Sampling and Estimation

```
Key Words: Probability and Non-probability Sampling, Sampling Error,CLT,
```

1. Sampling Methods, Central Limit Theorem, Standard Error

- Sampling Error:
  - Definition: differences between a `sample statistic` and relative `population parameter`
    - Eg: sampling error of mean=sample mean-population mean= $x-\mu$
  - Sampling Distribution:
    - sample statistics is a r.v., having a probability distribution
    - the distribution of frequencies of a range of different outcomes that could possibly occur for a statistic of a population.有点难理解，简单来说就是统计量的概率分布，从population中选取不同个等量的samples，每组samples得出一个统计值，整合起来就是统计量在不同组等量samples可能出现的统计值，遵从的一个概率分布。
- Several Sampling Methods:
  - Probability Sampling Methods: selecting a sample with given probability of each sample in population.
    - Simple Random Sampling: each item has equal probability of being selected.
    - Systematic Sampling: select every nth member from population
    - Stratified Random Sampling:
      - divide population into smaller groups or stratum, then *simple random sampling* is applied in each stratum.

        - size of samples from each stratum ~ size of stratum relative to the population（层数）
        - reduce sampling error
        - Application in Finance: bond indexing.
            - Reason: bonds are classified by certain risk factors: duration, maturity, coupon rate...
    - Cluster Sampling:
        - Definition:
            - divide population into several subsets(clusters), then use ***simple random sample*** to select samples in individual group.
            - each cluster is a **representative** of population in terms of the item we sampling.
        - Types:
            - One-stage cluster sampling
            - Two-stage cluster sampling:
                - has **greater** sampling error than one-stage
        - Evaluation:
            - have **greater** sampling error than simple random sampling
            - less time and lower cost required
- Non-probability Sampling Method: select samples using the ***judgement of researchers*** and lower cost or easy access to some data items, with greater sampling error.
    - Convenience Sampling
        - definition: select sample data based on the ease of access, using data readily available
    - Judgement Sampling
        - Definition: select larger dataset based on experts experience and judgment.
        - PS: ensure the distribution of data of interest/feature is constant for population sampled.

- Central Limit Theorem(CLT)

    - reference: https://www.probabilitycourse.com/chapter7/7_1_2_central_limit_theorem.php
    - Definition: for iid samples, the standardized sample mean→standard normal distribution, regardless the original population is normal or not, when n sample size is large enough($n >= 30$)

    - Math Expression:
        - $\mu_{\bar{x}} = \mu$
        - $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt[2]{n}}$
        - $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt[2]{n}})$
        - $Z_n = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt[2]{n}}} = \frac{X_1+X_2+..+X_n-n\mu}{\sqrt[2]{n}\sigma}$
        - cdf: $lim_{n\to\infty} P(Z_n \leq x) = \phi(x)$, where $\phi(x)$ is the standard normal cdf

- Standard Error of Sample Mean

    - Definition: std of the distribution of the ***sample means***
    - With known population std σ, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt[2]{n}}$
    - With unknown population std, we use sample std s to inference population statistic. $s_{\bar{x}} = \frac{s}{\sqrt[2]{n}}$
        - less widely dispersed around expected value than a single observation with σ of r.v.
    - relies heavily on sample size, when n increase, the ***standard error of sample mean*** decreases, more close to true population level.

- Estimator Properties Description

    - Unbiased:
        - $Bias(\tilde{\theta}) = E(\tilde{\theta}) - \theta = 0$
        - the `expected value of the estimator`= the `parameter` you are trying to estimate
    - Efficient
        - $MSE(\tilde{\theta}) = Var(\tilde{\theta}) + Bias(\tilde{\theta})^2$
        - has smaller variance of sampling distribution than all other unbiased estimators.
    - Consistent
        - $lim_{n\to\infty} P(|\tilde{\theta} - \theta| \geq \epsilon) = 0$, for all $\epsilon > 0$.
        - the accuracy of parameter increases as the sample size increases.

