Chapter 12
Extensions of the CLT to Sums of Dependent Random Variables

12.1 Introduction

In this chapter, we consider some extensions of the Central Limit Theorem to classes of sums (or averages) of dependent random variables. Many further extensions are possible, but we focus on ones that will be useful in the sequel. Section 12.2 considers sampling without replacement from a finite population. As an application, the potential outcomes framework is introduced in order to study treatment effects. The class of U-statistics is studied in Section 12.3, with applications to the classical one-sample signed-rank statistic and the two-sample Wilcoxon rank-sum statistic. Section 12.4 considers CLTs for stationary, mixing sequences, which provides a basis for understanding robustness of procedures under dependence, as in Section 13.2.2. These three sections may be read independently of each other. A very general approach is provided by Stein's method in Section 12.5.

12.2 Random Sampling Without Replacement from a Finite Population

Let $\Pi_N = \{x_{N,1}, \ldots, x_{N,N}\}$ denote a population consisting of N real-valued units. Let $X_1, \ldots, X_n$ be a random sample taken without replacement from Π_N (so $n \leq N$), and let $\bar{X}_n = \sum_{i=1}^{n} X_i/n$ be the sample mean. Let $\bar{x}_N = \sum_{j=1}^{N} x_{N,i}/N$ be the mean of population Π_N and let

$$s_N^2 = \frac{1}{N}\sum_{j=1}^{N}(x_{N,j} - \bar{x}_N)^2 \tag{12.1}$$

denote its variance, so that $Var(X_i) = s_N^2$ for simple random sampling. Then, it is easy to check (Problem 12.1) that

E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, Springer Texts in Statistics, https://doi.org/10.1007/978-3-030-70578-7_12

$$E(\bar{X}_n) = \bar{x}_N \tag{12.2}$$

and

$$Var(\bar{X}_n) = \frac{s_N^2}{n} \cdot \frac{N-n}{N-1} . \tag{12.3}$$

Under certain conditions where both n and N tend to infinity, one might expect that $(\bar{X}_n - \bar{x}_N)/\sqrt{Var(\bar{X}_n)}$ is asymptotically $N(0, 1)$. The following theorem gives a sufficient condition.

Theorem 12.2.1 *Under the above setup, assume*

$$\frac{1}{\min(n, N-n)} \cdot \frac{\max_{1 \le j \le N} (x_{N,j} - \bar{x}_N)^2}{s_N^2} \to 0 . \tag{12.4}$$

Then, as $N \to \infty$,

$$\frac{(\bar{X}_n - \bar{x}_N)}{\sqrt{Var(\bar{X}_n)}} \xrightarrow{d} N(0, 1) . \tag{12.5}$$

Example 12.2.1 (**Two-Sample Wilcoxon Rank-Sum Test**) Consider the case where $x_{N,j} = j$. In other words, $\bar{X}_n$ is the average of n numbers taken without replacement from $\{1, \ldots, N\}$. Then $\bar{x}_N = (N+1)/2$ and

$$s_N^2 = \frac{N^2 - 1}{12} . \tag{12.6}$$

Therefore, (12.3) reduces to

$$Var(\bar{X}_n) = \frac{(N+1)(N-n)}{12n} . \tag{12.7}$$

Also,

$$\max_{1 \le j \le N} \left(j - \frac{N+1}{2} \right)^2 = \frac{(N-1)^2}{4}$$

and so the left side of (12.4) reduces to

$$\frac{1}{\min(n, N-n)} \cdot \frac{3(N-1)^2}{N^2 - 1} \to 0 .$$

If $\min(n, N-n) \to \infty$, then (12.5) follows.

In a statistical context, assume that $Y_1, \ldots, Y_m$ are i.i.d. F and, independently, $Z_1, \ldots, Z_n$ are i.i.d. G. Let $N = m + n$. Under the null hypothesis $F = G$ and the assumption that F is continuous, the $\binom{N}{n}$ assignment of ranks are all equally

likely. The Wilcoxon statistic, W_n, denotes the sum of the ranks of the Y_is. So, the distribution of W_n under $F = G$ is that of the sum of n numbers taken at random from $\{1, \ldots, N\}$, or $W_n = n\bar{X}_n$ in the statement of Theorem 12.2.1. Therefore, we conclude

$$\frac{W_n - \frac{1}{2}n(N+1)}{\sqrt{mn(N+1)/12}} \xrightarrow{d} N(0,1) \,. \blacksquare$$

The necessary and sufficient conditions for (12.5) were obtained from Hájek (1960), stated next.

Theorem 12.2.2 *Under the above setup, let*

$$\delta_{N,j} = (x_{N,j} - \bar{x}_N)/s_N \,.$$

Assume $\min(n, N-n) \to \infty$. *Then, (12.5) holds if and only if, for every* $\epsilon > 0$,

$$\frac{1}{N} \sum_{j:\, |\delta_{N,j}| > \epsilon\sqrt{n(N-n)/N}} \delta_{N,j}^2 \to 0 \,. \tag{12.8}$$

PROOF. We only prove the sufficiency part. Without loss of generality, assume $\bar{x}_N = 0$. Write $\bar{X}_n = \sum_{j=1}^N H_j x_{N,j}/n$, where $(H_1, \ldots, H_N)$ are indicator variables, with $H_j = 1$ if and only if item j is chosen in the sample (so that $\sum_j H_j = n$). Let $I_1, \ldots, I_N$ be i.i.d. Bernoulli variables with success probability n/N. Let $\tilde{X}_n = \sum_{j=1}^N I_j x_{N,j}/n$. Note that $\tilde{X}_n$ is an average of independent random variables, whose limiting distribution (after normalization) can be obtained by the Lindeberg Central Limit Theorem. Indeed, (12.5) holds with $\bar{X}_n$ replaced by $\tilde{X}_n$. We will show that $\tilde{X}_n$ and $\bar{X}_n$ have the same limiting distributions. We will apply Lemma 11.3.1 to $S_n = n\bar{X}_n$ and $\tilde{S}_n = n\tilde{S}_n$, and therefore the result follows once we show that

$$\frac{E[(\tilde{S}_n - S_n)^2]}{Var(\tilde{S}_n)} \to 0 \,. \tag{12.9}$$

To do this, we will first construct $I = (I_1, \ldots, I_N)$ and then construct $H = (H_1, \ldots, H_N)$ so that H is uniform over all vectors of length N with exactly n ones and $N - n$ zeroes in which case H and I are appropriately close (or "coupled").

First, let $B_N = \sum_{j=1}^N I_j$, which has the binomial distribution with parameters N and n/N. If $B_N = n$, just take $H = I$. If $B_N < n$, then let $H_j = 1$ whenever $I_j = 1$, which generates B_N out of the required n observations in the sample. Then, choose $n - B_N$ remaining indices at random without replacement among the remaining $N - B_N$ observations, and set $H_j = 1$ for those chosen indices. Similarly, if $B_N > n$, then there are too many j for which $I_j = 1$, so choose a subset of size n from B_N randomly without replacement.

Next, note that if $B_N > n$, then

$$\tilde{S}_n - S_n = \sum_{j=1}^{N} x_{N,j} I\{I_j = 1, H_j = 0\}$$

(because the terms $x_{N,j}$ for which $I_j = H_j = 1$ cancel and we cannot have $H_j = 1$ and $I_j = 0$ if $B_N > n$). Conditional on a value of $B_N > n$, $\tilde{S}_n - S_n$ is a sum of $B_N - n$ observations taken without replacement from the $x_{N,j}$s, and hence has mean 0. Moreover, using (12.3) with n replaced by $B_n - n$,

$$Var(\tilde{S}_n - S_n | B_N) = (B_N - n) s_N^2 \cdot \frac{N - B_N + n}{N - 1} \leq |B_N - n| s_N^2 \,, \tag{12.10}$$

if $B_n > n$.

Similarly, if $B_N < n$, then $\tilde{S}_n$ is a sum of B_N variables and we need $n - B_N$ more observations to construct the sample of size n. So, $S_n - \tilde{S}_n$ is a sum of $n - B_N$ variables, yielding the same bound as in (12.10). Since the bound is clearly true when $B_N = n$, (12.10) holds for all B_N.

Therefore, we can conclude that

$$E[(\tilde{S}_n - S_n)^2] = E[Var(\tilde{S}_n - S_n)|B_N] \leq s_N^2 E(|B_n - n|)$$

$$\leq s_N^2 \sqrt{Var(B_N)} \leq s_N^2 \sqrt{n(1 - \frac{n}{N})} \,.$$

(Note that if $\tilde{S}_n$ and S_n were independent, this term would be order n, not $\sqrt{n}$, a consequence of the coupling.) Since

$$Var(\tilde{S}_n) = s_N^2 n(1 - \frac{n}{N}) \,,$$

the left side of (12.9) is $1/\tau_N$, where

$$\tau_N = \sqrt{n(1 - \frac{n}{N})} \,. \tag{12.11}$$

But, $\tau_N \to \infty$ as $\min(n, N - n) \to \infty$ (Problem 12.5), as required. ■

In the above results, the $x_{N,j}$ are fixed, but in the study of permutation tests later on, they may sometimes be considered as outcomes of random variables. One might then apply Theorems 12.2.1 and 12.2.2 conditional on the outcomes. For this, we develop a simple and perhaps more intuitive sufficient condition. First, let Δ_N denote a random variable which is uniform on the N standardized values $(x_{N,j} - \bar{x}_N)/s_N$, with its distribution denoted by G_N. (Note here and below that ties are allowed and G_N is just the distribution of Δ_N.) In an asymptotic framework where the $x_{N,j}$ are fixed, we may nevertheless envision them settling down in such a way that G_N is getting close to some G.

Theorem 12.2.3 *Under the above setup with* $\min(n, N-n) \to \infty$, *assume* $G_N \xrightarrow{d} G$, *where* G *is a distribution with variance one. Then, (12.5) holds.*

PROOF. Let Δ denote a random variable with distribution G. Let τ_N be defined in (12.11), so that $\tau_N \to \infty$. We need to check condition (12.8). Fix any $\beta > 0$. Then, as soon as N is large enough so that $\epsilon\tau_N \geq \beta$, the left side of (12.8) is bounded above by

$$\frac{1}{N} \sum_{j:\ |\delta_{N,j}|>\beta} \delta_{N,j}^2 = E(\Delta_N^2 I\{|\Delta_N| > \beta\}) \to E(\Delta^2 I\{|\Delta| > \beta\})$$

if β is a continuity point of G, by (11.40). Since the set of continuity points of G is dense, β can be chosen large enough, while being a continuity point, to make the last expression as small as desired. ■

Note that since G_N itself has variance one, the condition requires that G_N converges in distribution to G and the variance of G_N converges to that of G. In the study of treatment effects, it is useful to generalize Theorem 12.2.2 to the bivariate case. The first part just restates the theorem under the assumptions that s_N^2 converges to a finite value (as it would when the $x_{N,j}$ mimic the population setting).

Corollary 12.2.1 *Assume* $\min(n, N-n) \to \infty$ *with* $n/N \to p$.
(i) Let F_N *be the distribution placing mass* $1/N$ *at each of the* $x_{N,j}$ *in* Π_N. *Let* s_N^2 *be defined as in (12.1) and assume* $s_N^2 \to \sigma^2 < \infty$. *Further assume* $F_N \xrightarrow{d} F$, *where* F *has variance* σ^2. *Then,*

$$\sqrt{n}(\bar{X}_n - \bar{x}_N) \xrightarrow{d} N(0, (1-p)\sigma^2)\ . \tag{12.12}$$

(ii) Generalizing to the bivariate case, suppose the population Π_N *consists of paired units* $x_{N,j} = (u_{N,j}, v_{N,j})$, $j = 1, \ldots, N$. *Let* F_N *be the (joint) distribution placing mass* $1/N$ *at each* $(u_{N,j}, v_{N,j})$. *Let*

$$\bar{u}_N = \frac{1}{N}\sum_{j=1}^N u_{N,j} \quad \text{and} \quad \bar{v}_N = \frac{1}{N}\sum_{j=1}^N v_{N,j}\ .$$

Let $s_{N,u}^2$ *denote the population variance of the* $u_{N,j}$, *as defined in (12.1), and let* $s_{N,v}^2$ *denote the population variance of the* $v_{N,j}$. *Define the population covariance as*

$$s_{N,uv} = \frac{1}{N}\sum_{j=1}^N (u_{N,j} - \bar{u}_N)(v_{N,j} - \bar{v}_N)\ .$$

Let $(U_1, V_1)^\top, \ldots, (U_n, V_n)^\top$ *denote a random sample taken without replacement from* Π_N, *and let* $\bar{U}_n$ *and* $\bar{V}_n$ *be the corresponding sample means. Assume*

$$s_{N,u}^2 \to \sigma_u^2 \quad \text{and} \quad s_{N,v}^2 \to \sigma_v^2 \, .$$

Finally, assume $F_N \xrightarrow{d} F$*, where* F *is a bivariate distribution with marginal variances* σ_u^2 *and* σ_v^2*. Then,* $s_{N,uv}$ *converges to a limit* σ_{uv} *(which is the covariance of the bivariate distribution* F*) and*

$$\sqrt{n}(\bar{U}_n - u_N, \bar{V}_n - v_N)^\top \xrightarrow{d} N(0, (1-p)\Sigma) \, , \tag{12.13}$$

where Σ *is the covariance matrix*

$$\Sigma = \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix} .$$

PROOF. Since Theorem 12.2.3 applies, (12.5) holds. But

$$nVar(\bar{X}_n) = s_N^2 \cdot \frac{N-n}{N-1} \to \sigma^2(1-p) \, ,$$

and (i) follows. To prove (ii), note the fact that $s_{N,uv} \to \sigma_{uv}$, where σ_{uv} is the covariance of F, follows by Problem 11.79. The bivariate asymptotic normality follows by using (i) together with the Cramér-Wold Device. ■

Example 12.2.2 (**Potential Outcomes and Treatment Effects**) Consider an experiment where Y_j is the observed outcome of interest for unit j, for $j = 1, \ldots, N$. Some units are treated while others serve as controls, and the outcomes could potentially vary under the two scenarios. Denote by $Y_j(1)$ the potential outcome for unit j if treated and $Y_j(0)$ the potential outcome for unit j if untreated. Let D_j denote an indicator variable that is 1 if unit j is treated and 0 if not. The observed outcomes can be expressed in terms of the potential outcomes and treatment assignments by the relationship

$$Y_j = Y_j(1)D_j + Y_j(0)(1 - D_j) \, . \tag{12.14}$$

It is assumed that the Y_j and D_j are observed, so that $Y_j(1)$ is observed if $D_j = 1$ but $Y_j(0)$ is observed if $D_j = 0$. Such a framework dates back to Neyman (1923) and is expanded upon by Rubin (1974).

