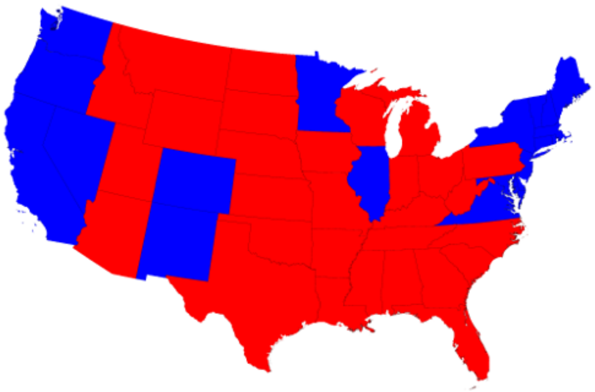
Slides 2: Stratified random sampling

- Stratified population
- Sample allocation
- Optimal allocation
- Proportional allocation
- Random allocation



Assume that a population consists of k strata with known strata fractions $(w_1, \ldots, w_k)$ such that

$$w_1 + \ldots + w_k = 1.$$

Suppose each stratum is characterised by its mean μ_j and standard deviation σ_j. The population population mean

$$\mu = w_1\mu_1 + \ldots + w_k\mu_k$$

is the parameter we would like to know. Denote by

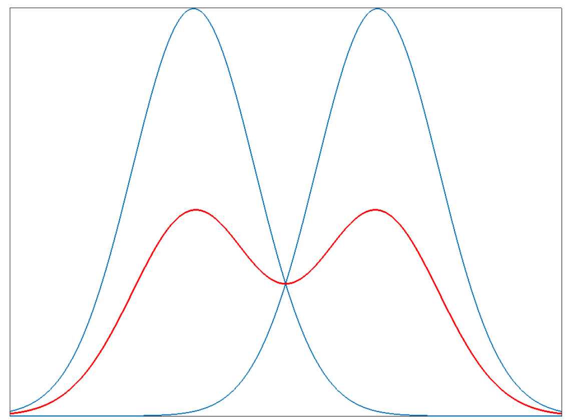
$$\overline{\sigma^2} = w_1\sigma_1^2 + \ldots + w_k\sigma_k^2$$

the weighted average variance. Then, the total population variance

$$\sigma^2 = \overline{\sigma^2} + \sum_{j=1}^{k} w_j(\mu_j - \mu)^2$$

takes into account two sources of variation: within strata variation and between strata variation.

For example, a mixture of two normal distributions ($k = 2$) with $w_1 = w_2 = 0.5$ would result in camel-curve for the population distribution



Notice that the camel-curve has a larger standard deviation than any of the strata curves.

Question. Think of the heights of people in a population with two equally large strata (women and men). Then the population distribution will look like a camel-curve. Why the central limit theorem fails in such a setting?

The random sampling of size n leads to $\bar{x}$ as an unbiased estimate of μ with standard error $s_{\bar{x}} = \frac{s}{\sqrt{n}}$, where s is the sample standard deviation.

A stratified random sampling consists of allocating n observations among k strata. A natural allocation is proportional $n_j = nw_j$ to the strata size.

Assume we collected k independent random samples of sample sizes $(n_1, \ldots, n_k)$, and let $(\bar{x}_1, \ldots, \bar{x}_k)$ be the corresponding sample means.

Define a stratified sample mean by

$$\bar{x}_{\mathrm{s}} = w_1\bar{x}_1 + \ldots + w_k\bar{x}_k$$

Observe that for any allocation $(n_1, \ldots, n_k)$, the stratified sample mean is an unbiased estimate of μ since

$$\mathrm{E}(\bar{X}_{\mathrm{s}}) = w_1\mathrm{E}(\bar{X}_1) + \ldots + w_k\mathrm{E}(\bar{X}_k) = w_1\mu_1 + \ldots + w_k\mu_k = \mu.$$

Question. Political polls have failed to accurately predict the USA presidential election results. What types of stratification of the USA population might improve the prediction performance of the polls?

Female heights $n_1 = 24$

155, 158, 160,160,162,162,162,163,165,165,168,168,

170,170,170,170,171,172,172,173,173,174,175,178, 180,182,188

Sample mean, standard deviation, and standard error

$$\bar{x}_1 = 169.11, \quad s_1 = 7.71, \quad s_{\bar{x}_1} = 1.48$$

Male heights $n_2 = 67$

170,170,170,170,173,173,173,174,174,174, 175,176,176,177,178,178,178,178,

180,180,180,180,180,180,180,180,180,180,181, 182,182,182,182,183,183,183,183,183,183,

185,185,185,185,185,186,186,187,187,187,187,188,188,189,

190,190,190,190,190,190,190,191,191,191,193,194,194, 195

Sample mean, standard deviation, and standard error

$$\bar{x}_2 = 182.58, \quad s_2 = 9.19, \quad s_{\bar{x}_1} = 0.80$$

Stratified sample mean (assuming $w_1 = w_2 = 0.5$)

$$\bar{x}_{\mathrm{s}} = \frac{169.111 + 182.582}{2} = 175.85, \quad s_{\bar{x}_{\mathrm{s}}} = \tfrac{1}{2}\sqrt{s_{\bar{x}_1}^2 + s_{\bar{x}_2}^2} = 0.84$$

