SAMPLING THEORY AND STUDY DESIGN

DATA COLLECTION AND DATA PROCESSING

> “The latest survey shows that 3 out of 4 people make up 75% of the population”
>
> D. Letterman

THE GOAL OF GOOD STUDY/SAMPLING DESIGN

We need data that can:

- provide legitimate insight into our system of interest;
- provide correct, accurate answers to relevant questions;
- support the drawing of legitimate, valid conclusions, with the ability to qualify these conclusions in terms of scope and precision.

This starts with **study design** – what data to collect and how it should be collected

[G. Smith, The Exaggerated Promise of So-Called Unbiased Data Mining]

"A Dartmouth graduate student used an MRI machine to study the brain activity of a salmon as it was shown photographs and asked questions. The most interesting thing about the study was not that a salmon was studied, but that the salmon was dead. Yep, a dead salmon purchased at a local market was put into the MRI machine, and some patterns were discovered. There were inevitably patterns—and they were invariably meaningless."

NPS AND PATTERN FISHING

Two separate issues can be combined to cause **problems** with data analysis:

- drawing conclusions (inferences) from a sample about a population that are not warranted by the sample collection method (symptomatic of NPS);
- looking for any available patterns in the data and then coming up with *post hoc* explanations for these patterns.

Alone or in combination, these lead to poor (and **potentially harmful**) conclusions.

STUDIES AND SURVEYS

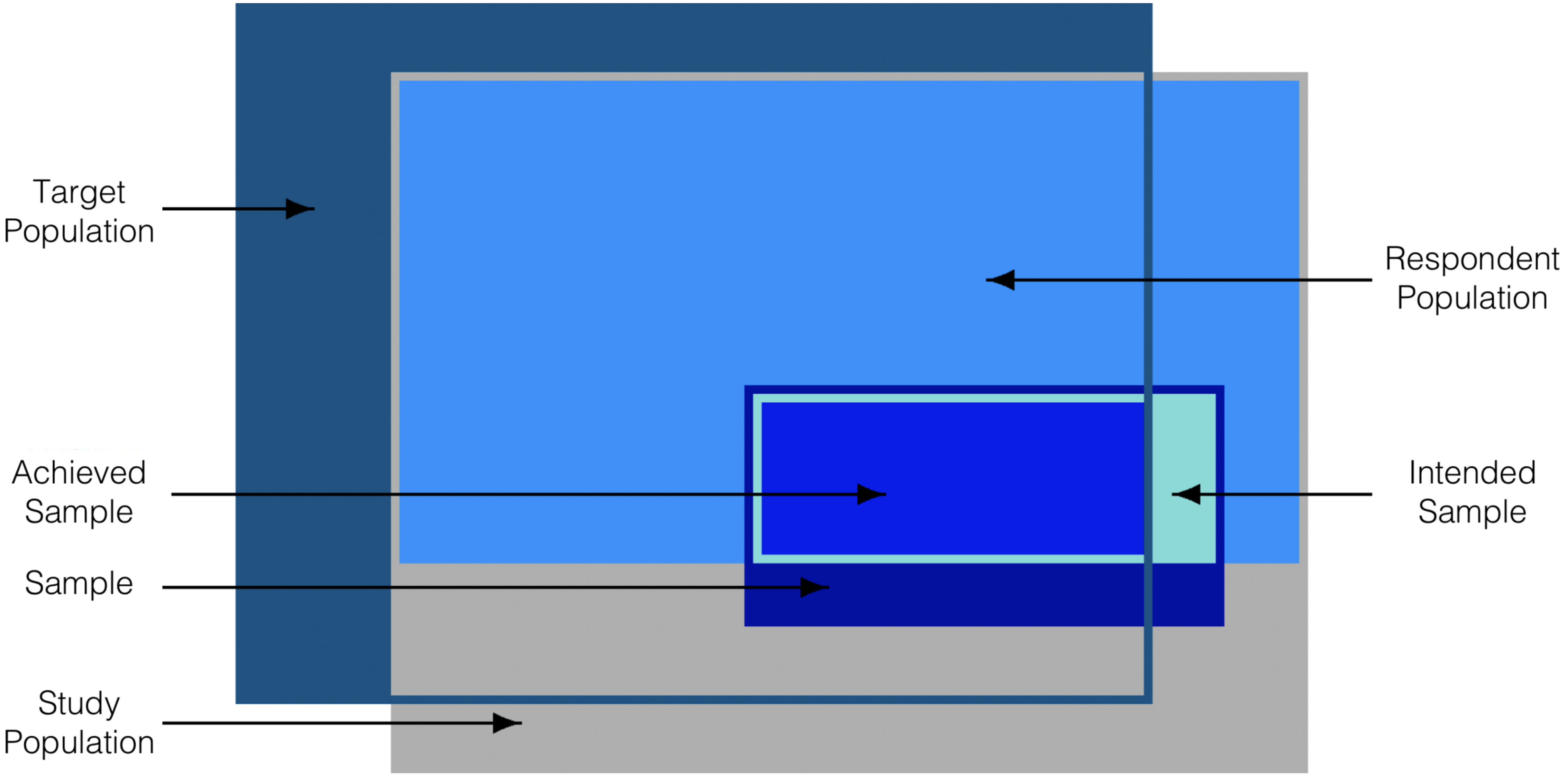
A **survey** is any activity that collects information about characteristics of interest:

