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。【上一届＋gpt】

（1）有关分类的统计方法：

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）重要意义：

①数据理解和描述：通过分类，我们能更好地理解数据的结构和特征，从而为数据提供更准确、更有说服力的描述。

②预测和决策：分类方法可以用于预测未知数据的类别，从而支持决策制定。例如，在医学领域，可以使用分类模型预测患者是否患有某种疾病。

③模式识别：通过分类，我们可以识别数据中的模式和趋势，这有助于更深入地理解数据生成的过程。

④优化资源分配：在商业和工业应用中，分类可用于优化资源的分配，例如客户分类、产品质量控制等。

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。【李金昌-统计学＋苏为华-统计指标理论与方法研究＋gpt】

（1）类型：

①统计指标按其计算的范围不同，可以分为总体指标和样本指标。

总体指标根据（有限）总体中所有个体的标志表现综合计算而得，反映总体数量特征；样本指标则仅根据总体中部分个体的标志表现综合计算而得，反映样本数量特征。总体指标也称总体参数，对于某一确定的总体，任何一个总体指标的数值是惟一的，但在非全面观测的情况下是未知的。样本指标也称样本统计量，对于所抽的一个样本来说，任何一个样本指标都有一个可知的数值，但由于样本是随机抽取的、非惟一的，因此样本指标的数值随样本不同而不同，样本指标是随机变量。统计研究的一大任务，就是要用可知但非惟一的样本指标数值去推断惟一却未知的总体指标数值。

②统计指标按其反映现象的内容不同，可以分为数量指标和质量指标两种。

•数量指标也称为总量指标，它是反映现象总体某一方面绝对数量特征的指标，表明现象所达到的总规模、总水平或工作总量。例如人口数、企业数、总产量、总产值、土地面积、投资额等，都属于数量指标。数量指标的计量单位有实物单位、价值单位和时间单位三种，其中实物单位又有自然单位、度量衡单位、双重单位和复合单位等。数量指标按照其反映现象内容的不同，可以分为总体标志总量和总体容量两种：总体标志总量是总体中所有个体的某个标志的标志值之和，也即某变量的所有变量值之和；而总体容量则是总体所包含的个体数，也即某变量的变量值个数。显然，总体标志总量中的标志值（变量值）与总体容量中的个体是一一对应的，所以总体容量越大（小），总体标志总量一般也就越大（小）。在通常情况下，数量指标主要是指总体标志总量。数量指标按照其反映现象时间状况的不同，又可以分为时期指标与时点指标两种：时期指标是反映现象在一定时期内累计达到的总量，其数值大小与时间长短有直接关系，不同时间上的数值可以累加，数值需要通过连续登记取得，例如企业产量、地区 GDP 等指标；时点指标是反映现象在某一时点（时刻、瞬间）所达到的总量，其数值大小与时间长短无直接关系，不同时间上的数值不可以累加，数值通常不需要通过连续登记取得，例如企业人数、地区居民存款余额等指标。数量指标的具体结果就是前述的绝对数。

•质量指标是反映现象总体内在对比关系或总体间对比

关系的指标，表明现象所达到的相对水平、平均水平、工作质量或相互依存关系。例如人口性别比例、职工平均工资、产品合格率、人均土地面积、产值增长速度、资金利润率等，都属于质量指标。质量指标又可分为相对指标和平均指标两种。相对指标是反映事物内部或相关事物之间相对数量关系的指标，是两个有联系的统计指标对比的结果，包括结构相对指标（总体中部分总量与总体总量之比）、比例相对指标（总体中某部分总量与其他部分总量之比）、比较相对指标（两个同类指标之比）、动态相对指标（同一指标在不同时间之比）、强度相对指标（两个性质不同但有联系的总量指标之比）和计划完成程度相对指标（实际指标与计划指标之比）等；平均指标是反映变量分布集中趋势或中心位置的指标，表明变量的一般数量水平，包括算术平均指标、几何平均指标、调和平均指标、众数指标和中位数指标等。平均指标和相对指标中的一部分强度相对指标有计量单位，其他相对指标则没有具体的计量单位。相对指标和平均指标的具体结果分别就是前述的相对数和平均数。

• 由于数量指标（主要是总体标志总量）的数值大小一般与总体容量大小有关，所以又称为外延指标；而质量指标的数值大小一般与总体容量大小无直接关系，所以又称为内涵指标。在三大类统计指标中，总量指标是基础，相对指标和平均指标由总量指标派生而来。

③统计指标按其反映现象的时间状态不同，可以分为静态指标和动态指标两种。

静态指标是反映现象总体在某一时点或相对静止时间上数量特征的指标，包括一般的总量指标、静态相对指标和一般平均指标。动态指标是反映现象总体在不同时期或时点上发展变化情况的指标，包括增长量指标、动态相对指标和序时平均指标等。

（2）重要地位：

• 统计指标是社会经济统计学的四大基本范畴(总体与总体单位、指标与标志)之一，也是政府统计活动成果的基本信息单元。研究统计指标理论，不仅对于社会经济统计学原理的发展与完善有着重要的理论意义，对于社会科学理论研究中的实证分析(如经济问题、社会问题、人口问题等的实证分析)同样具有非常重要的意义。

• 在社会经济统计活动实践中，指标是最基本的统计成果，也是最基本的统计分析方法。无论是一般问题的研究(如经济学家、社会学家、人口学家等从事社会经济问题的实证分析)，还是官方统计活动，都离不开统计指标。统计指标理论就是要寻求设计统计指标的一般方法，为统计实践提供理论指导。

• 对于以下方面统计指标理论拥有重要意义：

①定义和测量社会经济概念： 统计指标理论提供了一个框架，帮助研究人员和政策制定者明确定义和测量社会经济概念，如生产总值（GDP）、失业率、贫困率等。这些指标的清晰定义和测量对于准确描述社会经济状况至关重要。

②数据收集和采样方法： 统计指标理论涉及到如何设计有效的数据收集和采样方法，以确保得到具有代表性的样本。这对于获取可靠和准确的社会经济统计数据至关重要，从而支持对整体经济和社会状况的评估。

