

通用经济系列教材

统 计 学

(含练习题及参考答案)

金勇进 编著

 中国人民大学出版社

通用经济系列教材

统 计 学

(含练习题及参考答案)

金勇进 编著

中国人民大学出版社

· 北京 ·

图书在版编目(CIP)数据

统计学/金勇进编著.

北京:中国人民大学出版社,2009

(通用经济系列教材)

ISBN 978-7-300-11529-0

I. 统…

II. 金…

III. 统计学-教材

IV. C8

中国版本图书馆 CIP 数据核字(2009)第 219954 号

通用经济系列教材

统计学

金勇进 编著

Tongjixue

出版发行 中国人民大学出版社

社 址 北京中关村大街 31 号

邮政编码 100080

电 话 010-62511242(总编室)

010-62511398(质管部)

010-82501766(邮购部)

010-62514148(门市部)

010-62515195(发行公司)

010-62515275(盗版举报)

网 址 <http://www.crup.com.cn>

<http://www.ttrnet.com> (人大教研网)

经 销 新华书店

印 刷 北京东方圣雅印刷有限公司

规 格 170mm×228mm 16 开本

版 次 2010 年 1 月第 1 版

印 张 25

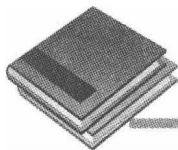
印 次 2010 年 1 月第 1 次印刷

字 数 455 000

定 价 29.80 元

版权所有 侵权必究

印装差错 负责调换



前言

翻开报纸，打开电视，登录网络，我们就会发现越来越多的数字，我们正生活在数字的海洋中。其实，每个人的生活都与这些数字息息相关，也许我们会想：

- 股市有回暖之势，股价指数不断上扬，想要投资现在是否应该进入？

- 官方公布全国职工平均工资比上年同期增长 18%，我的工资上涨了多少？

- 有关方面公布的央视春节联欢晚会调查中的观众满意度达到 83%，这个数字是否可信？

类似的数据比比皆是，在好奇之余，我们不禁要询问，这些数字是怎么得出的，它说明了什么，怎样对这些数字进行恰当的分析？

统计学是一门关于数据的科学，是关于数据搜集和数据分析的科学。对于统计专业的学生，毫无疑问，统计学课程是他们最重要的专业基础课；对于非统计专业的学生，统计学课程也非常重要，它是分析和解读数据最重要的方法和工具。具备了统计学的知识，我们就可以在数字的海洋中遨游，并发现数字背后许多有趣的现象和规律。

统计学与数学有联系，因为分析数据、寻找数字背后的规律、用



已有数据对未知现象进行推论、对未来的情况进行预测，都需要利用统计模型。但统计学不是数学，数学研究的是抽象的数量关系，采用的是演绎思维；统计学研究的是具体的数量关系，具有不确定性，主要采用归纳思维。本书是统计学的基础性教材，笔者在写作过程中尽量避免复杂的数学推导，具有高中水平的读者就可以读懂。

全书共十二章，具体内容如下：

第一章引论，主要讨论了数据的类型，介绍了统计学中的最基本概念和几种最常用的统计软件，使读者对什么是统计学有了一个大致地了解。

第二章数据的搜集，主要介绍统计数据是怎么来的。其实，数据的来源主要有两个，通过调查和实验可以获得不同特征的数据，而通过误差分析，就可以从某个角度解读数据的质量。本章的核心是通过什么样的技术手段来获取高质量的统计数据。

第三章和第四章是姐妹章，名称分别为“数据的描述 1——数据的直观显示”和“数据的描述 2——重要的统计量”。第三章主要介绍了统计表和统计图，不同统计表、统计图的特征和应用场合，以及如何制作规范的统计表、统计图；第四章介绍了一些重要的统计量以及每个统计量的作用和特点。数据的描述既是统计分析的一个组成部分，也为后面要讲到的统计推断做了铺垫。

第五章概率和概率分布，这是统计推断的理论基础。在高等学校中，概率论课程是一门重要的基础课程，学过概率论课程的读者可以略过本章。

第六章参数估计，本章讲授利用样本统计量对总体参数进行估计的方法，包括参数估计的基本原理，进行点估计、区间估计的方法，以及进行区间估计时确定合适样本量的方法。参数估计是统计推断的重要内容。