- in an **organized** and **methodical** manner;
- from some or all **units** of a population;
- using **well-defined** concepts, methods, and procedures, and
- compiles such information into a **meaningful** summary form.

SAMPLING MODELS

A **census** is a survey where information is collected from all units of a population, whereas a **sample survey** uses only a fraction of the units.

When survey sampling is done properly, we may be able to use various **statistical methods** to make **inferences** about the **target population** by sampling a (comparatively) small number of units in the **study population**.


Target Population
Respondent Population
Achieved Sample
Intended Sample
Sample
Study Population

SURVEY FRAMES

The ideal frame contains identification data, contact data, classification data, maintenance data, and linkage data, and must minimize the risk of **undercoverage** or **overcoverage**, as well as the number of duplications and misclassifications (although some issues that arise can be fixed at the data processing stage).

A statistical sampling approach is contraindicated unless the selected frame is

- **relevant** (which is to say, it corresponds, and permits accessibility to, the target population),
- **accurate** (the information it contains is valid),
- **timely** (it is up-to-date), and
- **competitively priced**.

SURVEY ERROR

$$\text{Total Error} = \underbrace{\text{Sampling Error}}_{\substack{\text{survey, not}\\ \text{census}}} + \underbrace{\text{Measurement Error}}_{\substack{\text{observations not}\\ \text{measured accurately}}} + \underbrace{\text{Non-Response Error}}_{\substack{\text{non-respondents}\\ \text{having systematic}\\ \text{observation differences}}} + \underbrace{\text{Coverage Error}}_{\substack{\text{frame decay}\\ \text{and/or}\\ \text{corruption}}}$$

Statistical sampling can help provide estimates, but importantly, it can also provide some control over the **total error** (TE) of the estimates.

Ideally, TE= 0. In practice, there are two main contributions to TE: **sampling errors** (due to the choice of sampling scheme), and **nonsampling errors** (everything else).

NONSAMPLING ERROR

Nonsampling error can be controlled, to some extent:

- **coverage error** can be minimized by selecting high quality, up-to-date survey frames;
- **non-response error** can be minimized by careful choice of the data collection mode and questionnaire design, and by using "call-backs" and "follow-ups";
- **measurement error** can be minimized by careful questionnaire design, pre-testing of the measurement apparatus, and cross-validation of answers.

In practice, these suggestions are not that useful in modern times (landline-based survey frames are becoming irrelevant due to demographics, response rates for surveys that are not mandated by law are low, etc.). This explains, in part, the over-use of **web scraping** and **non-probabilistic sampling**.