③数据处理和分析： 一旦收集到数据，统计指标理论提供了一系列的方法和技术，用于数据的处理和分析。这包括了各种统计指标的计算、趋势分析、相关性分析等，帮助揭示社会经济现象的模式和关系。

④比较和国际标准： 统计指标理论提供了一种标准化的方法，使得不同国家、地区或时间点的社会经济指标可以进行比较。这有助于国际间的经济和社会问题的研究，以及制定国际政策。

⑤决策支持： 统计指标是政策制定的基础。在社会经济领域，政府和其他组织使用各种指标来监测经济状况、社会福祉和不平等，以便制定政策、规划资源分配和评估政策效果。统计指标的准确性和可解释性对于有效的政策决策至关重要。

3. 谈谈你对统计思维的理解以及提升统计思维的路径。【新，文献＋gpt】

（1）理解：

统计思维是在获取数据、从数据中提取信息、论证结论可靠性等过程中表现出来的一种思维模式，对于人类提高认知起到巨大的作用。以下是统计思维的一些关键特征：

①问题导向：统计思维的起点通常是一个问题或者假设。它强调从实际问题出发，通过收集和分析数据来回答问题或验证假设。

②数据的重要性：统计思维认为数据是解答问题的关键。了解数据的生成过程、收集方法以及数据的特征有助于更好地理解问题和提高分析的准确性。

③不确定性的考虑：统计思维接受和处理不确定性，它考虑到随机性和变异性，通过概率和统计方法来描述和量化不确定性，从而做出更为可靠的推断。

④模型的使用：统计思维借助统计模型来概括和描述数据背后的规律。这些模型可以是简单的描述性统计，也可以是复杂的统计学模型，用以理解和预测现象。

⑤推断和泛化：统计思维强调从样本推断到总体，并且致力于从局部现象推广到更广泛的背景。这包括参数估计、假设检验等统计推断的方法。

⑥实证主义的观点：统计思维注重实证主义的研究方法，即通过实证证据来验证理论或假设，强调实际观察和实验的结果。

⑦跨学科性：统计思维通常需要在多学科背景中应用。它与科学、工程、医学、社会科学等领域相互交融，帮

助解决复杂的现实问题。

（2）大数据下的统计思维：大数据时代的统计首先要适应三个重大的思维转变。

①分析与事物相关的所有数据，而不是依靠分析少量的样本数据。统计往往希望用尽可能少的数据来证实可能重大的发现、假设等，小数据时代一般采用随机采样，用最少的数据获得最多的信息。在处理大数据时不再采用随机抽样的方法，而利用所有数据进行分析。分析整个数据库，而不是对一个样本进行分析，能够提高微观层面分析的准确性，甚至能够推测出任何特定尺度的数据特征。

②乐于接受数据的纷繁复杂，而不再追求精确性。对小数据而言，最基本、最重要的要求是减少误差，保证数据质量。大数据时代，随着数据规模的扩大，人们对数据精确度的痴迷将逐步减弱。

③不再探求难以捉摸的因果关系，转而关注事物的相关关系。在小数据时代，人们往往乐此不疲地想知道现象背后的原因。大数据时代，由于坐拥海量数据和良好的机器计算能力，相关关系分析为人们提供了一系列新的视野和有用的预测，能够找出新种类数据间的相互联系来解决日常需要。

（3）路径：

①理解统计思维的本质：统计思维并不仅仅是计算和应用统计方法，更重要的是理解数据的背后含义。统计思维是一种推断和决策的过程，通过收集、分析和解释数据来理解现象并做出合理的结论。

②注重问题背后的数据：统计思维强调问题导向，你需要学会看待问题并理解数据如何帮助你回答问题。了解数据的生成过程，收集和整理数据是培养统计思维的第一步。

③学会提出问题：不仅仅是回答问题，更是提出合适的问题。良好的统计思维始于对问题的深刻理解，因此你需要学会通过问题来引导你的数据分析。

④深入了解统计原理：虽然有各种现成的统计工具，但深入理解统计学原理对培养统计思维至关重要。学习统计理论和模型的基础，能够帮助你更好地理解何时以及为何使用特定的统计方法。

⑤实践是提升的关键：只有通过实际的数据分析项目，你才能真正将理论知识转化为实际技能。参与实际的研究项目、实习或者解决实际生活中的问题，这将有助于巩固你的统计思维。

⑥跨学科学习：统计学与其他学科密切相关。了解与你的研究方向相关的其他学科，可以帮助你更好地理解问题，同时也为你提供了更多思考问题的角度。

⑦利用工具和技术：学习使用统计软件（如 R、Python 等）是必不可少的。这些工具能够帮助你更高效地进行数据分析，从而更好地培养统计思维。

⑧参与学术社区：加入统计学或相关领域的学术社区，与同行交流，参与研讨会和会议。这有助于拓展你的视野，了解最新的研究动态，同时也能够建立有益的学术网络。

•总之，统计思维的养成不但需要学习一些具体的指示，还要能够从发展的眼光，把这些指示连缀成一个有机的、清晰的图景，获得一种历史的厚重感，透过统计思维理解科学事实的应有态度。

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。【文献】

（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。现有的测验方法体系包括：时间(基位)互换测验和弱时间(基位)互换测验，因素互换测验和弱因素互换测验，循环测验，比例性(甲均数)测验，中值测验，恒等(一致性)测验，单调性测验，公度性(单位共通性)测验，扩展性测验，正定性(非负性)测验，线性齐次测验，增删(进退)测验，确定性测验。这一体系可能存在的问题是：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。

（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关于指数的数学形式及其性质的讨论有一个共同的特征，就是带有浓厚的数学形式主义和主观主义色彩。因此，它们不能科学地解释各种测验的实际经济分析涵

义，更无法正确地阐明评价经济指数的基本标准是什么，以及不同性质的浏验在指数的评价方法体系中占有何种地位。指数的数学性质测验只有与具体的经济分析问题联系起来，才有实际意义，从而构成经济指数评价方法体系中的一个有机部分。

