Portfolio Construction and Analytics 读书笔记（第二章）

1　资产管理的介绍

2　随机变量、概率分布和重要的统计概念

2.1　随机变量

随机变量：定义在样本空间ω 上的实值函数。

2.2　伯努利试验和概率分布函数

设伯努利试验一次成功的概率为p 那么一重伯努利试验有如下的概率分布函数：

$$\Pr(\widetilde{X}=x)=\begin{cases}1-p & \text{x=0}\\ p & \text{x=1}\end{cases}$$

2.3　n重伯努利试验

设伯努利试验一次成功的概率为p 那么n重伯努利试验成功x次的概率为：

$$\Pr(\widetilde{X}=x)=\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x},x=0,\ldots,n$$

2.4　正态分布和概率分布函数

正态分布：

$$f(x|\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

概率分布函数(PDF): 表示随机变量在样本空间上的概率分布：

$$\Pr[a\leq X\leq b]=\int_a^b f_X(x)\,dx$$

2.5　累积分布函数

累积分布函数(CDF):

$$F(x)=\Pr[X\leq b]=\int_{-\infty}^b f_X(x)\,dx$$

2.6　描述分布

2.6.1集中趋势的度量

2.6.1.1　均值：

$$\mu=\mathrm{E}\left[X\right]=\int_{-\infty}^{\infty}xP(x)\,dx,$$

2.6.1.2 方差：

$$\operatorname{Var}(X) = \operatorname{E}\left[(X - \operatorname{E}[X])^2\right]$$

2.6.1.3 k阶中心矩:

$$\mu_k = \operatorname{E}\left[(X-\mu)^k\right] = \int_{-\infty}^{\infty} (x-\mu)^k P(x)\,dx,$$

4.矩量母函数

$$M_t(X) = \operatorname{E}\left[e^{tX}\right] = \int_{-\infty}^{\infty} e^{tx} P(x)\,dx,$$

2.6.2风险的度量

2.6.2.1 方差和标准差

在度量投资组合的风险时，首要考虑的就是投资组合的方差和标准差：

1.方差：

$$\begin{aligned}\operatorname{Var}(X) &= \operatorname{E}\left[(X - \operatorname{E}[X])^2\right] \\ &= \operatorname{E}\left[X^2 - 2X\operatorname{E}[X] + \operatorname{E}[X]^2\right] \\ &= \operatorname{E}\left[X^2\right] - 2\operatorname{E}[X]\operatorname{E}[X] + \operatorname{E}[X]^2 \\ &= \operatorname{E}\left[X^2\right] - \operatorname{E}[X]^2\end{aligned}$$

2.标准差：$\sigma_X = \sqrt{Var(X)}$

2.6.2.2 变异系数

当需要比较两组数据离散程度大小的时候，如果两组数据的测量尺度相差太大，或者数据量纲的不同，直接使用标准差来进行比较不合适，此时就应当消除测量尺度和量纲的影响，而变异系数可以做到这一点，它是原始数据标准差与原始数据平均数的比。CV没有量纲，这样就可以进行客观比较了。事实上，可以认为变异系数和极差、标准差和方差一样，都是反映数据离散程度的绝对值。其数据大小不仅受变量值离散程度的影响，而且还受变量值平均水平大小的影响。

变异系数的定义：

$$c_{\mathrm{v}} = \frac{\sigma}{\mu}$$

例如：投资组合A的变异系数为0.8,投资组合B的变异系数为0.5，那么我们可以认为投资组合A的风险比较大。

2.6.2.3 范围

随机变量的范围：即随机变量的取值范围。例如正态分布的取值范围是负无穷到正无穷。

2.6.2.4 百分位数

随机变量X或它的概率分布的分位数Z_a，是指满足条件$\Pr(X \leq Z_a) = \alpha$ 的实数

2.6.2.5　风险价值

风险价值VaR(Value at Risk)：在市场正常波动下，某一金融资产或证券组合的最大可能损失。更为确切的是指，在一定概率水平（置信度）下，某一金融资产或证券组合价值在未来特定时期内的最大可能损失。

给定置信度α:

$$\mathrm{VaR}_\alpha(X) = \inf\left\{x \in \mathbb{R} : F_X(x) > \alpha\right\} = F_Y^{-1}(1-\alpha).$$

2.6.2.6　条件风险价值

在投资组合超过某个给定VaR值的条件下，该投资组合的平均损失值。

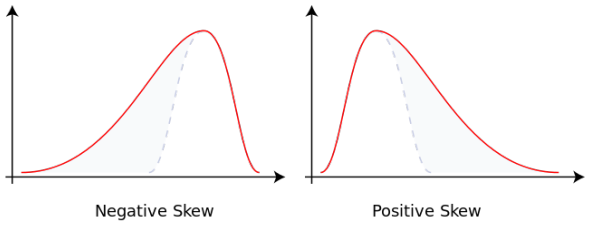
$$CVaR_\alpha(X) = E[-X \mid X \leq -VaR_\alpha(X)]$$

