《统计学基础》考试题库

1.有关分类的统计方法有哪些？举例说明分类研究思想的重要意义。

2.统计指标有哪些类型？谈谈统计指标理论在社会经济统计学中的重要地位。

3.试述统计调查与统计实验的区别与联系。

4.试就统计指数偏误理论与指数测验理论谈谈自己的看法。

5.时间序列分析常用的指标有哪类？当前有关时间序列分析的一些新思路或新方法有哪些？

6.从统计学方法（体系）的内容等角度，谈谈经济统计学与数理统计学或计量经济学之间的关系。

7.为什么会出现基础比率谬误？如何避免该谬误？

8.请阐述空间计量模型的主要模型形式及其内在联系。

9.试述统计研究过程中（或某一环节）的缺失值产生原因与处理方法。

10.当前常用的 P 值存在什么问题，如何改进？

11.大数据会对相关分析带来什么样的影响？如何对相关分析进行拓展？

12.数据预处理有哪些步骤，可采取哪些方法来提升数据质量？

13.统计平均方法有哪些类型或方法，常用平均数公式有什么特点或应用条件？

14.试述敏感性问题的抽样调查技术研究进展及其应用。

15.大数据只要相关不要因果，谈谈你对这句话的理解及看法。

16.社会网络分析方法的数据表现是什么？基于社群图，社会网络有哪些重要的统计特征？如何测量？

17.试对统计指数编制方法与指数因素分析体系的最新进展进行评述。

18.从利益相关者视角谈谈统计数据质量的维度、评估方法及提升途经。

19.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？

20.谈谈可视化技术主要内容及其在统计学中的可能应用

1.有关分类的统计方法有哪些？举例说明分类研究思想的重要意义。

答：（1）①传统的分类方法主要是聚类分析和判别分析。

聚类分析是根据变量间的相似程度聚类，聚类的依据是利用样品相似程度所反映出来的量，显然描述变量关系的数学方法不同，产生的分类结果一般也会有所不同，常用相似（相关系数）和距离描述变量间的关系。聚类分析把分类对象按一定规则分成组或类，这些组和类不是事先给定的而是根据特征而定的。

判别分析是判别样品所属类别的一种统计方法，是在已知分类情况之下，遇到有新的样本时，可以利用此方法选定以判别准则，以判定将新样品放置于哪个类中。判别分析可以从不同角度提出问题，因此有不同的判别准则，如马氏距离最小准则、Fisher 准则、平均损失最小准则、最小平方准则、最大似然准则、最大概率准则等，按判别规则的不同又提出多种判别方法，常用的方法有距离判别法、Fisher 判别法、Bayes 判别法和逐步判别法。

②可以将描述统计阶段的统计分组过程也视作是一种分类方法。

统计分组就是根据统计研究的需要，按照一定的标志，将统计总体划分为若干个组成部分的一种统计方法。总体的这些组成部分，称为“组”，也就是大总体中的小总体。通过统计分组,使同一组内的各单位在分组标志的性质相同，不同组之间的性质相异。对统计总体进行分组，是由统计总体中各个总体单位所具有的“差异性”特征所决定的。统计总体中的各个单位，一方面，在某一个或几个标志上具有相同的性质，可以被结合在同一性质的总体中;另一方面，又在其他标志上具有彼此相异的性质，从而又可以被区分为性质不同的若干个组成部分。

③随着大数据时代的到来，涌现了许多新的分类方法，或者改进的传统分类方法。

主要分类方法介绍解决分类问题的方法很多，单一的分类方法主要包括：决策树、贝叶斯、人工神经网络、K-近邻、支持向量机和基于关联规则的分类等；另外还有用于组合单一分类方法的集成学习算法，如 Bagging 和 Boosting 等。

（2）例子：

①通过分类算法，对目标进行自动分类，能够极大的节省人力成本。比如，垃圾邮件的分类问题、自然语言处理中的文本分类问题、人脸识别中的联合贝叶斯模型等。

②通过分类讨论个体的异质性对某一因素的影响。在研究城市规模对于经济增长质量的影响中，有必要对不同等级的城市进行分别回归，以讨论城市自身的异质性对经济增长质量的影响。

③基于 Beyes 分类器的手写汉字识别系统，应对传统方法难以应对的问题。

能够和那后应对脱机手写识别文字识别中最有挑战性的问题，一是手写文字变化很大难以辨认；二是难以提取出对识别很有帮助的笔顺信息

2.统计指标有哪些类型？谈谈统计指标理论在社会经济统计学中的重要地位。

答：统计指标简称‘指标’，是反映社会经济总体现象数量特征的概念和数值。

（1）统计指标的类型

①统计指标按照其反映的内容或其数值表现形式，可以分为总量指标、相对指标和平均指标。

总量指标：反映现象总体规模的统计指标，按其反映的时间状况不同又可以分为时期指标和时点指标。时期指标又称时期数，它反映的是现象在一段时期内的总量，时期数通常可以累积，从而得到更长时期内的总量。时点指标通常又称为时点数，它反映的是现象在某一时刻上的总量，时点数通常不能累积。

相对指标：又称为相对数，是两个绝对数之比，其表现形式通常为比例和比率两种。

平均指标：又称为平均数或均值，它反映的是现象在某一空间或时间上的平均数量状况。

②统计指标按其所反映总体现象的数量特性的性质不同可分为数量指标和质量指标。

数量指标：反映社会经济现象总规模水平和工作总量的统计指标，一般用绝对数表示。
质量指标：反映总体相对水平或平均水平的统计指标，一般用相对数或平均数表示。

③按管理功能作用不同，可以分为描述指标、评价指标和预警指标。

描述指标:反映社会经济运行的状况、过程和结果，提供对社会经济总体现象的基本认识，是统计信息的主题。

评价指标:是用于对社会经济运行的结果进行比较、评估和考核，以检查工作质量或其他定额指标的结合作用，包括国民经济评价指标和企业经济活动评价指标。

预警指标:用于对宏观经济运行进行监测，对国民经济运行中即将发生的失衡、失控等进行预报、警示。

（2）统计指标理论在社会经济统计学中的重要地位

统计指标是社会经济统计学中的基本范畴,它和统计分组作为统计的两大要素贯穿统计工作的全过程。统计指标理论是统计理论的重要组成，是制定计划、实行宏观调控的基础，是制定政策的依据、实行管理的手段等等，被应用于经济发展、经济效益、生活质量、综合国力、社会发展水平的综合研究。如 GDP 指标反映一个国家或地区的经济发展规模并据以进行经济结构分析，数据来源依计算方法不同逐项分别获取，GDP 这一指标科学概括了经济活动最终成果这一属性，客观反映了国民经济运行的完整过程和内在关系，是最终产品（包括货物和服务）总规模的价值表现。

3.试述统计调查与统计实验的区别与联系。

答：（1）统计调查与统计实验的区别：

①概念不同：统计调查是指研究者通过客观地观察、记录、描述调查对象来收集数据，包括资料的收集、整理和分析及整个设计过程的统计设想和科学安排。

统计实验是指研究者根据研究目的，对研究对象实施干预措施，观察总结研究效应的结果，回答研究假设所提出的问题的整个过程。

②特点不同：统计调查是探索性研究，研究者不能人为安排研究因素；统计实验是实验研究，研究者能够主动安排实验因素。

统计调查不能随机分组，不能平衡或消除非研究因素对研究结果的影响；统计实验可以随机分组，控制实验条件，排除非研究因素的干扰。

统计调查有较为固定的基本程序，统计实验则没有太固定的基本程序。

根据不同的要素它们的分类也不同。

③研究目的不同：统计调查观察和记录对象本身的固有属性及某些现象，通过对数据的整理分析，比较不同组之间某种现象或相关特征的差异，但是不能说明研究因素与观察之间的因果关系。实验设计通过积极干预，观察总结研究效应的结果，目的是说明研究因素与观察现象之间的因果关系，回答研究假设所提出的问题。

（2）统计调查与统计实验的联系：

二者都是通过对资料的收集、分析，解决一定的问题，为研究提供数据支撑，保证研究的顺利进行。但是在实际科研过程中，为了提高研究效率，常常需要研究者准确地选择恰当的设计方法。但是在实际科研过程中，为了提高研究效率，常常需要研究者准确地选择恰当的设计方法。在科学探究过程中，有些问题单凭调查或者单凭实验是难以得出结论的. 这时就需要通过两者相结合，相辅相成，共同发现和验证科学结论。

