STAT 6720

Mathematical Statistics II

Spring Semester 2001

Dr. Jürgen Symanzik

Utah State University

Department of Mathematics and Statistics

3900 Old Main Hill

Logan, UT 84322–3900

Tel.: (435) 797–0696

FAX: (435) 797–1822

e-mail: `symanzik@sunfs.math.usu.edu`

Contents

Acknowledgements

I would like to thank my students, Hanadi B. Eltahir, Rich Madsen, and Bill Morphet, who helped during the Fall 1999 and Spring 2000 semesters in typesetting these lecture notes using LaTeX and for their suggestions how to improve some of the material presented in class. Thanks are also due to my Fall 2000 and Spring 2001 semester students, Weiping Deng, David B. Neal, Emily Simmons, Shea Watrin, Aonan Zhang, and Qian Zhao, for their comments that helped to improve these lecture notes.

In addition, I particularly would like to thank Mike Minnotte and Dan Coster, who previously taught this course at Utah State University, for providing me with their lecture notes and other materials related to this course. Their lecture notes, combined with additional material from Rohatgi (1976) and other sources listed below, form the basis of the script presented here.

The textbook required for this class is:

- Rohatgi, V. K. (1976): *An Introduction to Probability Theory and Mathematical Statistics*, John Wiley and Sons, New York, NY.

A Web page dedicated to this class is accessible at:

`http://www.math.usu.edu/~symanzik/teaching/2001_stat6720/stat6720.html`

This course closely follows Rohatgi (1976) as described in the syllabus. Additional material originates from the lectures from Professors Hering, Trenkler, Gather, and Kreienbrock I have attended while studying at the Universität Dortmund, Germany, the collection of Masters and PhD Preliminary Exam questions from Iowa State University, Ames, Iowa, and the following textbooks:

- Bandelow, C. (1981): *Einführung in die Wahrscheinlichkeitstheorie*, Bibliographisches Institut, Mannheim, Germany.

- Büning, H., and Trenkler, G. (1978): *Nichtparametrische statistische Methoden*, Walter de Gruyter, Berlin, Germany.

- Casella, G., and Berger, R. L. (1990): *Statistical Inference*, Wadsworth & Brooks/Cole, Pacific Grove, CA.

- Fisz, M. (1989): *Wahrscheinlichkeitsrechnung und mathematische Statistik*, VEB Deutscher Verlag der Wissenschaften, Berlin, German Democratic Republic.

- Gibbons, J. D., and Chakraborti, S. (1992): *Nonparametric Statistical Inference* (Third Edition, Revised and Expanded), Dekker, New York, NY.

- Johnson, N. L., and Kotz, S., and Balakrishnan, N. (1994): *Continuous Univariate Distributions, Volume 1* (Second Edition), Wiley, New York, NY.
- Johnson, N. L., and Kotz, S., and Balakrishnan, N. (1995): *Continuous Univariate Distributions, Volume 2* (Second Edition), Wiley, New York, NY.
- Kelly, D. G. (1994): *Introduction to Probability*, Macmillan, New York, NY.
- Lehmann, E. L. (1983): *Theory of Point Estimation* (1991 Reprint), Wadsworth & Brooks/Cole, Pacific Grove, CA.
- Lehmann, E. L. (1986): *Testing Statistical Hypotheses* (Second Edition – 1994 Reprint), Chapman & Hall, New York, NY.
- Mood, A. M., and Graybill, F. A., and Boes, D. C. (1974): *Introduction to the Theory of Statistics* (Third Edition), McGraw-Hill, Singapore.
- Parzen, E. (1960): *Modern Probability Theory and Its Applications*, Wiley, New York, NY.
- Searle, S. R. (1971): *Linear Models*, Wiley, New York, NY.
- Tamhane, A. C., and Dunlop, D. D. (2000): *Statistics and Data Analysis – From Elementary to Intermediate*, Prentice Hall, Upper Saddle River, NJ.

Additional definitions, integrals, sums, etc. originate from the following formula collections:

- Bronstein, I. N. and Semendjajew, K. A. (1985): *Taschenbuch der Mathematik* (22. Auflage), Verlag Harri Deutsch, Thun, German Democratic Republic.
- Bronstein, I. N. and Semendjajew, K. A. (1986): *Ergänzende Kapitel zu Taschenbuch der Mathematik* (4. Auflage), Verlag Harri Deutsch, Thun, German Democratic Republic.
- Sieber, H. (1980): *Mathematische Formeln — Erweiterte Ausgabe E*, Ernst Klett, Stuttgart, Germany.

Jürgen Symanzik, May 7, 2001

Lecture 01:
Mo 01/08/01

6 Limit Theorems (Continued from Stat 6710)

6.3 Strong Laws of Large Numbers

Definition 6.3.1:

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of rv's. Let $T_n = \sum_{i=1}^{n} X_i$. We say that $\{X_i\}$ obeys the SLLN with respect to a sequence of **norming constants** $\{B_i\}_{i=1}^{\infty}$, $B_i > 0, B_i \uparrow \infty$, if there exists a sequence of **centering constants** $\{A_i\}_{i=1}^{\infty}$ such that

$$B_n^{-1}(T_n - A_n) \xrightarrow{a.s.} 0.$$

■

Note:

Unless otherwise specified, we will only use the case that $B_n = n$ in this section. ■

Theorem 6.3.2:

$X_n \xrightarrow{a.s.} X \iff \lim_{n\to\infty} P(\sup_{m\geq n} \mid X_m - X \mid > \epsilon) = 0 \ \ \forall \epsilon > 0.$

Proof: (see also Rohatgi, page 249, Theorem 11)

WLOG, we can assume that $X = 0$ since $X_n \xrightarrow{a.s.} X$ implies $X_n - X \xrightarrow{a.s.} 0$. Thus, we have to prove:

$$X_n \xrightarrow{a.s.} 0 \iff \lim_{n\to\infty} P(\sup_{m\geq n} \mid X_m \mid > \epsilon) = 0 \ \ \forall \epsilon > 0$$

Choose $\epsilon > 0$ and define

$$\begin{aligned} A_n(\epsilon) &= \{\sup_{m\geq n} \mid X_m \mid > \epsilon\} \\ C &= \{\lim_{n\to\infty} X_n = 0\} \end{aligned}$$

"$\Longrightarrow$":

Since $X_n \xrightarrow{a.s.} 0$, we know that $P(C) = 1$ and therefore $P(C^c) = 0$.

Let $B_n(\epsilon) = C \cap A_n(\epsilon)$. Note that $B_{n+1}(\epsilon) \subseteq B_n(\epsilon)$ and for the limit set $\bigcap_{n=1}^{\infty} B_n(\epsilon) = \emptyset$. It follows that

$$\lim_{n\to\infty} P(B_n(\epsilon)) = P(\bigcap_{n=1}^{\infty} B_n(\epsilon)) = 0.$$

We also have

$$\begin{aligned} P(B_n(\epsilon)) &= P(A_n \cap C) \\ &= 1 - P(C^c \cup A_n^c) \\ &= 1 - \underbrace{P(C^c)}_{=0} - P(A_n^c) + \underbrace{P(C^c \cap A_n^C)}_{=0} \\ &= P(A_n) \end{aligned}$$

$$\Longrightarrow \lim_{n\to\infty} P(A_n(\epsilon)) = 0$$

"$\Longleftarrow$":
Assume that $\lim\limits_{n\to\infty} P(A_n(\epsilon)) = 0 \ \ \forall \epsilon > 0$ and define $D(\epsilon) = \{\overline{\lim\limits_{n\to\infty}} \mid X_n \mid > \epsilon\}$.
Since $D(\epsilon) \subseteq A_n(\epsilon) \ \ \forall n \in I\!N$, it follows that $P(D(\epsilon)) = 0 \ \ \forall \epsilon > 0$. Also,

$$C^c = \{\lim_{n\to\infty} X_n \neq 0\} \subseteq \bigcup_{k=1}^{\infty} \{\overline{\lim_{n\to\infty}} \mid X_n \mid > \frac{1}{k}\}.$$

$$\Longrightarrow 1 - P(C) \leq \sum_{k=1}^{\infty} P(D(\frac{1}{k})) = 0$$

$$\Longrightarrow X_n \xrightarrow{a.s.} 0$$

■

<u>Note:</u>

(i) $X_n \xrightarrow{a.s.} 0$ implies that $\forall \epsilon > 0 \ \forall \delta > 0 \ \exists n_0 \in I\!N \ : \ P(\sup\limits_{n \geq n_0} \mid X_n \mid > \epsilon) < \delta$.

(ii) Recall that for a given sequence of events $\{A_n\}_{n=1}^{\infty}$,

$$A = \overline{\lim_{n\to\infty}} A_n = \lim_{n\to\infty} \bigcup_{k=n}^{\infty} A_k = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$$

is the event that infinitely many of the A_n occur. We write $P(A) = P(A_n \ i.o.)$ where $i.o.$ stands for "infinitely often".

(iii) Using the terminology defined in (ii) above, we can rewrite Theorem 6.3.2 as

$$X_n \xrightarrow{a.s.} 0 \iff P(\mid X_n \mid > \epsilon \ i.o.) = 0 \ \ \forall \epsilon > 0.$$

■

Lecture 02: We 01/10/01

Theorem 6.3.3: **Borel–Cantelli Lemma**

(i) 1^{st} **BC–Lemma**:

Let $\{A_n\}_{n=1}^{\infty}$ be a sequence of events such that $\sum\limits_{n=1}^{\infty} P(A_n) < \infty$. Then $P(A) = 0$.

(ii) 2^{nd} **BC–Lemma**:

Let $\{A_n\}_{n=1}^{\infty}$ be a sequence of independent events such that $\sum\limits_{n=1}^{\infty} P(A_n) = \infty$. Then $P(A) = 1$.

Proof:

(i):

$$\begin{aligned} P(A) &= P(\lim_{n\to\infty} \bigcup_{k=n}^{\infty} A_k) \\ &= \lim_{n\to\infty} P(\bigcup_{k=n}^{\infty} A_k) \\ &\leq \lim_{n\to\infty} \sum_{k=n}^{\infty} P(A_k) \\ &= \lim_{n\to\infty} \left(\sum_{k=1}^{\infty} P(A_k) - \sum_{k=1}^{n-1} P(A_k) \right) \\ &\overset{(*)}{=} 0 \end{aligned}$$

$(*)$ holds since $\sum\limits_{n=1}^{\infty} P(A_n) < \infty$.

(ii): We have $A^c = \bigcup\limits_{n=1}^{\infty} \bigcap\limits_{k=n}^{\infty} A_k^c$. Therefore,

$$P(A^c) = P(\lim_{n\to\infty} \bigcap_{k=n}^{\infty} A_k^c) = \lim_{n\to\infty} P(\bigcap_{k=n}^{\infty} A_k^c).$$

If we choose $n_0 > n$, it holds that

$$\bigcap_{k=n}^{\infty} A_k^c \subseteq \bigcap_{k=n}^{n_0} A_k^c.$$

Therefore,

$$\begin{aligned} P(\bigcap_{k=n}^{\infty} A_k^c) &\leq \lim_{n_0\to\infty} P(\bigcap_{k=n}^{n_0} A_k^c) \\ &\overset{indep.}{=} \lim_{n_0\to\infty} \prod_{k=n}^{n_0} (1 - P(A_k)) \end{aligned}$$

$$\begin{aligned} &\overset{(+)}{\leq} \quad \lim_{n_0 \to \infty} \exp\left(- \sum_{k=n}^{n_0} P(A_k) \right) \\ &= \quad 0 \end{aligned}$$

$\Longrightarrow P(A) = 1$

(+) holds since

$$1 - \exp\left(- \sum_{j=n}^{n_0} \alpha_j \right) \leq 1 - \prod_{j=n}^{n_0} (1 - \alpha_j) \leq \sum_{j=n}^{n_0} \alpha_j \quad \text{for } n_0 > n \text{ and } 0 \leq \alpha_j \leq 1$$

■

Example 6.3.4:
Independence is necessary for 2^{nd} BC–Lemma:

Let $\Omega = (0,1)$ and P a uniform distribution on Ω.

Let $A_n = I_{(0, \frac{1}{n})}(\omega)$. Therefore,

$$\sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty.$$

But for any $\omega \in \Omega$, A_n occurs only for $1, 2, \ldots, \lfloor \frac{1}{\omega} \rfloor$, where $\lfloor \frac{1}{\omega} \rfloor$ denotes the largest integer ("floor") that is $\leq \frac{1}{\omega}$. Therefore, $P(A) = P(A_n \ i.o.) = 0$. ■

Lemma 6.3.5: **Kolmogorov's Inequality**
Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of independent rv's with common mean 0 and variances σ_i^2. Let $T_n = \sum_{i=1}^{n} X_i$. Then it holds:

$$P(\max_{1 \leq k \leq n} \mid T_k \mid \geq \epsilon) \leq \frac{\sum_{i=1}^{n} \sigma_i^2}{\epsilon^2} \quad \forall \epsilon > 0$$

Proof:
See Rohatgi, page 268, Lemma 2. ■

Lemma 6.3.6: **Kronecker's Lemma**
For any real numbers x_n, if $\sum_{n=1}^{\infty} x_n$ converges to $s < \infty$ and $B_n \uparrow \infty$, then it holds:

$$\frac{1}{B_n} \sum_{k=1}^{n} B_k x_k \to 0 \quad \text{as } n \to \infty$$

Proof:

See Rohatgi, page 269, Lemma 3. ■

Theorem 6.3.7: **Cauchy Criterion**

$X_n \xrightarrow{a.s.} X \iff \lim_{n\to\infty} P(\sup_m \mid X_{n+m} - X_n \mid \leq \epsilon) = 1 \;\; \forall \epsilon > 0.$

Proof:

See Rohatgi, page 270, Theorem 5. ■

Theorem 6.3.8:

If $\sum_{n=1}^{\infty} Var(X_n) < \infty$, then $\sum_{n=1}^{\infty} (X_n - E(X_n))$ converges almost surely.

Proof:

See Rohatgi, page 272, Theorem 6. ■

Corollary 6.3.9:

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of independent rv's. Let $\{B_i\}_{i=1}^{\infty}$, $B_i > 0, B_i \uparrow \infty$, be a sequence of norming constants. Let $T_n = \sum_{i=1}^{n} X_i$.

If $\sum_{i=1}^{\infty} \frac{Var(X_i)}{B_i^2} < \infty$, then it holds:

$$\frac{T_n - E(T_n)}{B_n} \xrightarrow{a.s.} 0$$

Proof:

This Corollary follows directly from Theorem 6.3.8 and Lemma 6.3.6. ■

Lemma 6.3.10: **Equivalence Lemma**

Let $\{X_i\}_{i=1}^{\infty}$ and $\{X_i'\}_{i=1}^{\infty}$ be sequences of rv's. Let $T_n = \sum_{i=1}^{n} X_i$ and $T_n' = \sum_{i=1}^{n} X_i'$.

If the series $\sum_{i=1}^{\infty} P(X_i \neq X_i') < \infty$, then the series $\{X_i\}$ and $\{X_i'\}$ are **tail–equivalent** and T_n and T_n' are **convergence–equivalent**, i.e., for $B_n \uparrow \infty$ the sequences $\frac{1}{B_n} T_n$ and $\frac{1}{B_n} T_n'$ converge on the same event and to the same limit, except for a null set.

Proof:

See Rohatgi, page 266, Lemma 1. ■

Lemma 6.3.11:
Let X be a rv with $E(\mid X \mid) < \infty$. Then it holds:

$$\sum_{n=1}^{\infty} P(\mid X \mid \geq n) \leq E(\mid X \mid) \leq 1 + \sum_{n=1}^{\infty} P(\mid X \mid \geq n)$$

Proof:
Continuous case only:
Let X have a pdf f. Then it holds:

$$E(\mid X \mid) = \int_{-\infty}^{\infty} \mid x \mid f(x)dx = \sum_{k=0}^{\infty} \int_{k \leq |x| \leq k+1} \mid x \mid f(x)dx$$

$$\Longrightarrow \sum_{k=0}^{\infty} kP(k \leq \mid X \mid \leq k+1) \leq E(\mid X \mid) \leq \sum_{k=0}^{\infty} (k+1)P(k \leq \mid X \mid \leq k+1)$$

It is

$$\begin{aligned} \sum_{k=0}^{\infty} kP(k \leq \mid X \mid \leq k+1) &= \sum_{k=0}^{\infty} \sum_{n=1}^{k} P(k \leq \mid X \mid \leq k+1) \\ &= \sum_{n=1}^{\infty} \sum_{k=n}^{\infty} P(k \leq \mid X \mid \leq k+1) \\ &= \sum_{n=1}^{\infty} P(\mid X \mid \geq n) \end{aligned}$$

Similarly,

$$\begin{aligned} \sum_{k=0}^{\infty} (k+1)P(k \leq \mid X \mid \leq k+1) &= \sum_{n=1}^{\infty} P(\mid X \mid \geq n) + \sum_{k=0}^{\infty} P(k \leq \mid X \mid \leq k+1) \\ &= \sum_{n=1}^{\infty} P(\mid X \mid \geq n) + 1 \end{aligned}$$

■

Lecture 03:
Fr 01/12/01

Theorem 6.3.12:
Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of independent rv's. Then it holds:

$$X_n \xrightarrow{a.s.} 0 \iff \sum_{n=1}^{\infty} P(\mid X_n \mid > \epsilon) < \infty \quad \forall \epsilon > 0$$

Proof:
See Rohatgi, page 265, Theorem 3. ■

Theorem 6.3.13: **Kolmogorov's SLLN**

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of iid rv's. Let $T_n = \sum\limits_{i=1}^{n} X_i$. Then it holds:

$$\frac{T_n}{n} = \overline{X}_n \xrightarrow{a.s.} \mu < \infty \iff E(\mid X \mid) < \infty \text{ (and then } \mu = E(X))$$

Proof:

"$\Longrightarrow$":

Suppose that $\overline{X}_n \xrightarrow{a.s.} \mu < \infty$. It is

$$T_n = \sum_{i=1}^{n} X_i = \sum_{i=1}^{n-1} X_i + X_n = T_{n-1} + X_n.$$

$$\Longrightarrow \tfrac{X_n}{n} = \underbrace{\frac{T_n}{n}}_{\xrightarrow{a.s.} \mu} - \underbrace{\frac{n-1}{n}}_{\to 1} \underbrace{\frac{T_{n-1}}{n-1}}_{\xrightarrow{a.s.} \mu} \xrightarrow{a.s.} 0$$

By Theorem 6.3.12, we have

$$\sum_{n=1}^{\infty} P(\mid \frac{X_n}{n} \mid \geq 1) < \infty, \text{ i.e., } \sum_{n=1}^{\infty} P(\mid X_n \mid \geq n) < \infty.$$

$$\overset{Lemma\ 6.3.11}{\Longrightarrow} E(\mid X \mid) < \infty$$

$$\overset{Th.\ 6.2.7\ (WLLN)}{\Longrightarrow} \overline{X}_n \xrightarrow{p} E(X)$$

Since $\overline{X}_n \xrightarrow{a.s.} \mu$, it follows by Theorem 6.1.26 that $\overline{X}_n \xrightarrow{p} \mu$. Therefore, it must hold that $\mu = E(X)$.

"$\Longleftarrow$":

Let $E(\mid X \mid) < \infty$. Define truncated rv's:

$$\begin{aligned} X_k' &= \begin{cases} X_k, & \text{if } \mid X_k \mid \leq k \\ 0, & \text{otherwise} \end{cases} \\ T_n' &= \sum_{k=1}^{n} X_k' \\ \overline{X}_n' &= \frac{T_n'}{n} \end{aligned}$$

Then it holds:

$$\begin{aligned} \sum_{k=1}^{\infty} P(X_k \neq X_k') &= \sum_{k=1}^{\infty} P(\mid X_k \mid > k) \\ &\leq \sum_{k=1}^{\infty} P(\mid X_k \mid \geq k) \end{aligned}$$

$$
\begin{aligned}
&\stackrel{iid}{=} && \sum_{k=1}^{\infty} P(\mid X \mid \geq k) \\
&\stackrel{Lemma\ 6.3.11}{\leq} && E(\mid X \mid) \\
&< && \infty
\end{aligned}
$$

By Lemma 6.3.10, it follows that T_n and T_n' are convergence–equivalent. Thus, it is sufficient to prove that $\overline{X}_n' \stackrel{a.s.}{\longrightarrow} E(X)$.

We now establish the conditions needed in Corollary 6.3.9. It is

$$
\begin{aligned}
Var(X_n') &\leq E((X_n')^2) \\
&= \int_{-n}^{n} x^2 f_X(x) dx \\
&= \sum_{k=0}^{n-1} \int_{k \leq |x| < k+1} x^2 f_X(x) dx \\
&\leq \sum_{k=0}^{n-1} (k+1)^2 \int_{k \leq |x| < k+1} f_X(x) dx \\
&= \sum_{k=0}^{n-1} (k+1)^2 P(k \leq \mid X \mid < k+1) \\
&\leq \sum_{k=0}^{n} (k+1)^2 P(k \leq \mid X \mid < k+1)
\end{aligned}
$$

$$
\begin{aligned}
\Longrightarrow \sum_{n=1}^{\infty} \frac{1}{n^2} Var(X_n') &\leq \sum_{n=1}^{\infty} \sum_{k=0}^{n} \frac{(k+1)^2}{n^2} P(k \leq \mid X \mid < k+1) \\
&= \sum_{n=1}^{\infty} \sum_{k=1}^{n} \frac{(k+1)^2}{n^2} P(k \leq \mid X \mid < k+1) + \sum_{n=1}^{\infty} \frac{1}{n^2} P(0 \leq \mid X \mid < 1) \\
&\stackrel{(*)}{\leq} \sum_{k=1}^{\infty} (k+1)^2 P(k \leq \mid X \mid < k+1) \left(\sum_{n=k}^{\infty} \frac{1}{n^2} \right) + 2P(0 \leq \mid X \mid < 1) \quad (A)
\end{aligned}
$$

$(*)$ holds since $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} \approx 1.65 < 2$ and the first two sums can be rearranged as follows:

$$
\begin{array}{c|c}
n & k \\
\hline
1 & 1 \\
2 & 1,2 \\
3 & 1,2,3 \\
\vdots & \vdots
\end{array}
\quad \Longrightarrow \quad
\begin{array}{c|c}
k & n \\
\hline
1 & 1,2,3,\ldots \\
2 & 2,3,\ldots \\
3 & 3,\ldots \\
\vdots & \vdots
\end{array}
$$

It is

$$\begin{aligned}
\sum_{n=k}^{\infty}\frac{1}{n^2} &= \frac{1}{k^2}+\frac{1}{(k+1)^2}+\frac{1}{(k+2)^2}+\ldots \\
&\leq \frac{1}{k^2}+\frac{1}{k(k+1)}+\frac{1}{(k+1)(k+2)}+\ldots \\
&= \frac{1}{k^2}+\sum_{n=k+1}^{\infty}\frac{1}{n(n-1)}
\end{aligned}$$

From Bronstein, page 30, # 7, we know that

$$\begin{aligned}
1 &= \frac{1}{1\cdot 2}+\frac{1}{2\cdot 3}+\frac{1}{3\cdot 4}+\ldots+\frac{1}{n(n+1)}+\ldots \\
&= \frac{1}{1\cdot 2}+\frac{1}{2\cdot 3}+\frac{1}{3\cdot 4}+\ldots+\frac{1}{(k-1)\cdot k}+\sum_{n=k+1}^{\infty}\frac{1}{n(n-1)}
\end{aligned}$$

$$\begin{aligned}
\Longrightarrow \sum_{n=k+1}^{\infty}\frac{1}{n(n-1)} &= 1-\frac{1}{1\cdot 2}-\frac{1}{2\cdot 3}-\frac{1}{3\cdot 4}-\ldots-\frac{1}{(k-1)\cdot k} \\
&= \frac{1}{2}-\frac{1}{2\cdot 3}-\frac{1}{3\cdot 4}-\ldots-\frac{1}{(k-1)\cdot k} \\
&= \frac{1}{3}-\frac{1}{3\cdot 4}-\ldots-\frac{1}{(k-1)\cdot k} \\
&= \frac{1}{4}-\ldots-\frac{1}{(k-1)\cdot k} \\
&= \ldots \\
&= \frac{1}{k}
\end{aligned}$$

$$\begin{aligned}
\Longrightarrow \sum_{n=k}^{\infty}\frac{1}{n^2} &\leq \frac{1}{k^2}+\sum_{n=k+1}^{\infty}\frac{1}{n(n-1)} \\
&= \frac{1}{k^2}+\frac{1}{k} \\
&\leq \frac{2}{k}
\end{aligned}$$

Using this result in (A), we get

$$\begin{aligned}
\sum_{n=1}^{\infty} \frac{1}{n^2} Var(X_n') &\leq 2 \sum_{k=1}^{\infty} \frac{(k+1)^2}{k} P(k \leq \mid X \mid < k+1) + 2P(0 \leq \mid X \mid < 1) \\
&= 2 \sum_{k=0}^{\infty} kP(k \leq \mid X \mid < k+1) + 4 \sum_{k=1}^{\infty} P(k \leq \mid X \mid < k+1) \\
&\quad + 2 \sum_{k=1}^{\infty} \frac{1}{k} P(k \leq \mid X \mid < k+1) + 2P(0 \leq \mid X \mid < 1) \\
&\overset{(B)}{\leq} 2E(\mid X \mid) + 4 + 2 + 2 \\
&< \infty
\end{aligned}$$

To establish (B), we use an inequality from the Proof of Lemma 6.3.11, i.e.,

$$\sum_{k=0}^{\infty} kP(k \leq \mid X \mid < k+1) \overset{Proof}{\leq} \sum_{n=1}^{\infty} P(\mid X \mid \geq n) \overset{Lemma\ 6.3.11}{\leq} E(\mid X \mid)$$

Thus, the conditions needed in Corollary 6.3.9 are met. With $B_n = n$, it follows that

$$\frac{1}{n} T_n' - \frac{1}{n} E(T_n') \xrightarrow{a.s.} 0 \quad (C)$$

Since $E(X_n') \to E(X)$ as $n \to \infty$, it follows by Kronecker's Lemma (6.3.6) that $\frac{1}{n} E(T_n') \to E(X)$. Thus, when we replace $\frac{1}{n} E(T_n')$ by $E(X)$ in (C), we get

$$\frac{1}{n} T_n' \xrightarrow{a.s.} E(X) \quad \overset{Lemma\ 6.3.10}{\Longrightarrow} \quad \frac{1}{n} T_n \xrightarrow{a.s.} E(X)$$

since T_n and T_n' are convergence–equivalent. ■

Lecture 04: We 01/17/01

6.4 Central Limit Theorems

Let $\{X_n\}_{n=1}^\infty$ be a sequence of rv's with cdf's $\{F_n\}_{n=1}^\infty$. Suppose that the mgf $M_n(t)$ of X_n exists.

Questions: Does $M_n(t)$ converge? Does it converge to a mgf $M(t)$? If it does converge, does it hold that $X_n \xrightarrow{d} X$ for some rv X?

Example 6.4.1:
Let $\{X_n\}_{n=1}^\infty$ be a sequence of rv's such that $P(X_n = -n) = 1$. Then the mgf is $M_n(t) = E(e^{tX}) = e^{-tn}$. So

$$\lim_{n\to\infty} M_n(t) = \begin{cases} 0, & t > 0 \\ 1, & t = 0 \\ \infty, & t < 0 \end{cases}$$

So $M_n(t)$ does not converge to a mgf and $F_n(x) \to F(x) = 1 \ \ \forall x$. But $F(x)$ is not a cdf. ■

Note:
Due to Example 6.4.1, the existence of mgf's $M_n(t)$ that converge to something is not enough to conclude convergence in distribution.

Conversely, suppose that X_n has mgf $M_n(t)$, X has mgf $M(t)$, and $X_n \xrightarrow{d} X$. Does it hold that

$$M_n(t) \to M(t)?$$

Not necessarily! See Rohatgi, page 277, Example 2, as a counter example. Thus, convergence in distribution of rv's that all have mgf's does not imply the convergence of mgf's.

However, we can make the following statement in the next Theorem: ■

Theorem 6.4.2: **Continuity Theorem**
Let $\{X_n\}_{n=1}^\infty$ be a sequence of rv's with cdf's $\{F_n\}_{n=1}^\infty$ and mgf's $\{M_n(t)\}_{n=1}^\infty$. Suppose that $M_n(t)$ exists for $\mid t \mid \leq t_0 \ \ \forall n$. If there exists a rv X with cdf F and mgf $M(t)$ which exists for $\mid t \mid \leq t_1 < t_0$ such that $\lim_{n\to\infty} M_n(t) = M(t) \ \ \forall t \in [-t_1, t_1]$, then $F_n \xrightarrow{w} F$, i.e., $X_n \xrightarrow{d} X$. ■

Example 6.4.3:
Let $X_n \sim Bin(n, \frac{\lambda}{n})$. Recall (e.g., from Theorem 3.3.12 and related Theorems) that for $X \sim Bin(n,p)$ the mgf is $M_X(t) = (1-p+pe^t)^n$. Thus,

$$M_n(t) = (1 - \frac{\lambda}{n} + \frac{\lambda}{n}e^t)^n = (1 + \frac{\lambda(e^t-1)}{n})^n \overset{(*)}{\longrightarrow} e^{\lambda(e^t-1)} \text{ as } n \to \infty.$$

In $(*)$ we use the fact that $\lim\limits_{n\to\infty}(1+\frac{x}{n})^n = e^x$. Recall that $e^{\lambda(e^t-1)}$ is the mgf of a rv X where $X \sim Poisson(\lambda)$. Thus, we have established the well–known result that the Binomial distribution approaches the Poisson distribution, given that $n \to \infty$ in such a way that $np = \lambda > 0$. ■

Note:
Recall Theorem 3.3.11: Suppose that $\{X_n\}_{n=1}^{\infty}$ is a sequence of rv's with characteristic fuctions $\{\Phi_n(t)\}_{n=1}^{\infty}$. Suppose that

$$\lim_{n\to\infty} \Phi_n(t) = \Phi(t) \quad \forall t \in (-h,h) \text{ for some } h > 0,$$

and $\Phi(t)$ is the characteristic function of a rv X. Then $X_n \xrightarrow{d} X$. ■

Theorem 6.4.4: **Lindeberg–Lévy Central Limit Theorem**
Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of iid rv's with $E(X_i) = \mu$ and $0 < Var(X_i) = \sigma^2 < \infty$. Then it holds for $\overline{X}_n = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$ that

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} Z$$

where $Z \sim N(0,1)$.

Proof:
Let $Z \sim N(0,1)$. According to Theorem 3.3.12 (v), the characteristic function of Z is $\Phi_Z(t) = \exp(-\frac{1}{2}t^2)$.

Let $\Phi(t)$ be the characteristic function of X_i. We now determine the characteristic function $\Phi_n(t)$ of $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$:

$$\begin{aligned}
\Phi_n(t) &= E\left(\exp\left(it\frac{\sqrt{n}(\frac{1}{n}\sum\limits_{i=1}^{n} X_i - \mu)}{\sigma}\right)\right) \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(it\frac{\sqrt{n}(\frac{1}{n}\sum\limits_{i=1}^{n} X_i - \mu)}{\sigma}\right) dF_{\underline{X}}
\end{aligned}$$

$$
\begin{aligned}
&= \exp(-\frac{it\sqrt{n}\mu}{\sigma}) \int_{-\infty}^{\infty} \exp(\frac{itx_1}{\sqrt{n}\sigma}) dF_{X_1} \ldots \int_{-\infty}^{\infty} \exp(\frac{itx_n}{\sqrt{n}\sigma}) dF_{X_n} \\
&= \left(\Phi(\frac{t}{\sqrt{n}\sigma}) \exp(-\frac{it\mu}{\sqrt{n}\sigma})\right)^n
\end{aligned}
$$

Recall from Theorem 3.3.5 that if the k^{th} moment exists, then $\Phi^{(k)}(0) = i^k E(X^k)$. Also recall the definition of a **Taylor series** in MacLaurin's form:

$$
f(x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \ldots + \frac{f^{(n)}(0)}{n!}x^n + \ldots,
$$

e.g.,

$$
f(x) = e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots
$$

Thus, if we develop a Taylor series for $\Phi(\frac{t}{\sqrt{n}\sigma})$ around $t = 0$, we get:

$$
\begin{aligned}
\Phi(\frac{t}{\sqrt{n}\sigma}) &= \Phi(0) + t\Phi'(0) + \frac{1}{2}t^2\Phi''(0) + \frac{1}{6}t^3\Phi'''(0) + \ldots \\
&= 1 + t\frac{i\mu}{\sqrt{n}\sigma} - \frac{1}{2}t^2\frac{\mu^2 + \sigma^2}{n\sigma^2} + o\left((\frac{t}{\sqrt{n}\sigma})^2\right)
\end{aligned}
$$

Here we make use of the **Landau symbol** "o". In general, if we write $u(x) = o(v(x))$ for $x \to L$, this implies $\lim_{x \to L} \frac{u(x)}{v(x)} = 0$, i.e., $u(x)$ goes to 0 faster than $v(x)$ or $v(x)$ goes to ∞ faster than $u(x)$. We say that $u(x)$ *is of smaller order than* $v(x)$ as $x \to L$. Examples are $\frac{1}{x^3} = o(\frac{1}{x^2})$ and $x^2 = o(x^3)$ for $x \to \infty$. See Rohatgi, page 6, for more details on the **Landau symbols** "O" and "o".

Similarly, if we develop a Taylor series for $\exp(-\frac{it\mu}{\sqrt{n}\sigma})$ around $t = 0$, we get:

$$
\exp(-\frac{it\mu}{\sqrt{n}\sigma}) = 1 - t\frac{i\mu}{\sqrt{n}\sigma} - \frac{1}{2}t^2\frac{\mu^2}{n\sigma^2} + o\left((\frac{t}{\sqrt{n}\sigma})^2\right)
$$

Combining these results, we get:

$$
\begin{aligned}
\Phi_n(t) &= \left(\left(1 + t\frac{i\mu}{\sqrt{n}\sigma} - \frac{1}{2}t^2\frac{\mu^2 + \sigma^2}{n\sigma^2} + o\left((\frac{t}{\sqrt{n}\sigma})^2\right)\right)\left(1 - t\frac{i\mu}{\sqrt{n}\sigma} - \frac{1}{2}t^2\frac{\mu^2}{n\sigma^2} + o\left((\frac{t}{\sqrt{n}\sigma})^2\right)\right)\right)^n \\
&= \left(1 - t\frac{i\mu}{\sqrt{n}\sigma} - \frac{1}{2}t^2\frac{\mu^2}{n\sigma^2} + t\frac{i\mu}{\sqrt{n}\sigma} + t^2\frac{\mu^2}{n\sigma^2} - \frac{1}{2}t^2\frac{\mu^2 + \sigma^2}{n\sigma^2} + o\left((\frac{t}{\sqrt{n}\sigma})^2\right)\right)^n \\
&= \left(1 - \frac{1}{2}\frac{t^2}{n} + o\left((\frac{t}{\sqrt{n}\sigma})^2\right)\right)^n \\
&\overset{(*)}{\longrightarrow} \exp(-\frac{t^2}{2}) \quad \text{as } n \to \infty
\end{aligned}
$$

Thus, $\lim_{n\to\infty} \Phi_n(t) = \Phi_Z(t) \;\; \forall t$. For a proof of $(*)$, see Rohatgi, page 278, Lemma 1. According to the Note above, it holds that

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} Z.$$

■

Lecture 05: Fr 01/19/01

Definition 6.4.5:
Let X_1, X_2 be iid non–degenerate rv's with common cdf F. Let $a_1, a_2 > 0$. We say that F is **stable** if there exist constants A and B (depending on a_1 and a_2) such that $B^{-1}(a_1 X_1 + a_2 X_2 - A)$ also has cdf F. ■

Note:
When generalizing the previous definition to sequences of rv's, we have the following examples for stable distributions:

- X_i iid Cauchy. Then $\frac{1}{n}\sum_{i=1}^{n} X_i \sim$ Cauchy (here $B_n = n, A_n = 0$).
- X_i iid $N(0,1)$. Then $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \sim N(0,1)$ (here $B_n = \sqrt{n}, A_n = 0$).

■

Definition 6.4.6:
Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of iid rv's with common cdf F. Let $T_n = \sum_{i=1}^{n} X_i$. F belongs to the **domain of attraction** of a distribution V if there exist norming and centering constants $\{B_n\}_{n=1}^{\infty}, B_n > 0$, and $\{A_n\}_{n=1}^{\infty}$ such that

$$P(B_n^{-1}(T_n - A_n) \leq x) = F_{B_n^{-1}(T_n - A_n)}(x) \to V(x) \text{ as } n \to \infty$$

at all continuity points x of V. ■

Note:
A very general Theorem from Loève states that only stable distributions can have domains of attraction. From the practical point of view, a wide class of distributions F belong to the domain of attraction of the Normal distribution. ■

Theorem 6.4.7: **Lindeberg Central Limit Theorem**
Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of independent non–degenerate rv's with cdf's $\{F_i\}_{i=1}^{\infty}$. Assume that $E(X_k) = \mu_k$ and $Var(X_k) = \sigma_k^2 < \infty$. Let $s_n^2 = \sum\limits_{k=1}^{n} \sigma_k^2$.

If the F_k are absolutely continuous with pdf's $f_k = F_k'$, assume that it holds for all $\epsilon > 0$ that

$$(A) \quad \lim_{n\to\infty} \frac{1}{s_n^2} \sum_{k=1}^{n} \int_{\{|x-\mu_k|>\epsilon s_n\}} (x-\mu_k)^2 F_k'(x)dx = 0.$$

If the X_k are discrete rv's with support $\{x_{kl}\}$ and probabilities $\{p_{kl}\}$, $l = 1, 2, \ldots$, assume that it holds for all $\epsilon > 0$ that

$$(B) \quad \lim_{n\to\infty} \frac{1}{s_n^2} \sum_{k=1}^{n} \sum_{|x_{kl}-\mu_k|>\epsilon s_n} (x_{kl}-\mu_k)^2 p_{kl} = 0.$$

The conditions (A) and (B) are called **Lindeberg Condition** (LC). If either LC holds, then

$$\frac{\sum\limits_{k=1}^{n}(X_k - \mu_k)}{s_n} \xrightarrow{d} Z$$

where $Z \sim N(0,1)$.

Proof:
Similar to the proof of Theorem 6.4.4, we can use characteristic functions again. An alternative proof is given in Rohatgi, pages 282–288. ■

Note:
Feller shows that the LC is a necessary condition if $\frac{\sigma_n^2}{s_n^2} \to 0$ and $s_n^2 \to \infty$ as $n \to \infty$. ■

Corollary 6.4.8:
Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of iid rv's such that $\frac{1}{\sqrt{n}} \sum\limits_{i=1}^{n} X_i$ has the same distribution for all n. If $E(X_i) = 0$ and $Var(X_i) = 1$, then $X_i \sim N(0,1)$.

Proof:
Let F be the common cdf of $\frac{1}{\sqrt{n}} \sum\limits_{i=1}^{n} X_i$ for all n (including $n = 1$). By the CLT,

$$\lim_{n\to\infty} P(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \leq x) = \Phi(x),$$

where $\Phi(x)$ denotes $P(Z \leq x)$ for $Z \sim N(0,1)$. Also, $P(\frac{1}{\sqrt{n}} \sum\limits_{i=1}^{n} X_i \leq x) = F(x)$ for each n. Therefore, we must have $F(x) = \Phi(x)$. ■

Note:
In general, if $X_1, X_2, \ldots,$ are independent rv's such that there exists a constant A with $P(\mid X_n \mid \leq A) = 1 \ \ \forall n$, then the LC is satisfied if $s_n^2 \to \infty$ as $n \to \infty$. Why??

Suppose that $s_n^2 \to \infty$ as $n \to \infty$. Since the $\mid X_k \mid$'s are uniformly bounded (by A), so are the rv's $(X_k - E(X_k))$. Thus, for every $\epsilon > 0$ there exists an N_ϵ such that if $n \geq N_\epsilon$ then

$$P(\mid X_k - E(X_k) \mid < \epsilon s_n, \quad k = 1, \ldots, n) = 1.$$

This implies that the LC holds since we would integrate (or sum) over the empty set, i.e., the set $\{\mid x - \mu_k \mid > \epsilon s_n\} = \emptyset$.

The converse also holds. For a sequence of uniformly bounded independent rv's, a necessary and sufficient condition for the CLT to hold is that $s_n^2 \to \infty$ as $n \to \infty$. ■

Example 6.4.9:
Let $\{X_i\}_{i=1}^\infty$ be a sequence of independent rv's such that $E(X_k) = 0$, $\alpha_k = E(\mid X_k \mid^{2+\delta}) < \infty$ for some $\delta > 0$, and $\sum\limits_{k=1}^n \alpha_k = o(s_n^{2+\delta})$.

Does the LC hold? It is:

$$\begin{aligned}
\frac{1}{s_n^2} \sum_{k=1}^n \int_{\{\mid x \mid > \epsilon s_n\}} x^2 f_k(x) dx &\overset{(A)}{\leq} \frac{1}{s_n^2} \sum_{k=1}^n \int_{\{\mid x \mid > \epsilon s_n\}} \frac{\mid x \mid^{2+\delta}}{\epsilon^\delta s_n^\delta} f_k(x) dx \\
&\leq \frac{1}{s_n^2 \epsilon^\delta s_n^\delta} \sum_{k=1}^n \int_{-\infty}^{\infty} \mid x \mid^{2+\delta} f_k(x) dx \\
&= \frac{1}{s_n^2 \epsilon^\delta s_n^\delta} \sum_{k=1}^n \alpha_k \\
&= \frac{1}{\epsilon^\delta} \left(\frac{\sum\limits_{k=1}^n \alpha_k}{s_n^{2+\delta}} \right) \\
&\overset{(B)}{\longrightarrow} 0 \quad \text{as } n \to \infty
\end{aligned}$$

(A) holds since for $\mid x \mid > \epsilon s_n$, it is $\frac{|x|^\delta}{\epsilon^\delta s_n^\delta} > 1$. (B) holds since $\sum\limits_{k=1}^n \alpha_k = o(s_n^{2+\delta})$.

Thus, the LC is satisfied and the CLT holds. ■

Note:

(i) In general, if there exists a $\delta > 0$ such that

$$\frac{\sum_{k=1}^{n} E(\mid X_k - \mu_k \mid^{2+\delta})}{s_n^{2+\delta}} \longrightarrow 0 \text{ as } n \to \infty,$$

then the LC holds.

(ii) Both the CLT and the WLLN hold for a large class of sequences of rv's $\{X_i\}_{i=1}^n$. If the $\{X_i\}$'s are independent uniformly bounded rv's, i.e., if $P(\mid X_n \mid \leq M) = 1 \;\; \forall n$, the WLLN (as formulated in Theorem 6.2.3) holds. The CLT holds provided that $s_n^2 \to \infty$ as $n \to \infty$.

If the rv's $\{X_i\}$ are iid, then the CLT is a stronger result than the WLLN since the CLT provides an estimate of the probability $P(\frac{1}{n} \mid \sum_{i=1}^{n} X_i - n\mu \mid \geq \epsilon) \approx 1 - P(\mid Z \mid \leq \frac{\epsilon}{\sigma}\sqrt{n})$, where $Z \sim N(0,1)$, and the WLLN follows. However, note that the CLT requires the existence of a 2^{nd} moment while the WLLN does not.

(iii) If the $\{X_i\}$ are independent (but not identically distributed) rv's, the CLT may apply while the WLLN does not.

(iv) See Rohatgi, pages 289–293, for additional details and examples.

■

7 Sample Moments

7.1 Random Sampling

Definition 7.1.1:
Let $X_1, \ldots, X_n$ be iid rv's with common cdf F. We say that $\{X_1, \ldots, X_n\}$ is a **(random) sample** of size n from the **population distribution** F. The vector of values $\{x_1, \ldots, x_n\}$ is called a **realization** of the sample. A rv $g(X_1, \ldots, X_n)$ which is a Borel–measurable function of $X_1, \ldots, X_n$ and does not depend on any unknown parameter is called a **(sample) statistic**.

■

Definition 7.1.2:
Let $X_1, \ldots, X_n$ be a sample of size n from a population with distribution F. Then

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

is called the **sample mean** and

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

is called the **sample variance**. ■

Definition 7.1.3:
Let $X_1, \ldots, X_n$ be a sample of size n from a population with distribution F. The function

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(X_i)$$

is called **empirical cumulative distribution function** (**empirical cdf**). ■

Note:
For any fixed $x \in I\!R$, $\hat{F}_n(x)$ is a rv. ■

Theorem 7.1.4:
The rv $\hat{F}_n(x)$ has pmf

$$P(\hat{F}_n(x) = \frac{j}{n}) = \binom{n}{j} (F(x))^j (1 - F(x))^{n-j}, \quad j \in \{0, 1, \ldots, n\},$$

with $E(\hat{F}_n(x)) = F(x)$ and $Var(\hat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}$.

Proof:

It is $I_{(-\infty,x]}(X_i) \sim Bin(1, F(x))$. Then $n\hat{F}_n(x) \sim Bin(n, F(x))$.

