**3.统计学学科性质：综合性的数量信息管理科学**

统计学是一门什么性质的科学，是属于方法论科学，还是属于实质性科学；是属于数学范畴，还是属于经济学范畴，或是属于管理科学或综合性科学，这是统计学科定位的重要问题。我们认为统计学是一门研究客观现象总体的具体时空条件下的数量信息的综合性的管理科学，它既是综合性科学，又是数量信息管理科学。

1.统计学是一门综合性科学。科学按照研究对象是属于自然现象还是社会现象，可分为自然科学和社会科学；同时，随着科学的发展和各门科学的相互渗透和交叉，又出现了综合性科学或交叉性科学。由于统计学研究现象总体的数量信息，既可以是社会现象总体的数量信息，也可以是自然现象总体的数量信息；既可以是确定性现象的数量信息，也可以是随机现象，模糊现象，突变现象等等现象的数量信息；既涉及到数量信息获取、处理、传递、存贮、分析和应用的一般理论和方法；又涉及到对具体现象的数量信息进行研究的专门理论和方法；既涉及到数量信息研究的基本原理和基本方法，又涉及到信息科学、信息技术、网络技术、系统科学、数学、管理科学等多学科知识在数量信息研究方面的应用。因此，统计学是一门综合性或交叉性很强的科学。

2.统计学是一门数量信息管理科学。统计学是研究客观现象总体的具体时空条件下的数量信息，从性质上讲，它显然不属于数学范畴，也不属于经济学范畴。统计学作为收集、处理和分析数量信息的科学与艺术，是属于信息管理科学范畴的，而不是属于诸如计算机科学与技术、网络技术等信息技术科学（硬科学）范畴的，是一门数量信息管理科学。信息资源管理理论认为，为了有效地利用信息，必须处理诸如信息搜集、组织、加工、存贮、传递、分析等程序化工作，以解决信息资源利用中的各种问题。信息资源管理涉及到信息资源开发、信息组织、信息整序、信息传递、信息交流、信息系统、信息咨询诸多方面的管理。统计信息是一种重要的数量信息资源，统计信息实践活动的过程既是数量信息开发与利用的过程，又是一种数量信息管理的过程；统计方法既是获取、处理、分析数量信息的方法，又是科学管理中经常运用的定量管理方法；统计不仅要做好数量信息的开发与管理，而且本身具有参与各项管理的功能。因此，将统计学定位于一门数量信息管理科学是符合逻辑的。但这并不意味着，它是属于工商管理、公共管理学科门类的，而是属于管理科学与管理工程门类的学科。

**4.试述经济统计与数理统计的关系**

在我国，由于特定的历史条件存在着两门并列而性质各异（一属社会科学，一属自然科学）的统计学，过去有过“统计学是一门还是两门”的讨论，如今意见也未能达到一致。依中国科学院院士陈希孺之见，此问题目前当存而不论，但在认同“大统计”主张的前提下，有必要明确二者的关系。

从数理统计学工作者的角度，看与社会经济统计工作者的合作，有利于扩大他们的研究范围和视野，从近年的发展趋势看，这一点尤其显得重要。

近年来有不少迹象表明：生命科学与社会经济科学将是未来数理统计方法的两个主要用武之地，我国以往的统计应用工作偏重于工业和国防方面，而以上两方面正是弱项，近年来生物统计（包括医学统计）在我国开始受到注意，也出现了我国数理统计学者与生物医学工作者联合研究的事例，但在社经统计问题研究方面情况还不理想，有待我国数理统计工作者的努力。

从根本上说统计学是研究变异的，群体内所含个体的性状各各不同，体现出变异只能从统计的观点去研究其群体的规律性，生物个体与社会中各单元的变异是固有的，并不因科学技术的进步而缩小（反之由测量工具的不精密而产生的测量值的变异，可随测量工具的改进而缩小，从而使统计方法的使用成为不那么必要），对包含极大数目个体的群体的研究，一般只能通过抽样来进行，样本的偶然性也是固有的，不随科技的进步而缩小，因此统计方法在这两个领域中的使用恒将是必要的，至少在可以预见的未来不会因时间的推移而减弱。

对社经统计学者而言，与数理统计工作者的合作有利于他们了解数理统计方法的作用与局限性，使他们能对数据作更深度的分析，有利于开展一些需要使用更专门和更复杂的数理统计工具的专题性研究。

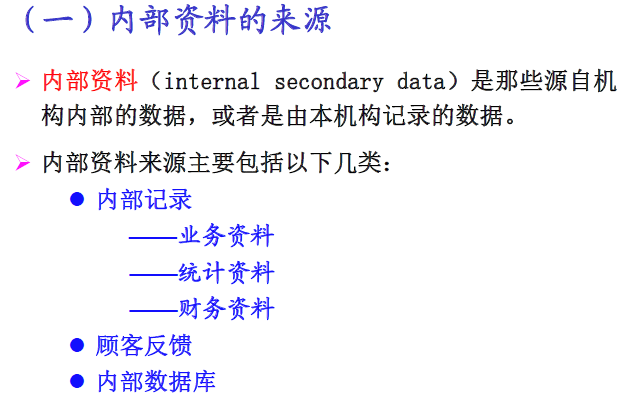
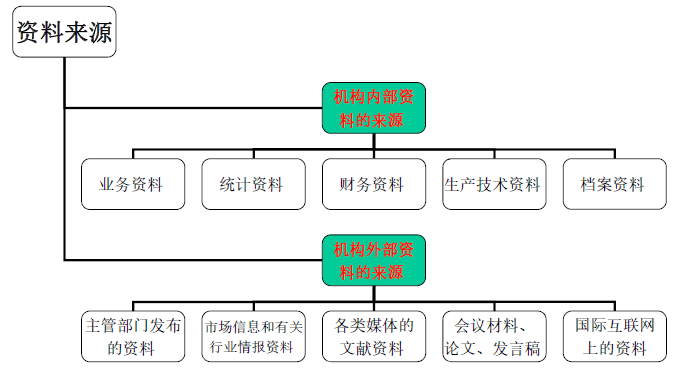
为实现这种合作，有一些组织上和技术性的问题，但在原则上不应有何障碍，因为数理统计方法是一种工具，它不构成一个学派因此也不存在学派间对立和竞争的问题，当然对一个特定问题是否适宜，或在哪一方面哪种程度上宜于使用数理统计学方法，可以有不同看法和主张，这可以通过实践和对比来解决，这也不影响数理统计方法是一种工具这个基本观点。

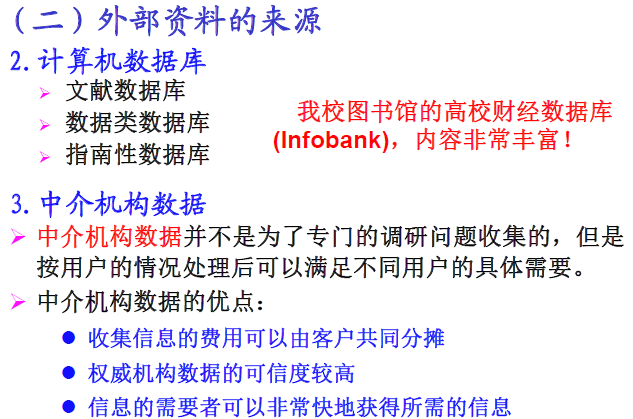
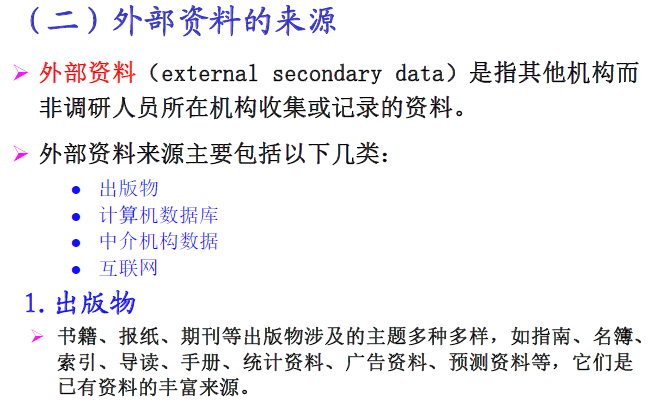
**（1）社会经济统计学和数理统计学的联系。**统计学本身起源于对社会经济问题的研究，最早可以追溯到原始社会末期。在奴隶社会制度的早起，需要针对奴隶社会的人口数量以及土地进行丈量与统计，虽然只是简单的登记和计数，但最初的社会经济统计学可以说已经出现了萌芽。之后，随着社会的不断发展以及经济的进步，人们开展了大量的社会经济统计活动，在资本主义社会时期，社会经济统计学的应用已经到达了一个巅峰。在资本主义环境之下，社会资源更加炙手可热，社会分工更加明确。随着人们社会实践活动的展开，在实践的过程中，社会经济统计学的使用也更加成熟。早期的概率论所研究的问题基本都来自于当时比较泛滥的赌博活动。17—19 世纪，不少数学家都对概率论的发展做出了贡献。数理统计学随着概率论的发展而迅速发展起来，虽然从时间上看，形成要晚于社会经济统计学，但发展飞速。数理统计并非完全独立于社会经济统计，它是在统计学的发展阶段中形成的一种分析数据的方法，社会经济统计学在分析问题时同样需要概率论与数理统计知识的支持。近代，数理统计学的发展势头迅猛，甚至有一些看法认为统计学几乎就是数理统计学。然而社会经济统计，作为对社会经济现象的一种调查研究活动，在社会发展中绝对有其存在的意义。有学者提出，社会经济统计学和数理统计学之间的关系与牛顿的力学和相对论力学的关系十分相似。相对论力学在接近光速时使用，而大多数情况是远离光速的，此时使用牛顿力学既准确又方便，社会经济统计学在描述变量时使用，数理统计学在描述随机变量时使用[3]。随机变量是随机现象下的变量，变量与随机变量的这种联系揭示了在一定条件下，社会经济统计学与数理统计学之间存在可以相互转化的关系。

