Lecture Notes on Asymptotic Statistics

Changliang Zou

Prologue

Why asymptotic statistics? The use of asymptotic approximation is two-fold. First, they enable us to find approximate tests and confidence regions. Second, approximations can be used theoretically to study the quality (efficiency) of statistical procedures— Van der Vaart

Approximate statistical procedures

To carry out a statistical test, we need to know the critical value of the test statistic. Roughly speaking, this means we must know the distribution of the test statistic under the null hypothesis. Because such distributions are often analytically intractable, only approximations are available in practice.

Consider for instance the classical t-test for location. Given a sample of iid observations $X_1, \ldots, X_n$, we wish to test $H_0 : \mu = \mu_0$. If the observations arise from a normal distribution with mean μ_0, then the distribution of t-test statistic, $\sqrt{n}(\bar{X}_n - \mu_0)/S_n$, is exactly known, say $t(n-1)$. However, we may have doubts regarding the normality. If the number of observations is not too small, this does not matter too much. Then we may act as if $\sqrt{n}(\bar{X}_n - \mu_0)/S_n \sim N(0,1)$. The theoretical justification is the limiting result, as $n \to \infty$,

$$\sup_x \left| P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \leq x \right) - \Phi(x) \right| \to 0,$$

provided that the variables X_i have a finite second moment. Then, a "large-sample" or "asymptotical" level α test is to reject H_0 if $|\sqrt{n}(\bar{X}_n - \mu_0)/S_n| > z_{\alpha/2}$. When the underlying distribution is exponential, the approximation is satisfactory if $n \geq 100$. Thus, one aim of asymptotic statistics is to derive the asymptotical distribution of many types of statistics.

There are similar benefits when obtaining confidence intervals. For instance, consider maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n$ of dimension p based on a sample of size n from a density $f(X;\boldsymbol{\theta})$. A major result in asymptotic statistic is that in many situations $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ is asymptotically normally distributed with zero mean and covariance matrix $\mathbf{I}_{\boldsymbol{\theta}}^{-1}$, where

$$\mathbf{I}_{\boldsymbol{\theta}} = E_{\boldsymbol{\theta}} \left[\left(\frac{\partial \log f(X;\theta)}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \log f(X;\theta)}{\partial \boldsymbol{\theta}} \right)^T \right]$$

is the Fisher information matrix. Thus, acting as if $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \sim N_p(0, \mathbf{I}_{\boldsymbol{\theta}}^{-1})$, we can find

the following ellipsoid

$$\left\{ \boldsymbol{\theta} : (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n)^T \mathbf{I}_{\boldsymbol{\theta}} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n) \leq \frac{\chi^2_{p,\alpha}}{n} \right\}$$

is an approximate $1 - \alpha$ confidence region.

Efficiency of statistical procedures

For a relatively small number of statistical problems, there exists an exact, optimal solution. For example, the Neyman-Pearson lemma to find UMP tests, the Rao-Blackwell theory to find MVUE, and Cramer-Rao Theorem.

However, there are not always exact optimal theory or procedure, then asymptotic optimality theory may help. For instance, to compare two tests, we might compare approximations to their power functions. Consider the foregoing hypothesis problem for location. A well-known nonparametric test statistic is the *sign statistic* $T_n = n^{-1} \sum_{i=1}^{n} I_{X_i > \theta_0}$, where the null hypothesis is $H_0 : \theta = \theta_0$ and θ denotes the median associated the distribution of X. To compare the efficiency of sign and t-test is rather difficult because the exact power functions of two tests are untractable. However, by the definitions and methods introduced later, we can obtain the asymptotic relative efficiency of the sign test versus the t-test is equal to

$$4f^2(0) \int x^2 f(x) dx.$$

To compare estimators, we might compare asymptotic variances rather than exact variances. A major result in this area is that for smooth parametric models maximum likelihood estimators are asymptotically optimal. This roughly means the following. First, MLE are asymptotically consistent; Second, the rate at which MLE converge to the true value is the fastest possible, typically $\sqrt{n}$; Third, the asymptotic variance, attain the C-R bound. Thus, asymptotic justify the use of MLE in certain situations. (Even though in general it does not lead to best estimators for finite sample in many cases, it is always not a worst one and always leads to a reasonable estimator.

Contents

- The basic sample statistics: distribution function, moment, quantiles, and order statistics (3)
- Asymptotic theory in parametric inference: MLE, likelihood ratio test, etc (6)
- U-statistic, M-estimates and R-estimates (6)
- Asymptotic relative efficiency (6)
- Asymptotic theory in nonparametric inference: rank and sign tests (6)
- Goodness of fit (3)
- Nonparametric regression and density estimation (4)
- Advanced topic selected: bootstrap and empirical likelihood (4)

Text books

Billingsley, P. (1995). *Probability and Measure*, 3rd edition, John Wiley, New York.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*, Springer.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*, John Wiley, New York.

Shao, J. (2003). *Mathematical Statistics*, 2nd ed. Springer, New York.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, Cambridge University Press.

Chapter 1

Basic convergence concepts and preliminary theorems

Throughout this course, there will usually be an underlying probability space $(\Omega, \mathcal{F}, P)$, where Ω is a set of points, $\mathcal{F}$ is a σ-field of subsets of Ω, and P is a probability distribution or measure defined on the element of $\mathcal{F}$. A random variable $X(w)$ is a transformation of Ω into the real line $\mathbb{R}$ such that images $X^{-1}(B)$ of Borel sets B are elements of $\mathcal{F}$. A collection of random variables $X_1(w), X_2(w), \ldots$ on a given $(\Omega, \mathcal{F})$ will typically be denoted by $X_1, X_2, \ldots$.

1.1 Modes of convergence of a sequence of random variables

Definition 1.1.1 (convergence in probability) *Let $\{X_n, X\}$ be random variables defined on a common probability space. We say X_n converges to X in probability if, for any $\epsilon > 0$, $P(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$, or equivalently*

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1, \quad \textit{every} \ \ \epsilon > 0.$$

This is usually written as $X_n \xrightarrow{p} X$. Extensions to the vector case: for random p-vectors $\mathbf{X}_1, \mathbf{X}_2 \ldots$ and $\mathbf{X}$, we say $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ if $||\mathbf{X}_n - \mathbf{X}|| \xrightarrow{p} 0$, where $||\mathbf{z}|| = (\sum_{i=1}^{p} z_i^2)^{1/2}$ denotes the Euclidean distance (L_2-norm) for $\mathbf{z} \in \mathbb{R}^p$. It is easily to seen that $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ iff the corresponding component-wise convergence holds.

Example 1.1.1 For iid Bernoulli trials with a success probability $p = 1/2$, let X_n denote the number of times in the first n trials that a success is followed by a failure. Denoting $T_i = I\{i\text{th trial is success and } (i+1)\text{st trial is a failure}\}$, $X_n = \sum_{i=1}^{n-1} T_i$, and therefore $E[X_n] = (n-1)/4$, and $\mathrm{Var}[X_n] = \sum_{i=1}^{n-1} \mathrm{Var}[T_i] + 2\sum_{i=1}^{n-2} \mathrm{Cov}[T_i, T_{i+1}] = 3(n-1)/16 - 2(n-2)/16 = (n+1)/16$. It then follows by an application of Chebyshev's inequality that $X_n/n \xrightarrow{p} 1/4$. $[P(|x - \mu| \geq \epsilon) \leq \sigma^2/\epsilon^2]$

Definition 1.1.2 (bounded in probability) *A sequence of random variables X_n is said to be bounded in probability if, for any $\epsilon > 0$, there exists a constant k such that $P(|X_n| > k) \leq \epsilon$ for all n.*

Any random variable (vector) is bounded in probability. It is convenient to have short expressions for terms that converge or bounded in probability. If $X_n \xrightarrow{p} 0$, then we write $X_n = o_p(1)$, pronounced by "small oh-P-one"; The expression $O_p(1)$ ("big oh-P-one") denotes a sequence that is bounded in probability, say, write $X_n = O_p(1)$. These are so-called stochastic $o(\cdot)$ and $O(\cdot)$. More generally, for a given sequence of random variables R_n,

$$X_n = o_p(R_n) \quad \text{means} \quad X_n = Y_n R_n \text{ and } Y_n \xrightarrow{p} 0;$$
$$X_n = O_p(R_n) \quad \text{means} \quad X_n = Y_n R_n \text{ and } Y_n = O_p(1).$$

This expresses that the sequence X_n converges in probability to zero or is bounded in probability "at the rate R_n". For deterministic sequences X_n and R_n, $O_p(\cdot)$ and $o_p(\cdot)$ reduce to the usual $o(\cdot)$ and $O(\cdot)$ from calculus. Obviously, $X_n = o_p(R_n)$ implies that $X_n = O_p(R_n)$. An expression we will often used is: for some sequence a_n, if $a_n X_n \xrightarrow{p} 0$, then we write $X_n = o_p(a_n^{-1})$; if $a_n X_n = O_p(1)$, then we write $X_n = O_p(a_n^{-1})$.

Definition 1.1.3 (convergence with probability one) *Let $\{X_n, X\}$ be random variables*

defined on a common probability space. We say X_n converges to X with probability 1 (or almost surely, strongly, almost everywhere) if

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

This can be written as $P(\omega : X_n(\omega) \to X(\omega)) = 1$. We denote this mode of convergence as $X_n \overset{wp1}{\to} X$ or $X_n \overset{\text{a.s}}{\to} X$. Extensions to random vector case is straightforward.

Almost sure convergence is a stronger mode of convergence than convergence in probability. In fact, a characterization of $wp1$ is that

$$\lim_{n\to\infty} P(|X_m - X| < \epsilon, \text{all } m \geq n) = 1, \quad \text{every} \;\; \epsilon > 0. \tag{1.1}$$

It is clear from this equivalent condition that $wp1$ is stronger than convergence in probability. Its proof can be found on page 7 in Serfling (1980).

Example 1.1.2 Suppose $X_1, X_2, \ldots$ is an infinite sequence of iid $U[0,1]$ random variables, and let $X_{(n)} = \max\{X_1, \ldots, X_n\}$. See $X_{(n)} \overset{wp1}{\to} 1$. Note that

$$\begin{aligned} P(|X_{(n)} - 1| \leq \epsilon, \; \forall n \geq m) &= P(X_{(n)} \geq 1 - \epsilon, \; \forall n \geq m) \\ &= P(X_{(m)} \geq 1 - \epsilon) = 1 - (1-\epsilon)^m \to 1, \quad \text{as } m \to \infty. \end{aligned}$$

Definition 1.1.4 (convergence in rth mean) *Let $\{X_n, X\}$ be random variables defined on a common probability space. For $r > 0$, we say X_n converges to X in rth mean if*

$$\lim_{n\to\infty} E|X_n - X|^r = 0.$$

This is written $X_n \overset{rth}{\to} X$. It is easily shown that

$$X_n \overset{rth}{\to} X \Rightarrow X_n \overset{sth}{\to} X, \; 0 < s < r,$$

by Jensen's inequality (If $g(\cdot)$ is a convex function on $\mathbb{R}$, and X and $g(X)$ are integrable r.v.'s, then $g(E[X]) \leq E[g(X)]$).

Definition 1.1.5 (convergence in distribution) *Let $\{X_n, X\}$ be random variables. Consider their distribution functions $F_{X_n}(\cdot)$ and $F_X(\cdot)$. We say that X_n converges in distribution (in law) to X if $\lim_{n\to\infty} F_{X_n}(t) = F_X(t)$ at every point that is a continuity point of F_X.*

This is written as $X_n \xrightarrow{d} X$ or $F_{X_n} \Rightarrow F_X$.

Example 1.1.3 Consider $X_n \sim \text{Uniform}\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1\}$. Then, it can be shown easily that the sequence X_n converges in law to $U[0,1]$. Actually, consider any $t \in [\frac{i}{n}, \frac{i+1}{n})$, the difference between $F_{X_n}(t) = \frac{i}{n}$ and $F_X(t) = t$ can be arbitrarily small if n is sufficiently large ($|\frac{i}{n} - t| < n^{-1}$). The result follows from the definition of $\xrightarrow{d}$.

Example 1.1.4 Let $\{X_n\}_{n=1}^{\infty}$ is a sequence of random variables where $X_n \sim N(0, 1 + n^{-1})$. Taking the limit of the distribution function of X_n as $n \to \infty$ yields $\lim_n F_{X_n}(x) = \Phi(x)$ for all $x \in \mathbb{R}$. Thus, $X_n \xrightarrow{d} N(0,1)$.

According to the assertion below the definition of $\xrightarrow{p}$, we know that $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ is equivalent to convergence of every one of the sequences of components. The analogous statement for convergence in distribution is false: Convergence in distribution of the sequence $\mathbf{X}_n$ is stronger than convergence of every one of the sequences of components X_{ni}. The point is that the distribution of the components X_{ni} separately does not determine their distribution (they might be independent or dependent in many ways). We speak of *joint convergence* in law versus *marginal convergence*.

Example 1.1.5 *If $X \sim U[0,1]$ and $X_n = X$ for all n, and $Y_n = X$ for n odd and $Y_n = 1 - X$ for n even, then $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} U[0,1]$, yet (X_n, Y_n) does not converge in law.*

Suppose $\{X_n, X\}$ are integer-valued random variables. It is not hard to show that

$$X_n \xrightarrow{d} X \Leftrightarrow P(X_n = k) \to P(X = k)$$

for every integer k. This is a useful characterization of convergence in law for integer-valued random variables.

1.2 Fundamental results and theorems on convergence

1.2.1 Relationship

The results describes the relationship among four convergence modes are summarized as follows.

Theorem 1.2.1 *Let $\{X_n, X\}$ be random variables (vectors).*

(i) If $X_n \overset{wp1}{\to} X$, then $X_n \overset{p}{\to} X$.

(ii) If $X_n \overset{rth}{\to} X$ for a $r > 0$, then $X_n \overset{p}{\to} X$.

(iii) If $X_n \overset{p}{\to} X$, then $X_n \overset{d}{\to} X$.

(iv) If, for every $\epsilon > 0$, $\sum_{n=1}^{\infty} P(|X_n - X| > \epsilon) < \infty$, then $X_n \overset{wp1}{\to} X$.

Proof. (i) is an obvious consequence of the equivalent characterization (1.1); (ii) for any $\epsilon > 0$,

$$E|X_n - X|^r \geq E[|X_n - X|^r I(|X_n - X| > \epsilon)] \geq \epsilon^r P(|X_n - X| > \epsilon)$$

and thus

$$P(|X_n - X| > \epsilon) \leq \epsilon^{-r} E|X_n - X|^r \to 0, \text{ as } n \to \infty.$$

(iii) This is a direct application of Slutsky Theorem; (iv) Let $\epsilon > 0$ be given. We have

$$P(|X_m - X| \geq \epsilon, \text{ for some } m \geq n) = P\left(\bigcup_{m=n}^{\infty} \{|X_m - X| \geq \epsilon\}\right) \leq \sum_{m=n}^{\infty} P(|X_m - X| \geq \epsilon).$$

The last term in the equation above is the tail of a convergent series and hence goes to zero as $n \to \infty$. □

Example 1.2.1 Consider iid $N(0,1)$ random variables $X_1, X_2, \ldots$, and suppose $\bar{X}_n$ is the mean of the first n observations. For an $\epsilon > 0$, consider $\sum_{n=1}^{\infty} P(|\bar{X}_n| > \epsilon)$. By Markov's inequality, $P(|\bar{X}_n| > \epsilon) \leq \frac{E[\bar{X}_n^4]}{\epsilon^4} = \frac{3}{\epsilon^4 n^2}$. Since $\sum_{n=1}^{\infty} n^{-2} < \infty$, from Theorem 1.2.1-(iv) it follows that $X_n \overset{wp1}{\to} 0$.

1.2.2 Transformation

It turns out that continuous transformations preserve many types of convergence, and this fact is useful in many applications. We record it next. Its proof can be found on page 24 in Serfling (1980).

Theorem 1.2.2 (Continuous Mapping Theorem) *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots$ and $\mathbf{X}$ be random p-vectors defined on a probability space, and let $g(\cdot)$ be a vector-valued (including real-valued) continuous function defined on $\mathbb{R}^p$. If $\mathbf{X}_n$ converges to $\mathbf{X}$ in probability, almost surely, or in law, then $g(\mathbf{X}_n)$ converges to $\mathbf{X}$ in probability, almost surely, or in law, respectively.*

Example 1.2.2 (i) If $X_n \xrightarrow{d} N(0,1)$, then χ^2_1; (ii) If $(X_n, Y_n) \xrightarrow{d} N_2(\mathbf{0}, \mathbf{I}_2)$, then

$$\max\{X_n, Y_n\} \xrightarrow{d} \max\{X, Y\},$$

which has the CDF $[\Phi(x)]^2$.

The most commonly considered functions of vectors converging in some stochastic sense are linear and quadratic forms, which is summarized in the following result.

Corollary 1.2.1 *Suppose that the p-vector $\mathbf{X}_n$ converge to the p-vector $\mathbf{X}$ in probability, almost surely, or in law. Let $\mathbf{A}_{q\times p}$ and $\mathbf{B}_{p\times p}$ be matrices. Then $\mathbf{A}\mathbf{X}_n \to \mathbf{A}\mathbf{X}$ and $\mathbf{X}_n^T\mathbf{B}\mathbf{X}_n \to \mathbf{X}^T\mathbf{B}\mathbf{X}$ in the given mode of convergence.*

Proof. The vector-valued function

$$\mathbf{A}\mathbf{x} = \left(\sum_{i=1}^{p} a_{1i}x_i, \ldots, \sum_{i=1}^{p} a_{qi}x_i\right)^T$$

and the real-valued function

$$\mathbf{x}^T\mathbf{B}\mathbf{x} = \sum_{i=1}^{p}\sum_{j=1}^{p} b_{ij}x_i x_j$$

are continuous function of $\mathbf{x} = (x_1, \ldots, x_p)^T$. □

Example 1.2.3 (i) If $\mathbf{X}_n \xrightarrow{d} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{C}\mathbf{X}_n \xrightarrow{d} N(\mathbf{C}\boldsymbol{\mu}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^T)$ where $\mathbf{C}_{q\times p}$ is a matrix; Also, $(\mathbf{X}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{X}_n - \boldsymbol{\mu}) \xrightarrow{d} \chi^2_p$; (ii) (*Sums and products of random variables converging wp1 or in probability*) If $X_n \xrightarrow{wp1} X$ and $Y_n \xrightarrow{wp1} Y$, then $X_n + Y_n \xrightarrow{wp1} X + Y$ and $X_nY_n \xrightarrow{wp1} XY$. Replacing the wp1 with in probability, the foregoing arguments also hold.

Remark 1.2.1 The condition that $g(\cdot)$ is continuous function in Theorem 1.2.2 can be further relaxed to that $g(\cdot)$ is continuous a.s., i.e., $P(\mathbf{X} \in C(g)) = 1$ where $C(g) = \{\mathbf{x} : g \text{ is continuous at } \mathbf{x}\}$ is called the continuity set of g.

Example 1.2.4 (i) If $X_n \xrightarrow{d} X \sim N(0, 1)$, then $1/X_n \xrightarrow{d} Z$, where Z has the distribution of $1/X$, even though the function $g(x) = 1/x$ is not continuous at 0. This is due to $P(X = 0) = 0$. However, if $X_n = 1/n$ (degenerate distribution) and

$$g(x) = \begin{cases} 1, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

then $X_n \xrightarrow{d} 0$ but $g(X_n) \xrightarrow{d} 1 \neq g(0)$; (ii)If $(X_n, Y_n) \xrightarrow{d} N_2(\mathbf{0}, \mathbf{I}_2)$ then $X_n/Y_n \xrightarrow{d}$ Cauchy.

Example 1.2.5 Let $\{X\}_{n=1}^{\infty}$ be a sequence of independent random variables where X_n has a Poi(θ) distribution. Let $\bar{X}_n$ be the sample mean computed on $X_1, \ldots, X_n$. By definition, we can see that $\bar{X}_n \xrightarrow{p} \theta$ as $n \to \infty$. If we wish to find a consistent estimator of the standard deviation of X_n which is $\theta^{1/2}$ we can consider $\bar{X}_n^{1/2}$. CMT implies that the square root transformation is continuous at θ if $\theta > 0$ that $\bar{X}_n^{1/2} \xrightarrow{p} \theta^{1/2}$ as $n \to \infty$.

In Example 1.2.2, the condition that $(X_n, Y_n) \xrightarrow{d} N_2(\mathbf{0}, \mathbf{I}_2)$ cannot be relaxed to $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$ where X and Y are independent, i.e., we need the convergence of the joint CDF of (X_n, Y_n). This is different when $\xrightarrow{d}$ is replaced by $\xrightarrow{p}$ or $\xrightarrow{wp1}$, such as in Example 1.2.3-(ii). The following result, which plays an important role in probability and statistics, establishes the convergence in distribution of $X_n + Y_n$ or X_nY_n when no information regarding the joint CDF of (X_n, Y_n) is provided.

Theorem 1.2.3 (Slutsky's Theorem) *Let $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, where c is a finite constant. Then,*

(i) $X_n + Y_n \xrightarrow{d} X + c$;

(ii) $X_n Y_n \xrightarrow{d} cX$;

(iii) $X_n / Y_n \xrightarrow{d} X/c$ *if* $c \neq 0$.

Proof. The method of proof of the theorem is demonstrated sufficiently by proving (i). Choose and fix t such that $t - c$ is a continuity point of F_X. Let $\varepsilon > 0$ be such that $t - c + \varepsilon$ and $t - c - \varepsilon$ are also continuity points of F_X. Then

$$\begin{aligned} F_{X_n+Y_n}(t) &= P(X_n + Y_n \leq t) \\ &\leq P(X_n + Y_n \leq t, |Y_n - c| < \varepsilon) + P(|Y_n - c| \geq \varepsilon) \\ &\leq P(X_n \leq t - c + \varepsilon) + P(|Y_n - c| \geq \varepsilon) \end{aligned}$$

and, similarly

$$F_{X_n+Y_n}(t) \geq P(X_n \leq t - c - \varepsilon) - P(|Y_n - c| \geq \varepsilon).$$

It follows from the previous two inequalities and the hypotheses of the theorem that

$$F_X(t - c - \varepsilon) \leq \liminf_n F_{X_n+Y_n}(t) \leq \limsup_n F_{X_n+Y_n}(t) \leq F_X(t - c + \varepsilon).$$

Since $t - c$ is a continuity point of F_X, and since ε can be taken arbitrary small, the above equation yields

$$\lim_n F_{X_n+Y_n}(t) = F_X(t - c).$$

The result follows from $F_X(t - c) = F_{X+c}(t)$. □

Extensions to the vector case is straightforward. (iii) is valid provided $\mathbf{C} \neq \mathbf{0}$ is understood as $\mathbf{C}$ being invertible.

A straightforward but often used result by this theorem is that $X_n \xrightarrow{d} X$ and $X_n - Y_n \xrightarrow{p} 0$, then $Y_n \xrightarrow{d} X$. In asymptotic practice, we often firstly derive the result such as $Y_n = X_n + o_p(1)$ and then investigate the asymptotic distribution of X_n.

Example 1.2.6 (i) Theorem 1.2.1-(iii); Furthermore, convergence in probability to a constant is equivalent to convergence in law to the given constant. "⇒" follows from the part (i). "⇐" can be proved by definition. Because the degenerate distribution function of constant c is continuous everywhere except for point c, for any $\epsilon > 0$,

$$\begin{aligned} P(|X_n - c| \geq \epsilon) &= P(X_n \geq c + \epsilon) + P(X_n \leq c - \epsilon) \\ &\to 1 - F_X(c + \epsilon) + F_X(c - \epsilon) = 0 \end{aligned}$$

The results follows from the definition of convergence in probability.

Example 1.2.7 Let $\{X_n\}_{n=1}^{\infty}$ is a sequence of independent random variables where $X_n \sim \text{Gamma}(\alpha_n, \beta_n)$, where α_n and β_n are sequences of positive real numbers such that $\alpha_n \to \alpha$ and $\beta_n \to \beta$ for some positive real numbers α and β. Also, let $\hat{\beta}_n$ be a consistent estimator of β. We can conclude that $X_n/\hat{\beta}_n \xrightarrow{d} \text{Gamma}(\alpha, 1)$.

Example 1.2.8 (t-statistic) Let $X_1, X_2, \ldots$ be iid random variables with $EX_1 = 0$ and $EX_1^2 < \infty$. Then the t-statistic $\sqrt{n}\bar{X}_n/S_n$, where $S_n^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ is the sample variance, is asymptotically standard normal. To see this, first note that by two applications of WLLN and CMT

$$S_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2\right) \xrightarrow{p} 1(EX_1^2 - (EX_1)^2) = \text{Var}(X_1).$$

Again, by CMT, $S_n \xrightarrow{p} \sqrt{\text{Var}(X_1)}$. By the CLT, $\sqrt{n}\bar{X}_n \xrightarrow{d} N(0, \text{Var}(X_1))$. Finally, Slutsky's Theorem gives that the sequence of t-statistics converges in law to $N(0, \text{Var}(X_1))/\sqrt{\text{Var}(X_1)} = N(0, 1)$.

1.2.3 WLLN and SLLN

We next state some theorems known as the *laws of large numbers*. It concerns the limiting behavior of sums of independent random variables. The weak law of large numbers (WLLN) refers to convergence in probability, whereas the strong of large numbers (SLLN) refers to a.s. convergence. Our first result gives the WLLN and SLLN for a sequence of iid random variables.

Theorem 1.2.4 *Let $X_1, X_2, \ldots,$ be iid random variables having a CDF F.*

(i) **The WLLN** *The existence of constants a_n for which*

$$\frac{1}{n}\sum_{i=1}^{n} X_i - a_n \xrightarrow{p} 0$$

holds iff $\lim_{x\to\infty} x[1 - F(x) + F(-x)] = 0$, in which case we may choose $a_n = \int_{-n}^{n} x dF(x)$.

(ii) **The SLLN** *The existence of a constant c for which*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{wp1} c$$

holds iff $E[X_1]$ is finite and equals c.

Example 1.2.9 Suppose $\{X_i\}_{i=1}^{\infty}$ is a sequence of independent random variables where $X_i \sim t(2)$. The variance of X_i does not exist, but Theorem 1.2.4 still applies to this case and we can still therefore conclude that $\bar{X}_n \xrightarrow{p} 0$ as $n \to \infty$.

The next result is for sequences of independent but not necessarily identically distributed random variables.

Theorem 1.2.5 *Let $X_1, X_2, \ldots,$ be random variables with finite expectations.*

(i) **The WLLN** *Let $X_1, X_2, \ldots,$ be uncorrelated with means $\mu_1, \mu_2, \ldots$ and variances $\sigma_1^2, \sigma_2^2, \ldots$. If $\lim_{n\to\infty} \frac{1}{n^2}\sum_{i=1}^{n} \sigma_i^2 = 0$, then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} \mu_i \xrightarrow{p} 0.$$

(ii) **The SLLN** *Let $X_1, X_2, \ldots,$ be independent with means $\mu_1, \mu_2, \ldots$ and variances $\sigma_1^2, \sigma_2^2, \ldots$. If $\sum_{i=1}^{\infty} \sigma_i^2/c_i^2 < \infty$ where c_n ultimately monotone and $c_n \to \infty$, then*

$$c_n^{-1}\sum_{i=1}^{n}(X_i - \mu_i) \xrightarrow{wp1} 0.$$

(iii) **The SLLN with common mean** *Let $X_1, X_2, \ldots,$ be independent with common mean μ and variances $\sigma_1^2, \sigma_2^2, \ldots.$ If $\sum_{i=1}^{\infty} \sigma_i^{-2} = \infty$, then*

$$\sum_{i=1}^{n} \frac{X_i}{\sigma_i^2} / \sum_{i=1}^{n} \sigma_i^{-2} \overset{wp1}{\to} \mu.$$

A special case of Theorem 1.2.5-(ii) is to set $c_i = i$ in which we have

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} \mu_i \overset{wp1}{\to} 0.$$

The proof of Theorems 1.2.4 and 1.2.5 can be found in Billingsley (1995).

Example 1.2.10 Suppose $X_i \overset{\text{indep}}{\sim} (\mu, \sigma_i^2)$. Then, by simple calculus, the BLUE (best linear unbiased estimate) of μ is $\sum_{i=1}^{n} \sigma_i^{-2} X_i / \sum_{i=1}^{n} \sigma_i^{-2}$. Suppose now that the σ_i^2 do not grow at a rate faster than i; i.e., for some constant K, $\sigma_i^2 \leq iK$. Then, $\sum_{i=1}^{n} \sigma_i^{-2}$ clearly diverges as $n \to \infty$, and so by Theorem 1.2.5-(iii) the BLUE of μ is strongly consistent.

Example 1.2.11 Suppose (X_i, Y_i), $i = 1, \ldots, n$ are iid bivariate samples from some distribution with $E(X_1) = \mu_1$, $E(Y_1) = \mu_2$, $\mathrm{Var}(X_1) = \sigma_1^2$, $\mathrm{Var}(Y_1) = \sigma_2^2$, and $\mathrm{corr}(X_1, Y_1) = \rho$. Let r_n denote the sample correlation coefficient. The almost sure convergence of r_n to ρ follow very easily. We write

$$r_n = \frac{\frac{1}{n}\sum X_i Y_i - \bar{X}\bar{Y}}{\sqrt{(\sum \frac{X_i^2}{n} - \bar{X}^2)(\sum \frac{Y_i^2}{n} - \bar{Y}^2)}},$$

then from the SLLN for iid random variables (Theorem 1.2.4) and continuous mapping theorem (Theorem 1.2.2; Example 1.2.3-(ii)),

$$r_n \overset{wp1}{\to} \frac{E(X_1 Y_1) - \mu_1 \mu_2}{\sigma_1^2 \sigma_2^2} = \rho.$$

1.2.4 Characterization of convergence in law

Next we provide a collection of basic facts about convergence in distribution. The following theorems provide methodology for establishing convergence in distribution.

Theorem 1.2.6 *Let* $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \ldots$ *random p-vectors.*

(i) **(The Portmanteau Theorem)** $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ *is equivalent to the following condition:* $E[g(\mathbf{X}_n)] \to E[g(\mathbf{X})]$ *for every bounded continuous function* g*.*

(ii) **(Levy-Cramer continuity theorem)** *Let* $\Phi_{\mathbf{X}}$, $\Phi_{\mathbf{X}_1}$, $\Phi_{\mathbf{X}_2}, \ldots$ *be the character functions of* $\mathbf{X}, \mathbf{X_1}, \mathbf{X_2}, \ldots$*, respectively.* $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ *iff* $\lim_{n\to\infty} \Phi_{\mathbf{X}_n}(\mathbf{t}) = \Phi_{\mathbf{X}}(\mathbf{t})$ *for all* $\mathbf{t} \in \mathbb{R}^p$*.*

(iii) **(Cramer-Wold device)** $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ *iff* $\mathbf{c}^T\mathbf{X}_n \xrightarrow{d} \mathbf{c}^T\mathbf{X}$ *for every* $\mathbf{c} \in \mathbb{R}^p$*.*

Proof. (i) See Serfling (1980), page 16; (ii) Shao (2003), page 57; (iii) Assume $\mathbf{c}^T\mathbf{X}_n \xrightarrow{d} \mathbf{c}^T\mathbf{X}$ for any $\mathbf{c}$, then by Theorem 1.2.6-(ii)

$$\lim_{n\to\infty} \Phi_{\mathbf{X}_n}(tc_1, \ldots, tc_p) = \Phi_{\mathbf{X}}(tc_1, \ldots, tc_p), \text{ for all } t.$$

With $t = 1$, and since $\mathbf{c}$ is arbitrary, it follows by Theorem 1.2.6-(ii) again that $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$. The converse can be proved by a similar argument. [$\Phi_{\mathbf{c}^T\mathbf{X}_n}(t) = \Phi_{\mathbf{X}_n}(t\mathbf{c})$ and $\Phi_{\mathbf{c}^T\mathbf{X}}(t) = \Phi_{\mathbf{X}}(t\mathbf{c})$ for any $t \in \mathbb{R}$ and any $\mathbf{c} \in \mathbb{R}^p$.] □

A straightforward application of Theorem 1.2.6 is that if $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ and $\mathbf{Y}_n \xrightarrow{d} \mathbf{c}$ for constant vector $\mathbf{c}$, then $(\mathbf{X}_n, \mathbf{Y}_n) \xrightarrow{d} (\mathbf{X}, \mathbf{c})$.

Example 1.2.12 Example 1.1.3 revisited. Consider now the function $g(x) = x^{10}, 0 \leq x \leq 1$. Note that g is continuous and bounded. Therefore, by the Portmanteau theorem, $E(g(X_n)) = \sum_{i=1}^n \frac{i^{10}}{n^{11}} \to E(g(X)) = \int_0^1 x^{10}dx = \frac{1}{11}$.

Example 1.2.13 For $n \geq 1, 0 \leq p \leq 1$, and a given continuous function $g : [0, 1] \to \mathbb{R}$, define the sequence

$$B_n(p) = \sum_{k=0}^{n} g(\frac{k}{n}) C_n^k p^k (1-p)^{n-k},$$

which is so-called Bernstein polynomials. Note that $B_n(p) = E[g(\frac{X}{n})|X \sim \text{Bin}(n, p)]$. As $n \to \infty$, $\frac{X}{n} \xrightarrow{p} p$ (WLLN), and it follows that $\frac{X}{n} \xrightarrow{d} \delta_p$, the point mass at p. Since g is continuous and hence bounded (compact interval), it follows from the Portmanteau theorem that $B_n(p) \to g(p)$.

Example 1.2.14 (i) Let $X_1, \ldots, X_n$ be independent random variables having a common CDF and $T_n = X_1 + \ldots + X_n, n = 1, 2, \ldots$. Suppose that $E|X_1| < \infty$. It follows from the property of CHF and Taylor expansion that the CHF of X_1 satisfies $[\frac{\partial \Phi_X(t)}{\partial t}]|_{t=0} = \sqrt{-1}EX, [\frac{\partial^2 \Phi_X(t)}{\partial t^2}]|_{t=0} = -EX^2]$

$$\Phi_{X_1}(t) = \Phi_{X_1}(0) + \sqrt{-1}\mu t + o(|t|)$$

as $|t| \to 0$, where $\mu = EX_1$. Then, it follows that the CHF of T_n/n is

$$\Phi_{T_n/n}(t) = \left[\Phi_{X_1}\left(\frac{t}{n}\right)\right]^n = \left[1 + \frac{\sqrt{-1}\mu t}{n} + o(|t|n^{-1})\right]^n$$

for any $t \in \mathbb{R}$ as $n \to \infty$. Since $(1 + c_n/n)^n \to \exp\{c\}$ for any complex sequence c_n satisfying $c_n \to c$, we obtain that $\Phi_{T_n/n}(t) \to \exp\{\sqrt{-1}\mu t\}$, which is the CHF of the distribution degenerated at μ. By Theorem 1.2.6-(ii), $T_n/n \xrightarrow{d} \mu$. From 1.2.6-(i), this also shows that $T_n/n \xrightarrow{p} \mu$ (an informal proof of WLLN); (ii) Similarly, $\mu = 0$ and $\sigma^2 = \mathrm{Var}(X_1) < \infty$ imply [second-order Taylor expansion]

$$\Phi_{T_n/\sqrt{n}}(t) = \left[1 - \frac{\sigma^2 t^2}{2n} + o(t^2 n^{-1})\right]^n$$

for any $t \in \mathbb{R}$ as $n \to \infty$, which implies that $\Phi_{T_n/\sqrt{n}}(t) \to \exp\{-\sigma^2 t^2/2\}$, the CHF of $N(0, \sigma^2)$. Hence, $T_n/\sqrt{n} \xrightarrow{d} N(0, \sigma^2)$; (iii) Suppose now that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are random p-vectors and $\boldsymbol{\mu} = E\mathbf{X}_1$ and $\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{X}_1)$ are finite. For any fixed $\mathbf{c} \in \mathbb{R}^p$, it follows from the previous discussion that $(\mathbf{c}^T T_n - n\mathbf{c}^T\boldsymbol{\mu})/\sqrt{n} \xrightarrow{d} N(0, \mathbf{c}^T\boldsymbol{\Sigma}\mathbf{c})$. From Theorem 1.2.6-(iii), we conclude that $(T_n - n\boldsymbol{\mu})/\sqrt{n} \xrightarrow{d} N_p(\mathbf{0}, \boldsymbol{\Sigma})$.

