Large Sample Statistical Methods

July 31, 2020

General instructions

- The textbook is Serfling: Approximation Theorems in Mathematical Statistics.
- Go to the library and consult other books on Asymptotics.
- Try working out problems from the reference books.
- For a better understanding, you should all discuss among yourselves.
- You should not only workout the problems but also write down your solutions. This will help you learn how to communicate your argument.
- **Grading** Final will be cumulative. Homework will not be graded thoroughly. 10% of homework score is on class involvement including attendance.
- My email is rsen@isichennai.res.in and phone no is 9176620249. Please email me unless it is a very urgent situation.
- You are encouraged to meet Arnab on Wednesdays between 5pm and 6pm for further discussions.

Syllabus

1. Review of various modes of convergence of random variables and central limit theorems. Continuous mapping theorem. Cramer-Wold device and multivariate central limit theorem. Scheffe's theorem. Polya's theorem. Slutsky's theorem. Law of iterated logarithm (statement only).

2. Asymptotic distribution of transformed statistics. Delta method. Derivation of the variance stabilizing formula. Asymptotic distribution of functions of sample moments like sample correlation coefficient, coefficient of variation, measures of skewness and kurtosis.

3. Asymptotic distribution of order statistics including extreme order statistics. Asymptotic representation of sample quantiles.

4. Large sample properties of maximum likelihood estimates and the method of scoring.

5. Large sample properties of parameter estimates in linear models.

6. Pearson's chi-square statistic. Chi-square and likelihood ratio test statistics for simple hypotheses related to contingency tables. Heuristic proof for composite hypothesis with contingency tables as examples.

7. Large sample nonparametric inference (e.g., asymptotics of U-statistics and related rank based statistics).

8. Brief introduction to asymptotic efficiency of estimators.

9. (if time permits) Edgeworth expansions.

References

RJ Serfling, Approximation Theorems in Mathematical Statistics
CR Rao Linear, Statistical Inference and its Applications
AW van der Vaart, Asymptotic Statistics
EL Lehmann, Elements of Large Sample Theory
TS Ferguson, A Course in Large Sample Theory.

Review of basic concepts

0.1 Preliminary notations and definitions

Read Serfling 1.1, particularly 1.1.2, definition of quantile function and the accompanying Lemma in 1.1.4.

0.2 Expectation

1. Suppose X is a random variable. Define 2 new rvs as follows:

$$\begin{aligned} X^+(\omega) &= X(\omega) \quad \text{if} \quad X(\omega) > 0 \\ &= 0 \quad \text{otherwise} \\ X^-(\omega) &= -X(\omega) \quad \text{if} \quad X(\omega) < 0 \\ &= 0 \quad \text{otherwise} \\ X &= X^+ - X^- \end{aligned}$$

   We will define expectation of non-negative random variables. For a random variable X, we define $E(X) = E(X^+) - E(X^-)$. when both are finite. Otherwise, $E(X)$ is not defined.

2. If X is a non-negative simple random variable, that is, it takes countably many values $x_1, x_2, \cdots$ with probabilities $p_1, p_2, \cdots$, define $E(X) = \sum_i x_i p_i$

3. For a general non-negative random variable X, define a sequence of random variables X_n as follows

$$X_n(\omega) = \frac{k}{2^n} \quad \text{if} \frac{k}{2^n} < X(\omega) < \frac{k+1}{2^n} \tag{1}$$

   $X_n(\omega) \uparrow X(\omega)$ for each ω and X_n is a non-negative simple random variable. (Exercise) Now we define $E(X) = \lim_{n\to\infty} E(X_n)$

In case of a discrete random variable, the new definition of expectation coincides with our old definition.
If X is a random variable with density f, then it follows from the above definition that $E(X) = \int xf(x)dx$.(Exercise)

0.3 Multiple Random Variables

Properties of multivariate cdf $F_{X_1,X_2,\cdots,X_m} = F$

1. $\lim_{x_j\to-\infty} F(x_1, \cdots, x_m) = 0 \quad \forall j \in 1, \cdots, m$

2. $\lim_{x_1 \to \infty, \cdots, x_m \to \infty} F(x_1, \cdots, x_m) = 1$
3. $F(x_1, \cdots, x_m)$ is increasing in each argument.
4. (Rectangle inequality)For all $(a_1, \cdots, a_m), (b_1, \cdots, b_m)$ with $a_i < b_i \forall i \in 1, \cdots, m$,

$$\sum_{i_1=1}^{2} \cdots \sum_{i_m=1}^{2} (-1)^{i_1 + \cdots + i_m} F(x_{1i_1}, \cdots, x_{mi_m}) \geq 0$$

where $x_{j1} = a_j$ and $x_{j2} = b_j$ for all $j \in 1, \cdots m$.

0.4 Change of variables formula

Suppose $y = g(x)$ is a differentiable and either increasing or decreasing function of x for all values in the support of the random variable X having density $f_X(.)$. Consider the inverse transform $x = h(y)$. Then the density of Y exists and is given by $f_Y(y) = f(h(y))h'(y)$.

Multivariate version: Suppose $\underline{X}$ has density $f_{\underline{X}}(\underline{x})$ and $\underline{Y} = g(\underline{X})$, where $g : \mathbf{R}^k \to \mathbf{R}^k$ is differentiable and has a well-defined inverse. We are interested in determining the density of $\underline{Y}$. Note that $\dot{g}(\underline{x})$ is a $k \times k$ matrix, namely, the matrix of partial derivatives $\partial g_i(\underline{x})/\partial x_j$. This square matrix is called the Jacobian matrix. We will use the term | *Jacobian* | to refer to the determinant of this matrix. Note that the Jacobian of $g(\underline{x})$ is a real-valued function of $\underline{x}$. We denote this function $J_g : \mathbf{R}^k \to \mathbf{R}$. Thus, $J_g(\underline{x}) \overset{\text{def}}{=} \text{Det}\{\dot{g}(\underline{x})\}$.

Suppose now that $h(\underline{y})$ denotes the inverse of $g(\underline{x})$, so that $h \circ g(\underline{x}) = \underline{x}$ for all x and $g \circ h(\underline{y}) = \underline{y}$ for all y.

Then we may write the density function for $\underline{Y}$ in terms of the density function for $\underline{X}$ as follows:

$$f_{\underline{Y}}(\underline{y}) =| J_h(\underline{y}) | f_{\underline{X}} \circ h(\underline{y}) \tag{2}$$

It is important to note, that the Jacobian referred to is the Jacobian of the inverse transformation, not the transformation itself.

0.5 Sums of independent random variables

- Convolution formula (pg 56 Ross)

$$f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(a-y) f_Y(y) dy$$

- Through MGF

0.6 Exercises

(Submission on Jan 14th)

1. Prove Lemma 1.1.4 of Serfling

2. Show that when X is a continuous random variable with density f, then the new definition of expectation in section 1.4 of notes coincides with the old definition $E(X) = \int sf(x)dx$.

3. Prove the properties of monotonicity, subadditivity and continuity, starting from the definition of probability.

4. For any n events, $A_1, \cdots, A_n$, show that $P(\bigcup A_i) = S_1 - S_2 + S_3 - \cdots \pm S_n$. where $S_1 = \sum P(A_i), S_2 = \sum_{i<j} P(A_i \bigcap A_j)$ etc. This is known as Poincare's theorem or the inclusion-exclusion principle. (Jules Henri Poincare was a French Mathematician/Mathemetical Physicist/Philosopher; one of the greatest Mathematicians; 1854 - 1912).

5. Show that

   (a) If $f(x) = o(g(x))$ then $f(x) = O(g(x))$;

   (b) If $f(x) = O(g(x))$ then $O(f(x)) + O(g(x)) = O(g(x))$;

   (c) If $f(x) = O(g(x))$ then $o(f(x)) + o(g(x)) = o(g(x))$;

   (d) $O(f(x))O(g(x)) = O(f(x)g(x))$;

   (e) $o(f(x))O(g(x)) = o(f(x)g(x))$;

6. Show that any function with the properties

   (a) F is non-decreasing

   (b) F is right-continuous

   (c) $\lim_{x\to\infty} F(x) = 1$

   (d) $\lim_{x\to-\infty} F(x) = 0$

   is the cdf of some random variable.

7. Using the definition $F(b) = P(X \leq b)$, show that F need not be left-continuous. Can you think of an alternate definition that will make F left-continuous. Is the new function right continuous?

8. Suppose that f and g are densities on an interval. Show that $f + g$ can not be a density. Show that for any number $0 \leq c \leq 1, cf + (1 - c)g$ is a density.

9. For a non-negative continuous rv X show that $E(X) = \int(1 - F(x))dx$

10. If X is a random variable that takes only non-negative values, then for any $a > 0$, $P(X \geq a) \leq E(X)/a$. Show this for a continuous rv. This is known as Markov inequality.

11. If X is a random variable with mean μ and variance σ^2, then for any value of $k > 0$, $P(\mid X - \mu \mid \geq k) \leq \sigma^2/k^2$. Prove this using Markov inequality. This is known as Chebyshev's inequality.

12. Show that $(E(XY))^2 \leq E(X^2)E(Y^2)$. [Hint: Start with $E(tX+Y)^2$.] This is known as the Cauchy-Schwartz inequality and is useful in statistics to show things like correlation coefficient is less than 1 in absolute value.

13. If g is a convex function then $E(g(X)) \geq g(E(X))$ provided the expectations exist and are finite. This is called Jensen's inequality.[Hint: use Taylor's series expansion]

14. $X \sim N(0,1), Y = X^2$, then use the change of variables formula to show that $Y \sim \chi^2_1$

15. X is a continuous random variable with cdf F and $Y = F(X)$, then what is the distribution of Y?

16. An urn contains nr balls numbered $1, 2, \cdots, n$ where each number $i, 1 \leq i \leq n$, appears on r balls. From the urn, N balls are drawn without replacement. Find the expectation and variance of the number of distinct symbols that appear in the sample.

17. Suppose that $F_1, \cdots, F_k$ are k distribution functions each of one variable. Define $F : \mathbf{R}^k \mapsto \mathbf{R}$ by $F(a) = \prod F_i(a_i)$ where $a = (a_1, \cdots, a_k)$. Show that F is a distribution function.

18. Let X_1, X_2, X_3 be independent with common exponential distribution. Find the joint density of $(X_2 - X_1, X_3 - X_1)$.

19. Suppose that $X_1, X_2, \cdots, X_k$ have joint pdf of the form $g(x'Ax)$ where A is a symmetric positive definite matrix. Let $Y = X'AX$. Show that Y has pdf

$$\frac{\pi^{k/2}}{\sqrt{\mid A \mid}\gamma(k/2)} y^{k/2-1} g(y) \quad \text{for} \quad y > 0$$

20. Suppose that X is a random variable with mgf $M(t)$ finite for all t. Show that $P(X \geq x) \leq e^{-tx}M(t)$ for all $t \geq 0$. Hence deduce that $P(X \geq x) \leq \min_{t \geq 0} e^{-tx}M(t)$. In particular if $X \sim \Gamma(\alpha, \lambda)$, then show that $P(X \geq 2\alpha/\lambda) \leq (2/e)^\alpha$

21. Use the polar coordinate transformation to show that bivariate normal$(0, I)$ is a density.
    The distribution of the distance from the origin is called the Raleigh density and is useful in electrical engineering.

22. Suppose that ϕ_1 and ϕ_2 are bivariate normal densities with means zero, variances 1 and different correlation coefficients. Show that their mixture $\frac{1}{2}(\phi_1+\phi_2)$ is a nonnormal bivariate density with normal marginals.

23. Let $X_1, \cdots, X_n$ be iid uniform (0, 1) and S_n is their sum. Show that its density is given by

$$f_n(x) = \frac{1}{(n-1)!}\left[x^{n-1} - \left(\begin{array}{c} n \\ 1 \end{array}\right)(x-1)_+^{n-1} + \left(\begin{array}{c} n \\ 2 \end{array}\right)(x-2)_+^{n-1} - \cdots\right]$$

    for $0 < x < n$. Here a_+ denotes a or 0 according as $a > 0$ or not. f_1 is discontinuous at some points. f_2 is continuous but not differentiable at some points. f_3 is differentiable at all points but second derivative does not exist at some points.

1 Concepts of convergence of a sequence of r.v.s.

1.1 Topics covered

1. Different modes of convergence: definition
2. Relationships: implications and counterexamples
   (a) Convergence in probability does not imply convergence with probability one. Counterexample
   (b) O_P and o_P (1.2.5). Convergence in distribution implies bounded in probability.
   (c) Sum and product of sequences of RV's converging with probability one.
   (d) Convergence in probability implies convergence in distribution.
   (e) Sum and product of sequences of RV's converging in probability.
   (f) Sum and product of sequences of RV's converging in r-th moment. Minkowski inequality.
   (g) Convergence in probability does not imply convergence in r-th moment. Counterexample
   (h) Convergence in distribution does not imply convergence in probability. Counterexample
   (i) Almost sure convergence in union intersection notation
   (j) Convergence with probability one implies convergence in probability.
   (k) Convergence in r-th moment implies convergence in probability.
   (l) Continuous function of sequence of RV's converging in probability.
   (m) MCT, DCT, Fatou, Levy
3. Scheffe 1.5.1C
4. Cramer-Wold device 1.5.2
5. Polya 1.5.3
6. Slutsky 1.5.4
7. LLN 1.8
8. Continuous mapping theorem 1.7 (iii) [needs Skorohod 1.6.3 and countable discontinuities of F^{-1} 1.5.6].
9. CLT (univariate and multivariate) 1.9
10. Law of iterated logarithms 1.10

1.2 Exercises

(Submission on Jan 21st)

1. Read 1.5.5 and 1.15 of textbook.

2. If for each $\epsilon > 0, \sum P(|X_n - X| > \epsilon) < \infty$, then show that $X_n \to X$ a.e. If $\sum E|X_n - X|^2 < \infty$, then show that $X_n \to X$ a.e.

3. If $X_n \to X$ in probability and $E(X_n - Y_n)^2 \to 0$ then show that $Y_n \to X$ in probability.

4. If $X_n \to X$ in probability and $X_n \to Y$ in probability, then show that $P(X = Y) = 1$.

5. Let P_n put mass $1/n$ at each of the points $0, 1/n, \cdots, (n-1)/n$. Show $P_n \Rightarrow U$, the Uniform (0,1) probability. What if P_n puts mass $1/(n+1)$ at each of the points $0, 1/n, \cdots, n/n$.

6. Suppose P_n puts mass $2k/n(n+1)$ at the point k/n for $k = 1, 2, \cdots, n$. Does this sequence of probabilities converge?

7. If F is discrete, then show that $F_n \Rightarrow F$ iff for each x in the support of F, $P(X_n = x) \to P(X = x)$.

8. $X_n \Rightarrow X$. Show by examples that each X_n may be integrable but X may not be. X may be integrable but NONE of the X_n is integrable.

9. Give an example of a sequence of densities that converge point-wise to a function that is NOT a density.

2 Asymptotic Distribution of Transformed Statistics

2.1 Delta method

(pp 240-245 Casella & Berger;pp 58-59 Lehmann & Casella) In the simplest form of the central limit theorem, we consider a sequence $X_1, X_2, ...X_n$ of independent and identically distributed (univariate) random variables with mean μ and finite variance σ^2. In this case, the central limit theorem states that

$$\sqrt{n}(\bar{X_n} - \mu) \Rightarrow \mathcal{N}(0, \sigma^2) \qquad (3)$$

In this section, we wish to consider the asymptotic distribution of, say, some function of $\bar{X_n}$. In the simplest case, the answer depends on results already known: Consider a linear function $h(t) = at+b$ for some known constants a and b. Clearly $\mathrm{E}(h(\bar{X_n})) = a\mu+b = h(\mu)$ by the linearity of the expectation operator. Therefore, it is reasonable to ask whether $\sqrt{n}(h(\bar{X_n}) - h(\mu))$ tends to some distribution as $n \to \infty$. But the linearity of $h(t)$ allows one to write

$$\sqrt{n}(h(\bar{X_n}) - h(\mu)) = a\sqrt{n}(\bar{X_n} - \mu) \tag{4}$$

We conclude that

$$\sqrt{n}(h(\bar{X_n}) - h(\mu)) \Rightarrow \mathcal{N}(0, a^2\sigma^2) \tag{5}$$

None of the preceding development is especially deep; one might even say that it is obvious that a linear transformation of the random variable T_n alters its asymptotic distribution by a constant multiple. Yet what if the function $h(t)$ is nonlinear? It is in this nonlinear case that a strong understanding of the argument above, as simple as it may be, pays real dividends. For if T_n is consistent for θ (say), then we know that, roughly speaking, T_n will be very close to θ for large n. Therefore, the only meaningful aspect of the behavior of $h(t)$ is its behavior in a small neighborhood of θ. And in a small neighborhood of θ, $h(\theta)$ may be considered to be roughly a linear function. Formally we use the Taylor expansion to obtain the following result:

Theorem 1 (First Order Delta Method)**.** *If*

$$\sqrt{n}(T_n - \theta) \Rightarrow \mathcal{N}(0, \tau^2) \tag{6}$$

then

$$\sqrt{n}(h(T_n) - h(\theta)) \Rightarrow \mathcal{N}(0, \tau^2(h'(\theta))^2) \tag{7}$$

provided $h'(\theta)$ exists and is not zero.

Proof. Step 1: It follows from equation (6) that $T_n \to \theta$ in probability.
For a fixed $\epsilon > 0$, $\quad P(\mid T_n - \theta \mid > \epsilon) = P(\mid X_n \mid > \sqrt{n}\epsilon)$ where $X_n = \sqrt{n}(T_n - \theta)$.
From equation (6), $X_n \Rightarrow X$ where $X \sim \mathcal{N}(0, \tau^2)$.
For given δ, one can find N_1, so that for $n > N_1$,

$$P(\mid X_n \mid > \sqrt{n}\epsilon) < P(\mid X \mid > \sqrt{n}\epsilon) + \delta/2$$

One can find $N_2 > N_1$ so that for $n > N_2$, $\quad P(\mid X \mid > \sqrt{n}\epsilon) < \delta/2$.
Hence for $n > N_2$, $\quad P(\mid T_n - \theta \mid > \epsilon) < \delta$.
Step 2: Consider the Taylor expansion of h around θ.

$$h(x) = h(\theta) + (x - \theta)(h'(\theta) + r) \tag{8}$$

where $r \to 0$ as $x \to \theta$.
Define R_n as the remainder in

$$h(T_n) = h(\theta) + (T_n - \theta)(h'(\theta) + R_n) \tag{9}$$

By step 1, $T_n \to \theta$ in probability.
Hence $R_n \to 0$ in probability.
This implies $h'(\theta) + R_n \to h'(\theta)$ in probability.

Step 3: The result follows by applying Slutsky's theorem to $\sqrt{n}(h(T_n) - h(\theta))$.
$\sqrt{n}(h(T_n) - h(\theta)) = \sqrt{n}(T_n - \theta) \times (h'(\theta) + R_n)$.
Let $Y_n = (h'(\theta) + R_n)$ and $X_n = \sqrt{n}(T_n - \theta)$ as above.
$X_n \Rightarrow X$ and $Y_n \to c$ in probability where $c = h'(\theta), X \sim \mathcal{N}(0, \tau^2)$.
By Slutsky's theorem, $\sqrt{n}(h(T_n) - h(\theta)) = Y_n X_n \Rightarrow cX$.
The distribution of cX is $\mathcal{N}(0, \tau^2(h'(\theta))^2)$. □

Example 1(Estimating the odds) Let $X_i, i = 1, 2, \cdots, n$ be independent Bernoulli(p) random variables and let $T_n = \frac{1}{n}\sum_{i=1}^n X_i$. A popular parameter is the odds $\frac{p}{1-p}$. For example, if the data represent the outcomes of a medical treatment with $p = 2/3$, then a person has odds 2:1 of getting better. We consider the estimate $\frac{T_n}{1-T_n}$ for the parameter $h(p) = \frac{p}{1-p}$. Since $h'(p) = \frac{1}{(1-p)^2}$, it follows from Theorem 1 that

$$\sqrt{n}\left[\frac{T_n}{1-T_n} - \frac{p}{1-p}\right] \Rightarrow \mathcal{N}(0, \left(\frac{1}{(1-p)^2}\right)^2 p(1-p) = \frac{p}{(1-p)^3}) \tag{10}$$

Example 2(Exponential Rate) Let $X_i, i = 1, 2, \cdots, n$ be independent Exponential(λ) random variables and let $T_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then by CLT,

$$\sqrt{n}(T_n - \lambda) \Rightarrow \mathcal{N}(0, \lambda^2) \tag{11}$$

Suppose we are now interested in the large sample behavior of the estimate $\frac{1}{T_n}$ of the rate $h(\lambda) = \frac{1}{\lambda}$.
Since $h'(\lambda) = -\frac{1}{\lambda^2}$, it follows from Theorem 1 that

$$\sqrt{n}(\frac{1}{T_n} - \frac{1}{\lambda}) \Rightarrow \mathcal{N}(0, \left(-\frac{1}{\lambda^2}\right)\lambda^2 = \frac{1}{\lambda^2}) \tag{12}$$

Example 3 (Binomial Variance) Let $X_i, i = 1, 2, \cdots, n$ be independent Bernoulli random variables and let $T_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then by CLT,

$$\sqrt{n}(T_n - p) \Rightarrow \mathcal{N}(0, p(1-p)) \tag{13}$$

Suppose we are now interested in the large sample behavior of the estimate $T_n(1-T_n)$ of the variance $h(p) = p(1-p)$.
Since $h'(p) = 1 - 2p$, it follows from Theorem 1, when $p \neq 1/2$, that

$$\sqrt{n}(T_n(1-T_n) - p(1-p)) \Rightarrow \mathcal{N}(0, (1-2p)^2 p(1-p)) \tag{14}$$

What happens when $h'(\theta) = 0$?