The results follow immediately. ■

Corollary 7.1.5:

By the WLLN, it follows that

$$\hat{F}_n(x) \xrightarrow{p} F(x).$$

■

Corollary 7.1.6:

By the CLT, it follows that

$$\frac{\sqrt{n}(\hat{F}_n(x) - F(x))}{\sqrt{F(x)(1 - F(x))}} \xrightarrow{d} Z,$$

where $Z \sim N(0, 1)$. ■

Theorem 7.1.7: **Glivenko–Cantelli Theorem**

$\hat{F}_n(x)$ converges uniformly to $F(x)$, i.e., it holds for all $\epsilon > 0$ that

$$\lim_{n\to\infty} P\left(\sup_{-\infty<x<\infty} \mid \hat{F}_n(x) - F(x) \mid > \epsilon\right) = 0.$$

■

Lecture 06:
Mo 01/22/01

Definition 7.1.8:

Let $X_1, \ldots, X_n$ be a sample of size n from a population with distribution F. We call

$$a_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

the **sample moment of order** k and

$$b_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - a_1)^k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$$

the **sample central moment of order** k. ■

Note:

It is $b_1 = 0$ and $b_2 = \frac{n-1}{n}S^2$. ■

Theorem 7.1.9:
Let $X_1, \ldots, X_n$ be a sample of size n from a population with distribution F. Assume that $E(X) = \mu$, $Var(X) = \sigma^2$, and $E((X-\mu)^k) = \mu_k$ exist. Then it holds:

(i) $E(a_1) = E(\overline{X}) = \mu$

(ii) $Var(a_1) = Var(\overline{X}) = \frac{\sigma^2}{n}$

(iii) $E(b_2) = \frac{n-1}{n}\sigma^2$

(iv) $Var(b_2) = \frac{\mu_4 - \mu_2^2}{n} - \frac{2(\mu_4 - 2\mu_2^2)}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}$

(v) $E(S^2) = \sigma^2$

(vi) $Var(S^2) = \frac{\mu_4}{n} - \frac{n-3}{n(n-1)}\mu_2^2$

Proof:
(i)

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{n}{n}\mu = \mu$$

(ii)

$$Var(\overline{X}) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n}$$

(iii)

$$\begin{aligned}
E(b_2) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n} X_i\right)^2\right) \\
&= E(X^2) - \frac{1}{n^2}E\left(\sum_{i=1}^{n} X_i^2 + \sum\sum_{i \neq j} X_i X_j\right) \\
&\overset{(*)}{=} E(X^2) - \frac{1}{n^2}(nE(X^2) + n(n-1)\mu^2) \\
&= \frac{n-1}{n}(E(X^2) - \mu^2) \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}$$

$(*)$ holds since X_i and X_j are independent and then, due to Theorem 4.5.3, it holds that $E(X_i X_j) = E(X_i)E(X_j)$.

See Rohatgi, page 303–306, for the proof of parts (iv) through (vi) and results regarding the 3^{rd} and 4^{th} moments and covariances. ■

7.2 Sample Moments and the Normal Distribution

Theorem 7.2.1:

Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ rv's. Then $\overline{X} = \frac{1}{n} \sum\limits_{i=1}^{n} X_i$ and $(X_1 - \overline{X}, \ldots, X_n - \overline{X})$ are independent.

Proof:

By computing the joint mgf of $(\overline{X}, X_1 - \overline{X}, X_2 - \overline{X}, \ldots, X_n - \overline{X})$, we can use Theorem 4.6.3 (iv) to show independence. We will use the following two facts:

(1):

$$
\begin{aligned}
M_{\overline{X}}(t) &= M_{\left(\frac{1}{n}\sum\limits_{i=1}^{n} X_i\right)}(t) \\
&\overset{(A)}{=} \prod_{i=1}^{n} M_{X_i}\left(\frac{t}{n}\right) \\
&\overset{(B)}{=} \left[\exp\left(\frac{t}{n}\mu + \frac{\sigma^2 t^2}{2n^2}\right)\right]^n \\
&= \exp\left(t\mu + \frac{\sigma^2 t^2}{2n}\right)
\end{aligned}
$$

(A) holds by Theorem 4.6.4 (i). (B) follows from Theorem 3.3.12 (vi) since the X_i's are iid.

(2):

$$
\begin{aligned}
M_{X_1-\overline{X}, X_2-\overline{X}, \ldots, X_n-\overline{X}}(t_1, t_2, \ldots, t_n) &\overset{Def.4.6.1}{=} E\left[\exp\left(\sum_{i=1}^{n} t_i (X_i - \overline{X})\right)\right] \\
&= E\left[\exp\left(\sum_{i=1}^{n} t_i X_i - \overline{X} \sum_{i=1}^{n} t_i\right)\right] \\
&= E\left[\exp\left(\sum_{i=1}^{n} X_i (t_i - \overline{t})\right)\right], \qquad \text{where } \overline{t} = \frac{1}{n}\sum_{i=1}^{n} t_i \\
&= E\left[\prod_{i=1}^{n} \exp\left(X_i (t_i - \overline{t})\right)\right] \\
&\overset{(C)}{=} \prod_{i=1}^{n} E(\exp(X_i (t_i - \overline{t}))) \\
&= \prod_{i=1}^{n} M_{X_i}(t_i - \overline{t})
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(D)}{=} && \prod_{i=1}^{n} \exp\left(\mu(t_i - \overline{t}) + \frac{\sigma^2 (t_i - \overline{t})^2}{2}\right) \\
&= && \exp\left(\mu \underbrace{\sum_{i=1}^{n}(t_i - \overline{t})}_{=0} + \frac{\sigma^2}{2}\sum_{i=1}^{n}(t_i - \overline{t})^2\right) \\
&= && \exp\left(\frac{\sigma^2}{2}\sum_{i=1}^{n}(t_i - \overline{t})^2\right)
\end{aligned}
$$

(C) follows from Theorem 4.5.3 since the X_i's are independent. (D) holds since we evaluate $M_X(h) = \exp(\mu h + \frac{\sigma^2 h^2}{2})$ for $h = t_i - \overline{t}$.

From (1) and (2), it follows:

$$
\begin{aligned}
M_{\overline{X}, X_1 - \overline{X}, \ldots, X_n - \overline{X}}(t, t_1, \ldots, t_n) &\overset{Def. 4.6.1}{=} E\left[\exp(t\overline{X} + t_1(X_1 - \overline{X}) + \ldots + t_n(X_n - \overline{X}))\right] \\
&= E\left[\exp(t\overline{X} + t_1 X_1 - t_1\overline{X} + \ldots + t_n X_n - t_n\overline{X})\right] \\
&= E\left[\exp\left(\sum_{i=1}^{n} X_i t_i - (\sum_{i=1}^{n} t_i - t)\overline{X}\right)\right] \\
&= E\left[\exp\left(\sum_{i=1}^{n} X_i t_i - \frac{(t_1 + \ldots + t_n - t)\sum_{i=1}^{n} X_i}{n}\right)\right] \\
&= E\left[\exp\left(\sum_{i=1}^{n} X_i (t_i - \frac{t_1 + \ldots + t_n - t}{n})\right)\right] \\
&= E\left[\prod_{i=1}^{n} \exp\left(X_i \frac{n t_i - n\overline{t} + t}{n}\right)\right], \quad \text{where } \overline{t} = \frac{1}{n}\sum_{i=1}^{n} t_i \\
&\overset{(E)}{=} \prod_{i=1}^{n} E\left[\exp\left(\frac{X_i[t + n(t_i - \overline{t})]}{n}\right)\right] \\
&= \prod_{i=1}^{n} M_{X_i}\left(\frac{t + n(t_i - \overline{t})}{n}\right) \\
&\overset{(F)}{=} \prod_{i=1}^{n} \exp\left(\frac{\mu[t + n(t_i - \overline{t})]}{n} + \frac{\sigma^2}{2}\frac{1}{n^2}[t + n(t_i - \overline{t})]^2\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \exp\left(\frac{\mu}{n}\left(nt + n\underbrace{\sum_{i=1}^{n}(t_i - \bar{t})}_{=0}\right) + \frac{\sigma^2}{2n^2}\sum_{i=1}^{n}\left(t + n(t_i - \bar{t})\right)^2\right) \\
&= \exp(\mu t)\exp\left(\frac{\sigma^2}{2n^2}\left(nt^2 + 2nt\underbrace{\sum_{i=1}^{n}(t_i - \bar{t})}_{=0} + n^2\sum_{i=1}^{n}(t_i - \bar{t})^2\right)\right) \\
&= \exp\left(\mu t + \frac{\sigma^2}{2n}t^2\right)\exp\left(\frac{\sigma^2}{2}\sum_{i=1}^{n}(t_i - \bar{t})^2\right) \\
&\overset{(1)\&(2)}{=} M_{\overline{X}}(t)M_{X_1-\overline{X},\ldots,X_n-\overline{X}}(t_1,\ldots,t_n)
\end{aligned}
$$

Thus, $\overline{X}$ and $(X_1 - \overline{X}, \ldots, X_n - \overline{X})$ are independent by Theorem 4.6.3 (iv). (E) follows from Theorem 4.5.3 since the X_i's are independent. (F) holds since we evaluate $M_X(h) = \exp(\mu h + \frac{\sigma^2 h^2}{2})$ for $h = \frac{t+n(t_i-\bar{t})}{n}$. ■

Corollary 7.2.2:
$\overline{X}$ and S^2 are independent.

Proof:
This can be seen since S^2 is a function of the vector $(X_1 - \overline{X}, \ldots, X_n - \overline{X})$, and $(X_1 - \overline{X}, \ldots, X_n - \overline{X})$ is independent of $\overline{X}$, as previously shown in Theorem 7.2.1. We can use Theorem 4.2.7 to formally complete this proof. ■

Corollary 7.2.3:

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

Proof:
Recall the following facts:

- If $Z \sim N(0,1)$ then $Z^2 \sim \chi^2_1$.
- If $Y_1, \ldots, Y_n \sim$ iid χ^2_1, then $\sum_{i=1}^{n} Y_i \sim \chi^2_n$.
- For χ^2_n, the mgf is $M(t) = (1-2t)^{-n/2}$.
- If $X_i \sim N(\mu, \sigma^2)$, then $\frac{X_i - \mu}{\sigma} \sim N(0,1)$ and $\frac{(X_i-\mu)^2}{\sigma^2} \sim \chi^2_1$.

  Therefore, $\sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2} \sim \chi^2_n$ and $\frac{(\overline{X}-\mu)^2}{(\frac{\sigma}{\sqrt{n}})^2} = n\frac{(\overline{X}-\mu)^2}{\sigma^2} \sim \chi^2_1$. $\quad (*)$

Now consider

$$\begin{aligned}
\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}((X_i-\overline{X})+(\overline{X}-\mu))^2 \\
&= \sum_{i=1}^{n}((X_i-\overline{X})^2+2(X_i-\overline{X})(\overline{X}-\mu)+(\overline{X}-\mu)^2) \\
&= (n-1)S^2+0+n(\overline{X}-\mu)^2
\end{aligned}$$

Therefore,

$$\underbrace{\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2}}_{W}=\underbrace{\frac{n(\overline{X}-\mu)^2}{\sigma^2}}_{U}+\underbrace{\frac{(n-1)S^2}{\sigma^2}}_{V}$$

We have an expression of the form: $W=U+V$

Since U and V are functions of $\overline{X}$ and S^2, we know by Corollary 7.2.2 that they are independent and also that their mgf's factor by Theorem 4.6.3 (iv). Now we can write:

$$\begin{aligned}
M_W(t) &= M_U(t)M_V(t) \\
\Longrightarrow M_V(t) &= \frac{M_W(t)}{M_U(t)} \\
&\overset{(*)}{=} \frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}} \\
&= (1-2t)^{-(n-1)/2}
\end{aligned}$$

Note that this is the mgf of χ^2_{n-1} by the uniqueness of mgf's. Thus, $V=\frac{(n-1)S^2}{\sigma^2}\sim\chi^2_{n-1}$. ■

<u>Corollary 7.2.4:</u>

$$\frac{\sqrt{n}(\overline{X}-\mu)}{S}\sim t_{n-1}.$$

<u>Proof:</u>
Recall the following facts:

- If $Z\sim N(0,1)$, $\quad Y\sim\chi^2_n$ and Z,Y independent, then $\frac{Z}{\sqrt{\frac{Y}{n}}}\sim t_n$.
- $Z_1=\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}\sim N(0,1)$, $\quad Y_{n-1}=\frac{(n-1)S^2}{\sigma^2}\sim\chi^2_{n-1}$, and Z_1,Y_{n-1} are independent.

Therefore,

$$\frac{\sqrt{n}(\overline{X}-\mu)}{S}=\frac{\frac{(\overline{X}-\mu)}{\sigma/\sqrt{n}}}{\frac{S/\sqrt{n}}{\sigma/\sqrt{n}}}=\frac{\frac{(\overline{X}-\mu)}{\sigma/\sqrt{n}}}{\sqrt{\frac{S^2(n-1)}{\sigma^2(n-1)}}}=\frac{Z_1}{\sqrt{\frac{Y_{n-1}}{(n-1)}}}\sim t_{n-1}.$$

■

Corollary 7.2.5:
Let $(X_1, \ldots, X_m) \sim$ iid $N(\mu_1, \sigma_1^2)$ and $(Y_1, \ldots, Y_n) \sim$ iid $N(\mu_2, \sigma_2^2)$. Let X_i, Y_j be independent $\forall i, j$.
Then it holds:

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{[(m-1)S_1^2/\sigma_1^2] + [(n-1)S_2^2/\sigma_2^2]}} \cdot \sqrt{\frac{m+n-2}{\sigma_1^2/m + \sigma_2^2/n}} \sim t_{m+n-2}$$

In particular, if $\sigma_1 = \sigma_2$, then:

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{(m-1)S_1^2 + (n-1)S_2^2}} \cdot \sqrt{\frac{mn(m+n-2)}{m+n}} \sim t_{m+n-2}$$

Proof:
Homework. ■

Corollary 7.2.6:
Let $(X_1, \ldots, X_m) \sim$ iid $N(\mu_1, \sigma_1^2)$ and $(Y_1, \ldots, Y_n) \sim$ iid $N(\mu_2, \sigma_2^2)$. Let X_i, Y_j be independent $\forall i, j$.
Then it holds:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{m-1,n-1}$$

In particular, if $\sigma_1 = \sigma_2$, then:

$$\frac{S_1^2}{S_2^2} \sim F_{m-1,n-1}$$

Proof:
Recall that, if $Y_1 \sim \chi_m^2$ and $Y_2 \sim \chi_n^2$, then

$$F = \frac{Y_1/m}{Y_2/n} \sim F_{m,n}.$$

Now, $C_1 = \frac{(m-1)S_1^2}{\sigma_1^2} \sim \chi_{m-1}^2$ and $C_2 = \frac{(n-1)S_2^2}{\sigma_2^2} \sim \chi_{n-1}^2$. Therefore,

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\frac{(m-1)S_1^2}{\sigma_1^2(m-1)}}{\frac{(n-1)S_2^2}{\sigma_2^2(n-1)}} = \frac{C_1/(m-1)}{C_2/(n-1)} \sim F_{m-1,n-1}.$$

If $\sigma_1 = \sigma_2$, then

$$\frac{S_1^2}{S_2^2} \sim F_{m-1,n-1}.$$

■

Lecture 07:
We 01/24/01

8 The Theory of Point Estimation

8.1 The Problem of Point Estimation

Let $\underline{X}$ be a rv defined on a probability space (Ω, L, P). Suppose that the cdf F of $\underline{X}$ depends on some set of parameters and that the functional form of F is known except for a finite number of these parameters.

Definition 8.1.1:
The set of admissible values of θ is called the **parameter space** Θ. If F_θ is the cdf of $\underline{X}$ when θ is the parameter, the set $\{F_\theta \, : \, \theta \in \Theta\}$ is the **family of cdf's**. Likewise, we speak of the **family of pdf's** if $\underline{X}$ is continuous, and the **family of pmf's** if $\underline{X}$ is discrete. ■

Example 8.1.2:
$X \sim Bin(n,p)$, p unknown. Then $\theta = p$ and $\Theta = \{p : 0 < p < 1\}$.

$X \sim N(\mu, \sigma^2)$, (μ, σ^2) unknown. Then $\theta = (\mu, \sigma^2)$ and $\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$. ■

Definition 8.1.3:
Let $\underline{X}$ be a sample from F_θ, $\theta \in \Theta \subseteq I\!R$. Let a statistic $T(\underline{X})$ map $I\!R^n$ to Θ. We call $T(\underline{X})$ an **estimator** of θ and $T(\underline{x})$ for a realization $\underline{x}$ of $\underline{X}$ an **(point) estimate** of θ. In practice, the term *estimate* is used for both. ■

Example 8.1.4:
Let $X_1, \ldots, X_n$ be iid $Bin(1,p)$, p unknown. Estimates of p include:

$$T_1(\underline{X}) = \overline{X}, \; T_2(\underline{X}) = X_1, \; T_3(\underline{X}) = \frac{1}{2}, T_4(\underline{X}) = \frac{X_1 + X_2}{3}$$

Obviously, not all estimates are equally good. ■

8.2 Properties of Estimates

Definition 8.2.1:

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of iid rv's with cdf F_θ, $\theta \in \Theta$. A sequence of point estimates $T_n(X_1, \ldots, X_n) = T_n$ is called

- **(weakly) consistent** for θ if $T_n \xrightarrow{p} \theta$ as $n \to \infty$ $\forall \theta \in \Theta$
- **strongly consistent** for θ if $T_n \xrightarrow{a.s.} \theta$ as $n \to \infty$ $\forall \theta \in \Theta$
- **consistent in the r^{th} mean** for θ if $T_n \xrightarrow{r} \theta$ as $n \to \infty$ $\forall \theta \in \Theta$

■

Example 8.2.2:

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of iid $Bin(1,p)$ rv's. Let $\overline{X}_n = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$. Since $E(X_i) = p$, it follows by the WLLN that $\overline{X}_n \xrightarrow{p} p$, i.e., consistency, and by the SLLN that $\overline{X}_n \xrightarrow{a.s.} p$, i.e, strong consistency.

However, a consistent estimate may not be unique. We may even have infinite many consistent estimates, e.g.,

$$\frac{\sum\limits_{i=1}^{n} X_i + a}{n+b} \xrightarrow{p} p \quad \forall \text{ finite } a, b \in I\!R.$$

■

Theorem 8.2.3:

If T_n is a sequence of estimates such that $E(T_n) \to \theta$ and $Var(T_n) \to 0$ as $n \to \infty$, then T_n is consistent for θ.

Proof:

$$\begin{aligned}
P(\mid T_n - \theta \mid > \epsilon) &\overset{(A)}{\leq} \frac{E(T_n - \theta)^2)}{\epsilon^2} \\
&= \frac{E[((T_n - E(T_n)) + (E(T_n) - \theta))^2]}{\epsilon^2} \\
&= \frac{Var(T_n) + 2E[(T_n - E(T_n))(E(T_n) - \theta)] + (E(T_n) - \theta)^2}{\epsilon^2} \\
&= \frac{Var(T_n) + (E(T_n) - \theta)^2}{\epsilon^2} \\
&\overset{(B)}{\longrightarrow} 0 \quad \text{as } n \to \infty
\end{aligned}$$

(A) holds due to Corollary 3.5.2 (Markov's Inequality). (B) holds since $Var(T_n) \to 0$ as $n \to \infty$ and $E(T_n) \to \theta$ as $n \to \infty$. ■

Definition 8.2.4:

Let $\mathcal{G}$ be a group of Borel–measurable functions of $I\!R^n$ onto itself which is closed under composition and inverse. A family of distributions $\{P_\theta : \theta \in \Theta\}$ is **invariant** under $\mathcal{G}$ if for each $g \in \mathcal{G}$ and for all $\theta \in \Theta$, there exists a unique $\theta' = \overline{g}(\theta)$ such that the distribution of $g(\underline{X})$ is $P_{\theta'}$ whenever the distribution of $\underline{X}$ is P_θ. We call $\overline{g}$ the **induced function** on θ since $P_\theta(g(\underline{X}) \in A) = P_{\overline{g}(\theta)}(\underline{X} \in A)$. ■

Example 8.2.5:

Let $(X_1, \ldots, X_n)$ be iid $N(\mu, \sigma^2)$ with pdf

$$f(x_1, \ldots, x_n) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right).$$

The group of linear transformations $\mathcal{G}$ has elements

$$g(x_1, \ldots, x_n) = (ax_1 + b, \ldots, ax_n + b), \quad a > 0, \quad -\infty < b < \infty.$$

The pdf of $g(\underline{X})$ is

$$f^*(x_1^*, \ldots, x_n^*) = \frac{1}{(\sqrt{2\pi}a\sigma)^n} \exp\left(-\frac{1}{2a^2\sigma^2}\sum_{i=1}^{n}(x_i^* - a\mu - b)^2\right), \quad x_i^* = ax_i + b, \; i = 1, \ldots, n.$$

So $\{f : -\infty < \mu < \infty, \; \sigma^2 > 0\}$ is invariant under this group $\mathcal{G}$, with $\overline{g}(\mu, \sigma^2) = (a\mu + b, a^2\sigma^2)$, where $-\infty < a\mu + b < \infty$ and $a^2\sigma^2 > 0$. ■

Definition 8.2.6:

Let $\mathcal{G}$ be a group of transformations that leaves $\{F_\theta : \theta \in \Theta\}$ invariant. An estimate T is **invariant** under $\mathcal{G}$ if

$$T(g(X_1), \ldots, g(X_n)) = T(X_1, \ldots, X_n) \quad \forall g \in \mathcal{G}.$$

■

Definition 8.2.7:
An estimate T is:

- **location invariant** if $T(X_1 + a, \ldots, X_n + a) = T(X_1, \ldots, X_n), \quad a \in I\!R$
- **scale invariant** if $T(cX_1, \ldots, cX_n) = T(X_1, \ldots, X_n), \quad c \in I\!R - \{0\}$
- **permutation invariant** if $T(X_{i_1}, \ldots, X_{i_n}) = T(X_1, \ldots, X_n)$ $\forall$ permutations $(i_1, \ldots, i_n)$ of $1, \ldots, n$

■

Example 8.2.8:
Let $F_\theta \sim N(\mu, \sigma^2)$.

S^2 is location invariant.

$\overline{X}$ and S^2 are both permutation invariant.

Neither $\overline{X}$ nor S^2 is scale invariant. ■

Note:
Different sources make different use of the term *invariant*. Mood, Graybill & Boes (1974) for example define *location invariant* as $T(X_1 + a, \ldots, X_n + a) = T(X_1, \ldots, X_n) + a$ (page 332) and *scale invariant* as $T(cX_1, \ldots, cX_n) = cT(X_1, \ldots, X_n)$ (page 336). According to their definition, $\overline{X}$ is location invariant and scale invariant. ■

8.3 Sufficient Statistics

Definition 8.3.1:
Let $\underline{X} = (X_1, \ldots, X_n)$ be a sample from $\{F_\theta \ : \ \theta \in \Theta \subseteq I\!R^k\}$. A statistic $T = T(\underline{X})$ is **sufficient** for θ (or for the family of distributions $\{F_\theta \ : \ \theta \in \Theta\}$) iff the conditional distribution of $\underline{X}$ given $T = t$ does not depend on θ (except possibly on a null set A where $P_\theta(T \in A) = 0 \ \ \forall \theta$). ■

Note:

(i) The sample $\underline{X}$ is always sufficient but this is not particularly interesting and usually is excluded from further considerations.

(ii) Idea: Once we have "reduced" from $\underline{X}$ to $T(\underline{X})$, we have captured all the information in $\underline{X}$ about θ.

(iii) Usually, there are several sufficient statistics for a given family of distributions

■

Example 8.3.2:
Let $\underline{X} = (X_1, \ldots, X_n)$ be iid $Bin(1,p)$ rv's. To estimate p, can we ignore the order and simply count the number of "successes"?

Let $T(\underline{X}) = \sum\limits_{i=1}^n X_i$. It is

$$
\begin{aligned}
P(X_1 = x_1, \ldots X_n = x_n \mid \sum_{i=1}^n X_i = t) &= \frac{P(X_1 = x_1, \ldots, X_n = x_n, T = t)}{P(T = t)} \\
&= \begin{cases} \dfrac{P(X_1 = x_1, \ldots, X_n = x_n)}{P(T = t)}, & \sum\limits_{i=1}^n x_i = t \\ 0, & \text{otherwise} \end{cases} \\
&= \begin{cases} \dfrac{p^t(1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}}, & \sum\limits_{i=1}^n x_i = t \\ 0, & \text{otherwise} \end{cases} \\
&= \begin{cases} \dfrac{1}{\binom{n}{t}}, & \sum\limits_{i=1}^n x_i = t \\ 0, & \text{otherwise} \end{cases}
\end{aligned}
$$

This does not depend on p. Thus, $T = \sum\limits_{i=1}^n X_i$ is sufficient for p. ■

Example 8.3.3:

Let $\underline{X} = (X_1, \ldots, X_n)$ be iid Poisson(λ). Is $T = \sum_{i=1}^{n} X_i$ sufficient for λ? It is

$$\begin{aligned}
P(X_1 = x_1, \ldots, X_n = x_n \mid T = t) &= \begin{cases} \dfrac{\prod_{i=1}^{n} \dfrac{e^{-\lambda}\lambda^{x_i}}{x_i!}}{\dfrac{e^{-n\lambda}(n\lambda)^t}{t!}}, & \sum_{i=1}^{n} x_i = t \\ 0, & \text{otherwise} \end{cases} \\
&= \begin{cases} \dfrac{\dfrac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod x_i!}}{\dfrac{e^{-n\lambda}(n\lambda)^t}{t!}}, & \sum_{i=1}^{n} x_i = t \\ 0, & \text{otherwise} \end{cases} \\
&= \begin{cases} \dfrac{t!}{n^t \prod_{i=1}^{n} x_i!}, & \sum_{i=1}^{n} x_i = t \\ 0, & \text{otherwise} \end{cases}
\end{aligned}$$

This does not depend on λ. Thus, $T = \sum_{i=1}^{n} X_i$ is sufficient for λ. ■

Example 8.3.4:

Let X_1, X_2 be iid Poisson(λ). Is $T = X_1 + 2X_2$ sufficient for λ? It is

$$\begin{aligned}
P(X_1 = 0, X_2 = 1 \mid X_1 + 2X_2 = 2) &= \frac{P(X_1 = 0, X_2 = 1)}{P(X_1 + 2X_2 = 2)} \\
&= \frac{P(X_1 = 0, X_2 = 1)}{P(X_1 = 0, X_2 = 1) + P(X_1 = 2, X_2 = 0)} \\
&= \frac{e^{-\lambda}(e^{-\lambda}\lambda)}{e^{-\lambda}(e^{-\lambda}\lambda) + (\frac{e^{-\lambda}\lambda^2}{2})e^{-\lambda}} \\
&= \frac{1}{1 + \frac{\lambda}{2}}
\end{aligned}$$

This still depends on λ. Thus, $T = X_1 + 2X_2$ is not sufficient for λ. ■

Note:

Definition 8.3.1 can be difficult to check. In addition, it requires a candidate statistic. We need something constructive that helps in finding sufficient statistics without having to check Definition 8.3.1. The next Theorem helps in finding such statistics. ■

Lecture 08: Fr 01/26/01

Theorem 8.3.5: **Factorization Criterion**
Let $X_1, \ldots, X_n$ be rv's with pdf (or pmf) $f(x_1, \ldots, x_n \mid \theta), \quad \theta \in \Theta$. Then $T(X_1, \ldots, X_n)$ is sufficient for θ iff we can write

$$f(x_1, \ldots, x_n \mid \theta) = h(x_1, \ldots, x_n) \; g(T(x_1, \ldots, x_n) \mid \theta),$$

where h does not depend on θ and g does not depend on $x_1, \ldots, x_n$ except as a function of T.

Proof:
Discrete case only.

"$\Longrightarrow$":
Suppose $T(\underline{X})$ is sufficient for θ. Let

$$\begin{aligned} g(t \mid \theta) &= P_\theta(T(\underline{X}) = t) \\ h(\underline{x}) &= P(\underline{X} = \underline{x} \mid T(\underline{X}) = t) \end{aligned}$$

Then it holds:

$$\begin{aligned} f(\underline{x} \mid \theta) &= P_\theta(\underline{X} = \underline{x}) \\ &\overset{(*)}{=} P_\theta(\underline{X} = \underline{x}, T(\underline{X}) = T(\underline{x}) = t) \\ &= P_\theta(T(\underline{X}) = t) \; P(\underline{X} = \underline{x} \mid T(\underline{X}) = t) \\ &= g(t \mid \theta) h(\underline{x}) \end{aligned}$$

$(*)$ holds since $\underline{X} = \underline{x}$ implies that $T(\underline{X}) = T(\underline{x}) = t$.

"$\Longleftarrow$":
Suppose the factorization holds. For fixed t_0, it is

$$\begin{aligned} P_\theta(T(\underline{X}) = t_0) &= \sum_{\{\underline{x} \,:\, T(\underline{x}) = t_0\}} P_\theta(\underline{X} = \underline{x}) \\ &= \sum_{\{\underline{x} \,:\, T(\underline{x}) = t_0\}} h(\underline{x}) g(T(\underline{x}) \mid \theta) \\ &= g(t_0 \mid \theta) \sum_{\{\underline{x} \,:\, T(\underline{x}) = t_0\}} h(\underline{x}) \quad (A) \end{aligned}$$

If $P_\theta(T(\underline{X}) = t_0) > 0$, it holds:

$$
\begin{aligned}
P_\theta(\underline{X} = \underline{x} \mid T(\underline{X}) = t_0) &= \frac{P_\theta(\underline{X} = \underline{x}, T(\underline{X}) = t_0)}{P_\theta(T(\underline{X}) = t_0)} \\
&= \begin{cases} \dfrac{P_\theta(\underline{X} = \underline{x})}{P_\theta(T(\underline{X}) = t_0)}, & \text{if } T(\underline{x}) = t_0 \\ 0, & \text{otherwise} \end{cases} \\
&\overset{(A)}{=} \begin{cases} \dfrac{g(t_0 \mid \theta) h(\underline{x})}{g(t_0 \mid \theta) \sum\limits_{\{\underline{x} \,:\, T(\underline{x}) = t_0\}} h(\underline{x})}, & \text{if } T(\underline{x}) = t_0 \\ 0, & \text{otherwise} \end{cases} \\
&= \begin{cases} \dfrac{h(\underline{x})}{\sum\limits_{\{\underline{x} \,:\, T(\underline{x}) = t_0\}} h(\underline{x})}, & \text{if } T(\underline{x}) = t_0 \\ 0, & \text{otherwise} \end{cases}
\end{aligned}
$$

This last expression does not depend on θ. Thus, $T(\underline{X})$ is sufficient for θ. ■

Note:

(i) In the Theorem above, θ and T may be vectors.

(ii) If T is sufficient for θ, then also any 1–to–1 mapping of T is sufficient for θ. However, this does not hold for arbitrary functions of T.

■

Example 8.3.6:

Let $X_1, \ldots, X_n$ be iid $Bin(1, p)$. It is

$$P(X_1 = x_1, \ldots, X_n = x_n \mid p) = p^{\sum x_i} (1-p)^{n - \sum x_i}.$$

Thus, $h(x_1, \ldots, x_n) = 1$ and $g(\sum x_i \mid p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$.

Hence, $T = \sum\limits_{i=1}^{n} X_i$ is sufficient for p. ■

Example 8.3.7:
Let $X_1, \ldots, X_n$ be iid Poisson(λ). It is

$$P(X_1 = x_1, \ldots, X_n = x_n \mid \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod x_i!}.$$

Thus, $h(x_1, \ldots, x_n) = \frac{1}{\prod x_i!}$ and $g(\sum x_i \mid \lambda) = e^{-n\lambda}\lambda^{\sum x_i}$.

Hence, $T = \sum_{i=1}^{n} X_i$ is sufficient for λ. ■

Example 8.3.8:
Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ where $\mu \in I\!R$ and $\sigma^2 > 0$ are both unknown. It is

$$f(x_1, \ldots, x_n \mid \mu, \sigma^2) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{\sum(x_i - \mu)^2}{2\sigma^2}\right) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{\sum x_i^2}{2\sigma^2} + \mu\frac{\sum x_i}{\sigma^2} - \frac{n\mu^2}{2\sigma^2}\right).$$

Hence, $T = (\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2)$ is sufficient for (μ, σ^2). ■

Example 8.3.9:
Let $X_1, \ldots, X_n$ be iid $U(\theta, \theta+1)$ where $-\infty < \theta < \infty$. It is

$$\begin{aligned} f(x_1, \ldots, x_n \mid \theta) &= \begin{cases} 1, & \theta < x_i < \theta + 1 \ \ \forall i \in \{1, \ldots, n\} \\ 0, & \text{otherwise} \end{cases} \\ &= \prod_{i=1}^{n} I_{(\theta,\infty)}(x_i)\ I_{(-\infty,\theta+1)}(x_i) \\ &= I_{(\theta,\infty)}(\min(x_i))\ I_{(-\infty,\theta+1)}(\max(x_i)) \end{aligned}$$

Hence, $T = (X_{(1)}, X_{(n)})$ is sufficient for θ. ■

Definition 8.3.10:
Let $\{f_\theta(x) \ : \ \theta \in \Theta\}$ be a family of pdf's (or pmf's). We say the family is **complete** if

$$E_\theta(g(X)) = 0 \ \ \forall \theta \in \Theta$$

implies that

$$P_\theta(g(X) = 0) = 1 \ \ \forall \theta \in \Theta.$$

We say a statistic $T(X)$ is **complete** if the family of distributions of T is complete. ■

Example 8.3.11:

Let $X_1, \ldots, X_n$ be iid $Bin(1,p)$. We have seen in Example 8.3.6 that $T = \sum_{i=1}^n X_i$ is sufficient for p. Is it also complete?

We know that $T \sim Bin(n,p)$. Thus,

$$E_p(g(T)) = \sum_{t=0}^n g(t) \binom{n}{t} p^t (1-p)^{n-t} = 0 \quad \forall p \in (0,1)$$

implies that

$$(1-p)^n \sum_{t=0}^n g(t) \binom{n}{t} (\frac{p}{1-p})^t = 0 \quad \forall p \in (0,1) \quad \forall t.$$

However, $\sum_{t=0}^n g(t) \binom{n}{t} (\frac{p}{1-p})^t$ is a polynomial in $\frac{p}{1-p}$ which is only equal to 0 for all $p \in (0,1)$ if all of its coefficients are 0.

Therefore, $g(t) = 0$ for $t = 0, 1, \ldots, n$. Hence, T is complete. ■

Lecture 09: Mo 01/29/01

Example 8.3.12:

Let $X_1, \ldots, X_n$ be iid $N(\theta, \theta^2)$. We know from Example 8.3.8 that $T = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is sufficient for θ. Is it also complete?

We know that $\sum_{i=1}^n X_i \sim N(n\theta, n\theta^2)$. Therefore,

$$\begin{aligned} E((\sum_{i=1}^n X_i)^2) &= n\theta^2 + n^2\theta^2 = n(n+1)\theta^2 \\ E(\sum_{i=1}^n X_i^2) &= n(\theta^2 + \theta^2) = 2n\theta^2 \end{aligned}$$

It follows that

$$E\left(2(\sum_{i=1}^n X_i)^2 - (n+1)\sum_{i=1}^n X_i^2\right) = 0 \quad \forall \theta.$$

But $g(x_1, \ldots, x_n) = 2(\sum_{i=1}^n x_i)^2 - (n+1)\sum_{i=1}^n x_i^2$ is not identically to 0.

Therefore, T is not complete. ■

Note:

Recall from Section 5.2 what it means if we say the family of distributions $\{f_\theta : \theta \in \Theta\}$ is a one–parameter (or k–parameter) exponential family. ■

Theorem 8.3.13:

Let $\{f_\theta : \theta \in \Theta\}$ be a k–parameter exponential family. Let $T_1, \ldots, T_k$ be statistics. Then the family of distributions of $(T_1(\underline{X}), \ldots, T_k(\underline{X}))$ is also a k–parameter exponential family given by

$$g_\theta(\underline{t}) = \exp\left(\sum_{i=1}^{k} t_i Q_i(\theta) + D(\theta) + S^*(\underline{t})\right)$$

for suitable $S^*(t)$.

Proof:

The proof follows from our Theorems regarding the transformation of rv's. ■

Theorem 8.3.14:

Let $\{f_\theta : \theta \in \Theta\}$ be a k–parameter exponential family with $k \leq n$ and let $T_1, \ldots, T_k$ be statistics as in Theorem 8.3.13. Suppose that the range of $\underline{Q} = (Q_1, \ldots, Q_k)$ contains an open set in $I\!R^k$. Then $\underline{T} = (T_1(\underline{X}), \ldots, T_k(\underline{X}))$ is a complete sufficient statistic.

Proof:

Discrete case and $k = 1$ only.

Write $Q(\theta) = \theta$ and let $(a, b) \subseteq \Theta$.
It follows from the Factorization Criterion (Theorem 8.3.5) that T is sufficient for θ. Thus, we only have to show that T is complete, i.e., that

$$\begin{aligned} E_\theta(g(T(\underline{X}))) &= \sum_t g(t) P_\theta(T(\underline{X}) = t) \\ &\overset{(A)}{=} \sum_t g(t) \exp(\theta t + D(\theta) + S^*(t)) = 0 \quad \forall \theta \quad (B) \end{aligned}$$

implies $g(t) = 0 \ \ \forall t$. Note that in (A) we make use of a result established in Theorem 8.3.13.

We now define functions g^+ and g^- as:

$$g^+(t) = \begin{cases} g(t), & \text{if } g(t) \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

$$g^-(t) = \begin{cases} -g(t), & \text{if } g(t) < 0 \\ 0, & \text{otherwise} \end{cases}$$

It is $g(t) = g^+(t) - g^-(t)$ where both functions, g^+ and g^-, are non–negative functions. Using g^+ and g^-, it turns out that (B) is equivalent to

$$\sum_t g^+(t) \exp(\theta t + S^*(t)) = \sum_t g^-(t) \exp(\theta t + S^*(t)) \quad \forall \theta \quad (C)$$

where the term $\exp(D(\theta))$ in (A) drops out as a constant on both sides.

If we fix $\theta_0 \in (a, b)$ and define

$$p^+(t) = \frac{g^+(t)\exp(\theta_0 t + S^*(t))}{\sum_t g^+(t)\exp(\theta_0 t + S^*(t))}, \quad p^-(t) = \frac{g^-(t)\exp(\theta_0 t + S^*(t))}{\sum_t g^-(t)\exp(\theta_0 t + S^*(t))},$$

it is obvious that $p^+(t) \geq 0 \quad \forall t$ and $p^-(t) \geq 0 \quad \forall t$ and by construction $\sum_t p^+(t) = 1$ and $\sum_t p^-(t) = 1$. Hence, p^+ and p^- are both pmf's.

From (C), it follows for the mgf's M^+ and M^- of p^+ and p^- that

$$\begin{aligned}
M^+(\delta) &= \sum_t e^{\delta t} p^+(t) \\
&= \frac{\sum_t e^{\delta t} g^+(t)\exp(\theta_0 t + S^*(t))}{\sum_t g^+(t)\exp(\theta_0 t + S^*(t))} \\
&= \frac{\sum_t g^+(t)\exp((\theta_0 + \delta)t + S^*(t))}{\sum_t g^+(t)\exp(\theta_0 t + S^*(t))} \\
&\overset{(C)}{=} \frac{\sum_t g^-(t)\exp((\theta_0 + \delta)t + S^*(t))}{\sum_t g^-(t)\exp(\theta_0 t + S^*(t))} \\
&= \frac{\sum_t e^{\delta t} g^-(t)\exp(\theta_0 t + S^*(t))}{\sum_t g^-(t)\exp(\theta_0 t + S^*(t))} \\
&= \sum_t e^{\delta t} p^-(t) \\
&= M^-(\delta) \quad \forall \delta \in (\underbrace{a - \theta_0}_{<0}, \underbrace{b - \theta_0}_{>0}).
\end{aligned}$$

By the uniqueness of mgf's it follows that $p^+(t) = p^-(t) \quad \forall t$.

$\Longrightarrow g^+(t) = g^-(t) \quad \forall t$

$\Longrightarrow g(t) = 0 \quad \forall t$

$\Longrightarrow T$ is complete $\blacksquare$

8.4 Unbiased Estimation

Definition 8.4.1:
Let $\{F_\theta \,:\, \theta \in \Theta\}$, $\Theta \subseteq I\!R$, be a nonempty set of cdf's. A Borel–measurable function T from $I\!R^n$ to Θ is called **unbiased** for θ (or an unbiased estimate for θ) if

$$E_\theta(T) = \theta \quad \forall \theta \in \Theta.$$

Any function $d(\theta)$ for which an unbiased estimate T exists is called an **estimable function**.

If T is biased,

$$b(\theta, T) = E_\theta(T) - \theta$$

is called the **bias** of T. ■

Example 8.4.2:
If the k^{th} population moment exists, the k^{th} sample moment is an unbiased estimate. If $Var(X) = \sigma^2$, the sample variance S^2 is an unbiased estimate of σ^2.

However, note that for $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$, S is not an unbiased estimate of σ:

$$\begin{aligned}
\frac{(n-1)S^2}{\sigma^2} &\sim \chi^2_{n-1} = Gamma(\frac{n-1}{2}, 2) \\
\Longrightarrow E\left(\sqrt{\frac{(n-1)S^2}{\sigma^2}}\right) &= \int_0^\infty \sqrt{x}\frac{x^{\frac{n-1}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})}dx \\
&= \frac{\sqrt{2}\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}\int_0^\infty \frac{x^{\frac{n}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}dx \\
&\overset{(*)}{=} \frac{\sqrt{2}\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \\
\Longrightarrow E(S) &= \sigma\sqrt{\frac{2}{n-1}}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}
\end{aligned}$$

$(*)$ holds since $\frac{x^{\frac{n}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}$ is the pdf of a $Gamma(\frac{n}{2}, 2)$ distribution and thus the integral is 1.

So S is biased for σ and

$$b(\sigma, S) = \sigma\left(\sqrt{\frac{2}{n-1}}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} - 1\right).$$

■

Note:
If T is unbiased for θ, $g(T)$ is not necessarily unbiased for $g(\theta)$ (unless g is a linear function).

■

Lecture 10:
We 01/31/01

Example 8.4.3:

Unbiased estimates may not exist (see Rohatgi, page 351, Example 2) or they me be absurd as in the following case:

Let $X \sim$ Poisson(λ) and let $d(\lambda) = e^{-2\lambda}$. Consider $T(X) = (-1)^X$ as an estimate. It is

$$\begin{aligned} E_\lambda(T(X)) &= e^{-\lambda} \sum_{x=0}^{\infty} (-1)^x \frac{\lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(-\lambda)^x}{x!} \\ &= e^{-\lambda} e^{-\lambda} \\ &= e^{-2\lambda} \\ &= d(\lambda) \end{aligned}$$

Hence T is unbiased for $d(\lambda)$ but since T alternates between -1 and 1 while $d(\lambda) > 0$, T is not a good estimate. ■

Note:

If there exist 2 unbiased estimates T_1 and T_2 of θ, then any estimate of the form $\alpha T_1 + (1-\alpha) T_2$ for $0 < \alpha < 1$ will also be an unbiased estimate of θ. Which one should we choose? ■

Definition 8.4.4:

The **mean square error** of an estimate T of θ is defined as

$$\begin{aligned} MSE(\theta, T) &= E_\theta((T - \theta)^2) \\ &= Var_\theta(T) + (b(\theta, T))^2. \end{aligned}$$

Let $\{T_i\}_{i=1}^{\infty}$ be a sequence of estimates of θ. If

$$\lim_{i \to \infty} MSE(\theta, T_i) = 0 \quad \forall \theta \in \Theta,$$

then $\{T_i\}$ is called a **mean–squared–error consistent** (**MSE–consistent**) sequence of estimates of θ. ■

Note:

(i) If we allow all estimates and compare their MSE, generally it will depend on θ which estimate is better. For example $\hat{\theta} = 17$ is perfect if $\theta = 17$, but it is lousy otherwise.

(ii) If we restrict ourselves to the class of unbiased estimates, then $MSE(\theta, T) = Var_\theta(T)$.

(iii) MSE–consistency means that both the bias and the variance of T_i approach 0 as $i \to \infty$.

■

<u>Definition 8.4.5:</u>
Let $\theta_0 \in \Theta$ and let $U(\theta_0)$ be the class of all unbiased estimates T of θ_0 such that $E_{\theta_0}(T^2) < \infty$. Then $T_0 \in U(\theta_0)$ is called a **locally minimum variance unbiased estimate** (**LMVUE**) at θ_0 if

$$E_{\theta_0}((T_0 - \theta_0)^2) \leq E_{\theta_0}((T - \theta_0)^2) \quad \forall T \in U(\theta_0).$$

■

<u>Definition 8.4.6:</u>
Let U be the class of all unbiased estimates T of $\theta \in \Theta$ such that $E_\theta(T^2) < \infty \quad \forall \theta \in \Theta$. Then $T_0 \in U$ is called a **uniformly minimum variance unbiased estimate** (**UMVUE**) of θ if

$$E_\theta((T_0 - \theta)^2) \leq E_\theta((T - \theta)^2) \quad \forall \theta \in \Theta \quad \forall T \in U.$$

■

An Excursion into Logic II

In our first "Excursion into Logic" in Lectures 10 & 11 of Stat 6710 Mathematical Statistics I on 9/20/2000 and 9/22/2000, we have established the following results:

$A \Rightarrow B$ is equivalent to $\neg B \Rightarrow \neg A$ is equivalent to $\neg A \vee B$:

A	B	$A \Rightarrow B$	$\neg A$	$\neg B$	$\neg B \Rightarrow \neg A$	$\neg A \vee B$
1	1	1	0	0	1	1
1	0	0	0	1	0	0
0	1	1	1	0	1	1
0	0	1	1	1	1	1

When dealing with formal proofs, there exists one more technique to show $A \Rightarrow B$. Equivalently, we can show $(A \wedge \neg B) \Rightarrow 0$, a technique called **Proof by Contradiction**. This means, assuming that A and $\neg B$ hold, we show that this implies 0, i.e., something that is always false, i.e., a contradiction. And here is the corresponding truth table:

A	B	$A \Rightarrow B$	$\neg B$	$A \wedge \neg B$	$(A \wedge \neg B) \Rightarrow 0$
1	1	1	0	0	1
1	0	0	1	1	0
0	1	1	0	0	1
0	0	1	1	0	1

Note:
We make use of this proof technique in the Proof of the next Theorem. ■

Example:
Let $A: \quad x = 5$ and $B: \quad x^2 = 25$. Obviously $A \Rightarrow B$.

