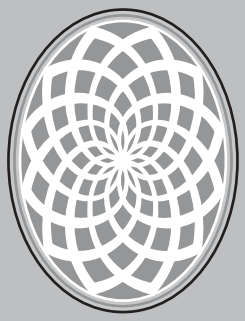

Industrial Mathematics Institute

2006:15

Some central limit theorems pertinent to estimating the effectiveness of information retrieval

E. Czabarka and J.L. Spouge

IMI

Preprint Series

Department of Mathematics
University of South Carolina

Some Central Limit Theorems Pertinent to Estimating the Effectiveness of Information Retrieval

Abbreviated Title: **Gaussian Limits and Information Retrieval**

Eva Czabarka
Department of Mathematics, University of South Carolina, Columbia, SC 29208
(email: czabarka@math.sc.edu)
John L. Spouge*
National Library of Medicine, National Institutes of Health, Bethesda MD 20894
(email: spouge@ncbi.nlm.nih.gov).
* Corresponding Author
Phone: (301) 435-5915
Fax: (301) 480-2484

Version Date: October 2, 2006

MSC 2000 Subject Classification: 60F05, 68P99, 92B99
Key words and phrases: Information Retrieval, ROC score, Central Limit Theorem
Acknowledgement of Support: Support for this work was provided by the intramural research program of the National Institutes of Health.

SUMMARY

An urn contains white and black balls numbered arbitrarily from 1 to d. The tendency of white balls to have lower numbers than black balls can be measured by a "ROC_n score". Now, let each ball have not one but two numbers on it, in red and green. Sample the balls with replacement from the urn into a "bag", and consider the difference between the ROC_ns in the bag for the two sets of numbers, red and green. The ROC_n difference has an approximate normal distribution, with mean equaling the difference in the urn. Now, bootstrap by sampling balls with replacement from the bag into a "sack", and consider the ROC_n difference in the sack. Again, the difference has an approximate normal distribution, with mean equaling the difference in the bag. Moreover, the difference has approximately the same variance in the bag and sack, the condition required to justify bootstrap inferences about sampling from the urn. The results have practical relevance, because researchers use the ROC_n score to measure the efficacy of database retrieval. They then bootstrap the database to assign a P-value to the ROC_n difference between two retrieval algorithms.

1. INTRODUCTION

Entire computing centers like the National Center for Biotechnology Information, the European Molecular Biology Laboratory, and the DNA Data Bank of Japan are presently devoted to the collection, storage, retrieval, and analysis of biological information. Much effort in bioinformatics is accordingly preoccupied with constructing algorithms for information retrieval.

In most retrieval applications, several competing algorithms are available. To compare the algorithms, the amount of information they retrieve from a database must be measured. The ROC_n score [10] (pronounced "receiver operating characteristic truncated at n", defined below, and closely related to the ROC score [8, 20, 21]) has become a popular measure of information retrieval in bioinformatics [1, 9, 11-14, 18, 19, 23]. Of two competing algorithms, the one with the higher ROC_n might be deemed superior, at least at first glance.

The assembly of a database involves chance events, however (e.g., grant acquisition), so chance alone could cause one ROC_n to be higher than another. Moreover, compared to its alternatives, the "superior" algorithm might incur extra costs in programming effort, running time, or code maintenance. The algorithm might not be preferable, then, unless statistical tests confirm that the higher ROC_n is not due to chance alone.

It has been suggested [22] that the bootstrap [6] can assess chance effects on a measure of information retrieval. Accordingly, many workers in bioinformatics routinely bootstrap their databases [9, 18]. Bootstrap inferences are sometimes wrong [3, 4], however, and their implications for information retrieval have not been investigated.

With this motivation, consider an urn containing d balls, with f of the balls black and the rest white. Every ball, white or black, has a distinct number in red paint upon it, from 1 to d. The balls' numbers have no methodical association with their color. The numbers implicitly order the black balls, giving each an "index" $i = 1,..., f$ (so the corresponding numbers $b(1) < ... < b(f)$ on the black balls are between 1 and d). Below, we refer to the black ball with index i as "the i-th black ball".

Sample D balls from the urn independently with replacement, and assemble copies of the sampled balls into a bag, each copy retaining color and number. If the probability of sampling the ball with red number j is p_j ($\sum_{j=1}^{d} p_j = 1$), the sampled balls in the bag follow a Multinomial$(D; p_1,..., p_d)$ distribution. Let the variate $\widehat{T}$ count the white balls in the bag; $\widehat{F}_i$, the copies of the i-th black ball; and $\widehat{T}_i$, the white balls in the bag with a lower number than the i-th black ball. Denote the corresponding expectations by $t := \mathbb{E}\widehat{T}$, $f_i := \mathbb{E}\widehat{F}_i$, and $t_i := \mathbb{E}\widehat{T}_i$. (The symbol ":=" denotes a definition.)

(Notation posed considerable difficulties for this article. The database application suggested as mnemonics "f" for "false" (black); "t" for "true" (white); and "d" for "document" (either white or black). "Urn variables" are usually lower case (e.g., t); corresponding "bag variates", upper case (e.g., $\widehat{T}$). The curved over-strokes on the bag variates differentiate them from the Poisson variates that supplant them throughout the proofs.)

Fix any deterministic cut-off c, so the notation below can suppress dependence on c where desirable. Define the index

$$n := n(c) := \min\left\{ j : c \le \sum_{i=1}^{j} f_i \right\}, \tag{1.1}$$

with default $n := f$, the total number of black balls in the urn, if the set in Eq (1.1) is empty. In the urn, let

$$r_m := r_{c;m} := \sum_{i=1}^{m-1} t_i f_i + t_m \left(c - \sum_{i=1}^{m-1} f_i \right). \tag{1.2}$$

Define the "non-normalized $\mathrm{ROC_n}$" $r_n := r_{c;n(c)}$ and the corresponding $\mathrm{ROC_n}$ score (also called simply "the $\mathrm{ROC_n}$") $roc_n := roc_{c;n(c)} := r_n / (ct)$. Intuitively, the $\mathrm{ROC_n}$ score measures whether white balls have lower numbers than black balls. For example, $roc_n = 0$ if the black balls are numbered $1,..., f$; and $roc_n = 1$ if they are numbered $d - f + 1,..., d$.

In the bag, define

$$\hat{N} := \hat{N}(c) := \min\left\{ j : c \le \sum\nolimits_{i=1}^{j} \hat{F}_i \right\}, \tag{1.3}$$

with default $\hat{N} := f$. (Because $f_i := \mathbb{E}\hat{F}_i$ for all i, under mild conditions $\hat{N} n^{-1} \approx 1$ for n large.) Let

$$\hat{R}_m := \hat{R}_{c;m} := \sum_{i=1}^{m-1} \hat{T}_i \hat{F}_i + \hat{T}_m \left(c - \sum_{i=1}^{m-1} \hat{F}_i \right). \tag{1.4}$$

For the bag, define the non-normalized $\mathrm{ROC_n}$ $\hat{R}_{\hat{N}} := \hat{R}_{c;\hat{N}(c)}$ and the $\mathrm{ROC_n}$ score $\widehat{ROC}_{\hat{N}} := \widehat{ROC}_{c;\hat{N}(c)} := \hat{R}_{\hat{N}} / (c\hat{T})$.

(The equation $\widehat{ROC}_{\hat{N}} := \hat{R}_{\hat{N}} / (c\hat{T})$ is the standard formula for the ROC_n [10], although the sampling context disguises it somewhat. To produce the standard formula, take c to be

a natural number, and let $\widehat{B}(1) < \widehat{B}(2) < ...$ be the indices of the black balls in the bag, so $\widehat{F}_i \geq 1$ for $i = \widehat{B}(1), \widehat{B}(2), ...$ and $\widehat{F}_i = 0$ otherwise. Under the hypotheses of the theorems below, it is overwhelmingly probable that $\widehat{F}_i = 1$ for $i = \widehat{B}(1), ..., \widehat{B}(c)$, yielding the standard formula $\widehat{ROC}_{\hat{N}} := \sum_{i=1}^{c} \widehat{T}_{\widehat{B}(i)} / \left(c\widehat{T} \right)$.)

All urn variables in this article are defined explicitly, but some definitions of bag variates remain implicit, for brevity. Any implicit definition can be recovered as follows: (1) take the definition of the corresponding urn variable; and (2) replace the "fundamental urn variables", $\{ f_i \}$ and $\{t_i\}$, with the "fundamental bag variates", $\left\{\widehat{F}_i\right\}$ and $\left\{\widehat{T}_i\right\}$. Eqs (1.1)-(1.4) exemplify the substitution.

Our interest lies in comparing different ROC_n s, so in addition to its red number, every ball has a distinct number in green paint upon it, also from 1 to d. The red and green numbers have no methodical association with each other or with ball color. In the following, quantities pertinent to green numbers are primed. Every unprimed equation has (possibly tacitly) a primed counterpart, where the primed quantities reverse the two colors, red and green, in the relevant definitions. Like the red numbers, the green numbers impose an order on the black balls, indicated by a "green index" $i = 1, ..., f$ (so the corresponding green numbers $b'(1) < ... < b'(f)$ on the black balls again are between 1 and d). For the green numbers in the bag, $\widehat{N}' := \min\left\{ j : c \leq \sum_{i=1}^{j} \widehat{F}_i' \right\}$, e.g., so the corresponding ROC_n s are $\widehat{R}'_{\hat{N}'} = \sum_{i=1}^{\hat{N}'-1} \widehat{T}_i' \widehat{F}_i' + \widehat{T}'_{\hat{N}'} \left(c - \sum_{i=1}^{\hat{N}'-1} \widehat{F}_i' \right)$ and $\widehat{ROC}'_{\hat{N}'} := \widehat{R}'_{\hat{N}'} / \left(c\widehat{T}' \right)$.

In this notation, our interest lies in the event $\left[\widehat{ROC'}_{\hat{N}'} \geq \widehat{ROC}_{\hat{N}} \right]$. The red and green numbers do not affect the total number of white balls, so $\hat{T} = \hat{T}'$, yielding

$$\left[\widehat{ROC'}_{\hat{N}'} \geq \widehat{ROC}_{\hat{N}} \right] = \left[\hat{R}'_{\hat{N}'} \geq \hat{R}_{\hat{N}} \right]. \tag{1.5}$$

The simpler quantities $\hat{R}_{\hat{N}}$ and $\hat{R}'_{\hat{N}'}$ therefore appear in our theorems.

Our article is organized as follows. In the Statement of Results in Section 2, Theorem 2.1 makes the trite observation that for balls within a subset having a small total sampling probability, a multivariate-Poisson distribution approximates multinomial sampling from the urn. Theorem 2.1 permits a deep analysis of the sampling behavior of the ROC_n, however. Theorem 2.2 and Theorem 2.3 provide central limit theorems (CLTs) for the bag variates $\hat{R}_{\hat{N}}$ and $\hat{R}'_{\hat{N}'} - \hat{R}_{\hat{N}}$. Theorem 2.4 and Theorem 2.5 show that bootstrapping the bag can mimic the Gaussian distributions of the bag variates $\hat{R}_{\hat{N}}$ and $\hat{R}'_{\hat{N}'} - \hat{R}_{\hat{N}}$. The end of Section 2 indicates the theorems' practical implications. Finally, Sections 3, 4, and 5 prove Theorem 2.1, Theorem 2.2, and Theorem 2.3, respectively; Section 6 proves Theorem 2.4 and Theorem 2.5.

2. STATEMENT OF RESULTS

Theorem 2.1 approximates multinomial sampling with independent Poisson variates. (If only one set of numbers is under scrutiny, no mention is made of color, red or green.)

Theorem 2.1: *Sample D balls with replacement from an urn, as in the Introduction. Consider any fixed subset of m balls in the urn, with numbers $\mathfrak{M} := \{k(1),...,k(m)\}$. The ball with number $k(i)$ has sampling probability $p_{k(i)}$. It is sampled $\widehat{D}_{k(i)}$ times, so the expected number of its copies in the bag is $\mathbb{E}\widehat{D}_{k(i)} = Dp_{k(i)}$. Then, there exist m independent Poisson variates $\left(D_{k(1)},...,D_{k(m)}\right)$, with $D_{k(i)} \sim \text{Poisson}\left(Dp_{k(i)}\right)$, such that* $\mathbb{P}\left\{\left(D_{k(1)},...,D_{k(m)}\right) \neq \left(\widehat{D}_{k(1)},...,\widehat{D}_{k(m)}\right)\right\} \leq p_{\mathfrak{M}} := \sum_{i=1}^{m} p_{k(i)}$.

In this article, "Poisson (re)sampling" refers to using the Poisson variates constructed in Theorem 2.1 to approximate multinomial sampling or bootstrap resampling. Because the fundamental bag variates ($\widehat{F}_i$, $\widehat{T}_i$, $\widehat{F}_i'$, and $\widehat{T}_i'$) count specific random sets of balls, Theorem 2.1 yields "fundamental Poisson variates" (F_i, T_i, F_i', and T_i') to approximate them. "Secondary Poisson variates" (e.g., N and R_N) can then be defined from F_i, T_i, F_i', and T_i' by dropping curved over-strokes in the formulas for the corresponding bag variates, e.g., Eq (1.3) yields $N := \min\left\{j : c \leq \sum_{i=1}^{j} F_i\right\}$, and Eq (1.4) yields $R_m := \sum_{i=1}^{m-1} T_i F_i + T_m\left(c - \sum_{i=1}^{m-1} F_i\right)$ and $R_N := \sum_{i=1}^{N-1} T_i F_i + T_N\left(c - \sum_{i=1}^{N-1} F_i\right)$, etc.