Theorem 2 (Second Order Delta Method)**.** *If*

$$\sqrt{n}(T_n - \theta) \Rightarrow \mathcal{N}(0, \tau^2) \quad \text{and} \quad h'(\theta) = 0 \tag{15}$$

then

$$n(h(T_n) - h(\theta)) \Rightarrow \frac{1}{2}\tau^2 h''(\theta)\chi_1^2 \tag{16}$$

Proof. Consider the Taylor expansion of $h(T_n)$ around $h(\theta)$ upto the second term.

$$h(T_n) = h(\theta) + (T_n - \theta)h'(\theta) + \frac{1}{2}(T_n - \theta)^2(h''(\theta) + R_n) \tag{17}$$

where $R_n \to 0$ as $T_n \to \theta$.

Step 1: It follows from equation (15) that $T_n \to \theta$ in probability.
Hence $R_n \to 0$ in probability. This implies $h''(\theta) + R_n \to h''(\theta)$ in probability.

Step 2: $\frac{1}{\tau^2}n(T_n - \theta)^2 \Rightarrow \chi_1^2$.
This follows from equation (15) after dividing by τ and squaring a standard normal random variable.

Step 3: The result follows by applying Slutsky's theorem to $n(h(T_n) - h(\theta))$.
$n(h(T_n) - h(\theta)) = n(T_n - \theta)^2 \times (h''(\theta) + R_n)$ since $h'(\theta) = 0$.
Let $Y_n = \tau^2(h''(\theta) + R_n)$ and $X_n = \frac{1}{\tau^2}n(T_n - \theta)^2$.
$X_n \Rightarrow X$ and $Y_n \to c$ in probability where $c = \tau^2 h''(\theta), X \sim \chi_1^2$.
By Slutsky's theorem, $n(h(T_n) - h(\theta)) = Y_n X_n \Rightarrow cX$.
The distribution of cX is $\tau^2 h''(\theta)\chi_1^2$. □

Example 3'(Binomial Variance at $p = 1/2$**)** For $h(p) = p(1-p)$, we have at $p = 1/2, \quad h'(1/2) = 0$ and $h''(1/2) = -2$. Hence from theorem 2, at $p = 1/2$,

$$n\left[T_n(1 - T_n) - \frac{1}{4}\right] \Rightarrow -\frac{1}{4}\chi_1^2 \tag{18}$$

Although the equation (18) might appear strange, note that $T_n(1 - T_n) \leq 1/4$, so the left side is always negative. An equivalent form is

$$4n\left[\frac{1}{4} - T_n(1 - T_n)\right] \Rightarrow \chi_1^2 \tag{19}$$

Corollary 1 (Bias of $h(\theta)$)**.** *The asymptotic bias of* $h(T_n)$ *is*

$$h'(\theta)\text{ABias}(T_n) + \frac{h''(\theta)}{2}\text{AMSE}(T_n) \tag{20}$$

where ABias and AMSE are the asymptotic bias and MSE respectively.

Proof. It follows from equation (17) that

$$\mathrm{Bias}(h(T_n)) = h'(\theta)\mathrm{Bias}(T_n) + \frac{1}{2}h''(\theta)\mathrm{MSE}(T_n) + \frac{1}{2}E\{R_n(T_n-\theta)^2\} \tag{21}$$

The result follows from this intuitively. A rigorous proof would require uniform integrability conditions and we refrain from that in this course. □

One implication is that even if T_n is unbiased, $h(T_n)$ is not: taking a non-linear function typically introduces a bias of the order of AMSE. For eg, for $T_n = \bar{X}$, for iid observations, bias=0, variance=$\frac{\sigma^2}{n}$. So the AMSE of T_n, hence bias of $h(T_n)$ is order of $\frac{1}{n}$.

We now present a result on multivariate Delta method without proof. The proof relies on Wold device. The interested reader can find more details in Lehmann and Casella (Sec 1.8).

Theorem 3 (Multivariate Delta Method)**.** *Let* $(X_{1\nu}, \cdots, X_{s\nu}), \quad, \nu = 1, \cdots, n$ *be* n *independent* s*-tuples of random variables with* $\mathrm{E}(X_{i\nu}) = \xi_i$ *and* $\mathrm{Cov}(X_{i\nu}, X_{j\nu}) = \sigma_{ij}$*. Let* $\bar{X}_i = \sum_{\nu=1}^{n} X_{i\nu}/n$*, and suppose that* h *is a real valued function of* s *arguments with continuous first partial derivatives. Then*

$$\sqrt{n}\left[h(\bar{X}_1, \cdots, \bar{X}_s) - h(\xi_1, \cdots, \xi_s)\right] \Rightarrow \mathcal{N}(0, v^2), \quad \text{where} \quad v^2 = \sum_{i=1}^{s}\sum_{j=1}^{s}\sigma_{ij}\frac{\partial h}{\partial \xi_i}\frac{\partial h}{\partial \xi_j} \tag{22}$$

Corollary 2 (Multi-dimensional Delta method)**.** *Assume the conditions of Theorem 3, but now* h *is a* k*-dimensional function. Let* $\mathbf{D}h_{k\times s} = ((\frac{\partial h_i}{\partial \xi_j}))_{i,j}$ *be the matrix of partial derivatives of* h *and* $\Sigma_{s\times s} = ((\sigma_{ij}))_{i,j}$ *be the covariance matrix of* $(X_{1\nu}, \cdots, X_{s\nu})$*. It follows from Theorem 3 that*

$$\sqrt{n}\left[h(\bar{X}_1, \cdots, \bar{X}_s) - h(\xi_1, \cdots, \xi_s)\right] \Rightarrow \mathcal{N}_k(0, v^2), \quad \text{where} \quad v^2 = \mathbf{D}h\Sigma\mathbf{D}h^T \tag{23}$$

Example 4 (Variance of Variance estimator) Suppose $X_1, \cdots, X_n$ are iid random variables with mean μ and variance σ^2. We are interested in the joint distribution of $(\bar{X}, s^2 = \frac{1}{n}\sum(X_i - \bar{X})^2)$, the estimator of (μ, σ^2). Denoting $\mathrm{E}(X^k)$ by m_k, we have

$$\begin{aligned}
\mathrm{E}(\bar{X}) &= m_1 \\
\mathrm{E}(\bar{X^2}) &= m_2 \\
\mathrm{Cov}(\bar{X}, \bar{X^2}) &= (m_3 - m_1 m_2)/n \\
\mathrm{Var}(\bar{X}) &= (m_2 - m_1^2)/n \\
\mathrm{Var}(\bar{X^2}) &= (m_4 - m_2^2)/n
\end{aligned}$$

The parameter of interest is $(\mu, \sigma^2) = h(m_1, m_2) = (m_1, m_2 - m_1^2)$. The derivatives of h are $\frac{\partial h}{\partial m_1} = (1, -2m_1)$ and $\frac{\partial h}{\partial m_2} = (0, 1)$.

$$\sqrt{n}\left[h(\bar{X}, \bar{X^2}) - h(m_1, m_2)\right] \Rightarrow \mathcal{N}(0, \upsilon^2), \quad \text{where} \tag{24}$$

$$\begin{aligned}
\upsilon^2 = \mathbf{D}h\Sigma\mathbf{D}h^T &= \begin{pmatrix} 1 & 0 \\ -2m_1 & 1 \end{pmatrix} \begin{pmatrix} m_2 - m_1^2 & m_3 - m_1 m_2 \\ m_3 - m_1 m_2 & m_4 - m_2^2 \end{pmatrix} \begin{pmatrix} 1 & -2m_1 \\ 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} m_2 - m_1^2 & 2m_1^3 + m_3 - 3m_1 m_2 \\ 2m_1^3 + m_3 - 3m_1 m_2 & -4m_1^4 + 8m_1^2 m_2 + m_4 - m_2^2 - 4m_1 m_3 \end{pmatrix}
\end{aligned}$$

2.2 Transformations and variance stabilizing formula

(Rao 6g; pp 76 Lehmann & Casella; pp87 Lehmann)

Usual Assumption in ANOVA and Regression is that the variance of each observation is the same. In many cases, the variance is not constant, but is related to the mean. For example,

- Poisson Data (Counts of events): $\mathrm{E}(X) = \mathrm{Var}(X) = \mu$
- Binomial Data (and Percents): $\mathrm{E}(X) = np, \quad \mathrm{Var}(X) = np(1-p)$
- Power relationship: $\mathrm{E}(X) = \mu, \quad \mathrm{Var}(X) = \sigma^2 = \alpha^2 \mu^{2\beta}$
- General Case: $\mathrm{E}(X) = \theta, \quad \mathrm{Var}(X) = \sigma^2(\theta)$

Random variable X has mean θ and variance $\sigma(\theta)$. We want a transformation $h(X)$ that has constant (does not depend on θ) variance. In the general case above, writing out a first-order Taylor series expansion:

$$\begin{aligned}
h(X) &\approx h(\theta) + (X - \theta)h'(\theta) \\
\Rightarrow h(X) - h(\theta) &\approx (X - \theta)h'(\theta) \\
\Rightarrow [h(X) - h(\theta)]^2 &\approx (X - \theta)^2 (h'(\theta))^2
\end{aligned}$$

Now taking expectations on both sides, we get $\mathrm{Var}[h(X)] \approx \mathrm{Var}(X)[h'(\theta)]^2 = \sigma^2(\theta)[h'(\theta)]^2$. Now, let

$$h(\theta) = c \int \frac{1}{\sigma(\theta)} d\theta. \tag{25}$$

$\Rightarrow \mathrm{Var}[h(X)] = \sigma^2(\theta) c^2 [\frac{d}{d\theta} \int \frac{1}{\sigma(\theta)} d\theta]^2 = c^2$. Thus, taking this transformation on X gives a random variable with an approximately constant variance.

In particular, if one has asymptotic normality, that is $\sqrt{n}(T_n - \theta) \Rightarrow \mathcal{N}(0, \sigma^2(\theta))$, then from the Delta method, $\sqrt{n}(h(T_n) - h(\theta)) \Rightarrow \mathcal{N}(0, (h'(\theta))^2 \sigma^2(\theta))$. Same argument above implies taking $h(.)$ given by equation (25) will yield $\sqrt{n}(h(T_n) - h(\theta)) \Rightarrow \mathcal{N}(0, c^2)$.

The extensive literature on variance-stabilizing transformations and transformations to approximate normality is reviewed in Hoyle(International Statistical Review, 1973). Two later references are Efron(Annals of Statistics, 1982) and Bar-Lev and Enis (Statistics and Probability Letters, 1990).

Example 5 (Power Relationship) Suppose $\sigma^2(\theta) = \alpha^2\theta^{2\beta}$.
Case 1: $\beta \neq 1$
$h(\theta) = \int \frac{1}{\sigma(\theta)} d\theta = \int \frac{1}{\alpha\theta^\beta} d\theta = \frac{1}{\alpha}\frac{\theta^{-\beta+1}}{-\beta+1} = c\theta^{1-\beta}$
Case 2: $\beta = 1$
$h(\theta) = \int \frac{1}{\sigma(\theta)} d\theta = \int \frac{1}{\alpha\theta} d\theta = \frac{1}{\alpha}\text{log}(\theta) = c\text{log}(\theta)$

Note This family of transformations are known as the **Box-Cox family of power transformations** and play an important role in applied statistics. For more details and interesting discussions, see Bickel and Doksum, Box and Cox 1982, Hinkley and Runger 1984. To estimate β:

$$\sigma = \alpha\theta^\beta, \Rightarrow \text{log}(\sigma) = \text{log}(\alpha) + \beta\text{log}(\theta)$$

Thus, we can estimate β by regressing $\text{log}(s_i)$ on $\text{log}(\bar{X}_i)$. See example in Kuehl (p. 137), for this, where the logarithmic transformation is suggested. ($\hat{\beta} = 0.99$).

Example 6 (Poisson data) $X_i \sim \text{Poisson}(\lambda), T_n = \bar{X}_n$. Then $\theta = \lambda, \sigma(\theta) = \sqrt{\lambda}$. Then (25) implies

$$h(\theta) = c\int \frac{1}{\sigma(\theta)} d\theta = c\int \frac{1}{\sqrt{\theta}} d\theta = 2c\sqrt{\theta}$$

Thus $h(T_n) = 2c\sqrt{T_n}$. Putting $c = 1$, we have

$$2\sqrt{n}(\sqrt{\bar{X}} - \sqrt{\lambda}) \Rightarrow \mathcal{N}(0,1)$$

Example 6' (Poisson Confidence Intervals) It follows from Example 2 that for any λ,

$$P(\mid \sqrt{\bar{X}} - \sqrt{\lambda} \mid < \frac{z_{\alpha/2}}{2\sqrt{n}}) \to 1-\alpha, \quad \text{as} \quad n \to \infty \quad \text{where} \quad z \text{ denotes the Normal quantile.}$$

This provides the following approximate (the results are asymptotic, so for any finite n they are approximate) confidence intervals for $\sqrt{\lambda}$:

$$\sqrt{\bar{X}} - \frac{z_{\alpha/2}}{2\sqrt{n}} < \sqrt{\lambda} < \sqrt{\bar{X}} + \frac{z_{\alpha/2}}{2\sqrt{n}}$$

The lower end point can be negative since $\bar{X}$ can be arbitrarily close to zero. However, for any positive λ we have $\bar{X} \xrightarrow{P} \lambda$ and $\frac{z_{\alpha/2}}{2\sqrt{n}} \to 0$. So the probability of a negative end-point tends to 0 as $n \to \infty$. When this does occur one would replace the left end point by 0. From this modified interval, one can obtain the corresponding interval for λ at the same level by squaring. This leads to the approximate level α confidence interval $\underline{\lambda} < \lambda < \overline{\lambda}$ where

$$\underline{\lambda} = \begin{cases} \left(\sqrt{\bar{X}} - \frac{z_{\alpha/2}}{2\sqrt{n}}\right)^2 & \text{if} \quad \sqrt{\bar{X}} - \frac{z_{\alpha/2}}{2\sqrt{n}} > 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \overline{\lambda} = \left(\sqrt{\bar{X}} + \frac{z_{\alpha/2}}{2\sqrt{n}}\right)^2$$

Example 7 (Binomial data) $X_i \sim \text{Bernoulli}(p)$, $T_n = \bar{X}_n$. Then $\theta = p, \sigma(\theta) = \sqrt{p(1-p)}$. Then (25) implies

$$h(\theta) = c\int \frac{1}{\sigma(\theta)} d\theta = c\int \frac{1}{\sqrt{\theta(1-\theta)}} d\theta = c \arcsin\sqrt{\theta}$$

Thus $h(T_n) = c\,\text{arcsin}(\sqrt{T_n})$.

Example 8 (Chi-squared data) Let $Y_i = X_i^2$ where $X_i \sim \mathcal{N}(0, \sigma^2)$ and $T_n = \bar{Y_n}$. Then $\theta = \sigma^2, \sigma(\theta) = \text{Var}(Y_i) = \text{E}(X_i^4) - (\text{E}(X_i^2))^2 = 3\theta^2 - \theta^2 = 2\theta^2$. Then (25) implies

$$h(\theta) = c\int \frac{1}{\sigma(\theta)} d\theta = c\int \frac{1}{\sqrt{2}\theta} d\theta = \frac{c}{\sqrt{2}} \log(\theta)$$

Thus $h(T_n) = \frac{c}{\sqrt{2}} \log(T_n)$. Putting $c = 1$, we have

$$\sqrt{\frac{n}{2}}(\log(\bar{Y}) - \log\sigma^2) \Rightarrow \mathcal{N}(0, 1)$$

2.3 Asymptotic distributions of functions of sample moments

The weak law of large numbers tells us that if $X_1, X_2, \cdots, X_n$ are independent and identically distributed with $\text{E} \mid X_1 \mid^k < 1$, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} \text{E}X_1^k. \tag{26}$$

That is, sample moments are (weakly) consistent. For example, the sample variance, which we define as

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2 \tag{27}$$

is consistent for $\text{Var}(X_i) = \text{E}X_i^2 - (\text{E}(X_i))^2$. However, consistency is not the end of the story. The central limit theorem and the delta method will prove very useful in deriving asymptotic distribution results about sample moments.

Example 9 (Distribution of sample T statistic) Suppose $X_1, X_2, \cdots, X_n$ are iid with $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Define s_n^2 as in Equation (27), and let

$$T_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n}. \tag{28}$$

Letting $A_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ and $B_n = \sigma/s_n$, we obtain $Tn = A_n B_n$. Therefore, since $A_n \Rightarrow \mathrm{N}(0,1)$ by the central limit theorem and $B_n \overset{P}{\to} 1$ by the weak law of large numbers, Slutsky's theorem implies that $T_n \Rightarrow \mathcal{N}(0,1)$. In other words, T statistics are asymptotically normal under the null hypothesis.

Example 10 (Sample Correlation) Suppose that $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ are iid vectors with $\mathrm{E}(X_i^4) < \infty$ and $\mathrm{E}(Y_i^4) < \infty$. For the sake of simplicity, we will assume without loss of generality that $\mathrm{E}(X_i) = \mathrm{E}(Y_i) = 0$ (alternatively, we could base all of the following derivations on the centered versions of the random variables). We wish to find the asymptotic distribution of the sample correlation coefficient, r. If we let

$$\begin{pmatrix} m_x \\ m_y \\ m_{xx} \\ m_{yy} \\ m_{xy} \end{pmatrix} = \frac{1}{n} \begin{pmatrix} \sum_{i=1}^n X_i \\ \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i^2 \\ \sum_{i=1}^n Y_i^2 \\ \sum_{i=1}^n X_i Y_i \end{pmatrix} \tag{29}$$

and $s_x^2 = m_{xx} - m_x^2$, $s_y^2 = m_{yy} - m_y^2$, and $s_{xy} = m_{xy} - m_x m_y$. Then $r = s_{xy}/(s_x s_y)$. According to the central limit theorem,

$$\sqrt{n} \left[\begin{pmatrix} m_x \\ m_y \\ m_{xx} \\ m_{yy} \\ m_{xy} \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ \sigma_x^2 \\ \sigma_y^2 \\ \sigma_{xy} \end{pmatrix} \right] \Rightarrow \mathcal{N}_5 \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \cdots & \mathrm{Cov}(X_1, X_1 Y_1) \\ \mathrm{Cov}(Y_1, X_1) & \cdots & \mathrm{Cov}(Y_1, X_1 Y_1) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_1 Y_1, X_1) & \cdots & \mathrm{Cov}(X_1 Y_1, X_1 Y_1) \end{pmatrix} \right) \tag{30}$$

Let Σ denote the covariance matrix in expression (30). Define a function $h : \mathbf{R}^5 \to \mathbf{R}^3$ such that h applied to the vector of moments in Equation (29) yields the vector (s_x^2, s_y^2, s_{xy}). Then

$$Dh \begin{pmatrix} a \\ b \\ c \\ d \\ e \end{pmatrix} = \begin{pmatrix} -2a & 0 & -b \\ 0 & -2b & -a \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}^T \tag{31}$$

Therefore, if we let

$$\Sigma^* = \left[Dh \begin{pmatrix} 0 \\ 0 \\ \sigma_x^2 \\ \sigma_y^2 \\ \sigma_{xy} \end{pmatrix} \right] \Sigma \left[Dh \begin{pmatrix} 0 \\ 0 \\ \sigma_x^2 \\ \sigma_y^2 \\ \sigma_{xy} \end{pmatrix} \right]^T = \begin{pmatrix} \mathrm{Cov}(X^2, X^2) & \mathrm{Cov}(X^2, Y^2) & \mathrm{Cov}(X^2, XY) \\ \mathrm{Cov}(Y^2, X^2) & \mathrm{Cov}(Y^2, Y^2) & \mathrm{Cov}(Y^2, X_1 Y_1) \\ \mathrm{Cov}(XY, X^2) & \mathrm{Cov}(XY, Y^2) & \mathrm{Cov}(XY, XY) \end{pmatrix} \tag{32}$$

then by the delta method,

$$\sqrt{n} \left[\begin{pmatrix} s_x^2 \\ s_y^2 \\ s_{xy} \end{pmatrix} - \begin{pmatrix} \sigma_x^2 \\ \sigma_y^2 \\ \sigma_{xy} \end{pmatrix} \right] \Rightarrow \mathcal{N}_3(0, \Sigma^*) \tag{33}$$