4.试就统计指数偏误理论与指数测验理论谈谈自己的看法。

答：（1）定义：

偏误理论是指数理论重要的组成部分,是指数理论工作者寻求最优指数的基点和依据，偏误理论认为，统计指数的偏误包括型偏误和权偏误这两大类型：(1)不满足时间互换检验的指数有型偏误；(2)不满足因子互换检验的指数有权偏误。“偏误”一词在指数理论中的提出缘于鲍利,而将整个偏误理论发扬光大的当属费雪。

指数常常需要测验，有关的测验实质上就是对指数分析性质的评价。这些性质可以由问题的具体内容中离析出来，经过归纳和提炼，仅就其数学形式进行专门研究。在费雪的经典名著《指数的构造》当中,费雪为对指数的优劣进行甄别,在前人的基础上归纳并独创地提出了一套检验方法,形成了费雪的统计指数检验理论，条目共有八项:确定性检验、恒等性检验、同度量检验、比例性检验、时间互换检验、因子互换检验、循环检验、进退检验。

（2）看法：

①统计指数偏误理论问题的提出，在整个统计指数理论中,型偏误理论是重要的组成部分,然而在我国目前关于指数的论著中,针对该问题的阐释却显得相对比较薄弱,甚至有些篇章中的论述还出现了错误。

②在谈及关于型偏误的指数文献中,有一部分论著指出经济学所考虑的经济现象严格沿时间一维变化,任何经济现象的发生都是在特定的环境、地点下出现的结果,时间互换检验成立的前提是时间维的可逆性,不符合经济学的研究方法。

③指数的型偏误理论确实具有实际的效用,例如在国民经济核算领域的年度和季度核算中,美国和加拿大都已经使用链式费雪指数,它是多方考虑了价格指数理论和经济理论的成果;另外,在国际经济对比中的价格换算指数的编制,和反映地域差别的生活费用指数的编制也都考虑了型偏误理论。

④测验本身可能存在着诸多问题，比如测验的种类过多；个别测验排斥了某些可用的指数公式；各种测验标准并非处于同一层次上；它们彼此之间有包含关系；另有个别测验缺乏实际意义。

5.时间数列分析常用的指标有哪类？当前有关时间数列分析的一些新思路或新方法有哪些？

答：（1）常用指标：

①水平分析指标

1）发展水平：数列中的具体指标数值为发展水平，可以是绝对数、相对数或平均数。可以分为：最初水平、最末水平、中间水平、基期水平、报告期水平等

2）平均发展水平：对不同时期的发展水平加以平均；

3）增长水平：一定时期内增长的绝对数量；

4）平均增长水平：对不同时期的增长水平加以平均；

5）边际倾向：两个增量的比值，用于说明 A 现象增加一个单位对 B 现象增量的拉动效率。

②速度分析指标

1）发展速度：报告期水平/基期水平；

2）增长速度：反映现象增长程度的相对指标；

3）平均发展速度：发展速度的平均值；

4）平均增长速度：增长速度的平均值；

5）超过系数：两个现象发展的同步性指标；

6）弹性系数：两个现象增长的同步性指标。

（2）新思路或新方法：

①灰色系统理论

灰色预测法是一种对含有不确定因素的系统进行预测的方法。灰色预测是对既含有已知信息又含有不确定信息的系统进行预则，就是对在一定范围内变化的、与时间有关的灰色过程进行预测。其特征包括：

1）通过鉴别系统因素之间发展趋势的相异程度，即进行关联分析，并对原始数据进行生成处理来寻找系统变动的规律，生成有较强规律性的数据序列,然后建立相应的微分方程模型，从而预测事物未来发展趋势的状况

2）用等时距观测到的反映预测对象特征的一系列数量值构造灰色预测模型，预测未来某一时刻的特征量，或达到某一特征量的时间。
②基于 EMD 和神经网络的非线性时间序列预测方法
综合经验模态分解(EMD)和人工神经网络，提出一种改进对非线性时间序列进行预测分析的方法。这种方法的优势在于:
1）相比于主流的预测模型，其预测精度更高。
2）对于神经网络在学习时相关参数难确定的问题，提出了一种有效的解决方案。
3）对时间序列变动背后的逻辑，提出了一种有数据依据的解释。
③指数平滑模型的改进
指数平滑法是基于滑动平均 MA（q）模型发展而来的，是时间序列分析的重要分支。指数平滑模型对于平滑初值确认存在固有缺陷，其参数都是静态的，随着时间的推移预测的效果因滞后性而逐渐下降，且其参数往往依赖经验获得，无法得到最优值。
优化指数平滑模型，以时间序列聚合理论为经典代表，对时间序列进行聚合的优点在于，不同的聚合阶数的序列，将呈现出不同的特征，有利于对其分别进行拟合和预测。针对时间序列的指数平滑方法研究仍相对滞后，缺少对一族方法进行整合并进行系统的研究，很少有理论推导。指数平滑方法是一族方法，而不是一种统计模型。基于此，进行族之间的组合、优化、借鉴，是提升时间序列指数平滑算法对未来预测效果的可行途径。时间序列变化有其复杂的机理，反映了序列内本质的属性，进行时间序列指数平滑算法优化，还需要基于序列本身的特点开展，而不是盲目的优化。

6.从统计学方法（体系）的内容等角度，谈谈经济统计学与数理统计学或计量经济学之间的关系。

答：（1）定义：
①社会经济统计学分科包括农业统计、工业统计、人口统计、社会统计、金融统计、国民经济核算等，是一门涉及范围相对广泛的学科。
②数理统计学派的产生与概率论的发展紧密相关。数理统计学是研究社会和自然界中大量随机现象数量变化基本规律的一种方法，可分为描述统计和推断统计。
（2）关系：
①社会经济统计学和数理统计学的相同之处：
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。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，并且两者都能对统计规律进行分析和探究。从研究对象角度来看，都能将人、事物、项目作为研究对象。
②社会经济统计学和数理统计学的不同之处：
1）社会经济统计学和数理统计学两者研究范围不同。
社会经济统计学，一般是针对社会经济现状内容进行分析。而数理统计不仅可以对社会经济现象进行分析，而且还可以有效针对自然现象进行数据分析处理，预测和体现出随机现象的可能性。数理统计学所涉及到的应用问题相对比较广泛。社会经济统计学虽然研究范围相对狭隘，但其所涵盖的内容广泛。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。社会经济统计学涉及到人们日常生活的各个层次领域。

2）社会经济统计学和数理统计学的理论基础存在差异。
数理统计学的核心理论基础内涵是概率论、统计推断理论。社会经济统计学和数理统计学两者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。
（3）总结：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。在统计学受到冲击的大数据时代，我们应科学合理地挖掘社会经济统计学和数理统计学的实际优势，确保社会经济统计学和数理统计学两者的高效运用。

7.为什么会出现基础比率谬误？如何避免该谬误？

答：（1）定义

是一种概率谬误，多因没有考虑统计学上的基础概率而导致推论谬误。人类在进行主观概率判断时，如果所获取的信息中既有一般信息(基础概率)又有具体信息(诊断信息)，那么他们往往倾向于根据诊断信息来进行判断，而忽略掉基础概率，导致判断结果与贝叶斯定理所给出的结论不符，这一现象被称为“基础比率谬误”。也就是说，当人们拥有两种类型的信息时，倾向于根据具体信息来进行直观判断，而把基础概率抛之脑后，导致判断结果与贝叶斯定理所给出的结论大相径庭，从而产生“谬误”。）

心理学家们开展了许多实验研究并提出各自的解释，其中较具代表性的解释包括以下几类。

（2）①代表性启发策略。卡尼曼和特维尔斯基试图用“代表性”来解释基础比率谬误产生的原因。他们认为代表性是一种启发式判断策略，在人们的直观判断和预测中被广泛使用。这种策略表明，当人们根据代表性策略来对某个事件的概率进行判断和预测时，主要是根据这个事件与某个范型的相似性或代表性来进行的，通过比较然后选择那种与这段描述最相似或最具代表性的选项。人们根据代表性启发原则来进行判断时，是通过相似性来确定概率的，而相似性不会受到基础概率的影响。这样一来，直观判断或预测对于证据的可靠性或结果的 初始概率就不“敏感”了，从而与规范性决策理论所给出的结果相违背，出现“基础比率谬误”。由此，卡尼曼和特维尔斯基得出结论：“根据代表性假设，当存在个别信息的时候，先验概率就被大大地忽略掉了。”这表明，在不确定条件下进行判断和预测时，人们通常不会严格遵守概率计算规则或统计预测理论，往往依据一些启发式策略来进行判断，容易导致系统性错误。