We now prepare to state the CLTs. Consider an infinite sequence of urns $k = 1, 2, ...$. Let notation and structure for every urn in the sequence follow the Introduction (so the

notation usually suppresses k). Each variable becomes part of a sequence, with its dependence on the corresponding urn left implicit, e.g., $f_i := f_i^{(k)}$, $t_i := t_i^{(k)}$, $c := c^{(k)}$, $n := n^{(k)}$, $r_n := r_n^{(k)}$, etc. The explicit limit $k \to \infty$ appears only occasionally below.

All asymptotic statements refer to the limit $k \to \infty$, unless stated otherwise. Let "$X_k \Rightarrow \Phi$" denote the convergence of $\{X_k\}$ to a Normal$(0,1)$ distribution; and "$\to_{\mathbb{P}}$", convergence in probability. We use the asymptotic notations $\sim$, O, and o (e.g., [7, p.5]), and the probabilistic counterparts $\sim_{\mathbb{P}}$, $\bar{O}_{\mathbb{P}}$, and $o_{\mathbb{P}}$. (For $Y_k \geq 0$, let $Z_k := X_k Y_k^{-1}$ if $Y_k > 0$; $Z_k := 0$ if $X_k = Y_k = 0$; and $Z_k := \infty \cdot \operatorname{sgn} X_k$ if $X_k \neq 0$ and $Y_k = 0$. Then $X_k \sim_{\mathbb{P}} Y_k \Leftrightarrow Z_k \to_{\mathbb{P}} 1$; $X_k = \bar{O}_{\mathbb{P}}(Y_k) \Leftrightarrow \limsup_{m\to\infty} \limsup_{k\to\infty} \mathbb{P}(|Z_k| > m) = 0$; and $X_k = o_{\mathbb{P}}(Y_k) \Leftrightarrow Z_k \to_{\mathbb{P}} 0$). Although $\bar{O}_{\mathbb{P}}$ (compactness for Z_k in the topology of distributional convergence) is not a standard notation, it is appropriate for our purposes here.

By definition, the CLTs below involve convergence in distribution. In contrast, the probabilistic notations $\sim_{\mathbb{P}}$, $\bar{O}_{\mathbb{P}}$, and $o_{\mathbb{P}}$ involve convergence in probability. Probabilistic notations compress the proofs, however, so to accommodate them, take the random samples from each urn ($k = 1, 2, \ldots$) to be mutually independent, so a single (product) probability space $(\Omega, \mathcal{F}, \mathbb{P})$ simultaneously supports all the samples from all the urns.

(Although mutual independence is an unnecessary restriction, it is irrelevant in practice. To apply the CLTs below to any particular urn with a large number of balls, the urn must be embedded in a hypothetical infinite sequence of urns. The sequence needs to satisfy the hypotheses of the CLTs, but other details of its construction are irrelevant to the conclusions. Mutual independence can therefore be regarded as one such irrelevant detail.)

To extend the notation, let $g_0 := 0$, $g_i := \sum_{j=1}^{i} f_j$, $t_0 := 0$, and $u_i := t_i - t_{i-1}$ (so $t_i := \sum_{j=1}^{i} u_j$). With n replacing $\hat{N}$, $\hat{R}_{\hat{N}}$ becomes $\hat{R}_n := \sum_{i=1}^{n-1} \hat{T}_i \hat{F}_i + \hat{T}_n \left(c - \sum_{i=1}^{n-1} \hat{F}_i \right)$. Define

$$\sigma_n^2 := \sum_{i=1}^{n} \left(c - g_{i-1} \right)^2 u_i + \sum_{i=1}^{n} \left(t_n - t_i \right) \left(t_n - t_i + 1 \right) f_i \,. \tag{2.1}$$

The proof of Theorem 2.2 shows that $r_n = \mathbb{E} R_n$ and $\sigma_n^2 = \operatorname{var} R_n$ (where R_n is the Poisson variate corresponding to $\hat{R}_n$). The approximation $\hat{R}_{\hat{N}} \approx R_n$ motivates the CLTs below.

Define $\hat{T}_0 := 0$. The urn variables g_i, u_i, and σ_n^2 correspond to the bag variates $\hat{G}_i := \sum_{j=1}^{i} \hat{F}_j$, $\hat{U}_i := \hat{T}_i - \hat{T}_{i-1}$, and

$$\hat{\Sigma}_{\hat{N}}^2 := \sum_{i=1}^{\hat{N}} \left(c - \hat{G}_{i-1} \right)^2 \hat{U}_i + \sum_{i=1}^{\hat{N}} \left(\hat{T}_n - \hat{T}_i \right) \left(\hat{T}_n - \hat{T}_i + 1 \right) \hat{F}_i \,, \tag{2.2}$$

Assume that c, t_n, g_n, and σ_n tend to ∞ as $k \to \infty$. Two variants of the following condition appear in the theorems, with different values of θ.

Condition O-θ: *Let a sequence of urns satisfy* $c = o\left(\sigma_n\right)$ *and* $t_n = o\left(\sigma_n\right)$. *Assume for each* k, *there exist integers* $\underline{j}$ *and* $\overline{j}$ *with the following properties. First,* $\overline{j} \to \infty$, *with* $\overline{f} := \max_{i=1,\ldots,\overline{j}} f_i$ *and* $\overline{u} := \max_{i=1,\ldots,\overline{j}} u_i$ *(which might not be integral) satisfying* $\overline{f} \log\log \overline{j} \to \infty$, $\overline{u} \log\log \overline{j} \to \infty$, $t_n^2 \theta \overline{f} = o\left(\sigma_n^2\right)$, $\left(\log \overline{j}\right) \theta \overline{f} = O\left(c^{1/2}\right)$, *and* $c^2 \theta\, \overline{u} = o\left(\sigma_n^2\right)$. *Second,* $g_{\underline{j}}^{-1/2} \left(c - g_{\underline{j}} \right) \to \infty$, $g_{\overline{j}}^{-1/2} \left(g_{\overline{j}} - c \right) \to \infty$, $t_{\overline{j}} + g_{\overline{j}} = o\left(D\right)$, *and* $c \left(t_{\overline{j}} - t_{\underline{j}} \right)^2 = o\left(\sigma_n^2\right)$.

Theorem 2.2 below requires in its hypothesis "Condition O-1", which is Condition O-θ with $\theta \equiv 1$ identically for all k. Later, Theorem 2.4 later requires a slightly stronger condition, "Condition O-$\log \overline{j}$", which is Condition O-θ with $\theta = \log \overline{j}$. The end of the section discusses the feasibility of the conditions.

Theorem 2.2: *If Condition O-1 holds,* $\hat{\Sigma}_{\tilde{N}} \sim_{\mathbb{P}} \sigma_n$ *and* $\left(\hat{R}_{\tilde{N}} - r_n\right)/\sigma_n \Rightarrow \Phi$ *as* $k \to \infty$.

Remark: The proof of Theorem 2.2 shows $\lim_{k\to\infty} \mathbb{P}\left(R_N \neq \hat{R}_{\tilde{N}}\right) = \lim_{k\to\infty} \mathbb{P}\left(\Sigma_N \neq \hat{\Sigma}_{\tilde{N}}\right) = 0$. Thus, $\Sigma_N \sim_{\mathbb{P}} \sigma_n$ and $\left(R_N - r_n\right)/\sigma_n \Rightarrow \Phi$, as might be expected, because $r_n = \mathbb{E}R_n$ and $\sigma_n^2 = \operatorname{var} R_n$.

The first conclusion in Theorem 2.2 ($\hat{\Sigma}_{\tilde{N}} \sim_{\mathbb{P}} \sigma_n$) resembles the second moment hypotheses for confirming bootstrap inferences [3]; the other asserts a Gaussian limit for $\hat{R}_{\tilde{N}}$.

We now state a CLT for the ROC_n difference $\hat{R}'_{\tilde{N}'} - \hat{R}_{\tilde{N}}$. First, we count the balls corresponding to the intersection of certain sets of red and green numbers. In the standard notation, define the set $[i] := \{1, 2, ..., i\}$. Define the permutation $\pi : [f] \mapsto [f]$ where $\pi(i) = j$, if and only if a black ball has red index i and green index j. For each pair of integers $m, m' > 0$, define the set $\mathfrak{C}_{mm'} := \left\{i : 1 \le i \le m \text{ and } 1 \le \pi(i) \le m'\right\}$ of red indices and the set $\mathfrak{C}'_{m'm} := \left\{\pi(i) : 1 \le i \le m \text{ and } 1 \le \pi(i) \le m'\right\}$ of green indices. (Usually, $\mathfrak{C}_{mm'} \neq \mathfrak{C}'_{m'm}$. The definitions of $\mathfrak{C}_{mm'}$ and $\mathfrak{C}'_{m'm}$ involve the same black balls, but the balls usually have different sets of red and green indices.) For $m = n = n(c)$ and $m' = n' = n'(c)$,

the set $\mathfrak{C}_{nn'}$ contains the red indices of black balls contributing to both $\widehat{R}_n$ and $\widehat{R}'_{n'}$; $\mathfrak{C}'_{n'n}$ contains the corresponding green indices. For $m = \widehat{N} = \widehat{N}(c)$ and $m' = \widehat{N}' = \widehat{N}'(c)$, the random set $\mathfrak{C}_{\widehat{N}\widehat{N}'}$ contains the red indices of black balls contributing to both $\widehat{R}_{\widehat{N}}$ and $\widehat{R}'_{\widehat{N}'}$; $\mathfrak{C}'_{\widehat{N}'\widehat{N}}$ contains the corresponding green indices.

To ease the language, the statement "ball A comes before (after) ball B" provides a shorthand for "the number on ball A is less (greater) than or equal to the number on ball B". Let the variate $\widehat{V}_{ij}$ count the white balls that in red numbering come after the $i-1$-st black ball but before the i-th black ball, and in green numbering come after the $j-1$-st black ball but before the j-th black ball. Define $v_{ij} \coloneqq \mathbb{E}\widehat{V}_{ij}$ and

$$\gamma_{nn'} \coloneqq \sum_{i=1}^{n}\sum_{j=1}^{n'}\left(c-g_{i-1}\right)\left(c-g'_{j-1}\right)v_{ij} + \sum_{i\in\mathfrak{C}_{nn'}}\sum_{j=i+1}^{n}\sum_{m=\pi(i)+1}^{n'} v_{jm}f_i + \sum_{i\in\mathfrak{C}_{nn'}}\left(t_n-t_i\right)\left(t'_{n'}-t'_{\pi(i)}\right)f_i . \tag{2.3}$$

Define also $\sigma^2_{nn'} \coloneqq \sigma^2_n + \sigma'^2_{n'} - 2\gamma_{nn'}$, $\rho_{nn'} \coloneqq \gamma_{nn'}/\left(\sigma_n\sigma'_{n'}\right)$, and $\rho \coloneqq \limsup_{k\to\infty}\rho_{nn'}$. If R_n and $R'_{n'}$ are the Poisson variates corresponding to $\widehat{R}_n$, and $\widehat{R}'_{n'}$, the proof of Theorem 2.3 shows that $\operatorname{cov}\left(R_n, R'_{n'}\right) = \gamma_{nn'}$, so that $\operatorname{var}\left(R'_{n'} - R_n\right) = \sigma^2_{nn'}$ and the correlation coefficient of R_n and $R'_{n'}$ is $\rho_{nn'}$.

Theorem 2.3 follows the pattern of Theorem 2.2 and asserts a Gaussian limit for $\widehat{R}'_{\widehat{N}'} - \widehat{R}_{\widehat{N}}$. Let the bag variates $\widehat{\Sigma}_{\widehat{N}\widehat{N}'}$, $\widehat{\Gamma}_{\widehat{N}\widehat{N}'}$, and $\widehat{\mathrm{P}}_{\widehat{N}\widehat{N}'}$ correspond to the urn variables $\sigma_{nn'}$, $\gamma_{nn'}$, and $\rho_{nn'}$. For example,

$$\widehat{\Gamma}_{\tilde{N}\tilde{N}'} := \sum_{i=1}^{\tilde{N}} \sum_{j=1}^{\tilde{N}'} \left(c - \widehat{G}_{i-1}\right)\left(c - \widehat{G}'_{j-1}\right)\widehat{V}_{ij} + \sum_{i \in \mathfrak{C}_{\tilde{N}\tilde{N}'}} \sum_{j=i+1}^{\tilde{N}} \sum_{m=\pi(i)+1}^{\tilde{N}'} \widehat{V}_{jm}\widehat{F}_i + \sum_{i \in \mathfrak{C}_{\tilde{N}\tilde{N}'}} \left(\widehat{T}_{\tilde{N}} - \widehat{T}_i\right)\left(\widehat{T}'_{\tilde{N}'} - \widehat{T}'_{\pi(i)}\right)\widehat{F}_i . \quad (2.4)$$

As usual, "a.s." abbreviates "almost surely". Loosely, the hypothesis $\rho < 1$ in Theorem 2.3 below ensures that $\widehat{R}_{\tilde{N}}$ and $\widehat{R}'_{\tilde{N}'}$ do not correlate perfectly in the limit $k \to \infty$.

