ST2132 Survey Sampling I

Point estimation

Semester 1 2022/2023

If printing, do DOUBLE-SIDED, each side TWO slides.

Why this topic?

- Useful skill
- Instructive prototype of estimation and statistical modeling

Key concepts: parameter, simple random sampling, estimate, estimator, standard error, hat notation.

Introduction

Suppose we want to know the average amount of daily exercise in a city, or the percentage of diabetic patients in a country.

- If the population is large, we study a small part: a sample, then try to infer about the population.
- The best sampling methods use chance carefully. Such data can be analysed using appropriate random variables.
- We focus on using the simple random sample to estimate the mean of a large population.

Populations and parameters

- We are interested in a variable v in a population of N individuals, where individual i has a fixed value v_i.
- Mean and variance of v:

$$
\begin{aligned}
\mu &= \frac{1}{N}\sum_{i=1}^{N} v_i \\
\sigma^2 &= \frac{1}{N}\sum_{i=1}^{N}(v_i - \mu)^2
\end{aligned}
$$

  These are called **parameters**. The SD σ is another one.

- Parameters are quantities calculated from all individuals in the population.

Mean and proportion

Examples from the Singapore Census in 2020.

- Among resident ever-married females age 50 years or more, the mean number of children is 2.44. Between 40 and 49 years, the mean is 1.76.
- Among residents age 25 years or more, 33% had university qualification.
- In a population, let p be the proportion of category c. In the derived population where c is replaced by 1, and any other category is replaced by 0,

$$\mu = p, \qquad \sigma^2 = p(1-p)$$

One random draw

Let X be the result of a random draw: every individual has equal probability of being chosen.

- What can you say about its distribution, expectation and variance?

$$\mathrm{E}(X) = \underline{\hspace{2cm}}, \qquad \mathrm{var}(X) = \underline{\hspace{2cm}}$$

- X is around _____, give or take _____ or so.

  In this sense, a random sample is **representative** of the population. Not intuitive: a realisation could be far from μ.

E(X) and var(X)

Let M be the index of the chosen individual. Then $X = v_M$.

- What is the distribution of M?
- $X = g(M)$, where $g(m) =$ _____.
- Now compute E(X) and var(X).

Is X uniformly distributed on $v_1, \ldots, v_N$?

Simple random sampling (SRS)

SRS of size n: make n random draws without replacement.

- Let the results be denoted $X_1, \ldots, X_n$. For $i = 1, \ldots, n$,

$$E(X_i) = \mu, \qquad \mathrm{var}(X_i) = \sigma^2 \tag{1}$$

- Intuitively, are X_1 and X_2 correlated positively or negatively?

- For $i \neq j$,

$$\mathrm{cov}(X_i, X_j) = -\frac{\sigma^2}{N-1} \tag{2}$$

Justification of (1) and (2)

Let $M_1, \ldots, M_N$ be the successive random indices.

Theorem: $(M_1, \ldots, M_N)$ is uniformly distributed on all permutations of $\{1, \ldots, N\}$.

Proof: combinatorics!

- Theorem implies M_i is uniformly distributed on $\{1, \ldots, N\}$, hence (1) holds.
- Theorem implies (M_i, M_j) is uniformly distributed on ________________________. (2) follows from calculating $\mathrm{cov}(X_1, X_2)$.

$M_1, \ldots, M_N$ and $X_1, \ldots, X_N$ are examples of exchangeable RV's.

Exchangeable RV's

- RV's $Y_1, \ldots, Y_k$ are exchangeable if all reordered vectors have the same distribution as $(Y_1, \ldots, Y_k)$.

  I.e., for any permutation π on $\{1, \ldots, K\}$,

  $$(Y_{\pi(1)}, \ldots, Y_{\pi(k)}) \stackrel{d}{=} (Y_1, \ldots, Y_k)$$

- Are IID RV's exchangeable?
- Are exchangeable RV's IID?
- Are exchangeable RV's identically distributed?

SRS mean

Let $X_1, \ldots, X_n$ be SRS from a large population of size N, with mean μ and variance σ^2.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$\mathrm{E}(\bar{X}) = \mu, \qquad \mathrm{var}(\bar{X}) = \frac{N-n}{N-1} \frac{\sigma^2}{n}$$

If $n \ll N$, $\frac{N-n}{N-1} \approx 1$, so an SRS is like draws with replacement.

