Inference

- Sampling and Estimation
- Hypothesis Testing

Sampling and Estimation

Data Types 数据类型

- **Time-series data 时间序列**
  - Observations taken over a period of time at specific and equally spaced time intervals
  - 一个主体、一个属性、**多个**时间
- **Cross-sectional data 截面数据**
  - Observations taken at a single point in time.
  - **多个主体**、一个属性、一个时间
- **Longitudinal data 纵向数据 （时间+属性）**
  - Observations over time of **multiple** characteristics of the **same** entity
  - 一个主体、**多个属性、多个时间**
  - Example: GDP, inflation for a country over 10 years
- **Panel data 面板数据 （时间+主体/截面）**
  - Observations over time of the **same** characteristic for **multiple** entities
  - **多个主体**、一个属性、**多个时间**
  - **Couple (c+p): multiple cross-sectional -> panel**

Sampling 采样

- **Simple Random** sampling: select a sample according to its probability
- **Systematic** sampling: select every n-th member from a population
- **Stratified** Random Sampling
  - Divide a population into groups (**stratum**), and sampling from each group
  - The size of samples from each **stratum** is based on the size of stratum relative to the population
  - Often used in **bond index**, group by duration, maturity, coupon rate

Sampling Distribution 采样分布

- **Sampling distribution of the sample statistic**
  - The sampling distribution of **sample statistic** is done by repeating this n times
    - Take a sample
    - Compute the sample **statistic (such as mean)**
- **Sampling Error 采样误差**
  - The difference between a sample statistic and its population parameter
    - Sampling error of the mean = sample mean - population mean

Sampling Distribution of the mean

- Define
  - Original distribution (μ, σ^2)
  - Repeat n times
  - New distribution $(\bar{X}, \frac{\sigma^2}{n})$

- **Central Limit Theorem**
    - As long as the number of samples is large $n \geq 30$, the sampling distribution approaches a **normal** distribution regardless of the **original** distribution
- **Standard error of the sample mean 标准误差**
    - The **standard deviation** of the distribution of the sample means
    - Known population variance σ^2
        - $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
    - Unknow population variance s^2
        - $s_{\bar{x}} = \frac{s}{\sqrt{n}}$

Estimator – Desirable properties

- Unbiased 无偏 （估计的平均值等于种群的参数）
    - Expected value of the estimator is the population parameter
    - $E(\bar{x}) = u$
- Efficient 有效（方差是估计中最小的）
    - An unbiased estimator is efficient if the variance of the sampling distribution is **smaller** than all other unbiased estimators of the parameter you are trying to estimate
- Consistent 一致 （精度随着数量而提升）
    - Accuracy of the estimation increase as the sample size increases
    - Standard error decreases as n increases

Estimation 估计

- Point estimation 点估计
    - mean
- Confidence Interval estimation 区间估计
    - Mean and standard error

Student's t-distribution

- Properties
    - Symmetric
    - Degree of freedom $df = n - 1$
    - fat tail than normal distribution
    - Flatter but have thicker tails
    - As n increase, it approaches a **standard** normal distribution
        - more peaked and having less fat tails
- Application
    - **unknown** variance, **Small** sample $n < 30$ from population and a (approximately) **normal** distribution 未知方差，小样本、近似正态分布
    - Unknown variance, large sample with any distribution 未知方差，大样本

Confidence Interval 置信区间

- Confidence interval

- The range of values within which the actual value of parameter will lie, with a probability of $1 - \alpha$
  - **point estimate $\pm$ reliability factor $\times$ standard error**
- Degree of Confidence $1 - \alpha$
- Level of Significance α

- **Known variance and normal population, confidence interval**
  - $\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
  - $P(X < -Z_{\alpha/2}) = \frac{\alpha}{2}$ lower tail probability
  - $P(X > Z_{\alpha/2}) = \frac{\alpha}{2}$ upper tail probability
- **Commonly used reliability factors**
  - 90% -> 1.645
  - 95% -> 1.960
  - 99% -> 2.575
- **Interpretation**
  - Probabilistic: 99% of the confidence intervals will, in the long run, include the population mean
  - Practical: 99% confident that the population mean is within the range
- **Unknown variance and normal population, confidence interval**
  - $\bar{x} \pm t_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ (t-distribution, df=n-1)
- **Rules**
  - **Large Sample 大样本都可以用 z-statistic**
    - Variance known -> Z-statistic
    - Variance unknow -> T-statistic or Z-statistic (but Z-statistic yield conservative range)
  - **Small Sample and Normal Distribution 小样本正态分布**
    - Variance known -> Z-statistic
    - Variance unknow -> T-statistic
  - **Small Sample and Non-Normal Distribution 小样本非正态分布**
    - No valid test statistic

Sampling Mean Test Statistic

Sample Size & Distribution	**Variance**	**Large Sample**
Large	Known	Z-statistic
Large	Unknown	T-statistic or Z-statistic*
Small & Normal	Known	Z-statistic
Small & Normal	Unknown	T-statistic
Small & Non-Normal		×

