# Chapter 1 Rates and Returns

## Interest Rate作用：要求收益率/折现率/机会成本

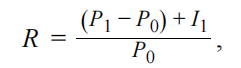
* **Nominal interest rate =** Nominal risk-free interest rate+ Default risk premium（违约） + Liquidity premium（变现） + Maturity premium（期限）

**Nominal risk-free interest rate =** Real risk-free interest rate + Inflation Rate【以加法或者乘法表示均可】

* 

## Holding Period Return

### 包括价格变动与利息收入（股息/债息）



### 计算整个收益期时：

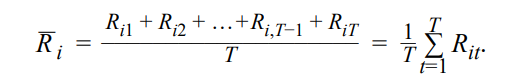


## 三种计算mean return的方式

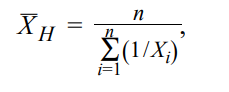
### Geometric【一般是同一资产+相等周期】



### Arithmetic【不同资产但投资初始金额相同】

****

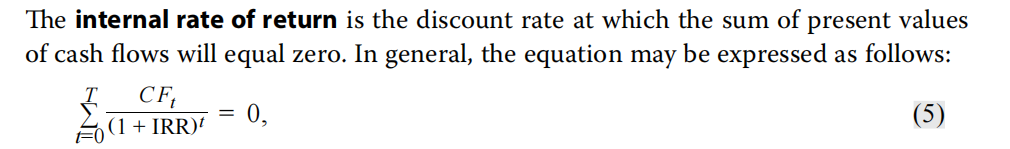
### Harmonic【cost averaging成本平均投资策略，固定金额的定期投资】

****

* **使用Harmonic调和平均数可以更好地去除样本中的极端值影响**

## Money-weighted return

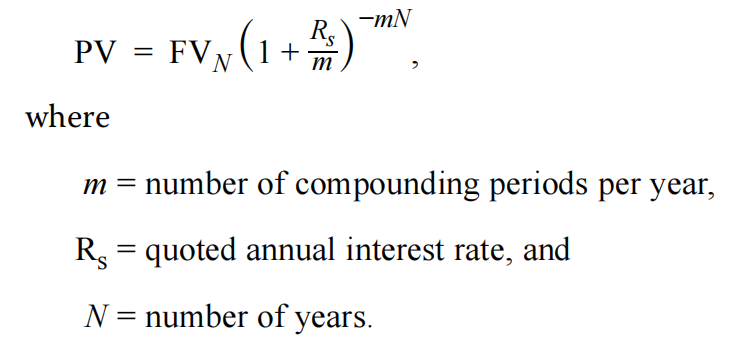
* 与IRR的计算方式一致，使得未来现金流现值为0的利率。



## Time-Weighted Returns时间加权收益率

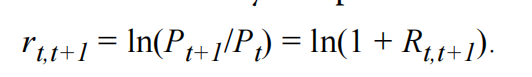
****

### 对于非连续compounding

****

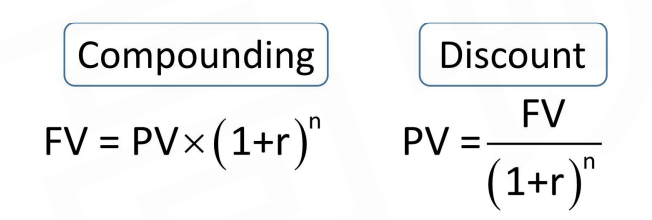
### 对于连续compounding



****

# Chapter 2 Time value of money

## Future Value and Present Value

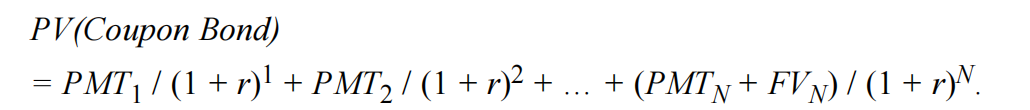


* **Ordinary annuity:** all constant cash flows occurring **at the end of each period**
* **Annuity due:** all constant cash flows occurring **at the beginning of each period**

## 债类

* **会在固收章节详细介绍，此处大致了解即可**
  1. **对于付息债券**

PMT（每期要支付的利息）

* + 如果PV小于面值：Price at discount折价发行。票面利息coupon rate＜实际利率r
  + 
  1. **对于永续债券perpetual bond/ Perpetuity**

每期都支付固定利息PMT，永续支付。

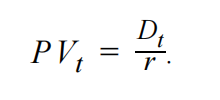
* + **Perpetuity:** a set of constant **never-ending sequential cash flows** occurring at the end of each period
  + 

## 股类

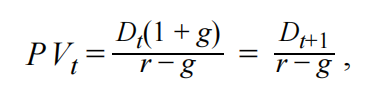
* **会在权益章节详细介绍，此处大致了解即可**

——认为股票的价格反应股票的未来现金流（股息）折现

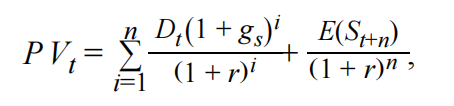
* 1. **对于优先股或者恒定股息的普通股**

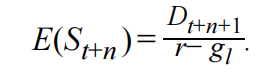


* 1. **对于股息以g增长的股在t时刻**

****

* 1. **对于股息增长有差异的股**

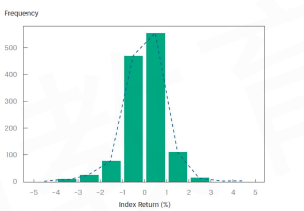
****

其中：

## Cash flow additivity principle:只有在相同时点的钱数量上才可以相加

# Chapter 3 Statistical Measures of Asset Returns

## 【一阶矩】中心趋势central tendency

* 一般先用直方图Histogram观测
* 

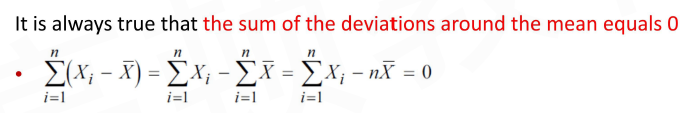
### Mean

**Arithmetic Mean ≥Geometric Means≥Harmonic Means**

【算几调】--所有数都相等时取等号，差距越大数据越分散

#### **Arithmetic Mean:** Population Mean和Sample Mean

* + - 1. 容易被极端值（Outliers）影响
      2. 使用deviation计算

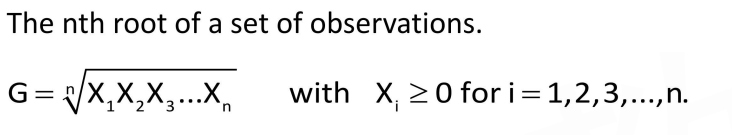


* + - 1. 应对outlier：
         1. 如果观测无偏，反映全部样本的分布，可以不调整
         2. 使用**trimmed mean切位平均数**——去掉最高**和**最低的样本值
         3. 使用**winsorized mean**——将最高**和**最低值替换

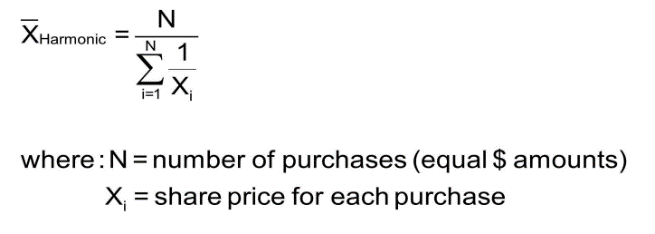
#### **Weighted Mean**

* 常用于计算portfolio return or the expected value based on probabilities

#### **Geometric Mean**

* 
* 常用于计算多期收益率

#### **Harmonic Mean调和平均数**

* + - * 倒数的平均数的倒数
      * Eg：计算每只股票买100块的平均成本

### Median

* 奇数是（n+1）/2位置的数，偶数是n/2和（n+2）/2两个位置的平均数

### Mode（众数）

* 可能没有众数或者多个众数，适用于大多数
* 连续性数据可能没有modal outcome，使用区间bins，找到modal interval
* 对于定性的数据nominal data（分类的），众数是唯一可以测量中心趋势的工具

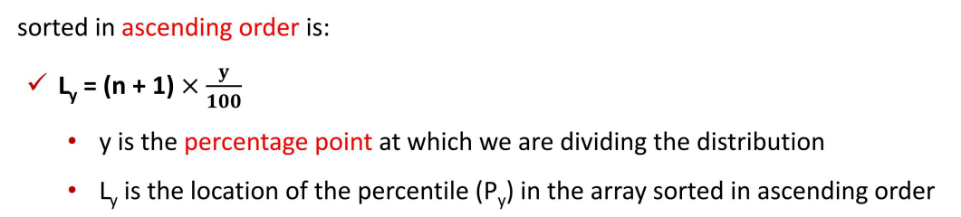
## 【一阶】Measures of location--- Quantiles（分位数）

