Replace all lower case x_i with c_i on pages 1–5

Let U_i be a random variable with

$$U_i = \begin{cases} 1 & \text{if } i\text{'th population member is in a sample} \\ 0 & \text{otherwise} \end{cases}$$

$i = 1, \ldots, N$

value of X for unit i in population

whether or not unit i from population is in sample

Define $W = \frac{1}{n} \sum_{i=1}^{N} x_i U_i$

For a particular sample we have realized values $u_1, \ldots, u_N$:

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

$$= \frac{1}{n} \left[\underbrace{\sum_{i:u_i=1} x_i u_i}_{\text{obs. in sample}} + \underbrace{\sum_{i:u_i=0} x_i u_i}_{\text{obs. not in sample}} \right]$$

$$= \frac{1}{n} \left[\sum_{i:u_i=1} x_i \cdot 1 + \cancel{\sum_{i:u_i=0} x_i \cdot 0} \right]$$

$$= \frac{1}{n} \sum_{i:u_i=1} x_i$$

$$= \frac{1}{n} \sum_{\substack{\text{obs. in} \\ \text{sample}}} x_i$$

$$= \bar{x}$$

So $\bar{X} = \frac{1}{n} \sum_{i=1}^{N} x_i U_i$ (random variable)

$\bar{x} = \frac{1}{n} \sum_{i=1}^{N} x_i u_i$ (realized value for a particular sample)

Goal: Find $E(\bar{X})$ $Var(\bar{X})$ based on a simple random sample of size n from a population of size N.

Collect some facts about the U_i's:

1) $E(U_i) = \frac{n}{N}$

2) $Var(U_i) = \frac{n}{N}(1 - \frac{n}{N})$

3) $E(U_iU_j) = \frac{n(n-1)}{N(N-1)}$ for $i \neq j$

4) $Cov(U_i, U_j) = \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2$ ~~for~~ $i \neq j \longrightarrow \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2$

---

1) $U_i = \begin{cases} 0 & \text{if obs. } i \text{ in sample} \\ 1 & \text{otherwise} \end{cases}$

$$E[U_i] = 0 \cdot P(U_i = 0) + 1 \cdot P(U_i = 1)$$
$$= P(U_i = 1)$$
$$= P(\text{obs. } i \text{ in sample})$$

Simple random sample: all samples of size n equally likely

$$P(\text{obs. } i \text{ in sample}) = \frac{\#\text{ samples of size n where obs. } i \text{ is included}}{\#\text{ samples of size n}}$$

$$= \frac{\binom{N-1}{n-1}}{\binom{N}{n}} \longleftarrow \text{we know obs. } i \text{ is in sample, how many ways to choose the remaining } n-1\text{ ?}$$

$$= \frac{\left\{\frac{(N-1)!}{(N-1-(n-1))!\,(n-1)!}\right\}}{\frac{N!}{(N-n)!\,n!}}$$

$$= \frac{(N-1)!}{N!} \; \frac{\cancel{(N-n)!}}{\cancel{(N-n)!}} \; \frac{n!}{(n-1)!}$$

$$= \frac{n}{N}$$

2) $U_i \sim \text{Bernoulli}(\frac{n}{N})$

If $X \sim \text{Bernoulli}(p)$, then $Var(X) = p(1-p)$

$$Var(U_i) = \frac{n}{N}(1 - \frac{n}{N})$$

3) $U_iU_j = \begin{cases} 1 & \text{if both observation } i \text{ and } j \text{ are in sample} \\ 0 & \text{otherwise} \end{cases}$

$$
\begin{aligned}
E(U_iU_j) &= P(U_iU_j = 1) \\
&= P(\text{both } i \text{ and } j \text{ are in sample}) \\
&= \frac{\#\text{ samples where both } i \text{ and } j \text{ are included}}{\#\text{ samples of size } n} \\
&= \frac{\binom{N-2}{n-2}}{\binom{N}{n}} \\
&= \frac{\left\{\frac{(N-2)!}{(N-2-(n-2))!\,(n-2)!}\right\}}{\frac{N!}{(N-n)!\,n!}} \\
&= \frac{n(n-1)}{N(N-1)}
\end{aligned}
$$

4) $$
\begin{aligned}
Cov(U_i, U_j) &= E(U_iU_j) - E(U_i)E(U_j) \\
&= \frac{n(n-1)}{N(N-1)} - \frac{n}{N} \cdot \frac{n}{N}
\end{aligned}
$$

Now,

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{N} x_i U_i\right)$$

$$= \frac{1}{n}\sum_{i=1}^{N}\{x_i E(U_i)\}$$

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

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

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

$$= \mu$$

Def.: ~~An e. The bias of an estimator~~
Suppose $\hat{\theta}$ is an estimator for a parameter θ.
The bias of $\hat{\theta}$ is

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Example: If we regard $\bar{X}$ as an estimator of ~~μ~~, the population parameter μ,

$$\text{Bias}(\bar{X}) = E(\bar{X}) - \mu$$

$$= \mu - \mu$$

$$= 0$$

Def.: An estimator is unbiased if it has bias 0.