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？ 【李金昌-统计学＋gpt】

（1）指标：

①水平指标：

• 发展水平指标：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

• 平均发展水平指标：对不同时期的发展水平加以平均

• 增长量指标：现象在一定时期内发展水平增加或减少的绝对数量，反映现象数量变动

• 平均增长量指标：各期逐期增长量相加除以其个数，说明现象在一定期内平均每期增加的数量

②速度指标：

• 发展速度指标：报告期发展水平与基期发展水平的比

• 增长速度指标：报告期增长量与基期发展水平之比

• 平均发展速指标：计算以发展速度构成的序列的平均水平

• 平均增长速度指标：环境增长速度的统计平均，说明现象在较长时期内逐期平均增长的相对程度

（2）新思路或新方法：

①深度学习在时间序列中的应用：随着深度学习的兴起，尤其是循环神经网络（RNN）和长短时记忆网络（LSTM）等模型的发展，深度学习在时间序列分析中的应用越来越受到关注。这些模型能够捕捉序列中的长期依赖关系，对于预测和分类等任务表现出色。

②因果推断和 Granger 因果关系：因果推断在时间序列分析中变得越来越重要。Granger 因果关系测试是一种常见的方法，近年来有关如何更好地理解和应用因果关系的研究也在增加。

③非线性时间序列分析：非线性时间序列模型的研究日益增多，因为一些现象在其动力学中可能包含非线性成分。这包括使用分岔理论等方法来理解非线性系统中的不稳定性和复杂性。

④高维时间序列分析：随着大数据时代的到来，高维时间序列的分析变得更加关键。这涉及到如何有效地处理大规模和高维度的时间序列数据，以便提取有意义的信息。

⑤时空数据集成：一些新的研究涉及时空数据的集成，即如何有效地结合时间序列和空间信息。这对于城市规划、气象学和环境科学等领域的应用具有重要意义。

⑥无监督学习在异常检测中的应用：无监督学习方法在时间序列异常检测方面的应用也越来越受到关注。这包括使用聚类、自编码器等技术来发现时间序列中的异常模式。

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。【孙敬水-计量经济学＋gpt】

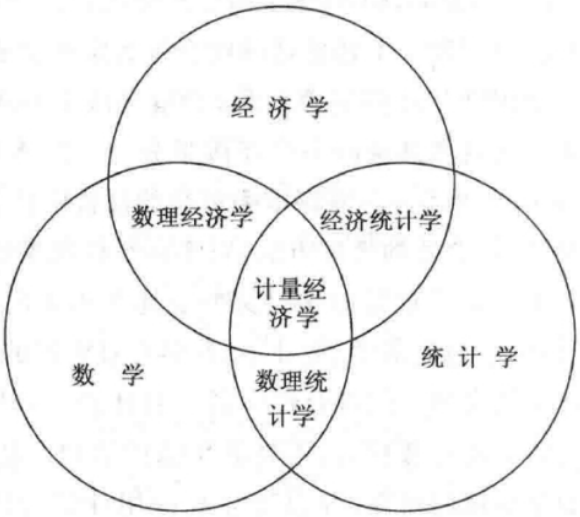


图 1.1.1 计量经济学与相关学科的关系

• 图 1.1.1 表明计量经济学是数理经济学、经济统计学和数理统计学的交集，而数理经济学是经济学与数学的交集，数理统计学是数学和统计学的交集，经济统计学是经济学与统计学的交集。显然，每一交集都形成了一门特定的学科，有其独立的研究对象或特点，这些特定学科彼此不能混淆或替代。

• 经济学与统计学结合形成了经济统计学。经济统计所关心的是描述性的统计量，着重于收集，整理并以图表的形式表达数据，并不利用所收集的数据来验证经济理论。而计量经济学则利用经济统计所提供的数据来估计经济变量之间的数量关系并加以验证。

• 数理统计学为各种类型数据的收集、整理与分析提供切实可靠的数学方法，是计量经济学建立计量经济模型的主要工具。但是数理统计学在研究变量之间的关系时，要求各种变量必须服从某种规律，即服从某种分布。在现实经济生活中，各经济变量很难完全满足这一假定，但又必须研究经济变量之间的关系，所以计量经济学必须在数理统计方法技术的基础上，开发出特有的分析方法技术。

• 计量经济学的主要特征在于利用由数理经济学提出的数学方程及实际数据来验证经济理论，模型所反映的经济变量间的关系是非确定性的、随机的相关关系。

• 经济统计学提供了计量经济学所需的实证数据。

• 数理统计学为经济统计学和计量经济学提供了理论和方法的支持，帮助研究者从数据中得出结论。

• 计量经济学利用数理统计学的方法，将经济理论转化

为可测试的模型，并运用经济统计学的数据进行实证验证。

• 计量经济学是经济理论、统计学和数学三者的统一。计量经济模型建立的过程，是综合应用经济理论、统计学、数学方法的过程。理论模型的设定、样本数据的收集是直接以经济理论为依据，建立在对所研究经济现象的透彻认识基础上的，而模型参数的估计和模型有效性的检验则是统计学和数学方法在具体经济研究中的具体应用。没有理论模型和样本数据，统计学和数学方法将没有发挥作用的“对象”和“原料”；反过来，没有统计学和数学所提供的方法，原料将无法成为“产品”。因此，计量经济学广泛涉及了经济学、统计学、数学这三门学科的理论、原则和方法，缺一不可。

7. 为什么会出现基础比率谬误？如何避免该谬误？【文献】

（1）定义：

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”。

（2）原因：

①代表性启发策略。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。

②相关性原则。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。

③因果基础概率与偶然基础概率的差异。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。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。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）措施：

①用贝叶斯定理来约束直觉。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。

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。

③判断与决策时不要被很细节的情境所迷惑。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。【上一届】

（1）模型：

①$SEM: y = X\beta + \mu, \mu = \lambda W\mu + \varepsilon, \varepsilon \sim N(0, \sigma^2 I_n)$

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

③$SEC: y = X\beta + \mu, \mu = W\eta + \varepsilon, \eta \sim N(0, \sigma_\eta^2 I_n), \varepsilon \sim N(0, \sigma_\varepsilon^2 I_n)$