But we can also prove this in the following way:

$A: \quad x = 5$ and $\neg B: \quad x^2 \neq 25$

$$\Longrightarrow x^2 = 25 \wedge x^2 \neq 25$$

This is impossible, i.e., a contradiction. Thus, $A \Rightarrow B$. ■

Theorem 8.4.7:

Let U be the class of all unbiased estimates T of $\theta \in \Theta$ with $E_\theta(T^2) < \infty \ \ \forall \theta$, and suppose that U is non–empty. Let U_0 be the set of all unbiased estimates of 0, i.e.,

$$U_0 = \{\nu \ : \ E_\theta(\nu) = 0, E_\theta(\nu^2) < \infty \ \ \forall \theta \in \Theta\}.$$

Then $T_0 \in U$ is UMVUE iff

$$E_\theta(\nu T_0) = 0 \ \ \forall \theta \in \Theta \ \forall \nu \in U_0.$$

Proof:

Note that $E_\theta(\nu T_0)$ always exists.

"$\Longrightarrow$:"

We suppose that $T_0 \in U$ is UMVUE and that $E_{\theta_0}(\nu_0 T_0) \neq 0$ for some $\theta_0 \in \Theta$ and some $\nu_0 \in U_0$.

It holds

$$E_\theta(T_0 + \lambda \nu_0) = E_\theta(T_0) = \theta \ \ \forall \lambda \in I\!R \ \ \forall \theta \in \Theta.$$

Therefore, $T_0 + \lambda \nu_0 \in U \ \ \forall \lambda \in I\!R$.

Also, $E_{\theta_0}(\nu_0^2) > 0$ (since otherwise, $P_{\theta_0}(\nu_0 = 0) = 1$ and then $E_{\theta_0}(\nu_0 T_0) = 0$).

Now let

$$\lambda = -\frac{E_{\theta_0}(T_0 \nu_0)}{E_{\theta_0}(\nu_0^2)}.$$

Then,

$$\begin{aligned}
E_{\theta_0}((T_0 + \lambda \nu_0)^2) &= E_{\theta_0}(T_0^2 + 2\lambda T_0 \nu_0 + \lambda^2 \nu_0^2) \\
&= E_{\theta_0}(T_0^2) - 2\frac{(E_{\theta_0}(T_0 \nu_0))^2}{E_{\theta_0}(\nu_0^2)} + \frac{(E_{\theta_0}(T_0 \nu_0))^2}{E_{\theta_0}(\nu_0^2)} \\
&= E_{\theta_0}(T_0^2) - \frac{(E_{\theta_0}(T_0 \nu_0))^2}{E_{\theta_0}(\nu_0^2)} \\
&< E_{\theta_0}(T_0^2),
\end{aligned}$$

and therefore,

$$Var_{\theta_0}(T_0 + \lambda \nu_0) < Var_{\theta_0}(T_0).$$

This means, T_0 is not UMVUE, i.e., a contradiction!

"$\Longleftarrow$:"

Let $E_\theta(\nu T_0) = 0$ for some $T_0 \in U$ for all $\theta \in \Theta$ and all $\nu \in U_0$.

We choose $T \in U$, then also $T_0 - T \in U_0$ and

$$E_\theta(T_0(T_0 - T)) = 0 \quad \forall \theta \in \Theta.$$

It follows by the Cauchy–Schwarz–Inequality (Theorem 4.5.7 (ii)) that

$$E_\theta(T_0^2) = E_\theta(T_0 T) \leq (E_\theta(T_0^2))^{\frac{1}{2}} (E_\theta(T^2))^{\frac{1}{2}}.$$

This implies

$$(E_\theta(T_0^2))^{\frac{1}{2}} \leq (E_\theta(T^2))^{\frac{1}{2}}$$

and

$$Var_\theta(T_0) \leq Var_\theta(T),$$

where T is an arbitrary unbiased estimate of θ. Thus, T_0 is UMVUE. ■

Lecture 11:
Mo 02/05/01

Theorem 8.4.8:

Let U be the non–empty class of unbiased estimates of $\theta \in \Theta$ as defined in Theorem 8.4.7. Then there exists at most one UMVUE $T \in U$ for θ.

Proof:

Suppose $T_0, T_1 \in U$ are both UMVUE.

Then $T_1 - T_0 \in U_0$, $Var_\theta(T_0) = Var_\theta(T_1)$, and $E_\theta(T_0(T_1 - T_0)) = 0 \quad \forall \theta \in \Theta$

$$\begin{aligned}
&\Longrightarrow E_\theta(T_0^2) = E_\theta(T_0 T_1) \\
&\Longrightarrow Cov_\theta(T_0, T_1) = E_\theta(T_0 T_1) - E_\theta(T_0) E_\theta(T_1) \\
&\qquad\qquad = E_\theta(T_0^2) - (E_\theta(T_0))^2 \\
&\qquad\qquad = Var_\theta(T_0) \\
&\qquad\qquad = Var_\theta(T_1) \quad \forall \theta \in \Theta \\
&\Longrightarrow \rho_{T_0 T_1} = 1 \quad \forall \theta \in \Theta \\
&\Longrightarrow P_\theta(aT_0 + bT_1 = 0) = 1 \quad \text{for some } a, b \quad \forall \theta \in \Theta \\
&\Longrightarrow \theta = E_\theta(T_0) = E_\theta(-\tfrac{b}{a} T_1) = E_\theta(T_1) \quad \forall \theta \in \Theta \\
&\Longrightarrow -\tfrac{b}{a} = 1 \\
&\Longrightarrow P_\theta(T_0 = T_1) = 1 \quad \forall \theta \in \Theta
\end{aligned}$$

■

Theorem 8.4.9:

(i) If an UMVUE T exists for a real function $d(\theta)$, then λT is the UMVUE for $\lambda d(\theta)$, $\lambda \in I\!R$.

(ii) If UMVUE's T_1 and T_2 exist for real functions $d_1(\theta)$ and $d_2(\theta)$, respectively, then $T_1 + T_2$ is the UMVUE for $d_1(\theta) + d_2(\theta)$.

Proof:
Homework. ■

Theorem 8.4.10:
If a sample consists of n independent observations $X_1, \ldots, X_n$ from the same distribution, the UMVUE, if it exists, is permutation invariant.

Proof:
Homework. ■

Theorem 8.4.11: **Rao–Blackwell**
Let $\{F_\theta : \theta \in \Theta\}$ be a family of cdf's, and let h be any statistic in U, where U is the non–empty class of all unbiased estimates of θ with $E_\theta(h^2) < \infty$. Let T be a sufficient statistic for $\{F_\theta : \theta \in \Theta\}$. Then the conditional expectation $E_\theta(h \mid T)$ is independent of θ and it is an unbiased estimate of θ. Additionally,

$$E_\theta((E(h \mid T) - \theta)^2) \leq E_\theta((h - \theta)^2) \quad \forall \theta \in \Theta$$

with equality iff $h = E(h \mid T)$.

Proof:
By Theorem 4.7.3, $E_\theta(E(h \mid T)) = E_\theta(h) = \theta$.

Since $\underline{X} \mid T$ does not depend on θ due to sufficiency, neither does $E(h \mid T)$ depend on θ.

Thus, we only have to show that

$$E_\theta((E(h \mid T))^2) \leq E_\theta(h^2) = E_\theta(E(h^2 \mid T)).$$

Thus, we only have to show that

$$(E(h \mid T))^2 \leq E(h^2 \mid T).$$

But the Cauchy–Schwarz–Inequality (Theorem 4.5.7 (ii)) gives us

$$(E(h \mid T))^2 \leq E(h^2 \mid T) E(1 \mid T) = E(h^2 \mid T).$$

Equality holds iff

$$
\begin{aligned}
&E_\theta((E(h \mid T))^2) = E_\theta(h^2) \\
&\qquad \Longleftrightarrow E_\theta(E(h^2 \mid T) - (E(h \mid T))^2) = 0 \\
&\qquad \Longleftrightarrow E_\theta(Var(h \mid T)) = 0 \\
&\qquad \Longleftrightarrow Var(h \mid T) = 0 \\
&\qquad \Longleftrightarrow E(h^2 \mid T) = (E(h \mid T))^2 \\
&\qquad \Longleftrightarrow h \text{ is a function of } T \text{ and } h = E(h \mid T).
\end{aligned}
$$

For the proof of the last step, see Rohatgi, page 170–171, Theorem 2, Corollary, and Proof of the Corollary. ■

<u>Theorem 8.4.12:</u> **Lehmann–Scheffée**
If T is a complete sufficient statistic and if there exists an unbiased estimate h of θ, then $E(h \mid T)$ is the (unique) UMVUE.

<u>Proof:</u>
Suppose that $h_1, h_2 \in U$. Then $E(h_1 \mid T) = E(h_2 \mid T) = \theta$ by Theorem 8.4.11.

Therefore,

$$E_\theta(E(h_1 \mid T) - E(h_2 \mid T)) = 0 \quad \forall \theta \in \Theta.$$

Since T is complete, $E(h_1 \mid T) = E(h_2 \mid T)$.

Therefore, $E(h \mid T)$ must be the same for all $h \in U$ and $E(h \mid T)$ improves all $h \in U$. Therefore, $E(h \mid T)$ is UMVUE by Theorem 8.4.11. ■

<u>Note:</u>
We can use Theorem 8.4.12 to find the UMVUE in two ways if we have a complete sufficient statistic T:

(i) If we can find an unbiased estimate $h(T)$, it will be the UMVUE since $E(h(T) \mid T) = h(T)$.

(ii) If we have any unbiased estimate h and if we can calculate $E(h \mid T)$, then $E(h \mid T)$ will be the UMVUE. The process of determining the UMVUE this way often is called *Rao–Blackwellization*.

(iii) Even if a complete sufficient statistic does not exist, the UMVUE may still exist (see Rohatgi, page 357–358, Example 10).

■

Example 8.4.13:

Let $X_1, \ldots, X_n$ be iid $Bin(1,p)$. Then $T = \sum_{i=1}^{n} X_i$ is a complete sufficient statistic as seen in Examples 8.3.6 and 8.3.11.

Since $E(X_1) = p$, X_1 is an unbiased estimate of p. However, due to part (i) of the Note above, since X_1 is not a function of T, X_1 is not the UMVUE.

We can use part (ii) of the Note above to construct the UMVUE. It is

$$P(X_1 = x \mid T) = \begin{cases} \frac{T}{n}, & x = 1 \\ \frac{n-T}{n}, & x = 0 \end{cases}$$

$$\Longrightarrow E(X_1 \mid T) = \tfrac{T}{n} = \overline{X}$$

$$\Longrightarrow \overline{X} \text{ is the UMVUE for } p$$

If we are interested in the UMVUE for $d(p) = p(1-p) = p - p^2 = Var(X)$, we can find it in the following way:

$$\begin{aligned}
E(T) &= np \\
E(T^2) &= E\left(\sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} X_i X_j\right) \\
&= np + n(n-1)p^2 \\
\Longrightarrow E\left(\frac{nT}{n(n-1)}\right) &= \frac{np}{n-1} \\
E\left(-\frac{T^2}{n(n-1)}\right) &= -\frac{p}{n-1} - p^2 \\
\Longrightarrow E\left(\frac{nT - T^2}{n(n-1)}\right) &= \frac{np}{n-1} - \frac{p}{n-1} - p^2 \\
&= \frac{(n-1)p}{n-1} - p^2 \\
&= p - p^2 \\
&= d(p)
\end{aligned}$$

Thus, due to part (i) of the Note above, $\frac{nT - T^2}{n(n-1)}$ is the UMVUE for $d(p) = p(1-p)$. ∎

Lecture 12:
We 02/07/01

8.5 Lower Bounds for the Variance of an Estimate

Theorem 8.5.1: **Cramér–Rao Lower Bound (CRLB)**
Let Θ be an open interval of $I\!R$. Let $\{f_\theta \ : \ \theta \in \Theta\}$ be a family of pdf's or pmf's. Assume that the set $\{\underline{x} \ : \ f_\theta(\underline{x}) = 0\}$ is independent of θ.

Let $\psi(\theta)$ be defined on Θ and let it be differentiable for all $\theta \in \Theta$. Let T be an unbiased estimate of $\psi(\theta)$ such that $E_\theta(T^2) < \infty \ \ \forall \theta \in \Theta$. Suppose that

(i) $\frac{\partial f_\theta(\underline{x})}{\partial \theta}$ is defined for all $\theta \in \Theta$,

(ii) for a pdf f_θ

$$\frac{\partial}{\partial \theta}\left(\int f_\theta(\underline{x}) d\underline{x}\right) = \int \frac{\partial f_\theta(\underline{x})}{\partial \theta} d\underline{x} = 0 \ \ \forall \theta \in \Theta$$

or for a pmf f_θ

$$\frac{\partial}{\partial \theta}\left(\sum_{\underline{x}} f_\theta(\underline{x})\right) = \sum_{\underline{x}} \frac{\partial f_\theta(\underline{x})}{\partial \theta} = 0 \ \ \forall \theta \in \Theta,$$

(iii) for a pdf f_θ

$$\frac{\partial}{\partial \theta}\left(\int T(\underline{x}) f_\theta(\underline{x}) d\underline{x}\right) = \int T(\underline{x}) \frac{\partial f_\theta(\underline{x})}{\partial \theta} d\underline{x} \ \ \forall \theta \in \Theta$$

or for a pmf f_θ

$$\frac{\partial}{\partial \theta}\left(\sum_{\underline{x}} T(\underline{x}) f_\theta(\underline{x})\right) = \sum_{\underline{x}} T(\underline{x}) \frac{\partial f_\theta(\underline{x})}{\partial \theta} \ \ \forall \theta \in \Theta.$$

Let $\chi : \Theta \to I\!R$ be any measurable function. Then it holds

$$(\psi'(\theta))^2 \leq E_\theta((T(\underline{X}) - \chi(\theta))^2) \ E_\theta\left(\left(\frac{\partial \log f_\theta(\underline{X})}{\partial \theta}\right)^2\right) \quad \forall \theta \in \Theta \quad (A).$$

Further, for any $\theta_0 \in \Theta$, either $\psi'(\theta_0) = 0$ and equality holds in (A) for $\theta = \theta_0$, or we have

$$E_{\theta_0}((T(\underline{X}) - \chi(\theta_0))^2) \geq \frac{(\psi'(\theta_0))^2}{E_{\theta_0}\left(\left(\frac{\partial \log f_\theta(\underline{X})}{\partial \theta}\right)^2\right)} \quad (B).$$

Finally, if equality holds in (B), then there exists a real number $K(\theta_0) \neq 0$ such that

$$T(\underline{X}) - \chi(\theta_0) = K(\theta_0) \left.\frac{\partial \log f_\theta(\underline{X})}{\partial \theta}\right|_{\theta = \theta_0} \quad (C)$$

with probability 1, provided that T is not a constant. ■

Note:

(i) Conditions (i), (ii), and (iii) are called **regularity conditions**. Conditions under which they hold can be found in Rohatgi, page 11–13, Parts 12 and 13.

(ii) The right hand side of inequality (B) is called *Cramér–Rao Lower Bound* of θ_0, or, in symbols $CRLB(\theta_0)$.

■

Proof:
From (ii), we get

$$\begin{aligned}
E_\theta\left(\frac{\partial}{\partial\theta}\log f_\theta(\underline{X})\right) &= \int\left(\frac{\partial}{\partial\theta}\log f_\theta(\underline{x})\right)f_\theta(\underline{x})d\underline{x} \\
&= \int\left(\frac{\partial}{\partial\theta}f_\theta(\underline{x})\right)\frac{1}{f_\theta(\underline{x})}f_\theta(\underline{x})d\underline{x} \\
&= \int\left(\frac{\partial}{\partial\theta}f_\theta(\underline{x})\right)d\underline{x} \\
&= 0 \\
\Longrightarrow E_\theta\left(\chi(\theta)\frac{\partial}{\partial\theta}\log f_\theta(\underline{X})\right) &= 0
\end{aligned}$$

From (iii), we get

$$\begin{aligned}
E_\theta\left(T(\underline{X})\frac{\partial}{\partial\theta}\log f_\theta(\underline{X})\right) &= \int\left(T(\underline{x})\frac{\partial}{\partial\theta}\log f_\theta(\underline{x})\right)f_\theta(\underline{x})d\underline{x} \\
&= \int\left(T(\underline{x})\frac{\partial}{\partial\theta}f_\theta(\underline{x})\right)\frac{1}{f_\theta(\underline{x})}f_\theta(\underline{x})d\underline{x} \\
&= \int\left(T(\underline{x})\frac{\partial}{\partial\theta}f_\theta(\underline{x})\right)d\underline{x} \\
&\overset{(iii)}{=} \frac{\partial}{\partial\theta}\left(\int T(\underline{x})f_\theta(\underline{x})d\underline{x}\right) \\
&= \frac{\partial}{\partial\theta}E(T(\underline{X})) \\
&= \frac{\partial}{\partial\theta}\psi(\theta) \\
&= \psi'(\theta) \\
\Longrightarrow E_\theta\left((T(\underline{X})-\chi(\theta))\frac{\partial}{\partial\theta}\log f_\theta(\underline{X})\right) &= \psi'(\theta)
\end{aligned}$$

$$\begin{aligned}\Longrightarrow (\psi'(\theta))^2 &= \left(E_\theta\left((T(\underline{X})-\chi(\theta))\frac{\partial}{\partial\theta}\log f_\theta(\underline{X})\right)\right)^2\\ &\overset{(*)}{\leq} E_\theta\left((T(\underline{X})-\chi(\theta))^2\right)E_\theta\left(\left(\frac{\partial}{\partial\theta}\log f_\theta(\underline{X})\right)^2\right),\end{aligned}$$

i.e., (A) holds. $(*)$ follows from the Cauchy–Schwarz–Inequality (Theorem 4.5.7 (ii)).

If $\psi'(\theta_0)\neq 0$, then the left–hand side of (A) is >0. Therefore, the right–hand side is >0. Thus,

$$E_{\theta_0}\left(\left(\frac{\partial}{\partial\theta}\log f_\theta(\underline{X})\right)^2\right)>0,$$

and (B) follows directly from (A).

If $\psi'(\theta_0)=0$, but equality does not hold in (A), then

$$E_{\theta_0}\left(\left(\frac{\partial}{\partial\theta}\log f_\theta(\underline{X})\right)^2\right)>0,$$

and (B) follows directly from (A) again.

Finally, if equality holds in (B), then $\psi'(\theta_0)\neq 0$ (because T is not constant). Thus, $MSE(\chi(\theta_0),T(\underline{X}))>0$. The Cauchy–Schwarz–Inequality (Theorem 4.5.7 (iii)) gives equality iff there exist constants $(\alpha,\beta)\in I\!R^2-\{(0,0)\}$ such that

$$P\left(\alpha(T(\underline{X})-\chi(\theta_0))+\beta\left(\frac{\partial}{\partial\theta}\log f_\theta(\underline{X})\Big|_{\theta=\theta_0}\right)=0\right)=1.$$

This implies $K(\theta_0)=-\frac{\beta}{\alpha}$ and (C) holds. Since T is not a constant, it also holds that $K(\theta_0)\neq 0$. ■

Example 8.5.2:
If we take $\chi(\theta)=\psi(\theta)$, we get from (B)

$$Var_\theta(T(\underline{X}))\geq\frac{(\psi'(\theta))^2}{E_\theta\left((\frac{\partial\log f_\theta(\underline{X})}{\partial\theta})^2\right)}\quad(*).$$

If we have $\psi(\theta)=\theta$, the inequality $(*)$ above reduces to

$$Var_\theta(T(\underline{X}))\geq\left(E_\theta\left(\left(\frac{\partial\log f_\theta(\underline{X})}{\partial\theta}\right)^2\right)\right)^{-1}.$$

Finally, if $\underline{X}=(X_1,\ldots,X_n)$ iid with identical $f_\theta(x)$, the inequality $(*)$ reduces to

$$Var_\theta(T(\underline{X}))\geq\frac{(\psi'(\theta))^2}{nE_\theta\left((\frac{\partial\log f_\theta(X_1)}{\partial\theta})^2\right)}.$$

■

Example 8.5.3:
Let $X_1, \ldots, X_n$ be iid $Bin(1,p)$. Let $X \sim Bin(n,p)$, $p \in \Theta = (0,1) \subset I\!R$. Let

$$\psi(p) = E(T(X)) = \sum_{x=0}^{n} T(x) \binom{n}{x} p^x (1-p)^{n-x}.$$

$\psi(p)$ is differentiable with respect to p under the summation sign since it is a finite polynomial in p.

Since $X = \sum_{i=1}^{n} X_i$ with $f_p(x_1) = p^{x_1}(1-p)^{1-x_1}, \quad x_1 \in \{0,1\}$

$$\begin{aligned}
\Longrightarrow \log f_p(x_1) &= x_1 \log p + (1-x_1)\log(1-p) \\
\Longrightarrow \frac{\partial}{\partial p} \log f_p(x_1) &= \frac{x_1}{p} - \frac{1-x_1}{1-p} = \frac{x_1(1-p) - p(1-x_1)}{p(1-p)} = \frac{x_1 - p}{p(1-p)} \\
\Longrightarrow E_p\left(\left(\frac{\partial}{\partial p} \log f_p(X_1)\right)^2\right) &= \frac{Var(X_1)}{p^2(1-p)^2} = \frac{1}{p(1-p)}
\end{aligned}$$

So, if $\psi(p) = \chi(p) = p$ and if T is unbiased for p, then

$$Var_p(T(X)) \geq \frac{1}{n\frac{1}{p(1-p)}} = \frac{p(1-p)}{n}.$$

Since $Var(\overline{X}) = \frac{p(1-p)}{n}$, $\overline{X}$ attains the CRLB. Therefore, $\overline{X}$ is the UMVUE. ■

Lecture 13:
Fr 02/09/01

Example 8.5.4:
Let $X \sim U(0,\theta)$, $\theta \in \Theta = (0,\infty) \subset I\!R$.

$$\begin{aligned}
f_\theta(x) &= \frac{1}{\theta} I_{(0,\theta)}(x) \\
\Longrightarrow \log f_\theta(x) &= -\log\theta \\
\Longrightarrow (\frac{\partial}{\partial\theta} \log f_\theta(x))^2 &= \frac{1}{\theta^2} \\
\Longrightarrow E_\theta\left(\left(\frac{\partial}{\partial\theta} \log f_\theta(X)\right)^2\right) &= \frac{1}{\theta^2}
\end{aligned}$$

Thus, the CRLB is $\frac{\theta^2}{n}$.

We know that $\frac{n+1}{n}X_{(n)}$ is the UMVUE since it is a function of a complete sufficient statistic $X_{(n)}$ (see Homework) and $E(X_{(n)}) = \frac{n}{n+1}\theta$. It is

$$Var\left(\frac{n+1}{n}X_{(n)}\right) = \frac{\theta^2}{n(n+2)} < \frac{\theta^2}{n} \quad ???$$

How is this possible? Quite simple, since one of the required conditions for Theorem 8.5.1 does not hold. The support of X depends on θ. ■

Theorem 8.5.5: **Chapman, Robbins, Kiefer Inequality (CRK Inequality)**
Let $\Theta \subseteq I\!R$. Let $\{f_\theta \ : \ \theta \in \Theta\}$ be a family of pdf's or pmf's. Let $\psi(\theta)$ be defined on Θ. Let T be an unbiased estimate of $\psi(\theta)$ such that $E_\theta(T^2) < \infty \ \ \forall \theta \in \Theta$.

If $\theta \neq \vartheta$, θ and $\vartheta \in \Theta$, assume that $f_\theta(x)$ and $f_\vartheta(x)$ are different. Also assume that there exists such a $\vartheta \in \Theta$ such that $\theta \neq \vartheta$ and

$$S(\theta) = \{\underline{x} : f_\theta(\underline{x}) > 0\} \supset S(\vartheta) = \{\underline{x} : f_\vartheta(\underline{x}) > 0\}.$$

Then it holds that

$$Var_\theta(T(\underline{X})) \geq \sup_{\{\vartheta \ : \ S(\vartheta) \subset S(\theta), \ \vartheta \neq \theta\}} \frac{(\psi(\vartheta) - \psi(\theta))^2}{Var_\theta\left(\frac{f_\vartheta(\underline{X})}{f_\theta(\underline{X})}\right)} \ \ \forall \theta \in \Theta.$$

Proof:
Since T is unbiased, it follows

$$E_\vartheta(T(\underline{X})) = \psi(\vartheta) \ \ \forall \vartheta \in \Theta.$$

For $\vartheta \neq \theta$ and $S(\vartheta) \subset S(\theta)$, it follows

$$\int_{S(\theta)} T(\underline{x}) \frac{f_\vartheta(\underline{x}) - f_\theta(\underline{x})}{f_\theta(\underline{x})} f_\theta(\underline{x}) d\underline{x} = E_\vartheta(T(\underline{X})) - E_\theta(T(\underline{X})) = \psi(\vartheta) - \psi(\theta)$$

and

$$0 = \int_{S(\theta)} \frac{f_\vartheta(\underline{x}) - f_\theta(\underline{x})}{f_\theta(\underline{x})} f_\theta(\underline{x}) d\underline{x} = E_\theta\left(\frac{f_\vartheta(\underline{X})}{f_\theta(\underline{X})} - 1\right).$$

Therefore

$$Cov_\theta\left(T(\underline{X}), \frac{f_\vartheta(\underline{X})}{f_\theta(\underline{X})} - 1\right) = \psi(\vartheta) - \psi(\theta).$$

It follows by the Cauchy–Schwarz–Inequality (Theorem 4.5.7 (ii)) that

$$\begin{aligned}
(\psi(\vartheta) - \psi(\theta))^2 &= \left(Cov_\theta\left(T(\underline{X}), \frac{f_\vartheta(\underline{X})}{f_\theta(\underline{X})} - 1\right)\right)^2 \\
&\leq Var_\theta(T(\underline{X})) Var_\theta\left(\frac{f_\vartheta(\underline{X})}{f_\theta(\underline{X})} - 1\right) \\
&= Var_\theta(T(\underline{X})) Var_\theta\left(\frac{f_\vartheta(\underline{X})}{f_\theta(\underline{X})}\right).
\end{aligned}$$

Thus,

$$Var_\theta(T(\underline{X})) \geq \frac{(\psi(\vartheta) - \psi(\theta))^2}{Var_\theta\left(\frac{f_\vartheta(\underline{X})}{f_\theta(\underline{X})}\right)}.$$

Finally, we take the supremum of the right–hand side with respect to $\{\vartheta \ : \ S(\vartheta) \subset S(\theta)$, $\vartheta \neq \theta\}$, which completes the proof. ■

Note:

(i) The CRK inequality holds without the previous regularity conditions.

(ii) An alternative form of the CRK inequality is:

Let $\theta, \theta+\delta,\ \delta \neq 0$, be distinct with $S(\theta+\delta) \subset S(\theta)$. Let $\psi(\theta) = \theta$. Define

$$J = J(\theta, \delta) = \frac{1}{\delta^2}\left(\left(\frac{f_{\theta+\delta}(\underline{X})}{f_\theta(\underline{X})}\right)^2 - 1\right).$$

Then the CRK inequality reads as

$$Var_\theta(T(\underline{X})) \geq \frac{1}{\inf\limits_{\delta} E_\theta(J)}$$

with the infimum taken over $\delta \neq 0$ such that $S(\theta+\delta) \subset S(\theta)$.

(iii) The CRK inequality works for discrete Θ, the CRLB does not work in such cases.

■

Example 8.5.6:
Let $X \sim U(0, \theta),\ \theta > 0$. The required conditions for the CRLB are not met. Recall from Example 8.5.4 that $\frac{n+1}{n}X_{(n)}$ is UMVUE with $Var(\frac{n+1}{n}X_{(n)}) = \frac{\theta^2}{n(n+2)} < \frac{\theta^2}{n} = CRLB$.

Let $\psi(\theta) = \theta$. If $\vartheta < \theta$, then $S(\vartheta) \subset S(\theta)$. It is

$$\begin{aligned}
E_\theta\left(\left(\frac{f_\vartheta(X)}{f_\theta(X)}\right)^2\right) &= \int_0^\vartheta \left(\frac{\theta}{\vartheta}\right)^2 \frac{1}{\theta}dx = \frac{\theta}{\vartheta} \\
E_\theta\left(\frac{f_\vartheta(X)}{f_\theta(X)}\right) &= \int_0^\vartheta \frac{\theta}{\vartheta}\frac{1}{\theta}dx = 1 \\
\Longrightarrow Var_\theta(T(X)) &\geq \sup_{\{\vartheta\ :\ 0<\vartheta<\theta\}} \frac{(\vartheta-\theta)^2}{\frac{\theta}{\vartheta} - 1} = \sup_{\{\vartheta\ :\ 0<\vartheta<\theta\}} (\vartheta(\theta-\vartheta)) = \frac{\theta^2}{4}
\end{aligned}$$

Since X is complete and sufficient and $2X$ is unbiased for θ, so $T(X) = 2X$ is the UMVUE. It is

$$Var_\theta(2X) = 4Var_\theta(X) = 4\frac{\theta^2}{12} = \frac{\theta^2}{3} > \frac{\theta^2}{4}.$$

Since the CRK lower bound is not achieved by the UMVUE, it is not achieved by any unbiased estimate of θ. ■

Definition 8.5.7:
Let T_1, T_2 be unbiased estimates of θ with $E_\theta(T_1^2) < \infty$ and $E_\theta(T_2^2) < \infty \quad \forall \theta \in \Theta$. We define the **efficiency** of T_1 relative to T_2 by

$$eff_\theta(T_1, T_2) = \frac{Var_\theta(T_1)}{Var_\theta(T_2)}$$

and say that T_1 is **more efficient** than T_2 if $eff_\theta(T_1, T_2) < 1$. ■

Definition 8.5.8:
Assume the regularity conditions of Theorem 8.5.1 are satisfied by a family of cdf's $\{F_\theta : \theta \in \Theta\}$, $\Theta \subseteq I\!R$. An unbiased estimate T for θ is **most efficient** for the family $\{F_\theta\}$ if

$$Var_\theta(T) = \left(E_\theta \left((\frac{\partial \log f_\theta(\underline{X})}{\partial \theta})^2 \right) \right)^{-1}.$$

■

Definition 8.5.9:
Let T be the most efficient estimate for the family of cdf's $\{F_\theta : \theta \in \Theta\}$, $\Theta \subseteq I\!R$. Then the **efficiency** of any unbiased T_1 of θ is defined as

$$eff_\theta(T_1) = eff_\theta(T_1, T) = \frac{Var_\theta(T_1)}{Var_\theta(T)}.$$

■

Definition 8.5.10:
T_1 is **asymptotically (most) efficient** if T_1 is asymptotically unbiased, i.e., $\lim\limits_{n \to \infty} E_\theta(T_1) = \theta$, and $\lim\limits_{n \to \infty} eff_\theta(T_1) = 1$, where n is the sample size. ■

Lecture 14:
Mo 02/12/01

Theorem 8.5.11:
A necessary and sufficient condition for an estimate T of θ to be most efficient is that T is sufficient and

$$\frac{1}{K(\theta)}(T(\underline{x}) - \theta) = \frac{\partial \log f_\theta(\underline{x})}{\partial \theta} \quad \forall \theta \in \Theta \quad (*),$$

where $K(\theta)$ is defined as in Theorem 8.5.1 and the regularity conditions for Theorem 8.5.1 hold.

Proof:
"$\Longrightarrow$:"
Theorem 8.5.1 says that if T is most efficient, then $(*)$ holds.

Assume that $\Theta = I\!R$. We define

$$C(\theta_0) = \int_{-\infty}^{\theta_0} \frac{1}{K(\theta)} d\theta, \quad \psi(\theta_0) = \int_{-\infty}^{\theta_0} \frac{\theta}{K(\theta)} d\theta, \text{ and } \lambda(\underline{x}) = \lim_{\theta \to -\infty} \log f_\theta(\underline{x}) - c(\underline{x}).$$

Integrating $(*)$ with respect to θ gives

$$\begin{aligned}
\int_{-\infty}^{\theta_0} \frac{1}{K(\theta)} T(\underline{x}) d\theta - \int_{-\infty}^{\theta_0} \frac{\theta}{K(\theta)} d\theta &= \int_{-\infty}^{\theta_0} \frac{\partial \log f_\theta(\underline{x})}{\partial \theta} d\theta \\
\Longrightarrow T(\underline{x}) C(\theta_0) - \psi(\theta_0) &= \log f_\theta(\underline{x})|_{-\infty}^{\theta_0} + c(\underline{x}) \\
\Longrightarrow T(\underline{x}) C(\theta_0) - \psi(\theta_0) &= \log f_{\theta_0}(\underline{x}) - \lim_{\theta \to -\infty} \log f_\theta(\underline{x}) + c(\underline{x}) \\
\Longrightarrow T(\underline{x}) C(\theta_0) - \psi(\theta_0) &= \log f_{\theta_0}(\underline{x}) - \lambda(\underline{x})
\end{aligned}$$

Therefore,

$$f_{\theta_0}(\underline{x}) = \exp(T(\underline{x}) C(\theta_0) - \psi(\theta_0) + \lambda(\underline{x}))$$

which belongs to an exponential family. Thus, T is sufficient.

"$\Longleftarrow$:"

From $(*)$, we get

$$E_\theta \left(\left(\frac{\partial \log f_\theta(\underline{X})}{\partial \theta} \right)^2 \right) = \frac{1}{(K(\theta))^2} Var_\theta(T(\underline{X})).$$

Additionally, it holds

$$E_\theta \left((T(\underline{X}) - \theta) \frac{\partial \log f_\theta(\underline{X})}{\partial \theta} \right) = 1$$

as shown in the Proof of Theorem 8.5.1.

Using $(*)$ in the line above, we get

$$K(\theta) E_\theta \left(\left(\frac{\partial \log f_\theta(\underline{X})}{\partial \theta} \right)^2 \right) = 1,$$

i.e.,

$$K(\theta) = \left(E_\theta \left(\left(\frac{\partial \log f_\theta(\underline{X})}{\partial \theta} \right)^2 \right) \right)^{-1}.$$

Therefore,

$$Var_\theta(T(\underline{X})) = \left(E_\theta \left(\left(\frac{\partial \log f_\theta(\underline{X})}{\partial \theta} \right)^2 \right) \right)^{-1},$$

i.e., T is most efficient for θ. ■

<u>Note:</u>

Instead of saying "*a necessary and sufficient condition for an estimate T of θ to be most efficient ...*" in the previous Theorem, we could say that "*an estimate T of θ is most efficient iff ...*", i.e., "*necessary and sufficient*" means the same as "*iff*". ■

8.6 The Method of Moments

Definition 8.6.1:

Let $X_1, \ldots, X_n$ be iid with pdf (or pmf) f_θ, $\theta \in \Theta$. We assume that first k moments $m_1, \ldots, m_k$ of f_θ exist. If θ can be written as

$$\theta = h(m_1, \ldots, m_k),$$

where $h : I\!R^k \to I\!R$ is a Borel–measurable function, the **method of moments estimate (mom)** of θ is

$$\hat{\theta}_{mom} = T(X_1, \ldots, X_n) = h(\frac{1}{n}\sum_{i=1}^n X_i, \frac{1}{n}\sum_{i=1}^n X_i^2, \ldots, \frac{1}{n}\sum_{i=1}^n X_i^k).$$

■

Note:

(i) The Definition above can also be used to estimate joint moments. For example, we use $\frac{1}{n}\sum\limits_{i=1}^n X_i Y_i$ to estimate $E(XY)$.

(ii) Since $E(\frac{1}{n}\sum\limits_{i=1}^n X_i^j) = m_j$, method of moment estimates are unbiased for the population moments. The WLLN and the CLT say that these estimates are consistent and asymptotically Normal as well.

(iii) If θ is not a linear function of the population moments, $\hat{\theta}_{mom}$ will, in general, not be unbiased. However, it will be consistent and (usually) asymptotically Normal.

(iv) Method of moments estimates do not exist if the related moments do not exist.

(v) Method of moments estimates may not be unique. If there exist multiple choices for the mom, one usually takes the estimate involving the lowest–order sample moment.

(vi) Alternative method of moment estimates can be obtained from central moments (rather than from raw moments) or by using moments other than the first k moments.

■

Example 8.6.2:

Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$.

Since $\mu = m_1$, it is $\hat{\mu}_{mom} = \overline{X}$.

This is an unbiased, consistent and asymptotically Normal estimate.

Since $\sigma = \sqrt{m_2 - m_1^2}$, it is $\hat{\sigma}_{mom} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2}$.

This is a consistent, asymptotically Normal estimate. However, it is not unbiased. ■

Example 8.6.3:

Let $X_1, \ldots, X_n$ be iid Poisson(λ).

We know that $E(X_1) = Var(X_1) = \lambda$.

Thus, $\overline{X}$ and $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ are possible choices for the mom of λ. Due to part (v) of the Note above, one uses $\hat{\lambda}_{mom} = \overline{X}$. ■

8.7 Maximum Likelihood Estimation

Definition 8.7.1:
Let $(X_1, \ldots, X_n)$ be an n–rv with pdf (or pmf) $f_\theta(x_1, \ldots, x_n)$, $\theta \in \Theta$. We call the function

$$L(\theta; x_1, \ldots, x_n) = f_\theta(x_1, \ldots, x_n)$$

of θ the **likelihood function**. ■

Note:

(i) Often θ is a vector of parameters.

(ii) If $(X_1, \ldots, X_n)$ are iid with pdf (or pmf) $f_\theta(x)$, then $L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i)$.

■

Definition 8.7.2:
A **maximum likelihood estimate** (**MLE**) is a non–constant estimate $\hat{\theta}_{ML}$ such that

$$L(\hat{\theta}_{ML}; x_1, \ldots, x_n) = \sup_{\theta \in \Theta} L(\theta; x_1, \ldots, x_n).$$

■

Note:
It is often convenient to work with $\log L$ when determining the maximum likelihood estimate. Since the log is monotone, the maximum is the same. ■

Example 8.7.3:
Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$, where μ and σ^2 are unknown.

$$\begin{aligned} L(\mu, \sigma^2; x_1, \ldots, x_n) &= \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right) \\ \Longrightarrow \log L(\mu, \sigma^2; x_1, \ldots, x_n) &= -\frac{n}{2} \log \sigma^2 - \frac{n}{2} \log(2\pi) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \end{aligned}$$

The MLE must satisfy

$$\begin{aligned} \frac{\partial \log L}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0 \quad (A) \\ \frac{\partial \log L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0 \quad (B) \end{aligned}$$

These are the two likelihood equations. From equation (A) we get $\hat{\mu}_{ML} = \overline{X}$. Substituting this for μ into equation (B) and solving for σ^2, we get $\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Note that $\hat{\sigma}^2_{ML}$ is biased for σ^2.

Formally, we still have to verify that we found the maximum (and not a minimum) and that there is no parameter θ at the edge of the parameter space Θ such that the likelihood function does not take its absolute maximum which is not detectable by using our approach for local extrema. ■

Lecture 15: We 02/14/01

Example 8.7.4:

Let $X_1, \ldots, X_n$ be iid $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$.

$$L(\theta; x_1, \ldots, x_n) = \begin{cases} 1, & \text{if } \theta - \frac{1}{2} \leq x_i \leq \theta + \frac{1}{2} \ \ \forall i = 1, \ldots, n \\ 0, & \text{otherwise} \end{cases}$$

Therefore, any $\hat{\theta}(\underline{X})$ such that $\max(\underline{X}) - \frac{1}{2} \leq \hat{\theta}(\underline{X}) \leq \min(\underline{X}) + \frac{1}{2}$ is an MLE. Obviously, the MLE is not unique. ■

Example 8.7.5:

Let $X \sim Bin(1, p), \quad p \in [\frac{1}{4}, \frac{3}{4}]$.

$$L(p; x) = \begin{cases} p, & \text{if } x = 1 \\ 1 - p, & \text{if } x = 0 \end{cases}$$

This is maximized by

$$\hat{p} = \begin{cases} \frac{3}{4}, & \text{if } x = 1 \\ \frac{1}{4}, & \text{if } x = 0 \end{cases} = \frac{2x+1}{4}$$

It is

$$\begin{aligned}
E_p(\hat{p}) &= \frac{3}{4}p + \frac{1}{4}(1-p) = \frac{1}{2}p + \frac{1}{4} \\
MSE_p(\hat{p}) &= E_p((\hat{p} - p)^2) \\
&= E_p((\frac{2X+1}{4} - p)^2) \\
&= \frac{1}{16}E_p((2X + 1 - 4p)^2) \\
&= \frac{1}{16}E_p(4X^2 + 2 \cdot 2X - 2 \cdot 8pX - 2 \cdot 4p + 1 + 16p^2) \\
&= \frac{1}{16}(4(p(1-p) + p^2) + 4p - 16p^2 - 8p + 1 + 16p^2) \\
&= \frac{1}{16}
\end{aligned}$$

So $\hat{p}$ is biased with $MSE_p(\hat{p}) = \frac{1}{16}$. If we compare this with $\tilde{p} = \frac{1}{2}$ regardless of the data, we have

$$MSE_p(\frac{1}{2}) = E_p((\frac{1}{2} - p)^2) = (\frac{1}{2} - p)^2 \leq \frac{1}{16} \quad \forall p \in [\frac{1}{4}, \frac{3}{4}].$$

Thus, in this example the MLE is worse than the trivial estimate when comparing their MSE's.

■

<u>Theorem 8.7.6:</u>
Let T be a sufficient statistic for $f_\theta(\underline{x})$, $\theta \in \Theta$. If a unique MLE of θ exists, it is a function of T.

<u>Proof:</u>
Since T is sufficient, we can write

$$f_\theta(\underline{x}) = h(\underline{x}) g_\theta(T(\underline{x}))$$

due to the Factorization Criterion (Theorem 8.3.5). Maximizing the likelihood function with respect to θ takes $h(\underline{x})$ as a constant and therefore is equivalent to maximizing $g_\theta(\underline{x})$ with respect to θ. But $g_\theta(\underline{x})$ involves $\underline{x}$ only through T. ■

<u>Note:</u>

(i) MLE's may not be unique (however they frequently are).

(ii) MLE's are not necessarily unbiased.

(iii) MLE's may not exist.

(iv) If a unique MLE exists, it is a function of a sufficient statistic.

(v) Often (but not always), the MLE will be a sufficient statistic itself.

■

<u>Theorem 8.7.7:</u>
Suppose the regularity conditions of Theorem 8.5.1 hold and θ belongs to an open interval in $I\!R$. If an estimate $\hat{\theta}$ of θ attains the CRLB, it is the unique MLE.

<u>Proof:</u>
If $\hat{\theta}$ attains the CRLB, it follows by Theorem 8.5.1 that

$$\frac{\partial \log f_\theta(\underline{X})}{\partial \theta} = \frac{1}{K(\theta)}(\hat{\theta}(\underline{X}) - \theta) \quad \text{w.p. 1.}$$

Thus, $\hat{\theta}$ satisfies the likelihood equations.

We define $A(\theta) = \frac{1}{K(\theta)}$. Then it follows

$$\frac{\partial^2 \log f_\theta(\underline{X})}{\partial \theta^2} = A'(\theta)(\hat{\theta}(\underline{X}) - \theta) - A(\theta).$$

The Proof of Theorem 8.5.11 gives us

$$A(\theta) = E_\theta \left(\left(\frac{\partial \log f_\theta(\underline{X})}{\partial \theta} \right)^2 \right) > 0.$$

So

$$\left. \frac{\partial^2 \log f_\theta(\underline{X})}{\partial \theta^2} \right|_{\theta = \hat{\theta}} = -A(\theta) < 0,$$

i.e., $\log f_\theta(\underline{X})$ has a maximum in $\hat{\theta}$. Thus, $\hat{\theta}$ is the MLE. ■

Note:
The previous Theorem does not imply that every MLE is most efficient. ■

Theorem 8.7.8:
Let $\{f_\theta \ : \ \theta \in \Theta\}$ be a family of pdf's (or pmf's) with $\Theta \subseteq I\!R^k, \quad k \geq 1$. Let $h : \Theta \to \Delta$ be a mapping of Θ onto $\Delta \subseteq I\!R^p, \quad 1 \leq p \leq k$. If $\hat{\theta}$ is an MLE of θ, then $h(\hat{\theta})$ is an MLE of $h(\theta)$.

