• Normal approximation to the sampling distribution of sample mean

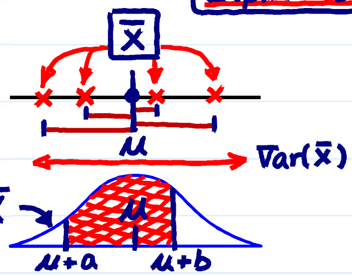
Recall. the shape of $F_{\bar{X}}$ in LNp.14~15

We have known
• $E(\bar{X})=\mu$
• $Var(\bar{X}) \to 0$, $n\to\infty$

how about the shape of $F_{\bar{X}}$?

Recall, dichotomous case (Thm4, LNp 20)

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



Theorem 12 (central limit theorem, CLT, for i.i.d. case) ← S.R.S with repl.

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 $\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$, we have $E(\bar{X}_n)=\mu$, $\sigma^2_{\bar{X}_n} = Var(\bar{X}_n) = \sigma^2/n$, and for any fixed value z,

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

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

$\frac{\bar{X}_n-\mu}{\sigma_{\bar{x}}} \overset{D}{\approx} N(0,1)$

cdf of $\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}$

standardization

Then, mean = 0, var = 1

shape of $F_{\bar{x}}$

shape of $F_{\bar{x}}$ (depend on shape of F_0) — n small

shape of $F_{\bar{x}}$ (irrelevant to F_0) — normal pdf — n large

μ $Var(\bar{x})$

(★) $\frac{\bar{X}_n-\mu}{S_{\bar{x}}} \overset{D}{\approx} N(0,1)$

(**cf.**) Law of large number (LLN) guarantees that $\bar{X}_n \xrightarrow{P} \mu$ and $s^2 \xrightarrow{P} \sigma^2$ ← TBp.179~180
as $n\to\infty$, i.e., $\bar{X}_n$ 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 $\bar{X}_n$ is approximately normally distributed with mean μ and variance $\sigma_{\bar{X}_n}$ (check graphs in Ex.3, LNp.15).

$0 \ll n \ll N$ — CLT ← — without ≈ with

$= (\sigma/\sqrt{n})\times\sqrt{1-\frac{n-1}{N-1}}$, not $\sigma/\sqrt{n}$

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

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

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

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

|error|

$1-\Phi(\delta/\sigma_{\bar{x}})$

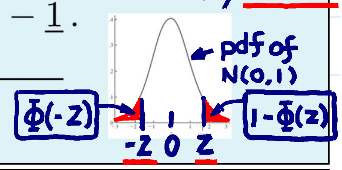
$\overset{D}{\approx} N(0,1)$

usually unknown

$\sigma_{\bar{x}}$ as a standard

$\pm\delta$	$\pm 2\sigma_{\bar{x}}$	$\pm 1.96\sigma_{\bar{x}}$	$\pm 1.64\sigma_{\bar{x}}$	$\pm 1\sigma_{\bar{x}}$
Φ	0.954	0.950	0.900	0.685

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

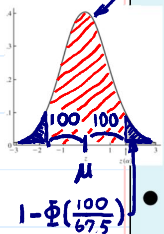


Example 9 (probability of estimation error more than δ, cont. Ex.2 in LNp.4)

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

×: unknown in sampling survey

pdf of $\overline{X}$



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

$$P(|\overline{X}-\mu| > 100) = 2\times P(\overline{X}-\mu > 100) = 2\times P\left(\frac{\overline{X}-\mu}{\sigma_{\overline{X}}} > \frac{100}{\sigma_{\overline{X}}}\right)$$

$$\approx 2\left[1-\Phi(100/67.5)\right] = 2\times 0.069 = 0.14.$$

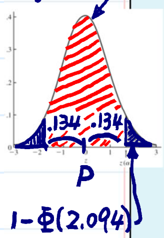
$N(0,1) \overset{D}{\approx}$; $\overset{D}{\sim} N(0, \sigma^2_{\overline{X}})$; cf. ; ×: unknown in sampling survey

