目　录

第一章 导论

1.什么是统计学？

答：统计学是关于数据的科学，它所提供的是一套有关数据收集、处理、分析、解释并从数据中得出结论的方法， 统计研究的是来自各领域的数据。

具体来说，（1）数据收集：取得统计数据；
（2）数据处理：将数据用图表等形式展示出来；
（3）数据分析：选择适当的统计方法研究数据；
（4）数据解释：结果的说明；
（5）得到结论：从数据分析中得出客观结论。

2. 解释描述统计和推断统计。

答：（1）描述统计研究的是数据收集、处理、汇总、图表描述、概括与分析等统计方法。

（2）推断统计研究如何利用样本数据来推断总体特征的统计方法。推断统计是在搜集、整理观测样本数据的基础上，对有关总体作出推断，其特点是根据随机样本数据以及问题的条件和假定，对未知事物作出的以概率形式表述的推断。推断统计的重要作用是通过从总体中抽取样本构造适当的统计量，由样本性质去推断总体性质。

3.统计数据可分为哪几种类型？不同类型的数据各有什么特点？

答：（1）根据所采用的计量尺度的不同（计量尺度是指对计量对象量化时采用的具体标准）：

①分类数据（定类尺度）：只能归于某一类别的非数字型数据，它是对事物进行分类的结果，数据表现为类别，是用文字来表述的。分类数据中，各类别之间是平等的并列关系，无法区分优劣或大小，各类别之间的顺序是可以改变的。

②顺序数据（定类尺度）：也叫等级数据，只能归于某一有序类别的非数字型数据。顺序数据虽然也是类别，但这些类别是有序的。

③数值型数据（定距尺度或定比尺度）：按数字测度测量的观察值，其结果表现为具体的数值，现实中所处理的大多数都是数值型数据。

也可分为：

①定性数据：也叫品质数据，包括分类数据和顺序数据。说明的是事物的品质特征，通常用文字来表述，其结果均表现为类别。

②定量数据：也叫数量数据，主要指数值型数据，说明现象的数量特征，用数值来表现。

（2）按照统计数据收集方法的不同：

①观测数据：通过调查或观测而收集到的数据，这类数据是在没有对事物人为控制的条件下得到的，有关社会经济现象的统计数据几乎都是观测数据。

②实验数据：在实验中控制实验对象而收集到的数据，自然科学领域的大多数数据都是实验数据。

（3）按照时间状况的不同：

①截面数据：通常是在不同的空间上获得的，在相同或近似相同的时间点上收集的数据。用于描述现象在某一时刻的变化情况。

②时间序列数据：按照时间顺序在不同时间上收集到的数据。用于描述现象随时间变化的情况。

③面板数据：对若干单位在不同时间进行重复跟踪、调查所形成的数据。

4.举例说明总体、样本、参数、统计量、变异、变量、变量值几个概念。

答：（1）总体是包含所研究的全部个体（数据）的集合，它通常由所研究的一些个体组成。组成总体的每个元素称为个体。比如，要检验一批灯泡的使用寿命，这批灯泡构成的集合就是总体，每个灯泡就是一个个体。

（2）样本是从总体中抽取的一部分元素的集合，构成样本的元素的数目称为样本量。比如，从一批灯泡中随机抽取 100 个，这 100 个灯泡就构成了一个样本，100 就是这个样本的样本量。

（3）参数是用来描述总体特征的概括性数字度量，它是研究者想要了解的总体的某种特征值。研究者所关心的参数通常由总体平均数、总体标准差、总体比例等。

（4）统计量是用来描述样本特征的概括性数字度量。它是根据样本数据计算出来的一个量，由于抽样是随机的，因此统计量是样本的函数。研究者所关心的统计量主要有样本平均数、样本标准差、样本比例等。

（5）变异是标志在同一总体不同总体单位之间的差别。例如：人的性别标志表现为男、女；年龄标志表现为 20 岁、30 岁等。

（6）变量是变异标志，是说明现象某种特征的概念，其特点是从一次观察到下一次观察结果会呈现出差别或变化。

（7）变量值是变量的具体取值，具体包括：

①分类变量，如“性别”就是分类变量，其变量值为“男”或“女”；

②顺序变量，如“产品等级”就是顺序变量，其变量值可以为“一等品”、“二等品”、“三等品”、“次品”等。

③数值型变量，如“年龄”是连续数值型变量，变量值为非负数；“企业数”是离散数值型变量，变量值为 1,2，...

5.变量可分为哪几类。

答：（1）分类变量：说明事物类别的一个名称，其取值是分类数据。

（2）顺序变量：说明事物有序类别的一个名称，其取值是顺序数据。

（3）数值型变量：说明事物数字特征的一个名称，其取值是数值型数据。数值型变量根据其取值的不同，又可以分为：

①离散型变量：只能取可数值的变量，它只能取有限个值，而且其取值都以整位数断开，可以一一列举，如“企业数”、“产品数量”等就是离散型变量；

②连续型变量：可以在一个或多个区间中取任何值的变量，它的取值是连续不断的，不能一一列举，如“年龄”“温度”“零件尺寸的误差”等都是连续型变量。

从其他角度变量还可以分为随机变量和非随机变量、经验变量和理论变量。

6. 请举出应用统计的几个领域。

答：统计是适用于所有学科领域的通用数据分析方法，是一种通用的数据分析语言。主要有几个研究方向：包括国民经济核算与研究，市场调查分析，社会公共事业统计领域，金融市场领域等等。统计学研究所涉及领域相当广泛，但归结起来三方面。核算、计量、市场调查。

统计学的应用

（1）企业发展战略

发展策略是一个企业长远的发展方向。控制发展战略一方面需要及时的了解和把握整个宏观经济的状况及发展变化趋势，另一方面还要对企业进行合理的市场定位，把握企业自身的优势和劣势。所有这些都需要统计提供可靠的数据，利用统计方法进行科学的数据分析和预测。

（2）产品质量管理

质量是企业的生命，是企业持续发展的基础。质量管理中离不开统计的应用。在一些知名的跨国公司，准则已经成为一种重要的管理理念。质量控制应经成为统计学在生产领域中的一项重要应用。各种统计质量控制图被广泛应用于监测生产过程。

（3）市场研究

企业要在激烈的市场竞争中取得优势，首先必须了解市场，要了解市场就需要进行广泛的市场统计调查，取得所需信息，并对这些信息进行统计分析，以便作为生产和营销的依据。

（4）财务分析

上市公司的财务数据是股民投资的重要参考依据。一些投资咨询公司主要是根据上市公司提供的财务和统计数据进行分析，为股民提供参考。企业自身的投资也离不开对财务数据的分析，其中要用到大量的统计方法。