Proof:
For each $\delta \in \Delta$, we define

$$\Theta_\delta = \{\theta \ : \ \theta \in \Theta, h(\theta) = \delta\}$$

and

$$M(\delta; \underline{x}) = \sup_{\theta \in \Theta_\delta} L(\theta; \underline{x}),$$

the likelihood function induced by h.

Let $\hat{\theta}$ be an MLE and a member of $\Theta_{\hat{\delta}}$, where $\hat{\delta} = h(\hat{\theta})$. It holds

$$M(\hat{\delta}; \underline{x}) = \sup_{\theta \in \Theta_{\hat{\delta}}} L(\theta; \underline{x}) \geq L(\hat{\theta}; \underline{x}),$$

but also

$$M(\hat{\delta}; \underline{x}) \leq \sup_{\delta \in \Delta} M(\delta; \underline{x}) = \sup_{\delta \in \Delta} \left(\sup_{\theta \in \Theta_\delta} L(\theta; \underline{x}) \right) = \sup_{\theta \in \Theta} L(\theta; \underline{x}) = L(\hat{\theta}; \underline{x}).$$

Therefore,

$$M(\hat{\delta}; \underline{x}) = L(\hat{\theta}; \underline{x}) = \sup_{\delta \in \Delta} M(\delta; \underline{x}).$$

Thus, $\hat{\delta} = h(\hat{\theta})$ is an MLE. ■

Example 8.7.9:
Let $X_1, \ldots, X_n$ be iid $Bin(1, p)$. Let $h(p) = p(1-p)$.

Since the MLE of p is $\hat{p} = \overline{X}$, the MLE of $h(p)$ is $h(\hat{p}) = \overline{X}(1 - \overline{X})$. ■

Theorem 8.7.10:
Consider the following conditions a pdf f_θ can fulfill:

(i) $\frac{\partial \log f_\theta}{\partial \theta}, \frac{\partial^2 \log f_\theta}{\partial \theta^2}, \frac{\partial^3 \log f_\theta}{\partial \theta^3}$ exist for all $\theta \in \Theta$ for all x. Also,

$$\int_{-\infty}^{\infty} \frac{\partial f_\theta(x)}{\partial \theta} dx = E_\theta \left(\frac{\partial \log f_\theta(X)}{\partial \theta} \right) = 0 \quad \forall \theta \in \Theta.$$

(ii) $\int_{-\infty}^{\infty} \frac{\partial^2 f_\theta(x)}{\partial \theta^2} dx = 0 \quad \forall \theta \in \Theta.$

(iii) $-\infty < \int_{-\infty}^{\infty} \frac{\partial^2 \log f_\theta(x)}{\partial \theta^2} f_\theta(x) dx < 0 \quad \forall \theta \in \Theta.$

(iv) There exists a function $H(x)$ such that for all $\theta \in \Theta$:

$$\left| \frac{\partial^3 \log f_\theta(x)}{\partial \theta^3} \right| < H(x) \text{ and } \int_{-\infty}^{\infty} H(x) f_\theta(x) dx = M(\theta) < \infty.$$

(v) There exists a function $g(\theta)$ that is positive and twice differentiable for every $\theta \in \Theta$ and there exists a function $H(x)$ such that for all $\theta \in \Theta$:

$$\left| \frac{\partial^2}{\partial \theta^2} \left[g(\theta) \frac{\partial \log f_\theta(x)}{\partial \theta} \right] \right| < H(x) \text{ and } \int_{-\infty}^{\infty} H(x) f_\theta(x) dx = M(\theta) < \infty.$$

In case that multiple of these conditions are fulfilled, we can make the following statements:

(i) (Cramér) Conditions (i), (iii), and (iv) imply that, with probability approaching 1, as $n \to \infty$, the likelihood equation has a consistent solution.

(ii) (Cramér) Conditions (i), (ii), (iii), and (iv) imply that a consistent solution $\hat{\theta}_n$ of the likelihood equation is asymptotically Normal, i.e.,

$$\frac{\sqrt{n}}{\sigma}(\hat{\theta}_n - \theta) \xrightarrow{d} Z$$

where $Z \sim N(0,1)$ and $\sigma^2 = \left(E_\theta \left(\left(\frac{\partial \log f_\theta(X)}{\partial \theta} \right)^2 \right) \right)^{-1}$.

(iii) (Kulldorf) Conditions (i), (iii), and (v) imply that, with probability approaching 1, as $n \to \infty$, the likelihood equation has a consistent solution.

(iv) (Kulldorf) Conditions (i), (ii), (iii), and (v) imply that a consistent solution $\hat{\theta}_n$ of the likelihood equation is asymptotically Normal.

■

Note:

In case of a pmf f_θ, we can define similar conditions as in Theorem 8.7.10. ■

Lecture 16:
Fr 02/16/01

8.8 Decision Theory — Bayes and Minimax Estimation

Let $\{f_\theta : \theta \in \Theta\}$ be a family of pdf's (or pmf's). Let $X_1, \ldots, X_n$ be a sample from f_θ. Let $\mathcal{A}$ be the set of possible **actions** (or decisions) that are open to the statistician in a given situation , e.g.,

$\mathcal{A} = \{\text{reject } H_0, \text{ do not reject } H_0\}$ (Hypothesis testing, see Chapter 9)

$\mathcal{A} =$ artefact found is of {Greek, Roman} origin (Classification)

$\mathcal{A} = \Theta$ (Estimation)

Definition 8.8.1:
A **decision function** d is a statistic, i.e., a Borel–measurable function, that maps $I\!R^n$ into $\mathcal{A}$. If $\underline{X} = \underline{x}$ is observed, the statistician takes action $d(\underline{x}) \in \mathcal{A}$. ■

Note:
For the remainder of this Section, we are restricting ourselves to $\mathcal{A} = \Theta$, i.e., we are facing the problem of estimation. ■

Definition 8.8.2:
A non–negative function L that maps $\Theta \times \mathcal{A}$ into $I\!R$ is called a **loss function**. The value $L(\theta, a)$ is the loss incurred to the statistician if he/she takes action a when θ is the true parameter value. ■

Definition 8.8.3:
Let $\mathcal{D}$ be a class of decision functions that map $I\!R^n$ into $\mathcal{A}$. Let L be a loss function on $\Theta \times \mathcal{A}$. The function R that maps $\Theta \times \mathcal{D}$ into $I\!R$ is defined as

$$R(\theta, d) = E_\theta(L(\theta, d(\underline{X})))$$

and is called the **risk function** of d at θ. ■

Example 8.8.4:
Let $\mathcal{A} = \Theta \subseteq I\!R$. Let $L(\theta, a) = (\theta - a)^2$. Then it holds that

$$R(\theta, d) = E_\theta(L(\theta, d(\underline{X}))) = E_\theta((\theta - d(\underline{X}))^2) = E_\theta((\theta - \hat{\theta})^2).$$

Note that this is just the MSE. If $\hat{\theta}$ is unbiased, this would just be $Var(\hat{\theta})$. ■

Note:
The basic problem of decision theory is that we would like to find a decision function $d \in \mathcal{D}$ such that $R(\theta, d)$ is minimized for all $\theta \in \Theta$. Unfortunately, this is usually not possible. ■

Definition 8.8.5:
The **minimax principle** is to choose the decision function $d^* \in \mathcal{D}$ such that

$$\max_{\theta \in \Theta} R(\theta, d^*) \leq \max_{\theta \in \Theta} R(\theta, d) \quad \forall d \in \mathcal{D}.$$

■

Note:
If the problem of interest is an estimation problem, we call a d^* that satisifies the condition in Definition 8.8.5 a **minimax estimate** of θ. ■

Example 8.8.6:
Let $X \sim Bin(1, p), \quad p \in \Theta = \{\frac{1}{4}, \frac{3}{4}\} = \mathcal{A}$.

We consider the following loss function:

p	a	$L(p,a)$
$\frac{1}{4}$	$\frac{1}{4}$	0
$\frac{1}{4}$	$\frac{3}{4}$	2
$\frac{3}{4}$	$\frac{1}{4}$	5
$\frac{3}{4}$	$\frac{3}{4}$	0

The set of decision functions consists of the following four functions:

$$\begin{aligned} d_1(0) &= \frac{1}{4}, & d_1(1) &= \frac{1}{4} \\ d_2(0) &= \frac{1}{4}, & d_2(1) &= \frac{3}{4} \\ d_3(0) &= \frac{3}{4}, & d_3(1) &= \frac{1}{4} \\ d_4(0) &= \frac{3}{4}, & d_4(1) &= \frac{3}{4} \end{aligned}$$

First, we evaluate the loss function for these four decision functions:

$$\begin{aligned} L(\frac{1}{4}, d_1(0)) &= L(\frac{1}{4}, \frac{1}{4}) = 0 \\ L(\frac{1}{4}, d_1(1)) &= L(\frac{1}{4}, \frac{1}{4}) = 0 \\ L(\frac{3}{4}, d_1(0)) &= L(\frac{3}{4}, \frac{1}{4}) = 5 \end{aligned}$$

$$\begin{aligned}
L(\frac{3}{4}, d_1(1)) &= L(\frac{3}{4}, \frac{1}{4}) = 5 \\[2ex]
L(\frac{1}{4}, d_2(0)) &= L(\frac{1}{4}, \frac{1}{4}) = 0 \\
L(\frac{1}{4}, d_2(1)) &= L(\frac{1}{4}, \frac{3}{4}) = 2 \\
L(\frac{3}{4}, d_2(0)) &= L(\frac{3}{4}, \frac{1}{4}) = 5 \\
L(\frac{3}{4}, d_2(1)) &= L(\frac{3}{4}, \frac{3}{4}) = 0 \\[2ex]
L(\frac{1}{4}, d_3(0)) &= L(\frac{1}{4}, \frac{3}{4}) = 2 \\
L(\frac{1}{4}, d_3(1)) &= L(\frac{1}{4}, \frac{1}{4}) = 0 \\
L(\frac{3}{4}, d_3(0)) &= L(\frac{3}{4}, \frac{3}{4}) = 0 \\
L(\frac{3}{4}, d_3(1)) &= L(\frac{3}{4}, \frac{1}{4}) = 5 \\[2ex]
L(\frac{1}{4}, d_4(0)) &= L(\frac{1}{4}, \frac{3}{4}) = 2 \\
L(\frac{1}{4}, d_4(1)) &= L(\frac{1}{4}, \frac{3}{4}) = 2 \\
L(\frac{3}{4}, d_4(0)) &= L(\frac{3}{4}, \frac{3}{4}) = 0 \\
L(\frac{3}{4}, d_4(1)) &= L(\frac{3}{4}, \frac{3}{4}) = 0
\end{aligned}$$

Then, the risk function

$$R(p, d_i(X)) = E_p(L(p, d(X))) = L(p, d(0)) \cdot P_p(X = 0) + L(p, d(1)) \cdot P_p(X = 1)$$

takes the following values:

i	$p = \frac{1}{4}$: $R(\frac{1}{4}, d_i)$	$p = \frac{3}{4}$: $R(\frac{3}{4}, d_i)$	$\max_{p \in \{1/4,\ 3/4\}} R(p, d_i)$
1	0	5	5
2	$\frac{3}{4} \cdot 0 + \frac{1}{4} \cdot 2 = \frac{1}{2}$	$\frac{1}{4} \cdot 5 + \frac{3}{4} \cdot 0 = \frac{5}{4}$	$\frac{5}{4}$
3	$\frac{3}{4} \cdot 2 + \frac{1}{4} \cdot 0 = \frac{3}{2}$	$\frac{1}{4} \cdot 0 + \frac{3}{4} \cdot 5 = \frac{15}{4}$	$\frac{15}{4}$
4	2	0	2

Hence,

$$\min_{i \in \{1,\ 2,\ 3,\ 4\}} \quad \max_{p \in \{1/4,\ 3/4\}} R(p, d_i) = \frac{5}{4}.$$

Thus, d_2 is the minimax estimate. ■

Note:
Minimax estimation does not require any unusual assumptions. However, it tends to be very conservative. ■

Definition 8.8.7:
Suppose we consider θ to be a rv with pdf $\pi(\theta)$ on Θ. We call π the **a priori distribution** (or **prior distribution**). ■

Note:
$f(\underline{x} \mid \theta)$ is the conditional density of $\underline{x}$ given a fixed θ. The joint density of $\underline{x}$ and θ is

$$f(\underline{x}, \theta) = \pi(\theta) f(\underline{x} \mid \theta),$$

the marginal density of $\underline{x}$ is

$$g(\underline{x}) = \int f(\underline{x}, \theta) d\theta,$$

and the **a posteriori distribution** (or **posterior distribution**), which gives the distribution of θ after sampling, has pdf (or pmf)

$$h(\theta \mid \underline{x}) = \frac{f(\underline{x}, \theta)}{g(\underline{x})}.$$

■

Definition 8.8.8:
The **Bayes risk** of a decision function d is defined as

$$R(\pi, d) = E_\pi(R(\theta, d)),$$

where π is the a priori distribution. ■

Note:
If θ is a continuous rv and $\underline{X}$ is of continuous type, then

$$\begin{aligned} R(\pi, d) &= E_\pi(R(\theta, d)) \\ &= \int R(\theta, d) \ \pi(\theta) \ d\theta \\ &= \int E_\theta(L(\theta, d(\underline{X}))) \ \pi(\theta) \ d\theta \\ &= \int \left(\int L(\theta, d(\underline{x})) \ f(\underline{x} \mid \theta) \ d\underline{x} \right) \ \pi(\theta) \ d\theta \\ &= \int \int L(\theta, d(\underline{x})) \ f(\underline{x} \mid \theta) \ \pi(\theta) \ d\underline{x} \ d\theta \end{aligned}$$

$$= \int \int L(\theta, d(\underline{x}))\ f(\underline{x}, \theta)\ \ d\underline{x}\ d\theta$$

$$= \int g(\underline{x}) \left(\int L(\theta, d(\underline{x})) h(\theta \mid \underline{x})\ d\theta \right)\ d\underline{x}$$

Similar expressions can be written if θ and/or $\underline{X}$ are discrete. ■

Lecture 17: Tu 02/20/01

Definition 8.8.9:
A decision function d^* is called a **Bayes rule** if d^* minimizes the Bayes risk, i.e., if

$$R(\pi, d^*) = \inf_{d \in \mathcal{D}} R(\pi, d).$$

■

Theorem 8.8.10:
Let $\mathcal{A} = \Theta \subseteq I\!R$. Let $L(\theta, d(\underline{x})) = (\theta - d(\underline{x}))^2$. In this case, a Bayes rule is

$$d(\underline{x}) = E(\theta \mid \underline{X} = \underline{x}).$$

Proof:
Minimizing

$$R(\pi, d) = \int g(\underline{x}) \left(\int (\theta - d(\underline{x}))^2 h(\theta \mid \underline{x})\ d\theta \right)\ d\underline{x},$$

where g is the marginal pdf of $\underline{X}$ and h is the conditional pdf of θ given $\underline{x}$, is the same as minimizing

$$\int (\theta - d(\underline{x}))^2 h(\theta \mid \underline{x})\ d\theta.$$

However, this is minimized when $d(\underline{x}) = E(\theta \mid \underline{X} = \underline{x})$ as shown in Stat 6710, Homework 4 (ii), for the unconditional case. ■

Note:
Under the conditions of Theorem 8.8.10, $d(\underline{x}) = E(\theta \mid \underline{X} = \underline{x})$ is called the **Bayes estimate**.

■

Example 8.8.11:
Let $X \sim Bin(n, p)$. Let $L(p, d(x)) = (p - d(x))^2$.

Let $\pi(p) = 1\ \ \forall p \in (0, 1)$, i.e., $\pi \sim U(0, 1)$, be the a priori distribution of p.

Then it holds:

$$\begin{aligned}
f(x,p) &= \binom{n}{x} p^x (1-p)^{n-x} \\
g(x) &= \int f(x,p)dp \\
&= \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} dp \\
h(p \mid x) &= \frac{f(x,p)}{g(x)} \\
&= \frac{\binom{n}{x} p^x (1-p)^{n-x}}{\int_0^1 \binom{n}{x} p^x (1-p)^{n-x} dp} \\
&= \frac{p^x (1-p)^{n-x}}{\int_0^1 p^x (1-p)^{n-x} dp} \\
E(p \mid x) &= \int_0^1 p h(p \mid x) dp \\
&= \frac{\int_0^1 p \binom{n}{x} p^x (1-p)^{n-x} dp}{\int_0^1 \binom{n}{x} p^x (1-p)^{n-x} dp} \\
&= \frac{\int_0^1 p^{x+1} (1-p)^{n-x} dp}{\int_0^1 p^x (1-p)^{n-x} dp} \\
&= \frac{B(x+2, n-x+1)}{B(x+1, n-x+1)} \\
&= \frac{\frac{\Gamma(x+2)\Gamma(n-x+1)}{\Gamma(x+2+n-x+1)}}{\frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(x+1+n-x+1)}} \\
&= \frac{x+1}{n+2}
\end{aligned}$$

Thus, the Bayes rule is

$$\hat{p}_{Bayes} = d^*(X) = \frac{X+1}{n+2}.$$

The Bayes risk of $d^*(X)$ is

$$
\begin{aligned}
R(\pi, d^*(X)) &= E_\pi(R(p, d^*(X))) \\
&= \int_0^1 \pi(p) R(p, d^*(x)) dp \\
&= \int_0^1 \pi(p) E_p(L(p, d^*(x))) dp \\
&= \int_0^1 \pi(p) E_p((p - d^*(x))^2) dp \\
&= \int_0^1 E_p\left((\frac{X+1}{n+2} - p)^2 \right) dp \\
&= \int_0^1 E_p\left((\frac{X+1}{n+2})^2 - 2p\frac{X+1}{n+2} + p^2 \right) dp \\
&= \frac{1}{(n+2)^2} \int_0^1 E_p\left((X+1)^2 - 2p(n+2)(X+1) + p^2(n+2)^2 \right) dp \\
&= \frac{1}{(n+2)^2} \int_0^1 E_p\left(X^2 + 2X + 1 - 2p(n+2)(X+1) + p^2(n+2)^2 \right) \ dp \\
&= \frac{1}{(n+2)^2} \int_0^1 (np(1-p) + (np)^2 + 2np + 1 - 2p(n+2)(np+1) + p^2(n+2)^2) \ dp \\
&= \frac{1}{(n+2)^2} \int_0^1 (np - np^2 + n^2p^2 + 2np + 1 - 2n^2p^2 - 2np - 4np^2 - 4p + \\
&\qquad p^2n^2 + 4np^2 + 4p^2) \ dp \\
&= \frac{1}{(n+2)^2} \int_0^1 (1 - 4p + np - np^2 + 4p^2) \ dp \\
&= \frac{1}{(n+2)^2} \int_0^1 (1 + (n-4)p + (4-n)p^2) \ dp \\
&= \frac{1}{(n+2)^2} \ (p + \frac{n-4}{2}p^2 + \frac{4-n}{3}p^3)\Big|_0^1 \\
&= \frac{1}{(n+2)^2}(1 + \frac{n-4}{2} + \frac{4-n}{3}) \\
&= \frac{1}{(n+2)^2} \frac{6 + 3n - 12 + 8 - 2n}{6} \\
&= \frac{1}{(n+2)^2} \frac{n+2}{6} \\
&= \frac{1}{6(n+2)}
\end{aligned}
$$

Now we compare the Bayes rule $d^*(X)$ with the MLE $\hat{p}_{ML} = \frac{X}{n}$. This estimate has Bayes risk

$$\begin{aligned}
R(\pi, \frac{X}{n}) &= \int_0^1 E_p((\frac{X}{n} - p)^2) dp \\
&= \int_0^1 \frac{1}{n^2} E_p((X - np)^2) dp \\
&= \int_0^1 \frac{np(1-p)}{n^2} dp \\
&= \int_0^1 \frac{p(1-p)}{n} dp \\
&= \frac{1}{n} \left(\frac{p^2}{2} - \frac{p^3}{3} \right) \Bigg|_0^1 \\
&= \frac{1}{6n}
\end{aligned}$$

which is, as expected, larger than the Bayes risk of $d^*(X)$. ■

Theorem 8.8.12:

Let $\{f_\theta \ : \ \theta \in \Theta\}$ be a family of pdf's (or pmf's). Suppose that an estimate d^* of θ is a Bayes estimate corresponding to some prior distribution π on Θ. If the risk function $R(\theta, d^*)$ is constant on Θ, then d^* is a minimax estimate of θ.

Proof:

Homework. ■

9 Hypothesis Testing

9.1 Fundamental Notions

We assume that $\underline{X} = (X_1, \ldots, X_n)$ is a random sample from a population distribution F_θ, $\theta \in \Theta \subseteq I\!R^k$, where the functional form of F_θ is known, except for the parameter θ. We also assume that Θ contains at least two points.

Definition 9.1.1:
A **parametric hypothesis** is an assumption about the unknown parameter θ.

The **null hypothesis** is of the form

$$H_0: \ \theta \in \Theta_0 \subset \Theta.$$

The **alternative hypothesis** is of the form

$$H_1: \ \theta \in \Theta_1 = \Theta - \Theta_0.$$

■

Definition 9.1.2:
If Θ_0 (or Θ_1) contains only one point, we say that H_0 and Θ_0 (or H_1 and Θ_1) are **simple**. In this case, the distribution of $\underline{X}$ is completely specified under the null (or alternative) hypothesis.

If Θ_0 (or Θ_1) contains more than one point, we say that H_0 and Θ_0 (or H_1 and Θ_1) are **composite**. ■

Example 9.1.3:
Let $X_1, \ldots, X_n$ be iid $Bin(1, p)$. Examples for hypotheses are $p = \frac{1}{2}$ (simple), $p \geq \frac{1}{2}$ (composite), $p \neq \frac{1}{4}$ (composite), etc. ■

Note:
The problem of testing a hypothesis can be described as follows: Given a sample point $\underline{x}$, find a decision rule that will lead to a decision to accept or reject the null hypothesis. This means, we partition the space $I\!R^n$ into two disjoint sets C and C^c such that, if $\underline{x} \in C$, we reject $H_0: \ \theta \in \Theta_0$ (and we accept H_1). Otherwise, if $\underline{x} \in C^c$, we accept H_0 that $\underline{X} \sim F_\theta, \ \ \theta \in \Theta_0$.

■

Definition 9.1.4:
Let $\underline{X} \sim F_\theta, \quad \theta \in \Theta$. Let C be a subset of $I\!R^n$ such that, if $\underline{x} \in C$, then H_0 is rejected (with probability 1), i.e.,

$$C = \{\underline{x} \in I\!R^n \ : \ H_0 \text{ is rejected for this } \underline{x}\}.$$

The set C is called the **critical region**. ■

Definition 9.1.5:
If we reject H_0 when it is true, we call this a **Type I error**. If we fail to reject H_0 when it is false, we call this a **Type II error**. Usually, H_0 and H_1 are chosen such that the Type I error is considered more serious. ■

Example 9.1.6:
We first consider a non–statistical example, in this case a jury trial. Our hypotheses are that the defendant is innocent or guilty. Our possible decisions are guilty or not guilty. Since it is considered worse to punish the innocent than to let the guilty go free, we make innocence the null hypothesis. Thus, we have

Truth (unknown) / Decision (known)	Innocent (H_0)	Guilty (H_1)
Not Guilty (H_0)	Correct	Type II Error
Guilty (H_1)	Type I Error	Correct

The jury tries to make a decision "beyond a reasonable doubt", i.e., it tries to make the probability of a Type I error small. ■

Definition 9.1.7:
If C is the critical region, then $P_\theta(C), \ \theta \in \Theta_0$, is a probability of Type I error, and $P_\theta(C^c), \ \theta \in \Theta_1$, is a probability of Type II error. ■

Note:
We would like both error probabilities to be 0, but this is usually not possible. We usually settle for fixing the probability of Type I error to be small, e.g., 0.05 or 0.01, and minimizing the Type II error. ■

Lecture 18:
We 02/21/01

Definition 9.1.8:
Every Borel–measurable mapping ϕ of $\mathbb{R}^n \to [0,1]$ is called a **test function**. $\phi(\underline{x})$ is the probability of rejecting H_0 when $\underline{x}$ is observed.

If ϕ is the indicator function of a subset $C \subseteq \mathbb{R}^n$, ϕ is called a **nonrandomized test** and C is the critical region of this test function.

Otherwise, if ϕ is not an indicator function of a subset $C \subseteq \mathbb{R}^n$, ϕ is called a **randomized test**. ■

Definition 9.1.9:
Let ϕ be a test function of the hypothesis $H_0: \ \theta \in \Theta_0$ against the alternative $H_1: \ \theta \in \Theta_1$. We say that ϕ has a **level of significance** of α (or ϕ is a **level–α–test** or ϕ is of **size** α) if

$$E_\theta(\phi(\underline{X})) = P_\theta(\text{rejct } H_0) \leq \alpha \quad \forall \theta \in \Theta_0.$$

In short, we say that ϕ is a test for the problem $(\alpha, \Theta_0, \Theta_1)$. ■

Definition 9.1.10:
Let ϕ be a test for the problem $(\alpha, \Theta_0, \Theta_1)$. For every $\theta \in \Theta$, we define

$$\beta_\phi(\theta) = E_\theta(\phi(\underline{X})) = P_\theta(\text{rejct } H_0).$$

We call $\beta_\phi(\theta)$ the **power function** of ϕ. For any $\theta \in \Theta_1$, $\beta_\phi(\theta)$ is called the **power** of ϕ against the alternative θ. ■

Definition 9.1.11:
Let Φ_α be the class of all tests for $(\alpha, \Theta_0, \Theta_1)$. A test $\phi_0 \in \Phi_\alpha$ is called a **most powerful (MP) test** against an alternative $\theta \in \Theta_1$ if

$$\beta_{\phi_0}(\theta) \geq \beta_\phi(\theta) \quad \forall \phi \in \Phi_\alpha.$$

■

Definition 9.1.12:
Let Φ_α be the class of all tests for $(\alpha, \Theta_0, \Theta_1)$. A test $\phi_0 \in \Phi_\alpha$ is called a **uniformly most powerful (UMP) test** if

$$\beta_{\phi_0}(\theta) \geq \beta_\phi(\theta) \quad \forall \phi \in \Phi_\alpha \quad \forall \theta \in \Theta_1.$$

■

Example 9.1.13:

Let $X_1, \ldots, X_n$ be iid $N(\mu, 1)$, $\mu \in \Theta = \{\mu_0, \mu_1\}$, $\mu_0 < \mu_1$.

Let $H_0: X_i \sim N(\mu_0, 1)$ vs. $H_1: X_i \sim N(\mu_1, 1)$.

Intuitively, reject H_0 when $\overline{X}$ is too large, i.e., if $\overline{X} \geq k$ for some k.

Under H_0 it holds that $\overline{X} \sim N(\mu_0, \frac{1}{n})$.

For a given α, we can solve the following equation for k:

$$P_{\mu_0}(\overline{X} > k) = P\left(\frac{\overline{X} - \mu_0}{1/\sqrt{n}} > \frac{k - \mu_0}{1/\sqrt{n}}\right) = P(Z > z_\alpha) = \alpha$$

Here, $\frac{\overline{X} - \mu_0}{1/\sqrt{n}} = Z \sim N(0,1)$ and z_α is defined such in such a way that $P(Z > z_\alpha) = \alpha$, i.e., z_α is the upper α–quantile of the $N(0,1)$ distribution. It follows that $\frac{k - \mu_0}{1/\sqrt{n}} = z_\alpha$ and therefore, $k = \mu_0 + \frac{z_\alpha}{\sqrt{n}}$.

Thus, we obtain the nonrandomized test

$$\phi(\underline{x}) = \begin{cases} 1, & \text{if } \overline{x} > \mu_0 + \frac{z_\alpha}{\sqrt{n}} \\ 0, & \text{otherwise} \end{cases}$$

ϕ has power

$$\begin{aligned}
\beta_\phi(\mu_1) &= P_{\mu_1}\left(\overline{X} > \mu_0 + \frac{z_\alpha}{\sqrt{n}}\right) \\
&= P\left(\frac{\overline{X} - \mu_1}{1/\sqrt{n}} > (\mu_0 - \mu_1)\sqrt{n} + z_\alpha\right) \\
&= P(Z > z_\alpha - \sqrt{n}\underbrace{(\mu_1 - \mu_0)}_{>0}) \\
&> \alpha
\end{aligned}$$

The probability of a Type II error is

$$P(\text{Type II error}) = 1 - \beta_\phi(\mu_1).$$

$\blacksquare$

Example 9.1.14:

Let $X \sim Bin(6, p), \quad p \in \Theta = (0, 1)$.

$H_0: \ p = \frac{1}{2}, \quad H_1 : p \neq \frac{1}{2}$.

Desired level of significance: $\alpha = 0.05$.

Reasonable plan: Since $E_{p=\frac{1}{2}}(X) = 3$, reject H_0 when $\mid X - 3 \mid \geq c$ for some constant c. But how should we select c?

x	$c = \mid x - 3 \mid$	$P_{p=\frac{1}{2}}(X = x)$	$P_{p=\frac{1}{2}}(\mid X - 3 \mid \geq c)$
$0, 6$	3	0.015625	0.03125
$1, 5$	2	0.093750	0.21875
$2, 4$	1	0.234375	0.68750
3	0	0.312500	1.00000

Thus, there is no nonrandomized test with $\alpha = 0.05$.

What can we do instead? — Three possibilities:

(i) Reject if $\mid X - 3 \mid = 3$, i.e., use a nonrandomized test of size $\alpha = 0.03125$.

(ii) Reject if $\mid X - 3 \mid \geq 2$, i.e., use a nonrandomized test of size $\alpha = 0.21875$.

(iii) Reject if $\mid X - 3 \mid = 3$, do not reject if $\mid X - 3 \mid \leq 1$, and reject with probability $\dfrac{0.05 - 0.03125}{2 \cdot 0.093750} = 0.1$ if $\mid X - 3 \mid = 2$. Thus, we obtain the randomized test

$$\phi(\underline{x}) = \begin{cases} 1, & \text{if } x = 0, 6 \\ 0.1, & \text{if } x = 1, 5 \\ 0, & \text{if } x = 2, 3, 4 \end{cases}$$

This test has size

$$\begin{aligned} \alpha &= E_{p=\frac{1}{2}}(\phi(X)) \\ &= 1 \cdot 0.015625 \cdot 2 + 0.1 \cdot 0.093750 \cdot 2 + 0 \cdot (0.24375 \cdot 2 + 0.3125) \\ &= 0.05 \end{aligned}$$

as intended. The power of ϕ can be calculated for any $p \neq \frac{1}{2}$ and it is

$$\beta_\phi(p) = P_p(X = 0 \text{ or } X = 6) + 0.1 \cdot P_p(X = 1 \text{ or } X = 5)$$

■

Lecture 19:
Fr 02/23/01

9.2 The Neyman–Pearson Lemma

Let $\{f_\theta \; : \; \theta \in \Theta = \{\theta_0, \theta_1\}\}$ be a family of possible distributions of $\underline{X}$. f_θ represents the pdf (or pmf) of $\underline{X}$. For convenience, we write $f_0(\underline{x}) = f_{\theta_0}(\underline{x})$ and $f_1(\underline{x}) = f_{\theta_1}(\underline{x})$.

Theorem 9.2.1: **Neyman–Pearson Lemma (NP Lemma)**
Suppose we wish to test $H_0: \; \underline{X} \sim f_0(\underline{x})$ vs. $H_1: \; \underline{X} \sim f_1(\underline{x})$, where f_i is the pdf (or pmf) of $\underline{X}$ under $H_i, \; i = 0, 1$, where both, H_0 and H_1, are simple.

(i) Any test of the form

$$\phi(\underline{x}) = \begin{cases} 1, & \text{if } f_1(\underline{x}) > k f_0(\underline{x}) \\ \gamma(\underline{x}), & \text{if } f_1(\underline{x}) = k f_0(\underline{x}) \\ 0, & \text{if } f_1(\underline{x}) < k f_0(\underline{x}) \end{cases} \qquad (*)$$

for some $k \geq 0$ and $0 \leq \gamma(\underline{x}) \leq 1$, is most powerful of its significance level for testing H_0 vs. H_1.

If $k = \infty$, the test

$$\phi(\underline{x}) = \begin{cases} 1, & \text{if } f_0(\underline{x}) = 0 \\ 0, & \text{if } f_0(\underline{x}) > 0 \end{cases} \qquad (**)$$

is most powerful of size (or significance level) 0 for testing H_0 vs. H_1.

(ii) Given $0 \leq \alpha \leq 1$, there exists a test of the form $(*)$ or $(**)$ with $\gamma(\underline{x}) = \gamma$ (i.e., a constant) such that

$$E_{\theta_0}(\phi(\underline{X})) = \alpha.$$

Proof:
We prove the continuous case only.

(i):
Let ϕ be a test satisfying $(*)$. Let ϕ^* be any test with $E_{\theta_0}(\phi^*(\underline{X})) \leq E_{\theta_0}(\phi(\underline{X}))$.

It holds that

$$\begin{aligned}\int (\phi(\underline{x}) - \phi^*(\underline{x}))(f_1(\underline{x}) - k f_0(\underline{x})) d\underline{x} \;\; &= \;\; \left(\int_{f_1 > k f_0} (\phi(\underline{x}) - \phi^*(\underline{x}))(f_1(\underline{x}) - k f_0(\underline{x})) d\underline{x} \right) \\ &\quad + \left(\int_{f_1 < k f_0} (\phi(\underline{x}) - \phi^*(\underline{x}))(f_1(\underline{x}) - k f_0(\underline{x})) d\underline{x} \right)\end{aligned}$$

since

$$\int_{f_1 = k f_0} (\phi(\underline{x}) - \phi^*(\underline{x}))(f_1(\underline{x}) - k f_0(\underline{x})) d\underline{x} = 0.$$

It is

$$\int_{f_1 > kf_0} (\phi(\underline{x}) - \phi^*(\underline{x}))(f_1(\underline{x}) - kf_0(\underline{x}))d\underline{x} = \int_{f_1 > kf_0} \underbrace{(1 - \phi^*(\underline{x}))}_{\geq 0} \underbrace{(f_1(\underline{x}) - kf_0(\underline{x}))}_{\geq 0} d\underline{x} \geq 0$$

and

$$\int_{f_1 < kf_0} (\phi(\underline{x}) - \phi^*(\underline{x}))(f_1(\underline{x}) - kf_0(\underline{x}))d\underline{x} = \int_{f_1 < kf_0} \underbrace{(0 - \phi^*(\underline{x}))}_{\leq 0} \underbrace{(f_1(\underline{x}) - kf_0(\underline{x}))}_{\leq 0} d\underline{x} \geq 0.$$

Therefore,

$$\begin{aligned} 0 &\leq \int (\phi(\underline{x}) - \phi^*(\underline{x}))(f_1(\underline{x}) - kf_0(\underline{x}))d\underline{x} \\ &= E_{\theta_1}(\phi(\underline{X})) - E_{\theta_1}(\phi^*(\underline{X})) - k(E_{\theta_0}(\phi(\underline{X})) - E_{\theta_0}(\phi^*(\underline{X}))). \end{aligned}$$

Since $E_{\theta_0}(\phi(\underline{X})) \geq E_{\theta_0}(\phi^*(\underline{X}))$, it holds that

$$\beta_\phi(\theta_1) - \beta_{\phi^*}(\theta_1) \geq k(E_{\theta_0}(\phi(\underline{X})) - E_{\theta_0}(\phi^*(\underline{X}))) \geq 0,$$

i.e., ϕ is most powerful.

If $k = \infty$, any test ϕ^* of size 0 must be 0 on the set $\{\underline{x} \mid f_0(\underline{x}) > 0\}$. Therefore,

$$\beta_\phi(\theta_1) - \beta_{\phi^*}(\theta_1) = E_{\theta_1}(\phi(\underline{X})) - E_{\theta_1}(\phi^*(\underline{X})) = \int_{\{\underline{x} \mid f_0(\underline{x}) = 0\}} (1 - \phi^*(\underline{x}))f_1(\underline{x})d\underline{x} \geq 0,$$

i.e., ϕ is most powerful of size 0.

<u>(ii):</u>

If $\alpha = 0$, then use $(**)$. Otherwise, assume that $0 < \alpha \leq 1$ and $\gamma(\underline{x}) = \gamma$. It is

$$\begin{aligned} E_{\theta_0}(\phi(\underline{X})) &= P_{\theta_0}(f_1(\underline{X}) > kf_0(\underline{X})) + \gamma P_{\theta_0}(f_1(\underline{X}) = kf_0(\underline{X})) \\ &= 1 - P_{\theta_0}(f_1(\underline{X}) \leq kf_0(\underline{X})) + \gamma P_{\theta_0}(f_1(\underline{X}) = kf_0(\underline{X})) \\ &= 1 - P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} \leq k\right) + \gamma P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} = k\right). \end{aligned}$$

Note that the last step is valid since $P_{\theta_0}(f_0(\underline{X}) = 0) = 0$.

Therefore, given $0 < \alpha \leq 1$, we want to find k and γ such that $E_{\theta_0}(\phi(\underline{X})) = \alpha$, i.e.,

$$P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} \leq k\right) - \gamma P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} = k\right) = 1 - \alpha.$$

Note that $\frac{f_1(\underline{X})}{f_0(\underline{X})}$ is a rv and, therefore, $P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} \leq k\right)$ is a cdf.

If there exists a k_0 such that

$$P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} \leq k_0\right) = 1 - \alpha,$$

we choose $\gamma = 0$ and $k = k_0$.

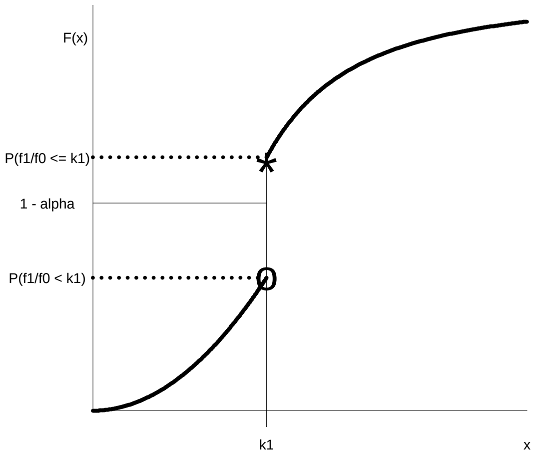
Otherwise, if there exists no such k_0, then there exists a k_1 such that

$$P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} < k_1\right) \leq 1 - \alpha < P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} \leq k_1\right),$$

i.e., the cdf has a jump at k_1. In this case, we choose $k = k_1$ and

$$\gamma = \frac{P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} \leq k_1\right) - (1-\alpha)}{P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} = k_1\right)}.$$

Theorem 9.2.1



Let us verify that these values for k_1 and γ meet the necessary conditions:

Obviously,

$$P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} \leq k_1\right) - \frac{P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} \leq k_1\right) - (1-\alpha)}{P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} = k_1\right)} P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} = k_1\right) = 1 - \alpha.$$

Also, since

$$P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} \leq k_1\right) > (1-\alpha),$$

it follows that $\gamma \geq 0$ and since

$$\begin{aligned}
\frac{P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} \leq k_1\right) - (1-\alpha)}{P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} = k_1\right)} &\leq \frac{P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} \leq k_1\right) - P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} < k_1\right)}{P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} = k_1\right)} \\
&= \frac{P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} = k_1\right)}{P_{\theta_0}\left(\frac{f_1(\underline{X})}{f_0(\underline{X})} = k_1\right)} \\
&= 1,
\end{aligned}$$

it follows that $\gamma \leq 1$. Overall, $0 \leq \gamma \leq 1$ as required. ■

Theorem 9.2.2:
If a sufficient statistic T exists for the family $\{f_\theta : \theta \in \Theta = \{\theta_0, \theta_1\}\}$, then the Neyman–Pearson most powerful test is a function of T.

Proof:
Homework ■

Example 9.2.3:
We want to test $H_0 : X \sim N(0,1)$ vs. $H_1 : X \sim \text{Cauchy}(1,0)$, based on a single observation. It is

$$\frac{f_1(x)}{f_0(x)} = \frac{\frac{1}{\pi}\frac{1}{1+x^2}}{\frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})} = \sqrt{\frac{2}{\pi}}\frac{\exp(\frac{x^2}{2})}{1+x^2}.$$

The MP test is

$$\phi(x) = \begin{cases} 1, & \text{if } \sqrt{\frac{2}{\pi}}\frac{\exp(\frac{x^2}{2})}{1+x^2} > k \\ 0, & \text{otherwise} \end{cases}$$

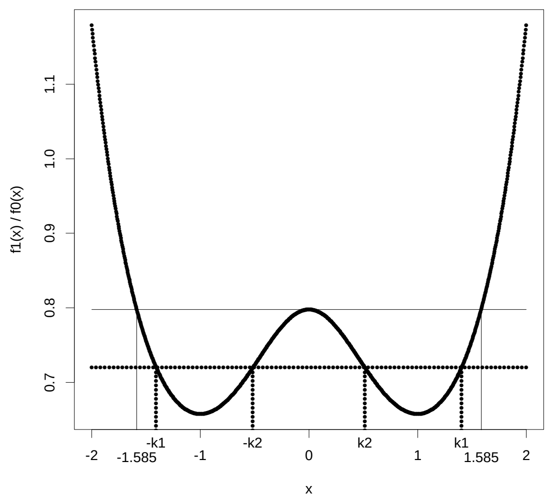
where k is determined such that $E_{H_0}(\phi(X)) = \alpha$.

If $\alpha < 0.113$, we reject H_0 if $| x |> z_{\frac{\alpha}{2}}$, where $z_{\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ quantile of a $N(0,1)$ distribution.

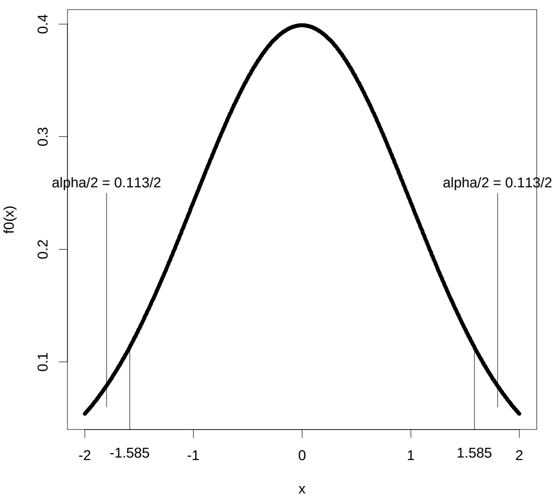
If $\alpha > 0.113$, we reject H_0 if $| x |> k_1$ or if $| x |< k_2$, where $k_1 > 0$, $k_2 > 0$, such that

$$\frac{\exp(\frac{k_1^2}{2})}{1+k_1^2} = \frac{\exp(\frac{k_2^2}{2})}{1+k_2^2} \quad \text{and} \quad \int_{k_2}^{k_1} \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})dx = \frac{1-\alpha}{2}.$$

Example 9.2.3a



Example 9.2.3b



Why is $\alpha = 0.113$ so interesting?

For $x = 0$, it is

$$\frac{f_1(x)}{f_0(x)} = \sqrt{\frac{2}{\pi}} \approx 0.7979.$$

Similarly, for $x \approx -1.585$ and $x \approx 1.585$, it is

$$\frac{f_1(x)}{f_0(x)} = \sqrt{\frac{2}{\pi}} \frac{\exp(\frac{(\pm 1.585)^2}{2})}{1 + (\pm 1.585)^2} \approx 0.7979 \approx \frac{f_1(0)}{f_0(0)}.$$

More importantly, $P_{H_0}(\mid X \mid > 1.585) = 0.113$.

■

9.3 Monotone Likelihood Ratios

Suppose we want to test $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$ for a family of pdf's $\{f_\theta : \theta \in \Theta \subseteq I\!R\}$. In general, it is not possible to find a UMP test. However, there exist conditions under which UMP tests exist.

<u>Definition 9.3.1:</u>
Let $\{f_\theta : \theta \in \Theta \subseteq I\!R\}$ be a family of pdf's (pmf's) on a one–dimensional parameter space. We say the family $\{f_\theta\}$ has a **monotone likelihood ratio (MLR)** in statistic $T(\underline{X})$ if for $\theta_1 < \theta_2$, whenever f_{θ_1} and f_{θ_2} are distinct, the ratio $\frac{f_{\theta_2}(\underline{x})}{f_{\theta_1}(\underline{x})}$ is a nondecreasing function of $T(\underline{x})$ for the set of values $\underline{x}$ for which at least one of f_{θ_1} and f_{θ_2} is > 0. ■

<u>Note:</u>
We can also define families of densities with nonincreasing MLR in $T(\underline{X})$, but such families can be treated by symmetry. ■

<u>Example 9.3.2:</u>
Let $X_1, \ldots, X_n \sim U[0, \theta], \ \theta > 0$. Then the joint pdf is

$$f_\theta(\underline{x}) = \begin{cases} \frac{1}{\theta^n}, & 0 \leq x_{max} \leq \theta \\ 0, & \text{otherwise} \end{cases} = I_{[0,\theta]}(x_{max}),$$

where $x_{max} = \max\limits_{i=1,\ldots,n} x_i$.

