

1 The Hájek Central Limit Theorem for Simple Random Sampling without Replacement

Suppose we are given a sequence of finite populations, on each of which is defined an \mathbb{R} -valued population characteristic. Suppose on each of the finite populations, we use SRSWOR (of fixed sample size) to select a sample, observe the values of the corresponding population characteristics on the selected elements, and use the Horvitz-Thompson estimator for the population mean. We seek to determine a necessary and sufficient condition for the (associated sequence of) “standardized deviations from the mean” of the Horvitz-Thompson estimator for population mean to converge in distribution to the standard Gaussian distribution.

Theorem 1.1 (The Hájek Central Limit Theorem for SRSWOR)

Suppose we have the following:

- Let $\{U_\nu\}_{\nu \in \mathbb{N}}$ be a sequence of finite populations, and $N_\nu = |U_\nu|$ be the population size of U_ν . Let the elements of U_ν be indexed by $1, 2, 3, \dots, N_\nu$.
- For each $\nu \in \mathbb{N}$, let $y^{(\nu)} : U_\nu \rightarrow \mathbb{R}$ be an \mathbb{R} -valued population characteristic. For each $i \in U_\nu$, let $y_i^{(\nu)}$ denote $y^{(\nu)}(i)$, the value of $y^{(\nu)}$ evaluated at the i^{th} element of U_ν .
- For each $\nu \in \mathbb{N}$, let $n_\nu \in \{1, 2, 3, \dots, N_\nu\}$ be given, and let \mathcal{S}_ν be the set of all n_ν -element subsets of U_ν . Let \mathcal{S}_ν be endowed with the uniform probability function, namely

$$P(s) = \frac{1}{\binom{N_\nu}{n_\nu}}, \text{ for each } s \in \mathcal{S}_\nu.$$

- For each $\nu \in \mathbb{N}$, let $\widehat{Y}_\nu : \mathcal{S}_\nu \rightarrow \mathbb{R}$ be the random variable defined as follows:

$$\widehat{Y}_\nu(s) := \frac{1}{n_\nu} \sum_{i \in s} y_i^{(\nu)}, \text{ for each } s \in \mathcal{S}_\nu$$

Let

$$\mu_\nu := E[\widehat{Y}_\nu] = \frac{1}{N_\nu} \sum_{i \in U_\nu} y_i^{(\nu)} \quad \text{and} \quad \sigma_\nu^2 := \text{Var}[\widehat{Y}_\nu] = \left(1 - \frac{n_\nu}{N_\nu}\right) \frac{S_\nu^2}{n_\nu},$$

where

$$S_\nu^2 := \frac{1}{N_\nu - 1} \sum_{i \in U_\nu} \left(y_i^{(\nu)} - \mu_\nu\right)^2$$

- For each $\nu \in \mathbb{N}$ and each $\delta > 0$ define:

$$U_\nu(\delta) := \left\{ i \in U_\nu \mid |y_i^{(\nu)} - \mu_\nu| > \delta \sqrt{\sigma_\nu^2} \right\} \subset U_\nu.$$

Suppose $n_\nu \rightarrow \infty$ and $N_\nu - n_\nu \rightarrow \infty$. Then,

$$\lim_{\nu \rightarrow \infty} P \left\{ s \in \mathcal{S}_\nu \mid \frac{\widehat{Y}_\nu(s) - \mu_\nu}{\sqrt{\sigma_\nu^2}} < x \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$$

if and only if

$$\lim_{\nu \rightarrow \infty} \frac{\sum_{i \in U_\nu(\delta)} \left(y_i^{(\nu)} - \mu_\nu \right)^2}{\sum_{i \in U_\nu} \left(y_i^{(\nu)} - \mu_\nu \right)^2} = 0, \text{ for every } \delta > 0.$$

Lemma 1.2

Bernoulli sampling from a finite population U of size N with individual selection probability n/N , where $n = 1, 2, \dots, N$, is equivalent to the following two-step sampling scheme:

- **Step 1:** Sample k from $\text{Binomial}(N, n/N)$.
- **Step 2:** Take an *SRSWOR* sample s of size k from U .