②相关性原则。巴希勒认为，卡尼曼和特维尔斯基的解释是不充分的，并试图通过“相关性”来解释基础比率谬误现象。首先，人们忽略掉基础概率信息是因为觉得它与当下的判断无关。其次，当人们面对多条信息的时候会进行判断和筛选，依据是相关性的大小，相关性小的信息容易被相关性大的信息所支配或掩盖。人们在进行主观概率判断时，与问题不太相关的基础概率往往被忽略。相关性的大小是如何体现出来的呢？巴希勒认为，“相对于所需要判断的事件来说，如果一条信息比另外一条信息更加明确、特殊或个别，那么这条信息就比另外那条信息的相关性大。”当然，人们在做出判断的时候，并不是一味地忽略基础 概率，而且忽略基础概率并不等于“谬误”。当人们觉得基础概率的相关性并不比具体信息的相关性小时，基础概率就不会被忽略掉，从而不会导致基础概率谬误的产生。

③因果基础概率与偶然基础概率的差异。代表性启发策略提出来数年之后，特维尔斯基和卡尼曼又提出了一种新的理论，来对其之前的理论进行辩护和改进。他们认为，基础概率其实可以分为两种：一种是因果基础概率，另一种是偶然基础概率。如果基础概率存在一个因果因子来解释为什么某个特殊情况更有可能产生这种结果而不是其他结果的话，那么这个基础概率就是因果基础概率；否则，是偶然基础概率。卡尼曼和特维尔斯基认为，人们进行主观概率判断时所忽略掉的主要是偶然基础概率，而因果基础概率不太容易被忽略掉。

④思维定式。“思维定式”是指人们会将自己对某个团体的看法延伸到这个团体中每个成员的身上，通常与“因果基础概率”相联系。

（3）人们在做判断时，如果忽略基础比率和证据可靠性的影响，注定会犯错。现实中，做出判断的时候应该“像统计学家那样思考”，注重对基础概率信息的利用，采用贝叶斯定理来进行分析，把最强烈的信念与证据分析相结合，以做出更为合理的判断及决策。①用贝叶斯定理来约束直觉。对不确定性事件进行主观概率判断是认知决策领域一项重要内容，多年来贝叶斯主义者一直致力于用贝叶斯定理来提高人们判断和预测的准确性。注重先验概率与后验概率相结合，利用贝叶斯定理来约束“根据典型性进行判断”的直觉，从而将“基础比率”也考虑进来，以相对合理的基础比率对结果的可能性做出准确判断。当然，某些情况下对基础概率的忽略并不像贝叶斯定理信奉者所想象的那么糟糕，甚至还可能是一件好事。

②选择与需要解答的问题直接相关的对象作为参照系。参照系的选择对于人们的判断和选择至关重要，一般说来，参照系越小其中元素所具有的共同点就越多，最后得到的答案也就越准确。当然，如果参照系过小，可能就不具有参考作用； 而任意扩大参照系，由其给出的基础概率可能与问题的相关性并不大。

③判断与决策时不要被很细节的情境所迷惑。通常，正是情境中的细节使得整个情境看起来更加具有代表性，但其实也减少了其发生的可能性。一般而言，情境越是具体，其发生的可能性越低——即使这样的情境看起来能够非常好地代表最有可能发生的结果。所以，我们在做出判断和决策的时候，要尽量避免受到那些细节情境的影响而忽略基础比率。

8.请阐述空间计量模型的主要模型形式及其内在联系。

答：现有的空间计量模型可以列为以下十种：

$$SEM: y = X\beta + \mu, \mu = \lambda W\mu + \varepsilon, \varepsilon \sim N(0, \sigma^2 I_n)$$

$$SMA: y = X\beta + \mu, \mu = \varepsilon + \lambda W\varepsilon, \varepsilon \sim N(0, \sigma^2 I_n)$$

$$SEC: y = X\beta + \mu, \mu = W\eta + \varepsilon, \eta \sim N(0, \sigma_\eta^2 I_n), \varepsilon \sim N(0, \sigma_\varepsilon^2 I_n)$$

$$SLX: y = X\beta_1 + WX\beta_2 + \varepsilon, \varepsilon \sim N(0, \sigma_\varepsilon^2 I_n)$$

$$FAR: y = \rho Wy + \varepsilon, \varepsilon \sim N(0, \sigma_\varepsilon^2 I_n)$$

$$SAR: y = \rho Wy + X\beta + \varepsilon, \varepsilon \sim N(0, \sigma_\varepsilon^2 I_n)$$

$$SARMA: y = \rho W_1 y + X\beta + \mu, \mu = \varepsilon + \lambda W_2\varepsilon, \varepsilon \sim N(0, \sigma_\varepsilon^2 I_n)$$

$$SAC: y = \rho W_1 y + X\beta + \mu, \mu = \lambda W_2\mu + \varepsilon, \varepsilon \sim N(0, \sigma_\varepsilon^2 I_n)$$

$$SDM: y = \rho Wy + X\beta + WX\theta + \varepsilon, \varepsilon \sim N(0, \sigma_\varepsilon^2 I_n)$$

$$SDEM: y = X\beta + WX\theta + \mu, \mu = \lambda W\mu + \varepsilon, \varepsilon \sim N(0, \sigma_\varepsilon^2 I_n)$$

把模型（1）至（10）统称为空间模型族中的模型，其中模型（1）至（3）仅在误差项中存在空间相关性，模型（4）仅在解释变量中存在空间相关性，模型（5）至（6）仅在被解释变量中存在空间相关性，模型（7）至（10）存在混合的空间相关性。模型（1）（6）（9）是最常见的空间计量模型，分别称为空间误差模型（SEM）、空间自回归模型（SAR）、空间杜宾模型（SDM），SAC 模型既包括了空间滞后，又包括了空间误差项的一般空间模型。一阶空间自回归模型（FAR）、空间误差分量模型（SEC）和空间杜宾误差模型（SDEM）并不常见,FAR 类似于时间序列分析中的一阶自回归模型，主要用于研究相邻地区的被解释变量的变动如何影响被研究地区的被解释变量。SEC 与 SEM、SMA 的最大不同是误差项中不含有空间相关性系数，且误差项由两个独立误差分量构成。SDEM 是空间杜宾误差模型，只是对 SEM 模型中增加了解释变量的空间滞后项。

9.试述统计研究过程中（或某一环节）的缺失值产生原因与处理方法。

答：（1）缺失值产生的原因

缺失值产生的原因只要分为机械原因和人为原因。机械原因是由于机械原因导致的数据收集或保存的失败造成的数据缺失，比如数据存储的失败、存储器损坏、机械故障导致某段时间数据未能收集等。人为原因是由于人的主观失误、历史局限或有意隐瞒造成的数据缺失。

（2）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再抽样。又包括以下三种情况：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。②替换被调查单位。在出现无回答的情况下，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。③对无回答进行子抽样。当后继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的方法。

③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获得数据身上，主要用于单元数据缺失情况下的调整。

④插补。该方法的基本思想是利用辅助信息，为每个缺失值寻找替代值。 具体操作时，可采用以下几种策略：使用一个固定的值代替缺失值、使用均值代替缺失值、使用同一类别的均值代替缺失值、使用成数推导值代替缺失值以及使用最可能的值代替缺失值。
1）均值插补。数据的属性分为定距型和非定距型。如果缺失值是定距型的，就以该属性存在值的平均值来插补缺失的值；如果缺失值是非定距型的，就根据统计学中的众数原理，用该属性的众数(即出现频率最高的值)来补齐缺失的值。
2）利用同类均值插补。它用层次聚类模型预测缺失变量的类型，再以该类型的均值插补。假设 X=(X1,X2…Xp)为信息完全的变量，Y 为存在缺失值的变量，那么首先对 X 或其子集行聚类，然后按缺失个案所属类来插补不同类的均值。如果在以后统计分析中还需以引入的解释变量和 Y 做分析，那么这种插补方法将在模型中引入自相关，给分析造成障碍。
3）极大似然估计。在缺失类型为随机缺失的条件下，假设模型对于完整的样本是正确的，那么通过观测数据的边际分布可以对未知参数进行极大似然估计。对于极大似然的参数估计实际中常采用的计算方法是 EM 算法。该方法的一个重要前提：适用于大样本。