The following two simple results are frequently useful in calculations.

Theorem 1.2.7 *(i)* **(Prohorov's Theorem)** *If $X_n \xrightarrow{d} X$ for some X, then $X_n = O_p(1)$.*

(ii) **(Polya's Theorem)** *If $F_{X_n} \Rightarrow F_X$ and F_X is continuous, then as $n \to \infty$,*

$$\sup_{-\infty < x < \infty} |F_{X_n} - F_X| \to 0.$$

Proof. (i) For any given $\varepsilon > 0$, fix a constant M such that $P(X \geq M) < \varepsilon$. By the definition of convergence in law, $P(|X_n| \geq M)$ exceeds $P(|X| \geq M)$ arbitrarily small for

sufficiently large n. Thus, there exists N such that $P(|X_n| \geq M) < 2\varepsilon$, for all $n \geq N$. The results follows from the definition of $O_p(1)$. (ii) Firstly, fix $k \in \mathbb{N}$. By the continuity of F there exists points $-\infty = x_0 < x_1 < \cdots < x_k = \infty$ with $F(x_i) = i/k$. By monotonicity, we have, for $x_{i-1} \leq x \leq x_i$,

$$\begin{aligned} F_{X_n}(x) - F_X(x) &\leq F_{X_n}(x_i) - F_X(x_{i-1}) = F_{X_n}(x_i) - F_X(x_i) + 1/k \\ &\geq F_{X_n}(x_{i-1}) - F_X(x_i) = F_{X_n}(x_{i-1}) - F_X(x_{i-1}) - 1/k. \end{aligned}$$

Thus, $F_{X_n}(x) - F_X(x)$ is bounded above by $\sup_i |F_{X_n}(x_i) - F_X(x_i)| + 1/k$, for every x. The latter, finite supremum converges to zero because each term converges to zero due to the condition, for each fixed k. Because k is arbitrary, the result follows. □

The following result can be used to check whether $X_n \xrightarrow{d} X$ when X has a PDF f and X_n has a PDF f_n.

Theorem 1.2.8 (Scheffe Theorem) *Let f_n be a sequence of densities of absolutely continuous functions,, with $\lim_n f_n(\mathbf{x}) = f(\mathbf{x})$, each $\mathbf{x} \in \mathbb{R}^p$. If f is a density function, then $\lim_n \int |f_n(\mathbf{x}) - f(\mathbf{x})| d\mathbf{x} = 0$.*

Proof. Put $g_n(\mathbf{x}) = [f(\mathbf{x}) - f_n(\mathbf{x})] I_{f(\mathbf{x}) \geq f_n(\mathbf{x})}$. By noting that $\int [f_n(\mathbf{x}) - f(\mathbf{x})] d\mathbf{x} = 0$,

$$\int |f_n(\mathbf{x}) - f(\mathbf{x})| d\mathbf{x} = 2 \int g_n(\mathbf{x}) d\mathbf{x}.$$

Since $0 \leq g_n(\mathbf{x}) \leq f(\mathbf{x})$ for all $\mathbf{x}$. Hence, by dominated convergence, $\lim_n \int g_n(\mathbf{x}) d\mathbf{x} = 0$. [Dominated convergence theorem. If $\lim_{n\to\infty} f_n = f$ and there exists an integrable function g such that $|f_n| \leq g$, then $\lim_n \int f_n(x) dx = \int \lim_n f_n(x) dx$ holds] □

As an example, consider the PDF f_n of the t- distribution $t_n, n = 1, 2, \ldots$. One can show (exercise) that $f_n \to f$, where f is the standard normal PDF.

The following result provides a convergence of moments criterion for convergence in law.

Theorem 1.2.9 (Frechet and Shohat Theorem) *Let the distribution function F_n possess finite moments $\alpha_{nk} = \int t^k dF_n(t)$ for $k = 1, 2, \ldots$ and $n = 1, 2, \ldots$. Assume that the limits $\alpha_k = \lim_n \alpha_{nk}$ exist (finite) for each k. Then,*

(i) the limits α_k are the moments of some a distribution function F;

(ii) if the F given by (i) is unique, then $F_n \Rightarrow F$.

[A sufficient condition: the moment sequence α_k determines the distribution F uniquely if the *Carleman condition* $\sum_{i=1}^{\infty} \alpha_{2i}^{-1/(2i)} = \infty$ holds.]

1.2.5 Results on o_p and O_p

There are many rules of calculus with o and O symbols, which we will apply without comment. For instance,

$$o_p(1) + o_p(1) = o_p(1),\ o_p(1) + O_p(1) = O_p(1),\ O_p(1)o_p(1) = o_p(1)$$
$$(1 + o_p(1))^{-1} = O_p(1),\ o_p(R_n) = R_n o_p(1),\ O_p(R_n) = R_n O_p(1),\ o_p(O_p(1)) = o_p(1).$$

Two more complicated rules are given by the following lemma.

Lemma 1.2.1 *Let g be a function defined on $\mathbb{R}^p$ such that $g(\mathbf{0}) = 0$. Let $\mathbf{X}_n$ be a sequence of random vectors with values on $\mathbb{R}$ that converges in probability to zero. Then, for every $r > 0$,*

(i) if $g(\mathbf{t}) = o(||\mathbf{t}||^r)$ as $t \to 0$, then $g(\mathbf{X}_n) = o_p(||\mathbf{X}_n||^r)$;

(ii) if $g(\mathbf{t}) = O(||\mathbf{t}||^r)$ as $t \to 0$, then $g(\mathbf{X}_n) = O_p(||\mathbf{X}_n||^r)$.

Proof. Define $f(\mathbf{t}) = g(\mathbf{t})/||\mathbf{t}||^r$ for $\mathbf{t} \neq 0$ and $f(\mathbf{0}) = 0$. Then $g(\mathbf{X}_n) = f(\mathbf{X}_n)||\mathbf{X}_n||^r$.

(i) Because the function f is continuous at zero by assumption, $f(\mathbf{X}_n) \xrightarrow{p} f(\mathbf{0}) = 0$ by Theorem 1.2.2.

(ii) By assumption there exists M and $\delta > 0$ such that $|f(\mathbf{t})| \leq M$ whenever $||\mathbf{t}|| \leq \delta$. Thus

$$P(|f(\mathbf{X}_n)| > M) \leq P(||\mathbf{X}_n|| > \delta) \to 0,$$

and the sequence $f(\mathbf{X}_n)$ is bounded. □

1.3 The central limit theorem

The most fundamental result on convergence in law is the central limit theorem (CLT) for sums of random variables. We firstly state the case of chief importance, iid summands.

Definition 1.3.1 *A sequence of random variables $X_n$is asymptotically normal with μ_n and σ_n^2 if $(X_n - \mu_n)/\sigma_n \xrightarrow{d} N(0,1)$, written by X_n is $AN(\mu_n, \sigma_n^2)$.*

1.3.1 The CLT for the iid case

Theorem 1.3.1 (Lindeberg-Levy) *Let X_i be iid with mean μ and finite variance σ^2. Then*

$$\frac{\sqrt{n}\left(\bar{X} - \mu\right)}{\sigma} \xrightarrow{d} N(0,1).$$

By Slutsky's Theorem, we can write $\sqrt{n}\left(\bar{X} - \mu\right) \xrightarrow{d} N(0, \sigma^2)$. Also, $\bar{X}$ is $AN(\mu, \sigma^2/n)$. See Billingsley (1995) for a proof.

Example 1.3.1 (Confidence intervals) This theorem can be used to approximate $P(\bar{X} \leq \mu + \frac{k\sigma}{\sqrt{n}})$ by $\Phi(k)$. This is very useful because the sampling distribution of $\bar{X}$ is not available except for some special cases. Then, setting $k = \Phi^{-1}(1-\alpha) = z_\alpha$, $[\bar{X}_n - \sigma/\sqrt{n}z_\alpha, \bar{X}_n + \sigma/\sqrt{n}z_\alpha]$ is a confidence interval for μ of asymptotic level $1 - 2\alpha$. More precisely, we have that the probability that μ is contained in this interval converges to $1 - 2\alpha$ (how accurate?).

Example 1.3.2 (**Sample variance**) Suppose $X_1, \ldots, X_n$ are iid with mean μ, variance σ^2 and $E(X_1^4) < \infty$. Consider the asymptotic distribution of $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. Write

$$\sqrt{n}(S_n^2 - \sigma^2) = \sqrt{n}\left(\frac{1}{n-1}\sum_{i=1}^n (X_i - \mu)^2 - \sigma^2\right) - \sqrt{n}\frac{n}{n-1}(\bar{X}_n - \mu)^2.$$

The second term converges to zero in probability and the first term is asymptotically normal by the CLT. The whole expression is asymptotically normal by the Slutsky' Theorem, i.e.,

$$\sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{d} N(0, \mu_4 - \sigma^4),$$

where μ_4 denotes the centered fourth moment of X_1 and $\mu_4 - \sigma^4$ comes certainly from computing the variance of $(X_1 - \mu)^2$.

Example 1.3.3 (**Level of the Chi-square test**) Normal theory prescribes to reject the null hypothesis $H_0 : \sigma^2 \leq 1$ for values of nS_n^2 exceeding the upper α point $\chi^2_{n-1,\alpha}$ of the χ^2_{n-1} distribution. If the observations are sample from a normal distribution, the test has exactly level α. However, this is not approximately the case of the underlying distribution is not normal. The CLT and the Example 1.3.2 yield the following two statements

$$\frac{\chi^2_{n-1} - (n-1)}{\sqrt{2(n-1)}} \xrightarrow{d} N(0,1), \ \sqrt{n}\left(\frac{S_n^2}{\sigma^2} - 1\right) \xrightarrow{d} N(0, \kappa + 2),$$

where $\kappa = \mu_4/\sigma^4 - 3$ is the kurtosis of the underlying distribution. The first statement implies that $(\chi^2_{n-1,\alpha} - (n-1))/\sqrt{2(n-1)}$ converges to the upper α point z_α of $N(0,1)$. Thus, the level of the chi-square test satisfies

$$P_{H_0}(nS_n^2 > \chi^2_{n-1,\alpha}) = P\left(\sqrt{n}\left(\frac{S_n^2}{\sigma^2} - 1\right) > \frac{\chi^2_{n-1,\alpha} - n}{\sqrt{n}}\right) \to 1 - \Phi\left(\frac{z_\alpha\sqrt{2}}{\sqrt{k+2}}\right)$$

So, the asymptotic level reduces to $1-\Phi(z_\alpha) = \alpha$ iff the kurtosis of the underlying distribution is 0. If the kurtosis goes to infinity, then the asymptotic level approaches to $1 - \Phi(0) = 1/2$. We conclude that the level of the chi-square test is nonrobust against departures of normality that affect the value of the kurtosis. If, instead, we would use a normal approximation to the distribution $\sqrt{n}(S_n^2/\sigma^2 - 1)$ the problem would not arise, provided that the asymptotic variance $\kappa + 2$ is estimated accurately.

Theorem 1.3.2 (Multivariate CLT for iid case) *Let $\mathbf{X}_i$ be iid random p-vectors with mean $\boldsymbol{\mu}$ and and covariance matrix $\boldsymbol{\Sigma}$. Then*

$$\sqrt{n}\left(\bar{\mathbf{X}} - \boldsymbol{\mu}\right) \xrightarrow{d} N_p(\mathbf{0}, \boldsymbol{\Sigma}).$$

Proof. By the Cramer-Wold device, this can be proved by finding the limit distribution of the sequences of real variables

$$\mathbf{c}^T\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu})\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{c}^T\mathbf{X}_i - \mathbf{c}^T\boldsymbol{\mu}).$$

Because the random variables $\mathbf{c}^T\mathbf{X}_i - \mathbf{c}^T\boldsymbol{\mu}$ are iid with zero mean and variance $\mathbf{c}^T\boldsymbol{\Sigma}\mathbf{c}$, this sequence is $AN(0, \mathbf{c}^T\boldsymbol{\Sigma}\mathbf{c})$ by Theorem 1.3.1. This is exactly the distribution of $\mathbf{c}^T\mathbf{X}$ if $\mathbf{X}$ possesses an $N_p(\mathbf{0}, \boldsymbol{\Sigma})$. □

Example 1.3.4 Suppose that $X_1, \ldots, X_n$ is a random sample from the Poisson distribution with mean θ. Let Z_n be the proportions of zero observed, i.e., $Z_n = 1/n \sum_{i=1}^n I_{\{X_j=0\}}$. Let us find the joint asymptotic distribution of $(\bar{X}_n, Z_n)$. Note that $E(X_1) = \theta$, $EI_{\{X_1=0\}} = e^{-\theta}$, $\mathrm{Var}(X_1) = \theta$, $\mathrm{Var}(I_{\{X_1=0\}}) = e^{-\theta}(1 - e^{-\theta})$, and $EX_1I_{\{X_1=0\}} = 0$. So, $\mathrm{Cov}(X_1, I_{\{X_1=0\}}) = -\theta e^{-\theta}$. Hence, $\sqrt{n}\left((\bar{X}_n, Z_n) - (\theta, e^{-\theta})\right) \xrightarrow{d} N_2(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \theta & -\theta e^{-\theta} \\ -\theta e^{-\theta} & e^{-\theta}(1 - e^{-\theta}) \end{pmatrix}.$$

It is not as widely known that existence of a variance is not necessary for asymptotic normality of partial sums of iid random variables. A CLT without a finite variance can sometimes be useful. We present the general result below and then give an illustrative example. Feller (1966) contains detailed information on the availability of CLTs without the existence of a variance, along with proofs. First, we need a definition.

Definition 1.3.2 *A function $g : \mathbb{R} \to \mathbb{R}$ is called slowly varying at ∞ if, for every $t > 0$, $\lim_{x\to\infty} g(tx)/g(x) = 1$.*

Examples of slowly varying functions are $\log x$, $x/(1 + x)$, and indeed any function with a finite limit as $x \to \infty$. But, for example, x or e^{-x} are not slowly varying.

Theorem 1.3.3 *Let $X_1, X_2, \ldots$ be iid from a CDF F on $\mathbb{R}$. Let $v(x) = \int_{-x}^{x} y^2 dF(y)$. Then, there exist constants $\{a_n\}, \{b_n\}$ such that*

$$\frac{\sum_{i=1}^n X_i - a_n}{b_n} \xrightarrow{d} N(0, 1),$$

if and only if $v(x)$ is slowly varying at ∞.

If F has a finite second moment, then automatically $v(x)$ is slowly varying at ∞. We present an example below where asymptotic normality of the sample partial sums still holds, although the summands do not have a finite variance.

Example 1.3.5 Suppose $X_1, X_2, \ldots$ are iid from a t-distribution with 2 degrees of freedom ($t(2)$) that has a finite mean but not a finite variance. The density is given by $f(y) = c/(2+y^2)^{\frac{3}{2}}$ for some positive c. Hence, by a direct integration, for some other constant k,

$$v(x) = k\sqrt{\frac{1}{2+x^2}}\left[x - \sqrt{2+x^2}\text{arcsinh}(x/\sqrt{2})\right].$$

Therefore, on using the fact that $\text{arcsinh}(x) = \log(2x) + O(x^{-2})$ as $x \to \infty$, we get, for any $t > 0$, $\frac{v(tx)}{v(x)} \to 1$ on some algebra. It follows that for iid observations from a $t(2)$ distribution, on suitable centering and normalizing, the partial sums $\sum_{i=1}^n X_i$ converge to a normal distribution, although the X_i's do not have a finite variance. The centering can be taken to be zero for the centered t-distribution; it can be shown that the normalizing required is $b_n = \sqrt{n \log n}$ (why?).

1.3.2 The CLT for the independent not necessarily iid case

Theorem 1.3.4 (Lindeberg-Feller) *Suppose X_n is a sequence of independent variables with means μ_n and variances $\sigma_n^2 < \infty$. Let $s_n^2 = \sum_{i=1}^n \sigma_i^2$. If for any $\epsilon > 0$*

$$\frac{1}{s_n^2}\sum_{j=1}^n \int_{|x-\mu_j|>\epsilon s_n} (x-\mu_j)^2 dF_j(x) \to 0, \tag{1.2}$$

where F_i is the CDF of X_i, then

$$\frac{\sum_{i=1}^n (X_i - \mu_i)}{s_n} \xrightarrow{d} N(0,1).$$

A proof can be seen on page 67 in Shao (2003). The condition (1.2) is called Lindeberg-Feller condition.

Example 1.3.6 Let $X_1, X_2 \ldots,$ be independent variables such that X_j has the uniform distribution on $[-j, j], j = 1, 2, \ldots$. Let us verify the conditions of Theorem 1.3.4 are satisfied. Note that $EX_j = 0$ and $\sigma_j^2 = \frac{1}{2j}\int_{-j}^{j} x^2 dx = j^2/3$ for all j. Hence,

$$s_n^2 = \sum_{j=1}^{n} \sigma_j^2 = \frac{1}{3}\sum_{j=1}^{n} j^2 = \frac{n(n+1)(2n+1)}{18}.$$

For any $\epsilon > 0$, $n < \epsilon s_n$ for sufficiently large n, since $\lim_n n/s_n = 0$. Because $|X_j| \leq j \leq n$, when n is sufficiently large,

$$E(X_j^2 I_{\{|X_j| > \epsilon s_n\}}) = 0.$$

Consequently, $\lim_{n\to\infty} \sum_{j=1}^{n} E(X_j^2 I_{\{|X_j| > \epsilon s_n\}}) < \infty$. Considering $s_n \to \infty$, Lindeberg's condition holds.

The Lindeberg- Feller theorem is a landmark theorem in probability and statistics. Generally, it is hard to verify the Lindeberg-Feller condition. A simpler theorem is the following.

Theorem 1.3.5 (Liapounov) *Suppose X_n is a sequence of independent variables with means μ_n and variances $\sigma_n^2 < \infty$. Let $s_n^2 = \sum_{i=1}^{n} \sigma_i^2$. If for some $\delta > 0$*

$$\frac{1}{s_n^{2+\delta}} \sum_{j=1}^{n} E|X_j - \mu_j|^{2+\delta} \to 0 \tag{1.3}$$

as $n \to \infty$, then

$$\frac{\sum_{i=1}^{n}(X_i - \mu_i)}{s_n} \xrightarrow{d} N(0, 1).$$

A proof is given in Sen and Singer (1993). For instance, if $s_n \to \infty$, $\sup_{j\geq 1} E|X_j - \mu_j|^{2+\delta} < \infty$ and $n^{-1}s_n$ is bounded, then the condition of Liapounov's theorem is satisfied. In practice, usually one tries to work with $\delta = 1$ or 2 for algebraic convenience. It can be easily checked that if X_i is uniformly bounded and $s_n \to \infty$, the condition is immediately satisfied with $\delta = 1$.

Example 1.3.7 Let $X_1, X_2, \ldots$ be independent random variables. Suppose that X_i has the binomial distribution $\text{BIN}(p_i, 1), i = 1, 2, \ldots$. For each i, $EX_i = p_i$ and $E|X_i - EX_i|^3 =$

$(1-p_i)^3 p_i + p_i^3(1-p_i) \le 2p_i(1-p_i)$. Hence, $\sum_{i=1}^n E|X_i - EX_i|^3 \le 2s_n^2 = 2\sum_{i=1}^n E|X_i - EX_i|^2 = 2\sum_{i=1}^n p_i(1-p_i)$. Then Liapounov's condition (1.3) holds with $\delta = 1$ if $s_n \to \infty$. For example, if $p_i = 1/i$ or $M_1 \le p_i \le M_2$ with two constants belong to $(0,1)$, $s_n \to \infty$ holds. Accordingly, by Liapounov's theorem, $\frac{\sum_{i=1}^n (X_i - p_i)}{s_n} \xrightarrow{d} N(0,1)$.

A consequence especially useful in regression is the following theorem, which is also proved in Sen and Singer (1993).

Theorem 1.3.6 (Hajek-Sidak) *Suppose $X_1, X_2, \ldots$ are iid random variables with mean μ and variance $\sigma^2 < \infty$. Let $c_n = (c_{n1}, c_{n2}, \ldots, c_{nn})$ be a vector of constants such that*

$$\max_{1 \le i \le n} \frac{c_{ni}^2}{\sum_{j=1}^n c_{nj}^2} \to 0 \tag{1.4}$$

as $n \to \infty$. Then

$$\frac{\sum_{i=1}^n c_{ni}(X_i - \mu)}{\sigma \sqrt{\sum_{j=1}^n c_{nj}^2}} \xrightarrow{d} N(0,1).$$

The condition (1.4) is to ensure that no coefficient dominates the vector c_n, and is referred as Hajek-Sidak condition in the literatures. For example, if $c_n = (1, 0, \ldots, 0)$, then the condition would fail and so would the theorem. The Hajek-Sidak's theorem has many applications, including in the regression problem. Here is an important example.

Example 1.3.8 (**Simplest linear regression**) Consider the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where ε_i's are iid with mean 0 and variance σ^2 but are not necessarily normally distributed. The least squares estimate of β_1 based on n observations is

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} = \beta_1 + \frac{\sum_{i=1}^n \varepsilon_i (x_i - \bar{x}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}.$$

So, $\widehat{\beta}_1 = \beta_1 + \sum_{i=1}^n \varepsilon_i c_{ni} / \sum_{j=1}^n c_{nj}^2$, where $c_{ni} = x_i - \bar{x}_n$. Hence, by the Hajek-Sidak's Theorem

$$\sqrt{\sum_{j=1}^n c_{nj}^2} \frac{\widehat{\beta}_1 - \beta_1}{\sigma} = \frac{\sum_{i=1}^n \varepsilon_i c_{ni}}{\sigma \sqrt{\sum_{j=1}^n c_{nj}^2}} \xrightarrow{d} N(0,1),$$

provided

$$\frac{\max_{1\le i\le n}(x_i-\bar{x}_n)^2}{\sum_{j=1}^n (x_j-\bar{x}_n)^2}\to 0$$

as $n\to\infty$. For most reasonable designs, this condition is satisfied. Thus, the asymptotic normality of the LSE (least squares estimate) is established under some conditions on the design variables, an important result.

Theorem 1.3.7 (Lindeberg-Feller multivariate) *Suppose $\mathbf{X}_i$ is a sequence of independent vectors with means $\boldsymbol{\mu}_i$, covariances $\boldsymbol{\Sigma}_i$ and distribution function F_i. Suppose that $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\Sigma}_i \to \boldsymbol{\Sigma}$ as $n\to\infty$, and that for any $\epsilon>0$*

$$\frac{1}{n}\sum_{j=1}^n \int_{||\mathbf{x}-\boldsymbol{\mu}_j||>\epsilon\sqrt{n}} ||\mathbf{x}-\boldsymbol{\mu}_j||^2 dF_j(\mathbf{x})\to 0,$$

then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n (\mathbf{X}_i-\boldsymbol{\mu}_i)\xrightarrow{d} N(\mathbf{0},\boldsymbol{\Sigma}).$$

Example 1.3.9 (multiple regression) In the linear regression problem, we observe a vector $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ for a fixed or random matrix $\mathbf{X}$ of full rank, and an error vector $\boldsymbol{\varepsilon}$ with iid components with mean zero and variance σ^2. The least squares estimator of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. This estimator is unbiased and has covariance matrix $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. If the error vector $\boldsymbol{\varepsilon}$ is normally distributed, then $\widehat{\boldsymbol{\beta}}$ is exactly normally distributed. Under reasonable conditions on the design matrix, $\widehat{\boldsymbol{\beta}}$ is asymptotically normally distributed for a large range of error distributions. Here we fix p and let n tend to infinity. This follows from the representation

$$(\mathbf{X}^T\mathbf{X})^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})=(\mathbf{X}^T\mathbf{X})^{-1/2}\mathbf{X}^T\boldsymbol{\varepsilon}=\sum_{i=1}^n \mathbf{a}_{ni}\varepsilon_i,$$

where $\mathbf{a}_{n1},\ldots,\mathbf{a}_{nn}$ are the columns of the $(p\times n)$ matrix $(\mathbf{X}^T\mathbf{X})^{-1/2}\mathbf{X}^T=:\mathbf{A}$. This sequence is asymptotically normal if the vectors $\mathbf{a}_{n1}\varepsilon_1,\ldots,\mathbf{a}_{nn}\varepsilon_n$ satisfy the Lindeberg conditions. The norming matrix $(\mathbf{X}^T\mathbf{X})^{1/2}$ has been chosen to ensure that the vectors in the display have covariance matrix $\sigma^2\mathbf{I}_p$ for every n. The remaining condition is

$$\sum_{i=1}^n ||\mathbf{a}_{ni}||^2 E\varepsilon_i^2 I_{\{||\mathbf{a}_{ni}||\,|\varepsilon_i|>\epsilon\}}\to 0.$$

This can be simplified to other conditions in several ways. Because $\sum ||\mathbf{a}_{ni}||^2 = \text{tr}(\mathbf{A}\mathbf{A}^T) = p$, it suffices that $\max_i E\varepsilon_i^2 I_{\{||\mathbf{a}_{ni}|||\varepsilon_i|>\epsilon\}} \to 0$, which is also equivalent to $\max_i ||\mathbf{a}_{ni}|| \to 0$. Alternatively, the expectation $E\varepsilon_i^2 I_{\{||\mathbf{a}_{ni}|||\varepsilon_i|>\epsilon\}}$ can be bounded $\epsilon^{-k}E|\varepsilon_i|^{k+2}||\mathbf{a}_{ni}||^k$ and a second set of sufficient conditions is

$$\sum_{i=1}^{n} ||\mathbf{a}_{ni}||^k \to 0; \quad E|\varepsilon_1|^k < \infty, \ k > 2.$$

1.3.3 CLT for a random number of summands

The canonical CLT for the iid case says that if $X_1, X_2, \ldots$ are iid with mean zero and a finite variance σ^2, then the sequence of partial sums $T_n = \sum_{i=1}^n X_i$ obeys the central limit theorem in the sense $\frac{T_n}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1)$. There are some practical problems that arise in applications, for example in sequential statistical analysis, where the number of terms present in a partial sum is a random variable. Precisely, $\{N(t)\}, t \geq 0$, is a family of (nonnegative) integer-valued random variables, and we want to approximate the distribution of $T_{N(t)}$, where for each fixed n, T_n is still the sum of n iid variables as above. The question is whether a CLT still holds under appropriate conditions. Here is the Anscombe-Renyi theorem.

Theorem 1.3.8 (Anscombe-Renyi) *Let X_i be iid with mean μ and a finite variance σ^2, and let $\{N_n\}$, be a sequence of (nonnegative) integer-valued random variables and $\{a_n\}$ a sequence of positive constants tending to ∞ such that $N_n/a_n \xrightarrow{p} c, 0 < c < \infty$, as $n \to \infty$. Then,*

$$\frac{T_{N_n} - N_n\mu}{\sigma\sqrt{N_n}} \xrightarrow{d} N(0,1) \quad as \quad n \to \infty.$$

Example 1.3.10 (coupon collection problem) Consider a problem in which a person keeps purchasing boxes of cereals until she obtains a full set of some n coupons. The assumptions are that the boxes have an equal probability of containing any of the n coupons mutually independently. Suppose that the costs of buying the cereal boxes are iid with some mean μ and some variance σ^2. If it takes N_n boxes to obtain the complete set of all n coupons, then $N_n/(n \ln n) \xrightarrow{p} 1$ as $n \to \infty$ The total cost to the customer to obtain the

complete set of coupons is $T_{N_n} = X_1 + \ldots + X_{N_n}$. By the Anscombe-Renyi theorem and Slutsky's theorem, we have that $\frac{T_{N_n} - N_n\mu}{\sigma\sqrt{n \ln n}}$ is approximately $N(0,1)$.

[On the distribution of N_n. Let t_i be the boxes to collect the i-th coupon after $i-1$ coupons have been collected. Observe that the probability of collecting a new coupon given $i-1$ coupons is $p_i = (n-i+1)/n$. Therefore, t_i has a geometric distribution with expectation $1/p_i$ and $N_n = \sum_{i=1}^n t_i$. By Theorem 1.2.5, we know

$$\frac{1}{n \ln n} N_n \xrightarrow{p} \frac{1}{n \ln n} \sum_{i=1}^n p_i^{-1} = \frac{1}{n \ln n} \sum_{i=1}^n n \frac{1}{i} = \frac{1}{\ln n} \sum_{i=1}^n \frac{1}{i} =: \frac{1}{\ln n} H_n.$$

Note that H_n is the harmonic number and hence by using the asymptotics of the harmonic numbers ($H_n = \ln n + \gamma + o(1)$; γ is Euler-constant), we obtain $\frac{N_n}{n \ln n} \to 1$.]

1.3.4 Central limit theorems for dependent sequences

The assumption that observed data $X_1, X_2, \ldots$ form an independent sequence is often one of technical convenience. Real data frequently exhibit some dependence and at the least some correlation at small lags. Exact sampling distributions for fixed n are even more complicated for dependent data than in the independent case, and so asymptotics remain useful. In this subsection, we present CLTs for some important dependence structures. The cases of stationary m-dependence and without replacement sampling are considered.

Stationary m-dependence

We start with an example to illustrate that a CLT for sample means can hold even if the summands are not independent.

Example 1.3.11 Suppose $X_1, X_2, \ldots$ is a stationary Gaussian sequence with $E(X_i) = \mu$, $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Then, for each n, $\sqrt{n}(\bar{X}_n - \mu)$ is normally distributed and so $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \tau^2)$, provided $\tau^2 = \lim_{n\to\infty} \mathrm{Var}(\sqrt{n}(\bar{X}_n - \mu)) < \infty$. But

$$\mathrm{Var}(\sqrt{n}(\bar{X}_n - \mu)) = \sigma^2 + \frac{1}{n} \sum_{i \neq j} \mathrm{Cov}(X_i, X_j) = \sigma^2 + \frac{2}{n} \sum_{i=1}^n (n-i)\gamma_i,$$

where $\gamma_i = \mathrm{Cov}(X_1, X_{i+1})$. Therefore, $\tau^2 < \infty$ if and only if $\frac{1}{n}\sum_{i=1}^{n}(n-i)\gamma_i$ has a finite limit, say ρ, in which case $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2 + \rho)$.

What is going on qualitatively is that $\frac{1}{n}\sum_{i=1}^{n}(n-i)\gamma_i$ is summable when $|\gamma_i| \to 0$ adequately fast. Instances of this are when only a fixed finite number of the γ_i are nonzero or when γ_i is damped exponentially; i.e., $\gamma_i = O(a^i)$ for some $|a| < 1$. It turns out that there are general CLTs for sample averages under such conditions. The case of m-dependence is provided below.

Definition 1.3.3 *A stationary sequence $\{X_n\}$ is called m-dependent for a given fixed m if $(X_1, \ldots, X_i)$ and $(X_j, X_{j+1}, \ldots)$ are independent whenever $j - i > m$.*

Theorem 1.3.9 (m-dependent sequence) *Let $\{X_i\}$ be a stationary m-dependent sequence. Let $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Then $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \tau^2)$, where $\tau^2 = \sigma^2 + 2\sum_{i=2}^{m+1}\mathrm{Cov}(X_1, X_i)$.*

See Lehmann (1999) for a proof; m-dependent data arise either as standard time series models or as models in their own right. For example, if $\{Z_i\}$ are i.i.d. random variables and $X_i = a_1 Z_{i-1} + a_2 Z_{i-2}, i \geq 3$, then $\{X_i\}$ is 1-dependent. This is a simple moving average process of use in time series analysis. A more general m-dependent sequence is $X_i = h(Z_i, Z_{i+1}, \ldots, Z_{i+m})$ for some function h.

Example 1.3.12 Suppose Z_i are i.i.d. with a finite variance σ^2, and let $X_i = (Z_i + Z_{i+1})/2$. Then, obviously $\sum_{i=1}^{n} X_i = \frac{Z_1 + Z_{n+1}}{2} + \sum_{i=2}^{n} Z_i$. Then, by Slutsky's theorem, $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$. Notice we write $\sqrt{n}(\bar{X}_n - \mu)$ into two parts in which one part is dominant and produces the CLT, and the other part is asymptotically negligible. This is essentially the method of proof of the CLT for more general m-dependent sequences.

Sampling without replacement

Dependent data also naturally arise in sampling without replacement from a finite population. Central limit theorems are available and we will present them shortly. But let us start

with an illustrative example.

Example 1.3.13 Suppose, among N objects in a population, D are of type 1 and $N - D$ of type 2. A sample without replacement of size n is taken, and let X be the number of sampled units of type 1. We can regard these D type 1 units as having numerical values $X_1, \ldots, X_D = 1$ and the rest as having values $X_{D+1}, \ldots, X_N = 0$, $X = \sum_{i=1}^{n} X_{N_i}$, where $X_{N_1}, \ldots, X_{N_n}$ correspond to the sampled units.

Of course, X has the hypergeometric distribution

$$P(X = x) = \frac{C_D^x C_{N-D}^{n-x}}{C_N^n},\ 0 \le x \le D.$$

Two configurations can be thought of: (a) n is fixed, and $D/N \to p$, $0 < p < 1$ with $N \to \infty$. In this case, by applying Stirlings approximation to $N!$ and $D!$, $P(X = x) \to C_n^x p^x (1-p)^x$, and so $X \xrightarrow{d} \text{Bin}(n, p)$; (b) $n, N, N - n \to \infty$, $D/N \to p$, $0 < p < 1$. This is the case where convergence of X to normality holds.

Here is a general result; again, see Lehmann (1999) for a proof.