Theorem 2.3: *In a sequence of urns, assume the red and green numbers both satisfy the hypotheses of Theorem 2.2. In addition, assume that* $\rho < 1$. *Then,* $\widehat{\Sigma}_{\tilde{N}\tilde{N}'} \sim_{\mathbb{P}} \sigma_{nn'}$ *and* $\left\{\left(\widehat{R}'_{\tilde{N}'} - \widehat{R}_{\tilde{N}}\right) - \left(r'_{n'} - r_n\right)\right\} / \sigma_{nn'} \Rightarrow \Phi$ *as* $k \to \infty$. *In addition, there is a subsequence of* $k = 1, 2, \ldots$, *which can be chosen deterministically, so that* $\limsup \widehat{\mathrm{P}}_{\tilde{N}\tilde{N}'} = \rho$ *a.s.-*$\mathbb{P}$ *as* $k \to \infty$ *along the subsequence.*

Remark: The proof of Theorem 2.3 shows $\lim_{k\to\infty} \mathbb{P}\left(\Sigma_{NN'} \neq \widehat{\Sigma}_{\tilde{N}\tilde{N}'}\right) = 0$. As in the remark after Theorem 2.2, $\Sigma_{NN'} \sim_{\mathbb{P}} \sigma_{nn'}$ and $\left\{\left(R'_{N'} - R_N\right) - \left(r'_{n'} - r_n\right)\right\} / \sigma_{nn'} \Rightarrow \Phi$, where $\mathrm{var}\left(R'_{n'} - R_n\right) = \sigma^2_{nn'}$.

Next, Theorem 2.4 and Theorem 2.5 show that bootstrapping the bags mimics the Gaussian approximation for $\widehat{R}_{\tilde{N}}$ and $\widehat{R}'_{\tilde{N}'} - \widehat{R}_{\tilde{N}}$. To describe the bootstrap, consider the bag of $D = D^{(k)}$ balls sampled from urn k. Resample D balls from the bag uniformly, independently, and with replacement, to assemble copies of the resampled balls into a "sack", each copy retaining color and (red and green) numbers as above. The resampled balls in the sack therefore follow a Multinomial$\left(D; p^*_1, \ldots, p^*_D\right)$ distribution ($p^*_1 = \ldots = p^*_D = D^{-1}$). As usual, a superscript star denotes a bootstrap quantity. In the sack

after resampling has occurred, and for the red numbering, let the variate $\widehat{F}_i^*$ count the copies of the i-th black ball; $\widehat{T}_i^*$, the white balls before the i-th black ball. The analogs of Eqs (1.1)-(1.4) yield bootstrap variates $\widehat{N}^*$ and $\widehat{R}_{\tilde{N}^*}^*$ corresponding to $\widehat{N}$ and $\widehat{R}_{\tilde{N}}$; analogously, the green numbers yield bootstrap variates $\widehat{N}'^*$ and $\widehat{R}_{\tilde{N}'^*}'^*$, etc., and from all these variates, other secondary bootstrap variates like $\widehat{\Sigma}^*_{\tilde{N}^*\tilde{N}'^*}$ and $\widehat{\mathrm{P}}^*_{\tilde{N}^*\tilde{N}'^*}$ can be constructed.

Again to accommodate probabilistic notation, assume that the random samples from each pair of containers (urn and bag) are mutually independent ($k=1,2,\ldots$). Construct a single probability space $\left(\Omega^*,\mathscr{F}^*,\mathbb{P}^*\right)$ simultaneously supporting all samples: $\Rightarrow^*$ connotes the corresponding convergence in distribution ($k=1,2,\ldots$); and so forth for $\sim_{\mathbb{P}^*}$, etc.

Theorem 2.4 describes a typical bootstrap convergence. ([3] explain the mode of convergence in greater detail.)

Theorem 2.4: *Assume Condition O-*$\log \overline{j}$. *Then, there is a subsequence of* $k=1,2,\ldots$, *which can be chosen deterministically, so that a.s.-*$\mathbb{P}^*$, *both* $\widehat{\Sigma}^*_{\tilde{N}^*} \sim_{\mathbb{P}^*} \sigma_n$ *and* $\left(\widehat{R}^*_{\tilde{N}^*} - \widehat{R}_{\tilde{N}}\right)/\widehat{\Sigma}_{\tilde{N}} \Rightarrow^* \Phi$ *as* $k\to\infty$ *along this subsequence.*

Similarly, Theorem 2.3 has a bootstrap analog.

Theorem 2.5: *In a sequence of urns, assume the red and green numbers both satisfy the hypotheses of Theorem 2.4. In addition, assume* $\rho<1$. *Then, there is a subsequence of* $k=1,2,\ldots$, *which can be chosen deterministically, so that a.s.-*$\mathbb{P}^*$, $\limsup \widehat{\mathrm{P}}^*_{\tilde{N}^*\tilde{N}'^*} = \rho$,

$\widehat{\Sigma}^*_{\tilde{N}\tilde{N}'} \sim_{\mathbb{P}^*} \sigma_{nn'}$, *and* $\left\{\left(\widehat{R}'^*_{\tilde{N}'^*} - \widehat{R}^*_{\tilde{N}^*}\right) - \left(\widehat{R}'_{\tilde{N}'} - \widehat{R}_{\tilde{N}}\right)\right\} / \widehat{\Sigma}_{\tilde{N}\tilde{N}'} \Rightarrow^* \Phi$ *as* $k \to \infty$ *along this subsequence.*

Remark: As the remarks after Theorem 2.2 and Theorem 2.3 suggest, Theorem 2.4 and Theorem 2.5 remain true for Poisson variates as well, if all curved over-strokes in them are dropped. This remark motivates a "Poisson bootstrap" to approximate the distribution of $\widehat{R}^*_{\tilde{N}}$ on a computer.

A (sequential) Poisson bootstrap for $\widehat{R}^*_{\tilde{N}}$: Apply Theorem 2.1 in the context of resampling from a bag into a sack, rather than sampling from an urn into a bag. Accordingly, let the D balls sampled into the bag have red numbers $\widehat{K}(1) \le ... \le \widehat{K}(D)$. To identify specific balls easily, paint them in order (breaking ties arbitrarily) with blue numbers from 1 to D. After bootstrapping from the bag, let the variate $\widehat{D}^*_i$ count the copies of the ball with blue number i resampled into the sack ($\mathbb{E}\widehat{D}^*_i = 1$). Because Theorem 2.1 approximates $\left\{\widehat{D}^*_i\right\}$ with independent Poisson(1) variates $\left\{D^*_i\right\}$, a computer can simulate an approximate bootstrap distribution for $\widehat{R}^*_{\tilde{N}}$, as follows. Independently in the order $i = 1,...,D$, resample the ball with blue number i from the bag into the sack with a Poisson(1) distribution, but stop when the sack contains at least c black balls (because R^*_N is then determined).

Remark: Two points are noteworthy. First, Theorem 2.1 bounds the error due to the Poisson bootstrap. Second, a sequential Poisson bootstrap can be given for variates other than $\widehat{R}^*_{\tilde{N}}$, if (with high probability) they are determined by resampling a small fraction of

the bag's balls. (For example, the distribution of $\left(\widehat{R}^*_{\hat{N}}, \widehat{R}'^*_{\hat{N}'}\right)$ can be approximated by modifying the stopping rule above to include both red and green numbers.)

We now comment on the relevance of the theorems to applications.

Our urn model above is the first attempt ever to provide theoretical foundations for bootstrapping a database. Our basic hypothesis is that the assembly of a database involves random selection from a parent population, namely, the set of all records that could possibly enter the database. To see how the model is relevant to databases, consider a computer user interested in the database records relevant to some particular query. The following dictionary gives the translation from our urn model to the database retrieval application: urn = parent population of records; white ball = relevant record; black ball = irrelevant record; red number on a ball = a record's rank under Retrieval Algorithm 1; green number on a ball = a record's rank under Retrieval Algorithm 2; and bag = database of records. In applications, the database ROC_n is precisely $\widehat{ROC}_{\hat{N}} := \sum_{i=1}^{c} \widehat{T}_{\hat{B}(i)} / \left(c\widehat{T}\right)$, described after Eq (1.4).

In the language of databases, our urn model explicitly displays the population quantity r_n that the non-normalized database ROC_n $\widehat{R}_{\hat{N}} = \sum_{i=1}^{c} \widehat{T}_{\hat{B}(i)}$ estimates. Database ROC_ns and their differences are asymptotically normally distributed around the corresponding population quantities (Theorem 2.2 and Theorem 2.3). Moreover, bootstrapping the database mimics the corresponding normal distributions (Theorem 2.4 and Theorem 2.5), precisely the condition required to justify bootstrap inferences [5]. Although simplistic, our model therefore shows that inferences drawn from bootstrapping a database can not be summarily dismissed as lacking a statistical meaning.

If the utility of the bootstrap inferences is granted (as they are in bioinformatics, for better or worse), the theorems above become directly relevant to information retrieval.

For example, most measures of information retrieval are estimated from a subset of database records evaluated by experts. Because human time is limited, the subset is typically only a tiny fraction of the database. Thus, Theorem 2.1 and the Poisson bootstrap have relevance to measures of information retrieval in general, not just the ROC_n.

In addition, because databases nowadays are large, Poisson bootstraps typically incur only a small error. Poisson bootstraps are easier to program than the usual multinomial bootstrap. Because they require less computer space and time, they avoid in particular the computer storage problems that multinomial resampling of a large database can cause.

In the CLTs, the Condition O-$\log \bar{j}$ (despite its formidable appearance) is theoretically mild. Space precludes a detailed presentation, but many models of retrieval ranking [16] satisfy Condition O-$\log \bar{j}$, e.g., the "Exponential Model" alluded to by [18], where only the balls with numbers $\lfloor a^i \rfloor$ are white, for $a > 1$ and $i = 1, 2, \ldots$ (the floor function $\lfloor x \rfloor := \max\{ j \in \mathbb{N} : j \le x \}$).

Condition O-$\log \bar{j}$ also leads to conclusions accepted in practice. The Gaussian approximation in Theorem 2.4 has been observed in bioinformatics [9, 18]. Although researchers in bioinformatics have recognized $\widehat{R}_{\tilde{N}}$ as its mean, they bootstrap to estimate its variance. Eq (2.2) in Theorem 2.4 replaces the bootstrap estimate with the analytic approximation $\widehat{\Sigma}_{\tilde{N}}$, which eases computation. The approximation $\widehat{\Sigma}_{\tilde{N}}$ has already proved adequate in practice [18]. In addition, extensive simulations have confirmed the accuracy

of the CLTs in Theorem 2.4 and Theorem 2.5 when applied to data from [18] (results not shown).

We now proceed to the arduous technical task of proving our theorems.

3. THE PROOF OF THEOREM 2.1

Theorem 2.1 holds, because conditioned on the sums $D_{\mathfrak{M}} \coloneqq \sum_{i=1}^{m} D_{k(i)}$ and $\widehat{D}_{\mathfrak{M}} \coloneqq \sum_{i=1}^{m} \widehat{D}_{k(i)}$ being equal, $\left(D_{k(1)},...,D_{k(m)}\right)$ and $\left(\widehat{D}_{k(1)},...,\widehat{D}_{k(m)}\right)$ share a Multinomial$\left(\widehat{D}_{\mathfrak{M}}; p_{k(1)}p_{\mathfrak{M}}^{-1},..., p_{k(m)}p_{\mathfrak{M}}^{-1}\right)$ distribution. Because the total $\widehat{D}_{\mathfrak{M}} \coloneqq \sum_{i=1}^{m} \widehat{D}_{k(i)}$ has a Binomial$\left(D; p_{\mathfrak{M}}\right)$ distribution, the Chen-Stein method confirms the existence of a Poisson$\left(d_{\mathfrak{M}}\right)$ variate $D_{\mathfrak{M}}$ with $\mathbb{P}\left(D_{\mathfrak{M}} \neq \widehat{D}_{\mathfrak{M}}\right) \leq p_{\mathfrak{M}}$ [2, p. 8]. Theorem 2.1 follows. (We omit some technical details about augmenting the sample space, so that randomization can relate $\left(D_{k(1)},...,D_{k(m)}\right)$ to $\left(\widehat{D}_{k(1)},...,\widehat{D}_{k(m)}\right)$.)

4. THE PROOF OF THEOREM 2.2

Recall, the presence of a curved over-stroke indicates a multinomial sampling variate (e.g., $\widehat{R}_{\tilde{N}}$); its absence, the Poisson counterpart (e.g., R_N). First, Theorem 2.1 shows that for our purposes, Poisson and multinomial sampling become equivalent as $k \to \infty$, e.g., $\lim_{k\to\infty} \mathbb{P}\left(R_N \neq \widehat{R}_{\tilde{N}}\right) = \lim_{k\to\infty} \mathbb{P}\left(\Sigma_N^2 \neq \widehat{\Sigma}_{\tilde{N}}^2\right) = 0$.

To this end, consider some properties of $\underline{j}$ and $\overline{j}$ as $k \to \infty$. In particular,

$$\underline{j} < n \leq \overline{j} \quad \text{and} \quad \mathbb{P}\left(\underline{j} < \widehat{N} \leq \overline{j}\right) \to 1. \tag{4.1}$$

As in renewal theory, because of Eq (1.1), the inequality $\underline{j} < n \le \overline{j}$ is logically equivalent to $g_{\underline{j}} < c \le g_{\overline{j}}$. Likewise, to derive $\mathbb{P}\left(\underline{j} < \widehat{N} \le \overline{j}\right) \to 1$, we need to show $\mathbb{P}\left(\widehat{G}_{\underline{j}} < c \le \widehat{G}_{\overline{j}}\right) \to 1$. For future reference, we actually show more, that $\widehat{G}_{\underline{j}}^{-1/2}\left(c - \widehat{G}_{\underline{j}}\right) \to_{\mathbb{P}} \infty$ and $\widehat{G}_{\overline{j}}^{-1/2}\left(\widehat{G}_{\overline{j}} - c\right) \to_{\mathbb{P}} \infty$.

Denote the mean-centering of any random variate or expression X by $\dot{X} := \left(X\right)^{\cdot} := X - \mathbb{E}X$. Some consequences of Chebyshev's inequality appear so frequently that we name them.

Proposition 4.1 (The Expectation Bound): $X_k = \overline{O}_{\mathbb{P}}\left(\mathbb{E}\left|X_k\right|\right)$.

