System Security Group | Lancaster University

Some Numerical Methods

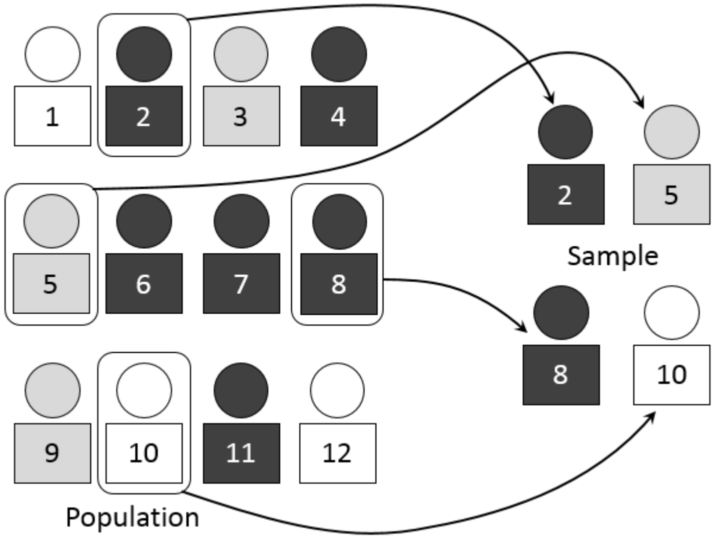
Topics

- Random sampling
- Linear regression
- Linear programming

System Security Group | Lancaster University

Random sampling

- Random sampling is the selection of a subset of a population to estimate characteristics of the population.
- Selected samples should be representative.
- Lower costs and faster data collection than measuring the entire population and can provide insights where it is infeasible to sample an entire population.
- Widely used to simulate potential risks to proactively secure systems.
- Refer to, e.g., *Cyber Security Risk Modelling and Assessment: A Quantitative Approach* & https://www.tcs.com/what-we-do/services/cybersecurity/white-paper/monte-carlo-method-quantify-cyber-risks#:~:text=Monte%20Carlo%20simulation%20constructs%20outcomes,the%20losses%20associated%20with%20them.

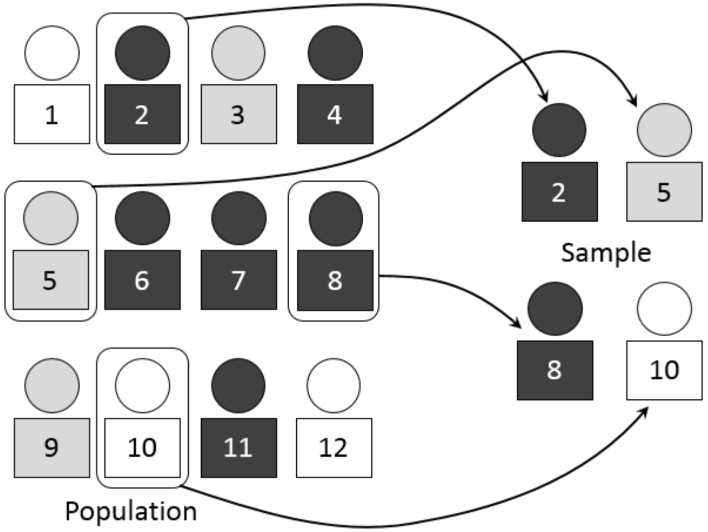


Population

- A population can be defined as including all items with the characteristics one wishes to understand.
- Sometimes necessary to sample over time, space, or some combination of these dimensions.
- The examined 'population' may be less tangible—often arises when seeking knowledge about the cause system of which the observed population is an outcome.
- The population from which the sample is drawn may not be the same as the population from which information is desired.

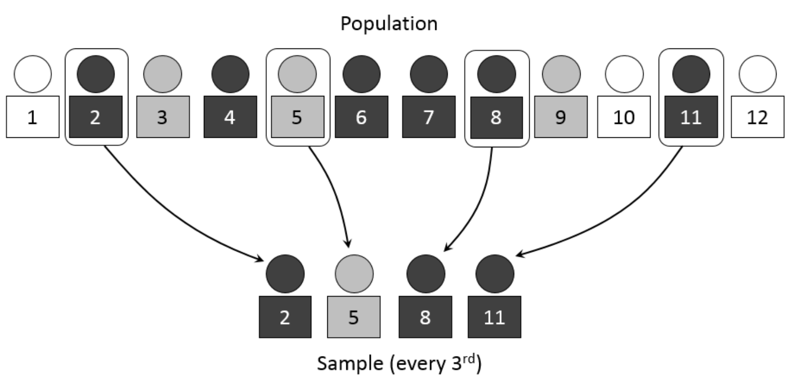
Simple random sampling

- All subsets of the same size have the same probability of being selected.
- This minimizes bias and simplifies analysis.
- Vulnerable to sampling error—selection randomness may result in a sample that doesn't reflect the makeup of the population.
- Cannot accommodate to cases where we are interested in questions specific to subgroups of the population.



