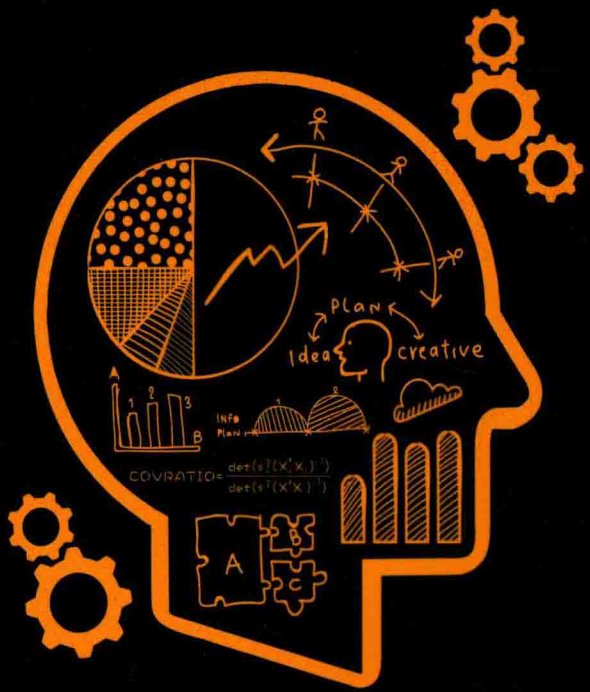

The Analysis and Application of Computational Finance with R

R语言计量金融分析与应用

·认识数据分析的统计原则·计量金融的 R 语言分析程序

·实例说明金融计量方法学·案例详解计量金融实践应用

何宗武 编著

清華大學出版社

R语言计量金融分析与应用

何宗武　编著

清華大學出版社
北京

内 容 简 介

计量金融专业兴起于20世纪90年代的西方，是专为金融市场而设的。随着中国金融业的崛起，这个专业越来越为大家所熟悉，也越来越热门。

本书编写主要侧重于用R来进行经济计量统计模型的运用和时间序列分析，以及计量金融中的数值分析，主要内容包括R的基本环境与安装、R的IDE模式、数据结构和数据处理、数据存取和基本处理、探索性数据分析和可视化、回归分析方法、时间序列入门、波动分析、非定态时间序列、时间序列的结构变动、价差与计量套利、R的金融工具箱、风险与投资组合分析和金融大数据的处理等。

如果你对计量金融感兴趣而且你已经具有一定的数学和计算机基础，那么本书就是一本引导你进入计量金融领域的参考书。书中各章均提供了丰富的范例程序，因此也可以作为大专院校计量金融专业R语言的上机实践教材。

本书为博硕文化股份有限公司授权出版发行的中文简体字版本。
北京市版权局著作权合同登记号　图字：01-2018-1221

图书在版编目（CIP）数据

R语言计量金融分析与应用 / 何宗武编著. —北京：清华大学出版社，2018
ISBN 978-7-302-50286-9

Ⅰ. ①R…　Ⅱ. ①何…　Ⅲ. ①程序语言－应用－金融－经济数学　Ⅳ. ①F830

中国版本图书馆CIP数据核字(2018)第112044号

责任编辑：夏毓彦
封面设计：王　翔
责任校对：闫秀华
责任印制：丛怀宇

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　**邮　　编：**100084
社 总 机：010-62770175　　**邮　　购：**010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：清华大学印刷厂
经　　销：全国新华书店
开　　本：190mm×260mm　　**印　张：**19.75　　**字　数：**506千字
版　　次：2018年8月第1版　　**印　次：**2018年8月第1次印刷
定　　价：79.00元

产品编号：079297-01

推荐序

计量金融专业兴起于 20 世纪 90 年代的西方，是专为金融市场而设的。随着中国金融业的崛起，这个专业越来越为大家所熟悉，也越来越热门。这个专业综合运用了数学（含统计学）和计算机编程技术来解决金融的问题。除了具有金融理论知识，该专业的人才需要具有一定的数学和计算机专业素质。这个专业的人才培养目标导向为投资银行、商业银行、基金公司、证券金融机构、保险公司、信托公司等金融机构。