Proof: $\limsup_{m\to\infty} \limsup_{k\to\infty} \mathbb{P}\left(\left|X_k\right| > m\mathbb{E}\left|X_k\right|\right) \le \limsup_{m\to\infty} \limsup_{k\to\infty} m^{-1} = 0$.

For future reference, note the mean-centered expectation bound $\dot{X}_k = \overline{O}_{\mathbb{P}}\left(\mathbb{E}\left|X_k\right|\right)$. An application of the expectation bound to $\dot{X}_k^2 := \left(\dot{X}_k\right)^2$ yields the next proposition.

Proposition 4.2 (The Variance Bound): $\dot{X}_k^2 = \overline{O}_{\mathbb{P}}\left(\operatorname{var} X_k\right)$.

The notation $\overline{O}_{\mathbb{P}}\left(\bullet\right)$ speeds some proofs, because Proposition 4.1 and Proposition 4.2 bound probabilities directly with expectations. In a pattern repeated many times below, $X_k = \overline{O}_{\mathbb{P}}\left(x_k\right)$ and $x_k = o\left(y_k\right)$ then imply $X_k = o_{\mathbb{P}}\left(y_k\right)$.

To show $\widehat{G}_{\underline{j}}^{-1/2}\left(c - \widehat{G}_{\underline{j}}\right) \to_{\mathbb{P}} \infty$, for arbitrary m consider

$$\mathbb{P}\left\{\widehat{G}_{\underline{j}}^{-1/2}\left(c - \widehat{G}_{\underline{j}}\right) \le m\right\} = \mathbb{P}\left\{g_{\underline{j}}^{-1/2}\left(\widehat{G}_{\underline{j}} - g_{\underline{j}}\right) \ge g_{\underline{j}}^{-1/2}\left(c - g_{\underline{j}}\right) - mg_{\underline{j}}^{-1/2}\widehat{G}_{\underline{j}}^{1/2}\right\}. \qquad (4.2)$$

Because of the expectation bound, $g_{\underline{j}}^{-½}\widehat{G}_{\underline{j}}^{½} = \overline{O}_{\mathbb{P}}(1)$, and because of the variance bound, $g_{\underline{j}}^{-½}\left(\widehat{G}_{\underline{j}} - g_{\underline{j}}\right) = \overline{O}_{\mathbb{P}}(1)$, so $g_{\underline{j}}^{-½}\left(c - g_{\underline{j}}\right) \to \infty$ implies $\mathbb{P}\left\{\widehat{G}_{\underline{j}}^{-½}\left(c - \widehat{G}_{\underline{j}}\right) \le m\right\} \to 0$ in Eq (4.2). Thus, $\widehat{G}_{\underline{j}}^{-½}\left(c - \widehat{G}_{\underline{j}}\right) \to_{\mathbb{P}} \infty$. A similar argument shows $\widehat{G}_{\overline{j}}^{-½}\left(\widehat{G}_{\overline{j}} - c\right) \to_{\mathbb{P}} \infty$.

In Theorem 2.1, let $\mathfrak{M}$ be the set of balls before the $\overline{j}$-th black ball. Because Eq (4.1) implies that $\mathbb{P}\left(\widehat{N} \le \overline{j}\right) \to 1$ and $n \le \overline{j}$, neither $\widehat{R}_{\widehat{N}}$ (with probability approaching 1) nor r_n depends on balls outside of $\mathfrak{M}$. By hypothesis in Theorem 2.2, $\mathfrak{M}$ has total sampling probability $\sum_{i=1}^{m} p_{k(i)} = \left(t_{\overline{j}} + g_{\overline{j}}\right) D^{-1} \to 0$, so Theorem 2.1 implies that the error in Poisson sampling from $\mathfrak{M}$ vanishes in probability as $k \to \infty$. Thus, the following proof of Theorem 2.2 for Poisson sampling applies to multinomial sampling, as well.

A. Proof of $\left(R_N - r_n\right)/\sigma_n \Rightarrow \Phi$ for Poisson sampling: In Poisson sampling, $F_i \sim \text{Poisson}\left(f_i\right)$, so $\mathbb{E}F_i = \text{var}\, F_i = f_i$. In addition, $t_0 \coloneqq T_0 \coloneqq 0$ and $u_i \coloneqq t_i - t_{i-1}$, while $U_i \coloneqq T_i - T_{i-1} \sim \text{Poisson}\left(u_i\right)$ counts (in the red numbering) the white balls in the bag after the $(i-1)$-st black ball but before the i-th black ball. Moreover, the $\left\{F_j\right\}$ and $\left\{U_i\right\}$ are mutually independent, because they count disjoint sets, unlike the $\left\{T_i\right\}$.

Because $\mathbb{E}R_n = r_n$,

$$\dot{R}_n = \sum_{i=1}^{n-1} \dot{T}_i f_i + \dot{T}_n \left(c - \sum_{i=1}^{n-1} f_i \right) - \sum_{i=1}^{n} \left(T_n - T_i\right) \dot{F}_i \,. \tag{4.3}$$

In addition,

$$R_N - R_n = (T_N - T_n)\left(c - \sum_{i=1}^{N} F_i\right) + \sum_{i=n}^{N} (T_i - T_n) F_i \,. \tag{4.4}$$

In Eq (4.4) and below, we extend the meaning of the summation sign with the convention that $\sum_{i=N}^{n} X_i := \sum_{i=1}^{n} X_i - \sum_{i=1}^{N-1} X_i$, regardless of the relative magnitudes of n and N.

Partial summation yields

$$\sum_{i=1}^{n-1} \dot{T}_i f_i + \dot{T}_n \left(c - \sum_{i=1}^{n-1} f_i\right) = \sum_{j=1}^{n} \dot{U}_j \left(c - \sum_{i=1}^{j-1} f_i\right) = \sum_{j=1}^{n} \dot{U}_j \left(c - g_{j-1}\right). \tag{4.5}$$

Substitution into Eq (4.3) implies Theorem 2.2, if we can show that

$$\sigma_n^{-1} \dot{R}_n = \sigma_n^{-1} \left\{ \sum_{i=1}^{n} \dot{U}_i \left(c - g_{i-1}\right) - \sum_{i=1}^{n} (T_n - T_i) \dot{F}_i \right\} \Rightarrow \Phi \tag{4.6}$$

and $R_N - R_n = o_{\mathbb{P}}(\sigma_n)$. First, we prove $\operatorname{var} R_n = \sigma_n^2$ and $\sigma_n^{-1} \dot{R}_n \Rightarrow \Phi$.

Two observations are used repeatedly without comment below to expand variances. If A and B are variates independent of X and Y, then $\operatorname{cov}\left(\dot{A}X, \dot{B}Y\right) = \operatorname{cov}(A, B)\,\mathbb{E}(XY)$. If they are also independent of each other then $\operatorname{cov}\left(\dot{A}X, BY\right) = 0$. The covariance of the sums in Eq (4.6) is 0, so

$$\begin{aligned} \operatorname{var} R_n &= \operatorname{var}\left\{ \sum_{i=1}^{n} \dot{U}_i \left(c - g_{i-1}\right) \right\} + \operatorname{var}\left\{ \sum_{i=1}^{n} (T_n - T_i) \dot{F}_i \right\} \\ &= \sum_{i=1}^{n} \left(c - g_{i-1}\right)^2 u_i + \sum_{i=1}^{n} f_i \mathbb{E}\left(T_n - T_i\right)^2 \\ &= \sigma_n^2 \end{aligned} \,. \tag{4.7}$$

Eq (2.1) gives the last equality, because $T_n - T_i \sim \text{Poisson}(t_n - t_i)$ implies $\mathbb{E}(T_n - T_i)^2 = (t_n - t_i)(t_n - t_i + 1)$. Now, the proof of $\sigma_n^{-1}\dot{R}_n \Rightarrow \Phi$ derives from the following martingale CLT.

Let $\mathbb{I}\left[|X| > \eta\right]$ denote an indicator function, equaling 1 on the event $\left[|X| > \eta\right]$ and 0 otherwise; let $\text{var}(X|\mathcal{F})$ denote the conditional variance $\mathbb{E}(X^2|\mathcal{F}) - \{\mathbb{E}(X|\mathcal{F})\}^2$.

Theorem 4.1 [15]**:** Consider any array (X_{ki}) satisfying the martingale difference condition $\mathbb{E}(X_{ki}|X_{k1}, X_{k2}, ..., X_{k,i-1}) = 0$ ($k = 1, 2, 3, ...$ and $i = 1, 2, ..., j_k$). As $k \to \infty$, if

$$\sum_{i=1}^{j_k} \mathbb{E}\left(X_{ki}^2 \mathbb{I}\left[|X_{ki}| > \eta\right] \middle| X_{k1}, X_{k2}, ..., X_{k,i-1}\right) \to_{\mathbb{P}} 0, \tag{4.8}$$

and

$$\sum_{i=1}^{j_k} \text{var}\left(X_{ki} \middle| X_{k1}, X_{k2}, ..., X_{k,i-1}\right) \to_{\mathbb{P}} 1, \tag{4.9}$$

then $S_k := \sum_{i=1}^{j_k} X_{ki} \Rightarrow \Phi$.

Our proofs require only the following corollary of Theorem 4.1, in the same notation.

Corollary 4.1: *Let* $X_{ki} = C_{ki} Z_{ki}$ *(*$i = 1, 2, ..., j_k$*). Assume that the* C_{ki} *are* $\sigma(X_{k1}, \ldots, X_{k,i-1})$*-measurable random variates, with the* Z_{ki} *independent of each other and any* C_{kj} *with* $j \le i$*. Assume also that the* Z_{ki} *have been standardized, so* $\mathbb{E}Z_{ki} = 0$ *with* $\mathbb{E}Z_{ki}^2 = 1$*. For each* k*, define* $C_k^* := \max_i \left(C_{ki}^2 \mathbb{E}Z_{ki}^4\right)$*. If* $\text{var}\, S_k \equiv 1$ *identically, with* $C_k^* \to_{\mathbb{P}} 0$ *and* $\sum_{i=1}^{j_k} C_{ki}^2 \to_{\mathbb{P}} 1$ *as* $k \to \infty$*, then* $S_k \Rightarrow \Phi$.

Remark: If $C_{ki} = 0$, the corresponding Z_{ki} may be chosen arbitrarily. Our proofs exploit this freedom, letting $Z_{ki} = 0$ without comment if $C_{ki} = 0$.

The independence properties of the Z_{ki} and their standardization yield

$$\mathbb{E}\sum_{i=1}^{j_k} C_{ki}^2 = \sum_{i=1}^{j_k} \operatorname{var}\left(C_{ki} Z_{ki}\right) = \operatorname{var}\left(\sum_{i=1}^{j_k} C_{ki} Z_{ki}\right) = \operatorname{var} S_k = 1. \tag{4.10}$$

Thus, if $\operatorname{var}\sum_{i=1}^{j_k} C_{ki}^2 \to 0$, then $\sum_{i=1}^{j_k} C_{ki}^2 \to_{\mathbb{P}} 1$ and $\left(\sum_{i=1}^{j_k} C_{ki}^2\right)^{\bullet} \to_{\mathbb{P}} 0$.

Proof of Corollary 4.1: We verify the conditions of Theorem 4.1 for the X_{ki} in Corollary 4.1. Because $\{X_{ki}\}$ forms a martingale difference array,

$$\begin{aligned}\sum_{i=1}^{j_k} \operatorname{var}\left(X_{ki} \middle| X_{k1}, \ldots, X_{k,i-1}\right) &= \sum_{i=1}^{j_k} \mathbb{E}\left(X_{ki}^2 \middle| X_{k1}, \ldots, X_{k,i-1}\right) \\ &= \sum_{i=1}^{j_k} \mathbb{E}\left(C_{ki}^2 \middle| X_{k1}, \ldots, X_{k,i-1}\right) \mathbb{E} Z_{ki}^2, \\ &= \sum_{i=1}^{j_k} C_{ki}^2 \to_{\mathbb{P}} 1\end{aligned} \tag{4.11}$$

yielding Eq. (4.9). Chebyshev's inequality also shows that for arbitrary $\eta > 0$,

$$\eta^2 \mathbb{P}\left(C_{ki} \middle| Z_{ki} \middle| > \eta \middle| C_{ki}\right) \le C_{ki}^2 \mathbb{E} Z_{ki}^2 = C_{ki}^2, \tag{4.12}$$

so

$$\begin{aligned} 0 &\le \sum_{i=1}^{j_k} \mathbb{E}\left\{ X_{ki}^2 \mathbb{I}\left(|X_{ki}| > \eta\right) \middle| X_{k1}, \ldots, X_{k,i-1} \right\} \\ &= \sum_{i=1}^{j_k} C_{ki}^2 \mathbb{E}\left\{ Z_{ki}^2 \mathbb{I}\left(C_{ki}|Z_{ki}| > \eta\right) \middle| X_{k1}, \ldots, X_{k,i-1} \right\} \\ &\le \sum_{i=1}^{j_k} C_{ki}^2 \mathbb{E} Z_{ki}^4 \mathbb{P}\left(C_{ki}|Z_{ki}| > \eta \middle| X_{k1}, \ldots, X_{k,i-1}\right) \\ &\le \left(\sum_{i=1}^{j_k} C_{ki}^2 \right) \eta^{-2} C_k^* \to_{\mathbb{P}} 0 \end{aligned} . \qquad (4.13)$$

yielding Eq. (4.8). Theorem 4.1 therefore gives Corollary 4.1.