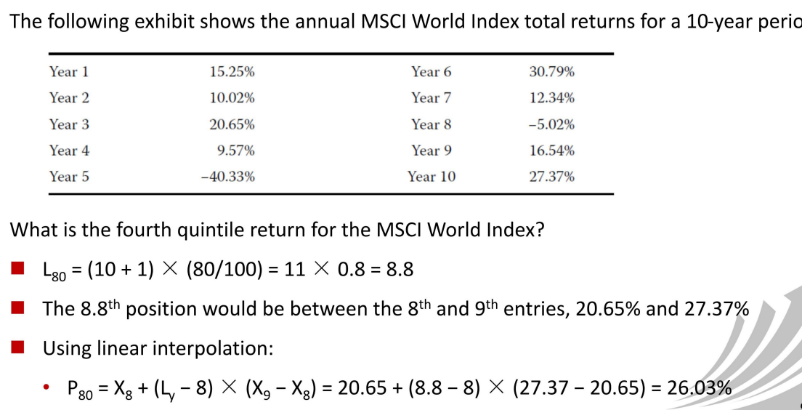
* + Quantiles（分位数）

### 多个名词

* Quartiles四分位——interquartile第一刀和第三刀之间的距离
* Quintiles 五段
* Deciles: 十分段
* Percentiles: 百分段

### 求分位数的值：先从小到大排序，代入公式求出所在位置数，利用同比例法求具体值





### 可视化：箱线图box and whisker plot

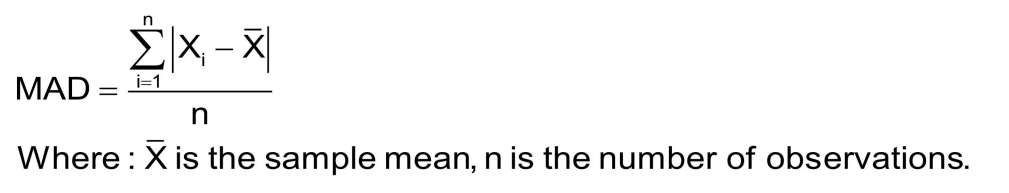


## 【二阶矩】Dispersion（离散程度）——为了讨论风险

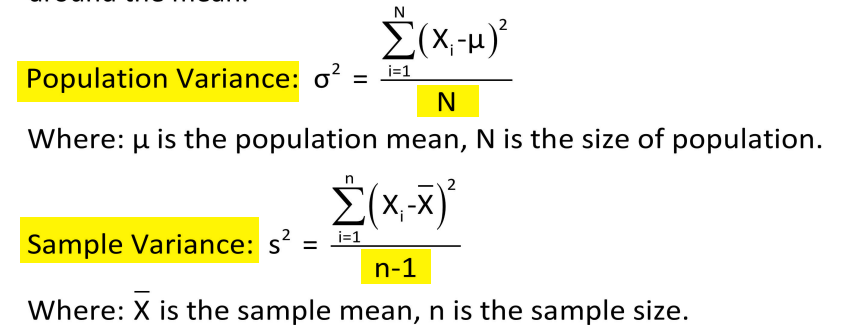
### Range 极差

* 只用极大极小值信息比较少，可以用分位数的极差去看离散

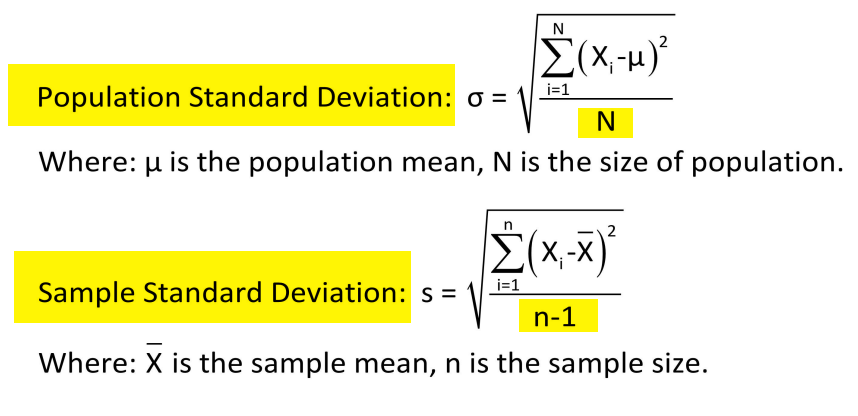
### Mean absolute deviation 平均绝对差



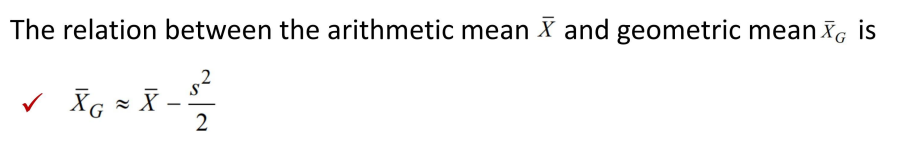
### Variance 方差

****

### Standard deviation 标准差

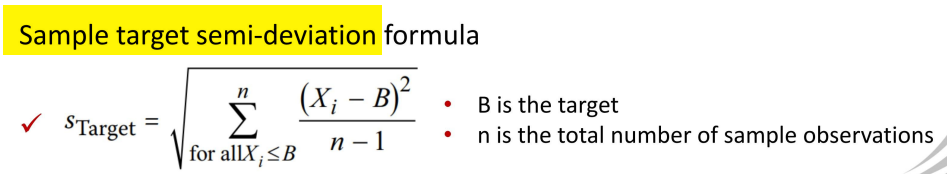


### 算数平均数、几何平均数、样本方差关系

* + 1. 
    2. 所以：离散程度越大（方差越大），平均数之间差值也越大

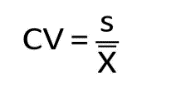
### Downside Risk

* 由于risk本身是不确定性，但在现实中大家更关注亏损
* 所以使用半方差：**Target semi-deviation**

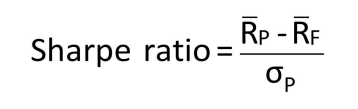


* 当B值变高，S变大
* 如果B＜算数平均数，S<样本方差；如果B＞算数平均数，不一定，当B足够大，才会超过样本方差

### 变异系数Coefficient of Variation 【CV】

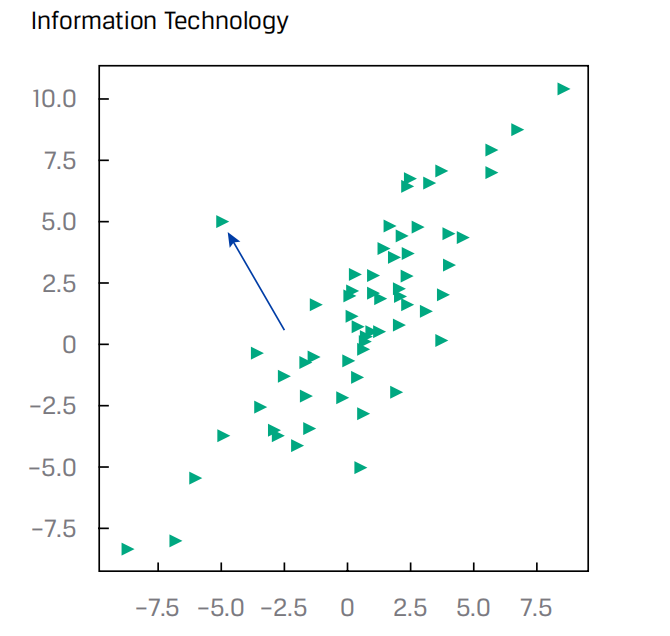
* + - * ****
      * CV没有单位，相当于是单位收益对应的风险，越低越好

### Sharpe ratio

* + - * 
      * 没有单位，承担单位风险获得的收益，越高越好

## 变量之间关系

### Scatter plot：先画出来大致观测关系

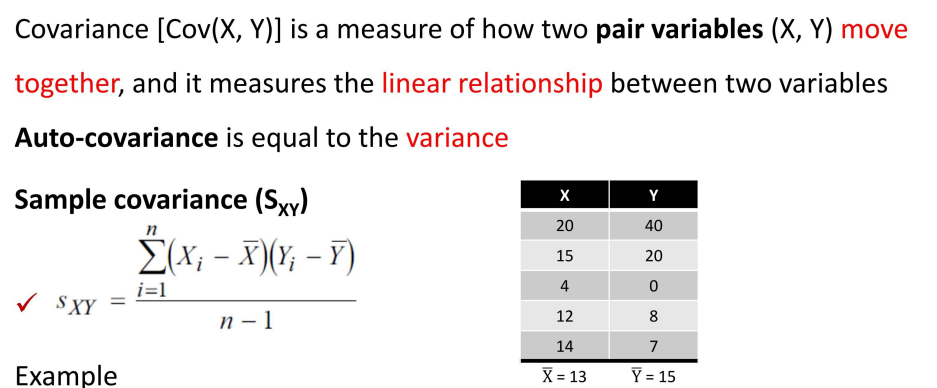


* + 展示和理解变量之间可能存在的关系
  + 关联程度强弱可以由数据分布是否聚集在直线附近看出
  + 可以用于发现变量之间的模式、计算数据范围，发现异常值

scatter plot matrix：多个变量，两两探究

### Covariance 协方差

* + 看成对的数据如何一同移动，衡量两个变量之间的线性关系
    1. **Auto-covariance自协方差** is equal to the variance
    2. **计算：**



