no cycles. The graph defines a set of distributions of the form

$$p(y_1, \ldots, y_k) = \prod p(y_j | \mathsf{parents}(y_j)$$

where $\mathsf{parents}(y_j)$ are the parents of $y_j$. A **causal graph** is a DAG with extra information. A DAG is a causal graph if it correctly encodes the effect of setting a variable to a fixed value.

Consider the graph $G$ in Figure (2). Here, $X$ denotes treatment, $Y$ is response and $Z$ is a confounding variable. To find the causal distribution $p(y|\mathsf{set}\ X = x)$ we do the following steps:

1. Form a new graph $G_*$ by removing all arrow into $X$. Now set $X$ equal to $x$. This corresponds to replacing the joint distribution $p(x, y, z) = p(z)p(x|z)p(y|x, z)$ with the new distribution $p_*(y, z) = p(z)p(y|x, z)$. The factor $p(x|z)$ is removed because we know regard $x$ as a fixed number.
2. Compute the distribution of $y$ from the new distribution:

$$p(y|\mathsf{set}\ X = x) \equiv p_*(y) = \int p_*(y, z)dz = \int p(z)p(y|x, z)dz.$$

Now we have that

$$\theta = p(y|\mathsf{set}\ X = 1) - p(y|\mathsf{set}\ X = 0) = \int p(z)p(y|1, z)dz - \int p(z)p(y|0, z)dz$$

This is precisely the same equation as (2). Both approaches lead to the same thing. If there were unobserved confounding variables, then the formula for $\theta$ would involve these variables and the causal effect would be non-estimable (as before).

In a randomized experiment, there would be no arrow from $Z$ to $X$. (That's the point of randomization). In that case the above calculations shows that $\theta = \mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0)$ just as we saw with the counterfactual approach.

To understand the difference between $p(y|x)$ and $p(y|\mathrm{set}\, x)$ more clearly, it is helpful to consider two different computer programs. Consider the DAG in Figure 2. The

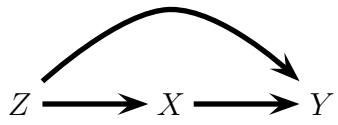



Figure 2: Conditioning versus intervening.

probability function for a distribution consistent with this DAG has the form $p(x, y, z) = p(x)p(y|x)p(z|x, y)$. The following is pseudocode for generating from this distribution.

$$\begin{aligned} \text{For } i &= 1, \ldots, n: \\ x_i &\leftarrow p_X(x_i) \\ y_i &\leftarrow p_{Y|X}(y_i|x_i) \\ z_i &\leftarrow p_{Z|X,Y}(z_i|x_i, y_i) \end{aligned}$$

Suppose we run this code, yielding data $(x_1, y_1, z_1), \ldots, (x_n, y_n, z_n)$. Among all the times that we observe $Y = y$, how often is $Z = z$? The answer to this question is given by the conditional distribution of $Z|Y$. Specifically,

$$\begin{aligned} \mathbb{P}(Z = z|Y = y) &= \frac{\mathbb{P}(Y = y, Z = z)}{\mathbb{P}(Y = y)} = \frac{p(y, z)}{p(y)} \\ &= \frac{\sum_x p(x, y, z)}{p(y)} = \frac{\sum_x p(x)\, p(y|x)\, p(z|x, y)}{p(y)} \\ &= \sum_x p(z|x, y) \frac{p(y|x)\, p(x)}{p(y)} = \sum_x p(z|x, y) \frac{p(x, y)}{p(y)} \\ &= \sum_x p(z|x, y)\, p(x|y). \end{aligned}$$

Now suppose we **intervene** by changing the computer code. Specifically, suppose we fix $Y$ at the value $y$. The code now looks like this:

$$\begin{aligned} \text{set } Y &= y \\ \text{for } i &= 1, \ldots, n \\ x_i &<- p_X(x_i) \\ z_i &<- p_{Z|X,Y}(z_i|x_i, y) \end{aligned}$$

Having **set** $Y = y$, how often was $Z = z$? To answer, note that the intervention has changed the joint probability to be

$$p^*(x, z) = p(x)p(z|x, y).$$

The answer to our question is given by the marginal distribution

$$p^*(z) = \sum_x p^*(x, z) = \sum_x p(x)p(z|x, y).$$

This is $p(z|\mathsf{set}\ Y = y)$.

**Example 2** *You may have noticed a correlation between rain and having a wet lawn, that is, the variable "Rain" is not independent of the variable "Wet Lawn" and hence* $p_{R,W}(r, w) \neq p_R(r)p_W(w)$ *where* $R$ *denotes Rain and* $W$ *denotes Wet Lawn. Consider the following two DAGs:*

$$\text{Rain} \longrightarrow \text{Wet Lawn} \qquad \text{Rain} \longleftarrow \text{Wet Lawn}.$$

*The first DAG implies that* $p(w, r) = p(r)p(w|r)$ *while the second implies that* $p(w, r) = p(w)p(r|w)$ *No matter what the joint distribution* $p(w, r)$ *is, both graphs are correct. Both imply that* $R$ *and* $W$ *are not independent. But, intuitively, if we want a graph to indicate causation, the first graph is right and the second is wrong. Throwing water on your lawn doesn't cause rain. The reason we feel the first is correct while the second is wrong is because the interventions implied by the first graph are correct.*

*Look at the first graph and form the intervention* $W = 1$ *where 1 denotes "wet lawn." Following the rules of intervention, we break the arrows into* $W$ *to get the modified graph:*

$$\text{Rain} \qquad \boxed{\textbf{set}\ \ \text{Wet Lawn} = 1}$$

*with distribution* $p^*(r) = p(r)$. *Thus* $\mathbb{P}(R = r \mid W := w) = \mathbb{P}(R = r)$ *tells us that "wet lawn" does not cause rain.*

*Suppose we (wrongly) assume that the second graph is the correct causal graph and form the intervention* $W = 1$ *on the second graph. There are no arrows into* $W$ *that need to be broken so the intervention graph is the same as the original graph. Thus* $p^*(r) = p(r|w)$ *which would imply that changing "wet" changes "rain." Clearly, this is nonsense.*

*Both are correct probability graphs but only the first is correct causally. We know the correct causal graph by using background knowledge.*

**Learning Casual Structure?** We could try to learn the correct causal graph from data but this is dangerous. In fact it is impossible with two variables. With more than two variables there are methods that can find the causal graph under certain assumptions but they are large sample methods and, furthermore, there is no way to ever know if the sample size you have is large enough to make the methods reliable.

**Randomization Again.** We can use DAGs to represent **confounding** variables. If $X$ is a treatment and $Y$ is an outcome, a confounding variable $Z$ is a variable with arrows into both $X$ and $Y$; see Figure 3. It is easy to check, using the formalism of interventions, that the following facts are true:

In a randomized study, the arrow between $Z$ and $X$ is broken. In this case, even with $Z$ unobserved (represented by enclosing $Z$ in a circle), the causal relationship between $X$ and $Y$ is estimable because it can be shown that $\mathbb{E}(Y|X := x) = \mathbb{E}(Y|X = x)$ which does not involve the unobserved $Z$. In an observational study, with all confounders observed, we get $\mathbb{E}(Y|X := x) = \int \mathbb{E}(Y|X = x, Z = z)p(z)$ which is just the **adjusted treatment effect.** If $Z$ is unobserved then we cannot estimate the causal effect because $\mathbb{E}(Y|X := x) = \int \mathbb{E}(Y|X = x, Z = z)dF_Z(z)$ involves the unobserved $Z$. We can't just use $X$ and $Y$ since in this case. $\mathbb{P}(Y = y|X = x) \neq \mathbb{P}(Y = y|X := x)$ which is just another way of saying that causation is not association.

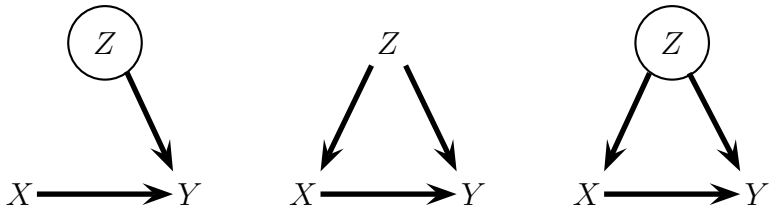



Figure 3: Randomized study; Observational study with measured confounders; Observational study with unmeasured confounders. The circled variables are unobserved.

# 3 Individual Sequence Prediction

The goal is to predict $y_t$ from $y_1, \ldots, y_{t-1}$ with no assumptions on the sequence. [2] The data are not assumed to be iid; they are not even assumed to be random. This is a version of *online learning.* For simplicity assume that $y_t \in \{0, 1\}$.

Suppose we have a set of *prediction algorithms* (or *experts*):

$$\mathcal{F} = \{F_1, \ldots, F_N\}$$

Let $F_{j,t}$ is the prediction of algorithm $j$ at time $t$ based on $y^{t-1} = (y_1, \ldots, y_{t-1})$. At time $t$:

1. You see $y^{t-1}$ and $(F_{1,t}, \ldots, F_{N,t})$.
2. You predict $P_t$.
3. $y_t$ is revealed.
4. You suffer loss $\ell(P_t, y_t)$.

We will focus on the loss $\ell(p_t, y_t) = |p_t - y_t|$ but the theory works well for any convex loss. The *cumulative loss* is

$$L_j(y^n) = \frac{1}{n}\sum_{i=1}^{n} |F_{j,t} - y_t| \equiv \frac{1}{n} S_j(y^n)$$

where $S_j(y^n) = \sum_{i=1}^{n} |F_{j,t} - y_t|$. The *maximum regret* is

$$R_n = \max_{y^t \in \{0,1\}^t} \left( L_P(y^n) - \min_j L_j(y^n) \right)$$

[2] Reference: *Prediction, Learning, and Games.* Nicolò Cesa-Bianchi and Gábor Lugosi, 2006.

and the *minimax regret* is

$$V_n = \inf_P \max_{y^t \in \{0,1\}^t} \left( L_P(y^n) - \min_j L_j(y^n) \right).$$

Let $P_t(y^{t-1}) = \sum_{j=1}^N w_{j,t-1}\ F_{j,t}$ where

$$w_{j,t-1} = \frac{\exp\{-\gamma S_{j,t-1}\}}{Z_t}$$

and $Z_t = \sum_{j=1}^N \exp\{-\gamma S_{j,t-1}\}$. The $w_j$'s are called *exponential weights*.

**Theorem 3** *Let $\gamma = \sqrt{8 \log N/n}$. Then*

$$L_P(y^n) - \min_{1 \le j \le N} L_j(y^n) \le \sqrt{\frac{\log N}{2n}}.$$

**Proof.** The idea is to place upper and lower bounds on $\log\left(\frac{Z_{n+1}}{Z_1}\right)$ then solve for $L_P(y^n)$.

*Upper bound:* We have

$$\begin{aligned} \log\left(\frac{Z_{n+1}}{Z_1}\right) &= \log\left(\sum_{j=1}^N \exp\{-\gamma n L_{j,n}\}\right) - \log N \\ &\ge \log\left(\max_j \exp\{-\gamma n L_{j,n}\}\right) - \log N \\ &= -\gamma n \min_j L_{j,n} - \log N. \end{aligned} \tag{3}$$

*Lower bound*: Note that

$$\begin{aligned} \log\left(\frac{Z_{t+1}}{Z_t}\right) &= \log\left(\frac{\sum_{j=1}^N w_{j,t-1}\ e^{-\gamma|F_{j,t}-y_t|}}{\sum_{j=1}^N w_{j,t-1}}\right) \\ &= \log \mathbb{E}\left(e^{-\gamma|F_{j,t}-y_t|}\right). \end{aligned}$$

This is a formal expectation with respect to the distribution over $j$ probability proportional to $e^{-\gamma|F_{j,t}-y_t|}$).

Recall Hoeffding's bound for mgf: if $a \le X \le b$

$$\log \mathbb{E}(e^{sX}) \le s\mathbb{E}(X) + \frac{s^2(b-a)^2}{8}.$$

So:

$$\begin{aligned}\log \mathbb{E}\left(e^{-\gamma|F_{j,t}-y_t|}\right) &\leq -\gamma\mathbb{E}|F_{j,t}-y_t| + \frac{\gamma^2}{8} \\ &= -\gamma|\mathbb{E}F_{j,t}-y_t| + \frac{\gamma^2}{8} \\ &= -\gamma|P_t(y^{t-1})-y_t| + \frac{\gamma^2}{8}.\end{aligned}$$

Summing over $t$:

$$\log\left(\frac{Z_{n+1}}{Z_1}\right) \leq -\gamma n L_P(y^n) + \frac{n\gamma^2}{8}. \tag{4}$$

Combining (3) and (4) we get

$$-\gamma n \min_j L_j(y^n) - \log N \leq \log\left(\frac{Z_{n+1}}{Z_1}\right) \leq -\gamma n L_P(y^n) + \frac{n\gamma^2}{8}.$$

Rearranging the terms we have:

$$L_P(y^n) \leq \min_j L_j(y^n) + \frac{\log N}{\gamma} + \frac{n\gamma}{8}.$$

Set $\gamma = \sqrt{8\log N/n}$ to get

$$L_P(y^n) - \min_{1\leq j\leq N} L_j(y^n) \leq \sqrt{\frac{\log N}{2n}}.$$

■

The result held for a specific time $n$. We can make the result uniform over time as follows. If we set $\gamma_t = \sqrt{8\log N/t}$ then we have:

$$L_P(y^n) \leq \min_j L_j(y^n) + \sqrt{\frac{1+12n\log N}{8}}$$

for all $n$ and for all $y_1, y_2, \ldots, y_n$.

Now suppose that $\mathcal{F}$ is an infinite class. A set $\mathcal{G} = \{G_1, \ldots, G_N\}$ is an $r$-covering if, for every $F$ and every $y^n$ there is a $G_j$ such that

$$\sum_{t=1}^{n} |F_t(y^{t-1}) - G_{j,t}(y^{t-1})| \leq r.$$

Let $N(r)$ denote the size of the smallest $r$-covering.

**Theorem 4 (Cesa-Bianchi and Lugosi)** *We have that*

$$V_n(\mathcal{F}) \leq \inf_{r>0} \left( \frac{r}{n} + \sqrt{\frac{\log N(r)}{2n}} \right)$$

Cesa-Bianchi and Lugosi also construct a predictor that nearly achieves the bound of the form

$$P_t = \sum_{k=1}^{\infty} a_k P_t^{(k)}$$

where $P_t^{(k)}$ is a predictor based on a finite subset of $\mathcal{F}$.

Using *batchification* it is possible to use online learning for non-online learning. Suppose we are given data: $(Z_1, \ldots, X_n)$ where $Z_i = (X_i, Y_i)$ and an arbitrary algorithm $A$ that takes data and outputs classifier $H$. We used uniform convergence theory to analyze $H$ but online methods provide an alternative analysis. [3] We apply $A$ sequentially to get classifiers $H_0, H_1, \ldots, H_n$. Let

$$M_n = \frac{1}{n} \sum_{i=1}^{n} \ell(H_{t-1}(X_t), Y_t)$$

To choose a final classifier:

1. usual batch method: use the last one $H_n$
2. average: $\overline{H} = \frac{1}{n} \sum_{i=1}^{n} H_{t-1}$
3. selection: choose $H_t$ to minimize

$$\frac{1}{t} \sum_{i=1}^{t} \ell(H_t(X_t), Y_t) + \sqrt{\frac{1}{2(n-t)} \log \left( \frac{n(n+1)}{\delta} \right)}$$

Analyzing $H_n$ requires assumptions on $A$, uniform convergence etc. This is not needed for the other two methods.

**Theorem 5** *If $\ell$ is convex:*

$$\mathbb{P}\left( R(\overline{H}) \geq M_n + \sqrt{\frac{2}{n} \log \left( \frac{1}{\delta} \right)} \right) \leq \delta.$$

*For any $\ell$,*

$$\mathbb{P}\left( R(\widehat{H}) \geq M_n + \sqrt{\frac{36}{n} \log \left( \frac{2(n+1)}{\delta} \right)} \right) \leq \delta.$$

[3] Reference: Cesa-Bianchi, Conconi and Gentile (2004).

Homework 1
36-705
Due: Thursday Sept 8 by 3:00

From Casella and Berger:

1. Chapter 1, problem 1.47.
2. Chapter 1, problem 1.49.
3. Chapter 2, problem 2.1.
4. Chapter 2, problem 2.3.
5. Chapter 2, problem 2.7a.
6. Chapter 2, problem 2.15.
7. Chapter 2, problem 2.30.
8. Chapter 3, problem 3.32.
9. Chapter 4, problem 4.4.
10. Chapter 4, problem 4.5.

Homework 2
36-705
Due: Thursday Sept 15 by 3:00

1. Let $X_n$ be a sequence of random variables such that $X_n \geq 0$ for all $n$. Suppose that $\mathbb{P}(X_n > t) \leq (\frac{1}{t})^k$ where $k > 1$. Derive an upper bound on $\mathbb{E}(X_n)$.

2. Let $X_1, \ldots, X_n \sim \text{Unif}(0,1)$. Let $Y = \max_{1 \leq i \leq n} X_i$.

   (i) Bound $\mathbb{E}(Y)$ using the method we derived in lecture notes 2.

   (ii) Find an exact expression for $\mathbb{E}(Y)$. Compare the result to part (i).

3. An improvement on Hoeffding's inequality is Bernstein's inequality. Let $X_1, \ldots, X_n$ be iid, with mean $\mu$, $\text{Var}(X_i) = \sigma^2$ and $|X_i| \leq c$. Then Bernstein's inequality says that

$$\mathbb{P}\left(|\overline{X}_n - \mu| > \epsilon\right) \leq 2\exp\left\{-\frac{n\epsilon^2}{2\sigma^2 + 2c\epsilon/3}\right\}.$$

   (When $\sigma$ is sufficiently small, this bound is tighter than Hoeffding's inequality.) Let $X_1, \ldots, X_n \sim \text{Uniform}(0,1)$ and $A_n = [0, 1/n]$. Let $p_n = \mathbb{P}(X_i \in A_n)$ and let

$$\widehat{p}_n = \frac{1}{n}\sum_{i=1}^{n} I_{A_n}(X_i).$$

   (i) Use Hoeffding's inequality and Bernstein's inequality to bound

$$\mathbb{P}(|\widehat{p}_n - p_n| > \epsilon).$$

   (ii) Show that the bound from Bernstein's inequality is tighter.

   (iii) Show that Hoeffding's inequality implies $\widehat{p}_n - p_n = O\left(\sqrt{\frac{1}{n}}\right)$ but that Bernstein's inequality implies $\widehat{p}_n - p_n = O_P(1/n)$.