④$SLX: y = X\beta_1 + WX\beta_2 + \varepsilon, \varepsilon \sim N(0, \sigma^2 I_n)$

⑤$FAR: y = \rho Wy + \varepsilon, \varepsilon \sim N(0, \sigma^2 I_n)$

⑥$SAR: y = \rho Wy + X\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2 I_n)$

⑦$SARMA: y = \rho W_1 y + X\beta + \mu, \mu = \varepsilon + \lambda W_2\varepsilon, \varepsilon \sim N(0, \sigma^2 I_n)$

⑧$SAC: y = \rho W_1 y + X\beta + \mu, \mu = \lambda W_2\mu + \varepsilon, \varepsilon \sim N(0, \sigma^2 I_n)$

⑨$SDM: y = \rho Wy + X\beta + WX\theta + \varepsilon, \varepsilon \sim N(0, \sigma^2 I_n)$

⑩$SDEM: y = X\beta + WX\theta + \mu, \mu = \varepsilon + \lambda W_2\mu, \varepsilon \sim N(0, \sigma^2 I_n)$

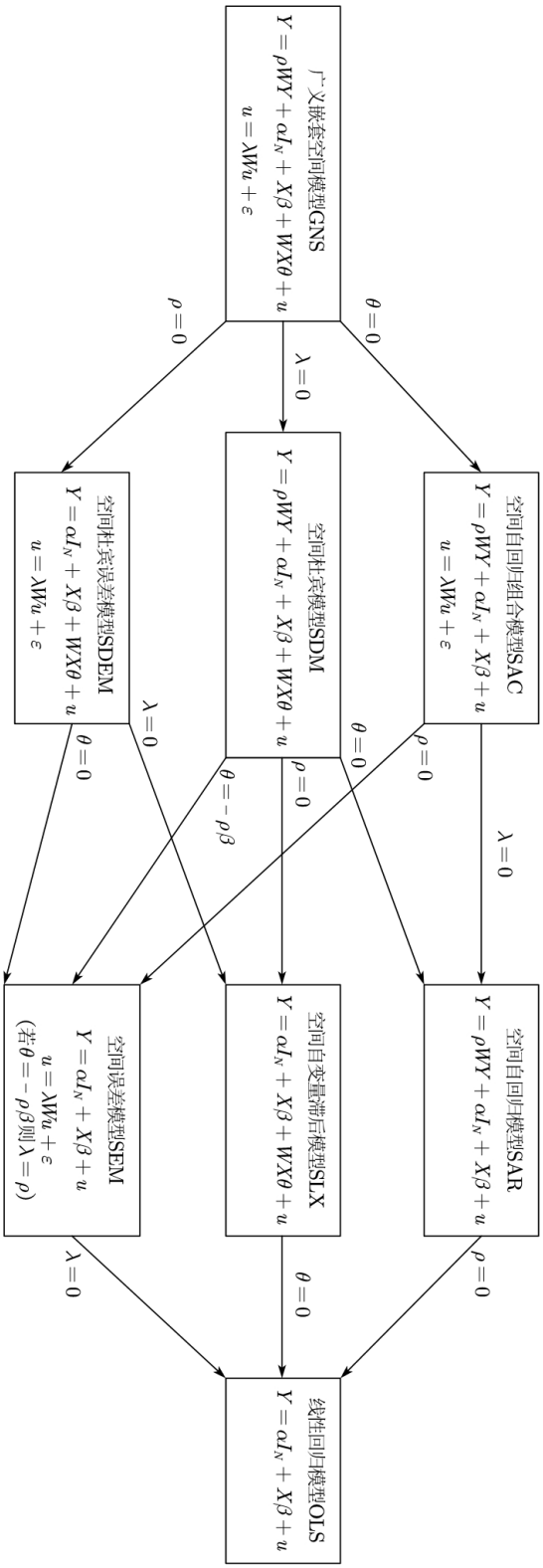
（2）联系：

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①⑥⑨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。【程开明-统计数据质量诊断与管理研究】

（1）原因：

缺失数据的产生机制是通过探讨缺失数据的出现与目标变量是否有关而界定的，如果缺失数据的出现是随机的，则将该类缺失数据的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，称之为不可忽略的。

如果缺失数据的产生不依赖于完全数据，则数据称为完全随机缺失（MCAR）。如果缺失数据的出现是随机的，则将该类缺失数据的产生机制定义为可忽略的，即随机缺失（MAR）。如果缺失数据的产生与研究变量有关，称之为不可忽略的，即非随机缺失（NMAR）。

金勇进等(2009)认为缺失数据的产生机制除以上三类之外，还可分为取决于协变量的缺失(CDM)、取决于随机影响缺失(REDM)和取决于前期数据的缺失等几种。

（2）处理方法：

①事前处理方法：数据收集过程中采取一定的事前预防措施，减少由于无回答、填报和汇总等原因造成的缺失数据。具体包括以下措施：首先，提高调查设计的质量，合理安排调查项目，问题由简单到复杂，由总体到局部，并且努力降低问题的敏感性；其次，加强调查过程的质量控制，特别是对调查员和数据录入、汇总人员的选拔和培训，增强责任心和业务能力；另外，对敏感性问题进行特殊处理。既使被调查者既愿意回答，又能保守个人秘密，可采用随机化回答技术——沃纳模型、西蒙斯模型。

②事后处理方法：1）忽略缺失数据，直接分析。2）开展再调查，补充缺失数据。包括：(1)多次访问；(2)替换被调查单位；(3)对无回答单位进行子抽样。3）利用辅助信息，进行间接估计。包括：(1)加权调整。(2)插补，基本思想是利用辅助信息，为每个缺失值寻找替代值。

具体来说，有 1.个案剔除法；2.加权调整法；3.均值插补法；4.热卡插补法；5.冷平台插补。

10. 当前常用的 *P* 值存在什么问题，如何改进？【文献】

（1）认识误区：

①***P*** 值是原假设成立的概率，**(1－*P*)**值则为备择假设成立的概率。实际上，***P*** 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 $P(D|H_0)$，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 $P(H_0|D)$。而 $P(D|H_0)$ 和 $P(H_0|D)$的现实差异可能很大。

