# 财富证券因子分析框架

# 一、多因子研究介绍

人们发现在证券市场上，特定的时间段内，某些指标与证券的收益存在一定的相关性，比如：2017年，大市值的个股相对比小市值个股，平均而言有一个更好的收益。那么？

在特定时间段内，如何寻找与证券收益高相关性的指标？  
如何判断指标与证券收益相关性的稳定性？

如果多因子研究最终能解决上述两个问题，并作出分析结果，形成因子库（指标库），后续以此选择股票的把握就会很大。这个因子分析的过程就是多因子研究。

# 二、因子处理流程

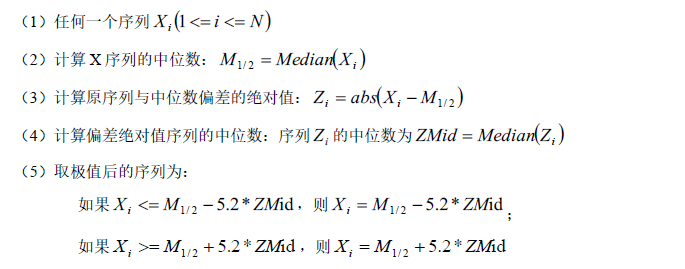
因子分析的过程是一个由样本分析系统的过程，样本有它的局限性：

1、样本数有限，通过有限的样本去分析系统的性质肯定会存在一些差异。  
2、样本中可能存在小概率事件，小概率事件会影响到对整体的分析。比如:分析小明的学习情况，他考了10次试，其中9次都是六十多分，1次是九十多分，那么这唯一一次的九十多分可能就是一个低概率事件（可能小明考100次才能考中一次），这个样本就会对我们的分析产生影响。

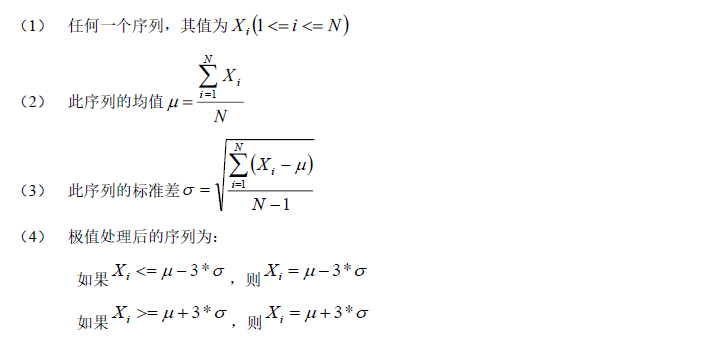
异常数据的判断和剔除是数据处理中必不可少的一环。尤其是金融数据处理中，不进行极值处理，会对数据最终的分析造成很大的误差和误判

# 三、因子极值处理方法

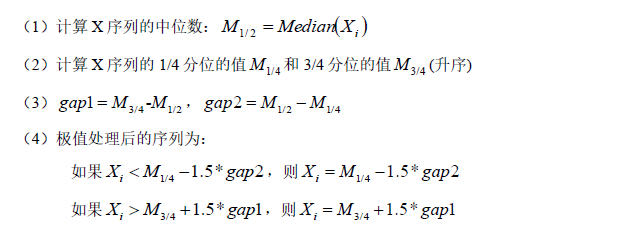
**A.中位数去极值法**



**B.3倍标准差法**



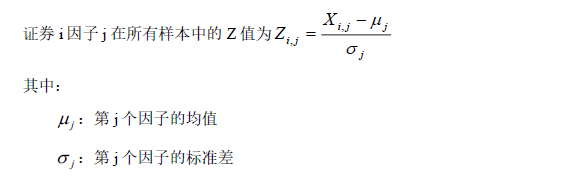
**C.四分位法**



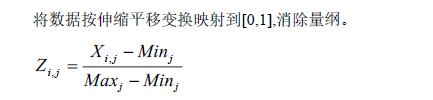
多因子研究过程，一般都会采取打分法确定综合因子。但每个单因子所代表的含义和数量级不尽相同。要对多组不同量纲数据进行比较，需要先将它们标准化转化成无量纲的标准化数据。比如：市盈率指标是整数，而净利润增长率是一个百分比，两个指标简单的相加是无法真正体现该综合因子的实际意义的。

