概率的游戏：不确定的世界 . 第6讲

渔阳 2017-12-29 22:41 字数 6,248 阅读需 16分钟

上一篇:
赌博与投资：赌场里真能走出投资达人吗？ . 第5讲

下一篇: 收藏
波动的惩罚：波动亦有规则？ . 第7讲

“欢迎来到量化小学”

概率的游戏：不确定的世界

▲ 订阅特辑 加入“量化小学”校友圈儿

来自特辑
量化小学
解放你的投资动手能力
最近更新
【学业总结】量化学习的脉络梳理，以及继续学习提高的路径
2019-04-12更新
进阶研究：集成学习和深度学习. 第31讲
2019-03-28更新

详细内容请观看PPT视频

校长语录

- “影响金融产品价格的因素很多，不可能完全预知”

- “未来不见得完全和过去无关…”

- “很多名牌学校出来的，简历上写了各种各样高大上理论的候选人，像这样简单的问题反而回答不上来”

内容梗概

大家好，欢迎来到量化小学。我是渔阳。

在上一期中我们讲到赌场的故事，也讲了怎样利用凯利公式做合理地赌注分配，以达到长期收益最大化的目的。

从这一期开始，我们将会讲一讲怎么在投资世界当中做类似的事情。

首先，要明确一个跟世界观有关的问题。在我看来，影响金融产品价格的因素非常多，不可能完全预知有的时候我们打开电视会听到股评家说：“我看准某一只股票一定会怎么样”，我认为这都是不正确的。

因为根据我们前面讲过的有效市场假说，大部分因素是没有办法去判断的，你所能够分析、判断的只是其中一小部分。因此投资世界是一个不确定的世界，投资的游戏是概率的游戏。

所以今天我们将谈论两方面的内容：第一，基础概率理论；第二，在金融当中的应用。

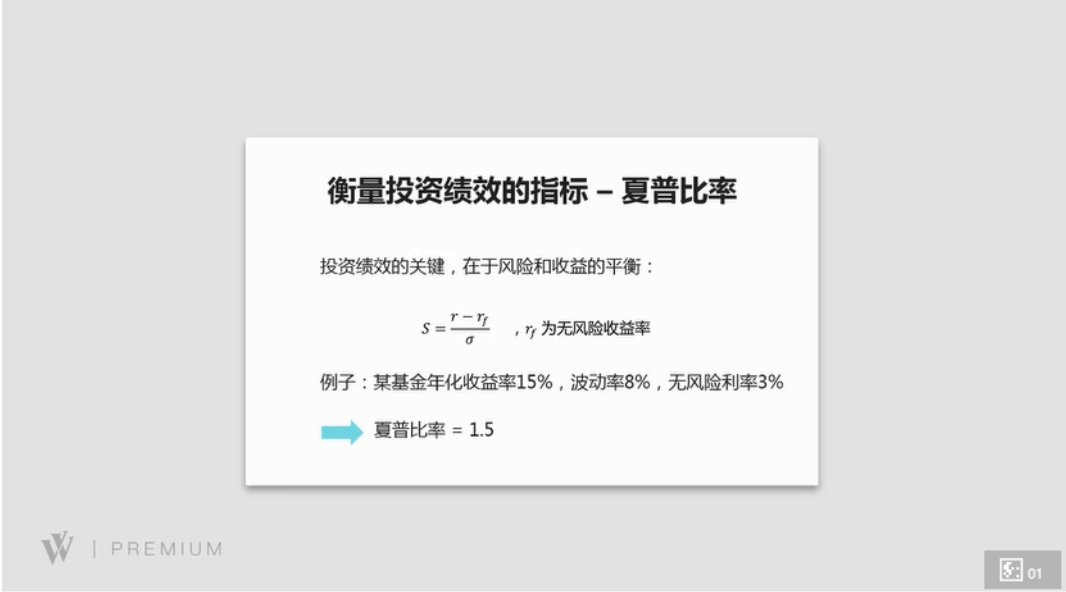
谈到概率理论、谈到数学，也许有的听众会有一种望而生畏的感觉，觉得比较枯燥，实际上我认为不是这样的。因为数学最关键的是要理解公式背后的意义、公式背后的逻辑。当你理解之后，你会觉得数学并不枯燥，而且非常优美，因为它用非常精确的方式去刻画客观的世界。

在量化小学里面，我们的重点不会放在数学公式上面，这在任何一本教科书上都有，大家可以自己去看。我们主要讲的是数学背后的逻辑，以及在金融当中的应用。大家应该将会感觉到，其实这些都是很有意思的。