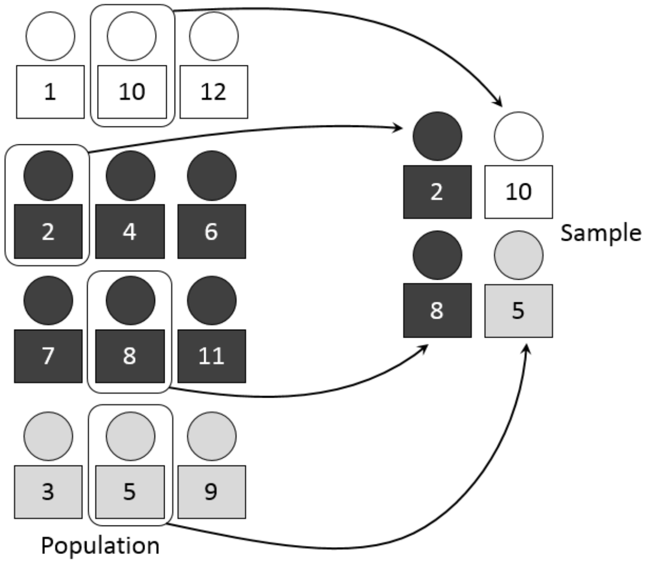
Systematic sampling

- Arrange the population by some ordering scheme. Start from a random position and proceed with selecting every *k*th element.
- It ensures that the sample is spread evenly along the list.
- Vulnerable to periodicities—unrepresentative if period is a multiple or factor of *k*.
- Difficult to quantify sampling accuracy.



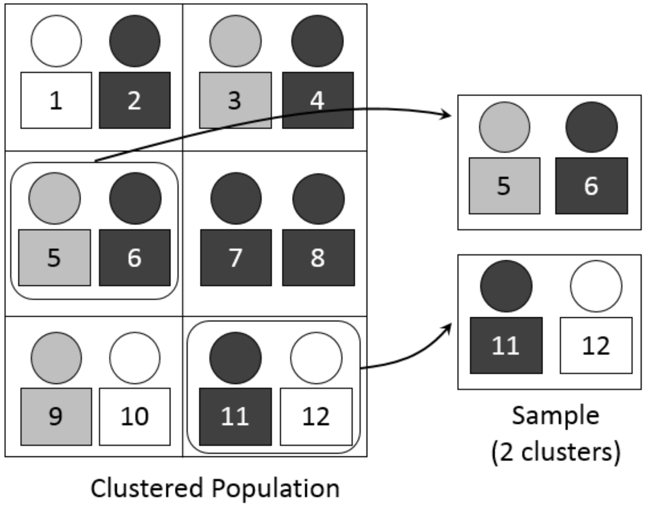
Stratified sampling

- Organize distinct categories of the population into separate 'strata', and each stratum is sampled as an independent sub-population.
- Most effective when:
  - Variability within strata are minimized;
  - Variability between strata are maximized;
  - The variables upon which the population is stratified are strongly correlated with the desired variable.
- Focus on subpopulations and ignores irrelevant ones.
- Selection of relevant stratification variables can be difficult.



Cluster sampling

- Separate the population into different clusters by geography or time, and do cluster-level sampling.
- Reduce travel and administrative costs.
- Require a larger sample than simple random sampling.