NONPROBABILISTIC SAMPLING

Nonprobabilistic sampling (NPS) methods (designs) select sampling units from the target population using subjective, non-random approaches.

- NPS are quick, relatively inexpensive and convenient (no survey frame required).
- NPS methods are ideal for exploratory analysis and survey development.

Unfortunately, NPS are often used instead of probabilistic designs (problematic)

- the associated selection bias makes NPS methods unsound when it comes to inferences (they cannot be used to provide reliable estimates of the sampling error, the only component of TE under the analyst's direct control);
- automated data collection often fall squarely in the NPS camp – we can still analyze data collected with a NPS approach, but may not generalize the results to the target population.

PROBABILISTIC SAMPLING

Probabilistic sample designs are usually more **difficult** and **expensive** to set-up (due to the need for a quality survey frame), and take longer to complete.

They provide **reliable estimates** for the attribute of interest and the **sampling error**, paving the way for small samples being used to draw inferences about larger target populations (in theory, at least; the non-sampling error components can still affect results and generalisation).

CONFIDENCE INTERVALS

If the estimate $\hat{\beta}$ is unbiased, $\mathrm{E}(\hat{\beta}-\beta)=0$, then an approximate **95% confidence interval** (95% CI) for β is given approximately by

$$\hat{\beta} \pm 2\sqrt{\widehat{\mathrm{V}}(\hat{\beta})},$$

where $\widehat{\mathrm{V}}(\hat{\beta})$ is a **sampling design-specific** estimate of $\mathrm{V}(\hat{\beta})$.

But what is a 95% CI, exactly?

SAMPLING DESIGN

Different **sampling designs** have distinct advantages and disadvantages.

They can be used to compute estimates

- for various population attributes: mean, total, proportion, ratio, difference, etc.
- for the corresponding 95% CI.

We might also want to compute sample sizes for a given **error bound** (an upper limit on the radius of the desired 95% CI), and how to determine the **sample allocation** (how many units to be sampled in various sub-population groups).

SAMPLING DESIGN – UNIVERSE OF DISCOURSE

Target population:

- N units and measurements $\mathcal{U} = \{u_1, \dots, u_N\}$

True population attributes:

- mean μ, variance σ^2, total τ, proportion p

Sample population:

- n units and measurements $\mathcal{Y} = \{y_1, \dots, y_n\} \subseteq \mathcal{U}$

Sample population attributes:

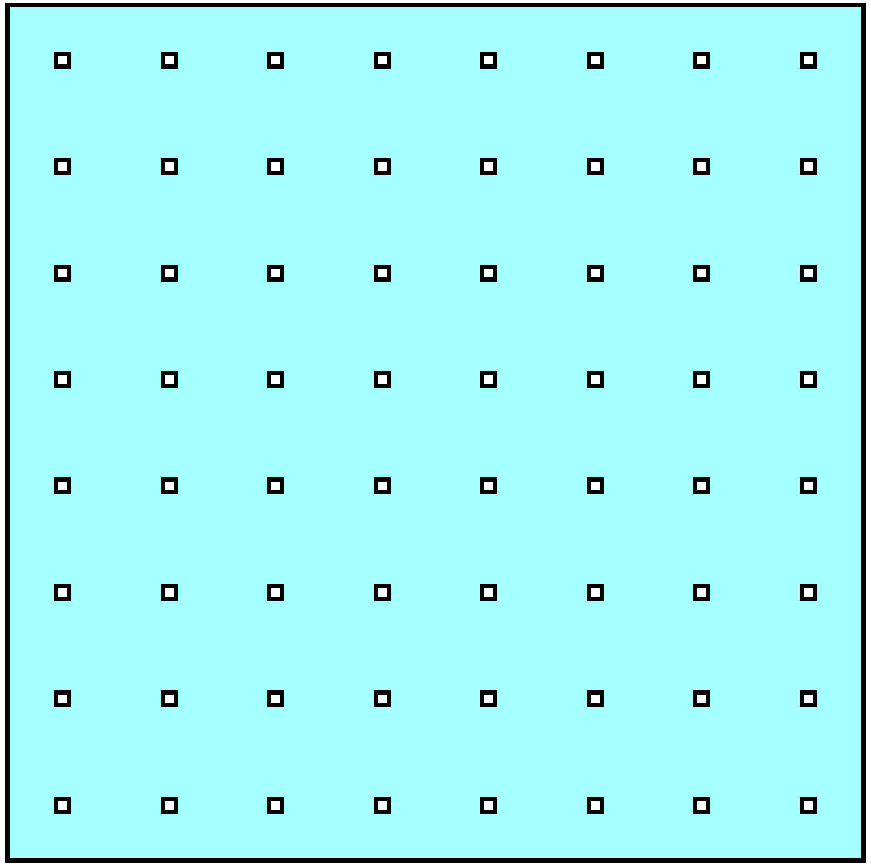
- sample mean $\bar{y}$, sample variance s^2, sample total $\hat{\tau}$, sample proportion $\hat{p}$