Corollary 4.1 proves Eq (4.6), as follows. Let $j_k = 2n$. For $i = 1, 2, ..., n$, if $u_i = 0$ define $C_{ki} = Z_{ki} = 0$, otherwise $C_{ki} = \sigma_n^{-1}\left(c - g_{i-1}\right) u_i^{1/2}$ and $Z_{ki} = u_i^{-1/2} \dot{U}_i$. For $i = j + n$ and $j = 1, 2, ..., n$, if $f_j = 0$ define $C_{ki} = Z_{ki} = 0$, otherwise $C_{ki} = -\sigma_n^{-1}\left(T_n - T_j\right) f_j^{1/2}$ and $Z_{ki} = f_j^{-1/2} \dot{F}_j$.

Notice that $S_k = \sigma_n^{-1} \dot{R}_n$ in Eq (4.6), so we only need to verify the hypotheses of Corollary 4.1. Because Eq (4.7) shows $\operatorname{var} S_k \equiv 1$ identically, all hypotheses are present by construction except $C_k^* \to_{\mathbb{P}} 0$ and $\sum_{i=1}^{j_k} C_{ki}^2 \to_{\mathbb{P}} 1$, which we now verify.

Consider the condition $C_k^* := \max_i \left(C_{ki}^2 \mathbb{E} Z_{ki}^4\right) \to_{\mathbb{P}} 0$. The assumptions in Theorem 2.2 yield $\max_{i=1,2,\ldots,n} \left(C_{ki}^2 \mathbb{E} Z_{ki}^4\right) = \max_{j=1,2,\ldots,n} \left\{ \sigma_n^{-2} \left(c - g_{j-1}\right)^2 \left(1 + 3u_j\right) \right\} \le \sigma_n^{-2} c^2 \left(1 + 3\bar{u}\right) \to 0$. Moreover, the expectation bound $\mathbb{E} T_n = t_n$ and the assumptions in Theorem 2.2 yield $\max_{i=n+1,\ldots,2n} \left(C_{ki}^2 \mathbb{E} Z_{ki}^4\right) = \max_{j=1,\ldots,n} \left\{ \sigma_n^{-2} \left(T_n - T_j\right)^2 \left(1 + 3 f_j\right) \right\} \le \sigma_n^{-2} T_n^2 \left(1 + 3\bar{f}\right) \to_{\mathbb{P}} 0$.

To derive $\operatorname{var} \sum_{i=1}^{j_k} C_{ki}^2 \to 0$, expand $\sigma_n^4 \operatorname{var}\left(\sum_{i=1}^{j_k} C_{ki}^2 \right) = \operatorname{var}\left\{ \sum_{i=1}^{n} \left(T_n - T_i\right)^2 f_i \right\}$:

$$\operatorname{var}\left\{\sum_{i=1}^{n}\left(T_n - T_i\right)^2 f_i\right\} = \sum_{i=1}^{n}\sum_{j=1}^{n} f_i f_j \operatorname{cov}\left\{\left(T_n - T_i\right)^2, \left(T_n - T_j\right)^2\right\}. \qquad (4.14)$$

Now, $T_n - T_i \sim \text{Poisson}\left(t_n - t_i\right)$ and $T_n - T_j = \left(T_n - T_i\right) + \left(T_i - T_j\right)$, where the terms in parentheses are independent for $i > j$. The expansion of $\left\{\left(T_n - T_i\right) + \left(T_i - T_j\right)\right\}^2$ in Eq (4.14), followed by explicit substitution of the corresponding covariances for independent Poisson variates, leads to a sum of terms similar to the square of the second sum in Eq (2.1), but with the degree of factors involving $\left\{t_n - t_i\right\}$ reduced by at least 1. The reduction in the degree suffices to show $\operatorname{var}\left\{\sum_{i=1}^{n}\left(T_n - T_i\right)^2 f_i\right\} = o\left(\sigma_n^4\right)$, as follows.

Because $c - \sum_{i=1}^{n} f_i \le 0 < c - \sum_{i=1}^{n-1} f_i$,

$$0 < c - g_{n-1} \le f_n \quad \text{and} \quad -f_n < c - g_n \le 0. \qquad (4.15)$$

The Poisson counterparts of Eq (4.15) follow in a similar fashion. Because $\bar{j} \ge n$, the assumption $\left(\log \bar{j}\right)\bar{f} = O\left(c^{1/2}\right)$ yields $f_n = O\left(c^{1/2}\right)$. Thus, Eq (4.15) implies that $g_n \sim c$.

Now,

$$\sum_{i=1}^{n}\left(t_n - t_i\right) f_i \le \sqrt{\left\{\sum_{i=1}^{n}\left(t_n - t_i\right)^2 f_i\right\}\sum_{i=1}^{n} f_i} = O\left(\sigma_n c^{1/2}\right) = o\left(\sigma_n^{3/2}\right) \qquad (4.16)$$

by the Cauchy-Schwartz inequality, and similarly,

$$\sum_{i=1}^{n}\left(c - g_{i-1}\right) u_i \le \sqrt{\left\{\sum_{i=1}^{n}\left(c - g_{i-1}\right)^2 u_i\right\}\sum_{i=1}^{n} u_i} = O\left(\sigma_n c^{1/2}\right) = o\left(\sigma_n^{3/2}\right), \qquad (4.17)$$

Frequently below, we bound an expression by expanding it and comparing to σ_n^{2m}, noting then that the degree of a factor $\{c-g_i\}$ or $\{t_n-t_i\}$ in σ_n^{2m} has been reduced. We can then show that the expression is $o\left(\sigma_n^{2m}\right)$, by noting the bounds in Eq (4.16) and (4.17). This "comparison estimate" obviates many detailed computations. It is most useful for mutually independent sets of variates $\{A_i\}$ and $\{B_i\}$, when $\operatorname{cov}\left(A_iB_i, A_jB_j\right) = \mathbb{E}\left(A_iA_j\right)\operatorname{cov}\left(B_i, B_j\right) + \operatorname{cov}\left(A_i, A_j\right)\mathbb{E}B_i\mathbb{E}B_j$. In addition, if the $\{B_i\}$ are mutually independent, then $\operatorname{cov}\left(A_iB_i, A_jB_j\right) = \operatorname{cov}\left(A_i, A_j\right)\mathbb{E}B_i\mathbb{E}B_j$.

The comparison estimate applied to σ_n^4 and the right side of Eq (4.14) shows $\operatorname{var}\left\{\sum_{i=1}^{n}\left(T_n-T_i\right)^2 f_i\right\} = o\left(\sigma_n^4\right)$, finishing the proof of Eq (4.6).

Having proved $\sigma_n^{-1}\dot{R}_n \Rightarrow \Phi$, we now prove that $R_N - R_n = o_{\mathbb{P}}\left(\sigma_n\right)$ by showing that both terms in $R_N - R_n$ from Eq (4.4) are $o_{\mathbb{P}}\left(\sigma_n\right)$.

First, $\underline{j} < n$, $\mathbb{P}\left(N \le \bar{j}\right) \to 1$, and the expectation $\mathbb{E}\left(T_{\bar{j}} - T_{\underline{j}}\right) = t_{\bar{j}} - t_{\underline{j}}$ yield

$$T_N - T_n = \bar{O}_{\mathbb{P}}\left(T_{\bar{j}} - T_{\underline{j}}\right) = \bar{O}_{\mathbb{P}}\left(t_{\bar{j}} - t_{\underline{j}}\right) = o_{\mathbb{P}}\left(\sigma_n c^{-1/2}\right). \tag{4.18}$$

Because $0 \le \sum_{i=1}^{N} F_i - c \le F_N$ in Eq (4.4), we next want to prove that $F_N = o_{\mathbb{P}}\left(c^{1/2}\right)$. The following useful proposition is the key.

Proposition 4.3: *Consider positive integer random variates J and deterministic bounds j satisfying $\lim_{k\to\infty}\mathbb{P}\left(J > j\right) = 0$, with $j \to \infty$. Let $\Lambda_i \sim \text{Poisson}\left(\lambda_i\right)$ be (possibly*

dependent) variates with $\bar{\lambda} := \max_{i=1,2,\ldots,j} \lambda_i$ *and* $\bar{\lambda} \log\log j \to \infty$. *Then* $\bar{\Lambda} := \max_{i=1,2,\ldots,J} \Lambda_i = o_{\mathbb{P}}\left(\bar{\lambda} \log j\right)$.

Proof: Let $\bar{\Lambda}_j := \max_{i=1,2,\ldots,j} \Lambda_i$. For $\Lambda \sim \text{Poisson}(\lambda)$, $\mathbb{E}\exp(\theta\Lambda) = \exp\left\{\lambda\left(e^{\theta}-1\right)\right\}$. For $\varepsilon \log j > 0$ and $\theta_0 = \log(\varepsilon \log j)$,

$$\begin{aligned}
\mathbb{P}\left(\bar{\Lambda}_j \geq \varepsilon\bar{\lambda} \log j\right) &\leq \sum_{i=1}^{j} \mathbb{P}\left(\Lambda_i \geq \varepsilon\bar{\lambda} \log j\right) \\
&\leq \inf_{\theta\geq 0} \sum_{i=1}^{j} \exp\left\{\lambda_i\left(e^{\theta}-1\right) - \left(\varepsilon\bar{\lambda} \log j\right)\theta\right\} \\
&\leq \inf_{\theta\geq 0} j \exp\left\{\bar{\lambda}\left(e^{\theta}-1\right) - \left(\varepsilon\bar{\lambda} \log j\right)\theta\right\} \qquad (4.19) \\
&\leq j \exp\left\{\bar{\lambda}\left(e^{\theta_0}-1\right) - \left(\varepsilon\bar{\lambda} \log j\right)\theta_0\right\} \\
&\leq \exp\left\{-\log j\left[\varepsilon\bar{\lambda}\left(\theta_0-1\right)-1\right]\right\} \to 0
\end{aligned}$$

by Chernoff's bound [17, p. 39]. Proposition 4.3 follows, because $\mathbb{P}\left(\bar{\Lambda} \geq \varepsilon\bar{\lambda} \log j\right) \leq \mathbb{P}\left(\bar{\Lambda}_j \geq \varepsilon\bar{\lambda} \log j\right) + \mathbb{P}\left(J > j\right) \to 0$.

Because Eq (4.1) shows that $\mathbb{P}\left(N > \bar{j}\right) \to 0$, Proposition 4.3 with $J = N$ and $j = \bar{j}$ yields $F_N = \bar{O}_{\mathbb{P}}\left(\max_{i=1,2,\ldots,\bar{j}} F_i\right) = o_{\mathbb{P}}\left\{\left(\log \bar{j}\right)\bar{f}\right\} = o_{\mathbb{P}}\left(c^{1/2}\right)$, because $\bar{f} \log\log \bar{j} \to \infty$. For the first term of $R_N - R_n$ in Eq (4.4), therefore,

$$\left|\left(T_N - T_n\right)\left(c - G_N\right)\right| \leq \left|T_N - T_n\right| F_N = o_{\mathbb{P}}\left(\sigma_n\right). \qquad (4.20)$$

For the second term of $R_N - R_n$, Eq (4.15) gives

$$\left|\sum_{i=n}^{N} F_i\right| \leq \left|G_N - c\right| + \left|-\dot{G}_{n-1}\right| + \left|c - g_{n-1}\right| \leq F_N + \left|\dot{G}_{n-1}\right| + f_n. \qquad (4.21)$$

The variance bound $\mathbb{E}\left(\sum_{i=1}^{n-1}\dot{F}_i\right)^2 = \sum_{i=1}^{n-1} f_i = g_{n-1}$ yields $\dot{G}_{n-1} = \bar{O}_{\mathbb{P}}\left(g_{n-1}^{1/2}\right) = \bar{O}_{\mathbb{P}}\left(c^{1/2}\right)$, so

$$(4.22)\qquad \left|\sum_{i=n}^{N}(T_i - T_n)F_i\right| \le |T_N - T_n|\left|\sum_{i=n}^{N} F_i\right| \le |T_N - T_n|\left(F_N + \left|\dot{G}_{n-1}\right| + f_n\right) = o_{\mathbb{P}}\left(\sigma_n\right).$$

B. Proof of $\Sigma_N \sim_{\mathbb{P}} \sigma_n$.

Our proof is a centering argument with three steps, namely: (1) $\Sigma_N^2 - \Sigma_n^2 = o_{\mathbb{P}}\left(\sigma_n^2\right)$; (2) $\Sigma_n^2 - \mathbb{E}\Sigma_n^2 = o_{\mathbb{P}}\left(\sigma_n^2\right)$; and (3) $\mathbb{E}\Sigma_n^2 - \sigma_n^2 = o\left(\sigma_n^2\right)$. The first step is as follows:

$$\begin{aligned}\Sigma_N^2 - \Sigma_n^2 &= \sum_{i=n+1}^{N}\left(c - G_{i-1}\right)^2 U_i + \sum_{i=1}^{N}(T_N - T_i)(T_N - T_i + 1)F_i - \\ &\qquad \sum_{i=1}^{n}(T_n - T_i)(T_n - T_i + 1)F_i \\ &= \sum_{i=n+1}^{N}\left(c - G_{i-1}\right)^2 U_i + \sum_{i=n+1}^{N}(T_N - T_i)(T_N - T_i + 1)F_i + \\ &\qquad (T_N - T_n)\sum_{i=1}^{n}\left\{(T_N - T_n + 1) + 2(T_n - T_i)\right\}F_i\end{aligned}, \qquad (4.23)$$

so

$$\begin{aligned}\left|\Sigma_N^2 - \Sigma_n^2\right| &\le \left|\sum_{i=n+1}^{N}\left(c - G_{i-1}\right)^2 U_i\right| + \\ &\qquad (T_N - T_n)(T_N - T_n + 1)G_N + 2|T_N - T_n|\sum_{i=1}^{n}(T_n - T_i)F_i\end{aligned}. \qquad (4.24)$$

The variance bound yields $G_{n-1} - g_{n-1} = \bar{O}_{\mathbb{P}}\left(g_{n-1}^{1/2}\right) = \bar{O}_{\mathbb{P}}\left(c^{1/2}\right)$, so $|c - g_{n-1}| \le f_n = O\left(c^{1/2}\right)$ implies $|c - G_{n-1}| = \bar{O}_{\mathbb{P}}\left(c^{1/2}\right)$. With $|c - G_{N-1}| \le F_N = \bar{O}_{\mathbb{P}}\left(c^{1/2}\right)$, we have $\left(c - G_{i-1}\right)^2 = \bar{O}_{\mathbb{P}}(c)$ for $i \in [N]\backslash[n]$ or $i \in [n]\backslash[N]$. In the first term of Eq (4.24), $\sum_{i=n+1}^{N}\left(c - G_{i-1}\right)^2 U_i = \bar{O}_{\mathbb{P}}\left(c|T_N - T_n|\right) = o_{\mathbb{P}}\left(c^{1/2}\sigma_n\right) = o_{\mathbb{P}}\left(\sigma_n^2\right)$ because $c = o\left(\sigma_n\right)$. Eq (4.18)

shows the second term $(T_N - T_n)(T_N - T_n + 1)G_N = \bar{O}_{\mathbb{P}}\left\{c\left(T_{\bar{j}} - T_{\underline{j}}\right)\left(T_{\bar{j}} - T_{\underline{j}} + 1\right)\right\} = o_{\mathbb{P}}\left(\sigma_n^2\right)$. Moreover, Eqs (4.16) and (4.18) together show that the third term $\left|T_N - T_n\right|\sum_{i=1}^{n}(T_n - T_i)F_i = \bar{O}_{\mathbb{P}}\left(\left|T_N - T_n\right|\sigma_n c^{1/2}\right) = o_{\mathbb{P}}\left(\sigma_n^2\right)$.