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

$n\hat{p} \sim$ binomial (with) or hypergeometric (without) ← Thm4 LNp.20

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

pdf of $\hat{p}$



- sample size $n = 50$, true $p = 0.654$, standard error of $\hat{p}$ is $\sigma_{\hat{p}} = 0.064$. estimate
- 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

estimator →
$$P(|\hat{p}-p| > 0.134) = 1 - P\left(|\hat{p}-p| \le 0.134\right)$$

Recall. Normal approximation to binomial (TBp.187)

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

$\overset{D}{\sim} N(0, \sigma^2_{\hat{p}})$; $N(0,1) \overset{D}{\approx}$

×: unknown in sampling survey

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

(✱) in LNp.29 → **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.,

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

cdf of $(\overline{X}-\mu)/s_{\overline{X}}$; $N(0,1) \overset{D}{\approx}$; (a) with (b) without

Definition 11 (interval estimator, coverage probability, interval estimate, confidence interval, and confidence level) ← Why construct confidence interval?

- 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*

Data

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

where

statistics (r.v.'s)

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. $P(\theta \in (\hat{\theta}_L(\mathbf{X}), \hat{\theta}_U(\mathbf{X}))) = 1-\alpha$.

θ is contained in the interval estimator

not-covered probability

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

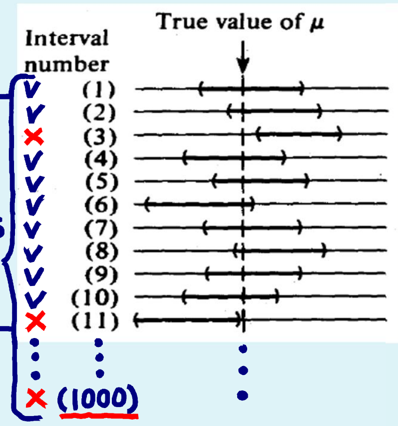
observed data

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

either contains μ or not, i.e., probability of containing μ is 1 or 0

is called an **interval estimate**.

Repeated construction of 95% confidence intervals



- 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.

significance level α in testing ; cf.

- The $100(1-\alpha)\%$ is also referred to as **confidence level**.

90% ; 95% ; 99%

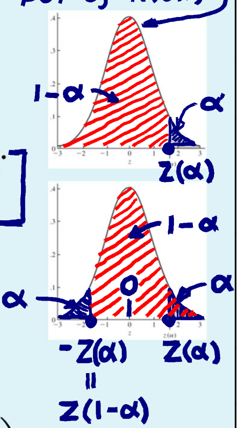
- **Note.** The α is usually assigned a small value, e.g. 0.1, 0.05, or 0.01.

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

quantile 分位數

- 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 $z(1-\alpha) = -z(\alpha)$.

pdf of N(0,1)



We know this so that we can construct C.I.

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

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

usually unknown, assume known here

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

(asymptotic) pivotal quantity of μ

if σ² known; $\frac{\bar{X}_n - \mu}{\sigma_{\bar{X}_n}} \overset{D}{\approx} N(0,1)$

★ $\frac{\bar{X}-\mu}{S_{\bar{X}}} \overset{D}{\approx} N(0,1)$ by (★) in LNp.29

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

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

×: unknown in sampling survey

check Def. 11, 1,2,3 in LNp.32 → $\bar{X}_n \pm z(\alpha/2)\sigma_{\bar{X}_n}$ ← replaced by $S_{\bar{X}}$

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. a r.v., but not a statistic

Note 9 (Some notes about confidence interval)

- In a sample survey, $\sigma_{\bar{X}_n}$ is unknown. In the case, $s_{\bar{X}_n}$ (or s^2, respectively) can be substituted for $\sigma_{\bar{X}_n}$ (or σ^2, respectively) if the sample size n is large enough, say $n \ge 25$ or 30 by a rule of thumb.