夏普比率：衡量投资绩效的指标

谈到投资，我们曾经反复指出，它是关于风险和收益之间的一个平衡，因此衡量投资绩效最基础的指标就是**夏普比率**，也就是收益的比值。

> **S=[E(Rp) - Rf] / σp**



这个公式我稍微解释一下，它的分母是比较简单的，就是波动率；它的分子是预期投资收益率减掉无风险收益率，为什么要减掉无风险收益率呢？

这是因为：要想让我去承受风险，我关心的是能够取得多少超额收益。我可以把钱放在银行里面，存款没有任何风险。所以无风险收益率通常也就是银行存款、国债收益率等等。

只有超过无风险收益率的这部分收益对我来说才是有意义的，才值得为它去承受风险。当然现在整个利率水平不高，因此在实际的计算当中，很多时候为了简化起见，把无风险收益率设成零，这个问题不大。

举个例子：如果某个基金年化收益率是15%，波动率是8%，而无风险利率是3%，那么它的夏普比率就是（15-3）÷8=1.5。

光讲数字可能比较枯燥，我们实际对比一下几种投资方式，这儿列了三种：第一种，我简单的长期投资于股票指数，那会怎么样？它的夏普比率在美国也好，在中国也好，预期应该在0.3-0.5这样的水平。

更具体的来说，长期投资于股票指数，预期收益率大致在6%到8%这样的水平，股票指数的波动率大概在20%，因此夏普比率就落在0.3-0.5的区间。

大家很自然地要去寻找一些夏普比率更高的投资方式，如果一个对冲基金能够长期做到夏普比率1.0以上，就算比较靠谱了，也可以长期的生存。如果有谁能够在实盘当中做到2.0以上的夏普比率，他不但是对冲基金界的常青树，而且通常规模都会做得很大。（为什么是这样？我们稍微留一点悬念，等到今天课程讲完之后，大家就会自然地都理解了）

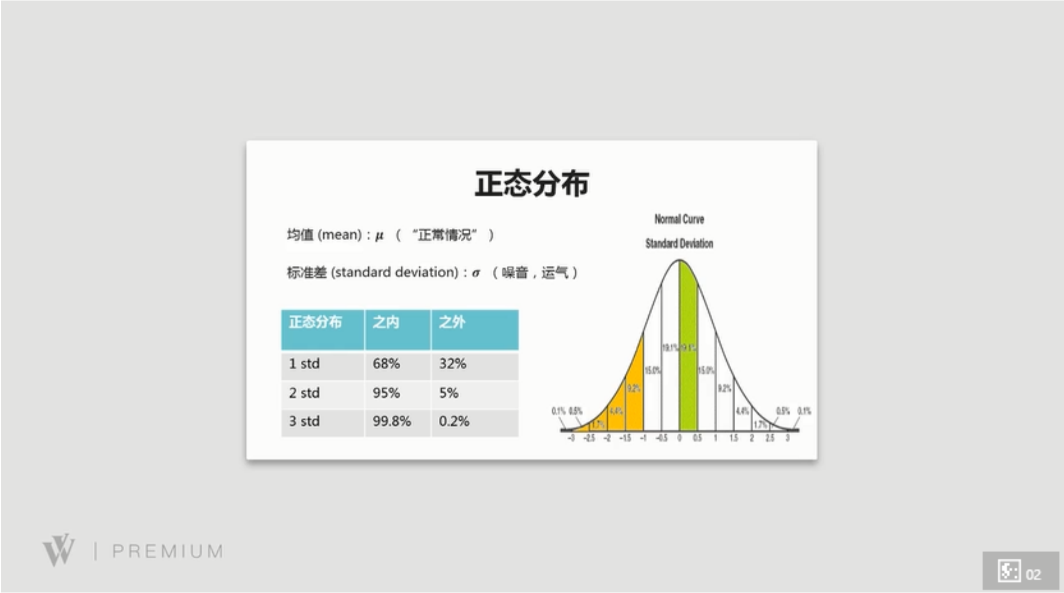
谈到夏普比率，我把最近发生的一个小故事作为**思考题**：

> 我和一个朋友聊天的时候，无意中谈到他所在的对冲基金非常牛，实盘夏普比率居然有5！我想了想就问：那你们是不是每年还有一个月是亏钱的？我的朋友一听就说：还真的和你说的差不多，每年就是有一个月亏钱！
>
> 那我是怎么算的呢？大家也可以想一想，你能不能心算出这个结果？如果不能也没关系，听完今天的课程就知道是怎么算的。