4. Show that $X_n = o_P(a_n)$ and $Y_n = O_P(b_n)$ implies that $X_n Y_n = o_P(a_n b_n)$.

Homework 3
36-705
Due: Thursday Sept 22 by 3:00

1. Let $\mathcal{A}$ be a class of sets. Let $\mathcal{B} = \{A^c : \ A \in \mathcal{A}\}$. Show that $s_n(\mathcal{B}) = s_n(\mathcal{A})$.

2. Let Let $\mathcal{A}$ and $\mathcal{B}$ be classes of sets. Let

$$\mathcal{C} = \left\{ A \bigcap B : \ A \in \mathcal{A}, B \in \mathcal{B} \right\}.$$

   Show that

$$s_n(\mathcal{C}) \leq s_n(\mathcal{A}) s_n(\mathcal{B}).$$

3. Show that $s_{n+m}(\mathcal{A}) \leq s_n(\mathcal{A}) s_m(\mathcal{A})$.

4. Let

$$\mathcal{A} = \left\{ A = [a,b] \cup [c,d] : \ a,b,c,d \in \mathbb{R}, \quad a \leq b \leq c \leq d \right\}.$$

   Find VC dimension of $\mathcal{A}$.

Homework 4
36-705
Due: Thursday September 29 by 3:00

1. 5.33
2. 5.34
3. 5.35
4. 5.36
5. 5.39

Homework 5
36-705
Due: Thursday October 6 by 3:00

1. 6.2

2. 6.4

3. 6.9 (b) and (e).

4. Write $(x_1, \ldots, x_n) \sim (y_1, \ldots, y_n)$ to mean that the likelihood function based on $(x_1, \ldots, x_n)$ is proportional to the likelihood function based on $(y_1, \ldots, y_n)$. The equivalence relation $\sim$ induces a partition $\Pi$ of the sample space: $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ are in the same element of the partition if and only if $(x_1, \ldots, x_n) \sim (y_1, \ldots, y_n)$. Show that $\Pi$ is a minimal sufficient partition.

5. 7.1

6. 7.5 (a).

7. 7.8.

8. 7.9.

9. In class, we found the minimax estimator for the Bernoulli. Here, you will fill in the details. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Let $L(p, \widehat{p}) = (p - \widehat{p})^2$.

   (a) Let $\widehat{p}$ be the Bayes estimator using a $\text{Beta}(\alpha, \beta)$ prior. Find the Bayes estimator.

   (b) Compute the risk function.

   (c) Compute the Bayes risk.

   (d) Find $\alpha$ and $\beta$ to make the risk constant and hence find the minimax estimator.

Homework 6
36-705
Due: Thursday October 20 by 3:00

1. 10.1
2. 10.2
3. 10.4
4. 10.18
5. 10.19

Homework 7
36-705
Due: Thursday October 27 by 3:00

1. 8.13 (a,b)
2. 8.14
3. 8.15
4. 8.17
5. 8.20
6. 10.31 (a,b,c,e)
7. Show that, when $H_0$ is true, then the p-value has a Uniform (0,1) distribution.

Homework 8
36-705
Due: Thursday November 10 2010 by 3:00

1. 9.1.

2. 9.4(a)

3. 9.33(a)

4. Let $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$. Find the $1-\alpha$ likelihood ratio confidence interval for $\theta$. Note: the limiting $\chi^2$ theory does not apply to this example. You need to find the cutoff value directly.

5. Let $X_1, \ldots, X_n \sim p$ and assume that $0 \leq X_i \leq 1$. The histogram density estimator is defined as follows. Divide $[0, 1]$ into $m$ bins $B_1 = [0, 1/m], B_2 = (1/m, 2/m], \ldots,$. Let $h = 1/m$ and let $\widehat{\theta}_j = n^{-1} \sum_{i=1}^n I(X_i \in B_j)$. Let

$$\widehat{p}(x) = \frac{\widehat{\theta}_j}{h}$$

when $x \in B_j$. Find the asymptotic MSE. Find the best $h$. Find the rate of convergence of the estimator.

Homework 9
10/36-705
Due: Thursday Nov 17 by 3:00

1. Let $X_1, \ldots, X_n \sim p$ and let $\widehat{p}_h$ denote the kernel density estimator with bandwidth $h$. Let $R(h) = \mathbb{E}[L(h)]$ denote the risk, where

$$L(h) = \int (\widehat{p}_h(x) - p(x))^2 dx.$$

(a) Define $\widetilde{R}(h) = \mathbb{E}[\widetilde{L}(h)]$ where

$$\widetilde{L}(h) = \int (\widehat{p}_h(x))^2 dx - 2\int \widehat{p}_h(x)p(x)dx.$$

Show that minimizing $\widetilde{R}(h)$ over $h$ is equivalent to minimizing $R(h)$.

(b) Let $Y_1, \ldots, Y_n$ be a second sample from $p$. Define

$$\widehat{R}(h) = \int (\widehat{p}_h(x))^2 dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{p}_h(Y_i)$$

where $\widehat{p}_h$ is still based on $X_1, \ldots, X_n$. Show that $\mathbb{E}\widehat{R}(h) = \widetilde{R}(h)$. (Hence, $\widehat{R}(h)$ can be used as an estimate of the risk.)

2. Again, let $\widehat{p}_h$ denote the kernel density estimator. Use Hoeffding's inequality to find a bound on $\mathbb{P}(|\widehat{p}_h(x) - p_h(x)| > \epsilon)$ where $p_h(x) = \mathbb{E}(\widehat{p}_h(x))$.

3. Let $X_1, \ldots, X_n \sim$ Bernoulli$(\theta)$. Let $\widehat{\theta}_n = n^{-1}\sum_{i=1}^n X_i$. Let $X_1^*, \ldots, X_n^*$ denote a bootstrap sample. Let $\widehat{\theta}^* = n^{-1}\sum_{i=1}^n X_i^*$. Find the following four quantities:

$$\mathbb{E}(\widehat{\theta}^*|X_1, \ldots, X_n), \quad \mathbb{E}(\widehat{\theta}^*), \quad \mathbb{V}(\widehat{\theta}^*|X_1, \ldots, X_n), \quad \mathbb{V}(\widehat{\theta}^*).$$

4. The bootstrap estimate of $\mathsf{Var}(\widehat{\theta}_n)$ is $\mathbb{V}(\widehat{\theta}^*|X_1, \ldots, X_n)$. (In other words, when $B \to \infty$, the bootstrap estimate of variance converges to $\mathbb{V}(\widehat{\theta}^*|X_1, \ldots, X_n)$.) Show that the bootstrap is consistent, in the sense that

$$\frac{\mathbb{V}(\widehat{\theta}^*|X_1, \ldots, X_n)}{\mathsf{Var}(\widehat{\theta}_n)} \xrightarrow{P} 1.$$

Homework 10
36/10-705
Due: Thursday December 1 2010 by 3:00

1. 7.23

2. 9.27

3. Suppose that $V = \sum_{j=1}^{k}(Z_j + \theta_j)^2$ where $Z_1, \ldots, Z_k$ are independent, standard Normal random variables. We say that $V$ has a non-central $\chi^2$ distribution with non-centrality parameter $\lambda = \sum_j \theta_j^2$ and $k$ degrees of freedom. We write $V \sim \chi^2_k(\lambda)$.

   (a) Show that if $V \sim \chi^2_k(\lambda)$ then $\mathbb{E}(V) = k + \lambda$ and $\mathrm{Var}(V) = 2(k + 2\lambda)$.

   (b) Let $Y_i \sim N(\mu_i, 1)$ for $i = 1, \ldots, n$. Find the posterior distribution of $\mu = (\mu_1, \ldots, \mu_k)$ using a flat prior.

   (c) Find the posterior distribution of $\tau = \sum_i \mu_i^2$.

   (d) Find the mean $\widehat{\tau}$ of the posterior.

   (e) Find the bias and variance of $\widehat{\tau}$.

   (f) Show that $\widehat{\tau}$ is not a consistent estimator of $\tau$. (Technically, the parameter $\tau$ is changing with $n$. You may assume that $\tau$ is bounded as $n$ increases.) Hint: you may use the fact that if $V \sim \chi^2_k(\lambda)$, then $(V - \mathbb{E}(V))/\sqrt{\mathsf{Var}(V)} \approx N(0, 1)$.

   (g) Find $c_n$ so that $P(\tau \in C_n | X_1, \ldots, X_n) = 1 - \alpha$ where $C_n = [c_n, \infty)$.

   (h) Construct an unbiased estimator of $\tau$. Compare this to the Bayes estimator.

   (i) Find a (frequentist) confidence interval $A_n = [a_n, \infty)$ such that $\mathbb{P}_\mu(\tau \in A_n) = 1 - \alpha$ for all $\mu$. Compare this to the Bayes posterior interval $C_n$.

# 2011 Fall 10-705 Homework 1 Solutions

1.47 All of the functions are continuous, hence right-continuous. Thus we only need to check the limit, and that they are nondecreasing

a. $\lim_{x\to-\infty} \frac{1}{2}+\frac{1}{\pi}\tan^{-1}(x) = \frac{1}{2}+\frac{1}{\pi}\left(\frac{-\pi}{2}\right) = 0$, $\lim_{x\to\infty} \frac{1}{2}+\frac{1}{\pi}\tan^{-1}(x) = \frac{1}{2}+\frac{1}{\pi}\left(\frac{\pi}{2}\right) = 1$, and $\frac{d}{dx}\left(\frac{1}{2}+\frac{1}{\pi}\tan^{-1}(x)\right) = \frac{1}{1+x^2} > 0$, so $F(x)$ is increasing.

b. See Example 1.5.5.

c. $\lim_{x\to-\infty} e^{-e^{-x}} = 0$, $\lim_{x\to\infty} e^{-e^{-x}} = 1$, $\frac{d}{dx}e^{-e^{-x}} = e^{-x}e^{-e^{-x}} > 0$.

d. $\lim_{x\to-\infty}(1-e^{-x}) = 0$, $\lim_{x\to\infty}(1-e^{-x}) = 1$, $\frac{d}{dx}(1-e^{-x}) = e^{-x} > 0$.

e. $\lim_{y\to-\infty} \frac{1-\epsilon}{1+e^{-y}} = 0$, $\lim_{y\to\infty} \epsilon+\frac{1-\epsilon}{1+e^{-y}} = 1$, $\frac{d}{dx}(\frac{1-\epsilon}{1+e^{-y}}) = \frac{(1-\epsilon)e^{-y}}{(1+e^{-y})^2} > 0$ and $\frac{d}{dx}(\epsilon+\frac{1-\epsilon}{1+e^{-y}}) > 0$, $F_Y(y)$ is continuous except on $y=0$ where $\lim_{y\downarrow 0}(\epsilon+\frac{1-\epsilon}{1+e^{-y}}) = F(0)$. Thus is $F_Y(y)$ right continuous.

1.49 For every $t$, $F_X(t) \le F_Y(t)$. Thus we have

$$P(X > t) = 1 - P(X \le t) = 1 - F_X(t) \ge 1 - F_Y(t) = 1 - P(Y \le t) = P(Y > t).$$

And for some $t^*$, $F_X(t^*) < F_Y(t^*)$. Then we have that

$$P(X > t^*) = 1 - P(X \le t^*) = 1 - F_X(t^*) > 1 - F_Y(t^*) = 1 - P(Y \le t^*) = P(Y > t^*).$$

2.1 a. $f_x(x) = 42x^5(1-x)$, $0 < x < 1$; $y = x^3 = g(x)$, monotone, and $\mathcal{Y} = (0,1)$. Use Theorem 2.1.5.

$$\begin{aligned} f_Y(y) &= f_x(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right| = f_x(y^{1/3})\frac{d}{dy}(y^{1/3}) = 42y^{5/3}(1-y^{1/3})(\frac{1}{3}y^{-2/3}) \\ &= 14y(1-y^{1/3}) = 14y - 14y^{4/3}, \quad 0 < y < 1. \end{aligned}$$

To check the integral,

$$\int_0^1 (14y - 14y^{4/3})dy = 7y^2 - 14\frac{y^{7/3}}{7/3}\Big|_0^1 = 7y^2 - 6y^{7/3}\Big|_0^1 = 1 - 0 = 1.$$

b. $f_x(x) = 7e^{-7x}$, $0 < x < \infty$, $y = 4x+3$, monotone, and $\mathcal{Y} = (3,\infty)$. Use Theorem 2.1.5.

$$f_Y(y) = f_x(\frac{y-3}{4})\left|\frac{d}{dy}(\frac{y-3}{4})\right| = 7e^{-(7/4)(y-3)}\left|\frac{1}{4}\right| = \frac{7}{4}e^{-(7/4)(y-3)}, \ 3 < y < \infty.$$

To check the integral,

$$\int_3^\infty \frac{7}{4}e^{-(7/4)(y-3)}dy = -e^{-(7/4)(y-3)}\Big|_3^\infty = 0 - (-1) = 1.$$

c. $F_Y(y) = P(0 \le X \le \sqrt{y}) = F_X(\sqrt{y})$. Then $f_Y(y) = \frac{1}{2\sqrt{y}} f_X(\sqrt{y})$. Therefore

$$f_Y(y) = \frac{1}{2\sqrt{y}} 30(\sqrt{y})^2(1-\sqrt{y})^2 = 15y^{\frac{1}{2}}(1-\sqrt{y})^2, \quad 0 < y < 1.$$

To check the integral,

$$\int_0^1 15y^{\frac{1}{2}}(1-\sqrt{y})^2 dy = \int_0^1 (15y^{\frac{1}{2}} - 30y + 15y^{\frac{3}{2}})dy = 15(\frac{2}{3}) - 30(\frac{1}{2}) + 15(\frac{2}{5}) = 1.$$

2.3 $P(Y = y) = P(\frac{X}{X+1} = y) = P(X = \frac{y}{1-y}) = \frac{1}{3}(\frac{2}{3})^{y/(1-y)}$, where $y = 0, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \ldots, \frac{x}{x+1}, \ldots$ .

2.7 a.

$$P(Y < y) = P(X^2 \le y) = \begin{cases} P(-\sqrt{y} \le X \le \sqrt{y}) & y \le 1 \\ P(-1 \le X \le \sqrt{y}) & y > 1 \end{cases} = \begin{cases} \int_{-\sqrt{y}}^{\sqrt{y}} f_X(x)dx & y \le 1 \\ \int_{-1}^{\sqrt{y}} f_X(x)dx & y > 1 \end{cases}$$

Differentiation gives

$$f_Y(y) = \begin{cases} \dfrac{2}{9\sqrt{y}} & y \le 1 \\ \dfrac{1}{9}\left(1 + \dfrac{1}{\sqrt{y}}\right) & y > 1 \end{cases}$$

2.15 Assume without loss of generality that $X \le Y$. Then $X \vee Y = Y$ and $X \wedge Y = X$. Thus $X + Y = (X \wedge Y) + (X \vee Y)$. Taking expectations

$$\mathrm{E}[X+Y] = \mathrm{E}[(X \wedge Y) + (X \vee Y)] = \mathrm{E}(X \wedge Y) + \mathrm{E}(X \vee Y).$$

Therefore $\mathrm{E}(X \vee Y) = \mathrm{E}X + \mathrm{E}Y - \mathrm{E}(X \wedge Y)$.

2.30 a. $\mathrm{E}(e^{tX}) = \int_0^c e^{tx}\frac{1}{c}dx = \frac{1}{ct}e^{tx}\Big|_0^c = \frac{1}{ct}e^{tc} - \frac{1}{ct}1 = \frac{1}{ct}(e^{tc}-1)$.

b. $\mathrm{E}(e^{tX}) = \int_0^c \frac{2x}{c^2}e^{tx}dx = \frac{2}{c^2t^2}(cte^{tc} - e^{tc} + 1)$. (integration-by-parts)

c.

$$\begin{aligned}\mathrm{E}(e^{tx}) &= \int_{-\infty}^{\alpha}\frac{1}{2\beta}e^{(x-\alpha)/\beta}e^{tx}dx + \int_{\alpha}^{\infty}\frac{1}{2\beta}e^{-(x-\alpha)/\beta}e^{tx}dx \\ &= \frac{e^{-\alpha/\beta}}{2\beta}\frac{1}{(\frac{1}{\beta}+t)}e^{x(\frac{1}{\beta}+t)}\Big|_{-\infty}^{\alpha} + -\frac{e^{\alpha/\beta}}{2\beta}\frac{1}{(\frac{1}{\beta}-t)}e^{-x(\frac{1}{\beta}-t)}\Big|_{\alpha}^{\infty} \\ &= \frac{4e^{\alpha t}}{4-\beta^2t^2}, \quad -2/\beta < t < 2/\beta.\end{aligned}$$

d. $\mathrm{E}\left(e^{tX}\right) = \sum_{x=0}^{\infty}e^{tx}\binom{r+x-1}{x}p^r(1-p)^x = p^r\sum_{x=0}^{\infty}\binom{r+x-1}{x}\left((1-p)e^t\right)^x$. Now use the fact that $\sum_{x=0}^{\infty}\binom{r+x-1}{x}\left((1-p)e^t\right)^x\left(1-(1-p)e^t\right)^r = 1$ for $(1-p)e^t < 1$, since this is just the sum of this pmf, to get $\mathrm{E}(e^{tX}) = \left(\frac{p}{1-(1-p)e^t}\right)^r$, $t < -\log(1-p)$.

3.32 a.