2.6.3偏度

偏度（skewness），是统计数据分布偏斜方向和程度的度量，是统计数据分布非对称程度的数字特征。我们可以从图片中（来自wiki百科）直观地看出正偏度和负偏度：

1.负偏度。密度函数左边的尾巴更加厚实，随机变量主要的取值分布在右边，通常我们也把他称为”右倾斜”。

2.正偏度。密度函数右边的尾巴更加厚实，随机变量主要的取值分布在左边，通常我们也把他称为”左倾斜”。



计算公式：

$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{\left(\mathrm{E}\left[(X-\mu)^2\right]\right)^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}}$$

2.6.4峰度

峰度是描述总体中所有取值分布形态陡缓程度的统计量。这个统计量需要与正态分布相比较，峰度为3表示该总体数据分布与正态分布的陡缓程度相同；峰度大于3表示该总体数据分布与正态分布相比较为陡峭，为尖顶峰；峰度小于3表示该总体数据分

布与正态分布相比较为平坦，为平顶峰。峰度的绝对值数值越大表示其分布形态的陡缓程度与正态分布的差异程度越大。

计算公式：

$$\operatorname{Kurt}[X] = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4} = \frac{\mathrm{E}[(X-\mu)^4]}{(\mathrm{E}[(X-\mu)^2])^2}$$

2.7 协方差和相关系数

协方差:用于衡量两个变量的总体误差。而方差是协方差的一种特殊情况，即当两个变量是相同的情况。期望值分别为E[X]与E[Y]的两个实随机变量X与Y之间的协方差Cov(X,Y)定义为：

$$\operatorname{cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right]$$

相关系数：研究变量之间线性相关程度的量。随机变量X和Y 的相关系数定义为：

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

2.8 随机变量和的性质

1.随机变量和的期望。

$$\mathrm{E}[aX + bY] = a \cdot \mathrm{E}[X] + b \cdot \mathrm{E}[Y]$$

2.随机变量和的方差。

$$Var[aX + bY] = a^2 \cdot Var[X] + b^2 \cdot Var[Y] + 2 \cdot a \cdot b \cdot Cov(X,Y)$$

2.随机变量和的分布。对于独立的随机变量X 和Y, 随机变量Z = X + Y 的密度函数就是X的密度函数和Y的密度函数的卷积。

$$f_Z(z) = \int_{-\infty}^{\infty} f_Y(z-x) f_X(x)\, dx f_Z(z) = \int_{-\infty}^{\infty} f_Y(z-x) f_X(x)\, dx$$

2.9 联合概率分布和条件概率

对于离散型随机变量X,Y，当X=x时，Y 的条件分布为：

$$P_Y(y \mid X = x) = P(Y = y \mid X = x) = \frac{P(X = x \cap Y = y)}{P(X = x)}$$

同样对于连续型随机变量，当X=x时，Y 的条件分布为：

$$f_Y(y \mid X = x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

一般的我们有如下结论：

1.重期望公式：$E(E(X \mid \mathcal{H})) = E(X)$

2.条件方差公式：$\operatorname{Var}(X) = \operatorname{E}(\operatorname{Var}(X \mid \mathcal{H})) + \operatorname{Var}(\operatorname{E}(X \mid \mathcal{H}))$

3.条件协方差公式：$\operatorname{cov}(X, Y) = \operatorname{E}(\operatorname{cov}(X, Y \mid Z)) + \operatorname{cov}(\operatorname{E}(X \mid Z), \operatorname{E}(Y \mid Z))$

边缘分布和密度函数

对于离散型随机变量：

$$\Pr(X = x) = \sum_y \Pr(X = x, Y = y) = \sum_y \Pr(X = x \mid Y = y) \Pr(Y = y)$$

对于连续型随机变量：

$$p_X(x) = \int_y p_{X,Y}(x, y)\, \mathrm{d}y = \int_y p_{X|Y}(x \mid y)\, p_Y(y)\, \mathrm{d}y$$

2.10　Copulas

copula函数描述的是变量间的相关性，实际上是一类将联合分布函数与它们各自的边缘分布函数连接在一起的函数，因此也有人将它称为连接函数。相关理论的提出可以追溯到1959年，Sklar通过定理形式将多元分布与Copula函数联系起来。

Copula函数的定义：$C : [0,1]^d \to [0,1]$ 称为是一个d维copula函数如果：

- $C(u_1, \ldots, u_{i-1}, 0, u_{i+1}, \ldots, u_d) = 0$, 如果某个分量为零，函数为零，
- $C(1, \ldots, 1, u, 1, \ldots, 1) = u$, 如果函数的d-1个分量为1，那么函数是把u映成u，
- C对于它的每一个分量是非减的。

Sklar定理

Sklar 定理（二元形式）：若$H(x, y)$是一个具有连续边缘分布的$F(x)$与$G(y)$的二元联合分布函数，那么存在唯一的copula函数C，使得$H(x, y) = C(F(x), G(y))$。反之，如果C是一个$copula$函数，而F和G是两个任意的概率分布函数，那么由上式定义的H函数一定是一个联合分布函数，且对应的边缘分布刚好就是F和G。