（5）经济预测

企业要对未来市场状况进行预测。比如：对产品的市场潜力进行预测，及时调整生产计划。这就需要利用统计方法进行收集、整理和分析数据。

（6）人力资源管理

利用统计方法对企业员工的年龄、性别、受教育程度、工资等进行分析，并作为企业制度工资计划、奖惩程度的依据。

第二章 数据的搜集

1.什么是二手资料？其优点和局限性分别是什么？使用二手资料需要注意些什么？

答：（1）与研究内容有关的原始信息已经存在，是由别人调查和实验得来的，只是对其进行加工整理使之成为进行统计分析可以使用的资料，称为“二手资料”。

（2）优点：收集容易，采集成本低，作用广泛。

（3）局限性：无法保证数据的准确性和及时性。

（4）注意：

①在使用之前要进行评估，即资料的原始搜集人、搜集资料的目的、搜集资料的途径、搜集资料的时间；

②数据的定义、含义、计算口径和计算方法，避免错用、误用、滥用。

③在引用二手资料时，要注明数据来源，尊重他人的劳动成果。

2.比较概率抽样和非概率抽样的定义和特点，举例说明什么情况下适合采用概率抽样，什么情况下适合采用非概率抽样？二者的区别是什么？

答：（1）概率抽样

①定义：指抽样时按一定概率以随机原则抽取样本，也称为随机抽样。

②特点：按一定的概率以随机原则抽取样本；抽取样本时使每个单位都有一定的机会被抽中，每个单位被抽中的概率是已知的或是可以计算出来的；当用样本对总体目标量进行估计时，要考虑到每个样本单位被抽中的概率；概率抽样的技术含量和成本都比较高。

③举例说明：如果调查的目的在于掌握和研究总体的数量特征，得到总体参数的置信区间，就使用概率抽样。

（2）非概率抽样

①定义：指抽取样本时不是依据随机原则，而是根据研究目的对数据的要求，采用某种方式从总体中抽出部分单位对其实施调查。

②特点：非概率抽样操作简单；实效快；成本低；而且对于抽样中的专业技术要求不是很高。

③举例说明：它适合探索性的研究，调查结果用于发现问题。为更深入的数量分析提供准备。非概率抽样也适合市场调查中的概念测试。

（3）概率抽样与非概率抽样的区别：

①是否遵循随机原则：

概率抽样也称随机抽样，是指遵循随机原则进行的抽样，总体中每个单位都有一定的机会被选入样本。非概率抽样是相对于概率抽样而言的，指抽取样本时不依据随机原则，而是根据研究目的对数据的要求，采用某种方式从总体中抽出部分单位对其实施调查。

②是否可以根据样本结果推断总体：

非概率抽样不是依据随机原则抽选样本，样本统计量的分布不确切，因而无法使用样本的结果对总体相应的参数进行推断；概率抽样是依据随机原则抽选样本，这时样本统计量的理论分布是存在的，因此可以根据调查的结果对总体的有关参数进行估计，计算估计误差，得到总体参数的置信区间，并且可以在进行抽样设计时，对估计的精度提出要求，计算为满足特定精度要求所需的样本量。

③其他特点的比较：

非概率抽样操作简单、时效快、成本低；对于抽样中的统计学专业技术要求不高；适合探索性研究或预备性研究，调查结果用于发现问题。

概率抽样调查成本高；抽样的技术含量更高，需要较高的统计专业知识；适合于各种推断总体的研究。

实际应用中二者往往结合使用。

3. 简述概率抽样和非概率抽样的分类及其各自的优缺点。

答：（1）概率抽样

①简单随机抽样

a.抽样框：对可以选择作为样本的总体单位列出名册或排序编号，以确定总体的抽样范围和结构。又称为抽样结构或抽样框架。

b.定义：从包括 N 个总体单位的抽样框中随机地、一个一个地抽取 n 个单位作为样本，每个单位入样的概率是相等的。简单随机抽样是一种最基本的抽样方法，是其他抽样方法的基础。

c.分类：重复抽样和不重复抽样

d.优点：简单、直观。在抽样框完整时，可以直接从中抽取样本，由于抽选的概率相同，用样本统计量对目标量进行估计及计算估计量误差都比较方便。

e.局限性：它要求将包含所有总体单位的名单作为抽样框，当 N 很大时，构造抽样框比较困难；抽出的单位很分散，给实施调查增加了困难；没有利用其他辅助信息以提高估计的效率。所以，在规模较大的调查中，很少直接采用简单随机抽样，一般是把这种方法和其他抽样方法结合起来使用。

②分层抽样

a.定义：将抽样单位按某种特征或某种规则划分为不同的层，然后从不同的层中独立、随机地抽取样本，将各层的样本结合起来，对总体的目标量进行估计。

b.优点：保证了样本中包含有各种特征的抽样单位，样本的结构比较相近，从而可以有效地提高估计的精度；在一定条件下，为组织实施调查提供了方便（当层是按行业或行政区进行划分时）；既可以对总体参数进行估计，也可以对层的目标量进行估计。

③整群抽样

a.定义：将总体中若干单位合并为组，这样的组称为群。抽样时直接抽取群，然后对中选群中的所有单位全部实施调查，这样的抽样方法称为整群抽样。

b.优点：抽取样本时只需要群的抽样框，而不必要求包括所有单位的抽样框，大大简化了编制抽样框的工作量；调查的地点相对集中，节省了调查费用，方便了调查的实施。

c.缺点：估计的精度较差。因为同一群内的单位或多或少有些相似，在样本量相同的条件下，整群抽样的抽样误差通常比较大。一般来说，要得到与简单随机抽样相同的精度，采用整群抽样需要增加基本调查单位。

④系统抽样

a.定义：将总体中的所有单位（抽样单位）按一定顺序排列，在规定的范围内随机地抽取一个单位作为初始单位，然后按事先规定好的规则确定其他样本单位。

b.优点：操作简便，如果有辅助信息，对总体内的单位进行有组织的排列，可以有效地提高估计的精度。

c.缺点：对估计量方差的估计比较困难。

⑤多阶段抽样

a.定义：采样类似整群抽样的方法，首先抽取群，但并不是调查群内所有单位，而是再进一步抽样，从选中的群中抽取出若干个单位进行调查。

b.优点：保证了样本相对集中，从而节约了调查费用；不需要包含所有低阶段抽样单位的抽样框；同时由于实行了再抽样，使调查单位在更广的范围内展开。在较大规模的抽样调查中，多阶段抽样是经常采用的方法。