Theorem 1.3.10 *For $N \ge 1$, let π_N be a finite population with numerical values $X_1, X_2, \ldots X_N$. Let $X_{N_1}, X_{N_2}, \ldots, X_{N_n}$ be the values of the units of a sample without replacement of size n. Let $\bar{X}_n = \sum_{i=1}^{n} X_{N_i}/n$ and $\bar{X}_N = \sum_{i=1}^{N} X_N/N$. Suppose $n, N - n \to \infty$, and*

(a)

$$\frac{\max_{1 \le i \le N}(X_i - \bar{X}_N)^2}{\sum_{i=1}^{N}(X_i - \bar{X}_N)^2} \to 0,$$

and $n/N \to 0 < \tau < 1$ as $N \to \infty$;

(b)

$$\frac{N \max_{1 \le i \le N}(X_i - \bar{X}_N)^2}{\sum_{i=1}^{N}(X_i - \bar{X}_N)^2} = O(1), \quad \text{as } N \to \infty.$$

Then,

$$\frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{\text{Var}(\bar{X}_n)}} \xrightarrow{d} N(0, 1).$$

Example 1.3.14 Suppose $X_{N_1}, \ldots, X_{N_n}$ is a sample without replacement from the set $\{1, 2, \ldots, N\}$, and let $\bar{X}_n = \sum_{i=1}^n X_{N_i}/n$. Then, by a direct calculation,

$$E(\bar{X}_n) = \frac{N+1}{2}, \ \mathrm{Var}(\bar{X}_n) = \frac{(N-n)(N+1)}{12n}.$$

Furthermore,

$$\frac{N \max_{1 \le i \le N}(X_i - \bar{X}_N)^2}{\sum_{i=1}^N (X_i - \bar{X}_n)^2} = \frac{3(N-1)}{N+1} = O(1).$$

Hence, by Theorem 1.3.10, $\frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{\mathrm{Var}\bar{X}_n}} \xrightarrow{d} N(0,1)$.

1.3.5 Accuracy of CLT

Suppose a sequence of CDFs $F_{X_n} \xrightarrow{d} F_X$ for some F_X. Such a weak convergence result is usually used to approximate the true value of $F_{X_n}(x)$ at some fixed n and x by $F_X(x)$. However, the weak convergence result by itself says absolutely nothing about the accuracy of approximating $F_{X_n}(x)$ by $F_X(x)$ for that particular value of n. To approximate $F_{X_n}(x)$ by $F_X(x)$ for a given finite n is a leap of faith unless we have some idea of the error committed; i.e., $|F_{X_n}(x) - F_X(x)|$. More specifically, if for a sequence of random variables $X_1, \ldots, X_n$

$$\frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{\mathrm{Var}(\bar{X}_n)}} \xrightarrow{d} Z \sim N(0,1),$$

then we need some idea of the error

$$\left| P\left(\frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{\mathrm{Var}(\bar{X}_n)}} \le x \right) - \Phi(x) \right|.$$

in order to use the central limit theorem for a practical approximation with some degree of confidence. The first result for the iid case in this direction is the classic Berry-Esseen theorem. Typically, these accuracy measures give bounds on the error in the appropriate CLT for any fixed n, making assumptions about moments of X_i.

In the canonical iid case with a finite variance, the CLT says that $\sqrt{n}(\bar{X} - \mu)/\sigma$ converges in law to the $N(0,1)$. By Polya's theorem, the uniform error $\Delta_n = \sup_{-\infty < x < \infty} |P(\sqrt{n}(\bar{X} - \mu)/\sigma \le x) - \Phi(x)| \to 0$ as $n \to \infty$. Bounds on Δ_n for any given n are called uniform bounds.

The following results are the classic Berry-Esseen uniform bound and an extension of the Berry-Esseen inequality to the case of independent but not iid variables.; a proof can be seen in Petrov (1975). Introducing higher-order moment assumptions (third), the Berry-Esseen inequality assert for this convergence the rate $O(n^{-1/2})$.

Theorem 1.3.11 *(i)* **(Berry-Esseen; iid case)** *Let $X_1, \ldots, X_n$ be iid with $E(X_1) = \mu$, $\mathrm{Var}(X_1) = \sigma^2$, and $\beta_3 = E|X_1 - \mu|^3 < \infty$. Then there exists a universal constant C, not depending on n or the distribution of the X_i, such that*

$$\sup_x \left| P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq x \right) - \Phi(x) \right| \leq \frac{C\beta_3}{\sigma^3 \sqrt{n}}.$$

(ii) **(independent but not iid case)** *Let $X_1, \ldots, X_n$ be independent with $E(X_i) = \mu_i$, $\mathrm{Var}(X_i) = \sigma_i^2$, and $\beta_{3i} = E|X_i - \mu_i|^3 < \infty$. Then there exists a universal constant C^*, not depending on n or the distribution of the X_i, such that*

$$\sup_x \left| P\left(\frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{\mathrm{Var}(\bar{X}_n)}} \leq x \right) - \Phi(x) \right| \leq \frac{C^* \sum_{i=1}^n \beta_{3i}}{(\sum_{i=1}^n \sigma_i^2)^{3/2}}.$$

It is the best possible rate in the sense of not being subject to improvement without narrowing the class of distribution functions considered. For some specific underlying CDFs F_X, better rates of convergence in the CLT may be possible. This issue will be clearer when we discuss asymptotic expansions for $P(\sqrt{n}(\bar{X}_n - \mu)/\sigma \leq x)$. In Theorem 1.3.11-(i), the universal constant C may be taken as $C = 0.8$.

Example 1.3.15 The Berry-Esseen bound is uniform in x, and it is valid for any $n \geq 1$. While these are positive features of the theorem, it may not be possible to establish that $\Delta_n \leq \epsilon$ for some preassigned $\epsilon > 0$ by using the Berry-Esseen theorem unless n is very large. Let us see an illustrative example. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{BIN}(p, 1)$ and $n = 100$. Suppose we want the CLT approximation to be accurate to within an error of $\Delta_n = 0.005$. In the Bernoulli case, $\beta_3 = pq(1 - 2pq)$, where $q = 1 - p$. Using $C = 0.8$, the uniform Berry-Esseen bound is

$$\Delta_n \leq \frac{0.8pq(1 - 2pq)}{(pq)^{3/2}\sqrt{n}}.$$

This is less than the prescribed $\Delta_n = 0.005$ iff $pq > 0.4784$, which does not hold for any $0 < p < 1$. Even for $p = 0.5$, the bound is less than or equal to $\Delta_n = 0.005$ only when $n > 25,000$, which is a very large sample size. Of course, this is not necessarily a flaw of the Berry-Esseen inequality itself because the desire to have a uniform error of at most $\Delta_n = 0.005$ is a tough demand, and a fairly large value of n is probably needed to have such a small error in the CLT.

Example 1.3.16 As an example of independent variables that are not iid, consider $X_i \sim \text{BIN}(i^{-1}, 1), i \geq 1$, and let $S_n = \sum_{i=1}^n X_i$. Then, $E(S_n) = \sum_{i=1}^n i^{-1}$, $\text{Var}(S_n) = \sum_{i=1}^n (i-1)/i^2$ and $\beta_{3i} = (i-1)(i^2 - 2i + 2)/i^4$. Therefore, from Theorem 1.3.11-(ii),

$$\Delta_n \leq C^* \frac{\sum_{i=1}^n (i-1)(i^2 - 2i + 2)/i^4}{\sum_{i=1}^n [(i-1)/i^2]^{3/2}}$$

Observe now $\sum_{i=1}^n (i-1)/i^2 = \log n + O(1)$ and $\sum_{i=1}^n (i-1)(i^2 - 2i + 2)/i^4 = \log n + O(1)$. Substituting these back into the Berry-Esseen bound, one obtains with some minor algebra that $\Delta_n = O(\log n)^{-1/2}$.

For x sufficiently large, while n remains fixed, the quantities $F_{X_n}(x)$ and $F_X(x)$each become so close to 1 that the bound given in Theorem 1.3.11 is too rude. There has been a parallel development on developing bounds on the error in the CLT at a particular x as opposed to bounds on the uniform error. Such bounds are called local Berry-Esseen bounds. Many different types of local bounds are available.We present here just one.

Theorem 1.3.12 *Let $X_1, \ldots, X_n$ be independent with $E(X_i) = \mu_i$, $\text{Var}(X_i) = \sigma_i^2$, and $E|X_i - \mu_i|^{2+\delta} < \infty$ for some $0 < \delta \leq 1$. Then*

$$\left| P\left(\frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{\text{Var}(\bar{X}_n)}} \leq x \right) - \Phi(x) \right| \leq \frac{D}{1 + |x|^{2+\delta}} \frac{\sum_{i=1}^n E|X_i - \mu_i|^{2+\delta}}{(\sum_{i=1}^n \sigma_i^2)^{1+\frac{\delta}{2}}}.$$

for some universal constant $0 < D < \infty$.

Such local bounds are useful in proving convergence of global error criteria such as $\int |F_{X_n}(x) - \Phi(x)|^p dx$ or for establishing approximations to the moments of F_{X_n}. Uniform error bounds would be useless for these purposes. If the third absolute moments are finite,

an explicit value for the universal constant D can be chosen to be 31. Good reference for local bounds is Serfling (1980).

Error bounds for normal approximations to many other types of statistics besides sample means are known, such as the result for statistics that are smooth functions of means. The order of the error depends on the conditions one assumes on the nature of the function. We will discuss this problem in Section 2 after we introduced the Delta method.

1.3.6 Edgeworth and Cornish-Fisher expansions

We now consider the important topic of writing asymptotic expansions for the CDFs of centered and normalized statistics. When the statistic is a sample mean, let $Z_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$ and $F_{Z_n}(x) = P(Z_n \leq x)$, where $X_1, \ldots, X_n$ are i.i.d with a CDF F having mean μ and variance $\sigma^2 < \infty$.

The CLT says that $F_{Z_n}(x) \to \Phi(x)$ for every x, and the Berry-Esseen theorem says $|F_{Z_n}(x) - \Phi(x)| = O(n^{-1/2})$ uniformly in x if X has three moments. If we change the approximation $\Phi(x)$ to $\Phi(x) + C_1(F)p_1(x)\phi(x)/\sqrt{n}$ for some suitable constant $C_1(F)$ and a suitable polynomial $p_1(x)$, we can assert that

$$|F_n(x) - \Phi(x) - \frac{C_1(F)p_1(x)\phi(x)}{\sqrt{n}}| = O(n^{-1}),$$

uniformly in x. Expansions of the form

$$F_n(x) = \Phi(x) + \sum_{s=1}^{k} \frac{q_s(x)}{\sqrt{n}^s} + o(n^{-k/2}) \text{ uniformly in } x,$$

are known as Edgeworth expansions for Z_n. One needs some conditions on F and enough moments of X to carry the expansion to k terms for a given k. Excellent references for the main results on Edgeworth expansions are Hall (1992). The coefficients in the Edgeworth expansion for means depend on the cumulants of F, which share a functional relationship with the sequence of moments of F. Cumulants are also useful in many other contexts, for example, the saddlepoint approximation.

We start with the definition and recursive representations of the sequence of cumulants of a distribution. The term cumulant was coined by Fisher (1931).

Definition 1.3.4 *Let $X \sim F$ have a finite m.g.f. $\psi_n(t)$ in some neighborhood of zero, and let $K(t) = \log \psi_n(t)$ when it exists. The rth cumulant of X (or of F) is defined as $\kappa_r = \frac{d^r}{dt^r} K(t)_{|t=0}$.*

Equivalently, the cumulants of X are the coefficients in the power series expansion $K(t) = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!}$ within the radius of convergence of $K(t)$. By equating coefficients in $e^{K(t)}$ with those in $\psi(t)$, it is easy to express the first few moments (and therefore the first few central moments) in terms of the cumulants. Indeed, letting $c_i = E(X^i)$, $\mu_i = E(X-\mu)^i$, one obtains the expressions

$$c_1 = \kappa_1,\ c_2 = \kappa_2 + \kappa_1^2,\ c_3 = \kappa_3 + 3\kappa_1\kappa_2 + \kappa_1^3,\ c_4 = \kappa_4 + 4\kappa_1\kappa_3 + 3\kappa_2^2 + 6\kappa_1^2\kappa_2 + \kappa_1^4$$

$$\mu_2 = \sigma^2 = \kappa_2,\ \mu_3 = \kappa_3,\ \mu_4 = \kappa_4 + 3\kappa_2^2.$$

In general, the cumulants satisfy the recursion relations

$$\kappa_n = c_n - \sum_{j=1}^{n-1} C_{n-1}^{j-1} c_{n-j}\kappa_j,$$

which results in

$$\kappa_1 = \mu,\ \kappa_2 = \sigma^2,\ \kappa_3 = \mu_3,\ \kappa_4 = \mu_4 - 3\mu_2^2.$$

The higher-order ones are quite complex but can be found from Kendall's *Advanced Theory of Statistics*.

Example 1.3.17 Suppose $X \sim N(\mu, \sigma^2)$. Of course, $\kappa_1 = \mu$, $\kappa_2 = \sigma^2$. Since $K(t) = t\mu + t^2\sigma^2/2$, a quadratic, all derivatives of $K(t)$ of order higher than 2 vanish. Consequently, $\kappa_r = 0$ for $r > 2$. If $X \sim \text{Poisson}(\lambda)$, then $K(t) = \lambda(e^t - 1)$, and therefore all derivatives of $K(t)$ are equal to λe^t. It follows that $\kappa_r = \lambda$ for $r \geq 1$. These are two interesting special cases with neat structure and have served as the basis for stochastic modeling.

Now let us consider the expansion for (function of) means. To illustrate the idea, let us consider Z_n. Assume that the m.g.f of $W = (X_1 - \mu)/\sigma$ is finite and positive in a neighborhood of 0. The m.g.f of Z_n is equal to

$$\psi_n(t) = \left[\exp\{K(t/\sqrt{n})\}\right]^n = \exp\left\{\frac{t^2}{2} + \sum_{j=3}^{\infty} \frac{\kappa_j t^j}{j! n^{(j-2)/2}}\right\},$$

where $K(t)$ is the cumulant generating function of W and κ_j's are the corresponding cumulants ($\kappa_1 = 0, \kappa_2 = 1, \kappa_3 = EW^3$ and $\kappa_4 = EW^4 - 3$). Using the series expansion for $e^{t^2/2}$, we obtain that

$$\psi_n(t) = e^{t^2/2} + n^{-1/2} r_1(t) e^{t^2/2} + \cdots + n^{-j/2} r_j(t) e^{t^2/2} + \cdots , \tag{1.5}$$

where r_j is a polynomial of degrees $3j$ depending on $\kappa_3, \ldots, \kappa_{j+2}$ but not on n, $j = 1, 2, \ldots$. For example, it can be shown that

$$r_1(t) = \frac{1}{6}\kappa_3 t^3, \; r_2(t) = \frac{1}{24}\kappa_4 t^4 + \frac{1}{72}\kappa_3^2 t^6.$$

Since $\psi_n(t) = \int e^{tx} dF_{Z_n}(x)$ and $e^{t^2/2} = \int e^{tx} d\Phi(x)$, expansions (1.5) suggests the inverse expansion

$$F_{Z_n}(x) = \Phi(x) + n^{-1/2} R_1(x) + \cdots + n^{-j/2} R_j(x) + \cdots ,$$

where $R_j(x)$ is a function satisfying $\int e^{tx} dR_j(x) = r_j(t) e^{t^2/2}, j = 1, 2, \ldots$. Thus, R_j's can be obtained once r_j's are derived. For example,

$$\begin{aligned} R_1(x) &= -\frac{1}{6}\kappa_3 (x^2 - 1)\phi(x) \\ R_2(x) &= -\left[\frac{1}{24}\kappa_4 (x^2 - 3) + \frac{1}{72}\kappa_3^2 x (x^4 - 10x^2 + 15)\right] \phi(x) \end{aligned}$$

The CLT for means fails to capture possible skewness in the distribution of the mean for a given finite n because all normal distributions are symmetric. By expanding the CDF to the next term, the skewness can be captured. Expansion to another term also adjusts for the kurtosis. Although expansions to any number of terms are available under existence of enough moments, usually an expansion to two terms after the leading term is of the most practical importance. Indeed, expansions to three terms or more can be unstable due to the presence of the polynomials in the expansions. We present the two-term expansion next. A rigorous statement of the Edgeworth expansion for a more general Z_n will be introduced in the next chapter after entailing the multivariate Delta theorem. The proof can be found in Hall (1992).

Theorem 1.3.13 (Two-term Edgeworth expansion) *Suppose F is absolutely continuous distributions and $E_F(X^4) < \infty$. Then*

$$F_{Z_n}(x) = \Phi(x) + \frac{C_1(F) p_1(x) \phi(x)}{\sqrt{n}} + \frac{C_2(F) p_2(x) + C_3(F) p_3(x)}{n} + O(n^{-3/2}),$$

uniformly in x, where

$$C_1(F) = \frac{E(X-\mu)^3}{6\sigma^3},\ C_2(F) = \frac{\frac{E(X-\mu)^4}{\sigma^4} - 3}{24},\ C_3(F) = \frac{C_1^2(F)}{72},$$
$$p_1(x) = 1 - x^2,\ p_2(x) = 3x - x^3,\ p_3(x) = 10x^3 - 15x - x^5.$$

Note that the terms $C_1(F)$ and $C_2(F)$ can be viewed as skewness and kurtosis correction of departure from normality for $F_{Z_n}(x)$, respectively. It is useful to mention here that the corresponding formal two-term expansion for the density of Z_n is given by

$$\phi(z)+n^{-1/2}C_1(F)(z^3-3z)\phi(z)+n^{-1}[C_3(F)(z^6-15z^4+45z^2-15)+C_2(F)(z^4-6z^2+3)]\phi(z).$$

One of the uses of an Edgeworth expansion in statistics is approximation of the power of a test. In the one-parameter regular exponential family, the natural sufficient statistic is a sample mean, and standard tests are based on this statistic. So the Edgeworth expansion for sample means of iid random variables can be used to approximate the power of such tests. Here is an example.

Example 1.3.18 Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$ and we wish to test $H_0 : \lambda = 1$ vs. $H_1 : \lambda > 1$. The UMP test rejects H_0 for large values of $\sum_{i=1}^n X_i$. If the cutoff value is found by using the CLT, then the test rejects H_0 for $\bar{X}_n > 1 + k/\sqrt{n}$, where $k = z_\alpha$. The power at an alternative λ equals

$$\begin{aligned}\text{Power} &= P_\lambda\left(\bar{X}_n > 1 + k/\sqrt{n}\right) = P_\lambda\left(\frac{\bar{X}_n - \lambda}{\lambda/\sqrt{n}} > \frac{1 + k/\sqrt{n} - \lambda}{\lambda/\sqrt{n}}\right) \\ &= 1 - P_\lambda\left(\frac{\bar{X}_n - \lambda}{\lambda/\sqrt{n}} \leq \frac{\sqrt{n}(1-\lambda)}{\lambda} + \frac{k}{\lambda}\right) \to 1.\end{aligned}$$

For a more useful approximation, the Edgeworth expansion is used. For example, the general one-term Edgeworth expansion for sample means

$$F_n(x) = \Phi(x) + \frac{C_1(F)(1-x^2)\phi(x)}{\sqrt{n}} + O(n^{-1}),$$

can be used to approximate the power expression above. Algebra reduces the one-term Edgeworth expression to the formal approximation

$$\text{Power} \approx \Phi\left(\frac{\sqrt{n}(\lambda-1)-k}{\lambda}\right) + \frac{1}{3\sqrt{n}}\left[\frac{(\sqrt{n}(\lambda-1)-k)^2}{\lambda^2} - 1\right]\phi\left(\frac{\sqrt{n}(\lambda-1)-k}{\lambda}\right).$$

This is a much more useful approximation than simply saying that for large n the power is close to 1.

For constructing asymptotically correct confidence intervals for a parameter on the basis of an asymptotically normal statistic, the first-order approximation to the quantiles of the statistic (suitably centered and normalized) comes from using the central limit theorem. Just as Edgeworth expansions produce more accurate expansions for the CDF of the statistic than does just the central limit theorem, higher-order expansions for the quantiles produce more accurate approximations than does just the normal quantile. These higher-order expansions for quantiles are essentially obtained from recursively inverted Edgeworth expansions, starting with the normal quantile as the initial approximation. They are called *Cornish-Fisher expansions*. We briefly present the case of sample means. Let the standardized cumulants are the quantities $\rho_r = \kappa_r/\sigma^r$.

Theorem 1.3.14 *Let $X_1, \ldots, X_n$ be i.i.d with absolutely continuous CDF F having a finite m.g.f in some open neighborhood of zero. Let $Z_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$ and $H_n(x) = P_F(Z_n \leq x)$. Then,*

$$H_n^{-1}(\alpha) = z_\alpha + \frac{(z_\alpha^2 - 1)\rho_3}{6\sqrt{n}} + \frac{(z_\alpha^3 - 3z_\alpha)\rho_4}{24n} - \frac{(2z_\alpha^3 - 5z_\alpha)\rho_3^2}{36n} + O(n^{-3/2}).$$

Using Taylor's expansions at z_α for $\Phi(w_{n\alpha})$, $p_1(w_{n\alpha})\phi(w_{n\alpha})$ and $p_2(w_{n\alpha})\phi(w_{n\alpha})$, and the fact that $\phi'(x) = -x\phi(x)$, we can obtain this theorem by inverting the Edgeworth expansion.

Example 1.3.19 Let $W_n \sim \chi_n^2$ and $Z_n = (W_n - n)/\sqrt{2n} \xrightarrow{d} N(0,1)$ as $n \to \infty$, so a first-order approximation to the upper αth quantile of W_n is just $n + z_\alpha\sqrt{2n}$. The Cornish-Fisher expansion should produce a more accurate approximation. To verify this, we will need the standardized cumulants, which are $\rho_3 = 2\sqrt{2}$ and $\rho_4 = 12$. Now substituting into the theorem above, we get the two-term Cornish-Fisher expansion $\chi_{n,\alpha}^2 = n + z_\alpha\sqrt{2n} + \frac{2}{3}(z_\alpha^2 - 1) + \frac{z_\alpha^3 - 7z_\alpha}{9\sqrt{2n}}$.

1.3.7 The law of the iterated logarithm

The law of the iterated logarithm (LIL) complements the CLT by describing the precise extremes of the fluctuations of the sequence of random variables

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

The CLT states that this sequence converges in law to $N(0,1)$, but does not otherwise provide information about the fluctuations of these random variables about the expected value 0. The LIL asserts that the extremes fluctuations of this sequence are essentially of the exact order of magnitude $(2 \log \log n)^{1/2}$. The classical iid case is covered by

Theorem 1.3.15 (Hartman and Wintner). *let $\{X_i\}$ be iid with mean μ and finite variance σ^2. Then*

$$\limsup_{n \to \infty} \frac{\sum_{i=1}^n (X_i - \mu)}{(2\sigma^2 n \log \log n)^{1/2}} = 1 \text{ wp1};$$
$$\liminf_{n \to \infty} \frac{\sum_{i=1}^n (X_i - \mu)}{(2\sigma^2 n \log \log n)^{1/2}} = -1 \text{ wp1}.$$

In other words: with probability 1, for any $\epsilon > 0$, only finitely many of the events

$$\frac{\sum_{i=1}^n (X_i - \mu)}{(2\sigma^2 n \log \log n)^{1/2}} > 1 + \epsilon, \quad n = 1, 2, \ldots;$$
$$\frac{\sum_{i=1}^n (X_i - \mu)}{(2\sigma^2 n \log \log n)^{1/2}} > -1 - \epsilon, \quad n = 1, 2, \ldots$$

are realized, whereas infinitely many of the events

$$\frac{\sum_{i=1}^n (X_i - \mu)}{(2\sigma^2 n \log \log n)^{1/2}} > 1 - \epsilon, \quad n = 1, 2, \ldots;$$
$$\frac{\sum_{i=1}^n (X_i - \mu)}{(2\sigma^2 n \log \log n)^{1/2}} > -1 + \epsilon, \quad n = 1, 2, \ldots,$$

occur. That is, with probability 1, for any $\epsilon > 0$, all but finitely many of these fluctuations fall within the boundaries $\pm(1+\epsilon)(2 \log \log n)^{1/2}$ and moreover, the boundaries $\pm(1-\epsilon)(2 \log \log n)^{1/2}$ are reached infinitely often.

In LIL theorem, what is going on is that, for a given n, there is some collection of sample points ω for which the partial sum $S_n - n\mu$ stays in a specific $\sqrt{n}$-neighborhood of zero.

But this collection keeps changing with changing n, and any particular ω is sometimes in the collection and at other times out of it. Such unlucky values of n are unbounded, giving rise to the LIL phenomenon. The exact rate $\sqrt{n \log\log n}$ is a technical aspect and cannot be explained intuitively.

The LIL also complements-indeed, refines- the SLLN (assuming existence of 2nd moments). It terms of the average dealt with the SLLN, $\frac{1}{n}\sum_{i=1}^{n} X_i - \mu$, the LIL assert that the extreme fluctuations are essentially of the exact order of magnitude

$$\frac{\sigma(2\log\log n)^{1/2}}{n^{1/2}}.$$

Thus, with provability 1, for any $\epsilon > 0$, the infinite sequence of "confidence intervals"

$$\left\{ \frac{1}{n}\sum_{i=1}^{n} X_i \pm (1+\epsilon)\frac{\sigma(2\log\log n)^{1/2}}{n^{1/2}} \right\}$$

contains μ with only finitely many exceptions. Say, in this asymptotic fashion, the LIL provides the basis for concepts of 100% confidence intervals. The LIL also provides an example of almost sure convergence being truly stronger than convergence in probability.

Example 1.3.20 Let $X_1, X_2, \ldots$ be iid with a finite variance. Then,

$$\frac{S_n - n\mu}{\sqrt{2n\log\log n}} = \frac{S_n - n\mu}{\sqrt{n}} \frac{1}{\sqrt{2\log\log n}} = O_p(1) \cdot o(1) = o_p(1).$$

But, by the LIL, $\frac{S_n - n\mu}{\sqrt{2n\log\log n}}$ does not converge a.s. to zero. Hence, convergence in probability is weaker than almost sure convergence, in general.

References

Billingsley, P. (1995). *Probability and Measure*, 3rd edition, John Wiley, New York.

Petrov, V. (1975). *Limit Theorems for Sums of Independent Random Variables* (translation from Russian), Springer-Verlag, New York.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*, John Wiley, New York.

Shao, J. (2003). *Mathematical Statistics*, 2nd ed. Springer, New York.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, Cambridge University Press.

Chapter 2

Transformations of given statistics: The delta method

Distributions of transformations of a statistic are of importance in applications. Suppose an estimator T_n for a parameter θ is available, but the quantity of interest is $g(\theta)$ for some known function g. A natural estimator is $g(T_n)$. The aim is to deduce the asymptotic behavior of $g(T_n)$ based on those of T_n.

A first result is an immediate consequence of the continuous-mapping theorem. Of greater interest is a similar question concerning limit distributions. In particular, if $\sqrt{n}(T_n - \theta)$ converges in law to a limit distribution, is the same true for $\sqrt{n}\,[g(T_n) - g(\theta)]$? If g is differentiable, then the answer is affirmative.

2.1 Basic result

The delta theorem says how to approximate the distribution of a transformation of a statistic in large samples if we can approximate the distribution of the statistic itself. We firstly treat the univariate case and present the basic delta theorem as follows.

Theorem 2.1.1 (Delta Theorem) *Let T_n be a sequence of statistics such that*

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2(\theta)). \tag{2.1}$$

Let $g : \mathbb{R} \to \mathbb{R}$ be once differentiable at θ with $g'(\theta) \neq 0$. Then

$$\sqrt{n}\,[g(T_n) - g(\theta)] \xrightarrow{d} N(0, [g'(\theta)]^2\sigma^2(\theta)).$$

Proof. First note that it follows from the assumed CLT for T_n that T_n converges in probability to θ and hence $T_n - \theta = o_p(1)$. The proof of the theorem now follows from a simple application of Taylor's theorem that says that

$$g(x_0 + h) = g(x_0) + hg'(x_0) + o(h)$$

if g is differentiable at x_0. Therefore

$$g(T_n) = g(\theta) + (T_n - \theta)g'(\theta) + o_p(T_n - \theta).$$

That the remainder term is $o_p(T_n - \theta)$ follows from our observation that $T_n - \theta = o_p(1)$ and Lemma 1.2.1. Taking $g(\theta)$ to the left and multiplying both sides by $\sqrt{n}$, we obtain

$$\sqrt{n}\,[g(T_n) - g(\theta)] = \sqrt{n}(T_n - \theta)g'(\theta) + \sqrt{n}o_p(T_n - \theta).$$

Observing that $\sqrt{n}(T_n - \theta) = O_p(1)$ by the assumption of the theorem, we see that the last term on the right-hand side is$\sqrt{n}o_p(T_n - \theta) = o_p(1)$. Hence, an application of Slutskys theorem to the above gives $\sqrt{n}\,[g(T_n) - g(\theta)] \xrightarrow{d} N(0, [g'(\theta)]^2\sigma^2(\theta))$. □

Remark 2.1.1 Assume that g is differentiable in a neighborhood of θ, and $g'(x)$ is continuous at θ. Further, if $\sigma(\theta)$ is a continuous function of θ, then we have the modified conclusion by replacing $g'(\theta)$ and $\sigma(\theta)$ with $g'(T_n)$ and $\sigma(T_n)$,

$$\frac{\sqrt{n}\,[g(T_n) - g(\theta)]}{\{[g'(T_n)]^2\sigma^2(T_n)\}^{1/2}} \xrightarrow{d} N(0, 1).$$

Remark 2.1.2 In fact, the Delta Theorem does not require the asymptotic distribution of T_n to be normal. By the foregoing proofs, we see that assuming $a_n(T_n - \theta) \xrightarrow{d} Y$ in which a_n is a sequence of positive numbers with $\lim_{n\to\infty} a_n = \infty$ and the conditions in the Delta Theorem hold, we have

$$a_n[g(T_n) - g(\theta)] \xrightarrow{d} [g'(\theta)]Y.$$

Example 2.1.1 Suppose $X_1, \ldots, X_n$ are iid with mean μ and variance σ^2. By taking $T_n = \bar{X}_n$, $\theta = \mu$, $\sigma^2(\theta) = \sigma^2$, and $g(x) = x^2$, one gets for $\mu \neq 0$

$$\sqrt{n}(\bar{X}_n^2 - \mu^2) \xrightarrow{d} N(0, 4\mu^2\sigma^2).$$

For $\mu = 0$, $n\bar{X}_n^2/\sigma^2 \xrightarrow{d} \chi_1^2$ by continuous mapping theorem.

Example 2.1.2 For estimating p^2, suppose that we have the choice between (a) $X \sim \text{Bin}(n, p^2)$; (b) $Y \sim \text{Bin}(n, p)$ and that as estimators of p^2 in the two cases, we would use respectively X/n and $(Y/n)^2$. Then we have

$$\sqrt{n}\left(\frac{X}{n} - p^2\right) \xrightarrow{d} N(0, p^2(1-p^2));$$
$$\sqrt{n}\left(\left(\frac{Y}{n}\right)^2 - p^2\right) \xrightarrow{d} N(0, pq \cdot 4p^2).$$

At least for large n, X/n will thus be more accurate than $(Y/n)^2$, provided

$$p^2(1-p^2) < pq \cdot 4p^2,$$

say X/n or Y^2/n^2 is preferable as $p > 1/3$ or $p < 1/3$.

Let us finally consider an example in which $g'(\cdot)$ does not exist.

Example 2.1.3 Suppose T_n is a sequence of statistics satisfying (2.1) and that we are interested in the limiting behavior of $|T_n|$. Since $g(\theta) = |\theta|$ is differentiable with derivative $g'(\theta) = \pm 1$ at all values of $\theta \neq 0$, it follows from Theorem 2.1.1 that

$$\sqrt{n}(|T_n| - |\theta|) \xrightarrow{d} N(0, \sigma^2) \text{ for all } \theta \neq 0.$$

When $\theta = 0$, Theorem 2.1.1 does not apply, but it is easy to determine the limit behavior of $|T_n|$ directly. With $|T_n| - |\theta| = |T_n|$, we have

$$\begin{aligned} P(\sqrt{n}|T_n| < a) &= P(-a < \sqrt{n}T_n < a) \\ &\to \Phi\left(\frac{a}{\sigma}\right) - \Phi\left(-\frac{a}{\sigma}\right) = P(\sigma\chi_1 < a), \end{aligned}$$

where $\chi_1 = \sqrt{\chi_1^2}$ is the distribution of the absolute value of a standard normal variable. The convergence rate of $|T_n|$ therefore continues to be $1/\sqrt{n}$, but the form of the limit distribution is χ_1 rather than normal.

2.2 Higher-order expansions

There are instances in which $g'(\theta) = 0$ (at least for some special value of θ), in which case the limiting distribution of $g(T_n)$ is determined by the third term in the Taylor expansion. Thus, if $g'(\theta) = 0$, then

$$g(T_n) = g(\theta) + \frac{(T_n - \theta)^2}{2} g''(\theta) + o_p\left((T_n - \theta)^2\right) \tag{2.2}$$

and hence

$$n(g(T_n) - g(\theta)) = n\frac{(T_n - \theta)^2}{2} g''(\theta) + o_p(1) \xrightarrow{d} \frac{g''(\theta)\sigma^2(\theta)}{2}\chi_1^2.$$

Formally, the following result generalizes Theorem 2.1.1 to include this case.

Theorem 2.2.1 *Let T_n be a sequence of statistics such that*

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2(\theta)).$$

Let g be a real-valued function differentiable $k(\geq 1)$ at θ with $g^{(k)}(\theta) \neq 0$ but $g^{(j)}(\theta) = 0$ for $j < k$. Then

$$(\sqrt{n})^k \left[g(T_n) - g(\theta)\right] \xrightarrow{d} \frac{1}{k!}[g^{(k)}(\theta)][N(0, \sigma^2(\theta))]^k.$$

Proof. The argument is similar to that for Theorem 2.1.1, this time using the higher-order Taylor expansions as in (2.2). The remaining details are left as an exercise. □

Example 2.2.1 (i) Example 2.1.1 revisited. For $\mu = 0$, $n\bar{X}_n^2/\sigma^2 \xrightarrow{d} \frac{1}{2} \cdot 2 \cdot [N(0,1)]^2 = \chi_1^2$; (ii) Suppose that $\sqrt{n}\bar{X}_n$ converges in law to a standard normal distribution. Now consider the limiting behavior of $\cos(\bar{X}_n)$. Because the derivative of $\cos(x)$ is zero at $x = 0$, the proof of Theorem 2.1.1 yields that $\sqrt{n}(\cos(\bar{X}_n) - 1)$ converges to zero in probability (or equivalently in law). Thus, it should be concluded that $\sqrt{n}$ is not the right norming rate for the random sequence $\cos(\bar{X}_n) - 1$. A more informative statement is that $-2n(\cos(\bar{X}_n) - 1)$ converges in law to χ_1^2.

2.3 Multivariate version of delta theorem

Next we state the multivariate delta theorem, which is similar to the univariate case.