概率分布

接下来我们讲一讲怎么用数学理论来解决实际问题。

谈到概率分布，首先课堂上一定学过**正态分布**，就是如下图右边的这条钟形曲线，它的样子像一个倒扣过来的钟，这是最基础的概率分布，我们很快会看到为什么它这么基本，英文直接就叫**正常的分布**（Normal distribution）。



关于概率分布，有两个非常重要的统计指标：第一个是**均值（mean）**，一般用希腊字母μ代替，它代表的是一种平均情况、一种正常的情况。比如以右面这个钟形曲线为例，是一个标准的正态分布，因此它的均值就是零。

第二个概念是**标准差**（Standard Deviation），希腊字母σ 表示，可以粗略理解它代表的是噪音的程度、运气的大小。如果σ 越大，意味着运气因素越大，越可能发生偏离平均值的情况，反之亦然。

有了这两个统计指标之后，我们就可以实际度量发生各种各样情况的概率有多大。我简单的列了一下，在正态分布的情况下，右边钟形曲线实际发生落在一个标准差（1std）以内

的概率是多大。图上标有数字，大家可以自己去看。在1std之内，它的概率约为68%，在之外约为32%；2std之内是95%，之外是5%；在3std之外，概率就非常小了。

这些数字其实对于平时做投资、资产管理是非常重要的，所以我希望大家能够在大脑中大概有个印象。接下来就会看到，其实这些是非常管用的，应该属于金融方面的常识。

掷骰子概率（动手实验）

正态分布我们有了一个基本的理解，下面我们就谈一谈为什么它如此的重要。可以从上一次讲到的掷骰子的例子谈起，上次有个思考题：掷10000次骰子，超过35500的概率是多大？我们就在这实际算一算。

首先掷一次骰子很简单，就是从1到6的一个均匀分布。很显然可以计算到它的平均值就是3.5，那么它的标准差呢？按照这个公式，如果掷出1，它离平均值差距就是2.5；然后平方，把它都加起来再取平方根，其实这个不难算，我觉得完全是可以心算的，就是约等于3开根号（稍微小一点），所以大概就是1.7。

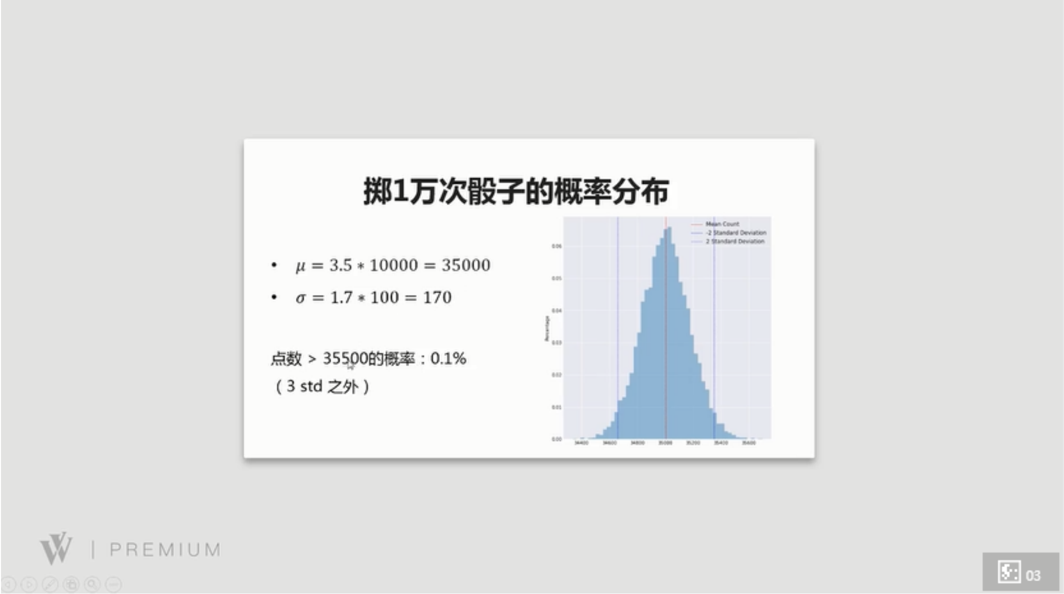
这是掷一次骰子，那么掷一万次骰子会怎么样呢？我们可以用Python做一个实验：现在切到Python，这一段没有用到金融数据的，所以也不用自己的用户名和密码！

case1：掷筛子概率实验

这段程序就是实际模拟掷骰子，扔10000次骰子，模拟10000次。这段是设置的一些参数，看看平均是一种什么样的情况。

这一行是利用Python的库函数生成一些随机的变量，有了这些之后，接下来这段程序就计算总和各种各样的统计量，底下就列了一下，其中包括刚才讲的平均值，这个是35004，离预期很近的。