有效样本的数量足够以保证 ML 估计值是渐近无偏的并服从正态分布。但是这种方法可能会陷入局部极值，收敛速度也不是很快，并且计算很复杂。
4）多重插补。多值插补的思想来源于贝叶斯估计，认为待插补的值是随机的，它的值来自于已观测到的值。具体实践上通常是估计出待插补的值，然后再加上不同的噪声，形成多组可选插补值。根据某种选择依据，选取最合适的插补值。

10.当前常用的 P 值存在什么问题，如何改进？

答：（1）对 P 值的认识误区：

①P 值是在原假设成立的前提下，统计量获得现有观测值或更极端观测值的概率，即 P(D|H0)，而原假设成立的概率则是在现有观测数据下零假设成立的可能性，即条件概率 P(H0|D)。而 P(D|H0)和 P(H0|D)的现实差异可能很大。

②P 值小于显著性水平α，说明原假设错误，拒绝原假设；显著性检验只提供假设检验的概率信息，不能证明某个假设为真或为假，以及假设为真或为假的概率。P 值小于显著性水平α，只说明有充分理由拒绝原假设，并不能说明原假设是完全错误的，仍然有α的概率错误地拒绝了原假设，造成第一类错误的发生。

③重复谬论——若某项研究重复多遍，则认为在(1-P)的场合下都能得到统计显著性结果。假设检验的 P 值只表示在零假设为真的条件下得到某个观测值或更极端值的概率，因此，一项研究中拒绝原假设并意味着在另一项重复性研究中一定能得到拒绝原假设的结果。对于同样的实验经过多次抽样会得到不同的样本，而假设检验对样本容量又具有较大的依赖性，随着样本不同 P 值也会发生变化。

④显著性水平α为 0.05。作为显著性水平，α是事先主观确定的，表示犯第一类错误概率，不同的显著性水平各有优缺点。在分析不同的问题时要根据实际情况和自己掌握的证据进行不同的考虑，选择适合的显著性水平。

⑤统计显著性结果总是有实际意义或在总体中存在很大效应，P 值越小代表检验总体的差异越大，即差异越不可能是因随机误差造成的。统计显著性只是告诉人们在特定条件下均值差异、变量相关性等是存在的，并不完全是由抽样误差造成的，但不意味着这种差异、相关性等就具有明确的实际意义，统计上的显著性不能等同于实际意义。

（2）P 值的局限性主要体现在以下方面：

①假设检验对样本量具有较强的依赖性。对于同一个检验，如果样本容量大，其自由度也大，更容易得到较小的 P 值。无论自变量的影响效应是大还是小，相较于小样本，大样本更容易拒绝原假设，也有足够的 统计功效保证得到具有统计显著性的结论。事实上，世事万物只要存在就会有差异，即原假设永 远不可能完全为真，只要样本容量足够大，就能得到拒绝 原假设的统计显著性结论，因此根据 P 值做出判断容易造成 逻辑上的不一致现象。

②显著性结论具有不确定性。检验统计量=效果量*样本容量。P 值是随机变量，混合了样本容量和效果量的影响，因此不能简单根据显著性结论而判定存在真实的效应。只有在控制了样本容量的条件下，才能得到 P 值越小效应越大的结论。统计显著性具有一定的不确定性。

③假设检验只注重结果的显著性，不考虑结果的可重复性。假设检验所计算的P值是在原假设为真时能获得当前样本数据的概率，逻辑上是由总体推断样本，而研究者希望由样本推断总体，唯有产生了对总体的推断才能够提供研究结果是否可以重复的信息。所以，P值代表的统计显著性并不意味着结果的可重复性，假设检验得到的是样本的可能性而不是总体的可能性，并没有考虑结果的可重复性。正是由于P值本身存在一定的局限性，使得研究人员、教师，以及不同领域的应用者对其存在一定的错误认识，进而导致实际应用中的误用甚至是滥用。

（2）对 P 值的改进策略

①效果量估计

零假设检验注重显著性差异的有无，并不探究差异的大小以及差异的实际意义。效果量代表的是自变量与因变量之间关系的强弱，反映了研究对象之间实际差异的大小，反映了实验效应大小的真实程度。其中，效果量分为标准化平均数差异效果量、未调校的考虑方差的效果、调校的考虑方差的效果量。

②统计功效检验

统计功效是在备择假设为真时拒绝错误原假设的概率。统计功效具有检验真实差异的能力，反映了假设检验正确侦查到真实处理效应的能力。可以检验方法的好坏，统计功效越强，方法越好。统计功效的影响因素包括不同总体的差异、效果量/样本容量/检验方向/显著性水平等。

③构建置信区间

置信区间是由样本统计量对总体参数做出的区间估计，可看作对点估计值信任程度的一种体现。P 值用来判断零假设的某个参数值能否具有合理性。譬如检验某个效果量是否与零具有显著性差异，可以构建一个 95%的置信区间，观察这个区间是否包含零，包含的话则差异不显著，从而获得估计值具体差异信息。

④计算贝叶斯因子（似然比）

贝叶斯因子用以描述与比较两个模型之间的相对确证性，在假设检验中反映当前数据对原假设与备择假设支持强度之间的比率，0 值是假设成立的条件下出现当前观测值或更极端观测值的概率，贝叶斯因子回答的是在当前数据条件下哪个模型相对更合理，贝叶斯因子相对于 P 值更有优势。

⑤计算错误发现率

在研究多重假设检验过程中，根据在 R 次拒绝原假设中错误拒绝次数 V 所占比例，计算错误发现率（FDR）。FDR 的定义为$FDR=\begin{cases}E\left(\frac{V}{R}\right) & ,R\neq 0\\ 0 & ,R=0\end{cases}$,FDR 对 P 值进行校正，试图在假阳性与假阴性之间找到平衡。FDR 相较于传统假设检验，有效降低了第一类错误对假设检验结果的影响，提高了检验的统计功效。

⑥重复性实验

首先，即使主观设定显著性水平α为 0.1，0.05 甚至 0.01，在拒绝正确原假设时仍然犯

了第一类错误，在接受错误原假设时也依然犯了第二类错误；其次，尽管得到了统计显著性结果，也不能完全确定样本差异不是由随机误差引起的，因为影响实验结果的因素包括抽样方法、样本容量等多个方面。所以，对于任何科学研究，重复性实验是必要的，是确保研究发现有效性的重要手段，为研究结果的可靠性提供保障。

⑦两阶段分析法等

两阶段分析法是探索性和证实性分析采用不同的处理方法，根据探索性研究的结果决定验证方法，进行重复研究，并一同公布探索性和证实性分析结果，保证了分析自由和灵活性，降低了公开发表结果的误报率，保证了研究结果的严谨性。

11. 大数据会对相关分析带来什么样的影响？如何对相关分析进行拓展？

答：大数据对相关分析带来的影响：

（1）传统相关分析理念面临挑战。大数据条件下数据体量和类型等方面的变化，使得一些传统相关分析理念面临挑战。①以样本代替总体，损失信息量。②对相关系数显著性的假设检验不充分。③数据不精确使得传统相关分析难以进行。④大数据的标准和类型不适用于传统相关分许。大数据时代，包括相关分析在内的众多统计思维都发生显著性变化，迫切需要转换相关分许理念。

（2）经典相关分析的应用局限。①当样本容量 n 趋向于无限大，使得传统相关系数的显著性检验失效。当 n 趋向于无限大时，t 检验或 Z 检验统计量将非常大，都倾向于拒绝原假设，认为相关性显著。②Pearson 相关系数只能度量变量间的线性相关，无法测度非线性关联。③传统相关分析法不适用于大数据的标准和类型。

对相关分析进行的拓展：

（1）针对更多数据类型的相关分析方法。大数据很多是认为行为的记录数据，通常表现为分类数据，需要能够适用于定类数据的相关分析方法。

①

不同类型变量之间相关性的测度方法

变量类型	测量方法
1.定类与定类	消减误差比率(PRE)、λ系数、τ系数、Φ系数、C系数
2.定类与定序	同上
3.定序与定序	γ系数、d系数、Kendall系数、rho系数
4.定距（定比）之间	Person　r系数
5.定类与定距(定比)	相关比率（E^2）
6.定序与定距(定比)	相关比率（E^2）

②分类数据的卡方检验。

分类数据的卡方检验

行(r_i) \ 列(c_j)	列(c_j)			合计
	$j=1$	$j=2$	…	
$i=1$	f_{11}	f_{12}	…	r_1
$i=2$	f_{21}	f_{22}	…	r_2
:	:	:	:	:
合计	c_1	c_2	…	n

