Definition 8 (sample mean)

a function of data

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

cf. population mean: μ of F_0 data

$X \sim F_0$
$\mu = E(X) = \frac{1}{N}\sum_{i=1}^{N} x_i = \sum_{j=1}^{m}\left(\frac{n_j}{N}\right)\zeta_j$

Note 4 (Some notes about sample mean)

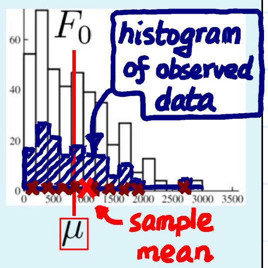
The statistic that contributes the most to the world.

- $\overline{X}$ is clearly a statistic, and hence a random variable.
- $\overline{X}$ is an intuitive estimator of μ. why?

In $x_1, \cdots, x_n$, observe
ζ_1 about $(n * \frac{n_1}{N})$ times
$\vdots$
ζ_m about $(n * \frac{n_m}{N})$ times
$\Rightarrow \sum_{k=1}^{n} X_k \approx n \times \sum_{j=1}^{m}\left(\frac{n_j}{N}\right)\zeta_j$

- In the dichotomous case, we have $\mu = p$ and population proportion $\hat{p} \equiv \overline{X}$ is the sample proportion. cf.

$\sum_{k=1}^{n} X_k$ = # of 1's in the sample

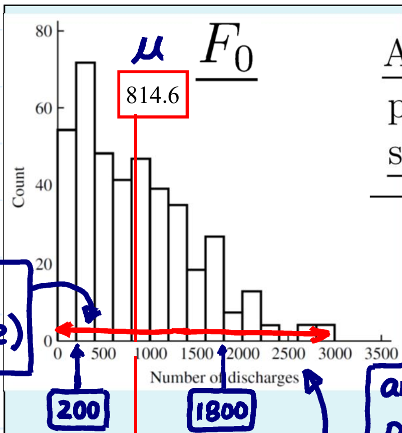


Example 3 (sampling distribution of sample mean, cont. Ex.2 in LNp.4)

- 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.

Q: What is the source of randomness in $\overline{X}$?
Ans: random sampling

$1 \vdots N$
$(\square, \square, \cdots, \square)$ 1st 2nd ... nth
$\{\square, \square, \cdots, \square\}$ 1st 2nd ... nth



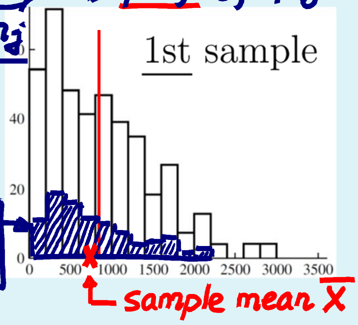
All $\binom{393}{16}$ possible samples

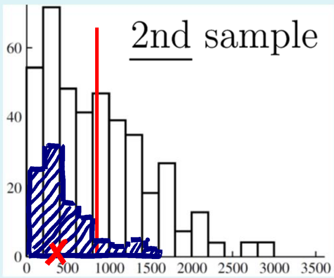
$\div N$ pmf of F_0
n_j

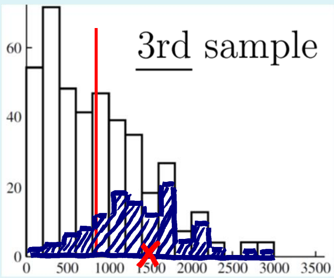
histogram of data

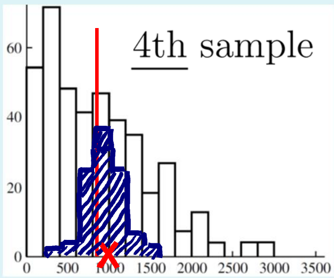
$\div n$

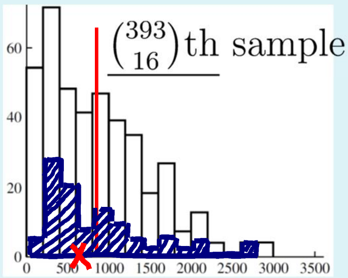
an empirical pmf of F_0, denoted by $\widehat{F_0}$ (a random function)