Let $\theta_2 > \theta_1$, then

$$\frac{f_{\theta_2}(\underline{x})}{f_{\theta_1}(\underline{x})} = (\frac{\theta_1}{\theta_2})^n \frac{I_{[0,\theta_2]}(x_{max})}{I_{[0,\theta_1]}(x_{max})}.$$

It is

$$\frac{I_{[0,\theta_2]}(x_{max})}{I_{[0,\theta_1]}(x_{max})} = \begin{cases} 1, & x_{max} \in [0, \theta_1] \\ \infty, & x_{max} \in (\theta_1, \theta_2] \end{cases}$$

since for $x_{max} \in [0, \theta_1]$, it holds that $x_{max} \in [0, \theta_2]$.

But for $x_{max} \in (\theta_1, \theta_2]$, it is $I_{[0,\theta_1]}(x_{max}) = 0$.

$\Longrightarrow$ as $T(\underline{X}) = X_{max}$ increases, the density ratio goes from $(\frac{\theta_1}{\theta_2})^n$ to ∞.

$\Longrightarrow$ $\frac{f_{\theta_2}}{f_{\theta_1}}$ is a nondecreasing function of $T(\underline{X}) = X_{max}$

$\Longrightarrow$ the family of $U[0, \theta]$ distributions has a MLR in $T(\underline{X}) = X_{max}$ ■

Theorem 9.3.3:
The one–parameter exponential family $f_\theta(\underline{x}) = \exp(Q(\theta)T(\underline{x}) + D(\theta) + S(\underline{x}))$, where $Q(\theta)$ is nondecreasing, has a MLR in $T(\underline{X})$.

Proof:
Homework. ■

Example 9.3.4:
Let $\underline{X} = (X_1, \cdots, X_n)$ be a random sample from the Poisson family with parameter $\lambda > 0$. Then the joint pdf is

$$f_\lambda(\underline{x}) = \prod_{i=1}^n \left(e^{-\lambda} \lambda^{x_i} \frac{1}{x_i!} \right) = e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i} \prod_{i=1}^n \frac{1}{x_i!} = \exp\left(-n\lambda + \sum_{i=1}^n x_i \cdot log(\lambda) - \sum_{i=1}^n log(x_i!) \right),$$

which belongs to the one–parameter exponential family.

Since $Q(\lambda) = log(\lambda)$ is a nondecreasing function of λ, it follows by Theorem 9.3.3 that the Poisson family with parameter $\lambda > 0$ has a MLR in $T(\underline{X}) = \sum_{i=1}^n X_i$.

We can verify this result by Definition 9.3.1:

$$\frac{f_{\lambda_2}(\underline{x})}{f_{\lambda_1}(\underline{x})} = \frac{\lambda_2^{\sum x_i} e^{-n\lambda_2}}{\lambda_1^{\sum x_i} e^{-n\lambda_1}} = \left(\frac{\lambda_2}{\lambda_1} \right)^{\sum x_i} e^{-n(\lambda_2 - \lambda_1)}.$$

If $\lambda_2 > \lambda_1$, then $\frac{\lambda_2}{\lambda_1} > 1$ and $\left(\frac{\lambda_2}{\lambda_1}\right)^{\sum x_i}$ is a nondecreasing function of $\sum x_i$.

Therefore, f_λ has a MLR in $T(\underline{X}) = \sum_{i=1}^n X_i$. ■

Theorem 9.3.5:
Let $X \sim f_\theta$, $\theta \in \Theta \subseteq I\!R$, where the family $\{f_\theta\}$ has a MLR in $T(\underline{X})$.

For testing $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$, $\theta_0 \in \Theta$, any test of the form

$$\phi(\underline{x}) = \begin{cases} 1, & \text{if } T(\underline{x}) > t_0 \\ \gamma, & \text{if } T(\underline{x}) = t_0 \\ 0, & \text{if } T(\underline{x}) < t_0 \end{cases} \qquad (*)$$

has a nondecreasing power function and is UMP of its size $E_{\theta_0}(\phi(\underline{X})) = \alpha$, if the size is not 0.

Also, for every $0 \leq \alpha \leq 1$ and every $\theta_0 \in \Theta$, there exists a t_0 and a γ ($-\infty \leq t_0 \leq \infty$, $0 \leq \gamma \leq 1$), such that the test of form $(*)$ is the UMP size α test of H_0 vs. H_1.

Proof:

"$\Longrightarrow$":

Let $\theta_1 < \theta_2, \;\; \theta_1, \theta_2 \in \Theta$ and suppose $E_{\theta_1}(\phi(\underline{X})) > 0$, i.e., the size is > 0. Since f_θ has a MLR in T, $\frac{f_{\theta_2}(\underline{x})}{f_{\theta_1}(\underline{x})}$ is a nondecreasing function of T. Therefore, any test of form $(*)$ is equivalent to a test

$$\phi(\underline{x}) = \begin{cases} 1, & \text{if } \frac{f_{\theta_2}(\underline{x})}{f_{\theta_1}(\underline{x})} > k \\ \gamma, & \text{if } \frac{f_{\theta_2}(\underline{x})}{f_{\theta_1}(\underline{x})} = k \\ 0, & \text{if } \frac{f_{\theta_2}(\underline{x})}{f_{\theta_1}(\underline{x})} < k \end{cases} \qquad (**)$$

which by the Neyman–Pearson Lemma (Theorem 9.2.1) is MP of size α for testing $H_0 : \theta = \theta_1$ vs. $H_1 : \theta = \theta_2$.

Lecture 22: Fr 03/02/01

Let Φ_α be the class of tests of size α, and let $\phi_t \in \Phi_\alpha$ be the trivial test $\phi_t(\underline{x}) = \alpha \;\; \forall \underline{x}$. Then ϕ_t has size and power α. The power of the MP test ϕ of form $(*)$ must be at least α as the MP test ϕ cannot have power less than the trivial test ϕ_t, i.e,

$$E_{\theta_2}(\phi(\underline{X})) \geq E_{\theta_2}(\phi_t(\underline{X})) = \alpha = E_{\theta_1}(\phi(\underline{X})).$$

Thus, for $\theta_2 > \theta_1$,

$$E_{\theta_2}(\phi(\underline{X})) \geq E_{\theta_1}(\phi(\underline{X})),$$

i.e., the power function of the test ϕ of form $(*)$ is a nondecreasing function of θ.

Now let $\theta_1 = \theta_0$ and $\theta_2 > \theta_0$. We know that a test ϕ of form $(*)$ is MP for $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_2 > \theta_0$, provided that its size $\alpha = E_{\theta_0}(\phi(\underline{X}))$ is > 0.

Notice that the test ϕ of form $(*)$ does not depend on θ_2. It only depends on t_0 and γ. Therefore, the test ϕ of form $(*)$ is MP for all $\theta_2 \in \Theta_1$. Thus, this test is UMP for simple $H_0 : \theta = \theta_0$ vs. composite $H_1 : \theta > \theta_0$ with size $E_{\theta_0}(\phi(\underline{X})) = \alpha_0$.

Since ϕ is UMP for a class Φ'' of tests ($\phi'' \in \Phi''$) satisfying

$$E_{\theta_0}(\phi''(\underline{X})) \leq \alpha_0,$$

ϕ must also be UMP for the more restrictive class Φ''' of tests ($\phi''' \in \Phi'''$) satisfying

$$E_\theta(\phi'''(\underline{X})) \leq \alpha_0 \;\;\; \forall \theta \leq \theta_0.$$

But since the power function of ϕ is nondecreasing, it holds for ϕ that

$$E_\theta(\phi(\underline{X})) \leq E_{\theta_0}(\phi(\underline{X})) = \alpha_0 \;\;\; \forall \theta \leq \theta_0.$$

Thus, ϕ is the UMP size α_0 test of $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$ if $\alpha_0 > 0$.

"$\Longleftarrow$":
Use the Neyman–Pearson Lemma (Theorem 9.2.1). ■

Note:
By interchanging inequalities throughout Theorem 9.3.5 and its proof, we see that this Theorem also provides a solution of the dual problem $H_0' : \theta \geq \theta_0$ vs. $H_1' : \theta < \theta_0$. ■

Theorem: 9.3.6
For the one–parameter exponential family, there exists a UMP two–sided test of $H_0 : \theta \leq \theta_1$ or $\theta \geq \theta_2$, (where $\theta_1 < \theta_2$) vs. $H_1 : \theta_1 < \theta < \theta_2$ of the form

$$\phi(\underline{x}) = \begin{cases} 1, & \text{if } c_1 < T(\underline{x}) < c_2 \\ \gamma_i, & \text{if } T(\underline{x}) = c_i, \ i = 1, 2 \\ 0, & \text{if } T(\underline{x}) < c_1, \text{ or if } T(\underline{x}) > c_2 \end{cases}$$

■

Note:
UMP tests for $H_0 : \theta_1 \leq \theta \leq \theta_2$ and $H_0' : \theta = \theta_0$ do not exist for one–parameter exponential families. ■

9.4 Unbiased and Invariant Tests

If we look at all size α tests in the class Φ_α, there exists no UMP test for many hypotheses. Can we find UMP tests if we reduce Φ_α by reasonable restrictions?

Definition 9.4.1:

A size α test ϕ of $H_0: \theta \in \Theta_0$ vs $H_1: \theta \in \Theta_1$ is **unbiased** if

$$E_\theta(\phi(\underline{X})) \geq \alpha \quad \forall \theta \in \Theta_1.$$

■

Note:

This condition means that $\beta_\phi(\theta) \leq \alpha \quad \forall \theta \in \Theta_0$ and $\beta_\phi(\theta) \geq \alpha \quad \forall \theta \in \Theta_1$. In other words, the power of this test is never less than α. ■

Definition 9.4.2:

Let U_α be the class of all unbiased size α tests of H_0 vs H_1. If there exists a test $\phi \in U_\alpha$ that has maximal power for all $\theta \in \Theta_1$, we call ϕ a **UMP unbiased** (**UMPU**) size α test. ■

Note:

It holds that $U_\alpha \subseteq \Phi_\alpha$. A UMP test $\phi_\alpha \in \Phi_\alpha$ will have $\beta_{\phi_\alpha} \geq \alpha \quad \forall \theta \in \Theta_1$ since we must compare all tests ϕ_α with the trivial test $\phi(\underline{x}) = \alpha$. Thus, if a UMP test exists in Φ_α, it is also a UMPU test in U_α. ■

Example 9.4.3:

Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$, where $\sigma^2 > 0$ is known. Consider $H_0: \mu = \mu_0$ vs $H_1: \mu \neq \mu_0$. From the Neyman–Pearson Lemma, we know that for $\mu_1 > \mu_0$, the MP test is of the form

$$\phi_1(\underline{X}) = \begin{cases} 1, & \text{if } \overline{X} > \mu_0 + \frac{\sigma}{\sqrt{n}} z_\alpha \\ 0, & \text{otherwise} \end{cases}$$

and for $\mu_2 < \mu_0$, the MP test is of the form

$$\phi_2(\underline{X}) = \begin{cases} 1, & \text{if } \overline{X} < \mu_0 - \frac{\sigma}{\sqrt{n}} z_\alpha \\ 0, & \text{otherwise} \end{cases}$$

If a test is UMP, it must have the same rejection region as ϕ_1 and ϕ_2. However, these 2 rejection regions are different (actually, their intersection is empty). Thus, there exists no UMP test.

We next state a helpful Theorem and then continue with this example and see how we can find a UMPU test. ■

Theorem 9.4.4:
Let $c_1, \ldots, c_n \in I\!R$ be constants and $f_1(\underline{x}), \ldots, f_{n+1}(\underline{x})$ be real–valued functions. Let $\mathcal{C}$ be the class of functions $\phi(\underline{x})$ satisfying $0 \leq \phi(\underline{x}) \leq 1$ and

$$\int_{-\infty}^{\infty} \phi(\underline{x}) f_i(\underline{x}) d\underline{x} = c_i \quad \forall i = 1, \ldots, n.$$

If $\phi* \in \mathcal{C}$ satisfies

$$\phi^*(\underline{x}) = \begin{cases} 1, & \text{if } f_{n+1}(\underline{x}) > \sum\limits_{i=1}^{n} k_i f_i(\underline{x}) \\ 0, & \text{if } f_{n+1}(\underline{x}) < \sum\limits_{i=1}^{n} k_i f_i(\underline{x}) \end{cases}$$

for some constants $k_1, \ldots, k_n \in I\!R$, then ϕ^* maximizes $\displaystyle\int_{-\infty}^{\infty} \phi(\underline{x}) f_{n+1}(\underline{x}) d\underline{x}$ among all $\phi \in \mathcal{C}$.

Proof:
Let $\phi^*(\underline{x})$ be as above. Let $\phi(\underline{x})$ be any other function in $\mathcal{C}$. Since $0 \leq \phi(\underline{x}) \leq 1 \;\; \forall \underline{x}$, it is

$$(\phi^*(\underline{x}) - \phi(\underline{x})) \left(f_{n+1}(\underline{x}) - \sum_{i=1}^{n} k_i f_i(\underline{x}) \right) \geq 0 \;\; \forall \underline{x}.$$

This holds since if $\phi^*(\underline{x}) = 1$, the left factor is ≥ 0 and the right factor is ≥ 0. If $\phi^*(\underline{x}) = 0$, the left factor is ≤ 0 and the right factor is ≤ 0.

Therefore,

$$\begin{aligned} 0 \;\; &\leq \;\; \int (\phi^*(\underline{x}) - \phi(\underline{x})) \left(f_{n+1}(\underline{x}) - \sum_{i=1}^{n} k_i f_i(\underline{x}) \right) d\underline{x} \\ &= \;\; \int \phi^*(\underline{x}) f_{n+1}(\underline{x}) d\underline{x} - \int \phi(\underline{x}) f_{n+1}(\underline{x}) d\underline{x} - \sum_{i=1}^{n} k_i \underbrace{\left(\int \phi^*(\underline{x}) f_i(\underline{x}) d\underline{x} - \int \phi(\underline{x}) f_i(\underline{x}) d\underline{x} \right)}_{= c_i - c_i = 0} \end{aligned}$$

Thus,

$$\int \phi^*(\underline{x}) f_{n+1}(\underline{x}) d\underline{x} \geq \int \phi(\underline{x}) f_{n+1}(\underline{x}) d\underline{x}.$$

■

Note:

(i) If f_{n+1} is a pdf, then ϕ^* maximizes the power.

(ii) The Theorem above is the Neyman–Pearson Lemma if $n = 1, f_1 = f_{\theta_0}, f_2 = f_{\theta_1}$, and $c_1 = \alpha$.

■

Example 9.4.3: (continued)
So far, we have seen that there exists no UMP test for $H_0: \ \mu = \mu_0$ vs $H_1: \ \mu \neq \mu_0$.

We will show that

$$\phi_3(\underline{x}) = \begin{cases} 1, & \text{if } \overline{X} < \mu_0 - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \text{ or if } \overline{X} > \mu_0 + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \\ 0, & \text{otherwise} \end{cases}$$

is a UMPU size α test.

Due to Theorem 9.2.2, we only have to consider functions of sufficient statistics $T(\underline{X}) = \overline{X}$.

Let $\tau^2 = \frac{\sigma^2}{n}$.

To be unbiased and of size α, a test ϕ must have

(i) $\displaystyle\int \phi(t) f_{\mu_0}(t) dt = \alpha$, and

(ii) $\frac{\partial}{\partial \mu} \displaystyle\int \phi(t) f_\mu(t) dt\Big|_{\mu=\mu_0} = \int \phi(t) \left(\frac{\partial}{\partial \mu} f_\mu(t) \right) \Big|_{\mu=\mu_0} dt = 0$, i.e., we have a minimum at μ_0.

We want to maximize $\displaystyle\int \phi(t) f_\mu(t) dt, \ \ \mu \neq \mu_0$ such that conditions (i) and (ii) hold.

Lecture 23:
Mo 03/05/01

We choose an arbitrary $\mu_1 \neq \mu_0$ and let

$$\begin{aligned} f_1(t) &= f_{\mu_0}(t) \\ f_2(t) &= \frac{\partial}{\partial \mu} f_\mu(t) \Big|_{\mu=\mu_0} \\ f_3(t) &= f_{\mu_1}(t) \end{aligned}$$

We now consider how the conditions on ϕ^* in Theorem 9.4.4 can be met:

$$\begin{aligned} & f_3(t) > k_1 f_1(t) + k_2 f_2(t) \\ \iff & \frac{1}{\sqrt{2\pi}\tau} \exp(-\frac{1}{2\tau^2}(\overline{x} - \mu_1)^2) > \frac{k_1}{\sqrt{2\pi}\tau} \exp(-\frac{1}{2\tau^2}(\overline{x} - \mu_0)^2) + \\ & \qquad \frac{k_2}{\sqrt{2\pi}\tau} \exp(-\frac{1}{2\tau^2}(\overline{x} - \mu_0)^2)(\frac{\overline{x} - \mu_0}{\tau^2}) \\ \iff & \exp(-\frac{1}{2\tau^2}(\overline{x} - \mu_1)^2) > k_1 \exp(-\frac{1}{2\tau^2}(\overline{x} - \mu_0)^2) + k_2 \exp(-\frac{1}{2\tau^2}(\overline{x} - \mu_0)^2)(\frac{\overline{x} - \mu_0}{\tau^2}) \\ \iff & \exp(\frac{1}{2\tau^2}((\overline{x} - \mu_0)^2 - (\overline{x} - \mu_1)^2)) > k_1 + k_2(\frac{\overline{x} - \mu_0}{\tau^2}) \\ \iff & \exp(\frac{\overline{x}(\mu_1 - \mu_0)}{\tau^2} - \frac{\mu_1^2 - \mu_0^2}{2\tau^2}) > k_1 + k_2(\frac{\overline{x} - \mu_0}{\tau^2}) \end{aligned}$$

Note that the left hand side of this inequality is increasing in $\overline{x}$ if $\mu_1 > \mu_0$ and decreasing in $\overline{x}$ if $\mu_1 < \mu_0$. Either way, we can choose k_1 and k_2 such that the linear function in $\overline{x}$ crosses the exponential function in $\overline{x}$ at the two points

$$\mu_L = \mu_0 - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \quad \mu_U = \mu_0 + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}.$$

Obviously, ϕ_3 satisfies (i). We still need to check that ϕ_3 satisfies (ii) and that $\beta_{\phi_3}(\mu)$ has a minimum at μ_0 but omit this part from our proof here.

ϕ_3 is of the form ϕ^* in Theorem 9.4.4 and therefore ϕ_3 is UMP in $\mathcal{C}$. But the trivial test $\phi_t(\underline{x}) = \alpha$ also satisfies (i) and (ii) above. Therefore, $\beta_{\phi_3}(\mu) \geq \alpha \;\; \forall \mu \neq \mu_0$. This means that ϕ_3 is unbiased.

Overall, ϕ_3 is a UMPU test of size α. ■

<u>Definition 9.4.5:</u>
A test ϕ is said to be α–**similar** on a subset Θ^* of Θ if

$$\beta_\phi(\theta) = E_\theta(\phi(\underline{X})) = \alpha \;\; \forall \theta \in \Theta^*.$$

A test ϕ is said to be **similar** on $\Theta^* \subseteq \Theta$ if it is α–similar on Θ^* for some α, $\;0 \leq \alpha \leq 1$. ■

<u>Note:</u>
The trivial test $\phi(\underline{x}) = \alpha$ is α–similar on every $\Theta^* \subseteq \Theta$. ■

<u>Theorem 9.4.6:</u>
Let ϕ be an unbiased test of size α for $H_0 : \;\theta \in \Theta_0$ vs $H_1 : \;\theta \in \Theta_1$ such that $\beta_\phi(\theta)$ is a continuous function in θ. Then ϕ is α–similar on the boundary $\Lambda = \overline{\Theta_0} \cap \overline{\Theta_1}$, where $\overline{\Theta_0}$ and $\overline{\Theta_1}$ are the closures of Θ_0 and Θ_1, respectively.

<u>Proof:</u>
Let $\theta \in \Lambda$. There exist sequences $\{\theta_n\}$ and $\{\theta'_n\}$ whith $\theta_n \in \Theta_0$ and $\theta'_n \in \Theta_1$ such that $\lim\limits_{n\to\infty} \theta_n = \theta$ and $\lim\limits_{n\to\infty} \theta'_n = \theta$.
By continuity, $\beta_\phi(\theta_n) \to \beta_\phi(\theta)$ and $\beta_\phi(\theta'_n) \to \beta_\phi(\theta)$.

Since $\beta_\phi(\theta_n) \leq \alpha$ implies $\beta_\phi(\theta) \leq \alpha$ and since $\beta_\phi(\theta'_n) \geq \alpha$ implies $\beta_\phi(\theta) \geq \alpha$ it must hold that $\beta_\phi(\theta) = \alpha \;\; \forall \theta \in \Lambda$. ■

Lecture 24:
We 03/07/01

Definition 9.4.7:

A test ϕ that is UMP among all α–similar tests on the boundary $\Lambda = \overline{\Theta_0} \cap \overline{\Theta_1}$ is called a **UMP α–similar** test. ■

Theorem 9.4.8:

Suppose $\beta_\phi(\theta)$ is continuous in θ for all tests ϕ of $H_0: \theta \in \Theta_0$ vs $H_1: \theta \in \Theta_1$. If a size α test of H_0 vs H_1 is UMP α–similar, then it is UMP unbiased.

Proof:

Let ϕ_0 be UMP α–similar and of size α. This means that $E_\theta(\phi(\underline{X})) \leq \alpha \;\; \forall \theta \in \Theta_0$.

Since the trivial test $\phi(\underline{x}) = \alpha$ is α–similar, it must hold for ϕ_0 that $\beta_{\phi_0}(\theta) \geq \alpha \;\; \forall \theta \in \Theta_1$ since ϕ_0 is UMP α–similiar. This implies that ϕ_0 is unbiased.

Since $\beta_\phi(\theta)$ is continuous in θ, we see from Theorem 9.4.6 that the class of unbiased tests is a subclass of the class of α–similar tests. Since ϕ_0 is UMP in the larger class, it is also UMP in the subclass. Thus, ϕ_0 is UMPU. ■

Note:

The continuity of the power function $\beta_\phi(\theta)$ cannot always be checked easily. ■

Example 9.4.9:

Let $X_1, \ldots, X_n \sim N(\mu, 1)$.

Let $H_0: \mu \leq 0$ vs $H_1: \mu > 0$.

Since the family of densities has a MLR in $\sum\limits_{i=1}^n X_i$, we could use Theorem 9.3.5 to find a UMP test. However, we want to illustate the use of Theorem 9.4.8 here.

It is $\Lambda = \{0\}$ and the power function

$$\beta_\phi(\mu) = \int_{I\!R^n} \phi(\underline{x}) \left(\frac{1}{\sqrt{2\pi}} \right)^n \exp\left(-\frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 \right) d\underline{x}$$

of any test ϕ is continuous in μ. Thus, due to Theorem 9.4.6, any unbiased size α test of H_0 is α–similar on Λ.

We need a UMP test of $H_0': \mu = 0$ vs $H_1: \mu > 0$.

By the NP Lemma, a MP test of $H_0'': \mu = 0$ vs $H_1'': \mu = \mu_1$, where $\mu_1 > 0$ is given by

$$\phi(\underline{x}) = \begin{cases} 1, & \text{if } \exp\left(\frac{\sum x_i^2}{2} - \frac{\sum (x_i - \mu)^2}{2} \right) > k' \\ 0, & \text{otherwise} \end{cases}$$

or equivalently, by Theorem 9.2.2,

$$\phi(\underline{x}) = \begin{cases} 1, & \text{if } T = \sum_{i=1}^{n} X_i > k \\ 0, & \text{otherwise} \end{cases}$$

Since under H_0, $T \sim N(0, n)$, k is determined by $\alpha = P_{\mu=0}(T > k) = P(\frac{T}{\sqrt{n}} > \frac{k}{\sqrt{n}})$, i.e., $k = \sqrt{n} z_\alpha$.

ϕ is independent of μ_1 for every $\mu_1 > 0$. So ϕ is UMP α–similar for H_0' vs. H_1.

Finally, ϕ is of size α, since for $\mu \leq 0$, it holds that

$$\begin{aligned} E_\mu(\phi(\underline{X})) &= P_\mu(T > \sqrt{n} z_\alpha) \\ &= P\left(\frac{T - n\mu}{\sqrt{n}} > z_\alpha - \sqrt{n}\mu\right) \\ &\overset{(*)}{\leq} P(Z > z_\alpha) \\ &= \alpha \end{aligned}$$

$(*)$ holds since $\frac{T-n\mu}{\sqrt{n}} \sim N(0,1)$ for $\mu \leq 0$ and $z_\alpha - \sqrt{n}\mu \geq z_\alpha$ for $\mu \leq 0$.

Thus all the requirements are met for Theorem 9.4.8, i.e., β_ϕ is continuous and ϕ is UMP α–similar and of size α, and thus ϕ is UMPU. ■

Note:
Rohatgi, page 428–430, lists Theorems (without proofs), stating that for Normal data, one– and two–tailed t–tests, one– and two–tailed χ^2–tests, two–sample t–tests, and F–tests are all UMPU. ■

Note:
Recall from Definition 8.2.4 that a class of distributions is invariant under a group $\mathcal{G}$ of transformations, if for each $g \in \mathcal{G}$ and for each $\theta \in \Theta$ there exists a unique $\theta' \in \Theta$ such that if $\underline{X} \sim P_\theta$, then $g(\underline{X}) \sim P_{\theta'}$. ■

Definition 9.4.10:
A group $\mathcal{G}$ of transformations on $\underline{X}$ leaves a hypothesis testing problem **invariant** if $\mathcal{G}$ leaves both $\{P_\theta : \theta \in \Theta_0\}$ and $\{P_\theta : \theta \in \Theta_1\}$ invariant, i.e., if $\underline{y} = g(\underline{x}) \sim h_\theta(\underline{y})$, then $\{f_\theta(\underline{x}) : \theta \in \Theta_0\} \equiv \{h_\theta(\underline{y}) : \theta \in \Theta_0\}$ and $\{f_\theta(\underline{x}) : \theta \in \Theta_1\} \equiv \{h_\theta(\underline{y}) : \theta \in \Theta_1\}$. ■

Note:

We want two types of invariance for our tests:

Measurement Invariance: If $\underline{y} = g(\underline{x})$ is a 1–to–1 mapping, the decision based on $\underline{y}$ should be the same as the decision based on $\underline{x}$. If $\phi(\underline{x})$ is the test based on $\underline{x}$ and $\phi^*(\underline{y})$ is the test based on $\underline{y}$, then it must hold that $\phi(\underline{x}) = \phi^*(g(\underline{x})) = \phi^*(\underline{y})$.

Formal Invariance: If two tests have the same structure, i.e, the same Θ, the same pdf's (or pmf's), and the same hypotheses, then we should use the same test in both problems. So, if the transformed problem in terms of $\underline{y}$ has the same formal structure as that of the problem in terms of $\underline{x}$, we must have that $\phi^*(\underline{y}) = \phi(\underline{x}) = \phi^*(g(\underline{x}))$.

We can combine these two requirements in the following definition: ■

Definition 9.4.11:

An **invariant test** with respect to a group $\mathcal{G}$ of tansformations is any test ϕ such that

$$\phi(\underline{x}) = \phi(g(\underline{x})) \quad \forall \underline{x} \quad \forall g \in \mathcal{G}.$$

■

Example 9.4.12:

Let $X \sim Bin(n,p)$. Let $H_0: \ p = \frac{1}{2}$ vs. $H_1: \ p \neq \frac{1}{2}$.

Let $\mathcal{G} = \{g_1, g_2\}$, where $g_1(x) = n - x$ and $g_2(x) = x$.

If ϕ is invariant, then $\phi(x) = \phi(n-x)$. Is the test problem invariant? For g_2, the answer is obvious.

For g_1, we get:

$$g_1(X) = n - X \sim Bin(n, 1-p)$$

$$H_0: \ p = \tfrac{1}{2} : \{f_p(x) : p = \tfrac{1}{2}\} = \{h_p(g_1(x)) : p = \tfrac{1}{2}\} = Bin(n, \tfrac{1}{2})$$

$$H_1: \ p \neq \tfrac{1}{2} : \underbrace{\{f_p(x) : p \neq \frac{1}{2}\}}_{=Bin(n, p \neq \frac{1}{2})} = \underbrace{\{h_p(g_1(x)) : p \neq \frac{1}{2}\}}_{=Bin(n, p \neq \frac{1}{2})}$$

So all the requirements in Definition 9.4.10 are met. If, for example, $n = 10$, the test

$$\phi(x) = \begin{cases} 1, & \text{if } x = 0,1,2,8,9,10 \\ 0, & \text{otherwise} \end{cases}$$

is invariant under $\mathcal{G}$. For example, $\phi(4) = 0 = \phi(10-4) = \phi(6)$, and, in general, $\phi(x) = \phi(10-x) \ \ \forall x \in \{0, 1, \ldots, 9, 10\}$. ■

Lecture 25:
Fr 03/09/01

Example 9.4.13:
Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ where both μ and $\sigma^2 > 0$ are unknown. It is $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$ and $\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$ and $\overline{X}$ and S^2 independent.

Let $H_0: \ \mu \leq 0$ vs. $H_1: \ \mu > 0$.

Let $\mathcal{G}$ be the group of scale changes:

$$\mathcal{G} = \{g_c(\overline{x}, s^2), c > 0: \quad g_c(\overline{x}, s^2) = (c\overline{x}, c^2 s^2)\}$$

The problem is invariant because, when $g_c(\overline{x}, s^2) = (c\overline{x}, c^2 s^2)$, then

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

(ii) $c\overline{X} \sim N(c\mu, \frac{c^2\sigma^2}{n}) \in \{N(\eta, \frac{\tau^2}{n})\}$.

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

So, this is the same family of distributions and Definition 9.4.10 holds because $\mu \leq 0$ implies that $c\mu \leq 0$ (for $c > 0$).

An invariant test satisfies $\phi(\overline{x}, s^2) \equiv \phi(c\overline{x}, c^2 s^2), \ c > 0, s^2 > 0, \overline{x} \in I\!R$.

Let $c = \frac{1}{s}$. Then $\phi(\overline{x}, s^2) \equiv \phi(\frac{\overline{x}}{s}, 1)$ so invariant tests depend on $(\overline{x}, s^2)$ only through $\frac{\overline{x}}{s}$.

If $\frac{\overline{x}_1}{s_1} \neq \frac{\overline{x}_2}{s_2}$, then there exists no $c > 0$ such that $(\overline{x}_2, s_2^2) \equiv (c\overline{x}_1, c^2 s_1^2)$. So invariance places no restrictions on ϕ for different $\frac{\overline{x}_1}{s_1} = \frac{\overline{x}_2}{s_2}$. Thus, invariant tests are exactly those that depend only on $\frac{\overline{x}}{s}$, which are equivalent to tests that are based only on $t = \frac{\overline{x}}{s/\sqrt{n}}$. Since this mapping is 1–to–1, the invariant test will use $T = \frac{\overline{X}}{S/\sqrt{n}} \sim t_{n-1}$ if $\mu = 0$. Note that this test does not depend on the nuisance parameter σ^2. Invariance often produces such results. ■

Definition 9.4.14:
Let $\mathcal{G}$ be a group of transformations on the space of $\underline{X}$. We say a statistic $T(\underline{x})$ is **maximal invariant** under $\mathcal{G}$ if

(i) T is invariant, i.e., $T(\underline{x}) = T(g(\underline{x})) \quad \forall g \in \mathcal{G}$, and

(ii) T is maximal, i.e., $T(\underline{x}_1) = T(\underline{x}_2)$ implies that $\underline{x}_1 = g(\underline{x}_2)$ for some $g \in \mathcal{G}$.

■

Example 9.4.15:
Let $\underline{x} = (x_1, \ldots, x_n)$ and $g_c(\underline{x}) = (x_1 + c, \ldots, x_n + c)$.

Consider $T(\underline{x}) = (x_n - x_1, x_n - x_2, \ldots, x_n - x_{n-1})$.

It is $T(g_c(\underline{x})) = (x_n - x_1, x_n - x_2, \ldots, x_n - x_{n-1}) = T(\underline{x})$, so T is invariant.

If $T(\underline{x}) = T(\underline{x}')$, then $x_n - x_i = x_n' - x_i' \ \ \forall i = 1, 2, \ldots, n-1$.

This implies that $x_i - x_i' = x_n - x_n' = c \ \ \forall i = 1, 2, \ldots, n-1$.

Thus, $g_c(\underline{x}') = (x_1' + c, \ldots, x_n' + c) = \underline{x}$.

Therefore, T is maximal invariant. ■

Definition 9.4.16:
Let I_α be the class of all invariant tests of size α of $H_0 : \ \theta \in \Theta_0$ vs. $H_1 : \ \theta \in \Theta_1$. If there exists a UMP member in I_α, it is called the **UMP invariant** test of H_0 vs H_1. ■

Theorem 9.4.17:
Let $T(\underline{x})$ be maximal invariant with respect to $\mathcal{G}$. A test ϕ is invariant under $\mathcal{G}$ iff ϕ is a function of T.

Proof:
"$\Longrightarrow$":
Let ϕ be invariant under $\mathcal{G}$. If $T(\underline{x}_1) = T(\underline{x}_2)$, then there exists a $g \in \mathcal{G}$ such that $\underline{x}_1 = g(\underline{x}_2)$. Thus, it follows from invariance that $\phi(\underline{x}_1) = \phi(g(\underline{x}_2)) = \phi(\underline{x}_2)$. Since ϕ is the same whenever $T(\underline{x}_1) = T(\underline{x}_2)$, ϕ must be a function of T.

"$\Longleftarrow$":
Let ϕ be a function of T, i.e., $\phi(\underline{x}) = h(T(\underline{x}))$. It follows that

$$\phi(g(\underline{x})) = h(T(g(\underline{x}))) \stackrel{(*)}{=} h(T(\underline{x})) = \phi(\underline{x}).$$

$(*)$ holds since T is invariant.

This means that ϕ is invariant. ■

Example 9.4.18:
Consider the test problem

$$H_0: \ \underline{X} \sim f_0(x_1-\theta,\ldots,x_n-\theta) \text{ vs. } H_1: \ \underline{X} \sim f_1(x_1-\theta,\ldots,x_n-\theta),$$

where $\theta \in I\!R$.

Let $\mathcal{G}$ be the group of transformations with

$$g_c(\underline{x}) = (x_1+c,\ldots,x_n+c),$$

where $c \in I\!R$ and $n \geq 2$.

As shown in Example 9.4.15, a maximal invariant statistic is $T(\underline{X}) = (X_1 - X_n, \ldots, X_{n-1} - X_n) = (T_1,\ldots,T_{n-1})$. Due to Theorem 9.4.17, an invariant test ϕ depends on $\underline{X}$ only through T.

Since the transformation

$$\begin{pmatrix} T' \\ Z \end{pmatrix} = \begin{pmatrix} T_1 \\ \vdots \\ T_{n-1} \\ Z \end{pmatrix} = \begin{pmatrix} X_1 - X_n \\ \vdots \\ X_{n-1} - X_n \\ X_n \end{pmatrix}$$

is 1–to–1, there exists inverses $X_n = Z$ and $X_i = T_i + X_n = T_i + Z \quad \forall i = 1,\ldots,n-1$. Applying Theorem 4.3.5 and integrating out the last component Z $(= X_n)$ gives us the joint pdf of $T = (T_1,\ldots,T_{n-1})$.

Thus, under $H_i, i = 0,1$, the joint pdf of T is given by $\displaystyle\int_{-\infty}^{\infty} f_i(t_1+z, t_2+z,\ldots,t_{n-1}+z,z)dz$ which is independent of θ. The problem is thus reduced to testing a simple hypothesis against a simple alternative. By the NP Lemma (Theorem 9.2.1), the MP test is

$$\phi(t_1,\ldots,t_{n-1}) = \begin{cases} 1, & \text{if } \lambda(\underline{t}) > c \\ 0, & \text{if } \lambda(\underline{t}) < c \end{cases}$$

where $\underline{t} = (t_1,\ldots,t_{n-1})$ and $\lambda(\underline{t}) = \dfrac{\displaystyle\int_{-\infty}^{\infty} f_1(t_1+z,t_2+z,\ldots,t_{n-1}+z,z)dz}{\displaystyle\int_{-\infty}^{\infty} f_0(t_1+z,t_2+z,\ldots,t_{n-1}+z,z)dz}$.

In the homework assignment, we use this result to construct a UMP invariant test of

$$H_0: \ \underline{X} \sim N(\theta,1) \text{ vs. } H_1: \ \underline{X} \sim \text{ Cauchy}(1,\theta),$$

where a Cauchy$(1,\theta)$ distribution has pdf $f(x;\theta) = \dfrac{1}{\pi}\dfrac{1}{1+(x-\theta)^2}$, where $\theta \in I\!R$. ■

Lecture 26:
Mo 03/19/01

10 More on Hypothesis Testing

10.1 Likelihood Ratio Tests

Definition 10.1.1:
The **likelihood ratio test statistic** for

$$H_0: \ \theta \in \Theta_0 \text{ vs. } H_1: \ \theta \in \Theta_1 = \Theta - \Theta_0$$

is

$$\lambda(\underline{x}) = \frac{\sup\limits_{\theta \in \Theta_0} f_\theta(\underline{x})}{\sup\limits_{\theta \in \Theta} f_\theta(\underline{x})}.$$

The **likelihood ratio test** (**LRT**) is the test function

$$\phi(\underline{x}) = I_{[0,c)}(\lambda(\underline{x})),$$

for some constant $c \in [0,1]$, where c is usually chosen in such a way to make ϕ a test of size α. ■

Note:

(i) We have to select c such that $0 \leq c \leq 1$ since $0 \leq \lambda(\underline{x}) \leq 1$.

(ii) LRT's are strongly related to MLE's. If $\hat{\theta}$ is the unrestricted MLE of θ over Θ and $\hat{\theta}_0$ is the MLE of θ over Θ_0, then $\lambda(\underline{x}) = \frac{f_{\hat{\theta}_0}(\underline{x})}{f_{\hat{\theta}}(\underline{x})}$.

■

Example 10.1.2:
Let $X_1, \ldots, X_n$ be a sample from $N(\mu, 1)$. We want to construct a LRT for

$$H_0: \ \mu = \mu_0 \text{ vs. } H_1: \ \mu \neq \mu_0.$$

It is $\hat{\mu}_0 = \mu_0$ and $\hat{\mu} = \overline{X}$. Thus,

$$\lambda(\underline{x}) = \frac{(2\pi)^{-n/2} \exp(-\frac{1}{2} \sum (x_i - \mu_0)^2)}{(2\pi)^{-n/2} \exp(-\frac{1}{2} \sum (x_i - \overline{x})^2)} = \exp(-\frac{n}{2}(\overline{x} - \mu_0)^2).$$

The LRT rejects H_0 if $\lambda(\underline{x}) \leq c$, or equivalently, $| \, \overline{x} - \mu_0 \, | \geq \sqrt{-2 \frac{\log c}{n}}$. This means, the LRT rejects $H_0: \ \mu = \mu_0$ if $\overline{x}$ is *too far from* μ_0. ■

Theorem 10.1.3:

If $T(\underline{X})$ is sufficient for θ and $\lambda^*(t)$ and $\lambda(\underline{x})$ are LRT statistics based on T and $\underline{X}$ respectively, then

$$\lambda^*(T(\underline{x})) = \lambda(\underline{x}) \quad \forall \underline{x},$$

i.e., the LRT can be expressed as a function of every sufficient statistic for θ.

Proof:

Since T is sufficient, it follows from Theorem 8.3.5 that its pdf (or pmf) factorizes as $f_\theta(\underline{x}) = g_\theta(T)h(\underline{x})$. Therefore we get:

$$\begin{aligned}
\lambda(\underline{x}) &= \frac{\sup\limits_{\theta \in \Theta_0} f_\theta(\underline{x})}{\sup\limits_{\theta \in \Theta} f_\theta(\underline{x})} \\
&= \frac{\sup\limits_{\theta \in \Theta_0} g_\theta(T)h(\underline{x})}{\sup\limits_{\theta \in \Theta} g_\theta(T)h(\underline{x})} \\
&= \frac{\sup\limits_{\theta \in \Theta_0} g_\theta(T)}{\sup\limits_{\theta \in \Theta} g_\theta(T)} \\
&= \lambda^*(T(\underline{x}))
\end{aligned}$$

Thus, our simplified expression for $\lambda(\underline{x})$ indeed only depends on a sufficient statistic T. ■

Theorem 10.1.4:

If for a given α, $0 \leq \alpha \leq 1$, and for a simple hypothesis H_0 and a simple alternative H_1 a non–randomized test based on the NP Lemma and a LRT exist, then these tests are equivalent. ■

Note:

Usually, LRT's perform well since they are often UMP or UMPU size α tests. However, this does not always hold. Rohatgi, Example 4, page 440–441, cites an example where the LRT is not unbiased and it is even worse than the trivial test $\phi(\underline{x}) = \alpha$. ■

Theorem 10.1.5:

Under some regularity conditions on $f_\theta(\underline{x})$, the rv $-2\log\lambda(\underline{X})$ under H_0 has asymptotically a chi–squared distribution with ν degrees of freedom, where ν equals the difference between the number of independent parameters in Θ and Θ_0, i.e.,

$$-2\log\lambda(\underline{X}) \xrightarrow{d} \chi^2_\nu \quad \text{under } H_0.$$

■

Note:

The regularity conditions required for Theorem 10.1.5 are basically the same as for Theorem 8.7.10. Under "independent" parameters we understand parameters that are unspecified, i.e., free to vary. ■

Example 10.1.6:

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ where $\mu \in I\!R$ and $\sigma^2 > 0$ are both unknown.

Let $H_0: \ \mu = \mu_0$ vs. $H_1: \ \mu \neq \mu_0$.

We have $\theta = (\mu, \sigma^2)$, $\Theta = \{(\mu, \sigma^2): \ \mu \in I\!R, \sigma^2 > 0\}$ and $\Theta_0 = \{(\mu_0, \sigma^2): \ \sigma^2 > 0\}$.

It is $\hat{\theta}_0 = (\mu_0, \frac{1}{n}\sum\limits_{i=1}^{n}(x_i - \mu_0)^2)$ and $\hat{\theta} = (\overline{x}, \frac{1}{n}\sum\limits_{i=1}^{n}(x_i - \overline{x})^2)$.

Now, the LR test statistic $\lambda(\underline{x})$ can be determined:

$$\begin{aligned}
\lambda(\underline{x}) &= \frac{f_{\hat{\theta}_0}(\underline{x})}{f_{\hat{\theta}}(\underline{x})} \\
&= \frac{\frac{1}{(\frac{2\pi}{n})^{\frac{n}{2}}(\sum(x_i-\mu_0)^2)^{\frac{n}{2}}} \exp\left(-\frac{\sum(x_i-\mu_0)^2}{2\frac{1}{n}\sum(x_i-\mu_0)^2}\right)}{\frac{1}{(\frac{2\pi}{n})^{\frac{n}{2}}(\sum(x_i-\overline{x})^2)^{\frac{n}{2}}} \exp\left(-\frac{\sum(x_i-\overline{x})^2}{2\frac{1}{n}\sum(x_i-\overline{x})^2}\right)} \\
&= \left(\frac{\sum(x_i - \overline{x})^2}{\sum(x_i - \mu_0)^2}\right)^{\frac{n}{2}} \\
&= \left(\frac{\sum x_i^2 - n\overline{x}^2}{\sum x_i^2 - 2\mu_0 \sum x_i + n\mu_0^2 - n\overline{x}^2 + n\overline{x}^2}\right)^{\frac{n}{2}} \\
&= \left(\frac{1}{1 + \left(\frac{n(\overline{x}-\mu_0)^2}{\sum(x_i-\overline{x})^2}\right)}\right)^{\frac{n}{2}}
\end{aligned}$$

Note that this is a decreasing function of

$$t(\underline{X}) = \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sqrt{\frac{1}{n-1}\sum(X_i - \overline{X})^2}} = \frac{\sqrt{n}(\overline{X} - \mu_0)}{S} \overset{(*)}{\sim} t_{n-1}.$$

$(*)$ holds due to Corollary 7.2.4.

So we reject H_0 if $t(\underline{x})$ is large. Now,

$$\begin{aligned}
-2\log\lambda(\underline{x}) &= -2(-\frac{n}{2})\log\left(1 + n\frac{(\overline{x} - \mu_0)^2}{\sum(x_i - \overline{x})^2}\right) \\
&= n\log\left(1 + n\frac{(\overline{x} - \mu_0)^2}{\sum(x_i - \overline{x})^2}\right)
\end{aligned}$$

Under H_0, $\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma} \sim N(0,1)$ and $\frac{\sum(X_i-\overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$ and both are independent.

Therefore, under H_0,

$$\frac{n(\overline{X}-\mu_0)^2}{\frac{1}{n-1}\sum(X_i-\overline{X})^2} \sim F_{1,n-1}.$$

Thus, the mgf of $-2\log\lambda(\underline{X})$ under H_0 is

$$\begin{aligned}
M_n(t) &= E_{H_0}(\exp(-2t\log\lambda(\underline{X}))) \\
&= E_{H_0}(\exp(nt\log(1+\frac{F}{n-1}))) \\
&= E_{H_0}(\exp(\log(1+\frac{F}{n-1})^{nt})) \\
&= E_{H_0}((1+\frac{F}{n-1})^{nt}) \\
&= \int_0^\infty \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})\Gamma(\frac{1}{2})(n-1)^{\frac{1}{2}}} \frac{1}{\sqrt{f}} \left(1+\frac{f}{n-1}\right)^{-\frac{n}{2}} \left(1+\frac{f}{n-1}\right)^{nt} df
\end{aligned}$$

Note that

$$f_{1,n-1}(f) = \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})\Gamma(\frac{1}{2})(n-1)^{\frac{1}{2}}} \frac{1}{\sqrt{f}} \left(1+\frac{f}{n-1}\right)^{-\frac{n}{2}} \cdot I_{[0,\infty)}(f)$$

is the pdf of a $F_{1,n-1}$ distribution.