第三行非常重要，就是标准差是多少？我们的模拟当中计算出来是172。接下来我们可以把分布画出来，看上去就比较有意思，它非常像我们刚才讲到的正态分布。



掷10000次骰子的概率分布，刚才我们用Python进行了模拟（如上图），它非常接近于正态分布。更具体的来看：它的平均值大概在35000这样的水平上。

它的标准差是172，约等于刚才算出的扔一次骰子的标准差：1.7×100，而100正好是10000的开平方。其实这里边是有数学规律的，马上就会看到数学公式就是这样写的。

咱们先把掷10000次骰子的概率问题解决掉，既然是一个正态分布，标准差是170，那么如果偏离500，基本是三倍标准差，它的概率可以从正态分布那张PPT上可以看到，是非常小的，只有0.1%。

因此从实践的角度来说：虽然35500距离预期的平均值只差500，但是它发生的概率是非常小的。我们上次也谈到，这作为一道基本的面试题，其实是可以心算的。

中心极限定理

那为什么掷1个骰子是均匀分布，掷10000次，它就变成正态分布呢？这就运用到概率里面很有名的中心极限定理。

讲到这儿，大家会不会有一种高大上的感觉，我们在大学当中学到的数学居然是有用的！应该给自己的概率老师发一封感谢信。

好，中心极限定理它是讲的是什么呢？就是如果有N个独立同分布的随机变量，所谓独立同分布，就是每一次试验跟其他试验没关系，比如一直掷色子，这就是独立同分布的；或者一直扔硬币，这也是独立同分布的（i.i.d）。

单次实验它的统计量均值是μ、标准差σ，中心极限定理就告诉我们：当N足够大，总和就会近似于正态分布。

另外，它的均值就等于n*μ，非常显然，因为一次是μ，试了n次当然就是乘以N倍。第二句话非常有意思：就是它的标准差，不是按照N线性放大，是按照N平方根线性放大的。

其实这一基本的数学公式就蕴含了很多金融交易中做风控、做投资组合的原理。中心极限定理，在金融当中我们怎么样来用这个公式呢？

收益率的概率分布

首先，我们可以考虑一下收益率的概率分布是什么样的。对于任何一个股票，想象一下在一个比较小的时间切片上的（对数）收益率。比如一分钟或者五分钟的收益率，可以大概认为它们是独立同分布的，也就是现在跟过去是没有什么关系的。

在这样的假设下，我们就可以运用到中心极限定理股票区间收益率是小的时间切片的收益相加，因此它就符合正态分布。

我们也可以深入的再思考一下：**为什么能够大致认为收益率是独立同分布的呢？**

其实这就用到了前面我们讲过的一个理论：弱有效市场假说，在特刊里面讲的，就是股票未来收益率和过去是没有关系的。既然没关系，当然就更加接近于独立同分布（i.i.d），所以金融理论和数学理论可以非常巧妙的结合在一起。

收益率既然是独立同分布，根据刚才讲的中心极限定理，如果把很多个区间收益率叠加起来，自然就是正态分布。而且：预期收益率是随着时间线性放大的，而它的波动率是随着时间平方根放大的。

比如，如果某个股票预期月收益率是0.5%，波动率是5%，那么年化一下收益率变成6%，这是线性放大的；而波动率是按照T的平方根放大，所以它的波动率是17.3%，只放大了三倍多（√12），因此在月的级别上，收益率和波动率的比例是1:10。到了年的级别上，就变成接近于1:3了。

这其实就蕴含了我们经常听说过的关于投资的故老相传的一个理念：**简单策略--“长线为王”**。因为时间周期拉得越长，波动率相对于收益率的占比越低，也就是说亏钱的概率其实是越小的，当然前提是投资的预期收益率必须是正的。