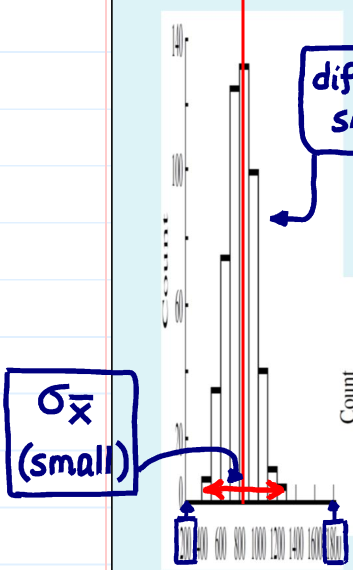




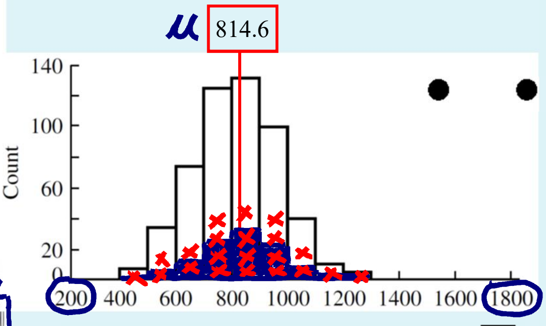




different shape



μ 814.6



• • •

sampling distribution of $\overline{X}$

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 $\overline{X}$.

perform an s.r.s on the $\binom{393}{16}$ samples

  – 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

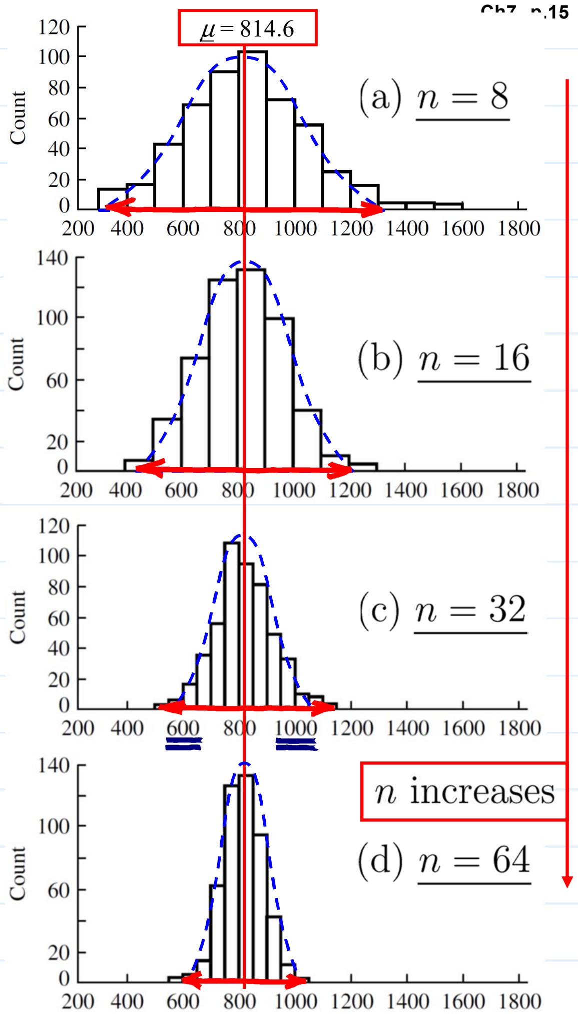
Why? This histogram will be an approximation to the sampling distribution of $\overline{X}$.

- Figure 7.2 (textbook) shows the results for sample size n=8, 16, 32, or 64.

Thm 1 LNp 16 – All the four histograms are centered at μ=814.6.

Thm 2-3 LNp.17-18 – As n increases, the histograms become less spread out.