So far, nothing has been assumed about the distribution of the variables introduced, observed or not. We now consider a specialized setting where the potential outcomes are nonrandom, and the randomness comes entirely from the treatment assignment vector $(D_1, \ldots, D_N)$. We will assume a fixed number n is allocated to treatment, with the remaining allocated to control, so that all $\binom{N}{n}$ combinations of treatment vectors are equally likely. The object of interest is the population average causal effect, θ_N, defined by

$$\theta_N = \frac{1}{N}\sum_{j=1}^{N}[Y_j(1) - Y_j(0)] = \bar{Y}_N(1) - \bar{Y}_N(0)\ , \tag{12.15}$$

where

$$\bar{Y}_N(k) = \frac{1}{N}\sum_{j=1}^{N} Y_j(k) \quad \text{for } k = 0, 1\ .$$

The usual unbiased estimator of the average treatment effect, θ_N, is then

$$\hat{\theta}_N = \frac{1}{n}\sum_{j=1}^{N} Y_j(1)D_j - \frac{1}{N-n}\sum_{j=1}^{N} Y_j(0)(1 - D_j)\ . \tag{12.16}$$

We would like to determine the limiting distribution of $\hat{\theta}_N$ as $N \to \infty$. Assume $n/N \to p \in (0, 1)$. Let the population variances and covariance be denoted by

$$s_N^2(k) = \frac{1}{N}\sum_{j=1}^{N}[Y_j(k) - \bar{Y}_N(k)]^2 \quad \text{for } k = 0, 1$$

and

$$s_N(0, 1) = \frac{1}{N}\sum_{j=1}^{N}[(Y_j(1) - \bar{Y}_N(1))(Y_j(0) - \bar{Y}_N(0))]\ .$$

Assume $s_N^2(k) \to s^2(k)$ and $s_N(0, 1) \to s(0, 1)$. Also, let

$$\tau_N^2 = \frac{1}{N}\sum_{j=1}^{N}[(Y_j(1) - Y_j(0)) - \theta_N]^2\ . \tag{12.17}$$

Simple algebra yields

$$\tau_N^2 = s_N^2(1) + s_N^2(0) - 2s_N(0, 1)$$

and so

$$\tau_N^2 \to \tau^2 = s^2(1) + s^2(0) - 2s(0, 1)\ .$$

Apply Corollary 12.2.1(ii) by taking $u_{N,j} = Y_j(1)$ and $v_{N,j} = Y_j(0)$. Then, let F_N denote the (empirical) distribution of the values $\{Y_j(1), Y_j(0)\}$. Assume F_N converges in distribution to F with covariance matrix Σ, and the diagonal elements of the covariance matrix of F_N, say Σ_N, converge to those of Σ. So, Σ has diagonal elements $s^2(1)$ and $s^2(0)$ with off-diagonal elements $s(0, 1)$. Then, Corollary 12.2.1(ii) yields that

$$\sqrt{n}\left[\frac{1}{n}(\sum_{j=1}^{N} Y_j(1)D_j, \sum_{j=1}^{N} Y_j(0)D_j) - (\bar{Y}_N(1), \bar{Y}_N(0))\right]$$
$$\xrightarrow{d} N(0, (1-p)\Sigma) \ . \tag{12.18}$$

Let the n units sampled correspond to those who receive treatment, and those not sampled the controls. Then, $\hat{\theta}_N$ can be expressed as

$$\frac{1}{n}\sum_{j=1}^{N} Y_j(1)D_j + \frac{n}{N-n}\frac{1}{n}\sum_{j=1}^{N} Y_j(0)D_j - \frac{1}{N-n}\sum_{j=1}^{N} Y_j(0) \ .$$

Let (U, V) denote a bivariate normal variable with mean vector 0 and covariance matrix $(1-p)\Sigma$. Apply the continuous mapping theorem, noting $n/(N-n) \to p/(1-p)$, to get

$$\sqrt{n}(\hat{\theta}_N - \theta_N) \xrightarrow{d} U + \frac{p}{1-p}V \ .$$

All that remains is to calculate the variance on the right-hand side. But,

$$\begin{aligned} Var(U + \frac{p}{1-p}V) &= Var(U) + \frac{p^2}{(1-p)^2}Var(V) + \frac{2p}{1-p}Cov(U, V) \\ &= (1-p)s^2(1) + \frac{p^2}{1-p}s^2(0) + 2ps(0, 1) \\ &= (1-p)s^2(1) + \frac{p^2}{1-p}s^2(0) + p[s^2(1) + s^2(0) - \tau^2] \\ &= s^2(1) + \frac{p}{1-p}s^2(0) - p\tau^2 \ . \end{aligned}$$

Therefore,

$$\sqrt{n}(\hat{\theta}_N - \theta_N) \xrightarrow{d} N\left(0, s^2(1) + \frac{p}{1-p}s^2(0) - p\tau^2\right) \ ,$$

(which actually holds even if $p = 0$ as long as $n \to \infty$) or equivalently

$$\sqrt{N}(\hat{\theta}_N - \theta_N) \xrightarrow{d} N\left(0, \frac{s^2(1)}{p} + \frac{s^2(0)}{1-p} - \tau^2\right) \ . \tag{12.19}$$

The above result is summarized in the following theorem.

Theorem 12.2.4 *Consider the above setup, where F_N is the distribution of the potential outcomes*

$$\{(Y_j(1), Y_j(0)), j = 1, \ldots, N\},$$

and F_N has covariance matrix Σ_N. Suppose n units are treated, where $n/N \to p \in (0, 1)$. Assume $F_N \xrightarrow{d} F$, where the diagonal elements of Σ_N converge to those of

Σ, where Σ is the covariance matrix of F. Then, the average treatment effect $\hat{\theta}_N$ defined in (12.16) satisfies (12.19).

Note that the limiting distribution depends on the limiting value of τ_N^2 defined in (12.17). The individual causal effects $Y_j(1) - Y_j(0)$, nor the average of their squares, cannot be estimated without further assumptions. Let us consider the implications of Theorem 12.2.4 for inference. First, consider Fisher's "sharp" null hypothesis H_F specified by

$$H_F : \ Y_j(1) = Y_j(0) , \quad j = 1, \ldots, N .$$

One can construct an exact level-α test by calculating a permutation test based on $\hat{\theta}_N$. That is, consider the permutation distribution defined as the empirical distribution of $\hat{\theta}_N$ recomputed over all $\binom{N}{n}$ treatment assignments. Theorem 12.2.4 gives its precise limiting behavior under H_F, where $\tau = 0$ and $s_1 = s_0$. Alternatively, one can apply a normal approximation. Under H_F, $\tau^2 = 0$ and the variance in the limiting distribution (12.19) simplifies, and only depends on $s^2(1)$ and $s^2(0)$. But, $s^2(1)$ can be estimated consistently (Problem 12.12) by

$$\hat{s}_N^2(1) = \frac{1}{n} \sum_{j=1}^{N} Y_j^2(1) D_i - \left[\frac{1}{n} \sum_{j=1}^{N} Y_j(1) D_i \right]^2 , \tag{12.20}$$

and similarly for $\hat{s}_N^2(0)$. Define $\hat{\sigma}_N^2$ by

$$\hat{\sigma}_N^2 = \frac{N\hat{s}_N^2(1)}{n} + \frac{N\hat{s}_N^2(0)}{N-n} \xrightarrow{P} \frac{s^2(1)}{p} + \frac{s^2(0)}{1-p} .$$

Then, the one-sided test that rejects H_F if $\sqrt{N}\hat{\theta}_N > \hat{\sigma}_N z_{1-\alpha}$ has limiting rejection probability equal to α under H_F. (Actually, Fisher proposed an alternative estimator of variance; see Ding (2017).)

On the other hand, consider Neyman's "weak" null hypothesis H_N specified by $\theta_N = 0$. In this case, Theorem 12.2.4 applies, but the limiting variance cannot be estimated consistently due to the presence of τ^2. Here, the one-sided test which rejects H_N if $\sqrt{N}\hat{\theta}_N > \hat{\sigma}_N z_{1-\alpha}$ has limiting rejection probability under H_N which is $\leq \alpha$ and is $< \alpha$ if $\tau > 0$. Thus, this approach, while valid, is conservative. Note, however, that improvements are possible by bounding τ^2, subject to constraints of the marginal distributions; see Aronow et al. (2015).

We can extend Theorem 12.2.4 to the setting where there is randomness, not only from treatment assignment, but also due to sampling. That is, assume there are N units, with N_0 of them sampled at random without replacement. Among the N_0 units in the experiment, a random sample of n are treated and $N_0 - n$ serve as controls.

Theorem 12.2.5 *Consider a population of size N. Let F_N be the distribution of the potential outcomes*

$$\{(Y_j(1), Y_j(0)), \ j = 1, \ldots, N\},$$

and let F_N have covariance matrix Σ_N. Assume $F_N \xrightarrow{d} F$, where the diagonal elements of Σ_N converge to those of Σ, where Σ is the covariance matrix of F and

Σ has diagonal elements $s^2(1)$ and $s^2(0)$. Sample N_0 without replacement, and then assign n at random to be treated and $N_0 - n$ to serve as controls. Assume $N_0/N \to f \in [0,1]$ and $n/N_0 \to p \in (0,1)$. Let $\hat{\theta}_{N_0}$ denote the estimated average treatment effect based on the N_0 units sampled, and let θ_N denote the population average treatment effect (for all N items). Then,

$$\sqrt{N_0}(\hat{\theta}_{N_0} - \theta_N) \xrightarrow{d} N\left(0, \frac{s^2(1)}{p} + \frac{s^2(0)}{1-p} - f\tau^2\right) , \tag{12.21}$$

where τ^2 is the limit of τ_N^2 defined in (12.17).

PROOF. Let $S_j = 1$ if item j is one of the N_0 sampled and 0 otherwise. Let $\bar{\theta}_{N_0}$ denote the average treatment effect for the N_0 items sampled; that is,

$$\bar{\theta}_{N_0} = \frac{1}{N_0}\sum_{j=1}^{N}[Y_j(1) - Y_j(0)]S_j .$$

Write

$$\sqrt{N_0}(\hat{\theta}_{N_0} - \theta_N) = A_N + B_N ,$$

where

$$A_N = \sqrt{N_0}(\hat{\theta}_{N_0} - \bar{\theta}_{N_0})$$

and

$$B_N = \sqrt{N_0}(\bar{\theta}_{N_0} - \theta_N) .$$

By Theorem 12.2.3 applied to $x_{N,j} = Y_j(1) - Y_j(0)$, $j = 1, \ldots, N$ and replacing n there with N_0, it follows that

$$B_n \xrightarrow{d} N(0, (1-f)\tau^2) .$$

But, conditional on the N_0 items sampled, we can apply Theorem 12.2.4 to conclude that (Problem 12.13)

$$A_N \xrightarrow{d} N\left(0, \frac{s^2(1)}{p} + \frac{s^2(0)}{1-p} - \tau^2\right) . \tag{12.22}$$

Note that, conditional on the N_0 items sampled, A_N and B_N are conditionally independent (since A_N is not even random). Therefore, we can apply Problem 11.73(i) to complete the proof by adding the limiting variances of A_N and B_N to get (12.21). ■

As expected, when N_0 is small relative to N, so that f =0, the limiting variance no longer depends on τ^2 and one can construct tests for Neyman's hypothesis that are no longer conservative. ■

12.3 U-Statistics

We begin by considering the one-sample case. Assume $X_1, \ldots, X_n$ are i.i.d. P on some general space. Suppose interest focuses on a real-valued parameter of the form

$$\theta(P) = E[h(X_1, \ldots, X_b)]$$

for some function $h(\cdot)$. The function $h(\cdot)$ is called the kernel of the U-statistic. It is assumed, without loss of generality, that $h(\cdot)$ is symmetric in its arguments. If it were not, it could be replaced by the average of h computed over all permutations of $X_1, \ldots, X_b$. We will also generally assume that

$$E[h^2(X_1, \ldots, X_b)] < \infty \, . \tag{12.23}$$

The corresponding U-statistic is defined (for $n \geq b$) by

$$U_n = \frac{1}{\binom{n}{b}} \sum_c h(X_{i_1}, \ldots, X_{i_b}) \, , \tag{12.24}$$

where $\sum_c$ denotes summation over the $\binom{n}{b}$ combinations of b-tuples $\{i_1, \ldots, i_b\}$ consisting of b distinct elements from $\{1, \ldots, n\}$. Of course, U_n is an unbiased estimator of $\theta(P)$. Notice that U_n is an average of identically distributed random variables, but the terms are independent only in the case $b = 1$. The goal will be to approximate the distribution of U_n, but first we consider some examples.