Note that $c^{*-1}(\eta) = \int h(x)\exp\left(\sum_i \eta_i t_i(x)\right)dx$. We have

$$\begin{aligned}&-\frac{\partial}{\partial\eta_j}\log c^*(\eta) = \frac{\partial}{\partial\eta_j}\log c^{*-1}(\eta) \\ &= c^*(\eta)\frac{\partial}{\partial\eta_j}\int h(x)\exp\left(\sum_i \eta_i t_i(x)\right)dx && (3.31(a)) \\ &= c^*(\eta)\int h(x)\frac{\partial}{\partial\eta_j}\exp\left(\sum_i \eta_i t_i(x)\right)dx && (\text{interchange integration and differentiation}) \\ &= c^*(\eta)\int h(x)\exp\left(\eta_j t_j(x)\right)t_j(x)dx \\ &= \int t_j(x)h(x)c^*(\eta)\exp\left(\eta_j t_j(x)\right)dx \\ &= \mathbb{E}\left(t_j(x)\right)\end{aligned}$$

and

$$-\frac{\partial^2}{\partial \eta_j^{\,2}}\log c^*(\eta)=\frac{\partial^2}{\partial \eta_j^{\,2}}\log c^{*-1}(\eta)$$

$$=c^*(\eta)\frac{\partial^2}{\partial \eta_j^{\,2}}c^{*-1}(\eta)-\left(\frac{\partial}{\partial \eta_j}\log c^{*-1}(\eta)\right)^2 \qquad (3.31\ (b))$$

$$=c^*(\eta)\int h(x)\frac{\partial^2}{\partial \eta_j^{\,2}}\exp\left(\sum_i \eta_i t_i(x)\right)dx-\mathbb{E}^2\left(t_j(x)\right)$$

$$=\int t_j^{\,2}(x)h(x)c^*(\eta)\,dx-\mathbb{E}^2\left(t_j(x)\right)$$

$$=\mathbb{E}\left(t_j^{\,2}(x)\right)-\mathbb{E}^2\left(t_j(x)\right)=Var\left(t_j(x)\right)$$

3.32 b.

$$\Gamma(x;\alpha,\beta)=x^{\alpha-1}\frac{\exp(-x/\beta)}{\beta^\alpha\Gamma(\alpha)}=\frac{1}{\beta^\alpha\Gamma(\alpha)}\exp\left(-\frac{1}{\beta}x+(\alpha-1)\ln x\right)$$

The natural parameters and sufficient statistics are $\eta=[-1/\beta,\alpha-1], t(x)=[x,\log x]$. Further,

$$-\log c^*(\eta)=\alpha\log\beta+\log\Gamma(\alpha)=\alpha\log(-1/\eta_1)+\log\Gamma(\alpha)$$

Therefore

$$\mathbb{E}(x)=-\frac{\partial}{\partial \eta_1}\log c^*(\eta)=\alpha\frac{\partial}{\partial \eta_1}\log(-1/\eta_1)=-\frac{\alpha}{\eta_1}=\alpha\beta$$

4.4 a. $\int_0^1\int_0^2 C(x+2y)dxdy=4C=1$, thus $C=\frac{1}{4}$.

b. $f_X(x)=\begin{cases}\int_0^1 \frac{1}{4}(x+2y)dy=\frac{1}{4}(x+1) & 0<x<2\\ 0 & \text{otherwise}\end{cases}$

c. $F_{XY}(x,y)=P(X\le x,Y\le y)=\int_{-\infty}^x\int_{-\infty}^y f(v,u)dvdu$. The way this integral is calculated depends on the values of $x$ and $y$. For example, for $0<x<2$ and $0<y<1$,

$$F_{XY}(x,y)=\int_{-\infty}^x\int_{-\infty}^y f(u,v)dvdu=\int_0^x\int_0^y\frac{1}{4}(u+2v)dvdu=\frac{x^2y}{8}+\frac{y^2x}{4}.$$

But for $0<x<2$ and $1\le y$,

$$F_{XY}(x,y)=\int_{-\infty}^x\int_{-\infty}^y f(u,v)dvdu=\int_0^x\int_0^1\frac{1}{4}(u+2v)dvdu=\frac{x^2}{8}+\frac{x}{4}.$$

The complete definition of $F_{XY}$ is

$$F_{XY}(x,y) = \begin{cases} 0 & x \le 0 \text{ or } y \le 0 \\ x^2y/8 + y^2x/4 & 0 < x < 2 \text{ and } 0 < y < 1 \\ y/2 + y^2/2 & 2 \le x \text{ and } 0 < y < 1 \\ x^2/8 + x/4 & 0 < x < 2 \text{ and } 1 \le y \\ 1 & 2 \le x \text{ and } 1 \le y \end{cases}.$$

d. The function $z = g(x) = 9/(x+1)^2$ is monotone on $0 < x < 2$, so use Theorem 2.1.5 to obtain $f_Z(z) = 9/(8z^2)$, $1 < z < 9$.

4.5 a. $P(X > \sqrt{Y}) = \int_0^1 \int_{\sqrt{y}}^1 (x+y)dxdy = \frac{7}{20}$.

b. $P(X^2 < Y < X) = \int_0^1 \int_y^{\sqrt{y}} 2xdxdy = \frac{1}{6}$.

# 36-705 Intermediate Statistics HW2

## Problem 1

As $X \geq 0$ with Prob 1, we have

$$E[X] = \int_0^\infty (1 - F(t))dt = \int_0^\infty P(X > t)dt = \int_0^a P(X > t)dt + \int_a^\infty P(X > t)dt$$

With $P(X > t) \leq 1$ and $P(X > t) \leq (1/t)^k$, we have the upper bound of $E[X]$,

$$E[X] = \int_0^a P(X > t)dt + \int_a^\infty P(X > t)dt \leq \int_0^a 1dt + \int_a^\infty (1/t)^k dt = a + \frac{1}{(k-1)a^{k-1}}.$$

Set the derivative of it to be 0, we get $1 + 1/a^k = 0$, so we get $a = 1$. The upper bound of $E[X]$ is $1 + \frac{1}{k-1} = \frac{k}{k-1}$.

## Problem 2

The cumulative density function of $Y$ is

$$P(Y \leq y) = P(X_i \leq y, 1 \leq i \leq n) = \prod_{i=1}^n P(X_i \leq y) = y^n, \quad 0 \leq y \leq 1.$$

So, the expected value of $Y$ is

$$E[Y] = \int_0^1 P(Y > t)dt = \int_0^1 (1 - t^n)dt = 1 - \frac{1}{n+1} = \frac{n}{n+1}.$$

## Problem 3

(i) Note that $p_n = P(X_i \in A_n) = 1/n$. Let $Y_i = I_{A_n}(X_i)$, then

$$E[Y_i] = E[I_{A_n}(X_i)] = P(X_i \in A_n) = 1/n,$$

$$\bar{Y_n} = \frac{1}{n}\sum_{i=1}^n Y_i = \frac{1}{n}\sum_{i=1}^n I_{A_n}(X_i) = \hat{p}_n,$$

therefore:

1. Hoeffding's inequality: $Y_i = 0$ or 1, thus the bound is $0 \leq Y_i \leq 1$, and finally

$$P(|\hat{p}_n - p_n| \geq \epsilon) = P(|\bar{Y_n} - E[Y]| \geq \epsilon) \leq 2\exp\{-\frac{2n\epsilon^2}{(1-0)^2}\} = 2e^{-2n\epsilon^2};$$

2. Bernstein's inequality: still $0 \leq Y_i \leq 1$, hence $|Y_i| \leq 1$, and the variance is

$$Var(Y_i) = 1/n(1-1/n) = \frac{n-1}{n^2},$$

as $Y_i \sim Bernoulli(1/n)$.

So, we have

$$P(|\hat{p}_n - p_n| \geq \epsilon) = P(|\bar{Y_n} - E[Y]| \geq \epsilon) \leq 2\exp\{-\frac{n\epsilon^2}{2(n-1)/n^2 + 2\epsilon/3}\}.$$

(ii) When $\epsilon$ is small and $n$ is large, $2(n-1)/n^2 + 2\epsilon/3$ will be very small, in the order of $1/n$, so $2(n-1)/n^2 + 2\epsilon/3 < 1/2$, and so we have

$$2\exp\{-\frac{n\epsilon^2}{2(n-1)/n^2 + 2\epsilon/3}\} \leq 2e^{-2n\epsilon^2}.$$

Therefore, Bernstein's inequality is tighter than Hoeffding's inequality.

(iii) Use Hoeffding's inequality,

$$P(\frac{|\hat{p}_n - p_n|}{1/\sqrt{n}} \geq C) = P(|\hat{p}_n - p_n| \geq C/\sqrt{n}) \leq 2e^{-2n(C/\sqrt{n})^2} = 2e^{-2C^2}.$$

So, for any $\delta$, when $C$ is large enough, there is $P(\frac{\hat{p}_n - p_n}{1/\sqrt{n}} \geq C) \leq \delta$, therefore, Hoeffiding's inequality implies $\hat{p}_n - p_n = O_p(1/\sqrt{n})$.

Use Bernstein's inequality, we have

$$P(\frac{|\hat{p}_n - p_n|}{1/n} \geq C) = P(|\hat{p}_n - p_n| \geq C/n) \leq 2\exp\{-\frac{nC^2/n^2}{2(n-1)/n^2 + 2C/3n}\}.$$

Simplify the exponential part, we have

$$-\frac{nC^2/n^2}{2(n-1)/n^2 + 2C/3n} = -\frac{C^2/n}{2(n-1)/n^2 + 2C/3n} = -\frac{C^2}{2(n-1)/n + 2C/3} \approx -\frac{3C}{2},$$

for large $n$ and large $C$. So, in all, we have

$$P(\frac{|\hat{p}_n - p_n|}{1/n} \geq C) \leq 2e^{-\frac{3C}{2}}.$$

For any $\delta$, there is $C$ large enough, such that the probability is smaller than $\delta$. So, Bernstein's inequality implies $\hat{p}_n - p_n = O_p(1/n)$.

## Problem 4

$Y_n = O_P(b_n)$ means $\forall \delta > 0$, $\exists C > 0$, s.t. $P(|\frac{Y_n}{b_n}| > C) < \delta$ for all large enough $n$.
$X_n = o_P(a_n)$ means $\forall \delta > 0$, $\forall \varepsilon > 0$, $P(|\frac{X_n}{a_n}| > \varepsilon/C) < \delta$ for all large enough $n$.
Therefore, for all $\delta > 0$, $\varepsilon > 0$ and large enough $n$,

$$\begin{aligned} & P(|\frac{X_n Y_n}{a_n b_n}| > \varepsilon) \leqslant P(|\frac{X_n}{a_n}| > \varepsilon/C \text{ or } |\frac{Y_n}{b_n}| > C) \\ \leqslant \quad & P(|\frac{X_n}{a_n}| > \varepsilon/C) + P(|\frac{Y_n}{b_n}| > C) < 2\delta \end{aligned}$$

1. The first $\leqslant$ is because:

$$|\frac{X_n}{a_n}| \leqslant \varepsilon/C \text{ and } |\frac{Y_n}{b_n}| \leqslant C \Longrightarrow |\frac{X_n Y_n}{a_n b_n}| \leqslant \varepsilon,$$

   taking the reverse negative statement,

$$|\frac{X_n Y_n}{a_n b_n}| > \varepsilon \Longrightarrow |\frac{X_n}{a_n}| > \varepsilon/C \text{ or } |\frac{Y_n}{b_n}| > C;$$

2. The second $\leqslant$ is because $P(A \text{ or } B) \leqslant P(A) + P(B)$.

That is, $P(|\frac{X_n Y_n}{a_n b_n}| > \varepsilon) \longrightarrow 0$ for all $\varepsilon > 0$. Therefore, $X_n Y_n = o_P(a_n b_n)$.

# 36-705 Intermediate Statistics HW3

## Problem 1

Notice that $(A\cap F)\cap(A^c\cap F) = (A\cap A^c)\cap F = \emptyset$. On the other hand, $(A\cap F)\cup(A^c\cap F) = (A\cup A^c)\cap F = \Omega\cap F = F$. This shows $A$ and $A^c$ pick different parts of $F$, that is, $(A^c\cap F) = F\backslash(A\cap F)$. For any finite set F w/ $n$ elements, say the total number of distinct $A\cap F$ is $m_1$, then for every distinct $A\cap F$, there is a corresponding $A^c\in\mathcal{B}$ such that $A^c\cap F$ picks the other part of $F$. Then the total number of distinct $A^c\cap F$ is also $m_1$. So we have $S(\mathcal{A},F) = S(\mathcal{B},F)$, taking "sup" on both sides, we have,

$$s_n(\mathcal{A}) = s_n(\mathcal{B})$$

## Problem 2

$\mathcal{C} = \{A\cap B : A\in\mathcal{A}, B\in\mathcal{B}\}$. Notice that for $C\in\mathcal{C}$, $C\cap F = (A\cap B)\cap F = (A\cap F)\cap(B\cap F)$, therefore, $A\cap F\subseteq C\cap F\subseteq F$, and $B\cap F\subseteq C\cap F\subseteq F$. For any finite set $F$ w/ $n$ elements, say the total number of **distinct** $A\cap F$ is $m_1$ and the total number of **distinct** $B\cap F$ is $m_2$. Then, the total number of **distinct** $C\cap F$ w/ $C = A\cap B$, i.e. the total number of distinct intersections $(A\cap F)\cap(B\cap F)$ is at most $m_1m_2$ (the maximum number of distinct pairs). That is $S(\mathcal{C},F)\leqslant S(\mathcal{A},F)S(\mathcal{B},F)$, taking "sup" on both sides,

$$\begin{aligned} s_n(\mathcal{C}) &\leqslant \sup_{F\in\mathcal{F}_n}[S(\mathcal{A},F)S(\mathcal{B},F)] \\ &\leqslant \sup_{F\in\mathcal{F}_n} S(\mathcal{A},F)\cdot\sup_{F\in\mathcal{F}_n} S(\mathcal{B},F) = s_n(\mathcal{A})s_n(\mathcal{B}). \end{aligned}$$

## Problem 3

Let $F_{n+m} = F_n\cup F_m$ where $F_n$ with $n$ elements and $F_m$ with $m$ elements are disjoint and $F_{n+m}$ have $m+n$ elements. For $A\in\mathcal{A}$, $A\cap F_{n+m} = A\cap(F_n\cup F_m) = (A\cap F_n)\cup(A\cap F_m)$. Therefore, $A\cap F_n\subseteq A\cap F_{n+m}\subseteq F_{n+m}$, and $A\cap F_m\subseteq A\cap F_{n+m}\subseteq F_{n+m}$. For any finite set $F_n$ w/ $n$ elements and $F_m$ w/ $m$ elements, say the total number of **distinct** $A\cap F_n$ is $n_1$ and

the total number of **distinct** $A \cap F_m$ is $m_1$. Then, the total number of **distinct** $A \cap F_{n+m}$ w/ $F_{n+m} = F_n \cup F_m$, which are subsets of distinct unions $(A \cap F_n) \cup (A \cap F_m)$ is at most $n_1 m_1$ (the maximum number of distinct pairs). That is $S(\mathcal{A}, F_{n+m}) \leqslant S(\mathcal{A}, F_n) S(\mathcal{A}, F_m)$, taking "sup" on both sides,

$$\begin{aligned} s_{n+m}(\mathcal{A}) &\leqslant \sup_{F_n \in \mathcal{F}_n, F_m \in \mathcal{F}_m} [S(\mathcal{A}, F_n) S(\mathcal{A}, F_m)] \\ &\leqslant \sup_{F_n \in \mathcal{F}_n} S(\mathcal{A}, F_n) \cdot \sup_{F_m \in \mathcal{F}_m} S(\mathcal{A}, F_m) = s_n(\mathcal{A}) s_m(\mathcal{A}). \end{aligned}$$

## Problem 4

$\mathcal{A}$ is the set of single intervals or joint of two separate intervals on the real line.

1. Let $F_4 = \{-1, 0, 1, 2\}$ with 4 elements. Then

$$\begin{aligned} &1).[-2, -1.5] \cap F = \emptyset, \\ &2).[-1.5, -0.5] \cap F = \{-1\}, \\ &3).[-1.5, -0.5] \cap F = \{-1, 0\}, \\ &4).[-1.5, 1.5] \cap F = \{-1, 0, 1\}, \\ &5).[-1.5, 2.5] \cap F = \{-1, 0, 1, 2\}, \\ &6).[-0.5, 0.5] \cap F = \{0\}, \\ &7).[-0.5, 1.5] \cap F = \{0, 1\}, \\ &8).[-0.5, 2.5] \cap F = \{0, 1, 2\}, \\ &9).[0.5, 1.5] \cap F = \{1\}, \\ &10).[0.5, 2.5] \cap F = \{1, 2\}, \\ &11).[1.5, 2.5] \cap F = \{2\}, \\ &12).[-1.5, -0.5] \cup [0.5, 1.5] \cap F = \{-1, 1\}, \\ &13).[-1.5, 0.5] \cup [1.5, 2.5] \cap F = \{-1, 2\}, \\ &14).[-0.5, 0.5] \cup [1.5, 2.5] \cap F = \{0, 2\}, \\ &15).[-1.5, 0.5] \cup [1.5, 2.5] \cap F = \{-1, 0, 2\}, \\ &16).[-1.5, -0.5] \cup [0.5, 2.5] \cap F = \{-1, 1, 2\} \end{aligned}$$

So $s(\mathcal{A}, F_4) = 2^4 = 16$ and $s_4(\mathcal{A}) = 16$. The VC dimension of $\mathcal{A}$, $d(\mathcal{A}) = max\{n : s_n(\mathcal{A}) = 2^n\} \geq 4$.

2. For set $F_n$, st. $n \geq 5$, eg $F_5 = \{-1, 0, 1, 2, 3\}$, it is impossible $A \cap F_5 = \{-1, 0, 2\}$, since any interval covering $\{-1, 1\}$ will also cover $\{0\}$, similarly, the interval covering $\{0, 2\}$ will also cover $\{1\}$. This is suffice to show that the VC dimension of $\mathcal{A}$ is less than 5. So we have $4 \leq d(\mathcal{A}) < 5$, that is $d(\mathcal{A}) = 4$.

# Test 1 Solutions

## Problem 1

$X_1$ and $X_2$ are iid Unif(0,2), then,

$$f_{X_1,X_2}(x_1,x_2) = f_{X_1}(x_1)f_{X_2}(x_2) = I(x_1 \in (0,2))I(x_2 \in (0,2))$$
$$= \begin{cases} \frac{1}{4} & 0 < x_1 < 2, 0 < x_2 < 2 \\ 0, & ow \end{cases}$$
$$F_Y(y) = P(Y \leq y) = P(X_1 - X_2 \leq y) = P(X_1 \leq X_2 + y) = \int_{\mathcal{A}} f_{X_1,X_2}(x_1,x_2)dx_1dx_2$$

where, $\mathcal{A} = \{(x_1,x_2) \in R^2 : x_1 \leq x_2 + y\}$. Since the integral is over a function which takes value $\frac{1}{4}$ over a square and 0 everywhere else, the value of the integral is equal to $\frac{1}{4}$ of the area of the region determined by the intersection of $\mathcal{A}$ with the square $0 < x_1 < 2, 0 < x_2 < 2$. The four different cases are shown in the last page.