②***P*** 值小于显著性水平α，说明原假设错误，拒绝原假设；***P*** 值大于显著性水平α，即说明原假设正确，接受原假设。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。***P*** 值大于显著性水平α，意味着没有充足的证据拒绝原假设，并不能说明原假设就是正确的；原假设中的参数具有存在的合理性，但不能排除其他参数存在的可能性，所以

不能证明原假设就是准确无误的。没有充分理由拒绝原假设也不意味着必然拒绝备择假设，在拒绝备择假设之前需要考虑第二类错误的严重性，若第二类错误很严重则不能轻易接受原假设。因此，不能依 *P* 值与显著性水平*α*的大小来确定是否一定拒绝或接受原假设。

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。

• 总之，对 *P* 值的误解总体上可概括为两个层面：一是基本层面，简单将 *P* 值视为原假设正确的概率，先将原假设正确引申至备择假设错误，再由原假设正确的概率推断备择假设错误的概率；二是应用层面，由于各领域实际问题的复杂性和不确定性，再加上很多讲授假设检验方法的教师本身对 *P* 值的认识存在偏差，导致应用者开展实际统计分析时对 *P* 值产生较多的认识误区。

（2）局限性：

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。有学者指出，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。

（3）改进：

①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。

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。

④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）。定义为$\mathrm{FDR}=\begin{cases}E\left(\frac{V}{R}\right), & R\neq 0\\ 0, & R=0\end{cases}$，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。

⑤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，*O* 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。

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？【上一届】

（1）影响：

• 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。

• 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。

（2）拓展：

• 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

列(c_j) / 行(r_i)	列(c_j) $j=1$	$j=2$	…	合计
$i=1$	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：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。

• 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。

• 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？【文献】

（1）步骤：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。

①数据审查：该步骤检查数据的数量（记录数）是否满足分析的最低要求，字段值的内容是否与调查要求一致，是否全面；还包括利用描述性统计分析检查各个字段的字段类型、字段值的最大值、最小值、平均数、中位数等，记录个数、缺失值或空值个数等。

②数据清理：该步骤针对数据审查过程中发现的明显错误值、缺失值、异常值、可疑数据，选用适当的方法进行“清理” 使“脏”数据变为“干净”数据，有利于后续的统计分析得出可靠的结论。当然，数据清理还包括对重复记录进行删除。

③数据转换：数据分析强调分析对象的可比性，但不同字段值由于计量单位等不同，往往造成数据不可比；对一些统计指标进行综合评价时，如果统计指标的性质、计量单位不同，也容易引起评价结果出现较大误差，再加上分析过程中的其他一些要求，需要在分析前对数据进行变换，包括无量纲化处理、线性变换、汇总和聚集、适度概化、规范化以及属性构造等。

④数据验证：该步骤的目的是初步评估和判断数据是否满足统计分析的需要，决定是否需要增加或减少数据量。利用简单的线性模型，以及散点图、直方图、折线图等图形进行探索性分析，利用相关分析、一致性检验等方法对数据的准确性进行验证，确保不把错误和偏差

的数据带入到数据分析中去。

（2）方法：

①描述及探索性分析。描述性统计技术主要是对数据开展频数、描述统计量及列联表分析。频数分析是利用非连续变量的频数表，报告出变量个数、记录数，以及缺失值数等；描述统计量分析主要是计算连续变量的均值、标准差、最小值、最大值、偏度、峰度等统计量，以便检查出超出范围的数据或极端值。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。

探索性分析利用图形直观地考察数据所具有的特征，反映数据的分布特征、发展趋势、集中和离散状况等，主要包括茎叶图、箱形图、散点图、直方图、折线图、条形图等。茎叶图把观测数据分为茎和叶两部分，使我们认识到数据接近对称的程度、是否有数据远离其它数据、数据是否集中、数据是否有间隙等特征。箱形图有助于直观地描述分布与离散状况，利用最大值、最小值、中位数、上四分位数和下四分位数等几个值反映出数据的实际分布。散点图用于直观地表现两个或多个变量之间有无相关关系，并反映数据的分布、集中、离散状况；直方图也是评估数据分布的常用图示法：*P*－*P* 图和 *Q*－*Q* 图则可用于展示数据是否符合正态分布，还有折线图、饼图、面积图、雷达图等，都从不同侧面直观地反映出数据的特征、趋势。

②缺失值处理。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；2．插补（替代）；3．再抽样；4．加权调整。

③异常值的处理。异常值又称为孤立点，异常值处理的首要任务是检测出孤立点。由于异常值可能是数据质量问题所致，也可能反映事物现象的真实发展变化，所以检测出异常值后必须判断其是否为真正的异常值。主要分为三类：统计学方法、基于距离的方法和基于偏离的方法。

④数据变换技术。数据变换是通过一定的方法将原始数据进行重新表达，以改变原始数据的某些特征 增进对数据的理解和分析。大致包括以下几类：1．对原始数据重新分类、编码、定义变量和修改变量。2．数据的代数运算。3．数据汇总和泛化。4．属性构造。5．加权处理。

⑤信度和效度检验。信度是指调查统计结果的稳定性或一致性，也就是对同一调查对象多次重复进行调查或测量，所得结果的一致程度，可表示为 *N* 次调查中有多少次是正确的，或每次调查属于正确的概率是多少。信度的度量通常是以相关系数来表示的，又称信度系数，包括重测信度、复本信度、折半信度等，分别可以利用相关分析、计算*α*系数等方法来进行检验。

调查统计资料的效度就是指调查结果反映客体的准确程度，反映出调查问卷本身设计的问题。如果问题设计的科学、合理，能够对调查对象进行很好地测量 那么效度就高，反之，则低。效度具体包括内容、准则和建构三个方面，分别对应内容效度、准则效度和构建效度，可以利用相关分析和因子分析等方法进行检验。