Example 12.3.1 (**Averages**) If $b = 1$, $U_n = \sum_{i=1}^n h(X_i)/n$ is indeed an average of i.i.d. random variables, and so pth sample moments are a special case with $h(x) = x^p$. Also, fixing t and letting $h(x) = I\{x \leq t\}$ yields the empirical c.d.f. evaluated at t. ■

Example 12.3.2 (**Sample Variance**) Consider the kernel

$$h(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2 \, .$$

Let $\sigma^2(P) = Var(X_i)$, assumed finite. Then,

$$\theta(P) = E[\frac{1}{2}(X_1 - X_2)^2] = \frac{1}{2}[Var(X_1) + Var(X_2)] = \sigma^2(P) \, .$$

Letting $\bar{X}_n = \sum_{i=1}^n X_i/n$, the corresponding U-statistic is

$$U_n = \frac{1}{2\binom{n}{2}} \sum_{i<j} (X_i - X_j)^2 = \frac{1}{2n(n-1)} \sum_{\text{all } i,j} (X_i - X_j)^2 =$$

$$\frac{1}{2n(n-1)} \sum_{\text{all } i,j} (X_i^2 - 2X_i X_j + X_j^2) = \frac{1}{2n(n-1)} \sum_{i=1}^{n} (2nX_i^2 - 2n^2 \bar{X}_n^2) =$$

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

the usual (unbiased version of the) sample variance. ■

Example 12.3.3 (**Gini's mean difference**) Let $h(x_1, x_2) = |x_1 - x_2|$, so that $\theta(P) = E(|X_1 - X_2|)$. The corresponding U-statistic

$$U_n = \frac{1}{\binom{n}{2}} \sum_{i<j} |X_i - X_j|$$

is known as Gini's mean difference. ■

In order to derive the limiting distribution of U_n, it is first helpful to derive its variance. Toward this end, for $1 \leq k \leq b$, define functions h_k as follows:

$$h_k(x_1, \ldots, x_k) = E[h(x_1, \ldots, x_k, X_{k+1}, \ldots, X_b)] . \tag{12.25}$$

Of course, $E[h_k(X_1, \ldots, X_k)] = \theta(P)$. Then, define

$$\zeta_k = Var[h_k(X_1, \ldots, X_k)] . \tag{12.26}$$

As we will soon see, asymptotic normality depends heavily on ζ_1 (and it being nonzero).

Example 12.3.4 (**Continuation of Example** 12.3.2) Here, $h_1(x_1)$ is given by

$$h_1(x_1) = \frac{1}{2} E[(x_1 - X_2)^2] = \frac{1}{2} [\sigma^2(P) + (x_1 - \mu(P))^2] .$$

Then,

$$\zeta_1 = \frac{1}{4} \left\{ E[(X - \mu(P))^4] - \sigma^4(P) \right\} .$$

Also, $h_2 = h$ and

$$\zeta_2 = Var[h(X_1, X_2)] = \frac{1}{4} E(X_1 - X_2)^4 - \sigma^4(P)$$

$$= \frac{1}{2} \{ E[(X - \mu(P))^4] + \sigma^4(P) \} . \blacksquare \tag{12.27}$$

Next, we consider a formula for the exact variance of a U-statistic.

Theorem 12.3.1 *Assume (12.23). Then, the variance of U_n is given by*

$$Var(U_n) = \frac{1}{\binom{n}{b}} \sum_{k=1}^{b} \binom{b}{k}\binom{n-b}{b-k}\zeta_k \,. \tag{12.28}$$

PROOF. Consider two combinations $(i_1, \ldots, i_b)$ and $(j_1, \ldots, j_b)$ of numbers between 1 and b, possibly overlapping. Suppose for the moment that in fact there are k indices in common. Then,

$$\begin{aligned} &Cov[h(X_{i_1}, \ldots, X_{i_b}), h(X_{j_1}, \ldots, X_{j_b})] \\ &= E[h(X_{i_1}, \ldots, X_{i_b})h(X_{j_1}, \ldots, X_{j_b})] - \theta^2(P) \\ &= E\left\{E[h(X_{i_1}, \ldots, X_{i_b}), h(X_{j_1}, \ldots, X_{j_b})|X_1, \ldots, X_k]\right\} - \theta^2(P). \end{aligned}$$

Since, by symmetry, we can assume that the common indices are $1, \ldots, k$, this becomes

$$E\left\{E[h(X_1, \ldots, X_k, X_{k+1}, \ldots, X_b)h(X_1, \ldots, X_k, X'_{k+1}, \ldots, X'_b)|X_1, \ldots, X_k]\right\} - \theta^2(P)\,,$$

where the X'_i are i.i.d. P and independent of the X_i. By (conditional) independence, this in turn becomes

$$E[h_k^2(X_1, \ldots, X_k)] - \theta^2(P) = \zeta_k \,.$$

Therefore, if there are k indices in common, we have

$$Cov[h(X_{i_1}, \ldots, X_{i_b}), h(X_{j_1}, \ldots, X_{j_b}] = \zeta_k \,. \tag{12.29}$$

If $\sum_c$ and $\sum_d$ both denote summation over all combinations, we can now calculate

$$\begin{aligned} Var(U_n) &= Cov\left[\binom{n}{b}^{-1} \sum_c h(X_{i_1}, \ldots, X_{i_b}), \binom{n}{b}^{-1} \sum_d h(X_{j_1}, \ldots, X_{j_b})\right] \\ &= \binom{n}{b}^{-2} \sum_c \sum_d Cov[h(X_{i_1}, \ldots, X_{i_b}), h(X_{j_1}, \ldots, X_{j_b})] \,. \end{aligned}$$

Each of the covariance terms that has k indices in common contributes ζ_k to the sum, by (12.29). But, the number of such terms sharing k indices is

$$\binom{n}{b}\binom{b}{k}\binom{n-b}{b-k} \,,$$

because there are $\binom{n}{b}$ ways to first pick $i_1, \ldots, i_b$, then $\binom{b}{k}$ ways to determine the indices in common, and then there are $\binom{n-b}{b-k}$ remaining ways to fill out $j_1, \ldots, j_b$. Putting this together yields

$$Var(U_n) = \binom{n}{b}^{-2} \sum_{k=1}^{b} \binom{n}{b}\binom{b}{k}\binom{n-b}{b-k}\zeta_k \ ,$$

and the result follows. ■

For any fixed nonnegative integers i and j, and n large enough so that $n - i > j$,

$$\binom{n-i}{j} = \frac{1}{j!}(n-i)(n-i-1)\cdots(n-i-j+1) \sim \frac{n^j}{j!} \ ,$$

where $p_n \sim q_n$ means $p_n/q_n \to 1$ as $n \to \infty$. It follows that the factor multiplying ζ_k in the kth term of (12.28) is equal to

$$\frac{1}{\binom{n}{b}}\binom{b}{k}\binom{n-b}{b-k} \sim \frac{b!}{n^b}\binom{b}{k}\frac{n^{b-k}}{(b-k)!} \sim k!\binom{b}{k}^2\frac{1}{n^k} \ .$$

Therefore, the following corollary is true.

Corollary 12.3.1 *Assume (12.23). Then, the variance of U_n satisfies*

$$Var(U_n) = \frac{b^2}{n}\zeta_1 + o\left(\frac{1}{n}\right) \ . \tag{12.30}$$

In fact, the error term in (12.30) is at most $O(1/n^2)$. Similarly, if $\zeta_1 = 0$, then

$$Var(U_n) = 2\binom{b}{2}^2\frac{1}{n^2}\zeta_2 + o\left(\frac{1}{n^2}\right) \ .$$

Clearly, the rate of convergence of $Var(U_n)$ to zero depends on the smallest value of j for which $\zeta_j > 0$. That is, if $\zeta_1 = \cdots \zeta_{j-1} = 0$ but $\zeta_j > 0$, then

$$n^j Var(U_n) \to j!\binom{b}{j}^2\zeta_j \ .$$

We are now in a position to prove asymptotic normality of U_n.

Theorem 12.3.2 *Assume (12.23) and $\zeta_1 > 0$. Then,*

$$\sqrt{n}[U_n - \theta(P)] - \frac{b}{n}\sum_{i=1}^{n}[h_1(X_i) - \theta(P)] \xrightarrow{P} 0 \tag{12.31}$$

and so

$$\sqrt{n}[U_n - \theta(P)] \xrightarrow{d} N(0, b^2\zeta_1) \tag{12.32}$$

PROOF. Define $\hat{U}_n$ so that

$$\hat{U}_n - \theta(P) = \frac{b}{n}\sum_{i=1}^{n}[h_1(X_i) - \theta(P)] \,. \tag{12.33}$$

By the Central Limit Theorem

$$\sqrt{n}[\hat{U}_n - \theta(P)] \xrightarrow{d} N(0, b^2\zeta_1) \,.$$

The result will follow by Slutsky's Theorem if we can show

$$\sqrt{n}[(U_n - \theta(P)) - (\hat{U}_n - \theta(P))] \xrightarrow{P} 0 \,. \tag{12.34}$$

But, $U'_n = U_n - \hat{U}_n$ is a U-statistic based on the kernel

$$h'(x_1, \ldots, x_b) = [h(x_1, \ldots, x_b) - \theta(P)] - \sum_{i=1}^{b}[h_1(x_i) - \theta(P)] \,.$$

Indeed, averaging h' over all combinations yields

$$U'_n = \frac{1}{\binom{n}{b}}\sum_c h'(X_{i_1}, \ldots, X_{i_b})$$

$$= U_n - \theta(P) - \frac{1}{\binom{n}{b}}\sum_c\sum_{j=1}^{b}[h_1(X_{i_j}) - \theta(P)]$$

$$= U_n - \theta(P) - \frac{\binom{n-1}{b-1}}{\binom{n}{b}}\sum_{i=1}^{n}[h_1(X_i) - \theta(P)]$$

$$= U_n - \theta(P) - \frac{b}{n}\sum_{i=1}^{n}[h_1(X_i) - \theta(P)] = U_n - \hat{U}_n \,,$$

as claimed. But, in obvious notation, the h'_1 corresponding to the kernel h' is zero, and so its variance ζ'_1 is 0 as well. By Corollary 12.3.1,

$$Var(U'_n) = Var(U_n - \hat{U}_n) = O(1/n^2) \,,$$

which certainly implies

$$E\{[\sqrt{n}(U_n - \hat{U}_n)]^2\} \to 0 \, .$$

By Chebychev's Inequality, (12.34) holds and the result follows. ■

Note the theorem is true as stated even if $\zeta_1 = 0$, if $N(0, 0)$ is interpreted as point mass at 0.

