Re-Sampling

Sampling Distibutions

Dr. Farshid Alizadeh-Shabdiz
Fall 2020

Proposition 7.31. *Let X_1, X_2, …, X_n be independent with respective population means μ_1, μ_2, …, μ_n and standard deviations σ_1, σ_2, …, σ_n. For given constants a_1, a_2, …,a_n define $Y = \sum_{i=1}^n a_i X_i$. Then the mean and standard deviation of Y are given by the formulas*

$$\mu_Y = \sum_{i=1}^{n} a_i \mu_i, \quad \sigma_Y = \left(\sum_{i=1}^{n} a_i^2 \sigma_i^2 \right)^{1/2}. \tag{7.8.8}$$

Proof. The mean is easy:

$$\mathbb{E}\, Y = \mathbb{E} \left(\sum_{i=1}^{n} a_i X_i \right) = \sum_{i=1}^{n} a_i\, \mathbb{E}\, X_i = \sum_{i=1}^{n} a_i \mu_i.$$

The variance is not too difficult to compute either. As an intermediate step, we calculate $\mathbb{E}\, Y^2$.

$$\mathbb{E}\, Y^2 = \mathbb{E} \left(\sum_{i=1}^{n} a_i X_i \right)^2 = \mathbb{E} \left(\sum_{i=1}^{n} a_i^2 X_i^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_i a_j X_i X_j \right).$$

Using linearity of expectation the $\mathbb{E}$ distributes through the sums. Now $\mathbb{E}\, X_i^2 = \sigma_i^2 + \mu_i^2$ and $\mathbb{E}\, X_i X_j = \mathbb{E}\, X_i\, \mathbb{E}\, X_j = \mu_i \mu_j$ when $i \neq j$ because of independence. Thus

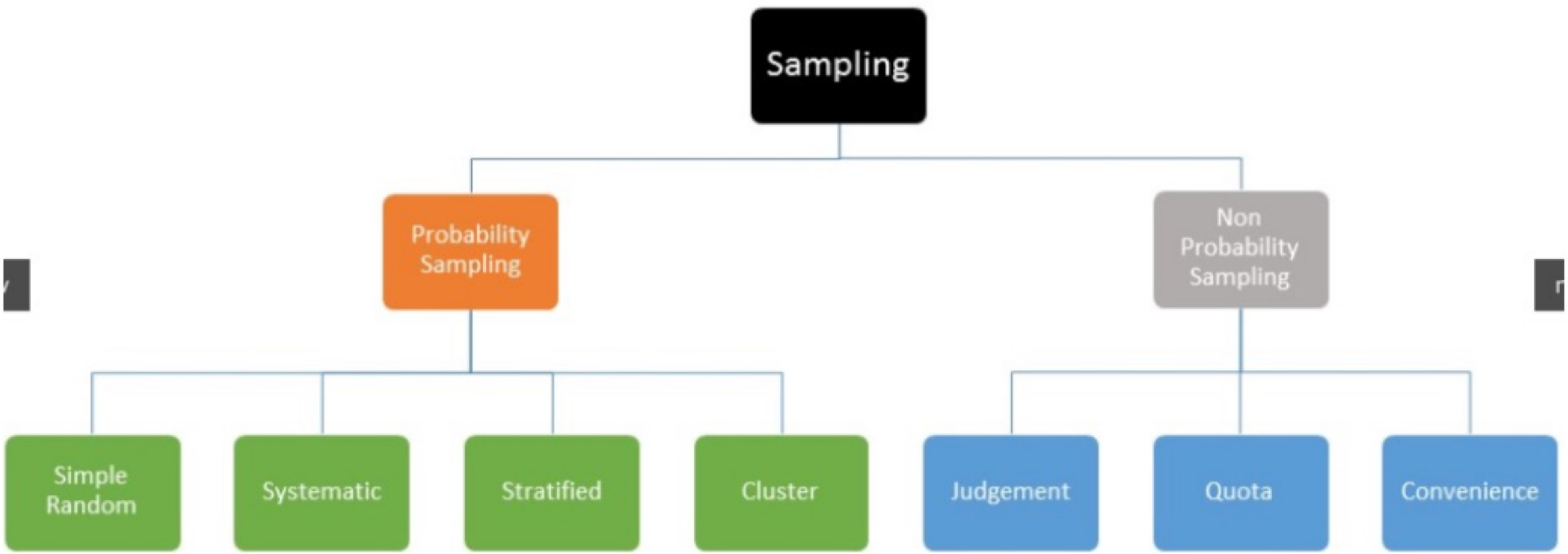
$$\begin{aligned}
\mathbb{E}\, Y^2 &= \sum_{i=1}^{n} a_i^2 (\sigma_i^2 + \mu_i^2) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_i a_j \mu_i \mu_j \\
&= \sum_{i=1}^{n} a_i^2 \sigma_i^2 + \left(\sum_{i=1}^{n} a_i^2 \mu_i^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_i a_j \mu_i \mu_j \right)
\end{aligned}$$