Analysis strategy for $n \ll N$

We assume that $X_1, \ldots, X_n$ are IID.

- To a high degree of accuracy,

$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}$$

- If concerned, multiply by correction factor

$$\frac{N - n}{N - 1}$$

Try $N = 5,000,000$, $n = 2,500$.

Estimate and estimator of μ, $n \ll N$

- Given data $x_1, \ldots, x_n$, estimate μ with the natural

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- $\bar{x}$ is an **estimate** of μ.
- $\bar{x}$ is a realisation of the **estimator** $\bar{X}$.
- $\bar{x}$ has an error of

$$\mu - \bar{x}$$

  Can the error be calculated? Can it be estimated?

Quantifying error in estimate of μ, $n \ll N$

- Quantify error in $\bar{x}$ by the **standard error**:

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}$$

- SE is defined as the SD of the estimator.

  Hence we use $\mathrm{SD}(\bar{X}) = \sigma/\sqrt{n}$.

  Since it is impossible to estimate $\bar{x} - \mu$, we settle for $\sigma/\sqrt{n}$, which indicates how much $\bar{X}$ fluctuates around μ.

- Can the SE be calculated? Can it be estimated?

How to estimate σ?

Bootstrap idea: use the data $x_1, \ldots, x_n$.

- Intuitive estimate of σ^2:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- "Sample variance" is preferred:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Why?

Estimation snapshot ($n \ll N$)

A population of size N has mean μ and variance σ^2, both unknown.

To estimate μ, we use an SRS $x_1, \ldots, x_n$.

- μ is estimated by $\bar{x}$.
- Error in $\bar{x}$ is measured by the SE: $\frac{\sigma}{\sqrt{n}}$.
- SE is estimated by $\frac{s}{\sqrt{n}}$.
- Conclusion: μ is estimated as $\bar{x}$, give or take $\frac{s}{\sqrt{n}}$.

Example 1

For 100000 newborn babies, their weights have mean μ g and variance σ^2 g^2. For 400 randomly selected babies:

$$\frac{1}{400}\sum_{i=1}^{400} x_i \approx 3531 \text{ g}, \qquad \frac{1}{400}\sum_{i=1}^{400}(x_i - 3531)^2 \approx 225136 \text{ g}^2$$

- μ is estimated as 3531 g.
- σ^2 is estimated by $s^2 = \frac{400}{399} \times 225136 = 225700$ g^2.
  SE is estimated as
  $$\frac{\sqrt{225700}}{\sqrt{400}} \approx 24 \text{ g}$$
- Conclusion: μ is estimated as 3531 g, give or take 24 g.

Using $\hat{\sigma}$ gives effectively the same SE estimate: $\sqrt{225136}/20 \approx 24$.

SE

- The estimate for μ, $\bar{x}$, is a realisation of $\bar{X}$. By definition,

$$\text{SE} = \text{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

- The estimate $\bar{x}$ is around μ, give or take SE or so.

  SE is fixed, but unknown. The need to estimate it gives the impression that it varies.

- Example 1: μ is estimated as 3531 ± 24 g. The study is replicated, and gives a new estimate 3512 ± 26 g.

  24 g and 26 g are both estimates of the SE.

Important terms

- μ, σ^2, σ are **parameters**.
- The **estimates** $\bar{x}, s^2, s$ are realisations of respective random variables $\bar{X}$, S^2, S, called **estimators**.
- The SE of $\bar{x}$ is SD($\bar{X}$). How about the SE of s^2 and s?
- Because $\text{E}(\bar{X}) = \mu$, $\bar{X}$ is an **unbiased** estimator, $\bar{x}$ is an unbiased estimate. Is s^2 or s unbiased?

On sampling assumption

- The estimation method above works well for SRS, provided $n \ll N$. Then $X_1, \ldots, X_n$ are effectively IID RV's.
- If n/N is relatively large, modify SE by correction factor $\sqrt{\frac{N-n}{N-1}}$.
- If data come from another probability sampling method, the formulae may not work, but other methods are avaiable.
- If a convenience sample, no estimation method can be justified.

Estimating proportion ($n \ll N$)

Although a proportion is a special case of a mean, it is important enough to be treated separately.

In a 0-1 population, let p be the proportion of 1's.

- $\mu = p, \sigma^2 = p(1-p)$.
- $\bar{x}$ is the realised proportion of 1's in the sample, which estimates p.
- SE is
$$\frac{\sqrt{p(1-p)}}{\sqrt{n}}$$
estimated by replacing p with its estimate.