Next, define the function $g(a, b, c) = c/\sqrt{ab}$, so that we have $g(s_x^2, s_y^2, s_{xy}) = r$. Then

$$[Dg(a, b, c)] = \frac{1}{2}(\frac{-c}{\sqrt{a^3 b}}, \frac{-c}{\sqrt{ab^3}}, \frac{2}{\sqrt{ab}}) \tag{34}$$

so that

$$A := [Dg(\sigma_x^2, \sigma_y^2, \sigma_{xy})] = (\frac{-\sigma_{xy}}{2\sigma_x^3 \sigma_y}, \frac{-\sigma_{xy}}{2\sigma_x \sigma_y^3}, \frac{1}{\sigma_x \sigma_y}) = (\frac{-\rho}{2\sigma_x^2}, \frac{-\rho}{2\sigma_y^2}, \frac{1}{\sigma_x \sigma_y}) \tag{35}$$

Therefore, using the delta method once again yields

$$\sqrt{n}(r - \rho) \Rightarrow \mathcal{N}(0, A\Sigma^* A^T). \tag{36}$$

Example 4 Consider the special case of bivariate normal (X_i, Y_i). In this case, we may derive

$$\Sigma^* = \begin{pmatrix} 2\sigma_x^4 & 2\rho^2 \sigma_x^2 \sigma_y^2 & 2\rho \sigma_x^3 \sigma_y \\ 2\rho^2 \sigma_x^2 \sigma_y^2 & 2\sigma_y^4 & 2\rho \sigma_x \sigma_y^3 \\ 2\rho \sigma_x^3 \sigma_y & 2\rho \sigma_x \sigma_y^3 & (1 + \rho^2)\sigma_x^2 \sigma^2 y \end{pmatrix} \tag{37}$$

In this case, $A\Sigma^* A^T = (1 - \rho^2)^2$, which implies that

$$\sqrt{n}(r - \rho) \Rightarrow \mathcal{N}(0, (1 - \rho^2)^2) \tag{38}$$

In the normal case, we may derive a variance-stabilizing transformation. According to Equation (36), we should find a function $f(x)$ satisfying $f'(x) = (1 - x^2)^{-1}$. Since

$$\frac{1}{1 - x^2} = \frac{1}{2(1 - x)} + \frac{1}{2(1 + x)},$$

we integrate to obtain

$$f(x) = \frac{1}{2} \log \frac{1 + x}{1 - x}.$$

This is called Fisher's transformation; we conclude that

$$\sqrt{n}(\frac{1}{2} \log \frac{1 + r}{1 - r} - \frac{1}{2} \log \frac{1 + \rho}{1 - \rho}) \Rightarrow \mathcal{N}(0, 1)$$

Exercises

1. **Ratio Estimator** Suppose X and Y are random variables with nonzero means μ_X and μ_Y respectively. The function to be estimated is $h(\mu_X, \mu_Y) = \frac{\mu_X}{\mu_Y}$ and we have random samples $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_n$. Conclude from theorem 3

$$\begin{aligned}\sqrt{n}\left[h(\bar{X}, \bar{Y}) - h(\mu_X, \mu_Y)\right] &\Rightarrow \mathcal{N}(0, v^2), \\ \text{where} \quad v^2 &= \text{Var}(X)\frac{1}{\mu_Y^2} + \text{Var}(Y)\frac{\mu_X^2}{\mu_Y^4} - 2\text{Cov}(X,Y)\frac{\mu_X}{\mu_Y^3}\end{aligned}$$

2. **k-th order Delta Method** If

$$\sqrt{n}(T_n - \theta) \Rightarrow \mathcal{N}(0, \tau^2) \quad \text{and} \quad h'(\theta) = h''(\theta) = \cdots = h^{(k-1)}(\theta) = 0 \tag{39}$$

then

$$n^{k/2}(h(T_n) - h(\theta)) \Rightarrow \frac{1}{k!}\tau^k h^{(k)}(\theta)\left[\mathcal{N}(0,1)\right]^k \tag{40}$$

3. **Direct Evaluation of Example 2'** In the setting of example 2', denote $Y_n = T_n - 1/2$. Show that $2\sqrt{n}Y_n \Rightarrow \mathcal{N}(0,1)$. Show that $4n\left[\frac{1}{4} - T_n(1-T_n)\right] = 4nY_n^2$ and hence has asymptotic χ_1^2 distribution.

4. Suppose that $X_1, X_2, \cdots, X_n$ are iid $\mathcal{N}(0, \sigma^2)$ random variables.
(a) Based on the result of Example 4, give an approximate test at $\alpha = 0.05$ for $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_a : \sigma^2 \neq \sigma_0^2$.
(Hint: For $\mathcal{N}(0, \sigma^2)$ random variable $m_1 = 0, m_2 = \sigma^2, m_3 = 0, m_4 = 3\sigma^4$.)
(b) For $n = 25$, estimate the true level of the test in part (a) for $\sigma_0^2 = 1$ by simulating 5000 samples of size $n = 25$ from the null distribution. Report the proportion of cases in which you reject the null hypothesis according to your test (ideally, this proportion will be about .05).

5. Verify expressions (37) and (38).

6. Derive the asymptotic distribution of the coefficient of variation, sample skewness and sample kurtosis.

7. Assume $(X_1, Y_1), \cdots, (X_n, Y_n)$ are iid from some bivariate normal distribution. Let ρ denote the population correlation coefficient and r the sample correlation coefficient.
(a) Describe a test of $H_0 : \rho = 0$ against $H_a : \rho \neq 0$ based on the fact that $\sqrt{n}[f(r) - f(\rho)] \Rightarrow \mathcal{N}(0,1)$, where $f(x)$ is Fisher's transformation $f(x) = (1/2)\text{log}[(1+x)/(1-x)]$. Use $\alpha = .05$.
(b) Based on 5000 repetitions each, estimate the actual level for this test in the case when $\text{E}(X_i) = \text{E}(Y_i) = 0, \text{Var}(X_i) = \text{Var}(Y_i) = 1$, and $n \in 3, 5, 10, 20$.

8. Suppose that X and Y are jointly distributed such that X and Y are Bernoulli (1/2) random variables with $\mathrm{P}(XY = 1) = \theta$ for $\theta \in (0, 1/2)$.
   Let $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ be iid with (X_i, Y_i) distributed as (X, Y).
   (a) Find the asymptotic distribution of $\sqrt{n}[(X_n, Y_n) - (1/2, 1/2)]$.
   (b) If r_n is the sample correlation coefficient for a sample of size n, find the asymptotic distribution of $\sqrt{n}(r_n - \rho)$.
   (c) Find a variance stabilizing transformation for r_n.
   (d) Based on your answer to part (c), construct a 95% confidence interval for ρ.
   (e) For each combination of $n \in 5, 20$ and $\rho \in .05, .25, .45$, estimate the true coverage probability of the confidence interval in part (d) by simulating 5000 samples and the corresponding confidence intervals. One problem you will face is that in some samples, the sample correlation coefficient is undefined because with positive probability each of the X_i or Y_i will be the same. In such cases, consider the confidence interval to be undefined and the true parameter therefore not contained therein.
   Hint: To generate a sample of (X, Y), first simulate the X's from their marginal distribution, then simulate the Y 's according to the conditional distribution of Y given X. To obtain this conditional distribution, find $\mathrm{P}(Y = 1|X = 1)$ and $\mathrm{P}(Y = 1|X = 0)$.

3 Asymptotic properties of sample quantiles and Order statistics

Let F be a distribution function. For $0 < p < 1$, the p-th quantile is defined as

$$\xi_p = \inf\{x : F(x) \geq p\} \tag{41}$$

and is alternately denoted by $F^{-1}(p)$. Note that ξ_p satistfies

$$F(\xi_p -) \leq p \leq F(\xi_p) \tag{42}$$

Corresponding to a sample $\{X_1, \cdots, X_n\}$ of observations from F, the sample p-th quantile $\hat{\xi}_p$ is defined as the p-th quantile of the sample distribution function F_n, that is

$$\hat{\xi}_p = F_n^{-1}(p). \tag{43}$$

Note that, $F_n(x) := \sharp X_i \leq x/n$.

Example 1 Suppose $X_1, \cdots, X_{n+1}$ are iid standard exponential random variables. For $j = 1, \cdots, n$ define

$$Y_j = \frac{\sum_{i=1}^{j} X_i}{\sum_{i=1}^{n+1} X_i}$$

We derive the joint density of $(Y_1, \cdots, Y_n)$ as follows. First, we observe that the joint density of $(X_1, \cdots, X_{n+1})$ is

$$f_{\underline{X}}(\underline{x}) = \exp\left\{-\sum_{i=1}^{n+1} x_i\right\} I\{\underline{x}_1 > 0, \cdots, \underline{x}_{n+1} > 0\}$$

As an intermediate step, define $Z_j = \sum_{i=1}^{j} X_i$ for $j = 1, \cdots, n+1$. Then the X_i may be expressed in terms of the Z_i as

$$X_i = \begin{cases} Z_i & \text{if} \quad i = 1 \\ Z_i - Z_i - 1 & \text{if} \quad i > 1 \end{cases} \tag{44}$$

The Jacobian is clearly 1 for this transformation, since the Jacobian matrix is lower triangular with ones on the diagonal. Therefore, we obtain $F_{\underline{Z}}(\underline{z}) = \exp\{-z_{n+1}\} I\{0 < z_1 < z_2 < \cdots < z_{n+1}\}$ as the density of Z. If we define $Y_{n+1} = Z_{n+1}$, then we may express the Z_i in terms of the Y_i as

$$Z_i = \begin{cases} Y_{n+1} Y_i & \text{if} \quad i < n+1 \\ Y_{n+1} & \text{if} \quad i = n+1 \end{cases} \tag{45}$$

The Jacobian matrix of the transformation in equation (45) is upper triangular, with y_{n+1} along the diagonal except for a 1 in the lower right corner. Thus, the Jacobian equals y_{n+1}^n, so the density of $\underline{Y}$ is

$$f_{\underline{Y}}(\underline{y}) = y_{n+1}^n \exp\{-y_{n+1}\} I\{y_{n+1} > 0\} I\{0 < y_1 < \cdots < y_n < 1\}. \tag{46}$$

Thus, $(Y_1, \cdots, Y_n)$ is independent of Y_{n+1} and the density of $(Y_1, \cdots, Y_n)$ is proportional to $I\{0 < y_1 < \cdots < y_n < 1\}$. Note that Y_{n+1} may be seen to have a Gamma$(n+1, 1)$ density, which is also evident when we consider that Y_{n+1} is the sum of $n+1$ iid standard exponential random variables.

Since the joint density of the order statistics from a sample of size n from Uniform(0,1) is $n!I\{0 < u_1 < \cdots < u_n < 1\}$, Example 1 proves the following lemma.

Lemma 1. *The joint distribution of the order statistics from a sample of size n from Uniform(0,1) is the same as the joint distribution of*

$$\frac{X_1}{\sum_{i=1}^{n+1} X_i}, \frac{X_1 + X_2}{\sum_{i=1}^{n+1} X_i}, \cdots, \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n+1} X_i} \tag{47}$$

where $X_1, \cdots, X_{n+1}$ are iid standard exponential random variables. Furthermore, the joint distribution of expression is independent of $\sum_{i=1}^{n+1} X_i$.

We now use Lemma 1 to achieve a powerful result, namely deriving the joint asymptotic distribution of a set of sample quantiles. Take $0 < p_1 < p_2 < 1$ and define $a_n = \lceil np_1 \rceil$ and $b_n = \lceil np_2 \rceil$, where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. Then for an iid sample $U_1, \cdots, U_n$, the a_nth and b_nth order statistics $U_{(a_n)}$ and $U_{(b_n)}$ are the p_1 and p_2 sample quantiles, respectively, as per definition (43).

Suppose that the U_i are Uniform(0,1). Then by Lemma 1, $(U_{(a_n)}, U_{(b_n)})$ has the same distribution as

$$\left(\frac{\sum_{i=1}^{a_n} X_i}{\sum_{i=1}^{n+1} X_i}, \frac{\sum_{i=1}^{b_n} X_i}{\sum_{i=1}^{n+1} X_i} \right) \tag{48}$$

where $X_1, \cdots, X_{n+1}$ are iid standard exponential random variables. Let $A = \sum_{i=1}^{a_n} X_i, B = \sum_{i=a_n+1}^{b_n} X_i$, and $C = \sum_{i=b_n+1}^{n+1} X_i$. Then the joint asymptotic distribution of $(U_{(a_n)}, U_{(b_n)})$ is the same as that of

$$g(A, B, C) \stackrel{\text{def}}{=} \left(\frac{A}{A+B+C}, \frac{A+B}{A+B+C} \right) \tag{49}$$

This joint asymptotic distribution may be easily determined using the delta method if we can determine the joint asymptotic distribution of (A, B, C). But this is easy, since A, B, and C are by definition sums of iid random variables and they are independent of one another. Consider, for example, the fact that a bit of algebra gives

$$\sqrt{n}\left(\frac{A}{n} - p_1\right) = \sqrt{\frac{a_n}{n}}\sqrt{a_n}\left(\frac{A}{a_n} - \frac{np_1}{a_n}\right) = \sqrt{\frac{a_n}{n}}\left\{\sqrt{a_n}\left(\frac{A}{a_n} - 1\right) + \sqrt{n}\left(1 - \frac{np_1}{a_n}\right)\right\} \tag{50}$$

By the central limit theorem, $\sqrt{a_n}(A/a_n - 1) \Rightarrow \mathcal{N}(0,1)$ because a standard exponential variable has mean 1 and variance 1. Furthermore, a_n was defined so that

$$\sqrt{n}\left(1 - \frac{np_1}{a_n}\right) \to 0,$$

which of course also implies that $a_n/n \to p_1$. Therefore, Slutsky's theorem gives

$$\sqrt{n}\left(\frac{A}{n} - p_1\right) \Rightarrow \mathcal{N}(0, p_1)$$

Similar arguments applied to B and to C, along with the fact that A, B, and C are independent, gives

$$\sqrt{n}\left\{\begin{pmatrix} A/n \\ B/n \\ C/n \end{pmatrix} - \begin{pmatrix} p_1 \\ p_2 - p_1 \\ 1 - p_2 \end{pmatrix}\right\} \Rightarrow \mathcal{N}_3\left\{\underline{0}, \begin{pmatrix} p_1 & 0 & 0 \\ 0 & p_2 - p_1 & 0 \\ 0 & 0 & 1 - p_2 \end{pmatrix}\right\}$$

by Theorem 15.3. Now recall the definition of $g(A,B,C)$ in equation (49). For this $g : \mathbf{R}^3 \to \mathbf{R}^2$, we obtain $\dot{g}(p_1, p_2 - p_1, 1 - p_2)$ so the delta method gives

$$\sqrt{n}\left\{\begin{pmatrix} U_{(a_n)} \\ U_{(b_n)} \end{pmatrix} - \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}\right\} \Rightarrow \mathcal{N}_2\left\{\underline{0}, \begin{pmatrix} p_1(1-p_1) & p_1(1-p_2) \\ p_1(1-p_2) & p_2(1-p_2) \end{pmatrix}\right\} \tag{51}$$

The method used above to derive the joint distribution (51) of two sample quantiles may be easily extended to any number of quantiles. We may thus state the following theorem that relies on a generalization of the above argument:

Theorem 4. *Suppose that for given constants* $p_1, \cdots, p_k$ *with* $0 < p_1 < \cdots < p_k < 1$, *there exist sequences* $\{a_{1n}\}, \cdots, \{a_{kn}\}$ *such that for all* $1 \le i \le k$,

$$\sqrt{n}\left(1 - \frac{np_i}{a_{in}}\right) \to 0. \tag{52}$$

Then if $U_1, \cdots, U_n$ *is a sample from Uniform(0,1),*

$$\sqrt{n}\left\{\begin{pmatrix} U_{(a_{1n})} \\ \vdots \\ U_{(a_{kn})} \end{pmatrix} - \begin{pmatrix} p_1 \\ \vdots \\ p_k \end{pmatrix}\right\} \Rightarrow \mathcal{N}_k\left\{\underline{0}, \begin{pmatrix} p_1(1-p_1) & \cdots & p_1(1-p_k) \\ \vdots & \ddots & \vdots \\ p_1(1-p_k) & \cdots & p_k(1-p_k) \end{pmatrix}\right\} \tag{53}$$

Note that in the covariance matrix in the above theorem, the (i,j) entry is either $p_i(1-p_j)$ or $p_j(1-p_i)$, depending on whether $i \le j$ or $j \le i$. As a corollary, we may restate Theorem 5.4.5 on page 314 of Lehmann:

Corollary 3. *Suppose that there exists a cdf F and points $\xi_1 < \cdots < \xi_k$ such that $F'(\xi_i)$ exists and is positive for all i. Let $p_i = F(\xi_i)$ for $1 \le i \le k$. If equation (51) is satisfied for sequences $\{a_{1n}\}, \cdots, \{a_{kn}\}$ and $X_1, \cdots, X_n$ is an iid sample from F, then*

$$\sqrt{n}\left\{\begin{pmatrix} X_{(a_{1n})} \\ \vdots \\ X_{(a_{kn})} \end{pmatrix} - \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_k \end{pmatrix}\right\} \Rightarrow \mathcal{N}_k\left\{\underline{0}, \begin{pmatrix} \frac{p_1(1-p_1)}{F'(\xi_1)^2} & \cdots & \frac{p_1(1-p_k)}{F'(\xi_1)F'(\xi_k)} \\ \vdots & \ddots & \vdots \\ \frac{p_1(1-p_k)}{F'(\xi_1)F'(\xi_k)} & \cdots & \frac{p_k(1-p_k)}{F'(\xi_k)^2} \end{pmatrix}\right\} \tag{54}$$

The corollary is proved by a simple application of the delta method, since the hypotheses imply that $F(t)$ is continuous and strictly increasing in a neighborhood of each ξ_i. Thus, $F^{-1}(u)$ is well-defined in a neighborhood of each p_i. Defining $h(u_1, \cdots, u_k) = (F^{-1}(u_1), \cdots, F^{-1}(u_k))$, the matrix $\dot{h}(u_1, \cdots, u_k)$ is a diagonal matrix with i-th element

$$\frac{\partial F^{-1}(ui)}{\partial u_i} = \frac{1}{F' \circ F^{-1}(u_i)}$$

- Read Lehmann Thm 5.4.5, Example 5.4.2 and references therein.
- Read Serfling Sections 2.3.4, 2.3.5, 2.3.6, 2.5.2, 2.5.3
- The last 2 problems of this section are on Asymptotic relative efficiency of estimators covered previously. Refer to Serfling 1.15 or Chapter 6 of Lehmann and Casella.
- Problems 1,2,3,5,6 of last chapter and 1,2,4,5,6 of this chapter are to be presented in class roll-number-wise on Monday Sept 1st.
- The other problems need to be submitted by each student separately by Wednesday Sept 3rd.

Exercises

1. Give a detailed proof of 51 and 54

2. Suppose $X_1, \cdots, X_n$ is an iid sample with $P(X_i \leq x) = F(x - \theta)$, where $F(x)$ is symmetric about zero. We wish to estimate θ by $(Q_p + Q_{1-p})/2$, where Q_p and Q_{1-p} are the p and $1-p$ sample quantiles, respectively. Find the smallest possible asymptotic variance for the estimator and the p for which it is achieved for each of the following forms of $F(x)$:

   (a) Standard Cauchy

   (b) Standard normal

   (c) Standard double exponential

   Hint: If you cannot solve a problem analytically, try attacking it numerically. Part (c) is a bit of a trick question.

3. When we use a boxplot to assess the symmetry of a distribution, one of the main things we do is visually compare the lengths of $Q_3 - Q_2$ and $Q_2 - Q_1$, where Q_i denotes the i-th sample quartile.

   (a) If we have a random sample of size n from $\mathcal{N}(0,1)$, find the asymptotic distribution of $(Q_3^{(n)} - Q_2^{(n)}) - (Q_2^{(n)} - Q_1^{(n)})$.

   (b) Repeat part (a) if the sample comes from a standard logistic distribution.

   (c) Using 1000 simulations from each distribution, use graphs to assess the accuracy of each of the asymptotic approximations above for $n = 5$ and $n = 13$. (For a sample of size $4n + 1$, define Q_i to be the $in + 1$ order statistic.) For each value of n and each distribution, plot the empirical distribution function against the theoretical limiting cdf.

   Hint: In parts (a) and (b), use Theorem 5.4.5 on p. 314 of Lehmann. If you are using R for part (c), the function cdf.compare is a very useful function for comparing an empirical cdf with another cdf. For example, to compare the empirical cdf of a vector x with a standard normal distribution function, type cdf.compare(x,dist=”normal”)

4. Let $X_1, \cdots, X_n$ be a random sample from Uniform$(0, 2\theta)$. Find the asymptotic distributions of the median, the midquartile range, and $\frac{2}{3}Q_3$. (The midquartile range is the mean of the 1st and 3rd quartiles.) Compare these three estimates of θ.

5. Problem 6.6.14 of Lehmann and Casella, pg 510.

6. Problem 6.6.15 of Lehmann and Casella, pg 510.

4 Non-parametrics

4.1 Introduction to Nonparametrics

Any statistical inference problem has the following basic structure. We have some random data having joint distribution F which is not entirely known. We want to make inference about the unknown aspects of F based on observed data. The inference is typically either an estimation problem or a testing problem. The difference between nonparametric and parametric problems has to do with how much we already assume known about F.