## 四、因子无量纲处理方法

**A.标准化法（Z 值法，默认方法，最常用）**



**B.极差正规化**



C.百分比打分法

采用百分比排名的方式打分。一般而言，百分比打分后的分数分布在0-100 之间   
**说明**：   
采用此种打分方式的目的是为了消除标准差非常大，而差距没有那么大的序列的影响。   
对于同名的采用向上打分，和excel 中的percentrank 向下打分略有不同。这种差异仅仅会在处理同名时体现。

# 五、因子分组

多因子研究一般要确认如下四点：

因子是否有效。即：按照因子，对样本分组，不同组别之间的样本，收益率是否有显著性差异。

* 因子的相关性如何？
* 因子是否有稳定的超额收益？
* 因子是否对未来有解释作用？解释能力如何？

　　因此需要按照因子的值，对样本进行分组。

* **不分组，具体为；因子值排序后，从前往后按次挑选，直到满足持仓上限数量。**

**三分位法分组，具体为；因子值处于前1/3为第一组，因子值处于中间1/3为第二组，因子值处于后1/3为第三组。**

## 1、分组时考虑行业因素

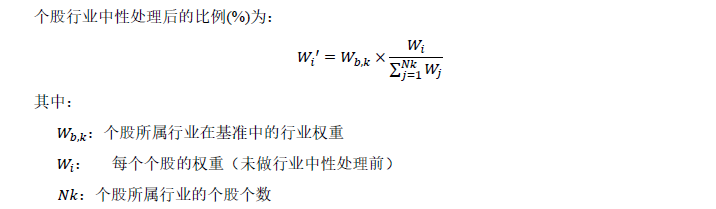
某些因子值和行业相关。如：金融行业的市值一般比较大，而电子行业的市值一般比较小。如果在分组的时候，不考虑行业因素，可能会集中把一些行业分到一个组中，从而使得组间的行业分布不均匀。

因子分组的目的是为了检验分组后各个组别之间的收益率是否有显著差异。如果组间收益率有显著差异，说明此因子可以把表现好的股票和表现差的股票分离出来

如果不考虑行业因素，各个组别的收益率差异可能不是由因子贡献的，可能是受到行业影响，由行业分布贡献。这不是符合多因子研究的初衷，因此在分组的时候，建议考虑行业因素的影响

## 2、行业资金分配方法

* 如果不做行业中性处理：只考虑个股的资金分配，不考虑不同行业在基准中的权重差异。
* 经过行业中性处理后，每个行业的配置比例和基准行业配置完全相同，个股行业中性处理后的比例(%)为：



# 六、因子评价体系

多因子研究从以下几个方面来评价因子是否有效，提供完整的分析报告：

因子收益率检验：检验每组是否可以取得正的收益率？收益率波动率如何？取得正的收益率概率有多大？

因子显著性检验：检验每组是否可以取得正的超额收益率？取得正的超额收益率概率多大？

因子区分度检验：第一组收益率是否显著大于最后一组？因子是否能显著把表现好的股票和表现差的股票分别出来？

因子延续性检验：由因子得到的分组是否能够在较长的一段时间保持比较好的方向性。即：本期表现的因子，在下期是否也取得好的收益率？

因子贡献度检验：因子的反映的信息是否有重叠？各个因子的打分对下期收益的贡献到底有多大？

## （1）因子组收益率检验

分组之后，各组在各个时间区间的表现如何？组与组之间在同一个时间区间的区分度如何？这些问题就是这个检验方法所探讨的问题。

### 1、检验目的

各组收益率均值、标准差、收益率大于0的概率

### 2、检验方法

统计了近一年、近2年、近3年和整体的数据。

检验对象：各组收益率均值、标准差、胜率

### 3、评价指标

|  |  |  |
| --- | --- | --- |
| 评价指标 | 说明 | 备注 |
| 平均收益(%) | 各组的平均收益 | 越大越好，说明第一组收益率显著大于最后一组 |
| 标准差(%) | 各组的标准差(%) |  |
| 夏普比率 | 各组的夏普比率 |  |
| 胜率(%) | 各组的胜率(%) | 各组取得正的收益率概 |