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-e_{ij})^2}{e_{ij}}$$

③关联规则挖掘

关联规则挖掘可发现大量数据中项集之间的关联或相关关系，是通过数据间隐含的依赖关系生成知识。在支持度、置信度及提升度的框架下，关联规则中只有同时满足支持度、置信度与提升度的规则才是强规则。即具有良好的预测性的规则。

（2）大数据条件下的非线性相关探测。变量之间存在的非线性相关的强弱，难以用简单相关系数去判断，一些改进方法依赖于统计学的相关知识，对传统相关系数计算中的某个薄弱环节进行改进。

①相关指数法

变量之间是否存在非线性相关以及相关的强弱，难以用传统相关系数来测度和反映，相关指数法可用以判断变量之间是否显著存在某种类型的非线性相关关系。相关指数的实质是对非线性回归模型进行拟合时所得到的决定系数，类似用以测度非线性相关的系数还包括最大相关系数、距离相关系数等。目前对相关系数法进行改进，以测度变量间非线性相关性的代表性方法主要是基于信息论基础的最大信息系数（MIC）。MIC 是一种普适性的关联挖掘方法，适用于检测各种类型的函数关系或非函数化的相关系数计算。

（3）改进方法的应用及检验

①关联规则挖掘的应用。

通过关联规则，进行相关产品推荐或者挑选相应的关联产品进行精准营销。

12.数据预处理有哪些步骤，可采取哪些方法来提升数据质量？

答：（1）数据预处理过程大致包括数据审查、数据清理、数据转换和数据验证四大步骤。

（2）①做一张所有非连续变量的频数表。在数据预处理过程中，利用频数分析可以直接发现变量值中是否存在明显的极端值。一般可利用现成的统计软件将所有的变量做频数分析，对超出变量取值范围的数据则查找出对应的原始资料，对照原始资料分析原因、改正错误。频数范围检查的对象主要是定类和定序数据。

②列联表逻辑检验。列联表主要起到交叉分类的作用，从中可轻易地发现逻辑上不一致

的数据。

③描述统计量诊断。频数分析和列联表都主要适用于定性数据，对于定量数据则通常开展数据的描述性统计分析，给出数据的描述性统计量，特别是通过最大值、最小值、中位数、上四分位数、下四分位数等几个典型描述性指标反映出的数据实际分布特征，可以初步判断数据中是否出现极端值。

④ 正态分布下异常值诊断及应用

1）三倍标准差法；2）方差比法；3）极值偏差法；4）极差比法；5）几种方法的功效比较

⑤异常值的平滑处理。面对异常数据，最直接的处理方法是在开展数据分析时应将这些异　常值剔除。在异常值诊断过程中，有时本身需要先删除某个数据点，然　后看该数据点的删除对于模型估计量是否有显著影响，两者在检验效果　上是否具有一致性等。如果数据点较少，剔除异常数据则不利于统计分　析或影响分析结果的稳定性;如果数据集为时间序列数据，剔除异常数　据后往往缺少一些年份的数据，很难开展相应的统计分析。面临这样的　情况，需要采取一定的方法对异常数据进行平滑处理。

1）截面数据的平滑处理。包括分箱、聚类、回归。

2）时序数据的平滑处理。包括移动平均法和指数平滑法。

⑥数据转换处理。统计数据的变换方法很多，包括综合指数法、均值化、标准化、比重法、初值化、功效系数法、极差变换法等等，大致可概括为四类:指标值域变换法、逆变换法、主成分变换和其他变换法。

⑦平衡协调性检验
利用宏观经济统计数据进行模型分析时，为保证分析结论的有效性，分析前必须对数据进行平衡协调性检验。具体可采用以下几种方法:
(1)相关指标据之间的匹配关系。利用国民经济各指标之间存在着一定比例、结构关系，以及一些指标的合理数量界限判定数据质量。
(2)分量指标对总量指标的支撑度判断。测算出分量指标数据所能支撑的总量数值，再将支撑数据与现实数据进行比较。
(3)相关指标间的因果性分析。如果某个变量的统计数据存在异常，利用与其存在因果关系的变量进行推论，并对其进行数据修正。
(4)基于模型的统计数据诊断方法。包括利用数据删除模型、均值漂移模型和方差扩大模型三种线性回归模型，构建学生化残差、预测残差、不相关残差和递推残差对异常点进行诊断。
(5)预测值与实际值的比较。通过模型对相应指标进行预测，即得到该指标在理论上应该达到的数值，然后将此数值与实际汇总的数据对比，以此评价汇总数据与理论值的接近程度。
(6)其他手段。全面调查与抽样调查的结果相验证，投入产出调查与国民经济核算资料相验证，利用统计执法检查的结果对数据进行调整等。

13.统计平均方法有哪些类型或方法，常用平均数公式有什么特点或应用条件？

答：（1）统计平均方法的类型：

统计方法在对社会经济现象进行综合描述与分析以及预测等方面被公认为是最科学、最先进的方法之一,而统计平均数是社会经济统计分析中应用最广泛、最重要的综合指标之一。统计指标的表现形式有三:总量指标、相对指标和平均指标，其中平均指标似乎是最容易受到质疑和责诘的指标。是因为如果平均指标方法运用不当,则会产生不准确的和片面的结果,把人们引入统计的误区。

平均指标是反映在一定时空条件下总体各单位标志值一般水平的代表值，其把各单位标志值的差异抽象化，用以反映具有概括性的总体水平。统计平均数就是将被研究的同类现象的某种数量指标的各个体数量差异抽象化,用一个概括的指标综合说明现象的有代表性的典型水平。统计平均数从随机的角度可以分为非随机平均数和随机平均数。通常意义上的平均数都指是的非随机平均数。在非随机平均数中,按被研究对象性质和反映时间状况不同可分为静态平均数和动态平均数(序时平均数)。静态平均数常用的有数值平均数和位置平均数之分。随机平均数包括期望、随机变量的各阶中心矩、各阶原点矩等。人们言及平均数，多是指算术平均数，其实平均数家族中还有调和平均数、几何平均数，中位数、众数以及平方均数、四分位数、十分位数等。其中包括算术平均数在内的前五种应用最为广泛。

（2）常用平均数公式的特点或应用条件：

①算术平均数：总体各单位标志值之和与总体单位数的比值，它是一种最为直接的“平均”，是把每个单位标志值平均分摊而得到的数值。

②调和平均数：亦称倒数平均数，其内容实质与算术平均相同，只是由于掌握的资料不同而采用不同的计算方法而已。在实际统计工作中即使采用调和平均数的计算方法（求各标志值倒数算术平均数的倒数），一般也冠以“算术平均数”的称谓。

③几何平均数：N 个标志值乘积的 N 次方根，它一般用来求平均比率、平均速度，只有当标志值具有乘积关系时，才使用这种方法。

④中位数：在按大小顺序排列的标志值中处于中点位置上的那个数值。有一半数值比它大，有一半数值比它小，这个排位居中的数值也能够反映总体标志值的一般水平，具有较好的代表性。

⑤众数：指在一群数值中出现次数最多的那个数。当总体各单位的标志值有明显的集中趋势时，众数可作为最为合理的代表值。众数是一个最为直观的平均数。

另外，测定和评价平均数的代表性，还可以计算度量各单位标志值离中程度的标志变异指标。其中标准差就是最常用的一个指标。标准差度量的是各标志值与平均数之间的平均距离，其计算方法是求各标志值与算术平均数离差平方的算术平均数的平方根。通过对统计活动的实际考查，我们发现应用平均数时容易发生两类偏误：一是“唯算术平均数是举”,在分析问题时，只是过分偏好算术平均数，没有把它和其他的平均指标或标志变异指标结合运用,容易使人产生疑问,影响了数字的可信度。二是“唯简单平均是用”,在计算相对数或平均数的平均数时，往往只是采用简单平均的方法，忽略了权数问题。统计平均数实际上代表的是一种平均值的思想,统计学上众多理论均运用到平均值的思想,而在实际生活中这种思想更是频繁地被应用。所以我们应当全面正确地理解、掌握和运用统计平均数的理论与方法,走出统计平均数应用的误区，让统计平均数真正发挥其重要的作用。