Example 1: (Parametric problem) Suppose $X_1, \cdots, X_n$ are iid $\mathrm{N}(\mu, \sigma^2)$. We want to estimate μ and σ^2. This is a typical problem from parametric statistics. Here we assume that the distribution of the X's is completely known except for only two unknown numbers μ and σ^2.

Definition 1. *If the distribution of the data is completely known except for finitely many unknown numbers, then the problem is called a* **parametric problem***. Otherwise, we have a* **nonparametric problem***. In a parametric situation each of the finitely many unknown numbers is called a* **parameter***.*

Example 2: (Nonparametric density estimation) Suppose $X_1, \cdots, X_n$ iid with continuous density f, which is unknown. We want to estimate f.

This is a non-parametric problem because the number of parameters, ie, values of $f(x)$ for all values of x on the real line, is infinite.

Example 3: (Nonparametric regression) Think of a nonparametric inference situation in regression. In the model $Y = \alpha + \beta X + \epsilon$, we may have ϵ's iid with some unknown distribution F. Or, we may have the model $Y = f(X) + \epsilon$, where f itself is some unknown continuous function.

Example 4: (Semi-parametric problem) Suppose that we are testing the efficacy of a sleeping pill. Let $X_1, \cdots, X_n$ be the amount of sleep of n patients before taking the pill, and let $Y_1, \cdots, Y_n$ be the corresponding amounts after taking it. We want to test if the pill really increases one's amount of sleep. Assuming that the patients behave independently we may reasonably assume that $(X_1, Y_1), \cdots, (X_n, Y_n)$ are independent, but not necessarily identically distributed. We model the effect of the drug as follows. There is an unknown number θ denoting the median increase of sleep, that is, $Z_i = Y_i - X_i$ have θ as its median. Note that we are not assuming that Z's all have the same distribution. We are merely assuming that they have a common median. We want to test $H_0 : \theta = 0$ Vs $H_1 : \theta > 0$. In this example we have not assumed any knowledge about the underlying distribution except for the existence of a common median θ for the Z's. Thus our ignorance cannot be summed up as finitely many unknown numbers. Hence this is a nonparametric statistical inference problem.

This is called a **semiparametric problem**, because here we are interested in only one unknown number, θ, though θ is not the only unknown quantity.

Why Nonparametric?

While in many situations parametric assumptions are reasonable (e.g. assumption of Normal distribution for the background noise, Poisson distribution for a photon counting signal of a nonvariable source), we often have no prior knowledge of the underlying distributions. In such situations, the use of parametric statistics can give misleading or even wrong results. We need statistical procedures which are insensitive to the model assumptions in the sense that the procedures retain their properties in the neighborhood of the model assumptions.

Insensitivity to model assumptions is called **Robustness**. Apart from this, we also need procedures which are robust against the presence of outliers in the data. Some common parametric procedures that are not robust include

- The sample mean is not robust against the presence of even one outlier in the data and is not variance robust as well. The sample median is robust against outliers and is variance robust.
- The t-test does not have t-distribution if the underlined distribution is not normal and the sample size is small. For large sample size, it is asymptotically level robust but is not power robust. Also, it is not robust against the presence of outliers.

The nonparametric tests described here are often called distribution free procedures because their significance levels do not depend on the underlying model assumption i.e., they are level robust. They are also power robust and robust against outliers.

4.2 Large sample behavior of Kolmogorov-Smirnov statistic

Sec 2.1 of Serfling

Exercises

1. Problem 2.3

2. Problem 2.4

3. Problem 2.5

4. Show that, for the class of continuous F's, the exact distribution of D_n does not depend on F.

5. Show that $D_n \geq 1/(2n)$.

6. (a) Prove/disprove/suitably modify : $F \succeq G$ iff $F(x) \geq G(x)$ for all x.
   (b) Show that $D+, D- \geq 0$.
   (c) What is the relation between $D, D+$ and $D-$?

7. (everybody needs to submit) The file data.txt contains iid data from some unknown continuous distribution, F. We want to test $H_0 : F = \mathcal{N}(0,1)$ Vs. $H_1 : F \neq \mathcal{N}(0,1)$. Perform KS test using the asymptotic distribution. Report the P-value.

8. Show that nD_+^2 is asymptotically distributed as Exponential(2).

5 Maximum Likelihood Estimator

Reading

- Lehman 7.1, 7.2
- Serfling 4.2
- Rao 5g pp366

If X is a random variable (or vector) with density or mass function $f_\theta(x)$ that depends on a parameter θ, then the function $f_\theta(x)$ viewed as a function of θ is called the likelihood function of θ. We often denote this function by $L(\theta)$. Note that $L(\theta) = f_\theta(x)$ is implicitly a function of x, but we suppress this fact in the notation. Let the set of possible values of θ be the set Ω. If $L(\theta)$ has a maximizer in Ω, say $\hat{\theta}$, then $\hat{\theta}$ is called the maximum likelihood estimator or MLE of θ. Since the logarithm function is a strictly increasing function, any maximizer of $L(\theta)$ also maximizes $l(\theta) \overset{\text{def}}{=} \log L(\theta)$. It is often much easier to maximize $l(\theta)$, called the loglikelihood function, than $L(\theta)$.

Example 1 Suppose $\Omega = (0, \infty)$ and $X \sim \text{Binomial}(n, e^{-\theta})$. Then

$$l(\theta) = \log \left(\begin{array}{c} n \\ x \end{array} \right) - x\theta + (n - x)\log(1 - e^{-\theta})$$

$$\text{so} \quad l'(\theta) = -x + \frac{n - x}{1 - e^{-\theta}}$$

Thus, setting $l'(\theta) = 0$ yields $\hat{\theta} = -\log(x/n)$. It isn't hard to verify that $-\log(x/n)$ is in fact the maximizer of $l(\theta)$.

This can also be derived using the following theorem:

Theorem 5 ((Invariance Property of MLEs)). *If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.*

This is Thm 7.2.10 of Casella and Berger. See proof there.

As the preceding example demonstrates, it is not always the case that an MLE exists, for if $X = 0$ or $X = n$, then $-\log(X/n)$ is not contained in Ω.

We will show that the maximum likelihood estimator is, in many cases, asymptotically normal. However, this is not always the case; in fact, it is not even necessarily true that the MLE is consistent, as shown in Exercise 1.

5.1 Consistency of MLE

We begin the discussion of the consistency of the MLE by defining the so-called Kullback-Liebler information.

Definition 2. *If $f_{\theta_0}(x)$ and $f_{\theta_1}(x)$ are two densities, the Kullback-Leibler information number equals $K(f_{\theta_0}, f_{\theta_1}) = E_{\theta_0} \log \frac{f_{\theta_0}(X)}{f_{\theta_1}(X)}$. If $P_{\theta_0}\left(f_{\theta_0}(X) > 0 \quad \text{and} \quad f_{\theta_1}(X) = 0\right) > 0$, then $K(f_{\theta_0}, f_{\theta_1})$ is defined to be 1.*

We may show that the Kullback-Leibler information must be nonnegative using Jensen's inequality.

Theorem (Jensen's inequality)**.** *If $g(t)$ is a convex function, then for any random variable $X, g(EX) \leq Eg(X)$. Furthermore, if $g(t)$ is strictly convex, then $Eg(X) = g(EX)$ only if $P(X = c) = 1$ for some constant c.*

Considering the Kullback-Leibler information once again, we first note that

$$E_{\theta_0} \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} = E_{\theta_1}\left(I_{f_{\theta_0}(X)>0}\right) \leq 1.$$

Therefore, by the strict convexity of the function $-\log x$,

$$K(f_{\theta_0}, f_{\theta_1}) = E_{\theta_0} - \log \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} \geq -\log E_{\theta_0} \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} \geq 0, \tag{55}$$

with equality if and only if $P_{\theta_0} f_{\theta_0}(X) = f_{\theta_1}(X) = 1$. Inequality (55) is sometimes called the Shannon- Kolmogorov information inequality.

If $X_1, \cdots, X_n$ are iid with density $f_{\theta_0}(x)$, then $l(\theta) = \sum_{i=1}^n \log f_{\theta_0}(x_i)$. Thus, the Shannon-Kolmogorov information inequality may be used to prove the consistency of the maximum likelihood estimator in the case of a finite parameter space.

Theorem 6 (Consistency of MLE)**.** *Suppose Ω is finite and that $X_1, \cdots, X_n$ are iid with density $f_{\theta_0}(x)$. Furthermore, suppose that the model is identifiable, which is to say that different values of θ lead to different distributions. Then if $\hat{\theta}_n$ denotes the maximum likelihood estimator, $\hat{\theta}_n \overset{P}{\to} \theta_0$.*

Proof: Notice that

$$\frac{1}{n}\sum_{i=1}^n \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} \overset{P}{\to} E_{\theta_0} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} = -K(f_{\theta_0}, f_\theta) \tag{56}$$

The value of $-K(f_{\theta_0}, f_\theta)$ is strictly negative for $\theta \neq \theta_0$ by the identifiability of the model. Therefore, since $\hat{\theta}_n$ is the maximizer of the left hand side of Equation (56),

$$P(\hat{\theta}_n \neq \theta_0) = P\left(\max_{\theta\neq\theta_0}\left(\frac{1}{n}\sum_{i=1}^n \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right) > 0\right) \leq \sum_{\theta\neq\theta_0} P\left(\frac{1}{n}\sum_{i=1}^n \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} > 0\right) \to 0. \tag{57}$$

This implies that $\hat{\theta}_n \xrightarrow{P} \theta_0$. The result of Theorem 6 may be extended in several ways; however, it is unfortunately not true in general that a maximum likelihood estimator is consistent, as seen in Exercise 1. We will present the extension given in Lehmann, but we do so without proof.

If we return to the simple Example 1, we found that the MLE was found by solving the equation

$$l'(\theta) = 0 \tag{58}$$

Equation (58) is called the likelihood equation, and naturally a root of the likelihood equation is a good candidate for a maximum likelihood estimator. However, there may be no root and there may be more than one. It turns out the probability that at least one root exists goes to 1 as $n \to \infty$. Consider Example 1, in which no MLE exists whenever $X = 0$ or $X = n$. In that case, both $P(X = 0) = (1 - e^{-\theta})^n$ and $P(X = n) = e^{-n\theta}$ go to zero as $n \to \infty$. In the case of multiple roots, one of these roots is typically consistent for θ_0, as stated in the following theorem.

Theorem 7. *Suppose that $X_1, \cdots, X_n$ are iid with density $f_{\theta_0}(x)$ for θ_0 in an open interval $\Omega \subset R$, where the model is identifiable (i.e., different values of $\theta \in \Omega$ give different distributions). Furthermore, suppose that the loglikelihood function $l(\theta)$ is differentiable and that the support $\{x : f_\theta(x) > 0\}$ does not depend on θ. Then with probability approaching 1 as $n \to \infty$, there exists $\hat{\theta}_n = \hat{\theta}_n(X_1, \cdots, X_n)$ such that $l'(\hat{\theta}_n) = 0$ and $\hat{\theta}_n \xrightarrow{P} \theta_0$.*

Stated succinctly, Theorem 7 says that under certain regularity conditions, there is a consistent root of the likelihood equation. It is important to note that there is no guarantee that this consistent root is the MLE. However, if the likelihood equation only has a single root, we can be more precise:

Corollary 4. *Under the conditions of Theorem 7, if for every n there is a unique root of the likelihood equation, and this root is a local maximum, then this root is the MLE and the MLE is consistent.*

Proof. The only thing that needs to be proved is the assertion that the unique root is the MLE. Denote the unique root by $\hat{\theta}_n$ and suppose there is some other point θ such that $l(\theta) \geq l(\hat{\theta}_n)$. Then there must be a local minimum between $\hat{\theta}_n$ and θ, which contradicts the assertion that $\hat{\theta}_n$ is the unique root of the likelihood equation. □

5.2 Asymptotic Normality of MLE

As seen in the preceding topic, the MLE is not necessarily even consistent, so the title of this topic is slightly misleading. However, "Asymptotic normality of the consistent root of the likelihood equation" is a bit too long!

It will be necessary to review a few facts regarding Fisher information before we proceed. For a density (or mass) function $f_\theta(x)$, we define the Fisher information function to be

$$I(\theta) = E_\theta \left\{ \frac{d}{d\theta} \log f_\theta(X) \right\}^2 \tag{59}$$

If $\eta = g(\theta)$ for some invertible and differentiable function $g(\cdot)$, then since

$$\frac{d}{d\eta} = \frac{d\theta}{d\eta} \frac{d}{d\theta} = \frac{1}{g'(\theta)} \frac{d}{d\theta} \tag{60}$$

by the chain rule, we conclude that

$$I(\eta) = \frac{I(\theta)}{\{g'(\theta)\}^2} \tag{61}$$

Loosely speaking, $I(\theta)$ is the amount of information about θ contained in a single observation from the density $f_\theta(x)$. However, this interpretation doesn't always make sense. For example, it is possible to have $I(\theta) = 0$ for a very informative observation (see Example 7.2.1 on page 462 of Lehmann). Although we do not dwell on this fact in this course, expectation may be viewed as integration. Suppose that $f_\theta(x)$ is twice differentiable with respect to θ and that the operations of differentiation and integration may be interchanged in the following sense:

$$E_\theta \left\{ \frac{d}{d\theta} \log f_\theta(X) \right\} = E_\theta \left\{ \frac{\frac{d}{d\theta} f_\theta(X)}{f_\theta(X)} \right\} = \int \frac{d}{d\theta} f_\theta(X) dx = \frac{d}{d\theta} \int f_\theta(X) dx = \frac{d}{d\theta} 1 = 0 \tag{62}$$

$$E_\theta \left\{ \frac{d}{d\theta} \frac{\frac{d}{d\theta} f_\theta(X)}{f_\theta(X)} \right\} = E_\theta \left\{ \frac{\frac{d^2}{d\theta^2} f_\theta(X)}{f_\theta(X)} \right\} - I(\theta) = \frac{d^2}{d\theta^2} \int f_\theta(X) dx - I(\theta) = -I(\theta) \tag{63}$$

Equations (62) and (63) give two additional expressions for $I(\theta)$. From Equation (62) follows

$$I(\theta) = \mathrm{Var}_\theta \left\{ \frac{d}{d\theta} \log f_\theta(X) \right\} \tag{64}$$

and Equation (63) implies

$$I(\theta) = -E_\theta \left\{ \frac{d^2}{d\theta^2} \log f_\theta(X) \right\}. \tag{65}$$

In many cases, Equation (65) is the easiest form of the information to work with. Equations (64) and (65) make clear a helpful property of the information, namely that for independent random variables, the information about θ contained in the joint sample is simply the sum of the individual information components. In particular, if we have an iid sample from $f_\theta(x)$, then the information about θ equals $nI(\theta)$. The reason that we need the Fisher information is that we will show that under certain regularity conditions,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}\left\{0, \frac{1}{I(\theta_0)}\right\}, \tag{66}$$

where $\hat{\theta}_n$ is the consistent root of the likelihood equation.

Example 1 (Poisson case) Suppose that $X_1, \cdots, X_n$ are iid Poisson(θ_0) random variables. Then the likelihood equation has a unique root, namely $\hat{\theta}_n = \bar{X}_n$, and we know that by the central limit theorem $\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}(0, \theta_0)$. However, the Fisher information for a single observation in this case is

$$-E_\theta\left\{\frac{d^2}{d\theta^2}\log f_\theta(X)\right\} = E_\theta \frac{X}{\theta^2} = \frac{1}{\theta} \tag{67}$$

Thus, in this example, equation (66) holds.

Rather than stating all of the regularity conditions necessary to prove Equation (64), we work backwards, figuring out the conditions as we go through the proof. The first step is to expand $\ell'(\hat{\theta}_n)$ in a power series around θ_0:

$$\ell'(\hat{\theta}_n) = \ell'(\theta_0) + (\hat{\theta}_n - \theta_0)\ell''(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2\ell'''(\theta_n^*) \tag{68}$$

for some θ_n^* between $\hat{\theta}_n$ and θ_0. Clearly, the validity of Equation (68) hinges on the existence of a continuous third derivative of $l(\theta)$. Rewriting equation (68) gives

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{\sqrt{n}\{\ell'(\hat{\theta}_n) - \ell'(\theta_0)\}}{\ell''(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)\ell'''(\theta_n^*)} = \frac{\frac{1}{\sqrt{n}}\{\ell'(\theta_0) - \ell'(\hat{\theta}_n)\}}{-\frac{1}{n}\ell''(\theta_0) - \frac{1}{2n}(\hat{\theta}_n - \theta_0)l'''(\theta_n^*)} \tag{69}$$

Let's consider the pieces of Equation (69) individually. If the conditions of Theorem 7 of last section are met, then $\ell'(\hat{\theta}_n) \xrightarrow{P} 0$. If Equation (62) holds and $I(\theta_0) < \infty$, then

$$\frac{1}{\sqrt{n}}\ell'(\theta_0) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\theta}\log f_{\theta_0}(X_i)\right) \Rightarrow \mathcal{N}(0, I(\theta_0)) \tag{70}$$

by the central limit theorem and Equation (64). If Equation (63) holds, then

$$\frac{1}{n}\ell''(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\theta^2}\log f_{\theta_0}(X_i) \xrightarrow{P} -I(\theta_0) \tag{71}$$

by the weak law of large numbers and Equation (65). Finally, we would like to have the term involving $\ell'''(\theta_n^*)$ disappear, so clearly it is enough to show that $\frac{1}{n}\ell'''(\theta)$ is bounded in probability in a neighborhood of θ_0. Putting all of these facts together gives a theorem.

Theorem 8. *Suppose that the conditions of Theorem 7 (of last section) are satisfied, and let $\hat{\theta}_n$ denote a consistent root of the likelihood equation. Assume also that* $\mathrm{l}'''(\theta)$ *exists and is continuous, that equations (62) and (63) hold, and that* $\frac{1}{n}\mathrm{l}'''(\theta)$ *is bounded in probability in a neighborhood of* θ_0*. Then if* $0 < I(\theta_0) < \infty$*, (66) holds.*

The theorem is proved by noting that under the stated regularity conditions, $l'(\hat{\theta}_n) \overset{P}{\to} 0$ so that the numerator in (69) converges in distribution to $\mathcal{N}\{0, I(\theta_0)\}$ by Slutsky's theorem. Furthermore, the denominator in (69) converges to $I(\theta_0)$, so another application of Slutsky's theorem gives the desired result.

Sometimes, it is not possible to find an exact zero of $l'(\theta)$. One way to get a numerical approximation to a zero of $l'(\theta)$ is to use Newton's method, in which we start at a point θ_0 and then set

$$\theta_1 = \theta_0 - \frac{l'(\theta_0)}{l''(\theta_0)}. \tag{72}$$

Ordinarily, after finding θ_1 we would set θ_0 equal to θ_1 and apply Equation (72) iteratively. However, we may show that by using a single step of Newton's method, starting from a $\sqrt{n}$-consistent estimator of θ_0, we may obtain an estimator with the same asymptotic distribution as $\hat{\theta}_n$. The proof of the following theorem is left as an exercise:

Theorem 9. *Suppose that $\tilde{\theta}_n$ is any $\sqrt{n}$-consistent estimator of θ_0 (i.e., $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ is bounded in probability). Then under the conditions of Theorem 7, if we set*

$$\delta_n = \tilde{\theta}_n - \frac{\mathrm{l}'(\tilde{\theta}_n)}{\mathrm{l}''(\tilde{\theta}_n)} \tag{73}$$

then

$$\sqrt{n}(\delta_n - \theta_0) \Rightarrow \mathcal{N}(0, \frac{1}{I(\theta_0)}) \tag{74}$$

5.3 Method of scoring

Recall that, under suitable regularity conditions, the maximum likelihood estimate is the solution to the score equation

$$S(\theta) = s(x; \theta) = \frac{\partial}{\partial \theta} l(\theta) = \frac{\partial}{\partial \theta} \log L(\theta; x) = 0, \tag{75}$$

where $S(\theta) = s(X; \theta)$ is the score statistic. Generally the solution to this equation must be calculated by iterative methods. One of the most common methods is the Newton-Raphson method and is based on successive approximations to the solution, using Taylor's theorem to approximate the equation. Thus, we take an initial value θ_0 and write

$$0 = S(\theta_0) - J(\theta_0)(\theta - \theta_0), \tag{76}$$

ignoring the remainder term. Here

$$J(\theta) = J(\theta; X) = -\frac{\partial}{\partial\theta} S(\theta) = -\frac{\partial^2}{\partial\theta^2} l(\theta). \tag{77}$$

Solving this equation for θ then yields a new value θ_1

$$\theta_1 = \theta_0 + J(\theta_0)^{-1} S(\theta_0) \tag{78}$$

and we keep repeating this procedure as long as $\mid S(\theta_j) \mid > \epsilon$, i.e. $\theta_{k+1} = \theta_k + J(\theta_k)^{-1} S(\theta_0)$. Clearly, $\hat{\theta}$ is a fixed point of this iteration as $S(\hat{\theta}) = 0$ and, conversely, any fixed point is a solution to the likelihood equation.