c.缺点：对估计量的方差的估计比较困难。

（2）非概率抽样

①方便抽样

a.定义：调查员依据方便的原则，自行确定作为样本的单位。

b.优点：容易实施；调查成本低。

c.缺点：样本单位的确定带有随意性，方便样本无法代表有明确定义的总体。若研究的目的是对总体的参数进行推断，使用方便样本是不合适的。

②判断抽样

a.定义：根据经验、判断和对研究对象的了解，有目的地选择一些单位作为样本，实施时根据不同的目的有：

重点抽样：从调查对象的全部单位中选择少数重点单位，对其实施调查。重点单位的数量虽不多，但在总体中地位重要。

典型抽样：从总体中选择若干个典型的单位进行深入的调研，目的是通过典型单位来描述或揭示所研究问题的本质和规律，因此，选择的典型单位应具有研究问题的本质或特征。

代表抽样：通过分析选择具有代表性的单位作为样本。

b.优点：成本低，操作容易。

c.缺点：由于样本的确定没有依据随机的原则，因而调查结果不能用于对总体有关参数进行估计。

③自愿样本

a.定义：被调查者自愿参加，成为样本中的一分子，而调查人员提供有关信息。

b.优点：可以给研究人员提供许多有价值的信息，可反映某类群体的一般看法。

c.缺点：不能依据样本的信息对总体的状况进行估计。

④滚雪球抽样

a.定义：用于对稀少群体的调查。首先选择一组调查单位，对其实施调查之后，再请他们提供另外一些属于研究总体的调查对象，调查人员根据所提供的线索，继续进行调查。这个过程持续下去，就会形成滚雪球效应。

b.优点：容易找到属于特定群体的被调查者，调查的成本也比较低。适合对特定群体进行资料的搜集和研究。

⑤配额抽样

a.定义：类似于概率抽样中的分层抽样，将总体中的所有单位按一定的标志（变量）分为若干类，然后再每个类中采用方便抽样或判断抽样的方式选取样本单位。

b.优点：操作简单，且可以保证总体中不同类别的单位都能包括在所抽的样本中，使得样本的结构和总体的结构类似。

c.缺点：抽取具体样本单位时并不依据随机原则，不能对总体进行估计。

4.调查中搜集数据的方法主要有自填式、面方式、电话式，除此之外，还有那些搜集数据的方法？

答：实验式、观察式等。

5. 搜集数据的方法有哪几种，这些方法各自有什么利弊？

答：（1）自填式

①优点：调查组织者管理容易，成本低，可以进行较大规模调查，对被调查者可以选择方便时间答卷，减少回答敏感问题的压力。

②缺点：返回率低，调查时间长，在数据搜集过程中遇到问题不能及时调整。

（2）面谈式

①优点：回答率高，数据质量高，在数据搜集过程中遇到问题可以及时调整可以充分发挥调查员的作用。

②缺点：成本比较高，对调查过程的质量控制有一定难度。对于敏感问题，被访者会有压力。

（3）电话式

①优点：速度快，对调查员比较安全，对访问过程的控制比较容易。

②缺点：实施地区有限，调查时间不宜过长，问卷要简单，被访者不愿回答时，不宜劝服。

6.简述抽样误差和非抽样误差的概念，并列举抽样误差的影响因素和非抽样误差的种类。

答：（1）抽样误差

①定义：是一种随机性误差，只存在于概率抽样中。是由抽样的随机性引起的样本结果与总体真值之间的误差。

②影响因素：

a.样本单位数目：通过增大样本量可以减少抽样误差，当样本量大到与总体单位相同时，也就是抽样调查变成普查时，抽样误差减小到零。

b.总体标志变异程度：在其他条件不变的情况下，总体标志变异程度越大，抽样误差越大；总体变异程度越小，抽样误差越小。

c.抽样方法：一般，不重复抽样 $\sqrt{\frac{N-n}{N-1}}\bullet\sqrt{\frac{\sigma^2}{n}}$ 的抽样误差要小于重复抽样 $\frac{\sigma}{\sqrt{\mathrm{n}}}$ 的抽样误差。当 n 相对 N 非常小时，两种抽样方法的抽样误差相差很小，可忽略不计。

d.抽样组织方式：采用不同的抽样组织方式，也会有不同的抽样误差。一般，分层抽样的抽样误差较小，而整群抽样的抽样误差较大。

（2）非抽样误差

①定义：非抽样误差是相对抽样误差而言的，是指除抽样误差之外的，由其他原因引起的样本观察结果与总体真值之间的差异。无论是概率抽样、非概率抽样，或者是在全面调查中，都有可能产生非抽样误差。

②种类：

a.抽样框误差：

在概率抽样中需要根据抽样框抽取样本。抽样框是有关总体全部单位的名录，在地域抽样中，抽样框也可以是地图。一个好的抽样框应该是，抽样框中的单位和研究总体中的单位有一一对应的关系。由于抽样框的不完善造成的这些统计推论的错误，把这种误差称为抽样框误差。

b.回答误差：被调查者在接受调查时给出的回答与真实情况不符。

理解误差：不同的被调查者对调查问题的理解程度不同，每个人都按自己的理解回答，大家的标准不一致，由此造成理解误差。

记忆误差：调查的问题是关于一段时期内的现象或事实，需要被调查者回忆。需要回忆的时间间隔越久，回忆的数据可能就越不准确。所以，缩短调查所涉及的时间间隔可以减少记忆误差。但是，有些时间是按一定周期发生的。

有意识误差：当调查的问题比较敏感，被调查者不愿意回答，迫于各种原因又必须回答时，就可能会提供一个不真实的数字。原因有二，一是调查问题涉及个人隐私，被调查者不愿意告知；二是受到利益驱动，进行数字造假。有意识误差比记忆误差的危害大，因为记忆误差具有随机性，调查结果还是具有趋中的倾向；有意识误差则不同，它往往偏向某一个方向，是一种系统性偏差。

c.无回答误差：

无回答的产生与调查内容有关时，属于系统性的误差，如被调查者拒绝接受调查，调查人员得到的是一份空白的答卷。无回答的产生与调查的内容无关时属于随机性的误差，如调查进行时被访者不在家的情况；电话调查中，拨通后无人接听；邮寄问卷调查中，地址写错，被调查者搬家，或被调查者虽然收到问卷，却把问卷遗忘或丢失。

d.调查员误差：

指由于调查员的原因而产生的调查误差。例如，调查员粗心，在记录调查结果时出现错误。调查员误差还产生于调查中的诱导，而调查员本人可能并没有意识到。例如，在调查过程中调查员有意无意地流露出对调查选项的看法或倾向，调查员表情的变化、语气变化、语速变化都可能对被调查者产生某种影响。

e.测量误差：

如果调查与测量工具有关，则很有可能产生测量误差。例如，对小学生的视力状况进行抽样调查，而视力的测定与现场的灯光、测试距离都有密切关系。调查在不同地点进行，如果各测试点的灯光、测试距离有所差异，就会给调查结果带来测量误差。调查有时也采用观察、记数的方式进行。

7.你认为应当如何控制调查中的回答误差？

答：（1）对于理解误差，要注意表述中的措辞，学习一定的心理学知识。

（2）对于记忆误差，尽量缩短所涉及问题的时间范围。

（3）对于有意识误差，调查人员要想法打消被调查者的思想顾虑，调查人员要遵守职业道德，为被调查者保密，尽量避免敏感问题。

8.怎样减少无回答？请通过一个例子，说明你所考虑到的减少无回答的具体措施。

答：（1）对于随机误差，可以通过增加样本容量来控制。

（2）对于系统误差，做好预防，在调查前做好各方面的准备工作，尽量把无回答率降到最低程度。无回答出现后，分析无回答产生的原因，采取补救措施。比如要收回一百份，就要做好一百二十份或一百三十份问卷的准备，当被调查者不愿意回答时，可以通过一定的方法劝服被访者，还可以通过馈赠小礼品等的方式提高回收率。