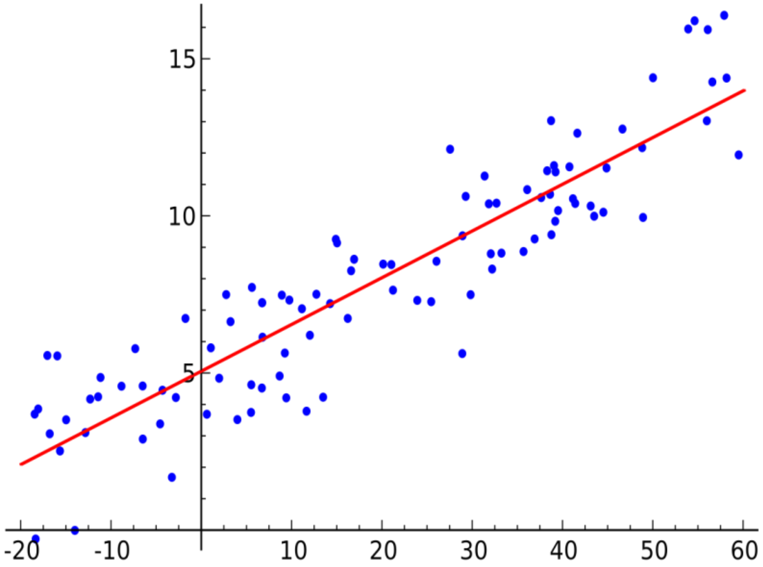
Monte Carlo Method

- Monte Carlo methods are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results.
- The underlying concept is to use randomness to solve problems that might be deterministic.
- A general pattern:
  - Define a domain of possible inputs;
  - Generate inputs randomly from a probability distribution over the domain;
  - Perform a deterministic computation on the inputs;
  - Aggregate the results.

System Security Group | Lancaster University

Linear regression

- Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables.
- In linear regression, the relationships are modelled using linear predictor functions whose unknown model parameters are estimated from the data.
- Widely used to understand relationship between multiple factors in cybersecurity applications.
- Refer to, e.g., *Identifying the Cyber Attack Origin with Partial Observation: A Linear Regression Based Approach* & https://cyberpedia.reasonlabs.com/EN/linear%20regression.html



Formulation

- Given a data set $\{y_i, x_{i1}, \dots, x_{ip}\}_{i=1}^{n}$ of n statistical units, a linear regression model assumes that the relationship between the dependent variable y and the p-vector of regressors $\mathbf{x}$ is linear.
- The relationship is modelled via a disturbance term ε that adds noise to the linear relationship between the dependent variable and regressors.
- The model takes the form $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$ for each $i = 1, \dots, n$.

System Security Group | Lancaster University

Solving for the weights

- Linear regression models are often fitted using the least squares approach. That is, we aim to find the weights $\beta_0, \beta_1, \ldots, \beta_p$ that minimize certain norm of the absolute deviations, e.g., $\sum_{i=1}^{n}(\varepsilon_i)^2$.
- Can be formulated as an optimization problem:
$$\min_{\beta_0,\beta_1,\ldots,\beta_n} \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))^2$$
- In Python, methods for solving the fitting problem have been implemented in the library *scipy.optimize*, see the function *curve_fit*.
- Again, be careful with the syntax of the function!

Applications

- Businesses often use LR to understand the relationship between advertising spending and revenue: $revenue = \beta_0 + \beta_1(£ad1) + \cdots + \beta_p(£ad1)$.
  - What does $\beta_i > 0$, $\beta_i < 0$ or $\beta_i \approx 0$ indicate?
- Medical researchers often use LR to understand the relationship between drug dosage and blood pressure: $blood\ pressure = \beta_0 + \beta_1(dosage)$.
- Agricultural scientists often use LR to measure the effect of fertilizer and water on crop yields: $crop\ yield = \beta_0 + \beta_1(amount\ of\ fertilizer) + \beta_2(amount\ of\ water)$.

System Security Group | Lancaster University

Linear programming

- Linear programming (LP), also called linear optimization, is a method to achieve the best outcome in a mathematical model whose requirements are represented by linear relationships.
- More specifically, LP is a technique for the optimization of a linear objective function, subject to linear equality and inequality constraints.
- LP is widely used to determine optimal cybersecurity investment.
- Refer to, e.g., *A linear model for optimal cybersecurity investment in Industry 4.0 supply chains*

System Security Group | Lancaster University

LP formulation

- Canonical form:

$$\max_x \; c^T x$$
$$\text{s.t. } Ax \leq b$$

- max=maximize, s.t.=subject to, x and c are both n-dimensional column vectors, A is an $m \times n$ matrix, and b is an m-dimensional column vector.
- c, A and b are problem parameters, which are given and fixed, x is called the decision variable, $c^T x$ is the objective function, and $Ax \leq b$ are the constraints.
- The purpose is to find a vector x^* such that $Ax^* \leq b$ (feasibility), and $c^T x^* \geq c^T x$ for all x such that $Ax \leq b$ (optimality).

Example

A company makes two products X and Y using two machines P and Q. Each unit of X needs 50 minutes on P and 30 minutes on Q. Each unit of Y needs 24 minutes on P and 33 minutes on Q.

At the start of the current week, there are 30 units of X and 90 units of Y in stock. Available processing time on P is 40 hours and on Q is 35 hours.