The cdf is,

$$F_Y(y) = \begin{cases} 0 & y \leq -2 \\ \frac{(2+y)^2}{8} & -2 < y < 0 \\ 1 - \frac{(2-y)^2}{8} & 0 \leq y \leq 2 \\ 1 & y > 2 \end{cases}$$

Differentiate it wrt $y$ to get the pdf,

$$f_Y(y) = \frac{dF_Y}{dy} = \begin{cases} \frac{2-y}{4} & 0 \leq y \leq 2 \\ \frac{2+y}{4} & -2 \leq y < 0 \\ 0 & ow \end{cases}$$

## Problem 2

Let $X_i \sim^{iid} Bernoulli(p)$ for $i = 1, 2, ..., n$. Then $X = \sum_{i=1}^n X_i$ has $Binomial(n,p)$ distribution and the MGF of $X$, $M_X = \prod_{i=1}^n M_{X_i} = (M_{X_i})^n$. We know the MGF of Bernoulli distribution is

$$M_{X_1} = E[e^{tX_1}] = e^t p + e^0(1-p)$$

Then we have

$$M_X = (M_{X_1})^n = (e^t p + (1-p))^n$$

## Problem 3

We know that,

$$E(g(X)|Y) = \int g(x)p(x|y)dx$$

Then,

$$\begin{aligned}
&E(E(g(X)|Y)) = \int E(g(X)|Y)p(y)dy \\
&= \int [\int g(x)p(x|y)dx]p(y)dy \\
&= \int \int g(x)p(x|y)p(y)dydx \\
&= \int g(x)[\int p(x,y)dy]dx \\
&= \int g(x)p(x)dx \\
&= E(g(X))
\end{aligned}$$

## Problem 4

We know that $X \sim Unif(-2,1)$ and $Y = e^{|X|}$, then

$$Y = \begin{cases} e^X & 0 \le x \le 1 \\ e^{-X} & -2 \le x < 0 \end{cases}$$

and $1 \le y \le e^2$. The attached figure shows how the function looks like.

The cdf is

$$F_Y(y) = P(Y \le y) = P(e^{|X|} \le y) = \begin{cases} 0 & y < 1 \\ P(-\log(y) \le x \le \log(y)) = \int_{-\log y}^{\log y} \frac{1}{3}dx = \frac{2}{3}\log y & 1 \le y \le e \\ P(-\log(y) \le x \le 1) = \int_{-\log y}^{1} \frac{1}{3}dx = \frac{\log y + 1}{3} & e \le y < e^2 \\ 1 & y \ge e^2 \end{cases}$$

Differentiate the cdf with respect to $y$, we get the pdf,

$$p_Y(y) = \begin{cases} \frac{2}{3y} & 1 \le y \le e \\ \frac{1}{3y} & e \le y < e^2 \\ 0 & ow \end{cases}$$

# Intermediate Statistics HW4

5.33

Since $\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \lim_{x\to-\infty} F_X(x) = 0$, for any $\varepsilon$, we can find an m and an $N_1$ such that $P(X_n > -m) > 1-\varepsilon/2$ for n> $N_1$. Then, since $\lim_{n\to\infty} P(Y_n > c+m) = 1$, we can find an $N_2$ such that $P(Y_n > c+m) > 1-\varepsilon/2$ for n> $N_2$.

Note that $P(A\cap B)+1 \geq P(A)+P(B)$, then
$P(X_n+Y_n > c) \geq P(X_n > -m, Y_n > c+m) \geq P(X_n > -m)+P(Y_n > c+m)-1$
$=1-\varepsilon/2+1-\varepsilon/2-1=1-\varepsilon$

for $n > \max(N_1, N_2)$. Thus $\lim_{n\to\infty} P(X_n+Y_n > c) = 1$.

5.34 Using $\mathrm{E}\bar{X}_n = \mu$ and $\mathrm{Var}\bar{X}_n = \sigma^2/n$, we obtain

$$\mathrm{E}\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma} = \frac{\sqrt{n}}{\sigma}\mathrm{E}(\bar{X}_n-\mu) = \frac{\sqrt{n}}{\sigma}(\mu-\mu) = 0.$$

$$\mathrm{Var}\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma} = \frac{n}{\sigma^2}\mathrm{Var}(\bar{X}_n-\mu) = \frac{n}{\sigma^2}\mathrm{Var}\bar{X} = \frac{n}{\sigma^2}\frac{\sigma^2}{n} = 1.$$

5.35 a. $X_i \sim$ exponential(1). $\mu_X = 1$, $\mathrm{Var}X = 1$. From the CLT, $\bar{X}_n$ is approximately $\mathrm{n}(1, 1/n)$. So

$$\frac{\bar{X}_n-1}{1/\sqrt{n}} \to Z \sim \mathrm{n}(0,1) \quad \text{and} \quad P\left(\frac{\bar{X}_n-1}{1/\sqrt{n}} \leq x\right) \to P(Z \leq x).$$

b.

$$\frac{d}{dx}P(Z \leq x) = \frac{d}{dx}F_Z(x) = f_Z(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$

$$\begin{aligned}
&\frac{d}{dx}P\left(\frac{\bar{X}_n-1}{1/\sqrt{n}} \leq x\right) \\
&= \frac{d}{dx}\left(\sum_{i=1}^n X_i \leq x\sqrt{n}+n\right) \qquad \left(W = \sum_{i=1}^n X_i \sim \mathrm{gamma}(n,1)\right) \\
&= \frac{d}{dx}F_W(x\sqrt{n}+n) = f_W(x\sqrt{n}+n)\cdot\sqrt{n} \quad = \quad \frac{1}{\Gamma(n)}(x\sqrt{n}+n)^{n-1}e^{-(x\sqrt{n}+n)}\sqrt{n}.
\end{aligned}$$

Therefore, $(1/\Gamma(n))(x\sqrt{n}+n)^{n-1}e^{-(x\sqrt{n}+n)}\sqrt{n} \approx \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ as $n \to \infty$. Substituting $x = 0$ yields $n! \approx n^{n+1/2}e^{-n}\sqrt{2\pi}$.

5.36 (a)

$$\mu = E(Y) = E\big(E(Y \mid N)\big) = E(2N) = 2E(N) = 2\theta$$
$$\sigma^2 = Var(Y) = E\big(Var(Y \mid N)\big) + Var\big(E(Y \mid N)\big) = E(4N) + Var(2N) = 4\theta + 4\theta = 8\theta$$

(b)

$$M_{(Y-\mu)/\sigma}(t) = E\left(e^{t(Y-\mu)/\sigma}\right) = e^{-t\mu/\sigma} E\left(e^{tY/\sigma}\right)$$
$$= e^{-t\mu/\sigma} E\left(E\left(e^{\frac{t}{\sigma}Y} \mid N\right)\right) \qquad \text{(total expectation)}$$
$$= e^{-t\mu/\sigma} E\left((1-2t/\sigma)^{-N}\right) = e^{-t\mu/\sigma} \sum_{n=0}^{\infty} (1-2t/\sigma)^{-n} \frac{\theta^n e^{-\theta}}{n!}$$
$$= e^{-t\mu/\sigma} e^{-\theta} e^{\frac{\theta}{1-2t/\sigma}} \sum_{n=0}^{\infty} \frac{\left(\frac{\theta}{1-2t/\sigma}\right)^n}{n!} e^{-\frac{\theta}{1-2t/\sigma}} \qquad \text{Poisson}\left(\frac{\theta}{1-2t/\sigma}\right)$$
$$= e^{-t\frac{\mu}{\sigma} - \theta + \frac{\theta}{1-2t/\sigma}} = e^{-t\frac{2\theta}{\sqrt{8\theta}} - \theta + \frac{\theta}{1-2t/\sqrt{8\theta}}} = e^{-t\frac{2\theta}{\sqrt{8\theta}} + \theta\left(\frac{1}{1-2t/\sqrt{8\theta}} - 1\right)}$$
$$= e^{-t\frac{2\theta}{\sqrt{8\theta}} + \theta\left(1 + \frac{2t}{\sqrt{8\theta}} + \left(\frac{2t}{\sqrt{8\theta}}\right)^2 + \left(\frac{2t}{\sqrt{8\theta}}\right)^3 + \cdots - 1\right)} = e^{-t\frac{2\theta}{\sqrt{8\theta}} + \theta\frac{2t}{\sqrt{8\theta}} + \theta\left(\frac{t^2}{2\theta} + \left(\frac{2t}{\sqrt{8\theta}}\right)^3 + \cdots\right)} \qquad \text{(Taylor expansion)}$$
$$= e^{\frac{t^2}{2} + \left(\theta\left(\frac{2t}{\sqrt{8\theta}}\right)^3 + \cdots\right)} \to e^{\frac{t^2}{2}} \text{ as } \theta \to \infty$$

which is the mgf of N(0,1).

5.39 a. If $h$ is continuous given $\epsilon > 0$ there exits $\delta$ such that $|h(x_n) - h(x)| < \epsilon$ for $|x_n - x| < \delta$. Since $X_1, \ldots, X_n$ converges in probability to the random variable $X$, then $\lim_{n\to\infty} P(|X_n - X| < \delta) = 1$. Thus $\lim_{n\to\infty} P(|h(X_n) - h(X)| < \epsilon) = 1$.

b. Define the subsequence $X_j(s) = s + I_{[a,b]}(s)$ such that in $I_{[a,b]}$, $a$ is always 0, i.e, the subsequence $X_1, X_2, X_4, X_7, \ldots$. For this subsequence

$$X_j(s) \to \begin{cases} s & \text{if } s > 0 \\ s+1 & \text{if } s = 0. \end{cases}$$

# 36-705 Intermediate Statistics HW5

## Problem 1 (C&B 6.2)

By the Factorization Theorem, $T(X) = \min_i(X_i/i)$ is sufficient because the joint pdf is

$$f(x_1,\ldots,x_n|\theta) = \prod_{i=1}^{n} e^{i\theta - x_i} I_{(i\theta,+\infty)}(x_i) = \underbrace{e^{in\theta} I_{(\theta,+\infty)}(T(\mathbf{x}))}_{g(T(\mathbf{x})|\theta)} \cdot \underbrace{e^{-\Sigma_i x_i}}_{h(\mathbf{x})}.$$

Notice, we use the fact that $i > 0$, and the fact that all $x_i$s $> i\theta$ if and only if $\min_i(x_i/i) > \theta$.

## Problem 2 (C&B 6.4)

The joint pdf is

$$\prod_{j=1}^{n}\left\{h(x_j)c(\boldsymbol{\theta})\exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(x_j)\right)\right\} = \underbrace{c(\boldsymbol{\theta})^n \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})\sum_{j=1}^{n} t_i(x_j)\right)}_{g(T(\mathbf{x})|\boldsymbol{\theta})} \cdot \underbrace{\prod_{j=1}^{n} h(x_j)}_{h(\mathbf{x})}.$$

By the Factorization Theorem, $\left(\sum_{j=1}^{n} t_1(X_j),\ldots,\sum_{j=1}^{n} t_k(X_j)\right)$ is a sufficient statistic for $\boldsymbol{\theta}$.

## Problem 3 (C&B 6.9)

(b)

Note, for $X \sim$ location exponential($\theta$), the range depends on the parameter. Now

$$\begin{aligned}\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} &= \frac{\prod_{i=1}^{n}\left(e^{-(x_i-\theta)}I_{(\theta,\infty)}(x_i)\right)}{\prod_{i=1}^{n}\left(e^{-(y_i-\theta)}I_{(\theta,\infty)}(y_i)\right)} \\ &= \frac{e^{n\theta}e^{-\Sigma_i x_i}\prod_{i=1}^{n} I_{(\theta,\infty)}(x_i)}{e^{n\theta}e^{-\Sigma_i y_i}\prod_{i=1}^{n} I_{(\theta,\infty)}(y_i)} = \frac{e^{-\Sigma_i x_i} I_{(\theta,\infty)}(\min x_i)}{e^{-\Sigma_i y_i} I_{(\theta,\infty)}(\min y_i)}.\end{aligned}$$

To make the ratio independent of $\theta$ we need the ratio of indicator functions independent of $\theta$. This will be the case if and only if $\min\{x_1,\ldots,x_n\} = \min\{y_1,\ldots,y_n\}$. So $T(\mathbf{X}) = \min\{X_1,\ldots,X_n\}$ is a minimal sufficient statistic.

(e)

Fix sample points $\mathbf{x}$ and $\mathbf{y}$. Define $A(\theta) = \{i : x_i \leq \theta\}$, $B(\theta) = \{i : y_i \leq \theta\}$, $a(\theta) =$ the number of elements in $A(\theta)$ and $b(\theta) =$ the number of elements in $B(\theta)$. Then the function $f(\mathbf{x}|\theta)/f(\mathbf{y}|\theta)$ depends on $\theta$ only through the function

$$\begin{aligned}
&\sum_{i=1}^{n} |\, x_i - \theta \,| - \sum_{i=1}^{n} |\, y_i - \theta \,| \\
&= \sum_{i \in A(\theta)} (\theta - x_i) + \sum_{i \in A(\theta)^c} (x_i - \theta) - \sum_{i \in B(\theta)} (\theta - y_i) - \sum_{i \in B(\theta)^c} (y_i - \theta) \\
&= (a(\theta) - [n - a(\theta)] - b(\theta) + [n - b(\theta)])\theta \\
&\quad + \left( - \sum_{i \in A(\theta)} x_i + \sum_{i \in A(\theta)^c} x_i + \sum_{i \in B(\theta)} y_i - \sum_{i \in B(\theta)^c} y_i \right) \\
&= 2(a(\theta) - b(\theta))\theta + \left( - \sum_{i \in A(\theta)} x_i + \sum_{i \in A(\theta)^c} x_i + \sum_{i \in B(\theta)} y_i - \sum_{i \in B(\theta)^c} y_i \right).
\end{aligned}$$

Consider an interval of $\theta$s that does not contain any $x_i$s or $y_i$s. The second term is constant on such an interval. The first term will be constant, on the interval if and only if $a(\theta) = b(\theta)$. This will be true for all such intervals if and only if the order statistics for $\mathbf{x}$ are the same as the order statistics for $\mathbf{y}$. Therefore, the order statistics are a minimal sufficient statistic.

## Problem 4

Refer to Notes 6 p.2 for the definition of minimal sufficient partition.

Let $\theta$ be the parameter of the distribution and $f$ be the joint pdf.

$$\frac{f(x_1, ..., x_n|\theta)}{f(y_1, ..., y_n|\theta)}$$

is independent of $\theta$ if and only if $(x_1, ..., x_n) \sim (y_1, ..., y_n)$.
Therefore, by C&B Theorem 6.2.13, $\prod$ is a minimal sufficient partition for $\theta$.

## Problem 5 (C&B 7.1)

1. $x = 0$, the likelihood $L(\theta) = \frac{1}{3}I(\theta = 1) + \frac{1}{4}I(\theta = 2) + 0 \cdot I(\theta = 3) = \frac{1}{3}I(\theta = 1) + \frac{1}{4}I(\theta = 2)$, therefore, the MLE $\hat{\theta} = 1$;
2. $x = 1$, $L(\theta) = \frac{1}{3}I(\theta = 1) + \frac{1}{4}I(\theta = 2)$, $\hat{\theta} = 1$;
3. $x = 2$, $L(\theta) = \frac{1}{4}I(\theta = 2) + \frac{1}{4}I(\theta = 3)$, $\hat{\theta} = 2$ or $\hat{\theta} = 3$;
4. $x = 3$, $L(\theta) = \frac{1}{6}I(\theta = 1) + \frac{1}{4}I(\theta = 2) + \frac{1}{2}I(\theta = 2)$, $\hat{\theta} = 3$;
5. $x = 4$, $L(\theta) = \frac{1}{6}I(\theta = 1) + \frac{1}{4}I(\theta = 3)$, $\hat{\theta} = 3$.

Finally,

$$\hat{\theta} = \begin{cases} 1 & X = 0, 1; \\ 2 \text{ or } 3 & X = 2; \\ 3 & X = 3, 4. \end{cases}$$

## Problem 6 (C&B 7.5(a))

The value $\hat{z}$ solves the equation

$$(1-p)^n = \prod_i (1 - x_i z),$$

where $0 \le z \le (\max_i x_i)^{-1}$. Let $\hat{k}$ = greatest integer less than or equal to $1/\hat{z}$. Then from Example 7.2.9, $\hat{k}$ must satisfy

$$[k(1-p)]^n \ge \prod_i (k - x_i) \qquad \text{and} \qquad [(k+1)(1-p)]^n < \prod_i (k + 1 - x_i).$$

Because the right-hand side of the first equation is decreasing in $\hat{z}$, and because $\hat{k} \le 1/\hat{z}$ (so $\hat{z} \le 1/\hat{k}$) and $\hat{k} + 1 > 1/\hat{z}$, $\hat{k}$ must satisfy the two inequalities. Thus $\hat{k}$ is the MLE.

## Problem 7 (C&B 7.8)

a. $\mathrm{E}\, X^2 = \mathrm{Var}\, X + \mu^2 = \sigma^2$. Therefore $X^2$ is an unbiased estimator of $\sigma^2$.

b.

$$\begin{aligned} L(\sigma|\mathbf{x}) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)}. \quad \log L(\sigma|\mathbf{x}) = \log(2\pi)^{-1/2} - \log\sigma - x^2/(2\sigma^2). \\ \frac{\partial \log L}{\partial \sigma} &= -\frac{1}{\sigma} + \frac{x^2}{\sigma^3} \overset{\text{set}}{=} 0 \Rightarrow \hat{\sigma} X^2 = \hat{\sigma}^3 \Rightarrow \hat{\sigma} = \sqrt{X^2} = |X|. \\ \frac{\partial^2 \log L}{\partial \sigma^2} &= \frac{-3x^2\sigma^2}{\sigma^6} + \frac{1}{\sigma^2}, \text{ which is negative at } \hat{\sigma} = |x|. \end{aligned}$$

Thus, $\hat{\sigma} = |x|$ is a local maximum. Because it is the only place where the first derivative is zero, it is also a global maximum.

c. Because $\mathrm{E}\, X = 0$ is known, just equate $\mathrm{E}\, X^2 = \sigma^2 = \frac{1}{n}\sum_{i=1}^1 X_i^2 = X^2 \Rightarrow \hat{\sigma} = |X|$.

## Problem 8 (C&B 7.9)

This is a uniform$(0,\theta)$ model. So $\mathrm{E}\,X = (0+\theta)/2 = \theta/2$. The method of moments estimator is the solution to the equation $\tilde{\theta}/2 = \bar{X}$, that is, $\tilde{\theta} = 2\bar{X}$. Because $\tilde{\theta}$ is a simple function of the sample mean, its mean and variance are easy to calculate. We have

$$\mathrm{E}\,\tilde{\theta} = 2\mathrm{E}\,\bar{X} = 2\mathrm{E}\,X = 2\frac{\theta}{2} = \theta, \quad \text{and} \quad \mathrm{Var}\,\tilde{\theta} = 4\mathrm{Var}\,\bar{X} = 4\frac{\theta^2/12}{n} = \frac{\theta^2}{3n}.$$

The likelihood function is

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\theta} I_{[0,\theta]}(x_i) = \frac{1}{\theta^n} I_{[0,\theta]}(x_{(n)}) I_{[0,\infty)}(x_{(1)}),$$

where $x_{(1)}$ and $x_{(n)}$ are the smallest and largest order statistics. For $\theta \geq x_{(n)}$, $L = 1/\theta^n$, a decreasing function. So for $\theta \geq x_{(n)}$, $L$ is maximized at $\hat{\theta} = x_{(n)}$. $L = 0$ for $\theta < x_{(n)}$. So the overall maximum, the MLE, is $\hat{\theta} = X_{(n)}$. The pdf of $\hat{\theta} = X_{(n)}$ is $nx^{n-1}/\theta^n$, $0 \leq x \leq \theta$. This can be used to calculate

$$\mathrm{E}\,\hat{\theta} = \frac{n}{n+1}\theta, \quad \mathrm{E}\,\hat{\theta}^2 = \frac{n}{n+2}\theta^2 \quad \text{and} \quad \mathrm{Var}\,\hat{\theta} = \frac{n\theta^2}{(n+2)(n+1)^2}.$$

$\tilde{\theta}$ is an unbiased estimator of $\theta$; $\hat{\theta}$ is a biased estimator. If $n$ is large, the bias is not large because $n/(n+1)$ is close to one. But if $n$ is small, the bias is quite large. On the other hand, $\mathrm{Var}\,\hat{\theta} < \mathrm{Var}\,\tilde{\theta}$ for all $\theta$. So, if $n$ is large, $\hat{\theta}$ is probably preferable to $\tilde{\theta}$.