14.试述敏感性问题的抽样调查技术研究进展及其应用。

答：（1）敏感性问题的抽样调查技术研究进展

敏感问题按照总体特征可以分为两类：属性特征的敏感性问题和数量特征的敏感性问题。属性特征的敏感性问题又有二项选择和多项选择两种情况，是否具有某种敏感属性的情况属于二项选择的属性特征敏感问题，1965 年 Warner 针对该情况提出了 Warner 模型，是随机化应答技术的起点，Warner 模型成功地保护了受访者的隐私。随后 Simmons 等人引入无关问题，提出了无关问题的随机化回答模型，增强了对受访者的隐私保护。Greenberg 等人针对 Simmons 模型中无关问题样本比例未知的情况提出了双无关问题模型，之后又有学者提出隐含的随机化回答模型及一系列改进的随机化模型，孙山泽等人在 2000 年对二项选择敏感性问题调查的基本方法和改进方法进行了总结。范大茵针对有多种备选的属性特征敏感问题提出了间接调查法，吕恕在 1994 年对其进行了改进，并运用蒙特卡罗法对两个方法进行了对比。此后，2000 年孙山泽等对多项选择敏感性问题的一样本模型和多样本模型做出了总结和介绍。数量特征敏感问题的解决是建立在属性特征敏感问题的基础上的，Greenberg 等人针对数量特征敏感性问题提出了无关问题模型、转移模型、加法模型、乘法模型、随机截尾模型，孙山泽等人在 2000 年前后对这些模型做出了总结。但随机截尾模型不能很有效地保护受访者的隐私，它并不是一种真正的随机化应答方法，因此顾震环等人提出了随机截尾的 Warner 与 Simmons 模型，既保留了随机截尾模型的优势，又保护了受访者的隐私，徐春梅在 2006 年改进了随机截尾模型。

近年来，随机化应答技术在各方面的研究都已经很充分，该技术在实际调查中的应用十分广泛，基于此的实证研究已有很多。Lensvelt-Mulders 等人在 2005 年讨论了随机化应答技术研究的两个元分析。但随机化应答技术本身有着一定的局限性：问卷调查缺乏再生性、随机装置不被受访者信任、随机装置使得调查成本增加等，为克服这些问题，学者们近年来从新的角度探讨敏感性问题的调查技术。2007 年田国梁等人提出了不需要随机化装置的非随机应答模型，这是一种新的敏感性问题问卷调查方法。接着他们又提出了三角模型和交叉模型这两个新的非随机化应答模型，不仅得到了敏感性问题中人群比重的极大似然估计和方差，还探究了这两个模型有效参数的取值范围，进一步比较两者效率。随后田国梁等人又将贝叶斯方法引入非随机化应答模型中对参数的估计，得到了可靠的后验分布及后验矩，并利用 EM 算法推导出后验模型，讨论获得独立同分布后验样本的方法。

非随机化应答技术克服了随机化应答技术的一些问题，不再需要随机化装置，也使得调查具有再生性，既可以用于面对面调查，还可以结合网络进行邮件问卷调查。非随机化应答技术近些年才发展起来，比之随机化应答技术还有很多方面需要研究，对模型本身的改进及研究和利用模型进行实证分析都有巨大的发展空间。

（2）敏感性问题调查方法的应用

分类特征敏感性问题，根据其所提供答案的数目又可进一步分为二项选择（两分类）敏感问题和多项选择（多分类）敏感问题。它常用来了解被调查者是否具有敏感问题的特征，并估计具有敏感问题特征的人在总体中所占的比重，故也可称为敏感性比例问题。数量特征的敏感性问题是指被调查者具有敏感性问题数量大小的特征,一般是估计敏感性数量的均值,也可称作敏感性均值问题。若采用直接提问的方式来调查敏感性问题,很难取得被调查者的合作而获得真实资料。被调查者为了保护自己的隐私或出于其他目的,

往往会拒绝回答或故意说谎。一方面，具有敏感特征的调查对象倾向于拒绝回答敏感性问题，如果直接用应答者的结果来推断整个研究人群的特征，就会产生偏倚，这种偏倚属于典型的无应答偏倚。另一方面，即使调查对象做出应答，在回答敏感问题时也往往受到特定社会倾向的影响，最常见的是社会期望反映定势,即应答者不是按自己的真实情况来回答，而是根据社会期望的取向来回答问题。一般而言，这些偏倚的方向多为负偏倚，对于敏感性问题调查，故意说谎偏倚程度往往比无应答偏倚更严重。这两种偏倚都会影响敏感性问题调查结果的可靠性、真实性。由此可见,对于敏感性问题,必须采取特殊的、科学的方法来提高调查对象的应答率、降低或消除不真实回答率，保证调查结果的真实可靠。目前随机应答技术被认为是最能有效保护被调查者隐私，提高其真实回答率的一种方法。随机应答技术是指在调查过程中使用特定的随机化装置，使被调查者以一个预定的基础概率 P 从两个或两个以上的问题中选择一个问题进行回答，除被调查者以外的所有人（包括调查者）均不知道被调查者的回答是针对哪一个问题，以便保护被调查者的隐私，最后根据概率论的知识计算出敏感问题特征在人群中的真实分布情况的一种调查方法。这一技术能够最大限度地保护被调查者的隐私,易取得被调查者的信任，从而获得真实可靠的资料。与敏感性问题的类型相对应，随机应答技术按其所调查的敏感性问题类型有二项选择敏感问题随机应答模型，多项选择敏感问题随机应答模型以及数量特征敏感问题随机应答模型。国外已将各种随机应答模型应用于流产率、女性饮酒量以及偷税漏税等各类敏感性问题的调查过程，并得到了较好的调查结果。国内关于随机应答技术的研究则主要集中于随机化设计和随机化装置的改进等理论研究，以及我们项目组进行的二分类敏感问题的分层整群抽样调查研究和数量特征敏感问题的整群抽样调查研究。

15.大数据只要相关不要因果，谈谈你对这句话的理解及看法。

答：（1）大数据的非精确性使得人们从对因果关系的追求中解脱出来，转而更多地去探究事物间的相关关系。由于人们的思维对因果关系的解释带有很强的偏见，所以当回归效应出现时，对其按照因果关系进行解释往往自动激活，用因果关系解释回归效应虽能得到现实的认同，实际上却是错误的。

（2）相关分析与因果分析不是互相对立的，而应相互补充。大数据时代，建立在相关分析基础上的预测正是大数据的核心议题，人们可以通过大数据技术挖掘出事物之间隐蔽的相关关系，获得更多的认知与洞见，进而捕捉当下特征和预测未来趋势。通过大数据关注线性相关关系及复杂的非线性相关关系，人们可以看到很多以前不曾注意的联系，掌握以前无法理解的 复杂社会经济现象，甚至可以超越因果关系，成为了解这个世界的更好视角。正如舍恩伯格指出，大数据让人们关注相关关系，只需知道“是什么”，而不用知道“为什么”。

（3）人们不必非得知晓事物或现象背后的复杂深层原因，只需通过大数据分析获知“是什么”就能提供一些新颖且有价值的观点、信息和知识。在大数据时代，人们的思维方式一定程度上从因果思维转向相关思维，颠覆千百年来人类形成的传统因果思维模式。大数据的价值在于预测，而预测正是建立于相关分析基础之上。相关关系通过识别有用的关联物来帮助人们分析一个现象，而不是通过揭示其内部的运作机制，故不具有必然性，只具有或然性。利用大数据对相关关系深入地分析和探究，可以预知将会发生什么，而一旦把因果关系考虑进来，其复杂性要求导致这些视角有可能被蒙蔽，相关关系提供的新视角可以帮助人们去发现在因果思维模式下无法知晓的新领域，这些新视角往往具

有重要的商业价值。

（4）大数据时代的相关分析利用机器计算能力来寻找到最优的关联物，在各个领域都涌现出一些很好的应用成果。例如亚马逊的推荐系统、可视化呈现的数据新闻等，这些应用通过数据挖掘实现从数据到价值的转变，创造出经济利润和社会效益。亚马逊的推荐算法能够根据消费记录来告诉用户可能会喜欢什么，这些消费记录有可能是别人的，也可能是该用户的历史记录。虽说它不能说出你为什么会喜欢，但通过及时推荐就能实现一定概率的购买行为转化，获取经济利润，这便是相关分析优势的最强说服力！相关关系能够创造利润，表明大数据相关分析已不再是计算、统计等学科的专宠，这只王谢堂前燕正式飞人寻常百姓家，为各行各业所广泛应用，以帮助企业盈利，帮助政府决策。

（5）大数据时代为什么人们强调相关性，而弱化因果性呢的原因可能在于，相关性更广泛，因果性更严格，相关性比较表象容易被识别，而因果性反映事物之间内在的本质关系， 不容易被认识和把握。很多日常生活与商业应用中，知晓相关关系就已足够， 相关分析不提供关于世界的真相和原理，只通过知其然而不知其所以然的一些判断来创造属于其自身的价值。在许多场合，只要知道事物之间具有依随性质的相关关系，大致能够推断出与之相关的另一个现象或变量可能会发生的变化，从而抓住商业应用的机会。