The demand for X in the current week is 75 units and for Y is 95 units. The company aims to maximize the combined sum of the units of X and Y in stock at the end of the week.

Question: Formulate the problem of deciding how many of each product to make in the current week as an LP.

Solution

- Let x and y be the number of units of X and Y to be produced in the current week, respectively.
- Constraints:
  - Machine P time: $50x + 24y \leq 40 \times 60$
  - Machine Q time: $30x + 33y \leq 35 \times 60$
  - Product X demand: $x + 30 \geq 75$
  - Product Y demand: $y + 90 \geq 95$
- Objective: maximize $(x + 30 - 75) + (y + 90 - 95)$
  - Effectively, maximize $(x + y)$

Solution

Direct formulation:

$$\max_{x,y} \ (x + y)$$
$$\text{s.t.}\ 50x + 24y \leq 2400$$
$$30x + 33y \leq 2100$$
$$x \geq 45$$
$$y \geq 5$$

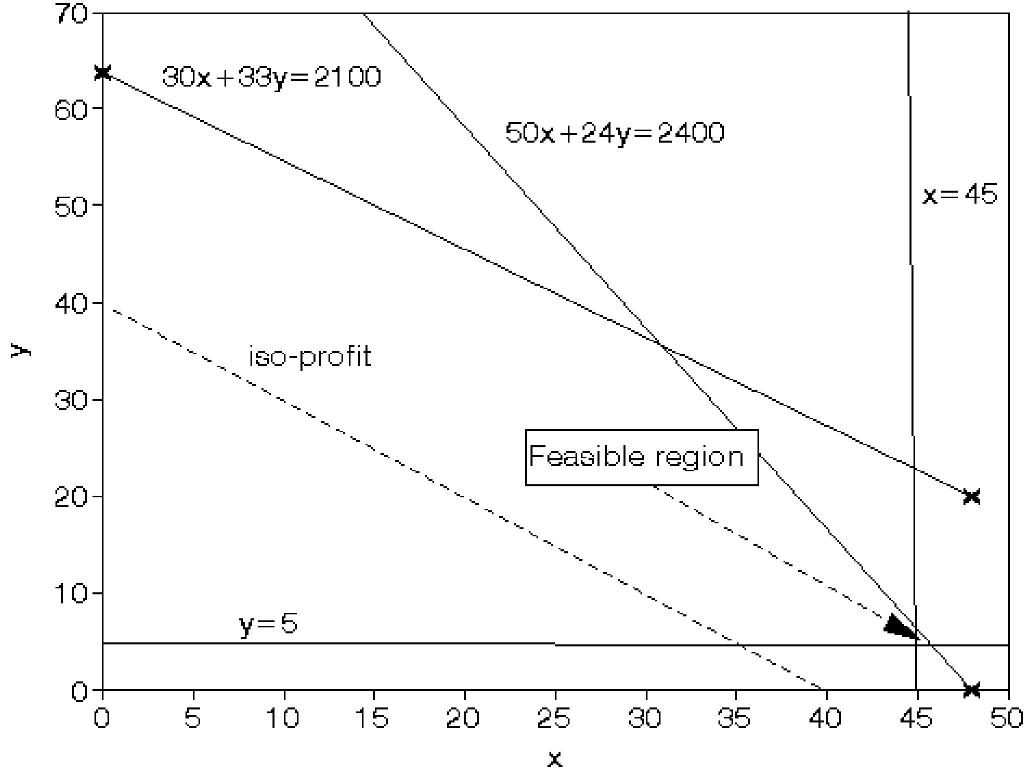
Canonical form:

$$\max_{x,y} \ \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$
$$\text{s.t.} \begin{bmatrix} 50 & 24 \\ 30 & 33 \\ -1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \leq \begin{bmatrix} 2400 \\ 2100 \\ -45 \\ -5 \end{bmatrix}$$

$$c = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, A = \begin{bmatrix} 50 & 24 \\ 30 & 33 \\ -1 & 0 \\ 0 & -1 \end{bmatrix}, b = \begin{bmatrix} 2400 \\ 2100 \\ -45 \\ -5 \end{bmatrix}.$$

Solving the LP graphically

- The maximum occurs at the intersection of $x = 45$ and $50x + 24y = 2400$.
- $x = 45$ and $y = 6.25$.
- If x and y must take integer values, then we take $x = 45$ and $y = 6$.



Solving methods for LP

- Simplex algorithms
- Interior point methods
- Details of these methods are out of our scope
- In Python, these methods have been implemented in the library *scipy.optimize*, see the function *linprog*.
- Be careful with the syntax of the function!

System Security Group | Lancaster University

Questions?