ST2132 Survey Sampling II

Interval estimation

Semester 1 2022/23

From point estimation to interval estimation

- We have seen how to use IID data to estimate a population mean or proportion, and to calculate an approximate SE for the estimate. This can be done for any sample size n.
- A confidence interval can be constructed using a formula, if
  (i) the population has a normal distribution. Or
  (ii) n is large, thanks to the Central Limit Theorem.
  For a real population, also need $n \ll N$, so that SRS is like sampling with replacement.
- Key concepts: random interval, confidence interval, bias, MSE

Normal approximation

$E(X_i)$ $\mathrm{var}(X_i)$

Let $X_1, \ldots, X_n$ be IID RV's with mean μ and variance σ^2. As $n \to \infty$, the distribution of

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

$E = 0$

$\mathrm{var} = 1$

converges to N(0,1).

- For sufficiently large n,

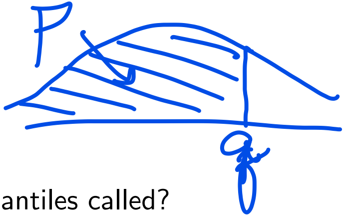
$$\boxed{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1) \qquad \text{approximately}}$$

$\Pr(-1.96 \le Z \le 1.96) = 0.95$

In particular, $\Pr\left(-1.96 \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le 1.96\right) \approx 0.95$.

- How large should n be? No fixed answer, unless an error margin is specified.

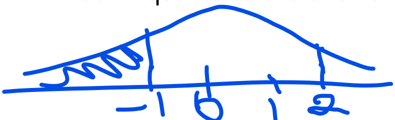
Suppose X has a strictly increasing CDF F. For $0 < p < 1$, the p-quantile of X is the number q such that



$$F(q) = \Pr(X \leq q) = p$$

Hence $q =$ $\underline{F^{-1}(p)}$.

- What are the 0.25-, 0.50- and 0.75-quantiles called?
- Let $Z \sim$ N(0,1). The p-quantile of Z can be written as $\underline{\Phi^{-1}(p)}$.
- What are the 0.4-, and 0.8-quantiles of Z? [qnorm()]
- What Z quantiles are the values -1, 2? [pnorm()]

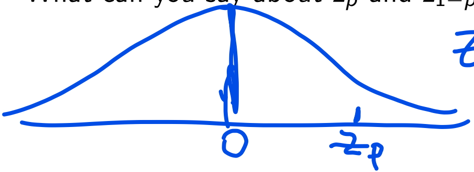


For $0 < p < 0.5$, let z_p be such that

$$\Pr(Z > z_p) = p$$

- $z_p = \underline{(1-p)}$-quantile of Z.
- Express z_p in terms of Φ. $z_p = \Phi^{-1}(1-p)$.
- What are the values of $z_{0.1}$ and $z_{0.05}$? [qnorm()]
- What can you say about z_p and z_{1-p}?

$z_{1-p} = -z_p$



Approximate distribution of $\bar{X}$

$$\Pr\left(-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

Let $0 < \alpha < 1$. Fo large n,

$$\Pr\left(-z_{\frac{\alpha}{2}} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{\frac{\alpha}{2}}\right) \approx 1-\alpha$$

- Consequently,

$$\Pr\left(\mu - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \bar{X} \le \mu + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \approx 1-\alpha$$

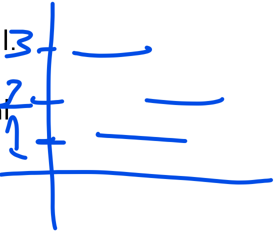
- Approximately, $\bar{X} \sim$ $N\left(\mu, \frac{\sigma^2}{n}\right)$.

$0 < \alpha < 1$. n large.

- Show that

$$\Pr\left(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \approx 1 - \alpha$$

$\left(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$ is a random interval.



- A realisation $\bar{x}$ of $\bar{X}$ gives the realised interval

$$\left(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

Imagine generating many such intervals, and marking the i-th interval on the line $y = i$.
How does this picture illustrate the meaning of the probability statement?

Confidence interval for μ

- Suppose μ is unknown but σ is known. An approximate $(1-\alpha)$-confidence interval for μ is

$$\left(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

- Almost always σ is also unknown. An approximate $(1-\alpha)$-CI for μ is

$$\left(\bar{x} - z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right)$$

- Since $s/\sqrt{n}$ is the estimated SE, we can write the $(1-\alpha)$-CI for μ in the form

$$\left(\text{estimate} - z_{\frac{\alpha}{2}}\text{SE}, \text{estimate} + z_{\frac{\alpha}{2}}\text{SE}\right)$$

Example 1 (Sample Survey I slide 17)

$n = 400$, $\bar{x} = 3531$ g, $s^2 = 225700$ g^2. μ is estimated as 3531 g, SE is estimated as $s/\sqrt{n} \approx 24$ g.

- An approximate 95%-CI for μ is

$$(3531 - 1.96 \times 24, 3531 + 1.96 \times 24) \approx (3484, 3578)$$

- Is it true?

$$\Pr(3484 \leq \mu \leq 3578) \approx 0.95$$

- Think of many CI's from independent SRS of size 400. What can you say about them?

Example 2 (Survey Sampling I slide 22)

$n = 100$, p is estimated as 0.78, SE is estimated as $\sqrt{0.78 \times 0.22)}/\sqrt{n} \approx 0.04$.

