Survey Sampling

- **What is survey sampling?** (cf. census survey)

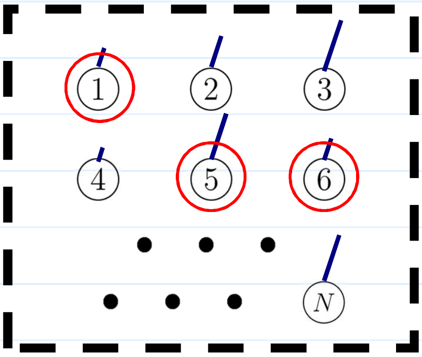
  - understanding the whole by a fraction (i.e., a sample)

population

Q: What is the population to survey?

quantity of interest x_i, i=1, …, N.

- x_i can be numerical or categorical
- Multivariate $(x_{i1}, x_{i2}, \ldots, x_{ik})$ i=1, …, N.



N: population size

a sample of size n: a subgroup of n members ($n < N$)

Q: Which n members should be included in the sample? i.e., how to produce a representative sample?

a unit/member (each labeled by an integer)

Definition 1 (survey sampling)

A technique to obtain information about a large population by examining only a small fraction of that population.

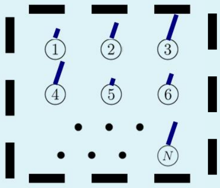
Example 1 (Applications of survey sampling)

- Governments conduct health survey of human populations.
- In agriculture, estimate total acreage of wheat in a state by surveying a sample of farms.
- Sample records of shipments of hoursehold goods by motor carriers to evaluate the accuracy of preshipment estimates of charges, claims for damages, and other variables.
- To control quality, the output of a manufacturing process may be sampled in order to examine the items for defects.
- During audits of the financial records of large companies, sampling techniques may be used when examination of the entire set of records is impractical.

• Population and population parameters

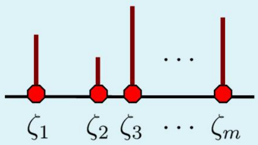
Formulation and some notations.

- N : size of the population (assumed **known**)
- $x_1, x_2, \ldots, x_N$: values associated with members of the population (x_i: value of the member labelled by i)
  - example of numerical x: age or weight
  - example of categorical x: values 1 and 0 denote the presence and absence, respectively, of some characteristic such as gender



(**Note**. $x_1, x_2, \ldots, x_N$ may not be distinct values)

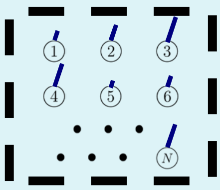
- Suppose that there are m distinct values in $x_1, x_2, \ldots, x_N$. Denote these distinct values by ζ_1, ζ_2, ..., ζ_m (WLOG, assume $\zeta_1 < \zeta_2 < \cdots < \zeta_m$).
- Denote the number of population members that have the value ζ_j by n_j, $j = 1, \ldots, m$.



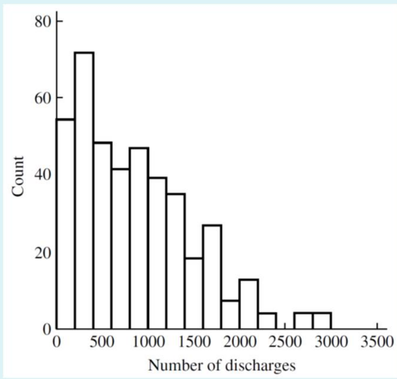
- The proportion of population members with value ζ_j is n_j/N (cf., a distribution of a random variable).
- Let F_0 be the distribution that assigns probability n_j/N on ζ_j, for $j = 1, \ldots, m$, called **population distribution**. (**Note**. F_0 can be known only in a census survey. It is unknown in a sample survey)

Example 2 (A survey of hospital discharge, [Herson, 1976])

- population: $N = 393$ short-stay hospitals
- x_i = the number of patients discharged from the ith hospital during January 1968



- A histogram of these $x_1, x_2, \ldots, x_N$ is given in Figure 7.1 of textbook. (**Note**. $x_1, x_2, \ldots, x_N$ are known here)
- A histogram is a convenient graphical representation of (the pmf of) F_0.
  - For example, this histogram indicates about 40 (40/393≈10%) hospitals discharged from 601 to 800 patients.



Definition 2 (population parameter)

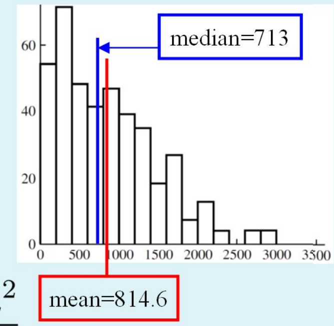
A population parameter θ is a value that describes some numerical characteristic of the population distribution F_0 (e.g., θ = mean or variance of F_0). When F_0 is *unknown*, the parameter θ is a fixed but unknown value.

θ can be estimated from a sample of data

Definition 3 (Some population parameters that are often of interest)

- population mean (mean of F_0): $\mu = \left(\sum_{i=1}^{N} x_i\right)/N = \left(\sum_{j=1}^{m} n_j \zeta_j\right)/N$ $= \sum_{j=1}^{m} (n_j/N)\,\zeta_j$
- population total: $\tau = \sum_{i=1}^{N} x_i = \sum_{j=1}^{m} n_j \zeta_j = N\mu$
- population variance (variance of F_0):

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 = \sum_{j=1}^{m}\frac{n_j}{N}(\zeta_j-\mu)^2$$



– a useful identity:

$$\sigma^2 = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - 2\mu\sum_{i=1}^{N} x_i + N\mu^2\right)$$

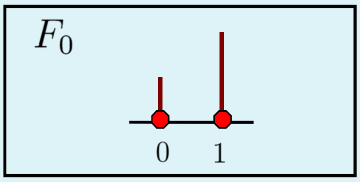
$$= \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - 2N\mu^2 + N\mu^2\right) = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2\right) - \mu^2$$

(**Note**. $X \sim F_0$, $Var(X) = E(X^2) - [E(X)]^2$)

- population standard deviation (st.d. of F_0): $\sigma = \sqrt{\sigma^2}$
  - σ is a measure of how spread out, dispersed, or scattered the x_i's are (**Note**. σ^2 also measures the spread of x_i's)
  - σ is given in the same units as are the x_i's and mean μ (cf., σ^2 is expressed in unit squared)

Note 1 (Some notes about dichotomous x_i's)

- In the **dichotomous** case, x_i takes on the value 1 or 0 to denote the presence or absence of some characteristic. In this case,
  - F_0 is Bernoulli(p) distribution, where p is the proportion of members in the population having the particular characteristic
  - $\mu = p$, since each x_i is either 0 or 1
  - $\sigma^2 = p(1-p)$, since



$$\sigma^2 = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2\right) - \mu^2 = \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right) - p^2 = p - p^2.$$

❖ **Reading**: textbook, 7.1, 7.2

- Simple random sampling

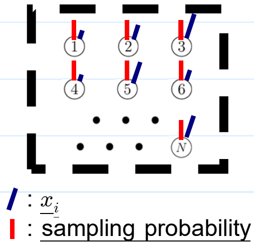
Q: Which members should be included in a sample, i.e., how to choose a representative sample?

Definition 4 (random sampling)

Each member of the population has a specified probability of being included in the sample, and the actual composition of the sample is random.

(cf.) Non-random sampling schemes: particular population members are included in the sample because the investigator thinks they are typical in some way. Such a scheme

- may be effective in some situations, but
- cannot (1) guarantee unbiasedness, or
  (2) estimate the magnitude of any error committed



Definition 5 (estimator, estimation error, bias, unbiased estimator, mean squared error)

Consider a parameter θ and an *estimator* $\hat{\theta}$ of θ, where θ is a fixed but unknown value, and $\hat{\theta}$, a function of sampled data, is a random variable (r.v.).

- estimation **error**: $\hat{\theta} - \theta$
- **bias** of an estimator $\hat{\theta}$: $E(\hat{\theta}) - \theta = E(\hat{\theta} - \theta)$
- $\hat{\theta}$ is an **unbiased** estimator if $E(\hat{\theta}) = \theta$, i.e., $E(\hat{\theta} - \theta) = 0$
- **mean squared error** (MSE): $E(\hat{\theta} - \theta)^2$ $(= Var(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2)$

  MSE = variance + bias^2

Note 2 (Advantages of random sampling)

- Unbiasedness of estimators can be guaranteed
- Probabilistic bounds on errors can be calculated.
- Other advantages:
  - guard against investigator biases
  - a small sample costs less than a complete enumeration

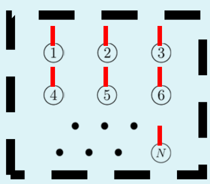
  - quality of a small data can be more easily monitored and controlled
  - random sampling makes possible the calculation of an estimate of the error due to sampling
  - it is possible to determine the sample size n necessary to obtain a prescribed error level

Definition 6 (simple random sampling, s.r.s)

- For a population of size N, each particular sample of size n $(n < N)$ has the *same* probability of occurrence.

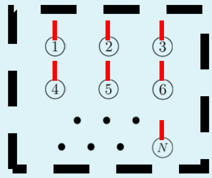
  (**Recall**. random process of drawing balls sequentially from an urn)
- Two versions of s.r.s.:
  - with replacement: duplicate members are allowed
    - number of all possible samples: N^n
    - occurence probability of each sample: $1/N^n$
  - without replacement: no duplication is allowed
    - number of all possible samples: $\binom{N}{n} n! = \frac{N!}{(N-n)!}$
    - occurence probability of each sample: $1/(\binom{N}{n} n!) = \frac{(N-n)!}{N!}$

  (**Note**. When the $n!$ permutations of a specific set of n members are considered identical, we say there are $\binom{N}{n}$ possible samples)



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
- s.r.s. with replacement
  - marginal distribution of I_k: $P(I_k = i_k) = 1/N$, for $i_k = 1, \ldots, N$.
  - conditional distribution:
  $$P\left(I_k = i_k \,\middle|\, I_1 = i_1, \ldots, I_{k-1} = i_{k-1}\right) = 1/N = P(I_k = i_k)$$
  $\Rightarrow$ It can be proved that $I_1, \ldots, I_n$ are independent.
  - joint distribution of $I_1, I_2, \ldots, I_n$:
  $$P(I_1 = i_1, \ldots, I_n = i_n) = \prod_{k=1}^{n} P(I_k = i_k) = 1/N^n$$



- 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,
  $$P\left(I_k = i_k \,\middle|\, I_1 = i_1, \ldots, I_{k-1} = i_{k-1}\right) = 1/(N - k + 1)$$
  $\Rightarrow$ $I_1, \ldots, I_n$ are not independent.
  - joint distribution of $I_1, I_2, \ldots, I_n$: for distinct i_k's,
  $$\begin{aligned} P(I_1 = i_1, \ldots, I_n = i_n) \\ &= P(I_1 = i_1)P(I_2 = i_2|I_1 = i_1) \cdots P(I_n = i_n|I_1 = i_1, \cdots, I_{n-1} = i_{n-1}) \\ &= (1/N)(1/(N-1)) \cdots (1/(N-n+1)) = (N-n)!/N! \end{aligned}$$

  - 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 $i_k \ne i_l$, and zero if $i_k = i_l$.

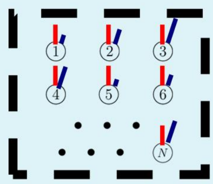
Note 3 (Some notes about s.r.s.)

- When $n \ll N$, s.r.s. with replacement $\approx$ s.r.s. without replacement
  - **Recall**. In dichotomous case, when $n \ll N$,

    binomial distribution $\approx$ hypergeometric distribution
- 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.

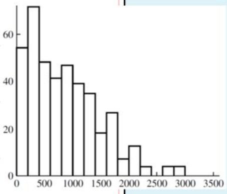
- Data $X_1, X_2, \ldots, X_n$. Let X_k, $k = 1, \ldots, n$, be the quantity of interest observed on the kth member in the sample. We have
$$X_k = x_{I_k},$$
and $X_1, \ldots, X_n$ are random variables.



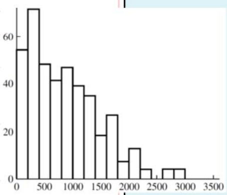
- **Recall**. The population distribution F_0 assigns probability n_j/N on ζ_j for $j = 1, \ldots, m$. (**Note**. F_0 is unknown in a sampling survey)

- 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
    - $P\left(X_k = \zeta_j\right) = P\left(I_k \in \{i_k | x_{i_k} = \zeta_j\}\right) = n_j/N$, $j = 1, \ldots, m$.

    That is, $X_k \sim F_0$, $k = 1, \ldots, n$.



  - Because $I_1, \ldots, I_n$ are independent, $X_1, \ldots, X_n$ are independent.
  - 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 $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F_0$.
- 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)
  - $X_1, \ldots, X_n$ are not independent.
  - joint distribution of X_k and X_l, $1 \le k < l \le n$:



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

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

• 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)

- A **statistic** is a function of data only, not involving any unknown parameter. Any statistic is a random variable.
- An **estimator** $\hat{\theta}$ of a parameter θ is a statistic used to estimate θ, and 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 $F_{\hat{\theta}}$.
- The **standard error** (st.e.) of an estimator $\hat{\theta}$ is the squared root of the variance of $\hat{\theta}$, i.e., $\sqrt{Var_\theta(\hat{\theta})}$.
- An estimate of the standard error of $\hat{\theta}$ is called an **estimated standard error** of $\hat{\theta}$.

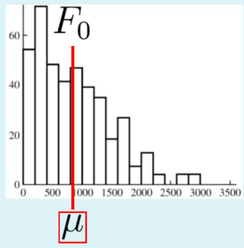
Definition 8 (sample mean)

The **sample mean** of $X_1, X_2, \ldots, X_n$ is $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$.

Note 4 (Some notes about sample mean)

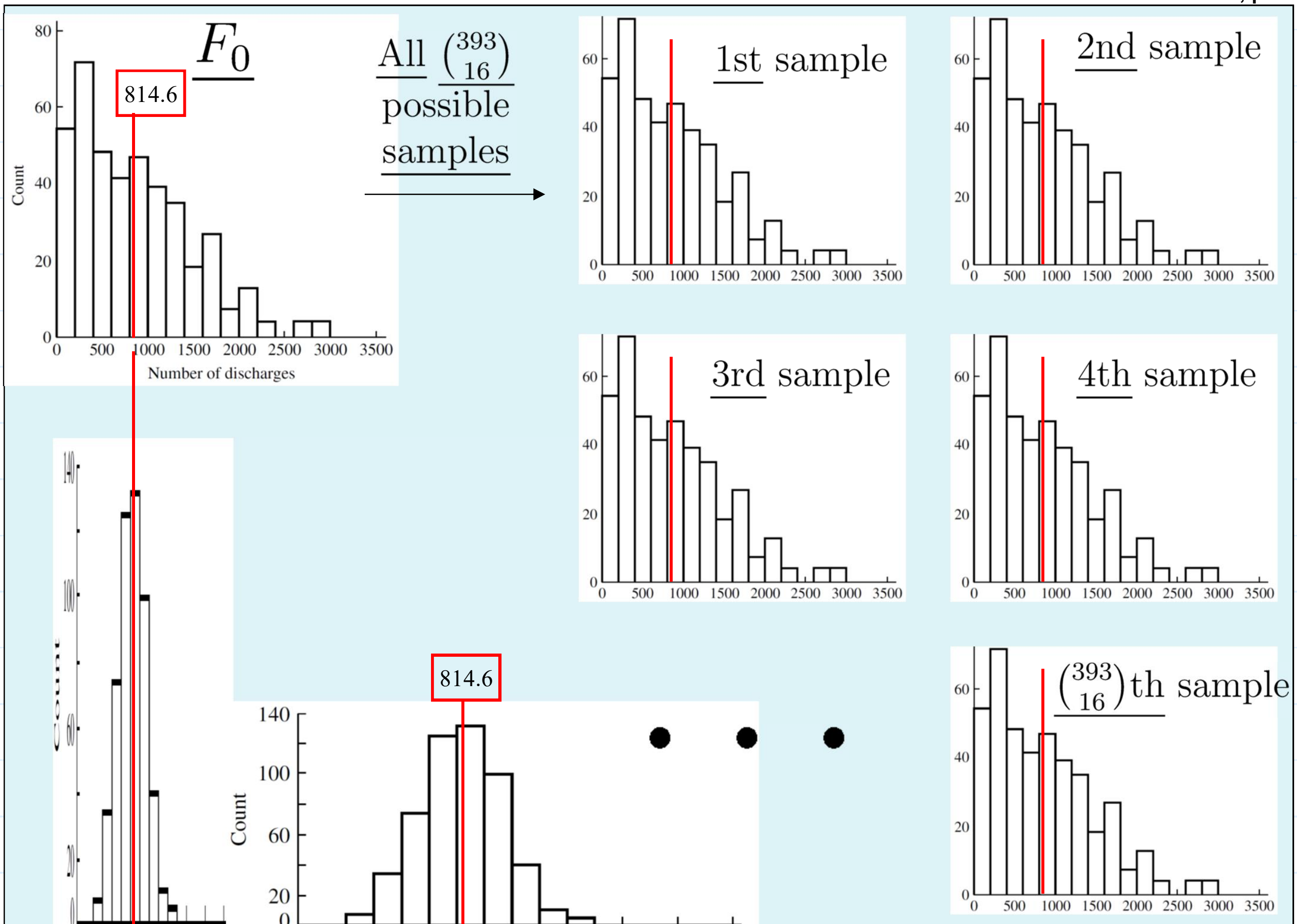
- $\overline{X}$ is clearly a statistic, and hence a random variable.
- $\overline{X}$ is an intuitive estimator of μ.
- In the dichotomous case, we have $\mu = p$ and

$$\hat{p} \equiv \overline{X} \text{ is the sample proportion.}$$



Example 3 (sampling distribution of sample mean, cont. Ex.2 in LNp.4)