协方差=0→ 没有**线性**关系，≠没有关系

协方差>0：两个变量tend to同时比它们各自的均值高/低

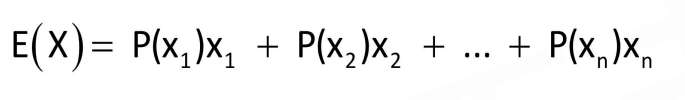
协方差的值在不同数组之间比较无意义

### Correlation：衡量两个变量的标准化线性关系

* + 
    - * 越靠近±1，线性关系越强/越明显
      * 两条都是直线的时候，都是ρ=1，与斜率无关
  + 缺点：
    - * 只反映线性关系，当非线性关系时，无法显示
      * 极端值的影响比较大
      * 相关性≠causation
      * Spurious correlation：只反映特定数据集，出现相关性可能是由于第三个变量的影响。

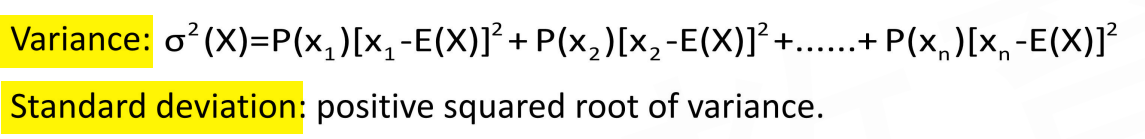
# Chapter 4 概率树和条件期望

## Expected Value E【X】



是变量特征

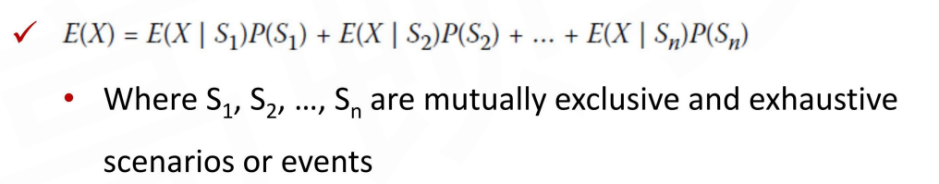
## Probability Weighted Variance and Standard Deviation

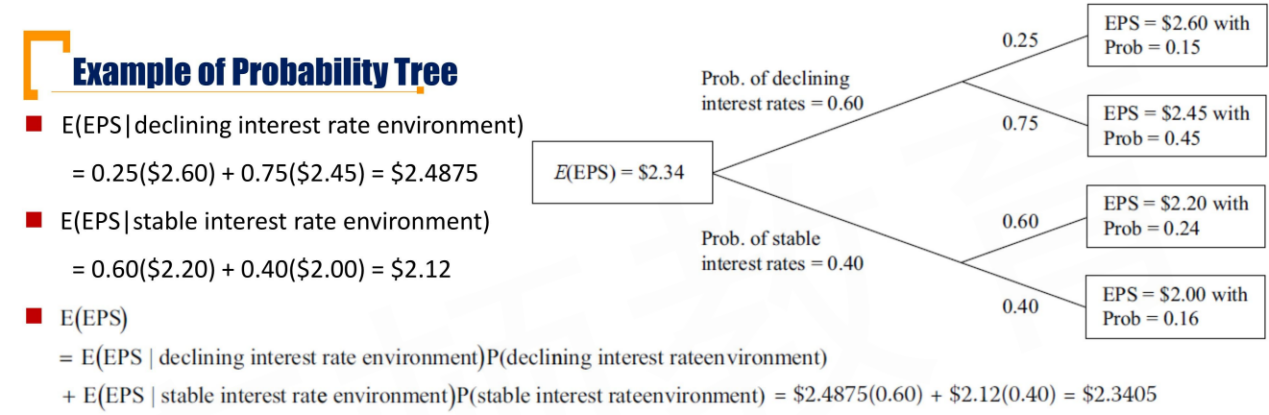


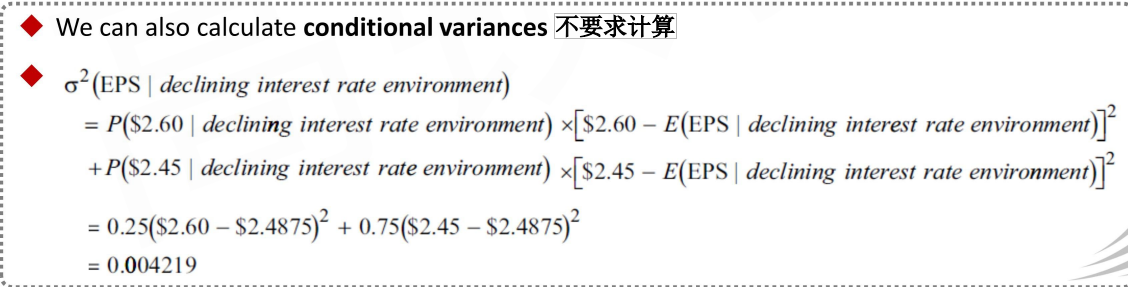
**Var（X）=E（X2）-（EX）2**

**= E（X- EX）2**

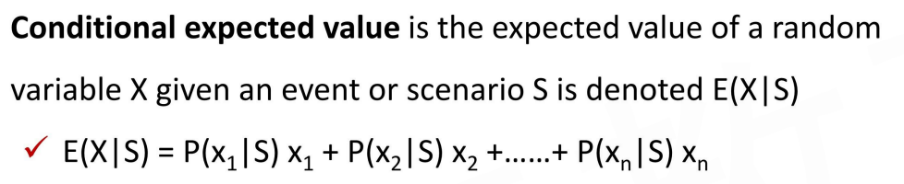
## Total probability rule for expected value/概率树





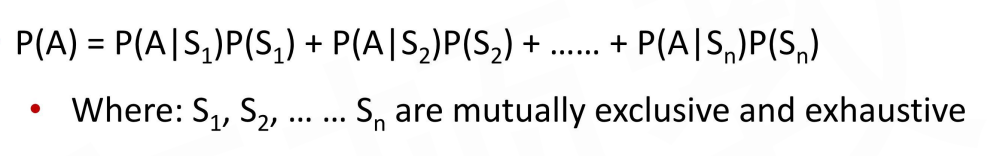


## Conditional expected value



已知S情况发生时，x的期望值

## Total Probability Rule全概率公式

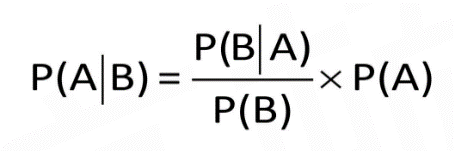


**当A的概率不能直接算，只能去分类讨论时使用**

## Bayes' Formula贝叶斯公式

### 定义

Given **a prior probabilities P（A）**for an event of interest, if you **receive new information（B）**，the rule for updating your probability （**updated probability P（A | B）**）of the event：