The variance bound on each term of Σ_n^2 in Eq (2.1) shows $\left(\Sigma_n^2\right)^{\bullet} := \Sigma_n^2 - \mathbb{E}\Sigma_n^2 = o_{\mathbb{P}}\left(\sigma_n^2\right)$, as follows. After expansion of the variances $\mathrm{var}\left\{\sum_{i=1}^{n}\left(c - G_{i-1}\right)^2 U_i\right\}$, $\mathrm{var}\left\{\sum_{i=1}^{n}(T_n - T_i)^2 F_i\right\}$ and $\mathrm{var}\left\{\sum_{i=1}^{n}(T_n - T_i)F_i\right\}$, the comparison estimate after Eq (4.16) applies to each variance and $\sigma_n^4 = \sigma_n^2\sigma_n^2$.

To show $\mathbb{E}\Sigma_n^2 - \sigma_n^2 = o\left(\sigma_n^2\right)$, apply the comparison estimate after Eq (4.16) again:

$$\mathbb{E}\Sigma_n^2 - \sigma_n^2 = \sum_{i=1}^{n} g_i u_i + \sum_{i=1}^{n}\left(t_n - t_i\right) f_i \le g_n t_n + \sum_{i=1}^{n}\left(t_n - t_i\right) f_i = o\left(\sigma_n^2\right). \qquad (4.25)$$

Eq (4.25) concludes the proof of Theorem 2.2

5. THE PROOF OF THEOREM 2.3

To begin, we extend the notations in Section 2 describing the balls common to specified sets of red and green numbers. Recall Section 2's definitions of π, $\mathfrak{C}_{nn'}$, and $\mathfrak{C}'_{n'n}$. Let $\mathfrak{D}_{nn'} := [n] \setminus \mathfrak{C}_{nn'}$ be the relative complement of $\mathfrak{C}_{nn'}$ within $[n] := \{1, 2, ..., n\}$; similarly, $\mathfrak{D}'_{n'n} := [n'] \setminus \mathfrak{C}'_{n'n}$. Thus, $\mathfrak{C}_{nn'}$ depends on <u>c</u>ommon balls; $\mathfrak{D}_{nn'}$, on <u>d</u>isparate. Recall also Section 2's definition of $\widehat{V}_{ij}$ and $v_{ij} := \mathbb{E}\widehat{V}_{ij}$. The variate $\widehat{V}_{ij}$ counts a set of balls in the bag, so it has a Poisson analog V_{ij}, whose expectation $\mathbb{E}V_{ij} = v_{ij}$. Similarly, under

Poisson sampling, let W_i count the white balls that come in the red numbering after the $i-1$-st black ball but before the i-th black ball that also come in the green numbering after the n'-th black ball. As usual, let $w_i = \mathbb{E}W_i$. Symmetry gives $V_{ij} = V'_{ji}$; conservation, $U_i = \sum_{j=1}^{n} V_{ij} + W_i$. Similar relations govern the expectations.

A proof of Theorem 2.3 for Poisson sampling proves it for multinomial sampling, as well. The only change to the start of Section 3 is to let $\mathfrak{M}$ in Theorem 2.1 be the union of two sets of balls: (1) those coming in the red numbering before the $\bar{j}$-th black ball; and (2) those coming in the green numbering before the $\bar{j}'$-th black ball.

A. Proof of $\left\{\left(R'_{N'} - R_N\right) - \left(r'_{n'} - r_n\right)\right\} / \sigma_{nn'} \Rightarrow \Phi$ for Poisson sampling: Recall that $\mathbb{E}R_n = r_n$ and $\mathbb{E}R'_n = r'_n$. Our proof follows the pattern of the proof of Theorem 2.2, but first we show that $\sigma_n = O\left(\sigma_{nn'}\right)$ and $\sigma'_{n'} = O\left(\sigma_{nn'}\right)$. Because Section 3 shows that $R_N - R_n = o_{\mathbb{P}}\left(\sigma_n\right)$ and $R'_{N'} - R'_{n'} = o_{\mathbb{P}}\left(\sigma'_{n'}\right)$, all that then remains is to show that $\sigma_{nn'}^{-1}\left(\dot{R}'_{n'} - \dot{R}_n\right) \Rightarrow \Phi$.

To show $\sigma_n = O\left(\sigma_{nn'}\right)$, note that $1 - 2rx + x^2 \geq 1 - \left(r^+\right)^2$ for $x \geq 0$, where $r^+ := \max\{0, r\}$. Thus,

$$\liminf_{k\to\infty} \frac{\sigma_{nn'}^2}{\sigma_n^2} = \liminf_{k\to\infty} \left\{1 - 2\left(\frac{\gamma_{nn'}}{\sigma_n \sigma'_{n'}}\right)\left(\frac{\sigma'_{n'}}{\sigma_n}\right) + \left(\frac{\sigma'_{n'}}{\sigma_n}\right)^2\right\} \geq \liminf_{k\to\infty} \left\{1 - \left(\rho^+_{nn'}\right)^2\right\} > 0, \quad (5.1)$$

where the final inequality follows from the assumption $\rho < 1$. Thus, $\sigma_n = O\left(\sigma_{nn'}\right)$, and by symmetry, $\sigma'_{n'} = O\left(\sigma_{nn'}\right)$.

Next, we show that $\operatorname{cov}(R_n, R'_{n'})$ equals $\gamma_{nn'}$. The $\{F_i\}$ and $\{F'_i\}$ are independent of the $\{U_i\}$ and $\{U'_i\}$, so Eq (4.6) yields

$$\begin{aligned}\operatorname{cov}(R_n, R'_{n'}) = \operatorname{cov}&\left\{\sum_{i=1}^{n}\dot{U}_i(c-g_{i-1}), \sum_{j=1}^{n'}\dot{U}'_j(c-g'_{j-1})\right\} + \\ &\operatorname{cov}\left\{\sum_{i=1}^{n}(T_n - T_i)\dot{F}_i, \sum_{j=1}^{n'}(T'_{n'} - T'_j)\dot{F}'_j\right\}\end{aligned}. \qquad (5.2)$$

To calculate the first term on the right in Eq (5.2), recall: $V_{ij} = V'_{ji}$, W_i, and W'_i have expectations $v_{ij} = v'_{ji}$, w_i, and w'_i; they are independent Poisson variates, independent also of the $\{F_i\}$ and $\{F'_i\}$. The symmetry $V_{ij} = V'_{ji}$ and the conservation law $U_i = \sum_{j=1}^{n} V_{ij} + W_i$ imply $\mathbb{E}(\dot{U}_i\dot{U}'_j) = \mathbb{E}\dot{V}_{ij}^2 = v_{ij}$. Therefore

$$\operatorname{cov}\left\{\sum_{i=1}^{n}\dot{U}_i(c-g_{i-1}), \sum_{j=1}^{n'}\dot{U}'_j(c-g'_{j-1})\right\} = \sum_{i=1}^{n}\sum_{j=1}^{n'}(c-g_{i-1})(c-g'_{j-1})v_{ij}\,. \qquad (5.3)$$

To calculate the second term in Eq (5.2), note that $\operatorname{var} F_i = f_i$, so

$$\operatorname{cov}\left\{\sum_{i=1}^{n}(T_n - T_i)\dot{F}_i, \sum_{j=1}^{n'}(T'_{n'} - T'_j)\dot{F}'_j\right\} = \sum_{i\in\mathfrak{C}_{nn'}} f_i\mathbb{E}\left\{(T_n - T_i)(T'_{n'} - T'_{\pi(i)})\right\}, \qquad (5.4)$$

The equation $\operatorname{cov}(\dot{U}_i, \dot{U}'_j) = \mathbb{E}(\dot{U}_i\dot{U}'_j) = v_{ij}$ yields $\mathbb{E}(U_iU'_j) = v_{ij} + u_iu'_j$. Thus,

$$\begin{aligned}\mathbb{E}\left\{(T_n - T_i)(T'_{n'} - T'_j)\right\} &= \sum_{a=i+1}^{n}\sum_{b=j+1}^{n'}\mathbb{E}(U_aU'_b) \\ &= \sum_{a=i+1}^{n}\sum_{b=j+1}^{n'}(v_{ab} + u_au'_b) \\ &= \sum_{a=i+1}^{n}\sum_{b=j+1}^{n'}v_{ab} + (t_n - t_i)(t'_{n'} - t'_j)\end{aligned}\,. \qquad (5.5)$$

Eqs (5.2)-(5.5) together show $\gamma_{nn'} = \mathrm{cov}\left(R_n, R'_{n'}\right)$ in Eq (2.3).

To show $\sigma_{nn'}^{-1}\left(\dot{R}'_{n'} - \dot{R}_n\right) \Rightarrow \Phi$, we show that $\sigma_{nn'}^{-1}\left(\dot{R}'_{n'} - \dot{R}_n\right) = \sum_{i=1}^{j_k} C_{ki} Z_{ki} = S_k$ and then use Corollary 4.1. Let $j_k = nn' + 3n + 2n'$. For $i = (j-1)n' + m$ with $j = 1, 2, ..., n$ and $m = 1, 2, ..., n'$, if $v_{jm} = 0$, let $C_{ki} = Z_{ki} = 0$, otherwise let $C_{ki} = \sigma_{nn'}^{-1}\left(g_{j-1} - g'_{m-1}\right) v_{jm}^{1/2}$ and $Z_{ki} = v_{jm}^{-1/2}\dot{V}_{jm}$. For $i = nn' + j$ and $j = 1, 2, ..., n'$, if $w'_j = 0$ define $C_{ki} = Z_{ki} = 0$, otherwise $C_{kj} = \sigma_{nn'}^{-1}\left(c - g'_{j-1}\right) w_j'^{1/2}$ and $Z_{kj} = w_j'^{-1/2}\dot{W}'_j$. For $i = nn' + n' + j$ and $j = 1, 2, ..., n$, if $w_j = 0$ define $C_{ki} = Z_{ki} = 0$, otherwise $C_{ki} = -\sigma_{nn'}^{-1}\left(c - g_{j-1}\right) w_j^{1/2}$ and $Z_{ki} = w_j^{-1/2}\dot{W}_j$. For $i = nn' + n + n' + j$ and $j = 1, 2, ..., n'$, if $j \in \mathfrak{D}'_{n'n}$ and $f'_j > 0$, define $C_{ki} = -\sigma_{nn'}^{-1}\left(T'_{n'} - T'_j\right) f_j'^{1/2}$, and $Z_{ki} = f_j'^{-1/2}\dot{F}'_j$, otherwise $C_{ki} = Z_{ki} = 0$. For $i = nn' + n + 2n' + j$ and $j = 1, 2, ..., n$, if $j \in \mathfrak{C}_{nn'}$ and $f_j > 0$, define $C_{ki} = \sigma_{nn'}^{-1}\left\{\left(T_n - T_j\right) - \left(T'_{n'} - T'_{\pi(j)}\right)\right\} f_j^{1/2}$ and $Z_{ki} = f_j^{-1/2}\dot{F}_j$, otherwise $C_{ki} = Z_{ki} = 0$. Finally, for $i = nn' + 2n + 2n' + j$ and $j = 1, 2, ..., n$, if $j \in \mathfrak{D}_{nn'}$ and $f_j > 0$, define $C_{ki} = \sigma_{nn'}^{-1}\left(T_n - T_j\right) f_j^{1/2}$ and $Z_{ki} = f_j^{-1/2}\dot{F}_j$, otherwise $C_{ki} = Z_{ki} = 0$. Now,

$$\begin{aligned}\sigma_{nn}\sum_{i=1}^{j_k} C_{ki} Z_{ki} = &\left\{\sum_{j=1}^{n}\sum_{m=1}^{n'}\left(g_{j-1} - g'_{m-1}\right)\dot{V}_{jm} + \sum_{j=1}^{n'}\left(c - g'_{j-1}\right)\dot{W}'_j - \sum_{j=1}^{n}\left(c - g_{j-1}\right)\dot{W}_j\right\} \\ &- \sum_{j\in\mathfrak{D}'_{n'n}}\left(T'_{n'} - T'_j\right)\dot{F}'_j + \sum_{j\in\mathfrak{C}_{nn'}}\left\{\left(T_n - T_j\right) - \left(T'_{n'} - T'_{\pi(j)}\right)\right\}\dot{F}_j + \sum_{j\in\mathfrak{D}_{nn'}}\left(T_n - T_j\right)\dot{F}_j\end{aligned} \quad .(5.6)$$

We now check that $S_k = \sigma_{nn'}^{-1}\left(\dot{R}'_{n'} - \dot{R}_n\right)$ by showing the common value in Eq (5.6) is $\dot{R}'_{n'} - \dot{R}_n$. Because of the conservation law, $U_i = \sum_{j=1}^{n} V_{ij} + W_i$ and the symmetry $V_{ij} = V'_{ji}$,

the first three terms in Eq (5.6) equal $\sum_{i=1}^{n'}(c-g'_{i-1})\dot{U}'_i - \sum_{i=1}^{n}(c-g_{i-1})\dot{U}_i$ in $\dot{R}'_{n'} - \dot{R}_n$. Because the last three terms equal $-\sum_{j=1}^{n'}(T'_{n'} - T'_j)\dot{F}'_j + \sum_{j=1}^{n}(T_n - T_j)\dot{F}_j$ in $\dot{R}'_{n'} - \dot{R}_n$, $S_k := \sum_{i=1}^{j_k} C_{ki} Z_{ki} = \sigma_{nn'}^{-1}\left(\dot{R}'_{n'} - \dot{R}_n\right)$.