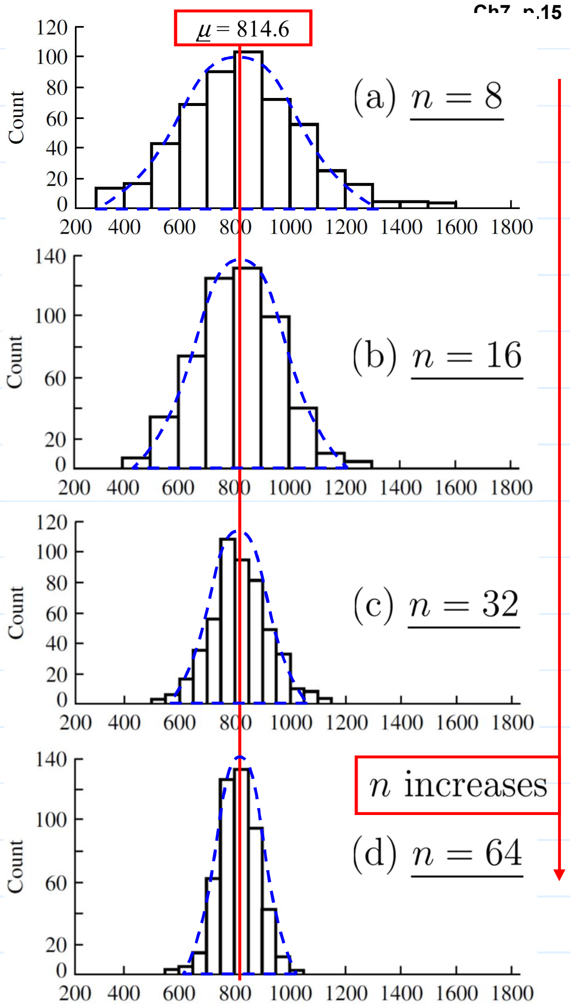
- Consider the population of N=393 hospitals.
- Suppose we want to know the sampling distribution of $\overline{X}$ of a s.r.s. without replacement of sample size n=16.
- There are $\binom{393}{16}$ possible samples. Note that $\binom{393}{16}$ is of order 10^{28} !
- Sampling distribution of $\overline{X}$ is formed by the (sample) mean of each of the possible samples along with their probabilities.



Note. If F_0 is unknown, the sampling distribution of $\overline{X}$ is unknown.

- $\binom{393}{16} = 10^{28}$ is too large
- To reduce computation, we can use the technique of **simulation** to understand the sampling distribution of $\overline{X}$.
  - randomly draw many (say, 500) s.r.s. of size n
  - compute the mean of each sample
  - form a histogram of the collection of these sample means

  This histogram will be an approximation to the sampling distribution of $\overline{X}$.
- Figure 7.2 (textbook) shows the results for sample size n=8, 16, 32, or 64.
  - All the four histograms are centered at μ=814.6.
  - As n increases, the histograms become less spread out.
  - Although shape of F_0 (population distribution) is not symmetric about μ, these histograms are nearly so.



sampling distribution of $\overline{X}$

Theorem 1 (expectation of sample mean)

(1) Under simple random sampling, with or without replacement,

$$E(X_k) = \mu \qquad \text{and} \qquad Var(X_k) = \sigma^2.$$

(2) Under simple random sampling, with or without replacement,

$$E(\overline{X}) = \mu.$$

So, $\overline{X}$ is an unbiased estimator of μ, i.e., the sampling distribution of $\overline{X}$ is *centered* at μ.

Proof: Under simple random sampling, no matter with or without replacement, the marginal distribution of X_k is F_0. Thus, we have

$$E(X_k) = \sum_{j=1}^{m} \zeta_j P(X_k = \zeta_j) = \sum_{j=1}^{m} \zeta_j (n_j/N) = \frac{1}{N} \sum_{j=1}^{m} n_j \zeta_j = \mu\,.$$

$$Var(X_k) = E(X_k^2) - [E(X_k)]^2 = \frac{1}{N}\left(\sum_{j=1}^{m} n_j \zeta_j^2\right) - \mu^2 = \sigma^2\,,$$

and

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \frac{1}{n}\sum_{k=1}^{n} E(X_k) = \frac{1}{n}(n\,\mu) = \mu\,.$$

Theorem 2 (variance of sample mean, s.r.s. with replacement)

Under simple random sampling with replacement, we have

$$Var(\overline{X}) = \sigma^2/n,$$

and the standard error (st.e.) of $\overline{X}$, denoted by $\sigma^*_{\overline{X}}$, is $\sigma/\sqrt{n}$.

Proof: Under simple random sampling with replacement, we have

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F_0.$$

Thus, $Cov(X_k, X_l) = 0$ for any $1 \le k < l \le n$, and

$$Var(\overline{X}) = Var\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2}\sum_{k=1}^{n} Var(X_k) = \frac{1}{n^2}(n\,\sigma^2) = \frac{\sigma^2}{n}.$$

Note 5 (Some notes about the st.e. of sample mean, with replacement)

- $\sigma^*_{\overline{X}} = \sigma/\sqrt{n}$ (a measure of how spread out $\overline{X}$ is) measures the precision of the estimator $\overline{X}$.
- $\sigma^*_{\overline{X}}$ is determined by n and σ, but not N.
- $\sigma^*_{\overline{X}}$ is inversely proportional to $\sqrt{n}$, i.e., in order to double the accuracy, n must be quadrupled (the contribution of each observation to the accuracy of $\overline{X}$ decays with the increase of n)

Theorem 3 (variance of sample mean, s.r.s. without replacement)

Under simple random sampling without replacement, we have

$$Var(\overline{X}) = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right),$$

and the standard error of $\overline{X}$, denoted by $\sigma_{\overline{X}}$, is $(\sigma/\sqrt{n})\sqrt{1 - \frac{n-1}{N-1}}$.

Proof: First, for $1 \le k < l \le n$,

$$\begin{aligned}
Cov(X_k, X_l) &= E(X_k X_l) - E(X_k)\,E(X_l) \\
&= \left(\sum_{s=1}^{m}\sum_{t=1}^{m} \zeta_s\,\zeta_t P(X_k = \zeta_s, X_l = \zeta_t)\right) - \mu^2 \\
&= \left[\sum_{s=1}^{m} \zeta_s^2 \left(\frac{n_s(n_s-1)}{N(N-1)}\right) + \sum_{s=1}^{m}\sum_{t\neq s} \zeta_s\,\zeta_t \left(\frac{n_s n_t}{N(N-1)}\right)\right] - \mu^2 \\
&= \left[\frac{N}{N-1}\sum_{s=1}^{m}\sum_{t=1}^{m} \zeta_s\zeta_t \left(\frac{n_s n_t}{N\cdot N}\right) - \frac{1}{N-1}\sum_{s=1}^{m}\zeta_s^2\left(\frac{n_s}{N}\right)\right] - \mu^2 \\
&= \frac{N}{N-1}\,E(X_k)\,E(X_l) - \frac{1}{N-1}\,E(X_k^2) - \mu^2 \\
&= \frac{N}{N-1}\,\mu^2 - \frac{1}{N-1}(\sigma^2 + \mu^2) - \mu^2 = \frac{-\sigma^2}{N-1}.
\end{aligned}$$

Then,

$$\begin{aligned} Var(\overline{X}) &= Var\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) \\ &= \frac{1}{n^2}\sum_{k=1}^{n} Var(X_k) + \frac{2}{n^2}\sum_{k=1}^{n-1}\sum_{l=k+1}^{n} Cov(X_k, X_l) \\ &= \frac{1}{n^2}\times(n\,\sigma^2) + \frac{2}{n^2}\times\frac{n(n-1)}{2}\times\frac{-\sigma^2}{N-1} = \frac{\sigma^2}{n}\left(1-\frac{n-1}{N-1}\right). \end{aligned}$$

Note 6 (Some notes about the st.e. of sample mean, without replacement)

- The variance of $\overline{X}$ in s.r.s. without replacement differs from that in s.r.s. with replacement by the factor $\left(1-\frac{n-1}{N-1}\right)$, which is called the **finite population correction**. $\left(\textbf{Note.}\ 1-\frac{n-1}{N-1}\to 1 \text{ when } N\to\infty\right)$
- n/N: sampling fraction $\left(\approx \frac{n-1}{N-1} \text{ in most cases}\right)$
- $\sigma_{\overline{X}} \approx \sigma^*_{\overline{X}} = \sigma/\sqrt{n}$ if the sampling fraction is very small (i.e., $n \ll N$).
- $\sigma_{\overline{X}}$ also depends on n and σ, i.e.,

  $$\sigma_{\overline{X}}\downarrow \text{ as } n\uparrow \quad\text{and}\quad \sigma_{\overline{X}}\uparrow \text{ as } \sigma\uparrow,$$

  and $\sigma_{\overline{X}}$ depends on N only through the sampling fraction.

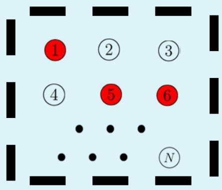
Example 4 (st.e. of sample mean, cont. Ex.2 in LNp.4)

- $N = 393$ hospitals. Consider s.r.s. without replacement of size $n = 32$.
- Because $\sigma = 589.7$ (of 393 hospitals), we have

  $$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n-1}{N-1}} \approx 100,$$

  where finite population correction $1-\frac{31}{392}\approx 0.92$ makes little difference.
- Most of sample means differ from the population mean 814 by less than $2\times\sigma_{\overline{X}} = 200$ (see graph (c) of Figure 7.2 in LNp.15).

Theorem 4 (mean and variance of sample mean for dichotomous x_i's)



In the dichotomous case, $\overline{X} = \hat{p}$ (sample proportion), and

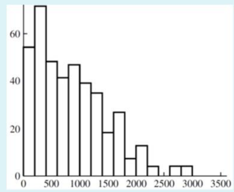
- under s.r.s. with or without replacement, $E(\hat{p}) = p$
- under s.r.s. with replacement, $Var(\hat{p}) = \dfrac{p(1-p)}{n}$,

  and $n\hat{p} = \sum_{k=1}^{n} X_k$ follows binominal(n, p) distribution
- under s.r.s. without replacement, $Var(\hat{p}) = \dfrac{p(1-p)}{n}\left(1-\dfrac{n-1}{N-1}\right)$,

  and $n\hat{p} = \sum_{k=1}^{n} X_k$ follows hypergeometric$(n, Np, N(1-p))$ distribution

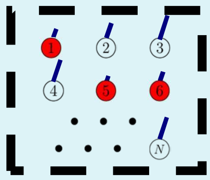
Example 5 (st.e. of sample mean, dichotomous case, cont. Ex.2 in LNp.4)

- In the population of 393 hospitals, a proportion of $p = 0.654$ had fewer than 1000 discharges.
- $y_i = 1$ if $x_i < 1000$ and $y_i = 0$ if $x_i \geq 1000$



- For an s.r.s. without replacement sample $Y_1, \ldots, Y_n$ of size $n = 32$, the estimator of p is $\hat{p} = \overline{Y}$ and

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{1 - \frac{n-1}{N-1}} = \sqrt{\frac{.654 \times .346}{32}}\sqrt{1 - \frac{31}{392}} = 0.08.$$



Definition 9 (estimator of population total)

Because $\tau = \sum_{i=1}^{N} x_i$ (population total) equals $N\mu$, an intuitive estimator of τ is

$$T = N\,\overline{X}.$$

Note. T is not $\sum_{k=1}^{n} X_k = n\,\overline{X}$.

Theorem 5 (mean of population total estimator)

Under simple random sampling, with or without replacement, we have

$$E(T) = \tau.$$

That is, T is an unbiased estimator of τ.

Theorem 6 (variance of population total estimator)

- Under simple random sampling with replacement, $Var(T) = N^2\left(\frac{\sigma^2}{n}\right)$.
- Under simple random sampling without replacement,

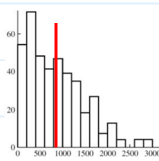
$$Var(T) = N^2\left(\frac{\sigma^2}{n}\right)\left(1 - \frac{n-1}{N-1}\right).$$

Note. The precision of the estimator T does depend on population size N.

• Estimation of population variance

Recall. When F_0 is unknown, the σ in the standard error of $\overline{X}$ is a parameter, i.e., it is unknown.

Q: how to estimate σ or σ^2?



Definition 10 (sample variance)

The **sample variance** of $X_1, X_2, \ldots, X_n$ is defined as $\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^2$.

Theorem 7 (expectation of sample variance, s.r.s. with replacement)

Under s.r.s. with replacement, we have $E(\hat{\sigma}^2) = \sigma^2\left(\frac{n-1}{n}\right)$.

Proof. From the identity

$$\sum_{k=1}^{n}(X_k-\mu)^2 = \sum_{k=1}^{n}(X_k-\overline{X})^2 + n\,(\overline{X}-\mu)^2, \qquad (\triangle)$$

by taking expectation on the both sides of $(\triangle)$, we have

$$\sum_{k=1}^{n} E\left[(X_k-\mu)^2\right] = E\left[\sum_{k=1}^{n}(X_k-\overline{X})^2\right] + n\,E\left[(\overline{X}-\mu)^2\right], \qquad (\triangledown)$$

which leads to

$$n\,\sigma^2 = E(n\,\hat{\sigma}^2) + n(\sigma^2/n).$$

Thus, we have $E(\hat{\sigma}^2) = ((n-1)\sigma^2)/n$.

Theorem 8 (expectation of sample variance, s.r.s. without replacement)

Under s.r.s. without replacement, we have $E(\hat{\sigma}^2) = \sigma^2\left(\frac{n-1}{n}\right)\left(\frac{N}{N-1}\right)$.

Proof: The identities $(\triangle)$ and $(\triangledown)$ in the above proof still hold, and $(\triangledown)$ leads to

$$n\,\sigma^2 = E(n\hat{\sigma}^2) + n\left[\frac{\sigma^2}{n}\left(1-\frac{n-1}{N-1}\right)\right].$$

After some algebra, this gives the desired result.

Note 7 (Some notes about the expectation of sample variance)

- No matter under s.r.s. with replacement or without replacement, the sample variance $\hat{\sigma}^2$ is a biased estimator of σ^2.
- Since $\frac{n-1}{n} < 1$ and $\left(\frac{n-1}{n}\right)\left(\frac{N}{N-1}\right) < 1$ (**note**. $n<N$), we have

  $$E(\hat{\sigma}^2) < \sigma^2.$$

  That is, $\hat{\sigma}^2$ tends to underestimate σ^2.

Theorem 9 (unbiased estimators of σ^2 and the variance of sample mean)

- Under s.r.s with replacement,
  - an unbiased estimator of σ^2 is

    $$s^2 = \left(\tfrac{n}{n-1}\right)\hat{\sigma}^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k-\overline{X})^2.$$

  - an unbiased estimator of $Var(\overline{X}) = \sigma^2/n$ is $s^2_{\overline{X}} = s^2/n$.
- Under s.r.s. without replacement,
  - an unbiased estimator of σ^2 is $\left(\frac{N-1}{N}\right)\left(\frac{n}{n-1}\right)\hat{\sigma}^2 = \left(\frac{N-1}{N}\right)s^2$.
  - an unbiased estimator of $Var(\overline{X}) = (\sigma^2/n)(1-\frac{n-1}{N-1})$ is

    $$s^2_{\overline{X}} = \frac{1}{n}\left(\frac{N-1}{N}s^2\right)\left(1-\frac{n-1}{N-1}\right) = \frac{s^2}{n}\left(\frac{N-n}{N}\right) = \frac{s^2}{n}\left(1-\frac{n}{N}\right).$$

Theorem 10 (unbiased est'ors of σ^2 and variance of sample mean, dichotomous x_i's)

In the dichotomous cases, $\overline{X} = \hat{p}$ and $\sigma^2 = p(1-p)$.