本书是面向计量金融（或金融工程）方向或专业学习 R 语言而撰写的实践教科书，并不是学习 R 语言的入门书籍，因而读者或者学习者应该具有 R 语言的基础（如语法、数据结构、数据读取、控制流等），以及具有一定的数学基础（含线性代数、微积分、概率统计和随机过程）。只有具有上述两大方面的基础知识，才能够比较透彻地学习到本书的精髓：使用 R 语言系统内建的工具和外加的软件包将统计建模运用于实际的经济计量和计量金融的数值分析与应用中。

R 语言并不是独立存在的程序设计语言，其实是 R 系统的一部分，R 系统中集统计分析、绘图语言和操作环境于一体的数值分析环境。R 系统中可以包含丰富的软件包（通过开源网站下载），将枯燥繁多的数值统计与分析工具化和可视化是它的最大特点。本书的内容主要侧重于用 R 来进行经济计量统计模型的运用和时间序列分析，以及计量金融中的数值分析（期货金融交易数据的分析、投资组合分析与回测等）。

如果你对计量金融感兴趣而且你已经具有一定的数学和计算机基础，那么本书就是一本引导你进入计量金融领域的参考书。本书各章均提供了丰富的范例程序，因而也可以作为大专院校计量金融专业 R 语言的上机实践教材。

资深架构师 赵军

2018 年 6 月

前　言

一般人都知道R语言是免费的开放源码（Open Source），而且功能十分强大。统计分析、统计计算和统计制图是它的强项，R强大的功能要运用程序才能充分释放。一个上百行的程序，往往会让一般的用户不容易接纳这个好工具。因此，本书按照以应用范例程序为中心的方式来编写，让用户一开始在阅读范例中就找到自己最需要的程序进行修改，并与自己要处理的数据结合，这应该是一个理想的学习方法。有别于“面面俱到”的程序设计语言，R语言是一门专注于数据分析的程序设计语言，R语言和它运行的整个系统是一款专注于数据分析的软件。目前业界有关于R的书已经不少，有感于财经学科在学术和实践应用的需要，本书的结构分为三个部分。

- 第一部分：本书在R基本功能部分，添加了“如何用R制作文件”和“使用R+MySQL数据库”等主题，也强化了网络提取数据的函数，让数据提取更为方便。
- 第二部分：在经济计量方法部分涵盖了“多变量GARCH”“阈值VAR（Threshold VAR）和阈值单位根/协整（Threshold Unit Root/Cointegration）”等模型，也详细介绍了结构变化（Structural Changes）等方法；这些方法对于进一步的时间序列分析相当有用。
- 第三部分：计量金融专题则介绍了“期货界常用的价差和统计套利原理”，以及“投资组合分析和回测方法”。这些主题，除了学术研究之外，对于金融从业人员所需要的数据分析相当有帮助。

另外，作者在CRAN上开发了两个程序包iClick和pdR，iClick.GARCH可以将8种概率分布进行一次性的估计，本书也做了如何使用的介绍。最后，简单介绍大数据的数据加载和输出。

何宗武

目 录

第 1 章　最简单的统计分析原理

数据分析的方法一般有两种：一个是统计学，另外一个是数据挖掘。本书重点在于以统计为基础的经济计量方法。

在学习数据统计分析之前，先要具体认识数据分析的统计原则。统计学是为了分析数据，所以，先有数据的概念。本章不谈概率统计的细节，从数据的两个概念开始：样本（sample）和总体（population）。

样本和总体是统计分析的核心概念。实际所收集的数据，无论量有多大，都称为样本，一个样本内记录的数据，称为观察值（observation）。例如，一间教室，内有 50 个学生，教室是样本，学生就是观察值。这种关系，利用代数的名词，样本可以视为一个集合（set），样本内的观察值则是集合内的元素（element）。一般我们会用下面的方式表示一个集合 X：

