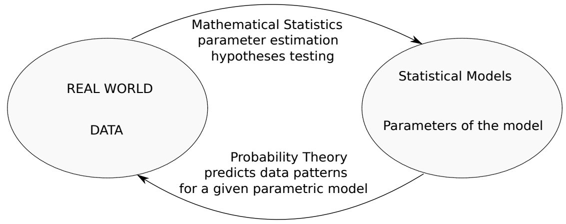
Serik Sagitov, Chalmers and GU, January 27, 2016




# Chapter 7. Survey sampling

## 1 Randomisation

Population is a set of $N$ elements characterised by values $\{x_1, x_2, \ldots, x_N\}$. We are interested in the population distribution of $x$-values. In many situations enumeration is expensive and even impossible. Solution: collect a random sample of $n$ random observations $(X_1, \ldots, X_n)$.

If we pick at random one element from the population, then its
$x$-value $X$ is a random variable whose distribution is the population distribution.

Types of data:
quantitative (continuous or discrete) and categorical.
General population parameters for quantitative data
population mean $\mu = \mathrm{E}(X)$,
population standard deviation $\sigma = \sqrt{\mathrm{Var}(X)}$.
Important special case of categorical data is dichotomous data. Example: $x_i \in \{\text{male}, \text{female}\}$. After converting to a quantitative form by $x_i \in \{0, 1\}$, the population mean turns into
population proportion $p = P(X = 1)$.

*Randomisation* is a guard against investigator's biases even unconscious.

## 2 Two basic ways of random sampling

Sampling without replacement produces so called Simple Random Sample:
negative dependence between observations $\sigma_{ij} := \mathrm{Cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}$, $i \neq j$,
indeed, since $X_1 + \ldots + X_N$ is a constant,
we get $\mathrm{Var}(X_1 + \ldots + X_N) = 0$, and $N\sigma^2 + N(N-1)\sigma_{12} = 0$.
Sampling with replacement produces an IID sample:
Independent Identically Distributed observations $(X_1, \ldots, X_n)$,
easier to analyse, good approximation of the simple random sample if $n/N$ is small.

**Example.** We collect data on students heights in cm and gender. Questions: is this a random sample, estimate the mean height of the Swedish population, what is the error of the estimate?

## 3 Point estimation

To estimate a population parameter $\theta$ we need a sensible point estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$. Sampling distribution of $\hat{\theta}$ around unknown $\theta$: different values $\hat{\theta}$ observed for different samples. The sampling distribution has mean $\mu_{\hat{\theta}} = \mathrm{E}(\hat{\theta})$ and variance $\sigma^2_{\hat{\theta}} = \mathrm{E}(\hat{\theta} - \mu_{\hat{\theta}})^2$.

The quality of the the point estimator is given by the mean square error
$\mathrm{E}(\hat{\theta} - \theta)^2 = \mathrm{E}(\hat{\theta} - \mu_{\hat{\theta}})^2 + 2\mathrm{E}(\hat{\theta} - \mu_{\hat{\theta}})(\mu_{\hat{\theta}} - \theta) + (\mu_{\hat{\theta}} - \theta)^2 = \sigma^2_{\hat{\theta}} + (\mu_{\hat{\theta}} - \theta)^2$.
The mean square error has two components
$\mu_{\hat{\theta}} - \theta$ is the size of systematic error, bias, lack of accuracy;
$\sigma_{\hat{\theta}}$ is the size of random error, lack of precision.
Desired properties of point estimates:
$\hat{\theta}$ is an unbiased estimate of $\theta$, if $\mu_{\hat{\theta}} = \theta$,
$\hat{\theta}$ is consistent, if the mean square error $\mathrm{E}(\hat{\theta} - \theta)^2 \to 0$ as $n \to \infty$.

The standard error for an estimator $\hat{\theta}$ is its standard deviation $\sigma_{\hat{\theta}} = \sqrt{\mathrm{Var}(\hat{\theta})}$.
The estimated standard error $s_{\hat{\theta}}$ is an estimator of $\sigma_{\hat{\theta}}$.