Example 2

A box has 10000 tickets; a proportion p are white.
SRS of size 100 has 78 white tickets.
Estimate p, and estimate the SE.

- p is estimated as 0.78.
- SE is estimated as

$$\frac{\sqrt{0.78 \times 0.22}}{\sqrt{100}} \approx 0.04$$

- Conclusion: p is estimated as 0.78 ± 0.04.
- Since $K = 10000p$, K is estimated as 7800 ± 400 (Introduction slide 3).

Behind the scene

- The estimator is the random proportion of 1's in the SRS, denoted by $\hat{p}$.

$$E(\hat{p}) = p, \qquad \text{var}(\hat{p}) = \frac{p(1-p)}{n}$$

- By definition, SE is

$$SD(\hat{p}) = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

which estimated by replacing p with the realisation of $\hat{p}$: the realised proportion of 1's.

Hat notation

- An estimator of p is denoted by $\hat{p}$, a random variable.

  Advantage: we can directly see the parameter to be estimated. Similarly, since $\bar{X}$ is an estimator of μ, we write $\hat{\mu} = \bar{X}$.

  Disadvantage: no small letter for realisation.

- Sometimes, a hat-symbol denotes a realisation. We will avoid it, by using words and numbers, like in slide 22. If it cannot be avoided, we shall make it very clear.

Example 3: Measurement

The US National Bureau of Standards has been measuring the weight of a checkweight NB 10 every week since the 1940's.

100 measurements have a mean of 404.6 μg below 10 g, and an SD of 6 μg. Estimate the weight of NB 10, and estimate the SE.

We estimate w, the amount that NB 10's weight is below 10 g.

The measurements of w, $x_1, \ldots, x_{100}$, have errors.

$$x_i = w + e_i, \qquad 1 \leq i \leq 100$$

Can the exact value of w be found?

Estimation of w

Assume the x's are realisations of random draws $X_1, \ldots, X_{100}$ from an imaginary population with mean w and variance σ^2.

- w is estimated as $\bar{x} = 404.6\ \mu\text{g}$.
- σ estimated as $s = \frac{\sqrt{100}}{\sqrt{99}} \times 6 \approx 6$.
  SE $= \sigma/10$ estimated as $6/10 = 0.6$.
- Conclusion: w is around $404.6 \pm 0.6\ \mu\text{g}$.
- $x_1 = 409\ \mu\text{g}$. If we use x_1 to estimate w, what is the SE?

Randomness assumption

- w was estimated assuming the data are realisations of random draws from an infinite hypothetical population. How to check the assumption?
- Compare plot of $(1, x_1)$, $(2, x_2), \ldots, (100, x_{100})$ with a typical plot from random samples.

Example 4: How fair is a coin?

While imprisoned in WWII, John Kerrich tossed a coin 10000 times using the same protocol. He observed a total of 5067 heads.

Estimate p, the probability of head, and estimate the SE.

Assume the data $x_1, \ldots, x_{10000}$ are realisations of random draws $X_1, \ldots, X_{10000}$ from __________________________.

Estimating p

- p is estimated as $5067/10000 = 0.5067$.
- SE is $\sqrt{p(1-p)}/100$, estimated as

$$\frac{\sqrt{0.5067 \times 0.4933}}{100} \approx 0.005$$

- Conclusion: p is around 0.507 ± 0.005.
- How to check assumption? Plot cumulative number of heads against number of tosses.

Conclusion (1)

- If an SRS (size n) is small compared to the population (size N), the data can be assumed to come from IID RV's.
- Population mean is estimated using $\bar{X}$, and the error is quantified by the SE, which is $\mathrm{SD}(\bar{X})$. The SE usually has to be estimated from the data (the bootstrap).
- The technique should be modified if n/N is not small, or if another probability sampling method is used. The technique does not work for convenience samples.
- Population proportion is a special case.

Conclusion (2)

- The estimation method applies to data which are not sampled from real populations, provided it is sensible to assume

  The data are like realisations of random draws from a hypothetical population (or distribution).

- Statistical modeling of a data set is a quest for a distribution so that the assumption is reasonable.

- Unlike parameters of real populations, the exact weight of NB 10 and the chance of getting a head are unknowable.

- Key concepts: parameter, simple random sampling, estimate, estimator, standard error, hat notation.