PROOF Note that the collection of possible samples for both schemes is the power set $\mathcal{P}(U)$ of U , i.e. all possible subsets of U . Let P_B and P_1 be the probability functions defined on $\mathcal{P}(U)$ under Bernoulli sampling and the two-step scheme, respectively. Then,

$$P_B(s) = \left(\frac{n}{N} \right)^k \left(1 - \frac{n}{N} \right)^{N-k}, \text{ for each } s \in \mathcal{P}(U), \text{ where } k = |s|.$$

On the other hand,

$$\begin{aligned} P_1(s) &= P(S = s \mid S \sim \text{SRSWOR}(k, N)) \cdot P(K = k \mid K \sim \text{Binomial}(N, n/N)) \\ &= \frac{1}{\binom{N}{k}} \cdot \binom{N}{k} \left(\frac{n}{N} \right)^k \left(1 - \frac{n}{N} \right)^{N-k} \\ &= \left(\frac{n}{N} \right)^k \left(1 - \frac{n}{N} \right)^{N-k}, \text{ for each } s \in \mathcal{P}(U), \text{ where } k = |s|. \end{aligned}$$

Thus, $P_B = P_1$ as (probability) functions on $\mathcal{P}(U)$. Hence, the two sampling schemes are equivalent. \square

Definition 1.3 (The Hájek Sampling Design)

Suppose U is a finite population of size $N \in \mathbb{N}$ with $N \geq 3$. Let $n \in \{2, \dots, N\}$ be fixed. Let $\mathcal{P}(U)$ be the power set of U . Let $\mathcal{S}(U, n)$ be the collection of all subsets of U with exactly n elements. The Hájek Sampling Design, by definition, selects an ordered pair of samples $(s^{(0)}, s^{(1)}) \in \mathcal{S}(U, n) \times \mathcal{P}(U)$ as follows:

- First, select $k \in \{0, 1, 2, \dots, N\}$ based on the binomial distribution $\text{Binom}(N, n/N)$.

More precisely, let $K \sim \text{Binomial}(N, n/N)$, i.e. let K be a random variable following the binomial distribution with number of trials N and probability of success n/N . In other words,

$$P(K = k) = \binom{N}{k} \cdot \left(\frac{n}{N} \right)^k \cdot \left(1 - \frac{n}{N} \right)^{N-k}, \text{ for each } k = 0, 1, 2, \dots, N.$$

Let $k \in \{0, 1, 2, \dots, N\}$ be a realization of the random variable $K \sim \text{Binomial}(N, n/N)$.

- If $k = n$, take an *SRSWOR* sample $s^{(0)} \subset U$ of size n , and let $s^{(1)} = s^{(0)}$.
- If $k > n$, take an *SRSWOR* sample $s^{(1)} \subset U$ of size k . Then, select an *SRSWOR* sample $s^{(0)}$ of $s^{(1)}$ of size n .
- If $k < n$, take an *SRSWOR* sample $s^{(0)} \subset U$ of size n . Then, select an *SRSWOR* sample $s^{(1)}$ of $s^{(0)}$ of size k .

Remark 1.4

Note that the Hájek Sampling Design defines implicitly a probability function P_H on $\mathcal{S}(U, n) \times \mathcal{P}(U)$, making it a finite probability space. More explicitly, for each $(s^{(0)}, s^{(1)}) \in \mathcal{S}(U, n) \times \mathcal{P}(U)$, writing $k = |s^{(1)}|$, we have

$$P_H(s^{(0)}, s^{(1)}) = \begin{cases} \binom{N}{n} \left(\frac{n}{N}\right)^n \left(1 - \frac{n}{N}\right)^{N-n} \cdot \frac{1}{\binom{N}{n}}, & \text{if } s^{(0)} = s^{(1)} \\ \binom{N}{k} \left(\frac{n}{N}\right)^k \left(1 - \frac{n}{N}\right)^{N-k} \cdot \frac{1}{\binom{N}{k}} \cdot \frac{1}{\binom{k}{n}}, & \text{if } s^{(0)} \subsetneq s^{(1)} \\ \binom{N}{k} \left(\frac{n}{N}\right)^k \left(1 - \frac{n}{N}\right)^{N-k} \cdot \frac{1}{\binom{N}{n}} \cdot \frac{1}{\binom{n}{k}}, & \text{if } s^{(0)} \supsetneq s^{(1)} \\ 0, & \text{otherwise} \end{cases}$$