**（2）社会经济统计学和数理统计学的相同之处。**社会经济统计学和数理统计学具有一定的相同点，都能够有效地针对客观的事物进行充分的统计，并且针对客观事物的发展趋势、发展规律进行研究。社会经济统计学和数理统计学两者在研究的方法上具有一定程度的共通性，都能够利用归纳、推理的研究手段分析问题，并针对问题提出相对客观，且具有建设性的解决建议。学术界在对于数理统计学进行表述时，都明确地说明了数理统计学是对随机现象的数量变化进行统计，并对研究对象进行规律性的研究和问题揭示。但是针对社会经济统计学来说，学术界对其的界定存在一定程度上的差异。一些学者认为，社会经济统计学是独立的社会学科，在应用的过程中，一般是将具体时间、具体地点、社会现象中所表现出的经济活动内涵进行阐述，有效地揭示其数量表现以及规律特点。另一部分学术工作人员认为，社会经济统计学应该被归纳到统计学当中，并在重大社会事件中详细地分析出事物发展的规律。在经济现象的积极引导之下，对数据进行收集、整理、分析。从长期的社会实践和社会发展的总体环境中来看，社会经济统计学和数理统计学两者的实际研究对象相同，并且两者都能够对统计规律进行详细的分析和探究。两者从研究对象的角度来看，都能够将某一人、某一事物、某一项目作为研究对象。研究对象还可以针对部分进行划分，分别是研究目标、研究客体，并能有效地分析出研究对象的客观发展规律。

**（3）社会经济统计学和数理统计学的不同之处。**社会经济统计学和数理统计学的区别与差异也是非常显著的，其主要的差异有以下几点：其一，社会经济统计学和数理统计学两者研究范围不同。针对社会经济统计学来说，一般是针对社会经济现状内容进行分析。针对数理统计来说，不仅可以对社会经济现象进行分析，而且还可以有效地针对自然现象进行数据分析处理。相较于社会经济统计学来说，数理统计学所涉及到的应用问题相对比较广泛。社会统计学虽然研究范围相对狭隘，但是社会经济当中所涵盖的内容非常广泛。从广义的层次上来说，社会经济不仅涵盖了人们的物质、精神、自然环境的再生产活动，而且社会经济统计学当中的各项内容又存在相辅相成、不可分离的特点。所以，在社会经济统计学在实际运用的过程中，还需要考虑到研究对象与物质、精神、自然环境等内容之间的关系。客观来说，社会经济统计学涉及到了人们日常生活的各个层次领域当中。而数理统计学一般是针对研究对象的自然现象进行研究，利用随机现象的手段，结合研究对象的实际情况，预测和体现出随机现象的可能性。其二，社会经济统计学和数理统计学的理论基础存在差异。客观来说，数理统计学的核心理论基础内涵便是概率论、统计推断理论。尤其是针对抽样推断来说，更是以概率论的大数法作为核心理论基础。总而言之，社会经济统计学和数理统计学作为我国社会学科的基础内容，都是在统计学的基础上所衍生出来的内容。客观来说，社会经济统计学和数理统计学两者在研究对象和研究方法的层面上，存在一定程度的相同之处，但是在研究范围以及理论基础层面当中，存在一定程度上的差异。在实际运用社会经济统计学和数理统计学的过程中，必须要清晰地认识到两者的优势和两者之间的区别，不能够将两者一概而论，更不能将两者进行分离。另外，社会经济统计学和数理统计学的研究范围，在一定的条件下其实是可以相互转化的。这对既有区别又有联系的学科，相信今后可以相辅相成，继续发展壮大。

**5.统计数据采集包括哪些方式？大数据时代的统计数据采集问题。**



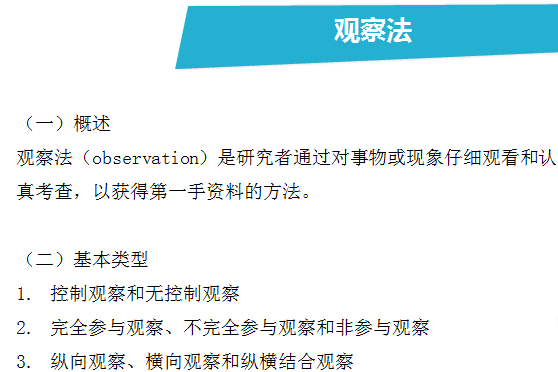




**汇报：**



**原始资料收集——观察法**



完全参与观察：观察者较长时期地生活在被观察者之中，甚至改变自己原有的身份。（卧底）

不完全参与观察：观察者不改变自己原有的身份，只在调查时置身于调查事项中取得资料。（生活体验）

非参与观察：观察者不参与调查活动之中，而以局外人的身份去观察事项发生和发展的情况。客观但无法深入（新闻调查）

完全参与观察：观察者较长时期地生活在被观察者之中，甚至改变自己原有的身份。（卧底）

不完全参与观察：观察者不改变自己原有的身份，只在调查时置身于调查事项中取得资料。（生活体验）

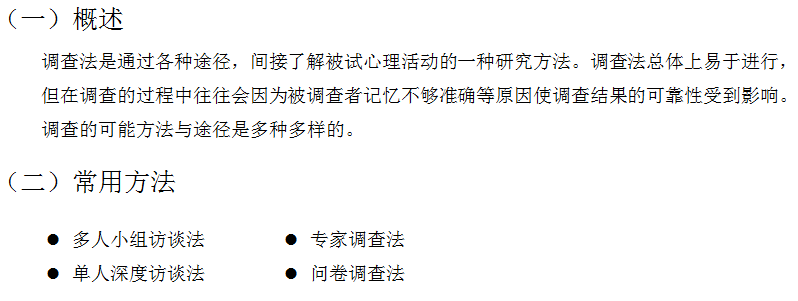
非参与观察：观察者不参与调查活动之中，而以局外人的身份去观察事项发生和发展的情况。客观但无法深入（新闻调查）

纵向观察：在不同时期加以观察，取得一连串的记录，能了解调查对象发展变化的过程和规律。

横向观察：在某一特定时间内对若干个调查对象所发生的事态同时加以记录，有利于不同对象间的对比分析。

纵横结合观察：对不同时期的多个调查对象同时进行观察、记录。

**原始资料收集——调查法**

多人小组访谈：

电话小组访谈：使用这种方法，被调查者会在一个特定的时间进行电话讨论。此时，无论他们在办公室、家里还是在任何地方，都可以通过电话进行访谈。

双向小组访谈：让目标小组观察另一个相关小组

电视会议小组访谈：这种方法是在配有摄像机和麦克风的会议室里进行的，委托方和调研人员不同在现场就可以通过视频终端观察到访谈的整个过程，委托方和调研人员还可以和主持人进行沟通，指明小组谈论的方向。

名义编组会议：根据目标消费者认为的重点问题进行研究，而不是让受访者讨论调研者所认为的重点，对于编制调研问卷和测定调研范围特别有用

在线小组访谈：是以互联网为媒介展开的焦点小组讨论。

单人深度访谈：

深度访谈是专业访谈人员和被调查者之间进行时间较长的（通常是30分钟到1小时），针对某一论题一对一方式的谈话。用以采集被调查者对某事物的看法，或做出某项决定的原因等。通常在被调查者家中或一个集中的访问地点进行。常用于了解个人是如何做出购买决策，产品或服务被如何使用，以及消费者生活中的情绪和个人倾向等。

深度访谈适合于并主要应用于实地研究。它的主要作用在于通过深入细致的访谈，获得丰富生动的定性资料，并通过研究者主观的、洞察性的分析，从中归纳和概括出某种结论。

正式访谈指的是研究者事先有计划、有准备、有安排、有预约的访谈。

非正式访谈指的是研究者在实地参与研究对象社会生活的过程中，随时碰上的、无事先准备的、更接近一般闲聊的交谈。

1、研究在自然情境下进行，通过较长时间（半小时至数小时）的访谈收集信息；

2、问题探察深入，通过连续询问鼓励访谈对象阐述、解释所作的回答；

3、研究思路主要是归纳方法，立足一手访谈信息进行归纳推论；

4、研究过程是一个递归循环过程。访谈问题常会在研究过程中变化、调整，数据分析与收集同时进行，结果推论是一个持续演化的过程。

专家调查法：

专家调查法是围绕某一主题或问题，征询有关专家或权威人士的意见和看法的调查方法。

这种调查的对象只限于专家这一层次。调查是多轮次的，每次都请调查对象回答内容基本一致的问卷，并要求他们简要陈述自己看法的理由根据。每轮次调查的结果经过整理后，都在下一轮调查时向所有被调查者公布，以便他们了解其他专家的意见，以及自己的看法与大多数专家意见的异同。

