Contents

Chapter 1   Quantitative Methods 数量分析

1.1 Reading 10 Sampling and Estimation 抽样估计

Framework

1. Simple random and stratified random sampling
2. Time-series and cross-sectional data
3. Central limit theorem
4. Standard error of the sample mean
5. The desirable properties of an estimator
6. Confidence interval estimate
7. Student's t-distribution
8. Five kinds of biases

1.1.1 Sampling 抽样

- 抽样方法
    - Simple random sampling 简单随机抽样法
    - Stratified random sampling 分层随机抽样法
- Sampling error 抽样误差
    - Sampling error of the mean=sample mean-population mean 抽样误差=样本统计值-总体值
      n↑ sampling error↓
- Sample statistic 样本统计量：描述样本的特征量→random variable 随机变量：取值不确定的量→probability distribution 概率分布

1.1.2 Types of Data 数据分类

- Time-series data 时间序列数据
  同一公司不同时间的一系列数据
- Cross-sectional data 横截图数据
  同一时间不同公司的一系列数据

1.1.3 Central Limit Theory 中心极限定理 ☆☆

The sampling distribution of the sample mean approaches N(μ, σ^2/n) if the sample size is sufficient large(n≥ 30).

- Central Limit Theory
    - 条件
    1. n≥ 30
    2. 总体均值μ、方差σ^2已知
    - 结论
    1. 服从正态分布
    2. $\mu_{\bar{X}} = \mu_X = \mu$, Var($\bar{X}$)=$\frac{\sigma_{\bar{X}}^2}{n}$
- Standard error of the sample mean
    - Known population variance $\sigma_{\bar{X}} = \sigma/\sqrt{n}$
      sample standard deviation 样本标准差 sample mean standard error 标准误

☐ Unknown population variance $S_{\bar{X}} = S/\sqrt{n}$

1.1.4 Properties of Estimator 估计量的性质

☐ The desirable properties of an estimator 估计量比较好的性质
  ☐ Unbiasedness 无偏性 $E(\bar{X}) = \mu$
  ☐ Efficiency 有效性 Var 最小
  ☐ Consistency 一致性
    1. $n \uparrow \rightarrow$ *准确性*↑
    2. $n \uparrow, sampling\ error \downarrow$

1.1.5 Estimation 估计

☐ 抽样推断的三种方法
  ☐ Point estimate 点估计
    $\mu = \bar{X}$
  ☐ Confidence interval estimate CI 置信区间估计
    1. 计算
    2. Width: $\left(\bar{X} + k\frac{\sigma}{\sqrt{n}}\right) - \left(\bar{X} - k\frac{\sigma}{\sqrt{n}}\right) = 2k\frac{\sigma}{\sqrt{n}}$
    3. $\bar{X} \sim N\left(\mu_{\bar{X}} = \mu_X, \sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}\right)$
  ☐ Interval Estimation 区间估计
    1. Level of Significance (alpha) 显著性水平
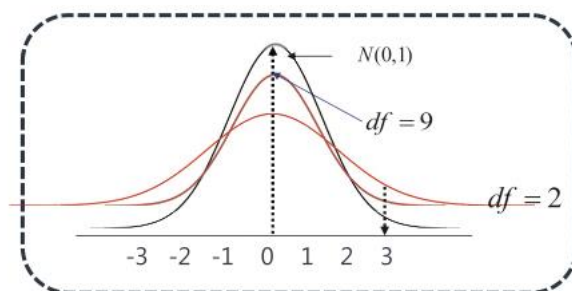   2. Degree of Confidence (1-alpha) 置信度

1.1.6 T-Distribution t 分布

☐ Student's t-distribution
  ☐ Symmetrical→skewness=0
  ☐ Degree of freedom(df): n-1
  ☐ Less peaked than a normal distribution("fatter tails")
    T 分布和正态分布相比，低峰肥尾(kurtosis<3)，方差更大($\sigma_t^2 > 1$)
  ☐ Student's t-distribution converges to the standard normal distribution as degrees of freedom goes to infinity.

$\mathrm{df}\uparrow \rightarrow n\uparrow, t\rightarrow N(0,1), \textit{峰}\uparrow \textit{尾}\downarrow, \sigma_t^2\downarrow\rightarrow 1$

- 相同α，CI$_t$更宽

- t 分布与 z 分布
  - 方差已知用 z
  - 方差未知用 t
  - 非正态总体小样本不可估计
  - 样本容量足够大，任何情况均可用 z

$$\overline{x}\pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad \overline{x}\pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

Sampling from:	Normal distribution with known variance	Normal distribution with unknown variance	Nonnormal distribution with known variance	Nonnormal distribution with unknown variance
Statistic for small sample size(n<30)	z- Statistic	t- Statistic	not available	not available
Statistic for large sample size(n>=30)	z- Statistic	t- Statistic/z	z- Statistic	t- Statistic/z

1.1.7 Types of Bias 各种偏差

- 五种偏差
  - Data-mining bias 把偶然当成必然
  - Sample selection bias 样本选择性偏差
    - Survivorship bias 生存性偏差
  - Look-ahead bias 前视性偏差
  - Time-period bias 时间段偏差