Lemma 1.5 (Properties of the Hájek Sampling Design)

Suppose U is a finite population of size $N \in \mathbb{N}$ with $N \geq 3$. Let $n \in \{2, \dots, N\}$ be fixed. Let $P_H : \mathcal{S}(U, n) \times \mathcal{P}(U) \rightarrow [0, 1]$ be the Hájek Sampling Design. Then, the following statements are true:

1. The marginal sampling design induced on $\mathcal{S}(U, n)$ by P_H is SRSWOR(U, n).
2. The marginal sampling design induced on $\mathcal{P}(U)$ by P_H is Bernoulli Sampling from U with unit selection probability n/N .
3. For each fixed $k \in \{n+1, n+2, \dots, N\}$, the sampling design induced on $\mathcal{S}(U, k-n)$ by pushing forward the conditional sampling design of $P_H|_{|s^{(1)}|=k}$ via the following map:

$$\left\{ (s^{(0)}, s^{(1)}) \in \mathcal{S}(U, n) \times \mathcal{P}(U) \mid |s^{(1)}| = k \right\} \longrightarrow \mathcal{S}(U, k-n) : (s^{(0)}, s^{(1)}) \longmapsto s^{(1)} \setminus s^{(0)}$$

is equivalent to SRSWOR($U, k-n$).

4. For each fixed $k \in \{0, 1, 2, \dots, n-1\}$, the sampling design induced on $\mathcal{S}(U, n-k)$ by pushing forward the pertinent restriction of P_H via the following map:

$$\left\{ (s^{(0)}, s^{(1)}) \in \mathcal{S}(U, n) \times \mathcal{P}(U) \mid |s^{(1)}| = k \right\} \longrightarrow \mathcal{S}(U, n-k) : (s^{(0)}, s^{(1)}) \longmapsto s^{(0)} \setminus s^{(1)}$$

is equivalent to SRSWOR($U, n-k$).

PROOF

1. For each $s^{(0)} \in \mathcal{S}(U, n)$, it suffices to show that the marginal probability $P_H(s^{(0)}, \cdot)$ is given by:

$$P_H(s^{(0)}, \cdot) = \frac{1}{\binom{N}{n}}$$

To this end,

$$\begin{aligned}
 P_H(s^{(0)}, \cdot) &= \sum_{s^{(1)}=s^{(0)}} P_H(s^{(0)}, s^{(1)}) + \sum_{s^{(1)} \supsetneq s^{(0)}} P_H(s^{(0)}, s^{(1)}) + \sum_{s^{(1)} \subsetneq s^{(0)}} P_H(s^{(0)}, s^{(1)}) \\
 &= \binom{N}{n} \left(\frac{n}{N}\right)^n \left(1 - \frac{n}{N}\right)^{N-n} \cdot \frac{1}{\binom{N}{n}} \\
 &\quad + \sum_{k=n+1}^N \binom{N}{k} \left(\frac{n}{N}\right)^k \left(1 - \frac{n}{N}\right)^{N-k} \cdot \frac{1}{\binom{N}{k}} \cdot \frac{1}{\binom{k}{n}} \cdot \binom{N-n}{k-n} \\
 &\quad + \sum_{k=0}^{n-1} \binom{N}{k} \left(\frac{n}{N}\right)^k \left(1 - \frac{n}{N}\right)^{N-k} \cdot \frac{1}{\binom{N}{k}} \cdot \frac{1}{\binom{n}{k}} \cdot \binom{n}{k}
 \end{aligned}$$