SAMPLING DESIGN – UNIVERSE OF DISCOURSE

Goal: estimate the true population attributes μ, σ^2, τ, p *via* the sample population attributes $\bar{y}$, s^2, $\hat{\tau}$, $\hat{p}$, n, and the size N of the target population.

For a given characteristic, we define δ_i as 1 or 0 depending on whether the sample unit y_i possesses the characteristic in question or not.

We use the error bound $B = 2\sqrt{\widehat{\mathrm{V}}}$.



SIMPLE RANDOM SAMPLING (SRS)

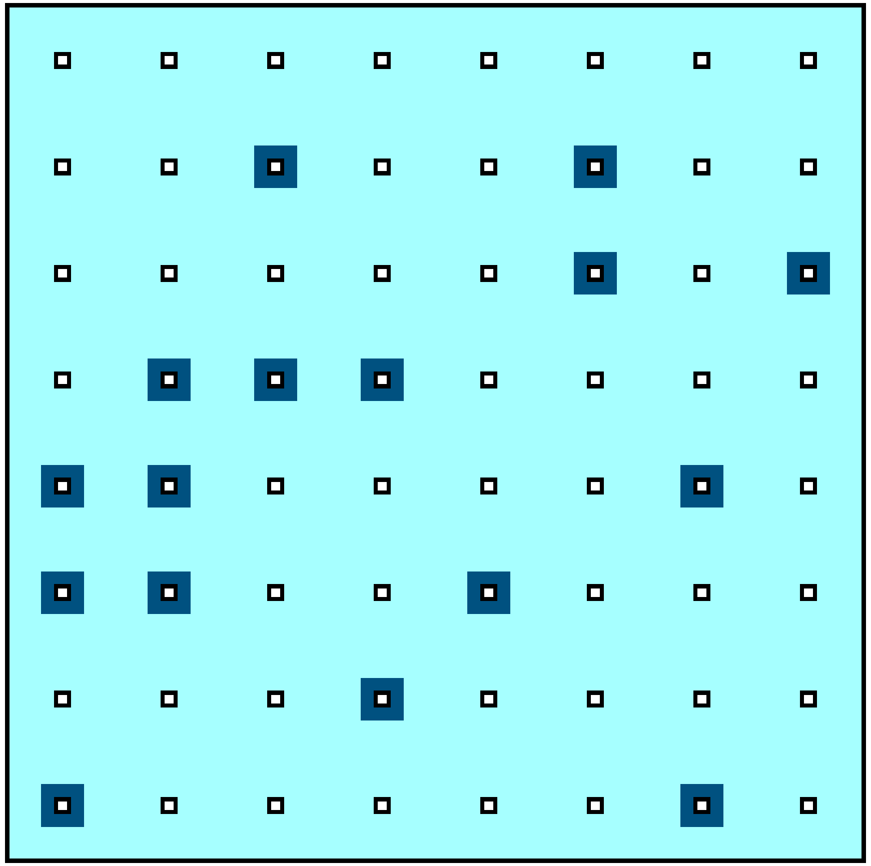
In SRS, n units are selected randomly from the frame.