第七章假设检验，这是统计推断的另一个重要内容。该方法先对研究总体的参数提出假设，然后利用样本数据检验事先所做假设是否合理，实质上是通过判断样本信息与原假设是否有显著性差异来决定原假设的可信度。

第八章列联分析，这是统计推断方法在定性数据上的应用。列联分析包括拟合优度检验和独立性检验，本章对这些定性数据的检验内容进行了比较详细的讨论。

第九章方差分析，这种统计方法用于分析分类型变量与数值型变量相互之间的关系，是假设检验方法在不同数据类型应用上的延伸。本章主要介绍了单因素方差分析和双因素方差分析，有助于加深读者对统计思想的理解。

第十章相关与回归，主要讨论变量之间的关系。本章通过统计建模的方法，将变量之间联系的密切程度数量化，并通过对模型系数的求解，实现统计预测的



功能。本章内容有非常广阔的应用空间。

上面的第六章至第十章都是统计推断中的经典内容，而第十一章和第十二章主要是对动态统计数据进行描述和分析的方法。

第十一章时间序列分析，本章内容既包括对动态数据进行描述的方法，也包括对动态数据进行统计推断的方法。本章介绍的统计方法也可用于预测。

第十二章指数，这是另一类描述事物动态变化的方法和工具，在社会经济生活中，指数方法有着广泛的应用。本章主要介绍指数理论，包括指数的编制、指数的分解以及如何利用指数进行分析。本章还介绍了经济生活中几种重要的指数。

本书的写作具有如下几个特点：

(1) 尽量通俗易懂，避免复杂的数学推导。统计学中有许多定理，而不同的统计方法又有不同的数学性质。本书在写作时，尽量以通俗易懂的方式，采用逻辑叙述和分析的方法，解释公式背后的统计思想，避免枯燥的数学公式推导。一些重要且必要的公式推导，均放到各章最后，并以附录形式表现，即使略去这些推导，也不妨碍读者对全书的理解。

(2) 计算机的迅速发展，解决了统计计算中的诸多问题，使不同的统计方法有了更广阔的应用空间。面对这种情况，本书在相应的计算部分都应用了计算机，同时介绍了统计软件的操作步骤和输出结果，而不必进行繁杂的手工计算。本书的统计软件以使用最广泛的 SPSS 软件为模版，相应计算机操作的具体步骤均在附录中说明，以不影响全书正文表达的紧凑。

(3) 本书的写作以应用为背景，在统计方法介绍中突出应用的特点，同时强调读者的动手能力，各章均有一些思考题和计算题，并附之计算题的参考答案，供读者在练习中参考。

本书可以作为非统计专业统计学课程的教学用书，也可以作为其他人员学习统计方法的参考书。

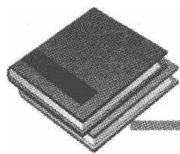
本书是集体劳动的成果，由我和我的研究生们共同完成。其中，金勇进负责全书编写大纲的设计，并负责第一章和第八章的写作；其余各章由我的学生分别提供初稿，具体为谢佳斌（第一章与金勇进合作，第六章），张璞（第二章，第九章与汤琰合作），陶然、王俊合作（第三章，第四章，第十一章），艾小青（第五章），汤琰（第七章，第九章与张璞合作），黄媛（第十章），孙欣（第十二章），金勇进对全书进行了最后的修改和总纂，张璞为全书进行了技术处理。在此，向这些同学表示深深的谢意！同时，感谢中国人民大学出版社为出版本书所付出的辛勤劳动。



尽管我们十分努力，但书中一定会有不少的错误和缺憾。对于书中的不足，恳请各位专家和读者提出宝贵意见。

金勇进





目 录

第 一 章	引论.....	1
	§ 1 统计数据与统计学	1
	§ 2 一些基本概念	4
	§ 3 统计与统计软件	8
第 二 章	数据的搜集	12
	§ 1 数据的来源.....	13
	§ 2 数据的误差.....	21
	§ 3 数据文件.....	26
第 三 章	数据的描述 1——数据的直观显示	31
	§ 1 用统计表描述数据.....	31
	§ 2 用统计图描述数据.....	38
	§ 3 用计算机实现绘制统计图.....	50
第 四 章	数据的描述 2——重要的统计量	60
	§ 1 集中趋势的描述.....	60