- Because
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^2 = \frac{1}{n}\left(\sum_{k=1}^{n} X_k^2\right) - \overline{X}^2 = \hat{p} - \hat{p}^2 = \hat{p}(1-\hat{p}),$$
we have,
$$s^2 = \left(\frac{n}{n-1}\right)\hat{\sigma}^2 = \frac{n}{n-1}\,\hat{p}(1-\hat{p})\,.$$
- Under s.r.s. with replacement, an unbiased estimator of $Var(\hat{p}) = \frac{p(1-p)}{n}$ is
$$s_{\hat{p}}^2 = s^2/n = [\hat{p}(1-\hat{p})]/n-1.$$
- Under s.r.s. without replacement, an unbiased estimator of $Var(\hat{p}) = \frac{p(1-p)}{n}\left(1 - \frac{n-1}{N-1}\right)$ is
$$s_{\hat{p}}^2 = \frac{s^2}{n}\left(1 - \frac{n}{N}\right) = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1 - \frac{n}{N}\right).$$

Theorem 11 (unbiased estimator of the variance of population total estimator)

An unbiased estimator of $Var(T) = N^2\, Var(\overline{X})$ is $s_T^2 = N^2\, s_{\overline{X}}^2$.

- The quantities $s_{\overline{X}}\ \left(= \sqrt{s_{\overline{X}}^2}\right)$, $s_T\ \left(= \sqrt{s_T^2}\right)$, and $s_{\hat{p}}\ \left(= \sqrt{s_{\hat{p}}^2}\right)$ are called **estimated standard errors**.

Example 6 (estimate population mean, cont. Ex.2 in LNp.4)

- An s.r.s. without replacement of size n=50 of the N=393 hospitals was taken.
- From this sample, $\overline{X} = 938.5$ (**recall**, $\mu = 814.6$), $s = \sqrt{s^2} = 614.53$ (**recall**, $\sigma = 590$), and an estimate of $Var(\overline{X})$ is
$$s_{\overline{X}}^2 = \frac{s^2}{n}\left(1 - \frac{n}{N}\right) = \frac{614.53^2}{50}\left(1 - \frac{50}{393}\right) = 6592.$$
- The estimated standard error of $\overline{X}$ is $s_{\overline{X}} = \sqrt{6592} = 81.19$,
$\left(\textbf{cf.}\text{ the (true) standard error of } \overline{X} \text{ is } \sigma_{\overline{X}} = \frac{590}{\sqrt{50}}\sqrt{1 - \frac{49}{392}} = 78\right)$
which gives a rough idea of how accurate the value of $\overline{X}$ (938.5) is. In this case, the magnitude of the error is of the order 80, as opposed to 8 or 800.
- The error of $\overline{X}$ is $938.5 - 814.9 = 123.9$, which is about $1.5 \times s_{\overline{X}}$.

Example 7 (estimate population total, cont. Ex.2 in LNp.4)

- For the same sample in Ex.6, the estimate of the total number of discharges τ in the population of hospitals is $T = N\,\overline{X} = 393 \times 938.5 = 368,831$ (**cf.** the true value of τ is 320,139).
- The estimated standard error of T is $s_T = N\, s_{\overline{X}} = 393 \times 81.19 = 31,908$ (**cf.** the (true) standard error of T is $\sigma_T = N\,\sigma_{\overline{X}} = 393 \times 78 = 30,654$).

Example 8 (estimate population proportion, dichotomous x_i's, cont. Ex.5 in LNp.21)

- $p = 0.654$: (true) proportion of hospitals in the population that had fewer than 1000 discharges ($\Rightarrow \sigma^2 = p(1-p) = 0.2263$).
- For the same sample in Ex.6 (LNp.26), 26 of 50 hospitals has fewer than 1000 discharges. The estimate of p is $\hat{p} = 26/50 = 0.52$, and an estimate of $Var(\hat{p})$ is

$$s_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1-\frac{n}{N}\right) = \frac{(.52)(.48)}{49}\left(1-\frac{50}{393}\right) = 0.0045$$

- The estimated standard error of $\hat{p}$ is $s_{\hat{p}} = \sqrt{0.0045} = 0.067$,

$$\left(\textbf{cf. } \text{(true) standard error of } \hat{p} \text{ is } \sigma_{\hat{p}} = \sqrt{\frac{(.654)(.346)}{50}}\sqrt{1-\frac{49}{392}} = 0.064\right)$$

which tells us that the error of $\hat{p}$ is in the 2nd or 1st decimal place — we are probably not so fortunate as to have an error in the 3rd decimal place.
- The true error of $\hat{p}$ is $0.52 - 0.654 = -0.134$, which is about $-2 \times s_{\hat{p}}$.

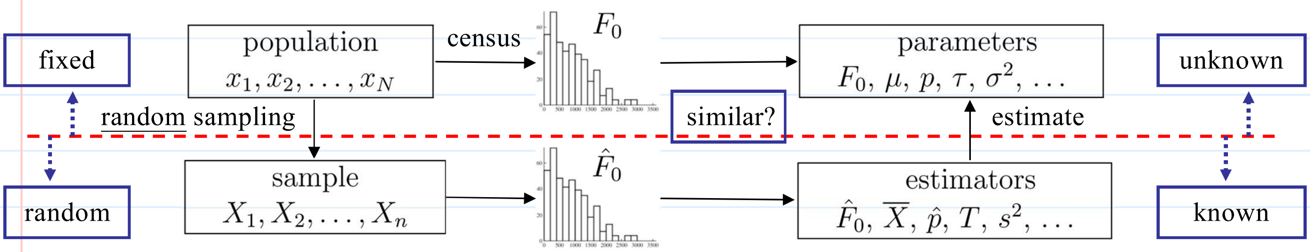
- **Note**. Examples 6-8 show how, in s.r.s., we can not only **form estimates** of unknown population parameters (e.g., use $\overline{X}$, T, $\hat{p}$ to estimate μ, τ, p, respectively), but also **gauge the likely size of the errors** of the estimates, by estimating their standard errors (e.g., $s_{\overline{X}}$, s_T, $s_{\hat{p}}$) using the data in the sample.

Note 8 (A summary of parameter estimation in s.r.s.)

- A summary table:

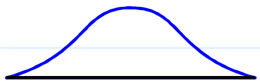
population parameter	estimator$^{(\dagger)}$	variance of estimator$^{(\dagger)(*)}$	estimated variance$^{(\dagger)(*)}$
μ	$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$	(a) $\sigma_{\overline{X}}^2 = \frac{\sigma^2}{n}$ (b) $\sigma_{\overline{X}}^2 = \frac{\sigma^2}{n}\left(1-\frac{n-1}{N-1}\right)$	(a) $s_{\overline{X}}^2 = \frac{s^2}{n}$ (b) $s_{\overline{X}}^2 = \frac{s^2}{n}\left(1-\frac{n}{N}\right)$
p	$\hat{p}$ = sample proportion	(a) $\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}$ (b) $\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}\left(1-\frac{n-1}{N-1}\right)$	(a) $s_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n-1}$ (b) $s_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1-\frac{n}{N}\right)$
τ	$T = N\,\overline{X}$	$\sigma_T^2 = N^2\,\sigma_{\overline{X}}^2$	$s_T^2 = N^2\,s_{\overline{X}}^2$
σ^2	(a) $s^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k-\overline{X})^2$ (b) $\left(1-\frac{1}{N}\right)s^2$		

- $(\dagger)$: (a) and (b) obtained under with and without replacement, repectively.
- $(*)$: the square root of entries in the 3rd column are **standard errors**, the square root of entries in the 4th column are **estimated standard errors**.



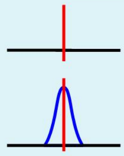
• Normal approximation to the sampling distribution of sample mean

- **Q**: without knowledge of the population distribution F_0, how to further characterize the sampling distribution $F_{\overline{X}}$ of $\overline{X}$ in addition to its mean and variance?
- Advantages if we (almostly) know the shape of $F_{\overline{X}}$?
  - accurately evaluate $P(\text{error} \in (a,b)) \approx ?$
    (**Note**. error $= \overline{X} - \mu$)
  - construct confidence interval for μ



Theorem 12 (central limit theorem, CLT, for i.i.d. case)

Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. r.v.'s and have common mean μ and variance $0 < \sigma^2 < \infty$. For the sample mean $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$, we have $E(\overline{X}_n) = \mu$, $\sigma^2_{\overline{X}_n} = Var(\overline{X}_n) = \sigma^2/n$, and for any fixed value z,



$$P\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma_{\overline{X}_n}} \le z\right) \longrightarrow \Phi(z)$$

as $n \to \infty$, where Φ is the cumulative distribution function (cdf) of the standard normal distribution $N(0,1)$. That is, $\overline{X}_n \overset{D}{\approx} N(\mu, \sigma^2/n)$.

(**cf.**) Law of large number (LLN) guarantees that $\overline{X}_n \xrightarrow{P} \mu$ and $s^2 \xrightarrow{P} \sigma^2$ as $n \to \infty$, i.e., $\overline{X}_n$ and s^2 are **consistent** estimators of μ and σ^2, respectively.

Theorem 13 (central limit theorem, CLT, for s.r.s. without replacement)

In s.r.s. without replacement, (1) $X_1, X_2, \ldots, X_n$ are not independent, and (2) there is no reason to have $n \to \infty$ while N remains fixed. But other CLTs are still appropriate, e.g.,

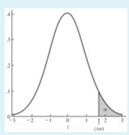
If n is large, but still small relative to N, then $\overline{X}_n$ is approximately normally distributed with mean μ and variance $\sigma_{\overline{X}_n}$ (check graphs in Ex.3, LNp.15).

Application 1 (CLT application on estimation error of population mean)

A use of CLT for estimation error $\overline{X}_n - \mu$ is

$$P\left(|\overline{X}_n - \mu| < \delta\right) = P\left(-\delta \le \overline{X}_n - \mu \le \delta\right) = P\left(-\frac{\delta}{\sigma_{\overline{X}_n}} \le \frac{\overline{X}_n - \mu}{\sigma_{\overline{X}_n}} \le \frac{\delta}{\sigma_{\overline{X}_n}}\right)$$

$$\approx \Phi\left(\frac{\delta}{\sigma_{\overline{X}_n}}\right) - \Phi\left(-\frac{\delta}{\sigma_{\overline{X}_n}}\right) = 2\,\Phi\left(\frac{\delta}{\sigma_{\overline{X}_n}}\right) - 1.$$

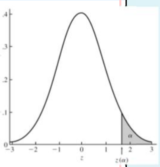


- **Note**. For the cdf Φ of $N(0,1)$, $\Phi(-z) = 1 - \Phi(z)$.

Example 9 (probability of estimation error more than δ, cont. Ex.2 in LNp.4)

- Consider the population of 393 hospitals and s.r.s. without replacement.
- For $n = 64$, $\quad \sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}\sqrt{1 - \dfrac{n-1}{N-1}} = \dfrac{589.7}{8}\sqrt{1 - \dfrac{63}{392}} = 67.5.$

- Apply CLT to approximate the probability that the sample mean $\overline{X}$ differs from μ by more than $\delta = 100$:

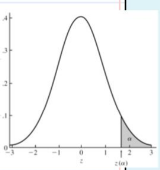


$$P(|\overline{X}-\mu| > 100) = 2 \times P(\overline{X}-\mu > 100) = 2 \times P\left(\frac{\overline{X}-\mu}{\sigma_{\overline{X}}} > \frac{100}{\sigma_{\overline{X}}}\right)$$
$$\approx \quad 2\left[1-\Phi\left(100/67.5\right)\right] = 2 \times 0.069 = 0.14.$$

- Among 500 samples of size 64 (Ex.3, LNp.15), 82 samples (or 16.4%) differed from μ more than 100.

Example 10 (estimation error more than δ, dichotomous x_i's, cont. Ex.8 in LNp.27)

- sample size $n = 50$, true $p = 0.654$, standard error of $\hat{p}$ is $\sigma_{\hat{p}} = 0.064$.
- From the sample in Ex.8, estimate of p is $\hat{p} = 0.52$ and $|\hat{p}-p| = 0.134$, the probability that the estimator will be off by an amount this large or larger is



$$P(|\hat{p}-p| > 0.134) = 1-P\left(|\hat{p}-p| \le 0.134\right)$$
$$= \quad 1-P\left(\frac{|\hat{p}-p|}{\sigma_{\hat{p}}} \le \frac{0.134}{\sigma_{\hat{p}}}\right) \approx 2\left[1-\Phi\left(2.094\right)\right] = 0.036.$$

- We see that the sample was rather "unlucky" — an error this large or larger would occur only about 3.6% of the time.

Note. In a sampling survey, σ^2 (or $\sigma^2_{\overline{X}_n}$) is not available because F_0 remains unknown. We can substitue s^2 for σ^2, and a similar CLT still holds, i.e.,

$$P\left(\frac{\overline{X}_n-\mu}{s_{\overline{X}_n}} < z\right) \longrightarrow \Phi(z) \quad \text{as } n \to \infty.$$

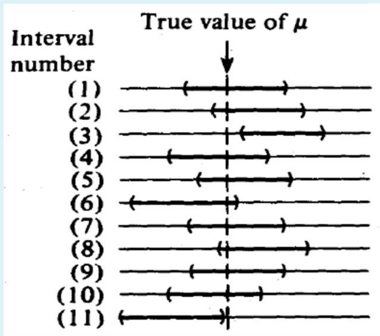
Definition 11 (interval estimator, coverage probability, interval estimate, confidence interval, and confidence level)

- For a random vector $\mathbf{X} = (X_1, \ldots, X_n)$, an **interval estimator** of a parameter θ with **coverage probability** $1-\alpha$ is a *random interval*

$$(\hat{\theta}_L(\mathbf{X}), \hat{\theta}_U(\mathbf{X})),$$

where
  1. $\hat{\theta}_L(\mathbf{X})$, $\hat{\theta}_U(\mathbf{X})$ are functions of data only,
  2. $\hat{\theta}_L(\mathbf{X}) < \hat{\theta}_U(\mathbf{X})$, and,
  3. $P(\theta \in (\hat{\theta}_L(\mathbf{X}), \hat{\theta}_U(\mathbf{X}))) = 1-\alpha$.

Repeated construction of **95%** confidence intervals



- If $\mathbf{X} = \mathbf{x}$ is observed, the interval

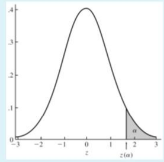
$$(\hat{\theta}_L(\mathbf{x}), \hat{\theta}_U(\mathbf{x}))$$

is called an **interval estimate**.

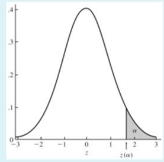
- The term "$100 \times (1-\alpha)$% **confidence interval**" (**C.I.**) is used to denote either an interval estimator with coverage probability $1-\alpha$ or an interval estimate.
- The $100(1-\alpha)$% is also referred to as **confidence level**.
- **Note**. The α is usually assigned a small value, e.g. 0.1, 0.05, or 0.01.

Application 2 (CLT application on the construction of confidence interval for μ)

- For $0 \le \alpha \le 1$, let $z(\alpha)$ be the $(1-\alpha)$-quantile of $N(0,1)$, i.e., $z(\alpha)$ is the number such that the area under the pdf of $N(0,1)$ to the right of $z(\alpha)$ is α and $\Phi(z(\alpha)) = 1-\alpha$. **Notice** that $z(1-\alpha) = -z(\alpha)$.



- For $Z \sim N(0,1)$, $P\big(-z(\alpha/2) \le Z \le z(\alpha/2)\big) = \Phi(z(\alpha/2)) - \Phi(-z(\alpha/2)) = 2 \times \Phi(z(\alpha/2)) - 1 = 1-\alpha$.



- Because $\overline{X}_n \overset{D}{\approx} N(\mu, \sigma^2_{\overline{X}_n})$ by CLT, we have

$$P\left(-z(\alpha/2) \le \frac{\overline{X}_n - \mu}{\sigma_{\overline{X}_n}} \le z(\alpha/2)\right) \approx 1-\alpha$$

$$\Leftrightarrow \quad P\left(\overline{X}_n - z(\alpha/2)\,\sigma_{\overline{X}_n} \le \mu \le \overline{X}_n + z(\alpha/2)\,\sigma_{\overline{X}_n}\right) \approx 1-\alpha$$

- The probability that μ lies in the random interval formed by data:

$$\overline{X}_n \pm z(\alpha/2)\,\sigma_{\overline{X}_n}$$

is $\approx 1-\alpha$, i.e., it is a $100(1-\alpha)\%$ (asymptotic) confidence interval of μ.

- **Recall**. A function $Q(\mathbf{X}, \theta)$ of the data $\mathbf{X}$ and a parameter, say θ, of interest is called a **pivotal quantity** for θ if the distribution of $Q(\mathbf{X}, \theta)$ is irrelevant to *all* parameters.