Example 12.3.5 (**Continuation of Example** 12.3.2) With U_n the (unbiased) version of the sample variance, we can conclude that

$$\sqrt{n}(U_n - \sigma^2(P)) \xrightarrow{d} N(0, 4\zeta_1) \, ,$$

where

$$\zeta_1 = \frac{1}{4}\left\{E[(X - \mu(P))^4] - \sigma^4(P)\right\} \, .$$

Note that it is possible that $\zeta_1 = 0$, as occurs when X_i is Bernoulli with success probability $1/2$. In this case, $U_n = n\hat{p}_n(1 - \hat{p}_n)/(n - 1)$, where $\hat{p}_n = \bar{X}_n$. Then, similar to Example 11.3.5, we can deduce in this case that (Problem 12.18)

$$n\left(U_n - \frac{1}{4}\right) \xrightarrow{d} -\frac{1}{4}\chi_1^2 + \frac{1}{4} \, . \blacksquare \tag{12.35}$$

Example 12.3.6 (**One-Sample Wilcoxon Signed-Rank Statistic**) Assume $X_1, \ldots, X_n$ are i.i.d. on the real line with c.d.f. F. Assume F is continuous (though the argument generalizes if F is not continuous). Let $h(x_1, x_2) = I\{x_1 + x_2 > 0\}$ and $\theta(F) = P\{X_1 + X_2 > 0\}$. Typically, U_n is used as a test of the null hypothesis H_0 that the center of the underlying distribution (assumed symmetric) is 0, in which case $\theta(F) = 1/2$. Then,

$$h_1(x) = 1 - F(-x) \, .$$

Under H_0,

$$\zeta_1 = Var[F(-X)] = VarF[(X)] = 1/12 \, ,$$

since $F(X)$ is distributed as $U(0, 1)$. Hence,

$$\sqrt{n}(U_n - \theta(F)) \xrightarrow{d} N(0, \frac{1}{3}) \, .$$

A variation is based on the usual one-sample Wilcoxon statistic V_n, which is described as follows. The assumption that F is continuous implies there are no ties with probability one. Rank $|X_1|, |X_2|, \ldots, |X_n|$ from smallest to largest and let R_i denote the rank of $|X_i|$. Define

$$V_n = \sum_{i=1}^{n} R_i I\{X_i > 0\} \, . \tag{12.36}$$

Then (Problem 12.19),

$$V_n = \binom{n}{2} U_n + S_n \ , \tag{12.37}$$

where S_n is the number of positive X_is. Since S_n is small compared with V_n, the limiting distribution of V_n can be obtained from U_n as

$$n^{-3/2}[V_n - E(V_n)] \xrightarrow{d} N(0, \frac{1}{12}) \ , \tag{12.38}$$

where $E(V_n) = n(n+1)/4$. ■

The proof of asymptotic normality of U_n was facilitated by introducing $\hat{U}_n$ in the proof, as defined in (12.33). At this point, this choice of definition may perhaps seem mysterious. But, we now investigate it as a special case of a general projection concept that has great utility. To begin, suppose $X_1, \ldots, X_n$ are mutually independent, though not necessarily i.i.d. The basic goal is to study the distribution of some statistic, say $T_n = T_n(X_1, \ldots, X_n)$. If one suspects that T_n is asymptotically normal, then it seems plausible that T_n can be approximated by a sum of independent variables of the form $\sum_{i=1}^n g_i(X_i)$. The following result describes an optimal choice of the g_i when minimizing the expected squared difference between T_n and the approximation.

Theorem 12.3.3 *Let $X_1, \ldots, X_n$ be independent, and $E(T_n^2) < \infty$.*
(i) The choice of $\sum_{i=1}^n g_i(X_i)$ minimizing

$$E\left\{[T_n - \sum_{i=1}^n g_i(X_i)]^2\right\} \tag{12.39}$$

is given by

$$\hat{T}_n = \sum_{i=1}^n E(T_n|X_i) - (n-1)E(T_n) \ ; \tag{12.40}$$

that is, taking $g_i(X_i) = E(T_n|X_i) - \frac{n-1}{n}E(T_n)$ minimizes (12.39).

(ii) For this choice, $E(\hat{T}_n) = E(T_n)$ and

$$E[(T_n - \hat{T}_n)^2] = Var(T_n) - Var(\hat{T}_n) \ . \tag{12.41}$$

PROOF. For any random variable Y with finite second moment, the choice of g minimizing $E\{[Y - g(Z)]^2\}$ is $g(Z) = E(Y|Z)$. Fix i, and apply this to (12.39) with $Y = T_n - \sum_{j \neq i} g_j(X_j)$ and $Z = X_i$. Then, g_i must satisfy

$$g_i(X_i) = E(T_n|X_i) - \sum_{j \neq i} E[g_j(X_j)]$$

or

$$g_i(X_i) - E[g_i(X_i)] = E(T_n|X_i) - \sum_{j=1}^{n} E[g_j(X_j)] \, .$$

Summing over i yields

$$\sum_{i=1}^{n} g_i(X_i) = \sum_{i=1}^{n} E(T_n|X_i) - (n-1) \sum_{j=1}^{n} E[g_j(X_j)] \, .$$

But certainly, $\sum_j E[g_j(X_j)]$ must be $E(T_n)$ because otherwise one could subtract the difference and further minimize (12.39).

To prove (ii), first calculate

$$Cov(T_n, \hat{T}_n) = Cov(T_n, \sum_{i=1}^{n} E(T_n|X_i)) = \sum_{i=1}^{n} Cov(T_n, E(T_n|X_i))$$

$$= \sum_{i=1}^{n} \left\{ E[T_n E(T_n|X_i)] - E^2(T_n) \right\} = \sum_{i=1}^{n} \left\{ E[E^2(T_n|X_i)] - E^2(T_n) \right\}$$

$$= \sum_{i=1}^{n} Var[E(T_n|X_i)] = Var(\sum_{i=1}^{n} E(T_n|X_i)) = Var(\hat{T}_n) \, .$$

Therefore,

$$Var(T_n - \hat{T}_n) = Var(T_n) + Var(\hat{T}_n) - 2Cov(T_n, \hat{T}_n) = Var(T_n) - Var(\hat{T}_n) \, ,$$

yielding (ii). ■

The function $\hat{T}_n$ is called the projection, because it projects T_n onto the linear space of all functions that are sums of independent random variables.

As a check, when $T_n = U_n$ is a U-statistic, then

$$E(U_n|X_i) = \frac{b}{n} h_1(X_i) + (1 - \frac{b}{n})\theta(P) \, , \tag{12.42}$$

and $\hat{T}_n$ agrees with $\hat{U}_n$ previously introduced in (12.33). Moreover, by Theorem 12.3.3(ii),

$$Var(U_n - \hat{U}_n) = Var(U_n) - Var(\hat{U}_n) = \frac{b^2}{n}\zeta_1 + O(\frac{1}{n^2}) - \frac{b^2}{n}\zeta_1 = O(\frac{1}{n^2}) \, .$$

Therefore, $\sqrt{n}(U_n - \theta(P))$ and $\sqrt{n}(\hat{U}_n - \theta(P))$ have the same normal limiting distribution as obtained before.

Next, we extend one-sample U-statistics to two-sample U-statistics. Suppose $X_1, \ldots, X_m$ are i.i.d. P and, independently, $Y_1, \ldots, Y_n$ are i.i.d. Q. The parameter of interest $\theta = \theta(P, Q)$ is given by

$$\theta(P, Q) = E[h(X_1, \ldots, X_a, Y_1, \ldots, Y_b)] ,$$

where the kernel h is a function of $a + b$ arguments, and assumed symmetric in its first a and its last b arguments. Also assume the kernel has a finite second moment. The corresponding U-statistic is then

$$U_{m,n} = \frac{1}{\binom{m}{a}\binom{n}{b}} \sum_c \sum_d h(X_{i_1}, \ldots, X_{i_a}, Y_{j_1}, \ldots, Y_{j_b}) .$$

Define

$$h_{1,0}(x) = E[h(x, X_2, \ldots, X_a, Y_1, \ldots .Y_b)]$$

and

$$h_{0,1}(y) = E[h(X_1, \ldots, X_a, y, Y_2, \ldots, Y_b)] .$$

Then, one can check (Problem 12.21) that the projection $\hat{U}_{m,n}$ of $U_{m,n}$ is given by

$$\hat{U}_{m,n} = \frac{a}{m} \sum_{i=1}^{m} [h_{1,0}(X_i) - \theta] + \frac{b}{n} \sum_{j=1}^{n} [h_{0,1}(Y_j) - \theta] + \theta . \tag{12.43}$$

Let $\zeta_{1,0} = Var[h_{1,0}(X)]$ and $\zeta_{0,1} = Var[h_{0,1}(Y)]$. If $\min(m, n) \to \infty$ with $m/n \to \lambda < \infty$, then by the CLT,

$$\sqrt{m}[\hat{U}_{m,n} - \theta(P, Q)] \xrightarrow{d} N(0, a^2\zeta_{1,0} + \lambda b^2 \zeta_{0,1}) . \tag{12.44}$$

The same is true if $\hat{U}_{m,n}$ is replaced by $U_{m,n}$. The argument requires showing that

$$\sqrt{m}(\hat{U}_{m,n} - U_{m,n}) \xrightarrow{P} 0$$

and is similar to the one-sample U-statistics case (Problem 12.22).

Example 12.3.7 (**Two-Sample Wilcoxon Statistic**) Let $h(x, y) = I\{x \le y\}$, so that $\theta(F, G) = P\{X \le Y\}$ when X and Y are independent, X has c.d.f. F and Y has c.d.f. G. Then,

$$U_{m,n} = \sum_{i=1}^{m} \sum_{j=1}^{n} I\{X_i \le Y_j\} .$$

The statistic $mnU_{m,n}$ is known as the Mann–Whitney statistic and is closely related to the Wilcoxon rank-sum statistic W_n in Example 12.2.1 (Problem 12.24). Now,

$$h_{1,0}(x) = 1 - G^-(x)$$

and

$$h_{0,1}(y) = F(y) \ ,$$

where $G^-(x) = P\{Y < x\}$ (so that $G^- = G$ if G is continuous). Then,

$$\zeta_{1,0} = Var[1 - G^-(X)] = E\left\{E[I\{X \le Y\}|X]^2\right\} - \theta^2(F, G)$$

$$= P\{X_1 \le Y_1 \ , X_1 \le Y_2\} - \theta^2(F, G) \ .$$

If F is continuous and $F = G$, $\zeta_{1,0} = 1/3 - 1/4 = 1/12$. Similarly,

$$\zeta_{0,1} = P\{X_1 \le Y_1, \ X_2 \le Y_1\} - \theta^2(F, G) \ , \tag{12.45}$$

which again reduces to $1/12$ when F is continuous and $F = G$. Note, however, that this method of proving asymptotic normality does not rely on the assumption $F = G$, unlike the method presented in Example 12.2.1.

Assume F is continuous. Under $H_0 : F = G$, the test that rejects H_0 when

$$\sqrt{m}|U_n - \frac{1}{4}| \ge \sqrt{\frac{1+\lambda}{12}} z_{1-\frac{\alpha}{2}} \tag{12.46}$$

has null rejection probability tending to α. On the other hand, for testing the null hypothesis $H_0' : P\{X \le Y\} = 1/2$ against two-sided alternatives where $P\{X \le Y\} \ne 1/2$, the same test does not control the probability of a Type 1 error, even in large samples. That is, there exists F and G satisfying H_0' such that the probability of a Type 1 error tends to some value $> \alpha$. Worse yet, the probability of a Type 3 or directional error can be large (Problem 12.25). ■

12.4 Stationary Mixing Processes

A stochastic process $\{X_t, \ t \in I\}$ is a collection of random variables, indexed by I, that are defined on some common probability space. In this section, we consider the case where I is the set of integers $\mathbf{Z}$, in which case the process may be referred to as a time series $\{X_j, \ j \in \mathbf{Z}\}$. Note that the X_j's may be random vectors, or more generally they may take values in some space S, though we focus on the case $S = \mathbf{R}$.

Dependence is typically the norm when considering random variables that evolve in time (or space). Data $X_1, \ldots, X_n$ may be regarded as a stretch of some time series.

In this section, we discuss some basic asymptotic theory for the normalized sum or average of such dependent random variables. Further references and historical notes are provided at the end of the chapter.

Any attempt to generalize central limit theorems from independent to dependent random variables must in some way rule out strong dependence. Indeed, in the example where $X_j = X_1$ for all j, asymptotic normality fails (unless X_1 is normal). Therefore, various types of weak dependence conditions have been used to capture the idea that observations separated far in time are approximately independent. In particular, Rosenblatt (1956) suggested the following notion of strong mixing, also called α-mixing.

Definition 12.4.1 For a time series $X = \{X_j,\ j \in \mathbf{Z}\}$, let $\mathcal{F}_m^n$ denote the σ-algebra generated by $\{X_j,\ m \le j \le n\}$. The corresponding mixing coefficients are defined by

$$\alpha_X(k) = \sup_n \sup_{A,B} |P(A \cap B) - P(A)P(B)|\ ,$$

where A and B vary over $\mathcal{F}_{-\infty}^n$ and $\mathcal{F}_{n+k}^{\infty}$, respectively. Then, the process X is called strong mixing (or α-mixing) if $\alpha_X(k) \to 0$ as $k \to \infty$.

A special case is the following.

Definition 12.4.2 The sequence $X = \{X_j,\ j \in \mathbf{Z}\}$ is m-dependent if $\alpha_X(k) = 0$ for all $k > m$.

Evidently, a 0-dependent sequence corresponds to a sequence of independent random variables.

Stochastic processes having a probabilistic structure that is invariant to shifts in time are called strictly stationary, or stationary for short.

Definition 12.4.3 The sequence $X = \{X_j,\ j \in \mathbf{Z}\}$ is stationary if, for any integers c, k and $j_1, \ldots, j_k$, the joint distribution of $(X_{j_1}, \ldots, X_{j_k})$ is the same as that of $(X_{j_1+c}, \ldots, X_{j_k+c})$.

In contrast, processes X satisfying $E(X_j)$ and $E(X_j^2)$ do not depend on j, as well as $Cov(X_j, X_k)$ depends on (j, k) only through $k - j$, are called weakly stationary or covariance stationary. In the case where $Cov(X_j, X_k) = \sigma^2 I\{j = k\}$, the process X is sometimes called a white noise process, or simply an uncorrelated sequence. Note that, for a covariance stationary process X, the function

$$R(k) = Cov(X_1, X_{k+1}) \tag{12.47}$$

is called the autocovariance (or just covariance) function of the process X.