特征：

函询：用通讯方式反复征求专家意见，调查人与调查对象之间的联系是通过书信来实现的。

多向：调查对象分布于不同的专业领域，在同一个问题上能了解到各方面专家的意见。

匿名：调查对象通过调查组织者的整理，可以了解到其他专家的意见。但他们是背靠背的,不记名的,互不了解对方为谁。这有助于他们发表独立的见解。

反复：有控制地进行反馈的迭代,使分散的意见逐步趋向一致,以发挥集体智慧。

集中：用统计方法集中所有调查对象的意见，把每个专家的个人判断尽可能反映在最后归纳的集体意见中。从上述特点可知专家调查法是比较科学的，有广泛的用途，但是交换信件费时间，不能面对面讨论,所提问题很难提得很明确而不需要进一步解释,最后得出的一致意见具有一定程度的人为强制性。若与其他调查方法配合使用，就能取得更好的效果。

适用情况：

在下列3种典型情况下，利用专家的知识和经验是有效的，也是唯一可选用的调查方法。

数据缺乏：数据是各种定量研究的基础。然而，有时数据不足，或数据不能反映真实情况，或采集数据的时间过长，或者付出的代价过高，因而无法采用定量方法。

新技术评估：对于一些崭新的科学技术，在没有或缺乏数据的条件下，专家的判断往往是唯一的评价根据。

非技术因素起主要作用：当决策的问题超出了技术和经济范围而涉及到生态环境、公众舆论以致政治因素时，这些非技术因素的重要性往往超过技术本身的发展因素，因而过去的数据和技术因素就处于次要地位，在这种情况下，只有依靠专家才能作出判断。

问卷调查法：

含义：指调查者利用事先设计好的问卷向被调查者了解情况或征求意见，是一种书面调查方法。

功能：

将所需信息转化为被调查者可以回答并愿意回答的一系列具体问题。

引导被调查者参与并完成调查，减少由被调查者引起的计量误差。

使调查人员的提问标准化，减少由调查人员引起的计量误差。

记录和反映被调查者的回答，提供正确的信息。

根据问卷进行编码 ，便于资料的整理和分析。

优点：

可应用于多种调查方式，方便且富于弹性，可用于邮寄，也可用于当面访谈，也可用于网络调查。

时间经济，可以在同一时间内收集众多被调查者的资料。

可做匿名处理，所以具有隐秘性，被调查者作答的意愿较高。

可根据调研的需要设计各式各样的问题，适用于多项调研主题的调查。

被调查者可自由表达自己的意见。

可用量化的统计分析方法来呈现研究结果。

题目缺乏弹性，被调查者对问题的回答可能受问卷设计的影响。

用邮寄调查，问卷的回收率一般会偏低。

常用一个问题代表一个观点、意见或态度，可能会比较武断和片面。

对于不识字或受教育水平较低者，实施问卷调查相当困难。

**大数据采集**

1、离线采集：

工具：ETL；

在数据仓库的语境下，ETL基本上就是数据采集的代表，包括数据的提取（Extract)、转换(Transform)和加载(Load)。在转换的过程中，需要针对具体的业务场景对数据进行治理，例如进行非法数据监测与过滤、格式转换与数据规范化、数据替换、保证数据完整性等。

2、实时采集：

工具：Flume/Kafka；

实时采集主要用在考虑流处理的业务场景，比如，用于记录数据源的执行的各种操作活动，比如网络监控的流量管理、金融应用的股票记账和 web 服务器记录的用户访问行为。在流处理场景，数据采集会成为Kafka的消费者，就像一个水坝一般将上游源源不断的数据拦截住，然后根据业务场景做对应的处理（例如去重、去噪、中间计算等），之后再写入到对应的数据存储中。这个过程类似传统的ETL，但它是流式的处理方式，而非定时的批处理Job，些工具均采用分布式架构，能满足每秒数百MB的日志数据采集和传输需求

3、互联网采集：

工具：Crawler, DPI等；

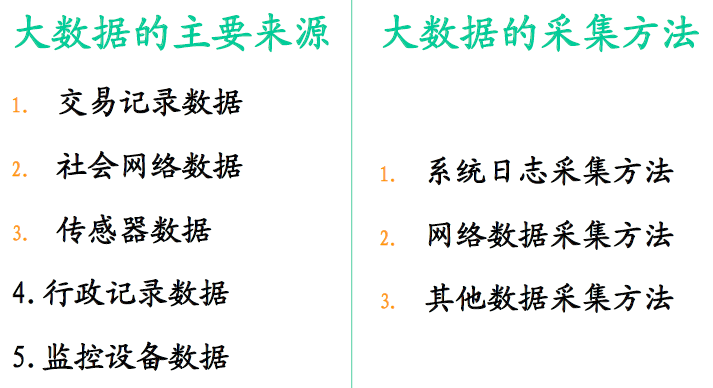
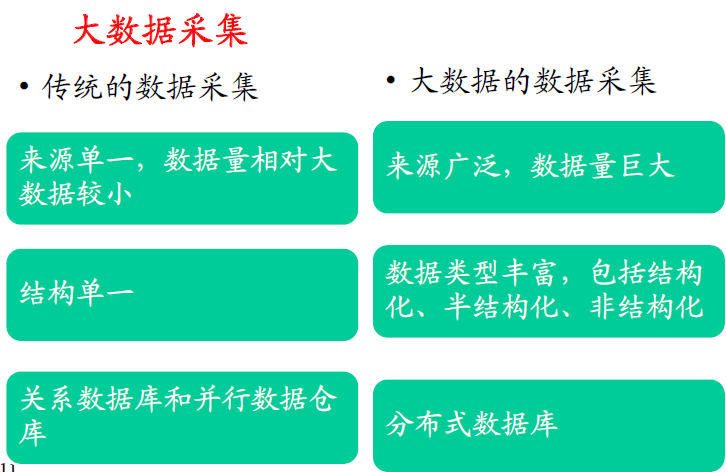
Scribe是Facebook开发的数据(日志)收集系统。又被称为网页蜘蛛，网络机器人，是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本，它支持图片、音频、视频等文件或附件的采集。

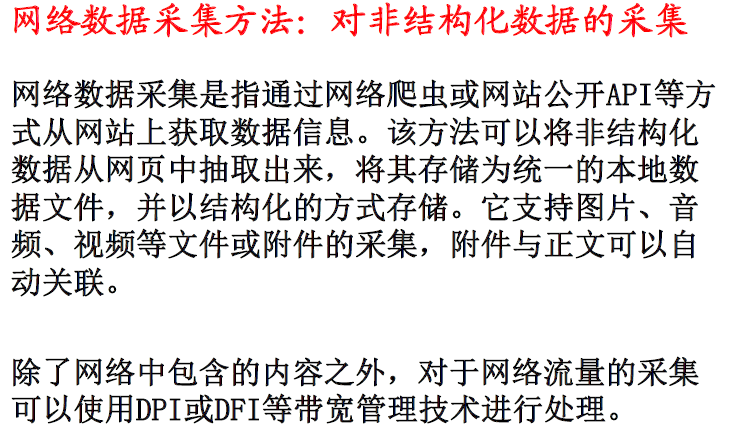
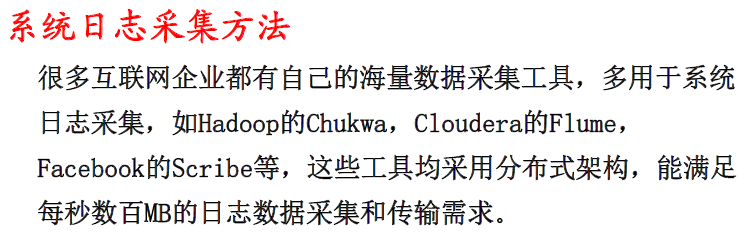
除了网络中包含的内容之外，对于网络流量的采集可以使用DPI或DFI等带宽管理技术进行处理。

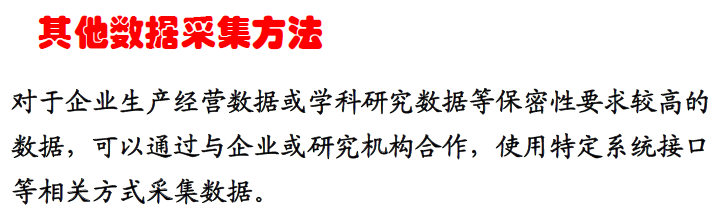
4、其他数据采集方法

对于企业生产经营数据上的客户数据，财务数据等保密性要求较高的数据，可以通过与数据技术服务商合作，使用特定系统接口等相关方式采集数据。比如八度云计算的数企BDSaaS，无论是数据采集技术、BI数据分析，还是数据的安全性和保密性，都做得很好。