Advantages:

- easiest sampling design to implement
- sampling errors are well-known and easy to estimate
- does not require auxiliary information

Disadvantages:

- makes no use of auxiliary information
- no guarantee that the sample is representative
- costly if sample is widely spread out, geographically



SRS ESTIMATORS

Estimators:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,, \qquad \hat{\tau} = N\bar{y}, \qquad \hat{p} = \frac{1}{n}\sum_{i=1}^{n} \delta_i$$

Sample Design-Specific Variance Estimates:

$$\widehat{\mathrm{V}}(\bar{y}) = \tfrac{s^2}{n}\left(1 - \tfrac{n}{N}\right), \qquad \widehat{\mathrm{V}}(\hat{\tau}) = N^2\widehat{\mathrm{V}}(\bar{y}), \qquad \widehat{\mathrm{V}}(\hat{p}) = \tfrac{\hat{p}(1-\hat{p})}{n}\left(1 - \tfrac{n}{N}\right)$$

Sample Allocation:

$$n_{\bar{y}} = \frac{4N\tilde{\sigma}^2}{(N-1)B^2+4\tilde{\sigma}^2}, \qquad n_{\hat{\tau}} = \frac{4N^3\tilde{\sigma}^2}{(N-1)B^2+4N^2\tilde{\sigma}^2}, \qquad n_{\hat{p}} = \frac{4\tilde{p}\,(1-\tilde{p})}{(N-1)B^2+4\tilde{p}\,(1-\tilde{p})}$$

STRATIFIED RANDOM SAMPLING (STS)

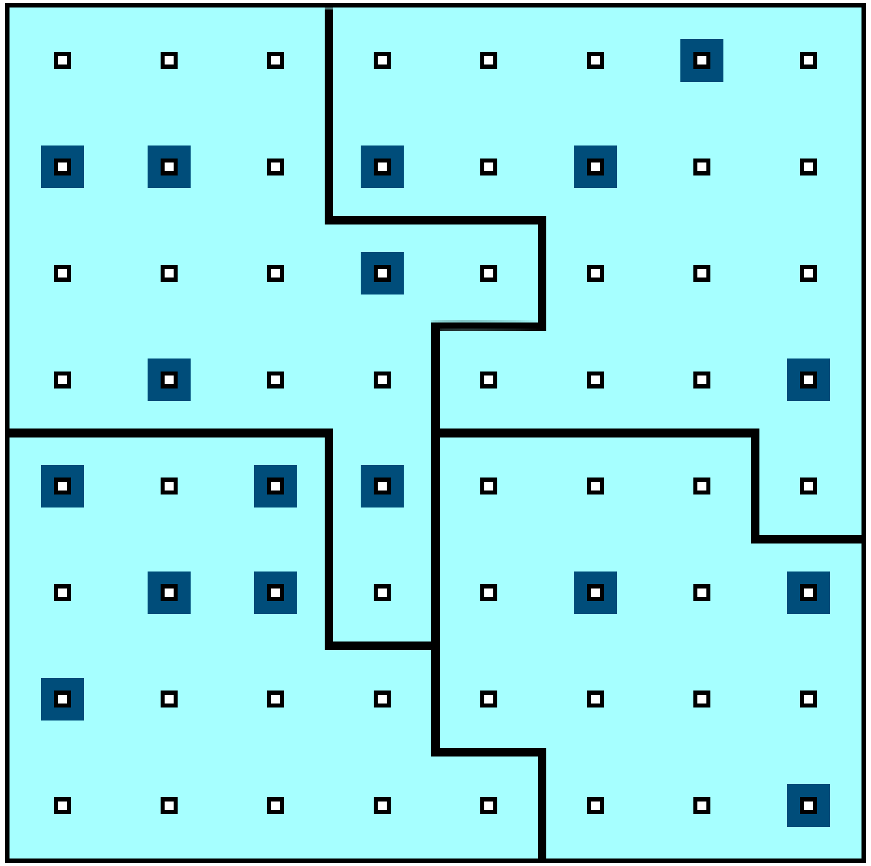
In StS, $n = n_1 + \cdots + n_k$ units are selected randomly from k frame **strata**.