We remark that, for a given $s^{(0)} \in \mathcal{S}(U, n)$ and $k > n$, the quantity $\binom{N-n}{k-n}$ is the number of elements in $\mathcal{P}(U)$ (i.e. number of subsets of U) of size k containing $s^{(0)}$ as a proper subset. Note also that, for $k > n$,

$$\frac{1}{\binom{N}{k}} \cdot \frac{1}{\binom{k}{n}} \cdot \binom{N-n}{k-n} = \frac{k!(N-k)!}{N!} \cdot \frac{n!(k-n)!}{k!} \cdot \frac{(N-n)!}{(k-n)!(N-k)!} = \frac{n!(N-n)!}{N!} = \frac{1}{\binom{N}{n}}.$$

Hence, we have

$$P_H(s^{(0)}, \cdot) = \frac{1}{\binom{N}{n}} \cdot \sum_{k=0}^N \binom{N}{k} \left(\frac{n}{N}\right)^k \left(1 - \frac{n}{N}\right)^{N-k} = \frac{1}{\binom{N}{n}} \cdot 1 = \frac{1}{\binom{N}{n}}$$

2. For each $s^{(1)} \in \mathcal{P}(U)$, it suffices to show that the marginal probability $P_H(\cdot, s^{(1)})$ is given by:

$$P_H(\cdot, s^{(1)}) = \left(\frac{n}{N}\right)^k \cdot \left(1 - \frac{n}{N}\right)^{N-k}, \quad \text{where } k = |s^{(1)}|.$$

To this end, first note that either $k = |s^{(1)}| \geq n$ holds, or $k = |s^{(1)}| < n$ holds. In the first case, i.e. $k = |s^{(1)}| \geq n$, we have

$$\begin{aligned}
 P_H(\cdot, s^{(1)}) &= P(S^{(1)} = s^{(1)} \mid K = k) \cdot P(K = k) \\
 &= \frac{1}{\binom{N}{k}} \cdot \binom{N}{k} \left(\frac{n}{N}\right)^k \left(1 - \frac{n}{N}\right)^{N-k} \\
 &= \left(\frac{n}{N}\right)^k \cdot \left(1 - \frac{n}{N}\right)^{N-k}.
 \end{aligned}$$

In the second case, i.e. $k = |s^{(1)}| < n$, we have

$$\begin{aligned}
 P_H(\cdot, s^{(1)}) &= \sum_{s^{(0)} \supsetneq s^{(1)}} P_H(s^{(0)}, s^{(1)}) = \sum_{s^{(0)} \supsetneq s^{(1)}} \binom{N}{k} \left(\frac{n}{N}\right)^k \left(1 - \frac{n}{N}\right)^{N-k} \cdot \frac{1}{\binom{N}{n}} \cdot \frac{1}{\binom{n}{k}} \\
 &= \binom{N-k}{n-k} \cdot \binom{N}{k} \left(\frac{n}{N}\right)^k \left(1 - \frac{n}{N}\right)^{N-k} \cdot \frac{1}{\binom{N}{n}} \cdot \frac{1}{\binom{n}{k}} \\
 &= \binom{N}{k} \left(\frac{n}{N}\right)^k \left(1 - \frac{n}{N}\right)^{N-k} \cdot \frac{(N-k)!}{(n-k)!(N-n)!} \cdot \frac{n!(N-n)!}{N!} \cdot \frac{k!(n-k)!}{n!} \\
 &= \binom{N}{k} \left(\frac{n}{N}\right)^k \left(1 - \frac{n}{N}\right)^{N-k} \cdot \frac{k!(N-k)!}{N!} = \binom{N}{k} \left(\frac{n}{N}\right)^k \left(1 - \frac{n}{N}\right)^{N-k} \cdot \frac{1}{\binom{N}{k}} \\
 &= \left(\frac{n}{N}\right)^k \cdot \left(1 - \frac{n}{N}\right)^{N-k}
 \end{aligned}$$

We remark that, for a given $s^{(1)} \in \mathcal{P}(U)$ with $|s^{(1)}| = k < n$, the quantity $\binom{N-k}{n-k}$ is the number of elements in $\mathcal{S}(U, n)$ containing $s^{(1)}$ as a proper subset.