The expression for the variance of $\bar{X}_\mathrm{s}$

$$\mathrm{Var}(\bar{X}_\mathrm{s}^2) = w_1^2 \mathrm{Var}(\bar{X}_1) + \ldots + w_k^2 \mathrm{Var}(\bar{X}_k) = \tfrac{w_1^2 \sigma_1^2}{n_1} + \ldots + \tfrac{w_k^2 \sigma_k^2}{n_k}$$

shows that the size of the random error depends on three vectors: $(w_1, \ldots, w_k)$, $(\sigma_1, \ldots, \sigma_k)$, and $(n_1, \ldots, n_k)$.

The last formula implies the next formula for the estimated standard error

$$s_{\bar{x}_\mathrm{s}} = \sqrt{\tfrac{w_1^2 s_1^2}{n_1} + \ldots + \tfrac{w_k^2 s_k^2}{n_k}},$$

as a function of the sample sizes $(n_1, \ldots, n_k)$. Here s_j is the sample standard deviation for strata j.

Using $(\bar{x}_\mathrm{s}, s_{\bar{x}_\mathrm{s}})$ we can build a 95% confidence interval for μ by the usual kind of formula

$$I_\mu \approx \bar{x}_\mathrm{s} \pm 1.96 \cdot s_{\bar{x}_\mathrm{s}}.$$

Question. How can you justify the use of the factor 1.96 in the stratified setting?

Optimisation problem: allocate $n = n_1 + \ldots + n_k$ observations among different strata to minimise the sampling error of $\bar{x}_{\mathrm{s}}$.

Solution: optimal allocation

$$n_j = n\frac{w_j \sigma_j}{\bar{\sigma}}.$$

The optimal allocation assigns more observations to larger strata and strata with larger variation.

The optimal allocation gives the theoretical minimal variance

$$\mathrm{Var}(\bar{X}_{\mathrm{so}}) = \frac{\bar{\sigma}^2}{n},$$

where $\bar{\sigma}^2$ is the squared average standard deviation

$$\bar{\sigma} = w_1\sigma_1 + \ldots + w_k\sigma_k.$$

The major drawback of the optimal allocation formula is that it requires knowledge of the standard deviations σ_j.

Question. If $\sigma_j = 0$, how large should be n_j?

If σ_j are unknown, then a common sense approach is to allocate observations proportionally to the strata sizes, so that

$$n_1 = nw_1, \quad \ldots, \quad n_k = nw_k.$$

The corresponding variance equals

$$\mathrm{Var}(\bar{X}_{\mathrm{sp}}) = \frac{w_1^2\sigma_1^2}{nw_1} + \ldots + \frac{w_k^2\sigma_k^2}{nw_k} = \frac{\overline{\sigma^2}}{n}$$

which is larger than $\mathrm{Var}(\bar{X}_{\mathrm{so}}) = \frac{\bar{\sigma}^2}{n}$, because

$$\overline{\sigma^2} - \bar{\sigma}^2 = \sum w_j(\sigma_j - \bar{\sigma})^2.$$

On the other hand, $\mathrm{Var}(\bar{X}_{\mathrm{sp}}) = \frac{\overline{\sigma^2}}{n}$ is smaller than $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$ since

$$\sigma^2 - \overline{\sigma^2} = \sum w_j(\mu_j - \mu)^2.$$

In summary, we have three unbiased estimated of μ with

$$\mathrm{Var}(\bar{X}_{\mathrm{so}}) \leq \mathrm{Var}(\bar{X}_{\mathrm{sp}}) \leq \mathrm{Var}(\bar{X}).$$

Variability of σ_j across strata makes optimal allocation more effective than proportional

$$\mathrm{Var}(\bar{X}_{\mathrm{sp}}) - \mathrm{Var}(\bar{X}_{\mathrm{so}}) = \tfrac{1}{n} \sum w_j (\sigma_j - \bar{\sigma})^2.$$

Variability in μ_j across strata makes proportional allocation more effective than the purely random sample

$$\mathrm{Var}(\bar{X}) - \mathrm{Var}(\bar{X}_{\mathrm{sp}}) = \tfrac{1}{n} \sum w_j (\mu_j - \mu)^2.$$

Question. Observe that with the proportional allocation $n_i = n w_i$, we formally get

$$\bar{x}_{\mathrm{sp}} = w_1 \bar{x}_1 + \ldots + w_n \bar{x}_k = \tfrac{n_1}{n} \bar{x}_1 + \ldots + \tfrac{n_k}{n} \bar{x}_k = \tfrac{x_1 + \ldots + x_n}{n} = \bar{x}.$$

However, this is not the mean of a truly random sample, since we know that usually,

$$\mathrm{Var}(\bar{X}_{\mathrm{sp}}) < \mathrm{Var}(\bar{X}).$$

Explain, what is going on here?