Again, we only need to verify the hypotheses of Corollary 4.1. Because Eqs (5.2)-(5.5) together show that $\operatorname{var} S_k \equiv 1$ identically, all hypotheses are present by construction except $C_k^* \to_{\mathbb{P}} 0$, and $\sum_{i=1}^{j_k} C_{ki}^2 \to_{\mathbb{P}} 1$, which we now verify.

Consider the condition $C_k^* := \max_i\left(C_{ki}^2 \mathbb{E} Z_{ki}^4\right) \to_{\mathbb{P}} 0$. For $j \in \mathfrak{C}_{nn'}$,

$$\begin{aligned}\sigma_{nn'}^{-2}\left\{\left(T_n - T_j\right) - \left(T'_{n'} - T'_{\pi(j)}\right)\right\}^2 \left(1+3f_j\right) \\ \le \sigma_{nn'}^{-2}\max\left\{T_n^2, T_{n'}'^2\right\}\left(1+3\max_{j=1,\ldots,n} f_j\right) \to_{\mathbb{P}} 0\end{aligned}. \tag{5.7}$$

For the other contributions to C_k^*, the inequalities $v_{ij} \le u_i$ and $w_i \le u_i$ lead to the same bounds found in the proof of Theorem 2.2. Thus, $C_k^* \to_{\mathbb{P}} 0$.

Finally, consider $\sum_{i=1}^{j_k} C_{ki}^2 \to_{\mathbb{P}} 1$. Let $\left(\sum_{i=1}^{j_k} C_{ki}^2\right)^{\bullet} = \sigma_{nn'}^{-2}\left(\dot{Y}_n + \dot{Y}'_{n'} - 2\dot{Y}_{nn'}\right)$, where $Y_n := \sum_{i=1}^{n}\left(T_n - T_i\right)^2 f_i$, $Y'_{n'} := \sum_{i=1}^{n'}\left(T'_{n'} - T'_i\right)^2 f'_i$, and $Y_{nn'} := \sum_{i\in\mathfrak{C}_{nn'}}\left(T_n - T_i\right)\left(T'_{n'} - T'_{\pi(i)}\right) f_i$. Because the end of Section 3 shows that $\operatorname{var} Y_n = o\left(\sigma_{nn'}^4\right)$ and $\operatorname{var} Y'_n = o\left(\sigma_{nn'}^4\right)$, and because the following shows that $\operatorname{var} Y_{nn'} = o\left(\sigma_{nn'}^4\right)$, the variance bound yields $\left(\sum_{i=1}^{j_k} C_{ki}^2\right)^{\bullet} \to_{\mathbb{P}} 0$, or equivalently, $\sum_{i=1}^{j_k} C_{ki}^2 \to_{\mathbb{P}} 1$.

To show $\operatorname{var} Y_{nn'} = o\left(\sigma_{nn'}^4\right)$, bound as follows the terms $c_{ij} := \operatorname{cov}\left\{\left(T_n - T_i\right)\left(T'_{n'} - T'_{\pi(i)}\right), \left(T_n - T_j\right)\left(T'_{n'} - T'_{\pi(j)}\right)\right\}$ from the expansion of $\operatorname{var} Y_{nn'}$. Write $T_n - T_i = t_n - t_i + \dot{T}_n - \dot{T}_i$, etc., and expand the factors in the covariance. Consider centered moments of $X \sim \text{Poisson}(x)$: $\mathbb{E}\dot{X}^2 = x$, $\mathbb{E}\dot{X}^3 = x$ and $\mathbb{E}\dot{X}^4 = 3x^2 + x$. Observe: all these centered moments are non-negative, and the last two have powers of x two less than the corresponding power of $\dot{X}$. Each term (both plus and minus) in the expansion of the covariance c_{ij} is therefore bounded by a bounded number of terms, each of which is either t_n^2 or the product of t_n with either one or two factors from $\left(t_n - t_i\right)$, $\left(t'_{n'} - t'_{\pi(i)}\right)$, $\left(t_n - t_j\right)$, or $\left(t'_{n'} - t'_{\pi(j)}\right)$. After multiplication by $f_i f_j$, the sum of the terms is $O\left(\sigma_n^2 \sigma_{n'}'^2\right)$, because $\sum_{i=1}^n f_i = O(c) = o(\sigma_n)$, $t_n = o(\sigma_n)$, $\sum_{i=1}^n \left(t_n - t_i\right) f_i = o\left(\sigma_n^{3/2}\right)$, and $\sum_{i \in \mathfrak{C}_{nn'}} \left(t_n - t_i\right)\left(t'_{n'} - t'_{\pi(i)}\right) f_i \le \sigma_n \sigma'_{n'}$, etc.

Because $\sum_{i=1}^{j_k} C_{ki}^2 \to_{\mathbb{P}} 1$ and $\sigma_{nn'}^{-1}\left(\dot{R}'_{n'} - \dot{R}_n\right) \Rightarrow \Phi$, the first part of Theorem 2.3 is proved.

B. Proof of $\Sigma_{NN'} \sim_{\mathbb{P}} \sigma_{nn'}$, with $\limsup_{k\to\infty} \mathrm{P}_{NN'} = \rho$ a.s.-$\mathbb{P}$ along a subsequence.

Assume for a moment that $\Gamma_{NN'} - \gamma_{nn'} = o_{\mathbb{P}}\left(\sigma_n \sigma'_{n'}\right)$. Because $\Sigma_N \sim_{\mathbb{P}} \sigma_n$ and $\Sigma_{N'} \sim_{\mathbb{P}} \sigma_{n'}$, we have $\mathrm{P}_{NN'} - \rho_{nn'} = o_{\mathbb{P}}(1)$, i.e., $\mathrm{P}_{NN'} - \rho_{nn'} \to_{\mathbb{P}} 0$. We can therefore select a subsequence of $k = 1, 2, \ldots$, so that $\mathrm{P}_{NN'} - \rho_{nn'} \to 0$ a.s.-$\mathbb{P}$. Along the subsequence,

$\limsup_{k\to\infty} \mathrm{P}_{NN'} = \limsup_{k\to\infty} \rho_{nn'} = \rho$ a.s.-$\mathbb{P}$. Moreover, we also have $\Sigma_{NN'} \sim_{\mathbb{P}} \sigma_{nn'}$, as follows. Eq (5.1) yields $\sigma_n = O(\sigma_{nn'})$ and $\sigma'_{n'} = O(\sigma_{nn'})$, so

$$\begin{aligned}\Sigma^2_{NN'} - \sigma^2_{nn'} &= \left(\Sigma^2_N - \sigma^2_n\right) + \left(\Sigma^2_{N'} - \sigma^2_{n'}\right) - 2\left(\Gamma_{NN'} - \gamma_{nn'}\right) \\ &= \sigma^2_n\left(\sigma_n^{-2}\Sigma^2_N - 1\right) + \sigma^2_{n'}\left(\sigma_{n'}^{-2}\Sigma^2_{N'} - 1\right) - 2\left(\Gamma_{NN'} - \gamma_{nn'}\right). \\ &= o_{\mathbb{P}}\left(\sigma^2_{nn'}\right)\end{aligned} \tag{5.8}$$

We therefore only need to show $\Gamma_{NN'} - \gamma_{nn'} = o_{\mathbb{P}}(\sigma_n\sigma'_{n'})$, which we do with the usual centering argument with three steps, namely: (1) $\Gamma_{NN'} - \Gamma_{nn'} = o_{\mathbb{P}}(\sigma_n\sigma'_{n'})$; (2) $\Gamma_{nn'} - \mathbb{E}\Gamma_{nn'} = o_{\mathbb{P}}(\sigma_n\sigma'_{n'})$; and (3) $\mathbb{E}\Gamma_{nn'} - \gamma_{nn'} = o(\sigma_n\sigma'_{n'})$. For notational convenience, let $\xi_{nn'}$, $\alpha_{nn'}$, and $\beta_{nn'}$ denote the first, second, and third terms on the right side of Eq (2.3); $\Xi_{NN'}$, $\mathrm{A}_{NN'}$, and $\mathrm{B}_{NN'}$, the corresponding first, second, and third terms on the right side of Eq (2.4).

First, we apply the Cauchy-Schwartz inequality to $\Gamma_{NN'} - \Gamma_{nn'}$, where by definition, $\Gamma_{NN'} = \Xi_{NN'} + \mathrm{A}_{NN'} + \mathrm{B}_{NN'}$. To bound $\Xi_{NN'} - \Xi_{nn'}$, rearrange the sums as

$$\sum_{i=1}^{N}\sum_{j=1}^{N'} - \sum_{i=1}^{n}\sum_{j=1}^{n'} = \sum_{i=n+1}^{N}\sum_{j=n'+1}^{N'} + \sum_{i=n+1}^{N}\sum_{j=1}^{n'} + \sum_{i=1}^{n}\sum_{j=n'+1}^{N'} \tag{5.9}$$

Now, $V_{ij}^2 \le \min\left\{U_i^2, U_j'^2\right\} \le U_i U'_j$, so Eq (5.9) and the Cauchy-Schwartz inequality yield

$$\begin{aligned}\left|\Xi_{NN'} - \Xi_{nn'}\right| &= \left|\left(\sum_{i=1}^{N}\sum_{j=1}^{N'} - \sum_{i=1}^{n}\sum_{j=1}^{n'}\right)\left(c - G_{i-1}\right)\left(c - G'_{j-1}\right)V_{ij}\right| \le \\ &\left(\left|\Sigma^2_N - \Sigma^2_n\right|\left|\Sigma'^2_{N'} - \Sigma'^2_{n'}\right|\right)^{1/2} + \left(\left|\Sigma^2_N - \Sigma^2_n\right|\Sigma'^2_{n'}\right)^{1/2} + \left(\Sigma^2_n\left|\Sigma'^2_{N'} - \Sigma'^2_{n'}\right|\right)^{1/2} = o_{\mathbb{P}}\left(\sigma_n\sigma'_{n'}\right)\end{aligned} \tag{5.10}$$

because $\Sigma_N \sim_{\mathbb{P}} \sigma_n$, with $\Sigma^2_N - \Sigma^2_n = o_{\mathbb{P}}\left(\sigma^2_n\right)$ from Eq (4.24) *et seq.*

To bound $\mathrm{A}_{NN'} - \mathrm{A}_{nn'}$ and $\mathrm{B}_{NN'} - \mathrm{B}_{nn'}$, we have $\sum_{j=i+1}^{n} \sum_{m=\pi(i)+1}^{n'} V_{jm} \le T_n - T_i$ and $\left| \sum_{i \in \mathfrak{C}_{NN'}} - \sum_{i \in \mathfrak{C}_{nn'}} \right| \le \left| \sum_{i=n+1}^{N} \right|$. The Cauchy-Schwartz inequality then yields

$$\left|\mathrm{A}_{NN'} - \mathrm{A}_{nn'}\right| \le \left|\Sigma_N^2 - \Sigma_n^2\right|^{1/2} \left|G_N - G_n\right|^{1/2} = o_{\mathbb{P}}\left(\sigma_n c^{1/2}\right) = o_{\mathbb{P}}\left(\sigma_n \sigma'_{n'}\right), \tag{5.11}$$

because the centering proof of $\Sigma_N \sim_{\mathbb{P}} \sigma_n$ shows $\Sigma_N^2 - \Sigma_n^2 = o_{\mathbb{P}}\left(\sigma_n^2\right)$. We also have $\left|\mathrm{B}_{NN'} - \mathrm{B}_{nn'}\right| \le \left|\Sigma_N^2 - \Sigma_n^2\right|^{1/2} \left|\Sigma_{N'}'^2 - \Sigma_{n'}'^2\right|^{1/2} = o_{\mathbb{P}}\left(\sigma_n \sigma'_{n'}\right)$.

