Serik Sagitov, Chalmers Tekniska Högskola, December 21, 2012

Chapter 7. Survey sampling

1 Random sampling

Population = set of elements $\{1, 2, \dots, N\}$ labeled by values $\{x_1, x_2, \dots, x_N\}$.
PD = population distribution of x-values. A single value of a random element $X \sim$ PD.
Types of x-values (data):
  continuous, discrete, categorical, dichotomous (2 categories).
General population parameters
  population mean $\mu = \mathrm{E}(X)$,
  population standard deviation $\sigma = \sqrt{\mathrm{Var}(X)}$,
  population proportion p (dichotomous data).
Two methods of studying PD and population parameters:
  enumeration - expensive, sometimes impossible,
  random sample: n random observations $(X_1, \dots, X_n)$.

Randomisation is a guard against investigator's biases even unconscious

IID sample, sampling with replacement: Independent Identically Distributed observations.
Simple random sample, sampling without replacement: negative dependence $\mathrm{Cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}$.
  Proof: $X_1 + \ldots + X_N = \text{const}$. Use the addition rule of variance.
Example. Students heights: height in cm = discrete data, gender = dichotomous data.

2 Point estimates

Population parameter θ estimation uses a point estimate $\hat\theta = \hat\theta(X_1, \dots, X_n)$.
Sampling distribution of $\hat\theta$ around unknown θ: different values $\hat\theta$ observed for different samples.
Mean square error
  $\mathrm{E}(\hat\theta - \theta)^2 = \Big[\mathrm{E}(\hat\theta) - \theta\Big]^2 + \sigma_{\hat\theta}^2$
  $\mathrm{E}(\hat\theta) - \theta$ = systematic error, bias, lack of accuracy; $\sigma_{\hat\theta}$ = random error, lack of precision.
Desired properties of point estimates:
  $\hat\theta$ is an unbiased estimate of θ, if $\mathrm{E}(\hat\theta) = \theta$,
  $\hat\theta$ is consistent, if $\mathrm{E}(\hat\theta - \theta)^2 \to 0$ as $n \to \infty$.

Sample mean $\bar X = \frac{X_1 + \ldots + X_n}{n}$ is an unbiased and consistent estimate of μ

$$\mathrm{Var}(\bar X) = \begin{cases} \sigma^2/n & \text{if IID sample} \\ \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right) & \text{if simple random sample} \end{cases}$$

Finite population correction $1 - \frac{n-1}{N-1}$ can be neglected if sample proportion $\frac{n}{N}$ is small.

Dichotomous data: $\mathrm{P}(X_i = 1) = p$, $\mathrm{P}(X_i = 0) = q$, $\mu = p$, $\sigma^2 = pq$, population proportion p. Sample proportion $\hat p = \bar X$ is an unbiased and consistent estimate of p.

Sample variance $s^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$, where s is the sample standard deviation.
Other formulae:

$s^2 = \frac{n}{n-1}(\overline{X^2} - \bar{X}^2)$, where $\overline{X^2} = \frac{1}{n}(X_1^2 + \ldots + X_n^2)$,

dichotomous data case $s^2 = \frac{n}{n-1}\hat{p}\hat{q}$.

Sample variance is an unbiased estimate of σ^2

$$\mathrm{E}(s^2) = \begin{cases} \sigma^2 & \text{if IID sample} \\ \sigma^2 \frac{N}{N-1} & \text{if simple random sample.} \end{cases}$$

Standard errors of $\bar{X}$ and $\hat{p}$ for simple random sample: $s_{\bar{X}} = \frac{s}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}$, $s_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n-1}}\sqrt{1 - \frac{n}{N}}$.

$$\boxed{\text{Standard errors for IID sampling } s_{\bar{X}} = \frac{s}{\sqrt{n}},\ s_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n-1}}}$$

3 Confidence intervals

Approximate sampling distribution $\bar{X} \overset{a}{\sim} \mathrm{N}(\mu, \frac{\sigma^2}{n})$

$\mathrm{P}(\bar{X} - zs_{\bar{X}} < \mu < \bar{X} + zs_{\bar{X}}) = \mathrm{P}(-z < \frac{\bar{X}-\mu}{s_{\bar{X}}} < z) \approx 2(1 - \Phi(z))$

Approximate 100(1–α)% two-sided CI for μ and p: $\bar{X} \pm z_{\alpha/2} \cdot s_{\bar{X}}$ and $\hat{p} \pm z_{\alpha/2} \cdot s_{\hat{p}}$, if n is large

100(1–α)%	68%	80%	90%	95%	99%	99.7%
$z_{\alpha/2}$	1.00	1.28	1.64	1.96	2.58	3.00

The higher is confidence level the wider is the CI, the larger is sample the narrower is the CI.