Example: $\text{Bias}(\bar{X}) = 0$, so
$\bar{X}$ is an unbiased estimator of μ.

Lemma (see Section 4.3 of Rice)

If a and $b_1, \ldots, b_n$ are constants and $X_1, \ldots, X_n$ are jointly distributed random variables then

$$Var\left(a + \sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} b_i b_j Cov(X_i, X_j)$$

$$\begin{aligned}
Var\left(\frac{1}{n}\sum_{i=1}^{N} x_i U_i\right) &= \frac{1}{n^2} \cdot Var\left(\sum_{i=1}^{N} x_i U_i\right) \\
&= \frac{1}{n^2}\left\{\sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j \cdot Cov(U_i, U_j)\right\} \\
&= \frac{1}{n^2} \cdot \left\{\sum_{i=1}^{N} x_i^2 \frac{n}{N}\left(1-\frac{n}{N}\right) + \sum_{i=1}^{N}\sum_{j \neq i} x_i x_j \left(\frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2\right)\right\} \\
&= \frac{1}{n^2}\left\{\sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j \frac{n}{N}\left(\frac{n-1}{N-1} - \frac{n}{N}\right) + \frac{n}{N}\left(1 - \frac{n-1}{N-1}\right)\sum_{i=1}^{N} x_i^2\right\} \\
&= \frac{1}{nN}\left\{\cancel{4}\, \frac{-1}{N}\left(1 - \frac{n-1}{N-1}\right)\sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j + \left(1 - \frac{n-1}{N-1}\right)\sum_{i=1}^{N} x_i^2\right\} \\
&= \frac{1}{nN}\left(1 - \frac{n-1}{N-1}\right)\left\{\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\sum_{i=1}^{N} x_i \sum_{j=1}^{N} x_j\right\} \\
&= \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{N}\sum_{i=1}^{N}\left\{x_i^2 - 2 \cdot \frac{1}{N} x_i \sum_{j=1}^{N} x_j + \frac{1}{N^2}\left(\sum_{j=1}^{N} x_j\right)^2\right\} \\
&= \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{j=1}^{N} x_j\right)^2 \\
&= \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\underbrace{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \mu\right)^2}_{\sigma^2} \\
&= \frac{\sigma^2}{n}\underbrace{\left(1 - \frac{n-1}{N-1}\right)}_{\text{finite population correction factor}}
\end{aligned}$$

finite population correction factor

If sample size n is small relative to population size N, $\frac{n-1}{N-1} \approx 0$ and $Var(\bar{X}) \approx \frac{\sigma^2}{n}$

$$\begin{aligned}
\left\{\frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2\right\} &= \left(\frac{n}{N}\right)\left\{\frac{n-1}{N-1} - \frac{n}{N}\right\} \\
&= \frac{n}{N}\left\{1 - \frac{n}{N} - \left(1 - \frac{n-1}{N-1}\right)\right\} \\
&= \frac{n}{N}\left(1 - \frac{n}{N}\right) - \frac{n}{N}\left(1 - \frac{n-1}{N-1}\right)
\end{aligned}$$

$$\begin{aligned}
\left(\frac{n-1}{N-1} - \frac{n}{N}\right) &= \frac{-1}{N}\left(\frac{-N(n-1)}{N-1} + \frac{n(N-1)}{N-1}\right) \\
&= \frac{-1}{N}\left(\frac{N+1-1-n}{N-1}\right) \\
&= \frac{-1}{N}\left(\frac{N-1}{N-1} + \frac{1-n}{N-1}\right) \\
&= \frac{-1}{N}\left(1 - \frac{n-1}{N-1}\right)
\end{aligned}$$

Estimation of the population variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(c_i-\mu)^2$

Consider the naive estimator

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

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

Proof: more of the same, see Section 7.3.2

Since $n<N$, $\left(\frac{n-1}{n}\right)\left(\frac{N}{N-1}\right) < 1$

(Example: $n=2$, $N=10 \Rightarrow \frac{1}{2}\cdot\frac{10}{9} = \frac{5}{9}$)

$\therefore E(\hat{\sigma}^2_{naive}) < \sigma^2$

So $\hat{\sigma}^2_{naive}$ is a biased estimator of σ^2

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

$$= \left(\frac{N-1}{N}\right)\left(\frac{n}{n-1}\right)\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \left(\frac{N-1}{N}\right)\left(\frac{n}{n-1}\right)\cdot\hat{\sigma}^2_{naive}$$

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

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

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

$$= \sigma^2$$

$\hat{\sigma}^2$ is an unbiased estimator of σ^2