Advantages:

- may produce smaller error bound on estimation than SRS
- may be less expensive if elements are conveniently strat.
- may provide estimates for sub-populations

Disadvantages:

- no major disadvantage
- if there are no natural ways to stratify the frame into homogeneous groupings, StS is roughly equivalent to SRS



STS ESTIMATORS

Estimators:

$$\bar{y}_{st} = \sum_{j=1}^{k} \frac{N_j}{N} \bar{y}_j \, , \qquad \hat{\tau}_{st} = N \bar{y}_{st}, \qquad \hat{p}_{st} = \sum_{j=1}^{k} \frac{N_j}{N} \hat{p}_j$$

Sample Design-Specific Variance Estimates:

$$\widehat{\mathbb{V}}(\bar{y}_{st}) = \frac{1}{N^2} \sum_{j=1}^{k} N_j^2 \widehat{\mathbb{V}}(\bar{y}_j) \, , \qquad \widehat{\mathbb{V}}(\hat{\tau}_{st}) = N^2 \widehat{\mathbb{V}}(\bar{y}_{st}), \qquad \widehat{\mathbb{V}}(\hat{p}_{st}) = \frac{1}{N^2} \sum_{j=1}^{k} N_j^2 \widehat{\mathbb{V}}(\hat{p}_j)$$

EXERCISES

You are charged with estimating the yearly salary of data scientists in Canada.

Identify potential:

- populations (target, study, respondent, sampling frames)
- samples (intended, achieved)
- unit information (unit, response variate, population attribute)
- sources of bias (coverage, nonresponse, sampling, measurement) and variability (sampling, measurement).

EXERCISES

The file `cities.txt` contains population information for a country's cities. A city is classified as "small" if its population is below 75K, as "medium" if it falls between 75K and 1M, and as "large" otherwise.

1. Locate and load the file into the workspace of your choice. How many cities are there? How many in each group?
2. Display summary population statistics for the cities, both overall and by group.
3. Compute a 95% C.I. for the 1999 population mean using a SRS of size $n = 10$.
4. Compute a 95% C.I. for the 1999 population mean using a StS of size $(n_s, n_m, n_l) = (5,3,2)$.
5. Compare the estimates with the true value. Are the results surprising? If not, could they have been?

Supplemental Material

DECIDING FACTORS

In some instances, information about the **entire** population is required in order to answer questions, whereas in others it is not necessary. The **survey type** depends on multiple factors:

- the type of question that needs to be answered;
- the required precision;
- the cost of surveying a unit;
- the time required to survey a unit;
- size of the population under investigation, and
- the prevalence of the attributes of interest.

STUDY/SURVEY STEPS

Studies or surveys follow the same general steps:

1. statement of objective
2. selection of survey frame
3. sampling design
4. questionnaire design
5. data collection
6. data capture and coding
7. data processing and imputation
8. estimation
9. data analysis
10. dissemination
11. documentation

The process is not always linear, but there is a definite movement from objective to dissemination.

SURVEY FRAMES

The **frame** provides the means of **identifying** and **contacting** the units of the study population. It is generally costly to create and to maintain (in fact, there are organisations and companies that specialise in building and/or selling such frames).

Useful frames contain:

- identification data,
- contact data,
- classification data,
- maintenance data, and
- linkage data.