95% CI is a random interval: out of 100 intervals computed for 100 samples
Bin(100,0.95) $\approx$ N(95,$(2.18)^2$) will cover the true value.

4 Stratified random sampling

Population consists of L strata with known L strata fractions $W_1 + \ldots + W_L = 1$ and
unknown strata means μ_l and standard deviations σ_l

Population mean $\mu = W_1\mu_1 + \ldots + W_L\mu_L$,

population variance $\sigma^2 = \overline{\sigma^2} + \sum W_l(\mu_l - \mu)^2$,

average variance $\overline{\sigma^2} = W_1\sigma_1^2 + \ldots + W_L\sigma_L^2$,

average standard deviation $\bar{\sigma} = W_1\sigma_1 + \ldots + W_L\sigma_L$.

Stratified random sampling: take L independent samples from each stratum with sample means $\bar{X}_1, \ldots, \bar{X}_L$

$$\boxed{\text{Stratified sample mean: } \bar{X}_s = W_1\bar{X}_1 + \ldots + W_L\bar{X}_L}$$

$\bar{X}_s$ is an unbiased and consistent estimate of μ: $\mathrm{E}(\bar{X}_s) = W_1\mathrm{E}(\bar{X}_1) + \ldots + W_L\mathrm{E}(\bar{X}_L) = \mu$.

Sample variance $s_{\bar{X}_s}^2 = (W_1 s_{\bar{X}_1})^2 + \ldots + (W_L s_{\bar{X}_L})^2$

$$\boxed{\text{Approximate CI for } \mu\text{: } \bar{X}_s \pm z_{\alpha/2} \cdot s_{\bar{X}_s}}$$

Pooled sample mean $\bar{X}_p = \frac{1}{n}(n_1\bar{X}_1 + \ldots + n_L\bar{X}_L)$, polled sample size $n = n_1 + \ldots + n_L$.

$\mathrm{E}(\bar{X}_p) = \frac{n_1}{n}\mu_1 + \ldots + \frac{n_L}{n}\mu_L = \mu + \sum(\frac{n_l}{n} - W_l)\mu_l$,

$\mathrm{bias}(\bar{X}_p) = \sum(\frac{n_l}{n} - W_l)\mu_l$.

Example. Students heights: $L = 2$, $W_1 = W_2 = 0.5$, compare $\bar{X}_s$ with $\bar{X}_p$.

$$\boxed{\text{Optimal allocation: } n_l = n\frac{W_l\sigma_l}{\bar{\sigma}},\ \mathrm{Var}(\bar{X}_{so}) = \frac{1}{n}\cdot\bar{\sigma}^2}$$

$\bar{X}_{so}$ minimizes standard error of X_s. Weakness: usually unknown σ_l and $\bar{\sigma}$.

$$\boxed{\text{Proportional allocation: } n_l = nW_l,\ \mathrm{Var}(\bar{X}_{sp}) = \frac{1}{n}\cdot\overline{\sigma^2}}$$

Compare three unbiased estimates of μ: $\mathrm{Var}(\bar{X}_{so}) \le \mathrm{Var}(\bar{X}_{sp}) \le \mathrm{Var}(\bar{X})$.
Variability in σ_l across strata:

$$\mathrm{Var}(\bar{X}_{sp}) - \mathrm{Var}(\bar{X}_{so}) = \frac{1}{n}(\overline{\sigma^2} - \bar{\sigma}^2)=\frac{1}{n}\sum W_l(\sigma_l - \bar{\sigma})^2.$$

Variability in means μ_l across strata:

$$\mathrm{Var}(\bar{X}) - \mathrm{Var}(\bar{X}_{sp}) = \frac{1}{n}(\sigma^2 - \overline{\sigma^2}) = \frac{1}{n}\sum W_l(\mu_l - \mu)^2.$$