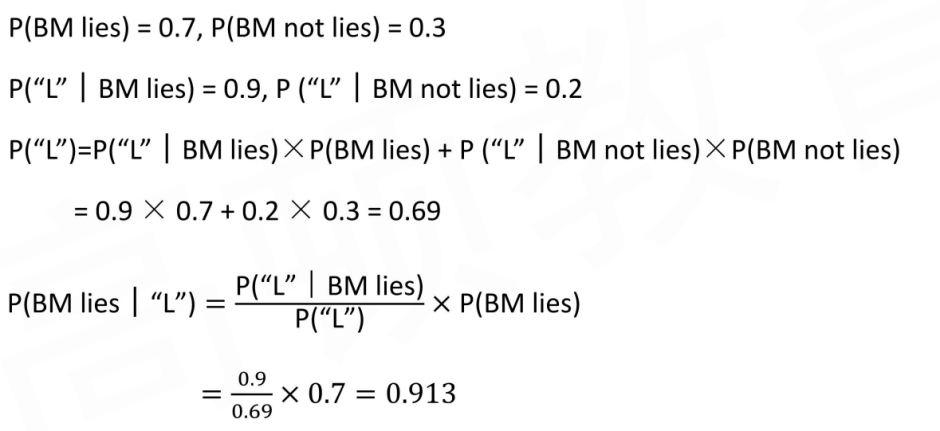
### 意义：

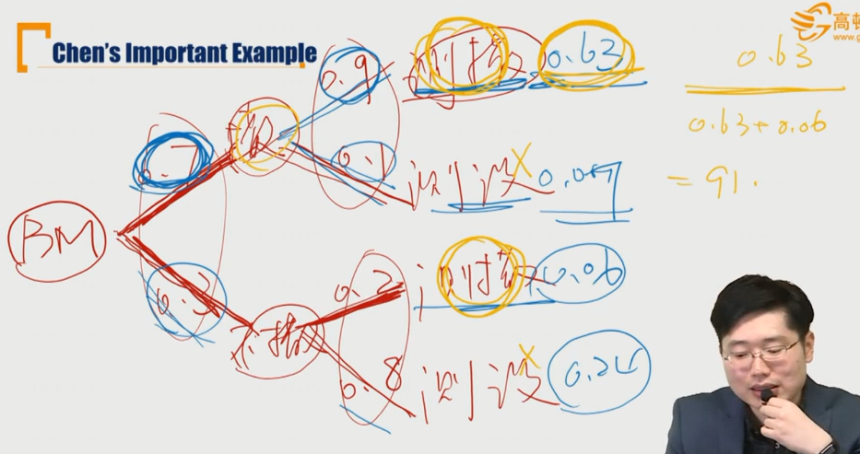
因果倒置之间的概率；得到新信息可以修正原来的概率

P（B）常常使用A是否发生使用全概率公式计算

### 例子：

BM撒谎概率是0.7，如果BM说谎测谎仪测谎出来的概率是0.9，BM没说谎但测出来说谎概率是0.2:。求测谎仪结果是说谎时，BM说谎的概率？



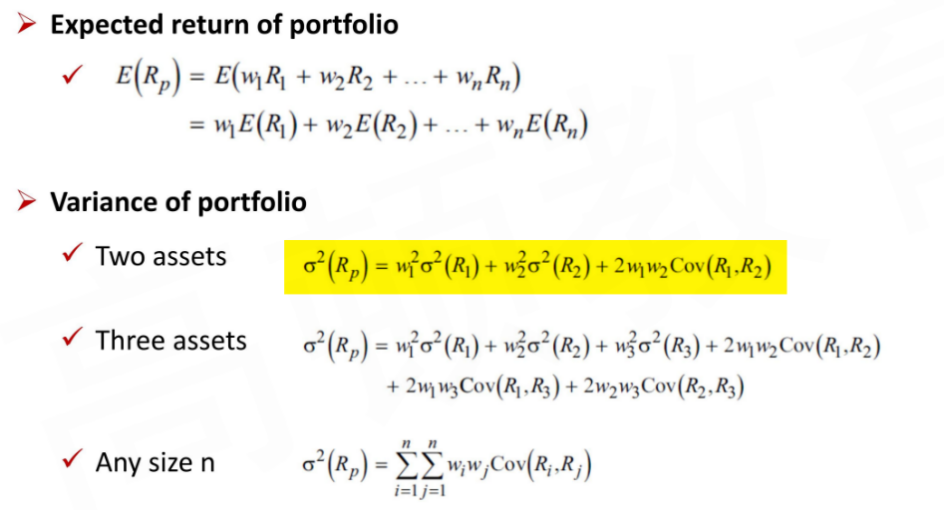
使用二叉树计算：

# Chapter 5 投资组合计算

## Covariance

**Cov（X,Y）**=E【（X-EX）（Y-EY）】**=E（XY）- EX\*EY**

## Expected Value and Variance of Portfolio



## Joint probability 联合概率

* Multiplication Rule乘法规则

P（AB）= P（A|B）\* P（B）

P（ABC）= P（A）x P（B |A）x P（C|AB）

* Addition Rule加法规则

P（A+B）=P（A）+P（B）- P（AB）

P（A+B+C）=P（A）+P（B）+P(C) - P（AB）-P（BC）-P（AC）+P（ABC）

|  |  |  |
| --- | --- | --- |
|  | 乘法 | 加法 |
| 一般公式 | P（AB）= P（A|B）\* P（B） | P（A+B）=P（A）+P（B）- P（AB） |
| 独立时 | P（AB）=P（A）\* P（B） | P（A+B）=P（A）+P（B）- P（A）\* P（B） |
| 互斥 | P（AB）=0 | P（A+B）=P（A）+P（B） |

**A/B独立↔P（AB）=P（A）x P（B）**

A、B互斥↔P（AB）=0

→P（A|B）=0

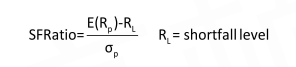
## Safety-first rules【注意名词&公式记忆】

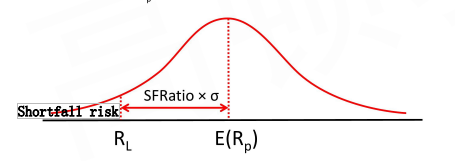
### Shortfall Risk：

* Shortfall risk is the risk that portfolio value or return will fall below **the minimum acceptable level （RL, shortfall level, threshold level）**over some time horizon
* 收益率低于底线亏损率的概率，越低越好

### Safety-First Ratio

* Safety-first ratio is the distance from the mean return to the shortfall level in units of standard deviation
* （均值-最低标准）/标准差---值越大说明安全垫越大，越好

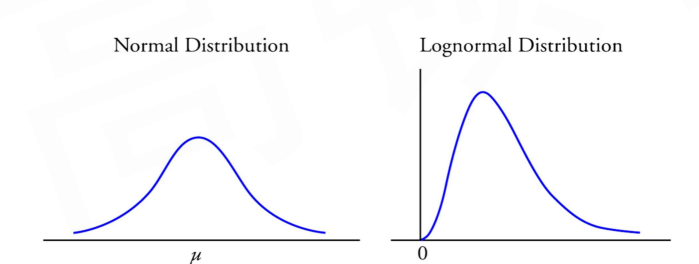


* Minimizing shortfall risk=Maximizing safety-first ratio
* 

# Chapter 6 Common Probability Distributions

## 对数正态分布Lognormal Distribution

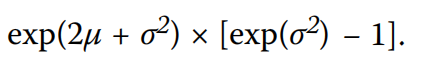
* X是正态分布的，那么Y=ex是服从对数正态分布的---Y取log之后服从正态分布
* **Y大于零，分布正偏**



### 期望与方差

如果X满足正态分布，均值为μ，标准差为σ

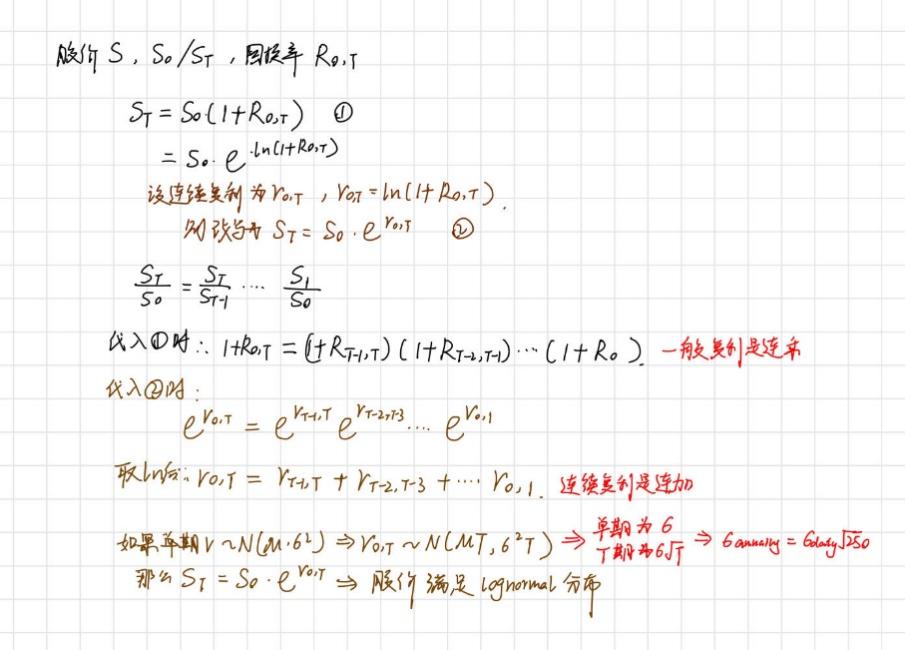
则LnY满足正态分布

* Y的均值为【**只需要记住比直接代入eμ大**】
* Y的方差为

### 股票回报分布与股票价格之间关系。

* **如果一只股票的连续复利收益是正态分布的，那么未来的股票价格必然是对数正态分布的。**
  + 连续复利收益率可以连加
  + 风险=

推导如下：**【推导大概了解即可，需要记住以上结论】**



## Monte Carlo Simulation【已知分布函数情况下】

* 定义：Use randomly generated values for risk factors, based on their assumed distributions, to produce a distribution of possible outcome. 从一个特定的概率分布或多个分布中生成大量的随机样本，以获得一系列结果的可能性
* **Inverse transformation method：**Random observations from any/distribution可以通过在excel中的随机数是【0,1】的均匀分布，经过inverse transformation method产生
* 

## Resampling【未知分布函数】

### Bootstrap

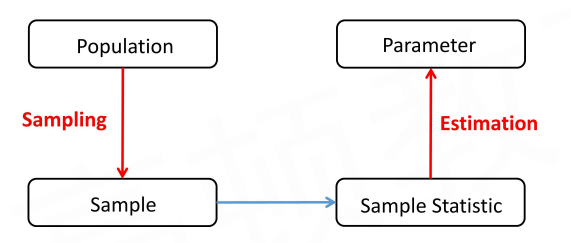
* + 是最流行的Resampling方法之一，
  + 通过替换的随机采样Random sampling with replacement，从原始训练数据集生成n个新的训练数据集（或数据包）。
  + ——换言之， 将原始训练数据当作总体进行n次抽样，从而得到n组新训练数据为抽样样本，**每个重采样的样本大小与原始样本大小相同。（抽取后会放回）**

