Theorem 6 (variance of population total estimator)

9/18

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

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

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

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

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

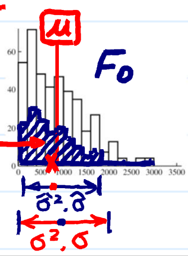
• Estimation of population variance

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

Recall. When F_0 is unknown, the σ in the standard error of $\bar{X}$ 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 $\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})^2$.

- 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

Proof. From the identity

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

$0 = +2\left[\sum_{k=1}^{n}(X_k - \bar{X})\right](\bar{X}-\mu)$

E E E

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

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

$E(X_k) = \mu$ $\quad$ $E(\bar{X}) = \mu$

which leads to $\boxed{Var(X_k)}$ $\boxed{Var(\bar{X})}$

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

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

Theorem 8 (expectation of sample variance, s.r.s. without replacement) not unbiased

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 $(\bigtriangledown)$ in the above proof still hold, and $(\bigtriangledown)$ leads to

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

$\boxed{Var(X_k)}$ $\boxed{Var(\bar{X})}$

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

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 $\frac{n-1}{n} \le 1$ and $\left(\frac{n-1}{n}\right)\left(\frac{N}{N-1}\right) \le 1$ (**note.** $n < N$), we have
  $$E(\hat{\sigma}^2) < \sigma^2.$$
  That is, $\hat{\sigma}^2$ tends to underestimate σ^2.

irrelevant to N

with repl.

without repl. $\frac{nN-N}{nN-n}$

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

$\frac{1}{n}$ (in $\hat{\sigma}^2$) $\approx \frac{1}{n-1}$ (in s^2) when n is large

$\hat{\sigma}^2$ s^2

$E(\hat{\theta}) = a\theta$ a known constant

$E\left(\frac{\hat{\theta}}{a}\right) = \theta$

although use same notation $S^2_{\overline{X}}$ different formula

Thm 2 (LNp.17)

depend on N

$\hat{\sigma}^2$ $\left(\frac{N-1}{N}\right)s^2$

Thm 3 (LNp.18)

cf.

sampling fraction

cf.

cf.

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
  $$s^2_{\hat{p}} = s^2/n = [\hat{p}(1-\hat{p})]/n-1.$$
- 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
  $$s^2_{\hat{p}} = \frac{s^2}{n}\left(1-\frac{n}{N}\right) = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1-\frac{n}{N}\right).$$

Note1 (LNp.6)

cf.

$= X_k$

sample variance

$= \frac{1}{n}\sum_{k=1}^n X_k = \overline{X}$

What is the without repl. version?

unbiased est'or of population variance under with repl.

functions of sample mean $\hat{p}$. When $\hat{p}$ is known, $\hat{\sigma}^2$, S^2 are known.

Thm 4 (LNp.20)

cf.

Note. same notation, different formula

cf.

Thm 4 (LNp.20)

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

An unbiased estimator of $Var(T) = N^2 Var(\overline{X})$ is $s^2_T = N^2 s^2_{\overline{X}}$.

$T = N\overline{X}$

Thm 6 (LNp.20)

unbiased estimator

without: $\frac{S^2}{n}\left(1-\frac{n}{N}\right)$

with: $\frac{S^2}{n}$

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

offer information about the magnitude of estimation error $\hat{\theta} - \theta$

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

discharge data: $x_1, \ldots, x_{50}$

- 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

sample mean; population mean; (unbiased) sample variance; population st.d.

$$s_{\overline{X}}^2 = \frac{s^2}{n}\left(1 - \frac{n}{N}\right) = \frac{614.53^2}{50}\left(1 - \frac{50}{393}\right) = 6592.$$

sampling fraction 12.72%

- The estimated standard error of $\overline{X}$ is $s_{\overline{X}} = \sqrt{6592} = 81.19$, → $2 \times s_{\overline{X}} = 162.38$

×: unknown in sampling survey

(**cf.** the (true) standard error of $\overline{X}$ is $\sigma_{\overline{X}} = \frac{590}{\sqrt{50}}\sqrt{1 - \frac{49}{392}} = 78$) ← $\sqrt{E(\overline{X}-\mu)^2}$ error

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. ← cf

- 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) ← population total
- 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)

population proportion

- $p = 0.654$: (true) proportion of hospitals in the population that had fewer than 1000 discharges ($\Rightarrow \sigma^2 = p(1-p) = 0.2263$).

population variance; $\sum_{k=1}^{50} I_{[0,1000)}(x_k)$

- 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

sample proportion

×: unknown in sampling survey; (unbiased) sample variance

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

- The estimated standard error of $\hat{p}$ is $s_{\hat{p}} = \sqrt{0.0045} = 0.067$, → $2 \times s_{\hat{p}} = 0.134$

$\sqrt{E(\hat{p}-p)^2}$ error → (**cf.** (true) standard error of $\hat{p}$ is $\sigma_{\hat{p}} = \sqrt{\frac{(.654)(.346)}{50}}\sqrt{1 - \frac{49}{392}} = 0.064$)

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. ← cf.

- 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 ① not only **form estimates** of unknown population parameters (e.g., use $\overline{X}$, T, $\hat{p}$ to estimate μ, τ, p, respectively), ② 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: parameter — statistic (random variable)

$F_0 \to$ population parameter	For ① → estimator(†)	variance of estimator(†)(*)	For ② → estimated variance(†)(*)
μ	$\overline{X}=\frac{1}{n}\sum_{k=1}^{n}X_k$	(a) $\sigma^2_{\overline{X}}=\frac{\sigma^2}{n}$ Smaller (Why?) (b) $\sigma^2_{\overline{X}}=\frac{\sigma^2}{n}\left(1-\frac{n-1}{N-1}\right)$	(a) $s^2_{\overline{X}}=\frac{s^2}{n}$ (b) $s^2_{\overline{X}}=\frac{s^2}{n}\left(1-\frac{n}{N}\right)$
p	$\hat{p}$ = sample proportion	(a) $\sigma^2_{\hat{p}}=\frac{p(1-p)}{n}$ (b) $\sigma^2_{\hat{p}}=\frac{p(1-p)}{n}\left(1-\frac{n-1}{N-1}\right)$	(a) $s^2_{\hat{p}}=\frac{\hat{p}(1-\hat{p})}{n-1}$ (b) $s^2_{\hat{p}}=\frac{\hat{p}(1-\hat{p})}{n-1}\left(1-\frac{n}{N}\right)$
τ	$T=N\,\overline{X}$	(a), (b) $\to \sigma^2_T=N^2\,\sigma^2_{\overline{X}}$	(a), (b) $\to s^2_T=N^2\,s^2_{\overline{X}}$
σ^2	(a) $s^2=\frac{1}{n-1}\sum_{k=1}^{n}(X_k-\overline{X})^2 \to \boxed{=\frac{n\hat{p}(1-\hat{p})}{n-1}, \text{ dichotomous case}}$ (b) $\left(1-\frac{1}{N}\right)s^2$		

Finite population correction

related to sampling fraction n/N

When $n \ll N$, almost identical

- (†): (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**.