Jan 9th, 2013

Instructor: Zhou Zhou
zhou@utstat.toronto.edu

TA: Alexander Shestopaloff alexander@utstat.utoronto.ca

Final 55% + Mid-term 30% + Quiz 15% (3 out of 4)

Today, we are gonna start from Ch. 7 and finish it.

Chapter 7 Sample Survey

Goal: To make inference about a large population via sampling just a small fraction of it.

Sample surveys are random in nature.
(should be) ↓ (doctors giving medications)
objective

1. Population Parameters

Think of a finite population of size N. We are interested in one characteristic of the population the values of which are ordered as $x_1, x_2, \cdots, x_N$
↓
Not Random

Population mean $\frac{1}{N}\sum_{i=1}^{N} x_i = \mu$ ← not random
← parameter

Population Variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}[x_i - \mu]^2$
↳ parameter

Simple Random Sample

(S.R.S) For each sample of size n (↑ sample size) in the population the chance of being selected are the same.

* Same as choosing randomly without replacement. X_i's are not independent

The selected sample is denoted by
$X_1, X_2, \cdots, X_n$

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

Sample variance $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Thm: $E[X_i] = \mu$ for any i.

Sampling without replacement is equivalent to grabbing n items at the same time from an urn of N items.

$E[X_i] = E[X_1] = E[X_2] = \cdots$ (because random labeling)

Now suppose that characteristic we are interested in has k different values $\beta_1, \beta_2, \beta_3 \cdots \beta_k$. Suppose in the population there are N_1 β_1's N_2 β_2's $\cdots N_k$ β_k's

$$E[X_i] = \sum_{i=1}^{k} \beta_i p[X_i = \beta_i] = \sum_{i=1}^{k} \beta_i \frac{N_i}{N} = \frac{1}{N} \sum_{i=1}^{k} \beta_i N_i$$

$$= \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$= \mu$$

Corollary $E[\bar{x}] = \mu$

Proof: $E[\bar{X}] = E[\frac{1}{n}\sum_{i=1}^{n} x_i] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu$

$\bar{x}$ is an unbiased estimate of μ.

Quality of Estimators

Commonly used measure of quality is the **mean square error** (MSE)

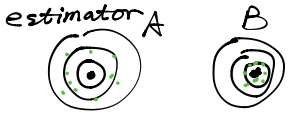
$MSE = E[\hat{\theta} - \theta]^2$ small MSE means good (accurate) estimate

$$= E[\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta]^2$$

$$= E[\hat{\theta} - E(\hat{\theta})]^2 + [E[\hat{\theta}] - \theta]^2 + 2E[(\hat{\theta} - E[\hat{\theta}])(\underbrace{E[\hat{\theta}] - \theta}_{\text{not random}})]$$

$$= E[\hat{\theta} - E(\hat{\theta})]^2 + [E[\hat{\theta}] - \theta]^2$$ two non-negative parts

variance of estimator; bias of estimator



B is better because of smaller variance

$$Var[X_i] = E[X_i^2] - (E[X_i])^2$$
$$= E[X_1^2] - (E[X_1])^2$$

$$= \sum_{i=1}^{k} \beta_i^2 p[X_1 = \beta_i] - \mu^2$$

$$= \sum_{i=1}^{k} \beta_i^2 \frac{N_i}{N} - \mu^2 = \frac{1}{N}\sum_{i=1}^{k} \beta_i^2 N_i - \mu^2$$

For each i, $\beta_i^2 N_i$ = total sum of squares in class, $\Rightarrow \sum_{i=1}^{k} \beta_i^2 N_i = \sum_{i=1}^{N} x_i^2$

Hence $Var[X_i] = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} = \sigma^2$

But we don't want $Var[x_i]$ but $Var[\bar{X}]$

$$Var[\bar{X}] = \frac{1}{n^2} Var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}(\sum_{i \neq j} 2cov[X_i, X_j] + \sum_{i=1}^{n} Var(X_i)) \quad \star$$

$= n\sigma^2$!!!