### Jackknife 刀切法：随机切出去一小部分，计算剩余的部分

* 降低整体估计的偏差
* Jackknife每次结果比较相似。Bootstrap每次结果差异相对大

# Chapter 7 Estimation and Inference

## Sampling



一般因为：总量太大或者试验具有破坏性（火柴着火率）

### 样本容量Sample Size的大小：

更大的样本容量有better precision，但可能导致额外的费用超过带来精确度的提升

### 抽样方式：

## 概率抽样probability sampling：总体中每一个个体都有相同的机会被选中

* 概率抽样相比非概率结果更具有accuracy and reliability
* 需要注意当总体不遵从单一的分布时，抽样需要更careful

#### Simple random sample简单随机抽样---前提是抽样过程的很难无偏

#### Systematic sampling系统抽样：排序后，先选定某个数，后每k个抽一个

#### Stratified random sampling分层随机抽样：每个单元格按比例随机抽样

#### Cluster sampling整群抽样：分成不同的群，抽就整个群抽出

## 非概率Non-Probability抽样：依据除了概率以外的其他因素抽

#### Convenience sampling：使用便利的方式，但不准确

#### Judgmental sampling：依据人的主观判断

### 样本统计量/总体参数

* 总体参数population parameter是常数，样本统计量sample statistics是随机变量
* sampling error是样本统计量和总体参数的差异
* sampling distribution：是样本统计量的分布
  + 样本抽出有差异，算出的样本统计量是随机变量，会有自己的分布。
  + Eg： sample mean 会有自己的期望和方差

## Central Limit Theorem中心极限定理

### Properties of an Estimator估计量的性质

#### Unbiasedness无偏性

E（估计量）=总体参数

#### Consistency一致性

随着样本容量sample size增加，精度越高，估计量的方差越小

#### Efficiency有效性

对于无偏估计量，有着最小方差的样本分布

### Central Limit Theorem

* 内容：简单随机抽样, 总体有一个任意分布，平均值μ和方差是σ2存在，通过简单随机抽样得到的，sampling distribution of the sample mean ，在n足够大的时候， ~N（μ，σ2/n）
* that is:
  + 前提：简单随机抽样+样本量足够大（n≥30）
  + 结论: ~N（μ，σ2/n）---可以看到样本均值满足无偏性和一致性
* Standard error

Standard error of sample mean标准误 σ（）:the standard deviation of the distribution of sample means

* + 当总体的标准差σ已知：The population standard deviation is σ
    - σ（）=σ/
    - ~ N（0，1）
  + 当总体的标准差σ未知：The standard deviation for this sample is s
    - S（）=s/
    - ~ t（n-1）
      * 当n很大时，也近似于正态分布

## 估计方式

### Point Estimate点估计

* 只做一次抽样，认为得到的估计值estimate=估计量estimator

### Interval Estimate区间估计

#### 观测值有自身的分布

#### Confidence interval置信区间：**A range** for which “一个给定的概率”【即置信水平（degree of confidence，1-α）】 of 所有观测值，这些观测值将落在这个区间

* 构造出的置信区间有1-α的概率覆盖总体参数 【√】Confidence interval for population mean is a range that contain the population mean with a given confidence level（1-a）
* 总体参数有95%的概率落在该区间里 【×】总体参数是先定好的，是常数

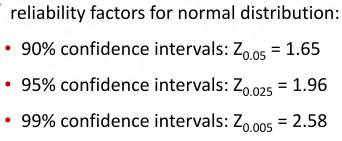
#### Significance level（α）显著区间：不落的概率

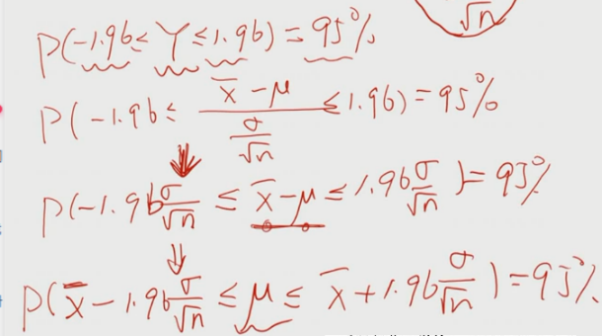
#### 对于总体均值的区间估计：【当总体方差σ已知】

Confidence Interval of population mean

=Point Estimate of population mean [μ]

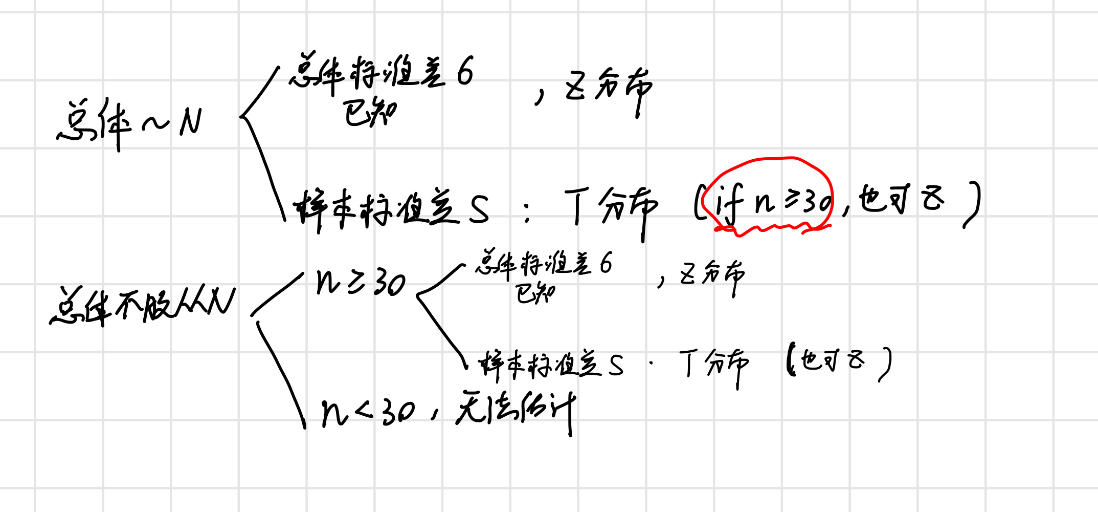
± Reliability Factor[Z0.05(90%) /Z0.025 (95%)/ Z0.005(99%)] x Standard Error of sample mean [σ/]

看最右边的小尾巴面积查表



P（σ/ ≤σ/）=1-α

#### 选择估计方式

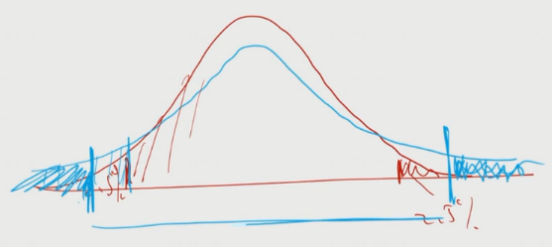


对于N分布：P（σ/ ≤σ/）=1-α

对于t分布：P（s/ ≤s/ ）=1-α 其中t的自由度是n-1

#### 影响置信区间：

* 根据s/ --置信度越大，样本数越小，样本方差越大，置信区间越大
* 使用t分布，置信区间越大----所以纳入总体参数的概率更大，更conservative



# Chapter 8 Hypothesis Testing

## 概念

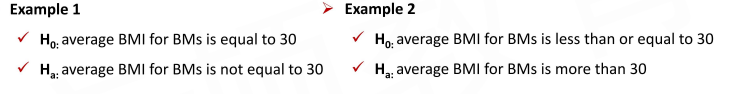
* is an act in statistics whereby an analyst tests an assumption regarding a population parameter
* 理论基础：小概率事件不会发生

|  |  |
| --- | --- |
| 估计： | 回答参数的值  构造区间，有95%的把握覆盖总体参数 |
| 假设检验： | 回答总体参数是否等于某个确定值  95%把握，判断原假设：进行抽样得到值如果落到小概率，拒绝原假设  ---区间估计和双边检验是一回事 |

## 步骤

1. 写明Hypothesis
2. 构造test statistics 和它的probability distribution
3. 确定significance level
4. 说明decision rule——只需要significance level
5. 收集数据并计算value of test statistic——这一步才需要critical value
6. 做出statistical decision，从而得到经济学或投资决策

## Null hypothesis原假设H0 & Alternative Hypothesis备则假设H1



* 等号只会出现在原假设中
* 能做出的决策只有：是否拒绝原假设H0，而不能说是否接受H1----证伪容易，证实难
* 想要拒绝的结论，放在原假设中

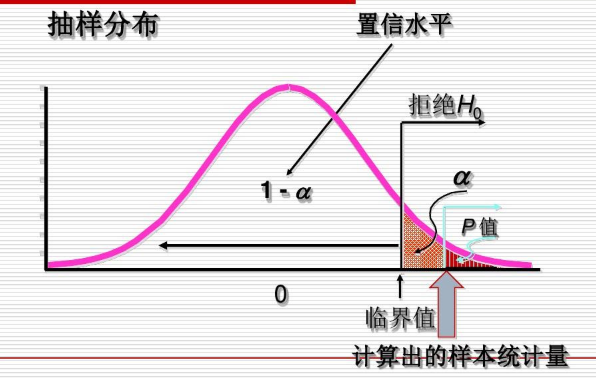
## Two-Tailed Test vs. One-Tailed Test

只要原假设是“=”是双边检验

|  |  |
| --- | --- |
| 双边  Two-Tailed Test | C:\Users\YYS\AppData\Local\Temp\tianruoocr\截图_20221201121129.png |
| 单边  One-Tailed Test | C:\Users\YYS\AppData\Local\Temp\tianruoocr\截图_20221201121203.png |