Note 9 (Some notes about confidence interval)

- In a sample survey, $\sigma_{\overline{X}_n}$ is unknown. In the case, $s_{\overline{X}_n}$ (or s^2, respectively) can be substituted for $\sigma_{\overline{X}_n}$ (or σ^2, respectively) if the sample size n is large enough, say $n \ge 25$ or 30 by a rule of thumb.
- **Recall**: duality between confidence interval and hypothesis testing.
  - Suppose for every parameter value θ_0, there is a level-α test for

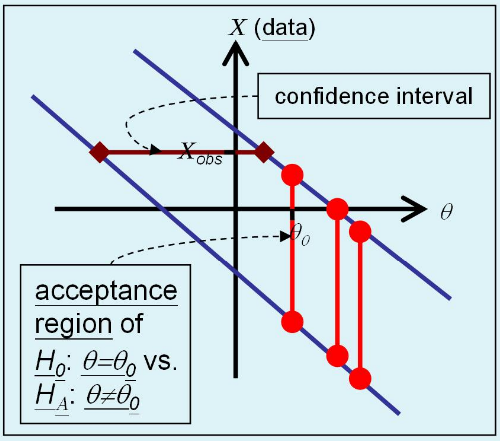
    $$H_0: \theta = \theta_0 \quad \text{vs.} \quad H_A: \theta \neq \theta_0.$$

    Denote the acceptance region of the test by $AR(\theta_0)$. Then, the set

    $$C(\mathbf{X}) = \{\,\theta \,|\, \mathbf{X} \in AR(\theta)\}$$

    is a $100(1-\alpha)\%$ C.I. for θ.
  - Suppose $C(\mathbf{X})$ is a $100(1-\alpha)\%$ C.I. for θ. Then, an acceptance region for a level-α test of H_0: $\theta = \theta_0$ is

    $$AR(\theta_0) = \{\,\mathbf{X} \,|\, \theta_0 \in C(\mathbf{X})\}.$$



- In a sample survey, for the population mean μ and the hypotheses H_0: $\mu = \mu_0$ vs. H_A: $\mu \neq \mu_0$, a test at (aymptotic) significance level α rejects H_0 if

$$\left|(\overline{X}_n - \mu_0)/\sigma_{\overline{X}_n}\right| > z(\alpha/2)$$

- Many confidence intervals have the form:

$$\text{estimate} \pm [\text{critical value}] \times [(\text{estimated})\ \text{standard error}]$$

$\Rightarrow$ C.I. combines information of estimate and (estimated) standard error

- The width of a confidence interval often depends on:
  - n : sample size
    $n \uparrow$, width $\downarrow$
  - σ: population standard deviation
    $\sigma \uparrow$, width $\uparrow$
  - $1-\alpha$: confidence level
    $(1-\alpha) \uparrow$, width $\uparrow$

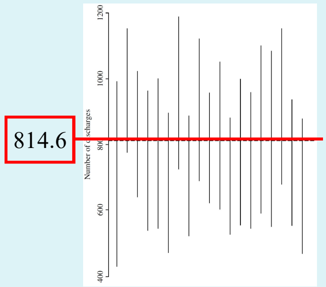
For example,
consider the C.I.:

$$\overline{X}_n \pm z(\alpha/2) \times \sigma_{\overline{X}_n}$$
$$= \overline{X}_n \pm z(\alpha/2) \times \frac{\sigma}{\sqrt{n}}$$

- If α is fixed and σ is (approximately) known, n can be chosen so as to obtain confidence intervals close to some desired length.
  $\Rightarrow$ a common way to determine an adequate survey sample size n

Example 11 (repeated construction of confidence intervals, cont. Ex.2 in LNp.4)

- 20 samples each of size $n = 25$ were drawn from the population of hospital discharges ($N = 393$).
- From each of the samples, an (approximate) 95% confidence interval for μ was computed and displayed in Figure 7.4 (textbook).
- On average 5%, or 1 out of 20, would not include μ.



Example 12 (construction of confidence intervals for μ, τ, p)

- A particular area contains 8000 (population size N) condominium units.
- To understand the numbers of motor vehecles owned by the units, a s.r.s. without replacement of size $n = 100$ was drawn.
- The sample yields that
  - the average number of motor vehicles per unit is $\overline{X} = 1.6$,
  - with a sample standard deviation $s = 0.8$.
  - So,

$$s_{\overline{X}} = \frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}} = \frac{0.8}{\sqrt{100}}\sqrt{1-\frac{100}{8000}} = 0.08.$$

- When $\alpha = 0.05$, we have $z(\alpha/2) = z(0.025) = 1.96$. Therefore, a 95% confidence interval for the population average μ is

$$\overline{X} \pm 1.96 \times s_{\overline{X}} = (1.44, 1.76).$$

- For the population total $\tau = N\mu$ (i.e., total number of motor vehicles owned by the 8000 units),
  - an estimate of τ is $T = N \times \overline{X} = 8000 \times 1.6 = 12,800$,
  - with an estimated standard error $s_T = N \times s_{\overline{X}} = 640$.
- So, a 95% confidence interval for τ is

$$T \pm 1.96 \times s_T = (11,546, 14,054).$$

- In the sample, 12% of the 100 (n) respondents said that they plan to sell their condos within the next year.
- For the proportion p of 8000 (N) units whose owners were planning to sell the units in next year,
  - an estimate of p is $\hat{p} = 0.12$,
  - with an estimated standard error $s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\sqrt{1-\frac{n}{N}} = 0.03$.
- So, a 95% confidence interval for p is $\hat{p} \pm 1.96 \times s_{\hat{p}} = (0.06, 0.18)$.
- A 95% confidence interval for the total number ($=N\times p$) of owners planning to sell is $(N\,\hat{p}) \pm 1.96 \times (N\,s_{\hat{p}}) = (451, 1469)$.

Example 13 (sample size determination, cont. Ex.12 in LNp.36)

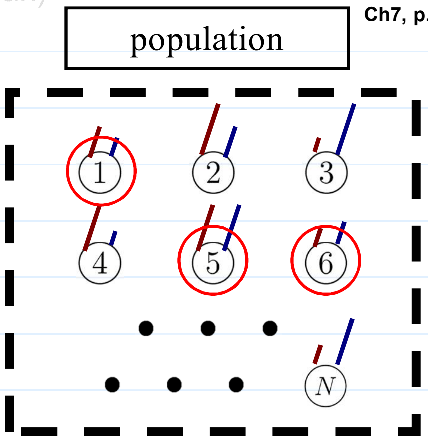
- Suppose a 95% C.I. of Np with a half-width of 200 is desired (cf., original half-width: (1469 – 451)/2=509).
- For a sample of size n^*, half-width of 95% C.I. of Np, neglecting the finite population correction (i.e., treated as s.r.s. with replacement), is

$$1.96 \times (N s_{\hat{p}}) \approx 1.96 \times N\sqrt{\frac{\hat{p}(1-\hat{p})}{n^*}} = \frac{5095}{\sqrt{n^*}}.$$

- Setting $5095/\sqrt{n^*}=200$ and solving for n^*, we have $n^* = (5095/200)^2 = 649$ (cf., original sample size n: 100).

❖ **Reading**: textbook, 7.3

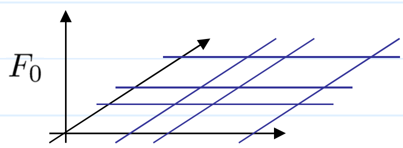
• Estimation of a ratio

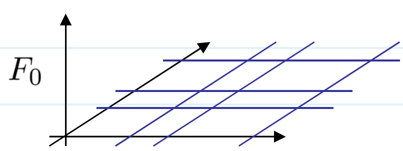


quantity of interest
(x_i, y_i)
i=1, …, N.

Some notations of the population

- (x_i, y_i), $i = 1, \ldots, N$: values associated with the members labelled by i in the population.
- Suppose that there are m_x distinct values in $x_1, x_2, \ldots, x_N$. Denote these distinct values by $\zeta_1, \zeta_2, \ldots, \zeta_{m_x}$.
- Suppose that there are m_y distinct values in $y_1, y_2, \ldots, y_N$. Denote these distinct values by $\eta_1, \eta_2, \ldots, \eta_{m_y}$.
- Denote the number of population members that have the value (ζ_s, η_u) by n_{su}, $s = 1, \ldots, m_x$, $u = 1, \ldots, m_y$.
- The proportion of population members with value (ζ_s, η_u) is n_{su}/N.
- Let $F_0(x, y)$, called **population distribution**, be the joint distribution that assigns probability n_{su}/N on (ζ_s, η_u) for $s = 1, \ldots, m_x$, $u = 1, \ldots, m_y$.





- marginal distributions of $F_0(x, y)$: Let

$$n_{s\cdot} = \sum\nolimits_{u=1}^{m_y} n_{su} \quad \text{and} \quad n_{\cdot u} = \sum\nolimits_{s=1}^{m_x} n_{su}$$

  - $F_{0,x}(x)$: assigning probability $n_{s\cdot}/N$ on ζ_s, $s = 1, \ldots, m_x$.
  - $F_{0,y}(y)$: assigning probability $n_{\cdot u}/N$ on η_u, $u = 1, \ldots, m_y$.

Definition 12 (Some population parameters that are often of interest for F_0)

- population mean, total, variance, and standard deviation of $F_{0,x}$ and $F_{0,y}$ similarly defined as in Definition 3 (LNp.5). Denote them respectively by μ_x, τ_x, σ_x^2, σ_x for $F_{0,x}$, and μ_y, τ_y, σ_y^2, σ_y for $F_{0,y}$.
- population covariance (covariance of F_0):

$$\sigma_{xy} = \frac{1}{N}\sum\nolimits_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y) = \sum\nolimits_{s=1}^{m_x}\sum\nolimits_{u=1}^{m_y}\frac{n_{su}}{N}(\zeta_s - \mu_x)(\eta_u - \mu_y)$$

- population correlation coefficient (correlation of F_0): $\rho_{xy} = \sigma_{xy}/(\sigma_x\sigma_y)$.

  Note. ρ_{xy} is a measure of the strength of the linear relationship between the x and y values in the population, and $-1 \le \rho_{xy} \le 1$.

- a population ratio: $r_{xy} = \dfrac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = \dfrac{\tau_y}{\tau_x} = \dfrac{\mu_y}{\mu_x}$. $\left(\textbf{Note}.\ r_{xy} \neq \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{x_i}\right)$

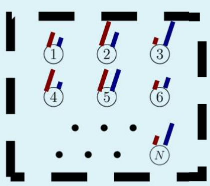
Example 14 (Applications of ratio estimation)

- survey of households
  - y: weekly food expenditure – x: number of inhabitants
  - $r_{xy} = \tau_y/\tau_x$: weekly food cost per inhabitant
- survey of households
  - y: number of unemployed males aged 20-30
  - x: number of males aged 20-30
  - $r_{xy} = \tau_y/\tau_x$: proportion of unemployed males aged 20-30
- survey of farms
  - y: acres of wheat planted – x: total acreage
  - $r_{xy} = \tau_y/\tau_x$: proportion of harvested acreage planted to wheat

Statistical modeling of (x, y)-data collected from an s.r.s. of size n.

- Define $I_1, \ldots, I_n$ as in LNp.9. The joint distribution of $I_1, \ldots, I_n$ is still as that given in LNp.9-10.
- Data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Let (X_k, Y_k), $k = 1, \ldots, n$, be the (x, y) quantity of interest observed on the kth member in the sample. We have

$$(X_k, Y_k) = (x_{I_k}, y_{I_k}),$$



  and $(X_1, Y_1), \ldots, (X_n, Y_n)$ are random variables.

- **Recall.** $F_0(x,y)$: assigning probability n_{su}/N on (ζ_s,η_u) for $s=1,\ldots,m_x$, $u=1,\ldots,m_y$. (**Note.** F_0 is unknown in a sampling survey)
- Statistical modeling of $(X_1,Y_1),\ldots,(X_n,Y_n)$ under s.r.s. with replacement
  - (exercise) marginal distribution: $(X_k,Y_k)\sim F_0(x,y)$, $k=1,\ldots,n$.
  - (exercise) the n pairs of data $(X_1,Y_1),\ldots,(X_n,Y_n)$ are independent.
  - joint distribution: $(X_1,Y_1),\ldots,(X_n,Y_n) \overset{\text{i.i.d.}}{\sim} F_0(x,y)$.
- Statistical modeling of $(X_1,Y_1),\ldots,(X_n,Y_n)$ under s.r.s. without replacement
  - (exercise) marginal distribution: $(X_k,Y_k)\sim F_0(x,y)$, $k=1,\ldots,n$ (same marginal distribution as in the with-replacement case)
  - the n pairs $(X_1,Y_1),\ldots,(X_n,Y_n)$ are not independent.
  - joint distribution of (X_k,Y_k) and (X_l,Y_l), $1\le k<l\le n$:

$$
\begin{aligned}
&P\big((X_k,Y_k)=(\zeta_s,\eta_u),(X_l,Y_l)=(\zeta_t,\eta_v)\big)\\
&= P\big((X_k,Y_k)=(\zeta_s,\eta_u)\big)\times P\big((X_l,Y_l)=(\zeta_t,\eta_v)\,\big|\,(X_k,Y_k)=(\zeta_s,\eta_u)\big)\\
&= P\big(\{(i_k,i_l)\,|\,(x_{i_k},y_{i_k})=(\zeta_s,\eta_u),\ (x_{i_l},y_{i_l})=(\zeta_t,\eta_v),\ i_k\neq i_l\}\big)\\
&= \begin{cases} \frac{n_{su}}{N}\times\frac{n_{su}-1}{N-1}=\frac{n_{su}(n_{su}-1)}{N(N-1)}, & \text{if } (\zeta_s,\eta_u)=(\zeta_t,\eta_v) \text{ (i.e., } s=t,\ u=v\text{)},\\ \frac{n_{su}}{N}\times\frac{n_{tv}}{N-1}=\frac{n_{su}n_{tv}}{N(N-1)}, & \text{otherwise.}\end{cases}
\end{aligned}
$$

  - joint distribution of $(X_1,Y_1),\ldots,(X_n,Y_n)$ is more complicated, but its derivation follows the same rule.

Definition 13 (some intuitive estimators of parameters of $F_0(x,y)$)

- population mean (μ_x,μ_y): $\overline{X}\overset{e}{\to}\mu_x$ and $\overline{Y}\overset{e}{\to}\mu_y$
- population variance (σ_x^2,σ_y^2):
  - With replacement, $s_x^2\overset{e}{\to}\sigma_x^2$. and $s_y^2\overset{e}{\to}\sigma_y^2$
  - Without replacement, $\left(\frac{N-1}{N}\right)s_x^2\overset{e}{\to}\sigma_x^2$ and $\left(\frac{N-1}{N}\right)s_y^2\overset{e}{\to}\sigma_y^2$.
- population covariance σ_{xy}: Define **sample covariance**

$$s_{xy}\equiv\frac{1}{n-1}\left[\sum_{k=1}^n(X_k-\overline{X})(Y_k-\overline{Y})\right]=\frac{1}{n-1}\left[\left(\sum_{k=1}^n X_kY_k\right)-n\,\overline{X}\,\overline{Y}\right].$$

  - (exercise) With replacement, s_{xy} is an unbiased estimator of σ_{xy}.
  - (exercise) Without replacement, $\left(\frac{N-1}{N}\right)s_{xy}$ is an unbiased estimator of σ_{xy}.
- population correlation coefficient $\rho_{xy}=\frac{\sigma_{xy}}{\sigma_x\sigma_y}$: Define **sample correlation coefficient** $\hat\rho_{xy}\overset{e}{\to}\rho_{xy}$, where

$$\hat\rho_{xy}\equiv\frac{s_{xy}}{s_xs_y}=\frac{1}{n-1}\sum_{k=1}^n\left(\frac{X_k-\overline{X}}{s_x}\right)\left(\frac{Y_k-\overline{Y}}{s_y}\right).$$

(**FYI.** $\hat\rho_{xy}$ is not an unbiased estimator of ρ_{xy} in general)

- the population ratio $r_{xy} = \frac{\mu_y}{\mu_x} = \frac{\tau_y}{\tau_x}$: a natural estimator of r_{xy} is

$$R \equiv \frac{\overline{Y}}{\overline{X}} = \frac{T_y}{T_x}.$$

Note 10 (Some notes about the mean and variance of R)

- To study the estimation properties of R, we wish to derive expressions for $E(R)$ and $Var(R)$.
- Since R is a *nonlinear* function of $\overline{X}$ and $\overline{Y}$, we cannot always do this in closed form.