2. Confidence Intervals, Resampling, and Sampling Biases

- Calculate CI for Population Mean with Given Variance a Normal Distribution

- Point estimates: single values to estimate population parameters
  - eg: the estimator of population mean μ: $\bar{x} = \frac{\sum x}{n}$
- Confidence Interval(CI):
  - Definition: range of values population parameter is expected to lie in, given the probability $1-\alpha$
    - α: level of significance
    - $1-\alpha$: degree of confidence
  - Calculation: `point estimate±(reliability factor * standard error)`
    - point estimate: value of a sample statistic of the population parameter
    - reliability factor: sampling distribution and probability lies in CI(通俗来说，就是标准统计量和α的取值)
  - Eg: for known variance normal distribution, CI for population mean can be expressed as follows:
    - $\bar{x} \pm z_{\alpha/2} * \frac{\sigma}{\sqrt[2]{n}}$
      - $\alpha = 10$, 5% for each tail, $z_{\alpha/2}$=**1.645** for 90% CI
      - $\alpha = 5$, 2.5% for each tail, $z_{\alpha/2}$=**1.960** for 95% CI
      - $\alpha = 1$, 0.5% for each tail, $z_{\alpha/2}$=**2.575** for 99% CI
      - ...For more, z-score table should be used
- Interpretation
  - Probabilistic interpretation: xx% of confidence intervals will include population parameter in the long run.
  - Practical interpretation: there are xx% confidence that the population parameter lies between $\bar{x} - z_{\alpha/2} * \frac{\sigma}{\sqrt[2]{n}}$ and $barx + z_{\alpha/2} * \frac{\sigma}{\sqrt[2]{n}}$ interval(比较常见)

- Calculate CI for Population Mean with Unknown Variance a Normal Distribution
  - Calculation: $\bar{x} \pm t_{\alpha/2} * \frac{s}{\sqrt[2]{n}}$
  - CI constructed by t-dist is more wider than those of z-dist.
  - the usage of t-table for getting the reliability factor values.
    - make sure we known the value of df(n-1) and α

- Calculate CI for Population Mean with Unknown Variance for Large Sample from Any Distribution(CLT applied)
  - non-normal distribution, n≥30, known population variance→z-score
  - non-normal distribution, n≥30, unknown population variance→t-value

- Resampling(to estimate the standard error of mean)
    - 这个概念可能比较陌生，reference: https://en.wikipedia.org/wiki/Resampling_(statistics)
    - Jackknife
        - calculate multiple sample means, each with **1 observation removed** from the sample for reducing bias.
        - it can be used when numbers of observation is relatively small
        - Evaluation: low cost, computationally easy
    - Bootstrap
        - Evaluation: computationally demanding, get more accuracy
        - repeatedly draw samples(size=n) from full dataset, then directly calculate std and standard error of the sample mean

- Sampling Bias:
    - Data Snooping:数据窥探
        - when repeatedly use the same dataset to search a trading rule until it is discovered.有可能具有偶然性
        - Data snooping bias:
            - the statistical significance of the pattern is overestimated, because the results are found through data snooping
            - Signal:
                1. lack of economic theory support the result
                2. many tested variables are not reported, until the rule is found.
            - Solution: change another dataset to test the rule we got whether adopt or not.(use out-of-sample data)
    - Sample Selection Bias
        - some data systemically excluded from analysis, thanks to lack of availability.
        - The observed samples are *not random*. Hence, the conclusion draw from the this sample set **cannot** applied to **population**.
        - Survivorship Bias
            - most common type of sample selection bias
            - 有点类似于幸存者偏差, the observed data only include the items currently exist, without any pervious abandoned or ceased samples.
            - eg: mutual fund performance studies
            - Solution: use samples that all started at the same time without any drop

- Look-ahead bias
  - test a relationship using sample data that wasn't available on the test date.
  - eg: price-to-book ratio calculation at the year end
  - Solution: the **estimation** of information lack variable is needed.
- Time-period bias
  - the time period for gathering data is too short or too long
    - too short:有偶然性，phenomena specific,isolated occurrence
    - too long: the fundamental economic underlie background may changed背景特征不同
      - Solution: data should divide into subgroups