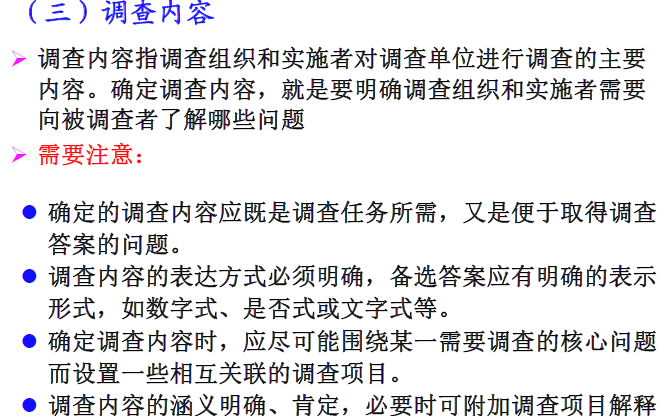
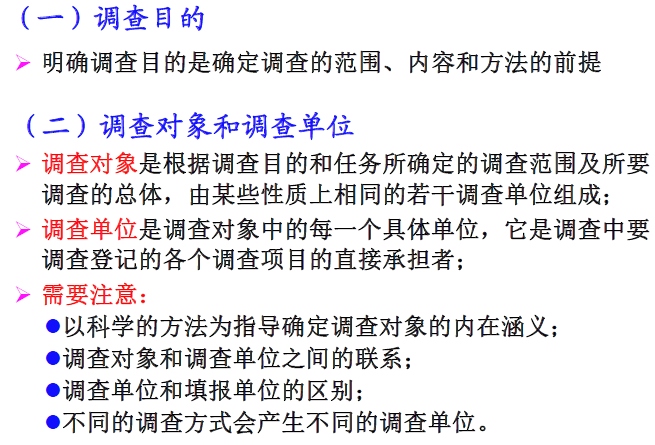
数据的采集是挖掘数据价值的第一步，当数据量越来越大时，可提取出来的有用数据必然也就更多。只要善用数据化处理平台，便能够保证数据分析结果的有效性，助力企业实现数据驱动。







**6.指出统计调查方案中需明确的最重要三项内容，加以详细分析。**



（1）调查目的是指社会调查研究活动所要达到的预期结果。也就是说为了什么进行调查和通过调查达到什么目标、发挥什么作用、解决什么问题。一般来说，进行调查研究的目的，或者是为了理论研究的需要，或者是为了解决和预测某一社会问题。调查目的在社会调查实践活动中具有十分重要的作用。首先，它是调查研究者进行社会调查活动的内在动机。正是这种目的性驱使人们去调查研究。其次，目的更是调查者进行调查活动的内在尺度。调查活动究竟如何进行，如何设计方案，选用什么方法，都要根据目的来确定。再次，目的是整个调研活动的“统帅”和目标。作为“统帅”，它使整个调研活动过程成为一个有机的系统，各个阶段、各个环节、各个步骤具有高度的严密的组织性；作为目标，目的对整个调研活动具有调节作用，以避免发生大的偏差。

在设计调查研究方案时，关于调查目的的阐述，至少应该包括这样一些问题：为什么要进行该项调研?调研要解决一些什么问题?解决到何种程度?用什么形式来反映调研的最终成果?调研的意义、价值何在?将起到怎样的社会作用?同时，还应写明将本着怎样的指导思想来进行调查研究。

（2）调查对象亦称“调查总体”。需要进行调查的总体范围。由许多性质相同的调查单位所组成。确定调查对象，要明确总体的界限，划清调查的范围，以防在调查工作中产生重复或遗漏。如，调查的目的是为了搜集某地区国有及国有控股企业经营活动情况的资料，则调查对象就是该地区所有国有及国有控股企业。 [1]

调查对象指接受调查的社会现象的总体。调查对象由性质相同的各个调查单位组成。例如，要调查研究全国运输企业的运输周转量、成本、燃料消耗、劳动生产率情况，则全国所有运输企业就是调查对象。确定调查对象的关键在于科学地确定调查对象的定义，明确地规定接受调查的总体的范围与界限。只有这样，才能避免因界限不清而导致调查登记的重复或遗漏，保证调查资料的准确性。

在某项调查中登记其具有特征的单位，即调查项目的承担者，是调查对象的具体单位。

**8、你如何理解空间计量经济学方法与传统计量经济学方法的本质区别？**

计量经济学是以一定的经济理论和统计资料为基础，运用数学、统计学方法与电脑技术，以建立经济计量模型为主要手段，定量分析研究具有随机性特性的经济变量关系的一门经济学学科。主要内容包括理论计量经济学和应用经济计量学。理论经济计量学主要研究如何运用、改造和发展数理统计的方法，使之成为经济关系测定的特殊方法。应用计量经济学是在一定的经济理论的指导下，以反映事实的统计数据为依据，用经济计量方法研究经济数学模型的实用化或探索实证经济规律。

空间计量经济学是计量经济学的一个分支，研究的是如何在横截面数据和面板数据的回归模型中处理空间相互作用（空间自相关）和空间结构（空间不均匀性）结构分析。它与地学统计和空间统计学相似。从某种程度上而言，空间计量经济学与空间统计学之间的不同和计量经济学与统计学之间的不同一样。由于对其理论上的关心以及将计量经济模型应用到新兴大型编码数据库中的要求，近年来这个领域获得了快速发展。空间数据分析和建模技巧与GIS（地理信息系统）的结合，现已广泛应用于经济政策分析中，尤其是实产和房地产经济、环境和资源经济、发展经济。当面临空间自相关时，标准的计量分析技巧通常会失效，而这种情形经常在地理或横截面数据集中出现，这也是空间计量得以迅速发展的原因之一。

经典经济计量模型，基于Gauss-Markov假设，假定变量相互独立，这一假定与社会经济事物间存在着广泛、普遍联系的现实并不吻合。近30年来，随着计算机技术的发展，引入空间效应的空间经济计量分析已经成为经济计量领域的一个重要分支和研究热点之一。简单地说，空间效应的存在是空间经济计量学作为一门独立学科产生的社会经济现实基础，在经济计量分析中引入空间效应是空间经济计量模型与经典经济计量模型的根本区别。

**10、地理加权模型与分位数回归模型存在什么样的区别与联系？**

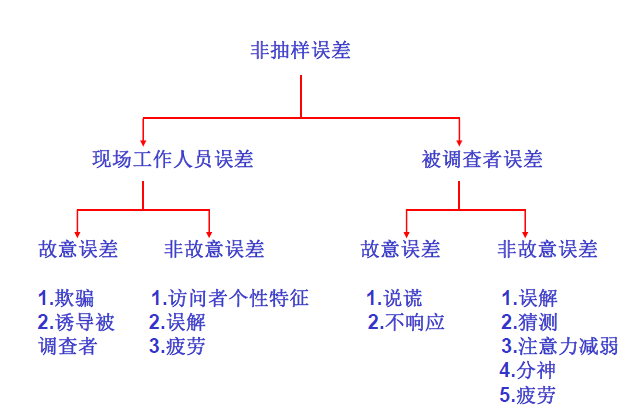
地理加权回归是一种非参数估计方法，由Brunsdon等在局部加权最小二乘法的基础上提出。该方法能有效处理空间异质性问题，应用领域日趋广泛，已涵盖了空间经济学、生物学、房地产经济学等多个学科，现成为空间经济计量研究关注的热点领域之一。

分位数回归是回归分析的方法之一，最早由Roger Koenker和Gilbert Bassett于1978年提出。一般地，传统的回归分析研究自变量与因变量的条件期望之间的关系，相应得到的回归模型可由自变量的估计因变量的条件期望；分位数回归研究自变量与因变量的条件分位数之间的关系，相应得到的回归模型可由自变量估计因变量的条件分位数。相较于传统回归分析仅能得到因变量的中央趋势，分量回归可以进一步推论因变量的条件概率分布。

**12.统计调查现场可能产生什么样的数据质量问题？如何解决？**

在数据收集过程中的误差可以分为两大类：第一大类是现场工作人员误差，即现场工作人员进行问

卷调查时所产生的误差，主要是指调查者误差。另一类是由被调查者造成的被调查者误差。当然此类误差的产生与采用何种数据收集方式无关。这两类都属于非抽样误差的范畴，对于每一大类又可作如下区分：故意误差和非故意误差，前者是有意造成的，后者则是无意中产生的。



(一) 现场工作人员故意误差

只要现场数据收集者故意违反调研人员的数据收集要求，就会产生现场工作人员故意误差；

1.访问者欺骗。当访问者故意谎报被调查者的情况时，就产生了访问者欺骗。因为可能访问员的报酬是与完成的问卷数量计算的。以完成的调查表的份数给付报酬的制度具有一定的合理性。访问者不像生产线工人，譬如商场调研者，他们可能很多时间段内无事可做，这要视购物者流量和被调查者是否符合要求而定，现场工作人员的酬金较低，工作时间较长，工作常常令人沮丧，这样就会诱使他们伪造调查问卷，有些访问者对这种诱使表示屈从。

2.诱导被调查者误差。

表现为访问者在遣词、评阅或形体与语言方面影响被调查者回答，最坏的情况是访问者可能改变问题的用词以造成诱导，例如：不说“你关心节约用电吗”，而说“难道你不关心节约用电吗”，从而影响被调查者的回答。还有其他不太明显的诱导的例子。

(1)一种方法是用比较微妙的方式向被调查者表明访问者所期望的回答。例如：一位被调查者对问题回答“是”，访问员可能对她说，“我知道你将回答是，因为超过90％的被调查者对此表示赞同”。这样的一句评论在被调查头脑中形成了一种观念，他或她应该继续和大多数人保持一致。

(2)另一种微妙的诱导来自访问者的暗示。例如：面对面访问中，访问者在提问时，可能对自己不赞同的问题微微摇头，而对赞同的问题则微微点头。被访问者则根据访问者头部运动所提供的信号去迎合访问者期望的回答。电话访问中可能使用“嗯”或“好“来表示不赞同和赞同，如果这种反方式一直持续下去，访问者就会很微妙地影响被访问者。