Copulas函数族

1.高斯Copula函数

高斯Copula函数是定义在单位立方体$[0,1]^d$.上面的函数。它是通过定义在$\mathbb{R}^d$上的多元正态函数构造的。

给定一个相关系数矩阵$R \in [-1,1]^{d\times d}$, 参数为R的高斯Copula函数可以写成：

$$C_R^{\text{Gauss}}(u) = \Phi_R\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right),$$

其中Φ^{-1}标准正态分布函数的逆而Φ_R 是均值为零、协方差为R的多元正态分布。

2.阿基米德Copula函数

函数C 被称为阿基米德copula 如果它满足如下条件：

$$C(u_1, \ldots, u_d; \theta) = \psi^{[-1]}\left(\psi(u_1;\theta) + \cdots + \psi(u_d;\theta);\theta\right)$$

其中ψ: $[0,1]\times\Theta \to [0,\infty)$是连续、严格递减的凸函数，并且满足：$\psi(1;\theta) = 0$。$\theta$ 从属于某个参数空间Θ.

常用的阿基米德copula函数：

名称	$C_\theta(u,v)$	参数θ
Ali-Mikhail-Haq	$\frac{uv}{1-\theta(1-u)(1-v)}$	$\theta \in [-1,1)$
Clayton	$\left[\max\left\{u^{-\theta}+v^{-\theta}-1;0\right\}\right]^{-1/\theta}$	$\theta \in [-1,\infty)\backslash\{0\}$
Frank	$-\frac{1}{\theta}\log\left[1+\frac{(\exp(-\theta u)-1)(\exp(-\theta v)-1)}{\exp(-\theta)-1}\right]$	$\theta \in \mathbb{R}\backslash\{0\}$
Gumbel	$\exp\left[-\left((-\log(u))^\theta + (-\log(v))^\theta\right)^{1/\theta}\right]$	$\theta \in [1,\infty)$
Independence	uv	
Joe	$1-\left[(1-u)^\theta+(1-v)^\theta-(1-u)^\theta(1-v)^\theta\right]^{1/\theta}$	$\theta \in [1,\infty)$

2.11 概率分布和取样

在通常情况下，我们无法知道总体的分布情况。所以我们通常通过样本来估计总体。例如，我们独立地观测到n个样本数据：$X_1, \ldots, X_n$,通常采用如下的公式估计总体的均值、方差等参数：

样本均值：

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

样本方差：

$$s_2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_n - \bar{X})^2$$

样本标准差：

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_n - \bar{X})^2}$$

样本协方差：

$$sCov(\bar{X},\bar{Y}) = \frac{1}{n-1}\sum_{i=1}^{n}(X_n - \bar{X})(Y_n - \bar{Y})$$

样本相关系数：

$$r(\bar{X},\bar{Y}) = \frac{sCoV(\bar{X},\bar{Y})}{s_X s_Y}$$

2.11.1中心极限定理

独立同分布的中心极限定理:

设随机变量$X_1X_2\ldots X_n$独立同分布，并且具有有限的数学期望和方差：$E(X_i) = \mu$，$Var[X_i] = \sigma^2 < \infty$，

则对任意z:

$$\lim_{n\to\infty}\Pr\left[\sqrt{n}(S_n - \mu) \le z\right] = \Phi\left(\frac{z}{\sigma}\right),$$

其中

$$S_n := \frac{X_1 + \cdots + X_n}{n}$$

为样本的均值。

2.11.2置信区间

置信区间是一种常用的区间估计方法，所谓置信区间就是分别以统计量的置信上限和置信下限为上下界构成的区间。对于一组给定的样本数据，其平均值为μ，标准偏差为σ，则其整体数据的平均值的$100(1-\alpha)\,\%$ 置信区间为$(\mu - Z_{\frac{\alpha}{2}}\sigma, \mu + Z_{\frac{\alpha}{2}}\sigma)$，其中 α 为非置信水平在正态分布内的覆盖面积$Z_{\frac{\alpha}{2}}$ 即为对应的标准分数。

2.11.3Bootstrapping

Bootstrapping算法，指的就是利用有限的样本资料经由多次重复抽样，重新建立起足以代表母体样本分布的新样本。我们会在后续的章节中结合蒙特卡洛模拟给出详细的介绍。

2.11.4假设检验

假设检验(Hypothesis Testing)是数理统计学中根据一定假设条件由样本推断总体的一种方法。具体作法是：根据问题的需要对所研究的总体作某种假设，记作H_0；选取合适的统计量，这个统计量的选取要使得在假设H_0成立时，其分布为已知；由实测的样本，计算出统计量的值，并根据预先给定的显著性水平进行检验，作出拒绝或接受假设H_0的判断。常用的假设检验方法有u检验法、t检验法、χ^2检验法(卡方检验)、F 检验法，秩和检验等。我们会在后续的章节中给出详细的介绍。