⑥宏观统计数据诊断。数据质量是计量经济模型赖以建立和成功应用的基础条件，确保进行计量分析的宏观数据质量，必须对数据进行严格的诊断。所谓数据诊断是通过适当的理论方法，发现对研究结果的可靠性产生显著不良影响的数据。对于横截面数据的质量诊断主要基于计量模型通过各种诊断统计量来进行，而对于时间序列数据则通过时间序列分析来进行。方法包括：1．分量指标对总量指标的支撑度判断。2．宏观统计数据的因果性分析。3．各专业数据之间的匹配关系判断。4．时间序列的预测值与实际值的比较。5．其他手段。

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？【李金昌-统计学】

（1）数值平均数

①算术平均数：也称为均值，是变量的所有取值的总和除以变量值个数的结果。

1.简单算术平均数 $\overline{x}=\dfrac{\sum x_i}{n}$：根据未分组数据计算的，即直接将变量的每个变量值相加，除以变量值的个数。

2.加权算术平均数 $\overline{x}=\dfrac{\sum x_i f_i}{\sum f_i}$：根据变量数列计算的，即以各组变量值（或组中值）乘以相应的频数求出各组标志总量，加总各组标志总量得出总体标志总量，再用总体标志总量除以总频数。

3.算术平均数具有以下几个优点：一是可以利用算术平均数来推算总体标志总量，因为算术平均数与变量值个数之乘积等于总体标志总量（变量值总和）；二是由算术平均数的两个数学性质可知，算术平均数在数理上具有无偏性与有效性（方差最小性）的特点，这使得算术平均数在统计推断中得到了极为广泛的应用。三是算术平均数具有良好的代数运算功能，即分组算术平均数的算术平均数等于总体算术平均数。

②调和平均数：变量值的倒数的算术平均数的倒数，也称为倒数平均数。但在实际应用中，它则是更多地作算术平均数的变形而存在。在计算平均数时，当我们不知道变量值个数（即总体总频数），而只知道各组变量值与各组标志总量（即各组变量总值）时，就要使用调和平均数。

1.简单调和平均数$H=\frac{k}{\sum\frac{1}{x_i}}$：当各组的标志总量相等时，所计算的调和平均数称为简单调和平均数。

2.加权调和平均数$H=\frac{\sum m_i}{\sum\frac{m_i}{x_i}}$：当各组的标志总量不相等时，所计算的调和平均数要以各组的标志总量为权数，其结果即为加权调和平均数。

③几何平均数：计算平均比率或平均速度常用的一种方法，例如用于计算水平法的平均发展速度、流水作业生产的产品平均合格率、复利法的平均利率等。

1.简单几何平均数$G=\sqrt[n]{\prod x_i}$：变量的 *n* 个变量值连乘积的 *n* 次方根。

2.加权几何平均数$G=\sqrt[\sum f_i]{\prod x_i^{f_i}}$：当计算几何平均数的各种变量值出现的次数不等，即数据经过了统计分组时，则应采用加权几何平均数。

• 在实际应用中这三者的比较往往没有意义，因为对于任何一个计算对象一般都只适合采用一种方法来计算平均数，也就是说不同的平均数计算方法适合于不同的计算条件，必须加以正确的选择。

（2）位置平均数

①中位数与分位数

1.中位数：变量的所有变量值按定序尺度排序后，处于中间位置的变量值。由于它居于数列的中间位置，所以在某些情况下可以用来代表变量值的一般水平。中位数既可用以测定定量变量的集中趋势，也可用以测定定序变量的集中趋势，但不适用于定类变量。

中位数将按顺序排列的变量值分为了两部分，使得至少一半数值不比它大，至少一半数值不比它小。

中位数具有以下一些优点：一是中位数作为一种位置平均数，概念较为清晰，只要排列数据顺序，就可比较容易地加以确定；二是中位数不受变量数列中特殊值的影响，遇有特大值或特小值时，用中位数来表示现象的一般水平更具有代表性；三是组距数列出现开口组时，对中位数无影响；四是当某些变量不能表现为数值但可以定序时，不能计算数值平均数而可以确定中位数。

当然中位数也有局限性，一是中位数不能如算术平均数那样可以进行代数运算；二是除了变量数列的中间部分数值外，其他数值的变化都不对中位数产生影响，因此中位数的灵敏度较低。

2.分位数：将变量的数值按大小顺序排列并等分为若干部分后，处于等分点位置的数值。

②众数：变量数列中出现次数最多、频率最高的变量值。在某些场合，众数可以用来反映现象的一般水平。

众数具有以下一些特点：一是众数也不受变量数列中特殊值的影响，用它来表示某些现象的一般水平会有较好的代表性；二是众数具有较广的应用面，可用于测定任何变量的集中趋势；三是众数只有在总频数充分多且某一组的频数明显高于其他组时才有意义，若各组的频数相差不多，则不能确定众数；四是有时一个变量数列会有两个组的频数明显最多，这就会有两个众数，该数列属于双众数数列。例如，英语专业与非英语专业的大学二年级学生参加同一英语水平测试，就可能出现双众数现象；再如现在一些高校招生，有的专业在第一批录取，有的专业在第二批次录取，那么全校新生的成绩分布也可能是双众数分布。五是众数也不能象算术平均数那样进行代数运算。

14. 试述敏感性问题抽样调查技术研究进展及其应用。【上一届＋gpt】

（1）概念：

在社会经济调查中经常会出现一些敏感性的或高度私人绝密的问题。譬如：收入、吸毒、作弊等敏感性问题，直接调查难以得到真实答案。调研工作者很希望设计一种办法既使被调查者不担心暴露隐私，又使调查者获得正确的资料。

（2）进展：

• 从最基本的来说，敏感性问题的调查方法主要有邮寄问卷调查法、网络技术调查法、实验法、观察法、投影技法等。

• 敏感性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。随后 Simmons 等人引入无关问题，以 Warner 模型为基础，提出了无关问题的随机化回答模型。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。