（6）尽管对相关关系的探测颇具价值，但相关分析只停留于数据表面，即使相关性很强的对象之间也可能并不存在本质上的关联性。因此，当面对具体的大数据应用时，因果思维仍会不由自主地走上台前，让人们自然而然地想去寻求对象之间的因果联系。

（7）大数据的相关性并不意味着两个变量具有因果联系，而具有因果联系的两个变量从大数据本身来看有时也并不相关。一般来说，相关关系不能确定两个变量 X、Y 之间是否存在因果关系，因为两个变量之间的相关性可能有三种解释：其 一，X 是 Y 的原因或一部分原因；其二，Y 是 X 的原因或一部分 原因；其三，X 和 Y 是第三个变量 Z 的原因(结果)或一部分原因 (结果)。特别是第三种情况的存在，使相关分析得到的相关性很可能是“伪相关”。

（8）很多情况下，相关关系并不是大数据洞察的终结目标。 因果分析是相关分析的深化，大数据的相关关系不仅没有替代因果关系，反而给因果关系的研究提供了更广泛的发展空 间。

16.社会网络分析方法的数据表现是什么？基于社群图，社会网络有哪些重要的统计特征？如何测量？

答：（1）社会网络分析方法的定义以及数据表现：

社会网络分析是研究一组行动者的关系的研究方法。社会网络分析是用于研究行动者及其之间的关系的一套规范和方法，是一种定量的群体交互行为研究方法。SNA 以数据挖掘为基础，采用可视化的图以及社会网络结构的形式表示。并利用此建立社会关系模型、发现社群内部行动者之间的各种社会关系。

数据表现形式为社会图群（可视化）以及矩阵代数（可测量）。

（2）统计特征

特征指标主要包括中心度、中间中心度、网络规模、网络密度、平均路径长度、聚集系

数、小世界值等，用来衡量网络的结构特征与网络各要素之间的关系。（主要从两个层面对社会网络模型进行特征分析:第一个层面为网络整体层面的特征，主要包括网络完备度、网络关联度等特征，前者主要通过网络规模与网络密度等指标反映，后者主要通过平均路径长度、聚集系数、小世界指数等指标反映。第二个层面为网络中构成要素层面的特征，主要包括节点的重要度或等级度，主要通过节点度、中间中心度等指标反映。）

（3）测量方法：

①离心度：从一个节点所有可以到达的节点中，找出最长的最短路径。即一个节点所能达到的最大的最短路径。

②特征向量中心性：一个节点的重要性既取决于其邻居节点的数量(即该节点的度)，也取决于其邻居节点的重要性

③图密度：实际有的边数与最大可能边数之比

④网络直径：一个网络中，所有最短路径的最大值。
⑤平均路径：一个网络中，所有最短路径之和的平均值等于这个网络的平均路径长度。平均路径长度是整个网络的一个指标。

⑥中心性：计算出网络直径等网络的边的特性，就可以计算出中介中心度、亲密中心度

⑦度中心性：单纯的数量来衡量。又叫点度中心度，度越多，就越大。

⑧接近中心性：一个节点能到达节点的数量除以所能到达节点的最短路径之和。

⑨中介中心性：统计某节点被其他节点以最短路径通过的数量与图中最短路径总数之比

17.试对统计指数编制方法与指数因素分析体系的最新进展进行评述。

答：（1）统计指数分析的重要性：

在统计学发展过程中，统计指数分析一直占据重要地位，是重要的宏观经济分析工具。指数在概念上有广义和狭义之分，其中，广义指数是指所有用以表明经济现象总体变动的相对数；狭义指数是用来综合反映在不同空间、时间上的复杂社会经济现象的变动相对数。

（2）统计指数编制方法与指数因素分析体系的最新进展

指数是国际统计学界和经济学界一个非常活跃的研究领域，近年来指数理论研究和编制实践取得了很大的进展。从指数构建方法、基本价格指数计算方法、链式指数、最佳指数、Hedonic 质量调整指数，大数据应用于价格指数的编制等方面是国际上指数理论与实践的动向。指数理论与实践还面临巨大的挑战，包括价格指数编制中的质量调整和大数据应用、房地产价格指数编制与 CPI 中自有住房的处理、季节性产品的处理、服务价格指数的编制等。

国外的因素分析法主要是基于十九世纪下半叶的拉氏指数和派氏指数，以及 20 世纪初的费暄理想指数等。国内的研究主要集中在函数指数理论、共变影响因素理论和共变影响分配理论，而共变影响分配理论又分为增长速率分解法和积分因素分析法等。

18. 从利益相关者视角谈谈统计数据质量的维度、评估方法及提升途经。

答：统计数据质量涉及到数据的生产者、使用者和提供者等直接参与主体及众多的外部监督主体，并非统计部门一家所能决定，开展统计数据质量控制需要综合考虑这些不同主体的影响。

概括起来，统计数据质量涉及的利益相关者主要包括统计机构、数据用户、被调查者及以媒体为代表的监督机构等，统计机构是统计数据的生产者，数据用户是统计数据的使用者，被调查者是统计数据的提供者，外部监督机构则是统计数据质量的监督者。

立足于利益相关者视角且将操作层面的数据质量评估方法与战略层面的数据质量管理相结合，由单一面向统计产品的质量评估扩展为将统计过程与统计产品相结合的全面质量管理，有利于拓宽提高统计数据质量的理论视野和现实途径。

（1）利益相关者视角的数据质量维度

不同的利益相关者对数据质量具有不一样的理解，所关注的质量维度也各有特点。

表1　利益相关者视角的统计数据质量维度

利益相关者	使用者	生产者	提供者	监督者
质量维度	准确性	客观性	保密性	前三类主体质量维度的综合
	及时性	经济性	简便性	
	完整性	可解释性	反馈性	
	适用性	有效性		
	可比性			
	可获取性			

①使用者视角的数据质量维度

统计数据质量因使用者的需求不同、角度不同而有不同的理解和看法，但广泛的共识包括数据的准确性、及时性、适用性、可比性、获取性和完整性等要素。准确性是统计数据使用者的首要要求，而及时性则是统计数据发挥信息功能的必要条件。完整性和可比性用户是统计数据质量的内在要求，适用性则是针对统计数据的最终需求而的必备品质，也是统计工作的最终目的。可获取性是指用户从统计部门取得统计信息的容易程度，完整性则是统计数据量上的要求。这几大质量特性相互联系、相互制约、相辅相成，共同构成了使用者视角统计数据质量的完整内涵，充分体现了统计的科学性。

②生产者视角的数据质量维度

使用者是作为统计数据的需求方提出对数据质量的要求和标准，而除了满足需求方的要求外，统计数据还要切实考虑到供给方即数据生产者－－统计机构的供给能力。从生产者视角提出的统计数据质量要求主要包括以下几个方面： 客观性、经济性、可解释性、有效性等。客观性是指统计调查机构在统计数据加工整理和公布过程中应该遵守客观性原则；经济性是指统计数据的生产要注意投入与产出、费用与效用的比较，以尽量少的劳动消耗提供更多更好符合社会需要的统计信息产品；可解释性是指在公布统计数据时，应同时公开相应的补充信息或称为“源数据” ，即关于统计数据的解释说明；有效性是指应降低统计工作的生产成本，提高效率。

③提供者视角的数据质量维度

好的质量不仅是满足用户的需要，而且要顾及到被调查者接受调查的负担和在保密性方面感到的顾虑。从被调查者角度看，统计认识主体与客体是一对矛盾，统计认识与被认识、反映与被反映都是这种矛盾的具体体现。从提供者视角提出的统计数据质量要求主要包括以下几个方面：保密性、简便性、反馈性等。保密性是与被调查者隐私相联系的概念，指防止已收集的被调查者信息未经授权被使用的程度；简便性是指统计机构应充分利用现有的行政记录资源，减少重复统计，统计调查表简单明了，并使用先进的电子技术和新的统计方法，最大限度地减轻社会调查负担；反馈性则指信息的提供者同时也应是信息的获取者，通过信息共享使所有参与者能够从共享中得到最大的收益。

④监督者视角的数据质量维度

监督者视角具有一定的综合性，既可从数据生产者角度开展统计质量管理措施、统计方法的宣传，以普及统计知识、增强社会大众对统计数据质量的认识、避免统计数据的滥用；也可从数据使用者和提供者的角度，对统计数据在准确性、及时性、可比性、一致性、客观性、经济性、可解释性等方面存在的问题进行曝光，达到对数据质量进行监督的效果。