Can use Z-statistic, but use t-statistic yields more conservative range

Sampling Mean Test Statistic

		Sample (N>30?)	
Distribution	**Variance**	**Small Sample**	**Large Sample**
Normal	Known	Z-statistic	Z-statistic
Normal	Unknown	T-statistic	T-statistic or Z-statistic*
Non-normal	Known	×	Z-statistic

Non-normal	Unknown	×	T-statistic or Z-statistic*

Can use Z-statistic, but use t-statistic yields more conservative range

Sampling Bias

- **Limitation of "Large is better"**
    - May select observations from another population
    - High cost
- **Data-mining Bias**
    - Find untrue pattern
    - Many different variables are tested, most of which are unreported, until significant ones were found
    - Lack of **economic theory** that is consistent with the results
    - Overcome: Use out-of-sample data
- **Sample Selection Bias**
    - Some data is systematically **excluded** from the analysis, lack of availability
- **Survivorship Bias**
    - Biased upward
    - Solution: funds started at the same time and no dropout
- **Look-Ahead Bias**
    - Use sample data that was not available on the test data
- **Time-period Bias**
    - The time period is too short (wont' appear in the future) or too long (relationship may have changed)

Hypothesis Testing

Hypothesis Testing Procedure

- Define **hypothesis**
- Choose test **statistic**
- Specific level of **significance**
- State the decision **rule** (i.e., reject area)
- Collect data and compute **sample** statistic
- Make a **decision**

Hypothesis Pair

- A statement about the value of a **population** parameter
- **Null Hypothesis $\mathbf{H_0}$ 原假设**
  - The hypothesis we want to reject, always contains the equal sign
- **Alternative Hypothesis $\mathbf{H_a}$ 备择假设**
  - The hypothesis we want to “accept”

Hypothesis Side

- **Two-sided 双边（等式）**
  - Test for equality
  - $\mathrm{H_0}: \mu = \mu_0 \; and \; H_a: \mu \neq \mu_0$
- **One-sided 单边（不等式）**
  - Test for inequality
  - One critical value
  - $\mathrm{H_0}: \mu \leq \mu_0 \; and \; H_a: \mu > \mu_0$
  - $\mathrm{H_0}: \mu \geq \mu_0 \; and \; H_a: \mu < \mu_0$

Decision Rule – Reject Area 拒绝域

- **Two-sided**
  - $\mathrm{H_0}: \mu = \mu_0 \; and \; H_a: \mu \neq \mu_0$
  - Two critical values: $\pm Z_{\alpha/2}$
  - Reject rule: $|\textbf{test statistic}| > \pm \boldsymbol{Z_{\alpha/2}}$
- **Upper Tail**
  - $\mathrm{H_0}: \mu \leq \mu_0 \; and \; H_a: \mu > \mu_0$
  - One critical value: Z_α
  - Reject rule: $\textbf{test statistic} > \boldsymbol{Z_\alpha}$
- **Lower Tail**
  - $\mathrm{H_0}: \mu \geq \mu_0 \; and \; H_a: \mu < \mu_0$
  - One critical value: $-Z_\alpha$
  - Reject rule: $\textbf{test statistic} < -\boldsymbol{Z_\alpha}$

Type I and Type II Errors 一类和二类错误

- **Type I error 拒绝真原假设**
  - Reject the null hypothesis when it is actually true
  - P(type I error) = α
- **Type II error 接受假原假设**

- Fail to reject the null hypothesis when it is actually false
- P(type II error) = β

- **Power of a test 功效（拒绝了假原假设）**
  - The probability of reject a false null hypothesis
  - Power= $1 - \beta$
- **Relation**
  - **inverse** relation between Type I error and Type II error (most of the time)
  - **inverse** relation between Type II error and test power
- **How to increase power of test?**
  - Fix sample size: increase test power also increase Type I error
  - Fix significance level: increase **sample** size to
    - **Decease** both type I and type II errors
    - increase test power

Decision	*True Condition*	
	H_0 is true	H_0 is false
Do not reject H_0	Correct decision	Incorrect decision **Type II error**
Reject H_0	Incorrect decision **Type I error** Significance level, α, = P(Type I error)	Correct decision Power of the test = 1 – P(Type II error)

Test Statistic - Mean

- **Test statistic**
  - $\text{test statistic} = \frac{sample\ statistic - hypothesized\ value}{standard\ error\ of\ sample\ statistic}$
- **Reject Region**: $|\text{test statistic}| > \text{critical vlaue}$
- **Accept Region**: $-critical\ value\ \leq\ \text{test statistic} < \text{critical vlaue}$
- **Confidence Interval 置信区间**
  - $sample\ statistic - critical\ value \times standard\ error \leq \text{hypothesized value} \leq \text{sample statistic} + critical\ value \times standard\ error$