第三章 数据的图表搜集

1.数据的预处理包括哪些内容？

答：（1）数据审核：检查数据中是否有错误。

①一手数据即通过调查取得的原始数据，主要从两个方面审核：

a.完整性审核：检查应调查的单位或个体是否有遗漏，所有调查项目是否填写齐全。

b.准确性审核：检查数据是否有错误，是否存在异常值。对于异常值要仔细鉴别：如果异常值属于记录时的错误，在分析之前应予以纠正；如果是一个正确的值，则应予以保留。

②二手数据，主要从两个方面审核：

a.适用性审核：应弄清数据的来源，数据的口径以及有关的背景资料，以便确定这些数据是否符合分析研究的需要，不能盲目生搬硬套。

b.时效性审核：对于时效性较强的问题，如果所取得的数据过于滞后，就可能失去研究的意义。

（2）数据筛选：根据需要找出符合特定条件的某类数据。

（3）数据排序：按一定顺序将数据排列，以便研究者通过浏览数据发现一些明显的特征或趋势，找到解决问题的线索，还有助于对数据进行检查纠错，以及为重新归类或分组提供方便。

2.分类数据、顺序数据和数值型数据的整理方法各有哪些？

答：（1）分类数据：由于定性数据不能用数字表示，一般都用频数来表示。

①频数：落在某一特定类别或组中的数据个数。

②频数分布：用表格的形式将频数表示出来。

③列联表：由两个或两个以上变量交叉分布的频数分布表。二维列联表也叫交叉表。

④比例：也称构成比，是一个样本（或总体）中各个部分的数据与全部数据之比，通常用于反映样本（或总体）的构成或结构。

⑤百分比：将比例乘以 100 得到的数值，用%表示。

⑥比率：样本（或总体）中不同类别数据之间的比值，由于比率不是部分与整体之间的对比关系，因而比值可能大于一。

（2）顺序数据：除了分类数据可以使用的方法以外，还可以计算累积频数和累积频率。

①累积频数：将各有序类别或组的频数逐级累加起来得到的频数，频数的累积方法有两种：一是向上累积；二是向下累积。通过累积频数，可以很容易看出某一类别（或数值）以下或某一类别（或数值）以上的频数之和。

②累积频率或累积百分比：将各有序类别或组的百分比逐级累加起来，也有向上累积和向下累积两种方法。

（3）数值型数据：处理数值型数据需要对其分组，数据分组的目的是观察数据的分布特征。

①分组的方法：

a.单变量值分组：适用于数据较少的离散型变量。

b.组距分组：适用数据多的连续型变量。各组数据如果在本组内呈均匀分布或在组中值两侧呈对称分布，那可以使用组中值来代表一组数据。

②分组步骤：

a.确定组数：一般来说组数不少于 5 组也不多于 15 组，确定组数的公式为：

$$k = 1 + \lg n \div \lg 2$$

b.确定组距：组距是一组上限和下限的差。一般将数据的最大值减去最小值除以组数等于组距。为了计算方便，组距要采用 5 或 10 的倍数，而且最低组的下限要低于全部值的最小值，最高组的上限要高于全部值的最大值。

c.根据分组整理出频数分布表，注意遵循“不重不漏”和“上限不在内”的原则。

③分组注意事项：

a.不重复不遗漏原则。不重复是指一个数据只能归为其中的某一个组；不遗漏原则是组别能够穷尽，指每个数据都必须属于同一个组。

解决措施：离散型变量采用的方法是采取两组组限间断；连续型变量采用的办法是“上组限不在内”的原则或对一个组的上限值采用小数点的形式，小数点的位数根据所要求的精度来确定。

b.避免空白组和个别极端值被遗漏。如果数据最大值和最小值离其他值相差悬殊，可以将最低组和最高组设立为“...以下”和“...以上”的开口组。

3.分类数据、顺序数据和数值型数据以及多变量数据的展示方法各有哪些？

答：（1）分类数据

①条形图

a.定义：条形图是用宽度相同的条形的宽度或长短来表示数据多少的图形，可反映不同类别的频数多少或分布状况。条形图可以横置或纵置，纵置时也称为柱形图。

b.分类：简单条形图和复式条形图，其中复式条形图可表示两个类别变量的情况，可分为并列条形图和堆叠条形图。条形图的变种为帕累托图、脊形图和马赛克图。

c.优点：能直观地、清楚地反映出各类别的频数多少。

②帕累托图

a.定义：按各类别数据出现的频数多少排序后绘制的条形图。图左侧纵轴给出了计数值，即频数，右侧的纵轴给出了累积百分比。

b.优点：通过对条形的排序，容易看出哪类数据出现得多，哪类数据出现得少。

③饼图

a.定义：用圆形及圆内扇形的角度来表示数值大小的图形，主要用于表示一个样本（或总体）各组成部分（不同类别）的数据占全部数据的比例。饼图的变种是扇形图。

b.分类：简单饼图的嵌套称为复式饼图，用于展示两个或多个分类变量的构成比较。

c.优点：对于研究结构性问题十分有用。

④环形图

a.定义：将多个简单饼图堆叠到同一张图上，挖去中间的部分，可显示多个总体各部分所占的比例。

b.优点：有利于研究结构性问题。

（2）顺序数据

除了上述分类数据的方法以外，顺序数据还可以通过绘制累积频数或频率分布图来展示数据。

（3）数值型数据

①直方图（分组数据）

a.定义：用于展示分组数据分布的一种图形，用矩形的宽度和高度（即面积）来表示频数分布，用横轴表示数据分组、纵轴表示频数或频率。直方图可以观察数据分布的大体形状，如分布是否对称。

b.分类：若纵轴为频数则称为频数直方图；若纵轴为频率则称为频率直方图。

②折线图（分组数据）

排列在工作表的列或行中的数据可绘制到折线图中。折线图可以显示随时间而变化的连续数据，因而非常适用于显示在相等时间间隔下数据的趋势。在折线图中，类别数据沿水平轴均匀分布，所有值数据沿垂直轴均匀分布。