Because of exchangeability, $\text{cov}[X_i, X_j] = \text{cov}[X_1, X_i]$ for $\forall i \neq j$

$$\text{cov}[X_1, X_2] = E[X_1 X_2] - E[X_1]E[X_2]$$
$$= E[X_1 X_2] - \mu^2$$

$$E[X_1 X_2] = E[\,E[X_1 X_2 \mid X_2]\,] \qquad *E[X] = E[E[X \mid i]]$$

$$E[X_1 X_2 \mid X_2 = \beta_t] = \beta_t E[X_1 \mid X_2 = \beta_t]$$

For $r \neq t$, $P[X_1 = \beta_r \mid X_2 = \beta_t] = \frac{N_r}{N-1}$

$$P[X_1 = \beta_t \mid X_2 = \beta_t] = \frac{N_t - 1}{N-1}$$

$$E[X_1 \mid X_2 = \beta_t] = \beta_t \frac{N_t - 1}{N-1} + \sum_{i \neq t} \beta_i \frac{N_i}{N-1} = \frac{1}{N-1} \sum_{i=1}^{k} \beta_i N_i - \frac{\beta_t}{N-1}$$
$$= \frac{1}{N-1}\mu - \frac{\beta_t}{N-1}$$

$$E[X_1 X_2 \mid X_2 = \beta_t] = \frac{\beta_t[N\mu - \beta_t]}{N-1}$$

$$E[X_1 X_2] = \sum_{t=1}^{k} \frac{\beta_t[N\mu - \beta_t]}{N-1} P[X_2 = \beta_t] = \sum_{t=1}^{k} \frac{\beta_t[N\mu - \beta_t]}{N-1} \cdot \frac{N_t}{N}$$

$$= \frac{1}{(N-1)N}\left[\sum_{t=1}^{k} (N\mu)\beta_t N_t - \sum_{t=1}^{k} \beta_t^2 N_t\right]$$
$$= \frac{1}{(N-1)N}\left[N^2\mu^2 - \sum_{t=1}^{N} x_i^2\right]$$

known: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \mu^2$

$$\Rightarrow \sum_{i=1}^{N} x_i^2 = (\sigma^2 + \mu^2)N$$

$$= \frac{1}{(N-1)N}\left[N^2\mu^2 - N(\mu^2 + \sigma^2)\right]$$
$$= \mu^2 - \frac{1}{N-1}\sigma^2$$
$$\text{cov}[X_i, X_j] = \mu^2 - \frac{1}{N-1}\sigma^2 - \mu^2$$
$$= -\frac{1}{N-1}\sigma^2$$

final answer

Note ☆ $\text{Var}(\bar{X}) = \frac{1}{n^2}\left[2\left(-\frac{1}{N-1}\sigma^2\right)n(n-1) + n\sigma^2\right]$

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

$$= \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\sigma^2 \quad \leftarrow \text{MSE of } \bar{X}$$

Frequently $n \ll N$

MSE of $\bar{X} \approx \frac{\sigma^2}{n}$

to make $\bar{X}$ more accurate :
① increase n
② or decrease σ^2.

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

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i^2] - E[\bar{X}^2] = [\sigma^2 + \mu^2] - [Var[\bar{X}] + E[\bar{X}]^2]$$

$$= \sigma^2 + \mu^2 - \left[\frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\sigma^2 + \mu^2\right]$$

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

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

expectation of sample variance

If $N \gg n$, $E[\hat{\sigma}^2] \approx \sigma^2\left(1 - \frac{1}{n}\right)$

Define $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ — sample variance statistics

If $n \ll N$, $E\,S^2 \approx \sigma^2$

$$E[\hat{\sigma}^2] \approx S^2\left(\frac{N}{N-1}\right)$$