{X|x1, x2,……}

样本从哪里来的呢？这不是一个脑筋急转弯的问题，这是一个科学研究方法论的问题。所收集的数据不管多大，终究不是“所有”的数据。所以，样本也意味着它只是部分数据。科学研究面对分析的对象，认为样本是由一个默认的总体产生而来的。总体就是种种理论上的概率分布，总体的性质，就是这些分布函数的性质。

通过对样本的研究，推导出总体的性质，这也就称为抽样（sampling）。举一个例子，某大学校园所有的学生为总体，已知总体中男生和女生的比率为 6:4。如果我们不能知道总体性别比率，要如何推导出这个比率呢？就是用抽样。

随机抽一次 100 位学生，记录男女生比率为 4:6，这个数字和总体相差太大，两者差距称为乖离率（bias，又称为偏离率）。要降低乖离率，有两个方法：

- 第 1：多抽几次，例如，100 次。
- 第 2：抽多一点观察值，例如，1000 人。

这样，计算 100 次记录的男女比率的平均数，这个平均数理论上会和总体的真实值很接近，因此，抽样乖离率（sampling bias）就会大大降低。

在上例中，已知总体男女比率其实是假设，而在统计的实际工作中是不可能的，所以，统计学的研究，提出了种种理论函数，通过假设来描述总体。概率学研究这些总体的性质，就是概率分布与随机过程；统计学则是从分析抽样，推导出其总体，就是估计和检验。

以“样本-总体”为基础的分析，类似“个人-人类”，在学习统计数据分析时，需牢牢记住。除此之外，统计学另一个发展分支，不从总体建立出发，称为非参数统计（Nonparametric Statistics）。本章将简单介绍。

1.1 统计分析原理

数据分析的核心是为了预测（Prediction），预测之前，要先恰当地描述数据，才能追踪数据的性质。我们先解释什么是“描述”一项数据。

假设我们有一组代号为 X 的数据，把它想象成 Excel 的 A 栏，有 1000 项观察值。令 X 的均值是 μ，描述这组数据的一个方法，就是先把它写成：

$$X = \mu + e$$

也就是说，数据可以由两部分组成：均值 μ 和剩余 e。数据分析的终极目的是预测（Prediction），如果我们可以用均值来描述这组数据，那么均值就有一个更专业的名称：样本期望值（sample expected value）。因为我们期待 1000 项观察值都围绕着均值随机变动，一个好的期望值，剩余的 e 就是随机散布于期望值的两端。把 X 的数据画成图 1-1-1，图 1-1-1 是一个分布图，y 轴的刻度是对应 x 轴数字的次数，比如 250，就是说数据中均值 μ 附近大约 55~60 的区间，数据个数约有 250 个。所以，这也称为次数分布。因为是真正的数据样本，图 1-1-1 称为样本直方图（Histogram）。

另一个看法是从预测误差（Prediction error）e 看：$e = X - \mu$，如图 1-1-2。直方图可以协助我们对自己数据的分布状况有一个基本的了解。例如：较多的数据靠中间，较少的数据在两端。

