cf. Non-random sampling

Some probabilities in s.r.s. scheme.

- Random variables $I_1, I_2, \ldots, I_n$: Let I_k, $k = 1, \ldots, n$, be the integer label on the kth member drawn from the population $\in \{1, 2, \ldots, N\}$ (sampling)

  $(\square, \square, \ldots, \square)$ = 1st 2nd nth; I_k kth

- s.r.s. with replacement
  - marginal distribution of I_k: $P(I_k = i_k) = 1/N$, for $i_k = 1, \ldots, N$.

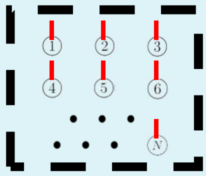
    Same marginal dist.
  - conditional distribution: (can be any k positions)

    $$P\left(I_k = i_k \,\middle|\, I_1 = i_1, \ldots, I_{k-1} = i_{k-1}\right) = 1/N = P(I_k = i_k)$$

    $\Rightarrow$ It can be proved that $I_1, \ldots, I_n$ are independent.
  - joint distribution of $I_1, I_2, \ldots, I_n$:

    $$P(I_1 = i_1, \ldots, I_n = i_n) = \prod_{k=1}^n P(I_k = i_k) = 1/N^n$$

    (Δ) in LNp.8



- s.r.s. without replacement
  - marginal distribution of I_k: $P(I_k = i_k) = 1/N$, for $i_k = 1, \ldots, N$.
  - conditional distribution: for distinct i_k's, (cf.) (can be any k positions)

    $$P\left(I_k = i_k \,\middle|\, I_1 = i_1, \ldots, I_{k-1} = i_{k-1}\right) = 1/(N-k+1)$$

    $\Rightarrow I_1, \ldots, I_n$ are not independent. $\xleftrightarrow{\text{cf.}}$ with repl.
  - joint distribution of $I_1, I_2, \ldots, I_n$: for distinct i_k's,

    $P(I_1 = i_1, \ldots, I_n = i_n)$ → (by multiplication law)

    $$\begin{aligned} &= P(I_1 = i_1)P(I_2 = i_2 | I_1 = i_1) \cdots P(I_n = i_n | I_1 = i_1, \cdots, I_{n-1} = i_{n-1}) \\ &= (1/N)(1/(N-1)) \cdots (1/(N-n+1)) = (N-n)!/N! \end{aligned}$$

    ← ($\square$) in LNp.8

  $(\square, \square, \ldots, \square, \ldots, \square)$ = 1st 2nd kth nth

  $$P(I_k = i_k) = \frac{\frac{(N-1)!}{(N-n)!}}{\frac{N!}{(N-n)!}} = \frac{1}{N}$$

  $(\square, \square, \ldots, \square)$ = 1st 2nd nth; I_k kth

Most s.r.s. are s.r.s. without replacement. Q: Why bother studying s.r.s. with repl.?

  - Similary, the joint distribution of I_k and I_l, $1 \le k < l \le n$, is

    $$P(I_k = i_k, I_l = i_l) = P(I_k = i_k)P(I_l = i_l | I_k = i_k) = \frac{1}{N(N-1)}$$

    if $i_k \neq i_l$, and zero if $i_k = i_l$.

Note 3 (Some notes about s.r.s.)

similar probabilities (Why?)

- When $n \ll N$, s.r.s. with replacement $\approx$ s.r.s. without replacement
  - **Recall.** In dichotomous case, when $n \ll N$, (black & white balls in an urn)

    (with repl.) → binomial distribution $\approx$ hypergeometric distribution ← (without repl.)

  probability related to s.r.s with repl. is easier to calculate

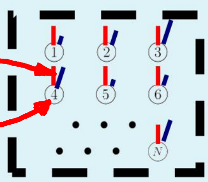
- The actual composition of an s.r.s. is usually determined by using a table of random numbers or a random number generator on a computer.

Statistical modeling of data collected from an s.r.s. of size n.

joint dist. of data

- Data $X_1, X_2, \ldots, X_n$. Let X_k, $k = 1, \ldots, n$, be the quantity of interest observed on the kth member in the sample. We have

  $$\boxed{f: i \to x_i} \qquad f(\underbrace{I_k}_{\text{r.v.}}) = X_k = x_{I_k},$$



  x_i, i

  and $X_1, \ldots, X_n$ are random variables.