Theorem 2.3.1 *Suppose* $\{\mathbf{T}_n\}$ *is a sequence of* k*-dimensional random vectors such that* $\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \xrightarrow{d} N_k(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$*. Let* $g : \mathbb{R}^k \to \mathbb{R}^m$ *be once differentiable at* $\boldsymbol{\theta}$ *with the gradient matrix* $\nabla g(\boldsymbol{\theta})$*. Then*

$$\sqrt{n}(g(\mathbf{T}_n) - g(\boldsymbol{\theta})) \xrightarrow{d} N_m\left(\mathbf{0}, \nabla^T g(\boldsymbol{\theta})\boldsymbol{\Sigma}(\boldsymbol{\theta})\nabla g(\boldsymbol{\theta})\right)$$

provided $\nabla^T g(\boldsymbol{\theta})\boldsymbol{\Sigma}(\boldsymbol{\theta})\nabla g(\boldsymbol{\theta})$ *is positive definite.*

Proof. This theorem can be proved by using the Cramer-Wold device. It suffices to show that for every $\mathbf{c} \in \mathbb{R}^m$, we have

$$\sqrt{n}\mathbf{c}^T(g(\mathbf{T}_n) - g(\boldsymbol{\theta})) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{c}^T\nabla^T g(\boldsymbol{\theta})\boldsymbol{\Sigma}(\boldsymbol{\theta})\nabla g(\boldsymbol{\theta})\mathbf{c}\right)$$

The first-order Taylor's expansion gives

$$g(\mathbf{T}_n) = g(\boldsymbol{\theta}) + \nabla^T g(\boldsymbol{\theta})(\mathbf{T}_n - \boldsymbol{\theta}) + o_p(||\mathbf{T}_n - \boldsymbol{\theta}||).$$

The remaining proofs are similar to the univariate case by the application of Corollary 1.2.1 and left to exercises. □

The multivariate delta theorem is useful in finding the limiting distribution of sample moments. We state next some examples most often used.

Example 2.3.1 (Sample variance revisited) Suppose $X_1, \ldots, X_n$ are iid with mean μ, variance σ^2 and $E(X_1^4) < \infty$. Then by taking

$$\mathbf{T}_n = (\bar{X}_n, \overline{X_n^2})^T, \quad \boldsymbol{\theta} = (EX_1, EX_1^2)^T, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_1^2) \\ \mathrm{Cov}(X_1^2, X_1) & \mathrm{Var}(X_1^2) \end{pmatrix}$$

and using the multivariate CLT theorem (Theorem 1.3.2), we have

$$\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \xrightarrow{d} N_2\left(\mathbf{0}, \boldsymbol{\Sigma}\right).$$

Taking the function $g(u,v) = v - u^2$ which is obviously differentiable at the point $\boldsymbol{\theta}$ with derivative $g'(u,v) = (-2u, 1)$, it follows that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 - \mathrm{Var}(X_1)\right) \xrightarrow{d} N_2\left(\mathbf{0}, (-2\mu, 1)\boldsymbol{\Sigma}(-2\mu, 1)^T\right).$$

Because the sample variance does not depend on location, we may as well assume $\mu = 0$ (or equivalently working with $X_i - \mu$). Thus, it is readily seen that

$$\sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{d} N(0, \mu_4 - \sigma^4),$$

where μ_4 denotes the centered fourth moment of X_1. If the parent distribution is normal, then $\mu_4 = 3\sigma^4$ and $\sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{d} N(0, 2\sigma^4)$. In view of Slutsky's Theorem, the same result is valid for the unbiased version $n/(n-1)S_n^2$ of the sample variance. From here, by another use of the univariate delta theorem, one sees that

$$\sqrt{n}(S_n - \sigma) \xrightarrow{d} N\left(0, \frac{\mu_4 - \sigma^4}{4\sigma^2}\right).$$

In the previous example the asymptotic distribution of $\sqrt{n}(S_n^2 - \sigma^2)$ was obtained by the delta method. Actually, it can also and more easily be derived by a direct application of CLT and Slutsky' theorem as we have illustrated in Example 1.3.2. Thus, it is not always a good idea to apply the general theorems. However, in many cases the delta method is a good way to package the mechanics of Taylor expansions in a transparent way. The followings are more examples.

Example 2.3.2 (**The joint limit distribution**) (i) Consider the joint limit distribution of the sample variance S_n^2 and the t-statistic $\bar{X}_n/S_n$. Again for the limit distribution it does not make a difference whether we use a factor n or $n-1$ to standardize S_n^2. For simplicity we use n. Then $(S_n^2, \bar{X}_n/S_n)$ can be written as $g(\bar{X}_n, \overline{X_n^2})$ for the map $g : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$g(u,v) = \left(v - u^2, \frac{u}{(v-u^2)^{1/2}}\right).$$

The joint limit distribution of $\sqrt{n}(\bar{X}_n - \alpha_1, \overline{X_n^2} - \alpha_2)$ is derived in the preceding example, where α_k denotes the kth moment of X_1. The function g is differentiable at $\boldsymbol{\theta} = (EX_1, EX_1^2)$

provided that σ^2 is positive, with derivative

$$[g'_{(\alpha_1,\alpha_2)}]^T = \begin{pmatrix} -2\alpha_1 & 1 \\ \frac{\alpha_1^2}{(\alpha_2-\alpha_1^2)^{3/2}} + \frac{1}{(\alpha_2-\alpha_1^2)^{1/2}} & \frac{-\alpha_1}{2(\alpha_2-\alpha_1^2)^{3/2}} \end{pmatrix}.$$

It follows that the sequence $\sqrt{n}(S_n^2 - \sigma^2, \bar{X}_n/S_n - \alpha_1/\sigma)$ is asymptotically bivariate normally distributed, with mean zero and covariance matrix,

$$[g'_{(\alpha_1,\alpha_2)}]^T \begin{pmatrix} \alpha_2 - \alpha_1^2 & \alpha_3 - \alpha_1\alpha_2 \\ \alpha_3 - \alpha_1\alpha_2 & \alpha_4 - \alpha_2^2 \end{pmatrix} g'_{(\alpha_1,\alpha_2)}.$$

It is easy but uninteresting to compute this explicitly; A direct application of this result is to analyze the so-called *effect size* $\theta = \mu/\sigma$. A natural estimator of θ is $\bar{X}_n/S_n$.

(ii) A more commonly seen case is to derive the joint limit distribution of $\bar{X}_n$ and S_n^2. Then, by using the multivariate delta theorem on some algebra,

$$\sqrt{n}\begin{pmatrix} \bar{X}_n - \mu \\ S_n^2 - \sigma^2 \end{pmatrix} \xrightarrow{d} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}\begin{pmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{pmatrix}\right)$$

Thus $\bar{X}_n$ and S_n^2 are asymptotically independent if the population skewness is 0 (i.e., μ_3=0).

2.4 Variance-stabilizing transformations

A principal use of parametric asymptotic theory is to construct asymptotically correct confidence intervals. More precisely, suppose $\widehat{\theta}$ is a reasonable estimate of some parameter θ. Suppose it is consistent and even asymptotically normal; i.e., $\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, \sigma^2(\theta))$ for some function $\sigma(\theta) > 0$. Then, a simple calculation shows that the confidence interval $\widehat{\theta} \pm \sigma(\widehat{\theta}) z_{\alpha/2}/\sqrt{n}$ is asymptotically correct; i.e., its limiting coverage is $1 - \alpha$ under every θ. A number of approximations have been made in using this interval. The exact distribution of $\widehat{\theta}$ has been replaced by a normal; the correct standard deviation has been replaced by another plug-in estimate $\sigma(\theta)$; and the true mean of $\widehat{\theta}$ has been replaced by θ. The plug-in standard deviation estimate is quite often an underestimate of the true standard deviation. And depending on the situation, $\widehat{\theta}$ may have a nontrivial bias as an estimate of θ. Interest

has centered on finding transformations, say $g(\widehat{\theta})$, that (i) have an asymptotic variance function free of θ, eliminating the annoying need to use a plugin estimate, (ii) have skewness≈ 0 in some precise sense, and (iii) have bias ≈ 0 as an estimate of $g(\theta)$, again in some precise sense.

Transformations of the first type are known as variance-stabilizing transformations (VSTs), those of the second type are known as symmetrizing transformations (STs), and those of the third type are known as bias-corrected transformations (BCTs). In this course, we only elaborate on the first one, i.e., VSTs since it is of greatest interest in practice and also a major use of the delta theorem. Unfortunately, the concept does not generalize to multiparameter cases, i.e., it is generally infeasible to find a dispersion-stabilizing transformation. It is, however, a useful tool in one-parameter problems.

Suppose T_n is sequence of statistics such that $\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2(\theta))$, $\sigma(\theta) > 0$. By the delta theorem, if $g(\cdot)$ is once differentiable at θ with $g'(\theta) \neq 0$, then

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{d} N(0, [g'(\theta)]^2\sigma^2(\theta)).$$

Therefore, if we want the variance in the asymptotic distribution of $g(T_n)$ to be constant, we set

$$[g'(\theta)]^2\sigma^2(\theta) = k^2.$$

for some constant k. Thus, a way of deriving $g(\cdot)$ from $\sigma(\cdot)$ is

$$g(\theta) = k \int \frac{1}{\sigma(\theta)} d\theta$$

if $\sigma(\theta)$ is continuous in θ. k can obviously be chosen as any nonzero real number. In the above, the integral is to be interpreted as a primitive. For such a $g(\cdot)$, $g(T_n)$ has an asymptotic distribution with a variance that is free of θ. Such a statistic or transformation of T_n is called a variance-stabilizing transformation. Note that the transformation is monotone. So, if we use $g(T_n)$ to make an inference for $g(\theta)$, then we can automatically retransform to make an inference for θ, which is the parameter of interest.

As long as there is an analytical formula for the asymptotic variance function in the limiting normal distribution for T_n, and as long as the reciprocal of its square root can be

integrated in closed form, a VST can be written down. Next, we work out some examples of VSTs and show how they are used to construct asymptotically correct confidence intervals for an original parameter of interest.

Example 2.4.1 Suppose $X_1, X_2, \ldots$, are iid Poisson(θ). Then $\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{d} N(0, \theta)$. Thus $\sigma(\theta) = \sqrt{\theta}$ and so a variance-stabilizing transformation is

$$g(\theta) = \int \frac{k}{\sqrt{\theta}} d\theta = 2k\sqrt{\theta}.$$

Taking $k = 1/2$ gives that $g(\theta) = \sqrt{\theta}$ is a variance-stabilizing transformation for the Poisson case. Indeed $\sqrt{n}(\sqrt{\bar{X}_n} - \sqrt{\theta}) \xrightarrow{d} N(0, 1/4)$. Thus, an asymptotically correct confidence interval for $\sqrt{\theta}$ is $\sqrt{\bar{X}_n} \pm \frac{z_\alpha}{2\sqrt{n}}$. This implies that an asymptotically correct confidence interval for θ is

$$\left\{ \left(\sqrt{\bar{X}_n} - \frac{z_\alpha}{2\sqrt{n}} \right)^2, \left(\sqrt{\bar{X}_n} + \frac{z_\alpha}{2\sqrt{n}} \right)^2 \right\}.$$

Of course, if $\sqrt{\bar{X}_n} - \frac{z_\alpha}{2\sqrt{n}} < 0$, that expression should be replaced by 0. This confidence interval is different from the more traditional interval, namely $\bar{X}_n \pm \frac{z_\alpha}{\sqrt{n}} \sqrt{\bar{X}_n}$, which goes by the name of the Wald interval. In fact, the actual coverage properties of the interval based on the VST are significantly better than those of the Wald interval.

Example 2.4.2 (**Sample correlation revisited**) Consider the same assumption in Example 1.2.11. Firstly, by using the multivariate delta theorem, we can derive the limiting distribution of the sample correlation coefficient r_n. By taking

$$\begin{aligned}
\mathbf{T}_n &= \left(\bar{X}_n, \bar{Y}_n, \frac{1}{n}\sum_{i=1}^n X_i^2, \frac{1}{n}\sum_{i=1}^n Y_i^2, \frac{1}{n}\sum_{i=1}^n X_i Y_i \right)^T, \\
\boldsymbol{\theta} &= (EX_1, EY_1, EX_1^2, EY_1^2, EX_1Y_1)^T, \\
\boldsymbol{\Sigma} &= \mathrm{Cov}(X_1, Y_1, X_1^2, Y_1^2, X_1Y_1),
\end{aligned}$$

and using the transformation $g(u_1, u_2, u_3, u_4, u_5) = (u_5 - u_1u_2)/\sqrt{(u_3 - u_1^2)(u_4 - u_2^2)}$, it follows that

$$\sqrt{n}(r_n - \rho) \xrightarrow{d} N(0, v^2)$$

for some $v > 0$, provided that the fourth moments of (X, Y) exist. It is not possible to write a clean formula for v in general. If (X_i, Y_i) are iid $N_2(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, then the calculation can be done in closed form and

$$\sqrt{n}(r_n - \rho) \xrightarrow{d} N(0, (1-\rho^2)^2).$$

However, it does not work well to base an asymptotic confidence interval directly on this result. The transformation

$$g(\rho) = \int \frac{1}{1-\rho^2} d\rho = \frac{1}{2} \log \frac{1+\rho}{1-\rho} = \operatorname{arctanh}(\rho)$$

is a VST for r_n. This is the famous arctanh transformation of Fisher, popularly known as *Fisher's z*. Thus, the sequence $\sqrt{n}(\operatorname{arctanh}(r_n) - \operatorname{arctanh}(\rho))$ converges in law to the $N(0,1)$ distribution. Confidence intervals for ρ are computed from the arctanh transformation as

$$\left(\tanh(\operatorname{arctanh}(r_n) - z_\alpha/\sqrt{n}), \tanh(\operatorname{arctanh}(r_n) + z_\alpha/\sqrt{n})\right).$$

rather than by using the asymptotic distribution of r_n itself. The arctanh transformation of r_n attains normality much quicker than r_n itself. (Interest students may run a small simulation to verify it by using R) .

2.5 Approximation of moments

The delta theorem is proved by an ordinary Taylor expansion of T_n around θ. The same method also produces approximations, with error bounds, on the moments of $g(T_n)$. The order of the error can be made smaller the more moments T_n has. To keep notation simple, we give approximations to the mean and variance of a function $g(T_n)$ below when T_n is a sample mean.

Before proceeding, we need to address the so-called moment convergence problem. Sometimes we need to establish that moments of some sequence $\{X_n\}$, or at least some lower-order moments, converge to moments of X when $X_n \xrightarrow{d} X$. Convergence in distribution by itself simply cannot ensure convergence of any moments. An extra condition that ensures convergence of appropriate moments is uniform integrability. However, direct verification of

its definition is usually cumbersome. Thus, here we choose to introduce some sufficient conditions which could ensure convergence of moments.

Theorem 2.5.1 *Suppose* $X_n \xrightarrow{d} X$ *for some* X. *If* $\sup_n E|X_n|^{k+\delta} < \infty$ *for some* $\delta > 0$, *then* $E(X_n^r) \to E(X^r)$ *for every* $1 \leq r \leq k$.

Another common question is the convergence of moments in the canonical CLT for iid random variables, which is stated in the following theorem.

Theorem 2.5.2 (von Bahr) *Suppose* $X_1, \ldots, X_n$ *are i.i.d. with mean* μ *and finite variance* σ^2, *suppose that, for some specific* k, $E|X_1|^k < \infty$. *Suppose* $Z \sim N(0,1)$. *Then,*

$$E\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}\right)^r = E(Z^r) + O(\frac{1}{\sqrt{n}}),$$

for every $r \leq k$.

By a similar arguments in the proof of Delta theorem, a direct application of this theorem is the following approximations to the mean and variance of a function $g(T_n)$.

Proposition 2.5.1 *Suppose* $X_1, X_2, \ldots$ *are iid observations with a finite fourth moment. Let* $E(X_1) = \mu$ *and* $\mathrm{Var}(X_1) = \sigma^2$. *Let* g *be a scalar function with four uniformly bounded derivatives. Then*

(i) $E(g(\bar{X}_n)) = g(\mu) + \frac{g''(\mu)\sigma^2}{2n} + O(n^{-2})$

(ii) $\mathrm{Var}(g(\bar{X}_n)) = \frac{(g'(\mu))^2\sigma^2}{n} + O(n^{-2})$.

The variance approximation above is simply what the delta theorem says. With more derivatives of g that are uniformly bounded, higher-order approximations can be given.

Example 2.5.1 Suppose $X_1, X_2, \ldots$ are iid $\mathrm{Poi}(\mu)$ and we wish to estimate $P(X_1 = 0) = e^{-\mu}$. The MLE is $e^{-\bar{X}_n}$, and suppose we want to find an approximation to the bias and variance of $e^{-\bar{X}_n}$. We apply Proposition 2.5.1 with the function $g(x) = e^{-x}$ so that $g'(x) =$

$-g''(x) = -e^{-x}$. Plugging into the proposition, we get the approximations $\text{Bias}(e^{-\bar{X}_n}) = \frac{\mu e^{-\mu}}{2n} + O(n^{-2})$, and $\text{Var}(e^{-\bar{X}_n}) = \frac{\mu e^{-2\mu}}{n} + O(n^{-2})$.

Note that it is in fact possible to derive exact expressions for the mean and variance of $e^{-\bar{X}_n}$ in this case, as $\sum_{i=1}^n X_i$ has a $\text{Poi}(n\mu)$ distribution and therefore its mgf (moment generating function) equals $\psi_n(t) = E(e^{t\bar{X}_n}) = (e^{\mu(e^{t/n}-1)})^n$. In particular, the mean of $e^{-\bar{X}_n}$ is $(e^{\mu(e^{-1/n}-1)})^n$. It is possible to recover the approximation for the bias given above from this exact expression. Indeed,

$$(e^{\mu(e^{-1/n}-1)})^n = e^{n\mu(e^{-1/n}-1)} = e^{n\mu(\sum_{k=1}^{\infty} \frac{(-1)^k}{k!n^k})} = e^{-\mu}(1 + \frac{\mu}{2n} + O(n^{-2}))$$

on collecting the terms of the exponentials together. On subtracting $e^{-\mu}$, this reproduces the bias approximation given above. The delta theorem produces it more easily than the direct calculation.

2.6 Multivariate-version Edgeworth expansion

In this section, we present the a more general result regarding Edgeworth expansions, which can be applied to many useful cases.

Theorem 2.6.1 (Edgeworth expansions) *Let m be a positive integer and $\mathbf{X}_1, \mathbf{X}_2, \ldots$ be i.i.d. random k-vectors having finite $m+2$ moments. Consider $W_n = \sqrt{n}h(\bar{\mathbf{X}}_n)/\sigma_h$, where $\bar{\mathbf{X}}_n = n^{-1}\sum_i \mathbf{X}_i$, h is a function being $m+2$ times continuous differentiable in a neighborhood of $\boldsymbol{\mu} = E\mathbf{X}_1$, $h(\boldsymbol{\mu}) = 0$, $\sigma_h^2 = [\nabla h(\boldsymbol{\mu})]^T \text{Var}(X_1) \nabla h(\boldsymbol{\mu}) > 0$. Assume the C.D.F. of $\mathbf{X}_1$ is absolutely continuous. Then F_{W_n} admits the Edgeworth expansion*

$$\sup_x \left| F_{W_n} - \Phi(x) - \sum_{j=1}^{m} \frac{p_j(x)\phi(x)}{n^{j/2}} \right| = o(\frac{1}{n^{m/2}}),$$

where $p_j(x)$ is a polynomial of degree at most $3j-1$, with coefficients depending on the first $m+2$ moments of X_1. In particular,

$$p_1(x) = -c_1\sigma_h^{-1} + \frac{1}{6}c_2\sigma_h^{-3}(x^2-1),$$

with $c_1 = \frac{1}{2}\sum_{i=1}^{k}\sum_{i=1}^{k} a_{ij}\mu_{ij}$ *and*

$$c_2 = \sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{l=1}^{k} a_i a_j a_l \mu_{ijl} + 3\sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{l=1}^{k}\sum_{q=1}^{k} a_i a_j a_{lq}\mu_{il}\mu_{jq},$$

where a_i *is the ith component of* $\nabla h(\boldsymbol{\mu})$, a_{ij} *is the (i,j)th element of the Hessian matrix* $\nabla^2 h(\boldsymbol{\mu})$, $\mu_{ij} = E(Y_i Y_j), \mu_{ijl} = E(Y_i Y_j Y_l)$, *and* Y_i *is the ith component of* $\mathbf{X}_1 - \boldsymbol{\mu}$.

Example 2.6.1 The t-test and the t confidence interval are among the most used tools of statistical methodology. As such, an Edgeworth expansion for the C.D.F. of the t-statistic for general populations is interesting and useful, and we can derive it according to Theorem 2.6.1. Consider the studentized random variable $W_n = \sqrt{n}(\bar{X}_n - \mu)/\widehat{\sigma}$, where $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. Assuming that $EX_1^{2m+4} < \infty$ and applying multivariate Delta theorem to random vectors $(X_i, X_i^2), i = 1, 2, \ldots,$ and $h(x, y) = (x - \mu)/\sqrt{y - x^2}$, we obtain the Edgeworth expansion with $\sigma_h = 1$

$$p_1(x) = \frac{1}{6}\kappa_3(2x^2 + 1).$$

Furthermore, it can be found in Hall (1992; p73) that

$$p_2(x) = \frac{1}{12}\kappa_4 x(x^2 - 3) - \frac{1}{18}\kappa_3^2 x(x^4 + 2x^2 - 3) - \frac{1}{4}x(x^2 + 3).$$

Chapter 3

The basic sample statistics

3.1 The sample distribution function

Consider $X_1, X_2 \ldots$ be iid with distribution function F. For each sample of size n, a corresponding *sample distribution function* F_n is constructed by placing at each observation X_i a mass $1/n$. Thus F_n can be represented as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq x\}},$$

which is always called *empirical cumulative distribution function; ECDF*. When $\mathbf{X}_i \in \mathbb{R}^p$, the inequality above is understood as componentwise version. For simplicity, here we only consider the case $d = 1$. The F_n can and does play a fundamental role in statistical inference. In this subsection, we discuss several aspects of the properties and applications of F_n.

3.1.1 Basic properties

The simplest aspect of F_n is that, for each fixed x, $F_n(x)$ serves as an estimator of $F(x)$.

Proposition 3.1.1 *For fixed x, $x \in (-\infty, \infty)$,*

(i) $F_n(x)$ *is unbiased and has variance*

$$\mathrm{Var}[F_n(x)] = \frac{F(x)[1-F(x)]}{n};$$

(ii) $F_n(x)$ *is consistent in mean square, i.e.,* $F_n(x) \overset{2nd}{\to} F(x)$*;*

(iii) $F_n(x) \overset{wp1}{\to} F(x)$*;*

(iv) $F_n(x)$ *is AN* $\left(F(x), \frac{F(x)[1-F(x)]}{n}\right)$.

Proof. Note that the exact distribution of $nF_n(x)$ is $\mathrm{BIN}(F(x), n)$. And, (i)-(ii) follows immediately; The third part is a direct application of SLLN; (iv) is a consequence of Lindeberg-Levy CLT and (i).

3.1.2 Kolmogorov-Smirnov distance

The ECDF is quite useful for estimation of the population distribution function F. Besides pointwise estimation of $F(x)$, it is also of interest to characterize globally the estimation of F by F_n. To this end, a popular useful measure of closeness of F_n to F is the *Kolmogorov-Smirnov distance*

$$D_n = \sup_{-\infty<x<\infty} |F_n(x) - F(x)|.$$

This measure is also known as the *sup-norm distance* between F_n and F, and denoted as $||F_n(x) - F(x)||_\infty$. The metrics such as D_n has many applications: (1) goodness-of-fit test; (2) confidence band; (3) theoretical investigation of many other statistics of interest which can be advantageously carried out by representing exactly and approximately as functions of ECDF. In this respect, the following results concerning the sup-norm distance is of interest on its own account but also provides a useful starting tool for the asymptotic analysis of other statistics, such as quantiles, order statistics and ranks.

The next results give useful explicit bounds on probabilities of large values for the deviation of F_n from F.

Theorem 3.1.1 (DKW's inequality) *Let F_n be the ECDF based on iid $X_1, \ldots, X_n$ from a CDF F defined on R. There exists a positive constant C (not depending on F) such that*

$$P(D_n > z) \leq Ce^{-2nz^2}, z > 0, \text{ for all } n = 1, 2, \ldots,$$

Note that this inequality may be expresses in the form:

$$P(\sqrt{n}D_n > z) \leq Ce^{-2z^2},$$

which clearly demonstrate that $\sqrt{n}D_n = O_p(1)$. The originally DKW inequality did not specify the constant C, however, Massart (1990) found that $C = 2$ which cannot be improved. It is stated next.

Theorem 3.1.2 (Massart) *If $nz^2 \geq \log 2/2$,*

$$P(\sqrt{n}D_n > z) \leq 2e^{-2z^2}, z > 0, \text{ for all } n = 1, 2, \ldots.$$

The following results useful in statistics are direct consequences of Theorem 3.1.1.

Corollary 3.1.1 *Let F and C be as in Theorem 3.1.1. Then for every $\epsilon > 0$,*

$$P\left(\sup_{m \geq n} D_m > \epsilon\right) \leq \frac{C}{1 - h_\epsilon} h_\epsilon^n,$$

where $h_\epsilon = \exp(-2\epsilon^2)$.

Proof.

$$P\left(\sup_{m \geq n} D_m > \epsilon\right) \leq \sum_{m=n}^{\infty} P(D_m > \epsilon) \leq C \sum_{m=n}^{\infty} h_\epsilon^m = \frac{C}{1 - h_\epsilon} h_\epsilon^n.$$

Theorem 3.1.3 (Glivenko-Cantelli) $D_n \overset{wp1}{\to} 0$.

Proof. Note that $\sum_{n=1}^{\infty} P(D_n > z) < \infty$ by DKW's inequality. Hence, the result follows from Theorem 1.2.1-(iv). □

From the Glivenko-Cantelli theorem, we know that $D_n = o_p(1)$. However, the statistic $\sqrt{n}D_n$ may have a nondegenerate limit distribution as suggested by DKW's inequality, and, this is true as revealed by the following result.

Theorem 3.1.4 (Kolmogorov) *Let F be continuous. Then*

$$\lim_{n\to\infty} P(\sqrt{n}D_n \le z) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j+1}e^{-2j^2z^2},\ z > 0.$$

A convenient feature of this asymptotic distribution is that it does not depend upon F. In fact, for every n, if the true CDF F is continuous, then D_n has the remarkable property that its exact distribution is completely independent of F which is stated as follows.

Proposition 3.1.2 *Let F be continuous. Then $\sqrt{n}D_n$ is distribution-free in the sense that its exact distribution does not depend on F for every fixed n.*

Proof. The quickest way to see this property is to notice the identity:

$$\sqrt{n}D_n \stackrel{d}{=} \sqrt{n}\max_{0\le i\le n}\max\left(\frac{i}{n} - U_{(i)}, U_{(i)} - \frac{i-1}{n}\right),$$

where $U_{(1)} \le \ldots \le U_{(n)}$ are order statistics of an independent sample from $U[0,1]$ and the relation $\stackrel{d}{=}$ denotes "equality in law". □

Example 3.1.1 (Kolmogorov-Smirnov confidence intervals) A method constructing asymptotically valid intervals for a mean is due to T. W. Anderson. The construction depends on the classical D_n distance due to Kolmogorov and Smirnov, summarized below. By Proposition 3.1.2, we know given $\alpha \in (0,1)$, there is a well-defined $d = d_{\alpha,n}$ such that, for any continuous CDF F, $P_F(\sqrt{n}D_n > d) = \alpha$. Thus,

$$\begin{aligned}
1 - \alpha = P_F(\sqrt{n}D_n \le d) &= P_F(\sqrt{n}||F_n - F||_\infty \le d) \\
&= P_F\left(|F_n - F| \le \frac{d}{\sqrt{n}}, \forall x\right) \\
&= P_F\left(F_n(x) - \frac{d}{\sqrt{n}} \le F(x) \le F_n(x) + \frac{d}{\sqrt{n}}, \forall x\right).
\end{aligned}$$

This gives us a "confidence band" for the true CDF F. More precisely, the $1-\alpha$ Kolmogorov-Smirnov confidence band for the CDF F is

$$KS_{n,\alpha} : \left\{\max(0, F_n(x) - \frac{d}{\sqrt{n}}) \le F(x) \le \min(1, F_n(x) + \frac{d}{\sqrt{n}})\right\}.$$

The computation of $d = d_{\alpha,n}$ is quite nontrivial, but tables are available which will be discussed later.

3.1.3 Applications: Kolmogorov-Smirnov and other ECDF-based GOF tests

We know that, for large n, F_n is "close" to the true F. So if $H_0 : F = F_0$ holds, then we should be able to test H_0 by studying the deviation between F_n and F_0. Any choice of a discrepancy measure between F_n and F_0 would result in a test. The utility of the test would depend on whether one can work out the distribution theory of the test statistic. Three most well-known discrepancy measures that have been proposed are the following

$$
\begin{aligned}
D_n &= \max(D_n^+, D_n^-) \equiv \max\left(\sup_{-\infty<x<\infty} (F_n(x) - F_0(x)), \sup_{-\infty<x<\infty} (F_0(x) - F_n(x)) \right), \\
C_n &= n \int (F_n(t) - F_0(t))^2 dF_0(t), \\
A_n &= n \int \frac{(F_n(t) - F_0(t))^2}{F_0(t)(1 - F_0(t))} dF_0(t),
\end{aligned}
$$

which are respectively known as the Kolmogorov-Smirnov, the Cramer-von Mises, and the Anderson-Darling test statistics.

Similar to Proposition 3.1.2, we have the following simple expressions for C_n and A_n.

Proposition 3.1.3 *Let F_0 be continuous.*

$$
\begin{aligned}
C_n &= \frac{1}{12n} + \sum_{i=1}^{n} \left(U_{(i)} - \frac{i - \frac{1}{2}}{n} \right)^2, \\
A_n &= -n - \frac{2}{n} \sum_{i=1}^{n} \left[\left(i - \frac{1}{2} \right) \log U_{(i)} + \left(n - i + \frac{1}{2} \right) \left(\log(1 - U_{(n-i+1)}) \right) \right].
\end{aligned}
$$

It is clear from these computational formulas that, for every fixed n, the sampling distributions of C_n and A_n under H_0 do not depend on F_0, provided F_0 is continuous. For small n, the true sampling distributions can be worked out exactly by discrete enumeration.

The tests introduced above based on the ECDF F_n all have the pleasant property that they are consistent against any alternative $F \neq F_0$. For example, the Kolmogorov-Smirnov statistic D_n has the property that $P_F(\sqrt{n} D_n > G_n^{-1}(1-\alpha)) \to 1, \forall F \neq F_0$, where $G_n^{-1}(1-\alpha)$ is the $(1-\alpha)$th quantile of the distribution of $\sqrt{n} D_n$ under F_0. To explain heuristically

why this should be the case, consider a CDF $F_1 \neq F_0$, so that there exists η such that $F_1(\eta) \neq F_0(\eta)$. Let us suppose that $F_1(\eta) > F_0(\eta)$. First note that $G_n^{-1}(1-\alpha) \to \lambda$ for some λ by Theorem 3.1.4, say $G_n^{-1}(1-\alpha) = O(1)$. So,

$$\begin{aligned}
&P_{F_1}\left(\sqrt{n}D_n > G_n^{-1}(1-\alpha)\right) \\
&\quad = P_{F_1}\left(\sup_t \left|\sqrt{n}(F_n(t) - F_0(t))\right| > G_n^{-1}(1-\alpha)\right) \\
&\quad = P_{F_1}\left(\sup_t \left|\sqrt{n}(F_n(t) - F_1(t)) + \sqrt{n}(F_1(t) - F_0(t))\right| > G_n^{-1}(1-\alpha)\right) \\
&\quad \geq P_{F_1}\left(\left|\sqrt{n}(F_n(\eta) - F_1(\eta)) + \sqrt{n}(F_1(\eta) - F_0(\eta))\right| > G_n^{-1}(1-\alpha)\right) \to 1,
\end{aligned}$$

as $n \to \infty$ since $\sqrt{n}(F_n(\eta) - F_1(\eta)) = O_p(1)$ under F_1, and $\sqrt{n}(F_1(\eta) - F_0(\eta)) \to \infty$. The same argument establishes the consistency of the other ECDF-based tests against all alternatives. In contrast, we will later see that chi-square goodness-of-fit tests cannot be consistent against all alternatives.

Example 3.1.2 (The Berk-Jones procedure) Berk and Jones (1979) proposed an intuitively appealing ECDF-base method of testing the simple goodness-of-fit null hypothesis $F = F_0$ for some specified continuous F_0 in the one-dimensional iid situation. It has also led to subsequent developments of other tests for the simple goodness-of-fit problem as generalizations of the Berk-Jones idea.

The Berk-Jones method is to transform the simple goodness-of-fit problem into a family of binomial testing problems. More specifically, if the true underlying CDF is F, then for any given x, as stated above, $nF_n(x) \sim \text{Bin}(n, F(x))$. Suppressing the x and writing p for $F(x)$ and p_0 for $F_0(x)$, for the given x, we want to test $p = p_0$. We can use a likelihood ratio test corresponding to a two-sided alternative to test this hypothesis. It will require maximization of the binomial likelihood function over all values of p that corresponds to maximization over $F(x)$, with x being fixed, while F is an arbitrary CDF. The likelihood is maximized at $F(x) = F_n(x)$, resulting in the likelihood ratio statistic

$$\lambda_n(x) = \frac{F_n(x)^{nF_n(x)}(1 - F_n(x))^{n-nF_n(x)}}{F_0(x)^{nF_n(x)}(1 - F_0(x))^{n-nF_n(x)}}.$$

But, of course, the original problem is to test that $F(x) = F_0(x), \forall x$. So, it would make sense to take a supremum of the log-likelihood ratio statistics over x. The Berk-Jones statistic is

$$R_n = n^{-1} \sup_x \log \lambda_n(x).$$

In recent literatures, some authors have found that an analog of the traditional Anderson-Darling rank test based on $\log \lambda_n(x)$

$$\int \frac{\log \lambda_n(x)}{F_n(t)(1 - F_n(t))} dF_n(t)$$

is much more powerful than the Anderson-Darling test and the foregoing Berk-Jones statistic.

Example 3.1.3 (The two-sample case) Suppose $X_i, i = 1, \ldots, n$ are iid samples from some continuous CDF F_1 and $Y_i, i = 1, \ldots, m$ are iid samples from some continuous CDF F_2, and all random variables are mutually independent. Let F_{n1} and F_{m2} denote the empirical CDFs of the X_i's and the Y_i's, respectively. Analogous to the one-sample case, one can define two-sided Kolmogorov- Smirnov test statistics and other ECDF based GOF tests, such as

$$\begin{aligned} D_{m,n} &= \sup_{-\infty<x<\infty} |F_{n1} - F_{m2}|. \\ A_{m,n} &= \frac{nm}{n+m} \int \frac{(F_{n1}(x) - F_{m2}(x))^2}{F_{n,m}(x)(1 - F_{n,m}(x))} dF_{n,m}(x), \end{aligned}$$

where $F_{n,m}(x)$ is the ECDF of the pooled sample $X_1, \ldots, X_n, Y_1, \ldots, Y_m$. Similar to Proposition 3.1.3, one can also show that neither the null distribution of $D_{m,n}$ nor that of the $A_{m,n}$ depends on F_1 or F_2.

3.1.4 The Chi-square test

Chi-square tests are well-known competitors to ECDF-based statistics. They discretize the null distribution in some way and assess the agreement of observed counts to the postulated counts, so there is obviously some loss of information and hence a loss in power. But they are versatile. Unlike ECDF-based tests, a chi-square test can be used for continuous as well as discrete data and in one dimension as well as many dimensions. Thus, a loss of information is being exchanged for versatility of the principle and ease of computation.