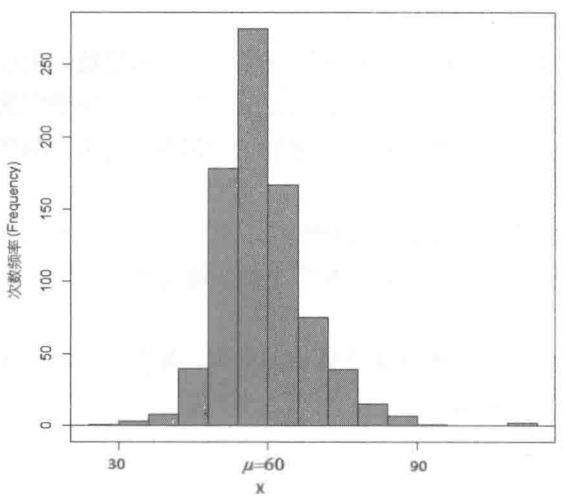

图 1-1-1　数据 X 的样本直方图（Histogram）

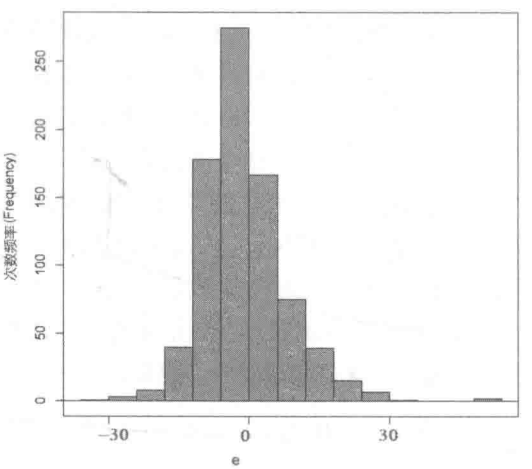

图 1-1-2　预测误差 e 的样本直方图

无论是哪一种图，概率都可以看成是样本直方图背后的一个理论模型。例如，我们可以把样本直方图想象成是“由一个总体所生出来的”，如图 1-1-3 的理论上的正态分布（也称为常态分布或高斯分布）。理论上的正态分布，是一个数学公式，所以可以呈现出连续且平滑的形状。例如，正态分布的公式如下：

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

概率就是一个通过数学公式描述样本分布的学问，一旦能够确认样本是总体的一个投影，这个公式内涵的数学性质就可以用来建立样本的预测。

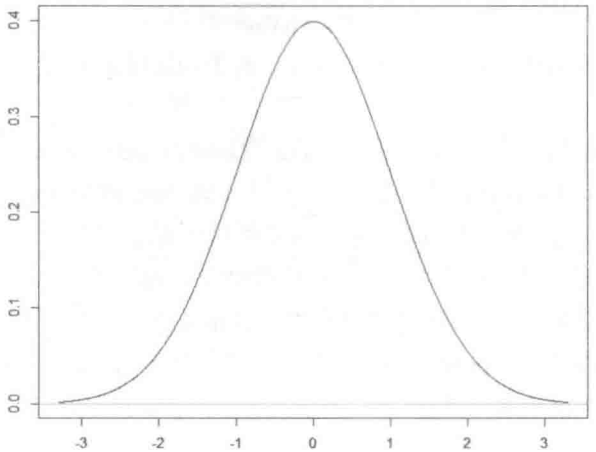

图 1-1-3 标准正态分布

对数据拟合概率分布，最好的做法就是对预测误差去画；所以，需要先产生期望值，然后计算样本数据和期望值的差距（e）。也就是说，我们要先估计（estimation） 期望值。基本估计的有“两种参数”：期望值和方差；以及“两类形态”样本和条件。更多内容，参考 1.1.1 节。

1.1.1 估计原理

统计分析的对象是数据，基本就是两件事：估计（estimation）和检验（test）。我们就分别解释这两件事。估计就是应用估计方法（estimation method）从数据计算出特定参数，检验就是用一个检验统计量（test statistic），检验所计算的参数是否具备我们期待的统计性质，例如，显著性。估计和检验两者息息相关。

具体来说，第一阶段，我们要从数据之中计算样本期望值、样本方差；高级阶段，我们要计算条件期望值和条件方差。遵循传统习惯，用 Y 表示被解释变量（注：一般数学中称为因变量），这个统计项目可以用一个 2×2 的表，如表 1-1-1 所示。

表1-1-1 统计项目表

	期望值	方差
样本（sample）	$\mu=E(Y)$	$\sigma^2=E(Y-\mu)^2$
条件（conditional）	$\mu y\|x=E(Y\|X)$	$\sigma^2 Y\|X=E((Y-\mu)^2\|X)$