Thm 12-13 LNp.29-30 – 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,

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

The kth observation. Note. $X_k \sim F_0$

population variance

population mean

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

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

parameter

So, $\overline{X}$ is an unbiased estimator of μ, i.e., the sampling distribution of $\overline{X}$ is *centered* at μ. Recall. graphs in LNp.15

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

LNp.11

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

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

Definition 3 in LNp.5

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

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

parameter

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

$n \to 4n$
$\sigma^*_{\overline{X}} \to \frac{1}{2}\sigma^*_{\overline{X}}$

Proof: Under simple random sampling with replacement, we have $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F_0$. ← LNp.11

∵ independent

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

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

	n	$\sigma^*_{\overline{X}}$	
+300	100	32	−16
+1200	400	16	−8
	1600	8	

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)

larger n, more accurate estimator

① same n → same precision
irrelevant to N
② No need to sample a certain proportion (i.e., n/N) of the population to reach same precision.

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 $(\sigma/\sqrt{n})\sqrt{1 - \frac{n-1}{N-1}}$.

What if $n = N$?

$\sigma^{*2}_{\overline{X}}$

parameter

a value between 0 and 1

$\sigma^*_{\overline{X}}$

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

$$Cov(X_k, X_l) = E(X_k X_l) - E(X_k)\,E(X_l)$$
$$= \left(\sum_{s=1}^{m}\sum_{t=1}^{m} \zeta_s \zeta_t P(X_k = \zeta_s, X_l = \zeta_t)\right) - \mu^2$$

$\zeta_s = \zeta_t$ iff $s = t$

$$\overset{\text{LNp.11}}{=} \left[\sum_{s=1}^{m} \zeta_s^2 \left(\frac{n_s(n_s-1)}{N(N-1)}\right) + \sum_{s=1}^{m}\sum_{t \ne s} \zeta_s\zeta_t \left(\frac{n_s n_t}{N(N-1)}\right)\right] - \mu^2$$

$= \zeta_s^2\left(\frac{n_s^2}{N(N-1)}\right)$ when $\zeta_s = \zeta_t$

When $\zeta_s = \zeta_t$ When $\zeta_s \ne \zeta_t$

$$= \left[\frac{N}{N-1}\sum_{s=1}^{m}\sum_{t=1}^{m} \zeta_s\zeta_t\left(\frac{n_s n_t}{N \cdot N}\right) - \frac{1}{N-1}\sum_{s=1}^{m}\zeta_s^2\left(\frac{n_s}{N}\right)\right] - \mu^2$$

$$= \frac{N}{N-1}E(X_k)\,E(X_l) - \frac{1}{N-1}E(X_k^2) - \mu^2$$

$$= \frac{N}{N-1}\cancel{\mu^2} - \frac{1}{N-1}(\sigma^2 + \cancel{\mu^2}) - \cancel{\mu^2} = \frac{-\sigma^2}{N-1}.$$ ← why negative?

$E(X_k) = \left(\sum_{s=1}^{m}\zeta_s \frac{n_s}{N}\right)$ × $\left(\sum_{t=1}^{m}\zeta_t \frac{n_t}{N}\right) = E(X_l)$

$\underbrace{Var(X_k)}_{\sigma^2} = E(X_k^2) - \underbrace{[E(X_k)]^2}_{\mu^2}$

Then,

$$\underline{Var}(\underline{\overline{X}}) = \underline{Var}\left(\frac{1}{n}\underline{\sum_{k=1}^{n} X_k}\right)$$

Note. $1 \le k < l \le n$

$= 0$ in with repl.

$$= \frac{1}{n^2}\underline{\sum_{k=1}^{n}}\underline{Var(X_k)} + \frac{2}{n^2}\underline{\sum_{k=1}^{n-1}}\underline{\sum_{l=k+1}^{n}}\underline{Cov(X_k, X_l)}$$

$$= \frac{1}{n^{\cancel{2}}} \times (\cancel{n}\sigma^2) + \frac{\cancel{2}}{n^{\cancel{2}}} \times \frac{\cancel{n}(n-1)}{\cancel{2}} \times \frac{-\sigma^2}{N-1} = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right).$$

$\frac{\sigma^2}{n} - \sigma^2 \times \overbrace{\frac{n-1}{n}}^{1} \times \frac{1}{N-1}$

$\approx \sigma^2/n - \sigma^2/N$

Why $\sigma^2_{\overline{X}}$ (without) $< \sigma^{*2}_{\overline{X}}$ (with) ?