Let $y = \left(1+\frac{f}{n-1}\right)^{-1}$, then $\frac{f}{n-1} = \frac{1-y}{y}$ and $df = -\frac{n-1}{y^2}dy$.

Thus,

$$\begin{aligned}
M_n(t) &= \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})\Gamma(\frac{1}{2})(n-1)^{\frac{1}{2}}} \int_0^\infty \frac{1}{\sqrt{f}} \left(1+\frac{f}{n-1}\right)^{nt-\frac{n}{2}} df \\
&= \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})\Gamma(\frac{1}{2})(n-1)^{\frac{1}{2}}} (n-1)^{\frac{1}{2}} \int_0^1 y^{\frac{n-3}{2}-nt}(1-y)^{-\frac{1}{2}}\, dy \\
&\overset{(*)}{=} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})\Gamma(\frac{1}{2})} B(\frac{n-1}{2}-nt, \frac{1}{2}) \\
&= \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})\Gamma(\frac{1}{2})} \frac{\Gamma(\frac{n-1}{2}-nt)\Gamma(\frac{1}{2})}{\Gamma(\frac{n}{2}-nt)} \\
&= \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \frac{\Gamma(\frac{n-1}{2}-nt)}{\Gamma(\frac{n}{2}-nt)}, \quad t < \frac{1}{2}-\frac{1}{2n}
\end{aligned}$$

$(*)$ holds since the integral represents the Beta function (see also Example 8.8.11).

As $n \to \infty$, we can apply Stirling's formula which states that

$$\Gamma(\alpha(n)+1) \approx (\alpha(n))! \approx \sqrt{2\pi}(\alpha(n))^{\alpha(n)+\frac{1}{2}} \exp(-\alpha(n)).$$

Lecture 27: We 03/21/01

So,

$$\begin{aligned}
M_n(t) &\approx \frac{\sqrt{2\pi}(\frac{n-2}{2})^{\frac{n-1}{2}} \exp(-\frac{n-2}{2}) \sqrt{2\pi}(\frac{n(1-2t)-3}{2})^{\frac{n(1-2t)-2}{2}} \exp(-\frac{n(1-2t)-3}{2})}{\sqrt{2\pi}(\frac{n-3}{2})^{\frac{n-2}{2}} \exp(-\frac{n-3}{2}) \sqrt{2\pi}(\frac{n(1-2t)-2}{2})^{\frac{n(1-2t)-1}{2}} \exp(-\frac{n(1-2t)-2}{2})} \\
&= \left(\frac{n-2}{n-3}\right)^{\frac{n-2}{2}} \left(\frac{n-2}{2}\right)^{\frac{1}{2}} \left(\frac{n(1-2t)-3}{n(1-2t)-2}\right)^{\frac{n(1-2t)-2}{2}} \left(\frac{n(1-2t)-2}{2}\right)^{-\frac{1}{2}} \\
&= \underbrace{\left((1+\frac{1}{n-3})^{n-2}\right)^{\frac{1}{2}}}_{\longrightarrow e^{\frac{1}{2}} \text{ as } n\to\infty} \underbrace{\left((1-\frac{1}{n(1-2t)-2})^{n(1-2t)-2}\right)^{\frac{1}{2}}}_{\longrightarrow e^{-\frac{1}{2}} \text{ as } n\to\infty} \underbrace{\left(\frac{n-2}{n(1-2t)-2}\right)^{\frac{1}{2}}}_{\longrightarrow (1-2t)^{-\frac{1}{2}} \text{ as } n\to\infty}
\end{aligned}$$

Thus,

$$M_n(t) \longrightarrow \frac{1}{(1-2t)^{\frac{1}{2}}} \text{ as } n \to \infty.$$

Note that this is the mgf of a χ_1^2 distribution. Therefore, it follows by the Continuity Theorem (Theorem 6.4.2) that

$$-2\log\lambda(\underline{X}) \xrightarrow{d} \chi_1^2.$$

Obviously, we could use Theorem 10.1.5 (after checking that the regularity conditions hold) as well to obtain the same result. Since under H_0 1 parameter (σ^2) is unspecified and under H_1 2 parameters (μ, σ^2) are unspecified, it is $\nu = 2 - 1 = 1$. ■

10.2 Parametric Chi–Squared Tests

Definition 10.2.1: **Normal Variance Tests**

Let $X_1, \ldots, X_n$ be a sample from a $N(\mu, \sigma^2)$ distribution where μ may be known or unknown and $\sigma^2 > 0$ is unknown. The following table summarizes the χ^2 tests that are typically being used:

			Reject H_0 at level α if	
	H_0	H_1	μ known	μ unknown
I	$\sigma \geq \sigma_0$	$\sigma < \sigma_0$	$\sum (x_i - \mu)^2 \leq \sigma_0^2 \chi^2_{n;1-\alpha}$	$s^2 \leq \frac{\sigma_0^2}{n-1} \chi^2_{n-1;1-\alpha}$
II	$\sigma \leq \sigma_0$	$\sigma > \sigma_0$	$\sum (x_i - \mu)^2 \geq \sigma_0^2 \chi^2_{n;\alpha}$	$s^2 \geq \frac{\sigma_0^2}{n-1} \chi^2_{n-1;\alpha}$
III	$\sigma = \sigma_0$	$\sigma \neq \sigma_0$	$\sum (x_i - \mu)^2 \leq \sigma_0^2 \chi^2_{n;1-\alpha/2}$	$s^2 \leq \frac{\sigma_0^2}{n-1} \chi^2_{n-1;1-\alpha/2}$
			or $\sum (x_i - \mu)^2 \geq \sigma_0^2 \chi^2_{n;\alpha/2}$	or $s^2 \geq \frac{\sigma_0^2}{n-1} \chi^2_{n-1;\alpha/2}$

■

Note:

(i) In Definition 10.2.1, σ_0 is any fixed positive constant.

(ii) Tests I and II are UMPU if μ is unknown and UMP if μ is known.

(iii) In test III, the constants have been chosen in such a way to give equal probability to each tail. This is the usual approach. However, this may result in a biased test.

(iv) $\chi^2_{n;1-\alpha}$ is the (lower) α quantile and $\chi^2_{n;\alpha}$ is the (upper) $1-\alpha$ quantile, i.e., for $X \sim \chi^2_n$, it holds that $P(X \leq \chi^2_{n;1-\alpha}) = \alpha$ and $P(X \leq \chi^2_{n;\alpha}) = 1 - \alpha$.

(v) We can also use χ^2 tests to test for equality of binomial probabilities as shown in the next few Theorems.

■

Theorem 10.2.2:

Let $X_1, \ldots, X_k$ be independent rv's with $X_i \sim Bin(n_i, p_i), i = 1, \ldots, k$. Then it holds that

$$T = \sum_{i=1}^{k} \left(\frac{X_i - n_i p_i}{\sqrt{n_i p_i (1 - p_i)}} \right)^2 \xrightarrow{d} \chi^2_k$$

as $n_1, \ldots, n_k \to \infty$.

Proof:

Homework ■

Corollary 10.2.3:
Let $X_1, \ldots, X_k$ be as in Theorem 10.2.2 above. We want to test the hypothesis that $H_0: p_1 = p_2 = \ldots = p_k = p$, where p is a known constant. An appoximate level–α test rejects H_0 if

$$y = \sum_{i=1}^{k} \left(\frac{x_i - n_i p}{\sqrt{n_i p(1-p)}} \right)^2 \geq \chi^2_{k;\alpha}.$$

■

Theorem 10.2.4:
Let $X_1, \ldots, X_k$ be independent rv's with $X_i \sim Bin(n_i, p), i = 1, \ldots, k$. Then the MLE of p is

$$\hat{p} = \frac{\sum\limits_{i=1}^{k} x_i}{\sum\limits_{i=1}^{k} n_i}.$$

Proof:
This can be shown by using the joint likelihood function or by the fact that $\sum X_i \sim Bin(\sum n_i, p)$ and for $X \sim Bin(n, p)$, the MLE is $\hat{p} = \frac{x}{n}$. ■

Theorem 10.2.5:
Let $X_1, \ldots, X_k$ be independent rv's with $X_i \sim Bin(n_i, p_i), i = 1, \ldots, k$. An approximate level–$\alpha$ test of $H_0: p_1 = p_2 = \ldots = p_k = p$, where p is unknown, rejects H_0 if

$$y = \sum_{i=1}^{k} \left(\frac{x_i - n_i \hat{p}}{\sqrt{n_i \hat{p}(1-\hat{p})}} \right)^2 \geq \chi^2_{k-1;\alpha},$$

where $\hat{p} = \frac{\sum x_i}{\sum n_i}$. ■

Theorem 10.2.6:
Let $(X_1, \ldots, X_k)$ be a multinomial rv with parameters $n, p_1, p_2, \ldots, p_k$ where $\sum\limits_{i=1}^{k} p_i = 1$ and $\sum\limits_{i=1}^{k} X_i = n$. Then it holds that

$$U_k = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \xrightarrow{d} \chi^2_{k-1}$$

as $n \to \infty$.

An approximate level–α test of $H_0: p_1 = p'_1, p_2 = p'_2, \ldots, p_k = p'_k$ rejects H_0 if

$$\sum_{i=1}^{k} \frac{(x_i - np'_i)^2}{np'_i} > \chi^2_{k-1;\alpha}.$$

Proof:
Case $k = 2$ only:

$$\begin{aligned} U_2 &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} \\ &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(n - X_1 - n(1-p_1))^2}{n(1-p_1)} \\ &= (X_1 - np_1)^2 \left(\frac{1}{np_1} + \frac{1}{n(1-p_1)} \right) \\ &= (X_1 - np_1)^2 \left(\frac{(1-p_1) + p_1}{np_1(1-p_1)} \right) \\ &= \frac{(X_1 - np_1)^2}{np_1(1-p_1)} \end{aligned}$$

By the CLT, $\dfrac{X_1 - np_1}{\sqrt{np_1(1-p_1)}} \xrightarrow{d} N(0,1)$. Therefore, $U_2 \xrightarrow{d} \chi_1^2$. ■

Lecture 28: Fr 03/23/01

Theorem 10.2.7:
Let $X_1, \ldots, X_n$ be a sample from X. Let $H_0: \ X \sim F$, where the functional form of F is known completely. We partition the real line into k disjoint Borel sets $A_1, \ldots, A_k$ and let $P(X \in A_i) = p_i$, where $p_i > 0 \ \ \forall i = 1, \ldots, k$.

Let $Y_j = \# X_i's$ in $A_j = \displaystyle\sum_{i=1}^{n} I_{A_j}(X_i), \ \ \forall j = 1, \ldots, k$.

Then, $(Y_1, \ldots, Y_k)$ has multinomial distribution with parameters $n, p_1, p_2, \ldots, p_k$. ■

Theorem 10.2.8:
Let $X_1, \ldots, X_n$ be a sample from X. Let $H_0: \ X \sim F_{\underline{\theta}}$, where $\underline{\theta} = (\theta_1, \ldots, \theta_r)$ is unknown. Let the MLE $\hat{\underline{\theta}}$ exist. We partition the real line into k disjoint Borel sets $A_1, \ldots, A_k$ and let $P_{\hat{\underline{\theta}}}(X \in A_i) = \hat{p}_i$, where $\hat{p}_i > 0 \ \ \forall i = 1, \ldots, k$.

Let $Y_j = \# X_i's$ in $A_j = \displaystyle\sum_{i=1}^{n} I_{A_j}(X_i), \ \ \forall j = 1, \ldots, k$.

Then it holds that

$$V_k = \sum_{i=1}^{k} \frac{(Y_i - n\hat{p}_i)^2}{n\hat{p}_i} \xrightarrow{d} \chi^2_{k-r-1}.$$

An approximate level–α test of $H_0: \ X \sim F_{\underline{\theta}}$ rejects H_0 if

$$\sum_{i=1}^{k} \frac{(y_i - n\hat{p}_i)^2}{n\hat{p}_i} > \chi^2_{k-r-1;\alpha},$$

where r is the number of parameters in $\underline{\theta}$ that have to be estimated. ■

10.3 t–Tests and F–Tests

Definition 10.3.1: **One– and Two–Tailed t-Tests**
Let $X_1, \ldots, X_n$ be a sample from a $N(\mu, \sigma^2)$ distribution where $\sigma^2 > 0$ may be known or unknown and μ is unknown. Let $\overline{X} = \frac{1}{n} \sum\limits_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1} \sum\limits_{i=1}^{n} (X_i - \overline{X})^2$.

The following table summarizes the z– and t–tests that are typically being used:

			Reject H_0 at level α if	
	H_0	H_1	σ^2 known	σ^2 unknown
I	$\mu \leq \mu_0$	$\mu > \mu_0$	$\overline{X} \geq \mu_0 + \frac{\sigma}{\sqrt{n}} z_\alpha$	$\overline{X} \geq \mu_0 + \frac{s}{\sqrt{n}} t_{n-1;\alpha}$
II	$\mu \geq \mu_0$	$\mu < \mu_0$	$\overline{X} \leq \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$	$\overline{X} \leq \mu_0 + \frac{s}{\sqrt{n}} t_{n-1;1-\alpha}$
III	$\mu = \mu_0$	$\mu \neq \mu_0$	$\mid \overline{X} - \mu_0 \mid \geq \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$	$\mid \overline{X} - \mu_0 \mid \geq \frac{s}{\sqrt{n}} t_{n-1;\alpha/2}$

■

Note:

(i) In Definition 10.3.1, μ_0 is any fixed constant.

(ii) These tests are based on just one sample and are often called *one sample t–tests*.

(iii) Tests I and II are UMP and test III is UMPU if σ^2 is known. Tests I, II, and III are UMPU and UMP invariant if σ^2 is unknown.

(iv) For large n (≥ 30), we can use z–tables instead of t-tables. Also, for large n we can drop the Normality assumption due to the CLT. However, for small n, none of these simplifications is justified.

■

Definition 10.3.2: **Two–Sample t-Tests**
Let $X_1, \ldots, X_m$ be a sample from a $N(\mu_1, \sigma_1^2)$ distribution where $\sigma_1^2 > 0$ may be known or unknown and μ_1 is unknown. Let $Y_1, \ldots, Y_n$ be a sample from a $N(\mu_2, \sigma_2^2)$ distribution where $\sigma_2^2 > 0$ may be known or unknown and μ_2 is unknown.

Let $\overline{X} = \frac{1}{m} \sum\limits_{i=1}^{m} X_i$ and $S_1^2 = \frac{1}{m-1} \sum\limits_{i=1}^{m} (X_i - \overline{X})^2$.

Let $\overline{Y} = \frac{1}{n} \sum\limits_{i=1}^{n} Y_i$ and $S_2^2 = \frac{1}{n-1} \sum\limits_{i=1}^{n} (Y_i - \overline{Y})^2$.

Let $S_p^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}$.

The following table summarizes the z– and t–tests that are typically being used:

			Reject H_0 at level α if	
	H_0	H_1	σ_1^2, σ_2^2 known	σ_1^2, σ_2^2 unknown, $\sigma_1 = \sigma_2$
I	$\mu_1 - \mu_2 \leq \delta$	$\mu_1 - \mu_2 > \delta$	$\overline{X} - \overline{Y} \geq \delta + z_\alpha \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$	$\overline{X} - \overline{Y} \geq \delta + t_{m+n-2;\alpha} S_p \sqrt{\frac{1}{m} + \frac{1}{n}}$
II	$\mu_1 - \mu_2 \geq \delta$	$\mu_1 - \mu_2 < \delta$	$\overline{X} - \overline{Y} \leq \delta + z_{1-\alpha} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$	$\overline{X} - \overline{Y} \leq \delta + t_{m+n-2;1-\alpha} S_p \sqrt{\frac{1}{m} + \frac{1}{n}}$
III	$\mu_1 - \mu_2 = \delta$	$\mu_1 - \mu_2 \neq \delta$	$\mid \overline{X} - \overline{Y} - \delta \mid \geq z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$	$\mid \overline{X} - \overline{Y} - \delta \mid \geq t_{m+n-2;\alpha/2} S_p \sqrt{\frac{1}{m} + \frac{1}{n}}$

■

Note:

(i) In Definition 10.3.2, δ is any fixed constant.

(ii) All tests are UMPU and UMP invariant.

(iii) If $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (which is unknown), then S_p^2 is an unbiased estimate of σ^2. We should check that $\sigma_1^2 = \sigma_2^2$ with an F–test.

(iv) For large $m + n$, we can use z–tables instead of t-tables. Also, for large m and large n we can drop the Normality assumption due to the CLT. However, for small m or small n, none of these simplifications is justified.

■

Definition 10.3.3: **Paired t-Tests**
Let $(X_1, Y_1) \ldots, (X_n, Y_N)$ be a sample from a bivariate $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ distribution where all 5 parameters are unknown.

Let $D_i = X_i - Y_i \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)$.

Let $\overline{D} = \frac{1}{n} \sum\limits_{i=1}^{n} D_i$ and $S_d^2 = \frac{1}{n-1} \sum\limits_{i=1}^{n} (D_i - \overline{D})^2$.

The following table summarizes the t–tests that are typically being used:

	H_0	H_1	Reject H_0 at level α if
I	$\mu_1 - \mu_2 \leq \delta$	$\mu_1 - \mu_2 > \delta$	$\overline{D} \geq \delta + \frac{S_d}{\sqrt{n}} t_{n-1;\alpha}$
II	$\mu_1 - \mu_2 \geq \delta$	$\mu_1 - \mu_2 < \delta$	$\overline{D} \leq \delta + \frac{S_d}{\sqrt{n}} t_{n-1;1-\alpha}$
III	$\mu_1 - \mu_2 = \delta$	$\mu_1 - \mu_2 \neq \delta$	$\mid \overline{D} - \delta \mid \geq \frac{S_d}{\sqrt{n}} t_{n-1;\alpha/2}$

■

Note:

(i) In Definition 10.3.3, δ is any fixed constant.

(ii) These tests are special cases of one–sample tests. All the properties stated in the Note following Definition 10.3.1 hold.

(iii) We could do a test based on Normality assumptions if $\sigma^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$ were known, but that is a very unrealistic assumption.

■

Definition 10.3.4: F**–Tests**

Let $X_1, \ldots, X_m$ be a sample from a $N(\mu_1, \sigma_1^2)$ distribution where μ_1 may be known or unknown and σ_1^2 is unknown. Let $Y_1, \ldots, Y_n$ be a sample from a $N(\mu_2, \sigma_2^2)$ distribution where μ_2 may be known or unknown and σ_2^2 is unknown.

Recall that

$$\frac{\sum\limits_{i=1}^{m}(X_i - \overline{X})^2}{\sigma_1^2} \sim \chi^2_{m-1}, \quad \frac{\sum\limits_{i=1}^{n}(Y_i - \overline{Y})^2}{\sigma_2^2} \sim \chi^2_{n-1},$$

and

$$\frac{\dfrac{\sum\limits_{i=1}^{m}(X_i - \overline{X})^2}{(m-1)\sigma_1^2}}{\dfrac{\sum\limits_{i=1}^{n}(Y_i - \overline{Y})^2}{(n-1)\sigma_2^2}} = \frac{\sigma_2^2}{\sigma_1^2}\frac{S_1^2}{S_2^2} \sim F_{m-1,n-1}.$$

The following table summarizes the F–tests that are typically being used:

			Reject H_0 at level α if	
	H_0	H_1	μ_1, μ_2 known	μ_1, μ_2 unknown
I	$\sigma_1^2 \leq \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	$\frac{\frac{1}{m}\sum(X_i-\mu_1)^2}{\frac{1}{n}\sum(Y_i-\mu_2)^2} \geq F_{m,n;\alpha}$	$\frac{S_1^2}{S_2^2} \geq F_{m-1,n-1;\alpha}$
II	$\sigma_1^2 \geq \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$	$\frac{\frac{1}{n}\sum(Y_i-\mu_2)^2}{\frac{1}{m}\sum(X_i-\mu_1)^2} \geq F_{n,m;\alpha}$	$\frac{S_2^2}{S_1^2} \geq F_{n-1,m-1;\alpha}$
III	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$\frac{\frac{1}{m}\sum(X_i-\mu_1)^2}{\frac{1}{n}\sum(Y_i-\mu_2)^2} \geq F_{m,n;\alpha/2}$ or $\frac{\frac{1}{n}\sum(Y_i-\mu_2)^2}{\frac{1}{m}\sum(X_i-\mu_1)^2} \geq F_{n,m;\alpha/2}$	$\frac{S_1^2}{S_2^2} \geq F_{m-1,n-1;\alpha/2}$ or $\frac{S_2^2}{S_1^2} \geq F_{n-1,m-1;\alpha/2}$

■

Note:

(i) Tests I and II are UMPU and UMP invariant if μ_1 and μ_2 are unknown.

(ii) Test III uses equal tails and therefore may not be unbiased.

(iii) If an F–test (at level α_1) and a t–test (at level α_2) are both performed, the combined test has level $\alpha = 1 - (1 - \alpha_1)(1 - \alpha_2) \geq max(\alpha_1, \alpha_2)$ ($\equiv \alpha_1 + \alpha_2$ if both are small).

■

Lecture 29: Mo 03/26/01

10.4 Bayes and Minimax Tests

Hypothesis testing may be conducted in a decision–theoretic framework. Here our action space $\mathcal{A}$ consists of two options: a_0 = fail to reject H_0 and a_1 = reject H_0.

Usually, we assume no loss for a correct decision. Thus, our loss function looks like:

$$L(\theta, a_0) = \begin{cases} 0, & \text{if } \theta \in \Theta_0 \\ a(\theta), & \text{if } \theta \in \Theta_1 \end{cases}$$

$$L(\theta, a_1) = \begin{cases} b(\theta), & \text{if } \theta \in \Theta_0 \\ 0, & \text{if } \theta \in \Theta_1 \end{cases}$$

We consider the following special cases:

0–1 loss: $a(\theta) = b(\theta) = 1$, i.e., all errors are equally bad.

Generalized 0–1 loss: $a(\theta) = c_{II}$, $b(\theta) = c_I$, i.e., all Type I errors are equally bad and all Type II errors are equally bad and Type I errors are worse than Type II errors or vice versa.

Then, the risk function can be written as

$$\begin{aligned} R(\theta, d(\underline{X})) &= L(\theta, a_0) P_\theta(d(\underline{X}) = a_0) + L(\theta, a_1) P_\theta(d(\underline{X}) = a_1) \\ &= \begin{cases} a(\theta) P_\theta(d(\underline{X}) = a_0), & \text{if } \theta \in \Theta_1 \\ b(\theta) P_\theta(d(\underline{X}) = a_1), & \text{if } \theta \in \Theta_0 \end{cases} \end{aligned}$$

The minimax rule minimizes

$$\max_\theta \{ a(\theta) P_\theta(d(\underline{X}) = a_0),\ b(\theta) P_\theta(d(\underline{X}) = a_1) \}.$$

<u>Theorem 10.4.1:</u>
The minimax rule d for testing

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

under the generalized 0–1 loss function rejects H_0 if

$$\frac{f_{\theta_1}(\underline{x})}{f_{\theta_0}(\underline{x})} \geq k,$$

where k is chosen such that

$$\begin{aligned} R(\theta_0, d(\underline{X})) &= R(\theta_1, d(\underline{X})) \\ \iff c_{II} P_{\theta_1}(d(\underline{X}) = a_0) &= c_I P_{\theta_0}(d(\underline{X}) = a_1) \\ \iff c_{II} P_{\theta_1}\left(\frac{f_{\theta_1}(\underline{X})}{f_{\theta_0}(\underline{X})} < k\right) &= c_I P_{\theta_0}\left(\frac{f_{\theta_1}(\underline{X})}{f_{\theta_0}(\underline{X})} \geq k\right). \end{aligned}$$

<u>Proof:</u>
Let d^* be any other rule.

- If $R(\theta_0, d) < R(\theta_0, d^*)$, then

$$R(\theta_0, d) = R(\theta_1, d) < \max\{R(\theta_0, d^*), R(\theta_1, d^*)\}.$$

So, d^* is not minimax.

- If $R(\theta_0, d) \geq R(\theta_0, d^*)$, i.e.,

$$c_I P_{\theta_0}(d = a_1) = R(\theta_0, d) \geq R(\theta_0, d^*) = c_I P_{\theta_0}(d^* = a_1),$$

then

$$P(\text{reject } H_0 \mid H_0 \text{ true}) = P_{\theta_0}(d = a_1) \geq P_{\theta_0}(d^* = a_1).$$

By the NP Lemma, the rule d is MP of its size. Thus,

$$\begin{aligned} & P_{\theta_1}(d = a_1) \geq P_{\theta_1}(d^* = a_1) \iff P_{\theta_1}(d = a_0) \leq P_{\theta_1}(d^* = a_0) \\ \implies & R(\theta_1, d) \leq R(\theta_1, d^*) \\ \implies & \max\{R(\theta_0, d), R(\theta_1, d)\} = R(\theta_1, d) \leq R(\theta_1, d^*) \leq \max\{R(\theta_0, d^*), R(\theta_1, d^*)\} \ \forall d^* \\ \implies & d \text{ is minimax} \end{aligned}$$

■

<u>Example 10.4.2:</u>
Let $X_1, \ldots, X_n$ be iid $N(\mu, 1)$. Let $H_0: \mu = \mu_0$ vs. $H_1: \mu = \mu_1 > \mu_0$.

As we have seen before, $\dfrac{f_{\theta_1}(\underline{x})}{f_{\theta_0}(\underline{x})} \geq k_1$ is equivalent to $\overline{x} \geq k_2$.

Therefore, we choose k_2 such that

$$\begin{aligned} c_{II} P_{\mu_1}(\overline{X} < k_2) &= c_I P_{\mu_0}(\overline{X} \geq k_2) \\ \iff c_{II} \Phi(\sqrt{n}(k_2 - \mu_1)) &= c_I(1 - \Phi(\sqrt{n}(k_2 - \mu_0))), \end{aligned}$$

where $\Phi(z) = P(Z \leq z)$ for $Z \sim N(0, 1)$.

Given $c_I, c_{II}, \mu_0, \mu_1$, and n, we can solve (numerically) for k_2 using Normal tables. ■

Note:
Now suppose we have a prior distribution $\pi(\theta)$ on Θ. Then the Bayes risk of a decision rule d (under the loss function introduced before) is

$$\begin{aligned}
R(\pi, d) &= E_\pi(R(\theta, d(\underline{X}))) \\
&= \int_\Theta R(\theta, d)\pi(\theta) d\theta \\
&= \int_{\Theta_0} b(\theta)\pi(\theta)P_\theta(d(\underline{X}) = a_1) d\theta + \int_{\Theta_1} a(\theta)\pi(\theta)P_\theta(d(\underline{X}) = a_0) d\theta
\end{aligned}$$

if π is a pdf.

The Bayes risk for a pmf π looks similar (see Rohatgi, page 461). ■

Theorem 10.4.3:
The Bayes rule for testing $H_0: \theta = \theta_0$ vs. $H_1: \theta = \theta_1$ under the prior $\pi(\theta_0) = \pi_0$ and $\pi(\theta_1) = \pi_1 = 1 - \pi_0$ and the generalized 0–1 loss function is to reject H_0 if

$$\frac{f_{\theta_1}(\underline{x})}{f_{\theta_0}(\underline{x})} \geq \frac{c_I \pi_0}{c_{II} \pi_1}.$$

Proof:
We wish to minimize $R(\pi, d)$. We know that

$$\begin{aligned}
R(\pi, d) &\overset{Def.\ 8.8.8}{=} E_\pi(R(\theta, d)) \\
&\overset{Def.\ 8.8.3}{=} E_\pi(E_\theta(L(\theta, d(\underline{X})))) \\
&\overset{Note}{=} \int \underbrace{g(\underline{x})}_{marginal} \left(\int L(\theta, d(\underline{x})) \underbrace{h(\theta \mid \underline{x})}_{posterior} d\theta \right) d\underline{x} \\
&\overset{Def.\ 4.7.1}{=} \int g(\underline{x}) E_\theta(L(\theta, d(\underline{X})) \mid \underline{X} = \underline{x}) \ d\underline{x} \\
&= E_{\underline{X}}(E_\theta(L(\theta, d(\underline{X})) \mid \underline{X})).
\end{aligned}$$

Therefore, it is sufficient to minimize $E_\theta(L(\theta, d(\underline{X})) \mid \underline{X})$.

The a posteriori distribution of θ is

$$\begin{aligned}
h(\theta \mid \underline{x}) &= \frac{\pi(\theta) f_\theta(\underline{x})}{\sum\limits_\theta \pi(\theta) f_\theta(\underline{x})} \\
&= \begin{cases} \dfrac{\pi_0 f_{\theta_0}(\underline{x})}{\pi_0 f_{\theta_0}(\underline{x}) + \pi_1 f_{\theta_1}(\underline{x})}, & \theta = \theta_0 \\ \dfrac{\pi_1 f_{\theta_1}(\underline{x})}{\pi_0 f_{\theta_0}(\underline{x}) + \pi_1 f_{\theta_1}(\underline{x})}, & \theta = \theta_1 \end{cases}
\end{aligned}$$

Therefore,

$$E_\theta(L(\theta, d(\underline{X})) \mid \underline{X} = \underline{x}) = \begin{cases} c_I h(\theta_0 \mid \underline{x}), & \text{if } \theta = \theta_0, d(\underline{x}) = a_1 \\ c_{II} h(\theta_1 \mid \underline{x}), & \text{if } \theta = \theta_1, d(\underline{x}) = a_0 \\ 0, & \text{if } \theta = \theta_0, d(\underline{x}) = a_0 \\ 0, & \text{if } \theta = \theta_1, d(\underline{x}) = a_1 \end{cases}$$

This will be minimized if we reject H_0, i.e., $d(\underline{x}) = a_1$, when $c_I h(\theta_0 \mid \underline{x}) \leq c_{II} h(\theta_1 \mid \underline{x})$

$$\Longrightarrow \quad c_I \pi_0 f_{\theta_0}(\underline{x}) \leq c_{II} \pi_1 f_{\theta_1}(\underline{x})$$

$$\Longrightarrow \quad \frac{f_{\theta_1}(\underline{x})}{f_{\theta_0}(\underline{x})} \geq \frac{c_I \pi_0}{c_{II} \pi_1}$$ ■

Note:
For minimax rules and Bayes rules, the significance level α is no longer predetermined. ■

Example 10.4.4:
Let $X_1, \ldots, X_n$ be iid $N(\mu, 1)$. Let $H_0: \mu = \mu_0$ vs. $H_1: \mu = \mu_1 > \mu_0$. Let $c_I = c_{II}$.

By Theorem 10.4.3, the Bayes rule d rejects H_0 if

$$\begin{aligned}
\frac{f_{\theta_1}(\underline{x})}{f_{\theta_0}(\underline{x})} &\geq \frac{\pi_0}{1 - \pi_0} \\
\Longrightarrow \quad \exp\left(-\frac{\sum(x_i - \mu_1)^2}{2} + \frac{\sum(x_i - \mu_0)^2}{2}\right) &\geq \frac{\pi_0}{1 - \pi_0} \\
\Longrightarrow \quad \exp\left((\mu_1 - \mu_0)\sum x_i + \frac{n(\mu_0^2 - \mu_1^2)}{2}\right) &\geq \frac{\pi_0}{1 - \pi_0} \\
\Longrightarrow \quad (\mu_1 - \mu_0)\sum x_i + \frac{n(\mu_0^2 - \mu_1^2)}{2} &\geq \ln\left(\frac{\pi_0}{1 - \pi_0}\right) \\
\Longrightarrow \quad \frac{1}{n}\sum x_i &\geq \frac{1}{n}\frac{\ln(\frac{\pi_0}{1-\pi_0})}{\mu_1 - \mu_0} + \frac{\mu_0 + \mu_1}{2}
\end{aligned}$$

If $\pi_0 = \frac{1}{2}$, then we reject H_0 if $\overline{x} \geq \dfrac{\mu_0 + \mu_1}{2}$. ■

Note:

We can generalize Theorem 10.4.3 to the case of classifying among k options $\theta_1, \ldots, \theta_k$. If we use the 0–1 loss function

$$L(\theta_i, d) = \begin{cases} 1, & \text{if } d(\underline{X}) = \theta_j \quad \forall j \neq i \\ 0, & \text{if } d(\underline{X}) = \theta_i \end{cases},$$

then the Bayes rule is to pick θ_i if

$$\pi_i f_{\theta_i}(\underline{x}) \geq \pi_j f_{\theta_j}(\underline{x}) \quad \forall j \neq i.$$

■

Lecture 30: We 03/28/01

Example 10.4.5:

Let $X_1, \ldots, X_n$ be iid $N(\mu, 1)$. Let $\mu_1 < \mu_2 < \mu_3$ and let $\pi_1 = \pi_2 = \pi_3$.

Choose $\mu = \mu_i$ if

$$\pi_i \exp\left(-\frac{\sum(x_k - \mu_i)^2}{2}\right) \geq \pi_j \exp\left(-\frac{\sum(x_k - \mu_j)^2}{2}\right), \quad j \neq i, j = 1, 2, 3.$$

Similar to Example 10.4.4, these conditions can be transformed as follows:

$$\overline{x}(\mu_i - \mu_j) \geq \frac{(\mu_i - \mu_j)(\mu_i + \mu_j)}{2}, \quad j \neq i, j = 1, 2, 3.$$

In our particular example, we get the following decision rules:

(i) Choose μ_1 if $\overline{x} \leq \frac{\mu_1+\mu_2}{2}$ (and $\overline{x} \leq \frac{\mu_1+\mu_3}{2}$).

(ii) Choose μ_2 if $\overline{x} \geq \frac{\mu_1+\mu_2}{2}$ and $\overline{x} \leq \frac{\mu_2+\mu_3}{2}$.

(iii) Choose μ_3 if $\overline{x} \geq \frac{\mu_2+\mu_3}{2}$ (and $\overline{x} \geq \frac{\mu_1+\mu_3}{2}$).

Note that in (i) and (iii) the condition in parentheses automatically holds when the other condition holds.

If $\mu_1 = 0,\ \mu_2 = 2$, and $\mu_3 = 4$, we have the decision rules:

(i) Choose μ_1 if $\overline{x} \leq 1$.

(ii) Choose μ_2 if $1 \leq \overline{x} \leq 3$.

(iii) Choose μ_3 if $\overline{x} \geq 3$.

We do not have to worry how to handle the boundary since the probability that the rv will realize on any of the two boundary points is 0. ■

11 Confidence Estimation

11.1 Fundamental Notions

Let X be a rv and a, b be fixed positive numbers, $a < b$. Then

$$\begin{aligned} P(a < X < b) &= P(a < X \text{ and } X < b) \\ &= P(a < X \text{ and } \frac{X}{b} < 1) \\ &= P(a < X \text{ and } \frac{aX}{b} < a) \\ &= P(\frac{aX}{b} < a < X) \end{aligned}$$

The interval $I(X) = (\frac{aX}{b}, X)$ is an example of a **random interval**. $I(X)$ contains the value a with a certain fixed probability.

For example, if $X \sim U(0,1), a = \frac{1}{4}$, and $b = \frac{3}{4}$, then the interval $I(X) = (\frac{X}{3}, X)$ contains $\frac{1}{4}$ with probability $\frac{1}{2}$.

Definition 11.1.1:

Let $P_{\underline{\theta}}, \underline{\theta} \in \Theta \subseteq I\!R^k$, be a set of probability distributions of a rv $\underline{X}$. A family of subsets $S(\underline{x})$ of Θ, where $S(\underline{x})$ depends on $\underline{x}$ but not on θ, is called a **family of random sets**. In particular, if $\theta \in \Theta \subseteq I\!R$ and $S(\underline{x})$ is an interval $(\underline{\theta}(\underline{x}), \overline{\theta}(\underline{x}))$ where $\underline{\theta}(\underline{x})$ and $\overline{\theta}(\underline{x})$ depend on $\underline{x}$ but not on θ, we call $S(\underline{X})$ a **random interval**, with $\underline{\theta}(\underline{X})$ and $\overline{\theta}(\underline{X})$ as lower and upper bounds, respectively. $\underline{\theta}(\underline{X})$ may be $-\infty$ and $\overline{\theta}(\underline{X})$ may be $+\infty$. ■

Note:

Frequently in inference, we are not interested in estimating a parameter or testing a hypothesis about it. Instead, we are interested in establishing a lower or upper bound (or both) for one or multiple parameters. ■

Definition 11.1.2:

A family of subsets $S(\underline{x})$ of $\Theta \subseteq I\!R^k$ is called a **family of confidence sets** at **confidence level** $1 - \alpha$ if

$$P_{\underline{\theta}}(S(\underline{X}) \ni \underline{\theta}) \geq 1 - \alpha \quad \forall \underline{\theta} \in \Theta,$$

where $0 < \alpha < 1$ is usually small.

The quantity

$$\inf_{\underline{\theta}} P_{\underline{\theta}}(S(\underline{X}) \ni \underline{\theta})$$

is called the **confidence coefficient**. ■

Definition 11.1.3:
For $k = 1$, we use the following names for some of the confidence sets defined in Definition 11.1.2:

(i) If $S(\underline{x}) = (\underline{\theta}(\underline{x}), \infty)$, then $\underline{\theta}(\underline{x})$ is called a level $1 - \alpha$ **lower confidence bound**.

(ii) If $S(\underline{x}) = (-\infty, \overline{\theta}(\underline{x}))$, then $\overline{\theta}(\underline{x})$ is called a level $1 - \alpha$ **upper confidence bound**.

(iii) $S(\underline{x}) = (\underline{\theta}(\underline{x}), \overline{\theta}(\underline{x}))$ is called a level $1 - \alpha$ **confidence interval** (**CI**).

■

Definition 11.1.4:
A family of $1 - \alpha$ level confidence sets $\{S(\underline{x})\}$ is called **uniformly most accurate** (**UMA**) if

$$P_{\underline{\theta}'}(S(\underline{X}) \ni \underline{\theta}) \leq P_{\underline{\theta}'}(S'(\underline{X}) \ni \underline{\theta}) \quad \forall \underline{\theta}, \underline{\theta}' \in \Theta$$

and for any $1 - \alpha$ level family of confidence sets $S'(\underline{X})$. ■

Theorem 11.1.5:
Let $X_1, \ldots, X_n \sim F_\theta$, $\theta \in \Theta$, where Θ is an interval on $I\!R$. Let $T(\underline{X}, \theta)$ be a function on $I\!R^n \times \Theta$ such that for each θ, $T(\underline{X}, \theta)$ is a statistic, and as a function of θ, T is strictly monotone (either increasing or decreasing) in θ at every value of $\underline{x} \in I\!R^n$.

Let $\Lambda \subseteq I\!R$ be the range of T and let the equation $\lambda = T(\underline{x}, \theta)$ be solvable for θ for every $\lambda \in \Lambda$ and every $\underline{x} \in I\!R^n$.

If the distribution of $T(\underline{X}, \theta)$ is independent of θ, then we can construct a confidence interval for θ at any level.

Proof:
Choose α such that $0 < \alpha < 1$. Then we can choose $\lambda_1(\alpha) < \lambda_2(\alpha)$ (which may not necessarily be unique) such that

$$P_\theta(\lambda_1(\alpha) < T(\underline{X}, \theta) < \lambda_2(\alpha)) \geq 1 - \alpha \quad \forall \theta.$$

Since the distribution of $T(\underline{X}, \theta)$ is independent of θ, $\lambda_1(\alpha)$ and $\lambda_2(\alpha)$ also do not depend on θ.

If $T(\underline{X}, \theta)$ is increasing in θ, solve the equations $\lambda_1(\alpha) = T(\underline{X}, \theta)$ for $\underline{\theta}(\underline{X})$ and $\lambda_2(\alpha) = T(\underline{X}, \theta)$ for $\overline{\theta}(\underline{X})$.

If $T(\underline{X}, \theta)$ is decreasing in θ, solve the equations $\lambda_1(\alpha) = T(\underline{X}, \theta)$ for $\overline{\theta}(\underline{X})$ and $\lambda_2(\alpha) = T(\underline{X}, \theta)$ for $\underline{\theta}(\underline{X})$.

In either case, it holds that

$$P_\theta(\underline{\theta}(\underline{X}) < \theta < \overline{\theta}(\underline{X})) \geq 1 - \alpha \quad \forall \theta.$$

■

Note:

(i) Solvability is guaranteed if T is continuous and strictly monotone as a function of θ.

(ii) If T is not monotone, we can still use this Theorem to get confidence sets that may not be confidence intervals.

■

Example 11.1.6:
Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, where μ and $\sigma^2 > 0$ are both unknown. We seek a $1 - \alpha$ level confidence interval for μ.

Note that

$$T(\underline{X}, \mu) = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

is independent of μ and monotone and decreasing in μ.

We choose $\lambda_1(\alpha)$ and $\lambda_2(\alpha)$ such that

$$P(\lambda_1(\alpha) < T(\underline{X}, \mu) < \lambda_2(\alpha)) = 1 - \alpha$$

and solve for μ which yields

$$P(\overline{X} - \frac{S\lambda_2(\alpha)}{\sqrt{n}} < \mu < \overline{X} - \frac{S\lambda_1(\alpha)}{\sqrt{n}}) = 1 - \alpha.$$

Thus,

$$(\overline{X} - \frac{S\lambda_2(\alpha)}{\sqrt{n}}, \overline{X} - \frac{S\lambda_1(\alpha)}{\sqrt{n}})$$

is a $1 - \alpha$ level CI for μ. We commonly choose $\lambda_2(\alpha) = -\lambda_1(\alpha) = t_{n-1;\alpha/2}$. ■

Example 11.1.7:
Let $X_1, \ldots, X_n \sim U(0, \theta)$.

We know that $\hat{\theta} = \max(X_i) = Max_n$ is the MLE for θ and sufficient for θ.

The pdf of Max_n is given by

$$f_n(y) = \frac{ny^{n-1}}{\theta^n} I_{(0,\theta)}(y).$$

Then the rv $T_n = \frac{Max_n}{\theta}$ has the pdf

$$h_n(t) = nt^{n-1} I_{(0,1)}(t),$$

which is independent of θ. T_n is monotone and decreasing in θ.

We now have to find numbers $\lambda_1(\alpha)$ and $\lambda_2(\alpha)$ such that

$$\begin{aligned} P(\lambda_1(\alpha) < T_n < \lambda_2(\alpha)) &= 1 - \alpha \\ \Longrightarrow \quad n \int_{\lambda_1}^{\lambda_2} t^{n-1} dt &= 1 - \alpha \\ \Longrightarrow \quad \lambda_2^n - \lambda_1^n &= 1 - \alpha \end{aligned}$$

If we choose $\lambda_2 = 1$ and $\lambda_1 = \alpha^{1/n}$, then $(Max_n, \alpha^{-1/n} Max_n)$ is a $1 - \alpha$ level CI for θ. This holds since

$$\begin{aligned} P(\alpha^{1/n} < \frac{Max_n}{\theta} < 1) &= P(\alpha^{-1/n} > \frac{\theta}{Max_n} > 1) \\ &= P(\alpha^{-1/n} Max_n > \theta > Max_n) \\ &= 1 - \alpha. \end{aligned}$$

■

Lecture 31:
Fr 03/30/01

11.2 Shortest–Length Confidence Intervals

In practice, we usually want not only an interval with coverage probability $1-\alpha$ for θ, but if possible the shortest (most precise) such interval.

Definition 11.2.1:
A rv $T(\underline{X}, \theta)$ whose distribution is independent of θ is called a **pivot**. ■

Note:
The methods we will discuss here can provide the shortest interval based on a given pivot. They will not guarantee that there is no other pivot with a shorter minimal interval. ■

Example 11.2.2:
Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, where $\sigma^2 > 0$ is known. The obvious pivot for μ is

$$T_\mu(\underline{X}) = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Suppose that (a, b) is an interval such that $P(a < Z < b) = 1 - \alpha$, where $Z \sim N(0,1)$.

A $1-\alpha$ level CI based on this pivot is found by

$$1 - \alpha = P\left(a < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < b\right) = P\left(\overline{X} - b\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} - a\frac{\sigma}{\sqrt{n}}\right).$$

The length of the interval is $L = (b-a)\frac{\sigma}{\sqrt{n}}$.

To minimize L, we must choose a and b such that $b - a$ is minimal while

$$\Phi(b) - \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx = 1 - \alpha,$$

where $\Phi(z) = P(Z \leq z)$.

To find a minimum, we can differentiate these expressions with respect to a. However, b is not a constant but is an implicit function of a. Formally, we could write $\frac{d\, b(a)}{da}$. However, this is usually shortened to $\frac{db}{da}$.