## Problem 9

**(a)** Bayes estimator under square error loss $L(p,\hat{p}) = (p-\hat{p})^2$ is the posterior mean. $X_i \overset{iid}{\sim} Bernoulli(p)$, $p \sim Beta(\alpha,\beta)$ are conjugate, the posterior is $p|\mathbf{X} \sim Beta(\alpha + \sum_i X_i, \beta + n - \sum_i X_i)$. Therefore, Bayes estimator $\hat{p} = \frac{\alpha+\sum_i X_i}{\alpha+\beta+n}$.

**(b)** Risk function for $\hat{p}$

$$\begin{aligned} & R(p,\hat{p}) = E_p[L(p,\hat{p})] = MSE(\hat{p}) \\ = \; & (E[\hat{p}] - p)^2 + V[\hat{p}] \\ = \; & \left(\frac{\alpha + np}{\alpha+\beta+n} - p\right)^2 + \frac{np(1-p)}{(\alpha+\beta+n)^2} \\ = \; & \frac{(\alpha(1-p) - \beta p)^2}{(\alpha+\beta+n)^2} + \frac{np(1-p)}{(\alpha+\beta+n)^2} \end{aligned}$$

**(c)** Bayes risk for $\hat{p}$

$$B(\pi,\hat{p}) = \int R(p,\hat{p})\pi(p)dp$$

$$
\begin{aligned}
&= \frac{1}{(\alpha+\beta+n)^2}\int[(\alpha+\beta)^2(p-\frac{\alpha}{\alpha+\beta})^2+np-np^2]\pi(p)dp \\
&= \frac{1}{(\alpha+\beta+n)^2}[(\alpha+\beta)^2\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}+\frac{n\alpha}{\alpha+\beta}-n(\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}+\frac{\alpha^2}{(\alpha+\beta)^2})] \\
&= \frac{1}{(\alpha+\beta+n)^2}[\frac{\alpha\beta}{\alpha+\beta+1}+\frac{n\alpha}{\alpha+\beta}-\frac{n\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}] \\
&= \frac{1}{(\alpha+\beta+n)^2}[\frac{\alpha\beta}{\alpha+\beta+1}+\frac{n\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}] \\
&= \frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)(\alpha+\beta+n)}
\end{aligned}
$$

**(d)** The risk

$$R(p,\hat{p}) = \frac{(\alpha(1-p)-\beta p)^2}{(\alpha+\beta+n)^2}+\frac{np(1-p)}{(\alpha+\beta+n)^2} = \frac{1}{(\alpha+\beta+n)^2}\{p^2[(\alpha+\beta)^2-n]+p[n-2\alpha(\alpha+\beta)]+\alpha^2\}$$

is a $2^{nd}$ order polynomial of $p$. To make it constant, set

$$\begin{cases} (\alpha+\beta)^2-n=0; \\ n-2\alpha(\alpha+\beta)=0. \end{cases} \Longrightarrow \begin{cases} \alpha=\frac{\sqrt{n}}{2}; \\ \beta=\frac{\sqrt{n}}{2}. \end{cases}$$

Thus $\hat{p}_m = \frac{\alpha+\sum_i X_i}{\alpha+\beta+n} = \frac{\sqrt{n}/2+\sum_i X_i}{\sqrt{n}+n}$ is the minimax estimator.

# 36705 Intermediate Statistics Homework 6 Solutions

## Problem 1 C & B 10.1

10.1 First calculate some moments for this distribution.

$$X = \theta/3, \qquad X^2 = 1/3, \quad \text{Var}X = \frac{1}{3} - \frac{\theta^2}{9}.$$

So $3\bar{X}_n$ is an unbiased estimator of $\theta$ with variance

$$\text{Var}(3\bar{X}_n) = 9(\text{Var}X)/n = (3 - \theta^2)/n \to 0 \text{ as } n \to \infty.$$

So by Theorem 10.1.3, $3\bar{X}_n$ is a consistent estimator of $\theta$.

## Problem 2 C & B 10.2

By theorem 10 in lecture 4,

$W_n \xrightarrow{P} \theta$, $a_n \xrightarrow{P} 1 \Longrightarrow a_n W_n \xrightarrow{P} 1\theta = \theta$.

$b_n \xrightarrow{P} 0 \Longrightarrow a_n W_n + b_n \xrightarrow{P} 1\theta + 0 = \theta$.

## Problem 3 C & B 10.4

a. Write

$$\frac{\sum X_i Y_i}{\sum X_i^2} = \frac{\sum X_i(X_i + \epsilon_i)}{\sum X_i^2} = 1 + \frac{\sum X_i \epsilon_i}{\sum X_i^2}.$$

From normality and independence

$$X_i \epsilon_i = 0, \quad \text{Var} X_i \epsilon_i = \sigma^2(\mu^2 + \tau^2), \qquad X_i^2 = \mu^2 + \tau^2, \quad \text{Var} X_i^2 = 2\tau^2(2\mu^2 + \tau^2),$$

and $\text{Cov}(X_i, X_i\epsilon_i) = 0$. Applying the formulas of xample 5.5.27, the asymptotic mean and variance are

$$E\left(\frac{\sum X_i Y_i}{\sum X_i^2}\right) \approx \beta$$

and

$$Var\left(\frac{\sum X_i Y_i}{\sum X_i^2}\right) \approx \frac{n\sigma^2(\mu^2 + \tau^2)}{[n(\mu^2 + \tau^2)]^2} = \frac{\sigma^2}{n(\mu^2 + \tau^2)}$$

b.

$$\frac{\sum Y_i}{\sum X_i} = \beta + \frac{\sum \epsilon_i}{\sum X_i}$$

with approximate mean $\beta$ and variance $\sigma^2/(n\mu^2)$.

c.

$$\frac{1}{n}\sum \frac{Y_i}{X_i} = \beta + \frac{1}{n}\sum \frac{\epsilon_i}{X_i}$$

with approximate mean $\beta$ and variance $\sigma^2/(n\mu^2)$.

## Problem 4, C & B 10.18

Denote the density of $n(\mu, \sigma^2)$ as $f_1(x)$, then

$$X_i \sim (1 - \delta) f_1(x) + \delta f(x)$$

Let $Y \sim Bernoulli(\delta)$, then $P(Y = 1) = \delta$ and $P(Y = 0) = 1 - \delta$ and,

$$Var(X_i) = E(Var(X_i|Y)) + Var(E(X_i|Y))$$

Note that, $Var(X_i|Y = 1) = \tau^2$ and $Var(X_i|Y = 0) = \sigma^2$,

$$E(Var(X_i|Y)) = \tau^2\delta + \sigma^2(1 - \delta)$$

Also $E(X_i|Y=1)=\theta$, $E(X_i|Y=0)=\mu$ and $E(E(X_i|Y))=\theta\delta+(1-\delta)\mu$,

$$\begin{aligned}
&Var(E(X_i|Y)) = \sum_Y (E(X_i|Y) - E(E(X_i|Y)))^2 P(Y) \\
&= (\theta - \theta\delta - (1-\delta)\mu)^2\delta + (\mu - \delta\theta - (1-\delta)\mu)^2(1-\delta) \\
&= \delta(1-\delta)(\theta-\mu)^2
\end{aligned}$$

By the fact that $X_i$'s are iid,

$$Var(\bar{X}) = Var(\frac{1}{n}\sum_i (X_i)) = \frac{1}{n}(\tau^2\delta + \sigma^2(1-\delta) + \delta(1-\delta)(\theta-\mu)^2)$$

Since the mean and variance of Cauchy distribution do not exist, any contaminate of cauchy distribution will make $(\theta-\mu)^2$ and $\tau^2$ infinite. So $Var(X_i)$ will be infinite.

## Problem 5, C & B 10.19

$X_i \sim n(\theta, \sigma^2)$.
a).

$$\begin{aligned}
&Var(\bar{(X)}) = Var(\frac{1}{n}\sum_i X_i) = \frac{1}{n^2} Var(\sum_i X_i) \\
&= \frac{1}{n^2}(\sum_i Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)) \\
&= \frac{1}{n^2}(n\sigma^2 + 2\frac{n(n-1)}{2}\rho\sigma^2) \\
&= \frac{1}{n}(\sigma^2 + (n-1)\rho\sigma^2)
\end{aligned}$$

So, as $n \to \infty$, $Var(\bar{X}) \nrightarrow 0$.
b).

$$\begin{aligned}
&Var(\bar{X}) = \frac{1}{n^2} Var(\sum_i X_i) \\
&= \frac{1}{n^2}(\sum_i Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)) \\
&= \frac{1}{n^2}(n\sigma^2 + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} Cov(X_i, X_j))
\end{aligned}$$

$$= \frac{1}{n^2}(n\sigma^2 + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n}\rho^{|i-j|}\sigma^2)$$
$$= \frac{1}{n}\sigma^2 + \frac{2\sigma^2}{n^2}\frac{\rho}{1-\rho}(n - \frac{1-\rho^n}{1-\rho})$$

c).

We know

$$Corr(X_1, X_i) = \frac{Cov(X_1, X_i)}{\sqrt{Var(X_1)Var(X_i)}}$$

And since $\delta_i \sim^{iid} n(0,1)$ we can use $\delta_1$ for all $\delta_i$'s,

$$X_2 = \rho X_1 + \delta_1$$
$$X_3 = \rho(\rho X_1 + \delta_1) + \delta_1$$
$$\dots$$
$$X_i = \rho^{i-1}X_1 + \sum_{j=0}^{i-2}\rho^j\delta_1$$

So,

$$Cov(X_1, X_i) = Cov(X_1, \rho^{i-1}X_1 + \sum_{j=0}^{i-2}\rho^j\delta_1)$$
$$= \rho^{i-1}Cov(X_1, X_1) + \sum_{j=1}^{i-2}Cov(X_1, \rho^j\delta_1)$$
$$= \rho^{i-1}Var(X_1)$$
$$= \rho^{i-1}\sigma^2$$

Also,

$$Var(X_i) = \rho^{2(i-1)}Var(X_1) + \sum_{j=0}^{i-2}\rho^{2j}Var(\delta_1)$$
$$= \rho^{2(i-1)}\sigma^2 + \frac{1-\rho^{2(i-1)}}{1-\rho^2}$$
$$= \frac{1}{1-\rho^2}$$

Given $\sigma^2 = \frac{1}{1-\rho^2}$,

$$Corr(X_1, X_i) = \rho^{i-1}$$

# 36-705 Intermediate Statistics Test 2 Solution

(1) Let $X_1, \ldots, X_n \sim$ Bernoulli($\theta$) where $0 < \theta < 1$. Let

$$W_n = \frac{1}{n}\sum_{i=1}^{n} X_i(1 - X_i).$$

(a) Show that there is a number $\mu$ such that $W_n$ converges in probability $\mu$.

**Solution:** As $X_i$ is either 0 or 1, so $X_i(1 - X_i) = 0$ with probability 1. Hence, $W_n$ has point mass probability 1 at 0. So, $E[W_n] = 0$, $Var(X_n) = 0$.

Obviously, if we set $\mu = 0$, we have $P(|W_n - \mu| > \epsilon) = 0$, for any $\epsilon > 0$. Hence $W_n$ converges to $\mu$ in probability.

(b) Find the limiting distribution of $\sqrt{n}(W_n - \mu)$.

**Solution:** $W_n - \mu$ has point mass 1 at 0, so $\sqrt{n}(W_n - \mu)$ also has point mass 1 at 0. The limiting distribution is

$$P(\sqrt{n}(W_n - \mu) = 0) = 1.$$

(2) Let $X_1, \ldots, X_n \sim$ Normal($\theta, 1$).

(a) Let $T = (X_1, \ldots, X_{n-1})$. Show that $T$ is not sufficient.

**Solution:** As $T = (X_1, \ldots, X_{n-1})$, the conditional distribution of $(X_1, \ldots, X_n | T = (t_1, \ldots, t_{n-1}))$ is

$$\frac{f(X_1, \ldots, X_n, T = t)}{f(T)} = \frac{f(X_1 = t_1, \ldots, X_{n-1} = t_{n-1}, X_n)}{f(X_1 = t_1, \ldots, X_{n-1} = t_{n-1})} = \frac{f(X_n)\prod_{i=1}^{n-1} f(t_i)}{\prod_{i=1}^{n-1} f(t_i)} = f(X_n),$$

where

$$f(X_n) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_n - \theta)^2}{2}}.$$

Obviously, the conditional pdf $f(X_1, \ldots, X_n | T = (t_1, \ldots, t_{n-1}))$ depends on $\theta$, so $T$ is not sufficient.

(b) Show that $U = \sum_{i=1}^{n} X_i$ is minimal sufficient.

**Solution:** From the solution above, the ratio between probability is

$$\frac{f(x_1, \ldots, x_n | T)}{f(y_1, \ldots, y_n | T)} = e^{-\frac{\sum_{i=1}^{n}(x_i^2 - y_i^2)}{2} + \theta \sum_{i=1}^{n}(x_i - y_i)}.$$

Obviously, when $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$, the ratio does not depend on $\theta$, which means that $T(X^n) = \sum_{i=1}^n x_i$ is sufficient. To make sure that the ratio does not depend on $\theta$, there must be $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$, so $T(X^n) = \sum_{i=1}^n x_i$ is also minimal.

In all, $T(X^n) = \sum_{i=1}^n x_i$ is minimal sufficient statistic.

(3) Let $X_1, \ldots, X_n$ be drawn from a uniform distribution on the set

$$[0, 1] \bigcup [2,\ 2 + \theta]$$

where $\theta > 0$.

(a) Find the method of moments estimator $\hat{\theta}$ of $\theta$.

**Solution:** The probability density function for $X$ is

$$f_X(x) = \begin{cases} \frac{1}{1+\theta} & 0 \le x \le 1, 2 \le x \le 2 + \theta, \\ 0 & ow. \end{cases}$$

So, the expectation for $X$ is

$$EX = \int_0^1 \frac{x}{1+\theta} dx + \int_2^{2+\theta} \frac{x}{1+\theta} dx = \frac{\theta^2 + 4\theta + 1}{2(1+\theta)}.$$

To find the moment estimator, let

$$E[X] = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X},$$

and we get the solution

$$\hat{\theta}_1 = \bar{X} - 2 + \sqrt{\bar{X}^2 - 2\bar{X} + 3}, \qquad \hat{\theta}_2 = \bar{X} - 2 - \sqrt{\bar{X}^2 - 2\bar{X} + 3}.$$

As $\theta > 0$, only the $\hat{\theta}_1$ is kept, and the estimator is

$$\hat{\theta} = \bar{X} - 2 + \sqrt{\bar{X}^2 - 2\bar{X} + 3}.$$

(b) Find the mean squared error of $\hat{\theta}$.

**Solution:** The form is too complicated, so we use Delta method to find the approximate MSE of $\hat{\theta}$.

In part (a), we have that

$$E[X] = \frac{\theta^2 + 4\theta + 1}{2(1+\theta)}.$$

The variance for $X$ can also be calculated, as

$$Var(X) = E[X^2] - (E[X])^2 = \frac{(\theta+2)^3 - 7}{3(1+\theta)} - (E[X])^2,$$

which ends up as

$$Var(X) = \frac{\theta^4 + 4\theta^3 + 18\theta^2 + 28\theta + 1}{12(1+\theta)^2}.$$

With CLT, we know that

$$\frac{1}{n}\bar{X} \sim N(\frac{\theta^2+4\theta+1}{2(1+\theta)}, \frac{1}{n}\frac{\theta^4 + 4\theta^3 + 18\theta^2 + 28\theta + 1}{12(1+\theta)^2})$$

Let $g(x) = x - 2 + \sqrt{x^2 - 2x + 3}$, so $\hat{\theta} = g(\bar{X})$, with

$$g(\frac{\theta^2+4\theta+1}{2(1+\theta)}) = \theta,$$

$$g'(\frac{\theta^2+4\theta+1}{2(1+\theta)}) = \frac{2(\theta+1)^2}{(1+\theta)^2+2},$$

and the approximate MSE is

$$\begin{aligned} MSE(\hat{\theta}) &= (E[(\theta - \hat{\theta})])^2 + Var(\hat{\theta}) \\ &= (\frac{2(\theta+1)^2}{(1+\theta)^2+2})^2 \frac{1}{n}\frac{\theta^4 + 4\theta^3 + 18\theta^2 + 28\theta + 1}{12(1+\theta)^2} \\ &= \frac{1}{n}(\frac{\theta+1}{(1+\theta)^2+2})^2 \frac{\theta^4 + 4\theta^3 + 18\theta^2 + 28\theta + 1}{3} \end{aligned}$$

(c) Show that $\hat{\theta}$ is consistent.

**Solution:** According to the result in part (b), the mean squared error of $\hat{\theta}$ goes to 0 in the order of $O(1/n)$, so $\hat{\theta} \to \theta$ in probability, which means that $\hat{\theta}$ is consistent.

(4) Let $X_1, \ldots, X_n \sim N(\theta, 1)$. Let $\tau = e^\theta + 1$.

(a) Find the maximum likelihood estimator $\hat{\tau}$ of $\tau$ and show that it is consistent.

**Solution:** The likelihood function for $\theta$ is

$$L(\theta; X^n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}},$$

hence the log-likelihood function is

$$l(\theta; X^n) = -\frac{n}{2}\log 2\pi - \sum_{i=1}^{n}\frac{(x_i-\theta)^2}{2}.$$

Take derivative of $l(\theta)$, we have

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

To find the MLE of $\theta$, let $\frac{\partial}{\partial\theta}l(\theta) = 0$, and the solution is

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

As this is the only solution for the derivative function, so this is global maximum, and MLE for $\theta$ is $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

As MLE of function $g(\theta)$ is function of MLE $g(\hat{\theta})$, so MLE for $\tau$ is

$$\hat{\tau} = e^{\frac{1}{n}\sum_{i=1}^{n} x_i} + 1.$$

Obviously, the distribution of $\hat{\theta}$ is $\hat{\theta} \sim N(\theta, 1/n)$, so $\hat{\theta} \to \theta$ in probability. As $\hat{\tau}$ is a continuous function of $\hat{\theta}$, according to continuous mapping theorem, $\hat{\tau} \to g(\theta) = \tau$ in probability, which means that MLE $\hat{\tau}$ is consistent.