讲到这儿，就有足够多的理论基础来回答刚才那道思考题了。

我的朋友告诉我：牛X对冲基金的夏普比率是5，所以就可以在脑子里面推导一下：

> 不妨假设它的年化收益率是60%，波动率是12%，所以相除正好是5。 很显然，月度收益率大约就是5%，而波动率是根据根号√T来放大缩小的，所以12%除以√12，约等于3.5。

> 如果学过小学奥数的话，应该知道3.5×3.5就12.25了，算是数学上的小常识，所以可以心算。
>
> 接下来就知道月度风险收益比就大约是1.4（5%÷3.5%）。根据刚才讲过的正态分布，落到-1std之外的概率是16%，在这种情况下，月度的风险收益比是1.4，就意味着如果落到了-1.4std之外，就亏钱了。大家自己对照算一下，这是非常显然的事实。
>
> 那么落到-1std之外的概率是-16%，落到2std之外是2.5%，简单地大致估算一下，根据正态分布的钟形曲线落到1.4std之外的概率大概是8%，正好一年12个月平均有一次。

因此如果你对数字足够熟练，就可以像我一样，在十几秒钟之内就把数字算出来，并不是那么难。我们也可以再审视一下刚才讲过关于投资收益夏普比率的观察。

夏普比率--再审视

我们讲了长期投资股指，夏普比率是0.3-0.5。如果放到钟形曲线里面对照一下，其实年度亏损概率大概是30%左右，也就是说如果发生了-0.5std以外的情况，你就变成亏钱了。

如果夏普比率到了1.0呢？同样根据钟形曲线，亏钱的概率就大概是在1/6或者更小。这就感觉比较保险，六年里面才有一年是亏钱的。

如果夏普比率能做到2.0，根据正态分布钟形曲线，亏损的概率就小于2.5%，也就意味着40年里面才有一年是亏钱的。所以如果能做到这样的对冲，基金投资人都会追着给钱。

当然讲到这，我也要特别做一个重要的注释：以上的分析全部是基于正态分布，但实际上收益率会有“尖峰肥尾”的现象，也就是说发生极端情况的概率会更高一些。

如果平时关注金融市场，就知道发生金融危机的时候，股市往往会连续下跌，产生所谓肥尾的现象。

虽然实际情况和正态分布的预测有所不同，还是应该把用正态分布作为出发点，对世界有一个基本的认知，然后在这个基础上再去研究实际情况，可能会有哪些偏差。

股票、基金实际收益率分布

讲了这么多，我们就用quant<OS>实际的观测一下股票和基金实际收益率的分布，看一看沪深300指数的日收益率分布，也看一看公募股票基金月度收益率的分布。

case2：股票基金实际收益率

接下来切到quant<OS>平台，这段代码就是帮助我们实际的观测，包括指数收益率以及股票型基金的收益率分布，开始和每次一样，需要自己的用户名还有密码，连接quant<OS>

数据库。

接下来我们首先研究一下沪深300在过去一年当中的日收益分布，这段程序是定义要取哪些数据、从什么时候到什么时候，下面计算每日收益率，然后把这段程序跟掷骰子的例子差不多，把统计指标都算出来了。

可以看到沪深300的平均日收益率就是0.06个百分点，它的平均波动率是每天0.71个百分点（过去一年多是股票指数波动率比较低的一段时间）。更直观一点，我们可以把过去一年多日收益率分布直方图画出来，看这个形状确实还是比较接近正态分布的。

右面是利用Python的库函数，先估算一下沪深300日收益概率分布的密度函数（蓝色线），这条黑色的线是正态分布的概率密度函数。

如果比较的话，确实看到沪深300日收益率有“尖峰”，尾巴（肥尾）比较肥的现象，就说明发生极端情况的概率会比正态分布预测的稍微多一些。

但是从总体来说，它还是接近钟形曲线的。其实关于尖峰肥尾也可以从上面数字当中看出来，实际测算沪深300日收益率的标准差是0.71个百分点，根据刚才讲的，如果是3std就是2.1个百分点。

如果是纯正态分布，基本上发生日涨跌幅超过2.1%的可能性是微乎其微的。但是从实际的角度来说，过去一年涨跌幅超过2.1%的得有五六天。 这是超出正态分布预期的，也就是我们刚才谈到“尖峰肥尾”的现象，值得注意。

好，接下来中间有两段程序，可以去看个股组合收益率的分布，都是跟正态分布相差不远，但是也有“尖峰肥尾”的现象。我们直接看一下第四部分就是股票型基金的收益率，

分析了规模在5亿以上的股票型基金从2016年以来月度收益率的信息。这段程序也是在设置参数，从数据库当中把所有股票型公募基金都取出来，同时要求它的份额规模大于五个亿。