Suppose $X_1, \ldots X_n$ are iid observations from some distribution F in and that we want to test $H_0 : F = F_0$, F_0 being a completely specified distribution. Let $\mathcal{S}$ be the support of F_0 and, for some given $k \geq 1$, $A_{ki}, i = 1, \ldots, k$ form a partition of $\mathcal{S}$. Let $p_{0i} = P_{F_0}(A_{ki})$ and $n_i = \#\{j : x_j \in A_{ki}\}$, i.e., the observed frequency of the partition set A_{ki}. Therefore, under H_0, $E(n_i) = np_{0i}$. K. Pearson suggested that as a measure of discrepancy between the observed sample and the null hypothesis, one compare $(n_1, \ldots, n_k)$ with $(np_{01}, \ldots, np_{0k})$. The Pearson chi-square statistic is defined as

$$K^2 = \sum_{i=1}^{k} \frac{(n_i - np_{0i})^2}{np_{0i}}.$$

For fixed n, certainly K^2 is not distributed as a chi-square, for it is just a quadratic form in a multinomial random vector. However, the asymptotic distribution of K^2 is χ^2_{k-1} if H_0 holds, which is stated in the following result.

Theorem 3.1.5 (The asymptotic null distribution) *Suppose $X_1, X_2, \ldots X_n$ are iid observations from some distribution F. Consider testing $H_0 : F = F_0$ (specified). $K^2 \xrightarrow{d} \chi^2_{k-1}$ under H_0.*

Proof. Define

$$\mathbf{Y} = (Y_1, \ldots, Y_k)^T = \left(\frac{n_1 - np_{01}}{\sqrt{np_{01}}}, \ldots, \frac{n_k - np_{0k}}{\sqrt{np_{0k}}} \right)^T.$$

By the multivariate CLT, we know $\mathbf{Y} \xrightarrow{d} N_k(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \mathbf{I}_k - \boldsymbol{\mu}\boldsymbol{\mu}^T$ and $\boldsymbol{\mu} = (\sqrt{p_{01}}, \ldots, \sqrt{p_{0k}})^T$. This can be easily seen by writing $\mathbf{n} = (n_1, \ldots, n_k)^T = \sum_{i=1}^{n} \mathbf{Z}_i$, where $\mathbf{Z}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ with a single nonzero component 1 located in the jth position if the ith trial yields the jth outcome. Note that $\mathbf{Z}_i$'s are iid with mean $\mathbf{p}_0 = (p_{01}, \ldots, p_{0k})^T$ and covariance matrix $\mathrm{diag}(\mathbf{p}_0) - \mathbf{p}_0\mathbf{p}_0^T$. By the multivariate CLT, $\frac{\mathbf{n} - n\mathbf{p}_0}{\sqrt{n}} \xrightarrow{d} N_k(\mathbf{0}, \mathrm{diag}(\mathbf{p}_0) - \mathbf{p}_0\mathbf{p}_0^T)$. Thus,

$$\begin{aligned}
\mathbf{Y} &= \mathrm{diag}^{-1}(\sqrt{p_{01}}, \ldots, \sqrt{p_{0k}}) \frac{\mathbf{n} - n\mathbf{p}_0}{\sqrt{n}} \\
&\xrightarrow{d} N_k\left(\mathbf{0}, \mathrm{diag}^{-1}(\sqrt{p_{01}}, \ldots, \sqrt{p_{0k}}) \left[\mathrm{diag}(\mathbf{p}_0) - \mathbf{p}_0\mathbf{p}_0^T\right] \mathrm{diag}^{-1}(\sqrt{p_{01}}, \ldots, \sqrt{p_{0k}})\right) \\
&\stackrel{d}{=} N_k(\mathbf{0}, \mathbf{\Sigma}).
\end{aligned}$$

Note that $\text{tr}(\boldsymbol{\Sigma}) = k-1$. Notice now that Pearson's $K^2 = \mathbf{Y}^T\mathbf{Y}$, and if $\mathbf{Y} \sim N_k(\mathbf{0}, \boldsymbol{\Sigma})$ for any general $\boldsymbol{\Sigma}$, then $\mathbf{Y}^T\mathbf{Y} \stackrel{d}{=} \mathbf{X}^T\mathbf{P}^T\mathbf{P}\mathbf{X} = \mathbf{X}^T\mathbf{X}$, where $\mathbf{X} \sim N_k(\mathbf{0}, \text{diag}(\lambda_1, \ldots, \lambda_k))$, λ_i are the eigenvalues of $\boldsymbol{\Sigma}$, and $\mathbf{P}^T\boldsymbol{\Sigma}\mathbf{P} = \text{diag}(\lambda_1, \ldots, \lambda_k)$ is the spectral decomposition of $\boldsymbol{\Sigma}$. Note that $\mathbf{X}$ has the same distribution as the vector $(\sqrt{\lambda_1}\eta_1, \ldots, \sqrt{\lambda_k}\eta_k)^T$, where η_j's are the iid standard normal variates. So, it follows that $\mathbf{X}^T\mathbf{X} \stackrel{d}{=} \sum_{i=1}^k \lambda_i w_i$ with $w_i \stackrel{\text{iid}}{\sim} \chi^2_1$. Because the eigenvalues of a symmetric and idempotent matrix ($\boldsymbol{\Sigma}$) are either 0 or 1, for our $\boldsymbol{\Sigma}$, $k-1$ of λ_i's are 1 and the remaining one is zero. Since a sum of independent chi-squares is again a chi-square, it follows that $K^2 \xrightarrow{d} \chi^2_{k-1}$ under H_0. □

Example 3.1.4 (The Hellinger statistic) We may consider a transformation $\mathbf{g}(\mathbf{x})$ that makes the denominator in Pearson's χ^2 a constant. Specially, the differentiable function of the form $\mathbf{g}(\mathbf{x}) = (g_1(x_1), \ldots, g_k(x_k))^T$, such that the jth component of the transformation is a function only of the jth component of $\mathbf{x}$. As a consequence, the gradient $\nabla\mathbf{g}(\mathbf{x}) = \text{diag}\{g_1'(x_1), \ldots, g_k'(x_k)\}$. As in the proof of Delta Theorem, $\sqrt{n}(\mathbf{g}(\bar{\mathbf{Z}}_n) - \mathbf{g}(\mathbf{p}_0))$ is asymptotically equivalent to $\sqrt{n}\nabla\mathbf{g}(\mathbf{p}_0)(\bar{\mathbf{Z}}_n - \mathbf{p}_0)$, so that in Pearson's χ^2, we may replace $\sqrt{n}(\bar{\mathbf{Z}}_n - \mathbf{p}_0)$ by $\sqrt{n}\nabla^{-1}\mathbf{g}(\mathbf{p}_0)(\mathbf{g}(\bar{\mathbf{Z}}_n) - g(\mathbf{p}_0))$ and obtain the transformed χ^2

$$\begin{aligned}\chi^2_g &= n\left(\mathbf{g}(\bar{\mathbf{Z}}_n) - \mathbf{g}(\mathbf{p}_0)\right)^T \nabla^{-1}\mathbf{g}(\mathbf{p}_0)\text{diag}(\mathbf{p})\nabla^{-1}\mathbf{g}(\mathbf{p}_0)\left(\mathbf{g}(\bar{\mathbf{Z}}_n) - \mathbf{g}(\mathbf{p}_0)\right) \\ &= n\sum_{i=1}^k \frac{\left(g_i(n_i/n) - g_i(p_{0i})\right)^2}{p_{0i}[g_i'(p_{0i})]^2} \xrightarrow{d} \chi^2_{k-1}.\end{aligned}$$

Naturally, we are led to investigate the transformed χ^2 with $\mathbf{g}(\mathbf{x}) = (\sqrt{x_1}, \ldots, \sqrt{x_k})^T$. The transformed χ^2, with $g_i'(p_{0i}) = \frac{1}{2\sqrt{p_{0i}}}$, becomes

$$\chi^2_H = 4n\sum_{i=1}^k \left(\sqrt{n_i/n} - \sqrt{p_{0i}}\right)^2.$$

This is known as the *Hellinger* χ^2 because of its relation to Hellinger distance. The Hellinger distance between two densities, $f(x)$ and $g(x)$, is $d(f,g)$, where

$$d^2(f,g) = \int \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx.$$

Let F_1 be a distribution different from F_0 and let $p_{1i} = P_{F_1}(A_{ki})$. Clearly, if by chance $p_{1i} = p_{0i}\ \forall i = 1, \ldots k$ (which is certainly possible), then a test based on the empirical

frequencies of A_{ki} cannot distinguish F_0 from F_1, even asymptotically. In such a case, the χ^2 test cannot be consistent against F_1. However, otherwise it will be consistent, as can be seen easily from the following result.

Proposition 3.1.4 *Under F_1,*

(i) $\frac{K^2}{n} \xrightarrow{p} \sum_{i=1}^{k} \frac{(p_{1i}-p_{0i})^2}{p_{0i}}$.

(ii) If $\sum_{i=1}^{k} \frac{(p_{1i}-p_{0i})^2}{p_{0i}} > 0$, *then* $K_P^2 \xrightarrow{p} \infty$ *and hence the Pearson* χ^2 *test is consistent against* F_1.

This is evident as $K^2 = \sum_{i=1}^{k} \frac{(n_i - np_{0i})^2}{np_{0i}} = n\sum_{i=1}^{k} \frac{(n_i/n - p_{0i})^2}{p_{0i}}$. But $\mathbf{n}/n \xrightarrow{p} (p_{11}, \ldots p_{1k}) \equiv \mathbf{p}_1$ under F_1. Therefore, by the continuous mapping theorem, $K^2 \xrightarrow{p} n\sum_{i=1}^{k} \frac{(p_{1i}-p_{0i})^2}{p_{0i}}$. Thus, for a fixed alternative F_1 such that the vector $\mathbf{p}_1 \neq \mathbf{p}_0$, Pearson's χ^2 cannot have a nondegenerate limit distribution under F_1. However, if the alternative is very close to the null, in the sense of being a Pitman alternative, there is a nondegenerate limit distribution.

To obtain an approximation to the power, we consider the behavior of K^2 under a sequence of *local alternatives* to the null hypothesis. In particular, take

$$p_{1i} = p_{0i} + \delta_i n^{-1/2}, \ 1 \leq i \leq k.$$

Note that because both $\mathbf{p}_1$ and $\mathbf{p}_0$ are probability vectors, $\mathbf{1}^T\boldsymbol{\delta} = \sum_i \delta_i = 0$. Then, we have the following result which allows us to approximate the power of the χ^2 test at a close alternative by using the noncentral χ^2 CDF as an approximation to the exact CDF of χ^2 under the alternative.

Theorem 3.1.6 (The asymptotic alternative distribution) *Under H_1, say $\mathbf{p} = \mathbf{p}_1 = \mathbf{p_0} + \boldsymbol{\delta} n^{-1/2}$. Then $K^2 \xrightarrow{d} \chi^2_{k-1}(\lambda)$, where $\lambda = \sum_{i=1}^{k} \delta_i^2/p_{0i}$ is the noncentrality parameter.*

Proof. Recall the definition in the proof of Theorem 3.1.5. It can be easily seen that $\mathbf{Y} \xrightarrow{d} N_k(\text{diag}^{-1}(\sqrt{p_{01}}, \ldots, \sqrt{p_{0k}})\boldsymbol{\delta}, \boldsymbol{\Sigma})$ by using the Slutsky's Theorem and CLT. Since $\boldsymbol{\Sigma}$ is symmetric and idempotent,

$$K^2 = \mathbf{Y}^T\mathbf{Y} \xrightarrow{d} \chi^2_{k-1}\left(\sum_{i=1}^{k} \delta_i^2/p_{0i}\right)$$

by using the Cochran Theorem (or derived by a similar arguments in the proof of Theorem 3.1.5). □

A direct application of this theorem is to calculate the approximate power of K^2 test. Suppose that the critical region is $\{K^2 > c\}$, where the choice of c for a level α test would be based on the null hypothesis asymptotic χ^2_{k-1} distribution of K^2. Then the approximate power of K^2 at the alternative H_1 is given by calculating the probability that a random variable having the distribution $\chi^2_{k-1}\left(n\sum_{i=1}^k \frac{1}{p_{0i}}(n_j/n - p_{0i})^2\right)$ exceeds the value c, in which the expected value p_{1i} in the noncentrality parameter is replaced by the observed frequencies n_j/n.

3.2 The sample moments

Let $X_1, X_2 \ldots$ be iid with distribution function F. For $k \in \mathbb{N}^+$, the kth *moment* and *central moment* of F are defined as

$$\alpha_k = \int_{-\infty}^{\infty} x^k dF(x) = EX_1^k$$
$$\mu_k = \int_{-\infty}^{\infty} (x - \alpha_1)^k dF(x) = E[(X_1 - \alpha_1)^k],$$

respectively. α_1 and μ_2 are certainly the mean and variance of F respectively. Also, $\mu_1 = 0$. α_k and μ_k represent important characteristics for describing F. Natural estimators of these parameters are given by the corresponding moments of the sample distribution function $F_n(x) = \frac{1}{n}\sum_{i=1}^n I_{\{X_i \le x\}}$, say

$$a_k = \int_{-\infty}^{\infty} x^k dF_n(x) = \frac{1}{n}\sum_{i=1}^n X_i^k, \quad k = 1, 2, \ldots,$$
$$m_k = \int_{-\infty}^{\infty} (x - \alpha_1)^k dF_n(x) = \frac{1}{n}\sum_{i=1}^n (X_i - a_1)^k, \quad k = 2, 3, \ldots.$$

Since F_n possesss desirable properties as an estimator of F, it could be expected that the sample moment a_k and m_k posses desirable features as estimators of α_k and μ_k. The first result is the strong mean square consistencies regarding the a_k.

Proposition 3.2.1 *(i)* $a_k \overset{wp1}{\to} \alpha_k$*; (ii)* $E(a_k) = \alpha_k$*; (iii)* $\mathrm{Var}(a_k) = \frac{\alpha_{2k}-\alpha_k^2}{n}$.

By noting that a_k is a mean of iid random variables having mean α_k and variance $\alpha_{2k}-\alpha_k^2$, the result follows immediately by SLLN. Furthermore, because the vector $(a_1, \ldots, a_k)^T$ is the mean of the iid vectors $(X_i, \ldots, X_i^k)^T, 1 \leq i \leq n$, we have the following asymptotically normal result.

Proposition 3.2.2 $\sqrt{n}(a_1 - \alpha_1, \ldots, a_k - \alpha_k)^T$ *is* $AN_k(\mathbf{0}, \mathbf{\Sigma})$*, where* $\mathbf{\Sigma} = (\sigma_{ij})_{k\times k}$ *with* $\sigma_{ij} = \alpha_{i+j} - \alpha_i\alpha_j$.

Certainly, it is implicitly assumed that all stated moments are finite. This proposition is a direct application of the multivariate CLT Theorem 1.3.2.

To deduce the the properties of m_k, as seen in Example 1.3.2, it is advantageous to consider the closely related random variables $b_k = \frac{1}{n}\sum_{i=1}^n (X_i - \alpha_1)^k, k = 1, 2, \ldots$. The same arguments employed in dealing with the a_k's immediately yield

Proposition 3.2.3 *(i)* $b_k \overset{wp1}{\to} \mu_k$*; (ii)* $E(b_k) = \mu_k$*; (iii)* $\mathrm{Var}(b_k) = \frac{\mu_{2k}-\mu_k^2}{n}$*; (iv)* $\sqrt{n}(b_1 - \mu_1, \ldots, b_k - \mu_k)^T$ *is* $AN_k(\mathbf{0}, \tilde{\mathbf{\Sigma}})$*, where* $\tilde{\mathbf{\Sigma}} = (\tilde{\sigma}_{ij})_{k\times k}$ *with* $\tilde{\sigma}_{ij} = \mu_{i+j} - \mu_i\mu_j$.

The following result concerns the consistency and the asymptotically normality of the vector $m_1, \ldots, m_k$.

Theorem 3.2.1 *Suppose that* $\mu_{2k} < \infty$.

(i) $m_k \overset{wp1}{\to} \mu_k$;

(ii) The random vector $\sqrt{n}(m_2-\mu_2, \ldots, m_k-\mu_k)^T$ *is* $AN_{k-1}(\mathbf{0}, \mathbf{\Sigma}^*)$*, where* $\mathbf{\Sigma}^* = (\sigma_{ij}^*)_{(k-1)\times(k-1)}$ *with* $\sigma_{ij}^* = \mu_{i+j+2} - \mu_{i+1}\mu_{j+1} - (i+1)\mu_i\mu_{j+2} - (j+1)\mu_{i+2}\mu_j + (i+1)(j+1)\mu_i\mu_j\mu_2$.

Proof. Instead of dealing with m_k directly, we exploit the connection between m_k and b_j's. Writing

$$m_k = \frac{1}{n}\sum_{i=1}^n (X_i - a_1)^k = \frac{1}{n}\sum_{i=1}^n \sum_{j=0}^k C_k^j (X_i - \alpha_1)^j (\alpha_1 - a_1)^{k-j},$$

we have

$$m_k = \sum_{j=0}^{k} C_k^j (-1)^{k-j} b_j b_1^{k-j},$$

where we define $b_0 = 1$. (i). By noting that $\mu_1 = 0$, this result follows from (i) of Proposition 3.2.3 and the CMT; (ii) This is again an application of the multivariate Delta Theorem. Consider the map $g : \mathbb{R}^k \to \mathbb{R}^{k-1}$ given by

$$g(t_1, \ldots, t_k) = \left(\sum_{j=0}^{2} C_2^j (-1)^{2-j} t_j t_1^{2-j}, \ldots, \sum_{j=0}^{k} C_k^j (-1)^{k-j} t_j t_1^{k-j} \right)^T.$$

Let $\boldsymbol{\theta} = (0, \mu_2, \ldots, \mu_k)^T$ and $g(\boldsymbol{\theta}) = (\mu_2, \ldots, \mu_k)^T$. A direct evaluation of ∇g at $\boldsymbol{\theta}$ yields

$$\nabla^T g_{|\boldsymbol{\theta}} = \begin{pmatrix} -2\mu_1 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \\ -(i+1)\mu_i & 0 & \cdots & 1 & \cdots \\ \vdots & & \vdots & & \\ -k\mu_{k-1} & 0 & \cdots & & 1 \end{pmatrix}.$$

It follows that the sequence $\sqrt{n}(m_2 - \mu_2, \ldots, m_k - \mu_k)^T$ is asymptotically normally distributed, with mean zero and covariance matrix,

$$\boldsymbol{\Sigma}^* = \nabla^T g_{|\boldsymbol{\theta}} \tilde{\boldsymbol{\Sigma}} \nabla g_{|\boldsymbol{\theta}}.$$

The assertion follows immediately from some simple algebras on $\nabla^T g_{|\boldsymbol{\theta}} \tilde{\boldsymbol{\Sigma}} \nabla g_{|\boldsymbol{\theta}}$. □.

A most direct result from this theorem is the asymptotical normality of the sample variance (by choosing $k = 2$ in (ii)) which is studied detailedly in Example 1.3.2.

3.3 The sample quantiles

A few selected sample percentiles provide useful diagnostic summaries of the full ECDF. For example, the three quartiles of the sample already provide some information about symmetry of the underlying population, and extreme percentiles give information about the tail. So asymptotic theory of sample percentiles is of great interest in statistics. In this section, we

present a selection of the fundamental results on the asymptotic theory for percentiles. The iid case and then an extension to the regression setup are discussed.

Suppose $X_1, \ldots, X_n$ are iid real-valued random variables with CDF F. We denote the order statistics of $X_1, \ldots, X_n$ by $X_{(1)}, \ldots, X_{(n)}$. For $0 < p < 1$, the pth quantile of F is defined as $F^{-1}(p) \equiv \xi_p = \inf\{x : F(x) \geq p\}$. Note that ξ_p satisfies $F(\xi_p-) \leq p \leq F(\xi_p)$. Correspondingly, the sample quantile is defined as the pth quantile of the ECDF F_n, that is, $F_n^{-1}(p) \equiv \widehat{\xi}_p = \inf\{x : F_n(x) \geq p\}$. Also, the sample quantile can be expressed as $X_{(\lceil np \rceil)}$ where $\lceil k \rceil$ denotes the smallest integer greater than or equal to k. Thus, the discussion of quantile could be carried out formally in terms of order statistics.

3.3.1 Basic results

The first result is a probability inequality for $|\widehat{\xi}_p - \xi_p|$ which implies that $\widehat{\xi}_p$ is strongly consistent, say $\widehat{\xi}_p \overset{wp1}{\to} \xi_p$.

Theorem 3.3.1 *Let $X_1, \ldots, X_n$ be iid random variables from a CDF F satisfying $p < F(\xi_p+\epsilon)$ for any $\epsilon > 0$. Then, for every $\epsilon > 0$ and $n = 1, 2, \ldots,$*

$$P(|\widehat{\xi}_p - \xi_p| > \epsilon) \leq 2Ce^{-2n\delta_\epsilon^2},$$

where $\delta_\epsilon = \min\{F(\xi_p + \epsilon) - p, p - F(\xi_p - \epsilon)\}$ and C is the same constant in DKW inequality.

Proof. Let $\epsilon > 0$ be fixed. Note that $G(x) \geq t$ iff $x \geq G^{-1}(t)$ for any CDF G on $\mathbb{R}$. Hence

$$\begin{aligned} P(\widehat{\xi}_p > \xi_p + \epsilon) &= P(p > F_n(\xi_p + \epsilon)) \\ &= P\left(F(\xi_p + \epsilon) - F_n(\xi_p + \epsilon) > F(\xi_p + \epsilon) - p\right) \\ &\leq P(D(F_n, F) > \delta_\epsilon) \leq Ce^{-2n\delta_\epsilon^2} \end{aligned}$$

where the last inequality follows from DKW's inequality (Theorem 3.1.1). Similarly,

$$P(\widehat{\xi}_p < \xi_p - \epsilon) \leq Ce^{-2n\delta_\epsilon^2}.$$

This proves the assertion. □

By this inequality, the strong consistency of $\widehat{\xi}_p$ can be established easily from Theorem 1.2.1-(iv).

Remark 3.3.1 The exact distribution of $\widehat{\xi}_p$ can be obtained as follows. Since $nF_n(t)$ has the binomial distribution $\text{BIN}(F(t), n)$ for any $t \in \mathbb{R}$,

$$\begin{aligned} P(\widehat{\xi}_p \le t) &= P(F_n(t) \ge p) \\ &= \sum_{i=l_p}^{n} C_n^i [F(t)]^i [1 - F(t)]^{n-i}, \end{aligned}$$

where $l_p = \lceil np \rceil$. If F has a PDF f, then $\widehat{\xi}_p$ has the PDF

$$\phi_n(t) = nC_{n-1}^{l_p-1} [F(t)]^{l_p-1} [1 - F(t)]^{n-l_p} f(t).$$

The following result provides an asymptotic distribution for $\sqrt{n}(\widehat{\xi}_p - \xi_p)$.

Theorem 3.3.2 *Let $X_1, \ldots, X_n$ be iid random variables from a CDF F. Suppose that F is continuous at ξ_p.*

(i) If there exists $F'(\xi_p-) > 0$, then for any $t < 0$,

$$\lim_{n\to\infty} P\left(\frac{\sqrt{n}(\widehat{\xi}_p - \xi_p)}{\sqrt{p(1-p)}/F'(\xi_p-)} \le t \right) = \Phi(t);$$

(ii) If there exists $F'(\xi_p+) > 0$, then for any $t > 0$,

$$\lim_{n\to\infty} P\left(\frac{\sqrt{n}(\widehat{\xi}_p - \xi_p)}{\sqrt{p(1-p)}/F'(\xi_p+)} \le t \right) = \Phi(t);$$

(iii) If $F'(\xi_p)$ exists and is positive, then

$$\sqrt{n}(\widehat{\xi}_p - \xi_p) \xrightarrow{d} N\left(0, \frac{p(1-p)}{[F'(\xi_p)]^2}\right).$$

Proof. If F is differentiable at ξ_p, then $F'(\xi_p-) = F'(\xi_p+) = F'(\xi_p)$. Thus, part (iii) is a direct consequence of (i) and (ii). Note that the proofs of (i) and (ii) are similar. Thus, we only give a proof for (ii).

Let $t > 0$, $p_{nt} = F(\xi_p + t\sigma_F^+ n^{-1/2})$, $c_{nt} = \sqrt{n}(p_{nt} - p)/\sqrt{p_{nt}(1-p_{nt})}$, and $Z_{nt} = [B_n(p_{nt}) - np_{nt}]/\sqrt{np_{nt}(1-p_{nt})}$, where $\sigma_F^+ = \sqrt{p(1-p)}/F'(\xi_p+)$ and $B_n(q)$ denotes a random variable having the binomial distribution BIN(q, n). Then,

$$P\left(\widehat{\xi}_p \le \xi_p + t\sigma_F^+ n^{-1/2}\right) = P\left(p \le F_n(\xi_p + t\sigma_F^+ n^{-1/2})\right)$$
$$= P(Z_{nt} \ge -c_{nt}).$$

Under the assumed conditions on F, $p_{nt} \to p$ and $c_{nt} \to t$. Hence, the result follows from

$$P(Z_{nt} < -c_{nt}) - \Phi(-c_{nt}) \to 0.$$

But this follows from the CLT and Polya's theorem (Theorem 1.2.7-(ii)). □

Remark 3.3.2 If both $F'(\xi_p+)$ and $F'(\xi_p-)$ exist and are positive, but $F'(\xi_p+) \neq F'(\xi_p-)$, then the asymptotic distribution of $\sqrt{n}(\widehat{\xi}_p - \xi_p)$ has the CDF $\Phi(t/\sigma_F^-)I_{\{-\infty<t<0\}} + \Phi(t/\sigma_F^+)I_{\{0\le t<\infty\}}$, a mixture of two normal distributions, where $\sigma_F^- = \sqrt{p(1-p)}/F'(\xi_p-)$. An example of such a case when $p = \frac{1}{2}$ is

$$F(x) = xI_{\{0\le x<\frac{1}{2}\}} + (2x - \frac{1}{2})I_{\{\frac{1}{2}\le x<\frac{3}{4}\}} + I_{\{\frac{3}{4}\le x<\infty\}}.$$

Example 3.3.1 Suppose $X_1, X_2, \ldots,$ are iid $N(\mu, 1)$. Let $M_n = \widehat{\xi}_{\frac{1}{2}}$ denote the sample median. Since the standard normal density $\phi(x)$ at zero equals $1/\sqrt{2\pi}$, it follows from Theorem 3.3.2 that $\sqrt{n}(M_n - \mu) \xrightarrow{d} N(0, \frac{\pi}{2})$. On the other hand, $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, 1)$. The ratio of the variances in the two asymptotic distributions, $2/\pi$, is called the ARE (asymptotic relative efficiency) of M_n relative to $\bar{X}_n$. Thus, for normal data, M_n is less efficient than $\bar{X}_n$.

3.3.2 Bahadur's representation

The sample median of an iid sample from some CDF F is clearly not a linear statistic; i.e., it is not a function of the form $\sum_{i=1}^n h_i(X_i)$. In 1966, Bahadur proved that the sample median, and more generally any fixed sample percentile, is almost a linear statistic. The result in Bahadur (1966) not only led to an understanding of the probabilistic structure of percentiles but also turned out to be an extremely useful technical tool. For example, as

we shall shortly see, it follows from Bahadur's result that, for iid samples from a CDF F, under suitable conditions not only are $\bar{X}_n$, $\widehat{\xi}_{\frac{1}{2}}$ marginally asymptotically normal but they are jointly asymptotically bivariate normal. The result derived in Bahadur (1966) is known as the *Bahadur representation of quantiles.*

Theorem 3.3.3 (Bahadur's representation) *Let $X_1, \ldots, X_n$ be iid random variables from a CDF F. Suppose that $F'(\xi_p)$ exists and is positive. Then*

$$\widehat{\xi}_p = \xi_p + \frac{F(\xi_p) - F_n(\xi_p)}{F'(\xi_p)} + o_p(\frac{1}{\sqrt{n}}).$$

Proof. Let $t \in \mathbb{R}$, $\xi_{nt} = \xi_p + tn^{-1/2}$, $Z_n(t) = \sqrt{n}[F(\xi_{nt}) - F_n(\xi_{nt})]/F'(\xi_p)$, and $U_n(t) = \sqrt{n}[F(\xi_{nt}) - F_n(\widehat{\xi}_p)]/F'(\xi_p)$. It can be shown that

$$Z_n(t) - Z_n(0) = o_p(1). \tag{3.1}$$

Note that $|p - F_n(\widehat{\xi}_p)| \leq n^{-1}$. Then,

$$\begin{aligned} U_n(t) &= \sqrt{n}[F(\xi_{nt}) - p + p - F_n(\widehat{\xi}_p)]/F'(\xi_p) \\ &= \sqrt{n}[F(\xi_{nt}) - p]/F'(\xi_p) + O(n^{-1/2}) \to t \end{aligned} \tag{3.2}$$

Let $\eta_n = \sqrt{n}(\widehat{\xi}_p - \xi_p)$. Then for any $t \in \mathbb{R}$ and $\epsilon > 0$,

$$\begin{aligned} P(\eta_n \leq t, Z_n(0) \geq t + \epsilon) &= P(Z_n(t) \leq U_n(t), Z_n(0) \geq t + \epsilon) \\ &\leq P(|Z_n(t) - Z_n(0)| \geq \epsilon/2) + P(|U_n(t) - t| \geq \epsilon/2) \to 0 \end{aligned}$$

by (3.1) and (3.2). Similarly,

$$P(\eta_n \geq t + \epsilon, Z_n(0) \leq t) \to 0.$$

It follows that $\eta_n - Z_n(0) = o_p(1)$ with Lemma 3.3.1 given below, which is the same as the assertion. □

Lemma 3.3.1 *Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables such that X_n is bounded in probability and, for any real number t and $\epsilon > 0$, $\lim_n[P(X_n \leq t, Y_n \geq t + \epsilon) + P(X_n \geq t + \epsilon, Y_n \leq t)] = 0$. Then $X_n - Y_n \xrightarrow{p} 0$.*

Proof. For any $\epsilon > 0$, there exists and $M > 0$ such that $P(|X_n| > M) \leq \epsilon$ for any n, since X_n is bounded in probability. For this fixed M, there exists an N such that $2M/N < \epsilon/2$. Let $t_i = -M + 2Mi/N, i = 0, 1, \ldots, N$. Then,

$$\begin{aligned} P(|X_n - Y_n| \geq \epsilon) &\leq P(|X_n| \geq M) + P(|X_n| < M, |X_n - Y_n| \geq \epsilon) \\ &\leq \epsilon + \sum_{i=1}^{N} P(t_{i-1} \leq X_n \leq t_i, |X_n - Y_n| \geq \epsilon) \\ &\leq \epsilon + \sum_{i=1}^{N} P(Y_n \leq t_{i-1} - \epsilon/2, t_{i-1} \leq X_n) + P(Y_n \geq t_i + \epsilon/2, X_n \leq t_i). \end{aligned}$$

This, together with the given condition, implies that

$$\limsup_n P(|X_n - Y_n| \geq \epsilon) \leq \epsilon.$$

Since ϵ is arbitrary, we conclude that $X_n - Y_n \xrightarrow{p} 0$. □

Remark 3.3.3 Actually, Bahadur gave an a.s. order for $o_p(n^{-1/2})$ under the stronger assumption that F is twice differentiable at ξ_p with $F'(\xi_p) > 0$. The theorem stated here is in the form later given in Ghosh (1971). The exact a.s. order was shown to be $n^{-3/4}(\log\log n)^{3/4}$ by Kiefer (1967) in a landmark paper. However, the weaker version presented here suffices for proving the following CLTs.

The Bahadur representation easily leads to the following two joint asymptotic distributions.

Corollary 3.3.1 *Let $X_1, \ldots, X_n$ be iid random variables from a CDF F having positive derivatives at ξ_{p_j}, where $0 < p_1 < \cdots < p_m < 1$ are fixed constants. Then*

$$\sqrt{n}[(\widehat{\xi}_{p_1}, \ldots, \widehat{\xi}_{p_m}) - (\xi_{p_1}, \ldots, \xi_{p_m})] \xrightarrow{d} N_m(0, \mathbf{D}),$$

where $\mathbf{D}$ is the $m \times m$ symmetric matrix with element

$$D_{ij} = p_i(1 - p_j)/[F'(\xi_{p_i})F'(\xi_{p_j})], \ i \leq j.$$

Proof. By Theorem 3.3.3, we know that the $\sqrt{n}[(\widehat{\xi}_{p_1}, \ldots, \widehat{\xi}_{p_m}) - (\xi_{p_1}, \ldots, \xi_{p_m})]^T$ is asymptotically equivalent to $\sqrt{n}[\frac{F(\xi_{p_1}) - F_n(\xi_{p_1})}{F'(\xi_{p_1})}, \ldots, \frac{F(\xi_{p_m}) - F_n(\xi_{p_m})}{F'(\xi_{p_m})}]^T$ and thus we only need to derive

the joint asymptotic distribution of $\frac{\sqrt{n}[F(\xi_{p_i})-F_n(\xi_{p_i})]}{F'(\xi_{p_i})}, i = 1, \ldots, m$. By the definition of ECDF, the sequence of $[F_n(\xi_{p_1}), \ldots, F_n(\xi_{p_m})]^T$ can be represented as the sum of independent random vectors

$$\frac{1}{n}\sum_{i=1}^{n}[I_{\{X_i \le \xi_{p_1}\}}, \ldots, I_{\{X_i \le \xi_{p_m}\}}]^T.$$

Thus, the result immediately follows from the multivariate CLT by using the fact that

$$E(I_{\{X_i \le \xi_{p_k}\}}) = F(\xi_{p_k}),\ \mathrm{Cov}(I_{\{X_i \le \xi_{p_k}\}}, I_{\{X_i \le \xi_{p_l}\}}) = p_k(1-p_l),\ k \le l.$$

Example 3.3.2 (Interquartile range; IQR) One application of Corollary 3.3.1 is the derivation of the derivation of the asymptotic distribution of the interquartile range $\widehat{\xi}_{0.75} - \widehat{\xi}_{0.25}$. It is widely used as a measure of the variability among X_i's. Use of such an estimate is quite common when normality is suspect. It can be shown that

$$\sqrt{n}[(\widehat{\xi}_{0.75} - \widehat{\xi}_{0.25}) - (\xi_{0.75} - \xi_{0.25})] \xrightarrow{d} N(0, \sigma_F^2)$$

with

$$\sigma_F^2 = \frac{3}{16[F'(\xi_{0.75})]^2} + \frac{3}{16[F'(\xi_{0.25})]^2} - \frac{1}{8F'(\xi_{0.75})F'(\xi_{0.25})}.$$

In particular, if $X_1, \ldots, X_n$ are iid $N(0, \sigma^2)$, then, by using the general result above, on some algebra, $\sqrt{n}(\mathrm{IQR} - 1.35\sigma) \xrightarrow{d} N(0, 2.48\sigma^2)$. Consequently, for normal data, IQR/1.35 is a consistent estimate of σ (the 1.35 value of course is an approximation) with asymptotic variance $2.48\sigma^2/1.35^2 = 1.36\sigma^2$. On the other hand, $\sqrt{n}(S_n - \sigma) \xrightarrow{d} N(0, 0.5\sigma^2)$. The ratio of the asymptotic variances, namely $0.5/1.36 = 0.37$, is the ARE of the IQR-based estimate relative to S_n. Thus, for normal data, one is better off using S_n. For populations with thicker tails, IQR-based estimates can be more efficient.