Example 12.4.1 (**Moving Averages**) Suppose $\{\epsilon_j,\ j \in \mathbf{Z}\}$ is a collection of independent random variables. Then, the sequence

$$X_j = h(\epsilon_j, \epsilon_{j+1}, \ldots, \epsilon_{j+m})$$

is a sequence of m-dependent random variables, where h is any (measurable) function from $\mathbf{R}^{m+1}$ to $\mathbf{R}$. In the special case where h is of the form

$$h(\epsilon_1, \ldots, \epsilon_{m+1}) = \sum_{i=1}^{m+1} w_i \epsilon_i$$

for constants $w_1, \ldots w_{m+1}$, the process $X = \{X_j, \ j \in \mathbf{Z}\}$ is known as a moving average process of order m. The case $w_i = 1/(m+1)$ is a simple moving average process. If the ϵ_j are also i.i.d., then the process X is stationary as well. If the ϵ_j sequence is weakly stationary, then so is X. Moreover, if the ϵ_j is an uncorrelated sequence with variance σ_ϵ^2, then an easy calculation (Problem 12.30) gives

$$R(k) = \sigma_\epsilon^2 (w_1 w_{1+k} + \cdots + w_{m+1-k} w_{m+1}) \quad \text{if} \ \ 0 \le k \le m \ , \tag{12.48}$$

and $R(k) = 0$ if $k > m$. ■

Example 12.4.2 (**Autoregressive Process**) Suppose

$$X_j = \rho X_{j-1} + \epsilon_j \ , \tag{12.49}$$

where the ϵ_j are i.i.d. with distribution F (though one can also consider the case where the ϵ_j are just weakly stationary). Such a process is known as an autoregressive process of order 1, denoted by AR(1). At this point, it may not be clear that a process X satisfying the $X_j - \rho X_{j-1}$ are i.i.d. with distribution F even exists. Assuming its existence for the moment, we may iterate (12.49) to get

$$X_j = \rho X_{j-1} + \epsilon_j = \rho(\rho X_{j-2} + \epsilon_{j-1}) + \epsilon_j = \rho^2 X_{j-2} + \rho \epsilon_{j-1} + \epsilon_j$$

and, in general,

$$X_j = \rho^m X_{j-m} + \sum_{i=0}^{m-1} \rho^i \epsilon_{j-i} \ . \tag{12.50}$$

Moreover, (12.50) suggests that, when $|\rho| < 1$, we can define

$$X_j = \sum_{i=0}^{\infty} \rho^i \epsilon_{j-i} \ , \tag{12.51}$$

where the infinite series is a well-defined random variable if the ϵ_j have a finite first moment. Indeed,

$$E\left(\sum_{i=0}^{\infty} |\rho^i \epsilon_{j-i}| \right) \le E(|\epsilon_1|) \sum_{i=0}^{\infty} |\rho|^i < \infty \ ,$$

which implies that $\sum_{i=0}^{\infty} |\rho^i \epsilon_{j-i}|$ is finite with probability one, and so is the right side of (12.51) with probability one. [1]

The representation (12.51) may be viewed as a moving average process of infinite order. Furthermore, with X_j defined as in (12.51), one may now check that the random variables $X_j - \rho X_{j-1} = \epsilon_j$ are i.i.d. F. It also follows from the representation (12.51) that $\{X_j,\ j \in \mathbf{Z}\}$ is stationary.

Assume the distribution F of the ϵ_j's has mean μ_ϵ and finite variance σ_ϵ^2. Then, (12.51) also implies that (Problem 12.31)

$$E(X_j) = \frac{\mu_\epsilon}{1-\rho}$$

and

$$R(k) = \sigma_\epsilon^2 \cdot \frac{\rho^k}{1-\rho^2} \tag{12.52}$$

so that the covariances decay geometrically (or exponentially) fast.

Finally, if the distribution F is absolutely continuous with respect to Lebesgue measure, then it is known that the process X is α-mixing, and the mixing coefficients decay geometrically fast as well; see Mokkadem (1988). Surprisingly, a stationary AR(1) process need not be mixing. A well-known counterexample can be obtained with $\rho = 1/2$ and taking F to be the distribution placing mass $1/2$ at both 0 and 1; see Section 2.3.1 of Doukhan (1995). ■

Strong mixing has important implications concerning the covariance between random variables separated in time.

Lemma 12.4.1 *Give a sequence $X = \{X_j,\ j \in \mathbf{Z}\}$ with mixing coefficients $\alpha_X(\cdot)$, assume U and V are $\mathcal{F}_{-\infty}^n$ and $\mathcal{F}_{n+k}^{\infty}$ measurable.*
(i) If U and V are bounded by one in absolute value, then

$$|Cov(U, V)| \le 4\alpha_X(k)\ .$$

(ii) If $E(|U|^p) < \infty$ and $E(|V|^q) < \infty$ for some p and q with $\frac{1}{p} + \frac{1}{q} < 1$, then

$$|Cov(U, V)| \le 8[E(|U|^p)]^{1/p}[E(|V|^q)]^{1/q}\alpha_X^{1-\frac{1}{p}-\frac{1}{q}}(k)\ .$$

A proof of Lemma 12.4.1 can be found in the appendix of Hall and Heyde (1980).

As usual, let $\bar{X}_n = \sum_{i=1}^n X_i/n$. Before considering asymptotic normality of $\sqrt{n}[\bar{X}_n - E(\bar{X}_n)]$, we consider its variance. If the X_i have a finite variance, then

[1] Alternatively, the celebrated Kolmogorov Three-Series Theorem may be used to easily show that the series (12.51) converges with probability one; see Billingsley (1995), Theorem 22.8. In addition, if $Var(\epsilon_j) < \infty$, we may write, $X_j = \lim_{m\to\infty} X_{m,j}$, where $X_{m,j} = \sum_{i=0}^{m-1} \rho^i \epsilon_{j-i}$, and the limit can be interpreted in the mean-squared sense; see Problem 11.65.

$$Var(\sqrt{n}\bar{X}_n) = \frac{1}{n}\sum_{i=1}^{n} Var(X_i) + \frac{2}{n}\sum_{i<j} Cov(X_i, X_j) \ . \tag{12.53}$$

Further assume weak stationarity and recall $R(k) = Cov(X_1, X_{k+1})$. Then, (12.53) simplifies to (Problem 12.32)

$$Var(\sqrt{n}\bar{X}_n) = Var(X_1) + 2\sum_{k=1}^{n-1}(1 - \frac{k}{n})R(k) \ . \tag{12.54}$$

If $R(k) \to 0$ sufficiently fast as $k \to \infty$, then we can expect

$$Var(\sqrt{n}\bar{X}_n) \to Var(X_1) + 2\sum_{k=1}^{\infty} R(k) < \infty \ . \tag{12.55}$$

Certainly, (12.55) holds if the process is also m-dependent because the infinite sum becomes a finite sum. If the process X is stationary and bounded in absolute value by one with α-mixing coefficients $\alpha_X(\cdot)$ that are summable, then by Lemma 12.4.1(i), $|R(k)| \leq 4\alpha_X(k)$. Thus,

$$\sum_{k=1}^{n-1}(1 - \frac{k}{n})|R(k)| \to \sum_{k=1}^{\infty}|R(k)| \leq 4\sum_{k=1}^{\infty}\alpha_X(k) < \infty \ ,$$

and then (12.55) holds. Finally, if X is stationary with $E(|X_1|^{2+\delta}) < \infty$ for some $\delta > 0$, then by Lemma 12.4.1(ii),

$$|R(k)| \leq C\alpha_X^{\frac{\delta}{2+\delta}}(k) \ ,$$

for some constant $C < \infty$ (which depends on δ). Thus, if the mixing coefficients satisfy

$$\sum_{k=1}^{\infty}\alpha_X^{\frac{\delta}{2+\delta}}(k) < \infty \ , \tag{12.56}$$

then we can also conclude that (12.55) holds (Problem 12.33). Assuming asymptotic normality holds, we can then expect

$$\sqrt{n}[\bar{X}_n - E(X_1)] \xrightarrow{d} N(0, \sigma_\infty^2) \ ,$$

where σ_∞^2 is given by the right-hand side of (12.55). In the m-dependent case, the following holds. (References for proofs are provided in the notes at the end of the chapter.)

Theorem 12.4.1 *Assume $X_1, X_2, \ldots$ is a stationary m-dependent sequence with mean μ and $Var(X_1) < \infty$. Then,*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma_\infty^2) , \tag{12.57}$$

where

$$\sigma_\infty^2 = Var(X_1) + 2\sum_{k=1}^{m} R(k) . \tag{12.58}$$

Example 12.4.3 (**Runs of Bernoulli Trials**) Suppose $\epsilon_1, \epsilon_2, \ldots$ is an i.i.d. sequence of Bernoulli trials, each trial having success probability p. Fix m and let

$$Y_{i,m} = I\{\epsilon_i = \cdots = \epsilon_{i+m} = 1\}$$

denote the indicator of the event of $m+1$ heads in a row starting at i. We would like to determine the limiting distribution of $\sum_{i=1}^{n-m} Y_{i,m}$, suitably normalized. But, the $Y_{i,m}$ form an m-dependent strictly stationary sequence, and so Theorem 12.4.1 applies. We need to calculate σ_∞^2 in (12.58). Note that $E(Y_{i,m}) = p^{m+1}$. Moreover, for $k \geq 0$,

$$Cov(Y_{1,m}, Y_{1+k,m}) = E(Y_{1,m}Y_{1+k,m}) - E(Y_{1,m})E(Y_{1+k,m}) = p^{m+1+k} - p^{2m+2} .$$

Therefore,

$$\sigma_\infty^2 = p^{m+1}(1 - p^{m+1}) + 2\sum_{k=1}^{m}(p^{m+1+k} - p^{2m+2})$$

$$= p^{m+1} - p^{2m+2} + (2\sum_{k=1}^{m} p^{m+1+k}) - 2mp^{2m+2}$$

$$= p^{m+1} - (2m+1)p^{2m+2} + 2p^{m+1}\left(\frac{p(1-p^m)}{1-p}\right)$$

$$= p^{m+1} - (2m+1)p^{2m+2} + \frac{2p^{m+2} - 2p^{2m+2}}{1-p} .$$

It follows that

$$(n-m)^{-1/2}\sum_{i=1}^{n-m}(Y_{i,m} - p^{m+1}) \xrightarrow{d} N(0, \sigma_\infty^2) .$$

Statistics like $\sum_{i=1}^{n-m} Y_{i,m}$ have been used to test alternatives to Bernoulli sequences; see Ritzwoller and Romano (2021) and their analysis of various tests as applied to the so-called hot hand fallacy. ■

Example 12.4.4 In Theorem 12.4.1, it is possible to have $\sigma_\infty^2 = 0$ even when $Var(X_1) > 0$. To see how, let $\{\epsilon_j,\ j \in \mathbf{Z}\}$ be i.i.d. N(0,1) and set $X_j = \epsilon_j - \epsilon_{j-1}$. Then, the X_js form a stationary 1-dependent sequence with mean $\mu = 0$ and

$$\sigma_\infty^2 = Var(X_1) + 2R(1) = 2 + 2Cov(\epsilon_2 - \epsilon_1, \epsilon_1 - \epsilon_0) = 0 .$$

The theorem still holds with the interpretation $\sqrt{n}\bar{X}_n \xrightarrow{P} 0$. ■

The following theorem, due to Ibragimov (1962), provides a Central Limit Theorem for stationary strong mixing sequences.

Theorem 12.4.2 *Suppose* $X = \{X_j,\ j \in \mathbf{Z}\}$ *is stationary, mean* μ*, with* $E(|X_1|^{2+\delta}) < \infty$ *for some* $\delta > 0$*. Assume the mixing coefficients* $\alpha_X(\cdot)$ *of* X *satisfy*

$$\sum_{j=1}^{\infty} [\alpha_X(j)]^{\frac{\delta}{2+\delta}} < \infty . \tag{12.59}$$

Then,

$$n^{1/2}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma_\infty^2) ,$$

where σ_∞^2 *is finite and given by (12.55).*

Example 12.4.5 (**Sample Autocovariance**) Let X be a stationary process with mean μ and covariance function $R(\cdot)$. Assume $E(|X_1|^{4+2\delta}) < \infty$ for some δ, and assume (12.59) as well. Let $\hat{R}_n(1)$ be the sample autocovariance at lag 1; that is,

$$\hat{R}_n(1) = \frac{1}{n-1} \sum_{i=1}^{n-1} (X_i - \bar{X}_n)(X_{i+1} - \bar{X}_n) .$$

In order to obtain the limiting distribution of $\sqrt{n}[\hat{R}_n(1) - R(1)]$, first consider $\bar{R}_n(1)$ defined by

$$\bar{R}_n(1) = \frac{1}{n-1} \sum_{i=1}^{n-1} Y_i ,$$

where $Y_i = (X_i - \mu)(X_{i+1} - \mu)$. Then, $Y_1, Y_2, \ldots$ is stationary with mean $R(1)$. We can apply Theorem 12.4.2 to deduce that

$$\sqrt{n}[\bar{R}_n(1) - R(1)] \xrightarrow{d} N\left(0, Var(Y_1) + 2\sum_{k=1}^{\infty} Cov(Y_1, Y_{1+k})\right) .$$

Then, by simple algebra and the fact that $\sqrt{n}(\bar{X}_n - \mu)^2 \xrightarrow{P} 0$, it follows that (Problem 12.37)

$$\sqrt{n}[\hat{R}_n(1) - \bar{R}_n(1)] \xrightarrow{P} 0 \,, \tag{12.60}$$

and so $\hat{R}_n(1)$ and $\bar{R}_n(1)$ have the same limiting distribution. ■

12.5 Stein's Method

In this section, Stein's (1972) method is introduced as a general technique for approximating the distribution of a sum (or average) of possibly dependent random variables. In particular, we focus on normal approximation, though the method is more general; see the notes at the end of the chapter. As will be seen, the method also produces error bounds, similar to that in the Berry–Esseen Theorem.