## （2） 因子显著性检验

分组之后，我们总是希望第一组显著的超越基准（说明分组可以选出好的股票），最后一组被基准超越（说明分组可以剔除差的股票）。

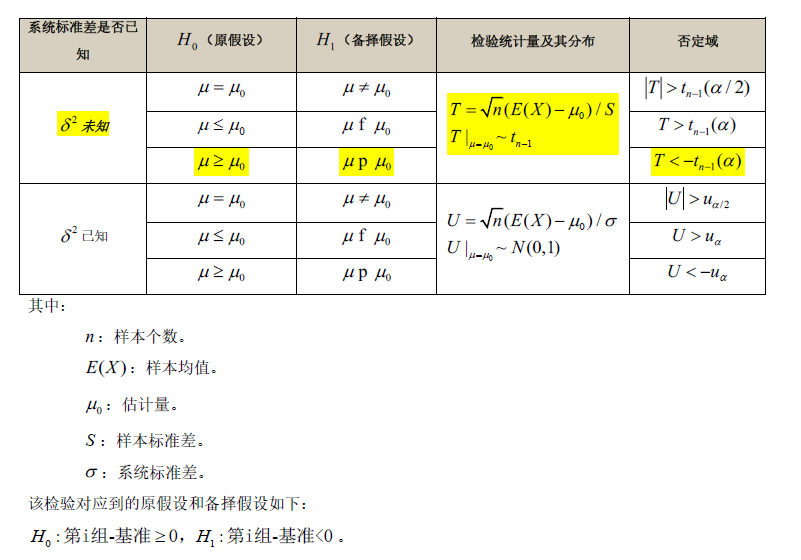
### 1、检验目的

研究各组超越基准的情况。我们希望第一组明显超越基准，最后一组明显被基准超越。

### 2、检验方法—t 检验法

统计样本：各分组的收益率序列-基准的收益率序列

原理：统计学单个正态总体均值的假设检验



### 3、评价指标

| **评价指标** | **说明** | **备注** |
| --- | --- | --- |
| 超额均值(%) | 各组收益率相对基准的超额收益率(%)的均值 | 越大说明该组超越基准越明显 |
| 跟踪误差(%) | 超额收益率的标准差，衡量各组偏离基准的程度 | 越小越好 |
| 信息比率 | 风险收益率指标，超额收益率均值/超额收益率标准差 | 越大说明超越基准越明显 |
| T-Stat | T 统计量 | 该值越大说明超越基准越多 |
| P-Value超基准收益率(%) | 显著性概率(%),T-Stat 对应到的T 累积分布函数的值中大于0的概率(%) | 越大说明该组超越基准越多超额概率(%) |

## （3） 因子区分度检验

分组之后，总是希望分组能把收益率高的分到前面的组，收益率低的分到后面的组。因子区分度检验评价方法就是为了检验这个问题而设计的。

### 1、检验目的

* 检验第一组收益率是否显著超越最后一组收益率

### 2、检验方法

时间维度：时间维度—近1年、近2年、近3年、整体。

检验对象：第一组相对最后一组的超额收益率(%)。一般用月度收益率（Monthly Extra Return）来度量

### 3、评价指标

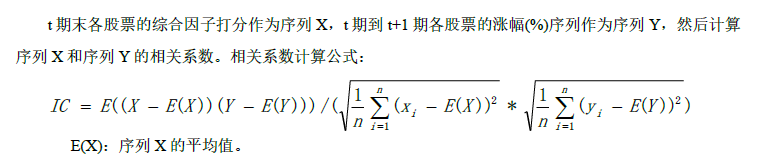
|  |  |  |
| --- | --- | --- |
| 评价指标 | 说明 | 备注 |
| 平均收益(%) | 第一组相对最后一组超额收益率的平均收益(%） | 越大越好，说明第一组收益率显著大于最后一组 |
| 标准差(%) | 第一组相对最后一组超额收益率的标准差。第一组相对最后一组超额收益率的波动 | 越小越好 |
| 胜率(%) | 第一组战胜最后一组的概率(%) | 越大越好，表示第一组收益率显著大于最后一组 |
| 夏普比率 | 超额收益的平均值/超额收益的标准差 | 越大越好 |
| T-Stat | 检验对象对应的T统计量 | 越大越好 |
| P-Value | T-Stat对应的T累计分布函数的值，是第一组超越最后一组的显著性概率(%) | 越大越好 |