  (**cf.** $\overline{X} = \frac{1}{n}\sum_{k=1}^n X_k$;
  $s_x^2 = \frac{1}{n-1}\sum_{k=1}^n (X_k - \overline{X})^2 = \frac{1}{n-1}\left[\left(\sum_{k=1}^n X_k^2\right) - n\overline{X}^2\right]$;
  $s_{xy} = \frac{1}{n-1}\sum_{k=1}^n (X_k - \overline{X})(Y_k - \overline{Y}) = \frac{1}{n-1}\left[\left(\sum_{k=1}^n X_k Y_k\right) - n\overline{X}\,\overline{Y}\right]$)
- **Q**: How can we derive approximate expressions for $E(R)$ and $Var(R)$?
- Consider the problem.
  - $Z = g(U)$, where U is a random variable and g is a known function.
  - Suppose we know only the mean μ_U and variance σ_U^2 of U, but **not** the exact distribution F_U of U (i.e., do not know the cdf or pdf/pmf of U).
  - **Q**: Can we derive the exact distribution of Z?

  - If not, can we "roughly" characterize the mean and variance of Z? (**Note.** $E(Z) = E\left[g(U)\right] \neq g\left[E(U)\right]$ in general.)

Theorem 14 (δ-method, propagation of error)

- univariate case $Z = g(U)$:

$$\begin{aligned} Z = g(U) &\approx g(\mu_U) + (U - \mu_U)g'(\mu_U) \qquad \text{(by Taylor expansion)} \\ \Rightarrow \quad E(Z) &\approx g(\mu_U) \\ Var(Z) &\approx Var(U)[g'(\mu_U)]^2 = \sigma_U^2\,[g'(\mu_U)]^2 \\ \text{or} \quad Z = g(U) &\approx g(\mu_U) + (U - \mu_U)g'(\mu_U) + (1/2)(U - \mu_U)^2\, g''(\mu_U) \\ \Rightarrow \quad E(Z) &\approx g(\mu_U) + (1/2)\sigma_U^2\, g''(\mu_U) \end{aligned}$$

  Note. How good these approximations are depends on whether g can be reasonably well approximated by the 1st- or 2nd-order polynomials in a neighborhood of μ_U and on the size of σ_U.
- case of 2 random variables U, V and $Z = g(U, V)$: Let $\boldsymbol{\mu} = (\mu_U, \mu_V)$.

$$\begin{aligned} Z = g(U, V) &\approx g(\boldsymbol{\mu}) + (U - \mu_U)\frac{\partial g(\boldsymbol{\mu})}{\partial u} + (V - \mu_V)\frac{\partial g(\boldsymbol{\mu})}{\partial v} \\ \Rightarrow \quad E(Z) &\approx g(\boldsymbol{\mu}) \\ Var(Z) &\approx \sigma_U^2\left[\frac{\partial g(\boldsymbol{\mu})}{\partial u}\right]^2 + \sigma_V^2\left[\frac{\partial g(\boldsymbol{\mu})}{\partial v}\right]^2 + 2\,\sigma_{UV}\left[\frac{\partial g(\boldsymbol{\mu})}{\partial u}\right]\left[\frac{\partial g(\boldsymbol{\mu})}{\partial v}\right] \end{aligned}$$

$$\text{or} \quad Z = g(U,V) \approx g(\boldsymbol{\mu}) + (U-\mu_U)\frac{\partial g(\boldsymbol{\mu})}{\partial u} + (V-\mu_V)\frac{\partial g(\boldsymbol{\mu})}{\partial v}$$
$$+ \frac{1}{2}(U-\mu_U)^2 \frac{\partial^2 g(\boldsymbol{\mu})}{\partial u^2} + \frac{1}{2}(V-\mu_V)^2 \frac{\partial^2 g(\boldsymbol{\mu})}{\partial v^2}$$
$$+ (U-\mu_U)(V-\mu_V)\frac{\partial^2 g(\boldsymbol{\mu})}{\partial u \partial v}$$
$$\Rightarrow \quad E(Z) \approx g(\boldsymbol{\mu}) + \frac{1}{2}\sigma_U^2\left[\frac{\partial^2 g(\boldsymbol{\mu})}{\partial u^2}\right] + \frac{1}{2}\sigma_V^2\left[\frac{\partial^2 g(\boldsymbol{\mu})}{\partial v^2}\right] + \sigma_{UV}\left[\frac{\partial^2 g(\boldsymbol{\mu})}{\partial u \partial v}\right]$$

- **Note**. A function g of k random variables can be worked out similarly.

Example 15 (Application of δ-method on the mean and variance of $g(U, V)=V/U$)

- Let $Z = g(U,V) = V/U$. Then, for $g(u,v) = v/u$,
$$\frac{\partial g}{\partial u} = \frac{-v}{u^2}, \quad \frac{\partial g}{\partial v} = \frac{1}{u}, \quad \frac{\partial^2 g}{\partial u^2} = \frac{2v}{u^3}, \quad \frac{\partial^2 g}{\partial v^2} = 0, \quad \frac{\partial^2 g}{\partial u \partial v} = \frac{-1}{u^2}.$$
- By δ-method, after substituting (μ_U, μ_V) for (u,v), we have
$$E(Z) \approx \frac{\mu_V}{\mu_U} + \frac{1}{2}\sigma_U^2\frac{2\mu_V}{\mu_U^3} + \frac{1}{2}\sigma_V^2 0 + \sigma_{UV}\frac{-1}{\mu_U^2} = \frac{\mu_V}{\mu_U} + \frac{1}{\mu_U^2}\left(\sigma_U^2\frac{\mu_V}{\mu_U} - \sigma_{UV}\right).$$
- Similarly, by δ-method,
$$Var(Z) \approx \sigma_V^2\frac{\mu_V^2}{\mu_U^4} + \sigma_V^2\frac{1}{\mu_U^2} + 2\sigma_{UV}\frac{-\mu_V}{\mu_U^2}\frac{1}{\mu_U} = \frac{1}{\mu_U^2}\left(\sigma_U^2\frac{\mu_V^2}{\mu_U^2} + \sigma_V^2 - 2\sigma_{UV}\frac{\mu_V}{\mu_U}\right).$$

Theorem 15 (Covariance of the two sample mean)

- Under s.r.s. with replacement
$$\sigma_{\overline{X}\,\overline{Y}} = Cov(\overline{X}, \overline{Y}) = \frac{\sigma_{xy}}{n}.$$
- Under s.r.s. without replacement
$$\sigma_{\overline{X}\,\overline{Y}} = Cov(\overline{X}, \overline{Y}) = \frac{\sigma_{xy}}{n}\left(1 - \frac{n-1}{N-1}\right).$$

Proof: First, under s.r.s., no matter with or without replacement, we have

$$Cov(\overline{X}, \overline{Y}) = E\left[(\overline{X}-\mu_x)(\overline{Y}-\mu_y)\right] = E\left[\left(\sum_{k=1}^n \frac{X_k-\mu_x}{n}\right)\left(\sum_{l=1}^n \frac{Y_l-\mu_y}{n}\right)\right]$$
$$= \frac{1}{n^2}\left(\sum_{k=1}^n\sum_{l=1}^n E\left[(X_k-\mu_x)(Y_l-\mu_y)\right]\right)$$
$$= \frac{1}{n^2}\sum_{k=1}^n E\left[(X_k-\mu_x)(Y_k-\mu_y)\right] + \frac{1}{n^2}\sum_{k=1}^n\sum_{\substack{l=1\\ l\neq k}}^n E\left[(X_k-\mu_x)(Y_l-\mu_y)\right]$$
$$= \frac{\sigma_{xy}}{n} + \frac{1}{n^2}\sum_{k=1}^n\sum_{\substack{l=1\\ l\neq k}}^n Cov(X_k, Y_l) \qquad (*)$$

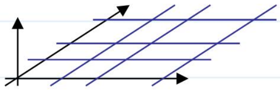
- Under s.r.s. with replacement, when $k \neq l$, X_k and Y_l are independent. Thus, for $k \neq l$, $Cov(X_k, Y_l) = 0$, and $(*)$ equals σ_{xy}/n.

- Under s.r.s. without replacement, for $k \neq l$, (X_k, Y_l) are correlated, and

$$Cov(X_k, Y_l) = E(X_k Y_l) - E(X_k)\,E(Y_l) = E(X_k Y_l) - \mu_x\,\mu_y$$

where

$$E(X_k Y_l) = \sum_{s=1}^{m_x}\sum_{v=1}^{m_y} \zeta_s\,\eta_v\,P(X_k = \zeta_s, Y_l = \eta_v)$$

F_0 

$$= \sum_{s=1}^{m_x}\sum_{v=1}^{m_y} \zeta_s\eta_v\left[\sum_{t=1}^{m_x}\sum_{u=1}^{m_y} P\Big((X_k, Y_k) = (\zeta_s, \eta_u), (X_l, Y_l) = (\zeta_t, \eta_v)\Big)\right]$$

$$= \sum_{s=1}^{m_x}\sum_{v=1}^{m_y} \zeta_s\eta_v\,P\Big((X_k, Y_k) = (\zeta_s, \eta_v),\ (X_l, Y_l) = (\zeta_s, \eta_v)\Big)$$

$$+ \sum_{s=1}^{m_x}\sum_{v=1}^{m_y} \zeta_s\eta_v\left[\underset{(\zeta_s,\eta_u)\neq(\zeta_t,\eta_v)}{\sum_{t=1}^{m_x}\sum_{u=1}^{m_y}} P\Big((X_k, Y_k) = (\zeta_s, \eta_u),\ (X_l, Y_l) = (\zeta_t, \eta_v)\Big)\right]$$

$$= \sum_{s=1}^{m_x}\sum_{v=1}^{m_y} \zeta_s\eta_v\left[\frac{n_{sv}(n_{sv}-1)}{N(N-1)}\right] + \sum_{s=1}^{m_x}\sum_{v=1}^{m_y} \zeta_s\eta_v\left[\underset{(\zeta_s,\eta_u)\neq(\zeta_t,\eta_v)}{\sum_{t=1}^{m_x}\sum_{u=1}^{m_y}}\left(\frac{n_{su}n_{tv}}{N(N-1)}\right)\right]$$

$$= \frac{N}{N-1}\sum_{s=1}^{m_x}\sum_{v=1}^{m_y} \zeta_s\eta_v\left(\sum_{t=1}^{m_x}\sum_{u=1}^{m_y}\frac{n_{su}n_{tv}}{N\cdot N}\right) - \frac{1}{N-1}\sum_{s=1}^{m_x}\sum_{v=1}^{m_y}\zeta_s\eta_v\left(\frac{n_{sv}}{N}\right)$$

$$= \frac{N}{N-1}\sum_{s=1}^{m_x}\sum_{v=1}^{m_y}\zeta_s\,\eta_v\left(\frac{n_{s\cdot}}{N}\right)\left(\frac{n_{\cdot v}}{N}\right) - \frac{1}{N-1}E(X_kY_k)$$

$$= \frac{N\big[E(X_k)\,E(Y_l)\big]}{N-1} - \frac{\sigma_{xy} + \mu_x\mu_y}{N-1} = -\frac{1}{N-1}\sigma_{xy} + \mu_x\mu_y.$$

Thus, $Cov(X_k, Y_l) = -\dfrac{\sigma_{xy}}{N-1}$ if $k \neq l$, and

$$(*) = \frac{\sigma_{xy}}{n} + \frac{1}{n^2}[n(n-1)]\left(-\frac{\sigma_{xy}}{N-1}\right) = \frac{\sigma_{xy}}{n}\left(1 - \frac{n-1}{N-1}\right).$$

Theorem 16 (approximate mean of R)

- Under s.r.s. with replacement,

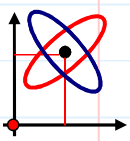
$$\mu_R \equiv E(R) \approx \frac{\mu_{\overline{Y}}}{\mu_{\overline{X}}} + \frac{1}{\mu_{\overline{X}}^2}\left(\sigma_{\overline{X}}^2\frac{\mu_{\overline{Y}}}{\mu_{\overline{X}}} - \sigma_{\overline{X}\,\overline{Y}}\right) = r_{xy} + \frac{1}{n}\times\frac{1}{\mu_x^2}(r_{xy}\,\sigma_x^2 - \sigma_{xy})$$

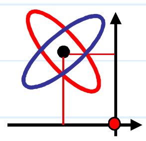
- Under s.r.s. without replacement,

$$\mu_R \equiv E(R) \approx r_{xy} + \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\times\frac{1}{\mu_x^2}(r_{xy}\,\sigma_x^2 - \sigma_{xy})$$

$$= r_{xy} + \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\times\frac{1}{\mu_x^2}(r_{xy}\sigma_x^2 - \rho_{xy}\sigma_x\sigma_y).$$

- with replacement,
  $\mu_{\overline{X}} = \mu_x$, $\mu_{\overline{Y}} = \mu_y$,
  $\sigma_{\overline{X}}^2 = \frac{\sigma_x^2}{n}$, $\sigma_{\overline{Y}}^2 = \frac{\sigma_y^2}{n}$,
  $\sigma_{\overline{X}\,\overline{Y}} = \frac{\sigma_{xy}}{n}$.





Proof: From δ-method, Ex.15 (LNp.45), and Theorem 15 (LNp.46), the results follows.

- without replacement,
  $\mu_{\overline{X}} = \mu_x$, $\mu_{\overline{Y}} = \mu_y$,
  $\sigma_{\overline{X}}^2 = \frac{\sigma_x^2}{n}\left(1 - \frac{n-1}{N-1}\right)$, $\sigma_{\overline{Y}}^2 = \frac{\sigma_y^2}{n}\left(1 - \frac{n-1}{N-1}\right)$,
  $\sigma_{\overline{X}\,\overline{Y}} = \frac{\sigma_{xy}}{n}\left(1 - \frac{n-1}{N-1}\right)$.

Note 11 (Some notes about the approximate mean of R)

- strong correlation ρ_{xy} of the same sign as $r_{xy} = \mu_y/\mu_x$ decreases the bias
- the bias is large if $|\mu_x|$ is small
- the bias is of the order $1/n$, denoted by "bias $\sim O(n^{-1})$," and its contribution to the MSE is of the order $1/n^2$, i.e., bias$^2 \sim O(n^{-2})$

Theorem 17 (approximate variance of R)

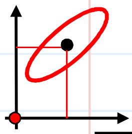
- Under s.r.s. with replacement,

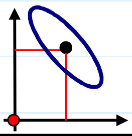
$$\sigma_R^2 \equiv Var(R) \approx \frac{1}{\mu_{\overline{X}}^2}\left(\sigma_{\overline{X}}^2 \frac{\mu_{\overline{Y}}^2}{\mu_{\overline{X}}^2} + \sigma_{\overline{Y}}^2 - 2\,\sigma_{\overline{X}\,\overline{Y}} \frac{\mu_{\overline{Y}}}{\mu_{\overline{X}}}\right) = \frac{1}{n} \times \frac{1}{\mu_x^2}\left(r_{xy}^2\,\sigma_x^2 + \sigma_y^2 - 2\,r_{xy}\,\sigma_{xy}\right).$$

- Under s.r.s. without replacement,

$$\sigma_R^2 \equiv Var(R) \approx \frac{1}{n}\left(1-\frac{n-1}{N-1}\right) \times \frac{1}{\mu_x^2}(r_{xy}^2\sigma_x^2 + \sigma_y^2 - 2r_{xy}\,\sigma_{xy})$$
$$= \frac{1}{n}\left(1-\frac{n-1}{N-1}\right) \times \frac{1}{\mu_x^2}\left(r_{xy}^2\sigma_x^2 + \sigma_y^2 - 2\,r_{xy}\,\rho_{xy}\sigma_x\sigma_y\right).$$

- with replacement, $\mu_{\overline{X}} = \mu_x$, $\mu_{\overline{Y}} = \mu_y$, $\sigma_{\overline{X}}^2 = \frac{\sigma_x^2}{n}$, $\sigma_{\overline{Y}}^2 = \frac{\sigma_y^2}{n}$, $\sigma_{\overline{X}\,\overline{Y}} = \frac{\sigma_{xy}}{n}$.





Proof: From δ-method, Ex.15 (LNp.45), and Theorem 15 (LNp.46), the results follows.

- without replacement, $\mu_{\overline{X}} = \mu_x$, $\mu_{\overline{Y}} = \mu_y$, $\sigma_{\overline{X}}^2 = \frac{\sigma_x^2}{n}\left(1-\frac{n-1}{N-1}\right)$, $\sigma_{\overline{Y}}^2 = \frac{\sigma_y^2}{n}\left(1-\frac{n-1}{N-1}\right)$, $\sigma_{\overline{X}\,\overline{Y}} = \frac{\sigma_{xy}}{n}\left(1-\frac{n-1}{N-1}\right)$.