筛选之后，跟刚才一样计算平均值、标准差这样的统计量，同时也看到筛选出来的基金一共有146只。如果大家有兴趣的话，可以直接把基金名字都给打出来。

看一下这146只公募基金在过去一年多时间里月度收益分布图（左图），和正态分布中心曲线还是有一些类似之处的，总体分布好像稍微偏左一些，右面有一些“肥尾”的现象，就是在个别月份会有非常良好的表现。

在这附带提一句，像这个所谓“尖峰肥尾”，在统计指标上叫Skewness和Kurtosis（偏度和峰度）。任何一本统计书里面都会讲到。看一下实际的计算结果，正态分布的Skewness是0，这里是0.6，说一它是左偏的。在正态分布的Kurtosis是0（注：Python函数定义，已减掉3）。这里的Kurtosis是1.24，就意味着它比正态分布的尾巴要稍微肥一点。

其实这些并非进阶的统计内容，我觉得做投资的都应该知道：观察任何一个收益率分布，想知道它是不是“尖峰肥尾”的时候，就看Skewness和Kurtosis。

再翻到上面看一下刚才沪深300日收益率分布的Skewness和Kurtosis，发现Skewness（偏度）是负的，就意味着是右偏的。同时它的Kurtosis比较高，就意味着它的肥尾现象比公募基金收益的月度分布要更肥一些。

现在咱们回到PPT。沪深300日收益率分布和公募基金月度收益率分布确实与正态分布总体较为接近，同时也确实有“尖峰肥尾”的现象，也就是说发生极端情况的概率会比正态分布要高一些。

为什么研究收益率？

在今天的节目结束之前，我们也稍微谈一点进阶的内容：今天我们一直在讲收益率，那**为什么我们要研究收益率，而不是价格呢？**这其实也是一个非常基本的面试题。

因为**从概率的角度来说，收益率的时间序列，它的概率分布性质比较稳定。**数学上叫**稳态（stationary）**也就意味着它的**均值和标准差不随着时间而改变**，因此今天的概率分布和一个星期之前、一个星期之后都是基本相同的，而价格的时间序列，它不满足这个条件。

我们可以想象：茅台的股价前几年只有200块钱，现在变成700块钱。一个200块的股票，一天波动十块属于很多的；对于700块钱的股票，一天波动十块就不怎么多。因此如果算价格的标准差，它会随着时间的变化而变化的，所以就会使整个研究比较困难，概率理论也用不上。如果用收益率就会好很多。一般来说，收益率是稳态的概率分布。

所以这也是一道非常基础的面试题，但是我感到比较奇怪的是：很多名牌学校出来的（中国、美国都算），简历上写了很多机器学习、各种各样高大上的理论的候选人，像这样简单的问题反而回答不上来。

因此我希望量化小学的同学们要在学习高大上的理论之前，先要掌握一些基础的东西，特别是要掌握这些基础的东西背后的逻辑。

最后，提出几个值得深入思考的问题，大家可以先想一想：

首先**正态分布的假设有没有潜在的问题?**

其实刚才已经指出它不少潜在的问题。第一，未来不见得完全和过去无关，可能是有一些序列相关性的，这也是造成尖峰肥尾问题的原因；

第二，刚才基本假设预期收益率、波动率是不随时间改变的，而实际上如果你对金融市场有一定观察，也会发现这个假设不完全正确，因为市场就是有高波动的时候，也有低波动的时候，那怎么样处理这样的问题？我们之后会讲到。

另外，我们一直在讲波动率小一点好，在同等收益的情况下，不要波动率那么大，那**为什么波动率大就不好呢**？这其实也是一个数学的问题，在下一期当中我们就会更深入的来探讨，所以请看下一期《波动的惩罚》。

今天的节目就到这，谢谢大家，我们下次再见。

-END-

量化小学

渔生

小学而大不遗，量化师生联谊会

感谢大家订阅《量化小学》，这里是学校见识社群，你可以随时提问、随时互动，我们一起投资，一起分享！

风险提示及免责条款

市场有风险，投资需谨慎。本文不构成个人投资建议，也未考虑到个别用户特殊的投资目标、财务状况或需要。用户应考虑本文中的任何意见、观点或结论是否符合其特定状况。据此投资，责任自负。

写评论

请发表您的评论

表情 图片 发布评论