(二)现场工作人员非故意误差

当访问者明明犯了错误却认为自己正确操作时，就产生了非故意误差。包括三类：访问者的个人性格特征、访问者的误解和疲劳。

（1）个人特性可以在访问者的口音、性别及举止行为中发现。口音称为“交流干扰”，访问者可能对口音产生定向思维，例如：在美国波士顿口音意味着上层阶层，得克萨斯就意味着大牧场和油井。

（2）访问者误解是指访问者自认为他知道怎样进行一次调查，但实际上操作上并不正确的情况。设计调研问卷的调研员和执行调查的访问者在受教育水平方面有很大的差异，这些差异会导致调查问卷上的说明可能会使访问者摸不着头脑，虽然访问者努力做到符合调研者的心愿，但也难免会有些不尽人意。

（3）第三类访问者非故意误差是疲劳误差，往往发生在访问者身心疲劳之时。因为访谈是一项重复性的工作，访问者容易产生疲倦感。访问者或误跳读了问题，或忘记检查被访问者的回答，或草草完成问卷的某一部分，或表现出厌烦，使得那些本愿接受调查者以拒绝参与访问而告终。

(三)被调查者故意误差

1.说谎。调查过程中，被调查者也许感到窘迫，也许想保护个人隐私，甚至怀疑访问者还有某种动机。例如收入水平是一个敏感话题，对单身妇女暴露婚姻状况是其担忧所在，某些人的年龄也是敏感话题；节约能源是一个道德话题，很多人知道应该节约能源，所有有些人会说谎，以免被人认为是在浪费或没有爱国心。

2.不响应。是指预期的被调查者不参与调查或不回答调查问卷中的某些特定问题。实际上，各种类型的不响应是调研人员所遇到的最常见的被调查者故意误差。

(四)被调查者非故意误差

1.被调查者误解。是指被调查者在没有理解问题或没有按照要求的情况回答问题。例如：问题是税前收入而很可能想得的是净收入。

2.猜测误差。是指在没有确切把握的情况下回答了问题。有些情况下，被调查者会被问到一些他们知之甚少或已遗忘的问题，但他们觉得有必要提供回答，这时就会猜测答案。所有的猜测都可能包含有错误。如你上个月用了多少度电？

3.注意力减弱。是指被调查者对调查的兴趣降低。一般的被调研者并不像调研者一样对调研感兴趣，而且一般感觉到参与调查的能力越来越弱。

4.干扰误差。问卷过程中因干扰而产生的误差，如购物中心调查时很可能遇到熟人而打招呼导致分神，电话访问时还要照看身边的小孩子，买东西时孩子在身边着急要走等等。

5.被调查者疲劳。指被调查者对参加调查感到厌烦，反应能力就会降低，为了快点答完问卷，很多时候选择“没有意见”来回答。

(五)现场数据收集质量控制

有很多办法来降低前面谈到的各类误差，但注意是“降低”而不是“消除”。

1.对现场工作人员故意误差的控制。有两种一般性策略方法来防止访问者故意犯错误：

（1）监督。是指派管理人员对现场工作人员的工作进行监督，如电话访问时采用监听。在私人访谈中，监督者可以跟随访问者，在管理现场调查问卷时观察访问者的行为。

（2）证实。是指对访问者的工作进行核实。这是针对篡改和欺骗问题的，一种方法是监督者重新与被调查者取得联系来证明他们是否接受了调查；另一种方式是检查问卷，以发现访问者问卷中的某种模式，这种模式会引起监督者对其是否做假的怀疑，那些伪造调查问卷的访问者不可能总是小心地模仿真正的被调查者而做到天衣无缝。

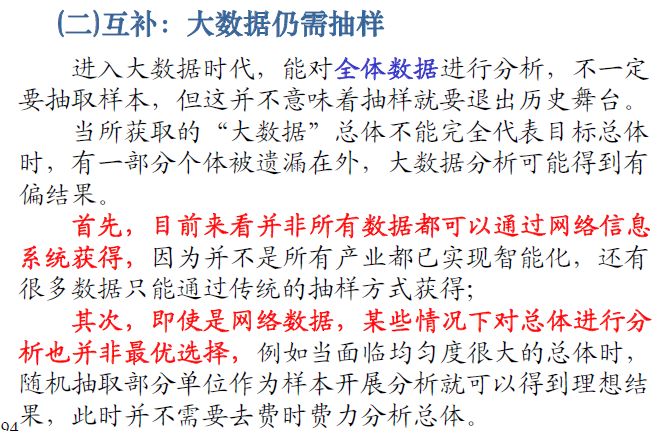
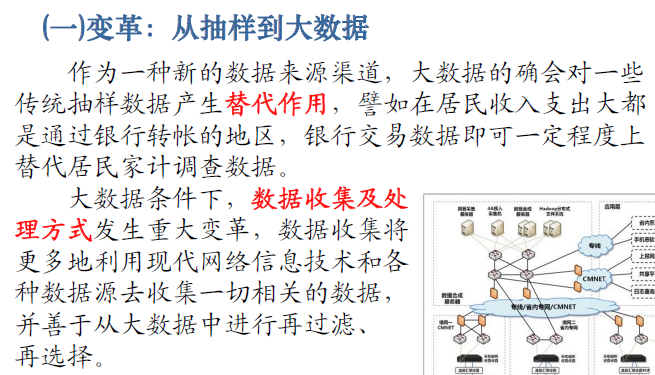
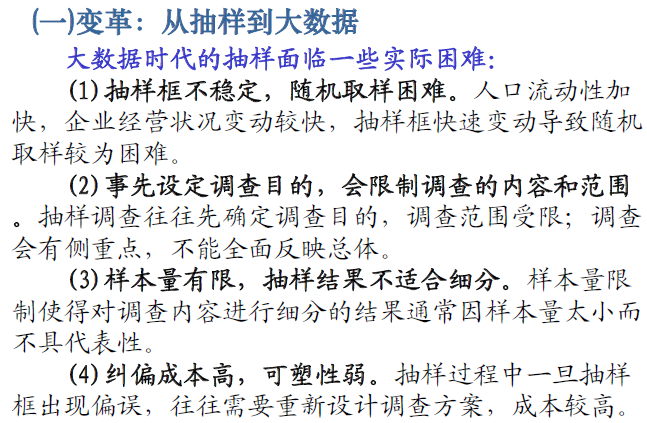
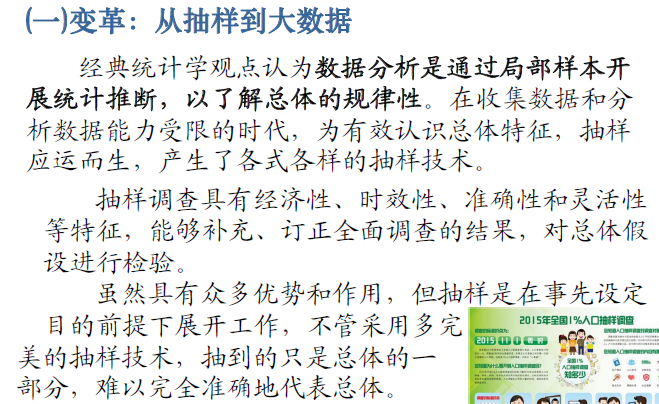
2.对现场工作人员非故意误差的控制。监督者或委托方可以进行选择、培训、角色训练，或要求每进行一定数量或一定时间的调查后休息一下等。