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 $(1 - \frac{n-1}{N-1})$, which is called the **finite population correction**. (**Note**. $1 - \frac{n-1}{N-1} \to 1$ when $N \to \infty$)

N

- n/N: sampling fraction $\left(\approx \frac{n-1}{N-1}\right.$ in most cases$\left.\right)$ $\Rightarrow$ $1 - \frac{n-1}{N-1} \approx$ unsampled fraction
- $\sigma_{\overline{X}} \approx \sigma^*_{\overline{X}} = \sigma/\sqrt{n}$ if the sampling fraction is very small (i.e., $n \ll N$).

When $n \ll N$, without repl. $\approx$ with repl.

- $\sigma_{\overline{X}}$ 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

st.d. of F_0, usually unknown, need to estimate using data

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

where finite population correction $1 - \frac{31}{392} \approx 0.92$ makes little difference.

×: unknown in sampling survey

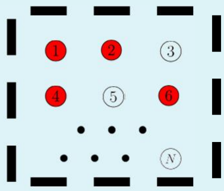
- Most of sample means differ from the population mean 814 by less than $2 \times \sigma_{\overline{X}} = 200$ (see graph (c) of Figure 7.2 in LNp.15).

$814 \pm 200 = (614, 1014)$

$Z \sim normal(\mu_Z, \sigma_Z^2)$
$P(Z \in \mu_Z \pm 2 \times \sigma_Z)$
≈ 0.95

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, $Var(\hat{p}) = \frac{p(1-p)}{n}$,

variance is a function of mean

check Note 1 in LNp.6

and $n\hat{p} = \sum_{k=1}^{n} X_k$ follows binominal(n, p) distribution

mean $= np$
variance $= np(1-p)$

$\because X_1, \cdots, X_n \overset{i.i.d.}{\sim}$ Bernoulli(p)

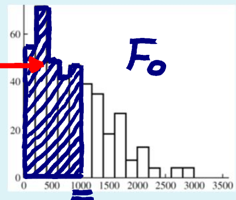
# of 1's in $X_1, \cdots, X_n$

- under s.r.s. without replacement, $Var(\hat{p}) = \frac{p(1-p)}{n}\left(1 - \frac{n-1}{N-1}\right)$,

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. parameter

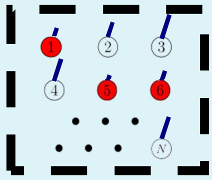


- $y_i = 1$ if $x_i < 1000$ and $y_i = 0$ if $x_i \geq 1000$

Data: $X_1, \cdots, X_n$. For $k = 1, \cdots, n$, $Y_k = I_{[0,1000)}(X_k)$

Indicator function
For a set A, $I_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$

- 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



×: unknown in sampling survey

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

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

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

known value · parameter · estimate

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

Theorem 5 (mean of population total estimator)

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

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

$E(T) = E(N\overline{X}) = NE(\overline{X}) = N\mu = \tau$

That is, T is an unbiased estimator of τ.

Theorem 6 (variance of population total estimator)

9/18

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

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

← $Var(\overline{X})$

$Var(T) = Var(N\overline{X}) = N^2\,Var(\overline{X})$

cf. the precision of $\overline{X}$
- Note 5 in LNp.17
- Note 6 in LNp.19

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

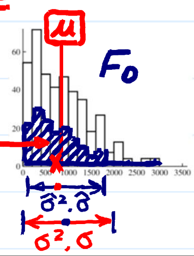
• Estimation of population variance

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

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?

histogram of data



Definition 10 (sample variance)

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

• a function of data • a r.v. • an estimator

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

Under s.r.s. with replacement, we have $E(\hat{\sigma}^2) = \sigma^2\left(\frac{n-1}{n}\right)$. ← not unbiased