③曲线图（分组数据）

表示矿体的有用组分品位变化曲线图沿某一方向的含量变化和各组分间消长关系的图。

④茎叶图（未分组数据）

a.定义：由茎和叶两部分构成，其图形是由数字组成的，反应原始数据分布的图形。茎叶图可看出数据的分布形状以及数据的离散状况，如分布是否对称，数据是否集中，是否有离群点。

b.绘制方法：关键是设计好树茎。制作茎叶图时，首先把一个数字分成两部分，通常以该组数据的高位数值作为树茎，而且叶上只保留该数值的最后一位数字。

⑤箱线图（未分组数据）

a.定义：根据一组数据的最大、最小值、中位数和上、下四分位数，这五个特征值绘制而成。可以反映原始数据分布的特征，也可以进行多组数据分布特征的比较。

b.绘制方法：找出一组数据的最大、最小值、中位数和上、下四分位数；然后，连接两个四分位数画出箱子；再将最大值和最小值与箱子相连接，中位数在箱子中间。

c.箱线图展示的数据分布特征：对称分布、左偏分布、右偏分布和 U 形分布。

⑥线图（时间序列数据）

a.定义：若数值型数据是在不同时间取得的，即时间序列数据，则可以绘制线图。线图主要用于反映现象随时间变化的特征。

b.绘制方法：时间一般绘在横轴上，观测值绘在纵轴，长宽比为 10:7。一般，纵轴数据下端应从“0”开始，以便做比较，如果数据与“0”之间的间距过大，可采取折断的符号将纵轴折断。

（4）多变量数据

①散点图

每组数据（x_i,y_i）在坐标系中用一个点表示，n 组数据在坐标系中形成的 n 个点称为散点，由坐标及散点形成的二维数据图称为散点图，可展示两个变量之间的关系。

②气泡图

一个变量在横轴，一个变量在纵轴，第三个变量用气泡大小表示，可用于展示三个变量之间的关系。

③雷达图

先画一个圆，然后将圆 P 等分，得到 P 个点，令这 P 个点分别对应 P 个变量，再将这 P 个点与圆心连线，得到 P 个辐射状的半径，这 P 个半径分别作为这 P 个变量的坐标轴，每个变量值的大小由半径上的点到圆心的距离表示，再将同一样本的值在 P 个坐标上的点连线，这样 n 个样本形成的 n 个多边形就是一张雷达图。可显示多个变量，在显示或对比各变量的数值总和时十分有用。假定各变量的取值具有相同的正负号，则总的绝对值与图形所围成的区域成正比，此外，利用雷达图可以研究多个样本之间的相似程度。

4.绘制线图应注意问题？

答：时间在横轴，观测值绘在纵轴。一般是长宽比例 10：7 的长方形，纵轴下端一般从 0 开始，数据与 0 距离过大的话用折断符号折断。

5.直方图和条形图有何联系与区别？

答：（1）区别

①条形图使用图形的长度表示各类别频数的多少，其宽度固定；直方图用面积表

示各组频数，矩形的高度表示每一组的频数或频率，宽度表示组距，高度与宽度都有意义；

②直方图各矩形连续排列，条形图分开排列；

③条形图主要展示分类数据，直方图主要展示数值型数据。

（2）联系

二者都是用来展示数据的分布情况；在平面直角坐标系中，二者的横轴都表示分组，纵轴都可表示频数或频率的大小。

6.茎叶图比直方图的优势，他们各自的应用场合？（或者说茎叶图与直方图的区别）

答：（1）图形特点

直方图是用于展示分组数据分布的一种图形，它是用矩形的宽度和高度（即面积）来表示频数分布的；茎叶图是反映原始数据分布的图形，它由茎、叶两部分组成，其图形是由数字组成的。

（2）数据分布状况

茎叶图类似于横置的直方图，与直方图相比，茎叶图既能给出数据的分布状况，又能给出每一个原始数据，即保留了原始数据的信息；而直方图虽然能很好地展示数据的分布，但不能保留原始的数据。

（3）应用方面

直方图通常适用于大批量数据，茎叶图适用于小批量数据。

7.饼图和环形图的区别是什么？

答：饼图只能显示一个样本或总体各部分所占比例，环形图可以同时绘制多个样本或总体各部分所占比例，其图形中间有个“空洞”，每个样本或总体的数据系类为一个环，样本中的每一部分数据用环中的一段表示。因此环形图可显示多个样本各部分所占的相应比例，有利于对构成做比较研究。

8.鉴别图标优劣的准则？

答：（1）一张好图应当精心设计，有助于洞察问题的实质；

（2）一张好图应当使复杂的观点得到简明、确切、高效的阐述；

（3）一张好图应当在最短的时间内以最少的笔墨给读者提供最大量的信息；

（4）一张好图应当是多维的；

（5）一张好图应当表述数据的真实性。

9.制作统计表应注意的问题？

答：（1）合理安排统计表结构；

（2）表头一般包括表号，总标题和表中数据的单位等内容；

（3）表中的上下两条横线一般用粗线，中间的其他用细线，两端开口，数字右对齐，不要有空白格；

（4）在使用统计表时，必要时可在下方加注释，注明数据来源。

第四章 数据的概括性度量

1.一组数据的分布特征可以从哪几个方面进行测度？

答：（1）分布的集中趋势：反映各数据向其中心值靠拢的程度；

（2）分布的离散程度：反映各数据远离其中心值的趋势；

（2）分布的形状：反映数据分布的偏态和峰态。

2.简述众数、中位数及四分位数和平均数的概念、特点和应用场合。

（如何对分类数据、顺序数据和数值型数据的集中趋势进行度量）

答：（1）众数（主要用于分类数据）

①定义：是一组数据中出现次数最多的变量值，用 M_O 表示。

②特点：

a.众数是一个位置代表值，它不受数据中极端值的影响。

b.从分布的角度看，众数是具有明显集中趋势点的数值，一组数据分布的最高峰点所对应的数值即为众数。

c.一组数据不一定有众数，也不一定只有一个众数。如果数据的分布没有明显的集中趋势或最高峰点，众数可能不存在；如果有两个或多个最高峰点，则可以有两个或多个众数。

③应用场合：主要用于测度分类数据的集中趋势，也可作为顺序数据以及数值型数据集中趋势的测度值，只有数据量较大的情况下，众数才有意义。

（2）中位数（主要用于顺序数据）

①定义：一组数据排序后处于中间位置上的变量值，用 M_e 表示。

②特点：中位数是一个位置代表值，不受极端值的影响。

③确定步骤：

a.将数据从小到大排序： $x_{(1)},x_{(2)},\ldots x_{(n)}$；

b.公式： $M_e=\dfrac{n+1}{2}=\begin{cases}x_{\left(\frac{n+1}{2}\right)}, & n\text{为奇数}\\ \dfrac{1}{2}\left\{x_{\frac{n}{2}}+x_{\left(\frac{n+1}{2}\right)}\right\}, & n\text{为偶数}\end{cases}$