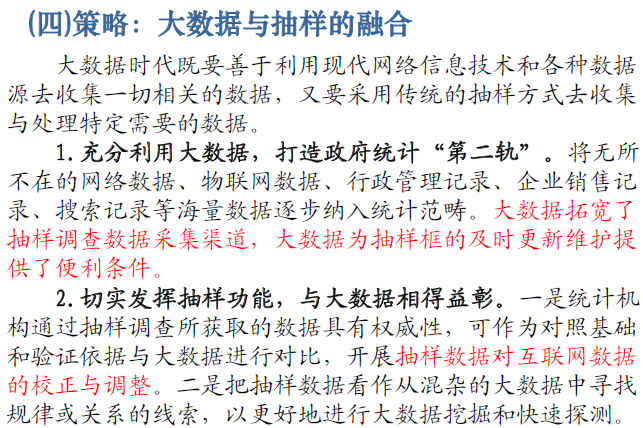
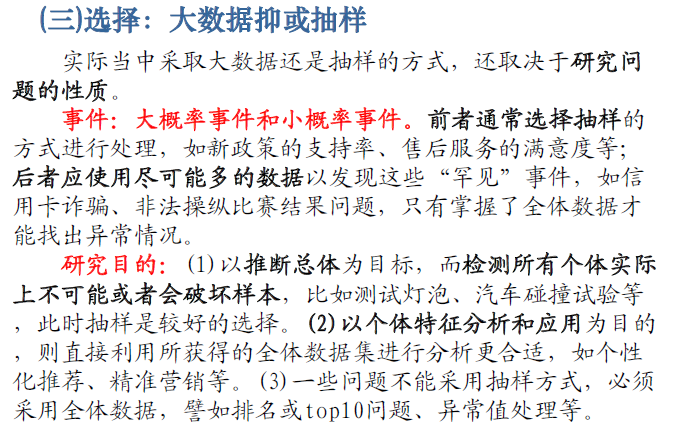
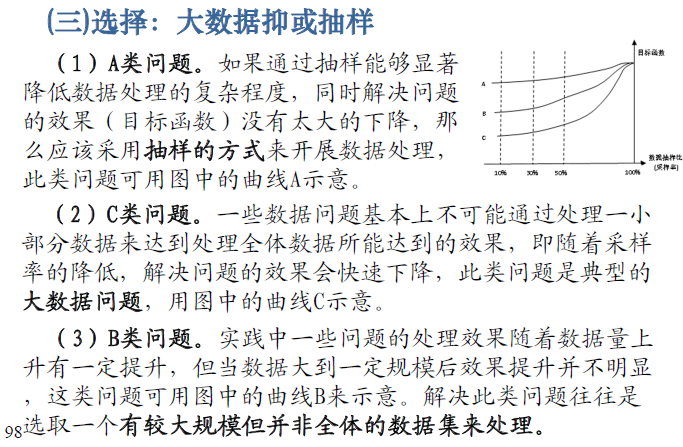
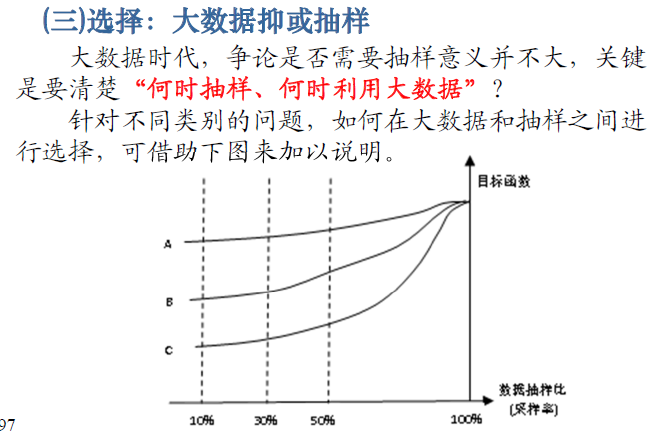
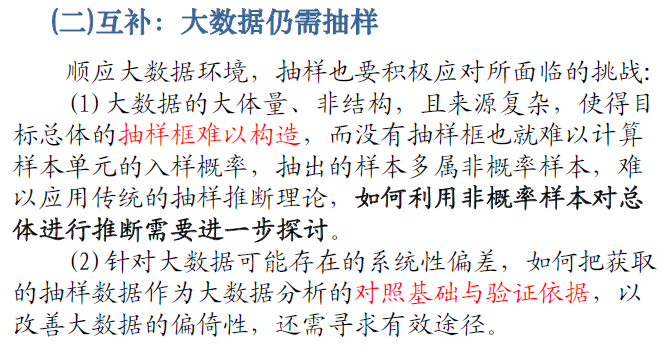
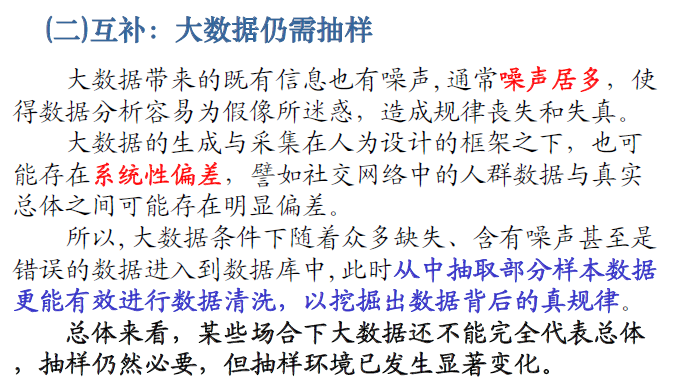
3.对被调查者故意误差的控制。可以采用的方法有：

(1)控制说谎。主要包括：匿名――是指向被调研者保证他们的名字不会与他们的回答联系在一起。保密――是指向被调查者保证他们的回答将不予公开。激励――是向被调查者许诺在其参与调查的情况下将给予现金支付、提供礼物或其他有价值的东西。证实检查――即对被调查者所提供的信息进行确认，如一个关于治疗脱发药品的家庭调查中，可以要求看一下被调查者的药品以求得证实。

(2)控制不响应。主要包括：物质激励。允许不回答。即使没有发生不响应，有一些问题的响应率也会低于其他问题，事实上，如果能够预料到某些问题可能会被拒绝，那么在问卷中加上“可拒绝回答”是明智的。当然在自填问卷中，没有设计这一选项可以提供准确答案的情况下，增添这样的设计是不明智的，因为有些被调查者利用这一选项来达到推诿的目的。

4.对被调查者非故意误差的控制可以采用的形式多样：周密设计的问卷说明和范例是一种常用的避免被调查者混淆的方法，可以减少误解。对于长的问卷可采用一些提示语，如我们快完成了，现在是最难回答的问题，或策略性地安排一些措辞来鼓励被调查者继续接受调查。

**13.你认为大数据条件下还需要统计调查吗？说明理由。**



**14.数据预处理过程中如何解决缺失值问题？**

缺失数据的产生机制通过探讨缺失数据的出现与目标变量是否有关而界定, 如果缺失数据是随机出现, 就将缺失数据产生机制定义为可忽略的, 如果缺失数据的产生与研究变量有关, 则称之为不可忽略的。对缺失数据的处理方法大体可以分为四类:

1 .忽略。若一条记录中有属性值缺失, 则将该条记录被排除在数据分析之外。该方法简单易行,但是容易导致严重的偏差, 仅适用于含有少量缺失数据的情况。

2 .插补(替代)。基本思想是利用辅助信息, 为每个缺失值寻找替代值。具体可采用以下几种策略:(1)使用一个固定的值代替缺失值:所有缺失值用一个常量代替, 譬如用字母“ N”代替缺失值。当某一属性的缺失值较多, 使用此方法可能导致结果出现偏差, 也只适合于缺失值不多的情况。(2)使用均值代替缺失值:对同一属性的所有缺失值都用其平均值代替。根据变量特征在简单及加权算术平均数、中位数、众数中选用合适的平均数, 尽量使替代值更接近缺失值, 减少误差。(3)使用同一类别的均值代替缺失值:对数据按某一标准分类, 分别计算各个类别的均值来代替相应类别的缺失值, 不同类别的均值可选用不同形式的平均数。(4)使用成数推导值代替缺失值:若同一属性的记录值只有少量几种, 可计算各种记录值在该属性中所占比例, 并对缺失值同比例赋值, 该方法较适合缺失属性为是非标志的情况。(5)使用最可能的值代替缺失值:利用回归分析、决策树或贝叶斯方法等建立一个预测模型, 利用模型的预测值代替缺失值。该方法相对复杂, 但能够最大程度地利用现存数据所包含的信息。

3 .再抽样。包括以下三种情况:(1)多次访问。对无回答单位进行再次补充调查, 尽可能多地获得调查数据。如果缺失数据是在不可忽略机制下产生, 由于积极回答者和不积极回答者之间的数量特征有较大差异, 多次访问很有必要, 且差异越大, 访问次数也需相应增加。(2)替换被调查单位。在出现无回答的情况下, 为使样本量不低于原设计要求,

补救方法之一是实行替换, 用总体中最初未被选入样本的其他单位去替代那些经过努力后仍未获得回答的单位, 替换时应尽可能保证替代者和被替代者的同质性。(3)对无回答进行子抽样。当后续访问的单位费用昂贵时, 子抽样可作为减少访问次数的一种现成方法。

4 .加权调整。基本思想是利用调整因子来调整包含缺失数据所进行的总体推断, 将调查设计中赋予缺失数据的权数分摊到已获取数据身上。该方法的前提上缺失数据在可忽略机制下产生, 即已获得数据与缺失数据之间没有显著差异, 主要用于单位数据缺失情况下的调整。

**15.试述统计分布或分类或代码在社会经济统计研究中的意义**

统计分布的意义：统计分布亦称“次数(频数)分布(分配)”。在统计分组的基础上，将总体中的所有单位按组归类整理，形成总体单位在各组间的分布。分布在各组中的单位数叫做次数或频数。各组次数与总次数(全部总体单位数)之比，称为比率或频率。将各组别与次数依次编排而成的数列就叫做统计分布数列，简称分布数列或分配数列。它可以反映总体中所有单位在各组间的分布状态和分布特征，研究这种分布特征是统计分析的一项重要内容。统计分布及其分布数列，可以用表格或图形来表示。

在统计分组的基础上，把总体的所有单位数按组归并排列，形成各组单位数在总体中的分布，称统计分布。统计分布的实质是，把总体的全部单位按某标志所分的组进行分配所形成的数列，也可称为分配数列或分布数列。在每次把某个单位分配到某一组时，人们常常说分配了一次，所以，分配数列又叫次数分布。分配数列有两个构成要素：一是总体按某标志所分的组：二是各组对应的单位数——次数。

统计分布形式十分简单，但在统计研究中却有着重要的意义。统计分布是统计分析结果的一种重要表现形式，也是统计分析的一种重要方法。它可以表明总体各单位的分布特征和结构状况，并有助于我们进一步研究标志的构成、平均水平及其变动规律。从文字含义看，统计分布理论性强一些，分配数列更通俗一点。