Example 3.3.3 (Gastwirth estimate) Suppose $X_1, \ldots, X_n$ are continuous and distributed as iid $F(x - \mu)$, where $F(-x) = 1 - F(x)$ and we wish to estimate the location parameter μ. An obvious idea is to use a convex combination of order statistics $\sum_{i=1}^{n} c_{ni} X_{(i)}$. Such statistics are called *L-statistics*. A particular L-statistic that was found to have attractive versatile performance is the Gastwirth estimate

$$\mu = 0.3X_{(\frac{n}{3})} + 0.4X_{(\frac{n}{2})} + 0.3X_{(\frac{2n}{3})}.$$

This estimate is asymptotically normal with an explicitly available variance formula since we know from our general theorem that $[X_{(\frac{n}{3})}, X_{(\frac{n}{2})}, X_{(\frac{2n}{3})}]^T$ is jointly asymptotically trivariate normal under mild conditions.

Corollary 3.3.2 *Let $X_1, \ldots X_n$ be iid from a CDF F. Let $0 < p < 1$ and suppose $\mathrm{Var}_F(X_1) < \infty$. If F is differentiable at ξ_p with $F'(\xi_p) = f(\xi_p) > 0$, then*

$$\sqrt{n}\left(\bar{X}_n - \mu, F_n^{-1}(p) - \xi_p\right) \xrightarrow{d} N_2(\mathbf{0}, \mathbf{\Sigma}),$$

where

$$\mathbf{\Sigma} = \begin{pmatrix} \mathrm{Var}(X_1) & \frac{p}{f(\xi_p)} E_F(X_1) - \frac{1}{f(\xi_p)} \int_{x \le \xi_p} x dF(x) \\ \frac{p}{f(\xi_p)} E_F(X_1) - \frac{1}{f(\xi_p)} \int_{x \le \xi_p} x dF(x) & \frac{p(1-p)}{f^2(\xi_p)} \end{pmatrix}.$$

The proof of this corollary is very similar to Corollary 3.3.1 and hence left as an exercise.

Example 3.3.4 As an application of this result, consider iid $N(\mu, 1)$ data. Take $p = \frac{1}{2}$ so that Corollary 3.3.2 gives the joint asymptotic distribution of the sample mean and the sample median. The covariance entry in the matrix $\mathbf{\Sigma}$ equals (assuming without any loss of generality that $\mu = 0$) $-\sqrt{2\pi} \int_{-\infty}^{0} x\phi(x)dx = 1$. Therefore, the asymptotic correlation between the sample mean and median in the normal case is $\sqrt{\frac{2}{\pi}} = 0.7979$, a fairly strong correlation.

3.3.3 Confidence intervals for quantiles

Since the population median and more generally population percentiles provide useful summaries of the population CDF, inference for them is of clear interest. Confidence intervals for population percentiles are therefore of interest in inference. Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$ and we wish to estimate $\xi_p = F^{-1}(p)$ for some $0 < p < 1$. The corresponding sample percentile $\widehat{\xi}_p = F_n^{-1}(p)$ is typically a fine point estimate for p. But how does one find a confidence interval of guaranteed coverage?

One possibility is to use the quantile transformation and observe that

$$(F(X_{(1)}), F(X_{(2)}), \ldots, F(X_{(n)})) \stackrel{d}{=} (U_{(1)}, U_{(2)}, \ldots, U_{(n)}),$$

where $U_{(i)}$ is the ith order statistic of a $U[0,1]$ random sample, provided F is continuous. Therefore, for given $1 \le i_1 < i_2 \le n$,

$$\begin{aligned} P_F\left(X_{(i_1)} \le \xi_p \le X_{(i_2)}\right) &= P_F\left(F(X_{(i_1)}) \le p \le F(X_{(i_2)})\right) \\ &= P\left(U_{(i_1)} \le p \le U_{(i_2)}\right) \ge 1 - \alpha \end{aligned}$$

if i_1, i_2 are appropriately chosen. The pair (i_1, i_2) can be chosen by studying the joint density of $(U_{(i_1)}, U_{(i_2)})$, which has an explicit formula. However, the formula involves incomplete Beta functions, and for certain n and α, the actual coverage can be substantially larger than $1-\alpha$. This is because no pair (i_1, i_2) may exist such that the event involving the two uniform order statistics has exactly or almost exactly $1 - \alpha$ probability. This will make the confidence interval $[X_{(i_1)}, X_{(i_2)}]$ larger than one wishes and therefore less useful.

Alternatively, under previously stated conditions,

$$\sqrt{n}(\widehat{\xi_p} - \xi_p) \stackrel{d}{\to} N\left(0, \frac{p(1-p)}{[F'(\xi_p)]^2}\right).$$

Hence, an asymptotically correct $1 - \alpha$ confidence interval for ξ_p is $\widehat{\xi_p} \pm \frac{z_\alpha}{\sqrt{n}} \frac{\sqrt{p(1-p)}}{F'(\xi_p)}$. This confidence interval typically will have an asymptotic $1-\alpha$ coverage probability. The interval has a simplistic appeal and is computed much more easily than the interval based on order statistics.

However, an obvious drawback of this procedure is that $F'(\xi_p)$ must be known in advance. Say, this method is *not* asymptotically distribution-free. A remedy is given as follows. Before proceeding, we need a refinement of Bahadur representation.

Theorem 3.3.4 *Let $X_1, \ldots, X_n$ be iid random variables from a continuous CDF F. Suppose that for $0 < p < 1$, $F'(\xi_p)$ exists and is positive. Let k_n be a sequence of integers satisfying $1 \le k_n \le n$ and $k_n/n = p + cn^{-1/2} + o(n^{-1/2})$ with a constant c. Then*

$$\sqrt{n}(X_{(k_n)} - \widehat{\xi_p}) \stackrel{p}{\to} \frac{c}{F'(\xi_p)}.$$

Proof. Let $t \in \mathbb{R}$, $\xi_{nt} = \xi_p + tn^{-1/2}$, $\eta_n = \sqrt{n}(\widehat{\xi}_{\frac{k_n}{n}} - \xi_p)$, $Z_n(t) = \sqrt{n}[F(\xi_{nt}) - F_n(\xi_{nt})]/F'(\xi_p)$, and $U_n(t) = \sqrt{n}[F(\xi_{nt}) - F_n(\widehat{\xi}_{\frac{k_n}{n}})]/F'(\xi_p)$. By using similar arguments in proving Theorem 3.3.3, it is not difficult to show that for any $t \in \mathbb{R}$ and $\epsilon > 0$,

$$P\left(\eta_n \leq t, Z_n(0) + \frac{c}{F'(\xi_p)} \geq t + \epsilon\right) \to 0$$

It follows that $\eta_n = Z_n(0) + \frac{c}{F'(\xi_p)} + o_p(1)$. Thus, we have

$$X_{(k_n)} - \xi_p = \frac{k_n/n - F_n(\xi_p)}{F'(\xi_p)} + o_p(n^{-1/2}).$$

By Theorem 3.3.3 again, we know

$$\widehat{\xi}_p - \xi_p = \frac{p - F_n(\xi_p)}{F'(\xi_p)} + o_p(n^{-1/2}).$$

The result follows by taking the difference of the two previous equations. □

Using this theorem, we can obtain an asymptotic $1 - \alpha$ confidence interval for ξ_p.

Corollary 3.3.3 *Assume the conditions in Theorem 3.3.4. Let* $\{k_{1n}\}$ *and* $\{k_{2n}\}$ *be two sequences of integers satisfying* $1 \leq k_{1n} < k_{2n} \leq n$,

$$k_{1n}/n = p - z_{\alpha/2}\sqrt{p(1-p)/n} + o(n^{-1/2})$$
$$k_{1n}/n = p + z_{\alpha/2}\sqrt{p(1-p)/n} + o(n^{-1/2}),$$

where $z_\alpha = \Phi^{-1}(1-\alpha)$. *Then, the confidence interval* $C(X) = [X_{(k_{1n})}, X_{(k_{2n})}]$ *has the property that* $P(\xi_p \in C(X))$ *does not depend on* F *and*

$$\lim_{n\to\infty} P(\xi_p \in C(X)) = 1 - \alpha.$$

Proof. Note that

$$\begin{aligned} P_F\left(X_{(k_{1n})} \leq \xi_p \leq X_{(k_{2n})}\right) &= P_F\left(F(X_{(k_{1n})}) \leq p \leq F(X_{(k_{2n})})\right) \\ &= P\left(U_{(k_{1n})} \leq p \leq U_{(k_{2n})}\right) \end{aligned}$$

and thus $P(\xi_p \in C(X))$ does not depend on F.

By Theorems 3.3.4, 3.3.2 and Slutsky's Theorem,

$$\begin{aligned} P(X_{(k_{1n})} > \xi_p) &= P\left(\widehat{\xi}_p - z_{\alpha/2}\frac{\sqrt{p(1-p)}}{F'(\xi_p)\sqrt{n}} + o_p(n^{-1/2}) > \xi_p\right) \\ &= P\left(\frac{\sqrt{n}(\widehat{\xi}_p - \xi_p)}{\sqrt{p(1-p)}/F'(\xi_p)} + o_p(1) > z_{\alpha/2}\right) \\ &\to 1 - \Phi(z_{\alpha/2}) = \alpha/2. \end{aligned}$$

Similarly, $P(X_{(k_{2n})} < \xi_p) \to \alpha/2$ which completes the proofs. □

3.3.4 Quantile regression

Least squares estimates in regression minimize the sum of squared deviations of the observed and the expected values of the dependent variable. In the location-parameter problem, this principle would result in the sample mean as the estimate. If instead one minimizes the sum of the absolute values of the deviations, one would obtain the median as the estimate. Likewise, one can estimate the regression parameters by minimizing the sum of the absolute deviations between the observed values and the regression function.

For example, if the model says $y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i$, then one can estimate the regression vector $\boldsymbol{\beta}$ by minimizing $\sum_{i=1}^n |y_i - \mathbf{x}_i^T\boldsymbol{\beta}|$, a very natural idea. This estimate is called the least absolute deviation (LAD) regression estimate. While it is not as good as the least squares estimate when the errors are exactly normal, it outperforms the least squares estimate for a variety of error distributions that are heavy-tailed. Generalizations of the LAD estimate, analogous to sample percentiles, are called quantile regression estimate. A good reference for the material in this section and proofs of theorems below is Koenker (2005).

Definition 3.3.1 *For $0 < p < 1$, the pth quantile regression estimate is defined as*

$$\widehat{\beta}_{QR} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n p|y_i - \mathbf{x}_i^T\boldsymbol{\beta}|I_{\{y_i \geq \mathbf{x}_i^T\boldsymbol{\beta}\}} + (1-p)|y_i - \mathbf{x}_i^T\boldsymbol{\beta}|I_{\{y_i < \mathbf{x}_i^T\boldsymbol{\beta}\}}.$$

We always write the equivalent definition

$$\widehat{\beta}_{QR} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n \rho_p(y_i - \mathbf{x}_i^T\boldsymbol{\beta}),$$

where $\rho_p(t) = pt_+ + (1-p)t_-$ is the so-called check function where subscripts + and − stand for the positive and negative parts, respectively. The following theorem describe the limiting distribution of quantile regression estimate. There are some neat analogies in this result to limiting distribution of the sample quantile for iid data.

Theorem 3.3.5 *Let* $y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i$, *where* $\epsilon_i \overset{iid}{\sim} F$, *with* F *having median zero. Let* $0 < p < 1$, *and let* $\widehat{\boldsymbol{\beta}}_{QR}$ *be any pth quantile regression estimate. Suppose* F *has a strictly positive derivative* $f(\xi_p)$ *at* ξ_p. *Then,*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{QR} - \boldsymbol{\beta} - \xi_p\mathbf{e}_1) \xrightarrow{d} N_p(\mathbf{0}, \nu\boldsymbol{\Sigma}^{-1}),$$

where $\mathbf{e}_1 = (1, 0, \ldots, 0)^T$, $\boldsymbol{\Sigma} = \lim_n \frac{1}{n}\mathbf{X}^T\mathbf{X}$ *(assumed to exist), and* $\nu = \frac{p(1-p)}{f^2(\xi_p)}$.

References

Bahadur, R. R. (1966). A note on quantiles in large samples, *Ann. Math. Stat.*, 37, 577–580.

Ghosh, J. K. (1971). A new proof of the Bahadur representation of quantiles and an application, *Ann. Math. Stat.*, 42, 1957–1961.

Kiefer, J. (1967). On Bahadurs representation of sample quantiles, *Ann. Math. Stat.*, 38, 1323–1342.

Koenker, R. (2005). *Quantile Regression*. Cambridge Univ. Press.

Chapter 4

Asymptotics in parametric inference

In this chapter, we treats asymptotic statistics which arises in connection with estimation or hypothesis testing relative to a parametric family of possible distributions for the data. In this respect, maximum likelihood inference might be one of the most popular methods. Many think that maximum likelihood is the greatest conceptual invention in the history of statistics. Although in some high-or infinite-dimensional problems, computation and performance of maximum likelihood estimates (MLEs) are problematic, in a vast majority of models in practical use, MLEs are about the best that one can do. They have many asymptotic optimality properties that translate into fine performance in finite samples. Before elaborating on maximum likelihood estimates and testings, we first consider the concept of asymptotic optimality of point estimators in parametric models.

4.1 Asymptotic efficient estimation

Let $\widehat{\boldsymbol{\theta}}_n$ be a sequence of estimators of $\boldsymbol{\theta}$ based on a sequence of samples $\mathbf{X} = \{X_1, \ldots, X_n\}$ whose distributions are in a parametric family indexed by $\boldsymbol{\theta}$. Suppose that as $n \to \infty$

$$(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \sim AN_k(\mathbf{0}, \mathbf{V}_n(\boldsymbol{\theta})), \tag{4.1}$$

where for each n, $\mathbf{V}_n(\boldsymbol{\theta})$ is a $k \times k$ positive definite matrix depending on $\boldsymbol{\theta}$. If $\boldsymbol{\theta}$ is one-dimensional ($k = 1$), then $\mathbf{V}_n(\theta)$ is the asymptotic variance as well as the asymptotic MSE of $\widehat{\theta}_n$. When $k > 1$, $\mathbf{V}_n(\boldsymbol{\theta})$ is called the *asymptotic covariance matrix* of $\widehat{\boldsymbol{\theta}}_n$ and can be used as a measure of asymptotic performance of estimators.

If $\widehat{\boldsymbol{\theta}}_{jn}$ satisfies (4.1) with asymptotic covariance matrix $\mathbf{V}_{jn}(\boldsymbol{\theta}), j = 1, 2$, and $\mathbf{V}_{1n}(\boldsymbol{\theta}) \leq \mathbf{V}_{2n}(\boldsymbol{\theta})$ (in the sense that $\mathbf{V}_{2n}(\boldsymbol{\theta}) - \mathbf{V}_{1n}(\boldsymbol{\theta})$ is nonnegative definite) for all $\boldsymbol{\theta} \in \Theta$, then $\widehat{\boldsymbol{\theta}}_{1n}$ is said to be *asymptotically more efficient* than $\widehat{\boldsymbol{\theta}}_{2n}$. When X_i's are iid, $\mathbf{V}_n(\boldsymbol{\theta})$ is usually of the form $n^{-\delta}\mathbf{V}(\boldsymbol{\theta})$ for some $\delta > 0$ (=1 in the majority of cases) and a positive definite matrix $\mathbf{V}(\boldsymbol{\theta})$ that does not depend on n.

Definition 4.1.1 *Assume that the Fisher information matrix*

$$\mathbf{I}_n(\boldsymbol{\theta}) = E\left\{\frac{\partial}{\partial \boldsymbol{\theta}} \sum_i \log f_{\boldsymbol{\theta}}(X_i) \left[\frac{\partial}{\partial \boldsymbol{\theta}} \sum_i \log f_{\boldsymbol{\theta}}(X_i)\right]^T\right\}$$

is well defined and positive definite for every n. A sequence of estimators $\widehat{\boldsymbol{\theta}}_n$ satisfying (4.1) is said to be asymptotically efficient or asymptotically optimal iff $\mathbf{V}_n(\boldsymbol{\theta}) = [\mathbf{I}_n(\boldsymbol{\theta})]^{-1}$.

Suppose that we are interested in estimating $\boldsymbol{\beta} = g(\boldsymbol{\theta})$, where g is a differentiable function from $\mathbb{R}^k$ to $\mathbb{R}^p, 1 \leq p \leq k$. If $\widehat{\boldsymbol{\theta}}_n$ satisfies (4.1), then by Delta Theorem, $\widehat{\boldsymbol{\beta}}_n = g(\widehat{\boldsymbol{\theta}}_n)$ is asymptotically distributed as $N_p(\boldsymbol{\beta}, [\nabla g(\boldsymbol{\theta})]^T \mathbf{V}_n(\boldsymbol{\theta}) \nabla g(\boldsymbol{\theta}))$. Thus, the information inequality

$$[\nabla g(\boldsymbol{\theta})]^T \mathbf{V}_n(\boldsymbol{\theta}) \nabla g(\boldsymbol{\theta}) \geq [\tilde{\mathbf{I}}_n(\boldsymbol{\beta})]^{-1},$$

where $\tilde{\mathbf{I}}_n(\boldsymbol{\beta})$ is the Fisher information matrix about $\boldsymbol{\beta}$. If $p = k$ and g is one-to-one, then

$$[\tilde{\mathbf{I}}_n(\boldsymbol{\beta})]^{-1} = [\nabla g(\boldsymbol{\theta})]^T [\mathbf{I}_n(\boldsymbol{\theta})]^{-1} \nabla g(\boldsymbol{\theta}),$$

and, therefore, $\widehat{\boldsymbol{\beta}}_n$ is asymptotically efficient iff $\widehat{\boldsymbol{\theta}}_n$ is asymptotically efficient. For this reason, we can focus on the estimation of $\boldsymbol{\theta}$ only.

Remark 4.1.1 (The super-efficiency and Hodges estimator)

It was first believed as folklore that the MLE under regularity conditions on the underlying distribution is asymptotically the best for every value of $\theta_0 \in \Theta$; i.e., if an MLE

$\widehat{\theta}_n$ exists and $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0))$, and if another competing sequence Tn satisfies $\sqrt{n}(T_n - \theta_0) \xrightarrow{d} N(0, V(\theta_0))$, then for every θ_0, $V(\theta_0) \geq I^{-1}(\theta_0)$.

It was a major shock when in 1952 Hodges gave an example that destroyed this belief and proved it to be false even in the normal case. Hodges, in a private communication to LeCam, produced an estimate T_n that beats the MLE $\bar{X}_n$ locally at some θ_0, say $\theta_0 = 0$. Later, in a very insightful result, LeCam (1953) showed that this can happen only on Lebesgue-null sets of θ. An excellent reference for this topic is van der Vaart (1998).

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$. Define an estimate T_n as

$$\tilde{\theta} = \begin{cases} \bar{X}_n, & \bar{X}_n \geq n^{-1/4}, \\ t\bar{X}_n, & \bar{X}_n < n^{-1/4}, \end{cases}$$

where we choose $0 < t < 1$. We are interested in estimating the population mean. If $\bar{X}_n$ is not close to 0, we simply take the sample mean as the estimator. If we know that it is pretty close to 0, we can shrink it further to make it closer to 0. Thus, the resulting estimator should be more efficient than the sample mean $\bar{X}_n$ at 0. Of course, we can take other values than 0, the same thing will happen too. Now, let's find the asymptotic distribution of $\tilde{\theta}$. If $\theta = 0$, then we can write

$$\begin{aligned} \sqrt{n}\tilde{\theta} &= \sqrt{n}\left[\bar{X}_n I_{\{|\bar{X}_n| \geq n^{-1/4}\}} + t\bar{X}_n I_{\{|\bar{X}_n| < n^{-1/4}\}}\right] \\ &= \sqrt{n}\left[t\bar{X}_n + (1-t)\bar{X}_n I_{\{|\bar{X}_n| \geq n^{-1/4}\}}\right] \\ &= tY_n + (1-t)Y_n I_{\{|Y_n| \geq n^{1/4}\}} \end{aligned}$$

where $Y_n = \sqrt{n}\bar{X}_n \sim N(0,1)$, and hence $tY_n \sim N(0, t^2)$. Now let us look at the second term $W_n = Y_n I_{\{|Y_n| \geq n^{1/4}\}}$. Since

$$\begin{aligned} (E|W_n|)^2 &\leq E(Y_n^2)E(I^2_{\{|Y_n| \geq n^{1/4}\}}) \\ &= P(|Y_n| \geq n^{1/4}) \leq E|Y_n|^2/n^{1/2} = n^{-1/2} \to 0, \end{aligned}$$

which implies $W_n \xrightarrow{p} 0$. By Slutsky's theorem, we get

$$\sqrt{n}\tilde{\theta} \xrightarrow{d} N(0, t^2), \quad \text{if} \quad \theta = 0.$$

Similarly, when $\theta \neq 0$, we can write

$$\sqrt{n}\tilde{\theta} = Y_n + (t-1)Y_n I_{\{|Y_n|<n^{1/4}\}}.$$

Again, $Y_n - \sqrt{n}\theta = \sqrt{n}(\bar{X}_n - \theta) \sim N(0,1)$. Now it remains to show that $Y_n I_{\{|Y_n|<n^{1/4}\}} \xrightarrow{p} 0$. For any $0 < \epsilon < n^{1/4}$,

$$\begin{aligned}
P(|Y_n I_{\{|Y_n|<n^{1/4}\}}| > \epsilon) &\leq P(n^{1/4} I_{\{|Y_n|<n^{1/4}\}} > \epsilon) \\
&= P(I_{\{|Y_n|<n^{1/4}\}} > \epsilon/n^{1/4}) \\
&= P(|Y_n| < n^{1/4}) \\
&= P(-n^{1/4} < Y_n < n^{1/4}) \\
&= \Phi(-\sqrt{n}\theta + n^{1/4}) - \Phi(-\sqrt{n}\theta - n^{1/4}) \to 0.
\end{aligned}$$

By Slutsky's theorem again, we get $\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{d} N(0,1)$. Combining the above two cases, we get

$$\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{d} \begin{cases} N(0, t^2), & \theta = 0, \\ N(0,1), & \theta \neq 0, \end{cases}$$

In the case of $\theta = 0$, the usual asymptotic Cramer-Rao theorem does not hold, since $t^2 < 1 = I^{-1}(\theta)$. It is clear, however, that T_n has certain undesirable features. First, as a function of $X_1, ..., X_n$, T_n is not smooth. Second, $V(\theta)$ is not continuous in θ.

4.2 Maximum likelihood estimation

Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be iid with distribution $F_{\boldsymbol{\theta}}$ belonging to a family $\mathcal{F} = \{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T \in \Theta\}$ and suppose that the distribution $F_{\boldsymbol{\theta}}$ posses densities $f_{\boldsymbol{\theta}}(x)$. The likelihood function of the sample $\mathbf{X}$ is defined as

$$L(\boldsymbol{\theta}; \mathbf{X}) = \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(X_i).$$

The maximum likelihood estimate (MLE) is given by $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \log L(\boldsymbol{\theta}; \mathbf{X})$. Often, the estimate $\widehat{\boldsymbol{\theta}}$ may be obtained by solving the system of likelihood equations (score function),

$$\frac{\partial \log L}{\partial \theta_i}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = 0$$

and check that the solution $\widehat{\boldsymbol{\theta}}$ indeed maximizes L. Since the solutions of likelihood equations may not be the MLE, we also always term the *root of the likelihood equation* as RLE.

Next, we will show that under regularity condition on $\mathcal{F}$, the MLE (RLE) are strongly consistent, asymptotically normal, and asymptotically efficient. For simplicity, we focus on the case of $k = 1$. The multivariate version will be discussed without giving it proof.

Regularity Condition of $\mathcal{F}$ Consider Θ to be an open interval in $\mathbb{R}$. Assume:

(C1) The third derivative with respect to θ, $\frac{\partial^3 \log f_\theta(x)}{\partial \theta^3}$, exists for all x, and also for each $\theta_0 \in \Theta$ there exists a function $H(x) \geq 0$ (possibly depending on θ_0) such that for $\theta \in N(\theta_0, \epsilon) = \{\theta : |\theta - \theta_0| < \epsilon\}$,

$$\left|\frac{\partial^3 \log f_\theta(x)}{\partial \theta^3}\right| \leq H(x), \quad E_\theta H(X_1) < \infty;$$

(C2) For $g_\theta(x) = f_\theta(x)$ or $g_\theta(x) = \frac{\partial f_\theta(x)}{\partial \theta}$, we have

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

(C3) For each $\theta \in \Theta$, we have

$$0 < I(\theta) = E_\theta\left(\frac{\partial \log f_\theta(x)}{\partial \theta}\right)^2 < \infty.$$

Remark 4.2.1 Condition (C1) ensures that $\frac{\partial \log f_\theta(x)}{\partial \theta}$, for any x, has a Taylor's expansion as a function of θ; Condition 2 means that $f_\theta(x)$ or $\frac{\partial f_\theta(x)}{\partial \theta}$ can be differentiated with respect to θ under the integral sign. That is, the integration and differentiation can be interchanged; A sufficient condition for Condition 2 is the following:

For each $\theta_0 \in \Theta$, there exists functions $g(x)$, $h(x)$, and $H(x)$ (possibly depending on θ_0) such that for $\theta \in N(\theta_0, \epsilon) = \{\theta : |\theta - \theta_0| < \epsilon\}$,

$$\left|\frac{\partial f_\theta(x)}{\partial \theta}\right| \leq g(x), \quad \left|\frac{\partial^2 f_\theta(x)}{\partial \theta^2}\right| \leq h(x), \quad \left|\frac{\partial^3 \log f_\theta(x)}{\partial \theta^3}\right| \leq H(x)$$

hold for all x and

$$\int g(x)dx < \infty, \quad \int h(x)dx < \infty, \; E_\theta H(X_1) < \infty;$$

Condition 3 ensures that the variance of $\frac{\partial \log f_\theta(x)}{\partial \theta}$ is finite.

Theorem 4.2.1 *Assume regularity conditions (C1)-(C3) on the family $\mathcal{F}$. Consider iid observations on F_{θ_0}, for θ_0 an element of Θ. Then with probability 1, the likelihood equations admit a sequence of solutions $\{\widehat{\theta}_n\}$ satisfying*

(i) strong consistency: $\widehat{\theta}_n \to \theta_0$, as $n \to \infty$;

(ii) asymptotic normality and efficiency: $\widehat{\theta}_n$ is $AN(\theta_0, [nI(\theta_0)]^{-1})$.

Proof. (i) Denote the score function by

$$s(\mathbf{X}, \theta) = \frac{1}{n}\frac{\partial \log L(\theta; \mathbf{X})}{\partial \theta} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f_\theta(X_i)}{\partial \theta}.$$

Then,

$$s'(\mathbf{X}, \theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f_\theta(X_i)}{\partial \theta^2}, \quad s''(\mathbf{X}, \theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^3 \log f_\theta(X_i)}{\partial \theta^3}.$$

Note that

$$|s''(\mathbf{X}, \theta)| \leq \frac{1}{n}\sum_{i=1}^{n} \left|\frac{\partial^3 \log f_\theta(X_i)}{\partial \theta^3}\right| \leq \frac{1}{n}\sum_{i=1}^{n} |H(X_i)| \equiv \bar{H}(\mathbf{X}),$$

where $\bar{H}(\mathbf{X}) = n^{-1}\sum_{i=1}^{n} H(X_i)$. By Taylor's expansion

$$\begin{aligned} s(\mathbf{X}, \theta) &= s(\mathbf{X}, \theta_0) + s'(\mathbf{X}, \theta_0)(\theta - \theta_0) + \frac{1}{2}s''(\mathbf{X}, \xi)(\theta - \theta_0)^2 \\ &= s(\mathbf{X}, \theta_0) + s'(\mathbf{X}, \theta_0)(\theta - \theta_0) + \frac{1}{2}\bar{H}(\mathbf{X})\eta^*(\theta - \theta_0)^2, \end{aligned}$$

where $|\eta^*| = |s''(\mathbf{X}, \xi)|/\bar{H}(\mathbf{X}) \leq 1$. By the SLLN, we have,

$$s(\mathbf{X}, \theta_0) \stackrel{wp1}{\to} E_{\theta_0} s(\mathbf{X}, \theta_0) = 0,$$
$$s'(\mathbf{X}, \theta_0) \stackrel{wp1}{\to} E_{\theta_0} s'(\mathbf{X}, \theta_0) = -I(\theta_0),$$
$$\bar{H}(\mathbf{X}) \stackrel{wp1}{\to} E_{\theta_0} H(X_i) < \infty,$$

where we use the fact that

$$E_\theta \left(\frac{\partial \log f_\theta(x)}{\partial \theta} \right) = \int \frac{1}{f_\theta(x)} \frac{\partial f_\theta(x)}{\partial \theta} f_\theta(x) dx = \frac{\partial}{\partial \theta} \int f_\theta(x) dx = 0$$
$$E_\theta \left(\frac{\partial^2 \log f_\theta(x)}{\partial \theta^2} \right) = \int \left[\frac{1}{f_\theta(x)} \frac{\partial^2 f_\theta(x)}{\partial \theta^2} - \left(\frac{1}{f_\theta(x)} \frac{\partial f_\theta(x)}{\partial \theta} \right)^2 \right] f_\theta(x) dx$$
$$= -E \left(\frac{\partial \log f_\theta(x)}{\partial \theta} \right)^2$$

provided Condition (C2) holds.

Clearly, for $\epsilon > 0$, we have with probability one,

$$s(\mathbf{X}, \theta_0 \pm \epsilon) = s(\mathbf{X}, \theta_0) + s'(\mathbf{X}, \theta_0)(\pm\epsilon) + \frac{1}{2} \bar{H}(\mathbf{X}) \eta^* (\pm\epsilon)^2$$
$$\approx \mp I(\theta_0)\epsilon + \frac{1}{2} E_{\theta_0} H(X_1) c\epsilon^2, \ |c| < 1.$$

In particular, we choose $0 < \epsilon < I(\theta_0)/E_{\theta_0} H(X_1)$. Then for large enough n, we have, with probability 1,

$$s(\mathbf{X}, \theta_0 + \epsilon) = s(\mathbf{X}, \theta_0) + s'(\mathbf{X}, \theta_0)\epsilon + \frac{1}{2} \bar{H}(\mathbf{X}) \eta^* \epsilon^2 \leq -I(\theta_0)\epsilon + \frac{1}{2} E_{\theta_0} H(X_1) c\epsilon^2 < 0$$
$$s(\mathbf{X}, \theta_0 - \epsilon) = s(\mathbf{X}, \theta_0) - s'(\mathbf{X}, \theta_0)\epsilon + \frac{1}{2} \bar{H}(\mathbf{X}) \eta^* \epsilon^2 \geq I(\theta_0)\epsilon - \frac{1}{2} E_{\theta_0} H(X_1) c\epsilon^2 > 0.$$

Therefore, by the continuity of $s(\mathbf{X}, \theta)$, for such n, the interval $[\theta_0 - \epsilon, \theta_0 + \epsilon]$ contains a solution of the likelihood equation $s(\mathbf{X}, \theta) = 0$. In particular, it contains the solution

$$\widehat{\theta}_{n,\epsilon} = \inf\{\theta : \theta_0 - \epsilon \leq \theta \leq \theta_0 + \epsilon, \ \text{ and } \ s(\mathbf{X}, \theta) = 0\}.$$

It can be shown that $\widehat{\theta}_{n,\epsilon}$ is a proper random variable. It can be also shown that we can obtain a sequence of $\widehat{\theta}_n$ not depending on the choice of ϵ. The details are omitted here but can be found in Serfling (1980). This proves (i).

(ii) For large n, we have seen that

$$0 = s(\mathbf{X}, \widehat{\theta}_n) = s(\mathbf{X}, \theta_0) + s'(\mathbf{X}, \theta_0)(\widehat{\theta}_n - \theta_0) + \frac{1}{2}\bar{H}(\mathbf{X})\eta^*(\widehat{\theta}_n - \theta_0)^2.$$

Thus,

$$\sqrt{n}s(\mathbf{X}, \theta_0) = \sqrt{n}(\widehat{\theta}_n - \theta_0)\left(-s'(\mathbf{X}, \theta_0) - \frac{1}{2}\bar{H}(\mathbf{X})\eta^*(\widehat{\theta}_n - \theta_0)\right).$$

Since $\sqrt{n}s(\mathbf{X}, \theta_0) \xrightarrow{d} N(0, I(\theta_0))$ by CLT, and $-s'(\mathbf{X}, \theta_0) - \frac{1}{2}\bar{H}(\mathbf{X})\eta^*(\widehat{\theta}_n - \theta_0) \xrightarrow{wp1} I(\theta_0)$, then it follows from Slutsky's theorem that

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \frac{\sqrt{n}s(\mathbf{X}, \theta_0)}{-s'(\mathbf{X}, \theta_0) - \frac{1}{2}\bar{H}(\mathbf{X})\eta^*(\widehat{\theta}_n - \theta_0)} \xrightarrow{d} N(0, I(\theta_0))/I(\theta_0) = N(0, I^{-1}(\theta_0)).$$

For the case of $\boldsymbol{\theta} \in \mathbb{R}^k$, under appropriate generalization of the Conditions (C1)-(C3), there exists a sequence of $\widehat{\boldsymbol{\theta}}_n$ of solutions to $s(\mathbf{X}, \boldsymbol{\theta})$ such that $\widehat{\boldsymbol{\theta}}_n \xrightarrow{wp1} \boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}_n$ is $AN(\boldsymbol{\theta}_0, \mathbf{I}_n^{-1}(\boldsymbol{\theta}_0))$, where $\mathbf{I}_n^{-1}(\boldsymbol{\theta}_0)$ is the information matrix defined in Definition 4.1.1. A succinct proof can be found on page 290 of Shao (2003).

Remark 4.2.2 This theorem does not say which sequence of roots of $s(\mathbf{X}; \theta) = 0$ should be chosen to ensure consistency in the case of multiple roots. It does not even guarantee that for any given n, however large, the likelihood function $\log L(\theta; \mathbf{X})$ has any local maxima at all. This specific theorem is useful in only those cases where $s(\mathbf{X}; \theta) = 0$ has a unique root for all n.

4.3 Improving the sub-efficient estimates

The method of moments ordinarily provides asymptotically normal estimates. Sometimes these estimates are asymptotically efficient. For example, in estimating (μ, σ^2) in $N(\mu, \sigma^2)$ by $(\bar{X}_n, S_n^2)$, the method of moments and MLE coincide but usually they are not. One would like to use MLE, but this has the disadvantage of being difficult to evaluate in general. The likelihood equations, $s(\mathbf{X}, \theta) = 0$, are generally highly nonlinear and one must to numerical approximation methods to solve them.