(b) Consider some loss function $L(\tau, \hat{\tau})$. Define what it means for an estimator to be a minimax estimator for $\tau$.

**Solution:** We say $\hat{\tau}$ is a minimax estimator of $\tau$, if for any other estimator $\tilde{\tau}$, there is

$$\sup_{\tau} R(\tau, \hat{\tau}) \leq \sup_{\tau} R(\tau, \tilde{\tau}),$$

where

$$R(\tau, \hat{\tau}) = E_\tau L(\tau, \hat{\tau}) = \int L(\tau, \hat{\tau}(x^n)) f(x^n; \tau) dx^n.$$

(c) Let $\pi$ be a prior for $\theta$. Find the Bayes estimator for $\tau$ under the loss $L(\tau, \hat{\tau}) = (\hat{\tau} - \tau)^2/\tau$.

**Solution:** To find the Bayes estimator, for any $x^n$, we want to choose $\hat{\tau}(x^n)$ to minimize

$$r(\hat{\tau}|x^n) = \int_{-\infty}^{\infty} L(\tau, \hat{\tau}(x^n))\pi(\theta|x^n)d\theta.$$

Introduce Loss function $L(\tau, \hat{\tau}) = \frac{(\hat{\tau}-\tau)^2}{\tau}$ in the equation, and take the derivative of $r(\hat{\tau}|x^n)$ with respect to $\hat{\tau}$, we have

$$\frac{\partial}{\partial \hat{\tau}} r(\hat{\tau}|x^n) = \int_{-\infty}^{\infty} \frac{2(\hat{\tau}-\tau)}{\tau} \pi(\theta|x^n) d\theta.$$

Let $\frac{\partial}{\partial \hat{\tau}} r(\hat{\tau}|x^n) = 0$, the equation is

$$\int_{-\infty}^{\infty} \frac{2(\hat{\tau}-\tau(\theta))}{\tau} \pi(\theta|x^n) d\theta = 0,$$

which is equivalent with

$$\hat{\tau} \int_{-\infty}^{\infty} \frac{1}{\tau} \pi(\theta|x^n) d\theta - \int_{-\infty}^{\infty} \pi(\theta|x^n) d\theta = 0,$$

hence the solution is

$$\hat{\tau}(x^n) = 1/\int_{-\infty}^{\infty} \frac{1}{\tau} \pi(\theta|x^n) d\tau = 1/E[1/\tau|x^n].$$

# 2011 Fall 10-705/36-705 Homework 7 Solutions

8.13 a. The size of $\phi_1$ is $\alpha_1 = P(X_1 > .95|\theta = 0) = .05$. The size of $\phi_2$ is $\alpha_2 = P(X_1+X_2 > C|\theta = 0)$. If $1 \le C \le 2$, this is

$$\alpha_2 = P(X_1 + X_2 > C|\theta = 0) = \int_{C-1}^{1} \int_{C-x_1}^{1} 1\, dx_2\, dx_1 = \frac{(2-C)^2}{2}.$$

Setting this equal to $\alpha$ and solving for $C$ gives $C = 2 - \sqrt{2\alpha}$, and for $\alpha = .05$, we get $C = 2 - \sqrt{.1} \approx 1.68$.

b. For the first test we have the power function

$$\beta_1(\theta) = P_\theta(X_1 > .95) = \begin{cases} 0 & \text{if } \theta \le -.05 \\ \theta + .05 & \text{if } -.05 < \theta \le .95 \\ 1 & \text{if } .95 < \theta. \end{cases}$$

Using the distribution of $Y = X_1 + X_2$, given by

$$f_Y(y|\theta) = \begin{cases} y - 2\theta & \text{if } 2\theta \le y < 2\theta + 1 \\ 2\theta + 2 - y & \text{if } 2\theta{+}1 \le y < 2\theta + 2 \\ 0 & \text{otherwise,} \end{cases}$$

we obtain the power function for the second test as

$$\beta_2(\theta) = P_\theta(Y > C) = \begin{cases} 0 & \text{if } \theta \le (C/2) - 1 \\ (2\theta + 2 - C)^2/2 & \text{if } (C/2) - 1 < \theta \le (C-1)/2 \\ 1 - (C - 2\theta)^2/2 & \text{if } (C-1)/2 < \theta \le C/2 \\ 1 & \text{if } C/2 < \theta. \end{cases}$$

8.14 The CLT tells us that $Z = (\sum_i X_i - np)/\sqrt{np(1-p)}$ is approximately n$(0, 1)$. For a test that rejects $H_0$ when $\sum_i X_i > c$, we need to find $c$ and $n$ to satisfy

$$P\left(Z > \frac{c - n(.49)}{\sqrt{n(.49)(.51)}}\right) = .01 \quad \text{and} \quad P\left(Z > \frac{c - n(.51)}{\sqrt{n(.51)(.49)}}\right) = .99.$$

We thus want

$$\frac{c - n(.49)}{\sqrt{n(.49)(.51)}} = 2.33 \quad \text{and} \quad \frac{c - n(.51)}{\sqrt{n(.51)(.49)}} = -2.33.$$

Solving these equations gives $n = 13{,}567$ and $c = 6{,}783.5$.

8.15 From the Neyman-Pearson lemma the UMP test rejects $H_0$ if

$$\frac{f(x \mid \sigma_1)}{f(x \mid \sigma_0)} = \frac{(2\pi\sigma_1^2)^{-n/2} e^{-\Sigma_i x_i^2/(2\sigma_1^2)}}{(2\pi\sigma_0^2)^{-n/2} e^{-\Sigma_i x_i^2/(2\sigma_0^2)}} = \left(\frac{\sigma_0}{\sigma_1}\right)^n \exp\left\{\frac{1}{2}\sum_i x_i^2\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)\right\} > k$$

for some $k \geq 0$. After some algebra, this is equivalent to rejecting if

$$\sum_i x_i^2 > \frac{2\log\left(k\left(\sigma_1/\sigma_0\right)^n\right)}{\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)} = c \quad \left(\text{because } \frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2} > 0\right).$$

This is the UMP test of size $\alpha$, where $\alpha = P_{\sigma_0}(\sum_i X_i^2 > c)$. To determine $c$ to obtain a specified $\alpha$, use the fact that $\sum_i X_i^2/\sigma_0^2 \sim \chi_n^2$. Thus

$$\alpha = P_{\sigma_0}\left(\sum_i X_i^2/\sigma_0^2 > c/\sigma_0^2\right) = P\left(\chi_n^2 > c/\sigma_0^2\right),$$

so we must have $c/\sigma_0^2 = \chi_{n,\alpha}^2$, which means $c = \sigma_0^2\chi_{n,\alpha}^2$.

8.17 a. The likelihood function is

$$L(\mu, \theta|\mathbf{x}, \mathbf{y}) = \mu^n \left(\prod_i x_i\right)^{\mu-1} \theta^n \left(\prod_j y_j\right)^{\theta-1}.$$

Maximizing, by differentiating the log-likelihood, yields the MLEs

$$\hat{\mu} = -\frac{n}{\sum_i \log x_i} \quad \text{and} \quad \hat{\theta} = -\frac{m}{\sum_j \log y_j}.$$

Under $H_0$, the likelihood is

$$L(\theta|\mathbf{x}, \mathbf{y}) = \theta^{n+m} \left(\prod_i x_i \prod_j y_j\right)^{\theta-1},$$

and maximizing as above yields the restricted MLE,

$$\hat{\theta}_0 = -\frac{n+m}{\sum_i \log x_i + \sum_j \log y_j}.$$

The LRT statistic is

$$\lambda(\mathbf{x}, \mathbf{y}) = \frac{\hat{\theta}_0^{m+n}}{\hat{\mu}^n \hat{\theta}^m} \left(\prod_i x_i\right)^{\hat{\theta}_0 - \hat{\mu}} \left(\prod_j y_j\right)^{\hat{\theta}_0 - \hat{\theta}}.$$

b. Substituting in the formulas for $\hat{\theta}$, $\hat{\mu}$ and $\hat{\theta}_0$ yields $(\prod_i x_i)^{\hat{\theta}_0-\hat{\mu}}\left(\prod_j y_j\right)^{\hat{\theta}_0-\hat{\theta}}=1$ and

$$\lambda(\mathbf{x},\mathbf{y})=\frac{\hat{\theta}_0^{m+n}}{\hat{\mu}^n\hat{\theta}^m}=\frac{\hat{\theta}_0^n}{\hat{\mu}^n}\frac{\hat{\theta}_0^m}{\hat{\theta}^m}=\left(\frac{m+n}{m}\right)^m\left(\frac{m+n}{n}\right)^n(1-T)^mT^n.$$

This is a unimodal function of $T$. So rejecting if $\lambda(\mathbf{x},\mathbf{y})\le c$ is equivalent to rejecting if $T\le c_1$ or $T\ge c_2$, where $c_1$ and $c_2$ are appropriately chosen constants.

c. Simple transformations yield $-\log X_i\sim$ exponential$(1/\mu)$ and $-\log Y_i\sim$ exponential$(1/\theta)$. Therefore, $T=W/(W+V)$ where $W$ and $V$ are independent, $W\sim$ gamma$(n,1/\mu)$ and $V\sim$ gamma$(m,1/\theta)$. Under $H_0$, the scale parameters of $W$ and $V$ are equal. Then, a simple generalization of Exercise 4.19b yields $T\sim$ beta$(n,m)$. The constants $c_1$ and $c_2$ are determined by the two equations

$$P(T\le c_1)+P(T\ge c_2)=\alpha\quad\text{and}\quad(1-c_1)^mc_1^n=(1-c_2)^mc_2^n.$$

8.20 By the Neyman-Pearson Lemma, the UMP test rejects for large values of $f(x|H_1)/f(x|H_0)$. Computing this ratio we obtain

| $x$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $\frac{f(x\|H_1)}{f(x\|H_0)}$ | 6 | 5 | 4 | 3 | 2 | 1 | .84 |

The ratio is decreasing in $x$. So rejecting for large values of $f(x|H_1)/f(x|H_0)$ corresponds to rejecting for small values of $x$. To get a size $\alpha$ test, we need to choose $c$ so that $P(X\le c|H_0)=\alpha$. The value $c=4$ gives the UMP size $\alpha=.04$ test. The Type II error probability is $P(X=5,6,7|H_1)=.82$.

10.31 a. By CLT, we have $\hat{p}_1\xrightarrow{d}\mathcal{N}\left(p_1,p_1(1-p_1)/n_1\right)$, $\hat{p}_2\xrightarrow{d}\mathcal{N}\left(p_2,p_2(1-p_2)/n_2\right)$. Stacking them together, and considering that they are independent, we have

$$\begin{bmatrix}\hat{p}_1\\ \hat{p}_2\end{bmatrix}\xrightarrow{d}\mathcal{N}\left(\begin{bmatrix}p_1\\ p_2\end{bmatrix},\begin{matrix}p_1(1-p_1)/n_1 & 0\\ 0 & p_2(1-p_2)/n_2\end{matrix}\right).$$ Using Delta's method, it is easy to show that

$\hat{p}_1-\hat{p}_2\xrightarrow{d}\mathcal{N}\left(p_1-p_2,p_1(1-p_1)/n_1+p_1(1-p_1)/n_2\right)$. Under $H_0:p_1=p_2=p$. $\hat{p}$ is the MLE of p, thus $\hat{p}\xrightarrow{p}p$. Combining these facts with Slutzkey's theorem, we get

$$\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\hat{p}(1-\hat{p})}}\to\mathrm{n}(0,1)$$

Therefore, $T\xrightarrow{d}\chi_1^2$.

b. Substitute $\hat{p}_i$s for $S_i$ and $F_i$s to get

$$\begin{aligned} T^* &= \frac{n_1^2(\hat{p}_1-\hat{p})^2}{n_1\hat{p}} + \frac{n_2^2(\hat{p}_2-\hat{p})^2}{n_2\hat{p}} \\ &\quad + \frac{n_1^2\left[(1-\hat{p}_1)-(1-\hat{p})\right]^2}{n_1(1-\hat{p})} + \frac{n_2^2\left[(1-\hat{p}_2)-(1-\hat{p})\right]^2}{n_2\hat{p}} \\ &= \frac{n_1(\hat{p}_1-\hat{p})^2}{\hat{p}(1-\hat{p})} + \frac{n_2(\hat{p}_2-\hat{p})^2}{\hat{p}(1-\hat{p})} \end{aligned}$$

Write $\hat{p} = (n_1\hat{p}_1 + n_2\hat{p}_2)/(n_1+n_2)$. Substitute this into the numerator, and some algebra will get

$$n_1(\hat{p}_1-\hat{p})^2 + n_2(\hat{p}_2-\hat{p})^2 = \frac{(\hat{p}_1-\hat{p}_2)^2}{\frac{1}{n_1}+\frac{1}{n_2}},$$

so $T^* = T$.

c. Under $H_0$,

$$\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)p(1-p)}} \to \mathrm{n}(0,1)$$

and both $\hat{p}_1$ and $\hat{p}_2$ are consistent, so $\hat{p}_1(1-\hat{p}_1) \to p(1-p)$ and $\hat{p}_2(1-\hat{p}_2) \to p(1-p)$ in probability. Therefore, by Slutsky's Theorem,

$$\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \to \mathrm{n}(0,1),$$

and $(T^{**})^2 \to \chi_1^2$. It is easy to see that $T^{**} \neq T$ in general.

e. We have $\hat{p}_1 = 34/40$, $\hat{p}_2 = 19/35$, $\hat{p} = (34+19)/(40+35) = 53/75$, and $T = 8.495$. Since $\chi^2_{1,.05} = 3.84$, we can reject $H_0$ at $\alpha = .05$.

7. Show that, when H0 is true, then the p-value has a Uniform (0,1) distribution.

Proof:

First, according to C&B Theorem 2.1.10, the cdf of a continuous r.v. follows Uniform(0, 1).

Now suppose that the reject region is $W(\mathbf{X}) < c$ for some test statistic $W(\mathbf{X})$ w/ some continuous cdf, then the p-value $p = \sup_{\theta \in \Theta_0} P_\theta(W(\mathbf{X}) < W(\mathbf{x}))$, where $\mathbf{x}$ are the (observed) data, and $W(\mathbf{x})$ is the observed statistic.

1. If $\Theta_0 = \{\theta_0\}$, then $p = P_{\theta_0}(W(\mathbf{X}) < W(\mathbf{x})) = F(W(\mathbf{x}); \theta_0)$, where $F$ is the cdf of $W$. Thus by the above theorem, $p \sim Unif(0, 1)$;

2. If $\Theta_0$ contains more than 1 point, then we need to take the sup on all points in $\Theta_0$, but remember that $P_\theta(W(\mathbf{X}) < W(\mathbf{x}))$ is a power function, thus it's monotonic, hence the sup is attained on the boundary. Assume the sup is attained at some boundary point $\theta^*$, then $p = P_{\theta^*}(W(\mathbf{X}) < W(\mathbf{x})) = F(W(\mathbf{x}); \theta^*)$, and $p \sim Unif(0, 1)$ same as case 1.

Same result holds for reject region $W(\mathbf{X}) > c$, where you'll end up w/ $p = 1 - F(W(\mathbf{x}); \theta)$ for some $\theta$. But since $F(W) \sim Unif(0, 1)$, $p \sim Unif(0, 1)$ too.
Therefore, under $H_0$, $p \sim Unif(0, 1)$.

36-705 Intermediate Statistics HW8

# Problem 1 (C & B 9.1)

**Solution:**

Denote $A = \{x : L(x) \leq \theta\}$ and $B = \{x : U(x) \geq \theta\}$. Then $A \cap B = \{x : L(x) \leq \theta \leq U(x)\}$ and $1 \geq P\{A \cup B\} = P\{L(X) \leq \theta \text{ or } \theta \not\leq U(X)\} \geq P\{L(X) \leq \theta \text{ or } \theta \leq L(X)\} = 1$, since $L(x) \leq U(x)$. Therefore, $P(A \cap B) = P(A) + P(B) - P(A \cup B) = 1 - \alpha_1 + 1 - \alpha_2 - 1 = 1 - \alpha_1 - \alpha_2$.

# Problem 2 (C & B 9.4(a))

**Solution:** (a).

$$\lambda(x, y) = \frac{\sup_{\lambda = \lambda_0} L\left(\sigma_X^2, \sigma_Y^2 \,|\, x, y\right)}{\sup_{\lambda \in (0, +\infty)} L\left(\sigma_X^2, \sigma_Y^2 \,|\, x, y\right)}$$

The unrestricted MLEs of $\sigma_X^2$ and $\sigma_Y^2$ are $\hat{\sigma}_X^2 = \frac{\Sigma X_i^2}{n}$ and $\hat{\sigma}_Y^2 = \frac{\Sigma Y_i^2}{m}$, as usual. Under the restriction, $\lambda = \lambda_0$, $\sigma_Y^2 = \lambda_0 \sigma_X^2$, and

$$\begin{aligned} L\left(\sigma_X^2, \lambda_0 \sigma_X^2 \,|\, x, y\right) &= \left(2\pi\sigma_X^2\right)^{-n/2} \left(2\pi\lambda_0\sigma_X^2\right)^{-m/2} e^{-\Sigma x_i^2/(2\sigma_X^2)} \cdot e^{-\Sigma y_i^2/(2\lambda_0\sigma_X^2)} \\ &= \left(2\pi\sigma_X^2\right)^{-(m+n)/2} \lambda_0^{-m/2} e^{-(\lambda_0 \Sigma x_i^2 + \Sigma y_i^2)/(2\lambda_0\sigma_X^2)} \end{aligned}$$

Differentiating the log likelihood gives

$$\begin{aligned} \frac{d \log L}{d\left(\sigma_X^2\right)^2} &= \frac{d}{d\sigma_X^2} \left[ -\frac{m+n}{2} \log \sigma_X^2 - \frac{m+n}{2} \log(2\pi) - \frac{m}{2} \log \lambda_0 - \frac{\lambda_0 \Sigma x_i^2 + \Sigma y_i^2}{2\lambda_0 \sigma_X^2} \right] \\ &= -\frac{m+n}{2} \left(\sigma_X^2\right)^{-1} + \frac{\lambda_0 \Sigma x_i^2 + \Sigma y_i^2}{2\lambda_0} \left(\sigma_X^2\right)^{-2} \overset{\text{set}}{=} 0 \end{aligned}$$

which implies

$$\hat{\sigma}_0^2 = \frac{\lambda_0 \Sigma x_i^2 + \Sigma y_i^2}{\lambda_0 (m+n)}.$$

To see this is a maximum, check the second derivative:

$$\begin{aligned} \frac{d^2 \log L}{d\left(\sigma_X^2\right)^2} &= \frac{m+n}{2} \left(\sigma_X^2\right)^{-2} - \frac{1}{\lambda_0} \left(\lambda_0 \Sigma x_i^2 + \Sigma y_i^2\right) \left(\sigma_X^2\right)^{-3} \Big|_{\sigma_X^2 = \hat{\sigma}_0^2} \\ &= -\frac{m+n}{2} (\hat{\sigma}_0^2)^{-2} < 0, \end{aligned}$$

therefore $\hat{\sigma}_0^2$ is the MLE. The LRT statistic is

$$\frac{\left(\hat{\sigma}_X^2\right)^{n/2}\left(\hat{\sigma}_Y^2\right)^{m/2}}{\lambda_0^{m/2}\left(\hat{\sigma}_0^2\right)^{(m+n)/2}},$$

and the test is: Reject $H_0$ if $\lambda(x,y) < k$, where $k$ is chosen to give the test size $\alpha$.