分类的意义：（1）划分类型。统计分组是确定社会经济现象各种类型的基础，例如将工业企业按所有制的不同、按轻重工业划分，居民按城镇、农村划分，从而说明不同的经济类型的特点。一般来说，社会经济类型的分组多采用品质标志来划分。（2）反映内部结构通过统计分组可以反映总体内部各部分之间的差别和相互关系，表明总体的内部结构。同时在各组的基础上计算各组所占总体的比重，从总体的构成上认识总体各部分的作用，并对总体作出正确的评价。业人员的分布情况，通过分组表明了从业人员在三次产业中的分布，也显示了人员在三次产业中的结构比重，说明这10年间中国的产业结构发生了很大的变化。（3）依存关系社会经济现象之间存在着相互制约、相互联系的关系，通过统计分组可以根据现象间的影响因素和结果因素的对应更好地揭示现象之间的这种依存关系。

**16、某种数值平均数（如幂平均）的性质研究及其统计学意义。**

算数平均数：算术平均数，又称均值，是统计学中最基本、最常用的一种平均指标，分为简单算术平均数、加权算术平均数。它主要适用于数值型数据，不适用于品质数据。根据表现形式的不同，算术平均数有不同的计算形式和计算公式。算术平均数是加权平均数的一种特殊形式（特殊在各项的权重相等）。在实际问题中，当各项权重不相等时，计算平均数时就要采用加权平均数；当各项权重相等时，计算平均数就要采用算术平均数。

简单算术平均：适用：主要用于未分组的原始数据。设一组数据为X1，X2，...，Xn，简单的算术平均数的计算公式为：https://bkimg.cdn.bcebos.com/pic/5882b2b7d0a20cf40b00fd2f77094b36adaf99a9

加权算术平均：适用：主要用于处理经分组整理的数据。设原始数据为被分成K组，各组的组中的值为X1，X2，...，Xk，各组的[频数](https://baike.baidu.com/item/%E9%A2%91%E6%95%B0" \t "_blank)分别为f1，f2，...，fk，加权算术平均数的计算公式为： https://bkimg.cdn.bcebos.com/pic/6c224f4a20a44623f523ae929922720e0df3d751

特点

1.加权算术平均数同时受到两个因素的影响，一个是各组数值的大小，另一个是各组分布频数的多少。在数值不变的情况下，一组的频数越多，该组的数值对平均数的作用就大，反之，越小。

频数在加权算术平均数中起着权衡轻重的作用，这也是加权算术平均数“加权”的含义。

2.算术平均数易受极端值的影响。例如有下列资料：5、7、5、4、6、7、8、5、4、7、8、6、20，全部资料的平均值是7.1，实际上大部分数据（有10个）不超过7，如果去掉20，则剩下的12个数的平均数为6。

由此可见，极端值的出现，会使平均数的真实性受到干扰。

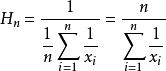
统计学意义：

1.算术平均数是一个良好的集中量数，具有反应灵敏、确定严密、简明易解、计算简单、适合进一步演算和较小受抽样变化的影响等优点。

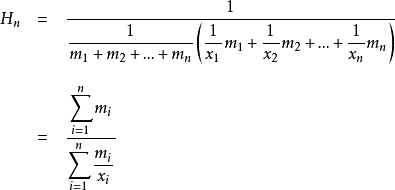
2.算术平均数易受极端数据的影响，这是因为平均数反应灵敏，每个数据的或大或小的变化都会影响到最终结果。

调和平均数：调和平均数又称倒数平均数，是总体各统计变量倒数的算术平均数的倒数。调和平均数是平均数的一种。但统计调和平均数，与数学调和平均数不同，它是变量倒数的算术平均数的倒数。由于它是根据变量的倒数计算的，所以又称倒数平均数。调和平均数也有简单调和平均数和加权调和平均数两种。

在数学中调和平均数与算术平均数都是独立的自成体系的。计算结果前者恒小于等于后者。因而数学调和平均数定义为：数值倒数的平均数的倒数。但统计加权调和平均数则与之不同，它是加权算术平均数的变形，附属于算术平均数，不能单独成立体系。且计算结果与加权算术平均数完全相等。 主要是用来解决在无法掌握总体单位数（频数）的情况下，只有每组的变量值和相应的标志总量，而需要求得平均数的情况下使用的一种数据方法。

。简单调和平均数是算术平均数的变形，它的计算公式：

。加权调和平均数是加权算术平均数的变形。它与加权算术平均数在实质上是相同的，而仅有形式上的区别，即表现为变量对称的区别、权数对称的区别和计算位置对称的区别。因而其计算公式为：



加权调和平均数的应用: 在很多情况下，由于只掌握每组某个标志的数值总和（M）而缺少总体单位数（f）的资料，不能直接采用加权算术平均数法计算平均数，则应采用加权调和平均数。

调和平均数具有以下几个主要特点：

①调和平均数易受极端值的影响，且受极小值的影响比受极大值的影响更大。

②只要有一个标志值为0，就不能计算调和平均数。

③当组距数列有开口组时，其组中值即使按相邻组距计算，假定性也很大，这时的调和平均数的代表性很不可靠。

④调和平均数应用的范围较小。在实际中，往往由于缺乏总体单位数的资料而不能直接计算算术平均数，这时需用调和平均法来求得平均数。

注意事项

（1）当变量数列有一变量X的值为零时，调和平均数公式的分母将等于无穷大，因而无法求出确定的平均值。

（2）调和平均数和算术平均数一样，易受两极端值影响。上端值越大，平均数向上偏离集中趋势就越大。反之，下端值越大，平均数向下偏离集中趋势越大。

（3）要注意区分调和平均数和算术平均数的使用条件，因事制宜。

几何平均数：对各变量值的连乘积开项数次方根。求几何平均数的方法叫做几何平均法。如果总水平、总成果等于所有阶段、所有环节水平、成果的连乘积总和时，求各阶段、各环节的一般水平、一般成果，要使用几何平均法计算几何平均数，而不能使用算术平均法计算算术平均数。根据所拿握资料的形式不同，其分为简单几何平均数和加权几何平均数两种形式。

简单几何平均数、加权几何平均数https://gss3.bdstatic.com/7Po3dSag_xI4khGkpoWK1HF6hhy/baike/pic/item/c2cec3fdfc039245790dc69d8b94a4c27c1e25e3.jpg

特点：1、几何平均数受极端值的影响较算术平均数小；2、如果变量值有负值，计算出的几何平均数就会成为负数或虚数；3、它仅适用于具有等比或近似等比关系的数据；4、几何平均数的对数是各变量值对数的算术平均数。

主要用途：计算几何平均数要求各观察值之间存在连乘积关系，它的主要用途是： 1、对比率、指数等进行平均；2、计算平均发展速度；其中：样本数据非负，主要用于对数正态分布。3、复利下的平均年利率；4、连续作业的车间求产品的平均合格率。

**18、关于抽样调查的组织形式分类与某种相应组织形式的抽样估计**

在进行抽样调查工作时，必须根据所研究总体本身的特点和抽样调查的目的要求，对抽取样本的操作程序和工作方式，进行周密的设计和安排，在抽样技术上称做抽样调查的组织方式，即抽样的实际工作方法。

由于抽样调查的目的不同，调查对象和特点不同，因此，抽样调查的组织方式也不同，在实际工作中，常用的抽样调查组织方式有：简单随机抽样、分层抽样、等距抽样、整群抽样和阶段抽样等。以下做简耍介绍。

1．简单随机抽样

简单随机抽样是按随机原理直接从总体N个单位中抽取n个单位作为样本，也称为纯随机抽样。前面介绍抽样技术的基本原理都是以简单随机抽样为基础的，它是抽样理论中最基本、最单纯的抽样组织方式。其他抽样方式都是以它的随机抽样原则为依据，采取分类、排队或分群等形式，来安排操作程序所形成的不同抽样组织方式。

简单随机抽样较适用于以下几种情况。

(1)总体单位分布比较均匀，各单位变量值差异不大。

(2)总体单位数较少，各单位排列五次序。

(3)抽到的样本单位数较分散时，不影响调查效果。

2．分层抽样

分层抽样又叫类型抽样。它是先对总体各单位按主要标志加以分组，其后再从各组中按随机原则抽取一定单位构成样本的抽样方式。例如，在农产品产量调查中按地势分为山区、丘陵、平原三类：在职工生活调查中按部门分为工业、商业、交通、文教等部门，然后在分类基础上，再抽取样本单位。

先按一定标志分组，再按各组频数占总体频数的比重来分配抽样数目，可使样本变量值的分布结构与总体变量值的分布结构完全一致，提高样本的代表性，使抽样误差进一步缩小。因此，类型抽样的优点是代表性高，抽样误差小。如果抽样误差是一定的，抽样数目可以减少。

3．等距抽样

等距抽样又称机械或系统抽样。它是事先把总体的全部单位按某一标志排列，然后按固定顺序和间隔来抽取调查单位的一种抽样方式。按等距抽样方式来抽取调查单位，能够使抽出的调查单位更均匀地分布在总体中。因此，等距抽样的误差一般较简单随机抽样的小，特别是当研究现象的变量差异程度大，而在实际工作中，又不可抽取更多的单位进行调查时，采用等距抽样是极好的选择。

例如，一个学校有3000多学生，需要抽出120人进行某项调查，可利用学生名册进行排队。从1号排到3000号，抽选的间隔是：3000÷120=25人。

先从第1组25人中随机确定第1抽选人，若假定是第15号，然后每隔25人抽出1个，即有15号，40号，65号，……，共120人。

这是实际工作常用的一种方法，这种抽取样本单位的工作可以由计算机完成。

4．整群抽样

前面三种抽样方式所抽取的样本单位都是个体，而整群抽样所抽取的样本，是由若干样本组成的群体。整群抽样是先将总体各单位划分成若干组群，然后以群为单位从中随机抽取一些群，对中选群的所有单位进行全面调查的抽样方式。例如，对冷库中箱装鲜蛋进行抽样调查时，就是以箱为单位抽出后进行观察。