One good strategy is to use *Newton's method* with one of simply computed estimates based on the method of moments or sample quantiles as the initial guess. This method takes the initial guess, $\widehat{\theta}^{(0)}$, and inductively generates a sequence of hopefully better and better estimates by

$$\widehat{\theta}^{(k+1)} = \widehat{\theta}^{(k)} - [s'(\mathbf{X}, \widehat{\theta}^{(k)})]^{-1} s(\mathbf{X}, \widehat{\theta}^{(k)}), k = 0, 1, 2, \ldots.$$

One simplification of this strategy can be made if the Fisher information is available. Ordinarily, $s'(\mathbf{X}, \widehat{\theta}^{(k)})$ will converge as $n \to \infty$ to $-I(\theta_0)$ and so can be replaced by $-I(\widehat{\theta}^{(k)})$ in the iterations,

$$\widehat{\theta}^{(k+1)} = \widehat{\theta}^{(k)} + [I(\widehat{\theta}^{(k)})]^{-1} s(\mathbf{X}, \widehat{\theta}^{(k)}), k = 0, 1, 2, \ldots.$$

As we know, this method is the *method of scoring*. The scores, $[I(\widehat{\theta}^{(k)})]^{-1} s(\mathbf{X}, \widehat{\theta}^{(k)})$ are increments added to an estimate to improve it.

Example 4.3.1 (**Logistic distribution**) Let $X_1, \ldots, X_n$ be a sample from density

$$f_\theta(x) = \frac{\exp\{-(x-\theta)\}}{(1+\exp\{-(x-\theta)\})^2}.$$

The log-likelihood function is given by

$$l_n(\theta) = -\sum_{j=1}^{n}(X_j - \theta) - 2\sum_{j=1}^{n}\log\left(1+\exp\{-(X_j-\theta)\}\right)$$

and the likelihood equations are

$$l_n'(\theta) = n - 2\sum_{j=1}^{n}\frac{1}{1+\exp\{X_j-\theta\}} = 0.$$

Newton's method is easy to apply here because

$$l_n''(\theta) = -2\sum_{j=1}^{n}\frac{\exp\{X_j-\theta\}}{(1+\exp\{X_j-\theta\})^2}.$$

Even easier is the method of scoring, since $I(\theta) = \frac{1}{3}$ [$I(\theta)$ is a constant for location parameter families of distributions.] As an initial guess we may use the sample median, m_n, or the sample mean, $\bar{X}_n$. The asymptotic distributions are

$$\sqrt{n}(m_n - \theta) \xrightarrow{d} N\left(0, \frac{1}{4f_\theta(\theta)^2}\right) = N(0, 4),$$

$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{d} N\left(0, \frac{\pi^2}{3}\right) \approx N(0, 3.2899).$$

Since for the MLE, $\widehat{\theta}_n$,

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, I(\theta)^{-1}) = N(0, 3),$$

it would seem worthwhile to improve m_n and $\bar{X}_n$ by once iteration or two of

$$\widehat{\theta}^{(k+1)} = \widehat{\theta}^{(k)} + 3\left[1 - \frac{2}{n}\sum_{j=1}^{n}\frac{1}{1+\exp\{X_j - \widehat{\theta}^{(k)}\}}\right].$$

Once is enough. In improving asymptotically normal estimates by scoring, one iteration is generally enough to achieve asymptotically efficiency! Let $\widehat{\theta}^{(0)}$ be an estimator that is sub-efficient. The estimator

$$\widehat{\theta}^{(1)} = \widehat{\theta}^{(0)} - [s'(\mathbf{X}, \widehat{\theta}^{(0)})]^{-1} s(\mathbf{X}, \widehat{\theta}^{(0)}), k = 0, 1, 2, \ldots.$$

is the first iteration in improving the estimator as discussed above. In fact, this is just the first iteration in computing and MLE using the Newton-Raphson iteration method with $\widehat{\theta}^{(0)}$ as the initial value and, hence, is often called the *one-step MLE*. Under some conditions, $\widehat{\theta}^{(1)}$ is asymptotically efficient, as the following result shows.

Theorem 4.3.1 *Assume the conditions in Theorem 4.2.1 hold and that $\widehat{\theta}^{(0)}$ is $\sqrt{n}$-consistent for θ. Then*

(i) The one-step MLE $\widehat{\theta}^{(1)}$ is asymptotically efficient;

(ii) The one-step MLE obtained by replacing $s'(\mathbf{X}, \widehat{\theta}^{(0)})$ with its expected value, $-I(\widehat{\theta}^{(0)})$, is asymptotically efficient.

Proof. Let $\widehat{\theta}_n$ be a $\sqrt{n}$-consistent sequence satisfying $s(\mathbf{X}; \widehat{\theta}_n) = 0$. In what follows, we suppress "$\mathbf{X}$" for simplicity. Expanding $\widehat{\theta}^{(0)}$ at $\widehat{\theta}_n$,

$$s(\widehat{\theta}^{(0)}) = s(\widehat{\theta}_n) + s'(\widehat{\theta}_n)(\widehat{\theta}^{(0)} - \widehat{\theta}_n) + \frac{1}{2}s''(\xi)(\widehat{\theta}^{(0)} - \widehat{\theta}_n)^2, \tag{4.2}$$

and using

$$(\widehat{\theta}^{(1)} - \widehat{\theta}_n) = (\widehat{\theta}^{(0)} - \widehat{\theta}_n) - [s'(\widehat{\theta}^{(0)})]^{-1} s(\widehat{\theta}^{(0)}),\ s(\widehat{\theta}_n) = 0$$

we find

$$\sqrt{n}(\widehat{\theta}^{(1)} - \widehat{\theta}_n) = \sqrt{n}s(\widehat{\theta}^{(0)}) \left\{ [s'(\widehat{\theta}_n)]^{-1} - [s'(\widehat{\theta}^{(0)})]^{-1} \right\} - \frac{\sqrt{n}}{2} s''(\xi)(\widehat{\theta}^{(0)} - \widehat{\theta}_n)^2 [s'(\widehat{\theta}_n)]^{-1}. \tag{4.3}$$

Now, we need to study the right hand of (4.3). Firstly, note that the term $\sqrt{n}(\widehat{\theta}^{(0)} - \widehat{\theta}_n) = \sqrt{n}(\widehat{\theta}^{(0)} - \theta_0) - \sqrt{n}(\widehat{\theta}_n - \theta_0)$, is bounded in probability because the second term is asymptotically normal from Theorem 4.2.1-(ii) and the first term is $O_p(1)$ by the assumption. By

$$|s''(\xi)| \leq \bar{H}(X) \overset{wp1}{\to} E_{\theta_0} H(X_i) < \infty$$

we have $s''(\xi) = O_p(1)$. Also, by CMT, we know $s'(\widehat{\theta}_n) = s'(\theta_0) + o_p(1) = -I(\theta_0) + o_p(1)$. Thus, the last term in (4.3) is of order $O_p(n^{-1/2})$. Similarly, $[s'(\widehat{\theta}_n)]^{-1} - [s'(\widehat{\theta}^{(0)})]^{-1} = o_p(1)$. Finally, by (4.2) again, we obtain $s(\widehat{\theta}^{(0)}) = O_p(n^{-1/2})$, which leads to

$$\sqrt{n}(\widehat{\theta}^{(1)} - \widehat{\theta}_n) = \sqrt{n} O_p(n^{-1/2}) o_p(1) + O_p(n^{-1/2}) = o_p(1).$$

Hence, $\sqrt{n}(\widehat{\theta}^{(1)} - \widehat{\theta}_n) \xrightarrow{p} 0$ as $n \to \infty$. Say, $\sqrt{n}(\widehat{\theta}^{(1)} - \widehat{\theta}_n) = \sqrt{n}(\widehat{\theta}^{(1)} - \theta_0) - \sqrt{n}(\widehat{\theta}_n - \theta_0) = o_p(1)$. $\sqrt{n}(\widehat{\theta}^{(1)} - \theta_0)$ is asymptotically equivalent to $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ which is asymptotically efficient according to Theorem 4.2.1. It follows that $\sqrt{n}(\widehat{\theta}^{(1)} - \theta_0)$ is $AN(\theta_0, [I(\theta_0)]^{-1})$ and thus asymptotically efficient. The agrement for estimate using scoring method is identical. □

4.4 Hypothesis testing by likelihood method

As we know, UMP and UMPU tests often do not exist in a particular problem. In this chapter, we shall introduce other tests. These tests may not be optimal, but they are very general methods, easy to use, and have intuitive appeal. They often coincide with optimal tests (UMP, UMPU tests). They play similar role to the MLE in the estimation theory. For all these reasons, a treatment of testing is essential. We discuss the asymptotic theory of likelihood ratio, Wald, and Rao score tests in the remainder of this chapter.

Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be iid with distribution $F_{\boldsymbol{\theta}}$ belonging to a family $\mathcal{F} = \{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T \in \Theta \subset \mathbb{R}^k\}$ and suppose that the distribution $F_{\boldsymbol{\theta}}$ possess densities $f_{\boldsymbol{\theta}}(x)$. The

testing problem is

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \in \Theta_1,$$

where $\Theta_0 \bigcup \Theta_1 = \Theta$ and $\Theta_0 \bigcap \Theta_1 = \emptyset$.

The likelihood ratio test (LRT) rejects H_0 for small values of

$$\Lambda_n = \frac{\sup_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta}; \mathbf{X})}{\sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; \mathbf{X})}$$

Equivalently, the test may be carried out in terms of the commonly used statistic

$$\lambda_n = -2 \log \Lambda_n,$$

which turns out to be more convenient for asymptotic derivation. The motivation for Λ_n comes from two sources: (a) The case where H_0, and H_1 are each simple, for which a UMP test is found from Λ_n by the Neyman-Pearson lemma; (b) The intuitive explanation that, for small values of Λ_n, we can better match the observed data with some value of $\boldsymbol{\theta}$ outside of Θ_0.

A null hypothesis H_0 will be specified as a subset Θ_0 of Θ, where Θ_0 is determined by a set of $r \leq k$ restrictions given by equations

$$R_i(\boldsymbol{\theta}) = 0, \quad 1 \leq i \leq r.$$

In the case of a *simple* hypothesis $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, the set $\Theta_0 = \{\boldsymbol{\theta}_0\}$, and the function $R_i(\boldsymbol{\theta})$ may be taken to be

$$R_i(\boldsymbol{\theta}) = \theta_i - \theta_{0i}, \quad 1 \leq i \leq k.$$

In the case of a *composite* hypothesis, the set Θ_0 contains more than one element and we must have $r < k$. For instance, $k = 3$, we might have $H_0 : \boldsymbol{\theta}_0 \in \Theta_0 = \{\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) : \theta_1 = \theta_{01}\}$. In this case, $r = 1$ and the function $R_1(\boldsymbol{\theta})$ may be taken to be $R_1(\boldsymbol{\theta}) = \theta_1 - \theta_{01}$. We start with a well-known but intuitive example that illustrate important aspects of the likelihood ratio method.

Example 4.4.1 Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, and consider testing $H_0 : \mu = 0$ versus $H_1 :$

$\mu \neq 0$. Let $\boldsymbol{\theta} = (\mu, \sigma^2)^T$. Then, $k = 2$, $r = 1$. Apparently,

$$\begin{aligned}\Lambda_n &= \frac{\sup_{\boldsymbol{\theta}\in\Theta_0}(1/\sigma^n)\exp\{-\frac{1}{2\sigma^2}\sum_i(X_i-\mu)^2\}}{\sup_{\boldsymbol{\theta}\in\Theta_1}(1/\sigma^n)\exp\{-\frac{1}{2\sigma^2}\sum_i(X_i-\mu)^2\}} \\ &= \left(\frac{\sum_i(X_i-\bar{X}_n)^2}{\sum_i X_i^2}\right)^{n/2}\end{aligned}$$

by an elementary calculation of MLEs of $\boldsymbol{\theta}$ under H_0 and in the general parameter space. By another elementary calculation, $\Lambda_n < c$ is seen to be equivalent to $t_n^2 > k$, where

$$t_n = \frac{\sqrt{n}\bar{X}_n}{\sqrt{\frac{1}{n-1}\sum_i(X_i-\bar{X}_n)^2}}.$$

is the t-statistic. In other words, the t-test is the LRT (equivalently). Also, observe that

$$t_n^2 = (n-1)\Lambda_n^{-2/n} - (n-1).$$

This implies

$$\begin{aligned}\Lambda_n &= \left(\frac{n-1}{t_n^2+n-1}\right)^{n/2} \\ \Rightarrow \lambda_n &= -2\log\Lambda_n = n\log\left(1+\frac{t_n^2}{n-1}\right) \\ &= n\left(\frac{t_n^2}{n-1} + o_p(\frac{t_n^2}{n-1})\right) \xrightarrow{d} \chi_1^2.\end{aligned}$$

under H_0 since $t_n \xrightarrow{d} N(0,1)$ as illustrated in Example 1.2.8.

As seen earlier, sometimes it is very difficult or impossible to find the exact distribution of λ_n. So approximations in these cases become necessary. The next celebrated theorem originally stated Wilks (1938), established the asymptotic chi-square distribution of λ_n under H_0. The degree of freedom is just the number of independent constraints specified by H_0; it is useful to remember this as a general rule. Before proceeding, to better derive the result, we need have a representation of Θ_0 given above. As we have r constraints on the parameter $\boldsymbol{\theta}$, then only $k-r$ components of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$ are free to change, and so it has $k-r$ degrees of freedom. Without loss of generality, we denote these $k-r$ dimension parameter by $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_{k-r})$. So, this specification of Θ_0 may equivalently be given as a transformation

$$H_0 : \boldsymbol{\theta} = g(\boldsymbol{\vartheta}), \tag{4.4}$$

where g is a continuously differentiable function from $\mathbb{R}^{k-r}$ to $\mathbb{R}^k$ with a full rank $\partial g(\boldsymbol{\vartheta})/\partial \boldsymbol{\vartheta}$. For example, consider again $H_0 : \boldsymbol{\theta}_0 \in \Theta_0 = \{\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) : \theta_1 = \theta_{01}\}$. Then, we can set $\vartheta_1 = \theta_2$, $\vartheta_2 = \theta_3$, $g_1(\boldsymbol{\vartheta}) = \theta_{01}$, $g_2(\boldsymbol{\vartheta}) = \theta_2$, $g_3(\boldsymbol{\vartheta}) = \theta_3$; Also, suppose $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T$ and $H_0 : \theta_1 = \theta_2$. Here, $\Theta = \mathbb{R}^3$, $k = 3$ and $r = 1$, and θ_2 and θ_3 are the two free changing parameters. Then we can take $\boldsymbol{\vartheta} = (\theta_2, \theta_3)^T \in \mathbb{R}^{k-r} = \mathbb{R}^2$, and $g_1(\boldsymbol{\vartheta}) = \theta_2$, $g_2(\boldsymbol{\vartheta}) = \theta_2$, $g_3(\boldsymbol{\vartheta}) = \theta_3$.

Theorem 4.4.1 *Assume the conditions in Theorem 4.2.1 hold and H_0 is determined by (4.4). Under H_0, $\lambda_n \xrightarrow{d} \chi^2_r$.*

Proof. Without loss of generality, we assume that there exists an MLE $\widehat{\boldsymbol{\theta}}_n$ and MLE $\widehat{\boldsymbol{\vartheta}}_n$ under H_0 such that

$$\Lambda_n = \frac{\sup_{\boldsymbol{\theta}\in\Theta_0} L(\boldsymbol{\theta}; \mathbf{X})}{\sup_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}; \mathbf{X})} = \frac{L(g(\widehat{\boldsymbol{\vartheta}}_n), \mathbf{X})}{L(\widehat{\boldsymbol{\theta}}_n; \mathbf{X})}.$$

Following the proof of Theorem 4.2.1-(ii), we can obtain that

$$\sqrt{n}\mathbf{I}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \sqrt{n}s(\boldsymbol{\theta}_0) + o_p(1),$$

and also, by Taylor's expansion,

$$\begin{aligned} 2\left[\log L(\widehat{\boldsymbol{\theta}}_n) - \log L(\boldsymbol{\theta}_0)\right] &= 2n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T s(\boldsymbol{\theta}_0) + n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T s'(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_p(1) \\ &= n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T \mathbf{I}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_p(1). \end{aligned}$$

Then,

$$2\left[\log L(\widehat{\boldsymbol{\theta}}_n) - \log L(\boldsymbol{\theta}_0)\right] = ns^T(\boldsymbol{\theta}_0)[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}s(\boldsymbol{\theta}_0) + o_p(1).$$

Similarly, under H_0,

$$2\left[\log L(g(\widehat{\boldsymbol{\vartheta}}_n)) - \log L(g(\boldsymbol{\vartheta}_0))\right] = n\tilde{s}^T(\boldsymbol{\vartheta}_0)[\tilde{\mathbf{I}}(\boldsymbol{\vartheta}_0)]^{-1}\tilde{s}(\boldsymbol{\vartheta}_0) + o_p(1)$$

where

$$\begin{aligned} \tilde{s}(\boldsymbol{\vartheta}) &= \frac{1}{n}\frac{\partial \log L(g(\boldsymbol{\vartheta}))}{\partial \boldsymbol{\vartheta}} = \mathbf{D}(\boldsymbol{\vartheta})s(g(\boldsymbol{\vartheta})), \\ \mathbf{D}(\boldsymbol{\vartheta}) &= \partial g(\boldsymbol{\vartheta})/\partial \boldsymbol{\vartheta}, \end{aligned}$$

and $\tilde{\mathbf{I}}(\boldsymbol{\vartheta})$ is the Fisher information matrix about $\boldsymbol{\vartheta}$. Combining these results, we can obtain

$$\begin{aligned}\lambda_n = -2\log\Lambda_n &= 2\left[\log L(\widehat{\boldsymbol{\theta}}_n) - \log L(g(\widehat{\boldsymbol{\vartheta}}_n))\right] \\ &= n[s(g(\boldsymbol{\vartheta}_0))]^T\mathbf{B}(\boldsymbol{\vartheta}_0)s(g(\boldsymbol{\vartheta}_0)) + o_p(1)\end{aligned}$$

under H_0, where

$$\mathbf{B}(\boldsymbol{\vartheta}) = [\mathbf{I}(g(\boldsymbol{\vartheta}))]^{-1} - [\mathbf{D}(\boldsymbol{\vartheta})]^T[\tilde{\mathbf{I}}(\boldsymbol{\vartheta})]^{-1}\mathbf{D}(\boldsymbol{\vartheta}).$$

By the CLT, $\sqrt{n}[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1/2}s(\boldsymbol{\theta}_0) \xrightarrow{d} \mathbf{Z}$, where $\mathbf{Z} = N_k(\mathbf{0}, \mathbf{I}_k)$. Then, it follows from the Slutsky's Theorem that, under H_0,

$$\lambda_n \xrightarrow{d} \mathbf{Z}^T[\mathbf{I}(g(\boldsymbol{\vartheta}_0))]^{1/2}\mathbf{B}(\boldsymbol{\vartheta}_0)[\mathbf{I}(g(\boldsymbol{\vartheta}_0))]^{1/2}\mathbf{Z}.$$

Finally, it remains to investigate the properties of the matrix $[\mathbf{I}(g(\boldsymbol{\vartheta}_0))]^{1/2}\mathbf{B}(\boldsymbol{\vartheta}_0)[\mathbf{I}(g(\boldsymbol{\vartheta}_0))]^{1/2}$. For notational convenience, let $D = \mathbf{D}(\boldsymbol{\vartheta})$, $B = \mathbf{B}(\boldsymbol{\vartheta})$, $A = \mathbf{I}(g(\boldsymbol{\vartheta}))$, and $C = \tilde{\mathbf{I}}(\boldsymbol{\vartheta})$. Then,

$$\begin{aligned}(A^{1/2}BA^{1/2})^2 &= A^{1/2}BABA^{1/2} \\ &= A^{1/2}(A^{-1} - D^TC^{-1}D)A(A^{-1} - D^TC^{-1}D)A^{1/2} \\ &= (\mathbf{I}_k - A^{1/2}D^TC^{-1}DA^{1/2})(\mathbf{I}_k - A^{1/2}D^TC^{-1}DA^{1/2}) \\ &= \mathbf{I}_k - 2A^{1/2}D^TC^{-1}DA^{1/2} + A^{1/2}D^TC^{-1}DAD^TC^{-1}A^{1/2} \\ &= \mathbf{I}_k - A^{1/2}D^TC^{-1}DA^{1/2} \\ &= A^{1/2}BA^{1/2},\end{aligned}$$

where the fourth equality follows from the fact that $C = DAD^T$. This shows that $A^{1/2}BA^{1/2}$ is a projection matrix. The rank of $A^{1/2}BA^{1/2}$ is

$$\begin{aligned}\mathrm{tr}(A^{1/2}BA^{1/2}) &= \mathrm{tr}(\mathbf{I}_k - D^TC^{-1}DA) \\ &= k - \mathrm{tr}(C^{-1}DAD^T) = k - \mathrm{tr}(C^{-1}C) = k - (k-r) = r.\end{aligned}$$

Thus, by using similar arguments in the proof of Theorem 3.1.5 (or more directly by Cochran Theorem), $\mathbf{Z}^T[\mathbf{I}(g(\boldsymbol{\vartheta}_0))]^{1/2}\mathbf{B}(\boldsymbol{\vartheta}_0)[\mathbf{I}(g(\boldsymbol{\vartheta}_0))]^{1/2}\mathbf{Z} \overset{d}{=} \chi^2_r$. □

Consequently, the LRT with rejection $\Lambda_n < e^{-\chi^2_{r,\alpha}/2}$ has asymptotic significance level α, where $\chi^2_{r,\alpha}$ is the $(1-\alpha)$th quantile of the chi-square distribution χ^2_r.

Under the first type of null hypothesis, say $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, the same result holds with the degree of freedom being k. This result can be easily derived in a similar fashion to Theorem 4.4.1 but with less algebras. We do not elaborate here but left as an exercise.

To find the power of the test that rejects H_0 when $\lambda_n > \chi^2_{r,\alpha}$, for some r, one would need to know the distribution of λ_n at the particular $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ value where we want to know the power. But the distribution under $\boldsymbol{\theta}_1$ of λ_n for a fixed n is also generally impossible to find, so we may appeal to asymptotics. However, there cannot be a nondegenerate limit distribution on $[0, \infty)$ for λ_n under a fixed $\boldsymbol{\theta}_1$ in the alternative. The following simple example illustrates this difficulty.

Example 4.4.2 Consider the testing problem in Example 4.4.1 again. We saw earlier that

$$\lambda_n = n \log \left(1 + \frac{\bar{X}_n^2}{\frac{1}{n}\sum (X_i - \bar{X}_n)^2} \right)$$

Consider now a value $\mu \neq 0$. Then, $\bar{X}_n^2 \overset{wp1}{\to} \mu^2 (> 0)$ and $\frac{1}{n}\sum (X_i - \bar{X}_n)^2 \overset{wp1}{\to} \sigma^2$. Therefore, clearly $\lambda_n \overset{wp1}{\to} \infty$ under each fixed $\mu \neq 0$. Thus, There cannot be a non-degenerate limit distribution for λ_n under a fixed alternative μ.

Instead, similar to the Pearson's Chi-square test discussed earlier, we may also consider the behavior of λ_n under "local" alternative, that is, for a sequence $\boldsymbol{\theta}_{1n} = \boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{\delta}$, where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_k)^T$. In this case, a non-central χ^2 approximation under the alternative could be achieved.

4.5 The Wald and Rao score tests

Two competitors to the LRT are available in the literature, see Wald (1943) and Rao (1948) for the first introduction of these procedures respectively. Both of them are general and can be applied to a wide selection of problems. Typically, the three procedures are asymptotically first-order equivalent. Recall the null hypothesis

$$H_0 : R(\boldsymbol{\theta}) = 0, \tag{4.5}$$

where $R(\boldsymbol{\theta})$ is continuously differentiable function from $\mathbb{R}^k$ to $\mathbb{R}^r$. The Wald test statistic is defined

$$W_n = [R(\widehat{\boldsymbol{\theta}}_n)]^T \left\{ [C(\widehat{\boldsymbol{\theta}}_n)]^T [\mathbf{I}_n(\widehat{\boldsymbol{\theta}}_n)]^{-1} C(\widehat{\boldsymbol{\theta}}_n) \right\}^{-1} R(\widehat{\boldsymbol{\theta}}_n),$$

where $\widehat{\boldsymbol{\theta}}_n$ is an MLE or RLE of θ, $\mathbf{I}_n(\widehat{\boldsymbol{\theta}}_n)$ is the Fisher information matrix based on $\mathbf{X}$ and $C(\boldsymbol{\theta}) = \partial R(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$. For testing a simple null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, $R(\boldsymbol{\theta})$ will become $\boldsymbol{\theta} - \boldsymbol{\theta}_0$ and W_n simplifies to

$$W_n = n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T \mathbf{I}(\widehat{\boldsymbol{\theta}}_n)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0).$$

Rao (1947) introduced a *score test* that rejects H_0 when the value of

$$R_n = n[s(\tilde{\boldsymbol{\theta}}_n)]^T [\mathbf{I}(\tilde{\boldsymbol{\theta}}_n)]^{-1} s(\tilde{\boldsymbol{\theta}}_n)$$

is large, where $\tilde{\boldsymbol{\theta}}_n$ is an MLE or RLE of $\boldsymbol{\theta}$ under H_0. For testing a simple null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, R_n will simplify to

$$R_n = n[s(\boldsymbol{\theta}_0)]^T [\mathbf{I}(\boldsymbol{\theta}_0)]^{-1} s(\boldsymbol{\theta}_0).$$

Here are the asymptotic chi-square results for these two statistics.

Theorem 4.5.1 *Assume the conditions in Theorem 4.2.1 hold. Under H_0 given by (4.5), then (i) $W_n \xrightarrow{d} \chi^2_r$ and (ii) also $R_n \xrightarrow{d} \chi^2_r$.*

Proof. (i) Using Theorem 4.2.1 and Delta Theorem,

$$\sqrt{n}(R(\widehat{\boldsymbol{\theta}}_n) - R(\boldsymbol{\theta}_0)) \xrightarrow{d} N_r(\mathbf{0}, [C(\boldsymbol{\theta}_0)]^T [\mathbf{I}(\boldsymbol{\theta}_0)]^{-1} C(\boldsymbol{\theta}_0)).$$

Under H_0, $R(\boldsymbol{\theta}_0) = 0$ and, therefore,

$$n[R(\widehat{\boldsymbol{\theta}}_n]^T \left\{ [C(\boldsymbol{\theta}_0)]^T [\mathbf{I}(\boldsymbol{\theta}_0)]^{-1} C(\boldsymbol{\theta}_0) \right\}^{-1} R(\widehat{\boldsymbol{\theta}}_n) \xrightarrow{d} \chi^2_r.$$

by CMT. Then the result follows from Slutsky's theorem and the fact that $\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$ and $\mathbf{I}(\boldsymbol{\theta})$ and $C(\boldsymbol{\theta})$ are continuous at $\boldsymbol{\theta}$.

(ii) From the Lagrange multipliers, $\tilde{\boldsymbol{\theta}}_n$ satisfies

$$ns(\tilde{\boldsymbol{\theta}}_n) + C(\tilde{\boldsymbol{\theta}}_n)\eta_n = 0 \quad \text{and} \quad R(\tilde{\boldsymbol{\theta}}_n) = 0.$$

Using Taylor's expansion, one can show that under H_0

$$[C(\boldsymbol{\theta}_0)]^T(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = o_p(n^{-1/2}) \tag{4.6}$$

and

$$ns(\boldsymbol{\theta}_0) - n\mathbf{I}(\boldsymbol{\theta}_0)(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + C(\boldsymbol{\theta}_0)\eta_n = o_p(n^{1/2}), \tag{4.7}$$

Multiplying $[C(\boldsymbol{\theta}_0)]^T[n\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}$ to the left-hand side of (4.7) and using (4.6), we obtain that

$$[C(\boldsymbol{\theta}_0)]^T[n\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}C(\boldsymbol{\theta}_0)\eta_n = -n[C(\boldsymbol{\theta}_0)]^T[n\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}s(\boldsymbol{\theta}_0) + o_p(n^{-1/2}),$$

which implies

$$\eta_n^T[C(\boldsymbol{\theta}_0)]^T[n\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}C(\boldsymbol{\theta}_0)\eta_n \xrightarrow{d} \chi_r^2.$$

Then, the result follows from the above equation and the fact that $C(\tilde{\boldsymbol{\theta}}_n)\eta_n = -ns(\tilde{\boldsymbol{\theta}}_n)$, and $I(\boldsymbol{\theta})$ is continuous at $\boldsymbol{\theta}_0$. □

Thus, Wald's test, Rao's tests and LRT are asymptotically equivalent. Note that Wald's test requires computing $\widehat{\boldsymbol{\theta}}_n$, whereas Rao's score test requires computing $\tilde{\boldsymbol{\theta}}_n$, not $\widehat{\boldsymbol{\theta}}_n$. On the other hand, the LRT requires computing both $\widehat{\boldsymbol{\theta}}_n$ and $\tilde{\boldsymbol{\theta}}_n$ (or solving two maximization problems). Hence, one may choose one of these tests that is easy to compute in a particular application.

4.6 Confidence sets based on likelihoods

The usual duality between testing and confidence intervals says that the acceptance region of a test with size α can be inverted to give a confidence set of coverage probability $(1-\alpha)$. In other words, suppose $A(\boldsymbol{\theta}_0)$ is the acceptance region of a size α test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, and define $C(\mathbf{X}) = \{\boldsymbol{\theta} : \mathbf{X} \in A(\boldsymbol{\theta})\}$. Then $P_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}_0 \in C(x)) = 1 - \alpha$ and hence $C(\mathbf{X})$ is

a $100(1-\alpha)\%$ confidence set for $\boldsymbol{\theta}$. For example, the acceptance region of the LRT with $\Theta_0 = \{\boldsymbol{\theta} : \boldsymbol{\theta} = \boldsymbol{\theta}_0\}$ is

$$A(\boldsymbol{\theta}_0) = \{\mathbf{x} : L(\boldsymbol{\theta}_0; \mathbf{x}) \geq e^{-\chi^2_{k,\alpha}/2} L(\widehat{\boldsymbol{\theta}}_n; \mathbf{x})\}$$

Consequently,

$$C(\mathbf{X}) = \{\boldsymbol{\theta} : L(\boldsymbol{\theta}; \mathbf{X}) \geq e^{-\chi^2_{k,\alpha}/2} L(\widehat{\boldsymbol{\theta}}_n; \mathbf{X})\}$$

is a $1-\alpha$ asymptotically correct confidence set.

This method is often called the inversion of a test. In particular, the LRT, the Wald test, and the Rao score test can all be inverted to construct confidence sets that have asymptotically a $100(1-\alpha)\%$ coverage probability. The confidence sets constructed from the LRT, the Wald test, and the score test are respectively called the likelihood ratio, Wald, and score confidence sets. Of these, the Wald and the score confidence sets are ellipsoids because of how the corresponding test statistics are defined. The likelihood ratio confidence set is typically more complicated but it is also ellipsoids from asymptotic viewpoints. Here is an example.

Example 4.6.1 Suppose $X_i \overset{\text{iid}}{\sim} \text{BIN}(p, 1), 1 \leq i \leq n$. For testing $H_0 : p = p_0$ versus $H_1 : p \neq p_0$, the LRT statistic is

$$\begin{aligned}\Lambda_n &= \frac{p_0^Y (1-p_0)^{n-Y}}{\sup_p p^Y (1-p)^{n-Y}} \\ &= \frac{p_0^Y (1-p_0)^{n-Y}}{\left(\frac{Y}{n}\right)^Y \left(1-\left(\frac{Y}{n}\right)\right)^{n-Y}} = \frac{p_0^Y (1-p_0)^{n-Y}}{\widehat{p}^Y (1-\widehat{p})^{n-Y}},\end{aligned}$$

where $Y = \sum_{i=1}^n X_i$ and $\widehat{p} = Y/n$. Thus, the likelihood ratio confidence set is of the form

$$C_1(X) = \left\{p : p^Y (1-p)^{n-Y} \geq e^{-\chi^2_{1,\alpha}/2} \widehat{p}^Y (1-\widehat{p})^{n-Y}\right\}.$$

The confidence set obtained by inverting acceptance regions of Wald's test is simply

$$\begin{aligned}C_2(X) &= \left\{p : |\widehat{p} - p| \leq \frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{\widehat{p}(1-\widehat{p})}\right\} \\ &= \left[\widehat{p} - z_{\alpha/2}\sqrt{\widehat{p}(1-\widehat{p})/n}, \widehat{p} + z_{\alpha/2}\sqrt{\widehat{p}(1-\widehat{p})/n}\right]\end{aligned}$$

since $(\chi^2_{1,\alpha})^{1/2} = z_{\alpha/2}$ and $W_n = n(\widehat{p} - p_0)^2 I(\widehat{p})$, where $I(p) = \frac{1}{p(1-p)}$. This is the textbook confidence interval for p.

For the score test statistic, we need

$$s(p) = \frac{\widehat{p}}{p} - \frac{1-\widehat{p}}{1-p} = \frac{\widehat{p} - p}{p(1-p)}$$

and

$$n[s(p)]^2[I(p)]^{-1} = n\frac{(\widehat{p}-p)^2}{p^2(1-p)^2}p(1-p) = \frac{n(\widehat{p}-p)^2}{p(1-p)}.$$

Hence, The confidence set obtained by inverting acceptance regions of Rao's score test is

$$C_3(X) = \left\{p : n(\widehat{p} - p)^2 \leq p(1-p)\chi^2_{1,\alpha}\right\} \equiv [l_C, u_C],$$

where l_C, u_C are the roots of the quadratic equation $p(1-p)\chi^2_{1,\alpha} - n(\widehat{p} - p)^2 = 0$.

Chapter 5

Asymptotics in nonparametric inference

5.1 Sign test (Fisher)

5.1.1 Test procedure

This is perhaps the earliest example of a nonparametric testing procedure. In fact, the test was apparently discussed by Laplace in the 1700s. The sign test is a test for the median of any continuous distribution without requiring any other assumptions.

Hypothesis The null hypothesis of interest here is that of zero shift in location due to the treatment, namely, $H_0 : \theta = 0$ versus $H_0 : \theta > 0$. This null hypothesis asserts that each of the distributions (not necessarily the same) for the difference (post-treatment minus pre-treatment observations) has median 0, corresponding to no shift in location due to treatment. Certainly, this is essentially equivalent to consider the null hypothesis $H_0 : \theta = \theta_0$ because we can simply use $H_0 : \theta - \theta_0 = 0$.

Procedure The test statistic is given by the total number of $X_1, X_2, \ldots, X_n$ that are greater

than θ_0, say

$$S_n = \sum_{i=1}^{n} I(X_i > \theta_0),$$

where $I(\cdot)$ is the indicator function. Then, small value of S_n leads to reject H_0. We now need to know the distribution of S_n. Obviously, the distribution of S_n under H_0:

$$S_n \sim \text{BIN}(n, 1/2),\ P(S_n = k) = C_n^k \left(\frac{1}{2}\right)^n$$

Thus, the p-value is

$$P(\text{Bin}(n, 1/2) \geq S_n) = \sum_{k=S_n}^{n} C_n^k \left(\frac{1}{2}\right)^n.$$

For simplicity, we may use the following large-sample approximation to obtain an approximated p-value. Note that

$$E_{H_0}(S_n) = \sum_{i=1}^{n} \left(\frac{1}{2}\right) = n/2$$
$$\text{Var}_{H_0}(S_n) = \sum_{i=1}^{n} \left(\frac{1}{4}\right) = n/4.$$

The asymptotic normality of the standardized form

$$S_n^* = \frac{S_n - E_{H_0}(S_n)}{\text{Var}_{H_0}^{1/2}(S_n)} = \frac{S_n - \frac{n}{2}}{(n/4)^{1/2}}.$$

follows from standard central limit theory for sums of mutually independent, identically distributed random variables.