Formally the iteration becomes

- Choose an initial value θ and calculate $S(\theta)$ and $J(\theta)$;
- While $\mid S(\theta) \mid > \epsilon$ Repeat
  1. $\theta \leftarrow \theta + J(\theta)^{-1} S(\theta)$
  2. Calculate $S(\theta)$ and $J(\theta)$ go to 1
- Return θ

Other criteria for terminating the iteration can be used. To get a criterion which is insensitive to scaling of the variables, one can instead use the criterion $\mid J(\theta)^{-1} S(\theta) \mid > \epsilon$.

If $\hat{\theta}$ is a local maximum for the likelihood function, we must have $J(\hat{\theta}) = -\frac{\partial^2}{\partial\theta^2} l(\hat{\theta}) > 0$. The quantity $J(\hat{\theta})$ determines the sharpness of the peak in the likelihood function around its maximum. It is also known as the observed information. Occasionally we also use this term for $J(\theta)$ where θ is arbitrary but strictly speaking this can be quite inadequate as $J(\theta)$ may well be negative (although positive in expectation).

Recall that the (expected) Fisher information is $I(\theta) = \mathrm{E}\{J(\theta)\}$ and that for large i.i.d. samples it holds approximately that $\hat{\theta} \sim \mathcal{N}(\theta, I(\theta)^{-1})$. But it is also approximately true, under the same assumptions that $\sqrt{J(\hat{\theta})}(\hat{\theta} - \theta) \sim \mathcal{N}(0, 1)$, so we could write $\hat{\theta} \sim \mathcal{N}(\theta, J(\hat{\theta})^{-1})$. Indeed, as $\hat{\theta}$ is approximately sufficient, $J(\hat{\theta})$ is approximately ancillary.

Note that, as a by-product of this algorithm, the final value of $J(\theta)$ is the observed information which can be used to assess the uncertainty of $\hat{\theta}$.

If θ_0 is chosen sufficiently near $\hat{\theta}$ convergence is very fast. It can be computationally expensive to evaluate $J(\theta)$ a large number of times. This is sometimes remedied by only changing J every 10 iterations or similar. Another problem with the Newton-Raphson method is its lack of stability. When the initial value θ_0 is far from θ it might wildly oscillate and not converge at all. This is sometimes remedied by making smaller steps as

$$\theta \leftarrow \theta + \gamma J(\theta)^{-1} S(\theta) \tag{79}$$

where $0 < \gamma < 1$ is a constant.

The iteration has a tendency to be unstable for many reasons, one of them being that $J(\theta)$ may be negative unless θ already is very close to the MLE $\hat{\theta}$. In addition, $J(\theta)$ might sometimes be hard to calculate. R. A. Fisher introduced the method of scoring which simply replaces the observed second derivative with its expectation to yield the iteration

$$\theta \leftarrow \theta + I(\theta)^{-1} S(\theta). \tag{80}$$

In many cases, $I(\theta)$ is easier to calculate and $I(\theta)$ is always positive.

In the case of n independent and identically distributed observations we have $I(\theta) = nI_1(\theta)$ so

$$\theta \leftarrow \theta + I_1(\theta)^{-1} S(\theta)/n \tag{81}$$

where $I_1(\theta)$ is the Fisher information in a single observation.

Example 1 (Exponential Family:) In a linear canonical one-parameter exponential family

$$f(x; \theta) = b(x) \exp\{\theta t(x) - c(\theta)\} \tag{82}$$

we get

$$J(\theta) = \frac{\partial^2}{\partial \theta^2}\{c(\theta) - \theta t(X)\} = c''(\theta) = I(\theta). \tag{83}$$

so for canonical exponential families the method of scoring and the method of Newton-Raphson coincide. If we let $v(\theta) = c''(\theta) = I(\theta) = V(t(X))$ the iteration becomes

$$\theta \leftarrow \theta + v(\theta)^{-1} S(\theta)/n. \tag{84}$$

The identity of Newton-Raphson and the method of scoring only holds for the canonical parameter.

If $\theta = g(\mu)$

$$\begin{aligned} J(\mu) &= \frac{\partial^2}{\partial \mu^2}\{c(g(\mu)) - g(\mu)t(X)\} \\ &= \frac{\partial}{\partial \mu}[g'(\mu)\tau\{g(\mu)\} - g'(\mu)t(X)] \\ &= v\{g(\mu)\}\{g'(\mu)\}^2 + g''(\mu)[\tau\{g(\mu)\} - t(X)] \end{aligned}$$

where we have let $\tau(\theta) = c'(\theta) = \mathrm{E}_\theta\{t(X)\}$ and $v(\theta) = c''(\theta) = V_\theta\{t(X)\}$. The method of scoring is simpler because the last term has expectation equal to 0:

$$I(\mu) = EJ(\mu) = v\{g(\mu)\}\{g'(\mu)\}^2.$$

Example 2 (Nonlinear Regression) Consider the case of nonlinear least squares (least squares is same as maximum likelihood for normally distributed residuals), in which context Fisher scoring has a very long history and is known as the Gauss-Newton algorithm. The objective function is $f(\beta) = \sum_{i=1}^{n}[y_i - \mu(t_i, \beta)]^2$, where the y_i are observations and $\mu(.,.)$ is a general function of covariates t_i and the unknown parameter β. Write $\underline{y}$ for the vector of y_i ,$\underline{\mu}$ for the vector of $\mu(t_i, \beta)$, and $\underline{\dot{\mu}}$ for the derivative vector of $\underline{\mu}$ with respect to β. The Fisher scoring iteration becomes

$$\beta_{k+1} = \beta_k + (\underline{\dot{\mu}}^T \underline{\dot{\mu}})^{-1} \underline{\dot{\mu}}^T (\underline{y} - \underline{\mu}), \tag{85}$$

where all terms on the right-hand size are evaluated at β_k. The updated estimate is obtained by adding to β_k the coefficients from the regression of the residuals $y_i - \mu_i$ on the derivatives $\dot{\mu}_i$. Gauss-Newton therefore solves the nonlinear least squares problem by way of a series of linear regressions.

Exercises

1. In this problem, we explore an example in which the MLE is not consistent. Suppose that for $\theta \in (0, 1), X$ is a continuous random variable with density $f_\theta(x) = \frac{3(1-\theta)}{\delta^3(\theta)}[\delta^2(\theta) - (x - \theta)^2]I\{|x - \theta| \le \delta(\theta)\} + \frac{\theta}{2}I\{|x| \le 1\}$ where $\delta(\theta) > 0$ for all θ.

   (a) Prove that $f_\theta(x)$ is a legitimate density.

   (b) What condition on $\delta(\theta)$ ensures that $\{x : f_\theta(x) > 0\}$ does not depend on θ?

   (c) With $\delta(\theta) = \exp\{-(1 - \theta)/4\}$, let $\theta = 0.125$. Take samples of sizes $n \in \{50, 250, 500\}$ from $f_\theta(x)$. In each case, graph the loglikelihood function and find the MLE. Also, try to identify the consistent root of the likelihood equation in each case. Hints: To generate a sample from $f_\theta(x)$, note that $f_\theta(x)$ is a mixture density, which means you can start by generating a standard uniform random variable. If it's less than θ, generate a uniform variable on $(-1, 1)$. Otherwise, generate a variable with density $3(\delta^2 - x^2)/\delta^3$ on $(-\delta, \delta)$ and then add θ. You should be able to do this by inverting the cdf. Be very careful when graphing the loglikelihood and finding the MLE. In particular, make sure you evaluate the loglikelihood analytically at each of the sample points in $(0, 1)$; if you fail to do this, you'll miss the point of the problem and you'll get the MLE incorrect. This is because the correct loglikelhood graph will have tall, extremely narrow spikes.

   (d) Prove that the MLE is inconsistent in this situation.

2. Suppose that $X_1, \cdots, X_n$ are iid with density $f_\theta(x)$, where $\theta \in (0, \infty)$. For each of the following forms of $f_\theta(x)$, prove that the likelihood equation has a unique solution and that this solution maximizes the likelihood function.

(a) Weibull: For some constant $a > 0$, $f_\theta(x) = a\theta^a x^{a-1} \exp\{-(\theta x)^a\} I\{x > 0\}$

(b) Cauchy: $f_\theta(x) = \frac{\theta}{\pi} \frac{1}{\pi^2 + \theta^2}$

(c) $f_\theta(x) = \frac{3\theta^2\sqrt{3}}{2\pi(x^3+\theta^3)} I\{x > 0\}$

3. Find the MLE and its asymptotic distribution given a random sample of size n from $f_\theta(x) = (1-\theta)\theta^x, x = 0, 1, 2, \cdots, \quad \theta \in (0, 1)$.

4. Problem 2.1 on p. 553

5. Problem 2.12 on p. 555

6. Prove Theorem 9
Hint: Start with $\sqrt{n}(\delta_n - \theta_0) = \sqrt{n}(\delta_n - \tilde{\theta}_n) + \sqrt{n}(\tilde{\theta}_n - \theta_0)$, then expand $\ell'(\tilde{\theta}_n)$ in a Taylor series about θ_0 and rewrite $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ using this expansion.

7. Suppose that the following is a random sample from a logistic density with cdf $F_\theta(x) = (1 + \exp\{\theta - x\})^{-1}$ (I'll cheat and tell you that I used $\theta = 2$.)

1.0944	6.4723	3.1180	3.8318	4.1262
1.2853	1.0439	1.7472	4.9483	1.7001
1.0422	0.1690	3.6111	0.9970	2.9438

(a) Evaluate the unique root of the likelihood equation numerically. Then, taking the sample median as our known $\sqrt{n}$-consistent estimator $\tilde{\theta}_n$ of θ, evaluate the estimator δ_n in equation (73) numerically.

(b) Find the asymptotic distributions of $\sqrt{n}(\tilde{\theta}_n - 2)$ and $\sqrt{n}(\delta_n - 2)$. Then, simulate 200 samples of size $n = 15$ from the logistic distribution with $\theta = 2$. Find the sample variances of the resulting sample medians and δ_n-estimators. How well does the asymptotic theory match reality?

6 Chi-square

6.1 Large sample properties of likelihood ratio test statistic

This is a simplified version of Serfling 4.4. You should know from Serfling Lemma 4.4.2A, Lemma 4.4.4C, Theorem 4.4.3 and Lemma 4.4.4A.

Let $X_1, \cdots, X_n$ be iid with density $f(x, \theta)$. We are interested in testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ where θ is $k \times 1$ using a likelihood ratio test. To carry out the test, we need to determine the appropriate critical value c. Recall that c is determined by the requirement that $P(\mathrm{LR} > c \mid H_0) = \alpha$. In order to determine the critical value, we thus need to determine the distribution of LR when the null hypothesis is true. We now develop a large sample approximation to solve this problem.

Let $\hat{\theta} = \mathrm{argmax}_\theta L(\theta)$ denote the mle, and write the maximized likelihood ratio statistic as

$$\mathrm{LR} = \frac{L(\hat{\theta})}{L(\theta_0)} \tag{86}$$

Define the statistic $\xi_{\mathrm{LR}} = 2\ln(LR) = 2(l(\hat{\theta}) - l(\theta_0))$ where $l(\theta) = \ln L(\theta)$. Since ξ_{LR} is a monotonic transformation of LR, the LR test can be implemented by rejecting the null hypothesis when ξ_{LR} is large.

To find the approximate distribution of ξ_{LR} under the null hypothesis, write

$$l(\theta_0) = l(\hat{\theta}) + (\theta_0 - \hat{\theta})' \frac{\partial l(\hat{\theta})}{\partial \theta} + \frac{1}{2}(\theta_0 - \hat{\theta})' \frac{\partial^2 l(\tilde{\theta})}{\partial \theta \partial \theta}(\theta_0 - \hat{\theta}) \tag{87}$$

where $\tilde{\theta}(\omega)$ is between θ_0 and $\hat{\theta}(\omega)$. Since mle is the root of the likelihood equation, $\frac{\partial l(\hat{\theta})}{\partial \theta} = 0$. We have

$$\begin{aligned} \xi_{\mathrm{LR}} &= -(\theta_0 - \hat{\theta})' \frac{\partial^2 l(\tilde{\theta})}{\partial \theta \partial \theta}(\theta_0 - \hat{\theta}) & (88) \\ &= \sqrt{n}(\theta_0 - \hat{\theta})' \left(-\frac{1}{n} \frac{\partial^2 l(\tilde{\theta})}{\partial \theta \partial \theta} \right) \sqrt{n}(\theta_0 - \hat{\theta}) & (89) \end{aligned}$$

Proceeding as in our derivations of the properties of the maximum likelihood estimator,

$$\begin{aligned} \sqrt{n}(\hat{\theta} - \theta_0) &\Rightarrow \mathcal{N}(0, I(\theta_0)^{-1}) & (90) \\ -\frac{1}{n} \frac{\partial^2 l(\tilde{\theta})}{\partial \theta \partial \theta'} &\xrightarrow{P} I(\theta_0) & (91) \end{aligned}$$

so that by Slutsky and the Continuous Mapping Theorem,

$$\xi_{\mathrm{LR}} \overset{H_0}{\Rightarrow} \chi^2_k \tag{92}$$

An asymptotically justified level $1-\alpha$ confidence set based on the LR statistic is hence of the form

$$\theta^* \mid (\hat{\theta}-\theta^*)' J(\hat{\theta})(\hat{\theta}-\theta^*) < c \tag{93}$$

where $J(\hat{\theta}) = \left(-\frac{\partial^2 l(\hat{\theta})}{\partial\theta\partial\theta'}\right)$ and c solves $P(\chi^2_k > c) = \alpha$. This confidence set may be recognized as the interior of an ellipse centered at $\theta = \hat{\theta}$. In the one-dimensional case, we obtain a confidence interval $(\hat{\theta} - c^* J(\hat{\theta})^{-1/2}, \hat{\theta} + c^* J(\hat{\theta})^{-1/2})$ where c^* is the positive number that solves $P(\mathcal{N}(0,1) > c^*) = \alpha/2$.

6.2 Pearson's chi-square

(Lehman 5.5, Ferguson 9,10, Rao 6b) Let $\underline{X}_1, \underline{X}_2, \cdots, \underline{X}_n$ be iid from a multinomial$_k(1, \underline{p})$ distribution, where $\underline{p}$ is a k-vector with nonnegative entries that sum to one. That is,

$$P(\underline{X}_i = e_j) = p_j \quad \text{for all} \quad 1 \le j \le k \tag{94}$$

where $e_j =$ the k vector with 1 at the j-th position and 0's everywhere else.

Note that the multinomial distribution is a generalization of the binomial distribution to the case in which there are k categories of outcome instead of only 2. Also note that we ordinarily do not consider a binomial random variable to be a 2-vector, but we could easily do so if the vector contained both the number of successes and the number of failures. Equation (94) implies that the random vector $\underline{X}_i$ has expectation $\underline{p}$ and covariance matrix

$$\Sigma = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_k \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_k \\ \vdots & \vdots & \ddots & \vdots \\ -p_1p_k & -p_2p_k & \cdots & p_k(1-p_k) \end{pmatrix} \tag{95}$$

Using the Cramer-Wold device, the multivariate central limit theorem implies

$$\sqrt{n}(\bar{\underline{X}}_n - \underline{p}) \Rightarrow \mathcal{N}_k(\underline{0}, \Sigma). \tag{96}$$

Note that the sum of the j-th column of Σ is $p_j - p_j(p_1 + \cdots + p_k) = 0$, which is to say that the sum of the rows of Σ is the zero vector, so Σ is not invertible.

We wish to derive the asymptotic distribution of Pearson's chi-square statistic

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j}, \tag{97}$$

where n_j is the random variable that is the j-th component if $n\bar{\underline{X}}_n$, the number of successes in the j-th category for trials $1, \cdots, n$. We will discuss two different ways to do this. One way avoids dealing with the singular matrix Σ, whereas the other does not.

In the first approach, define for each i, $\underline{Y}_i = (\underline{X}_{i1}, \cdots, \underline{X}_{ik-1})$. That is, let $\underline{Y}_i$ be the $k-1$-vector consisting of the first $k-1$ components of $\underline{X}_i$. Then the covariance matrix of $\underline{Y}_i$ is the upper-left $(k-1) \times (k-1)$ submatrix of Σ, which we denote by Σ^*. Similarly, let $\underline{p}^*$ denote the vector $(p_1, \cdots, p_{k-1})$. First, verify that Σ^* is invertible and that

$$\Sigma^{*-1} = \begin{pmatrix} \frac{1}{p_1} + \frac{1}{p_k} & \frac{1}{p_k} & \cdots & \frac{1}{p_k} \\ \frac{1}{p_k} & \frac{1}{p_2} + \frac{1}{p_k} & \cdots & \frac{1}{p_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{p_k} & \frac{1}{p_k} & \cdots & \frac{1}{p_{k-1}} + \frac{1}{p_k} \end{pmatrix} \tag{98}$$

Second, verify that

$$\chi^2 = n(\underline{\bar{Y}}_n - \underline{p}^*)^t (\Sigma^*)^{-1} (\underline{\bar{Y}}_n - \underline{p}^*) \tag{99}$$

The facts in equations (98) and (99) are checked in exercise 1. If we now define

$$\underline{Z}_n = \sqrt{n}(\Sigma^*)^{-1/2}(\underline{\bar{Y}}_n - \underline{p}^*), \tag{100}$$

then clearly the central limit theorem implies $\underline{Z}_n \Rightarrow \mathcal{N}_{k-1}(\underline{0}, I)$. By definition, the χ^2_{k-1} distribution is the distribution of the sum of the squares of $k-1$ independent standard normal random variables. Therefore,

$$\chi^2 = (\underline{Z}_n)^t \underline{Z}_n \Rightarrow \chi^2_{k-1}, \tag{101}$$

which is the result that leads to the familiar chi-square test.

In a second approach to deriving the limiting distribution (101), we use some properties of projection matrices.

Definition 3. *A matrix P is called a projection matrix if it is idempotent; that is, if $P^2 = P$.*

The following lemmas, to be proven in exercise 2, give some basic facts about projection matrices.

Lemma 2. *Suppose P is a projection matrix. Then every eigenvalue of P equals 0 or 1. Suppose that r denotes the number of eigenvalues of P equal to 1. Then if $Z \sim \mathcal{N}_k(\underline{0}, P)$, then, $Z^t Z \sim \chi^2_r$.*

This can be derived from the Fisher-Cochran Theorem.

Lemma 3. *The trace of a square matrix equals the sum of its eigenvalues. For matrices A and B whose sizes allow them to be multiplied in either order,* $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$.

Define $\Gamma = \text{diag}(\underline{p})$. Clearly, equation (96) implies

$$\sqrt{n}\Gamma^{-1/2}(\underline{\bar{X}}_n - \underline{p}) \Rightarrow \mathcal{N}_k(\underline{0}, \Gamma^{-1/2}\Sigma\Gamma^{-1/2}). \tag{102}$$

Since Σ may be written in the form $\Gamma - \underline{pp}^t$,

$$\Gamma^{-1/2}\Sigma\Gamma^{-1/2} = I - \Gamma^{-1/2}\underline{pp}^t\Gamma^{-1/2} = I - \sqrt{\underline{p}}\sqrt{\underline{p}}^t \tag{103}$$

clearly has trace $k-1$; furthermore, $(I - \sqrt{\underline{p}}\sqrt{\underline{p}}^t)(I - \sqrt{\underline{p}}\sqrt{\underline{p}}^t) = I - 2\sqrt{\underline{p}}\sqrt{\underline{p}}^t + \sqrt{\underline{p}}\sqrt{\underline{p}}^t\sqrt{\underline{p}}\sqrt{\underline{p}}^t = I - \sqrt{\underline{p}}\sqrt{\underline{p}}^t$ because $\sqrt{\underline{p}}^t\sqrt{\underline{p}} = 1$, so the covariance matrix (103) is a projection matrix.

Define $\Delta_n = \sqrt{n}\Gamma^{-1/2}(\underline{\bar{X}} - \underline{p})$. Then we may check (exercise 2) that

$$\chi^2 = (\Delta_n)^t\Delta_n \tag{104}$$

Therefore, since the covariance matrix (103) is a projection with trace $k-1$, Lemma 2 and Lemma 3 prove that $\chi^2 \Rightarrow \chi^2_{k-1}$ as desired.

6.3 Contingency tables

(Rao 6d pp403, Lehman 7.8) We are interested in testing H_0 : X indep of Y Vs H_1 : not indep in a two-way contingency table. Note that this is a composite null hypothesis in a multinomial setting, as opposed to the simple hypothesis of the previous section.

Let n_{ij} = the frquency in (i,j)-th cell and m_{ij} = estimate of $E(n_{ij})$ under H_0.

Then the χ^2 test rejects H_0 for large values of the following test statistic

$$\chi^2 = \sum_i (n_{ij} - m_{ij})^2/m_{ij}$$

The asymptotic distribution of this under H_0 is a χ^2 distribution. We shall not prove this here. It follows from similar calculations as Pearson's χ^2. The calculations are more tedious since now the rank of the covariance matrix is (number of classes - 1 - number of free parameters estimated).

