STAT 403 Tutorial · Week 11

Topics to be included in this tutorial are proportion estimator and ratio estimator for simple random sampling.

1. Proportion estimator

The sample proportion is basically a special case of sample mean when the variable of interest $y_i \in \{0,1\}$. Therefore, all the theory and results of sample mean under SRS still hold for the sample proportion. Besides proportion $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, sometimes we may also be interested in the number of units in the population with certain property, i.e., $T = \sum_{i=1}^{N} y_i$. The estimators and standard errors are given in the table below.

- parameter of interest :	$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$	$T = \sum_{i=1}^{N} y_i$
- estimator :	$\hat{p} = \frac{1}{n}\sum_{i\in u} y_i$	$\hat{T} = \frac{N}{n}\sum_{i\in u} y_i = N\hat{p}$
- variance of the estimator :	$Var(\hat{p}) = (1-\frac{n}{N})\frac{S^2}{n}$, $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{Y})^2 = \frac{N}{N-1}\bar{Y}(1-\bar{Y})$	$Var(\hat{T}) = N^2(1-\frac{n}{N})\frac{S^2}{n}$, $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{Y})^2 = \frac{N}{N-1}\bar{Y}(1-\bar{Y})$.
- standard error of the estimator :	$se(\hat{p}) = \sqrt{(1-\frac{n}{N})\frac{\hat{p}(1-\hat{p})}{n-1}}$	$se(\hat{T}) = N\sqrt{(1-\frac{n}{N})\frac{\hat{p}(1-\hat{p})}{n-1}}$
- $(1-\alpha)$ - C.I. :	$\hat{p} \pm z_{\alpha/2}\cdot se(\hat{p})$	$\hat{T} \pm z_{\alpha/2}\cdot se(\hat{T})$.

Example. Suppose that the governor wants to estimate the total number of people infected by a certain disease in a certain area. To this end, he takes an SRS of size 1000 and finds that 120 among them are infected. Try to give a point estimate as well as a 95% CI. for the total number of infected people in this area. (It's known that the population in this area is 10^5).

Solution. Let $y_i = \begin{cases} 1, & \text{if unit } i \text{ is infected,} \\ 0, & \text{otherwise} \end{cases}$ $(i=1,\dots,N=10^5)$, then we estimate $T = \sum_{i=1}^{N} y_i$ by $\hat{T} = N\hat{p} = N\cdot\frac{1}{n}\sum_{i\in u} y_i = 10^5 \times \frac{1.2\times10^2}{10^3} = 1.2\times10^4$. The 95% C.I. is given by $\hat{T} \pm z_{0.025}\times N\sqrt{(1-\frac{n}{N})\frac{\hat{p}(1-\hat{p})}{n-1}} = 12000 \pm 1.96\times10^5\sqrt{0.99\times\frac{0.12\times0.88}{999}}$

$\approx [9995, 14005]$.

2. Ratio estimator

In class, we have learned that the ratio estimator can be used to improve our estimation for $\bar{Y}$ if the variable of interest y_i and the auxiliary variable x_i are highly correlated. In this case, $\hat{\bar{Y}}_r = \bar{X} \cdot \frac{\bar{y}}{\bar{x}}$ with $se(\hat{\bar{Y}}_r) = \sqrt{(1-\frac{n}{N})\frac{S_e^2}{n}}$, $S_e^2 = \frac{1}{n-1}\sum_{i\in u}(y_i - Bx_i)^2$. Sometimes in practice, it is the ratio itself $R = \sum_{i=1}^{N} y_i / \sum_{i=1}^{N} x_i$ of interest and we can use the ratio estimator $\hat{R} = \bar{y}/\bar{x}$ to estimate R. The standard error of $\hat{R}$ is given by $se(\hat{R}) = \frac{1}{\bar{X}}\sqrt{(1-\frac{n}{N})\frac{S_e^2}{n}}$.

As usual, a $(1-\alpha)$-C.I. for R is $\hat{R} \pm z_{\alpha/2} \cdot se(\hat{R})$.

Example. Suppose the population consists of 100 agricultural fields of different sizes. Let

y_i = bushels of grain harvested in field i

x_i = acreage of field i

We want to estimate the average yield per acre, i.e. $R = \bar{Y}/\bar{X}$. Suppose we have selected an SRS of size 6 and the numbers are recorded as below.

y_i	251	150	387	166	480	308
x_i	3	2	5	2	6	4

Then we can estimate the average yield per acre by $\hat{R} = \bar{y}/\bar{x} = \sum_{i\in u} y_i / \sum_{i\in u} x_i = 79.18$

and a 95% CI is given by $\hat{R} \pm z_{\alpha/2} \cdot se(\hat{R}) = 79.18 \pm 1.96 \times \frac{1}{3.67}\sqrt{(1-0.06)\cdot\frac{S_e^2}{6}}$,

= [77.09, 81.27]

[redacted].

Note that S_e^2 = 1/(n-1) * \sum_{i \in u} (y_i - B x_i)^2, where B = \hat{R} = \bar{y} / \bar{x}.

Therefore, S_e^2 = 1/5 *[(251 - 79.18*3)^2 + (150 - 79.18*2)^2 + ... + (308 - 79.18*4)^2] = 97.79.