Note 12 (Some notes about the approximate variance of R)

- strong correlation ρ_{xy} of the same sign as $r_{xy} = \mu_y/\mu_x$ decreases the variance
- the variance is large if $|\mu_x|$ is small (**Note**. small values of $\overline{X}$ in the ratio $R = \overline{Y}/\overline{X}$ cause R to fluctuate wildly)
- the variance is of the order $1/n$, i.e., "Var $\sim O(n^{-1})$"
- the contributions of the Var and the bias^2 to the MSE (=Var+bias^2) are of the order $1/n$ and $1/n^2$, respectively $\Rightarrow$ for samples of large n, the bias is negligible compared to the standard error of the estimator

Definition 14 (an intuitive estimators of the standard error of R)

- Under s.r.s. with replacement, an estimator of the $\sigma_R^2 = Var(R)$ is

$$s_R^2 = \frac{1}{n} \times \frac{1}{\overline{X}^2}(R^2\,s_x^2 + s_y^2 - 2R\,s_{xy}).$$

$$\boxed{\sigma_R^2 \approx \frac{1}{n} \times \frac{1}{\mu_x^2}\left(r_{xy}^2\,\sigma_x^2 + \sigma_y^2 - 2\,r_{xy}\,\sigma_{xy}\right)}$$

The quantity $s_R \left(= \sqrt{s_R^2}\right)$ is an estimated standard error of R.

- Under s.r.s. without replacement, an estimator of the $\sigma_R^2 = Var(R)$ is

$$s_R^2 = \frac{1}{n}\left(1-\frac{n-1}{N-1}\right)\left(\frac{N-1}{N}\right) \times \frac{1}{\overline{X}^2}(R^2\,s_x^2 + s_y^2 - 2R\,s_{xy})$$
$$= \frac{1}{n}\left(1-\frac{n}{N}\right) \times \frac{1}{\overline{X}^2}(R^2 s_x^2 + s_y^2 - 2Rs_{xy})$$

The quantity $s_R \left(= \sqrt{s_R^2}\right)$ is an estimated standard error of R.

Theorem 18 (asymptotic sampling distribution of R)

For samples of large size n,

- truncating the Taylor series (in Thm 14, LNp.44) to the 1st order provides a good approximation, since the deviations $\overline{X}_n - \mu_x$ and $\overline{Y}_n - \mu_y$ are likely to be small (by LLN)
- to this order of approximation, $R \approx \frac{\mu_y}{\mu_x} - \frac{\mu_y}{\mu_x^2}(\overline{X}_n - \mu_x) + \frac{1}{\mu_x}(\overline{Y}_n - \mu_y)$ (from Ex. 15, LNp.45), where $\overline{X}_n \overset{D}{\approx} N\left(\mu_x, \sigma_{\overline{X}_n}^2\right)$ and $\overline{Y}_n \overset{D}{\approx} N\left(\mu_y, \sigma_{\overline{Y}_n}^2\right)$ (by CLT).

- an argument based on the CLT can be used to show that R is approximately normally distributed, i.e., $R \stackrel{D}{\approx} N\left(\mu_R, \sigma_R^2\right)$, when sample size n is large.
- **Applications**
  - probability of estimation error $\in [a, b]$, e.g., $P\left(\left|\frac{R - r_{xy}}{s_R}\right| > \delta\right) \approx 2[1 - \Phi(\delta)]$
  - approximate $100(1-\alpha)\%$ confidence interval of r_{xy}: $R \pm z(\alpha/2)\, s_R$

Example 16 (estimate population ratio r_{xy})

- Suppose that 100 people who recently bought houses are surveyed, and

  y: mortgage payment $\qquad$ x: gross income

  are observed. The $r_{xy} = \tau_y/\tau_x$ is the percentage of the total mortgage amount to the total gross income of all people who recently bought houses.
- Suppose that the population size N is missing, but it is known that $100 \ll N$.
- Suppose that $\overline{X} = 3100$, $s_x = 1200$, $\overline{Y} = 868$, $s_y = 250$, $\hat{\rho}_{xy} = 0.85$.

  We have $\quad R = 868/3100 = 0.28$.
- Neglecting the finite population correction, the estimated standard error of R is

$$s_R = \frac{1}{10} \times \frac{1}{3100}\sqrt{0.28^2(1200^2) + 250^2 - 2(0.28)(0.85)(250)(1200)} = 0.006.$$

  Note that s_R is small because x and y are highly positively correlated, $r_{xy} > 0$, and $\overline{X}$ is large.

- An approximate 95% confidence interval for r_{xy} is

$$0.28 \pm 1.96 \times 0.006 = 0.28 \pm 0.012 = (0.268, 0.292).$$

- Again, neglecting the finite population correction, an estimated bias of R using Thm 16 (LNp.48) is

$$\frac{1}{n} \times \frac{1}{\overline{X}^2}(R s_x^2 - \hat{\rho}_{xy} s_x s_y) = \frac{1}{100} \times \frac{1}{3100^2}\left[(0.28)(250^2) - (0.85)(250)(1200)\right] = -0.00025,$$

  which is negligible relative to s_R (=0.006). Note that the large $\hat{\rho}_{xy}$ (=0.85) and the large value of $\overline{X}$ (=3100) cause the bias to be small.

• Ratios used for estimating population means (and totals)

- Suppose μ_x is known, e.g., the example of 393 hospitals in Ex.2 (LNp.4),
  - y: number of discharges,
  - x: number of beds.

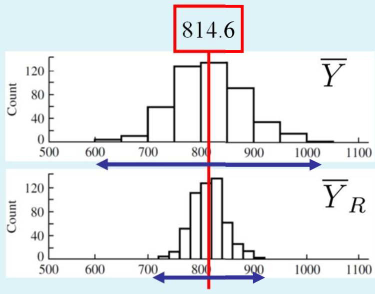
  Suppose the average (or total) number of beds μ_x (or τ_x) in the 393 hospitals is known (before a sample has been taken).
- **Q**: how to take advantage of this information in the estimation of μ_y?
- Select a random sample, and collect the data: (X_k, Y_k), $k = 1, \ldots, n$. For the parameter $\mu_y = \mu_x r_{xy}$, an intuitive **ratio estimator** of μ_y is

$$\overline{Y}_R = \mu_x R = \overline{Y}\left(\frac{\mu_x}{\overline{X}}\right) \quad \left(\stackrel{\text{cf.}}{\longleftrightarrow} \overline{Y};\ \textbf{Q}\text{: which estimator of } \mu_y \text{ is better?}\right).$$

- In the following discussion of this topic, we only consider the case of s.r.s. without replacement. The case of s.r.s. with replacement follows analogously.

Example 17 (Comparison of sample mean and ratio estimator, cont. Ex.2 in LNp.4)

- Consider the example of hospital discharges. For the population of 393 (N) hospitals and $1 \le i \le N$, let
  - x_i: number of beds in the ith hospital (known before sampling)
  - y_i: number of discharges in the ith hospital

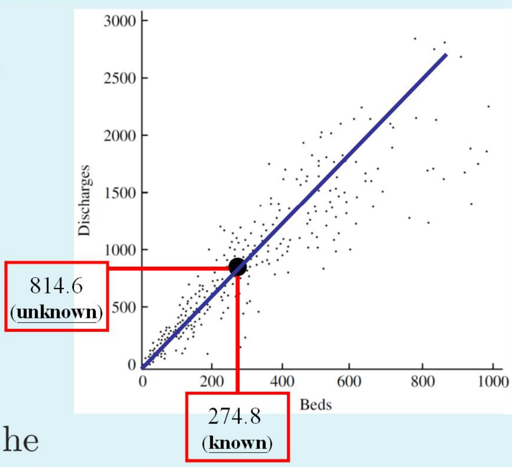


- In this population,

$$\mu_x = 274.8 \text{ (known)},$$
$$\sigma_x = 213.2 \text{ (known)},$$
$$\mu_y = 814.6, \quad \sigma_y = 589.7,$$
$$r_{xy} = 2.96, \quad \rho_{xy} = 0.91.$$

- To compare the preformance of $\overline{Y}$ and $\overline{Y}_R$, it was simulated (check Ex.3, LNp.15) 500 samples of size 64 (n) from the population of hospitals.
- The histograms of this result are shown in Figure 7.6 of textbook.
  - The histograms show that the ratio estimator $\overline{Y}_R$ of μ_y is less variable than the sample mean $\overline{Y}$.
  - The comparison shows the ratio estimator $\overline{Y}_R$ is effective at reducing variability $\Rightarrow$ $\overline{Y}_R$ is a more accurate estimator than $\overline{Y}$.

- **Q**: Why is $\overline{Y}_R$ better than $\overline{Y}$ in this case?
- An explanation. Check the scatterplot of (x_i, y_i) for the 393 hospitals in the population (Figure 7.5 of textbook) and consider a random sample (X_k, Y_k), $k = 1, \ldots, n$.



  - the population correlation $\rho_{xy} = 0.91$ is high $\Rightarrow$ a hospital with a large x_i tends to have a large y_i
  - if $\overline{X} > \mu_x$, the sample over-estimates the number of beds μ_x, and probably the number of discharges as well, i.e., probably $\overline{Y} > \mu_y$.
  - for this sample, multiplying $\overline{Y}$ by $\dfrac{\mu_x}{\overline{X}}$ decreases $\overline{Y}$ to $\overline{Y}_R$, which might be closer to μ_y than $\overline{Y}$.

Theorem 19 (approximate mean, bias, and variance of the ratio estimator)

Since $\overline{Y}_R = \mu_x R$, we have $E(\overline{Y}_R) = \mu_x E(R)$ and $Var(\overline{Y}_R) = \mu_x^2 Var(R)$. Under s.r.s. without replacement,

- the approximate bias of the ratio estimator $\overline{Y}_R$ of μ_y is

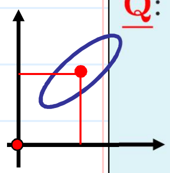
$$E(\overline{Y}_R) - \mu_y = \mu_x [E(R) - r_{xy}] \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right) \times \frac{1}{\mu_x}(r_{xy}\sigma_x^2 - \rho_{xy}\sigma_x\sigma_y),$$

- The approximate variance of the ratio estimator $\overline{Y}_R$ of μ_y is

$$\sigma^2_{\overline{Y}_R} = Var(\overline{Y}_R) = \mu_x^2\, Var(R) \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right) \times (r_{xy}^2\sigma_x^2 + \sigma_y^2 - 2r_{xy}\,\rho_{xy}\sigma_x\sigma_y)\,.$$

Proof: The results follows directly from the formulas for the approximate mean and variance of R given in Thm. 16 (LNp.48) and Thm. 17 (LNp.49).

Note 13 (A note about the approximate variance of the ratio estimator)

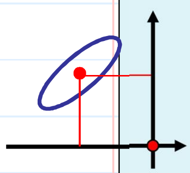


Q: When will the ratio estimator $\overline{Y}_R$ be better than the ordinary estimator $\overline{Y}$, i.e., $Var(\overline{Y}_R) < Var(\overline{Y})$?

- The ordinary estimator $\overline{Y}$ has variance $\sigma^2_{\overline{Y}} = Var(\overline{Y}) = \dfrac{\sigma_y^2}{n}\left(1 - \dfrac{n-1}{N-1}\right)$, (Thm. 3, LNp.18) and

$$\sigma^2_{\overline{Y}_R} - \sigma^2_{\overline{Y}} = Var(\overline{Y}_R) - Var(\overline{Y}) \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right) \times (r_{xy}^2\sigma_x^2 - 2r_{xy}\rho_{xy}\sigma_x\sigma_y).$$

- The ratio estimator $\overline{Y}_R$ has a smaller variance than $\overline{Y}$ if



$$r_{xy}^2\sigma_x^2 - 2r_{xy}\rho_{xy}\sigma_x\sigma_y < 0 \quad \Leftrightarrow \quad r_{xy}^2\,\sigma_x < 2\,r_{xy}\,\rho_{xy}\,\sigma_y$$

i.e.

$$\rho_{xy}\begin{cases} > \frac{1}{2}\left(\frac{\mu_y}{\mu_x}\right)\left(\frac{\sigma_x}{\sigma_y}\right) = \frac{1}{2}\left(\frac{CV_x}{CV_y}\right) & > 0, \quad \text{provided that } r_{xy} > 0, \\ < \frac{1}{2}\left(\frac{\mu_y}{\mu_x}\right)\left(\frac{\sigma_x}{\sigma_y}\right) = \frac{1}{2}\left(\frac{CV_x}{CV_y}\right) & < 0, \quad \text{provided that } r_{xy} < 0. \end{cases}$$

where $CV_x = \sigma_x/\mu_x$ and $CV_y = \sigma_y/\mu_y$ are the **coefficients of variation**.

Definition 15 (estimated standard error of ratio estimator; C.I. based on ratio estimator)

- By Thm. 19 (LNp.55) and Def. 13 (LNp.42), the variance of $\overline{Y}_R$ can be estimated by

$$s^2_{\overline{Y}_R} = \frac{1}{n}\left(1 - \frac{n}{N}\right)(R^2\,s_x^2 + s_y^2 - 2\,R\,s_{xy}),$$

The quantity $s_{\overline{Y}_R}$ $\left(= \sqrt{s^2_{\overline{Y}_R}}\right)$ is an estimated standard error of $\overline{Y}_R$.

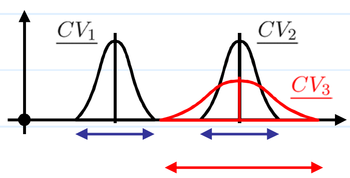
- Since $\overline{Y}_R = \mu_x\,R$, by Thm. 18 (LNp.51), an approximate $100(1-\alpha)\%$ confidence interval for μ_y $(= \mu_x\,r_{xy})$ is $\overline{Y}_R \pm z(\alpha/2)\,s_{\overline{Y}_R}$

(**cf.** the C.I. of μ_y based on $\overline{Y}$: $\overline{Y} \pm z(\alpha/2)\,s_{\overline{Y}}$ in App. 2, LNp.33.

Q: which C.I. of μ_y has a shorter width? under what condition?)

Definition 16 (coefficient of variation)

For a distribution F with mean $\mu \neq 0$ and variance σ^2, its **coefficient of variation** is defined as $CV = \sigma/\mu$, which gives σ as a proportion of μ.



Note 14 (Some notes about coefficient of variation)

- In some cases, CV is more meaningful in explaining variation than σ, e.g., communication systems.
- $CV = \sigma/\mu$ is sometimes called **noise-to-signal ratio**.
- The value of CV is free of unit.

Example 18 (Comparison of sample mean and ratio estimator, cont. Ex.17 in LNp.53)

- In the population of 393 hospitals,

  $\mu_x = 274.8,\ \sigma_x = 213.2,\ \mu_y = 814.6,\ \sigma_y = 589.7,\ r_{xy} = 2.96,\ \rho_{xy} = 0.91.$

- For a sample of size $n = 64$,
  - the standard error of the ratio estimator $\overline{Y}_R$ is (by Thm.19, LNp.55)

$$\sigma_{\overline{Y}_R} \approx \sqrt{\frac{1}{64}\left(1-\frac{63}{392}\right)} \times \sqrt{(2.96^2)(213.2^2) + 589.7^2 - 2(2.96)(0.91)(213.2)(589.7)} = 30.0$$

  - the standard error of the ordinary estimator $\overline{Y}$ is (by Thm.3, LNp.18)

$$\sigma_{\overline{Y}} = \sqrt{\frac{1}{64}\left(1-\frac{63}{392}\right)} \times 589.7 = 67.5$$

    The comparison of $\sigma_{\overline{Y}}$ to $\sigma_{\overline{Y}_R}$ is consistent with the substantial reduction in variability shown in the graph of Ex.17 (LNp.53).
  - the bias of the ratio estimator $\overline{Y}_R$ is (by Thm.19, LNp.54)

$$E(\overline{Y}_R) - \mu_y \approx \frac{1}{64}\left(1-\frac{63}{392}\right) \times \frac{1}{274.8}[(2.96)(213.2^2) - (0.91)(213.2)(589.7)] = 1.0,$$

    which is a slight and negligible bias compared to the variation reduction.

- An alternative interpretation of $\sigma^2_{\overline{Y}_R}/\sigma^2_{\overline{Y}}$. Neglecting finite population correction, an ordinary estimator $\overline{Y}_{n_1}$ from a sample of size n_1 will have about the same variance as a ratio estimator $\overline{Y}_{R,n_2}$ from a sample of size n_2 if

$$(1/n_1) \times 589.7^2 \approx (1/n_2) \times [(2.96^2)(213.2^2) + 589.7^2 - 2(2.96)(0.91)(213.2)(589.7)]$$

  Thus, $n_2/n_1 \approx (30.0/67.5)^2 = \sigma^2_{\overline{Y}_{R,n}}/\sigma^2_{\overline{Y}_n} = 0.198$, i.e., we can obtain *same precision* from $\overline{Y}_R$ using *a sample about* 80% *smaller* than the sample of $\overline{Y}$. Note that this comparison neglects the bias of $\overline{Y}_R$, justifiable in this case.

- This is a case in which a biased estimator performs substantially better than an unbiased estimator.
- In this case, the biased estimator is better because the bias is quite small and the reduction in variance is quite large.