Here we get

$$\frac{d}{da}(\Phi(b) - \Phi(a)) = \phi(b)\frac{db}{da} - \phi(a) = \frac{d}{da}(1-\alpha) = 0$$

and

$$\frac{dL}{da} = \frac{\sigma}{\sqrt{n}}(\frac{db}{da} - 1) = \frac{\sigma}{\sqrt{n}}(\frac{\phi(a)}{\phi(b)} - 1).$$

The minimum occurs when $\phi(a) = \phi(b)$ which happens when $a = b$ or $a = -b$. If we select $a = b$, then $\Phi(b) - \Phi(a) = \Phi(a) - \Phi(a) = 0 \neq 1 - \alpha$. Thus, we must have that $b = -a = z_{\alpha/2}$. Thus, the shortest CI based on T_μ is

$$(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}).$$

■

Definition 11.2.3:

A pdf $f(x)$ is **unimodal** iff there exists a x^* such that $f(x)$ is nondecreasing for $x \leq x^*$ and $f(x)$ is nonincreasing for $x \geq x^*$. ■

Theorem 11.2.4:

Let $f(x)$ be a unimodal pdf. If the interval $[a, b]$ satisfies

(i) $\int_a^b f(x)dx = 1 - \alpha$

(ii) $f(a) = f(b) > 0$, and

(iii) $a \leq x^* \leq b$, where x^* is a mode of $f(x)$,

then the interval $[a, b]$ is the shortest of all intervals which satisfy condition (i).

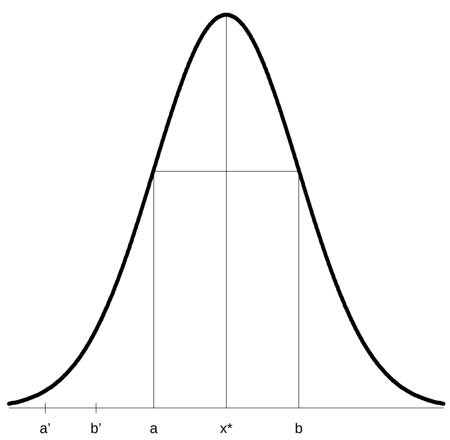
Proof:

Let $[a', b']$ be any interval with $b' - a' < b - a$. We will show that this implies $\int_a^b f(x)dx < 1 - \alpha$, i.e., a contradiction.

We assume that $a' \leq a$. The case $a < a'$ is similar.

• Suppose that $b' \leq a$. Then $a' \leq b' \leq a \leq x^*$.

Theorem 11.2.4a

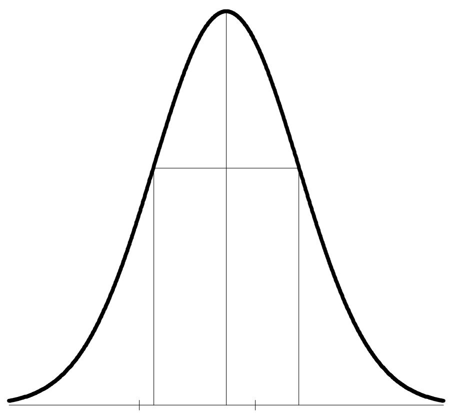


It follows

$$
\begin{aligned}
\int_{a'}^{b'} f(x)dx &\leq f(b')(b'-a') && |x \leq b' \leq x^* \Rightarrow f(x) \leq f(b') \\
&\leq f(a)(b'-a') && |b' \leq a \leq x^* \Rightarrow f(b') \leq f(a) \\
&< f(a)(b-a) && |b'-a' < b-a \text{ and } f(a) > 0 \\
&\leq \int_a^b f(x)dx && |f(x) \geq f(a) \text{ for } a \leq x \leq b \\
&= 1-\alpha && |\text{by (i)}
\end{aligned}
$$

• Suppose $b' > a$. We can immediately exclude that $b' > b$ since then $b'-a' > b-a$, i.e., $b'-a'$ wouldn't be of shorter length than $b-a$. Thus, we have to consider the case that $a' \leq a < b' < b$.

Theorem 11.2.4b



It holds that

$$\int_{a'}^{b'} f(x)dx = \int_a^b f(x)dx + \int_{a'}^{a} f(x)dx - \int_{b'}^{b} f(x)dx$$

Note that $\int_{a'}^{a} f(x)dx \leq f(a)(a-a')$ and $\int_{b'}^{b} f(x)dx \geq f(b)(b-b')$. Therefore, we get

$$
\begin{aligned}
\int_{a'}^{a} f(x)dx - \int_{b'}^{b} f(x)dx &\leq f(a)(a-a') - f(b)(b-b') \\
&= f(a)((a-a')-(b-b')) \quad \text{since } f(a) = f(b) \\
&= f(a)((b'-a')-(b-a)) \\
&< 0
\end{aligned}
$$

Thus,

$$\int_{a'}^{b'} f(x)dx < 1-\alpha.$$

■

Note:

Example 11.2.2 is a special case of Theorem 11.2.4. ■

Example 11.2.5:

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, where μ is known. The obvious pivot for σ^2 is

$$T_{\sigma^2}(\underline{X}) = \frac{\sum(X_i - \mu)^2}{\sigma^2} \sim \chi^2_n.$$

So

$$\begin{aligned} P\left(a < \frac{\sum(X_i - \mu)^2}{\sigma^2} < b\right) &= 1 - \alpha \\ \Longleftrightarrow \quad P\left(\frac{\sum(X_i - \mu)^2}{b} < \sigma^2 < \frac{\sum(X_i - \mu)^2}{a}\right) &= 1 - \alpha \end{aligned}$$

We wish to minimize

$$L = (\frac{1}{a} - \frac{1}{b}) \sum(X_i - \mu)^2$$

such that $\int_a^b f_n(t)dt = 1 - \alpha$, where $f_n(t)$ is the pdf of a χ^2_n distribution.

We get

$$f_n(b)\frac{db}{da} - f_n(a) = 0$$

and

$$\frac{dL}{da} = \left(\frac{1}{a^2} - \frac{1}{b^2}\frac{db}{da}\right) \sum(X_i - \mu)^2 = \left(\frac{1}{a^2} - \frac{1}{b^2}\frac{f_n(a)}{f_n(b)}\right) \sum(X_i - \mu)^2.$$

We obtain a minimum if $a^2 f_n(a) = b^2 f_n(b)$.

Note that in practice equal tails $\chi^2_{n;\alpha/2}$ and $\chi^2_{n;1-\alpha/2}$ are used, which do not result in shortest–length CI's. The reason for this selection is simple: When these tests were developed, computers did not exist that could solve these equations numerically. People in general had to rely on tabulated values. Manually solving the equation above for each case obviously wasn't a feasible solution. ■

Example 11.2.6:

Let $X_1, \ldots, X_n \sim U(0, \theta)$. Let $Max_n = \max X_i = X_{(n)}$. Since $T_n = \frac{Max_n}{\theta}$ has pdf $nt^{n-1}I_{(0,1)}(t)$ which does not depend on θ, T_n can be selected as a our pivot. The density of T_n is strictly increasing for $n \geq 2$, so we cannot find constants a and b as in Example 11.2.5.

If $P(a < T_n < b) = 1 - \alpha$, then $P(\frac{Max_n}{b} < \theta < \frac{Max_n}{a}) = 1 - \alpha$.

We wish to minimize

$$L = Max_n(\frac{1}{a} - \frac{1}{b})$$

such that $\int_a^b nt^{n-1}dt = b^n - a^n = 1 - \alpha$.

We get

$$nb^{n-1} - na^{n-1}\frac{da}{db} = 0 \implies \frac{da}{db} = \frac{b^{n-1}}{a^{n-1}}$$

and

$$\frac{dL}{db} = Max_n(-\frac{1}{a^2}\frac{da}{db} + \frac{1}{b^2}) = Max_n(-\frac{b^{n-1}}{a^{n+1}} + \frac{1}{b^2}) = Max_n(\frac{a^{n+1} - b^{n+1}}{b^2 a^{n+1}}) < 0 \quad \text{for } 0 \leq a < b \leq 1.$$

Thus, L does not have a local minimum. However, since $\frac{dL}{db} < 0$, L is strictly decreasing as a function of b. It is minimized when $b = 1$, i.e., when b is as large as possible. The corresponding a is selected as $a = \alpha^{1/n}$.

The shortest $1 - \alpha$ level CI based on T_n is $(Max_n, \alpha^{-1/n} Max_n)$. ■

Lecture 32: Mo 04/02/01

11.3 Confidence Intervals and Hypothesis Tests

Example 11.3.1:
Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, where $\sigma^2 > 0$ is known. In Example 11.2.2 we have shown that the interval

$$(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

is a $1-\alpha$ level CI for μ.

Suppose we define a test ϕ of $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ that rejects H_0 iff μ_0 does not fall in this interval. Then,

$$\begin{aligned} P_{\mu_0}(\text{Type I error}) &= P_{\mu_0}(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \geq \mu_0) + P_{\mu_0}(\overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0) \\ &= P_{\mu_0}\left(\frac{\mid \overline{X} - \mu_0 \mid}{\frac{\sigma}{\sqrt{n}}} \geq z_{\alpha/2}\right) = \alpha, \end{aligned}$$

i.e., ϕ has size α.

Conversely, if $\phi(\underline{x}, \mu_0)$ is a family of size α tests of $H_0 : \mu = \mu_0$, the set $\{\mu_0 \mid \phi(\underline{x}, \mu_0) \text{ fails to reject } H_0\}$ is a level $1-\alpha$ confidence set for μ_0. ■

Theorem 11.3.2:
Denote $H_0(\theta_0)$ for $H_0 : \theta = \theta_0$, and $H_1(\theta_0)$ for the alternative. Let $A(\theta_0), \theta_0 \in \Theta$, denote the acceptance region of a level–α test of $H_0(\theta_0)$. For each possible observation $\underline{x}$, define

$$S(\underline{x}) = \{\theta : x \in A(\theta), \theta \in \Theta\}.$$

Then $S(\underline{x})$ is a family of $1-\alpha$ level confidence sets for θ.

If, moreover, $A(\theta_0)$ is UMP for $(\alpha, H_0(\theta_0), H_1(\theta_0))$, then $S(\underline{x})$ minimizes $P_\theta(S(\underline{X}) \ni \theta')$ $\forall \theta \in H_1(\theta')$ among all $1-\alpha$ level families of confidence sets, i.e., $S(\underline{x})$ is UMA.

Proof:
It holds that $S(\underline{x}) \ni \theta$ iff $\underline{x} \in A(\theta)$. Therefore,

$$P_\theta(S(\underline{X}) \ni \theta) = P_\theta(\underline{X} \in A(\theta)) \geq 1 - \alpha.$$

Let $S^*(\underline{X})$ be any other family of $1-\alpha$ level confidence sets. Define $A^*(\theta) = \{\underline{x} : S^*(\underline{x}) \ni \theta\}$. Then,

$$P_\theta(\underline{X} \in A^*(\theta)) = P_\theta(S^*(\underline{X}) \ni \theta) \geq 1 - \alpha.$$

Since $A(\theta_0)$ is UMP, it holds that

$$P_\theta(\underline{X} \in A^*(\theta_0)) \geq P_\theta(\underline{X} \in A(\theta_0)) \quad \forall \theta \in H_1(\theta_0).$$

This implies that

$$P_\theta(S^*(\underline{X}) \ni \theta_0) \geq P_\theta(\underline{X} \in A(\theta_0)) = P_\theta(S(\underline{X}) \ni \theta_0) \quad \forall \theta \in H_1(\theta_0).$$

■

Example 11.3.3:

Let X be a rv that belongs to a one–parameter exponential family with pdf $f_\theta(x) = \exp(Q(\theta)T(x) + S'(x) + D(\theta))$, where $Q(\theta)$ is non–decreasing. We consider a test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta < \theta_0$. The acceptance region of a UMP size α test of H_0 has the form $A(\theta_0) = \{x : T(x) > c(\theta_0)\}$.

For $\theta \geq \theta'$,

$$P_{\theta'}(T(X) \leq c(\theta')) = \alpha = P_\theta(T(X) \leq c(\theta)) \leq P_{\theta'}(T(X) \leq c(\theta))$$

(since for a UMP test, it holds power $\geq$ size). Therefore, we can choose $c(\theta)$ as non–decreasing.

A level $1 - \alpha$ CI for θ is then

$$S(x) = \{\theta : x \in A(\theta)\} = (-\infty, c^{-1}(T(x))),$$

where $c^{-1}(T(x)) = \sup_\theta \{\theta : c(\theta) \leq T(x)\}$. ■

Example 11.3.4:

Let $X \sim Exp(\theta)$ with $f_\theta(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}} I_{(0,\infty)}(x)$. Then $Q(\theta) = -\frac{1}{\theta}$ and $T(x) = x$.

We want to test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta < \theta_0$.

The acceptance region of a UMP size α test of H_0 has the form $A(\theta_0) = \{x : x \geq c(\theta_0)\}$, where

$$\alpha = \int_0^{c(\theta_0)} f_{\theta_0}(x) dx = \int_0^{c(\theta_0)} \frac{1}{\theta_0} e^{-\frac{x}{\theta_0}} dx = 1 - e^{-\frac{c(\theta_0)}{\theta_0}}.$$

Thus,

$$e^{-\frac{c(\theta_0)}{\theta_0}} = 1 - \alpha \implies c(\theta_0) = \theta_0 \log(\frac{1}{1-\alpha})$$

Therefore, the UMA family of $1 - \alpha$ level confidence sets is of the form

$$\begin{aligned} S(x) &= \{\theta : x \in A(\theta)\} \\ &= \{\theta : \theta \leq \frac{x}{\log(\frac{1}{1-\alpha})}\} \\ &= \left[0, \frac{x}{\log(\frac{1}{1-\alpha})}\right]. \end{aligned}$$

■

Note:

Just as we frequently restrict the class of tests (when UMP tests don't exist), we can make the same sorts of restrictions on CI's. ■

Definition 11.3.5:

A family $S(\underline{x})$ of confidence sets for parameter θ is said to be **unbiased** at level $1-\alpha$ if

$$P_\theta(S(\underline{X}) \ni \theta) \geq 1-\alpha \text{ and } P_{\theta'}(S(\underline{X}) \ni \theta) \leq 1-\alpha \quad \forall \theta, \theta' \in \Theta.$$

If $S(\underline{x})$ is unbiased and minimizes $P_{\theta'}(S(\underline{X}) \ni \theta)$ among all unbiased CI's at level $1-\alpha$, it is called **uniformly most accurate unbiased** (**UMAU**). ■

Theorem 11.3.6:

Let $A(\theta_0)$ be the acceptance region of a UMPU size α test of $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ (for all θ_0). Then $S(\underline{x}) = \{\theta : \underline{x} \in A(\theta)\}$ is a UMAU family of confidence sets at level $1-\alpha$.

Proof:

Since $A(\theta)$ is unbiased, it holds that

$$P_{\theta'}(S(\underline{X}) \ni \theta) = P_{\theta'}(\underline{X} \in A(\theta)) \leq 1-\alpha.$$

Thus, S is unbiased.

Let $S^*(\underline{x})$ be any other unbiased family of level $1-\alpha$ confidence sets, where

$$A^*(\theta) = \{\underline{x} : S^*(\underline{x}) \ni \theta\}.$$

It holds that

$$P_{\theta'}(\underline{X} \in A^*(\theta)) = P_{\theta'}(S^*(\underline{X}) \ni \theta) \leq 1-\alpha.$$

Therefore, $A^*(\theta)$ is the acceptance region of an unbiased size α test. Thus,

$$\begin{aligned} P_{\theta'}(S^*(\underline{X}) \ni \theta) &= P_{\theta'}(\underline{X} \in A^*(\theta)) \\ &\overset{A \text{ UMPU}}{\geq} P_{\theta'}(\underline{X} \in A(\theta)) \\ &= P_{\theta'}(S(\underline{X}) \ni \theta) \end{aligned}$$

■

Lecture 33: We 04/04/01

Theorem 11.3.7:

Let Θ be an interval on $I\!R$ and f_θ be the pdf of $\underline{X}$. Let $S(\underline{X})$ be a family of $1-\alpha$ level CI's, where $S(\underline{X}) = (\underline{\theta}(\underline{X}), \overline{\theta}(\underline{X}))$, $\underline{\theta}$ and $\overline{\theta}$ increasing functions of $\underline{X}$, and $\overline{\theta}(\underline{X}) - \underline{\theta}(\underline{X})$ is a finite rv.

Then it holds that

$$E_\theta(\overline{\theta}(\underline{X}) - \underline{\theta}(\underline{X}))) = \int (\overline{\theta}(\underline{x}) - \underline{\theta}(\underline{x})) f_\theta(\underline{x}) d\underline{x} = \int_{\theta' \neq \theta} P_\theta(S(\underline{X}) \ni \theta') \; d\theta' \quad \forall \theta \in \Theta.$$

Proof:

It holds that $\overline{\theta} - \underline{\theta} = \int_{\underline{\theta}}^{\overline{\theta}} d\theta'$. Thus, for all $\theta \in \Theta$,

$$\begin{aligned}
E_\theta(\overline{\theta}(\underline{X}) - \underline{\theta}(\underline{X}))) &= \int_{I\!R^n} (\overline{\theta}(\underline{x}) - \underline{\theta}(\underline{x})) f_\theta(\underline{x}) d\underline{x} \\
&= \int_{I\!R^n} \left(\int_{\underline{\theta}(\underline{x})}^{\overline{\theta}(\underline{x})} d\theta' \right) f_\theta(\underline{x}) d\underline{x} \\
&= \int_{I\!R} \left(\int_{\underbrace{\underline{\theta}^{-1}(\theta')}_{\in I\!R^n}}^{\overbrace{\overline{\theta}^{-1}(\theta')}^{\in I\!R^n}} f_\theta(\underline{x}) d\underline{x} \right) d\theta' \\
&= \int_{I\!R} P_\theta \left(\underline{X} \in [\underline{\theta}^{-1}(\theta'), \overline{\theta}^{-1}(\theta')] \right) \; d\theta' \\
&= \int_{I\!R} P_\theta(S(\underline{X}) \ni \theta') \; d\theta' \\
&= \int_{\theta' \neq \theta} P_\theta(S(\underline{X}) \ni \theta') \; d\theta'
\end{aligned}$$

■

Note:

Theorem 11.3.7 says that the expected length of the CI is the probability that $S(\underline{X})$ includes the false θ, averaged over all false values of θ. ■

Corollary 11.3.8:

If $S(\underline{X})$ is UMAU, then $E_\theta(\overline{\theta}(\underline{X}) - \underline{\theta}(\underline{X}))$ is minimized among all unbiased families of CI's.

Proof:

In Theorem 11.3.7 we have shown that

$$E_\theta(\overline{\theta}(\underline{X}) - \underline{\theta}(\underline{X})) = \int_{\theta' \neq \theta} P_\theta(S(\underline{X}) \ni \theta') \; d\theta'.$$

Since a UMAU CI minimizes this probability for all θ', the entire integral is minimized. ■

Example 11.3.9:
Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, where $\sigma^2 > 0$ is known.

By Example 11.2.2, $(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$ is the shortest $1-\alpha$ level CI for μ.

By Example 9.4.3, the equivalent test is UMPU. So by Theorem 11.3.6 this interval is UMAU and by Corollary 11.3.8 it has shortest expected length as well. ■

Example 11.3.10:
Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, where μ and $\sigma^2 > 0$ are both unknown.

Note that
$$T(\underline{X}, \sigma^2) = \frac{(n-1)S^2}{\sigma^2} = T_\sigma \sim \chi^2_{n-1}.$$

Thus,
$$P_{\sigma^2}\left(\lambda_1 < \frac{(n-1)S^2}{\sigma^2} < \lambda_2\right) = 1-\alpha \iff P_{\sigma^2}\left(\frac{(n-1)S^2}{\lambda_2} < \sigma^2 < \frac{(n-1)S^2}{\lambda_1}\right) = 1-\alpha.$$

We now define $P(\gamma)$ as
$$P_{\sigma^2}\left(\frac{(n-1)S^2}{\lambda_2} < \sigma'^2 < \frac{(n-1)S^2}{\lambda_1}\right) = P\left(\frac{T_\sigma}{\lambda_2} < \gamma < \frac{T_\sigma}{\lambda_1}\right) = P(\lambda_1\gamma < T_\sigma < \lambda_2\gamma) = P(\gamma),$$
where $\gamma = \frac{\sigma'^2}{\sigma^2}$.

If our test is unbiased, then it follows from Definition 11.3.5 that
$$P(1) = 1-\alpha \text{ and } P(\gamma) < 1-\alpha \quad \forall \gamma < 1.$$

This implies that we can find λ_1, λ_2 such that $P(1) = 1-\alpha$ and
$$\left.\frac{dP(\gamma)}{d\gamma}\right|_{\gamma=1} = \lambda_2 f_{T_\sigma}(\lambda_2) - \lambda_1 f_{T_\sigma}(\lambda_1) = 0,$$
where f_{T_σ} is the pdf that relates to T_σ, i.e., a the pdf of a χ^2_{n-1} distribution. We can solve for λ_1, λ_2 numerically. Then,
$$\left(\frac{(n-1)S^2}{\lambda_2}, \frac{(n-1)S^2}{\lambda_1}\right)$$
is an unbiased $1-\alpha$ level CI for σ^2.

Rohatgi, Theorem 4(b), page 428–429, states that the related test is UMPU. Therefore, by Theorem 11.3.6 and Corollary 11.3.8, our CI is UMAU with shortest expected length among all unbiased intervals.

Note that this CI is different from the equal–tail CI based on Definition 10.2.1, III, and from the shortest–length CI obtained in Example 11.2.5. ■

11.4 Bayes Confidence Intervals

Definition 11.4.1:

Given a posterior distribution $h(\theta \mid \underline{x})$, a level $1-\alpha$ **credible set** (**Bayesian confidence set**) is any set A such that

$$P(\theta \in A \mid \underline{x}) = \int_A h(\theta \mid \underline{x}) d\theta = 1 - \alpha.$$

■

Note:

If A is an interval, we speak of a **Bayesian confidence interval**. ■

Example 11.4.2:

Let $X \sim Bin(n,p)$ and $\pi(p) \sim U(0,1)$.

In Example 8.8.11, we have shown that

$$\begin{aligned} h(p \mid x) &= \frac{p^x(1-p)^{n-x}}{\displaystyle\int_0^1 p^x(1-p)^{n-x}dp} I_{(0,1)}(p) \\ &= B(x+1, n-x+1)^{-1} p^x (1-p)^{n-x} I_{(0,1)}(p) \\ &= \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} p^x(1-p)^{n-x} I_{(0,1)}(p) \\ &\sim Beta(x+1, n-x+1), \end{aligned}$$

where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the beta function evaluated for a and b and $Beta(x+1, n-x+1)$ represents a Beta distribution with parameters $x+1$ and $n-x+1$.

Using the observed value for x and tables for incomplete beta integrals or a numerical approach, we can find λ_1 and λ_2 such that $P_{p|x}(\lambda_1 < p < \lambda_2) = 1-\alpha$. So (λ_1, λ_2) is a credible interval for p. ■

Note:

(i) The definitions and interpretations of credible intervals and confidence intervals are quite different. Therefore, very different intervals may result.

(ii) We can often use Theorem 11.2.4 to find the shortest credible interval.

■

Lecture 34: Mo 04/09/01

Example 11.4.3:

Let $X_1, \ldots, X_n$ be iid $N(\mu, 1)$ and $\pi(\mu) \sim N(0, 1)$. We want to construct a Bayesian level $1 - \alpha$ CI for μ.

By Definition 8.8.7, the posterior distribution of μ given $\underline{x}$ is

$$h(\mu \mid \underline{x}) = \frac{\pi(\mu) f(\underline{x} \mid \mu)}{g(\underline{x})}$$

where

$$\begin{aligned}
g(\underline{x}) &= \int_{-\infty}^{\infty} f(\underline{x}, \mu) d\mu \\
&= \int_{-\infty}^{\infty} \pi(\mu) f(\underline{x} \mid \mu) d\mu \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}\mu^2) \frac{1}{(\sqrt{2\pi})^n} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) d\mu \\
&= \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{\frac{n+1}{2}}} \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2\right) - \frac{1}{2}\mu^2\right) d\mu \\
&= \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{\frac{n+1}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2 + \mu n\overline{x} - \frac{n}{2}\mu^2 - \frac{1}{2}\mu^2\right) d\mu \\
&= \frac{1}{(2\pi)^{\frac{n+1}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right) \int_{-\infty}^{\infty} \exp\left(-\frac{n+1}{2}\left(\mu^2 - 2\mu\frac{n\overline{x}}{n+1}\right)\right) d\mu \\
&= \frac{1}{(2\pi)^{\frac{n+1}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right) \int_{-\infty}^{\infty} \exp\left(-\frac{n+1}{2}\left(\mu - \frac{n\overline{x}}{n+1}\right)^2 + \frac{n+1}{2}\left(\frac{n\overline{x}}{n+1}\right)^2\right) d\mu \\
&= \frac{1}{(2\pi)^{\frac{n+1}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2 + \frac{n^2\overline{x}^2}{2(n+1)}\right) \sqrt{2\pi\frac{1}{n+1}} \cdot \\
&\qquad \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\frac{1}{n+1}}} \exp\left(-\frac{1}{2}\frac{1}{1/(n+1)}\left(\mu - \frac{n\overline{x}}{n+1}\right)^2\right) d\mu}_{=1 \text{ since pdf of a } N(\frac{n\overline{x}}{n+1}, \frac{1}{n+1}) \text{ distribution}} \\
&= \frac{(n+1)^{-\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2 + \frac{n^2\overline{x}^2}{2(n+1)}\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}
h(\mu \mid \underline{x}) &= \frac{\pi(\mu) f(\underline{x} \mid \mu)}{g(\underline{x})} \\
&= \frac{\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}\mu^2) \frac{1}{(\sqrt{2\pi})^n} \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - \mu)^2\right)}{\frac{(n+1)^{-\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^2 + \frac{n^2\overline{x}^2}{2(n+1)}\right)} \\
&= \frac{\sqrt{n+1}}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^2 + \mu n\overline{x} - \frac{n}{2}\mu^2 - \frac{1}{2}\mu^2 + \frac{1}{2}\sum_{i=1}^n x_i^2 - \frac{n^2\overline{x}^2}{2(n+1)}\right) \\
&= \frac{\sqrt{n+1}}{\sqrt{2\pi}} \exp\left(-\frac{n+1}{2}\left(\mu^2 - \frac{2}{n+1}\mu n\overline{x} + \frac{2}{n+1}\frac{n^2\overline{x}^2}{2(n+1)}\right)\right) \\
&= \frac{1}{\sqrt{2\pi\frac{1}{n+1}}} \exp\left(-\frac{1}{2}\frac{1}{\frac{1}{n+1}}\left(\mu - \frac{n\overline{x}}{n+1}\right)^2\right),
\end{aligned}$$

i.e.,

$$h(\mu \mid \underline{x}) \sim N\left(\frac{n\overline{x}}{n+1}, \frac{1}{n+1}\right).$$

Therefore, a Bayesian level $1-\alpha$ CI for μ is

$$\left(\frac{n}{n+1}\overline{X} - \frac{z_{\alpha/2}}{\sqrt{n+1}},\ \frac{n}{n+1}\overline{X} + \frac{z_{\alpha/2}}{\sqrt{n+1}}\right).$$

The shortest ("classical") level $1-\alpha$ CI for μ (treating μ as fixed) is

$$\left(\overline{X} - \frac{z_{\alpha/2}}{\sqrt{n}},\ \overline{X} + \frac{z_{\alpha/2}}{\sqrt{n}}\right)$$

as seen in Example 11.3.9.

Thus, the Bayesian CI is slightly shorter than the classical CI since we use additional information in constructing the Bayesian CI. ■

12 Nonparametric Inference

12.1 Nonparametric Estimation

Definition 12.1.1:
A statistical method which does not rely on assumptions about the distributional form of a rv (except, perhaps, that it is absolutely continuous, or purely discrete) is called a **nonparametric** or **distribution–free** method. ■

Note:
Unless otherwise specified, we make the following assumptions for the remainder of this chapter: Let $X_1, \ldots, X_n$ be iid $\sim F$, where F is unknown. Let $\mathcal{P}$ be the class of all possible distributions of X. ■

Definition 12.1.2:
A statistic $T(\underline{X})$ is **sufficient** for a family of distributions $\mathcal{P}$ if the conditional distibution of $\underline{X}$ given $T = t$ is the same for all $F \in \mathcal{P}$. ■

Example 12.1.3:
Let $X_1, \ldots, X_n$ be absolutely continuous. Let $\underline{T} = (X_{(1)}, \ldots, X_{(n)})$ be the order statistics.

It holds that

$$f(\underline{x} \mid \underline{T} = \underline{t}) = \frac{1}{n!},$$

so $\underline{T}$ is sufficient for the family of absolutely continuous distributions on $I\!R$. ■

Definition 12.1.4:
A family of distributions $\mathcal{P}$ is **complete** if the only unbiased estimate of 0 is the 0 itself, i.e.,

$$E_F(h(\underline{X})) = 0 \;\; \forall F \in \mathcal{P} \;\; \Longrightarrow h(\underline{x}) = 0 \;\; \forall \underline{x}.$$

■

Definition 12.1.5:
A statistic $T(\underline{X})$ is **complete in relation to** $\mathcal{P}$ if the class of induced distributions of T is complete. ■

Theorem 12.1.6:
The order statistic $(X_{(1)}, \ldots, X_{(n)})$ is a complete sufficient statistic, provided that $X_1, \ldots, X_n$ are of either (pure) discrete of (pure) continuous type. ■

Definition 12.1.7:

A parameter $g(F)$ is called **estimable** if it has an unbiased estimate, i.e., if there exists a $T(\underline{X})$ such that

$$E_F(T(\underline{X})) = g(F) \quad \forall F \in \mathcal{P}.$$

■

Example 12.1.8:

Let $\mathcal{P}$ be the class of distributions for which second moments exist. Then $\overline{X}$ is unbiased for $\mu(F) = \int x dF(x)$. Thus, $\mu(F)$ is estimable. ■

Definition 12.1.9:

The **degree** m of an estimable parameter $g(F)$ is the smallest sample size for which an unbiased estimate exists for all $F \in \mathcal{P}$.

An unbiased estimate based on a sample of size m is called a **kernel**. ■

Lemma 12.1.10:

There exists a **symmetric kernel** for every estimable parameter.

Proof:

Let $T(X_1, \ldots, X_m)$ be a kernel of $g(F)$. Define

$$T_s(X_1, \ldots, X_m) = \frac{1}{m!} \sum_{\text{all permutations of}\{1,\ldots,m\}} T(X_{i_1}, \ldots, X_{i_m}).$$

where the summation is over all $m!$ permutations of $\{1, \ldots, m\}$.

Clearly T_s is symmetric and $E(T_s) = g(F)$. ■

Example 12.1.11:

(i) $E(X_1) = \mu(F)$, so $\mu(F)$ has degree 1 with kernel X_1.

(ii) $E(I_{(c,\infty)}(X_1)) = P_F(X > c)$, where c is a known constant. So $g(F) = P_F(X > c)$ has degree 1 with kernel $I_{(c,\infty)}(X_1)$.

(iii) There exists no $T(X_1)$ such that $E(T(X_1)) = \sigma^2(F) = \int (x - \mu(F))^2 dF(x)$.

But $E(T(X_1, X_2)) = E(X_1^2 - X_1 X_2) = \sigma^2(F)$. So $\sigma^2(F)$ has degree 2 with kernel $X_1^2 - X_1 X_2$. Note that $X_2^2 - X_2 X_1$ is another kernel.

(iv) A symmetric kernel for $\sigma^2(F)$ is

$$T_s(X_1, X_2) = \frac{1}{2}((X_1^2 - X_1 X_2) + (X_2^2 - X_1 X_2)) = \frac{1}{2}(X_1 - X_2)^2.$$

■

Lecture 35: We 04/11/01

Definition 12.1.12:

Let $g(F)$ be an estimable parameter of degree m. Let $X_1, \ldots, X_n$ be a sample of size $n, n \geq m$. Given a kernel $T(X_{i_1}, \ldots, X_{i_m})$ of $g(F)$, we define a U–**statistic** by

$$U(X_1, \ldots, X_n) = \frac{1}{\binom{n}{m}} \sum_c T_s(X_{i_1}, \ldots, X_{i_m}),$$

where T_s is defined as in Lemma 12.1.10 and the summation c is over all $\binom{n}{m}$ combinations of m integers $(i_1, \ldots, i_m)$ from $\{1, \cdots, n\}$. $U(X_1, \ldots, X_n)$ is symmetric in the X_i's and $E_F(U(\underline{X})) = g(F)$ for all F. ■

Example 12.1.13:

For estimating $\mu(F)$ with degree m of $\mu(F) = 1$:

Symmetric kernel:

$$T_s(X_i) = X_i, \quad i = 1, \ldots, n$$

U–statistic:

$$\begin{aligned} U_\mu(\underline{X}) &= \frac{1}{\binom{n}{1}} \sum_c X_i \\ &= \frac{1 \cdot (n-1)!}{n!} \sum_c X_i \\ &= \frac{1}{n} \sum_{i=1}^n X_i \\ &= \overline{X} \end{aligned}$$

For estimating $\sigma^2(F)$ with degree m of $\sigma^2(F) = 2$:

Symmetric kernel:

$$T_s(X_{i_1}, X_{i_2}) = \frac{1}{2}(X_{i_1} - X_{i_2})^2, \quad i_1, i_2 = 1, \ldots, n, i_1 \neq i_2$$

U–statistic:

$$\begin{aligned} U_{\sigma^2}(\underline{X}) &= \frac{1}{\binom{n}{2}} \sum_{i_1 < i_2} \frac{1}{2}(X_{i_1} - X_{i_2})^2 \\ &= \frac{1}{\binom{n}{2}} \frac{1}{4} \sum_{i_1 \neq i_2} (X_{i_1} - X_{i_2})^2 \\ &= \frac{(n-2)! \cdot 2!}{n!} \frac{1}{4} \sum_{i_1 \neq i_2} (X_{i_1} - X_{i_2})^2 \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2n(n-1)} \sum_{i_1} \sum_{i_2 \neq i_1} (X_{i_1}^2 - 2X_{i_1}X_{i_2} + X_{i_2}^2) \\
&= \frac{1}{2n(n-1)} \left[(n-1) \sum_{i_1=1}^{n} X_{i_1}^2 - 2(\sum_{i_1=1}^{n} X_{i_1})(\sum_{i_2=1}^{n} X_{i_2}) + 2 \sum_{i=1}^{n} X_i^2 + \right. \\
&\qquad \left. (n-1) \sum_{i_2=1}^{n} X_{i_2}^2 \right] \\
&= \frac{1}{2n(n-1)} \left[n \sum_{i_1=1}^{n} X_{i_1}^2 - \sum_{i_1=1}^{n} X_{i_1}^2 - 2(\sum_{i_1=1}^{n} X_{i_1})^2 + 2 \sum_{i=1}^{n} X_i^2 + \right. \\
&\qquad \left. n \sum_{i_2=1}^{n} X_{i_2}^2 - \sum_{i_2=1}^{n} X_{i_2}^2 \right] \\
&= \frac{1}{n(n-1)} \left[n \sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2 \right] \\
&= \frac{1}{n(n-1)} \left[n \sum_{i=1}^{n} \left(X_i - \frac{1}{n} \sum_{j=1}^{n} X_j \right)^2 \right] \\
&= \frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - \overline{X})^2 \\
&= S^2
\end{aligned}
$$

■

Theorem 12.1.14:

Let $\mathcal{P}$ be the class of all absolutely continuous or all purely discrete distribution functions on $I\!R$. Any estimable function $g(F), F \in \mathcal{P}$, has a unique estimate that is unbiased and symmetric in the observations and has uniformly minimum variance among all unbiased estimates.

Proof:

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F \in \mathcal{P}$, with $T(X_1, \ldots, X_n)$ an unbiased estimate of $g(F)$.

We define

$$T_i = T_i(X_1, \ldots, X_n) = T(X_{i_1}, X_{i_2}, \ldots, X_{i_n}), \quad i = 1, 2, \ldots, n!,$$

over all possible permutations of $\{1, \ldots, n\}$.

Let $\overline{T} = \frac{1}{n!} \sum_{i=1}^{n!} T_i$ and $T = \sum_{i=1}^{n!} T_i$.

Then

$$E_F(\overline{T}) = g(F)$$

and

$$\begin{aligned}
Var(\overline{T}) &= E(\overline{T}^2) - (E(\overline{T}))^2 \\
&= E\left[(\frac{1}{n!}\sum_{i=1}^{n!} T_i)^2\right] - [g(F)]^2 \\
&= E\left[(\frac{1}{n!})^2 \sum_{i=1}^{n!}\sum_{j=1}^{n!} T_i T_j\right] - [g(F)]^2 \\
&\leq E\left[\sum_{i=1}^{n!}\sum_{j=1}^{n!} T_i T_j\right] - [g(F)]^2 \\
&= E\left[\left(\sum_{i=1}^{n!} T_i\right)\left(\sum_{j=1}^{n!} T_j\right)\right] - [g(F)]^2 \\
&= E\left[\left(\sum_{i=1}^{n!} T_i\right)^2\right] - [g(F)]^2 \\
&= E(T^2) - [g(F)]^2 \\
&= Var(T)
\end{aligned}$$

Equality holds iff $T_i = T_j \quad \forall i, j = 1, \ldots, n!$

$\Longrightarrow$ T is symmetric in $(X_1, \ldots, X_n)$ and $\overline{T} = T$

$\Longrightarrow$ by Rohatgi, Problem 4, page 538, T is a function of order statistics

$\Longrightarrow$ by Rohatgi, Theorem 1, page 535, T is a complete sufficient statistic

$\Longrightarrow$ by Note (i) following Theorem 8.4.12, T is UMVUE ■

Corollary 12.1.15:

If $T(X_1, \ldots, X_n)$ is unbiased for $g(F), F \in \mathcal{P}$, the corresponding U–statistic is an essentially unique UMVUE. ■

Definition 12.1.16:

Suppose we have independent samples $X_1, \ldots, X_m \overset{\text{iid}}{\sim} F \in \mathcal{P},\ Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} G \in \mathcal{P}$ (G may or may not equal F.) Let $g(F, G)$ be an estimable function with unbiased estimator $T(X_1, \ldots, X_k,\ Y_1, \ldots, Y_l)$. Define

$$T_s(X_1, \ldots, X_k,\ Y_1, \ldots, Y_l) = \frac{1}{k!l!} \sum_{P_X} \sum_{P_Y} T(X_{i_1}, \ldots, X_{i_k},\ Y_{j_1}, \ldots, Y_{j_l})$$

(where P_X and P_Y are permutations of X and Y) and

$$U(\underline{X}, \underline{Y}) = \frac{1}{\binom{m}{k}\binom{n}{l}} \sum_{C_X} \sum_{C_Y} T_s(X_{i_1}, \ldots, X_{i_k},\ Y_{j_1}, \ldots, Y_{j_l})$$

(where C_X and C_Y are combinations of X and Y).

U is a called a **generalized U–statistic**. ■

Example 12.1.17:

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples from F and G, respectively, with $F, G \in \mathcal{P}$. We wish to estimate

$$g(F, G) = P_{F,G}(X \leq Y).$$

Let us define

$$Z_{ij} = \begin{cases} 1, & X_i \leq Y_j \\ 0, & X_i > Y_j \end{cases}$$

for each pair X_i, Y_j, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$.

Then $\sum_{i=1}^{m} Z_{ij}$ is the number of X's $\leq Y_j$, and $\sum_{j=1}^{n} Z_{ij}$ is the number of Y's $> X_i$.

$$E(I(X_i \leq Y_j)) = g(F, G) = P_{F,G}(X \leq Y),$$

and degrees k and l are $= 1$, so we use

$$\begin{aligned} U(\underline{X}, \underline{Y}) &= \frac{1}{\binom{m}{1}\binom{n}{1}} \sum_{C_X} \sum_{C_Y} T_s(X_{i_1}, \ldots, X_{i_k},\ Y_{j_1}, \ldots, Y_{j_l}) \\ &= \frac{(m-1)!(n-1)!}{m!n!} \sum_{C_X} \sum_{C_Y} \frac{1}{1!1!} \sum_{P_X} \sum_{P_Y} T(X_{i_1}, \ldots, X_{i_k},\ Y_{j_1}, \ldots, Y_{j_l}) \\ &= \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} I(X_i \leq Y_j). \end{aligned}$$

This **Mann–Whitney estimator** (or **Wilcoxin 2–Sample estimator**) is unbiased and symmetric in the X's and Y's. It follows by Corollary 12.1.15 that it has minimum variance. ■

Lecture 36: Fr 04/13/01

12.2 Single-Sample Hypothesis Tests

Let $X_1, \ldots, X_n$ be a sample from a distribution F. The **problem of fit** is to test the hypothesis that the sample $X_1, \ldots, X_n$ is from some specified distribution against the alternative that it is from some other distribution, i.e., $H_0 : F = F_0$ vs. $H_1 : F(x) \neq F_0(x)$ for some x.

Definition 12.2.1:
Let $X_1, \ldots, X_n \overset{iid}{\sim} F$, and let the corresponding empirical cdf be

$$F_n^*(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(X_i).$$

The statistic

$$D_n = \sup_x \mid F_n^*(x) - F(x) \mid$$

is called the **two–sided Kolmogorov–Smirnov statistic** (**K–S statistic**).

The **one–sided K–S statistics** are

$$D_n^+ = \sup_x [F_n^*(x) - F(x)] \quad \text{and} \quad D_n^- = \sup_x [F(x) - F_n^*(x)].$$

■

Theorem 12.2.2:
For any continuous distribution F, the K–S statistics D_n, D_n^-, D_n^+ are distribution free.

Proof:
Let $X_{(1)}, \ldots, X_{(n)}$ be the order statistics of $X_1, \ldots, X_n$, i.e., $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$, and define $X_{(0)} = -\infty$ and $X_{(n+1)} = +\infty$.

Then,

$$F_n^*(x) = \frac{i}{n} \text{ for } X_{(i)} \leq x < X_{(i+1)}, \ i = 0, \ldots, n.$$

Therefore,

$$\begin{aligned}
D_n^+ &= \max_{0 \leq i \leq n} \{ \sup_{X_{(i)} \leq x < X_{(i+1)}} [\frac{i}{n} - F(x)]\} \\
&= \max_{0 \leq i \leq n} \{ \frac{i}{n} - [\inf_{X_{(i)} \leq x < X_{(i+1)}} F(x)]\} \\
&\overset{(*)}{=} \max_{0 \leq i \leq n} \{ \frac{i}{n} - F(X_{(i)})\} \\
&= \max \{ \max_{1 \leq i \leq n} \left\{ \frac{i}{n} - F(X_{(i)}) \right\}, \ 0\}
\end{aligned}$$

$(*)$ holds since F is nondecreasing in $[X_{(i)}, X_{(i+1)})$.

Note that D_n^+ is a function of $F(X_{(i)})$. In order to make some inference about D_n^+, the distribution of $F(X_{(i)})$ must be known. We know from the *Probability Integral Transformation* (see Rohatgi, page 203, Theorem 1) that for a rv X with continuous cdf F_X, it holds that $F_X(X) \sim U(0,1)$.

Thus, $F(X_{(i)})$ is the i^{th} order statistic of a sample from $U(0,1)$, independent from F. Therefore, the distribution of D_n^+ is independent of F.

Similarly, the distribution of

$$D_n^- = \max \ \{ \max_{1 \le i \le n} \left\{ F(X_{(i)}) - \frac{i-1}{n} \right\}, \ 0 \}$$

is independent of F.

Since

$$D_n = \sup_x \mid F_n^*(x) - F(x) \mid = \max \ \{D_n^+, D_n^-\},$$

the distribution of D_n is also independent of F. ■

Theorem 12.2.3:
If F is continuous, then

$$P(D_n \le \nu + \frac{1}{2n}) = \begin{cases} 0, & \text{if } \nu \le 0 \\ \displaystyle\int_{\frac{1}{2n}-\nu}^{\nu+\frac{1}{2n}} \int_{\frac{3}{2n}-\nu}^{\nu+\frac{3}{2n}} \dots \int_{\frac{2n-1}{2n}-\nu}^{\nu+\frac{2n-1}{2n}} f(\underline{u}) d\underline{u}, & \text{if } 0 < \nu < \frac{2n-1}{2n} \\ 1, & \text{if } \nu \ge \frac{2n-1}{2n} \end{cases}$$

where

$$f(\underline{u}) = f(u_1, \dots, u_n) = \begin{cases} n!, & \text{if } 0 < u_1 < u_2 < \dots < u_n < 1 \\ 0, & \text{otherwise} \end{cases}$$

is the joint pdf of an order statistic of a sample of size n from $U(0,1)$. ■

Note:
As Gibbons & Chakraborti (1992), page 108–109, point out, this result must be interpreted carefully. Consider the case $n = 2$.