④应用场合：主要用于测度顺序数据的集中趋势，也适用于测度数值型数据的集中趋势，但不适合分类数据。

（3）四分位数（主要用于顺序数据）

①定义：是一组数据排序后处于 25%和 75%位置上的值。上四分位数是处在 75%位置上的值，下四分位数是处在 25%位置上的值。

②特点：排序数据中

a.至少 25%的数据将小于或等于 Q_L,至少 75%的数据将大于或等于 Q_L；

b.至少 75%的数据将小于或等于 Q_U,至少 25%的数据将大于或等于 Q_U；

c.Q_L 和 Q_U 之间包含了 50%的数据。

③确定步骤：

a.将数据从小到大排序： $x_{(1)},x_{(2)},\ldots x_{(n)}$；

b. Q_L位置 $=\dfrac{n}{4}$

Q_U位置 $=\dfrac{3n}{4}$

c.若结果为整数，那么四分位数就是该位置对应的值；如果是 0.5 位置，取该位置两侧值的平均数；如果是 0.25 或 0.75 位置，则四分位数等于该位置的下侧值加上按比例分摊位置两侧数值的差值。

④应用场合：主要用于测度顺序数据的集中趋势，也适用于测度数值型数据的集中趋势，但不适合分类数据。

（4）平均数（主要用于数值型数据）

①定义：也叫均值，是一组数据相加后除以数据的个数得到的结果。

②分类：

a.简单平均数：根据未分组数据计算的平均数，$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n}$。

b.加权平均数：根据分组数据计算的平均数，$\bar{x}=\frac{\sum_{i=1}^{k}M_i f_i}{n}$。

c.几何平均数：n 个变量值乘积的 n 次方根，用 G 表示，$G=\sqrt[n]{\prod_{i=1}^{n}x_i}$。

d.调和平均数：$A=\frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$。

③应用场合：平均数是集中趋势的最主要测度值，适用于数值型数据，而不适用于分类数据和顺序数据。其中几何平均数是一种适用于特殊数据的平均数，主要用于计算平均比率。当所掌握的变量值本身是比率形式时，采用几何平均法更合理。实际应用中，主要用于计算现象的平均增长率。

3.简述异众比率、四分位差、方差或标准差的概念、特点和应用场合。

（如何对分类数据、顺序数据、数值型数据的离散程度进行度量）

答：（1）异众比率（主要用于分类数据）

①定义：指非众数组的频数占总频数的比例，用 V_r 表示。

②公式：$V_r=\frac{\sum f_i-f_m}{\sum f_i}=1-\frac{f_m}{\sum f_i}$

③特点：V_r 越大，说明非众数组的频数占总频数的比重越大，众数的代表性越差；V_r 越小，说明非众数组的频数占总频数的比重越小，众数的代表性越好。

④应用场合：V_r 主要用于衡量众数对一组数据的代表程度。V_r 适合测度分类数据的离散程度，对于顺序数据以及数值型数据也可以计算 V_r。

（2）四分位差（主要用于顺序数据）

①定义：也叫内距或四分位距，是上四分位数与下四分位数之差，用 Q_d 表示。

②公式：$Q_d=Q_U-Q_L$

③特点：Q_d 反应了中间 50%的数据的离散程度，数值越小，说明中间的数据越集中；数值越大，说明中间的数据越分散。

④应用场合：Q_d 的大小一定程度上说明了中位数对一组数据的代表程度。Q_d 主要用于测度顺序数据的离散程度，对于数值型数据也可计算四分位差，但不适合分类数据。

（3）极差（用于数值型数据）

①定义：也叫全距，是一组数据的最大值与最小值之差，用 R 表示。

②公式： $R=\max(x_i)-\min(x_i)$

③优点：计算简单，易于理解。

④缺点：易受极端值的影响，极差只是利用了一组数据两端的信息，不能反映出中间数据的分散状况，因而不能准确描述出数据的分散程度。

（4）平均差（用于数值型数据）

①定义：也叫平均绝对离差，是各变量值与其平均数离差绝对值的平均数，用 M_d 表示。

②公式： 根据未分组数据，$M_d=\dfrac{\sum_{i=1}^{n}\left|x_i-\bar{x}\right|}{n}$

根据已分组数据：$M_d=\dfrac{\sum_{i=1}^{k}\left|M_i-\bar{x}\right|f_i}{n}$

③优点：M_d 以平均数为中心，反应了每个数据与平均数的平均差异程度，它能全面准确地反映一组数据的离散状况。M_d 越大，说明数据的离散程度越大；反之，则说明数据的离散程度越小。

④缺点：为了避免离差之和为零而无法计算平均差这一问题，以离差的绝对值来表示总离差，这给计算带来了不便。

（5）方差和标准差（用于数值型数据）

①定义：方差是各变量值与其平均数离差平方的平方根，二者都反应了数据离散程度的绝对值。但方差无量纲，而标准差有量纲且与变量值的计量单位相同，实际意义比方差更清楚。

②公式：

根据未分组数据：$s^2=\dfrac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{n-1},s=\sqrt{\dfrac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{n-1}}$

根据已分组数据：$s^2=\dfrac{\sum_{i=1}^{n}\left(M_i-\bar{x}\right)^2 f_i}{n-1},s=\sqrt{\dfrac{\sum_{i=1}^{n}\left(M_i-\bar{x}\right)^2 f_i}{n-1}}$

4.标准分数的概念及用途？

答：（1）定义：也叫标准化值或 z 分数，变量值与其平均数的离差除以标准差后的值，记为 z。

（2）公式：$z_i=\dfrac{x_i-\bar{x}}{s}$

（3）特点：标准分数的平均数为 0，标准差为 1。标准分数是对原始数据的线性变换，并没有改变一个数据在该组数据中的位置，也没有改变该组数据分布的形状。

（4）用途：标准分数给出了一组数据中各数据的相对位置，适用于多个具有不同量纲即单位不同的变量进行处理，可以用来判断是否有离群数据。

5.如何判断一组数据是否有离群数据？（即相对位置的度量）

答：（1）标准分数（答案同 4 题）

（2）经验法则：适合对称分布的数据。当一组数据对称分布时，经验法则表明：

①约有 68%的数据在平均数±1 个标准差的范围之内；

②约有 95%的数据在平均数±2 个标准差的范围之内；

③约有 99%的数据在平均数±3 个标准差的范围之内。

在平均数±3 个标准差的范围内几乎包含了全部数据，而在±3 个标准差之外的数据，统计上称为离群点。

（3）切比雪夫不等式：对任何分布形态的数据都适用，提供的是下界。至少有$\left(1-\frac{1}{k^2}\right)$的数据落在±k 个标准差之内，k 是大于 1 的任意值，也可以不是整数。