又如，第五次人口普查登记质量的抽样调查中，是按照一个居民小组或村民小组作为—个组群来观察的。

整群抽样是随机抽样的一种特殊组织方式，多用于研究对象较广、总体单位较多的抽样调查。

5．阶段抽样

当总体很大时，抽样调查直接抽选总体单位，在技术上有很大的困难，一般可以分阶段进行抽样。例如，我国农产量的抽样调查，第一阶段是从省抽县，第二阶段是从中选县抽乡，第三阶段是从中选乡抽村，第四阶段是从中选村抽地块，最后还可以从地块抽具体的样本点，并以样本点的实际资料推算平均亩产和总产量。

**19、敏感性问题抽样调查技术进展及其应用**

展望

由于敏感性问题具有复杂性、多样性和可变性的特点，在进行敏感性问题的随机化调查时，往往要对调查人员进行培训，每一个调查人员必须准确掌握随机化回答模型的原理，同时，要让被调查者对随机化装置进行确认，使每一个被调查者确信调查人员无法从其回答中推断自己对敏感性问题的真实回答。随机化回答抽样调查的时间过长，费用增大，不利于在实践中推广，有时被查者并未理解或不相信这种方法，使得敏感性问题的随机化回答调查法在应用中有一定的局限性。

对敏感性问题的抽样调查，Warner首先提出了随机化回答技术，此后，国内外学者先后提出了一些其他的抽样调查方法。方法一是Simith 等人1974年提出的“区组化总计回答技术”，它不是让被调查者根据随机化装置来决定回答哪个问题，而是将问题进行随机化，最简单的二元回答的情况下，把问题分为二类，一类包含敏感性问题和几个非敏感性问题；第二类仅有敏感性问题，在调查中只要求被调查者提供问卷中所答问题的答案之和，这样既可以对敏感性问题的答案保密，又可以根据两类问卷中平均总量的差异进行估计。方法二是Oh 和Scheuren1983年提出的“拟随机化回答技术”，它建立在被调查的总体可以分为回答层和无回答层的假设上，对回答层的每个个体来说，所调查的问题并不存在敏感性，被调查者以概率1真实回答敏感性问题；无回答层的个体以概率p 随机化回答敏感性问题，由二层的估计量对总体敏感性问题进行估计。方法三是涂光华和马岚提出的“分层弱化模型”，它根据敏感性调查的不同目的，研究造成该调查问题为敏感性问题的原因，由此选择与之相关的敏感分层标志以划分不同的调查层，再根据每层被调查者的不同心理特点或性格特征相应设计分层弱化或非弱化的敏感性问题，达到降低敏感性问题的敏感度争取被调查者合作的目的。

上述三种敏感性问题的抽样调查法都存在一定的局限性，方法一同随机化回答调查一样误差比直接调查大，在小样本的情况下，误差可能为负值；方法二是建立在假设的基础上，缺乏一定的可靠性；方法三分层弱化的效果如何衡量以及敏感性问题如何估计都有待进一步的研究。抽样调查的目标是设计最优的调查方案，找到参数的无偏估计，并使抽样调查的误差尽可能地小。事实上，在抽样调查中完全消除估计量的偏倚是不可能的，因此，在对敏感性问题进行调查时，可以考虑一个具有较小方差的有偏估计。

**21、社会网络分析中的数据格式及与传统统计数据形式的关系**

社会网络是指：群体中的行为人以及行为人之间的关系。行为人可以是自然个体，也可以是非正式群体、正式组织和集体，关系为两个行为人之间的特定接触、连接或联结。

社会网络的概念和分析过程不同于传统的统计分析和数据分析。斯科特认为，社会科学的数据主要分为“属性数据”、“关系数据”和“观念数据”三类

（1）属性数据

属性数据是关于行动者的自然情况、态度、观点以及行为等方面的数据，它们一般被视为个人或者群体所具有的财产、性质、特点等属性。“是人们、对象或者事件的内在特点”。适用于分析属性数据的方法主要是“变量分析”，如相关分析、回归分析、列联表分析等。各种属性被看成是特定变量（如收入、职业、教育程度等）的取值。

（2）关系数据

关系数据则是关于联系、接触、联络或者聚会等方面的数据。这类数据把一个行动者与另一个行动者连接在一起，因此不能还原为单个行动者的属性。关系不是行动者的属性，而是行动者系统的属性。这些关系把多对行动者联系成一个更大的关系系统。针对这些关系数据的分析方法就是网络分析。尽管对这些关系也可以进行常规的定量统计分析，但是，社会网络分析方法更适用于此类数据。社会网络分析者首先认为，应该利用关系数据来解释社会现象。关系是网络分析的基石。“关系数据”有如下类型：

。1行动者-行动者数据（方阵数据）。这种数据的手机方法可以是询问类似于如下问题：“谁是你的朋友？”得到的是类似于如下表格的数据。例如，假设“张三”认为“李四”是他的朋友，就在“张三”这一行和“李四”这一行的交叉处填上“1”，否则为“0”。

。2行动者-事件（长方阵数据）。问题：你是哪个班级的？得到的数据就是隶属关系数据。进一步说，传统数据关注的是行动者和属性，网络数据关注的是行动者和关系。网络学者所运用的研究工具很大程度上不同于其他社会科学家所使用的工具。

（3）观念数据

观念数据描述意义、动机、定义等。分析这类数据的技巧不如分析前两类数据的技巧那么强，尽管它们居于社会科学的核心。总的来说，韦伯提出的类型分析是最出色的研究，但是此类放啊需要进一步发展。

因此，社会网络分析中的数据类型主要以关系数据为主，其数据的表现格式一般为方阵或长方阵，这就与传统的统计数据形式有很大的差别。在传统的统计学中，数据一般是以个案及其特征的形式体现，例如学生的成绩、居民的收入等，或者对于时间序列数据，必定有一个属性是时间。社会网络数据可以看做是传统统计数据的延展，这种延展一般是横向的，即延展成“行动者-行动者”与“行动者-事件”这样的关系型数据。

**22、统计回归技术进展与趋势（及其应用）评述**

从高斯提出最小二乘法算起，回归分析已有200多年的历史。回归分析的应用非常广泛，我们大概很难找到不用它的领域，这也正是200多年来其经久不衰、生命力强大的根本原因。

计量经济学是线代经济学中影响最大的一门独立学科，诺贝尔经济学奖得者萨缪尔森曾经说过，第二次世界大战之后的经济学是计量经济学的时代。然而，计量经济学中的基本计量方法就是回归分析，计量经济学的一个重要理论支柱是回归分析理论。

自1969年设立诺贝尔经济学奖以来，已有近60位学者获奖，其中绝大部分获奖者是统计学家、计量经济学家、数学家。从大多数获奖者的论著看，他们对统计学及其回归分析方法的应用都有娴熟的技巧，这足以说明统计学方法在现代经济研究中的重要作用。

矩阵理论和计算机技术的发展为回归分析模型在经济研究中的应用提供了极大的方便。国民经济是一个错综复杂的系统，一个宏观经济问题常常要涉及几十个甚至几千个变量和方程，如果没有先进的计算机和求解线性方程组的矩阵计算理论，要研究复杂的经济问题是不可想象的。也正是由于计算方法的改进和现代计算机的发展，过去不可想象的事情变成了现实。

回归分析的理论和方法研究200多年来也得到不断发展，统计学中的许多重要方法都与回归分析有着密切的联系，如时间序列分析、判别分析、主成分分析、因子分析、典型相关分析等。这些都极大地丰富了统计学方法的宝库。

回归分析方法自身的完善和发展至今是统计学家研究的热点课题。例如自变量的选择、稳健回归、回归诊断、投影寻踪、分位回归、非参数回归模型等近年仍有大量研究文献出现。

对于满足基本假设的回归模型，它的理论已经成熟，但对于违背基本假设的回归模型的参数估计问题近年仍有较多研究。

在实际问题的研究应用中，人们发现经典的最小二乘估计的结果并不总是令人满意，统计学家从多方面进行努力，试图克服经典方法的不足，提出了许多新的估计，其中主要有岭估计、压缩估计、主成分估计、斯泰因估计，以及特征根估计。这些估计的共同点是有偏，即它们的均值并不等于待估参数，于是人们把这些估计成为有偏估计。当矩阵X呈病态是，这些估计都改进了最小二乘估计。

为了克服最小二乘估计对异常值的敏感性，人提出了各种稳健回归；为了研究模型假设条件的合理性及样本数据对统计推断影响的大小，产生了回归诊断；为了研究回归模型中未知参数非线性的问题，人们提出了许多非线性回归方法，这其中有利用数学规划理论提出的非线性回归参数估计方法、样条回归方法、微分几何方法等；为了分析和处理高维数据，特别是高维非正态数据，产生了投影寻踪回归、切片回归等。

近年来，新的研究方法层出不穷，如非参数统计、自助法、刀切法、经验贝叶斯估计等方法都对回归分析起着渗透和促进作用。由此看来，回归模型技术随着它自身的不断完善和发展以及应用领域的不断扩大，必将在统计学中占有更重要的位置，也必将为人类社会的发展发挥它独到的作用。