（3）应用：

• 敏感性问题的抽样调查技术在各个领域都有广泛的应用，尤其是在社会科学、医学研究、市场研究和公共政策领域。以下是一些敏感性问题抽样调查技术的应用情况：

①社会科学研究：在社会科学领域，敏感性问题的研究可能涉及到个体行为、社会态度和价值观等。随机化响应技术和限制性应答模型被广泛用于调查不法行为、性行为、药物滥用等敏感性较高的主题。这些方法有助于研究人员获取更真实的数据，而不受受访者社会期望和审查的影响。

②医学研究：在医学研究中，敏感性问题可能涉及到患者的疾病历史、生活方式、药物使用等。研究者可以使用随机化响应技术或其他非随机应答模型来获取患者更真实的信息，从而改善研究的可信度和有效性。

③市场研究：在市场研究中，了解消费者的偏好、购买行为和态度是至关重要的。有时，这些信息可能是敏感的，例如个人收入、购物习惯等。市场研究人员可以采用一系列非随机应答模型，以减轻受访者的担忧，同时仍然获得有关他们行为和态度的宝贵信息。

④公共政策制定：在制定公共政策时，对于一些社会问题，例如毒品滥用、性犯罪、家庭暴力等，了解真实的社会现状和受影响群体的需求至关重要。抽样调查技术的应用有助于政策制定者更好地了解社会问题的实质，以制定更为切实可行的政策。

⑤疾病控制和预防：在疾病控制和预防领域，了解人们的卫生行为和风险因素是至关重要的。采用非随机应答模型有助于获得更准确的数据，从而更好地指导疾病控制和预防策略。

• 总体而言，敏感性问题的抽样调查技术在各个领域的应用都在不断发展和拓展，以提高数据的准确性、可信度和实用性，同时保护受访者的隐私。

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。【文献】

（1）大数据下的相关和因果：

• 相关关系是指一个变量发生变化时，另一个变量也随之变化，而不论这两个变量有没有必然联系。

• 因果关系是指当一个作为原因的变量变化时，另一个作为结果的变量也一定程度上发生变化，两个变量之间存在着必然联系。

• 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。

• 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。

• 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。

• 强调相关性，而弱化因果性的原因可能在于，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。

（2）因果关系仍重要：

• 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。

• 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。

• 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。

• 如果只知道相关性而不清楚因果性，那么大数据分析的深度只达一半，一旦出现问题或疑问就无从下手。如果清楚了因果关系，则能更好地利用相关关系，更好地掌握预测未来的主动权，帮助我们更科学地进行决策。

（3）因果关系的独特价值：

• 大数据时代的丰富数据为从概率论的立场研究因果关系提供了新视角，其对因果关系的重构，使得因果关系重新焕发生命力。大数据作为总体或全部样本的数据，有助于从根本上克服由于抽样偏颇所引起的样本选择性偏误。若采用单一数据，变量遗漏问题往往非常严重，如果将不同来源的大数据匹配起来，可以克服或缓解变量遗漏问题；尽管在复杂、开放、动态的庞大系统中，因果关系的内生性问题仍难解决，但大数据对因果关系的检验比有限样本的抽样数据更为稳健和可靠，内生性问题也有所改进。另外，大数据通常表现为面板数据和分层数据，这对于确定因果效应也极为有利。

• 因果关系也是科学和哲学的主题，从因推导出果，找到两类对象之间的规律和相互关系一直是推动科学与哲学前进的动力之一。大数据中一个耳熟能详的说法是：大数据长于相关关系，而非因果关系，但这可能是一个伪命题。任何科学都想追求因果解释，缺乏因果关系的解释就没有规律；反过来，希望发现规律就必然要追求因果关系。舍恩伯格强调了相关性对大数据的重要性，但并不否定因果性的存在，更没有说要用相关性完全取代因果性，相关关系能为因果关系探求创新条件，因果关系往往能够比相关关系提供更加精确的干预建议。

（4）如何从相关关系进一步推断出因果关系

• 这是大数据的真正问题所在，因果推断也是当前热门的人工智能（AI）的技术基础。

• 。在大数据和人工智能时代，通常使用“数据驱动法”来设置模型和参数，用数量关系来刻画因果关系，在因与果之间架起数据连接。基于大数据的因果推断虽然只是对原因和结果关系的检验，却是一种基于相关性的因果关系量化把握，作为变量相互作用过程确定性关系的描述，因果性在更深层次关系到大数据的哲学意蕴。

• 随着因果关系哲学基础的建立，基于观测性数据的因果关系发现算法

或因果推断框架，逐渐成为数据科学中可能创造商业价值和进行科学发现的重要研究领域之一。

• 基于反事实理论框架和随机化实验思想的因果推断在政策评估中得到广泛应用。因果推断则为政策评估提供了基于观测数据开展实证研究的新方法，可以帮助人们进一步揭示变量间的因果关系，更好地识别政策效应。从已有文献看，巧妙开展因果推断的方法主要包括多元回归、倾向值匹配、工具变量法、双重差分、断点回归等，另外还有基于时序数据格兰杰因果关系检验及基于结构方程模型的因果关系发现方法。

• 大数据时代需要深层次的因果分析，当前开展因果推断的两种代表性方法是以唐纳德 • 鲁宾为代表的结构因果模型和以朱迪亚 • 珀尔为代表的因果图方法。结构因果模型尝试利用结构方程模型，对潜在结果开展建模，从因果作用机制引发的数据分布特性等视角发现事物间的因果关系。因果图方法是在有向无环图（DAG）上清晰地引入因果概念，提出 do 算子即“干预”，进而发展出因果推断的一整套理论和方法。随着因果关系假设和因果模型的不断发展，部分学者尝试利用两类方法的特性设计混合型因果关系发现方法，一定程度上实现了高维扩展性和较强因果发现能力的优势结合，成为实现高维数据因果关系发现的有效方法。

• 总体来看，因果关系的“形式化理论”不仅解决了困扰统计学家多年的一些悖论，更重要的是利用“干预”让人类和机器摆脱了被动观察，从而转向主动地去探索因果关系，以做出更好的决策；“反事实框架”则扩展了想象的空间，从而摆脱现实世界的束缚，更为有效地探求因果关系。这两点突破带来因果革命，分别构成了因果关系之梯的第二层级和第三层级内容，沿着珀尔所构造的因果关系之梯，机器有望拥有更强的人工智能。