TBp.338

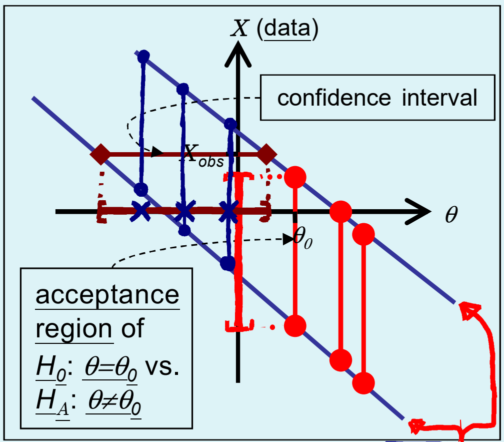
- **Recall**: duality between confidence interval and hypothesis testing.
  - Suppose for every parameter value θ_0, there is a level-α test for

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

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

    a set of parameter values → $$C(\mathbf{X}) = \{\theta \,|\, \mathbf{X} \in AR(\theta)\}$$

    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

    can use C.I. to perform a test ← $$AR(\theta_0) = \{\mathbf{X} \,|\, \theta_0 \in C(\mathbf{X})\}.$$

pivotal quantity → $\left|\frac{\bar{X}-\mu}{\sigma_{\bar{X}}}\right| < z(\frac{\alpha}{2})$ in LNp 33

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

cf.

$$\underbrace{\left|(\bar{X}_n - \mu_0)/\sigma_{\bar{X}_n}\right|}_{\text{distance difference}} > \underbrace{z(\alpha/2)}_{\text{scale marked on a ruler}}$$

known value

- Many confidence intervals have the form:

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

$\hat{\theta}$

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

- The width of a confidence interval often depends on:

Under same α, smaller width $\Leftrightarrow$ more accurate C.I.

– n : sample size

$n \uparrow$, width $\downarrow$

– σ: population standard deviation

$\sigma \uparrow$, width $\uparrow$

– $1-\alpha$: confidence level

$(1-\alpha)\uparrow$, width $\uparrow$

e.g., use previously collected information of population

For example, consider the C.I.:

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

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

S

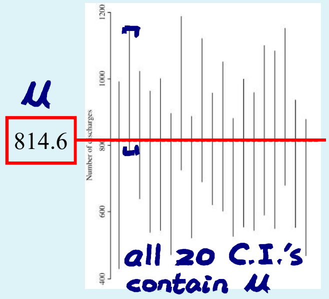
- If α is fixed and σ is (approximately) known, n can be chosen so as to obtain confidence intervals close to some desired length. → i.e., use estimated st.e. to determine n

$\Rightarrow$ 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.

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. x_i, $i = 1, \cdots, 8000$
- The sample yields that → Data: $X_1, \cdots, X_{100}$

parameter

– the average number of motor vehicles per unit is $\overline{X} = 1.6$, $\xrightarrow{\text{estimate}} \mu$

– with a sample standard deviation $s = 0.8$. ← s^2 estimates population variance

– So,

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

→ shows accuracy of $\overline{X}$

$s_{\overline{X}}^2$ estimates the variance of $\overline{X}$

- 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

C.I. combines 2 information

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

Why?

an interval estimate: a collection of many possible μ's

- For the population total $\tau = N\mu$ (i.e., total number of motor vehicles owned by the 8000 units),

parameter

– an estimate of τ is $T = N \times \overline{X} = 8000 \times 1.6 = 12,800$,

cf.

– with an estimated standard error $s_T = N \times s_{\overline{X}} = 640$. → shows accuracy of T

- So, a 95% confidence interval for τ is

$8000 \times (1.44, 1.76)$

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

← an interval estimate

Why?

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

$8000 \times (0.06, 0.18)$

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

$n=100$, 509, $N\hat{p}$; $n^*=?\ (>n)$, 200, $N\hat{p}$

- Suppose a 95% C.I. of Np with a half-width of 200 is desired (cf., original half-width: (1469 − 451)/2=509).

why can we do this? ⇒ assume $n^* \ll N$

- 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

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

previously collected information; s.r.s. with

Why use s.r.s. with? ⇒ easier to solve for n^*

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

$\ll N = 8000$

$\frac{649}{100} = 6.51$倍 $\approx \left(\frac{509}{200}\right)^2$ ∵ $s_{\hat{p}} \propto 1/\sqrt{n}$

❖ **Reading**: textbook, 7.3