样本期望值和样本方差的定义，用方程式来说明，一个样本 Y 可以表示如下：

$$Y=\mu+e$$

也就是样本期望值（μ）和剩余（e，也称为残差）的加总。这样可以知道，样本期望值是一个固定数，或称常数（constant）。条件均值（conditional mean）则意味着：期望值是否会与其他变量连动。其英文的表示就是 “conditional on other variables”。$E(Y|X)$ 毕竟只是一个符号，在实际应用的估计中，必须假设它的函数关系，例如，假设 Y 和 X 之间的关系是线性的（linear）。一个最简单也是最重要的函数模式，就是线性模式（Linear model）：

$$E(Y|X)=a+bX$$

代入数据时，就必须把期望值符号 E 拿掉，再补上残差：

$$Y=a+bX+e$$

这样做，是因为一般都假设残差的期望值是0。对于以这样表示的条件均值，必须估计参数 a 和 b。

统计上有三个常用的估计方法：最小二乘回归（Least-Square regression，LS），极大似然估计量（Maximum Likelihood Estimator，MLE）和矩量法（Method of Moments，MoM）。统计上的估计方法就这么多，都是以这三种为基础的架构。其余的估计方法都是这三种方法的修改或扩充。例如，一般矩量法 GMM 就是将矩量法常规化；两阶段最小二乘法（2-Stage LS）是在 LS 架构下，修改为两阶段引入工具变量。如果还要说有其他的，就是数值计算方法（Numerical methods，或称为数值方法）。例如，拔靴法（Bootstrapping）和贝叶斯统计方法（Bayesian method）。但是，数值方法，也会搭配三种基本架构进行计算。

1.1.2 检验原理

估计完之后，一定是检验某个特定的虚无假设，我们以一台计算机产生的回归报表为例，如表 1-1-2 所示。

表1-1-2 一台计算机产生的回归报表

Variable	Coefficient	Std. Error	t-Statistic	Prob.
Constant	1.740592	0.122648	14.19176	0.0000
Beta1	-1.923351	0.134289	-14.32245	0.0000
Beta2	0.080352	0.150364	0.534381	0.5932
Beta3	0.331608	0.186346	1.779535	0.0755
Beta4	1.312026	0.676490	1.939463	0.0528
R-squared	0.260821	Akaike info criterion		2.402585
Adjusted R-squared	0.257521	Schwarz criterion		2.429242
Log likelihood	-1077.365	Durbin-Watson stat		1.685408
F-statistic	79.03878	Prob(F-statistic)		0.000000

假设被解释变量为 Y，这个报表的第 2 列 Coefficient，估计的方程式如下：

$$Y=1.74-1.92\cdot\mathbf{Beta1}+0.08\cdot\mathbf{Beta2}+0.33\cdot\mathbf{Beta3}+1.31\cdot\mathbf{Beta4}$$

第 4 列的 t-Statistic 是 t-统计量。这个统计量检验的虚无假设 H_0，对立假设 H_a 以及 t 检验统计量，以 Beta1 为例，如下：

$$H_o:\mathbf{Beta1}=0 \qquad H_a:\mathbf{Beta1}\neq 0$$

$$t-Statistic=\frac{\mathbf{Beta1}-0}{\mathbf{Std.\ Err}}=\frac{-1.923351}{0.134}=-14.322$$

因为虚无假设 H_0 是对 0 进行检验，所以，也称为统计的显著性检验（Test for statistical significance）。t-Statistic 的原理是这样的：把虚无假设相减放在分子，标准偏差放在分母，相除之后，如果拒绝虚无假设，则这个相除之后的统计量会大到足以拒绝。是否大到足以拒绝虚无假设，就由这个检验统计量的 p-value，也就是表中最右一列的 Prob.来判断：如果 Prob. 小于基准的 5%，则我们可以拒绝虚无假设，也就是说这个参数 Beta1 显著不同于 0。以这个例子而言，Prob.很小，所以，估计的参数值-1.923351 是显著的。

这张报表是标准的回归估计结果，所有的软件都会这样产生。但是，如果要检验参数 Beta1

是不是特定的数字，就不能参考报表的 Prob.了，看下例。

如果我们要检验的问题如下：

$$H_o: \textbf{Beta1} = -2 \qquad H_a: \textbf{Beta1} \neq -2$$

$$t-Statistic = \frac{\textbf{Beta1} - (-2)}{\textbf{Std. Err}} = \frac{0.07}{0.134} \approx 0.55$$