# Problem 3 (C & B 9.33(a))

**Solution:**

a. Since $0 \in C_a(x)$ for every $x$, $P(0 \in C_a(X)|\,\mu = 0) = 1$. If $\mu > 0$,

$$\begin{aligned} P(\mu \in C_a(X)) &= P(\mu \le \max\{0, X+a\}) = P(\mu \le X+a) && (\text{since } \mu > 0) \\ &= P(Z \ge -a) && (Z \sim \mathrm{n}(0,1)) \\ &= .95 && (a = 1.645.) \end{aligned}$$

A similar calculation holds for $\mu < 0$.

# Problem 4

**Solution:** The likelihood function for $\theta$ is

$$L(\theta; X_1, \ldots, X_n) = \frac{1}{\theta^n} I_{X_{(n)} \le \theta}, \qquad X_{(n)} = \max\{X_1, \ldots, X_n\}.$$

So, the MLE is $\hat{\theta} = X_{(n)}$.

The likelihood ratio for data is

$$\frac{L(\theta)}{L(\hat{\theta})} = \frac{\hat{\theta}^n}{\theta^n} \frac{I_{X_{(n)} \le \theta}}{I_{X_{(n)} \le \hat{\theta}}} = \frac{X_{(n)}^n}{\theta^n} I_{X_{(n)} \le \theta}.$$

Hence, using LRT, we accept the $H_0$ when

$$\frac{L(\theta)}{L(\hat{\theta})} = \frac{X_{(n)}^n}{\theta^n} I_{X_{(n)} \le \theta} > C.$$

Choose a proper $C$ to make sure that the test has size $\alpha$. For Uniform distribution, the size could be calculated as

$$P_\theta\left(\frac{X_{(n)}^n}{\theta^n} I_{X_{(n)} \le \theta} \le C\right) = P_\theta(X_{(n)} \le C^{1/n}\theta) = \frac{(C^{1/n}\theta)^n}{\theta^n} = C.$$

So, take $C = \alpha$ to make sure that the LRT is with size $\alpha$.

In this sense, the acceptance region is

$$A(\theta) = \{X_1, \cdots, X_n : \theta \geq X_{(n)} > \alpha^{1/n}\theta\},$$

and the corresponding $1-\alpha$ confidence interval is $(X_{(n)}, \frac{X_{(n)}}{\alpha^{1/n}})$.

# Problem 5

**Solution:** Given $x$, say that $x \in B_j$, then the estimator is $\hat{p}(x) = \frac{\hat{\theta}_j}{h} = \frac{1}{nh}\sum_{i=1}^{n} I(X_i \in B_j)$.

For any $X_i$, the distribution of $I(X_i \in B_j)$ is Bernoulli Distribution with parameter $p = P(X \in B_j) = \int_{B_j} p(t)dt$. As $X_1, \cdots, X_n$ are $i.i.d$ samples, so $I(X_1 \in B_j), \cdots, I(X_n \in B_j)$ are also $i.i.d$ samples. Hence, we have the expectation and variance for $\hat{\theta}_j$ as

$$E[\hat{\theta}_j] = \int_{B_j} p(t)dt,$$

and

$$var(\hat{\theta}_j) = \frac{1}{n}\int_{B_j} p(t)dt(1 - \int_{B_j} p(t)dt).$$

Hence, the bias and variance for the estimator is

$$bias(\hat{p}(x)) = \frac{1}{h}\int_{B_j} p(t)dt - p(x),$$

and

$$var(\hat{p}(x)) = \frac{1}{nh^2}\int_{B_j} p(t)dt(1 - \int_{B_j} p(t)dt)$$

So, the MSE for this single point is

$$MSE(x) = b^2 + v = (\frac{1}{h}\int_{B_j} p(t)dt - p(x))^2 + \frac{1}{nh^2}\int_{B_j} p(t)dt(1 - \int_{B_j} p(t)dt).$$

Now try to estimate the MSE term by term.

Taylor expansion shows that

$$\int_{B_j} p(t)dt = hp(x) + p'(x)\int_{B_j}(t-x)dt + \int_{B_j}\frac{(t-x)^2}{2}p''(\tilde{x})dx = hp(x) + hp'(x)(h(j-\frac{1}{2}) - x) + O(h^3).$$

In the bin $B_j$, the integration over bias square is

$$\int_{B_j}(\frac{1}{h}\int_{B_j} p(t)dt - p(x))^2 dx = \int_{B_j} p'(x)^2(h(j-\frac{1}{2}) - x)^2 dx + O(h^3),$$

and by the mean value theorem,

$$\int_{B_j} (\frac{1}{h}\int_{B_j} p(t)dt - p(x))^2 dx \approx p'(\tilde{x}_j)^2 \int_{B_j} (h(j-\frac{1}{2}) - x)^2 dx = p'(\tilde{x}_j)^2 \frac{h^3}{12}.$$

Hence, we have

$$\int_0^1 bias(x)^2 dx = \sum_{j=1}^m \int_{B_j} bias(x)^2 dx \approx \sum_{j=1}^m p'(\tilde{x}_j)^2 \frac{h^3}{12} \approx \int p'(x_j)^2 dx \frac{h^2}{12}.$$

For the variance part, in the bin $B_j$, it does not change, so the integration is

$$\int_{B_j} v dt = h(\frac{1}{nh^2}\int_{B_j} p(t)dt(1 - \int_{B_j} p(t)dt)) = \frac{1}{nh}(\int_{B_j} p(t)dt - (\int_{B_j} p(t)dt)^2),$$

and on $[0,1]$ interval, the variance is

$$\int_0^1 v dt = \sum_{j=1}^m \frac{1}{nh}(\int_{B_j} p(t)dt - (\int_{B_j} p(t)dt)^2) = \frac{1}{nh} - \frac{1}{nh}\sum_{j=1}^m (\int_{B_j} p(t)dt)^2.$$

With mean value theorem, we have that $\int_{B_j} p(t)dt = p(\tilde{x}_j)h$, so it becomes

$$\int_0^1 v dt = \frac{1}{nh} - \frac{1}{nh}\sum_{j=1}^m h^2 p(\tilde{x}_j)^2 \approx \frac{1}{nh}(1 - \int p^2(x)dx).$$

So, the approximation of MSE on the density function is

$$MSE = \int b^2 + v \approx \int p'(x_j)^2 dx \frac{h^2}{12} + \frac{1}{nh}(1 - \int p^2(x)dx).$$

If we take $C_1 = \int p'(x_j)^2 dx/12$, and $C_2 = (1 - \int p^2(x)dx)$, then the approximate MSE is

$$MSE \approx C_1 h^2 + C_2 \frac{1}{nh},$$

so the best $h$ should be $O(n^{-1/3})$, and the corresponding convergence rate is $n^{-2/3}$.

# 2011 Fall 10-705/36-705 Test 3 Solutions

(1) Let $X_1,\cdots,X_n \sim Bernoulli(\theta), \theta \in (0,1)$

(a) Find MLE $\hat{\theta}$, score function, and Fisher information.

**Solution**:

$$L(\theta; X^n) = \theta^{\Sigma X_i}(1-\theta)^{n-\Sigma X_i}$$

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S(\theta) = \frac{\partial \log L}{\partial \theta} = \frac{\Sigma X_i - n\theta}{\theta(1-\theta)}$$

$$I_n(\theta) = -E_\theta\left(\frac{\partial^2 \log L}{\partial \theta^2}\right) = \frac{n}{\theta(1-\theta)}$$

(b) Find the limiting distribution of $\hat{\tau} = e^{\hat{\theta}}$.

**Solution**:

According to Thm 11 Lecture 9:

$$\sqrt{n}\left(\tau(\hat{\theta}) - \tau(\theta)\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{|\tau'(\theta)|^2}{I_1(\theta)}\right) = \mathcal{N}\left(0, e^{2\theta}\theta(1-\theta)\right)$$

$$\Rightarrow \tau(\hat{\theta}) \xrightarrow{d} \mathcal{N}\left(e^\theta, \frac{e^{2\theta}\theta(1-\theta)}{n}\right)$$

(c) Find the Wald test for $H_0: \theta = 1/2, H_1: \theta \neq 1/2$.

**Solution**:

$$se(\hat{\theta}) = \sqrt{\frac{\theta(1-\theta)}{n}}$$

Therefore the Wald test is: reject when $\left|\frac{\hat{\theta} - 1/2}{1/\sqrt{4n}}\right| > z_{\alpha/2}$.

(2) Let $X_1,\cdots,X_n \sim \mathcal{N}(\theta, 1)$.

(a) Find the level-α Neyman-Pearson test for $H_0 : \theta = 1, H_1 : \theta = 2$.

**Solution**:

The Neyman-Pearson test rejects when $T\left(X^n\right) > k_\alpha$.

$$T\left(x^n\right) = \frac{L\left(\theta_1\right)}{L\left(\theta_0\right)} = \frac{\exp\left(-\frac{1}{2}\sum_i \left(X_i - 2\right)^2\right)}{\exp\left(-\frac{1}{2}\sum_i \left(X_i - 1\right)^2\right)} = \exp\left(n\left(\bar{X} - 3/2\right)\right)$$

To get $k_\alpha$

$$P_0\left(T\left(x\right) > k_\alpha\right) = \alpha \Leftrightarrow P_0\left(\exp\left(n\left(\bar{X} - 3/2\right)\right) > k_\alpha\right) = \alpha \Leftrightarrow P_0\left(\bar{X} > c\right) = \alpha .$$

Since $\bar{X} \sim \mathcal{N}\left(1, 1/n\right)$, we know $P_0\left(\bar{X} > 1 + \frac{z_\alpha}{\sqrt{n}}\right) = \alpha$. Therefore the test is: reject when $\bar{X} > 1 + \frac{z_\alpha}{\sqrt{n}}$.

(b) In what sense is Neyman-Pearson optimal:

**Solution**:

Neyman-Pearson is optimal because it is UMP: among all the level-α tests, Neyman-Pearson has the largest power function for all $\theta \in \Theta_1$ i.e. has minimum type II error.

(c) Find the LRT of $H_0 : \theta = 1, H_1 : \theta \neq 1$.

**Solution**:

$$\hat{\theta}_0 = 1, \hat{\theta}_{MLE} = \bar{X}$$

$$\lambda\left(X^n\right) = \frac{L\left(\hat{\theta}_0\right)}{L\left(\hat{\theta}\right)} = \frac{\exp\left(-\frac{1}{2}\sum_i \left(X_i - \theta_0\right)^2\right)}{\exp\left(-\frac{1}{2}\sum_i \left(X_i - \hat{\theta}\right)^2\right)} = \frac{\exp\left(-\frac{1}{2}\sum_i \left(X_i - 1\right)^2\right)}{\exp\left(-\frac{1}{2}\sum_i \left(X_i - \bar{X}\right)^2\right)}$$

$$= \exp\left(-\frac{1}{2}\sum_i \left(X_i - 1\right)^2 - \left(X_i - \bar{X}\right)^2\right) = \exp\left(-\frac{n}{2}\left(\bar{X} - 1\right)^2\right)$$

We know under H0 $\sqrt{n}\left(\bar{X}-1\right)\sim\mathcal{N}\left(0,1\right)\Rightarrow n\left(\bar{X}-1\right)^2\sim\chi_1^2$. Therefore the LRT is: reject H0 when $n\left(\bar{X}-1\right)^2>\chi_{1,\alpha}^2$ (or equivalently $\sqrt{n}\left|\bar{X}-1\right|>z_{\alpha/2}$).

(3) Let $X_1,\cdots,X_n\sim Uniform\left(0,\theta\right),\theta>0$

(a) Find the likelihood and MLE.

**Solution**:

$$L=\frac{1}{\theta^n}\prod_i I\left(X_i\le\theta\right)=\frac{1}{\theta^n}I\left(X_{(n)}\le\theta\right)$$
$$\hat{\theta}=X_{(n)}=\max_i X_i$$

(b) Find the form of LRT for $H_0:\theta=1,H_1:\theta\neq1$

**Solution**:

The LRT rejects H0 if $\lambda\left(X^n\right)\le c$.

$$\lambda\left(X^n\right)=\frac{L\left(\hat{\theta}_0\right)}{L\left(\hat{\theta}\right)}=\frac{I\left(X_{(n)}\le1\right)}{\frac{1}{X_{(n)}^n}I\left(X_{(n)}\le X_{(n)}\right)}=X_{(n)}^nI\left(X_{(n)}\le1\right)$$

Therefore, if $X_{(n)}>1$, always reject H0. Otherwise reject H0 if $X_{(n)}^n$ is smaller than some value.

(c) Find the form of likelihood ratio confidence interval.

**Solution**:

$$C=\left\{\theta:\frac{L\left(\theta\right)}{L\left(\hat{\theta}\right)}\ge c\right\}$$
$$\frac{L\left(\theta\right)}{L\left(\hat{\theta}\right)}=\frac{\frac{1}{\theta^n}I\left(X_{(n)}\le\theta\right)}{\frac{1}{X_{(n)}^n}}=\frac{X_{(n)}^n}{\theta^n}I\left(X_{(n)}\le\theta\right)$$

When $\theta < X_{(n)}$, this ratio is always zero. When $\theta \geq X_{(n)}$, this ratio is monotonically decreasing with $\theta$. Therefore, C should have the form $\left[ X_{(n)}, U \right]$.

(4) Let $X_1, \cdots, X_n \sim p(x;\theta)$

(a) Let C($X^n$) be a 1-α confidence interval for θ. Consider testing $H_0 : \theta = \theta_0, H_1 : \theta \neq \theta_0$.

Suppose we reject H0 if $\theta_0 \in C\left(X^n\right)$. Show that this defines a level α test for H0.

**Solution**:

$$\inf_\theta P_\theta\left(\theta \in C\left(X^n\right)\right) \geq 1-\alpha$$
$$\Leftrightarrow 1-\inf_\theta P_\theta\left(\theta \in C\left(X^n\right)\right) \leq \alpha$$
$$\Leftrightarrow \sup_\theta 1-P_\theta\left(\theta \in C\left(X^n\right)\right) \leq \alpha$$
$$\Leftrightarrow \sup_\theta P_\theta\left(\theta \notin C\left(X^n\right)\right) \leq \alpha$$

which is the definition of a level-α test.

(b) $X_1, \cdots, X_n \sim Uniform(0,\theta), \theta > 0$. Let $C_n = \left(X_{(n)}, X_{(n)} / \alpha^{1/n}\right)$. Show that $C_n$ is a 1-α CI for θ.

**Solution**:

For $\forall \theta$,

$$P_\theta\left(\theta \in C_n\right) = P_\theta\left(X_{(n)} \leq \theta \leq X_{(n)} / \alpha^{1/n}\right) = P_\theta\left(\theta\alpha^{1/n} \leq X_{(n)}\right) = 1 - P_\theta\left(\theta\alpha^{1/n} \geq X_{(n)}\right).$$
$$= 1-\left(P_\theta\left(\theta\alpha^{1/n} \geq X_1\right)\right)^n = 1-\left(\frac{1}{\theta}\theta\alpha^{1/n}\right)^n = 1-\alpha$$

(c) Use (a) and (b) to define a level α test of $H_0 : \theta = 1, H_1 : \theta \neq 1$. Find its power function.

**Solution**:

The test is: reject $H_0$ if $1 \notin \left[ X_{(n)}, X_{(n)} / \alpha^{1/n} \right] \Leftrightarrow X_{(n)} > 1, or, X_{(n)} < \alpha^{1/n}$.

$$\beta(\theta) = P_\theta\left(X_{(n)} > 1 \cup X_{(n)} < \alpha^{1/n}\right) = P_\theta\left(X_{(n)} > 1\right) + P_\theta\left(X_{(n)} < \alpha^{1/n}\right)$$

$$= 1 - P_\theta\left(X_{(n)} < 1\right) + P_\theta\left(X_{(n)} < \alpha^{1/n}\right)$$

$$= 1 - \left(P_\theta\left(X_1 < 1\right)\right)^n + \left(P_\theta\left(X_1 < \alpha^{1/n}\right)\right)^n$$

$$= \begin{cases} 1 & \theta \le \alpha^{1/n} \\ \dfrac{\alpha}{\theta^n} & \alpha^{1/n} < \theta \le 1 \\ 1 + \dfrac{\alpha - 1}{\theta^n} & 1 < \theta \end{cases}$$

# 36705 Intermediate Statistics Homework 9 Solutions

## Problem 1

a).

$$R(h) = \mathbb{E}(L(h)) = \mathbb{E}\int (\hat{p}_h(x))^2 dx - 2\mathbb{E}\int \hat{p}_h(x)p(x)dx + \int (p(x))^2 dx$$

Since the last term of the rhs has nothing to do with "$h$", differentiate $R(h)$ with respective to $h$, $d\int (p(x))^2/dh = 0$. Then,

$$\min_h R(h) = \min_h \mathbb{E}(L(h)) = \min_h \left(\mathbb{E}\int (\hat{p}_h(x))^2 dx - 2\mathbb{E}\int \hat{p}_h(x)p(x)dx\right) = \min_h \tilde{R}(h)$$

b).

$$\begin{aligned}
\mathbb{E}\hat{R}(h) &= \mathbb{E}\int (\hat{p}_h(x))^2 dx - \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}(\hat{p}_h(Y_i)) \\
&= \mathbb{E}_X \int (\hat{p}_h(x))^2 dx - 2\mathbb{E}_{X,Y}(\hat{p}_h(Y_1)) \qquad (1) \\
&= \mathbb{E}_X \int (\hat{p}_h(x))^2 dx - 2\mathbb{E}_X \int (\hat{p}_h(y)p(y)dy) \qquad (2)
\end{aligned}$$

The second expectation in (1) is with respect both X's and Y's, while the second expectation in (2) is with respect to X's. So with prove that $\mathbb{E}\hat{R}(h) = \tilde{R}(h) = \mathbb{E}\int (\hat{p}_h(x))^2 dx - 2\mathbb{E}\int \hat{p}_h(x)p(x)dx$.