For $0 < \nu < \frac{3}{4}$, it holds that

$$P(D_2 \le \nu + \frac{1}{4}) = \int_{\frac{1}{4}-\nu}^{\nu+\frac{1}{4}} \ \int_{\substack{\frac{3}{4}-\nu \\ 0<u_1<u_2<1}}^{\nu+\frac{3}{4}} 2! \ du_2 \ du_1.$$

Note that the integration limits overlap if

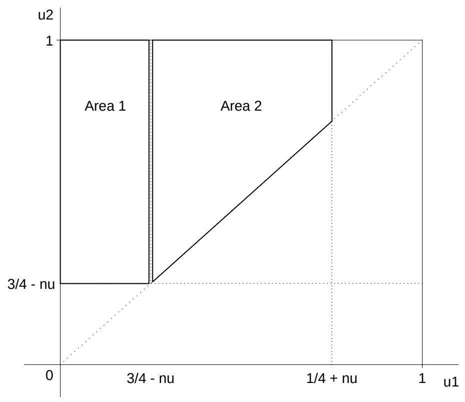
$$\begin{aligned} \nu + \frac{1}{4} &\ge -\nu + \frac{3}{4} \\ \iff \nu &\ge \frac{1}{4} \end{aligned}$$

When $0 < \nu < \frac{1}{4}$, it automatically holds that $0 < u_1 < u_2 < 1$. Thus, for $0 < \nu < \frac{1}{4}$, it holds that

$$
\begin{aligned}
P(D_2 \leq \nu + \frac{1}{4}) &= \int_{\frac{1}{4}-\nu}^{\nu+\frac{1}{4}} \int_{\frac{3}{4}-\nu}^{\nu+\frac{3}{4}} 2! \; du_2 \; du_1 \\
&= 2! \int_{\frac{1}{4}-\nu}^{\nu+\frac{1}{4}} \left(u_2 |_{\frac{3}{4}-\nu}^{\nu+\frac{3}{4}} \right) du_1 \\
&= 2! \int_{\frac{1}{4}-\nu}^{\nu+\frac{1}{4}} 2\nu \; du_1 \\
&= 2! \; (2\nu) \; u_1 |_{\frac{1}{4}-\nu}^{\nu+\frac{1}{4}} \\
&= 2! \; (2\nu)^2
\end{aligned}
$$

For $\frac{1}{4} \leq \nu < \frac{3}{4}$, the region of integration is as follows:

Note to Theorem 12.2.3



Thus, for $\frac{1}{4} \leq \nu < \frac{3}{4}$, it holds that

$$
\begin{aligned}
P(D_2 \leq \nu + \frac{1}{4}) &= \int_{\frac{1}{4}-\nu}^{\nu+\frac{1}{4}} {}_{0<u_1<u_2<1} \int_{\frac{3}{4}-\nu}^{\nu+\frac{3}{4}} 2! \; du_2 \; du_1 \\
&= \int_{\frac{3}{4}-\nu}^{\nu+\frac{1}{4}} \int_{u_1}^{1} 2! \; du_2 \; du_1 + \int_{0}^{\frac{3}{4}-\nu} \int_{\frac{3}{4}-\nu}^{1} 2! \; du_2 \; du_1 \\
&= 2 \left[\int_{\frac{3}{4}-\nu}^{\nu+\frac{1}{4}} \left(u_2 |_{u_1}^{1} \right) du_1 + \int_{0}^{\frac{3}{4}-\nu} \left(u_2 |_{\frac{3}{4}-\nu}^{1} \right) du_1 \right] \\
&= 2 \left[\int_{\frac{3}{4}-\nu}^{\nu+\frac{1}{4}} (1 - u_1) \; du_1 + \int_{0}^{\frac{3}{4}-\nu} (1 - \frac{3}{4} + \nu) \; du_1 \right]
\end{aligned}
$$

$$
\begin{aligned}
&= 2\left[\left(u_1 - \frac{u_1^2}{2}\right)\Bigg|_{\frac{3}{4}-\nu}^{\nu+\frac{1}{4}} + \left(\frac{u_1}{4} + \nu u_1\right)\Bigg|_0^{\frac{3}{4}-\nu}\right] \\
&= 2\left[(\nu + \frac{1}{4}) - \frac{(\nu+\frac{1}{4})^2}{2} - (-\nu + \frac{3}{4}) + \frac{(-\nu+\frac{3}{4})^2}{2} + \frac{(-\nu+\frac{3}{4})}{4} + \nu(-\nu+\frac{3}{4})\right] \\
&= 2\left[\nu + \frac{1}{4} - \frac{\nu^2}{2} - \frac{\nu}{4} - \frac{1}{32} + \nu - \frac{3}{4} + \frac{\nu^2}{2} - \frac{3}{4}\nu + \frac{9}{32} - \frac{\nu}{4} + \frac{3}{16} - \nu^2 + \frac{3}{4}\nu\right] \\
&= 2\left[-\nu^2 + \frac{3}{2}\nu - \frac{1}{16}\right] \\
&= -2\nu^2 + 3\nu - \frac{1}{8}
\end{aligned}
$$

Combining these results gives

$$
P(D_2 \leq \nu + \frac{1}{4}) = \begin{cases} 0, & \text{if } \nu \leq 0 \\ 2!\ (2\nu)^2, & \text{if } 0 < \nu < \frac{1}{4} \\ -2\nu^2 + 3\nu - \frac{1}{8}, & \text{if } \frac{1}{4} \leq \nu < \frac{3}{4} \\ 1, & \text{if } \nu \geq \frac{3}{4} \end{cases}
$$

■

Lecture 37: Mo 04/16/01

Theorem 12.2.4:

Let F be a continuous cdf. Then it holds $\forall z \geq 0$:

$$
\lim_{n\to\infty} P(D_n \leq \frac{z}{\sqrt{n}}) = L_1(z) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}\exp(-2i^2z^2).
$$

■

Theorem 12.2.5:

Let F be a continuous cdf. Then it holds:

$$
P(D_n^+ \leq z) = P(D_n^- \leq z) = \begin{cases} 0, & \text{if } z \leq 0 \\ \displaystyle\int_{1-z}^{1}\int_{\frac{n-1}{n}-z}^{u_n}\cdots\int_{\frac{2}{n}-z}^{u_3}\int_{\frac{1}{n}-z}^{u_2} f(\underline{u})d\underline{u}, & \text{if } 0 < z < 1 \\ 1, & \text{if } z \geq 1 \end{cases}
$$

where $f(\underline{u})$ is defined in Theorem 12.2.3. ■

Note:

It should be obvious that the statistics D_n^+ and D_n^- have the same distribution because of symmetry. ■

Theorem 12.2.6:
Let F be a continuous cdf. Then it holds $\forall z \geq 0$:

$$\lim_{n\to\infty} P(D_n^+ \leq \frac{z}{\sqrt{n}}) = \lim_{n\to\infty} P(D_n^- \leq \frac{z}{\sqrt{n}}) = L_2(z) = 1 - \exp(-2z^2)$$

■

Corollary 12.2.7:
Let $V_n = 4n(D_n^+)^2$. Then it holds $V_n \xrightarrow{d} \chi_2^2$, i.e., this transformation of D_n^+ has an asymptotic χ_2^2 distribution.

Proof:
Let $x \geq 0$. Then it follows:

$$\begin{aligned}
\lim_{n\to\infty} P(V_n \leq x) &\overset{x=4z^2}{=} \lim_{n\to\infty} P(V_n \leq 4z^2) \\
&= \lim_{n\to\infty} P(4n(D_n^+)^2 \leq 4z^2) \\
&= \lim_{n\to\infty} P(\sqrt{n} D_n^+ \leq z) \\
&\overset{Th.12.2.6}{=} 1 - \exp(-2z^2) \\
&\overset{4z^2=x}{=} 1 - \exp(-x/2)
\end{aligned}$$

Thus, $\lim\limits_{n\to\infty} P(V_n \leq x) = 1 - \exp(-x/2)$ for $x \geq 0$. Note that this is the cdf of a χ_2^2 distribution. ■

Definition 12.2.8:
Let $D_{n;\alpha}$ be the smallest value such that $P(D_n > D_{n;\alpha}) \leq \alpha$. Likewise, let $D_{n;\alpha}^+$ be the smallest value such that $P(D_n^+ > D_{n;\alpha}^+) \leq \alpha$.

The **Kolmogorov–Smirnov test** (**K–S test**) rejects $H_0 : F(x) = F_0(x) \ \ \forall x$ at level α if $D_n > D_{n;\alpha}$.

It rejects $H_0' : F(x) \geq F_0(x) \ \ \forall x$ at level α if $D_n^- > D_{n;\alpha}^+$ and it rejects $H_0'' : F(x) \leq F_0(x) \ \ \forall x$ at level α if $D_n^+ > D_{n;\alpha}^+$. ■

Note:
Rohatgi, Table 7, page 661, gives values of $D_{n;\alpha}$ and $D_{n;\alpha}^+$ for selected values of α and small n. Theorems 12.2.4 and 12.2.6 allow the approximation of $D_{n;\alpha}$ and $D_{n;\alpha}^+$ for large n. ■

Example 12.2.9:

Let $X_1, \ldots, X_n \sim C(1, 0)$. We want to test whether $H_0 : X \sim N(0, 1)$.

The following data has been observed for $x_{(1)}, \ldots, x_{(10)}$:

$$-1.42, -0.43, -0.19, 0.26, 0.30, 0.45, 0.64, 0.96, 1.97, \text{ and } 4.68$$

The results for the K–S test have been obtained through the following S–Plus session, i.e., $D_{10}^+ = 0.02219616$, $D_{10}^- = 0.3025681$, and $D_{10} = 0.3025681$:

```
> x _ c(-1.42, -0.43, -0.19, 0.26, 0.30, 0.45, 0.64, 0.96, 1.97, 4.68)
> FX _ pnorm(x)
> FX
 [1] 0.07780384 0.33359782 0.42465457 0.60256811 0.61791142 0.67364478
 [7] 0.73891370 0.83147239 0.97558081 0.99999857
> Dp _ (1:10)/10 - FX
> Dp
 [1]  2.219616e-02 -1.335978e-01 -1.246546e-01 -2.025681e-01 -1.179114e-01
 [6] -7.364478e-02 -3.891370e-02 -3.147239e-02 -7.558081e-02  1.434375e-06
> Dm _ FX - (0:9)/10
> Dm
 [1] 0.07780384 0.23359782 0.22465457 0.30256811 0.21791142 0.17364478
 [7] 0.13891370 0.13147239 0.17558081 0.09999857
> max(Dp)
[1] 0.02219616
> max(Dm)
[1] 0.3025681
> max(max(Dp), max(Dm))
[1] 0.3025681
>
> ks.gof(x, alternative = "two.sided", mean = 0, sd = 1)

        One-sample Kolmogorov-Smirnov Test
        Hypothesized distribution = normal

data:  x
ks = 0.3026, p-value = 0.2617
alternative hypothesis:
  True cdf is not the normal distn. with the specified parameters
```

Using Rohatgi, Table 7, page 661, we have to use $D_{10;0.20} = 0.323$ for $\alpha = 0.20$. Since $D_{10} = 0.3026 < 0.323 = D_{10;0.20}$, it is $p > 0.20$. The K–S test does not reject H_0 at level $\alpha = 0.20$. As S–Plus shows, the precise p–value is even $p = 0.2617$. ■

Note:

Comparison between χ^2 and K–S goodness of fit tests:

- K–S uses all available data; χ^2 bins the data and loses information
- K–S works for all sample sizes; χ^2 requires large sample sizes
- it is more difficult to modify K–S for estimated parameters; χ^2 can be easily adapted for estimated parameters
- K–S is "conservative" for discrete data, i.e., it tends to accept H_0 for such data
- the order matters for K–S; χ^2 is better for unordered categorical data

■

12.3 More on Order Statistics

Definition 12.3.1:

Let F be a continuous cdf. A **tolerance interval** for F with **tolerance coefficient** γ is a random interval such that the probability is γ that this random interval covers at least a specified percentage 100p% of the distribution. ■

Theorem 12.3.2:

If order statistics $X_{(r)} < X_{(s)}$ are used as the endpoints for a tolerance interval for a continuous cdf F, it holds that

$$\gamma = \sum_{i=0}^{s-r-1} \binom{n}{i} p^i (1-p)^{n-i}.$$

Proof:

According to Definition 12.3.1, it holds that

$$\gamma = P_{X_{(r)},X_{(s)}} \left(P_X(X_{(r)} < X < X_{(s)}) \geq p \right).$$

Since F is continuous, it holds that $F_X(X) \sim U(0,1)$. Therefore,

$$\begin{aligned} P_X(X_{(r)} < X < X_{(s)}) &= P(X < X_{(s)}) - P(X \leq X_{(r)}) \\ &= F(X_{(s)}) - F(X_{(r)}) \\ &= U_{(s)} - U_{(r)}, \end{aligned}$$

where $U_{(s)}$ and $U_{(r)}$ are the order statistics of a $U(0,1)$ distribution.

Thus,

$$\gamma = P_{X_{(r)},X_{(s)}} \left(P_X(X_{(r)} < X < X_{(s)}) \geq p \right) = P(U_{(s)} - U_{(r)} \geq p).$$

By Therorem 4.4.4, we can determine the joint distribution of order statistics and calculate γ as

$$\gamma = \int_p^1 \int_0^{y-p} \frac{n!}{(r-1)!(s-r-1)!(n-s)!} x^{r-1} (y-x)^{s-r-1} (1-y)^{n-s} \, dx \, dy.$$

Rather than solving this integral directly, we make the transformation

$$\begin{aligned} U &= U_{(s)} - U_{(r)} \\ V &= U_{(s)}. \end{aligned}$$

Then the joint pdf of U and V is

$$f_{U,V}(u,v) = \begin{cases} \frac{n!}{(r-1)!(s-r-1)!(n-s)!} (v-u)^{r-1} u^{s-r-1} (1-v)^{n-s}, & \text{if } 0 < u < v < 1 \\ 0, & \text{otherwise} \end{cases}$$

and the marginal pdf of U is

$$\begin{aligned}
f_U(u) &= \int_0^1 f_{U,V}(u,v)\ dv \\
&= \frac{n!}{(r-1)!(s-r-1)!(n-s)!} u^{s-r-1} I_{(0,1)}(u) \int_u^1 (v-u)^{r-1}(1-v)^{n-s}\ dv \\
&\overset{(A)}{=} \frac{n!}{(r-1)!(s-r-1)!(n-s)!} u^{s-r-1}(1-u)^{n-s+r} I_{(0,1)}(u) \underbrace{\int_0^1 t^{r-1}(1-t)^{n-s}\ dt}_{B(r,n-s+1)} \\
&= \frac{n!}{(r-1)!(s-r-1)!(n-s)!} u^{s-r-1}(1-u)^{n-s+r} \frac{(r-1)!(n-s)!}{(n-s+r)!} I_{(0,1)}(u) \\
&= \frac{n!}{(n-s+r)!(s-r-1)!} u^{s-r-1}(1-u)^{n-s+r} I_{(0,1)}(u) \\
&= n \binom{n-1}{s-r-1} u^{s-r-1}(1-u)^{n-s+r} I_{(0,1)}(u).
\end{aligned}$$

(A) is based on the transformation $t = \dfrac{v-u}{1-u}$, $\quad v-u = (1-u)t, \quad 1-v = 1-u-(1-u)t = (1-u)(1-t)$ and $dv = (1-u)dt$.

It follows that

$$\begin{aligned}
\gamma &= P(U_{(s)} - U_{(r)} \geq p) \\
&= P(U \geq p) \\
&= \int_p^1 n \binom{n-1}{s-r-1} u^{s-r-1}(1-u)^{n-s+r}\ du \\
&\overset{(B)}{=} P(Y < s-r) \quad \mid \text{ where } Y \sim Bin(n,p) \\
&= \sum_{i=0}^{s-r-1} \binom{n}{i} p^i (1-p)^{n-i}.
\end{aligned}$$

(B) holds due to Rohatgi, Remark 3 after Theorem 5.3.18, page 216, since for $X \sim Bin(n,p)$, it holds that

$$P(X < k) = \int_p^1 n \binom{n-1}{k-1} x^{k-1}(1-x)^{n-k}\ dx.$$

■

Lecture 38:
We 04/18/01

Example 12.3.3:
Let $s = n$ and $r = 1$. Then,

$$\gamma = \sum_{i=0}^{n-2} \binom{n}{i} p^i (1-p)^{n-i} = 1 - p^n - np^{n-1}(1-p).$$

If $p = 0.8$ and $n = 10$, then

$$\gamma_{10} = 1 - (0.8)^{10} - 10 \cdot (0.8)^9 \cdot (0.2) = 0.624,$$

i.e., $(X_{(1)}, X_{(10)})$ defines a 62.4% tolerance interval for 80% probability.

If $p = 0.8$ and $n = 20$, then

$$\gamma_{20} = 1 - (0.8)^{20} - 20 \cdot (0.8)^{19} \cdot (0.2) = 0.931,$$

and if $p = 0.8$ and $n = 30$, then

$$\gamma_{30} = 1 - (0.8)^{30} - 30 \cdot (0.8)^{29} \cdot (0.2) = 0.989.$$

■

Theorem 12.3.4:
Let k_p be the p^{th} quantile of a continuous cdf F. Let $X_{(1)}, \ldots, X_{(n)}$ be the order statistics of a sample of size n from F. Then it holds that

$$P(X_{(r)} \leq k_p \leq X_{(s)}) = \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i}.$$

Proof:
It holds that

$$\begin{aligned} P(X_{(r)} \leq k_p) &= P(\text{at least } r \text{ of the } X_i\text{'s are } \leq k_p) \\ &= \sum_{i=r}^{n} \binom{n}{i} p^i (1-p)^{n-i}. \end{aligned}$$

Therefore,

$$\begin{aligned} P(X_{(r)} \leq k_p \leq X_{(s)}) &= P(X_{(r)} \leq k_p) - P(X_{(s)} < k_p) \\ &= \sum_{i=r}^{n} \binom{n}{i} p^i (1-p)^{n-i} - \sum_{i=s}^{n} \binom{n}{i} p^i (1-p)^{n-i} \\ &= \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i}. \end{aligned}$$

■

Corollary 12.3.5:

$(X_{(r)}, X_{(s)})$ is a level $\sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i}$ confidence interval for k_p. ■

Example 12.3.6:

Let $n = 10$. We want a 95% confidence interval for the median, i.e., k_p where $p = \frac{1}{2}$.

We get the following probabilities $p_{r,s} = \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i}$ that $(X_{(r)}, X_{(s)})$ covers $k_{0.5}$:

$p_{r,s}$		s								
		2	3	4	5	6	7	8	9	10
	1	0.01	0.05	0.17	0.38	0.62	0.83	0.94	**0.99**	**0.998**
	2		0.04	0.16	0.37	0.61	0.82	0.93	**0.98**	**0.99**
	3			0.12	0.32	0.57	0.77	0.89	0.93	0.94
	4				0.21	0.45	0.66	0.77	0.82	0.83
r	5					0.25	0.45	0.57	0.61	0.62
	6						0.21	0.32	0.37	0.38
	7							0.12	0.16	0.17
	8								0.04	0.05
	9									0.01

Only the random intervals $(X_{(1)}, X_{(9)})$, $(X_{(1)}, X_{(10)})$, $(X_{(2)}, X_{(9)})$, and $(X_{(2)}, X_{(10)})$ give the desired coverage probability. Therefore, we use the one that comes closest to 95%, i.e., $(X_{(2)}, X_{(9)})$, as the 95% confidence interval for the median. ■

13 Some Results from Sampling

13.1 Simple Random Samples

Definition 13.1.1:

Let Ω be a population of size N with mean μ and variance σ^2. A sampling method (of size n) is called **simple** if the set S of possible samples contains all combinations of n elements of Ω (without repetition) and the probability for each sample $s \in S$ to become selected depends only on n, i.e., $p(s) = \frac{1}{\binom{N}{n}} \;\; \forall s \in S$. Then we call $s \in S$ a **simple random sample** (**SRS**) of size n. ■

Theorem 13.1.2:

Let Ω be a population of size N with mean μ and variance σ^2. Let $Y : \Omega \to I\!R$ be a measurable function. Let n_i be the total number of times the parameter $\tilde{y}_i$ occurs in the population and $p_i = \frac{n_i}{N}$ be the relative frequency the parameter $\tilde{y}_i$ occurs in the population. Let $(y_1, \ldots, y_n)$ be a SRS of size n with respect to Y, where $P(Y = \tilde{y}_i) = p_i = \frac{n_i}{N}$.

Then the components $y_i, \;\; i = 1, \ldots, n$, are identically distributed as Y and it holds for $i \neq j$:

$$P(y_i = \tilde{y}_k, y_j = \tilde{y}_l) = \frac{1}{N(N-1)} n_{kl}, \text{ where } n_{kl} = \begin{cases} n_k n_l, & k \neq l \\ n_k(n_k - 1), & k = l \end{cases}$$

■

Note:

(i) In Sampling, many authors use capital letters to denote properties of the population and small letters to denote properties of the random sample. In particular, x_i's and y_i's are considered as random variables related to the sample. They are not seen as specific realizations.

(ii) The following equalities hold in the scenario of Theorem 13.1.2:

$$\begin{aligned} N &= \sum_i n_i \\ \mu &= \frac{1}{N} \sum_i n_i \tilde{y}_i \\ \sigma^2 &= \frac{1}{N} \sum_i n_i (\tilde{y}_i - \mu)^2 \\ &= \frac{1}{N} \sum_i n_i \tilde{y}_i^2 - \mu^2 \end{aligned}$$

■

Theorem 13.1.3:

Let the same conditions hold as in Theorem 13.1.2. Let $\overline{y} = \frac{1}{n} \sum\limits_{i=1}^{n} y_i$ be the sample mean of a SRS of size n. Then it holds:

(i) $E(\overline{y}) = \mu$, i.e., the sample mean is unbiased for the population mean μ.

(ii) $Var(\overline{y}) = \dfrac{1}{n}\dfrac{N-n}{N-1}\sigma^2 = \dfrac{1}{n}(1-f)\dfrac{N}{N-1}\sigma^2$, where $f = \frac{n}{N}$.

Proof:

(i)

$$E(\overline{y}) = \frac{1}{n}\sum_{i=1}^{n} E(y_i) = \mu, \quad \text{since } E(y_i) = \mu \;\; \forall i.$$

(ii)

$$\begin{aligned}
Var(\overline{y}) &= \frac{1}{n^2}\left(\sum_{i=1}^{n} Var(y_i) + 2\sum_{i<j} Cov(y_i, y_j)\right) \\
Cov(y_i, y_j) &= E(y_i \cdot y_j) - E(y_i)E(y_j) \\
&= E(y_i \cdot y_j) - \mu^2 \\
&= \sum_{k,l} \tilde{y}_k \tilde{y}_l P(y_i = \tilde{y}_k, y_j = \tilde{y}_l) - \mu^2 \\
&\overset{Th. 13.1.2}{=} \frac{1}{N(N-1)}\left(\sum_{k \neq l} \tilde{y}_k \tilde{y}_l n_k n_l + \sum_k \tilde{y}_k^2 n_k (n_k - 1)\right) - \mu^2 \\
&= \frac{1}{N(N-1)}\left(\sum_{k,l} \tilde{y}_k \tilde{y}_l n_k n_l - \sum_k \tilde{y}_k^2 n_k\right) - \mu^2 \\
&= \frac{1}{N(N-1)}\left(\left(\sum_k \tilde{y}_k n_k\right)\left(\sum_l \tilde{y}_l n_l\right) - \sum_k \tilde{y}_k^2 n_k\right) - \mu^2 \\
&\overset{Note\ (ii)}{=} \frac{1}{N(N-1)}\left(N^2\mu^2 - N(\sigma^2 + \mu^2)\right) - \mu^2 \\
&= \frac{1}{N-1}\left(N\mu^2 - \sigma^2 - \mu^2 - \mu^2(N-1)\right) \\
&= -\frac{1}{N-1}\sigma^2, \quad \text{for } i \neq j
\end{aligned}$$

Lecture 39:
Fr 04/20/01

$$
\begin{aligned}
\Longrightarrow Var(\overline{y}) &= \frac{1}{n^2}\left(\sum_{i=1}^{n} Var(y_i) + 2\sum_{i<j} Cov(y_i, y_j)\right) \\
&= \frac{1}{n^2}\left(n\sigma^2 + n(n-1)\left(-\frac{1}{N-1}\sigma^2\right)\right) \\
&= \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\sigma^2 \\
&= \frac{1}{n}\frac{N-n}{N-1}\sigma^2 \\
&= \frac{1}{n}(1-\frac{n}{N})\frac{N}{N-1}\sigma^2 \\
&= \frac{1}{n}(1-f)\frac{N}{N-1}\sigma^2
\end{aligned}
$$

■

Theorem 13.1.4:
Let $\overline{y}_n$ be the sample mean of a SRS of size n. Then it holds that

$$\sqrt{\frac{n}{1-f}}\frac{\overline{y}_n - \mu}{\sqrt{\frac{N}{N-1}}\sigma} \xrightarrow{d} N(0,1),$$

where $N \to \infty$ and $f = \frac{n}{N}$ is a constant.

In particular, when the y_i's are 0–1–distributed with $E(y_i) = P(y_i = 1) = p \;\; \forall i$, then it holds that

$$\sqrt{\frac{n}{1-f}}\frac{\overline{y}_n - p}{\sqrt{\frac{N}{N-1}p(1-p)}} \xrightarrow{d} N(0,1),$$

where $N \to \infty$ and $f = \frac{n}{N}$ is a constant. ■

13.2 Stratified Random Samples

Definition 13.2.1:

Let Ω be a population of size N, that is split into m disjoint sets Ω_j, called **strata**, of size N_j, $j = 1, \ldots, m$, where $N = \sum\limits_{j=1}^{m} N_j$. If we independently draw a random sample of size n_j in each strata, we speak of a **stratified random sample**. ■

Note:

(i) The random samples in each strata are not always SRS's.

(ii) Stratified random samples are used in practice as a means to reduce the sample variance in the case that data in each strata is homogeneous and data among different strata is heterogeneous.

(iii) Frequently used strata in practice are gender, state (or county), income range, ethnic background, etc.

■

Definition 13.2.2:

Let $Y : \Omega \to I\!R$ be a measurable function. In case of a stratified random sample, we use the following notation:

Let $Y_{jk}, \quad j = 1, \ldots, m, \quad k = 1, \ldots, N_j$ be the elements in Ω_j. Then, we define

(i) $Y_j = \sum\limits_{k=1}^{N_j} Y_{jk}$ the total in the j^{th} strata,

(ii) $\mu_j = \frac{1}{N_j} Y_j$ the mean in the j^{th} strata,

(iii) $\mu = \frac{1}{N} \sum\limits_{j=1}^{m} N_j \mu_j$ the expectation (or grand mean),

(iv) $N\mu = \sum\limits_{j=1}^{m} Y_j = \sum\limits_{j=1}^{m} \sum\limits_{k=1}^{N_j} Y_{jk}$ the total,

(v) $\sigma_j^2 = \frac{1}{N_j} \sum\limits_{k=1}^{N_j} (Y_{jk} - \mu_j)^2$ the variance in the j^{th} strata, and

(vi) $\sigma^2 = \frac{1}{N} \sum\limits_{j=1}^{m} \sum\limits_{k=1}^{N_j} (Y_{jk} - \mu)^2$ the variance.

(vii) We denote an (ordered) sample in Ω_j of size n_j as $(y_{j1}, \ldots, y_{jn_j})$ and $\overline{y}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} y_{jk}$ the sample mean in the j^{th} strata.

■

Theorem 13.2.3:

Let the same conditions hold as in Definitions 13.2.1 and 13.2.2. Let $\hat{\mu}_j$ be an unbiased estimate of μ_j and $\widehat{Var(\hat{\mu}_j)}$ be an unbiased estimate of $Var(\hat{\mu}_j)$. Then it holds:

(i) $\hat{\mu} = \frac{1}{N} \sum_{j=1}^{m} N_j \hat{\mu}_j$ is unbiased for μ.

$Var(\hat{\mu}) = \frac{1}{N^2} \sum_{j=1}^{m} N_j^2 Var(\hat{\mu}_j)$.

(ii) $\widehat{Var(\hat{\mu})} = \frac{1}{N^2} \sum_{j=1}^{m} N_j^2 \widehat{Var(\hat{\mu}_j)}$ is unbiased for $Var(\hat{\mu})$.

Proof:

(i)

$$E(\hat{\mu}) = \frac{1}{N} \sum_{j=1}^{m} N_j E(\hat{\mu}_j) = \frac{1}{N} \sum_{j=1}^{m} N_j \mu_j = \mu$$

By independence of the samples within each strata,

$$Var(\hat{\mu}) = \frac{1}{N^2} \sum_{j=1}^{m} N_j^2 Var(\hat{\mu}_j).$$

(ii)

$$E(\widehat{Var(\hat{\mu})}) = \frac{1}{N^2} \sum_{j=1}^{m} N_j^2 E(\widehat{Var(\hat{\mu}_j)}) = \frac{1}{N^2} \sum_{j=1}^{m} N_j^2 Var(\hat{\mu}_j) = Var(\hat{\mu})$$

■

Theorem 13.2.4:

Let the same conditions hold as in Theorem 13.2.3. If we draw a SRS in each strata, then it holds:

(i) $\hat{\mu} = \frac{1}{N} \sum_{j=1}^{m} N_j \overline{y}_j$ is unbiased for μ, where $\overline{y}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} y_{jk}, \quad j = 1, \ldots, m$.

$Var(\hat{\mu}) = \frac{1}{N^2} \sum_{j=1}^{m} N_j^2 \frac{1}{n_j} (1 - f_j) \frac{N_j}{N_j - 1} \sigma_j^2$, where $f_j = \frac{n_j}{N_j}$.

(ii) $\widehat{Var(\hat{\mu})} = \frac{1}{N^2} \sum_{j=1}^{m} N_j^2 \frac{1}{n_j}(1 - f_j) s_j^2$ is unbiased for $Var(\hat{\mu})$, where

$$s_j^2 = \frac{1}{n_j - 1} \sum_{k=1}^{n_j} (y_{jk} - \overline{y}_j)^2.$$

Proof:
For a SRS in the j^{th} strata, it follows by Theorem 13.1.3:

$$\begin{aligned} E(\overline{y}_j) &= \mu_j \\ Var(\overline{y}_j) &= \frac{1}{n_j}(1 - f_j) \frac{N_j}{N_j - 1} \sigma_j^2 \end{aligned}$$

Also, we can show that

$$E(s_j^2) = \frac{N_j}{N_j - 1} \sigma_j^2.$$

Now the proof follows directly from Theorem 13.2.3. ■

Definition 13.2.5:
Let the same conditions hold as in Definitions 13.2.1 and 13.2.2. If the sample in each strata is of size $n_j = n \frac{N_j}{N}, \quad j = 1, \ldots, m$, where n is the total sample size, then we speak of **proportional selection**. ■

Note:

(i) In the case of proportional selection, it holds that $f_j = \frac{n_j}{N_j} = \frac{n}{N} = f, \quad j = 1, \ldots, m$.

(ii) Proportional strata cannot always be obtained for each combination of m, n, and N.

■

Theorem 13.2.6:
Let the same conditions hold as in Definition 13.2.5. If we draw a SRS in each strata, then it holds in case of proportional selection that

$$Var(\hat{\mu}) = \frac{1}{N^2} \frac{1 - f}{f} \sum_{j=1}^{m} N_j \tilde{\sigma}_j^2,$$

where $\tilde{\sigma}_j^2 = \frac{N_j}{N_j - 1} \sigma_j^2$.

Proof:
The proof follows directly from Theorem 13.2.4 (i). ■

Theorem 13.2.7:

If we draw (1) a stratified random sample that consists of SRS's of sizes n_j under proportional selection and (2) a SRS of size $n = \sum_{j=1}^{m} n_j$ from the same population, then it holds that

$$Var(\overline{y}) - Var(\hat{\mu}) = \frac{1}{n} \frac{N-n}{N(N-1)} \left(\sum_{j=1}^{m} N_j(\mu_j - \mu)^2 - \frac{1}{N} \sum_{j=1}^{m} (N - N_j)\tilde{\sigma}_j^2 \right).$$

Proof:

Homework. ■

Lecture 41:
We 04/25/01

14 Some Results from Sequential Statistical Inference

14.1 Fundamentals of Sequential Sampling

Example 14.1.1:
A particular machine produces a large number of items every day. Each item can be either "defective" or "non–defective". The unknown proportion of defective items in the production of a particular day is p.

Let $(X_1, \ldots, X_m)$ be a sample from the daily production where $x_i = 1$ when the item is defective and $x_i = 0$ when the item is non–defective. Obviously, $S_m = \sum_{i=1}^{m} X_i \sim Bin(m, p)$ denotes the total number of defective items in the sample (assuming that m is small compared to the daily production).

We might be interested to test $H_0 : \ p \leq p_0$ vs. $H_1 : \ p > p_0$ at a given significance level α and use this decision to trash the entire daily production and have the machine fixed if indeed $p > p_0$. A suitable test could be

$$\Phi_1(x_1, \ldots, x_m) = \begin{cases} 1, & \text{if } s_m > c \\ 0, & \text{if } s_m \leq c \end{cases}$$

where c is chosen such that Φ_1 is a level–α test.

However, wouldn't it be more beneficial if we sequentially sample the items (e.g., take item # 57, 623, 1005, 1286, 2663, etc.) and stop the machine as soon as it becomes obvious that it produces too many bad items. (Alternatively, we could also finish the time consuming and expensive process to determine whether an item is defective or non–defective if it is impossible to surpass a certain proportion of defectives.) For example, if for some $j < m$ it already holds that $s_j > c$, then we could stop (and immediately call maintenance) and reject H_0 after only j observations.

More formally, let us define $T = min\{j \mid S_j > c\}$ and $T' = min\{T, m\}$. We can now consider a decision rule that stops with the sampling process at random time T' and rejects H_0 if $T \leq m$. Thus, if we consider $R_0 = \{(x_1, \ldots, x_m) \mid t \leq m\}$ and $R_1 = \{(x_1, \ldots, x_m) \mid s_m > c\}$ as critical regions of two tests Φ_0 and Φ_1, then these two tests are equivalent. ■

Definition 14.1.2:

Let Θ be the parameter space and $\mathcal{A}$ the set of actions the statistician can take. We assume that the rv's $X_1, X_2, \ldots$ are observed sequentially and iid with common pdf (or pmf) $f_\theta(x)$. A **sequential decision procedure** is defined as follows:

(i) A **stopping rule** specifies whether an element of $\mathcal{A}$ should be chosen without taking any further observation. If at least one observation is taken, this rule specifies for every set of observed values $(x_1, x_2, \ldots, x_n)$, $n \geq 1$, whether to stop sampling and choose an action in $\mathcal{A}$ or to take another observation x_{n+1}.

(ii) A **decision rule** specifies the decision to be taken. If no observation has been taken, then we take action $d_0 \in \mathcal{A}$. If $n \geq 1$ observation have been taken, then we take action $d_n(x_1, \ldots, x_n) \in \mathcal{A}$, where $d_n(x_1, \ldots, x_n)$ specifies the action that has to be taken for the set $(x_1, \ldots, x_n)$ of observed values. Once an action has been taken, the sampling process is stopped.

■

Note:

In the remainder of this chapter, we assume that the statistician takes at least one observation.

■

Definition 14.1.3:

Let $R_n \subseteq I\!R^n, n = 1, 2, \ldots$, be a sequence of Borel–measurable sets such that the sampling process is stopped after observing $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ if $(x_1, \ldots, x_n) \in R_n$. If $(x_1, \ldots, x_n) \notin R_n$, then another observation x_{n+1} is taken. The sets $R_n, n = 1, 2, \ldots$ are called **stopping regions**. ■

Definition 14.1.4:

With every sequential stopping rule we associate a **stopping random variable** N which takes on the values $1, 2, 3, \ldots$. Thus, N is a rv that indicates the total number of observations taken before the sampling is stopped. ■

Note:

We use the (sloppy) notation $\{N = n\}$ to denote the event that sampling is stopped after observing exactly n values $x_1, \ldots, x_n$ (i.e., sampling is not stopped before taking n samples). Then the following equalities hold:

$$\{N = 1\} \quad = \quad R_1$$

$$
\begin{aligned}
\{N = n\} &= \{(x_1, \ldots, x_n) \in I\!R^n \mid \text{sampling is stopped after } n \text{ observations but not before}\} \\
&= (R_1 \cup R_2 \cup \ldots \cup R_{n-1})^c \cap R_n \\
&= R_1^c \cap R_2^c \cap \ldots \cap R_{n-1}^c \cap R_n \\
\{N \leq n\} &= \bigcup_{k=1}^{n} \{N = k\}
\end{aligned}
$$

Here we will only consider **closed** sequential sampling procedures, i.e., procedures where sampling eventually stops with probability 1, i.e.,

$$
\begin{aligned}
P(N < \infty) &= 1, \\
P(N = \infty) &= 1 - P(N < \infty) = 0.
\end{aligned}
$$

■

Theorem 14.1.5: **Wald's Equation**

Let $X_1, X_2, \ldots$ be iid rv's with $E(\mid X_1 \mid) < \infty$. Let N be a stopping variable. Let $S_N = \sum_{k=1}^{N} X_k$. If $E(N) < \infty$, then it holds

$$
E(S_N) = E(X_1)E(N).
$$

Proof:

Define a sequence of rv's $Y_i, \quad i = 1, 2, \ldots$, where

$$
Y_i = \begin{cases} 1, & \text{if no decision is reached up to the } (i-1)^{th} \text{ stage, i.e., } N > (i-1) \\ 0, & \text{otherwise} \end{cases}
$$

Then each Y_i is a function of $X_1, X_2, \ldots, X_{i-1}$ only and Y_i is independent of X_i.

Consider the rv

$$
\sum_{n=1}^{\infty} X_n Y_n.
$$

Obviously, it holds that

$$
S_N = \sum_{n=1}^{\infty} X_n Y_n.
$$

Thus, it follows that

$$
E(S_N) = E\left(\sum_{n=1}^{\infty} X_n Y_n\right). \qquad (*)
$$

It holds that

$$
\begin{aligned}
\sum_{n=1}^{\infty} E(\mid X_n Y_n \mid) &= \sum_{n=1}^{\infty} E(\mid X_n \mid) E(\mid Y_n \mid) \\
&= E(\mid X_1 \mid) \sum_{n=1}^{\infty} P(N \geq n)
\end{aligned}
$$

$$
\begin{aligned}
&= E(\mid X_1 \mid) \sum_{n=1}^{\infty} \sum_{k=n}^{\infty} P(N = k) \\
&\overset{(A)}{=} E(\mid X_1 \mid) \sum_{n=1}^{\infty} nP(N = n) \\
&= E(\mid X_1 \mid) E(N) \\
&< \infty
\end{aligned}
$$

(A) holds due to the following rearrangement of indizes:

n	k
1	$1, 2, 3, \ldots$
2	$2, 3, \ldots$
3	$3, \ldots$
$\vdots$	$\vdots$

We may therefore interchange the expectation and summation signs in $(*)$ and get

Lecture 42: Fr 04/27/01

$$
\begin{aligned}
E(S_N) &= E\left(\sum_{n=1}^{\infty} X_n Y_n\right) \\
&= \sum_{n=1}^{\infty} E(X_n Y_n) \\
&= \sum_{n=1}^{\infty} E(X_n) E(Y_n) \\
&= E(X_1) \sum_{n=1}^{\infty} P(N \geq n) \\
&= E(X_1) E(N)
\end{aligned}
$$

which completes the proof. ■

14.2 Sequential Probability Ratio Tests

Definition 14.2.1:

Let $X_1, X_2, \ldots$ be a sequence of iid rv's with common pdf (or pmf) $f_\theta(x)$. We want to test a simple hypothesis $H_0 : X \sim f_{\theta_0}$ vs. a simple alternative $H_1 : X \sim f_{\theta_1}$ when the observations are taken sequentially.

Let f_{0n} and f_{1n} denote the joint pdf's (or pmf's) of $X_1, \ldots, X_n$ under H_0 and H_1 respectively, i.e.,

$$f_{0n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{\theta_0}(x_i) \quad \text{and} \quad f_{1n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{\theta_1}(x_i).$$

Finally, let

$$\lambda_n(x_1, \ldots, x_n) = \frac{f_{1n}(\underline{x})}{f_{0n}(\underline{x})},$$

where $\underline{x} = (x_1, \ldots, x_n)$. Then a **sequential probability ratio test** (**SPRT**) for testing H_0 vs. H_1 is the following decision rule:

(i) If at any stage of the sampling process it holds that

$$\lambda_n(\underline{x}) \geq A,$$

then stop and reject H_0.

(ii) If at any stage of the sampling process it holds that

$$\lambda_n(\underline{x}) \leq B,$$

then stop and accept H_0, i.e., reject H_1.

(iii) If

$$B < \lambda_n(\underline{x}) < A,$$

then continue sampling by taking another observation x_{n+1}.

■

Note:

(i) It is usually convenient to define

$$Z_i = \log \frac{f_{\theta_1}(X_i)}{f_{\theta_0}(X_i)},$$

where $Z_1, Z_2, \ldots$ are iid rv's. Then, we work with

$$\log \lambda_n(\underline{x}) = \sum_{i=1}^{n} z_i = \sum_{i=1}^{n} (\log f_{\theta_1}(x_i) - \log f_{\theta_0}(x_i))$$

instead of using $\lambda_n(\underline{x})$. Obviously, we now have to use constants $b = \log B$ and $a = \log A$ instead of the original constants B and A.

(ii) A and B (where $A > B$) are constants such that the SPRT will have strength (α, β), where

$$\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 \mid H_0)$$

and

$$\beta = P(\text{Type II error}) = P(\text{Accept } H_0 \mid H_1).$$

If N is the stopping rv, then

$$\alpha = P_{\theta_0}(\lambda_N(\underline{X}) \geq A) \quad \text{and } \beta = P_{\theta_1}(\lambda_N(\underline{X}) \leq B).$$

■

Example 14.2.2:

Let $X_1, X_2, \ldots$ be iid $N(\mu, \sigma^2)$, where μ is unknown and $\sigma^2 > 0$ is known. We want to test $H_0: \ \mu = \mu_0$ vs. $H_1: \ \mu = \mu_1$, where $\mu_0 < \mu_1$.

If our data is sampled sequentially, we can constract a SPRT as follows:

$$\begin{aligned}
\log \lambda_n(\underline{x}) &= \sum_{i=1}^{n} \left(-\frac{1}{2\sigma^2}(x_i - \mu_1)^2 - (-\frac{1}{2\sigma^2}(x_i - \mu_0)^2) \right) \\
&= \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left((x_i - \mu_0)^2 - (x_i - \mu_1)^2) \right) \\
&= \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(x_i^2 - 2x_i\mu_0 + \mu_0^2 - x_i^2 + 2x_i\mu_1 - \mu_1^2 \right) \\
&= \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(-2x_i\mu_0 + \mu_0^2 + 2x_i\mu_1 - \mu_1^2 \right) \\
&= \frac{1}{2\sigma^2} \left(\sum_{i=1}^{n} 2x_i(\mu_1 - \mu_0) + n(\mu_0^2 - \mu_1^2) \right) \\
&= \frac{\mu_1 - \mu_0}{\sigma^2} \left(\sum_{i=1}^{n} x_i - n\frac{\mu_0 + \mu_1}{2} \right)
\end{aligned}$$

We decide for H_0 if

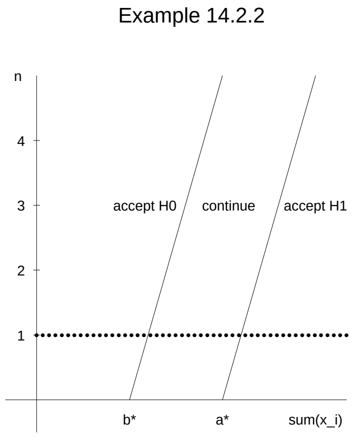
$$\begin{aligned}
& \log \lambda_n(\underline{x}) \leq b \\
\iff \quad & \frac{\mu_1 - \mu_0}{\sigma^2} \left(\sum_{i=1}^{n} x_i - n\frac{\mu_0 + \mu_1}{2} \right) \leq b \\
\iff \quad & \sum_{i=1}^{n} x_i \leq n\frac{\mu_0 + \mu_1}{2} + b^*,
\end{aligned}$$

where $b^* = \frac{\sigma^2}{\mu_1 - \mu_0} b$.

We decide for H_1 if

$$\begin{aligned} & \log \lambda_n(\underline{x}) \geq a \\ \Longleftrightarrow \quad & \frac{\mu_1 - \mu_0}{\sigma^2} \left(\sum_{i=1}^{n} x_i - n \frac{\mu_0 + \mu_1}{2} \right) \geq a \\ \Longleftrightarrow \quad & \sum_{i=1}^{n} x_i \geq n \frac{\mu_0 + \mu_1}{2} + a^*, \end{aligned}$$

where $a^* = \frac{\sigma^2}{\mu_1 - \mu_0} a$.



■

Theorem 14.2.3:

For a SPRT with stopping bounds A and B, $A > B$, and strength (α, β), we have

$$A \leq \frac{1-\beta}{\alpha} \quad \text{and} \quad B \geq \frac{\beta}{1-\alpha},$$

where $0 < \alpha < 1$ and $0 < \beta < 1$. ■

Theorem 14.2.4:

Assume we select for given $\alpha, \beta \in (0,1)$, where $\alpha + \beta \leq 1$, the stopping bounds

$$A' = \frac{1-\beta}{\alpha} \text{ and } B' = \frac{\beta}{1-\alpha}.$$

Then it holds that the SPRT with stopping bounds A' and B' has strength (α', β'), where

$$\alpha' \leq \frac{\alpha}{1-\beta}, \quad \beta' \leq \frac{\beta}{1-\alpha}, \text{ and } \alpha' + \beta' \leq \alpha + \beta.$$

■

Note:

(i) The approximation $A' = \frac{1-\beta}{\alpha}$ and $B' = \frac{\beta}{1-\alpha}$ in Theorem 14.2.4 is called **Wald–Approximation** for the optimal stopping bounds of a SPRT.

(ii) A' and B' are functions of α and β only and do not depend on the pdf's (or pmf's) f_{θ_0} and f_{θ_1}. Therefore, they can be computed once and for all f_{θ_i}'s, $i = 0, 1$.

■

THE END !!!

Index