## Decision rule及名词：

* **Test statistic's value**检验统计量的值—也是借助抽样得到的
* **Significance level**（α）显著性水平：拒绝原假设的概率
* **Critical value（Rejection point）**临界值：划分接受域/拒绝域的横轴值
* **p-value**：计算出Test statistic's value围出来的面积值，直接跟α比较——如果是双边的话，注意一下p是否要乘2再跟α比较



### 由此判断过程：

* + 如果Test statistic's value超出Critical value范围，拒绝原假设
  + 如果p-value ≤ Significance level（α），拒绝原假设

### 拓展：多次抽样修正假设检验

k次抽样，计算出k个p-value，从小到大记为排序i

* 单次的拒绝只需要满足p-value ≤ Significance level（α）
* 对于多次抽样需要满足单个p-value≤ Significance level（α）\* 排序i/ k -----让拒绝变的更严格

### 拓展：economic decision与statistical decision

当样本量足够大的时候，可能可以得到令人满意的拒绝原假设的统计学结论。但对于做经济投资决策，可能还有其他的因素没有考虑进模型，eg：交易成本，税、风险等——也就是做完satistical之后还要做economic

## 两种错误

### Type I error【拒真】

* rejecting null hypothesis when it is true
* P（Type I Error）=Significance level α
* 也叫作False Positive Result---因为拒绝有两种写法：the null is rejected （the positive）

### Type II error【不拒绝伪】

* failing to reject the null hypothesis when it is false
* P（Type ll Error）= β
* Power of test=1- P（Type ll Error）=1-β
* 给定实验设计，不可能同时降低两种错误的概率——负相关
* 但提高样本容量，可以同时降低两种错误的概率
* 应该优先避免Type I error---因为我们的实验都在拒真



## 对于总体均值做假设检验：

### 双边检验：

* 方法一：假设检验中的假设值μ，样本平均值直接代入计算（-）/σ/ 的值，看是否大于，大于落在拒绝域。

或者当样本数量很大且σ未知时，（-）/ s/的值，看是否大于；

* + 方法二（基本不会用）：将值代入P（σ/ ≤σ/）=1-α得到μ应该选取的区间，再看假设检验中的假设值μ是否落在这个区间里

### 单边检验：

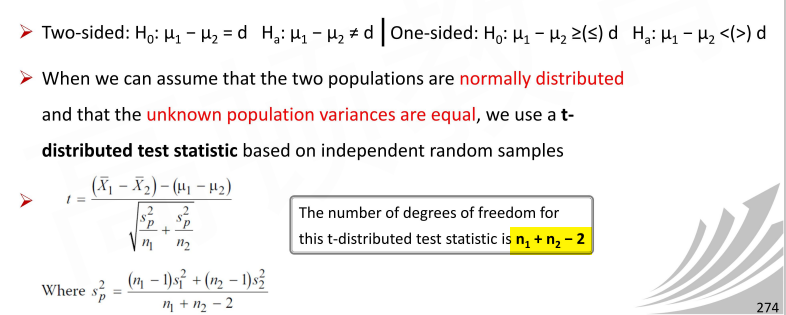
* 希望证明“中国人平均身高大于1.6米”，原假设写为H0：中国人平均身高≤1.6

## 样本均值差异：t分布

### 独立样本：

前提：两个样本集是相互独立的+各自正态分布+**方差相等**但未知——pooled estimator

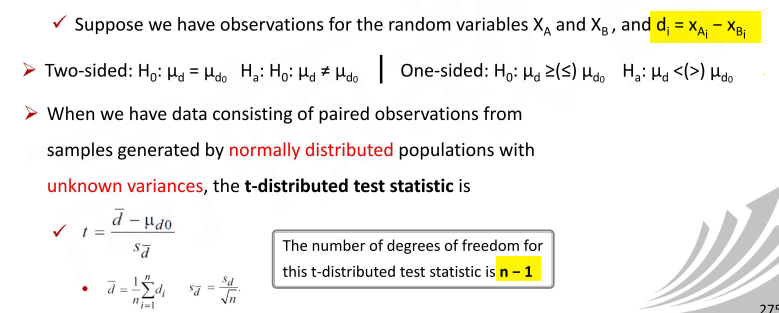
结论：使用t分布



### 相关联样本dependent：paired comparisons test

前提：两个样本集是相互关联的+各自正态分布+方差相等但未知

结论：使用t分布



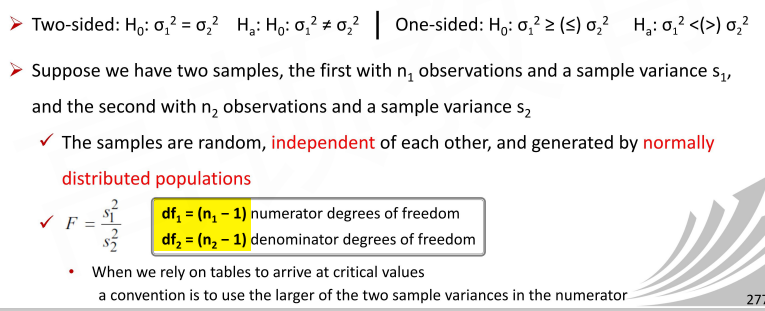
## 检验总体方差：卡方分布

* 前提：要求单个满足标准正态分布



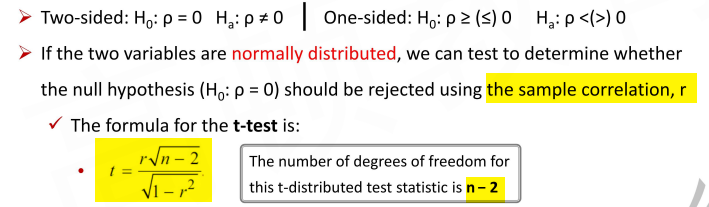
## 检验总体方差的差异：F分布

* 前提：两个个体独立+正态分布
* 一般用大的方差去除以小的，构建F参数；
* 自由度是一组数



# Chapter 9非参数检验【大致了解即可】

## 【重要一些】Test correlation检验线性相关关系

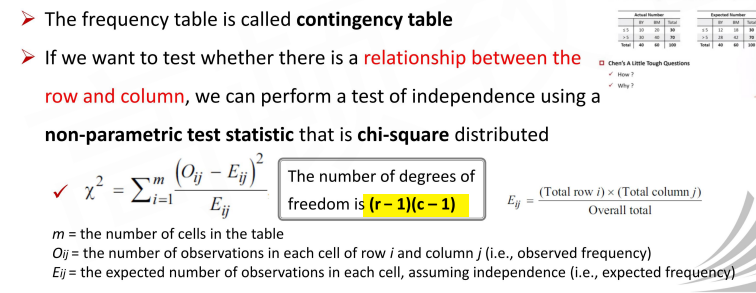


对于其中的r：

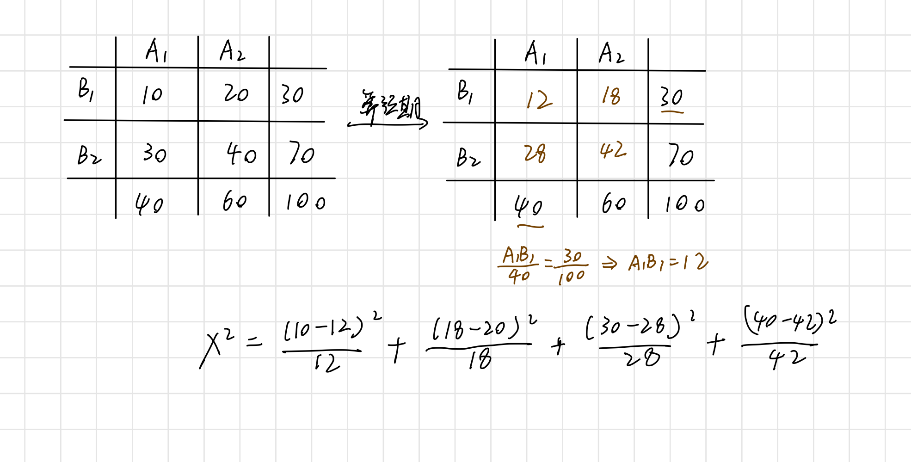
|  |  |
| --- | --- |
| Pearson Correlation  参数检验统计 | C:\Users\YYS\AppData\Local\Temp\tianruoocr\截图_20221201161447.png |
| Spearman Rank Correlation coefficient  **是非参数检验统计Non-Parametric**  数据代表的是排序，但没有绝对值意义 | C:\Users\YYS\AppData\Local\Temp\tianruoocr\截图_20221201161513.png其中d是各组排序的差  总体明显不服从正态分布时，且要求样本量大，此时可以代入使用以上的（n-2）的t检验 |