- For $\alpha = 0.1$, $z_{\frac{\alpha}{2}} \approx 1.64$. An approximate 90%-CI for p is

$$(0.78 - 1.64 \times 0.04, 0.78 + 1.64 \times 0.04) \approx (0.71, 0.85)$$

- Is it true?

$$\Pr(0.71 \leq p \leq 0.85) \approx 0.90$$

- How can a CI for p be interpreted?
- The $(1-\alpha)$-CI for p:

$$\left(\text{estimate} - z_{\frac{\alpha}{2}}\text{SE}, \text{estimate} + z_{\frac{\alpha}{2}}\text{SE}\right)$$

Examples: hypothetical populations

- (Survey Sampling I slide 25) NB 10 weighs 10 g $-$ w μg. Using 100 measurements, w was estimated as 404.6 ± 0.6 μg.

  An approximate 95%-CI for w is

- (Survey Sampling I slide 28) For Kerrich's coin, the probability of head is p. Using 10000 tosses, p was estimated as 0.507 ± 0.005.

  For $\alpha = 0.01$, $z_{\frac{\alpha}{2}} \approx 2.58$. An approximate 99%-CI for p is

Summary on "large-sample CI" (1)

Assume the data are realisations from IID RV's $X_1, \ldots, X_n$ with expectation μ or p (Bernoulli RV). Suppose n is large.

- An approximate $(1-\alpha)$-CI for μ or p is

$$\left(\text{estimate} - z_{\frac{\alpha}{2}}\text{SE}, \text{estimate} + z_{\frac{\alpha}{2}}\text{SE}\right)$$

- For a real population of size N, if n/N is not small, the method works, with corrected SE (multiply by $\sqrt{\frac{N-n}{N-1}}$).
- For a hypothetical population, $N = \infty$, so correction is irrelevant. Seems like studying an infinite population is easier?

Summary on “large-sample CI” (2)

- Confidence level is approximately $(1-\alpha)$, because
  (i) normal approximation is used.
  (ii) almost always, the SE is estimated.

- If n is small, the confidence level is typically less than $1-\alpha$. The actual level can be estimated by simulation: not in syllabus.

- If another probability sampling method is used, CI makes sense, but a different method is needed. Not in syllabus.

- CI method does not take care of sampling bias, such as in a convenience sample.

Normal data: exact CI for μ

Let $t_{\frac{\alpha}{2},n-1}$ be the number such that $\Pr(t_{n-1} > t_{\frac{\alpha}{2},n-1}) = \alpha/2$.

Let $x_1, \ldots, x_n$ be realisations from IID $\mathrm{N}(\mu, \sigma^2)$ RV's $X_1, \ldots, X_n$, with mean $\bar{x}$ and sample SD s.

- A $(1-\alpha)$-CI for μ is

$$\left(\bar{x} - t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}\right)$$

  This works for any sample size $n > 1$.

- What does it mean to say the CI is exact?

- In practice, we do not know with certainty if a population is normal, so this CI is also approximate.

Bias in survey

If a convenience sample, there is no good method for CI.

- To estimate the proportion of votes for Alfred Landon in the 1936 US presidential election, *Literary Digest* asked 10 million of its subscribers.

  2.4 million responded, of whom 0.57 favoured Landon.

- By the formulae, the estimate is 0.57, and the estimated SE is $\sqrt{0.57 \times 0.43/2400000} \approx 0.00$. But Landon got only 38% of the votes.

  The formulae went wrong, partly because of sampling bias.

Bias in measurement

Survey Sampling I: the weight of NB 10 is 10 g $-$ w μg. w was estimated assuming the measurements had no bias.

- Suppose the $x_1, \ldots, x_n$ are realisations of random draws $X_1, \ldots, X_n$ from a population with mean $w + b$ and variance σ^2. The bias b is a constant.
- Now the SE $\sigma/\sqrt{n}$ measures how far $\bar{x}$ is from $w + b$, not w. $\mathrm{E}(\bar{X}) = w + b$.
- If $b \neq 0$, $\bar{x}$ is a biased estimate of w. The estimated SE and the CI are misleading.
- Measurement bias is unlikely to be removed by smart manipulation of data. It is quite essential to use known standards to estimate bias.

MSE

The MSE of $\bar{X}$ as an estimator of w is

$$\begin{aligned} \mathrm{E}(\bar{X} - w)^2 &= \mathrm{var}(\bar{X}) + \{\mathrm{E}(\bar{X}) - w\}^2 \\ &= \frac{\sigma^2}{n} + b^2 \end{aligned}$$

"MSE $=$ SE^2 $+$ bias^2"

- As $n \to \infty$, MSE approaches b^2. Bias does not go away with infinite data, just like estimating support for Landon.
- If $b = 0$, then MSE $=$ SE^2.
- CI does not take care of measurement bias. Correcting it takes more careful observations than smart computations.

September 2011

- Researchers reported neutrinos that took 61 nanoseconds less than light would have taken to travel 732 km.
- Wikipedia *Faster-than-light neutrino anomaly*: "...two flaws in their equipment set-up that had caused errors far outside their original confidence interval...".
- Apparently, the scientists were convinced by their (very small) CI for measuring time, and neglected to consider bias.

On parameters

- The mean or SD of a large real population, is practically unknowable. It can be determined exactly via a census, which seeks every individual's value. A census takes a lot of resources.

- A parameter of a hypothetical population seems unknowable in principle. If there is no bias, MSE decreases with more samples, but is never 0. Bias makes it worse.

Looking forward

Current framework of statistical inference:

1. Parameter is a simple function of the population, real or hypothetical.
2. Data are realisations of IID RV's (if $n \ll N$ for a real population).
3. Estimate is a realisation of an estimator, whose SD is the SE. For large n, can construct CI.
4. $\text{MSE} = \text{SE}^2 + \text{bias}^2$.

General case: Parameter may not be a simple function of population, so need methods to construct estimators.