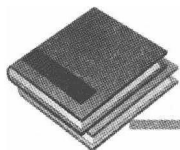


	§ 2 离散趋势的描述.....	72
	§ 3 偏态与峰度的描述.....	78
	§ 4 数据的标准化处理.....	82
第 五 章	概率和概率分布	89
	§ 1 概率的问题.....	89
	§ 2 离散变量的概率分布.....	99
	§ 3 连续变量的概率分布	105
	§ 4 抽样分布	114
第 六 章	参数估计.....	117
	§ 1 点估计	117
	§ 2 区间估计	124
	§ 3 一个总体参数的区间估计	124
	§ 4 两个总体参数的区间估计	128
	§ 5 关于样本量	133
第 七 章	假设检验.....	138
	§ 1 假设检验基本问题	139
	§ 2 一个总体参数的检验	143
	§ 3 两个总体参数的检验	150
第 八 章	列联分析.....	170
	§ 1 定类数据与列联表	170
	§ 2 拟合优度检验	174
	§ 3 独立性检验	179
	§ 4 列联表中的相关测量	181
	§ 5 列联分析中应注意的问题	185
第 九 章	方差分析.....	194
	§ 1 基本问题	195
	§ 2 单因素方差分析	198
	§ 3 双因素方差分析	205



第 十 章	相关与回归	217
	§ 1 相关分析	217
	§ 2 回归分析	222
	§ 3 用回归进行预测	236
	§ 4 Logistic 回归	239
第十一章	时间序列分析	255
	§ 1 时间序列的描述	255
	§ 2 时间序列的分解法	264
	§ 3 时间序列的平滑法	276
	§ 4 ARIMA 模型	282
第十二章	指数	301
	§ 1 基本问题	301
	§ 2 加权指数	305
	§ 3 指数体系	317
	§ 4 几种常用的指数	324
常用统计表.....		337
参考文献.....		377
习题参考答案.....		378





第一章

引 论

§ 1 统计数据与统计学

§ 1.1 统计数据

我们每天都可以从报纸上、网上、电视机中看到、听到各种数据。例如，“2009年7月，北京期房商品住宅的销量为12 840套，剔除其中的经济适用房和限价房外的期房商品住宅签约套数为10 862套，日均约350套，较6月份的日均385套下跌9.1%”；又如，“央行在公开市场发行了50亿元一年期央行票据，结果发行利率水平较前次发行直降近95个基点，至2.245 9%，达到了2006年5月时的水平”，等等。类似的股票行情、物价指数、房价波动、汇率变化，乃至交通事故、死亡人数等各种数据围绕着我们，我们生活在数据之中，想避开都难。

数据是统计的起点，因此我们有必要了解数据的类型。

统计数据按照所采用计量尺度的不同可分为三种类型：一是数值型数据，是指用数字尺度测量的观察值，如GDP等宏观经济运行数据、每天进出海关的旅游人数、某地流动人口的数量等。数值型数据表现为具体的数值，统计处理中的大多数数据都是数值型数据。二是



分类型数据，是指对数字进行分类的结果，如人口按性别分为男、女两类，受教育程度也可以按不同类别来区分。这种数据通常用频数（率）来表示分类的结果，如某地区男性常住人口占 52%，女性占 48%。三是顺序型数据，是指数据不仅是分类的，而且其类别是有序的。例如，满意度调查中的选项有“非常满意”、“比较满意”、“比较不满意”、“非常不满意”等。在这三类数据中，数值型数据由于说明了事物的数量特征，因此可归为定量数据，分类型数据和顺序型数据由于定义了事物所属的类别，说明了事物的品质特征，因而可统称为定性数据，见图 1—1。区分数据的类型非常重要，这是因为不同类型的数据在一些情况下需要用不同的统计方法进行处理。