Definition 17 (ratio estimator of population total τ_y)

Since $\tau_y = N\,\mu_y = N\,\mu_x r_{xy} = \tau_x\, r_{xy}$, an intuitive ratio estimator of τ_y is

$$T_R = \tau_x\left(\overline{Y}/\overline{X}\right) = N\ \overline{Y}\ (\mu_x/\overline{X}) = N\,\overline{Y}_R.$$

Note 15 (Some notes about the ratio estimator of population total)

- Since $E(T_R) = N\,E(\overline{Y}_R)$ and $Var(T_R) = N^2\,Var(\overline{Y}_R)$, the approximate bias and variance of T_R can be derived from Thm. 19 (LNp.54).
- The condition for $Var(T_R)$ to be smaller than $Var(T)$, where $T = N\,\overline{Y}$, is same as that given in Note 13 (LNp.55).
- An estimated standard error of T_R is $s_{T_R} = N\,s_{\overline{Y}_R}$, and an approximate $100(1-\alpha)\%$ C.I. of τ_y is $T_R \pm z(\alpha/2)\,s_{T_R}$ (following from Def.15, LNp.56).

❖ **Reading**: textbook, 7.4

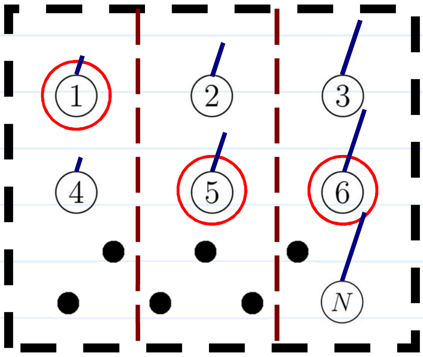
• Stratified random sampling

- **Recall**. In the discussion of ratio estimator, extra information is used to adjust the sample mean of a **biased sample** and increase accuracy.
- **Q**: Is it possible to **exclude** some biased samples in s.r.s.?

Some notations for stratified random sampling

- Let $\Omega = \{1, 2, \ldots, N\}$ be the population.
- Let $\mathbb{S}_l$, $l = 1, \ldots, L$, be a subset of Ω, and $\mathbb{S}_1, \ldots, \mathbb{S}_L$ form a partition of Ω, i.e.,
  $$\mathbb{S}_l\text{'s are disjoint and } \mathbb{S}_1 \cup \cdots \cup \mathbb{S}_L = \Omega.$$
- Each $\mathbb{S}_l$ is called a **stratum** of Ω, and the number of strata is L.
- Denote the number of members in $\mathbb{S}_l$ by N_l (subpopulation size), $l = 1, \ldots, L$. Then,
  $$N = N_1 + N_2 + \cdots + N_L.$$
- Let $W_l = N_l/N$, $l = 1, \ldots, L$, be the fraction of the population in the lth stratum.
- Let $x_{i,l}$, $i = 1, \ldots, N_l$, $l = 1, \ldots, L$, denote the value associated with the ith member in the stratum $\mathbb{S}_l$.

population



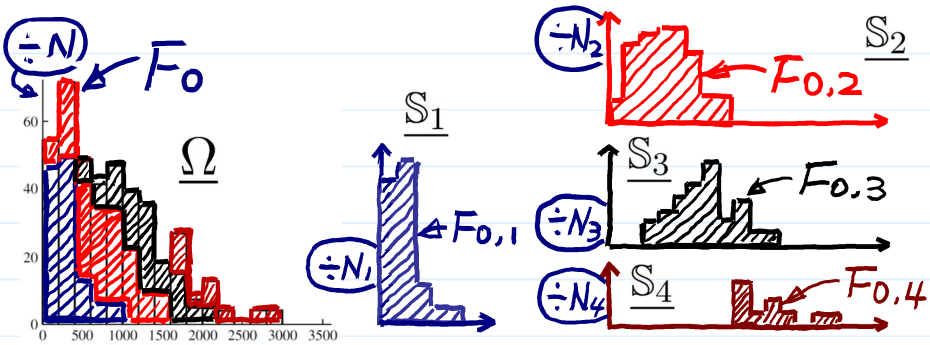
a stratum:
a sub-population (subset) of the N members in the population

the population is ***partitioned*** into L strata

- Let F_0 denote the population distribution of the whole population Ω.
- Let $F_{0,l}$, $l = 1, \ldots, L$, denote the population distribution of the subpopulation $\mathbb{S}_l$.

Ω: x_i's → distinct ζ_j's → n_j/N → F_0

$\mathbb{S}_l$: $x_{i,l}$'s → distinct $\zeta_{j,l}$'s → $n_{j,l}/N_l$ → $F_{0,l}$



Definition 18 (Some population parameters that are often of interest for F_0 and $F_{0,l}$)

- For Ω, population mean, total, variance, and standard deviation of F_0, denoted by μ, τ, σ^2, σ, respectively, are defined as in Def. 3 (LNp.5).
- For each $\mathbb{S}_l$, subpopulation mean, total, variance, and standard deviation of $F_{0,l}$, denoted by μ_l, τ_l, σ_l^2, σ_l, respectively, are similarly defined as above.

Theorem 20 (some relations between the parameters of population and subpopulation)

- The two sampling schemes are equivalent:
  [a]. Perform an s.r.s. from Ω to get one observation X
  [b]. (1) Randomly select a stratum Z, say $Z = l$, with probability proportional to the stratum size N_l; (2) Perform an s.r.s from $\mathbb{S}_l$ to get an X
- Thus, we have the distributions of X, Z, and $X|Z$ as follows:
  - $X \sim F_0$ (from [a])
  - $Z \in \{1, \ldots, L\}$ and $P(Z = l) = W_l = N_l/N$, $l = 1, \ldots, L$ (from [b].(1))
  - $X \big| Z = l \sim F_{0,l}$ (from [b].(2))
- mean: $\mu = \frac{1}{N} \sum_{l=1}^{L} \sum_{i=1}^{N_l} x_{i,l} = \frac{1}{N} \sum_{l=1}^{L} N_l \mu_l = \sum_{l=1}^{L} W_l \mu_l$

  $\Big($**Recall.** $E(X) = E_Z\Big[E_{X|Z}(X|Z)\Big]$ $\Big)$
- total: $\tau = N\mu = \sum_{l=1}^{L} \sum_{i=1}^{N_l} x_{i,l} = \sum_{l=1}^{L} N_l \mu_l = \sum_{l=1}^{L} \tau_l$
- variance: $\sigma^2 = \frac{1}{N} \sum_{l=1}^{L} \sum_{i=1}^{N_l} (x_{i,l} - \mu)^2$

  $$= \frac{1}{N} \sum_{l=1}^{L} N_l \sigma_l^2 + \frac{1}{N} \sum_{l=1}^{L} N_l (\mu_l - \mu)^2 = \sum_{l=1}^{L} W_l \sigma_l^2 + \sum_{l=1}^{L} W_l (\mu_l - \mu)^2$$

  $\Big($**Recall.** $Var(X) = E_Z\Big[Var_{X|Z}(X|Z)\Big] + Var_Z\Big[E_{X|Z}(X|Z)\Big]$ $\Big)$

Definition 19 (Stratified random sampling)

In a stratified random sampling, to obtain a sample of size n, a simple random sampling (either with replacement or without replacement, but consistent in all strata) is taken independently within each stratum $\mathbb{S}_l$ to draw a subsample of size n_l, $l = 1, \ldots, L$, where $n = n_1 + n_2 + \cdots + n_L$.

Results from the strata are combined to estimate the population parameters.

Theorem 21 (Q: how many different possible samples? how many s.r.s are excluded?)

- Under with replacement, the number of all possible stratified random samples of size n is
  $$N_1^{n_1} \times N_2^{n_2} \times \cdots \times N_L^{n_L} < N^n,$$
  where N^n = number of all possible s.r.s. of size n with replacement.
- Under without replacement, the number of all possible stratified random samples of size n is
  $$\binom{N_1}{n_1} \times \binom{N_2}{n_2} \times \cdots \times \binom{N_L}{n_L} < \binom{N}{n},$$
  where $\binom{N}{n}$ = number of all possible s.r.s. of size n without replacement.

- **Q**: What is a good way of partitioning the population Ω into strata?
- **Q**: How to choose a good sampling schemne? $\Rightarrow$ How to allocate the sample size n to each stratum, i.e., how to determine $n_1, \ldots, n_L$?

Example 19 (Applications of stratified random sampling)

- In auditing financial transactions, the trnasactions may be grouped into strata on the basis of their nominal values, e.g., high-value, medium-value, and low-value strata.
- In human populations, geographical area often form natural strata.

Note 16 (Advantages of stratified random sampling)

- It provides information about each subpopulation $\mathbb{S}_l$ in addition to the population Ω as a whole, e.g., in an industrial application,
  - population = all items produced by a manufacturing process;
  - subpopulations = items produced from different shifts or lots.
- It guarantees a prescribed number n_l of observations from each $\mathbb{S}_l$.
- Stratified sample mean can be considerably more precise than the mean of a simple random sample (shown in later slides), expecially if the partition of the population into strata
  - is homogeneous within each stratum, and
  - has large variation between strata.

$$\left(\textbf{Recall.} \quad \sigma^2 = \sum_{l=1}^{L} W_l\,\sigma_l^2 + \sum_{l=1}^{L} W_l\,(\mu_l - \mu)^2 \right)$$

Statistical modeling of data collected from a stratified random sampling.

- **Data**: $\underbrace{(X_{1,1}, X_{2,1}, \ldots, X_{n_1,1})}_{\in \mathbb{S}_1}, \ldots, \underbrace{(X_{1,L}, X_{2,L}, \ldots, X_{n_L,L})}_{\in \mathbb{S}_L}$,

  where $(X_{1,l}, \ldots, X_{n_l,l})$, $l = 1, \ldots, L$, is the data collected from the s.r.s. (either with or without replacement) taken within the lth stratum $\mathbb{S}_l$.
- distribution of data
  - $(X_{1,l}, \ldots, X_{n_l,l})$: since a s.r.s. is taken within each stratum, the joint distribution of the data from the stratum $\mathbb{S}_l$ is as that given in LNp.11-12, with F_0 replaced by $F_{0,l}$
  - data from different strata are independent

Definition 20 (some intuitive estimators of the parameters of population and stratum)

- subpopulation $\mathbb{S}_l$: since a s.r.s. is taken within each stratum,
  - mean μ_l: estimated by the subsample mean $\quad \overline{X}_l \equiv \frac{1}{n_l}\sum_{k=1}^{n_l} X_{k,l}$
  - total τ_l: estimated by subsample total $\quad T_l \equiv N_l \overline{X}_l$
  - variance σ_l^2: estimated by $s_l^2 \equiv \frac{1}{n_l - 1}\sum_{k=1}^{n_l}(X_{k,l} - \overline{X}_l)^2$ under with replacement, and by $\left(1 - \frac{1}{N_l}\right)s_l^2$ under without replacement

- (whole) population Ω: under a stratified random sample,
  - mean μ: estimated by the stratified sample mean
    $$\overline{X}_{\mathbb{S}} \equiv \frac{1}{N}\sum_{l=1}^{L} N_l \overline{X}_l = \sum_{l=1}^{L} W_l \overline{X}_l = \frac{1}{N}\sum_{l=1}^{L} \frac{1}{(n_l/N_l)}\left(\sum_{k=1}^{n_l} X_{k,l}\right),$$
    since $\mu = \sum_{l=1}^{L} W_l \mu_l$.
    $\Big($**Note**. $\overline{X}_{\mathbb{S}} \neq \frac{1}{n}\sum_{l=1}^{L}\sum_{k=1}^{n_l} X_{k,l} = \sum_{l=1}^{L} \frac{n_l}{n}\overline{X}_l$ in general,
    they are equal only when $\frac{n_l}{n} = \frac{N_l}{N}$, $l = 1, \ldots, L$. $\Big)$
  - total τ $(=N\mu)$: estimated by $T_{\mathbb{S}} \equiv N\overline{X}_{\mathbb{S}}$
  - **FYI**. An intuitive estimator of the population variance σ^2 can be developed, based on the relation between σ^2 and μ_l's, σ_l^2's (Thm. 20, LNp.61), by using the estimators $\overline{X}_l$'s and s_l^2's (or $(1-\frac{1}{N_l})s_l^2$'s).

Theorem 22 (mean and variance of the stratified estimator of population mean)

- Under stratified random sampling, with or without replacement, $E(\overline{X}_{\mathbb{S}}) = \mu$.
- Under stratified random sampling,
  - with replacement, $Var(\overline{X}_{\mathbb{S}}) = \sum_{l=1}^{L} W_l^2\left(\frac{\sigma_l^2}{n_l}\right)$.
  - without replacement, $Var(\overline{X}_{\mathbb{S}}) = \sum_{l=1}^{L} W_l^2\left(\frac{\sigma_l^2}{n_l}\right)\left(1-\frac{n_l-1}{N_l-1}\right)$.

Proof: The expectation of the stratified estimator $\overline{X}_{\mathbb{S}}$ is
$$E(\overline{X}_{\mathbb{S}}) = E\left(\sum_{l=1}^{L} W_l \overline{X}_l\right) = \sum_{l=1}^{L} W_l E(\overline{X}_l) = \sum_{l=1}^{L} W_l \mu_l = \mu.$$
Since the data from different strata are independent of one anther, the subsample means $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_L$ are independent random variables, and
$$Var(\overline{X}_{\mathbb{S}}) = Var\left(\sum_{l=1}^{L} W_l \overline{X}_l\right) = \sum_{l=1}^{L} W_l^2\, Var(\overline{X}_l).$$
Since s.r.s. is taken within each stratum, the results follows from Thm.2 (LNp.17) and Thm.3 (LNp.18) respectively for with and without replacement.

Note 17 (Some notes about the mean and variance of the stratified estimator of μ)

- Under stratified random sampling, $\overline{X}_{\mathbb{S}}$ is an unbiased estimator of μ.
- If the sampling fractions (i.e., (n_l/N_l)'s) within all strata are small, then
  with replacement $\approx$ without replacement,
  and
  $$\sum_{l=1}^{L} W_l^2\left(\frac{\sigma_l^2}{n_l}\right) \approx \sum_{l=1}^{L} W_l^2\left(\frac{\sigma_l^2}{n_l}\right)\left(1-\frac{n_l-1}{N_l-1}\right).$$

Definition 22 (estimated standard error of the stratified estimator of population mean)

- Under stratified random sampling with replacement, since s_l^2 is an unbiased estimator of σ_l^2, the $Var(\overline{X}_{\mathbb{S}})$ can be estimated by
  $$s_{\overline{X}_{\mathbb{S}}}^2 = \sum_{l=1}^{L} W_l^2\left(\frac{s_l^2}{n_l}\right). \qquad \left(\xrightarrow{\sqrt{}} s_{\overline{X}_{\mathbb{S}}}\right)$$

- Under stratified random sampling without replacement, since $(1-\frac{1}{N_l})s_l^2$ is an unbiased estimator of σ_l^2, the $Var(\overline{X}_{\mathbb{S}})$ can be estimated by

$$s^2_{\overline{X}_{\mathbb{S}}} = \sum_{l=1}^{L} W_l^2\left(\frac{s_l^2}{n_l}\right)\left(1-\frac{1}{N_l}\right)\left(1-\frac{n_l-1}{N_l-1}\right)$$
$$= \sum_{l=1}^{L} W_l^2\left(\frac{s_l^2}{n_l}\right)\left(1-\frac{n_l}{N_l}\right). \qquad \left(\xrightarrow{\sqrt{}} s_{\overline{X}_{\mathbb{S}}}\right)$$

Theorem 23 (mean and variance of the stratified estimator of population total)

Since $T_{\mathbb{S}} = N\,\overline{X}_{\mathbb{S}}$, we have $E(T_{\mathbb{S}}) = N\,E(\overline{X}_{\mathbb{S}})$ and $Var(T_{\mathbb{S}}) = N^2\,Var(\overline{X}_{\mathbb{S}})$.