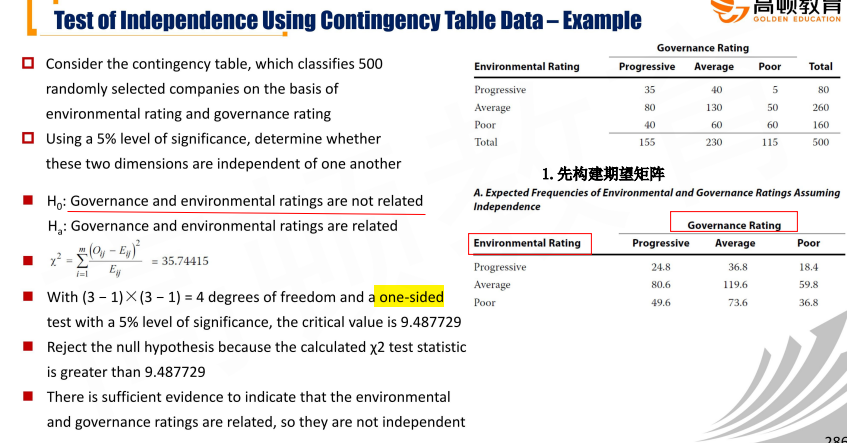
## 对于Contingency Table Data的独立性进行检验—使用卡方

* 回答横纵两个维度是否有关系。假设H0无关ρ=0
* 先算每个格子的预期数。
* 自由度是（横-1）\*（纵-1）
* 注意是**单边检验**



计算预期矩阵：

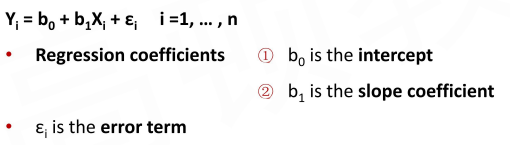


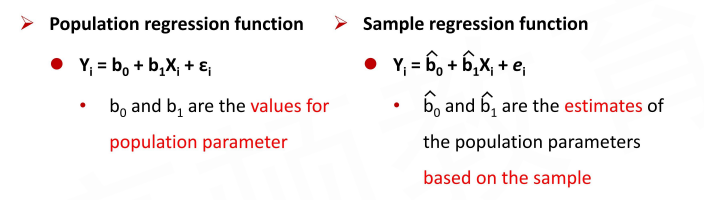


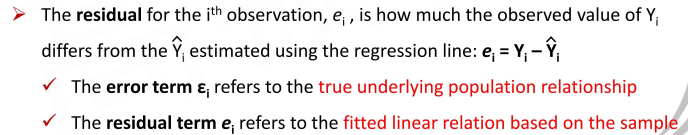
# Chapter 10 Linear Regression

## Linear Regression Analysis的基础

* Dependent/explained variable：Y
* Independent/explanatory/indicator variables：X（s）
* 对于只有一个Y的，是simple linear regression
* 应用于两种回归数据：cross sectional data；time series data
* 描述：Y is regressed on X





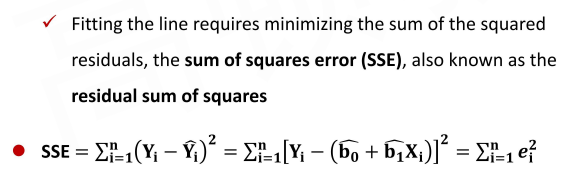


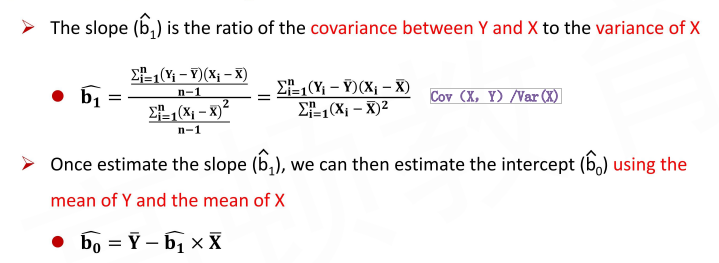
C:\Users\YYS\AppData\Local\Temp\tianruoocr\截图_20221201230906.png

## Simple Linear Regression

### 最小二乘法：Ordinary Least Suuarest OlS Regression

The goal is to fit a line to the observations on Y and X to minimize the squared deviations from the line。**使得残差平方和SSE最小——没有被解释的部分**





一定会经过（）

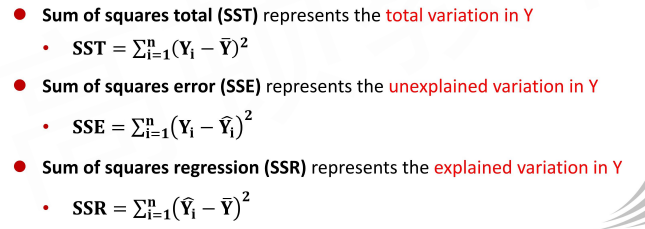
### 对于简单线性回归的前提假设：

* Linearity：
  + X和Y之间的关系应当是线性的+residuals是random+自变量X不是随机变量
* Homoscedasticity同方差：
  + 对于所有的观测值residuals的方差相同
* Independence：
  + 每组点之间是独立的→the regression residuals are uncorrelated across observations---cov（e1，ei）=0
* Normality正态性：
  + 残差项满足正态分布——对于大样本可以不需要这个假设

## 方差分析Analysis of Variance

### SST = SSE+ SSR

**SST总体变动=SSE不能被解释的变动（注意E是error）+SSR被线性解释的变动**



### **除**以自由度——可以直接看后一页的 [ANOVA] Table

* **Variance=SST/ degrees of freedom=SST/(n-1)**
* **Mean square regression（MSR）均方回归=SSR /degrees of freedom （df= number of independent variables（k））**

#### 对于简单线性回归，k=1，MSR=SSR

* **Mean square error（MSE）均方误差=SSE/degrees of freedom（df=n-1-k）**
* 对于简单线性回归，k=1

C:\Users\YYS\AppData\Local\Temp\tianruoocr\截图_20221201235317.png

## 衡量拟合效果Measures of Goodness of Fit

### Coefficient of determination（R2）

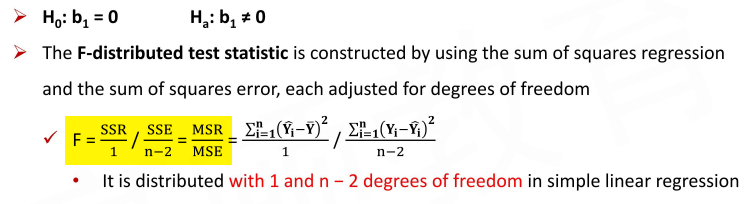
* 由SST = SSE+ SSR，SST/STT=SSE/STT+ SSR/STT，1= SSE/STT+ R2
* 衡量可以由自变量解释的比重

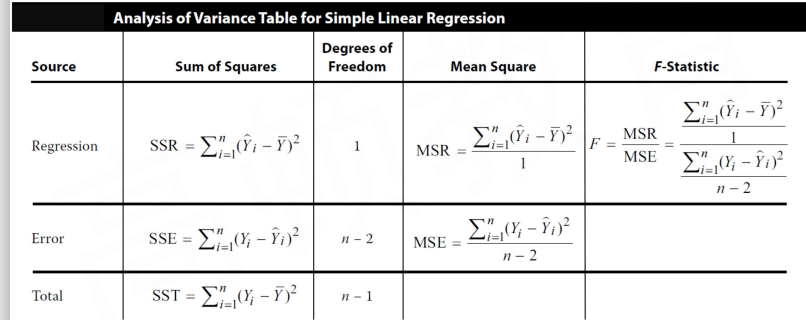


* 在简单线性回归中，the square of the **pairwise correlation** r2= the coefficient of determination R2——求两个参数的correlation可以直接用回归表中的R方

### F-statistic for the test of fit

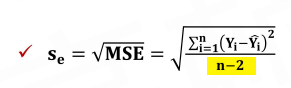
* 检验是否所有回归系数都等于0



* 由于拒绝域对应的临界值一般都大于1，所以需要MSR>MSE，即：被线性解释的部分远大于未被解释的
* Analysis of Variance [ANOVA] Table
* 

### Standard error of the regression（Se=）

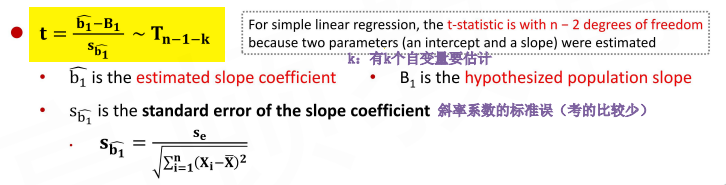
* Se： Standard error of the Estimate，也称为Standard error of the regression，或the root mean square error