## （4）因子延续性检验

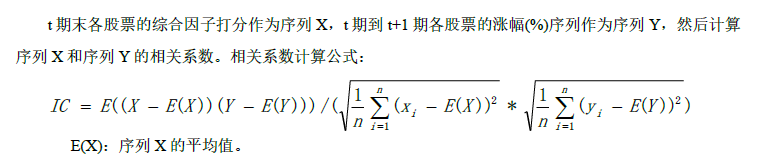
事实上，很难找到某个因子，一直保持某个的方向，不同时期因子的方向可能会改变。举一个类似例子，‘增发股票’这个事件可能在牛市的时候是一个利好消息，在熊市的时候确是一个利空消息，因为按常理，人的心理变化了，对同一事件的反应不同，最后反应到对股票的购买欲上面。

那么，既然不能找出方向一直保持不变的因子，我们还是希望能找出延续性很好的因子（即能在比较长时间保持方向不变的因子）。这个评价方法就是来解决这个问题的。

### 1、 IC计算流程

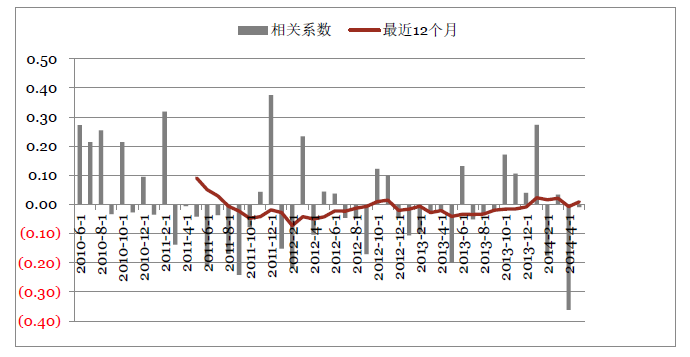


### 2、IC计算方法



### 3、IC（相关系数）的作用

通过各期的IC 值，判断出因子的延续性；比如：下图中，柱状的系列1代表对应日期的相关系数，红线系列2 是1年的IC均线，如果红线始终在x 轴上方（或者下方），那么说明延续性很好。如果红线总是在x轴上下波动，那么说明因子的延续性很差。比如下面的图显示在2011年8月份开始，到2012年8月份，所示因子都是表现出比较好的反向延续性。



# （5）因子贡献度检验

### 1、 检验目的

多因子框架可以对多个因子进行有效性的测试，那么就有一个问题，各个因子所起的作用究竟如何？各个因子分别对下期收益的影响到底怎么样？如何在之后调整各个因子的比例(%)以达到更好的效果？

那么贡献度检验方法就是检验各个因子的得分对下期收益的贡献度，以此作为之后调整因子比例(%)的依据。

### 2、 检验方法

用各个因子标准化后得分和下期收益做最小二乘参数估计（带常数项），即自变量通过某种线性变换去拟合因变量，寻找各个自变量与因变量的关系。

两种展现形式:

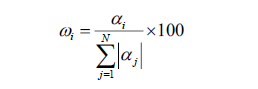
每期的数据分别去做小二乘参数估计  
用所有期的数据去做最小二乘，注意：如果每期变更因子，则不能使用这种方法。

### 3、 评价指标

检验结果得到各个因子对应的回归系数、常数项，以及基于此计算的推荐的配置比例(%)。详细含义解析如下：

回归系数的正负的含义：如果回归系数是负的，那么说明该系数对应的因子对下期收益的贡献是负的，应该在之后配置因子的时候把该因子的方向设置为跟原来相反。如果是正的说明该系数对应的因子对下期收益的贡献是正的。

回归系数的值的含义：令某个因子对应的回归系数为ai，那么ai越大说明对应到的因子对股票的区分度越好，下期收益高和下期收益低的股票被区分的很明显，所以对于ai高的股票建议配置较高的比例(%)。

评价结果中推荐比例(%)的计算方法：评价结果中给出了一个推荐比例(%),计算方法如下，其中，ai 是第ai 个因子对应的回归系数，N 是回归系数的总个数， wi 是多因子框架计算到的推荐比例(%)。当结果wi 为负数时代表，建议你给因子配|wi| 这么大的比例(%)，然后因子方向调成跟原来相反的。  


# 单因子测试—风格类因子