①当 k=2 时，至少有 75%的数据在平均数±2 个标准差的范围之内；

②当 k=3 时，至少有 89%的数据在平均数±3 个标准差的范围之内；

③当 k=4 时，至少有 94%的数据在平均数±4 个标准差的范围之内。

6.为什么要计算离散系数？并简述其概念。

答：（1）方差和标注差的缺点

①二者的数值受原变量自身水平高低即变量的平均数大小的影响。变量值绝对水平高，则离散程度的测度值也就大；变量值绝对水平低，测度值也就小。

②二者与原变量值的计量单位相同，采用不同计量单位计量的变量值，其离散程度的测度值也就不同。

为了消除变量值水平高低和计量单位不同对离散程度测度值的影响，需要计算离散系数。

（2）离散系数

①定义：也叫变异系数，是一组数据的标准差与其相应的平均数之比，记为 v_s。

②公式：$v_s=\frac{s}{\bar{x}}$

③特点：离散系数大，离散程度也就大；离散系数小，离散程度也就小。

④应用场合：离散系数是测度数据离散程度的统计量，主要用于比较不同样本（单位不同，相对量也不同）数据的离散程度。

7.测度数据分布形状的统计量有哪些？

答：（1）偏态

①定义：对数据分布对称性的测度。

②公式：

根据未分组数据：$SK=\dfrac{n\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^3}{(n-1)(n-2)s^3}$

根据已分组数据：$SK=\dfrac{\sum_{i=1}^{k}\left(M_i-\bar{x}\right)^3 f_i}{ns^3}$

③判断：若数据分布是对称的，则 SK=0；

$SK>1$或$SK<1$，则称为高度偏态分布；

若 SK 明显不等于 0，则$0.5<SK<1$或-1，则称为中度偏态分布；

SK 越接近 0，偏斜程度越小；SK 数值越大，偏斜程度越大。当 SK 为正值时，表示正离差值较大，可以判断为正偏或右偏；当 SK 为负值时，表示负离差值较大，可以判断为负偏或左偏。

（2）峰态

①定义：对数据分布平峰或尖峰程度的测度。

②公式：

根据未分组数据：$K=\dfrac{n(n+1)\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^4-3\left[\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right]^2(n-1)}{(n-1)(n-2)(n-3)s^4}$

根据已分组数据：$K=\dfrac{\sum_{i=1}^{k}\left(M_i-\bar{x}\right)^4 f_i}{ns^4}-3$

③判断：若数据服从标准正态分布，则 K=0；

若 K 明显不等于 0，则 $K>0$时为尖峰分布，数据的分布更集中；
$K<0$时为扁平分布，数据的分布更分散。

第六章 统计量与抽样分布

1.简述总体，个体，容量，抽样，样本和样本量的概念。

答：（1）总体：包含所调查或研究的全部个体（数据）的集合。

（2）个体：组成总体的每个元素。

（3）容量：总体中所含个体的个数。

（4）抽样：从总体中按一定的抽样技术抽取若干个个体的过程。

（5）样本：从总体中抽取的一部分元素的集合。

（6）样本量：构成样本的元素的数目。

2.什么是统计量？为什么要引进统计量？统计量中为什么不含任何未知参数？

答：（1）设$X_1,X_2,\cdots X_n$是从总体 X 中抽取的容量为 n 的一个样本，如果由

此样本构造一个函数$T(X_1, X_2,\cdots, X_n)$，不依赖于任何未知参数，则称函数$T(X_1, X_2,\cdots, X_n)$是一个统计量。又称为样本统计量，当观测值为$X_1, X_2,\cdots X_n$时，代入$T(X_1, X_2,\cdots, X_n)$，计算出其数值，就获得一个具体统计量。

（2）由样本构建具体的统计量，实际上是对样本所含的总体信息按某种要求进行加工处理，把分散在样本中的信息集中到统计量的取值上，不同的统计推断问题要求构造不同的统计量。

（3）构造统计量的主要目的就是对总体的未知参数进行推断，如果统计量中含有总体的未知参数就无法再对参数进行统计推断。

3.什么是次序统计量？什么是充分统计量？什么是自由度？

答：（1）次序统计量：设是从总体 X 中抽取的一个样本，$X_{(i)}$称为第 i 个次序统计量，它是样本$X_1, X_2,\cdots X_n$满足$X_1, X_2,\cdots X_n$如下条件的函数：每当样本得到一组观测值 $x_1,x_2,...,x_n$时，其由小到大的排序$x_{(1)} \le x_{(2)} \le \cdots \le x_{(i)} \le \cdots \le x_{(n)}$中第 i 个值$x_{(i)}$就作为次序统计量$X_{(i)}$的观测值，而$X_{(1)}, X_{(2)},\cdots, X_{(n)}$称为次序统计量。

（2）充分统计量：样本统计量$T(X_1, X_2,\cdots, X_n)$的构造过程是对样本$X_1, X_2,\cdots X_n$，根据要求进行加工处理的过程，这种加工处理是把原来杂乱无章的样本观测值用少数几个经过加工以后的统计量的值来代替。用统计量进行推断，而不是用样本观测值进行推断。所以如果统计量加工过程中一点信息都不损失，能够基本反映总体的信息，这样的统计量最好，称为充分统计量。

（3）统计学上的自由度指当以样本的统计量来估计总体的参数时，样本中独立和能自由变化的变量的个数。

4.简述χ^2分布、t 分布、F 分布及正态分布之间的关系。三大分布概念在自己整理的纸上。

答：（1）χ^2分布与正态分布的关系：$n \to +\infty$时，χ^2分布的极限分布是正态分布；

（2）t 分布与正态分布的关系：t 分布的密度函数曲线与标准正态分布的密度函数曲线非常相似，但 t 分布的密度函数在两侧的尾部都要比标准正太的两侧尾部粗一些，方差也比标准正态分布的方差大。随着自由度 n 的增加，t 分布的

密度函数越来越接近标准正态分布的密度函数。

（3）F 分布和正态分布的关系：若$X\sim t(n)$，则$X^2\sim F(1,\ n)$。随着自由度 n 的增加，X 也越来越接近于标准正态分布，若把 X 看成近似服从标准正态分布的一个随机变量，则$X^2\sim F(1,\ n)$。

5.什么是总体分布、样本分布和抽样分布？

答：（1）总体分布

①总体中各元素的观察值形成的分布；

②总体分布通常是未知的；

③可以假定它服从某种分布。

（2）样本分布

①一个样本中各观察值的分布；

②样本分布也称为经验分布；

③当样本容量 n 逐渐增大时，样本分布逐渐接近总体分布。

（3）抽样分布

①抽样分布是一种理论分布在重复选取容量为 n 的样本时，由样本统计量的所有可能取值形成的相对频数分布，是样本统计量的概率分布。

②样本统计量是随机变量，如样本均值、样本比例、样本方差等。

③其结果来自容量相同的所有可能样本。

④提供了样本统计量长远而稳定的信息，是进行推断的理论基础，也是抽样推断科学性的重要依据。

6.简述中心极限定理的意义。

答：设从均值为μ，方差为σ^2的任意一个总体中抽取样本量为 n 的样本，当 n 充分大时，样本均值$\bar{X}$的抽样分布近似服从均值为μ，方差为$\frac{\sigma^2}{n}$的正态分布。中心极限定理解决了在总体为非正态的情况下，样本平均数的抽样分布问题，为总体参数的推断提供了理论基础。