* Se是绝对值，Se越小越好

## Inference by Simple Linear Regression

### Hypothesis Tests of the Coefficient (b1):使用t检验

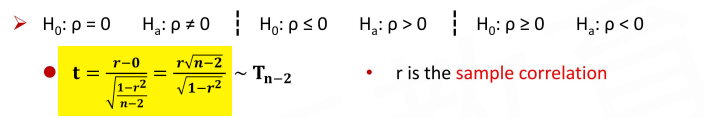


如果自变量X的离散程度变大，Sb1越小，t值越大，那么越可能落在拒绝域中——这时候如果不拒绝，更有说服力

前面的F检验——F=MSR/MSE，F是对整体的回归是否有解释力度；

T检验是对其中某一个X的系数，是否解释力度

### Hypothesis Tests of the Correlation (ρ):使用t检验【经常考】



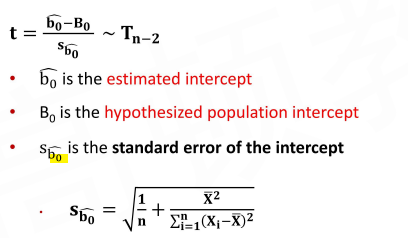
### 关系

* 对于单系数简单线性回归，a）和b）的两个T值可以在数学上证明，严格相等；但两个假设应当是一致的。



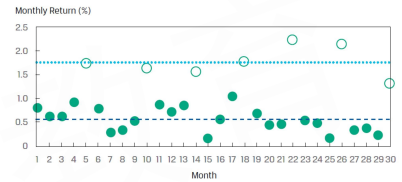
## 单一简单线性回归，F检验——F=MSR/MSE，的值与t2一致—F= t2

### 检验截距intercept （b0）：t检验——如果给了coefficient 和standard error可以算t值



## Indicator/Dummy Variable 指示变量/哑变量

* 只有0/1两个值
* 此时Y（x=1）=b0+b1
* Y（X=0）=b0
* 所以b1= Y（x=1）- Y（X=0）【b1代表了按照1/0分成两组之后的两组差异】



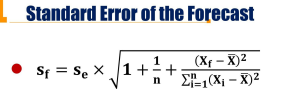
## 预测Prediction Interval of Dependent Variable

将未来的Xf代入之前拟合出来的直线，

C:\Users\YYS\AppData\Local\Temp\tianruoocr\截图_20221202151740.png

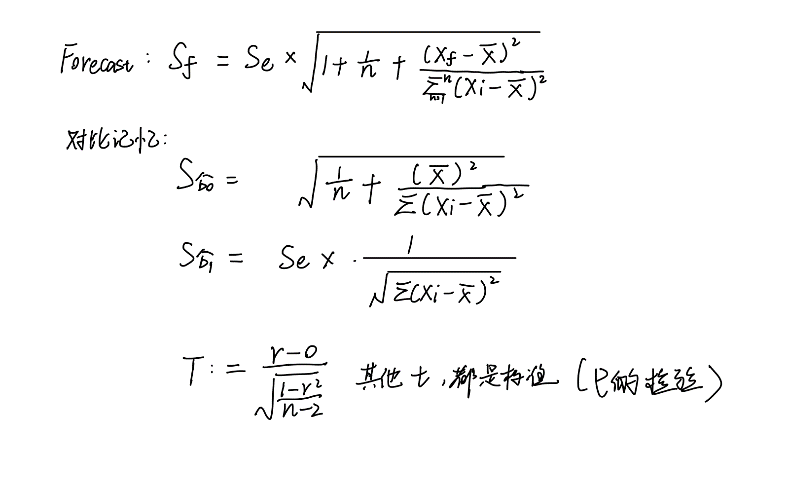
|  |  |
| --- | --- |
| 但residuals 并不均为零，所以预测的对应Y is an average of the relation between the two variables- |  |
| C:\Users\YYS\AppData\Local\Temp\tianruoocr\截图_20221202155951.png  --所以Yf的值会反应这种uncertainty | |

### Standard Error of the Forecast（Sf）



* **Sf降低意味着预测精度更高：**

1. Se降低：回归模型拟合程度越好
2. n增大：样本数量增加
3. Xf - 更小：预测取值的Xf距离近一些
4. ∑（Xi -）2更大：样本X的取值更分散



## 对于Non-linear relationship改变其函数形式来做线性回归【关注解读系数含义】

### Log变换----变成指数形式，谁在指数上，谁百分号。

|  |  |
| --- | --- |
| **The Log-Lin Model【LnY=F（X）】** | ：the absolute change in the dependent variable for a relative change in the independent variable  **即：X变动一单位，Y变动%** |
| **The Lin-Log Model【Y=F（LnX）】** | ：the absolute change in the dependent variable for a relative change in the independent variable  **即：X变动1%，Y变动** |
| **The Log-Log Model【LnY=F（LnX）】** | ：**弹性**the relative change in the dependent variable for a relative change in the independent variable  **即：X变动1%，Y变动** |

### 如何选择合适的函数变形

|  |  |
| --- | --- |
| [拟合效果三方法](#_衡量拟合效果Measures_of_Goodness)  1.Coefficient of determination（R2）  2.F-statistic for the test of fit  3.Standard error of the regression（Se=） | C:\Users\YYS\Documents\Tencent Files\935889113\FileRecv\MobileFile\238DAEF8B27579EAB56CB0ADA340002A.png |
| 4.与此同时还可以看patterns in the residuals【适用最广】 | —看残差能否更拟合白噪声的特点（random residuals） |

# Chapter 11 Big Data

## 大数据介绍

* 主要关注对于数据的收集、处理、清洗；
* 但选择模型、数据分析和应用主要靠人工

### 大数据数据来源：

* traditional sources ：金融市场、商业、政府
* non-traditional sources：个人文字语言类数据、sensors传感器数据、IoT物联网数据

### 结构化数据&非结构化数据

* Structured data：依据format处理过的数据
* Unstructured data：不是传统data model，例如：文字、图片、声音、视频

### 大数据特点3Vs+4th V

1. Volume多：高达TB、PB的数据量
2. Velocity快：数据产生、传播速度很快，几乎实时
3. Variety杂：数据来源、类型丰富
4. **Veracity可信度**：当在使用大数据进行推断和预测需要判断，第四个“V”特点

* 有20%左右都是无用/假数据

## 机器学习

* 实际上就是找数据之间的关系——建模：【1.收集并分析数据；2.选择并构建模型；3.**分析结果——模型是否符合假设、解释力；4.预测**】

### 定义：

* 机器学习包括多种方法，通过这些方法，计算机are programmed to improve performance in 特定任务 with experience
* Find the pattern，apply the pattern

### 优点：

* 对数据的分布没有严格的要求handle problems with a high degree of non-linearity.
  + 传统统计模型statistical approaches对数据分布有着严格的要求——一般要求线性
* 善于处理大数据handle problems with many variables（high dimensionality）

### 应用：

* 机器学习对投资管理具有理论和实践意义，从客户分析到资产配置，再到股票选择，再到投资组合构建和风险管理，再到交易。

### 类别

#### **Supervised learning（监督学习）**

* 数据是**事先打好标签**的requires a labeled data set，已经规定好inputs（Xs）和desired output（Y）使用这个数据集来推断输入和输出之间的模式被称为“training”算法。

#### **Unsupervised learning（非监督学习）**

* 数据是没有标签，有inputs（Xs），**但没有desired output（Y）**
* 用于去发现数据自身的结构

#### **Deep learning（深度学习）/reinforcement learning(强化学习）**

* 基于neural networks（神经网络），两者可能是supervised或者unsupervised
* NNs和DL非常适合处理高度复杂的机器学习任务，而不是简单的单变量OLS回归模型

##### Deep learning nets（DLNs）

* 深度学习**可以处理高度复杂的任务**，如图像分类、人脸识别、语音识别和自然语言处理。
* Deep learning nets（DLNs）：具有许多隐藏层hidden layers的神经网络

##### Reinforcement learning 强化学习

* 计算机通过与自身（或由相同算法生成的数据）的交互进行学习
* 与监督学习不同，强化学习既没有直接标记的数据，也**没有即时反馈 instantaneous feedback（这里即时反馈是指最终的结果值）**
* 该算法需要观察其环境，learn by testing new actions，并reuse其以前的经验。
  + 例如AlphaGo问题。

## 数据处理步骤

|  |  |  |
| --- | --- | --- |
| 步骤 | 传统数据 | 文本数据 |
| 1目标 | Conceptualization of the modeling task（明确建模目标）：输出什么、模型给谁用、用在什么阶段 | Text problem formulation（构建文本问题）：找到文本问题；定义输出/输入；决定输出如何运用 |
| 2收集 | Data collection（数据收集）  传统的数据通常已经是结构化的表格格式，包含特征的列和例子的行。 | Data（text）curation（数据护理）：  收集文本数据、用labels标注文本数据 |
| 3整理 | Data preparation and wrangling（数据准备与整理）  1)Data cleansing（数据清洗）：处理缺省值、超范围值等  2)Data preprocessing（数据预处理）：提取、聚合、筛选和选择相关的数据列。 | Text preparation and wrangling（文本准备与整理）：将非结构化转化为结构化 |
| 4探索 | Data exploration（数据探索）   1. Exploratory data analysis（探索性数据分析）——描述特征； 2. Feature selection（特征选择）——为了达成目标，选择一部分； 3. Feature engineering（特征工程）——计算一些比值等 | Text exploration〈文本探索）   1. Text visualization via word clouds（词库云） 2. text feature selection（文本特征选取），etc. |
| 5训练 | Model training（模型训练）：选择合适的ML方式；评价模型表现；调试模型 | Model training（模型训练） |