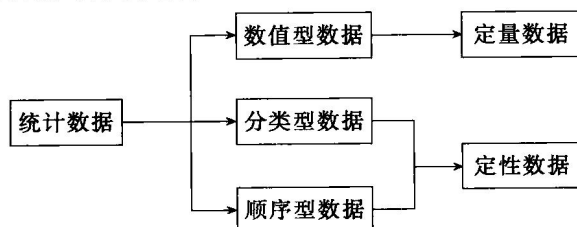


图 1—1 统计数据的分类

此外，还可以从其他角度对统计数据进行划分：

按照收集方法的不同，可将统计数据分为观测数据和实验数据两类。观测数据在没有对事物进行人为控制的条件下得到，是通过调查或观测收集到的数据，主要集中在社会经济领域。例如，对商品零售价格变动水平的测量可以得到商品零售价格指数、对股票价格变动水平的测量可以得到股票价格指数。实验数据是指通过在实验中控制实验对象而收集到的数据，主要集中在自然科学领域，如某种新型电池的使用寿命、一种新型降压药疗效的实验数据等。

按照是否与时间相联系，还可以将统计数据分成截面数据和时间序列数据。表 1—1 中的数据为截面数据，描述了现象在某一时刻的变化情况。表 1—2 中的数据则为时间序列数据，描述了现象随时间变化的情况。

表 1—1 2007 年我国部分地区国内生产总值及其构成数据 单位：亿元

地区	国内生产总值	第一产业	第二产业	第三产业
北京	9 353.32	101.26	2 509.40	6 742.66
天津	5 050.40	110.19	2 892.53	2 047.68
河北	13 709.50	1 804.72	7 241.80	4 662.98
山西	5 733.35	269.68	3 438.58	2 025.09
内蒙古	6 091.12	762.10	3 154.56	2 174.46



表 1—2

2001—2007 年我国国内生产总值及其构成数据

单位：亿元

年份	国内生产总值	第一产业	第二产业	第三产业
2001	109 655.2	15 781.3	49 512.3	44 361.6
2002	120 332.7	16 537.0	53 896.8	49 898.9
2003	135 822.8	17 381.7	62 436.3	56 004.7
2004	159 878.3	21 412.7	73 904.3	64 561.3
2005	183 217.4	22 420.0	87 364.6	73 432.9
2006	211 923.5	24 040.0	103 162.0	84 721.4
2007	249 529.9	28 095.0	121 381.3	100 053.5

§ 1.2 统计学

统计学是一门收集、整理和分析数据的科学，只要有数据的地方，就有统计的应用。收集数据，研究如何得到数据，与之对应的是统计学中的抽样调查和试验设计等理论；整理数据是指将数据用图或表的形式展现出来，与之对应的是描述统计的方法；分析数据是指选择适当的统计方法研究数据，并从数据中提取有用信息，进而得出结论，与之对应的是推断统计的理论与方法，包含了参数估计、假设检验、相关分析、回归分析、时间序列分析等诸多内容。

统计学的应用领域非常广泛，是一门适用于几乎所有学科领域的通用数据分析方法。无论是学术研究、政府管理，还是公司或企业的生产经营管理，都离不开统计方法的应用。表 1—3 列出了统计的一些应用领域，从中可以看到，统计无处不在。

表 1—3

统计的应用领域

Actuarial work (精算)	Hydrology (水文学)
Agriculture (农业)	Industry (工业)
Animal science (动物学)	Linguistics (语言学)
Anthropology (人类学)	Literature (文学)
Archaeology (考古学)	Manpower planning (劳动力计划)
Auditing (审计学)	Management science (管理科学)
Crystallography (晶体学)	Marketing (市场营销学)
Demography (人口统计学)	Medical diagnosis (医学诊断)
Dentistry (牙医学)	Meteorology (气象学)
Ecology (生态学)	Military science (军事科学)
Econometrics (经济计量学)	Nuclear material safeguards (核材料安全管理)
Education (教育学)	Ophthalmology (眼科学)



续前表

Election forecasting and projection (选举预测和策划)	Pharmaceutics (制药学)
Engineering (工程学)	Physics (物理学)
Epidemiology (流行病学)	Political science (政治学)
Finance (金融)	Psychology (心理学)
Fisheries research (水产渔业研究)	Psychophysics (心理物理学)
Gambling (博彩业)	Quality control (质量控制)
Genetics (遗传学)	Religious studies (宗教研究)
Geography (地理学)	Sociology (社会学)
Geology (地质学)	Survey sampling (抽样调查)
Historical research (历史研究)	Taxonomy (分类学)
Human genetics (人类遗传学)	Weather modification (气象改善)