  $(\square, \square, \ldots, \square)$ = 1st 2nd nth; X_k kth

- **Recall.** The population distribution F_0 (defined in LNp.3) assigns probability n_j/N on ζ_j for $j = 1, \ldots, m$. (**Note.** F_0 is unknown in a sampling survey)

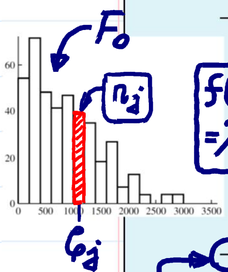
- Statistical modeling of $X_1, \ldots, X_n$ under s.r.s. with replacement
  - marginal distribution of X_k:
    - X_k can take values only on $\zeta_1, \ldots, \zeta_m$, and
    - $P\left(X_k = \zeta_j\right) = P\left(I_k \in \{i_k | x_{i_k} = \zeta_j\}\right) = n_j/N$, $j = 1, \ldots, m$.

    $f(I_k) = \chi_{I_k}$

    That is, $X_k \sim F_0$, $k = 1, \ldots, n$. → $X_1, \ldots, X_n$ have same marginal dist.

  - Because $I_1, \ldots, I_n$ are independent, $X_1, \ldots, X_n$ are independent.

    $\because X_1 = f(I_1)$, $X_2 = f(I_2)$, $\vdots$, $X_n = f(I_n)$
  - joint distribution of $X_1, \ldots, X_n$: $X_1, \ldots, X_n$ are **independent and identically distributed** (**i.i.d.**) random variables with the distribution F_0, denoted by $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F_0$.

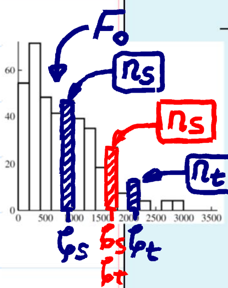


cf.

- Statistical modeling of $X_1, \ldots, X_n$ under s.r.s. without replacement
  - marginal distribution of X_k: $X_k \sim F_0$, $k = 1, \ldots, n$ (same marginal distribution as in the with-replacement case)

    $\because X_k = f(I_k) = \chi_{I_k}$ & check LNp.9
  - $X_1, \ldots, X_n$ are not independent. $\overset{\text{cf.}}{\longleftrightarrow}$ with repl.
  - joint distribution of X_k and X_l, $1 \le k < l \le n$:

$$P(X_k = \zeta_s, X_l = \zeta_t) = P(X_k = \zeta_s) P(X_l = \zeta_t | X_k = \zeta_s)$$
$$= P((I_k, I_l) \in \{(i_k, i_l) \,|\, x_{i_k} = \zeta_s,\ x_{i_l} = \zeta_t,\ i_k \neq i_l\})$$
$$= \begin{cases} \frac{n_s}{N} \times \frac{n_t}{N-1} = \frac{n_s n_t}{N(N-1)}, & \text{if } \zeta_s \neq \zeta_t \text{ (i.e., } s \neq t), \\ \frac{n_s}{N} \times \frac{n_s - 1}{N-1} = \frac{n_s(n_s-1)}{N(N-1)}, & \text{if } \zeta_s = \zeta_t \text{ (i.e., } s = t). \end{cases}$$



- joint distribution of $X_1, \ldots, X_n$ is more complicated, but its derivation follows the same rule.

  e.g. 3 r.v.'s $P(X_k = \zeta_s, X_l = \zeta_t, X_g = \zeta_u)$
  - $\zeta_s \neq \zeta_t \neq \zeta_u$
  - $\zeta_s = \zeta_t \neq \zeta_u$
  - $\zeta_s \neq \zeta_t = \zeta_u$
  - $\zeta_s = \zeta_u \neq \zeta_t$
  - $\zeta_s = \zeta_t = \zeta_u$

• Estimation of population mean (and population total)

- population mean: mean μ of F_0 (unknown parameter)
- data: $X_1, \ldots, X_n$ (random variables) with distribution related to F_0
- estimation of population mean: use (a function of) the data to estimate μ

Definition 7 (statistic, sampling distribution, estimator, estimate, standard error)

- 統計量 A **statistic** is a function of data only, not involving any unknown parameter. Any statistic is a random variable. ← $\because$ data are random
- 估計式 An **estimator** $\hat{\theta}$ of a parameter θ is a statistic used to estimate θ, and 估計值 an **estimate** is an observed value (an observation, a realization) of $\hat{\theta}$ computed based on a specific sample.
- The distribution of $\hat{\theta}$ is called **sampling distribution**, denoted by $F_{\hat{\theta}}$. ← pdf, pmf, cdf, mgf, ch.f
- standard deviation of $\hat{\theta}$ The **standard error** (st.e.) of an estimator $\hat{\theta}$ is the squared root of the variance of $\hat{\theta}$, i.e., $\sqrt{Var_\theta(\hat{\theta})}$. ← a function of F_0, involving parameters

  smaller variance, often more accurate estimator
- An estimate of the standard error of $\hat{\theta}$ is called an **estimated standard error** of $\hat{\theta}$.