Second, we show $\Gamma_{nn'} - \mathbb{E}\Gamma_{nn'} = o_{\mathbb{P}}\left(\sigma_n \sigma'_{n'}\right)$. For the first term $\Xi_{nn'}$ of $\dot{\Gamma}_{nn'}$, we prepare the comparison estimate again:

$$\begin{aligned} \operatorname{var}\left\{ \sum_{i=1}^{n} \sum_{j=1}^{n'} \left(c - G_{i-1}\right)\left(c - G'_{j-1}\right) V_{ij} \right\} &= \sum_{i=1}^{n} \sum_{j=1}^{n'} \mathbb{E}\left\{ \left(c - G_{i-1}\right)^2 \left(c - G'_{j-1}\right)^2 \right\} v_{ij} + \\ &\sum_{i=1}^{n} \sum_{i'=1}^{n'} \sum_{j=1}^{n} \sum_{j'=1}^{n'} \operatorname{cov}\left\{ \left(c - G_{i-1}\right)\left(c - G'_{i'-1}\right), \left(c - G_{j-1}\right)\left(c - G'_{j'-1}\right) \right\} v_{ii'} v_{jj'} \end{aligned} \tag{5.12}$$

We bound the covariance in Eq (5.12) with the same method used for bounding c_{ij} above. Write $G_{i-1} = g_{i-1} + \dot{G}_{i-1}$, etc., and expand the factors in the covariance. Each term (both plus and minus) resulting from the covariance in Eq (5.12) is therefore bounded by a finite number of terms, each of which is either c^2 or the product of c with either one or two factors from $\left(c - g_{i-1}\right)$, $\left(c - g'_{i'-1}\right)$, $\left(c - g_{j-1}\right)$, or $\left(c - g'_{j'-1}\right)$. After multiplication by $v_{ij} v_{i'j'}$ and summation, all terms are $o\left(\sigma_n^2 \sigma_{n'}'^2\right)$, because $\sum_{j=1}^{n'} v_{ij} \le u_i$, $\sum_{i=1}^{n} \left(c - g_{i-1}\right) u_i = o\left(\sigma_n^{3/2}\right)$, $\sum_{i=1}^{n} \sum_{j'=1}^{n'} v_{ij'} \le t_n = o\left(\sigma_n\right)$, and $\sum_{i=1}^{n} \sum_{j'=1}^{n'} \left(c - g_{i-1}\right)\left(c - g'_{j'-1}\right) v_{ij'} \le \sigma_n \sigma'_{n'}$, etc.

The double sum in Eq (5.12) is similarly bounded. Note

$$\mathbb{E}\left\{\left(c-G_{i-1}\right)^2\left(c-G'_{j-1}\right)^2\right\} = \operatorname{var}\left\{\left(c-G_{i-1}\right)\left(c-G'_{j-1}\right)\right\} + \left[\mathbb{E}\left\{\left(c-G_{i-1}\right)\left(c-G'_{j-1}\right)\right\}\right]^2, \quad (5.13)$$

where the variance can be bounded as a special case of the covariance in Eq (5.12). The expectation on the right of Eq (5.13) equals $\left(c-g_{i-1}\right)\left(c-g'_{j-1}\right)+\sum_{l\in\mathfrak{C}_{i-1,j-1}} f_l$. The new terms of degree four satisfy $\sum_{i=1}^{n}\sum_{j=1}^{n'}\left(c-g_{i-1}\right)^2\left(c-g'_{j-1}\right)^2 v_{ij} \le c^2\sigma_n^2 = o\left(\sigma_n^2\sigma_{n'}'^2\right)$.

Because the common value in Eq (5.12) is bounded by $o\left(\sigma_n^2\sigma_{n'}'^2\right)$, $\dot{\Xi}_{nn'} = o\left(\sigma_n\sigma'_{n'}\right)$.

For the second term $\dot{\mathrm{A}}_{nn'}$ of $\dot{\Gamma}_{nn'}$, the expectation $\alpha_{nn'} := \mathbb{E}\mathrm{A}_{nn'} \le t_n g_n = o\left(\sigma_n\sigma'_{n'}\right)$ yields $\dot{\mathrm{A}}_{nn'} = o_{\mathbb{P}}\left(\sigma_n\sigma'_{n'}\right)$.

For the third term $\dot{\mathrm{B}}_{nn'}$ of $\dot{\Gamma}_{nn'}$, the same techniques that applied to $Y_{nn'}$ above also show that $\operatorname{var}\dot{\mathrm{B}}_{nn'} = o\left(\sigma_n^2\sigma_{n'}'^2\right)$. Because $\dot{\mathrm{B}}_{nn'} = o_{\mathbb{P}}\left(\sigma_n\sigma'_{n'}\right)$, we have shown $\dot{\Gamma}_{nn'} = o_{\mathbb{P}}\left(\sigma_n\sigma'_{n'}\right)$.

Finally,

$$\mathbb{E}\Gamma_{nn'} - \gamma_{nn'} = \sum_{i=1}^{n}\sum_{j=1}^{n'}\operatorname{cov}\left(G_{i-1}, G'_{j-1}\right)v_{ij} + \sum_{i\in\mathfrak{C}_{nn'}}\operatorname{cov}\left(T_n - T_i, T'_{n'} - T'_{\pi(i)}\right) f_i \,. \quad (5.14)$$

A trite calculation shows that both terms actually equal $\alpha_{nn'} = o\left(\sigma_n\sigma'_{n'}\right)$. The proof of Theorem 2.3 is complete.

6. PROOFS OF THEOREM 2.4 AND THEOREM 2.5

To prove Theorem 2.4, this section first demonstrates the hypotheses of Theorem 2.2 with the sequence of bags replacing the sequence of urns, each asymptotic hypothesis being replaced by an equivalent or stronger hypothesis in probability (i.e., with $o(\bullet)$ or $O(\bullet)$ replaced by $o_{\mathbb{P}}(\bullet)$; or $\sim$, by $\sim_{\mathbb{P}}$). This demonstration suffices to prove Theorem 2.4, because we can then select a subsequence of $k = 1, 2, \ldots$, so that each asymptotic hypothesis in probability can be replaced by the corresponding a.s. hypothesis. Theorem 2.2 applied to the bootstrap distributions then shows that a.s., $\left(\widehat{R}^{*}_{\hat{N}^{*}} - \widehat{R}_{\hat{N}}\right) / \widehat{\Sigma}_{\hat{N}} \Rightarrow^{*} \Phi$ along this subsequence. We then extend the same method of proof to Theorem 2.5.

Now, we verify that the bags in Theorem 2.4 satisfy in probability the asymptotic hypotheses for the urns in Condition O-1 of Theorem 2.2. First, the controlling asymptotic parameters tend to ∞, as follows. Trivially, $c \to \infty$. Because $\mathbb{E}\widehat{T}_n = t_n$ and $\operatorname{var}\widehat{T}_n = O(t_n)$, we have $\widehat{T}_n \sim_{\mathbb{P}} t_n \to \infty$ and similarly, $\widehat{G}_n \sim_{\mathbb{P}} g_n \to \infty$. Theorem 2.2 yields $\widehat{\Sigma}_{\hat{N}} \to_{\mathbb{P}} \infty$, because it concludes that $\widehat{\Sigma}_{\hat{N}} \sim_{\mathbb{P}} \sigma_n$, so any hypotheses in probability can replace σ_n by $\widehat{\Sigma}_{\hat{N}}$.

Second, the hypotheses about $\overline{f}$ and $\overline{u}$ hold in probability for the corresponding bag variates $\overline{F} := \max_{i=1,\ldots,\bar{j}} \widehat{F}_i$ and $\overline{U} := \max_{i=1,\ldots,\bar{j}} \widehat{U}_i$. Recall that $\left\{\widehat{F}_i\right\}$ and $\left\{\widehat{U}_i\right\}$ are integral. A comparison of multinomial and Poisson sampling yields

$\mathbb{P}\left(\overline{F} \geq 1\right) > 1 - \exp\left(-\sum_{i=1}^{\overline{j}} f_i\right) \geq 1 - e^{-c} \to 1$, so $\overline{F} \log\log \overline{j} \to_{\mathbb{P}} \infty$. Similarly, $\overline{U} \log\log \overline{j} \to_{\mathbb{P}} \infty$.

Third, because of Proposition 4.3, with $\theta = \log \overline{j}$, each of the three hypotheses $t_n^2 \theta \overline{f} = o\left(\sigma_n^2\right)$, $\left(\log \overline{j}\right)\theta \overline{f} = O\left(c^{1/2}\right)$, and $c^2 \theta\, \overline{u} = o\left(\sigma_n^2\right)$ in Condition O-$\log \overline{j}$ of Theorem 2.4 yields an equivalent or stronger hypothesis in probability for Condition O-1 of Theorem 2.2. Fourth, Eq (4.1) *et seq.* show $\hat{G}_{\underline{j}}^{-1/2}\left(c - \hat{G}_{\underline{j}}\right) \to_{\mathbb{P}} \infty$ and $\hat{G}_{\overline{j}}^{-1/2}\left(\hat{G}_{\overline{j}} - c\right) \to_{\mathbb{P}} \infty$. Moreover, $t_{\overline{j}} + g_{\overline{j}} = o\left(D\right)$ implies $\hat{T}_{\overline{j}} + \hat{G}_{\overline{j}} = o_{\mathbb{P}}\left(D\right)$. Finally, $\underline{j} < n$, $\mathbb{P}\left(\hat{N} \leq \overline{j}\right) \to 1$, and $\mathbb{E}\left(\hat{T}_{\overline{j}} - \hat{T}_{\underline{j}}\right) = t_{\overline{j}} - t_{\underline{j}}$ in the expectation bound yield

$$c\left(\hat{T}_{\hat{N}} - \hat{T}_n\right)^2 = \overline{O}_{\mathbb{P}}\left\{c\left(\hat{T}_{\overline{j}} - \hat{T}_{\underline{j}}\right)^2\right\} = \overline{O}_{\mathbb{P}}\left\{c\left(t_{\overline{j}} - t_{\underline{j}}\right)^2\right\} = o_{\mathbb{P}}\left(\sigma_n^2\right) = o_{\mathbb{P}}\left(\hat{\Sigma}_{\hat{N}}^2\right). \qquad (6.1)$$

In the bags, $\underline{j}$ and $\overline{j}$ therefore have in probability the properties required by the hypotheses of Theorem 2.2, so Theorem 2.4 follows

The same type of proof applies to Theorem 2.5. Theorem 2.3 proves $\limsup \hat{\mathrm{P}}_{\hat{N}\hat{N}'} = \rho$ a.s.-$\mathbb{P}$ for some subsequence of $k = 1, 2, \ldots$. For this subsequence of bags, the hypothesis $\hat{\mathrm{P}} := \limsup \hat{\mathrm{P}}_{\hat{N}\hat{N}'} < 1$ a.s.-$\mathbb{P}^*$ holds. For a subsequence from this subsequence of bags, Theorem 2.5 then follows from Theorem 2.3 for the sequence of urns.

7. REFERENCES

[1] Agarwal, P. and States, D.J. (1998). Comparative accuracy of methods for protein sequence similarity search. *Bioinformatics.* **14**, 40-7.

[2] Barbour, A.D., Holst, L. and Janson, S. (1992). *Poisson Approximation*, Clarendon Press, Oxford.

[3] Bickel, P.J. and Freedman, D. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics.* **9**, 1196-1217.

[4] Bickel, P.J. and Freedman, D. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Annals of Statistics.* **12**, 470-482.

[5] Chernick, M.R. (1999). *Bootstrap Methods*, Wiley and Sons, New York.

[6] Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* **57**, Chapman and Hall, New York.

[7] Erdélyi, A. (1956). *Asymptotic Expansions*, Dover, New York.

[8] Green, D.M. and Swets, J.A. (1966). *Signal Detection Theory and Psychophysics*, Wiley, New York.

[9] Green, R.E. and Brenner, S.E. (2002). Bootstrapping and normalization for enhanced evaluations of pairwise sequence comparison. *Proceedings of the IEEE.* **90**, 1834-1847.

[10] Gribskov, M. and Robinson, N.L. (1996). Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Computers & Chemistry.* **20**, 25-33.

[11] Grundy, W.N. (1998). Homology detection via family pairwise search. *Journal of Computational Biology.* **5**, 479-491.

[12] Grundy, W.N. and Bailey, T.L. (1999). Family pairwise search with embedded motif models. *Bioinformatics.* **15**, 463-470.

[13] Grundy, W.N., Bailey, T.L., Elkan, C.P. and Baker, M.E. (1997). Meta-MEME: Motif-based hidden Markov models of protein families. *Computer Applications in the Biosciences.* **13**, 397-406.

[14] Halperin, E., Faigler, S. and Gill-More, R. (1999). FramePlus: aligning DNA to protein sequences. *Bioinformatics.* **15**, 867-873.

[15] Helland, I.S. (1982). Central Limit-Theorems for Martingales with Discrete or Continuous-Time. *Scandinavian Journal of Statistics.* **9**, 79-94.

[16] McCarn, D.B. and Lewis, C.M. (1990). A mathematical model of retrieval system performance. *Journal of the American Society for Information Science.* **41**, 495-500.

[17] Ross, S. (1996). *Stochastic Processes*, Wiley, New York.

[18] Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V. and Altschul, S.F. (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research.* **29**, 2994-3005.

[19] Spang, R., Rehmsmeier, M. and Stoye, J. (2002). A novel approach to remote homology detection: Jumping alignments. *Journal of Computational Biology.* **9**, 747-760.

[20] Swets, J.A. (1967). Effectiveness of information retrieval methods. *Federal Scientific and Technical Information*. Bolt, Beranek, and Newman, Inc, Cambridge, Massachusetts, 1-47.
[21] Swets, J.A. (1988). Measuring the Accuracy of Diagnostic Systems. *Science.* **240**, 1285-1293.
[22] Wilbur, W.J. (1994). Non-parametric significance tests of retrieval performance comparisons. *Journal of Information Science.* **20**, 270-284.
[23] Yu, Y.K., Bundschuh, R. and Hwa, T. (2002). Hybrid alignment: high-performance with universal statistics. *Bioinformatics.* **18**, 864-872.