资料来源：贾俊平、何晓群、金勇进：《统计学》，北京，中国人民大学出版社，2007。

当今世界，随着信息技术，特别是数据库技术的发展，社会各行业和领域收集、存储数据的能力有了很大的提高，同时也积累了大量的数据，这就为统计理论的进一步发展和在社会各领域的更广泛应用提供了机遇。在知识经济和信息时代里，统计学具有十分广阔的前景。

§ 2 一些基本概念

§ 2.1 随机性和规律性

许多社会现象都是随机的，带有不确定性。例如，买一注“双色球”福利彩票，能否中奖是不确定的；在犁好的土地中播下 1 000 颗种子，有多少会发芽也是不确定的。当然，也有许多社会现象是有规律的，比如人口密度高的地方犯罪率通常也比较高。

在现实中，随机性与规律性并非完全对立，社会现象通常是随机性和规律性的有机结合体，在随机之中带有规律性。抛一枚均匀的硬币，既可能出现正面，也可能出现反面，其结果是随机的。但是，在随机之中又有规律。如果重复抛该硬币若干次，则硬币正面出现的频率会随着重复次数的增加逐渐趋于 0.5 这个数值。图 1—2 记录了 500 次掷硬币试验中正面出现频率的波动情况。在重复次数较小时，该频率波动剧烈；随着重复次数的增加，该频率波动的幅度在逐渐减小，并逐渐向 0.5 靠拢。



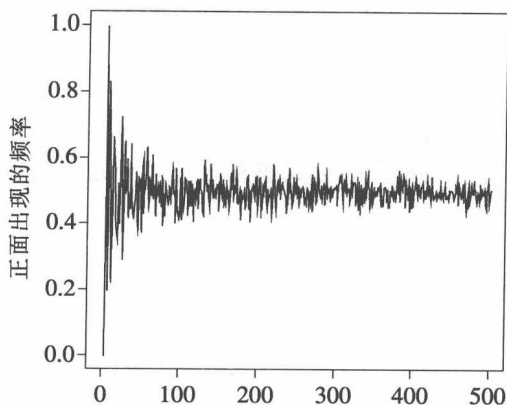


图 1—2 多次抛同一枚均匀硬币，正面出现频率的变化趋势

类似的例子还有很多：你很难预测股市在某一天是上涨还是下跌，但股市作为国民经济的晴雨表，在经济向好的那些年份，股票价格指数总的趋势是上涨的；人的身高很难预先确定，受生活习惯、基因、体育锻炼等因素的影响，存在较大的个体差异，带有一定的随机性。但从总体来说，我国公民的平均身高却是非常稳定的，并且随着生活水平的提高正在逐步增加。这都说明了随机之中有规律，这种规律称为统计规律。

对统计数据进行分析，就是利用数据产生的随机性和统计规律进行推断和决策。例如，要比较甲和乙两所高中的英语教学水平，可分别在这两所学校各随机抽取 30 名学生进行试卷测试。一般情况下，两个学校学生的成绩会有差异，这种差异有可能来自于抽取到水平不同的学生，也可能来自于两所高中英语教学水平的不同。在很多情况下，我们会面临不同背景的观测数据，需要研究这两组不同背景的数据是否来自于同一随机现象，即需要研究两组数据之间的差异是否大到超过随机性本身所能解释的地步，这种研究的理论是概率论。

§ 2.2 概率和机会

概率是对机会的描述，度量了某件事情发生的可能性，其取值在 0 和 1 之间（也可能是 0 和 1）。概率为 0 对应了那些绝对不可能发生的事情，比如在标准大气压下，水加热到 60°C 就沸腾，这是不可能的。概率为 1 的是那些一定会发生的事情，比如在真空状态下，自由落体在经过 t 秒钟后，落下的距离 s 必定是 $gt^2/2$ 。在现实中，概率为 0 或 1 的事件都比较少，绝大部分事件发生的概率都介于 0 和 1 之间，为随机事件。这样的例子有：

- (1) 随意抛掷一颗骰子，出现的点数为 6。



- (2) A、B 两队进行一场足球比赛，比赛结果为平局。
- (3) 在某交易日，上证指数以红盘报收。
- (4) 某对新婚夫妇生下的是一名男孩。
- (5) 某天出现雷雨天气。