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？【上一届】

（1）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）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：
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。【上一届】

（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. 谈谈你对均值回归效应的理解，并举例加以说明。【新，文献】

（1）理解：
均值回归效应是指进行重复观测时，前测中获得的极高值或极低值会在后测时倾向于向平均值靠拢，即随着时间的推移极高值趋于下降、极低值趋于上升这一自然倾向。
（2）举例：
• 英国人弗朗西斯• 高尔顿发现：个头非常高的人，其后代会稍矮一些，不像父亲那么高；个头非常矮的父亲，其后代身高会稍高一些。
• 谚语：日中则昃，月盈则亏；否极泰来，物极必反；祸兮福所倚，福兮祸所伏；有无相生，难易相成，长短相较，高下相顷，音声相和，前后相随。
• 《体育画报》封面魔咒：每当一个运动员或一支球队上了《体育画报》的封面（通常是因为他们取得了优异成绩）之后，这个人或这支球队就往往表现低迷。
• 当飞行员完成一项完美的飞行特技动作后，如果他给予赞许，这位飞行员在下一次完成同样的动作时表现会差一些；对于表现不好的飞行员，如果他
向其怒吼，这位飞行员在下一次飞行中通常会表现得更好。
（3）说明：
• 事物的表现经常围绕着平均值上下变动，极端情况在下一次就往往不那么极端。偏离平均值的异常出色或糟糕表现发生后，紧接着会出现普通表现或者不太极端的事件。如果这一次表现非常惹眼，下一次的表现就会稍逊一筹，而如果这一次表现不尽如人意，下一次就可能做得更好。极端值向平均数回归的现象，主要原因在于随机误差对极端值的影响比对普通值的影响要大得多。当然，回归均值并不是一个自然法则，仅仅是统计上的一种倾向，且发生可能需要较长的时间。
• 在日常判断中，很多现象本是回归效应的产物，人们往往未能认识到这一点，反而对这些现象发展出其他详尽的解释，采用一些不必要、通常也很复杂的因果论来解释所观察到的现象，产生“回归谬误”。
• 鉴于回归效应的存在，实际当中需要评估、检验变量之间的因果关系时，可通过科学的实验设计来避免自变量混淆。对于出现的问题，采取随机化、随机区组、正交设计等方法控制额外变量或使额外变量保持恒定，以消除其对因变量的影响。除了待测变量之外，对照组与实验组还需保持可比性，以确保没有额外的因素干扰结果。如果没有对照组，往往难以知晓实验处理究竟是否对结果产生影响。

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？【上一届＋gpt】

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

指数距离权值：$W_{ij} = \sqrt{\exp(-d_{ij}/q)}$

三次方距离权值：$W_{ij} = [1-(\theta/d_{ij})^3]^3$

②分位数回归模型（QR）：基于古典均值回归模型，使用残差绝对值的加权平均作为最小化的目标函数，不易受极端值的影响，还能提供关于条件分布 *Y|X* 的全面信息。

$$Y = X'\beta(\tau) + \varepsilon, Q_\tau(Y|X) = X'\beta(\tau), Q_\tau(\varepsilon|X) = 0$$

（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。

• 具体来说，可以从以下几个方面来比较：

①目的：GWR 主要用于处理空间异质性，强调地理位置对回归系数的影响；QR 主要用于研究因变量在不同分位数上的条件差异，强调不同条件下的回归关系。

②应用领域：GWR 在地理信息系统（GIS）、地理学等领域中应用较多；QR 则在经济学、社会学等领域中应用较为广泛。

③权重：GWR 使用地理权重，强调地理位置的影响；QR 则不涉及地理权重，其主要关注的是因变量在分位数上的不同条件。

④鲁棒性：QR 相对于异常值较为鲁棒，而 GWR 主要关注的是在地理空间中局部化的模式。

（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，GWR 和 QR 都在一定程度上都考虑了数据的异质性。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。同时也有学者将二者结合，建立了地理加权分位数回归模型（GWQR）。

20. 阐述统计调查与市场实验的区别与联系。【新，gpt】

（1）定义：

• 统计调查：根据统计研究的目的和任务， 运用科学有效的调查方法， 有计划有组织地向调查对象搜集原始资料和次级资料的活动过程。

• 市场实验：在既定条件下，通过实验对比，对变量之间的因果关系及其发展变化过程进行观察分析的一种调查方法，通常又称为因果关系调查。

（2）区别：

①设计和控制：

• 统计调查：通常是一种观察性研究方法，研究者并不主动介入或控制研究对象。调查数据的收集依赖于被调查者的回答，而研究者不能操纵变量。

•市场实验：通常涉及主动操纵或控制一个或多个变量，以观察其对目标变量的影响。实验是一种因果推断的强有力工具，因为它允许研究者通过控制变量来识别原因和效应之间的关系。

②随机化：

• 统计调查：通常无法进行随机分组，因此对于观察到的关系难以进行因果推断。

• 市场实验：通过随机分组的方式，实验可以有效控制潜在的混淆因素，增加因果推断的可靠性。

③外部效度和内部效度：

• 统计调查：通常具有较强的外部效度，能够反映真实世界的情况。但受到潜在混淆因素的影响，内部效度较弱。

• 市场实验：由于对变量的操控和随机化控制，实验通常具有较强的内部效度，能够更好地确定原因和效应之间的关系。但有时候可能牺牲一些外部效度。

（3）联系：

• 数据收集：统计调查和市场实验都涉及数据的收集，可以通过问卷、观察、实验等方式获取信息。

• 分析方法：在数据收集后，统计调查和市场实验都需要经过数据分析的过程，以提取有关变量之间关系的信息。

• 市场研究目的：两者都用于市场研究，以了解消费者行为、市场趋势、产品接受度等信息。

• 推断：虽然统计调查主要是一种描述性研究方法，但通过适当的统计分析仍然可以得出一些推断性的结论。市场实验则更侧重于因果推断。