（2）利益相关者视角的统计数据质量评估方法

表2　利益相关者视角的统计数据质量评估方法

利益相关者	质量维度	主要评估方法	说　明
使用者	准确性	误差分析、指标逻辑校验、匹配关系核查、结构比例趋势、因果关系检验、支撑度判断、反常结果判断	侧重于统计产品评估，兼顾统计生产过程评估；以定量评估方法为主，定性评估方法为辅
	及时性	发布及时率、数据发布时滞等	
	适用性	指标使用频率、指标一致性等	
	可比性	口径范围一致性、计算方法可比性等	
	可获取性	数据获取渠道多样性、获取方便程度等	
	完整性	数据供需缺口、数据缺失率等	
生产者	客观性	样本代表性检查、政策与方法透明度等	侧重于统计过程评估，兼顾产品评估；以定性评估方法为主，辅之以定量评估方法
	经济性	统计成本分析等	
	可解释性	指标解释覆盖率与清晰度、数据修订说明等	
	有效性	投入产出分析等	
提供者	保密性	保密措施、泄密率等	将产品与过程评估、定性与定量评估相结合
	简便性	负担率、年均接受调查次数、填表数、填报方式等	
	反馈性	有无反馈、反馈比率等	
监督者	前三类主体质量维度的综合	媒体宣传频率、数据发布及时性、媒体曝光率、媒体意见调查等	以产品评估为主，兼顾过程评估，将定性与定量评估相结合

①使用者视角的统计数据质量评估方法

数据使用者可在数据发布前和数据发布后两个阶段参与到统计数据质量评估中来，统计数据发布前引入使用者开展参与式评价，可了解公众、研究机构等用户对统计数据的需求，避免统计数据的一些明显性错误，提高公众对统计数据的认同度；统计数据发布后，从数据使用者角度，针对不同的质量维度，可利用譬如误差分析、指标逻辑校验、匹配关系核查、结构比例趋势、因果关系检验、支撑度判断、反常结果判断等方法开展统计数据质量的评估，将评估结果反馈到统计数据的周期性调整当中，提高统计数据质量。从使用者视角总体上反映数据质量状况，可开展用户满意度测评。统计数据“用户满意度”则是指统计数据用户对政府统计部门提供的统计数据产品和信息服务满足自己需求程度的判断，是对统计数据质量的一种主观感知。

②生产者视角的统计数据质量评估方法

生产者是作为统计数据的供给方提出数据质量标准，对统计数据质量的要求主要包括客观性、经济性、可解释性、有效性等方面。生产者对统计数据生产过程的熟悉，掌握的原始数据资料详实，由于存在发布的时间要求，考虑时间上的要求不可能充分进行相关性等考虑。从数据生产者角度看，统计数据发布之前主要依靠统计机构自身的力量，对照相应的数据质量维度，通过样本代表性、政策与方法透明度、统计成本分析、指标解释覆盖率与清晰度、数据修订说明、投入产出比及用户需求调查等方法与指标，对统计数据质量进行综合性评价。

③提供者视角的统计数据质量评估方法

统计数据的提供者一般是指调查对象即被调查者。从被调查者角度来看，统计机构必须考虑到被调查者的利益要求，为提供资料的被调查者保密，尽量减少被调查的负担，体现在数据质量特征上便是保密性、简便性的要求。从使用者视角考虑，一方面主要通过调查问卷中的敏感性问题脱敏处理、缺失值处理及被调查者态度调查等方法进行保密性评估；另一方面则积极加强统计信息管理系统建设、推行企业一套表制度、加强统计机构与其他政府部门的协调、计算平均每月报表处理天数等，减少重复统计，最大限度地减轻调查负担。

④监督者视角的统计数据质量评估方法

监督者视角具有一定的综合性，既可从数据生产者角度开展统计质量管理措施、统计方法的宣传，以普及统计知识、增强社会大众对统计数据质量的认识、避免统计数据的滥用；也可从数据使用者和提供者角度，对统计数据在准确性、及时性、可比性、一致性、客观性、经济性、可解释性等方面存在的问题进行曝光，达到对数据质量进行监督的效果。从监督者视角可计算统计数据的媒体宣传频率、数据发布及时性、数据质量问题的媒体曝光频率等指标，以及开展媒体意见调查等方法进行质量评估。

（3）利益相关者视角的统计数据质量评估形式

统计数据质量评估的具体组织形式包括自我评估、同行评议、质量认证、滚动评估等几种形式，实际中往往是多种形式的综合运用。

19.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？

答：(1)定义：

①地理加权回归（GWR）：试图利用空间的非稳态性，使得变量间的关系随着空间的变化而变化；模型通过对回归方程残差项加权体现其空间变化。

$$y_i = \beta_0(u_i, v_i) + \sum_{j=1}^{k} \beta_k(u_i, v_i) x_{ij} + \varepsilon_i$$

高斯距离权值(Gaussian Distance)

$$W_{ij} = \Phi(d_{ij}/\theta)$$

指数距离权值(Exponential Distance)

$$W_{ij} = \sqrt{\exp(-d_{ij}/q)}$$

三次方距离权值(Tricube Distance)

$$W_{ij} = [1-(\theta/d_{ij})^3]^3$$

②分位数回归模型：分位数回归具有普通最小二乘法与极大似然估计等均值回归方法不具备的优良性质（如稳健性与处理异质性时的优势）。

$$y_{ni} = \lambda_0(\tau)\overline{y}_{ni} + x'_{ni}\beta_0(\tau) + u_{ni}, i = 1, \cdots, n$$

$$\overline{y}_{ni} = \sum_{j=1}^{n} w_{n,ij} y_{nj}$$

$$Q_\tau(u) = 0$$

（2）区别：在扩展的地理加权模型模型中，特定区位的回归系数不再是利用全部信息获得的假定常数，而是利用邻近观测值的子样本数据信息进行局域回归估计而得到、随着空间上局域地理位置变化而变化的变数，较其它刻画空间异质性的方法相比，GWR 模型简单且易于操作，并能将模型系数的估计结果显示在图形上，直接和直观的刻画空间的非平稳性。分位数回归模型的其本质是通过分位数取 0－1 之间的任何小数，调节回归平面的位置和转向，从而让自变量估计不同分位数的因变量。分位数回归的回归系数的估计量具有最佳线性无偏性，并且能够更好地描述自变量对因变量的条件均值的影响过程。

（3）联系：地理加权模型与分位数回归模型都是对一般回归模型的扩展。他们的模型结构是建立在一般回归模型的基础上，然后针对不同的研究方向进行不同角度、不同方法的改进。

20.谈谈可视化技术主要内容及其在统计学中的可能应用

答：（1）可视化技术主要内容

可视化是利用计算机图形学和图像处理技术，将数据转换为图形或图像在屏幕上显示出来，并进行交互处理的理论、方法和技术。它涉及到计算机图形学、图像处理、计算机视觉、计算机辅助设计等多个领域，成为研究数据表示、数据处理、决策分析等一系列问题的综合技术。种类繁多的信息源产生的大量数据远远超出了人脑分析解释这些数据的能力，可视化技术作为解释大量数据最有效的手段而率先被科学与工程计算领域采用，并发展为当前热门的研究领域——科学可视化。科学可视化的主要过程是建模和渲染。

（2）可视化在统计学中的可能应用

①应用领域一：宏观态势可视化

态势可视化是在特定环境中对随时间推移而不断动作并变化的目标实体进行觉察、认知、理解，最终展示整体态势。此类大数据可视化应用通过建立复杂的仿真环境，通过大量数据多维度的积累，可以直观、灵活、逼真地展示宏观态势，从而让非专业人士很快掌握某一领域的整体态势、特征。如全球航班运行可视化、卫星分布运行可视化。

②应用领域二：设备仿真运行可视化

通过图像、三维动画以及计算机远程控制技术与实体模型相融合，实现对设备的可视化表达，使管理者对其所管理的设备有形象具体的概念，对设备所处的位置、外形及所有参数一目了然，会大大减少管理者的劳动强度，提高管理效率和管理水平。如工业设备运行可视化、军工领域战场设备可视化、卫星运行可视化。

③应用领域三：数据统计分析可视化

此领域是目前媒体大众提及最多的应用，可用于商业智能、政府决策、公共服务、市场营销等领域。如商业智能可视化、精准营销可视化、智能硬件数据可视化。