For large sample sizes, we can make use of the standard central limit theorem for sums of i.i.d. random variables to conclude that

$$\frac{S_n - np_\theta}{[np_\theta(1-p_\theta)]^{1/2}} = \frac{S_n - n(1-F(0))}{[n(1-F(0))(F(0))]^{1/2}}$$

has an asymptotic $N(0,1)$ distribution. Thus, for large n, we can approximate the exact power by

$$\text{Power}_\theta = 1 - \Phi\left(\frac{b_{\alpha,1/2} - np_\theta}{[np_\theta(1-p_\theta)]^{1/2}}\right)$$

We note that both the exact power and the approximate power against an alternative $\theta > 0$ depend on the common distribution only through the value of its distribution $F(z)$ at $z = 0$. Thus, if two distributions F_1 and F_2 have a common median $\theta > 0$ and $F_1(0) = F_2(0)$, then the exact power of the sign test against the alternative $\theta > 0$ will be the same for both F_1 and F_2.

5.1.2 Asymptotic Properties

Consistency of the sign Test

Definition 5.1.1 *Let $\{\phi_n\}$ be a sequence of tests for $H_0 : F \in \Omega_0$ versus $H_1 : F \in \Omega_1$. Then, $\{\phi_n\}$ is consistent against the alternatives Ω_1 if*

(i) $E_F(\phi_n) \to \alpha \in (0, 1)$, $\forall F \in \Omega_0$;

(ii) $E_F(\phi_n) \to 1$, $\forall F \in \Omega_1$.

As in estimation, consistency is a rather weak property of a sequence of tests. However, something must be fundamentally wrong with the test for it not to be consistent. If a test is inconsistent against a large class of alternatives, then it is considered an undesirable test.

Example 5.1.1 For a parametric example, let $X_1, \ldots, X_n$ be an i.i.d. sample from the Cauchy distribution, $C(\theta, 1)$. For all $n \geq 1$, we know that $\bar{X}_n$ also has the $C(\theta, 1)$ distribution. Consider testing the hypothesis $H_0 : \theta = 0$ versus $H_1 : \theta > 0$ by using a test that rejects for large $\bar{X}_n$. The cutoff point, k, is found by making $P_{H_0}(\bar{X}_n > k) = \alpha$. But k is simply the αth quantile of the $C(0, 1)$ distribution. Then the power of this test is given by

$$P_\theta(\bar{X}_n > k) = P(C(\theta, 1) > k) = P(\theta + C(0, 1) > k) = P(C(0, 1) > k - \theta).$$

This is a fixed number not dependent on n. Therefore, the power does not approach to 1 as $n \to \infty$, and so so the test is not consistent even against parametric alternatives. In contrast, a test based on the median would be consistent in the $C(\theta, 1)$ case (why?).

Theorem 5.1.1 *If F is a continuous C.D.F. with unique median θ, then the sign test is consistent for tests on θ.*

Proof. Recall that the sign test rejects H_0 if $S_n = \sum I(X_i > \theta_0) \geq k_n$. If we choose $k_n = \frac{n}{2} + z_\alpha \sqrt{\frac{n}{4}}$, then, by the ordinary central limit theorem, we have

$$P_{H_0}(S_n \geq k_n) \to \alpha.$$

The power of the test is

$$Q_n = P_F(S_n \geq k_n) = P_F\left(\frac{1}{n}S_n - p_\theta \geq \frac{1}{n}k_n - p_\theta\right),$$

where $p_\theta = P_\theta(X_1 > \theta_0)$. Since we assume $\theta > \theta_0$, it follows that $\frac{1}{n}k_n - p_\theta < 0$ for all large n. Also, $\frac{1}{n}S_n - p_\theta$ converges in probability to 0 under any F (WLLN), and so $Q_n \to 1$. Since the power goes to 1, the test is consistent against any alternative F satisfying $\theta > \theta_0$. □

Asymptotic relative efficiency (ARE)

We wish to compare the sign test with the t-test in terms of asymptotic relative efficiency. The point is that, at a fixed alternative θ, if α remains fixed, then, for large n, the power of both tests is approximately 1 (say, consistent) and there would be no way to practically compare the two tests. Perhaps we can see how the powers compare for $\theta \approx \theta_0$. The idea is to take $\theta = \theta_n \to \theta_0$ at such a rate that the limiting power of the tests is strictly between α and 1. If the two powers converge to different values, then we can take the ratio of the limits as a measure of efficiency. The idea is due to E.J.G. Pitman (Pitman 1948). We firstly give a brief introduction to the concept of ARE regarding the test.

In estimation, an agreed-on basis for comparing two sequences of estimates whose mean squared error each converges to zero as $n \to \infty$ is to compare the variances in their limit distributions. Thus, if $\sqrt{n}(\widehat{\theta}_{1n} - \theta) \xrightarrow{d} N(0, \sigma_1^2(\theta))$ and $\sqrt{n}(\widehat{\theta}_{2n} - \theta) \xrightarrow{d} N(0, \sigma_2^2(\theta))$, then the asymptotic relative efficiency (ARE) of $\widehat{\theta}_{2n}$ with respect to $\widehat{\theta}_{1n}$ is defined as $\sigma_1^2(\theta)/\sigma_2^2(\theta)$.

One can similarly ask what should be a basis for comparison of two sequences of tests based on statistics T_{1n} and T_{2n} of a hypothesis $H_0 : \theta = \theta_0$. Suppose we use statistics such that large values of them correspond to rejection of H_0; i.e., H_0 is rejected if $T_n > c_n$. Let α,

β denote the type 1 error probability and the power of the test, and let θ denote a specific alternative. Suppose $n(\alpha, \beta, \theta, T)$ is the smallest sample size such that

$$P_{H_0}(T_n \geq c_n) \leq \alpha, \quad P_\theta(T_n \geq c_n) \geq \beta,$$

Two tests based on T_{1n} and T_{2n} can be compared through the ratio

$$e(T_2, T_1) = n(\alpha, \beta, \theta, T_1)/n(\alpha, \beta, \theta, T_2),$$

and T_{1n} is preferred if this ratio is less than 1. The threshold sample size $n(\alpha, \beta, \theta, T)$ is difficult or impossible to calculate even in the simplest examples. Furthermore, the ratio can depend on particular choices of α, β, θ.

Fortunately, if $\alpha \to 0, \beta \to 1$, or $\theta \to \theta_0$ (an element of the boundary θ_0), then the ratio (generally) converges to something that depends on θ alone or is just a constant. The three respective measures of efficiency correspond to approaches by Bahadur, Hodges and Lehmann, and Pitman; see Pitman (1948), Hodges and Lehmann (1956), and Bahadur (1960). Typically, of these, Pitman ARE is the easiest to calculate in most applications by a fixed recipe under frequently satisfied conditions that we present below. It is also important to note that the Pitman efficiency works out to just the asymptotic efficiency in the point estimation problem, with T_{1n} and T_{2n} being considered as the respective estimates. Testing and estimation come together in the Pitman approach. We state a theorem describing the calculation of the Pitman efficiency, which is a simple one in form and suffices for many applications.

Theorem 5.1.2 *Let $-\infty < h < \infty$ and $\theta_n = \theta_0 + \frac{h}{\sqrt{n}}$. Consider the following conditions: (i) there exist functions $\mu(\theta)$, $\sigma(\theta)$, such that, for all h,*

$$\frac{\sqrt{n}(T_n - \mu(\theta_n))}{\sigma(\theta_n)} \xrightarrow{d} N(0, 1);$$

(ii) $\mu'(\theta_0) > 0$; (iii) $\sigma(\theta_0) > 0$ and $\sigma(\theta)$ is continuous at θ_0. Suppose T_{1n} and T_{2n} each satisfy conditions (i)-(iii). Then

$$e(T_2, T_1) = \frac{\sigma_1^2(\theta_0)}{\sigma_2^2(\theta_0)} \left[\frac{\mu_2'(\theta_0)}{\mu_1'(\theta_0)}\right]^2$$

See Serfling (1980) for a detailed proof. By this theorem, we are now ready to derive the ARE of the sign test with respect to the t-test.

Corollary 5.1.1 *Let $X_1, \ldots, X_n$ be i.i.d. observations from any symmetric continuous distribution function $F(x-\theta)$ with density $f(\cdot)$, where $f(0) > 0$, f is continuous at 0 and $F(0) = \frac{1}{2}$. The Pitman asymptotic relative efficiency of the one-sample test procedure (one- or two-sided) based on the sign test statistic S_n with respect to the corresponding normal theory test based on $\bar{X}_n$ is*

$$e(S_n, \bar{X}_n) = 4\sigma_F^2 f^2(0),$$

where $\sigma_F^2 = \mathrm{Var}_F(X) < \infty$.

Proof. For $T_{2n} = \frac{1}{n}S_n$, first notice that $E_\theta(T_{2n}) = P_\theta(X_1 > 0) = 1 - F(-\theta)$. Also $\mathrm{Var}_\theta(T_{2n}) = F(-\theta)(1-F(-\theta))/n$. We choose $\mu_n(\theta) = 1 - F(-\theta)$ and $\sigma_n^2(\theta) = F(-\theta)(1 - F(-\theta))/n$. Therefore, $\mu_n'(\theta) = f(-\theta)$ and $\mu_n'(\theta_0) = f(0) > 0$. For $T_{1n} = \bar{X}_n$, choose $\mu_n(\theta) = \theta$ and $\sigma_n^2(\theta) = \sigma_F^2/n$. Conditions (i)-(iii) are easily verified here, too, with these choices of $\mu_n(\theta)$ and $\sigma_n(\theta)$. Therefore, by Theorem 5.1.2, the result follows immediately. □

Some values of this ARE for selected $F(\cdot)$ are:

F:	Normal	Uniform	Logistic	DE	Cauchy	t_3	t_5
$e(S_n, T_n)$	0.637	0.333	0.822	2.000	∞	1.620	0.961

The sign test, however, cannot get arbitrarily bad with respect to the t-test under some restrictions on the C.D.F. F, as is shown by the following result, although the t-test can be arbitrarily bad with respect to the sign test. Hodges and Lehmann (1956) found that within a certain class of populations, $e(S_n, \bar{X}_n)$ is always at least 1/3 and the bound is attained when F is any symmetric uniform distribution. Of course, the minimum efficiency is not very good. We will later discuss alternative nonparametric tests for the location-parameter problem that have much better asymptotic efficiencies.

5.2 Signed rank test (Wilcoxon)

5.2.1 Procedure

Recall that Hodges and Lehmann proved that the sign test has a small positive lower bound of 1/3 on the Pitman efficiency with respect to the t-test in the class of densities with a finite variance, which is not satisfactory. The problem with the sign test is that it only uses the signs of $X_i - \theta_0$, not the magnitude of $X_i - \theta_0$. A nonparametric test that incorporates the magnitudes as well as the signs is called the Wilcoxon signed-rank test, under a little bit more assumption about the population distribution; see Wilcoxon (1945).

Suppose that $X_1, \ldots, X_n$ are the observed data from some location parameter distribution $F(x - \theta)$, and assume that F is symmetric. Let $\theta = \text{median}(F)$. We want to test $H_0 : \theta = 0$ against $H_1 : \theta > 0$. We start by ranking $|X_i|$ from the smallest to the largest, giving the units ranks $R_1, \ldots, R_n$ and order statistics $|X|_{(1)}, \ldots, |X|_{(n)}$.

Then, the Wilcoxon signed-rank statistic is defined to be the sum of these ranks that correspond to originally positive observations. That is,

$$T_n = \sum_{i=1}^{n} R_i I(X_i > 0),$$

where the term $R_i I(X_i > 0)$ is known as the positive signed rank of X_i.

When θ is greater than 0, there will tend to be a large proportion of positive X and they will tend to have the larger absolute values. Hence, we would expect a higher proportion of positive signed ranks with relatively large sizes. At the α level of significance, reject H_0 if $T_n \geq t_\alpha$, where the constant t_α is chosen to make the type I error probability equal to α. Lower-sided and two-sided tests can be constructed similarly.

Remark 5.2.1 It may appear that some of the information in the ranking of the sample is being lost by using only the positive signed ranks to compute T_n. Such is not the case. If we define $\tilde{T}_n$ to be the sum of ranks (of the absolute values) corresponding to the negative X observations, then $\tilde{T}_n = \sum_{i=1}^{n}(1 - I(X_i > 0)R_i$. It follows that $T_n + \tilde{T}_n = \sum_{i=1}^{n} R_i = n(n+1)/2$.

Thus, the test procedures defined above could be constructed equivalently based on $\tilde{T}_n = n(n+1)/2 - T_n$.

To do a test, we need the null distribution of T_n. If we define

$$W_i = I(|X|_{(i)} \text{ corresponds to some positive } X_j),$$

then we have an alternative expression for T_n, namely

$$T_n = \sum_{i=1}^{n} iW_i.$$

It turns out that, under H_0, the $\{W_i\}$ have a relatively simple joint distribution.

Proposition 5.2.1 *Under H_0, $W_1, \ldots, W_n$ are i.i.d.* BIN$(1, 1/2)$ *variables.*

Proof. By the symmetric assumption, $W_i \sim \text{BIN}(1, 1/2)$ is obvious. To show the independence, we define the so-called anti-rank,

$$D_k = \{i : R_i = k, 1 \le i \le n\},$$

say, the index of the observation whose absolute rank is k. Thus, $W_k = I(X_{D_k} > 0)$. Let $\mathbf{D} = (D_1, \ldots, D_n)$, $\mathbf{d} = (d_1, \ldots, d_n)$, and then we have

$$\begin{aligned}
&P(W_1 = w_1, \ldots, W_n = w_n) \\
&\quad = \sum_{\mathbf{d}} P\left(I(X_{D_1} > 0) = w_1, \ldots, I(X_{D_n} > 0) = w_n \mid \mathbf{D} = \mathbf{d}\right) P(\mathbf{D} = \mathbf{d}) \\
&\quad = \sum_{\mathbf{d}} P\left(I(X_{d_1} > 0) = w_1, \ldots, I(X_{d_n} > 0) = w_n\right) P(\mathbf{D} = \mathbf{d}) \\
&\quad = \left(\frac{1}{2}\right)^n \sum_{\mathbf{d}} P(\mathbf{D} = \mathbf{d}) = \left(\frac{1}{2}\right)^n,
\end{aligned}$$

where the second equality comes from the fact that $I(X_1 > 0), \ldots, I(X_n > 0)$ are independent with $(D_1, \ldots, D_n)$. The independence is therefore immediately obtained by noting that $P(W_i = w_i) = \frac{1}{2}$. The independence between $I(X_1 > 0), \ldots, I(X_n > 0)$ and $(D_1, \ldots, D_n)$ can be easily established as follows. Actually, $(D_1, \ldots, D_n)$ is the function of $|X_1|, \ldots, |X_n|$

and $(I(X_i > 0), |X_i|), i = 1, \ldots, n$ are independent each other. Thus, it suffices to show that $I(X_i > 0)$ is independent with $|X_i|$. In fact,

$$P(I(X_i > 0) = 1, |X_i| \leq x) = P(0 < X_i \leq x) = F(x) - F(0) = F(x) - \frac{1}{2}$$
$$= \frac{2F(x) - 1}{2} = P(I(X_i > 0) = 1)P(|X_i| \leq x). \qquad \square$$

When n is large, a large-sample approximation is sufficient to obtain an approximately correct signed-rank test. Proposition 5.2.1, together with the representation of T_n above and Hajek-Sidak's CLT, leads to the asymptotic null distribution of T_n. Clearly,

$$E_{H_0}(T_n) = \frac{n(n+1)}{4}, \text{ and } \mathrm{Var}_{H_0}(T_n) = \frac{n(n+1)(2n+1)}{24}.$$

The results above imply the following theorem.

Theorem 5.2.1 *Let $X_1, \ldots, X_n$ be i.i.d. observations from $F(x-\theta)$, where F is continuous and symmetric. Under $H_0 : \theta = 0$,*

$$\frac{T_n - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \xrightarrow{d} N(0, 1).$$

Therefore, the signed-rank test can be implemented by rejecting the null hypothesis, $H_0 : \theta = 0$ if

$$T_n > \frac{n(n+1)}{4} + z_\alpha \sqrt{\frac{n(n+1)(2n+1)}{24}}.$$

The other option would be to find the exact finite sample distribution of T_n under the null as illustrated above. This can be done in principle, but the CLT approximation works pretty well.

Unlike the null case, the Wilcoxon signed-rank statistic T_n does not have a representation as a sum of independent random variables under the alternative. So the asymptotic non-null distribution of T_n, which is very useful for approximating the power and establishing the consistency of the test, does not follow from the CLT for independent summands. However, T_n still belongs to the class of U-statistics, and hence the CLTs for U-statistics can be used to derive the asymptotic nonnull distribution of T_n and thereby get an approximation to the power of the Wilcoxon signed-rank test. The following proposition is useful for deriving its non-null distribution.

Proposition 5.2.2 *We have the following equivalent expression for* T_n,

$$T_n = \sum_{i \le j} I\left(\frac{X_i + X_j}{2} > 0\right).$$

Proof. Consider to use the antirank D_k again. Note that

$$\sum_{i \le j} I\left(\frac{X_i + X_j}{2} > 0\right) = \sum_{i=1}^{n} I(X_i > 0) + \sum_{i<j} I\left(\frac{X_{D_i} + X_{D_j}}{2} > 0\right). \tag{5.1}$$

For $i < j$, hence $|X_{D_i}| \le |X_{D_j}|$, consider the expression $I\left(\frac{X_{D_i}+X_{D_j}}{2} > 0\right)$. There are four cases to consider: where X_{D_i} and X_{D_j} are both positive; where they are both negative; and the two cases where they have mixed signs. In all these cases, though, it is easy to see that

$$I\left(\frac{X_{D_i} + X_{D_j}}{2} > 0\right) = I(X_{D_j} > 0).$$

Using this, we have that the right side of expression (5.1) is equal to

$$\begin{aligned}\sum_{i=1}^{n} I(X_i > 0) + \sum_{i<j} I\left(\frac{X_{D_i} + X_{D_j}}{2} > 0\right) &= \sum_{j=1}^{n} I(X_{D_j} > 0) + \sum_{j=1}^{n}(j-1) I(X_{D_j} > 0) \\ &= \sum_{j=1}^{n} j I(X_{D_j} > 0),\end{aligned}$$

and we are finished. □

To established the asymptotic normality of T_n under alternative cases, we present the basic results about U-statistics here. Suppose that $h(x_1, x_2, \ldots, x_r)$ is some real-valued function of r arguments $x_1, x_2, \ldots, x_r$. The arguments $x_1, x_2, \ldots, x_r$ can be real or vector valued. Now, suppose $X_1, \ldots, X_n$ are i.i.d. observations from some C.D.F. F, and for a given $r \ge 1$ we want to estimate or make inferences about the parameter $\theta = \theta(F) = E_F h(X_1, X_2, \ldots, X_r)$. We assume $n \ge r$. Of course, one unbiased estimate is $h(X_1, X_2, \ldots, X_r)$ itself. But one should be able to find a better unbiased estimate if $n > r$ because $h(X_1, X_2, \ldots, X_r)$ does not use all of the sample data. Indeed, in this case

$$\frac{1}{C_n^r} \sum_{1 \le i_1 < i_2 \cdots < i_r \le n} h(X_{i_1}, X_{i_2}, \ldots, X_{i_r})$$

may be a better unbiased estimate than $h(X_1, X_2, \ldots, X_r)$.

Statistics of this form are called U-statistics (U for unbiased), and h is called the kernel and r its order. We will assume that h is permutation symmetric in order that U has that property as well.

Example 5.2.1 Suppose, $r = 1$. Then the linear statistic $\frac{1}{n}\sum_{i=1}^{n} h(X_i)$ is clearly a U-statistic. In particular, $\frac{1}{n}\sum_{i=1}^{n} X_i^k$ is a U-statistic for any k; Let $r = 2$ and $h(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2$. Then, on calculation,

$$\frac{1}{C_n^2}\sum_{i<j}\frac{1}{2}(X_i - X_j)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Thus, the sample variance is a U-statistic; Let x_0 be a fixed real, $r = 1$, and $h(x) = I(x \leq x_0)$. Then $U = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x_0) = F_n(x_0)$, the empirical C.D.F. at x_0. Thus $F_n(x_0)$ for any specified x_0 is a U-statistic.

Example 5.2.2 Let $r = 2$ and $h(X_1, X_2) = I(X_1 + X_2 > 0)$. The corresponding $U = \frac{1}{C_n^2}\sum_{i<j} I(X_i + X_j > 0)$. Now, U is related to the one-sample Wilcoxon statistic, T_n

The summands in the definition of a U-statistic are not independent. Hence, neither the exact distribution theory nor the asymptotics are straightforward. Hajek had the brilliant idea of projecting U onto the class of linear statistics of the form $\frac{1}{n}\sum_{i=1}^{n} h(X_i)$. It turns out that the projection is the dominant part and determines the limiting distribution of U. The main theorems can be seen in Serfling (1980).

For $k = 1, \ldots, r$, let

$$\begin{aligned} h_k(x_1, \ldots, x_k) &= E[h(X_1, \ldots, X_r) \mid X_1 = x_1, \ldots, X_k = x_k] \\ &= E[h(x_1, \ldots, x_k, X_{k+1}, \ldots, X_r)]. \end{aligned}$$

Define $\zeta_k = \mathrm{Var}(h_k(X_1, \ldots, X_k))$.

Theorem 5.2.2 *Suppose that the kernel h satisfying $Eh^2(X_1, \ldots, X_r) < \infty$. Assume that $0 < \zeta_1 < \infty$. Then,*

$$\frac{U - \theta}{\sqrt{\mathrm{Var}(U)}} \xrightarrow{d} N(0, 1).$$

where $\mathrm{Var}(U) = \frac{1}{n}r^2\zeta_1 + O(n^{-2})$.

With these results, we are ready to present the asymptotic normality of T_n.

Theorem 5.2.3 *The Wilcoxon signed-rank statistic T_n is asymptotically normally distributed,*

$$\frac{T_n - E(T_n)}{\sqrt{\text{Var}(T_n)}} \xrightarrow{d} N(0,1).$$

Proof. By Proposition 5.2.2,

$$\begin{aligned}\frac{1}{C_n^2}T_n &= \frac{1}{C_n^2}\sum_{i\le j} I\left(\frac{X_i+X_j}{2}>0\right)\\ &= \frac{1}{C_n^2}\sum_{i=1}^{n} I(X_i>0) + \frac{1}{C_n^2}\sum_{i<j} I\left(\frac{X_i+X_j}{2}>0\right).\end{aligned}$$

Note that the first term is of smaller order ($O_p(n^{-1})$) and we need only consider the second term ($O_p(1)$). However, the second term, denoted as U_n is a U-statistic as defined above. Thus, by Theorem 5.2.2, $(U_n - E(U_n))/\text{Var}(U_n) \xrightarrow{d} N(0,1)$. The result immediately follows from the Slutsky's Theorem. □

With the help of this theorem, we can easily establish the consistency of the T_n test.

Theorem 5.2.4 *If F is a continuous symmetric C.D.F. with unique median θ, then the signed rank test is consistent for tests on θ.*

Proof. Recall that the signed-rank test rejects H_0 if $T_n = \sum_{i\le j} I(\frac{X_i+X_j}{2}>0) \ge t_n$. If we choose $t_n = \frac{n(n+1)}{4} + z_\alpha\sqrt{\frac{n(n+1)(2n+1)}{24}}$, then, by Theorem 5.2.1, we have

$$P_{H_0}(T_n \ge t_n) \to \alpha.$$

The power of the test is

$$Q_n = P_F(T_n \ge t_n) = P_F\left(\frac{1}{C_n^2}T_n - p_\theta \ge \frac{1}{C_n^2}t_n - p_\theta\right),$$

where $p_\theta = P_\theta(X_1+X_2>0)$. Since we assume $\theta>0$ under the alternative, it follows that $\frac{1}{C_n^2}k_n - p_\theta < 0$ for all large n. Also, $\frac{1}{C_n^2}T_n - p_\theta$ converges in probability to 0 under any F (Theorem 5.2.2), and so $Q_n \to 1$. Since the power goes to 1, the test is consistent against any alternative F satisfying $\theta>0$. □

Furthermore, Theorem 5.2.2 allows us to derive the relative efficiency of T_n with respect to other tests. Since T_n takes into account the magnitude as well as the sign of the sample observations, we expect that overall it may have better efficiency properties than the sign test. The following striking result was proved by Hodges and Lehmann in (1956).

Theorem 5.2.5 *Let $X_1, \ldots, X_n$ be i.i.d. observations from any symmetric continuous distribution function $F(x-\theta)$ with density $f(x-\theta)$,*

(i) The Pitman asymptotic relative efficiency of the one-sample test procedure based on the T_n with respect to the test based on $\bar{X}_n$ is

$$e(T_n, \bar{X}_n) = 12\sigma_F^2 \left(\int_{-\infty}^{\infty} f^2(u)du \right)^2,$$

where $\sigma_F^2 = \mathrm{Var}_F(X) < \infty$.

(ii) $\inf_{F\in\mathcal{F}} e(T_n, \bar{X}_n) = \frac{108}{125} \approx 0.864$, where $\mathcal{F}$ is the family of CDFs satisfying continuous, symmetric and $\sigma_F^2 < \infty$. The equality is attained at F such that $f(x) = b(a^2 - x^2), |x| < a$, where $a = \sqrt{5}$ and $b = 3\sqrt{5}/20$.

Proof. (i) Similar to the proof of Corollary 5.1.1, we need to verify the conditions in Theorem 5.1.2. Let $T_{2n} = \frac{1}{C_n^2} T_n$. By Theorem 5.2.3, we know the T_{2n} is asymptotically normally distributed. It suffices to study its expectation and variance. It is easily to see that

$$\begin{aligned} E(T_{2n}) &= \frac{1}{C_n^2} \left[n(1 - F(-\theta)) + \frac{n(n-1)}{2} P_\theta(X_1 + X_2 > 0) \right] \\ &= P_\theta(X_1 + X_2 > 0) + O(n^{-1}) \approx \int [1 - F(-x-\theta)] f(x-\theta)dx. \end{aligned}$$

The variance is more complicated, however, by using Theorem 5.2.2,

$$\begin{aligned} \mathrm{Var}(T_{2n}) &= \frac{1}{n} 2^2 \mathrm{Var}(h_1(X_1)) + O(n^{-2}) \\ &\approx \frac{4}{n} \left\{ E(E^2(h(X_1, X_2) \mid X_1)) - (E(E(h(X_1, X_2) \mid X_1)))^2 \right\}. \\ &= \frac{4}{n} \left\{ E\left[1 - F(-X_1)\right]^2 - E^2 h(X_1, X_2) \right\} \\ &= \frac{4}{n} \left\{ \int [1 - F(-x-\theta)]^2 f(x-\theta)dx - \left(\int [1 - F(-x-\theta)] f(x-\theta)dx \right)^2 \right\}. \end{aligned}$$

Thus, to apply Pitman efficiency theorem, we choose $\mu_n(\theta) = \int F(x+\theta)f(x-\theta)dx$ and

$$\sigma_n^2(\theta) = \frac{4}{n}\left\{\int F^2(x+\theta)f(x-\theta)dx - \left(\int F(x+\theta)f(x-\theta)dx\right)^2\right\}.$$

Therefore, some calculation yields $\mu_n'(\theta) = 2\int f(x+\theta)f(x-\theta)dx$ and $\mu_n'(0) = 2\int f^2(u)du > 0$, while $\sigma_n^2(0) = \frac{4}{n}\text{Var}_F[F(X)] = \frac{4}{n}\frac{1}{12} = \frac{1}{3n}$. For $T_{1n} = \bar{X}_n$, choose $\mu_n(\theta) = \theta$ and $\sigma_n^2(\theta) = \sigma_F^2/n$. With these choices of $\mu_n(\theta)$ and $\sigma_n(\theta)$, the results are immediately follows from Theorem 5.1.2.

(ii) It can be shown that $e(T_n, \bar{X}_n)$ is location and scale invariant, so, we can assume that h is symmetric about 0 and $\sigma_F^2 = 1$. The problem, then, is to minimize $\int f^2(u)du$ subject to $\int f(u)du = \int u^2 f(u) = 1$ and $\int uf(u) = 0$ (by symmetry). This is equivalent to minimizing

$$\int f^2 + 2b\int u^2 f - 2ba^2\int f, \tag{5.2}$$

where a and b are positive constants to be determined later. We now write as (5.2)

$$\int [f^2 + 2b(x^2-a^2)f] = \int_{|x|\le a}[f^2 + 2b(x^2-a^2)f] + \int_{|x|>a}[f^2 + 2b(x^2-a^2)f]. \tag{5.3}$$

First complete the square on the first term on the right side of (5.3) to get

$$\int_{|x|\le a}[f + b(x^2-a^2)]^2 - \int_{|x|\le a} b^2(x^2-a^2)^2. \tag{5.4}$$

Now (5.3) is equal to the two terms of (5.4) plus the second term on the right side of (5.3). We can now write the density that minimizes (5.3).

If $|x| > a$ take $f(x) = 0$, since $x^2 > a^2$, and if $|x| \le a$ take $f(x) = b(a^2 - x^2)$, since the integral in the first term of (5.4) is nonnegative. We can now determine the values of a and b from the side conditions. From $\int f = 1$, we have

$$\int_{-a}^{a} b(a^2-x^2)dx = 1,$$

which implies that $a^3b = 3/4$. Further, from $\int x^2 f = 1$, we have $\int_{-a}^{a} x^2 b(a^2-x^2)dx = 1$, from which $a^5 b = 15/4$. Hence solving for a and b yields $a = \sqrt{5}$ and $b = 3\sqrt{5}/100$. Now,

$$\int f^2 = \int_{-\sqrt{5}}^{\sqrt{5}}\left[\frac{3\sqrt{5}}{100}(5-x^2)\right]^2 dx = \frac{3\sqrt{5}}{25},$$

which leads to the result, $\inf_{F\in\mathcal{F}} e(T_n, \bar{X}_n) = 12\left(\frac{3\sqrt{5}}{25}\right)^2 = \frac{108}{125} \approx 0.864$. □

Remark 5.2.2 Notice that the worst-case density f is not one of heavy tails but one with no tails at all (i.e., it has a compact support). Also note that the minimum Pitman efficiency is 0.864 in the class of symmetric densities with a finite variance, a very respectable lower bound.

F:	Normal	Uniform	Logistic	DE	Cauchy	t_3	t_5
$e(S_n, T_n)$	0.955	1.000	1.097	1.500	∞	1.900	1.240

The following table shows the value of the Pitman efficiency for several distributions that belong to the family of CDFs F defined in Theorem 5.2.5. They are obtained by direct calculation using the formula given above. It is interesting that, even in the normal case, the Wilcoxon test is 95% efficient with respect to the t-test.

5.2.2 Point estimator and confidence interval associated with the Wilcoxon signed rank statistic

The Wilcoxon signed-rank statistic T_n can be used to construct a point estimate for the point of symmetry of a symmetric density, and from it one can construct a confidence interval.

Suppose $X_1, \ldots, X_n \sim F$, where F has a symmetric density centered at θ. Consider to estimate the θ. When $\theta = 0$, the distribution of the statistic $T_n = \sum_{i \le j} I\left(\frac{X_i + X_j}{2} > 0\right)$ is symmetric about its mean $n(n+1)/4$. A natural estimator of θ is the amount $\widehat{\theta}$ that should be subtracted from each X_i so that the value of T_n, when applied to the shifted sample $X_1 - \widehat{\theta}, \ldots, X_n - \widehat{\theta}$, is as close to $n(n+1)/4$ as possible. Intuitively, we estimate θ by the amount $(\widehat{\theta})$ that the X sample should be shifted in order that $X_1 - \widehat{\theta}, \ldots, X_n - \widehat{\theta}$ as a sample from a population with median 0.

For any pair i, j with $i \le j$, define the Walsh average $W_{ij} = \frac{1}{2}(X_i + X_j)$ (see Walsh (1959)). Then the Hodges-Lehmann estimate $\widehat{\theta}$ is defined as

$$\widehat{\theta} = \text{Median}\{W_{ij} : 1 \le i \le j \le n\}.$$

Theorem 5.2.6 *Let $X_1, \ldots, X_n \sim F(x-\theta)$, where f, the density of F, is symmetric around zero. Let $\widehat{\theta}$ be the Hodges-Lehmann estimator of θ. Then, if $\int_{-\infty}^{\infty} f^2(u)du < \infty$,*

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N\left(0, \frac{1}{12\left\{\int_{-\infty}^{\infty} f^2(u)du\right\}^2}\right).$$

The proof of this theorem can be found in Hettmansperger and McKean (1998). For symmetric distributions, by CLT, $\sqrt{n}(\bar{X} - \theta) \xrightarrow{d} N(0, \sigma_F^2)$. The ratio of the variances in the two asymptotic distributions, $12\sigma_F^2\left(\int_{-\infty}^{\infty} f^2(u)du\right)^2$, is the ARE of $\widehat{\theta}$ relative to $\bar{X}_n$. This ARE equals to the asymptotic relative efficiency of the Wilcoxon signed rank test with respect to t-test in the testing problem (Theorem 5.2.5).

A confidence interval for θ can be constructed using the distribution of T_n. The interval is found from the following connection with the null distribution of T_n. Let $M = \frac{n(n+1)}{2}$ be the total number of Walsh averages $W_{(1)} \leq, \cdots \leq W_{(M)}$.

Theorem 5.2.7 (Tukey's method of confidence interval) *Let k_α denote the positive integer such that: $P(T_n < k_\alpha) = \alpha/2$. Then, $[W_{(k_\alpha)}, W_{(M-k_\alpha+1)}]$ is a confidence interval for θ at confidence level $1-\alpha$ $(0 < \alpha < 1/2)$.*

Proof. Write

$$\begin{aligned}
&P(\theta \in [W_{(k_\alpha)}, W_{(M-k_\alpha+1)}]) \\
&\quad = 1 - P\left(\theta < W_{(k_\alpha)}\right) - P\left(\theta > W_{(M-k_\alpha+1)}\right) \\
&\quad = 1 - P(T_n \geq M - k_\alpha + 1) - P(T_n \leq k_\alpha - 1) \\
&\quad = 1 - 2P(T_n < k_\alpha) = 1 - \alpha,
\end{aligned}$$

where we use the fact that T_n follows a symmetric distribution about $n(n+1)/4$ (Remark **??**). □

In practice, we can approximate k_α by using the asymptotic normality of T_n:

$$k_\alpha = \frac{n(n+1)}{4} - z_{\alpha/2}\sqrt{\frac{n(n+1)(2n+1)}{24}}.$$

For any continuous symmetric distribution, this confidence interval all holds. Hence, we can control the coverage probability to be $1 - \alpha$ whiteout having any more specific knowledge about the forms of the underlying X distributions. Thus, it is a distribution-free confidence interval for θ over a very large class of populations.