- $E(T_{\mathbb{S}}) = N\,\mu = \tau$, i.e., $T_{\mathbb{S}}$ is an unbiased estimator of τ
- $$Var(T_{\mathbb{S}}) = \begin{cases} \sum_{l=1}^{L} N_l^2\left(\frac{\sigma_l^2}{n_l}\right), & \text{if with replacement,} \\ \sum_{l=1}^{L} N_l^2\left(\frac{\sigma_l^2}{n_l}\right)\left(1-\frac{n_l-1}{N_l-1}\right), & \text{if without replacement,} \end{cases}$$

Note. The $Var(T_{\mathbb{S}})$ can be estimated by $s^2_{T_{\mathbb{S}}} \equiv N^2\,s^2_{\overline{X}_{\mathbb{S}}}$. $\left(\xrightarrow{\sqrt{}} s_{T_{\mathbb{S}}} = N\,s_{\overline{X}_{\mathbb{S}}}\right)$

Example 20 (stratified random sampling, cont. Ex.17 in LNp.53)

- Consider the population of 393 hospitals.
- Assume that the number of beds in each hospital is known, and 4 strata are determined by the number of beds from small to large:

Stratum	N_l	W_l	μ_l	σ_l
A	98	0.249	182.9	103.4
B	98	0.249	526.5	204.8
C	98	0.249	956.3	243.5
D	99	0.252	1591.2	419.2

- For a without-replacement stratified random sample of size n, suppose we choose $n_1 = n_2 = n_3 = n_4 = n/4$. Neglecting the finite population correction, we have
$$\sigma_{\overline{X}_{\mathbb{S}}} = \sqrt{\frac{4}{n}\sum_{l=1}^{4} W_l^2\,\sigma_l^2} = \frac{268.4}{\sqrt{n}}.$$
- For a without-replacement s.r.s. of size n, neglecting the finite population correction, we have (see Ex.4, LNp.20)
$$\sigma_{\overline{X}} = \frac{589.7}{\sqrt{n}}.$$
- Note that the stratification has resulted in a tremendous gain in precision: $\sigma_{\overline{X}_{\mathbb{S}}} \approx 0.455 \times \sigma_{\overline{X}} \;\Rightarrow\; \sigma^2_{\overline{X}_{\mathbb{S}}}/\sigma^2_{\overline{X}} = 0.207$. The stratified estimator $\overline{X}_{\mathbb{S}}$ based on a total sample size of $n/5$ is as precise as $\overline{X}$ based on a s.r.s. of size n.
(**cf.** the reduction in variance due to stratification is comparable to that achieved by using a ratio estimator given in Ex.18, LNp.58).

• Methods of allocation in stratified random sampling

- **Q**: Why and when can a stratification produce a dramatic improvement in precision?

- In the following discussion of this topic, we consider the without-replacement case, but neglect the finite population correction. Actually, this is equivalent to the with-replacement case.

Theorem 24 (optimal allocation of the sample size n in a stratified random sampling)

Neglecting the finite population correction, the subsample sizes $n_1, \ldots, n_L$ that minimize $Var(\overline{X}_{\mathbb{S}})$ subject to the constraint $n_1 + n_2 + \cdots + n_L = n$ are

$$n_l = n \times \frac{W_l \sigma_l}{\sum_{l'=1}^{L} W_{l'} \sigma_{l'}} = n \times \frac{W_l \sigma_l}{\overline{\sigma}}, \qquad l = 1, 2, \ldots, L,$$

where $\overline{\sigma} = \sum_{l'=1}^{L} W_{l'} \sigma_{l'}$ is a weighted average of $\sigma_1, \ldots, \sigma_L$.

Proof. Introduce a Lagrange multiplier λ, and minimize

$$L(n_1, \cdots, n_L, \lambda) = Var(\overline{X}_{\mathbb{S}}) + \lambda\left(\sum_{l'=1}^{L} n_{l'} - n\right) = \sum_{l'=1}^{L} \frac{W_{l'}^2 \sigma_{l'}^2}{n_{l'}} + \lambda\left(\sum_{l'=1}^{L} n_{l'} - n\right).$$

Setting the partial derivatives equal to zero

$$0 = \frac{\partial L}{\partial n_l} = -\frac{W_l^2 \sigma_l^2}{n_l^2} + \lambda, \quad l = 1, \cdots, L, \quad \text{and} \quad \frac{\partial L}{\partial \lambda} = \sum_{l'=1}^{L} n_{l'} - n = 0$$

we have

$$n_l = \frac{W_l \sigma_l}{\sqrt{\lambda}}, \quad l = 1, \cdots, L \quad \Rightarrow \quad n = \sum_{l'=1}^{L} n_{l'} = \frac{1}{\sqrt{\lambda}} \sum_{l'=1}^{L} W_{l'} \sigma_{l'}.$$

Thus,

$$\frac{1}{\sqrt{\lambda}} = \frac{n}{\sum_{l'=1}^{L} W_{l'} \sigma_{l'}} \quad \Rightarrow \quad n_l = n \times \frac{W_l \sigma_l}{\sum_{l'=1}^{L} W_{l'} \sigma_{l'}}.$$

Note 18 (Some notes about the optimal allocation scheme)

- This theorem shows that those strata with large $W_l \sigma_l$ should be sampled heavily. This makes sense intuitively because
  - W_l is large $\Rightarrow$ $\mathbb{S}_l$ contains a large fraction of Ω $\Rightarrow$ sample more
  - σ_l is large $\Rightarrow$ $x_{i,l}$'s in $\mathbb{S}_l$ are quite variable $\Rightarrow$ a relatively large n_l is required to obtain a good determination of μ_l
- This optimal allocation scheme depends on the within-stratum variances σ_1^2, ..., σ_L^2, which generally is unknown before sampling.
- If a survey measures several attributes, it is usually impossible to find an allocation optimal for all attributes.

Definition 23 (optimal stratified estimator)

- This optimal allocation scheme is called **Neyman allocation**.
- Denote the stratified estimator under this optimal allocation scheme by
$$\overline{X}_{\mathbb{S},o}.$$

Theorem 25 (variance of the optimal stratified estimator)

Neglecting the finite population correction, and substituting the optimal values of n_l's in Thm. 24 (LNp.69) for the variance of the stratified estimator $\overline{X}_{\mathbb{S}}$ of μ presented in Thm. 22 (LNp.65) gives us

$$Var(\overline{X}_{\mathbb{S},o}) \approx \sum_{l=1}^{L} W_l^2 \left(\frac{\sigma_l^2}{n_l}\right) = \sum_{l=1}^{L} \frac{W_l^2 \sigma_l^2}{n(W_l \sigma_l)/\overline{\sigma}} = \frac{\overline{\sigma}^2}{n} = \frac{1}{n}\left(\sum_{l=1}^{L} W_l \sigma_l\right)^2.$$

Definition 24 (Proportional allocation and its stratified estimator)

- **Proportional allocation**. A simple and popular alternative method of allocation is to use the same sampling fraction in each stratum, i.e.,
$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \cdots = \frac{n_L}{N_L} \quad \left(= \frac{n}{N}\right),$$
which holds iff $n_l = n(N_l/N) = n\,W_l$ for $l = 1, \cdots, L$.
- Denote the stratified estimator under the proportional allocation scheme by $\overline{X}_{\mathbb{S},p}$.

  (**Note**. $\overline{X}_{\mathbb{S},o}$ and $\overline{X}_{\mathbb{S},p}$ are the estimator $\overline{X}_{\mathbb{S}}$ under two different allocation schemes (different n_l's, different possible samples (LNp.62), different joint distribution of data (LNp.64)). They are not different estimators.)

Note 19 (Some notes about the proportional allocation)

- Compared to the optimal allocation schemes, the proportional allocation neglects the difference in within-stratum variances σ_l's.
- If $\sigma_1 = \sigma_2 = \cdots = \sigma_L = \overline{\sigma}$, proportional allocation = optimal allocation, and $\overline{X}_{\mathbb{S},p}$ and $\overline{X}_{\mathbb{S},o}$ have same variance (accuracy).
- Under the proportional allocation,
$$\overline{X}_{\mathbb{S},p} = \sum\nolimits_{l=1}^{L} W_l \overline{X}_l = \sum\nolimits_{l=1}^{L} \left(\frac{n_l}{n}\right)\left(\frac{1}{n_l}\sum\nolimits_{k=1}^{n_l} X_{k,l}\right) = \frac{1}{n}\sum\nolimits_{l=1}^{L}\sum\nolimits_{k=1}^{n_l} X_{k,l},$$
which is the unweighted mean of the sampled data.

Example 21 (optimal and proportional allocations, cont. Ex.20 in LNp.68)

Consider the population of 393 hospitals. The sampling fractions of the 4 strata are

Stratum	A	B	C	D	
optimal allocation	0.106	0.210	0.250	0.434	$\Leftarrow W_l\,\sigma_l/(\sum_l W_l\sigma_l)$
proportional allocation	0.249	0.249	0.249	0.252	$\Leftarrow W_l$

Theorem 26 (variance of the stratified estimator under proportional allocation)

Ignoring the finite population correction, and substituting the proportional allocation, $n_l = nW_l$, for the variance of the stratified estimator $\overline{X}_{\mathbb{S}}$ of μ presented in Thm. 22 (LNp.65) gives us
$$Var(\overline{X}_{\mathbb{S},p}) \approx \sum\nolimits_{l=1}^{L} W_l^2\left(\frac{\sigma_l^2}{n_l}\right) = \sum\nolimits_{l=1}^{L} \frac{W_l^2\sigma_l^2}{nW_l} = \frac{1}{n}\left(\sum\nolimits_{l=1}^{L} W_l\,\sigma_l^2\right).$$

Theorem 27 (variance difference between the optimal and proportional allocations)

Ignoring the finite population correction, from Thm.25 (LNp.70) and Thm.26,
$$Var(\overline{X}_{\mathbb{S},p}) - Var(\overline{X}_{\mathbb{S},o}) \approx \frac{1}{n}\left(\sum\nolimits_{l=1}^{L} W_l\sigma_l^2\right) - \frac{1}{n}\,\overline{\sigma}^2 = \frac{1}{n}\sum_{l=1}^{L} W_l(\sigma_l - \overline{\sigma})^2 \geq 0.$$

Note that

- If $\sigma_1 = \sigma_2 = \cdots = \sigma_L$, then $Var(\overline{X}_{\mathbb{S},p}) = Var(\overline{X}_{\mathbb{S},o})$.
- The more variable these σ_l's are, the better it is to use (if feasible) the optimal allocation.

Example 22 (comparison of optimal and proportional allocations, cont. Ex.20, LNp.67)

- Consider the population of 393 hospitals and the 4 strata.
- Under the sampling fractions given in Ex.21 (LNp.72),

$$\frac{Var(\overline{X}_{\mathbb{S},p})}{Var(\overline{X}_{\mathbb{S},o})} = 1 + \frac{\sum_l W_l(\sigma_l - \bar{\sigma})^2}{\bar{\sigma}^2} = 1 + 0.218.$$

- Under proportional allocation, the variance of the stratified estimator $\overline{X}_{\mathbb{S}}$ of μ is about 20% larger than it is under the optimal allocation.

- **Q**: **When** can a stratified random sample based on proportional allocation perform better than a simple random sample? Note that
  - under proportional allocation, $\overline{X}_{\mathbb{S},p}$ and $\overline{X}$, where $\overline{X}$ is the sample mean of the data from a s.r.s., have the same functional form,
  - a s.r.s. has more possible samples than a stratified random sample (check Thm. 21, LNp.62).
- **Recall.**
  - Under a s.r.s. without replacement, neglecting the finite population correction, we have $Var(\overline{X}) \approx \sigma^2/n$ (check Thm.3, LNp.18).
  - From Thm.20 (LNp.61),

$$\sigma^2 = \sum_{l=1}^{L} W_l\,\sigma_l^2 + \sum_{l=1}^{L} W_l(\mu_l - \mu)^2.$$

Theorem 28 (variance difference between s.r.s. and proportional allocation)

Ignoring the finite population correction, from Thm.26 (LNp.71), we have

$$Var(\overline{X}) - Var(\overline{X}_{\mathbb{S},p}) \approx \frac{\sigma^2}{n} - \frac{1}{n}\left(\sum_{l=1}^{L} W_l\sigma_l^2\right) = \frac{1}{n}\left[\sum_{l=1}^{L} W_l(\mu_l - \mu)^2\right] \geq 0\,.$$

Note 20 (Some notes about s.r.s., proportional allocation, and optimal allocation)

- Stratified random sampling with proportional allocation is better than s.r.s., which is a result of excluding some unwanted simple random samples.
- Comparing the equations for the variances under s.r.s., proportional allocation, and optimal allocation, we see that

$$\boxed{\begin{aligned}&Var(\overline{X})\\ &\geq Var(\overline{X}_{\mathbb{S},p})\\ &\geq Var(\overline{X}_{\mathbb{S},o})\end{aligned}}$$

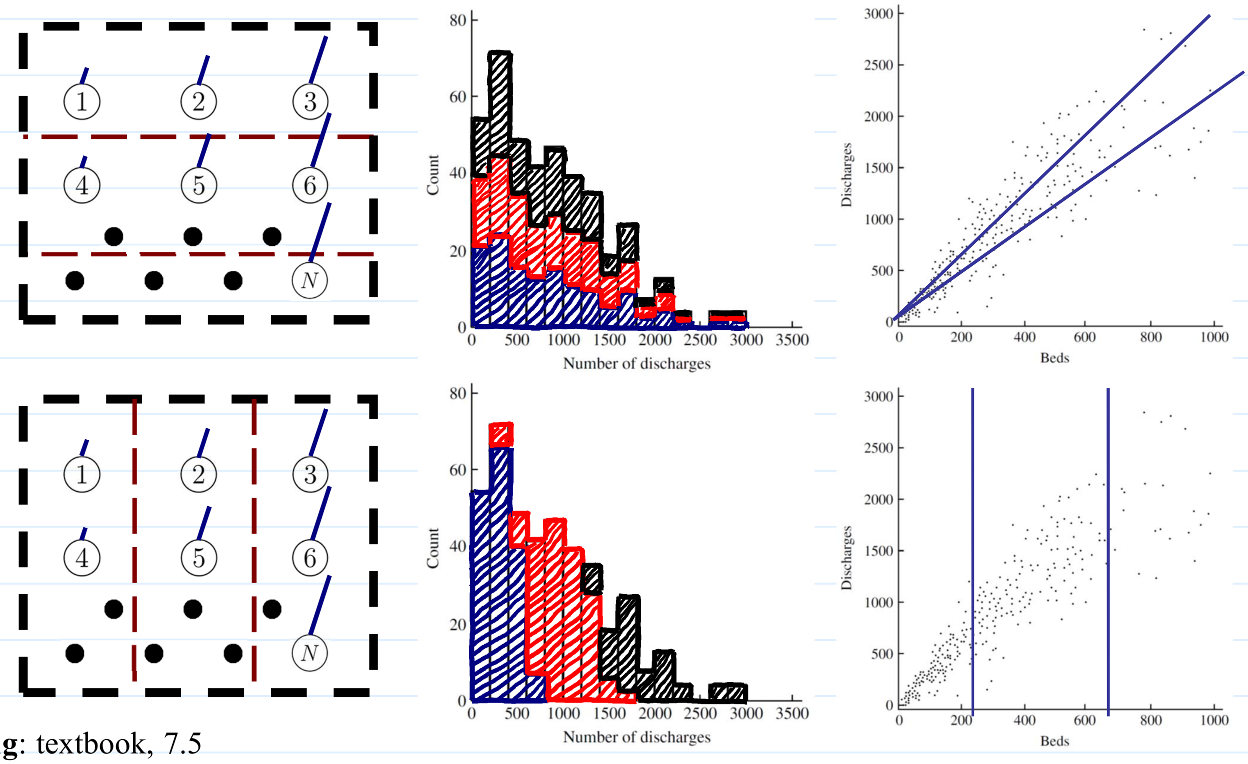
$$\underbrace{\frac{\sigma^2}{n}}_{\boxed{\overline{X}}} = \underbrace{\frac{1}{n}\sum_{l=1}^{L} W_l\,\sigma_l^2}_{\boxed{\overline{X}_{\mathbb{S},p}}} + \frac{1}{n}\sum_{l=1}^{L} W_l(\mu_l - \mu)^2$$

$$= \underbrace{\frac{\bar{\sigma}^2}{n}}_{\boxed{\overline{X}_{\mathbb{S},o}}} + \underbrace{\frac{1}{n}\sum_{l=1}^{L} W_l(\sigma_l - \bar{\sigma})^2}_{\boxed{\overline{X}_{\mathbb{S},p} \to \overline{X}_{\mathbb{S},o}}} + \underbrace{\frac{1}{n}\sum_{l=1}^{L} W_l(\mu_l - \mu)^2}_{\boxed{\overline{X} \to \overline{X}_{\mathbb{S},p}}}$$

  - $\boxed{\overline{X} \to \overline{X}_{\mathbb{S},p}}$: $\overline{X}_{\mathbb{S},p}$ much better than $\overline{X}$ if μ_l's are quite variable
  - $\boxed{\overline{X}_{\mathbb{S},p} \to \overline{X}_{\mathbb{S},o}}$: $\overline{X}_{\mathbb{S},o}$ much better $\overline{X}_{\mathbb{S},p}$ if σ_l's are quite variable
- The gain from $\overline{X} \to \overline{X}_{\mathbb{S},p}$ is often greater than the gain from $\overline{X}_{\mathbb{S},p} \to \overline{X}_{\mathbb{S},o}$.

- **Q**: Which one is a better way to form strata (i.e., to partition the population)?

$$\sigma^2 = \sum_{l=1}^{L} W_l \sigma_l^2 + \sum_{l=1}^{L} W_l (\mu_l - \mu)^2 \qquad \boxed{\overline{X} \to \overline{X}_{\mathbb{S},p}}$$



❖ **Reading**: textbook, 7.5

❖ **Further reading**: textbook, 7.6 (**systematic sampling**, **cluster sampling**, **practical difficult**)