这样的检验，就不能看报表后面的 Prob.，翻阅统计书后面的检验表，5%显著水平的临界值大约是 1.92，也就是说，要拒绝虚无假设的 t-Statistic 不是正的数字大于 1.92 就是负的数字小于 -1.92。上面计算出的 t-Statistic 约为 0.55，就是说，这个虚无假设是被接受的。也即表明，-1.925 这个参数的信任区间也包含了-2 这个数字。

其次就是 F 检验。报表下半部有一个 F-statistic 和这个统计量的 P-value，Prob(F-Statistic)，这个其实是一个联合检验，虚无假设如下：

$$H_o: \textbf{Beta1} = \textbf{Beta2} = \textbf{Beta3} = \textbf{Beta4} = 0$$

$$H_a: H_o \textbf{ is not true}$$

F 统计量的构建，和 t 不一样。它的分子是虚无假设为真的时候，估计一个只有截距的回归，计算经自由度修正的“残差平方和”（或称为剩余平方和）；分母则是以报表估计结果，计算经自由度修正的“残差平方和”。两者相除，如果虚无假设是错的，则虚无假设为真的模型，所产生的方差会很大。

1.2　函数原理和数据分析

虽然函数的应用及其基本概念在前文已经介绍了，但是，从数据分析的角度，在着重估计数据的统计章节中介绍应该更有体会。因为统计的一些实际范例都脱不了这个框架，其基础在于函数和极限的概念。函数的定义域和值域也是数的集合，所以要建立极限的概念，必须先从实数（real numbers）的概念着手。本节简单复习一下实数、函数的概念。

（1）实数（Real Numbers）

实数可以借助中学数学所学的坐标轴来理解，只要实际划得出来的线，我们就称之为实线（real line），在线上每个点的坐标值就称为实数。实线以 0 为原点，分为正、负两个区域，左边为负数、右边为正数。非负实数指的是：不是“正数”就是“零”。在实线上，越往左越小；反之越大。

实数由“有理数”（rational numbers）和“无理数”（irrational numbers）共同构成。任意一个实数，若能由一个分数表示，则为有理数；反之，则为无理数。有理数的表现形式有两种：

① 有限小数，例如：$0.4=\frac{2}{5}$；$0.875=\frac{7}{8}$。

② 无限循环的小数，例如：$0.333\ldots=0.\dot{3}=\frac{1}{3}$；$1.714285714285=1.\overline{714285}=\frac{12}{7}$。

无理数的例子有 $\sqrt{2}\approx1.4142135623\ldots$；$\pi\approx3.1415926535\ldots$；$e=2.71828\ldots$。

（2）函数（Function）

有了数的基本概念后，我们可以进一步认识函数。函数是指两个以上的集合间的对应关系。为方便起见，令这两个集合一个为 X、一个为 Y。这两个集合内的元素（elements）均是由实数所

构成。使用简单的符号，函数可表示如下：

$$f: X \rightarrow Y$$

在口语上，我们可说成 X 集合里的数字，通过 f 的转换（或运算），变成 Y 集合里的数字。因此，函数 f 事实上定义了一个“运算”；这个运算，将 X 变成 Y。例如：$y=2x$，是说所有的 x 乘上两倍，变成 y。在数学上，我们写成 $y=f(x)$，读成“y 是 x 的函数”。在这个表示下，y 由 x 所决定，故我们称 y 为“因变量”（dependent variable），x 为自行定义，故称 x 为“自变量”。以数学语言来说，我们称 X 是 Y 的定义域（domain），Y 是 X 的值域（range）。

（3）函数对问题的思考——经验研究第一步

函数 $y=f(x)$这样的形式表示出了两个变量的量化对应关系，在财务、经济、管理的学科扮演了一个理性思考的角色，也就是建立理论的数学基础。对于一个做研究的人而言，被观察的现象或问题就是 y，好比苹果从树上落下；根据逻辑分析或理论，心中臆测的答案就是 x；两者之间如何建立相互关系就是函数的运算。