There is a second method of deriving the asymptotic distribution. It relies on the fact that under the regularity conditions the null distribution of -2*log(likelihood ratio) converges to a χ^2 distribution.

If $n \geq 30$ and $n_{ij}, m_{ij} \geq 5$ for all i, j, then it is customary to consider the asymptotic distribution as a good approximation to the exact distribution. We reject H_0 for large values of the test statistic.

If the sample size is small, then one method is to perform Fisher's exact test, which finds the exact conditional distribution under H_0 given the marginals. This conditional is like hypergeometric distribution. It is obtained by conditioning multinomial distribution in the same way as hypergeometric distribution is obtained by conditioning binomial.

Another solution will be to list all possible tables with the given marginals. These are equally likely under H_0. Compute the test statistic for all these, and get a histogram. This is the true distribution of the test statistic under H_0. Now locate the test statistic value for the given data in this histogram to find p-value.

Exercises

1. Verify equations (98) and (99).

2. Prove Lemma 2 and Lemma 3; then verify equation (104).

3. The following example comes from genetics. There is a particular characteristic of human blood (the so-called MN blood group) that has three types: M, MN, and N. Under idealized circumstances, which we assume to be true for the purposes of this problem, these three types occur in the population with probabilities $p_1 = \pi_M^2, p_2 = 2\pi_M\pi_N$, and $p_3 = \pi_N^2$, respectively, where π_M is the frequency of the M allele in the population and $\pi_N = 1 - \pi_M$ is the frequency of the N allele. If the value of π_M were known, then the asymptotic distribution of the Pearson χ^2 statistic would be given in the development earlier in this topic. However, of course we usually don't know π_M. Instead, we estimate it using the maximum likelihood estimator $(2n_1 + n_2)/2n$.

   (a) Define $B_n = \sqrt{n}(\underline{\bar{X}} - \underline{\hat{p}})$, where $\underline{\hat{p}}$ is the MLE for p. Use the delta method to derive the asymptotic distribution of $\Gamma^{-1/2}B_n$.

   (b) Define $\hat{\Gamma}$ to be the diagonal matrix with entries $\hat{p}_1, \hat{p}_2, \hat{p}_3$ along its diagonal. Derive the asymptotic distribution of $\hat{\Gamma}^{-1/2}B_n$.

   (c) Derive the asymptotic distribution of the Pearson chi-square statistic

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - n\hat{p}_j)^2}{n\hat{p}_j}. \tag{105}$$

4. Take $\pi_M = .75$ and $n = 100$ in the situation described in exercise 3. Simulate 500 realizations of the data.

   (a) Compute $\sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j}$ for each of your 500 datasets. Compare the empirical cdf of these statistics with both the χ_1^2 and χ_2^2 cdf's. Comment on what you observe.

   (b) Compute the χ^2 statistic of equation (105) for each of your 500 datasets. Compare the empirical cdf of these statistics with both the χ_1^2 and χ_2^2 cdf's. Comment on what you observe.

5. Show that the LR statistic is invariant to reparametrizations of the parameters.

6. Under the alternative $\theta = \theta_1$, where θ_1 is fixed, the power of the LR test converges to 1. (Hint: Use Chebyshev and the moments)

7. If $\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow \mathcal{N}(g, I(\theta_0)^{-1})$, show that $\xi_{\text{LR}} \Rightarrow \chi^2_k(\delta)$ where δ is the noncentrality parameter and equals $g'Ig$.

8. Problem 10.31 of Casella and Berger, pg 511.

7 Nonparametrics Cont

7.1 U-statistic

Ch 5 of Serfling

7.2 Rank procedures

Ch 9 of Serfling

Exercises

1. Show that the kernel $h(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2$ leads to the U-statistic s^2.
2. Problem 5.P.3 of Serfling
3. Problem 5.P.4 of Serfling
4. Problem 5.P.5 of Serfling
5. Problem 5.P.6 of Serfling
6. Problem 5.P.9 of Serfling
7. Problem 5.P.10 of Serfling
8. Problem 5.P.16 of Serfling
9. Problem 5.P.22 of Serfling using result of Example 5.5.2A
10. Problem 5.P.22 of Serfling using delta method.
11. Problem 5.P.23 of Serfling
12. The Wilcoxon signed rank test is often used to test for symmetry of the distribution about the origin. This test is based on the statistic

$$W_n^+ = \sum_{i=1}^{n} R_i^+ I(Z_i > 0)$$

where R_i^+ is the rank of $\mid Z_i \mid$ among $\mid Z_1 \mid, \mid Z_2 \mid, \cdots, \mid Z_n \mid$. Although it is not a U-statistic, show that W_n^+ is a linear combination of two U-statistics,

$$W_n^+ = \sum_i I(Z_i > 0) + \sum_{i<j} I(Z_i + Z_j > 0) = nU^{(1)} + \binom{n}{2} U^{(2)}$$

The first U-statistic is based on the kernel, $h(z) = I(z > 0)$. This is the U-statistic used for the sign test. The second U-statistic is based on the kernel, $h(z_1, z_2) = I(z_1 + z_2 > 0)$. For large n the second term dominates the first, so asymptotically W_n^+ behaves like $n^2 U_n^{(2)}/2$. Show that when F is indeed symmetric,that is $P(X < 0) = 1/2$ then $\sqrt{n}(U_n^{(2)} - 1/2) \Rightarrow \mathcal{N}(0, 1/3)$.

13. Derive the asymptotic distribution of the rank test statistic 9.1.1(ii) under the null hypothesis using the Wald and Wolfowitz theorem of 9.2.2

14. Derive the asymptotic distribution of the above rank test statistic under the alternative specified in 9.1.1(ii) using the Chernoff and Savage theorem of 9.2.3

15. Prove lemma 9.2.5

16. Show that the projection of Kendall's tau on the ranks is Spearmen's rho upto a constant.

8 Bayesian

A Short Course on Bayesian Inference (based on *An Introduction to Bayesian Analysis: Theory and Methods* by **Ghosh, Delampady and Samanta**) **Module 2**

1 Large Sample Methods in Bayesian Inference

In order to make Bayesian inference about a parameter θ with model $f(\boldsymbol{x}|\theta)$, one needs to choose an appropriate prior $\pi(\theta)$ for θ. Exact or approximate computation of various features of the posterior $\pi(\theta|\boldsymbol{x})$ is a major challenge for Bayesians. Under some regularity conditions, the posterior can be approximated by a normal distribution with the MLE as the mean (or mode), and inverse of the Fisher information matrix as the posterior variance-covariance matrix. If more accuracy is needed, one may have to go for an asymptotic expansion of the posterior. Alternatively, one may sample from the approximated posterior (or some type of $t-$distribution) and use importance sampling. An intuitive rationale behind posterior normality is given below.

How the posterior inference is influenced by a particular prior depends on the relative magnitude of the amount of information in the data, which for iid observations can be measured by the sample size n or $nI(\theta)$ ($I(\theta)$ being the per unit Fisher information) or observed Fisher information

$$\hat{\boldsymbol{I}}_n = -\frac{\partial^2 \log f(\boldsymbol{x}|\theta)}{\partial\theta\partial\theta^T}|_{\theta=\hat{\theta}},$$

$\hat{\theta}$ being the MLE of θ. As the sample size grows, the influence of the prior diminishes. Thus, for large samples, a precise formulation of the prior is not necessary. In most instances when the parameter space is low-dimensional, the prior is washed away by the data. Another important asymptotic fact is consistency of the posterior which we now describe below. In general, the limiting results to be discussed provide a frequentist validation of Bayesian analysis.

Consistency of Posterior Distribution:

Suppose a data sequence $X_1, \ldots, X_n, \ldots$ is generated as iid with a common density $f(x|\theta_0)$. Would a Bayesian analyzing this data with a prior $\pi(\theta)$ be able to learn about θ_0? Ideally, the updated knowledge about θ represented by its posterior should become more and more concentrated near θ_0 as the sample size increases. This asymptotic property is known as the consistency of the posterior distribution at θ_0. Let $X_1, \ldots, X_n$ be iid with joint pdf $f(\boldsymbol{x}_n|\theta), \theta \in \Theta \subset R^p$. Let $\pi(\theta)$ denote the prior pdf and $\pi(\theta|\boldsymbol{X}_n)$ the posterior pdf. Let $\Pi(\cdot|\boldsymbol{X}_n)$ denote the corresponding posterior distribution of θ.

Definition 1. The sequence of posterior distributions $\Pi(\cdot|\boldsymbol{X}_n)$ of θ is said to be consistent at $\theta = \theta_0 \in \Theta$ if for every neighborhood U of θ_0, $\Pi(U|\boldsymbol{X}_n) \to 1$ as $n \to \infty$ with probability 1 wrt to the distribution (of $\boldsymbol{X}_n$) under θ_0.

From the definition of convergence in distribution, it follows that consistency of $\Pi(\cdot|\boldsymbol{X}_n)$ at θ_0 is equivalent to the fact that $\Pi(\cdot|\boldsymbol{X}_n) \overset{d}{\to}$ a distribution degenerate at θ_0 with probability 1 under θ_0.

Example 1. Let $X_1, \ldots, X_n$ be iid Bernoulli observations with $P_\theta(X_1 = 1) = \theta$. Consider a Beta$(\alpha, \beta)$ prior density for θ. The posterior density of θ given $X_1, \ldots, X_n$ is then a Beta$(\sum_{i=1}^n X_i + \alpha, n - \sum_{i=1}^n X_i + \beta)$ distribution with

$$E(\theta|\boldsymbol{X}_n) = \frac{n\bar{X}_n + \alpha}{n + \alpha + \beta}, \quad \text{Var}(\theta|\boldsymbol{X}_n) = \frac{(n\bar{X}_n + \alpha)(n - n\bar{X}_n + \beta)}{(n + \alpha + \beta)^2(n + \alpha + \beta + 1)}.$$

Note that $\bar{X}_n \overset{\text{a.s. } (P_{\theta_0})}{\longrightarrow} \theta_0$ as $n \to \infty$ by strong law of large numbers. Hence $E(\theta|\boldsymbol{X}_n) \overset{\text{a.s. } (P_{\theta_0})}{\longrightarrow} \theta_0$, $\text{Var}(\theta|\boldsymbol{X}_n) \overset{\text{a.s. } (P_{\theta_0})}{\longrightarrow} 0$. Then,

$$\begin{aligned} & P\{\theta \notin [\theta_0 - \epsilon, \theta_0 + \epsilon]|\boldsymbol{X}_n\} = P(|\theta - \theta_0| > \epsilon|\boldsymbol{X}_n) \\ \leq \; & \frac{E[(\theta - \theta_0)^2|\boldsymbol{X}_n]}{\epsilon^2} = \frac{\text{Var}(\theta|\boldsymbol{X}_n) + \{E(\theta|\boldsymbol{X}_n) - \theta_0\}^2}{\epsilon^2} \\ & \overset{\text{a.s. } (P_{\theta_0})}{\longrightarrow} 0 \quad \text{as} \quad n \to \infty. \end{aligned}$$

An important consequence of the consistency of the posterior is the robustness of Bayesian inference with respect to the choice of prior. Let $X_1, \ldots, X_n$ be iid and π_1 and π_2 be two prior pdf's positive and continuous at θ_0, an interior point of Θ such that $\Pi_1(\cdot|\boldsymbol{X}_n)$ and $\Pi_2(\cdot|\boldsymbol{X}_n)$ are both consistent at θ_0. Then with probability 1 under θ_0,

$$\int_\Theta |\pi_1(\theta|\boldsymbol{X}_n) - \pi_2(\theta|\boldsymbol{X}_n)| d\theta \to 0$$

or equivalently, $\sup_A |\Pi_1(A|\boldsymbol{X}_n) - \Pi_2(A|\boldsymbol{X}_n)| \to 0$ as $n \to \infty$. Thus two different choices of prior density lead approximately to the same posterior distribution.

Asymptotic Normality of the Posterior

Large sample Bayesian methods are primarily based on normal approximation to the posterior distribution of θ. As the sample size n increases, the posterior distribution approaches normality under certain regularity conditions and concentrates in the neighborhood of the posterior mode. Suppose $\tilde{\theta}_n$ is the posterior mode and the first-order partial derivatives of $\log \pi(\theta|\boldsymbol{X}_n)$ vanish at $\tilde{\theta}_n$. Define

$$\tilde{I}_n = -\frac{\partial^2 \log \pi(\theta|\boldsymbol{X}_n)}{\partial\theta\partial\theta^T}|\theta = \tilde{\theta}_n.$$

Then a formal Taylor expansion gives

$$\begin{aligned}
\log \pi(\theta|\boldsymbol{X}_n) &\doteq \log \pi(\tilde{\theta}|\boldsymbol{X}_n) - \frac{1}{2}(\theta - \tilde{\theta}_n)^T[-\frac{\partial^2 \log \pi(\theta|\boldsymbol{X}_n)}{\partial\theta\partial\theta^T}|\theta = \tilde{\theta}_n](\theta - \tilde{\theta}_n) \\
&= \log \pi(\tilde{\theta}|\boldsymbol{X}_n) - \frac{1}{2}(\theta - \tilde{\theta}_n)^T \tilde{I}_n(\theta - \tilde{\theta}_n).
\end{aligned}$$

Hence

$$\begin{aligned}
\pi(\theta|\boldsymbol{X}_n) &\doteq \pi(\tilde{\theta}|\boldsymbol{X}_n)\exp[-\frac{1}{2}(\theta - \tilde{\theta}_n)^T \tilde{I}_n(\theta - \tilde{\theta}_n)] \\
&\propto \exp[-\frac{1}{2}(\theta - \tilde{\theta}_n)^T \tilde{I}_n(\theta - \tilde{\theta}_n)],
\end{aligned}$$

which is the kernel of a $N_p(\theta|\tilde{\theta}_n, \tilde{I}_n^{-1})$ density (with p being the dimension of θ).

As the posterior density becomes highly concentrated in a neighborhood of the posterior mode where the prior $\pi(\theta)$ is nearly constant (this is true for a diffuse prior), the posterior is essentially the same as the likelihood $f(\boldsymbol{X}_n|\theta)$. Then we may replace, to the first order of approximation, $\tilde{\theta}_n$ by $\hat{\theta}_n$ and $\tilde{I}_n$ by $\hat{I}_n$ where $\hat{\theta}_n$ is the maximum likelihood estimator (MLE) of θ.

Remark 1. From the above discussion it follows that for iid $X_1, \ldots, X_n|\theta$, we have several ways to approximate the posterior density either by $N_p(\tilde{\theta}_n, \tilde{I}_n^{-1})$ or $N_p(\hat{\theta}_n, \hat{I}_n^{-1})$ or $N_p(\hat{\theta}_n, I^{-1}(\hat{\theta}_n))$, where $I(\theta)$ is the total Fisher information in $\boldsymbol{X}_n$. In particular, under suitable regularity conditions, $\hat{I}_n^{1/2}(\theta - \hat{\theta}_n)$ given $\boldsymbol{X}_n$ converges to $N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ with probability 1 (P_θ). This is comparable with the classical statistical theory where $\hat{I}_n^{1/2}(\theta - \hat{\theta}_n)|\theta \xrightarrow{d} N_p(\boldsymbol{0}, \boldsymbol{I}_p)$.

A Formal Result on Asymptotic Normality of the Posterior Distribution

Let $X_1, \ldots, X_n|\theta$ be iid with a cdf $F(x|\theta)$ and a pdf $f(x|\theta)$. For simplicity, let θ be a scalar with $\theta \in \Theta$ an open subset of R. Fix $\theta_0 \in \Theta$, the "true" value of θ, and all probability statements will be made under P_{θ_0}. Let $l(\theta, x) = \log f(x|\theta)$, $L_n(\theta) = \sum_{i=1}^n l(\theta, X_i)$ and $h^{(i)}$ a generic notation for the ith derivative of a function $h(X, \theta)$ with respect to θ. The function $h(\cdot)$ may not involve X explicitly. Assume the following regularity conditions.

I. The set $\{x : f(x|\theta) > 0\}$ is the same for all $\theta \in \Theta$, i.e., the support does not depend on the parameter.

II. The function $l(\theta, x)$ is thrice differentiable with respect to θ in a neighborhood $(\theta_0 - \delta, \theta_0 + \delta)$ of θ_0 and $\sup_{\theta\in(\theta_0-\delta,\theta_0+\delta)} |l^{(3)}(\theta, x)| \leq M(x)$ with $E_{\theta_0}[M(X_1)] < \infty$.

III. $E_{\theta_0}[l^{(1)}(\theta_0, X)] = 0, 0 < E_{\theta_0}[-l^{(2)}(\theta_0, X)] = E_{\theta_0}[l^{(1)}(\theta_0, X)]^2 < \infty$.

IV. For any $\delta > 0$, $\sup_{|\theta-\theta_0|>\delta} n^{-1}[L_n(\theta) - L_n(\theta_0)] < -\epsilon$ for some $\epsilon > 0$ and all n sufficiently large.

Remark 2. Suppose there exists a sequence of estimators $\{\theta_n^*\}$ of θ such that $\theta_n^* \to \theta_0$ with probability 1 (P_{θ_0}). Then there exists a solution $\hat{\theta}_n$ of the likelihood equation $L_n^{(1)}(\theta) = 0$, i.e., there exists a sequence $\hat{\theta}_n$ of statistics such that with probability 1 (P_{θ_0}), $L_n^{(1)}(\hat{\theta}_n) = 0$ for sufficiently large n and $\hat{\theta}_n \overset{\text{a.s. } (P_{\theta_0})}{\longrightarrow} \theta_0$.

Theorem 1. Suppose assumptions (I)-(IV) hold and $\hat{\theta}_n$ is a strongly consistent solution of the likelihood equations. Then for any prior density $\pi(\theta)$ which is continuous and positive at θ_0,

$$\lim_{n\to\infty} \int_R |\pi_n^*(t|\boldsymbol{X}_n) - \frac{\sqrt{I(\theta_0)}}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2 I(\theta_0)}|dt = 0 \tag{1}$$

with P_{θ_0}−probability one, where $\pi_n^*(t|\boldsymbol{X}_n)$ is the posterior density of $T_n = \sqrt{n}(\theta - \hat{\theta}_n)$ given $\boldsymbol{X}_n$. Also, under the same assumptions, (1) holds with $I(\theta_0)$ replaced by $-n^{-1}L_n^{(2)}(\hat{\theta}_n)$.

Proof. Recall that the posterior density of θ, $\pi_n(\theta|\boldsymbol{X}_n) \propto [\prod_{i=1}^n f(X_i|\theta)]\pi(\theta) \propto \exp[L_n(\theta) - L_n(\hat{\theta}_n)]\pi(\theta)$. Hence, the posterior pdf of $T_n = \sqrt{n}(\theta - \hat{\theta}_n)$ is given by

$$\pi_n^*(t|\boldsymbol{X}_n) = C_n^{-1} \exp[L_n(\hat{\theta}_n + n^{-\frac{1}{2}}t) - L_n(\hat{\theta}_n)]\pi(\hat{\theta}_n + n^{-\frac{1}{2}}t), \tag{2}$$

where $C_n = \int_R \exp[L_n(\hat{\theta}_n + n^{-\frac{1}{2}}t) - L_n(\hat{\theta}_n)]\pi(\hat{\theta}_n + n^{-\frac{1}{2}}t)dt$. Let

$$g_n(t) = \exp[L_n(\hat{\theta}_n + n^{-\frac{1}{2}}t) - L_n(\hat{\theta}_n)]\pi(\hat{\theta}_n + n^{-\frac{1}{2}}t) - \exp[-\frac{1}{2}t^2 I(\theta_0)]\pi(\theta_0). \tag{3}$$

Suppose we show that $\int_R |g_n(t)|dt \to 0$ as $n \to \infty$. Then $C_n \to \int_R \pi(\theta_0)\exp(-\frac{t^2}{2}I(\theta_0))dt = \pi(\theta_0)(2\pi)^{1/2}I^{-1/2}(\theta_0)$. Then the integral in (1) is dominated by

$$C_n^{-1}\int_R |g_n(t)|dt + \int_R |C_n^{-1}\pi(\theta_0)\exp[-\frac{1}{2}t^2 I(\theta_0)] - N(t|0, I^{-1}(\theta_0))|dt \to 0.$$

In order to prove that $\int_R |g_n(t)|dt \to 0$, we write $R = A_1 \cup A_2$, where $A_1 = \{t : |t| > \delta_0\sqrt{n}\}$ and $A_2 = A_1^c$. First,

$$\int_{A_1} |g_n(t)|dt \le \int_{A_1} \pi(\hat{\theta}_n + n^{-\frac{1}{2}}t)\exp[L_n(\hat{\theta}_n + n^{-\frac{1}{2}}t) - L_n(\hat{\theta}_n)]dt + \int_{A_1} \pi(\theta_0)\exp[-\frac{1}{2}t^2 I(\theta_0)]dt. \tag{4}$$

Now

$$\int_{A_1} \pi(\theta_0) \exp[-\frac{1}{2}t^2 I(\theta_0)]dt = \pi(\theta_0) \int_{|t|>\delta_0\sqrt{n}} \exp[-\frac{1}{2}t^2 I(\theta_0)]dt \to 0 \quad \text{as} \quad n \to \infty. \tag{5}$$

Moreover, since by (IV) for $t \in A_1$, $n^{-1}|L_n(\hat{\theta}_n + n^{-\frac{1}{2}}t) - L_n(\hat{\theta}_n)| < -\epsilon$ for all sufficiently large n,

$$\text{First term in (4)} < \exp(-n\epsilon) \int_{A_1} \pi(\hat{\theta}_n + n^{-\frac{1}{2}}t)dt \to 0 \quad \text{as } n \to \infty. \tag{6}$$

Combine (5) and (6) to get (4).