MODES OF DATA COLLECTION

Paper-based vs. computer-assisted

- **self-administered questionnaires** are used when the survey requires detailed information to allow the units to consult personal records; associated with high non-response rate.
- **interviewer-assisted questionnaires** use well-trained interviewers to increase the response rate and overall quality of the data; face-to-face vs. telephone.
- **computer-assisted interviews** combine data collection and data capture, which saves time.
- unobtrusive direct observation
- diaries to be filled (paper or electronic)
- omnibus surveys
- email, Internet, and social media

NPS METHODS

Haphazard

- man on the street, depends on availability of units and interviewer bias

Volunteer

- self-selection bias

Judgement

- biased by inaccurate preconceptions about the target population

Quota

- exit polling, ignores non-response bias

NPS METHODS

Modified

- starts probabilistic, switches to quota as a reaction to high non-response rates

Snowball

- “pyramid” scheme

There are contexts where NPS methods might fit a client’s or an organization’s need (and that remains their decision to make, ultimately), but they must be informed of the drawbacks, and presented with some probabilistic alternatives.

BASIC MATHEMATICAL CONCEPTS

Consider a finite population $\mathcal{U}$, with N units and measurements $\{u_1, \ldots, u_N\}$.

The **mean** and **variance** of the population for the variable of interest are given by

$$\mu = \frac{1}{N}\sum_{j=1}^{N} u_j\,, \qquad \sigma^2 = \frac{1}{N}\sum_{j=1}^{N}\left(u_j - \mu\right)^2.$$

If $\mathcal{Y} \subseteq \mathcal{U}$ is a **sample** of the population with n units and measurements $\{y_1, \ldots, y_n\}$, then the **sample mean** and **sample variance** are given by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

BASIC MATHEMATICAL CONCEPTS

Let $X_1, \dots, X_n$ be **random variables**, $b_1, \dots, b_n \in \mathbb{R}$, and E, V, Cov be the **expectation**, **variance**, and **covariance** operators, respectively, i.e.:

- $\mathrm{E}(X_i) = \mu_i$
- $\mathrm{Cov}(X_i, X_j) = \mathrm{E}(X_i X_j) - \mathrm{E}(X_i)\mathrm{E}(X_j)$
- $\mathrm{V}(X_i) = \mathrm{Cov}(X_i, X_i) = \mathrm{E}(X_i^2) - \mathrm{E}^2(X_i) = \mathrm{E}(X_i^2) - \mu_i^2 = \sigma_i^2$ and

$$\mathrm{E}\left(\sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} b_i \mathrm{E}(X_i) = \sum_{i=1}^{n} b_i \mu_i$$

$$\mathrm{V}\left(\sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} b_i^2 \mathrm{V}(X_i) + \sum_{i \neq j} b_i b_j \,\mathrm{Cov}(X_i, X_j)$$

BASIC MATHEMATICAL CONCEPTS

The **bias** in an error component is the average of that error component if the survey is repeated many times independently under the same conditions. An **unbiased** estimate is one for which the bias is nil.

The **variability** in an error component is the extent to which that component would vary about its average value in the ideal scenario described above.

The **mean square error** of an error component is a measure of its size:

$$\mathrm{MSE}(\hat{\beta}) = \mathrm{V}(\hat{\beta}) + \mathrm{Bias}^2(\hat{\beta}),$$

Where $\hat{\beta}$ is an estimator of β.

PROBABILISTIC SAMPLING DESIGNS

Simple random sampling (SRS)

Stratified random sampling (StS)

Systematic sampling (SyS)

Cluster sampling (ClS)