通过对这些随机事件以概率的形式进行表述，可以为进一步的统计推断提供基础。

§ 2.3 参数和统计量

统计学经常涉及参数和统计量的概念。为研究某一问题，通常需要对研究对象进行界定。统计学将所要研究的全部个体（数据的集合）称为总体，将对其特征的一些概括性数字度量称为参数。例如，某一地区的平均受教育年限、一批袋装牛奶的合格品率等。

作为总体特征的概括性数字度量，参数的种类可以有很多，但研究者关心的主要有以下几种：总体平均数（记为 μ ）、总体方差（记为 σ^2 ）、总体比例（记为 π ）等。

在现实中，研究对象的范围虽然容易界定（有时也不容易），但要把这些研究对象相关特征的数据全部收集到，通常却面临着很大的困难，涉及时间、人力、物力、财力等诸多因素。例如，想要知道全国的总人口数，若逐个进行统计或实施普查，则花费的时间和费用是惊人的，现阶段只能是每十年进行一次人口普查。有时，所对应的研究对象可能是无限总体，即该总体所包括的元素是无限、不可数的，此时收集总体的全部数据变得根本不可行。例如在科学试验中，每一个试验数据可以看做总体的一个元素，而试验可以无限地进行下去。为此，很自然的一个想法是，我们不必去收集总体的全部数据，而是从总体中随机抽取一小部分元素的集合作为样本，再根据样本提供的信息来推断总体的特征。比如，在我国每隔五年进行一次千分之一的人口调查，从出厂的某批次灯泡中随机抽出少量的几个检测其寿命，等等。

与参数相对应，用来描述样本特征的概括性数字度量称为统计量，它根据样本数据计算得出，是样本的函数。常用的统计量和参数类似，主要有样本平均数（记为 $\bar{x}$ ）、样本方差（记为 s^2 ）、样本比例（记为 p ）等。

统计学中的绝大多数问题都是研究如何根据统计量去推断参数的问题。例如，如何用样本平均数（ $\bar{x}$ ）去估计总体平均数（ μ ），如何用样本方差（ s^2 ）去估计总体方差（ σ^2 ），如何用样本比例（ p ）去估计总体比例（ π ）等。图 1—3 形象地展示了这一过程：

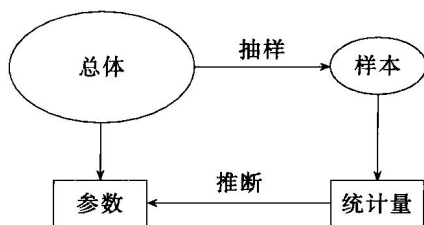


图 1—3 绝大多数统计问题的本质

§ 2.4 变量

1. 变量与变量的类型

变量与常量（也叫常数）相对，是说明随机现象某种特征的概念。例如，某个小学中每天上学的人数是不同的，可能今天多几个，明天少几个，因为有些人生病，有些人逃学，还有一些人因为其他原因没有来，因而上课学生人数就是一个变量，但学校登记注册的学生人数则是常量，是一个固定的已知数目。

事实上，所有随机取值的数据都归属于某个变量，是变量的某一次具体实现。数据可划分为定性数据和定量数据两类；类似地，变量的类型也区分为定性变量和定量变量。性别、受教育程度等都是定性变量，而商品销售额、职工工资等则属于定量变量。统计学所面对的是变量和变量之间的关系。

2. 变量之间的关系

事物是普遍联系的，作为对随机现象某种特征的表达，变量与变量之间同样存在着千丝万缕的联系。与区分数据的类型类似，这里同样需要区分不同变量类型之间的关系，因为研究不同变量类型之间的关系，往往对应着不同种类的统计模型。

按照变量所属类型的不同组合，可以将变量之间的关系分为定性变量之间的关系、定量变量之间的关系和定性变量与定量变量之间的关系。

性别与文化程度是否有关，不同国家的居民对陌生人的态度倾向是否存在差异，居民家庭订阅报纸和开通宽带上网之间是否有联系等，这些都属于定性变量之间的关系。研究定性变量之间关系的统计模型与方法主要有列联分析、对数线性模型等。

广告投入是否会影响销售额，城镇居民人均收入和人均支出之间的相互影响有多大，复习时间和考试成绩之间存在必然联系吗？这些都是定量变量之间的关系。研究定量变量之间关系的统计模型与方法主要有线性回归、非线性回归等。

手机品牌和手机销售量之间的关系，是否违约和信用卡用户年龄、月收入之