表 1-2-1 对应类型的前 3 个是比较倾向数学形式的称呼，后面则是从事问题思考时常常使用的逻辑架构。事实上，就整个财经学科所谓的理论，就修正为：“问题 y 受哪些 x 因素的影响（或决定）”。例如，了解资产回报率受哪些因素所决定，就是资产定价理论；资产回报率是 y，哪些因素就是 x。了解消费变化受哪些因素所决定，就是消费理论；消费变化是 y，哪些因素就是 x。

表1-2-1　函数类型

类型	y	x
1	因变量	自变量
2	依赖变量	独立变量
3	内生变量	外生变量
4	果	因
5	被解释变量	解释变量
6	产出	投入

学者研究的成果，可写成以函数表示的理论，例如：

消费理论的恒常所得假说，认为消费由恒常所得所决定，可以表示成：

$$\text{消费}=f(\text{恒常所得})$$

资本资产定价，认为权益回报率由风险因子所决定，可以表示成：

$$\text{权益回报率}=f(\text{风险因子})$$

学校课程里面所学习的各种知识多是在解说 x 的内容。学术研究的成果则告诉我们为什么恒常所得会影响消费、如何影响、风险因子有哪些、如何决定回报率等。同时也提供真实世界的数据给予某种程度的佐证，这些都牵涉到 f 的运算方式。

所以，学习使用函数形态对掌握问题的形式是开始训练理性思维的第一步。

1.3 再进一步

望文生义，期望值的意义就是说观察者期望他所观察的随机现象，有一个收敛或集中的位置（Location），只要掌握这个数字，就能掌握数据的特征，也可以说，对于所观察对象的变动，会

有比较高的预测力（Predictability）。期望值也称为数学期望值（Mathematical Expectation，μ），简单地说，期望值就是用概率去加权计算的平均数或均值（Weighted Mean）。例如，一个给小费的行为，某餐厅观察，客人给的小费（随机数）的样本空间是：给 100 元的概率是 0.6，给 150 元的概率是 0.3，给 250 元的概率是 0.1。因此，小费的期望值为：

$$\frac{6}{10}\times 100+\frac{3}{10}\times 150+\frac{1}{10}\times 250=60+45+25=130$$

如果不看概率加权，直接计算样本平均数（Sample Mean，或称为样本均值），则为：

$$\frac{100+150+250}{3}=\frac{500}{3}\approx 166.66$$

两者有着还不算小的误差。所以如何估计正确的期望值，就需要进一步的研究。因此，如果随机数是由一个概率分布在背后控制所产生的数据，忽略这个分布的概率性质对于预测是很危险的。用数学方程式来表示：

- 离散随机变量 y，概率质量函数 $f(y)$，则其期望值为：

$$\mu=\sum_{y\in S} y_i f(y_i)$$

- 连续随机变量 y，概率密度函数 $f(y)$，则其期望值为：

$$\mu=\int yf(y)\mathrm{d}y$$

另外，数据的第 2 个性质就是衡量样本观察值和期望值的差距程度有多大，这个差距也称为偏差（deviation，或称为离差）程度或散布（scatter）程度。每个样本和期望值相减，都可以得到一个偏差数字；这些数字，有的高于期望值，有的低于期望值，有的和期望值很接近。如果直接把偏差加起来算一个平均偏差，会因为正负互相抵消，而低估了观察值的离散度。因此，把所有的偏差平方以消除正负，再加总算平均值，就可以得到一个衡量数据值偏离算术平均值的程度，称为方差（variance），这个数字开根号就是标准偏差（standard deviation，或称为标准差）。

- 离散随机变量的方差：$\sigma^2=\sum_{y\in S}(y_i-\mu)^2 f(y_i)$
- 连续随机变量的方差：$\sigma^2=\int(y-\mu)^2 f(y)\mathrm{d}y$

期望值和方差是统计分析的核心。除此之外，数据分析还要看数据的集中程度和集中的形态，测量的方法就是峰度（kurtosis，或称为峰度系数）和偏度（skewness，或称为偏度系数）。另外，峰度也称为峰态，偏度也称为偏态。