A useful starting point is the following characterization of a random variable W having the standard normal distribution. A random variable W satisfies

$$E[f'(W)] = E[Wf(W)] \tag{12.61}$$

for all "smooth" f if and only if W has the standard normal distribution. We will formalize this characterization below. But, note that if f is bounded and absolutely continuous and ϕ is the standard normal density, then integration by parts yields

$$E[f'(W)] = \int_{-\infty}^{\infty} f'(w)\phi(w)dw = f(w)\phi(w)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} f(w)\phi'(w)dw$$

$$= 0 + \int_{-\infty}^{\infty} f(w)w\phi(w)dw = E[Wf(W)] \,.$$

A rough strategy will be to argue that W is approximately standard normal if

$$E[f'(W)] - E[Wf(W)] \approx 0$$

in some sense. Unlike classical Fourier methods, this will be accomplished by using local perturbations of W. In order to get a quick idea of how this may be possible, consider the simple case where $X_1, \ldots, X_n$ are i.i.d. with $E(X_i) = 0$ and $Var(X_i) = 1$. Let

$$W = \frac{X_1 + \cdots X_n}{\sqrt{n}}$$

and

$$W_i = W - \frac{X_i}{\sqrt{n}} \,,$$

so that W_i and X_i are independent. It follows that

$$E[X_i f(W_i)] = E(X_i)E[f(W_i)] = 0$$

and also

$$E[X_i f(W)] = E\{X_i[f(W) - f(W_i)]\}$$

$$\approx E[X_i(W - W_i)f'(W)] = \frac{1}{\sqrt{n}} E[X_i^2 f'(W)] \ . \tag{12.62}$$

Therefore,

$$E[Wf(W)] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E[X_i f(W)] \approx \frac{1}{n} \sum_{i=1}^{n} E[X_i^2 f'(W)]$$

$$= E\left[f'(W) \cdot \frac{1}{n} \sum_{i=1}^{n} X_i^2 \right] \approx E[f'(W)] \, ,$$

since by the law of large numbers, $\sum X_i^2/n \approx 1$. Such an approach will be made rigorous later, and it will also produce error bounds.

A more formal statement of the characterization (12.61) is given by the following lemma. For any function $g(\cdot)$ of a real-variable, let $\|g\| = \sup_w |g(w)\|$.

Lemma 12.5.1 *If Z has the standard normal distribution, denoted by $Z \sim N(0, 1)$, then $E[f'(Z)] = E[Zf(Z)]$ for all absolutely continuous f with $E|f'(Z)| < \infty$. Conversely, if a random variable W satisfies $E[f'(W)] = E[Wf(W)]$ for all absolutely continuous f with $\|f'\| < \infty$, then $W \sim N(0, 1)$.*

PROOF OF LEMMA 12.5.1. Assume $Z \sim N(0, 1)$ with $E|f'(Z)| < \infty$. By Fubini's Theorem,

$$E[f'(Z)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f'(z) e^{-z^2/2} dz$$

$$= \frac{1}{\sqrt{2\pi}} \int_0^{\infty} f'(z) \int_z^{\infty} x e^{-x^2/2} dx dz + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} f'(z) \int_{-\infty}^{z} -x e^{-x^2/2} dx dz$$

$$= \frac{1}{\sqrt{2\pi}} \int_0^{\infty} x e^{-x^2/2} \int_0^{x} f'(z) dz dx + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} -x e^{-x^2/2} \int_x^{0} f'(z) dz dx$$

$$= E\{Z[f(Z) - f(0)]\} = E[Zf(Z)] \, .$$

The proof of the second part of Lemma 12.5.1 is left as Problem 12.39, but follows easily from Lemma 12.5.2 below. ■

As usual, let $\Phi(\cdot)$ denote the standard normal c.d.f.

Lemma 12.5.2 *Fix $x \in \mathbf{R}$. The unique bounded solution $f_x(\cdot)$ to the differential equation*

$$f_x'(w) - wf_x(w) = I\{w \le x\} - \Phi(x) \tag{12.63}$$

is given by

$$f_x(w) = e^{w^2/2} \int_w^\infty e^{-t^2/2}[\Phi(x) - I\{t \le x\}]dt \tag{12.64}$$

$$= -e^{w^2/2} \int_{-\infty}^w e^{-t^2/2}[\Phi(x) - I\{t \le x\}]dt \ . \tag{12.65}$$

Hence,

$$f_x(w) = \begin{cases} \sqrt{2\pi}e^{w^2/2}\Phi(w)[1 - \Phi(x)] & \text{if } w \le x \\ \sqrt{2\pi}e^{w^2/2}\Phi(x)[1 - \Phi(w)] & \text{if } w > x \ . \end{cases} \tag{12.66}$$

PROOF. Multiply both sides of (12.63) by $e^{-w^2/2}$ to get

$$\frac{d(e^{-w^2/2} f_x(w))}{dw} = e^{-w^2/2}[I\{w \le x\} - \Phi(x)] \ .$$

Integration then yields, for an arbitrary constant C,

$$f_x(w) = e^{w^2/2} \int_{-\infty}^w [I\{t \le x\} - \Phi(x)]e^{-t^2/2}dt + Ce^{t^2/2} \ . \tag{12.67}$$

The only bounded solution requires $C = 0$, and then (12.67) agrees with (12.65). The equivalence of (12.64) and (12.65) is easy to check, as is (12.66) (Problem 12.40). ■

By replacing w with a random variable W in (12.63) and then taking expectations, it follows that

$$|P\{W \le x\} - \Phi(x)| = |E[f_x'(W) - Wf_x(W)]| \ .$$

Therefore, Stein's method is to bound the right side.

In general, normal approximation may be specified by $E[h(W)] \approx E[h(Z)]$ for certain functions h in some specified class $\mathcal{H}$. Define, for random variables X and Y, the distance $d_{\mathcal{H}}(X, Y)$ by

$$d_{\mathcal{H}}(X, Y) = \sup_{h \in \mathcal{H}} |E[h(X)] - E[h(Y)]| \ .$$

Note that d really is a measure of closeness on the space of distributions of X and Y and, depending on the choice of $\mathcal{H}$, is a metric in the usual sense. For example, the choice

$$\mathcal{H} = \{I\{\cdot \le x\] : \ x \in \mathbf{R}\}$$

corresponds to the Kolmogorov metric, denoted by d_K. The choice

$$\mathcal{H} = \{h : \mathbf{R} \to \mathbf{R} : \ |h(x) - h(y)| \le |x - y|\}$$

is the Wasserstein metric, denoted by d_W. Finally, the choice

$$\mathcal{H} = \{I\{\cdot \in A\} : \ A \in \text{Borel sets}\}$$

is the total variation metric d_{TV}.

For a random variable W and $Z \sim N(0, 1)$, in order to compare $E[h(W)]$ with $E[h(Z)]$, fix a function h with $E|h(Z)| < \infty$. Let $f = f_h$ be the solution to the *Stein equation* given by

$$f_h'(w) - w f_h(w) = h(w) - E[h(Z)] \ . \tag{12.68}$$

Then,

$$d_{\mathcal{H}}(W, Z) = \sup_{h \in \mathcal{H}} |E[f_h'(W)] - E[W f_h(W)]| \ . \tag{12.69}$$

(Thus, for a real number x, the notation f_x introduced earlier really corresponds to f_h with $h(\cdot) = I\{\cdot \le x\}$.)

Before we can exploit (12.69), it is helpful to record certain smoothness properties of the solution f_h to (12.68).

Lemma 12.5.3 *The solution f_h to the Stein equation (12.68) can be written as*

$$f_h(w) = -e^{w^2/2} \int_w^\infty e^{-t^2/2}[h(t) - Eh(Z)]dt \tag{12.70}$$

$$= e^{w^2/s} \int_{-\infty}^w e^{-t^2/2}[h(t) - Eh(Z)]dt \ . \tag{12.71}$$

If h is bounded, i.e. $\|h\| < \infty$, then

$$\|f_h\| = \sup_w |f_h(w)| \le \sqrt{\frac{\pi}{2}} \|h - Eh(Z)\|$$

and

$$\|f_h'\| \le 2\|h - Eh(Z)\| \ .$$

If h is absolutely continuous, then

$$\|f_h\| \le 2\|h'\| \ , \quad \|f_h'\| \le \sqrt{\frac{2}{\pi}} \|h'\| \ , \quad \textit{and} \quad \|f_h''\| \le 2\|h'\| \ . \tag{12.72}$$

The proof of (12.70) and (12.71) is analogous to the proof of (12.64) and (12.65). The rest of the proof is somewhat technical, but an argument can be found in the

appendix to Chapter 2 in Chen et al. (2011). At least intuitively, the solution f_h should be smoother than h since Equation (12.68) equates a function of f_h and f_h' to a shift of h. In what follows, the bounds in (12.72) will be important, but otherwise the arguments will be self-contained.

The next goal is to prove a result in the spirit of Berry–Esseen, except that, for simplicity, we work with the metric d_W. Of course, when $W = W_n$ is indexed by n, then $d_W(W_n, Z) \to 0$ implies weak convergence. Results for d_K may be found in Ross (2011) or Chen et al. (2011). The first step is the following lemma, where independence is assumed.

Lemma 12.5.4 *Assume $X_1, \ldots, X_n$ are independent with $E(X_i) = 0$ and $Var(X_i) = 1$. Let $W = n^{-1/2} \sum_{i=1}^n X_i$. Let f satisfy $\|f''\| \le C < \infty$. Then,*

$$|E[Wf(W)] - E[f'(W)]| \le \frac{3C}{2n^{3/2}} \sum_{i=1}^n E(|X_i|^3) \,. \tag{12.73}$$

PROOF. Let $W_i = W - n^{-1/2} X_i$, so that W_i and X_i are independent. Then $E[X_i f(W_i)] = 0$ and so

$$E[X_i f(W)] = E\{X_i[f(W) - f(W_i)]\} =$$

$$E\{X_i[f(W) - f(W_i)] - X_i(W - W_i) f'(W_i)\} + E[X_i(W - W_i) f'(W_i)] \,.$$

Summing over i and dividing by $\sqrt{n}$ yields

$$E[Wf(W)] = \frac{1}{\sqrt{n}} \sum_{i=1}^n E\{X_i[f(W) - f(W_i) - (W - W_i) f'(W_i)]\}$$

$$+ \frac{1}{n} \sum_{i=1}^n E[X_i^2 f'(W_i)]$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^n E\{X_i[f(W) - f(W_i) - (W - W_i) f'(W_i)]\} \tag{12.74}$$

$$+ \frac{1}{n} \sum_{i=1}^n E[f'(W_i)] \,. \tag{12.75}$$

Therefore, to get a bound on the left side of (12.73), we get a bound on the absolute value of (12.74) and a bound on the absolute difference between $E[f'(W)]$ and (12.75). But, by Taylor's Theorem,

$$|X_i[f(W) - f(W_i) - (W - W_i) f'(W_i)]| \le |X_i| \frac{1}{2} (W - W_i)^2 C = \frac{C}{2} \cdot \frac{|X_i|^3}{n} \,.$$

Therefore, the absolute value of (12.74) is bounded above by

$$\frac{C}{2n^{3/2}} \sum_{i=1}^{n} E(|X_i|^3) \,. \tag{12.76}$$

Also, by Taylor's Theorem,

$$|f'(W_i) - f'(W)| \le |W_i - W|C = \frac{C|X_i|}{\sqrt{n}} \,,$$

and so the absolute difference between (12.75) and $E[f'(W)]$ is

$$|\frac{1}{n} E[f'(W_i) - f'(W)] \le \frac{C}{n^{3/2}} \sum_{i=1}^{n} E|X_i| \,. \tag{12.77}$$

Combining (12.76) and (12.77) yields

$$|E[Wf(W) - f'(W)]| \le \frac{C}{2n^{3/2}} \sum_{i=1}^{n} E(|X_i|^3) + \frac{C}{n^{3/2}} \sum_{i=1}^{n} E|X_i| \,. \tag{12.78}$$

But, $E|X_i| \le E(|X_i|^3)$ if $E(X_i^2) = 1$ (Problem 12.45), so that the right side of (12.78) is bounded by the right side of the result (12.73). ■

Theorem 12.5.1 *Under the assumptions of Lemma 12.5.4,*

$$d_W(W, Z) \le \frac{3}{n^{3/2}} \sum_{i=1}^{n} E[|X_i|^3] \,.$$

PROOF. The proof follows from Lemma 12.5.4 upon recalling from Lemma 12.5.3 that $\|f''\| \le C = 2$ for any f_h such that h is Lipschitz (with Lipschtiz constant one). ■

Next, we consider sums of certain classes of dependent random variables.