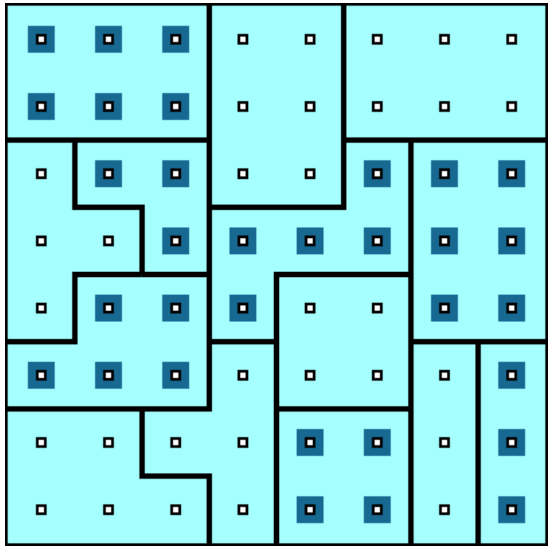
Probability proportional-to-size sampling (PPS)

Replicated sampling (ReS)

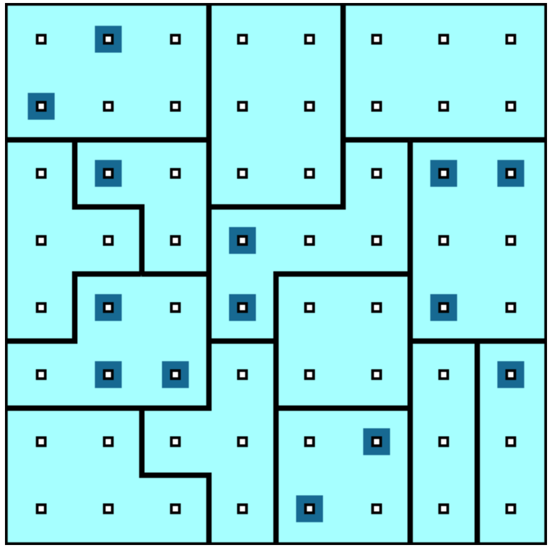
Multi-stage sampling (MSS)

Multi-phase sampling (MPS)

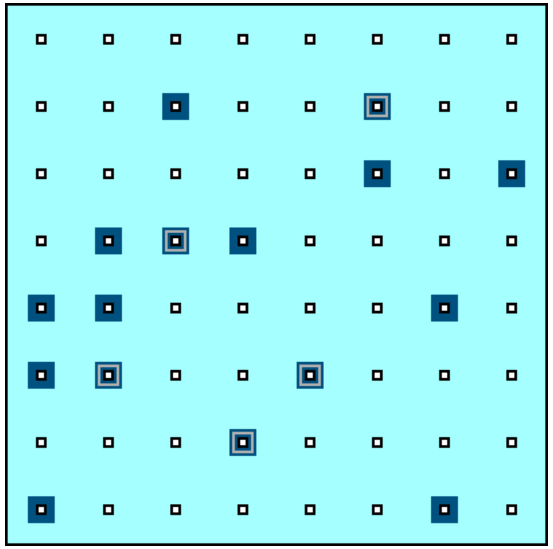
OTHER EXAMPLES OF SAMPLING DESIGNS



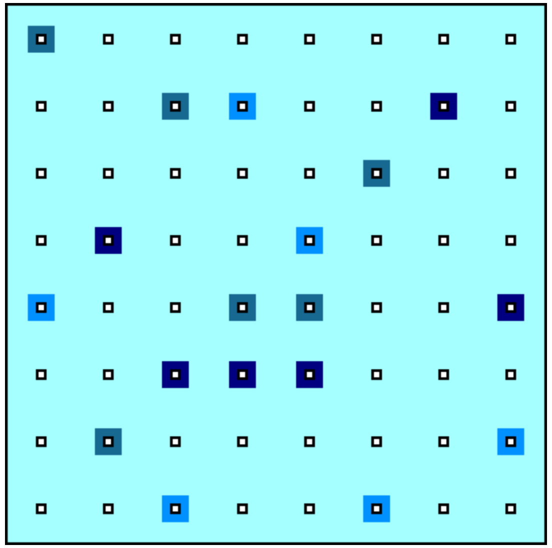
Cluster Sampling (ClS)



Multi-Stage Sampling (MSS)



Multi-Phase Sampling (MPS)



Replicated Sampling (ReS)