Significance

- Statistical significance
- Economic significance
  - Transaction cost
  - Tax
  - Long period, more samples, add additional risk

P-Value

- The probability of observing a test statistic larger than the critical value assuming the null hypothesis is true
- $\mathbf{P\ Value} = \mathbf{P}(|\boldsymbol{test\ stastic}| > \boldsymbol{critical\ value}\ |\mathbf{H_0}\ \boldsymbol{is\ true})$
- It is the probability of the critical value
- It is the **smallest** significance level for which the null hypothesis to be rejected

Level of Significance	*Two-Tailed Test*	*One-Tailed Test*
0.10 = 10%	±1.65	+1.28 or −1.28
0.05 = 5%	±1.96	+1.65 or −1.65
0.01 = 1%	±2.58	+2.33 or −2.33

Test Statistic

- Z -statistic
  - $\text{Z Test} = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}$
  - Condition
    - Large sample size (unknown variance t-statistic is better)
    - Small sample size, Normal distribution, Known Variance
- T-Statistic
  - $\text{T test} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}, \text{df} = \text{n} - 1$
  - Condition
    - Large sample size and unknown variance
    - Small sample size, Normal distribution, Unknown Variance

T-test (Difference in Means)

- **Assumption**
  - Unknown Variance
  - Independent and normally distributed
- **Equal Variance - Pooled T-test**
  - Assume variance is the same
  - $\text{H}_0: \mu_1 - \mu_2 = 0$
  - $\text{T test} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$
  - $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ is the pooled variance
  - Degree of freedom: $n_1 + n_2 - 2$
- **Unequal Variance**
  - $\text{H}_0: \mu_1 - \mu_2 = 0$
  - $\text{T test} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

- Degree of freedom: $\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2 / \left(\frac{(s_1^2/n_1)^2}{n_1}+\frac{(s_2^2/n_2)^2}{n_2}\right)^{1/2}$

Paired T-test (mean of difference)

- **Assumption**
  - Dependent and **normally** distributed
- **Paired difference**
  - $\mathrm{H_0}: \mu_\mathrm{d} = 0$
  - Use the difference $\mathrm{d_i} = \mathrm{X_i} - \mathrm{Y_i}$, convert to **one population** test
  - Mean of difference $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$
  - Sample variance of difference $s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$
  - Test statistic $t = \frac{\bar{d}-\mu_d}{\mathrm{s_d}/\sqrt{n}}$
  - Degree of freedom $n-1$

Chi-Square Test

- **Normally** distributed
- Test variance
- $\mathrm{H_0}: \sigma^2 = \sigma_0^2$
- $\chi_{n-1}^2 = (n-1)\frac{s^2}{\sigma_0^2}$, s is the sample variance
- Degree of freedom $n-1$

F-test

- **Independent** and **normally** distributed two populations
- $\mathrm{H_0}: \sigma_1^2 = \sigma_2^2$
- $F = \frac{s_1^2}{s_2^2}$
- Degree of freedom $(n_1 - 1, n_2 - 1)$
- Trick
  - Fact: Lower critical value = 1 / upper critical value
  - Always put the **larger** variance in the numerator, so we only need to test the upper tail, and get the conclusion for the lower tail 测试方便
  - For two-sided test, still need to use $\frac{\alpha}{2}$ in the upper tail 还是双边测试的

Parametric and Non-parametric

- Parametric test
  - Assume population distribution: Most rely on normal distribution or a large sample size to use the central limit theorem
- Non-parametric test
  - No assumption about the population distribution
  - Situations
    - Distribution **assumption** is not met. Such as a small sample size and non-normal distribution
    - data are **ranks** rather than values
    - hypothesis does not involve population parameters, such as testing whether a variable is normally distributed.

- **Run test**
    - A series of changes are random
- **Spearman** rank **correlation** test
    - Data is not normally distributed

<table>
<tr><th>Parameter</th><th>#Population</th><th>Condition</th><th>Test</th></tr>
<tr><td rowspan="2">Mean</td><td rowspan="2">One</td><td>Known Variance
• Large sample
• Small Sample and Normal</td><td>Z-Test</td></tr>
<tr><td>Unknown Variance
• Large Sample
• Small Sample and Normal</td><td>T-test</td></tr>
<tr><td rowspan="3">Mean</td><td rowspan="3">Two</td><td>Independent and Normal
Equal Unknown Variance</td><td>Pooled T-test</td></tr>
<tr><td>Independent and Normal
Unequal Unknown Variance</td><td>Independent T-test</td></tr>
<tr><td>Dependent (use the difference)</td><td>Paired T-test</td></tr>
<tr><td>Variance</td><td>Two</td><td>Independent and Normal</td><td>Chi-Square Test</td></tr>
<tr><td>Variance</td><td>Two</td><td>Independent and Normal</td><td>F-test</td></tr>
</table>