Definition 12.5.1 We say that $(X_1, \ldots, X_n)$ has dependency neighborhoods $N_i \subseteq \{1, \ldots, n\}$ if X_i is independent of $\{X_j\}_{j \notin N_i}$.

An immediate example is an m-dependent time series, where N_i is the set of indices in $\{1, \ldots, n\}$ such that $|j - i| \le m$.

The dependence structure of the X_i may be represented in terms of a graph with vertices $\{1, \ldots, n\}$, where i and j are connected with an edge if $j \in N_i$. If the cardinality of N_i, $|N_i|$, is not too big, then a normal approximation can be obtained. One such version is given in Ross (2011), which is stated next.

Theorem 12.5.2 *Suppose $(X_1, \ldots, X_n)$ has dependency neighborhoods N_i with $D = \max_i |N_i|$. Assume $E(X_i) = 0$, $E(X_i^4) < \infty$ and set*

$$\sigma^2 = Var\left(\sum_{i=1}^{n} X_i\right) = E\left(\sum_{i=1}^{n}\sum_{j\in N_i} X_i X_j\right) .$$

Let $W = \sum_{i=1}^{n} X_i/\sigma$. *Then,*

$$d_W(W, Z) \le \frac{D^2}{\sigma^3}\sum_{i=1}^{n} E(|X_i|^3) + \frac{\sqrt{28}D^{3/2}}{\sqrt{\pi}\sigma^2}\sqrt{\sum_{i=1}^{n} E(X_i^4)} . \tag{12.79}$$

PROOF. From (12.69) and the inequalities (12.72), it is enough to bound $|E[f'(W) - Wf(W)]|$ for f satisfying $\|f'\| \le \sqrt{2/\pi}$ and $\|f''\| \le 2$. Let $W_i = \sum_{j\notin N_i} X_j/\sigma$. Then, X_i and W_i are independent, and so $E[X_i f(W_i)] = 0$. By the triangle inequality,

$$|E[f'(W) - Wf(W)]| \le A + B ,$$

where

$$A = \left| E\left\{\frac{1}{\sigma}\sum_{i=1}^{n} X_i[f(W) - f(W_i) - (W - W_i)f'(W)]\right\}\right|$$

and

$$B = \left| E\left\{f'(W)[1 - \frac{1}{\sigma}\sum_{i=1}^{n} X_i(W - W_i)]\right\}\right| .$$

We claim that these two terms are bounded above by the corresponding two terms on the right side of (12.79). But by bringing the absolute value inside the expectation and applying Taylor's Theorem, it follows that

$$A \le \frac{1}{\sigma}\sum_{i=1}^{n} E\left|X_i\frac{(W - W_i)^2}{2}f''(W_i^*)\right| ,$$

where W_i^* is between W_i and W. Since $\|f''\| \le 2$,

$$A \le \frac{1}{\sigma}\sum_{i=1}^{n}|X_i(W - W_i)^2| = \frac{1}{\sigma^3}\sum_{i=1}^{n} E\left|X_i\left(\sum_{j\in N_i} X_j\right)^2\right|$$

$$\le \frac{1}{\sigma^3}\sum_{i=1}^{n}\sum_{j,k\in N_i} E|X_i X_j X_k|$$

$$\le \frac{1}{\sigma^3}\sum_{i=1}^{n}\sum_{j,k\in N_i}\frac{1}{3}[E(|X_i|^3) + E(|X_j|^3) + E(|X_k|^3)] ,$$

where the last inequality follows from the arithmetic-geometric mean inequality.[2] But,

$$\sum_{i=1}^{n} \sum_{j,k \in N_i} E(|X_i|^3) \leq D^2 \sum_{i=1}^{n} E(|X_i|^3)$$

and also

$$\sum_{i=1}^{n} \sum_{j \in N_i} \sum_{k \in N_i} E(|X_j|^3) = \sum_{j=1}^{n} \sum_{i \in N_j} \sum_{k \in N_i} E(|X_j|^3) \leq D^2 \sum_{j=1}^{n} E(|X_j|^3) \, .$$

Therefore,

$$A \leq \frac{D^2}{\sigma^3} \sum_{i=1}^{n} E(|X_i|^3) \, ,$$

as previously announced. Next,

$$B \leq \frac{\|f'\|}{\sigma^2} E \left| \sigma^2 - \sum_{i=1}^{n} \sum_{j \in N_i} X_i X_j \right| \, .$$

By the Cauchy–Schwarz Inequality, this is bounded above by

$$\sqrt{\frac{2}{\pi}} \frac{1}{\sigma^2} \sqrt{Var(\sum_{i=1}^{n} \sum_{j \in N_i} X_i X_j)} \, .$$

The remainder of the proof consists of bounding the variance term in the last expression; see Problem 12.47. ■

Example 12.5.1 (**U-statistics**) As in Section 12.3, consider the U-statistic, U_n, given by (12.24) based on a symmetric kernel h. Then, we can write $W = \binom{n}{b} U_n / \sigma$, where $\sigma^2 = \binom{n}{b}^2 Var(U_n)$. Theorem 12.5.2 applies if n is changed to $\binom{n}{b}$ and we identify an index i in the sum with a particular b-tuple $\{i_1, \ldots, i_b\}$ of b distinct indices. The number of terms in the sum for U_n, say $N_{\{i_1,\ldots,i_b\}}$, that share one of its b indices in common with $\{i_1, \ldots, i_b\}$, can be bounded above by $b\binom{n-1}{b-1}$, and so we can set D equal to this bound. Theorem 12.5.2 applies to yield a bound for d_W (Problem 12.48). ■

[2] The arithmetic-geometric mean inequality says that, for $y_i \geq 0$, $(y_1 + \cdots + y_k)/k \geq (y_1 \cdots y_k)^{1/k}$.

Example 12.5.2 (**Erdös-Rényi Random Graph**) Consider an Erdös-Rényi random graph, constructed as follows. Fix $0 < p < 1$. There are n vertices, and any pair of vertices is connected with an edge with probability p, independently for each pair of edges. For the sake of argument, assume the vertices are labeled from 1 to n. Let T be the number of triangles formed; that is, a particular triple of distinct vertices $\{i, j, k\}$ forms a triangle if all three pairs i and j, j and k, and i and k are connected with an edge. So, each of the $N = \binom{n}{3}$ possible triangles occurs with probability p^3. Let Y_i be the indicator of the event that the ith triple of N possible triangles is formed. Note that i indexes the possible N triangles that may be formed (in any specified order). So, $T = \sum_{i=1}^{N} Y_i$. Let $W = (T - ET)/\sigma$, where $\sigma^2 = Var(T)$. Let $N_i \setminus \{i\}$ be the triples of indices which share exactly two edges with those specified by the index i. Then, Theorem 12.5.2 applies to with $|N_i| = 3(n-3) + 1$, and so $D = 3n - 8$. In order to compute σ^2, let $X_i = Y_i - p^3$. Suppose $j \neq i$ and $j \in N_i$. Then,

$$Cov(Y_i, Y_j) = p^5 - p^6 .$$

To see why, if, for example, i corresponds to the vertices $\{1, 2, 3\}$ and j to the vertices $\{1, 2, 4\}$, then both triangles are formed if the 5 edges connecting 1 and 2, 2 and 3, 1 and 3, 1 and 4, and 2 and 4 are all present and so $E(Y_i Y_j) = p^5$. An easy calculation (Problem 12.51) then gives that, for any positive integer k,

$$E(|X_i|^k) = p^3(1-p^3)[(1-p^3)^{k-1} + p^{3(k-1)}] \tag{12.80}$$

and

$$\sigma^2 = \binom{n}{3} p^3[1 - p^3 + 3(n-3)p^2(1-p)] . \tag{12.81}$$

It follows that

$$d_W(W, Z) \leq \frac{(3n-8)^2}{\sigma^3}\binom{n}{3} p^3(1-p^3)[(1-p^3)^2 + p^6] \tag{12.82}$$

$$+\frac{\sqrt{28}(3n-8)^{3/2}}{\sqrt{\pi}\sigma^2}\sqrt{\binom{n}{3} p^3(1-p^3)[(1-p^3)^3 + p^9]} .$$

For fixed p, $\sigma^2 = O(n^4)$ and so the bound (12.82) tends to zero and a central limit theorem for T holds. One may even let $p \to 0$ and still derive a normal approximation to the distribution of T; see Problem 12.51. ■

The power of Stein's method goes significantly beyond the introduction presented here. For further information, see the notes at the end of the chapter.

12.6 Problems

Section 12.2

Problem 12.1 Show (12.2) and (12.3).

Problem 12.2 Show (12.6) and (12.7).

Problem 12.3 Use Theorem 12.2.1 to prove an asymptotic normal approximation to the hypergeometric distribution.

Problem 12.4 Show why Theorem 12.2.1 is a special case of Theorem 12.2.2.

Problem 12.5 Show that τ_N defined in the proof of Theorem 12.2.3 satisfies $\tau_N \to \infty$ as $\min(n, N-n) \to \infty$.

Problem 12.6 In the context of Example 12.2.1, find the limiting distribution of the W_n using Theorem 12.2.3. Identify G_n and G.

Problem 12.7 In Example 12.2.1, rather than considering the sum of the ranks of the Y_is, consider the statistic given by the sum of the squared ranks of the Y_is. Find its limiting distribution, properly normalized, under $F = G$.

Problem 12.8 In the setting of Section 12.2, assume $N = m + n$ and

$$(x_{N,1}, \ldots, x_{N,N}) = (y_1, \ldots, y_m, z_1, \ldots, z_n) \, .$$

Let $\bar{y}_m = \sum_{i=1}^m y_i/m$ and $\bar{z}_n = \sum_{j=1}^n z_j/n$. Let $\bar{x}_N = \sum_{j=1}^N x_{N,j}/N$. Also let $s^2_{m,y} = \sum_{i=1}^m (y_i - \bar{y}_m)^2/m$ and similarly define $s^2_{n,z}$. Let $Y_1, \ldots, Y_m$ denote a sample obtained without replacement from the N values, with sample mean $\bar{Y}_m$. Assume $m/n \to \lambda < \infty$. Assume $\bar{y}_m \to \bar{y}$ and $\bar{z}_n \to \bar{z}$, as well as $s_{m,y} \to s_y$ and $s_{n,z} \to s_z$. Finally assume the uniform distribution on $y_1, \ldots, y_m$ converges weakly to a c.d.f. G_y with variance s_y^2, and similarly the uniform distribution on $z_1, \ldots, z_n$ converges weakly to a c.d.f. G_z with variance s_z^2.
(i) Find the limiting distribution of $\sqrt{m}(\bar{Y}_m - \bar{x}_N)$.
(ii) If $Z_1, \ldots, Z_n$ denote the outcomes in Π_N not sampled by $Y_1, \ldots, Y_m$, and $\bar{Z}_n = \sum_{j=1}^n Z_j/n$, then find the limiting distribution of $\sqrt{m}(\bar{Y}_m - \bar{Z}_n)$.
(iii) Simplify your answers in the case $\bar{y}_m = \bar{z}_n$ and so $\bar{y} = \bar{z}$.

Problem 12.9 Complete the proof of Corollary 12.2.1(ii) using the Cramér-Wold Device.

Problem 12.10 In the setting of Corollary 12.2.1(ii), find an exact formula for $Cov(\bar{U}_n, \bar{V}_n)$ and then calculate the limit of $nCov(\bar{U}_n, \bar{V}_n)$.

Problem 12.11 The limiting expression for $NVar(\hat{\theta}_N)$ is given in (12.19). Find an exact expression for $NVar(\hat{\theta}_N)$ that has a similar representation.

Problem 12.12 Consider the estimator $\hat{s}_N^2(1)$ defined in (12.20). Show that $\hat{s}_N^2(1) \xrightarrow{P} s^2(1)$. State your assumptions.

Problem 12.13 Provide the details to show (12.22). *Hint:* Use Theorem 12.2.4 and Problem 12.12.

Problem 12.14 Prove an analogous result to Theorem 12.2.5 when sampling from an infinite population, where the asymptotic variance has the same form as (12.21) with $f = 0$. Assuming $s^2(1)$ and $s^2(0)$ are known, how would you allocate treatment among the N_0 units to minimize the asymptotic variance? (The solution is known as Neyman allocation.)