Next to prove $\int_{A_2} |g_n(t)|dt \to 0$ as $n \to \infty$, first by Taylor expansion, and $L_n^{(1)}(\hat{\theta}_n) = 0$,

$$L_n(\hat{\theta}_n + n^{-\frac{1}{2}}t) - L_n(\hat{\theta}_n) = -\frac{t^2}{2}\hat{I}_n + R_n(t), \tag{7}$$

where $R_n(t) = (1/6)(t/\sqrt{n})^3 L_n^{(3)}(\theta'_n)$, $|\theta'_n - \hat{\theta}_n| < |t|/\sqrt{n}$. Now by assumption (II), for each real t, $R_n(t) \overset{\text{a.s. } (P_{\theta_0})}{\longrightarrow} 0$ as $n \to \infty$. So, $g_n(t) \overset{\text{a.s. } (P_{\theta_0})}{\longrightarrow} 0$. Next for suitably chosen δ_0, for any $t \in A_2$,

$$|R_n(t)| \leq \frac{1}{6}\delta_0 t^2 n^{-1} \sum_{i=1}^n M(X_i) < \frac{1}{4}t^2 \hat{I}_n \ a.s.(P_{\theta_0})$$

for sufficiently large n so that from (7),

$$\exp[L_n(\hat{\theta}_n + n^{-\frac{1}{2}}t) - L_n(\hat{\theta}_n)] < \exp(-\frac{1}{4}t^2 \hat{I}_n) < \exp[-\frac{t^2}{8}I(\theta_0)],$$

a.s. for large n. Hence, for a suitably chosen $\delta_0 > 0$, $|g_n(t)|$ is dominated by an integrable function on A_2. Applying the dominated convergence theorem, $\int_{A_2} |g_n(t)|dt \to 0$ as $n \to \infty$.

Remark 3. We assume in the proof that $\pi(\theta)$ is a proper pdf. However, the result continues to hold even for improper prior $\pi(\theta)$ provided there exists n_0 such that the "posterior" $\pi(\theta|X_1, \ldots, X_{n_0})$ is proper a.e.

We next show that if $\hat{\theta}_n^B = \int_R \theta \pi_n(\theta|\boldsymbol{X}_n)d\theta$ is finite, then $\sqrt{n}(\hat{\theta}_n^B - \hat{\theta}_n) \to 0$ with probability 1 (P_{θ_0}) as $n \to \infty$ under some conditions.

Theorem 2. Suppose in addition to (I)-(IV), $\int \theta\pi(\theta)d\theta < \infty$. Then

$$\int_R |t||\pi_n^*(t|\boldsymbol{X}_n) - N(t|0, I^{-1}(\theta_0))|dt \to 0 \text{ with probability } 1(P_{\theta_0}).$$

Remark 4. The above result implies that

$$\int_R t\pi_n^*(t|\boldsymbol{X}_n)dt \to \int_R tN(t|0, I^{-1}(\theta_0))dt = 0.$$

Hence, $\hat{\theta}_n^B = E(\theta|\boldsymbol{X}_n) = E[\hat{\theta}_n + \frac{t}{\sqrt{n}}|\boldsymbol{X}_n] = \hat{\theta}_n + E[\frac{t}{\sqrt{n}}|\boldsymbol{X}_n]$. Hence, $\sqrt{n}(\hat{\theta}_n^B - \hat{\theta}_n) = \int_R t\pi_n^*(t|\boldsymbol{X}_n)dt \to 0$ as $n \to \infty$.

Laplace Approximation

Bayesian analysis requires evaluation of integrals of the form $\int g(\theta)f(\boldsymbol{x}|\theta)\pi(\theta)d\theta$. For example, when $g(\theta) = 1$, the integral reduces to the marginal likelihood of $\boldsymbol{X}$. The posterior mean requires evaluation of two integrals $\int \theta f(\boldsymbol{x}|\theta)\pi(\theta)d\theta$ and $\int f(\boldsymbol{x}|\theta)\pi(\theta)d\theta$. Laplace's method is a technique for approximating integrals when the integrand has a sharp maximum in the interior of the domain of integration.

Laplaces's method

Consider an integral of the form $I = \int_{-\infty}^{\infty} q(\theta)\exp[nu(\theta)]d\theta$ where q and u are smooth functions of θ with u having a unique maximum at $\hat{\theta}$. In applications, $nu(\theta) = \sum_{i=1}^n l(X_i, \theta)$, the log-likelihood function or the logarithm of the unnormalized posterior density $f(\boldsymbol{x}|\theta)\pi(\theta)$ with corresponding $\hat{\theta}$ equal to the posterior mode. The idea is that if u has a unique sharp maximum at $\hat{\theta}$, the most contribution to the integral I comes from the integral over a small neighborhood $(\hat{\theta}-\delta, \hat{\theta}+\delta)$ of $\hat{\theta}$. We study the behavior of I as $n \to \infty$. As $n \to \infty$,

$$I \approx I_1 = \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} q(\theta)\exp[nu(\theta)]d\theta.$$

Laplace's method involves Taylor series expansion of q and u about $\hat{\theta}$ which gives

$$\begin{aligned} I &\approx \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} [q(\hat{\theta}) + (\theta-\hat{\theta})q'(\hat{\theta}) + \frac{1}{2}(\theta-\hat{\theta})^2 q''(\hat{\theta}) + \text{ smaller terms}] \\ &\quad \times \exp[nu(\hat{\theta}) + nu'(\hat{\theta})(\theta-\hat{\theta}) + \frac{n}{2}u''(\hat{\theta})(\theta-\hat{\theta})^2 + \text{ smaller terms}]d\theta \\ &\approx \exp[nu(\hat{\theta})]q(\hat{\theta})\int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta}[1 + \frac{q'(\hat{\theta})}{q(\hat{\theta})}(\theta-\hat{\theta}) + \frac{1}{2}(\theta-\hat{\theta})^2\frac{q''(\hat{\theta})}{q(\hat{\theta})}]\exp[\frac{n}{2}u''(\hat{\theta})(\theta-\hat{\theta})^2]d\theta. \end{aligned}$$

Assume that $c = -u''(\hat{\theta}) > 0$ (e.g., when $u(\theta) = n^{-1}\log f(\boldsymbol{x}|\theta)$) and letting $t = \sqrt{nc}(\theta -$

$\hat{\theta}$), we have

$$\begin{aligned} I &\approx \exp[nu(\hat{\theta})]q(\hat{\theta})\frac{1}{\sqrt{nc}}\int_{-\delta\sqrt{nc}}^{\delta\sqrt{nc}}[1+\frac{t}{\sqrt{nc}}\frac{q'(\hat{\theta})}{q(\hat{\theta})}+\frac{t^2}{2nc}\frac{q''(\hat{\theta})}{q(\hat{\theta})}]\exp(-\frac{t^2}{2})dt \\ &\approx \exp[nu(\hat{\theta})]q(\hat{\theta})\frac{1}{\sqrt{nc}}\int_{-\infty}^{\infty}[1+\frac{t}{\sqrt{nc}}\frac{q'(\hat{\theta})}{q(\hat{\theta})}+\frac{t^2}{2nc}\frac{q''(\hat{\theta})}{q(\hat{\theta})}]\exp(-\frac{t^2}{2})dt \\ &= \exp[nu(\hat{\theta})]q(\hat{\theta})\frac{\sqrt{2\pi}}{\sqrt{nc}}[1+\frac{q''(\hat{\theta})}{2ncq(\hat{\theta})}] = \exp[nu(\hat{\theta})]q(\hat{\theta})\frac{\sqrt{2\pi}}{\sqrt{nc}}[1+O(n^{-1})]. \end{aligned}$$

In general, for the case with a $p-$dimensional parameter vector θ,

$$I = \exp[nu(\hat{\theta})]q(\hat{\theta})\frac{(2\pi)^{p/2}}{n^{p/2}}|\Delta_u(\hat{\theta})|^{-\frac{1}{2}}[1+O(n^{-1})],$$

where $\Delta_u(\theta) = (-\frac{\partial^2 u(\theta)}{\partial\theta_i\partial\theta_j})_{p\times p}$.

The Bayesian Information Criterion (BIC)

Consider a model with a likelihood $f(\boldsymbol{x}|\theta)$ and prior $\pi(\theta)$. Letting $q(\theta) = \pi(\theta)$ and $nu(\theta) = \sum_{i=1}^n l(X_i,\theta)$, the log-likelihood, one can find an approximation to the marginal $\int f(\boldsymbol{x}|\theta)\pi(\theta)d\theta$. This approximation is

$$\exp[\sum_{i=1}^n l(X_i,\hat{\theta})]\pi(\hat{\theta})\frac{(2\pi)^{p/2}}{n^{p/2}}|\Delta_u(\hat{\theta})|^{-\frac{1}{2}}[1+O(n^{-1})].$$

Its logarithm simplifies to

$$\sum_{i=1}^n l(X_i,\hat{\theta}) + \log\pi(\hat{\theta}) + \frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\Delta_u(\hat{\theta})| - \frac{p}{2}\log n + \log[1+O(n^{-1})].$$

Ignoring all the terms which stay bounded as $n \to \infty$, we get

$$BIC = \sum_{i=1}^n l(X_i,\hat{\theta}) - \frac{p}{2}\log n.$$

Laplace Approximation and Posterior Normality

Let $X_1,\ldots,X_n|\theta$ be iid with common pdf $f(x|\theta)$. Also, let $\hat{\theta}$ denote the MLE of θ. Write $T_n = \sqrt{n}(\theta-\hat{\theta})$. Let $\pi(\theta)$ denote the prior pdf, $\pi(\theta|\boldsymbol{X}_n)$ the posterior pdf and $\Pi(\cdot|\boldsymbol{X}_n)$ the posterior distribution. Then for $a > 0$, $\Pi(-a < T_n < a|\boldsymbol{X}_n) = \Pi(\hat{\theta}-\frac{a}{\sqrt{n}} < \theta < \hat{\theta}+\frac{a}{\sqrt{n}}|\boldsymbol{X}_n) = J_n/I_n$, where

$$J_n = \int_{\hat{\theta}-\frac{a}{\sqrt{n}}}^{\hat{\theta}+\frac{a}{\sqrt{n}}}\exp[nu(\theta)]\pi(\theta)d\theta, \quad I_n = \int\exp[nu(\theta)]\pi(\theta)d\theta,$$

with $u(\theta) = n^{-1}\sum_{i=1}^{n}\log f(X_i|\theta)$. By the Laplace approximation, $I_n \approx \exp[nu(\hat{\theta})]\pi(\hat{\theta})\frac{\sqrt{2\pi}}{\sqrt{nc}}$, $c = -u''(\hat{\theta})$, the observed Fisher information per unit observation.

Next by the Laplace method,

$$\begin{aligned} J_n &\approx \exp[nu(\hat{\theta})]\int_{\hat{\theta}-\frac{a}{\sqrt{n}}}^{\hat{\theta}+\frac{a}{\sqrt{n}}}[\pi(\hat{\theta}) + (\theta-\hat{\theta})\pi'(\hat{\theta}) + \text{ smaller terms}]\exp[-\frac{nc}{2}(\theta-\hat{\theta})^2]d\theta \\ &\approx \exp[nu(\hat{\theta})]\pi(\hat{\theta})\int_{\hat{\theta}-\frac{a}{\sqrt{n}}}^{\hat{\theta}+\frac{a}{\sqrt{n}}}\exp[-\frac{nc}{2}(\theta-\hat{\theta})^2]d\theta \\ &= \exp[nu(\hat{\theta})]\pi(\hat{\theta})n^{-1/2}\int_{-a}^{a}\exp(-\frac{ct^2}{2})dt. \end{aligned}$$

Thus, for $a > 0$,

$$\begin{aligned} \Pi(-a < T_n < a|\boldsymbol{X}_n) &\approx \frac{\sqrt{c}}{\sqrt{2\pi}}\int_{-a}^{a}\exp(-\frac{ct^2}{2})dt \\ &= P(-a < Z < a),\ Z \sim N(0, c^{-1}). \end{aligned}$$

Tierney-Kadane-Kass Refinements

Suppose we are interested in finding

$$\begin{aligned} E^{\pi}[g(\theta)|\boldsymbol{x}] &= \frac{\int g(\theta)f(\boldsymbol{x}|\theta)\pi\theta)d\theta}{f(\boldsymbol{x}|\theta)\pi(\theta)d\theta} \\ &= \frac{\int g(\theta)\exp[nu(\theta)]d\theta}{\int \exp[nu(\theta)]d\theta}, \end{aligned} \tag{8}$$

where $nu(\theta) = \log f(\boldsymbol{x}|\theta) + \log\pi(\theta)$. A simple first order approximation to this moment is given by $g(\hat{\theta})[1 + O(n^{-1})]$.

Suppose now $g(\theta) > 0$ for all $\theta \in \Theta$. Let $nu^*(\theta) = nu(\theta) + \log g(\theta) = nu(\theta) + G(\theta)$, (say). Now apply Laplace method to both the numerator and the denominator of (8). Let $\hat{\theta}_*$ denote the mode of $u^*(\theta)$,

$$\boldsymbol{\Sigma}^{-1} = -\frac{\partial^2 u}{\partial\theta\partial\theta^T}|_{\theta=\hat{\theta}} \text{ and } \boldsymbol{\Sigma}_*^{-1} = -\frac{\partial^2 u_*}{\partial\theta\partial\theta^T}|_{\theta=\hat{\theta}_*}.$$

Tierney and Kadane (JASA, 1986) obtained the approximation

$$E^{\pi}[g(\theta)|\boldsymbol{x}] = \frac{|\boldsymbol{\Sigma}_*|^{1/2}\exp[nu_*(\hat{\theta}_*)]}{|\boldsymbol{\Sigma}|^{1/2}\exp[nu(\hat{\theta})]}[1 + O(n^{-2})]. \tag{9}$$

We will give an informal proof of (9) when θ is a real-valued parameter. To this end, let $u_k \equiv u_k(\hat{\theta})$, the kth derivative of $u(\theta)$ evaluated at $\hat{\theta}$. Similarly, $u_{*k} \equiv u_{*k}(\hat{\theta}_*)$, the kth derivative of $u_*(\theta)$ evaluated at $\hat{\theta}_*$. Also, write

$$\sigma^2 = -\{u_2\}^{-1}, \text{ and } \sigma_*^2 = -\{u_{*2}\}^{-1}.$$

Under the usual regularity conditions, $\sigma, \sigma_*, u_k, u_{*k}$ are all $O(1)$. First get

$$\begin{aligned}\int \exp[nu(\theta)]d\theta &= \int \exp[nu(\hat{\theta}) - \frac{n}{2\sigma^2}(\theta-\hat{\theta})^2 + R_n(\theta)]d\theta \\ &= \exp[nu(\hat{\theta})]\sqrt{2\pi}\frac{\sigma}{\sqrt{n}}\int \exp[R_n(\theta)]N(\theta|\hat{\theta}, \frac{\sigma^2}{n})d\theta, \qquad (10)\end{aligned}$$

where

$$\begin{aligned}R_n(\theta) &= nu(\theta) - nu(\hat{\theta}) + \frac{n}{2\sigma^2}(\theta-\hat{\theta})^2 \\ &= \frac{n}{6}(\theta-\hat{\theta})^3u_3 + \frac{n}{24}(\theta-\hat{\theta})^4u_4 + \frac{n}{120}(\theta-\hat{\theta})^5u_5 + \frac{n}{720}(\theta-\hat{\theta})^6u_6 + \ldots \qquad (11)\end{aligned}$$

By Taylor expansion,

$$\begin{aligned}\exp[R_n(\theta)] &= 1 + \{\frac{n}{6}(\theta-\hat{\theta})^3u_3 + \frac{n}{24}(\theta-\hat{\theta})^4u_4 + \frac{n}{120}(\theta-\hat{\theta})^5u_5\} \\ &\quad + \frac{1}{2}\{\frac{n}{6}(\theta-\hat{\theta})^3u_3 + \frac{n}{24}(\theta-\hat{\theta})^4u_4\}^2 + \frac{1}{6}\{\frac{n}{6}(\theta-\hat{\theta})^3u_3\}^3 + \ldots \\ &= 1 + \frac{n}{6}(\theta-\hat{\theta})^3u_3 + [\frac{n}{24}(\theta-\hat{\theta})^4u_4 + \frac{n^2(\theta-\hat{\theta})^6u_3^2}{72}] \\ &\quad + [\frac{n(\theta-\hat{\theta})^5u_5}{120} + \frac{n^2(\theta-\hat{\theta})^7u_3u_4}{144} + \frac{n^3(\theta-\hat{\theta})^9u_3^3}{1296}] + O(n^{-2}). \qquad (12)\end{aligned}$$

On integration,

$$\begin{aligned}\int_R \exp[R_n(\theta)]N(\theta|\hat{\theta}, \frac{\sigma^2}{n})d\theta &= 1 + \frac{u_4}{24}\int_R n(\theta-\hat{\theta})^4 N(\theta|\hat{\theta}, \frac{\sigma^2}{n})d\theta \\ &\quad + \frac{u_3^2}{72}\int_R n^2(\theta-\hat{\theta})^6 N(\theta|\hat{\theta}, \frac{\sigma^2}{n})d\theta + O(n^{-2}) \\ &= 1 + \frac{nu_4}{24}\{3(\frac{\sigma^2}{n})^2\} + \frac{n^2u_3^2}{72}\{15(\frac{\sigma^2}{n})^3\} + O(n^{-2}) \\ &= 1 + \frac{a}{n} + O(n^{-2}), \qquad (13)\end{aligned}$$

where $a = \frac{1}{8}\sigma^4u_4 + \frac{5}{24}\sigma^6u_3^2$. Hence,

$$\int \exp[nu(\theta)]d\theta = \exp[nu(\hat{\theta})]\sqrt{2\pi}\frac{\sigma}{\sqrt{n}}[1 + \frac{a}{n} + O(n^{-2})]. \qquad (14)$$

Similarly,

$$\int \exp[nu_*(\theta)]d\theta = \exp[nu_*(\hat{\theta}_*)]\sqrt{2\pi}\frac{\sigma_*}{\sqrt{n}}[1 + \frac{a_*}{n} + O(n^{-2})], \qquad (15)$$

where $a_* = \frac{1}{8}\sigma_*^4u_{*4} + \frac{5}{24}\sigma_*^6u_{*3}^2$. Hence,

$$\begin{aligned}E^{\pi}[g(\theta)|\boldsymbol{x}] &= \frac{\sigma_*}{\sigma}\exp[nu_*(\hat{\theta}_*) - nu(\hat{\theta})]\frac{1 + \frac{a_*}{n} + O(n^{-2})}{1 + \frac{a}{n} + O(n^{-2})} \\ &= \frac{\sigma_*}{\sigma}\exp[nu_*(\hat{\theta}_*) - nu(\hat{\theta})][1 + \frac{a_* - a}{n} + O(n^{-2})]. \qquad (16)\end{aligned}$$

Next observe that

$$\begin{aligned} 0 = u_{*1}(\hat{\theta}_*) &= u_1(\hat{\theta}_*) + n^{-1}G'(\hat{\theta}_*) \\ &\approx u_1(\hat{\theta}) + (\hat{\theta}_* - \hat{\theta})u_2(\hat{\theta}) + n^{-1}G'(\hat{\theta}) + n^{-1}(\hat{\theta}_* - \hat{\theta})G''(\hat{\theta}) \\ &= (\hat{\theta}_* - \hat{\theta})[u_2(\hat{\theta}) + n^{-1}G''(\hat{\theta})] + n^{-1}G'(\hat{\theta}), \end{aligned}$$

implying $\hat{\theta}_* - \hat{\theta} \doteq -\{n^{-1}G'(\hat{\theta})\}/[u_2(\hat{\theta}) + n^{-1}G''(\hat{\theta})] = O(n^{-1})$. Hence, since $u_{*k}(\hat{\theta}) = u_k(\hat{\theta}) + n^{-1}G_k(\hat{\theta})$, $u_{*k}(\hat{\theta}_*) - u_k(\hat{\theta}) = O(n^{-1})$. So, $a_* - a = O(n^{-1})$. This leads to

$$E^\pi[g(\theta)|\boldsymbol{x}] = \frac{\sigma_*}{\sigma}\exp[nu_*(\hat{\theta}_*) - nu(\hat{\theta})][1 + O(n^{-2})]. \tag{17}$$

Asymptotic Expansion of the Posterior Distribution

Let $F_n(u) = P^\pi[\sqrt{n}\hat{I}_n^{1/2}(\theta - \hat{\theta}_n) \le u|\boldsymbol{X}_n]$ be the posterior distribution function of $\sqrt{n}\hat{I}_n^{1/2}(\theta - \hat{\theta}_n)$ given $\boldsymbol{X}_n$. We showed earlier that under a prior π which is continuous and positive at θ_0,

$$\lim_{n\to\infty}\sup_u |F_n(u) - \Phi(u)| = 0 \text{ a.s. } P_{\theta_0}$$

when θ_0 is the true value of the parameter, $\Phi(u)$ being the standard normal cdf.

