CS544 Module5

Suresh Kalathur

Module5

- Central Limit Theorem
- Sampling Methods
- Sample Term Project
- Review Final Exam Topics

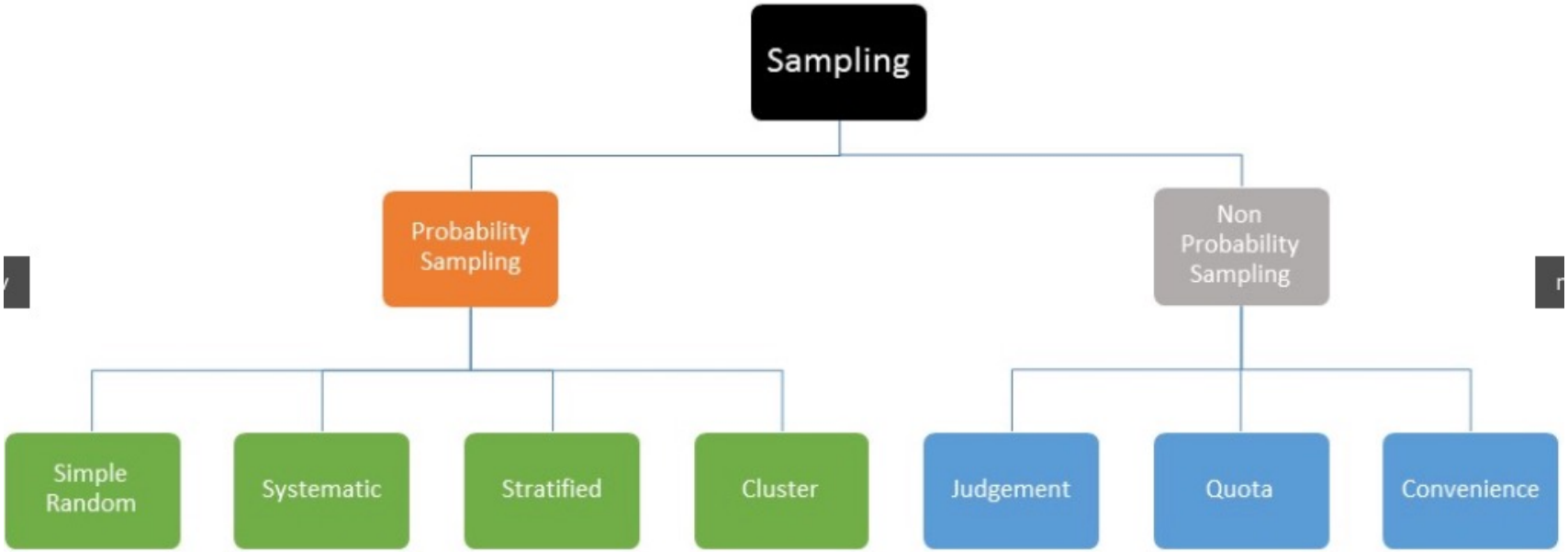
Central Limit Theorem

- Given
  - A Random variable, *x*, and a sample size (*n*)
  - Draw (all) samples of the given sample size
  - Compute the means of all the samples ($\bar{x}$)
  - Find the distribution of the sample means
    - This distribution follows a normal distribution

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

Sampling Methods

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



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*

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 (say, *r*)
- Remaining items of the sample
  - Select r^{th} item from each of the remaining groups

Systematic Sampling – **Unequal** Probabilities

- Select a numeric attribute, *x*, for inclusion probability
  - inclusionprobabilities(x, n)
- Inclusion probabilities sum will be equal to sample size
  - Max probability for any item will be 1
- Use Systematic sampling with these probabilities
  - UPsystematic(pik)
    - pik is the vector of inclusion probabilities

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 by the strata

Cluster Sampling

- Population divided into groups (clusters)
- Each cluster mirrors the population
- Single stage
  - A random sample of clusters is selected
- Two stage
  - Random sampling from each selected cluster

Errors

- Coverage errors
- Nonresponse errors
- Sampling errors
- Measurement errors
- Noise
- Data dredging

Data Visualization - Plotly

- R Graphing Library
  - https://plot.ly/r/
- Alternative to Basic R plots for Project

R Markdown Notebooks

- Weave Text and Code
- Produce output in different formats
- https://rmarkdown.rstudio.com