Problem 12.15 Prove a Glivenko–Cantelli Theorem (Theorem 11.4.2) for sampling without replacement from a finite population. Specifically, assume $X_1, \ldots, X_n$ are sampled at random without replacement from the population with $N = N_n$ elements given by $\{x_{N,1}, \ldots, x_{N,N}\}$. Let $\hat{F}_n(t) = n^{-1} \sum_{i=1}^n I\{X_i \le t\}$ and let $F_N(t) = N^{-1} \sum_{j=1}^N I\{x_{N,j} \le t\}$. Show that $\sup_t |\hat{F}_n(t) - F_N(t)| \xrightarrow{P} 0$. (First, consider the case where F_N converges in distribution to some F, but is this needed?)

Section 12.3

Problem 12.16 Verify (12.27).

Problem 12.17 Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. P, with $E(X_i^2) < \infty$ and $E(Y_i^2) < \infty$. The parameter of interest is $\theta(P) = Cov(X_i, Y_i)$. Find a kernel for which the corresponding U-statistic U_n is an unbiased estimator of $\theta(P)$. Under an appropriate moment assumption, find the limiting distribution of U_n. *Hint:* Compute $E[(X_1 - X_2)(Y_1 - Y_2)]$.

Problem 12.18 Verify (12.35).

Problem 12.19 In Example 12.3.6, show (12.37). Verify the limiting distribution of V_n in (12.38).

Problem 12.20 Show (12.42).

Problem 12.21 Show (12.43)

Problem 12.22 Show that (12.44) holds if $\hat{U}_n$ is replaced by U_n.

Problem 12.23 Verify (12.45).

Problem 12.24 Show that W_n in Example 12.2.1 and $U_{m,n}$ in Example 12.3.7 are related by $W_n = mnU_{m,n} + n(n+1)/2$, at least in the case of no ties in the data.

Problem 12.25 In Example 12.3.7, find F and G so that $\zeta_{0,1}$ and $\zeta_{1,0}$ are not $1/12$, even when $P\{X \le Y\} = 1/2$. Explore how large the rejection probability of the test with rejection region (12.46) can be under H_0'. What does this imply about a Type 3 or directional error? That is, if the test rejects H_0' and then declares $\theta(F, G) > 1/2$ because

$$\sqrt{m}(U_n - 1/2) \ge \sqrt{\frac{1+\lambda}{12}} z_{1-\frac{\alpha}{2}} ,$$

then how large can this probability be even if $P\{X \le Y\} < 1/2$?

Problem 12.26 Consider testing the null hypothesis that a sample $X_1, \ldots, X_n$ is i.i.d. against the alternative that the distributions of the X_i are stochastically increasing. Mann (1945) proposed the test which rejects for large values of N, where N is the number of pairs (X_i, X_j) with $i < j$ and $X_i < X_j$. Determine the limiting distribution of N, suitably normalized. How large should we choose the critical value (for large n) in order to control the Type 1 error at α?

Problem 12.27 Let $X_1, \ldots, X_n$ be i.i.d. P. Consider estimating $\theta(P)$ defined by

$$\theta(P) = E[h(X_1, \ldots, X_b)] ,$$

where h is a symmetric kernel. Assume P is such that $E|h(X_1, \ldots, X_b)| < \infty$, so that $\theta(P)$ is also well-defined. Show that $U_n \xrightarrow{P} \theta(P)$. In fact, $E|U_n - \theta(P)| \to 0$. *Hint: First how U_n is consistent by comparing U_n with a consistent estimator obtained by averaging the kernel over nonoverlapping subsets of the data of size b, and then apply Rao-Blackwell. Then, use Problem* 11.76.

Problem 12.28 Let $X_1, \ldots, X_n$ be i.i.d. P. Consider estimating $\theta(P)$ defined by

$$\theta(P) = E[h(X_1, \ldots, X_b)] ,$$

where h is a symmetric kernel. Assume P is such that $E[h^2(X_1, \ldots, X_b)] < \infty$, so that $\theta(P)$ is also well-defined. Let U_n be the corresponding U-statistic defined by (12.24). Let $\hat{P}_n$ be the empirical measure, and also consider the estimator

$$\theta(\hat{P}_n) = \frac{1}{n^b} \sum_{i_1=1}^{n} \cdots \sum_{i_b=1}^{n} h(X_{i_1}, \ldots, X_{i_b}) .$$

Do $\sqrt{n}[U_n - \theta(P)]$ and $\sqrt{n}[\theta(\hat{P}_n) - \theta(P)]$ converge to the same limiting distribution? If further conditions are needed, state them. Find the limiting behavior of $n[U_n - \theta(\hat{P}_n)]$. Again, state any conditions you might need.

Problem 12.29 Consider a U-statistic of degree 2, based on a kernel h. Let $h_1(x) = E[h(x, X_2)]$ and $\zeta_1 = Var[h_1(X_1)]$. Assume $\zeta_1 > 0$, so that we know that $\sqrt{n}[U_n - \theta(P)]$ converges in distribution to the normal distribution with mean 0 and variance

$4\zeta_1$. Consider estimating the limiting variance $4\zeta_1$. Since U_n averages $h(X_i, X_j)$ over the $\binom{n}{2}$ pairs (X_i, X_j) with $i \neq j$, one might use the sample variance of these $\binom{n}{2}$ pairs as an estimator. That is, define

$$S_n^2 = \frac{1}{\binom{n}{2}} \sum_{i<j} [h(X_i, X_j) - U_n]^2 .$$

Determine whether or not S_n^2 is a consistent estimator. State any added conditions you might need. Generalize to U-statistics of degree b.

Section 12.4

Problem 12.30 Verify (12.48).

Problem 12.31 In Example 12.4.2 with the ϵ_j having finite variance, derive the formulae for the mean and covariance (12.52) of the process.

Problem 12.32 Verify (12.54).

Problem 12.33 Assume X is stationary, $E(|X_1|^{2+\delta}) < \infty$ for some $\delta > 0$ and (12.56) holds. Show that (12.55) holds, and hence $R(k) \to 0$ as $k \to \infty$.

Problem 12.34 Suppose X is a stationary process with mean μ and covariance function $R(k)$. Assume $R(k) \to 0$ as $k \to \infty$. Show $\bar{X}_n \xrightarrow{P} \mu$. (A sufficient condition for $R(k) \to 0$ is X is strongly mixing with $E(|X_1|^{2+\delta}) < \infty$; see Problem 12.33.)

Problem 12.35 Generalize Theorem 12.4.1 to the case where the X_i are vector-valued.

Problem 12.36 Consider the setup of Example 12.4.3.
(i) Find the joint limiting distribution of $\sum_{i=1}^{n-1} (Y_{i,1}, Y_{i,0})^\top$, suitably normalized.
(ii) Let $\hat{R}_n = \sum_{i=1}^{n-1} Y_{i,1} / \sum_{i=1}^{n-1} Y_{i,0}$, which is the proportion of successes following a success. Show that $\sqrt{n}(\hat{R}_n - p) \xrightarrow{d} N(0, 1-p)$.

Problem 12.37 In Example 12.4.5, show the convergence (12.60).

Section 12.5

Problem 12.38 If $W \sim N(0, \sigma^2)$ with $\sigma \neq 1$, what is the generalization of the characterization (12.61)?

Problem 12.39 Complete the proof of the converse in Lemma 12.5.1. *Hint: Use Lemma 12.5.2.*

Problem 12.40 Complete the proof of Lemma 12.5.2 by showing that (12.64) and (12.65) are equivalent, and then showing that (12.66) follows.

Problem 12.41 Show that, for $w > 0$,

$$1 - \Phi(w) \leq \min\left(\frac{1}{2}, \frac{1}{w\sqrt{2\pi}}\right) e^{-w^2/2} .$$

Show that this inequality implies $\|f_x\| \leq \sqrt{\pi/2}$ and $\|f_x'\| \leq 2$.

Problem 12.42 Show that the Wasserstein metric implies weak convergence; that is, if $d_W(X_n, X) \to 0$, then $X_n \xrightarrow{d} X$. Give a counterexample to show the converse is false. Prove or disprove the following claim: For random variables X_n and X with finite first moments, show that $d_W(X_n, X) \to 0$ if and only if $X_n \xrightarrow{d} X$ and $E(X_n) \to E(X)$.

Problem 12.43 Investigate the relationships between d_W , d_K and d_{TV}, as well as the bounded Lipschitz metric introduced in Problem 11.24. Does convergence of one of them imply convergence of any of the others? If not, illustrate by finding counterexamples.

Problem 12.44 If Z is a real-valued random variable with density bounded by C, then show that, for any random variable W,

$$d_K(W, Z) \leq \sqrt{2C d_W(W, Z)} ,$$

where d_K is the Kolmogorov–Smirnov (or sup or uniform) metric between distribution functions, and d_W is the Wasserstein metric.

Problem 12.45 Show that, if $E(X^2) = 1$, then $E|X| \leq E(|X|^3)$.

Problem 12.46 Theorem 12.5.1 provides a bound for $d_W(W, Z)$ where $W = n^{-1/2} \sum_{i=1}^n X_i$ and the X_i are independent with mean 0 and variance one. Extend the result so that $Var(X_i) = \sigma_i^2$ may depend on i.

Problem 12.47 Finish the proof of Theorem 12.5.2 by showing

$$Var\left(\sum_{i=1}^n \sum_{j \in N_i} X_i X_j\right) \leq 14 D^3 \sum_{i=1}^n E(|X_i|^4) .$$

Hint: Use the arithmetic–geometric mean inequality.

Problem 12.48 Complete the details in Example 12.5.1 to get an explicit bound from Theorem 12.5.2 for d_W. What conditions are you assuming?

Problem 12.49 Use Theorem 12.5.2 to derive a Central Limit Theorem for the sample mean of an m-dependent stationary process. State your assumptions and compare with Theorem 12.4.1.

Problem 12.50 An alternative characterization of the Wasserstein metric is the following (which you do not have to show): $d_W(X, Y)$ is the infimum of $E|X' - Y'|$ over all possible joint distributions of (X', Y') such that the marginal distributions of X' and Y' are those of X and Y, respectively. However, do show that

$$d_W(X, Y) = \int_0^1 |F^{-1}(u) - G^{-1}(u)|du = \int_{-\infty}^{\infty} |F(x) - G(x)|dx \ ,$$

where F and G are the c.d.f.s of X and Y, respectively.

Problem 12.51 Verify (12.80), (12.81) and (12.82). Based on the bound (12.82), consider an asymptotic regime where $p \sim n^{-\beta}$ for some $\beta \geq 0$. For what β does the bound tend to zero, so that a central limit theorem for T holds?

Problem 12.52 Consider points on a lattice of the form (i, j) where i and j are integers from 0 to n. Each of these $(n + 1)^2$ points can be considered a vertex of a graph. Consider connecting edges adjoining (i, j) and $(i + 1, j)$ or (i, j) and $(i, j + 1)$, so that only edges between nearest vertices are considered in the graph (and each edge is either horizontal or vertical). Suppose each edge appears with probability p, independently of all other edges. For each little square of area one on the lattice, the square is colored red if all four edges appear. So for example, the region in the square with vertices $(0, 0)$, $(1, 0)$, $(0, 1)$ and $(1, 1)$ is colored red if all four edges appear, which has probability p^4, meaning the edge connecting $(0, 0)$ and $(1, 0)$, the edge connecting $(1, 0)$ and $(1, 1)$, the edge connecting $(0, 1)$ and $(1, 1)$, and the edge connecting $(0, 0)$ and $(0, 1)$. Let $\hat{\theta}_n$ be the proportion of the big square with area n^2 that is colored red. Find the limiting distribution of $\hat{\theta}_n$, suitably normalized.

12.7 Notes

Theorem 12.2.2 is due to Hájek (1960). For further discussion on the literature of CLTs for sampling from a finite population, see Li and Ding (2017). An interesting comparison of Fisher's sharp null versus Neyman's weak null is studied in Ding (2017). The potential outcomes framework has been used extensively in causal inference; Imbens and Rubin (2015), and the references therein.

The foundational paper on U-statistics is due to Hoeffding (1948). Further results on "projections" are due to Hájek (1968). Full length treatments on U-statistics can be found in Lee (1990) and Kowalski and Tu (2008).

The notion of α-mixing or strong mixing is due to Rosenblatt (1956). Various types of mixing are discussed in Doukhan (1995), Bradley (2007), and Dedecker, et al. (2007). Proofs of the mixing inequalities in Lemma 12.4.1 can be found in the appendix in Hall and Heyde (1980), though the results date back to Wolkonoski and Rozanov (1959) and Davydov (1979). A Central Limit Theorem under m-dependence appeared in Hoeffding and Robbins (1948) (assuming $2 + \delta$ moments), and under mixing in Ibragimov (1962). A quite general Central Limit Theorem under weak dependence, as well as references to others, is given in Neumann (2013). Theorem 12.4.1 is included as a special case of Neumann's result.

Section 12.5 on Stein's method was inspired by Stein (1972, 1986), Chen et al. (2011), and Ross (2011). See also the survey by Chatterjee (2014). The scope of application of Stein's method has been ever-expanding. Stein's method is actually a collection of tools, some of which are based on what Stein (1986) calls "auxiliary randomization", exchangeable pairs, and zero bias coupling. In particular, one can develop approximations in the Kolmogorov–Smirnov metric, and even in high dimensions. Moreover, Stein's method applies to distributional approximation beyond normality; see Ross (2011).