## 4 Sample mean and variance

Two most basic summary statistics: the sample mean and sample variance

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}, \quad s^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2.$$

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

Consider an IID sample. The sample mean $\bar{X}$ and sample variance $s^2$ are unbiased and consistent estimators for the population mean and variance respectively

$$\mathrm{E}(\bar{X}) = \mu, \quad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}, \quad \mathrm{E}(s^2) = \sigma^2, \quad \mathrm{Var}(s^2) = \frac{\sigma^4}{n}\Big(\mathrm{E}(Y^4) - \frac{n-3}{n-1}\Big), \quad Y = \frac{X - \mu}{\sigma}.$$

The sample standard deviation $s$ is a biased estimate of the population standard deviation $\sigma$:
$\mathrm{E}(s) = \sqrt{\sigma^2 - \mathrm{Var}(s)}$.
In the dichotomous case, when $X$ takes values 1 or 0 with probabilities $p$ and $q = 1 - p$, we have $\mu = p$ and $\sigma^2 = pq$. Then the sample proportion $\hat{p} = \bar{X}$ is an unbiased and consistent estimate of $p$. In this case $s^2 = \frac{n}{n-1}\hat{p}\hat{q}$.

Estimated standard errors for the sample mean and proportion $s_{\bar{X}} = \frac{s}{\sqrt{n}}$, $s_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n-1}}$

## 5 Finite population correction

Now consider simple random sampling, when there is dependence between observations. In this case the sample mean $\bar{X}$ is again an unbiased and consistent estimator for the population mean. However, the sample variance $s^2$ is a biased estimator of $\sigma^2$, since $\mathrm{E}(s^2) = \sigma^2 \frac{N}{N-1}$, where $N$ is the finite population size.

We have $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}\Big(1 - \frac{n-1}{N-1}\Big)$, so that the formulas for the estimated standard errors of $\bar{X}$ and $\hat{p}$ for the simple random sample take new form $s_{\bar{X}} = \frac{s}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}$, $s_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n-1}}\sqrt{1 - \frac{n}{N}}$.

## 6 Confidence interval, CI

By the Central Limit Theorem, the sample mean distribution is approximately normal $\bar{X} \overset{a}{\sim} \mathrm{N}(\mu, \frac{\sigma^2}{n})$

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

Approximate 100(1–$\alpha$)% two-sided CI for $\mu$ and $p$: $\bar{X} \pm z_{\alpha/2} \cdot s_{\bar{X}}$ and $\hat{p} \pm z_{\alpha/2} \cdot s_{\hat{p}}$

| 100(1–$\alpha$)% | 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 only

$Z \sim \mathrm{Bin}(100, 0.95) \approx \mathrm{N}(95, (2.18)^2)$ intervals will cover the true value.

## 7 Stratified random sampling

Population consists of $L$ strata with known $L$ strata fractions $W_1 + \ldots + W_L = 1$ and

unknown strata means $\mu_l$ and standard deviations $\sigma_l$.

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

Example: for students heights, we have $L = 2$ and $W_1 = W_2 = 0.5$.

Stratified random sampling:

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

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

$\mathrm{E}(\bar{X}_p) = \frac{n_1}{n}\mu_1 + \ldots + \frac{n_L}{n}\mu_L = \mu + \sum(\frac{n_l}{n} - W_l)\mu_l$ bias in the non-proportional allocation.

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

$\mathrm{E}(\bar{X}_s) = W_1\mathrm{E}(\bar{X}_1) + \ldots + W_L\mathrm{E}(\bar{X}_L) = \mu$ no bias.

The variance of $\bar{X}_s$ and its estimate

$\sigma^2_{\bar{X}_s} = W_1^2\sigma^2_{\bar{X}_1} + \ldots + W_L^2\sigma^2_{\bar{X}_L}$, $\quad s^2_{\bar{X}_s} = W_1^2 s^2_{\bar{X}_1} + \ldots + W_L^2 s^2_{\bar{X}_L} = \frac{W_1^2 s_1^2}{n_1} + \ldots + \frac{W_L^2 s_L^2}{n_L}$.

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

Question: how shall we allocate $n = n_1 + \ldots + n_L$ observations among different strata?

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

Optimal allocation $(n_1, \ldots, n_L)$ minimises the error size: $\sigma^2_{\bar{X}_s} = \frac{W_1^2\sigma_1^2}{n_1} + \ldots + \frac{W_L^2\sigma_L^2}{n_L} \to \min$.

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

Compare three unbiased estimates of the population mean $\mathrm{Var}(\bar{X}_{so}) \le \mathrm{Var}(\bar{X}_{sp}) \le \mathrm{Var}(\bar{X})$

variability in $\sigma_l$ across strata makes optimal allocation more effective than proportional

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

variability in $\mu_l$ across strata makes proportional allocation more effective than IID sample

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