## Problem 2

Kernel density estimator is defined as

$$\hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{X_i - x}{h}\right)$$

Let $Y_i = \frac{1}{h}K\left(\frac{X_i - x}{h}\right)$. Since kernel $K$ is a symmetric density with expectation 0. Then $a < Y_i < b$ with $a = min(K(\frac{X_i-x}{h}))/h$ and $b = max(K(\frac{X_i-x}{h}))/h$. We can apply Hoeffding inequality that

$$P(|\hat{p}_h(x) - p_h(x)| > \epsilon) = P(|\frac{1}{n}\sum_{i=1} Y_i - \mathbb{E}(Y)| > \epsilon) \leq 2e^{\frac{-2n\epsilon^2}{(b-a)^2}}$$

## Problem 3

$$\mathbb{E}(\hat{\theta}^*|X_1, ..., X_n) = n^{-1}\mathbb{E}[Y \sim Binomial(n, \overline{X})] = n^{-1}n\overline{X} = \overline{X}$$
$$\mathbb{E}(\hat{\theta}^*) = \mathbb{E}(\mathbb{E}(\hat{\theta}^*|X_1, ..., X_n)) = \mathbb{E}\overline{X} = \theta$$

$$\mathbb{V}ar(\hat{\theta}^*|X_1, ..., X_n) = n^{-2}n\overline{X}(1-\overline{X}) = n^{-1}\overline{X}(1-\overline{X})$$
$$\mathbb{V}ar(\hat{\theta}^*) = \mathbb{V}ar(\mathbb{E}(\hat{\theta}^*|X_1, ..., X_n)) + \mathbb{E}(\mathbb{V}ar(\hat{\theta}^*|X_1, ..., X_n)) =$$
$$= \mathbb{V}ar(\overline{X}) + \mathbb{E}(n^{-1}\overline{X}(1-\overline{X})) =$$
$$= n^{-1}\theta(1-\theta) + n^{-1}(\mathbb{E}\overline{X} - \mathbb{E}\overline{X}^2) =$$
$$= n^{-1}(\theta(1-\theta) + \theta - n^{-1}\theta(1-\theta) - \theta^2) =$$
$$= n^{-1}2\theta(1-\theta) - n^{-2}\theta(1-\theta) =$$
$$= \theta(1-\theta)\left(\frac{2n-1}{n^2}\right)$$

## Problem 4

$$\frac{\mathbb{V}(\hat{\theta^*}|X_1, \ldots, X_n)}{\mathbb{V}ar(\hat{\theta_n})} = \frac{\bar{X}(1-\bar{X})}{p(1-p)}$$

$$\bar{X} \xrightarrow{P} p \text{ by law of large number}$$

Then,

$$\bar{X}(1-\bar{X}) \xrightarrow{P} p(1-p) \text{ by Theorem 10 in Lecture 4}$$

So

$$\frac{\bar{X}(1-\bar{X})}{p(1-p)} \xrightarrow{P} 1$$

36-705 Intermediate Statistics HW10

# Problem 1 (C & B 7.23)

**Solution:**

Let $t = s^2$ and $\theta = \sigma^2$. Because $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, we have

$$f(t|\theta) = \frac{1}{\Gamma\left((n-1)/2\right) 2^{(n-1)/2}} \left(\frac{n-1}{\theta} t\right)^{[(n-1)/2]-1} e^{-(n-1)t/2\theta} \frac{n-1}{\theta}.$$

With $\pi(\theta)$ as given, we have (ignoring terms that do not depend on $\theta$)

$$\begin{aligned} \pi(\theta|t) &\propto \left[\left(\frac{1}{\theta}\right)^{((n-1)/2)-1} e^{-(n-1)t/2\theta} \frac{1}{\theta}\right] \left[\frac{1}{\theta^{\alpha+1}} e^{-1/\beta\theta}\right] \\ &\propto \left(\frac{1}{\theta}\right)^{((n-1)/2)+\alpha+1} \exp\left\{-\frac{1}{\theta}\left[\frac{(n-1)t}{2} + \frac{1}{\beta}\right]\right\}, \end{aligned}$$

which we recognize as the kernel of an inverted gamma pdf, $\text{IG}(a, b)$, with

$$a = \frac{n-1}{2} + \alpha \qquad \text{and} \qquad b = \left[\frac{(n-1)t}{2} + \frac{1}{\beta}\right]^{-1}.$$

Direct calculation shows that the mean of an $\text{IG}(a, b)$ is $1/((a-1)b)$, so

$$\text{E}(\theta|t) = \frac{\frac{n-1}{2} t + \frac{1}{\beta}}{\frac{n-1}{2} + \alpha - 1} = \frac{\frac{n-1}{2} s^2 + \frac{1}{\beta}}{\frac{n-1}{2} + \alpha - 1}.$$

This is a Bayes estimator of $\sigma^2$.

# Problem 2 (C & B 9.27)

**Solution:**

a. $Y = \sum X_i \sim \text{gamma}(n, \lambda)$, and the posterior distribution of $\lambda$ is

$$\pi(\lambda|y) = \frac{(y+\frac{1}{b})^{n+a}}{\Gamma(n+a)} \frac{1}{\lambda^{n+a+1}} e^{-\frac{1}{\lambda}(y+\frac{1}{b})},$$

an IG $\left(n+a, (y+\frac{1}{b})^{-1}\right)$. The Bayes HPD region is of the form $\{\lambda\colon \pi(\lambda|y) \geq k\}$, which is an interval since $\pi(\lambda|y)$ is unimodal. It thus has the form $\{\lambda\colon a_1(y) \leq \lambda \leq a_2(y)\}$, where $a_1$ and $a_2$ satisfy

$$\frac{1}{{a_1}^{n+a+1}} e^{-\frac{1}{a_1}(y+\frac{1}{b})} = \frac{1}{{a_2}^{n+a+1}} e^{-\frac{1}{a_2}(y+\frac{1}{b})}.$$

b. The posterior distribution is IG$(((n-1)/2)+a, (((n-1)s^2/2)+1/b)^{-1})$. So the Bayes HPD region is as in part a) with these parameters replacing $n+a$ and $y+1/b$.

c. As $a \to 0$ and $b \to \infty$, the condition on $a_1$ and $a_2$ becomes

$$\frac{1}{{a_1}^{((n-1)/2)+1}} e^{-\frac{1}{a_1}\frac{(n-1)s^2}{2}} = \frac{1}{{a_2}^{((n-1)/2)+1}} e^{-\frac{1}{a_2}\frac{(n-1)s^2}{2}}.$$

# Problem 3

**Solution:**

(a) Say that $X_j = Z_j + \theta_j$, then $X_j \sim N(\theta_j, 1)$. The moment generating function for $X_j \sim N(\theta, 1)$ is $M_X(t) = e^{\theta t + t^2/2}$. So, we have the moments for $X$ as

$$E[X] = \theta, \qquad E[X^2] = 1+\theta^2, \qquad E[X^3] = 3\theta+\theta^3, \qquad E[X^4] = 3+6\theta^2+\theta^4,$$

and

$$Var(X^2) = E[X^4] - (E[X^2])^2 = 3+12\theta^2+\theta^4-(1+\theta^2)^2 = 2+4\theta^2.$$

Because $X_j$s are independent, so we have

$$E[V] = \sum_{j=1}^{k} E[X_j^2] = \sum_{j=1}^{k} 1+\theta_j^2 = k+\lambda,$$

$$Var[V] = \sum_{j=1}^{k} Var[X_j^2] = \sum_{j=1}^{k} 2+4\theta_j^2 = 2(k+2\lambda).$$

(b) The posterior distribution of $\mu$ is

$$f(\mu) = f(y|\mu)f(\mu) = \Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i-\mu_i)^2}{2}}.$$

So, the posterior distribution of $\mu$ is $N(y, I_n)$.

(c) In this case, the distribution of $\tau = \sum_i \mu_i^2$ is $\chi_n^2(\lambda)$, where $\lambda = \sum_i y_i^2$.
(d) According to the result in (a), the posterior mean of $\tau$ is $\hat{\tau} = n + \sum_i y_i^2$.
(e) Let $W = \sum_i y_i^2$ and thus $\hat{\tau} = n + W$. By definition, $W \sim \chi_n^2(\sum_i \mu_i^2)$. Therefore

$$\begin{aligned} bias(\hat{\tau}) &= E(\hat{\tau}) - \sum_i \mu_i^2 = E(n+W) - \sum_i \mu_i^2 \\ &= n + (n + \sum_i \mu_i^2) - \sum_i \mu_i^2 \\ &= 2n \\ Var(\hat{\tau}) &= Var(n+W) = Var(W) \\ &= 2n + 4\sum_i \mu_i^2 \end{aligned}$$

(f) According to the hint, $W \approx N\left(E(W), Var(W)\right) = N\left(n + \sum_i \mu_i^2, 2n + 4\sum_i \mu_i^2\right)$. Now consider the probability $P(|\hat{\tau} - \tau| > \epsilon)$ for an arbitrarily small $\epsilon$. This probability will never approach 1 since $\hat{\tau} - \tau = W + n - \sum_i \mu_i^2 \approx N\left(2n, 2n + 4\sum_i \mu_i^2\right)$. In other words, the density of $\hat{\tau} - \tau$ will never concentrate around zero. Therefore, $\hat{\tau}$ is not consistent.

(g) From (c), $\tau = \sum_i \mu_i^2$ is $\chi_n^2(\lambda)$, then the $1 - \alpha$ confidence interval for $\tau$ is $C_n = [\chi_{n,\alpha}^2(\sum_i y_i^2), +\infty)$.

(h) From (e), $bias(\hat{\tau}) = 2n$, then $E(\hat{\tau} - 2n) = 0$. $\tilde{\tau} = \hat{\tau} - 2n = W - n$ is an unbiased estimator. $Var(\tilde{\tau}) = 2n + 4\sum_i \mu_i^2$. As $n \to \infty$, $Var(\tilde{\tau}) \nrightarrow 0$, so $\tilde{\tau}$ is not consistent either.

(i) From (e) we have $W \sim \chi_n^2(\sum_i \mu_i^2) = \chi_n^2(\tau)$. Suppose we want to test $H_0 : \tau = \tau_0$ vs. $H_1 : \tau \neq \tau_0$, then the rejection region of a level $\alpha$ test is $R = \{W : W \leq \chi_{n,\alpha}^2(\tau_0)\}$. By inverting this test, we have a size $1 - \alpha$ confidence interval $A_n = \{\tau : W \geq \chi_{n,\alpha}^2(\tau)\}$. The interval in (g) is actually Bayesian credible interval where the parameter $\tau$ is random and the interval is determined by the posterior distribution of $\tau$. The interval in (i) is the frequentist confidence interval which we assume it is fixed and the interval is determined from the distribution of the estimator of $\tau$.

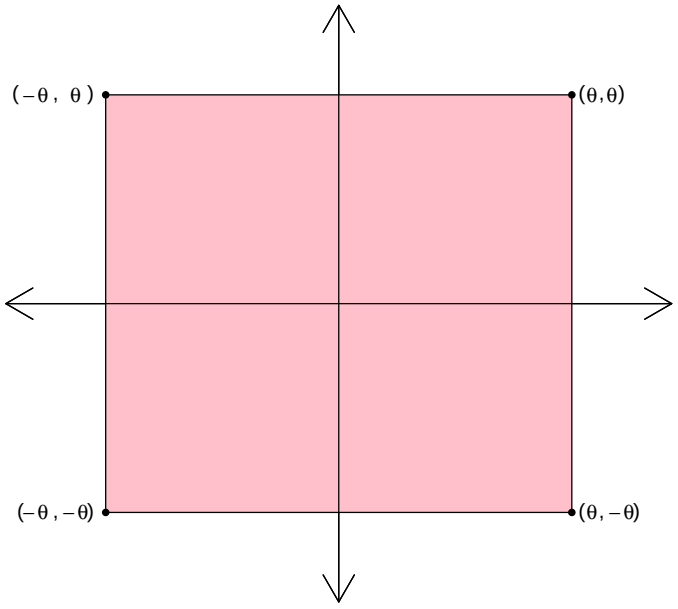
**Practice Final Exam**

1. Let $X_1, \ldots, X_n$ be iid from a distribution with mean $\mu$ and variance $\sigma^2$. Let

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

where $\overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. Prove that $S_n \xrightarrow{P} \sigma$.

2. Let $\theta > 0$. Let $S_\theta$ denote the square in the plane whose four corners are $(\theta, \theta), (-\theta, \theta), (-\theta, -\theta)$ and $(\theta, -\theta)$. Let $X_1, \ldots, X_n$ be iid data from a uniform distribution over $S_\theta$. (Note that each $X_i \in \mathbb{R}^2$.)




(a) Find a minimal sufficient statistic.

(b) Find the maximum likelihood estimate (mle).

(c) Show that the mle is consistent.

3. Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$ and let $Y_1, \ldots, Y_m \sim \text{Poisson}(\gamma)$. Assume that the two samples are independent.

   (a) Find the Wald test for testing

   $$H_0 : \lambda = \gamma \quad \text{versus} \quad H_1 : \lambda \neq \gamma.$$

   (b) Find the likelihood ratio test for testing

   $$H_0 : \lambda = \gamma \quad \text{versus} \quad H_1 : \lambda \neq \gamma.$$

   What is the (approximate) level $\alpha$ critical value?

   (c) Find an approximate $1 - \alpha$ confidence interval for $\lambda - \gamma$.

   (d) Find the BIC criterion for deciding between the two models:

   Model I: $\nu = \gamma$.

   Model II: $\nu \neq \gamma$.

4. Let $X_1, \ldots, X_n \sim \text{Unif}(0, \theta)$.

   (a) Let $\widehat{\theta} = a\overline{X}_n$ where $a > 0$ is a constant. Find the risk of $\widehat{\theta}$ under squared error loss.

   (b) Find the posterior mean using the (improper) prior $\pi(\theta) \propto 1/\theta$.

   (c) Suppose now that $0 \leq \theta \leq B$ where $B > 0$ is given. Hence the parameter space is $\Theta = [0, B]$. Let $\widehat{\theta}$ be the Bayes estimator (assuming squared error loss) assuming that the prior puts all its mass at $\theta = 0$. In other words, the prior is a point mass at $\theta = 0$. Prove that the posterior mean is **not** minimax. (Hint: You need only find some other estimator $\widetilde{\theta}$ such that $\sup_{\theta \in \Theta} R(\theta, \widetilde{\theta}) < \sup_{\theta \in \Theta} R(\theta, \widehat{\theta})$.

5. Suppose that $(Y, X)$ are random variables where $Y \in \{0, 1\}$ and $X \in \mathbb{R}$. Suppose that

   $$X|Y = 0 \sim \text{Unif}(-5, 5)$$

   and that

   $$X|Y = 1 \sim \text{Unif}(-1, 1).$$

   Further suppose that $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1) = 1/2$.

   (a) Find $m(x) = \mathbb{P}(Y = 1|X = x)$.

   (b) Let $\mathcal{A} = \{(a, b) : \ a, b \in \mathbb{R} \ , \ a \leq b\}$. Find the VC dimension of $\mathcal{A}$.

(c) Let $\mathcal{H} = \{h_A : \ A \in \mathcal{A}\}$ where $h_A(x) = 1$ if $x \in A$ and $h_A(x) = 0$ if $x \notin A$. Show that the Bayes rule $h_*$ is in $\mathcal{H}$.

(d) Let $\widehat{h}$ be the empirical risk minimizer based on data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Show that $R(\widehat{h}) - R(h_*) \leq \epsilon$ with high probability.

6. Let $X_1, X_2$ be iid Uniform$(0, 1)$. Find the density of $Y = X_1 + X_2$.

7. Let $X_1, \ldots, X_n$ be iid data from a uniform distribution over the disc of radius $\theta$ in $\mathbb{R}^2$. Thus, $X_i \in \mathbb{R}^2$ and

$$f(x; \theta) = \begin{cases} \frac{1}{\pi\theta^2} & \text{if } ||x|| \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

where $||x|| = \sqrt{x_1^2 + x_2^2}$.

(a) Find a minimal sufficient statistic.

(b) Find the maximum likelihood estimate (mle).

(c) Show that the mle is consistent.

8. Let $X \sim$ Binomial$(n, p)$ and $Y \sim$ Binomial$(m, q)$. Assume that $X$ and $Y$ are independent.

(a) Find the Wald test for testing

$$H_0 : p = q \quad \text{versus} \quad H_1 : p \neq q.$$

(b) Find the likelihood ratio test for testing

$$H_0 : p = q \quad \text{versus} \quad H_1 : p \neq q.$$

(c) Find an approximate $1 - \alpha$ confidence interval for $\theta = p - q$.

9. Let $X \sim f(x; \theta)$ where $\theta \in \Theta$. Let $L(\widehat{\theta}, \theta)$ be a loss fuctions.

(a) Define the following terms: *risk function*, *minimax estimator*, *Bayes estimator*.

(b) Show that a Bayes estimator with constant risk is minimax.

10. Let $X_1, \ldots, X_n \sim N(\theta, 1)$. Let $\pi$ be a $N(0,1)$ prior:

$$\pi(\theta) = \frac{1}{\sqrt{2\pi}} e^{-\theta^2/2}.$$

(a) Find the posterior distribution for $\theta$.

(b) Find the posterior mean $\overline{\theta}$.

(c) Find the mean squared error $R(\theta, \overline{\theta}) = \mathbb{E}_\theta(\overline{\theta} - \theta)^2$.

11. Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$.

(a) Find the mle $\widehat{\lambda}$.

(b) Find the score function.

(c) Find the Fisher informtion.

(d) Find the limiting distribution of the mle.

(e) Show that $\widehat{\lambda}$ is consistent.

(f) Let $\psi = e^\lambda$. Find the limiting distribution of $\widehat{\psi} = e^{\widehat{\lambda}}$.

(g) Show that $\widehat{\psi}$ is a consistent estimate of $\psi$.

12. Let $X_1, \ldots, X_n$ be a sample from $f(x;\theta) = (1/2)(1 + \theta x)$ where $-1 < x < 1$ and $-1 < \theta < 1$.

(a) Find the mle $\widehat{\theta}$. Show that it is consistent.

(b) Find the method of moments estimator and show that it is consistent.