To complete the proof, note that the expression in the parentheses is exactly $(\mathbb{E}\, Y)^2$, and recall the identity $\sigma_Y^2 = \mathbb{E}\, Y^2 - (\mathbb{E}\, Y)^2$. □

There is a corresponding statement of Fact 7.16 for the multivariate case. The proof is also omitted here.

Sampling Methods

- Population, Frame, Sample
- Probability Samples and nonprobability samples



Reference: Prof Katathur – CS544 course

SRS – Simple Random Sampling

- R package – sampling
- srswr(n, N)
  - Simple random sample of size *n* **with** replacement from a frame of size *N*
- srswor(n, N)
  - Simple random sample of size *n* **without** replacement from a frame of size *N*

Reference: Prof Katathur – CS544 course

Systematic Sampling

- Frame partitioned into *n* groups
- Each group has $\frac{N}{n}$ items (k)
- First item of the sample
  - Randomly selected from the first group, i.e., the first k items
- Remaining items of the sample
  - Select every k^{th} item after the first selection
- Review unequal probabilities case

Reference: Prof Katathur – CS544 course

Stratified Sampling

- Data divided into subgroups (strata)
- Simple random sampling from each strata
- Strata selections proportional to size of each strata
  - Another approach to select the same number from each strata
- Strata based on one/more than one attributes
- Data should be ordered first

Reference: Prof Katathur – CS544 course

Clustering Sampling

- Example choosing students from universities across US
- Cluster sampling is defined as a sampling method where the researcher creates multiple clusters of people from a population where they are indicative of homogeneous characteristics and have an equal chance of being a part of the sample.

Testing Methods

- Hold-Out
- Cross validation
  - What is cross validation error – general term if segments have different size.

$$CV\ Error\ Rate = \sum_{i=1}^{K} \frac{N_i}{N} MSE_i$$

  - Leave-one out CV (LOOCV)

$$CV\ Error\ Rate = \frac{1}{N} \sum_{i=1}^{N} \frac{MSE_i}{1 - hi}$$

Note: hi shows impact of the sample on variance.

$$hi = \frac{1}{N} + \frac{x_i - \bar{x}}{\sum(xj - \bar{x})}$$

  - CV of K between 5 and 10 and LOOCV doesn't have variance
  - Calculate standard deviation of CV
  - Common mistake –
    - Step 1 – correlation selects the best parameters
    - Step 2 training the data
    - You should cross validation to both steps

Bootstrap Sampling

- Bootstrap is sampling with replacement
- Bootstrap application – when uncertainty is getting calculated, but we have only small sample size, like 100, bootstrap can help to generate 1000 set and calculate standard error of accuracy

Bootstrap variations

- Block bootstrap - Time series solution by looking at a block of data as a unit in sampling
- Using bootstrap, distribution of results also can be found. As a result, the confidence interval of the result also can be found
- Testing set in Bootstrapping method is the samples left over after selecting Bootstrap sample
- Bootstrapping and CV or training/testing