3. Let $\tilde{P} : \mathcal{S}(U, k-n)$ be the induced sampling design on $\mathcal{S}(U, k-n)$. Then, for each $s^{(2)} \in \mathcal{S}(U, k-n)$, we have

$$\begin{aligned}
 \tilde{P}(s^{(2)}) &= \sum_{s^{(1)} \setminus s^{(0)} = s^{(2)}} P_H(s^{(0)}, s^{(1)} \mid K=k) = \sum_{s^{(1)} \setminus s^{(0)} = s^{(2)}} \frac{1}{\binom{N}{k}} \cdot \frac{1}{\binom{k}{n}} \\
 &= \binom{N-k+n}{n} \cdot \frac{1}{\binom{N}{k}} \cdot \frac{1}{\binom{k}{n}} = \frac{(N-k+n)!}{n!(N-k)!} \cdot \frac{k!(N-k)!}{N!} \cdot \frac{n!(k-n)!}{k!} \\
 &= \frac{(k-n)!(N-k+n)!}{N!} = 1 / \binom{N}{k-n}
 \end{aligned}$$

This proves that \tilde{P} is indeed equivalent to $\text{SRSWOR}(U, k-n)$.

4. Let $P' : \mathcal{S}(U, n-k)$ be the induced sampling design on $\mathcal{S}(U, n-k)$. Then, for each $s^{(2)} \in \mathcal{S}(U, n-k)$, we have

$$\begin{aligned}
 P'(s^{(2)}) &= \sum_{s^{(0)} \setminus s^{(1)} = s^{(2)}} P_H(s^{(0)}, s^{(1)} \mid K=k) = \sum_{s^{(0)} \setminus s^{(1)} = s^{(2)}} \frac{1}{\binom{N}{n}} \cdot \frac{1}{\binom{n}{k}} \\
 &= \binom{N-n+k}{k} \cdot \frac{1}{\binom{N}{n}} \cdot \frac{1}{\binom{n}{k}} = \frac{(N-n+k)!}{k!(N-n)!} \cdot \frac{n!(N-n)!}{N!} \cdot \frac{k!(n-k)!}{n!} \\
 &= \frac{(n-k)!(N-n+k)!}{N!} = 1 / \binom{N}{n-k}
 \end{aligned}$$

This proves that P' is indeed equivalent to $\text{SRSWOR}(U, n-k)$.

The proof of this Lemma is complete. □

Theorem 1.6 (Håjek's Fundamental Lemma)

Suppose U is a finite population of size $N \in \mathbb{N}$ with $N \geq 3$, and $y : U \rightarrow \mathbb{R}$ is a population characteristic. Let $n \in \{2, \dots, N\}$ be fixed. Let $\bar{y}_U := \frac{1}{N} \sum_{i \in U} y_i$. Let $\mathcal{S}(U, n) \times \mathcal{P}(U)$ be endowed with the probability function P_H defined

by the Hájek Sampling Design. Define the \mathbb{R}^2 -valued random variable $X = (X^{(0)}, X^{(1)}) : \mathcal{S}(U, n) \times \mathcal{P}(U) \rightarrow \mathbb{R}^2$ as follows: For any $(s^{(0)}, s^{(1)}) \in \mathcal{S}(U, n) \times \mathcal{P}(U)$,

$$X^{(0)}(s^{(0)}) := \frac{1}{n} \sum_{i \in s^{(0)}} (y_i - \bar{y}_U), \quad \text{and} \quad X^{(1)}(s^{(1)}) := \frac{1}{n} \sum_{i \in s^{(1)}} (y_i - \bar{y}_U).$$

Then,

$$E \left[\left(\frac{X^{(0)}}{\sqrt{\text{Var}[X^{(1)}]}} - \frac{X^{(1)}}{\sqrt{\text{Var}[X^{(1)}]}} \right)^2 \right] = \frac{E[(X^{(0)} - X^{(1)})^2]}{\text{Var}[X^{(1)}]} \leq \sqrt{\frac{1}{n} + \frac{1}{N-n}}$$

References