Johnson (1970) proved the following result refining the original results of Lindley:

$$\sup_u |F_n(u) - \Phi(u) - \phi(u)\sum_{j=1}^{k} n^{-j/2}\psi_j(u, \boldsymbol{X}_n)| \le M_k n^{-\frac{1}{2}(k+1)} \text{ a.s. } P_{\theta_0}$$

for some $M_k > 0$ depending on k, where $\phi(u)$ is the standard normal density and $\psi_j(u, \boldsymbol{X}_n)$ is a jth degree polynomial in u with coefficients bounded in $\boldsymbol{X}_n$. Ghosh, Sinha and Joshi (1982) proved a stronger version of the result.

We now present an informal argument to obtain the expansion for $k = 2$ without the formal rigor of Johnson (1970) or Ghosh et al. (1982).

Let $t = \sqrt{n}(\theta - \hat{\theta}_n)$ and $a_i = \frac{1}{n}\frac{d^iL_n(\theta)}{d\theta^i}|_{\theta=\hat{\theta}_n}, i \ge 1$ so that $a_2 = -\hat{I}_n$. Then by Taylor expansion,

$$\pi(\theta) = \pi(\hat{\theta}_n + t/\sqrt{n}) = \pi(\hat{\theta}_n)[1 + \frac{t}{\sqrt{n}}\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)} + \frac{t^2}{2n}\frac{\pi''(\hat{\theta}_n)}{\pi(\hat{\theta}_n)}] + o(n^{-1})$$

and

$$L_n(\hat{\theta}_n + t/\sqrt{n}) - L_n(\hat{\theta}_n) = \frac{1}{2}t^2a_2 + \frac{t^3}{6\sqrt{n}}a_3 + \frac{t^4}{24n}a_4 + o(n^{-1}).$$

Hence,

$$\begin{aligned} &\pi(\hat{\theta}_n + t/\sqrt{n})\exp[L_n(\hat{\theta}_n + t/\sqrt{n}) - L_n(\hat{\theta}_n)] \\ &= \pi(\hat{\theta}_n)\exp(\frac{a_2t^2}{2})(1 + \frac{\alpha_1}{\sqrt{n}} + \frac{\alpha_2}{n}) + o(n^{-1}), \end{aligned} \tag{18}$$

where

$$\alpha_1 \equiv \alpha_1(t;\boldsymbol{X}_n) = \frac{t^3}{6}a_3 + t\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)},$$

$$\alpha_2 \equiv \alpha_2(t;\boldsymbol{X}_n) = \frac{t^4}{24}a_4 + \frac{t^6}{72}a_3^2 + \frac{t^2}{2}\frac{\pi''(\hat{\theta}_n)}{\pi(\hat{\theta}_n)} + \frac{t^4}{6}a_3\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)}.$$

Then

$$\begin{aligned} C_n &= \int \pi(\hat{\theta}_n + t/\sqrt{n})\exp[L_n(\hat{\theta}_n + t/\sqrt{n}) - L_n(\hat{\theta}_n)]dt \\ &= \pi(\hat{\theta}_n)\sqrt{\frac{2\pi}{-a_2}}[1 + \frac{a_4}{8a_2^2} - \frac{5}{24}\frac{a_3^2}{a_2^3} - \frac{1}{2a_2}\frac{\pi''(\hat{\theta}_n)}{\pi(\hat{\theta}_n)} + \frac{a_3}{2a_2^2}\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)}] + o(n^{-1}). \end{aligned} \tag{19}$$

Hence the posterior π_n^* of T_n is

$$\begin{aligned} \pi_n^*(t|\boldsymbol{X}_n) &= C_n^{-1}\pi(\hat{\theta}_n + t/\sqrt{n})\exp[L_n(\hat{\theta}_n + t/\sqrt{n}) - L_n(\hat{\theta}_n)] \\ &= (2\pi)^{-\frac{1}{2}}\hat{I}_n^{1/2}\exp(-\frac{\hat{I}_n t^2}{2})[1 + \frac{\gamma_1}{\sqrt{n}} + \frac{\gamma_2}{n}] + o(n^{-1}), \end{aligned} \tag{20}$$

where

$$\gamma_1 \equiv \gamma_1(t;\boldsymbol{X}_n) = \alpha_1(t;\boldsymbol{X}_n) = \frac{t^3}{6}a_3 + t\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)},$$

and

$$\gamma_2 \equiv \gamma_2(t;\boldsymbol{X}_n) = \alpha_2(t;\boldsymbol{X}_n) - \frac{a_4}{8a_2^2} + \frac{5}{24}\frac{a_3^2}{a_2^3} + \frac{1}{2a_2}\frac{\pi''(\hat{\theta}_n)}{\pi(\hat{\theta}_n)} - \frac{a_3}{2a_2^2}\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)}.$$

Let $S_n = \hat{I}_n^{1/2}T_n = \sqrt{n}\hat{I}_n^{1/2}(\theta - \hat{\theta}_n)$. Then the posterior density of S_n is given by

$$\begin{aligned} \pi_n(s|\boldsymbol{X}_n) &= \phi(s)[1 + \frac{1}{\sqrt{n}}\{\frac{a_3 s^3}{6\hat{I}_n^{3/2}} + \frac{s}{\hat{I}_n^{1/2}}\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)}\} \\ &\quad + \frac{1}{n}\{\frac{a_4 s^4}{24\hat{I}_n^2} - \frac{a_3^2 s^6}{72\hat{I}_n^3} + \frac{s^2}{2\hat{I}_n}\frac{\pi''(\hat{\theta}_n)}{\pi(\hat{\theta}_n)} + \frac{a_3 s^4}{6\hat{I}_n^2}\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)} \\ &\quad - \frac{a_4}{8\hat{I}_n^2} + \frac{5a_3^2}{24\hat{I}_n^3} - \frac{1}{2\hat{I}_n}\frac{\pi''(\hat{\theta}_n)}{\pi(\hat{\theta}_n)} - \frac{a_3}{2\hat{I}_n^2}\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)}\}] + o(n^{-1}). \end{aligned} \tag{21}$$

The expansion given in (21) will be useful later in deriving probability matching priors.

9 Optimality

9.1 Asymptotic Relative Efficiency

Already done from Sec 1.15.4 of Serfling

9.2 Asymptotic relative efficiency of estimators

(Casella and Berger 476-477)

Definition 4. *A sequence of estimators $\tilde{\theta_n}$ is* **consistent** *for θ if $\tilde{\theta_n} \xrightarrow{P} \theta$.*

Definition 5. *A sequence of estimators $\tilde{\theta_n}$ is* $\sqrt{n}$**-consistent** *for θ if $\sqrt{n}(\underset{\sim}{\tilde{\theta_n}} - \theta)$ is bounded in probability. This will be satisfied in particular by any estimator $\tilde{\theta_n}$ for which $\sqrt{n}(\tilde{\theta_n} - \theta)$ tends in law to a non-degenerate limit distribution.*

In Chapters 1 and 3 we have considered various cases where the distribution of estimators converged at rate $\sqrt{n}$ to the normal distribution. If there are multiple estimators of the same parameter with this property, then all of them are $\sqrt{n}$ consistent. We can use the asymptotic variance as a means of comparing such estimators. This is the idea of asymptotic relative efficiency.

Definition 6. *If two estimators W_n and V_n satisfy*

$$\sqrt{n}[V_n - \theta] \Rightarrow \mathcal{N}(0, \sigma_V^2)$$
$$\sqrt{n}[W_n - \theta] \Rightarrow \mathcal{N}(0, \sigma_W^2)$$

The **asymptotic relative efficiency(ARE)** *of V_n with respect to W_n is*

$$\text{ARE}(V_n, W_n) = \frac{\sigma_W^2}{\sigma_V^2} \tag{106}$$

Example 1(ARE of Poisson Estimators) Suppose $X_1, \cdots, X_n$ are iid Poisson(λ), and we are interested in estimating $\tau = P_\lambda(X_1 = 0) = \exp(-\lambda)$. For example number of customers who come into a bank in a given time period is modeled as a Poisson random variable and we are interested in the probability that no one will enter the bank in one time period. A natural (but somewhat naive) estimator comes from defining $Y_i = I(X_i = 0)$. The Y_is are iid Bernoulli($\exp(-\lambda)$) and hence it follows that

$$\sqrt{n}(\bar{Y}_n - \exp(-\lambda)) \Rightarrow \mathcal{N}(0, \exp(-\lambda)(1 - \exp(-\lambda)))$$

Additionally,the MLE of $\exp(-\lambda)$ is $\hat{\tau} = \exp(-\hat{\lambda})$ where $\hat{\lambda} = \bar{X}_n$ is the MLE of λ. Using the Delta method, we have

$$\sqrt{n}(\hat{\tau} - \tau) \Rightarrow \mathcal{N}(0, \lambda \exp(-2\lambda))$$

The ARE of $\bar{Y}_n$ wrt the MLE is

$$\mathrm{ARE}(\bar{Y}, \exp(-\bar{X})) = \frac{\lambda \exp(-2\lambda)}{\exp(-\lambda)(1-\exp(-\lambda))} = \frac{\lambda \exp(-\lambda)}{(1-\exp(-\lambda))}$$

Examination of this function shows that it is strictly decreasing with a maximum of 1 at $\lambda = 0$ and tailing off rapidly (< 0.1 when $\lambda = 4$) to 0 as $\lambda \to \infty$. So in this case the MLE is better in terms of ARE.

Example 2 (Mean vs Median of a symmetric distribution) Consider a distribution function F with density function f symmetric about an unknown point θ to be estimated. For $X_1, \cdots, X_n$ a sample from F, put $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\mathrm{Med}_n = \mathrm{median}\{X_1, \cdots, X_n\}$. Each of $\bar{X}_n$ and Med_n is a consistent estimator of θ.

For $\bar{X}_n$, the classical central limit theorem tells us: if F has finite variance σ_F^2 , then the sampling distribution of $\bar{X}_n$ is approximately $\mathcal{N}(\theta, \sigma_F^2/n)$. For Med_n, from Section 4 we get that if the density f is continuous and positive at θ, then the sampling distribution of Med_n is approximately $\mathcal{N}(\theta, (4[f(\theta)]^2 n)^{-1})$. The asymptotic relative efficiency (ARE) of Med to $\bar{X}$ is $\mathrm{ARE}(\mathrm{Med}, \bar{X}, F) = 4[f(\theta)]^2\sigma_F^2$.

Example 2' (Mean vs Median: Different distributions) With $F = \mathcal{N}(\theta, \sigma_F^2)$, it is seen that $\mathrm{ARE}(\mathrm{Med}, \bar{X}, \mathcal{N}(\theta, \sigma_F^2)) = 2/\pi = 0.64$. For sampling from a double exponential (or Laplace) distribution with density $f(x) = \frac{\theta}{2}\exp(-\theta \mid x - \theta \mid), -\infty < x < \infty$ (and thus variance $2/\theta^2$), we get $\mathrm{ARE}(\mathrm{Med}, \bar{X}, \mathrm{Laplace}) = 2$. Thus depending on the distribution, median can be more or less efficient (asymptotically) than the mean. A very interesting solution to this dilemma is given by an estimator that has excellent overall performance, the so-called Hodges-Lehmann location estimator (Hodges and Lehmann(1963) Annals of Mathematical Statistics)

$$HL_n = \mathrm{Median}(\frac{X_i + X_j}{2}) \tag{107}$$

the median of all pairwise averages of the sample observations. We have that HL_n is asymptotically $\mathcal{N}(\theta, (12[\int f^2(x)dx]^2 n)^{-1})$, which yields that $\mathrm{ARE}(HL, \bar{X}, \mathcal{N}(\theta, \sigma_F^2)) = 3/\pi = 0.955$ and $\mathrm{ARE}(HL, \bar{X}, \mathrm{Laplace}) = 1.5$. Also, for the Logistic distribution with density

$$f(x) = \frac{1}{\sigma}\frac{\exp\{(x-\theta)/\sigma\}}{(1+\exp\{(x-\theta)/\sigma\})^2}, -\infty < x < \infty,$$

for which HL_n is the MLE of θ and thus optimal, we have $\mathrm{ARE}(HL, \bar{X}, \mathrm{Logistic}) = \pi^2/9 = 1.097$. Further, for $\mathcal{F}$ the class of all distributions symmetric about θ and having finite variance, we have $\inf_{\mathcal{F}}\mathrm{ARE}(HL, \bar{X}, F) = 108/125 = 0.864$ (see Lehmann). The estimator HL_n is highly competitive with X at Normal distributions, can be infinitely more efficient at some other symmetric distributions F, and is never much less efficient at any distribution F in $\mathcal{F}$.

9.3 Asymptotic Bias and Efficiency

(Casella and Berger 470-471, Lehmann and Casella sec 6.1, 6.2)

There are two ways in which we can look at the bias as sample size goes to infinity. We can look at the finite sample bias $\mathrm{Bias}(T_n)$ and take the limit as $n \to \infty$. This is called the limiting bias. We can also look for a suitably scaled version of the estimator converges in distribution to a non-degenerate random variable and look at the bias of that limiting distribution. This is the asymptotic bias. Here are the precise definitions:

Definition 7. *An estimator T_n of $\tau(\theta)$ is unbiased in the limit, if* $\lim_{n\to\infty} \mathrm{E}(T_n) = \tau(\theta)$.

Definition 8. *For an estimator T_n, suppose that $k_n(T_n - \tau(\theta)) \Rightarrow \mathcal{H}$. The estimator T_n is asymptotically unbiased if the expectation of $\mathcal{H}$ is zero.*

Example 1 (Asymptotically biased estimator) Let $X_1, \cdots, X_n$ are iid $U(0, \theta)$.

$$\text{The MLE of} \quad \theta \quad \text{is} \quad X_{(n)} \tag{108}$$

$$P(X_{(n)} \leq a) = (a/\theta)^n \quad \text{and} \quad \mathrm{E}(X_{(n)}) = \theta \tag{109}$$

Hence $P(n(\theta - X_{(n)}) \leq a) = P(X_{(n)} \geq \theta - a/n) = 1 - (1 - a/n\theta)^n \to 1 - e^{-a/\theta}$. Thus $n(\theta - X_{(n)}) \Rightarrow \mathrm{Exp}(\frac{1}{\theta})$. The expectation of the limiting random variable is not zero. So $X_{(n)}$ is not asymptotically unbiased. From (**??**) $X_{(n)}$ is unbiased in the limit.

Similar concepts exist for efficiency, which concerned with the asymptotic variance of the estimator.

Definition 9. *For an estimator T_n, if* $\lim_{n\to\infty} k_n \mathrm{Var}(T_n) = \tau^2 < \infty$, *where k_n is a sequence of constants, then τ^2 is called the limiting variance.*

Definition 10. *For an estimator T_n, suppose that $k_n(T_n - \tau(\theta)) \Rightarrow \mathcal{N}(0, \sigma^2)$. The parameter σ^2 is called the asymptotic variance of T_n.*

In most cases these two are the same. But in complicated cases, this may not hold. It is always the case that the asymptotic variance is smaller than the limiting variance (Lehmann and Casella Sec 6.1).

Example 2 Let us consider the mean $\bar{X}_n$ of n iid normal observations with mean μ and variance σ^2. Suppose we are interested in estimating $\frac{1}{\mu}$ and we use the estimator $T_n = \frac{1}{\bar{X}_n}$. For each finite n the distribution of $\sqrt{n}\bar{X}_n$ is $\mathcal{N}(0, \sigma^2)$.

$$\mathrm{Var}(\sqrt{n}T_n) = \infty, \quad \text{by direct integral of} \quad \frac{1}{x^2} \quad \text{with respect to the normal pdf.} \tag{110}$$

So, the limiting variance of T_n is infinity. On the other hand, by Delta method,

$$\sqrt{n}(T_n - \frac{1}{\mu}) \Rightarrow \mathcal{N}(0, \frac{\sigma^2}{\mu^4})$$

So the asymptotic variance of T_n is $\frac{\sigma^2}{\mu^4}$.

In the spirit of the Cramer Rao lower bound, there is an optimal asymptotic variance.

Definition 11. *A sequence of estimators W_n is asymptotically efficient for a parameter $\tau\theta$ if $\sqrt{n}(W_n - \tau(\theta) \Rightarrow \mathcal{N}(0, \nu(\theta)$ and*

$$\nu(\theta) = \frac{(\tau'(\theta)^2)}{\mathrm{E}_\theta((\frac{\partial}{\partial\theta}\log f(X \mid \theta))^2)} = \frac{(\tau'(\theta)^2)}{I(\theta)}, \tag{111}$$

that is the asymptotic variance of W_n achieves the Cramer-Rao lower bound.

For a long time it was believed that if

$$\sqrt{n}(W_n - \tau(\theta) \Rightarrow \mathcal{N}(0, \nu(\theta), \tag{112}$$

then

$$\nu(\theta) \geq \frac{(\tau'(\theta)^2)}{I(\theta)} \tag{113}$$

under regularity conditions on the densities. This belief was shattered by the example (due to hodges; see LaCam 1953) below:

Example 3 (Superefficient Estimator): Let $X_1, \cdots, X_n$ be iid $\mathcal{N}(\theta, 1)$ and the parameter of interest is θ. In this case, $h(\theta) = \theta$, and

$$\begin{aligned} I(\theta) &= \mathrm{E}_\theta((\frac{\partial}{\partial\theta}\log f(X \mid \theta))^2) \\ &= \mathrm{E}_\theta((\frac{\partial}{\partial\theta}\frac{1}{2}(X-\theta)^2)^2) \\ &= \mathrm{E}_\theta(X-\theta)^2 \\ &= 1 \end{aligned}$$

Thus equation(**??**) reduces $\nu(\theta) \geq 1$. Now consider the sequence of estimators

$$T_n = \begin{cases} \bar{X} & \text{if } \mid \bar{X} \mid \geq 1/n^{1/4} \\ a\bar{X} & \text{if } \mid \bar{X} \mid < 1/n^{1/4} \end{cases}$$

$$\text{Then,} \quad \sqrt{n}(T_n - \theta) \Rightarrow \mathcal{N}(0, \nu(\theta)), \tag{114}$$

$$\text{where} \quad \nu(\theta) = 1 \quad \text{when} \quad \theta \neq 0 \quad \text{and} \quad \nu(\theta) = a^2 \quad \text{when} \quad \theta = 0. \tag{115}$$

If $a < 1$, inequality (**??**) is violated at $\theta = 0$.

This phenomenon is quite common and is called superefficiency. There will typically exist estimators satisfying (**??**) but with $\nu(\theta)$ violating (**??**) at least for some values of θ. However, it was shown by LaCam(1953) that for any sequence of estimators satisfying (**??**), the set S of points of super-efficiency has Lebesgue measure zero.

9.4 ARE of tests

Under the alternative $\theta = \theta_1$, where θ_1 is fixed, the power of the LR test converges to 1 (homework: show this). Such tests are said to be consistent. There are three approaches to studying the asymptotic power of tests in order to obtain nontrivial asymptotic result:

1. let $\alpha \to 0$ (Bayes)

2. look at rates at which power → 1 (Bahadur)

3. let the alternative shrink toward θ_0 (Pitman)

We will focus on the third. Let us consider power against alternatives of the form $\theta_{1n} = \theta_0 + g/\sqrt{n}$ for some nonzero $k \times 1$ vector g. Proceeding as above, we now find

$$\sqrt{n}(\hat{\theta} - \theta_{1n}) \Rightarrow \mathcal{N}(0, I(\theta_0)^{-1}) \tag{116}$$

which implies $\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow \mathcal{N}(g, I(\theta_0)^{-1})$ so that ξ_W converges to a quadratic form in multivariate normals with a nonzero mean. The limiting distribution is $\chi^2_k(\delta)$ where δ is the noncentrality parameter and equals $g'Ig$.(homework)

Exercises

1. Show that posterior consistency implies robustness of Bayesian inference with respect to the prior.

2. Problem 6.6.14 of Lehmann and Casella, pg 510.

3. Problem 6.6.15 of Lehmann and Casella, pg 510.

4. Complete example 1 by proving (**??**) and (**??**)

5. Complete example 2 by proving (**??**)

6. Complete example 3 by proving (**??**) and (**??**)

7. Show that the likelihood ratio test is consistent

8. Consider the simple linear regression model

$$y_i = \beta x_i + \epsilon_i$$

with slope zero. ϵ_i are iid with mean 0 and variance σ^2. Find the asymptotic distribution of $\hat{\beta}$, the least squares estimator of β under suitable assumptions on x_i, namely, $\overline{x_n} \to 0$, max $\frac{x_i}{\sum x_j^2} \to 0$, $\frac{1}{n} \sum x_j^2 \to t < \infty$.