7.简述单个总体的抽样分布。

答：（1）样本均值的分布

①当总体服从正态分布时，来自该总体的所有容量为 n 的样本的均值$\bar{x}$也服从正态分布，$\bar{x}$的均值为μ，重复抽样时方差为$\frac{\sigma^2}{n}$，不重复抽样时方差为

$\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$。

②当样本容量足够大时$(n\geq 30)$，无论总体是正态分布还是非正态分布，样本均值的抽样分布均逐渐趋于正态分布。

（2）样本比例分布

①定义：从一个总体中重复选取样本量为 n 的样本，由样本比例的所有可能取值形成的分布是样本比例的概率分布。

②总体比例：总体中具有某种属性的单位与全部单位总数之比，$\pi=\frac{N_0}{N}$ 或 $1-\pi=\frac{N_1}{N}$

③样本比例：样本中具有某种属性的单位与全部单位总数之比，$p=\frac{n_0}{n}$ 或 $1-p=\frac{n_1}{n}$

④当样本量很大时（通常要求$np\geq 10, n(1-p)\geq 10$），样本比例分布可用正态分布近似，即 $p\sim N\left(\pi,\frac{\pi(1-\pi)}{n}\right)$（重复抽样），不重复抽样的方差为 $\sigma_p^2=\frac{\pi(1-\pi)}{n}\left(\frac{N-n}{N-1}\right)$。

（3）样本方差的分布

①定义：在重复选取容量为 n 的样本时，由样本方差的所有可能取值形成的相对频数分布。

②对于来自正态总体的简单随机样本，则比值 $\frac{(n-1)s^2}{\sigma^2}$ 的抽样分布服从自由度为（n-1）的 χ^2 分布，即 $\frac{(n-1)s^2}{\sigma^2}\sim\chi^2(n-1)$。

8.简述两个总体的抽样分布。

答：（1）两个样本平均值之差的分布：

①两个总体都为正态分布，$X_1\sim N(\mu_1,\sigma_1^2), X_2\sim(\mu_2,\sigma_2^2)$。

②两个样本均值之差 $\overline{x_1}-\overline{x_2}$ 的抽样分布服从正态分布，其分布的数学期望为两个总体均值之差 $E(\overline{x_1}-\overline{x_2})=\mu_1-\mu_2$。

③两个样本均为独立的大样本$(n_1 \geq 30, n_2 \geq 30)$方差为各自的方差之和为$\sigma^2_{\overline{x_1}-\overline{x_2}} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$，即$\overline{X_1} - \overline{X_2} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$。

④两个样本均为独立的小样本$(n_1 < 30, n_2 < 30)$

a.两个总体的方差未知但相等，即$\sigma_1^2 = \sigma_2^2 = \sigma^2$，此时对$\sigma^2$的合并估计量$s_p^2$的计算公式为$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$，这时，两个样本均值之差经标准化后服从自由度为$(n_1 + n_2 - 2)$的 t 分布，即$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$。

b.两个总体方差未知且不相等时，即$\sigma_1^2 \neq \sigma_2^2$，两个样本均值之差经标准化后近似服从自由度为 v 的 t 分布，自由度 v 的计算公式为$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$，即$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(v)$。

（2）两个样本比例之差的分布：

①两个总体都服从正态分布；

②分别从两个总体中抽取容量为 n_1 和 n_2 的独立样本，当两个样本都为大样本时，两个样本比例之差的抽样分布可用正态分布来近似。

③分布的数学期望为$E(P_1 - P_2) = \pi_1 - \pi_2$。

④方差为各自的方差之和$\sigma^2_{p_1 - p_2} = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$。

（3）两个样本方差比的分布：

①两个总体都为正态分布，即$X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$。

②从两个总体中分别抽取容量为 n_1 和 n_2 的独立样本。

③两个样本的方差比的抽样分布，服从分子自由度为(n_1-1)，分母自由度为(n_2-1)的 F 分布，$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}\sim F(n_1-1,n_2-1)$。

第七章 参数估计

1.什么是参数估计？

答：参数估计时推断统计的重要内容之一，它是在抽样及抽样分布的基础上，根据样本统计量来推断所关心的总体参数。参数估计就是用样本统计量去估计总体的参数。如果将总体参数用θ表示，用于估计总体参数的估计量用$\hat{\theta}$表示，参数估计也就是如何用$\hat{\theta}$来估计θ。

2.解释估计量与估计值。

答：（1）估计量：用于估计总体参数的统计量（随机变量），用符号$\hat{\theta}$表示。如样本均值、样本比例、样本方差等。例如：样本均值$\bar{x}$就是总体均值μ的一个估计量。

（2）估计值：根据一个具体的样本计算出来的估计量的数值。

3.简述点估计和区间估计的概念及其区别。

答：（1）点估计

①用样本统计量$\hat{\theta}$的某个取值直接作为总体参数θ的估计值。

②点估计无法给出估计值接近总体参数程度的信息。虽然在重复抽样条件下，点估计的均值可望等于总体真值，但由于样本是随机的，抽出一个具体的样本得到的估计值很可能不同于总体真值。一个点估计量的可靠性是由它的抽样标准误差$\left(s=\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2/n-1}\right)$来衡量的，这表明一个具体的点估计值无法给出估计的可靠性的度量。

（2）区间估计

①在点估计的基础上，给出总体参数估计的一个区间范围，该区间通常由样本统计量加减估计误差得到。

②根据样本统计量的抽样分布能够对样本统计量与总体参数的接近程度给出一个概率度量。

（3）点估计与区间估计的区别

①点估计也称定值估计，它是以抽样得到的样本统计量的某个取值$\hat{\theta}$直接作为总体参数θ的估计值；

②区间估计也是参数估计的一种形式。区间估计是在点估计的基础上，给出总体参数估计的一个区间范围，该区间通常由样本统计量加减估计误差得到；

③与点估计不同，进行区间估计时，根据样本统计量的抽样分布可以对样本统计量与总体参数的接近程度给出一个概率度量。点估计得出的是一个具体的值，而区间估计是一个区间。

4.简述评价估计量好坏的标准。

答：（1）无偏性：无偏性是指估计量抽样分布的数学期望等于被估计的总体参数。设总体参数为θ，所选择的估计量为$\hat{\theta}$，如果$E(\hat{\theta})=\theta$，则称$\hat{\theta}$为θ的无偏估计量。其实际意义是无系统误差。

（2）有效性：一个无偏的估计量并不意味着它就非常接近被估计的参数，还必须与总体参数的离散程度比较小。有效性是指对同一总体参数的两个无偏估计量，有更小标准差的估计量更有效。

（3）一致性：一致性是指随着样本量的增大，点估计量的值越来越接近被估总体的参数。换言之，一个大样本给出的估计量要比一个小样本给出的估计量更接近总体的参数。

5.什么是置信区间？怎样理解置信区间？

答：（1）含义：在区间估计中，由样本统计量所构造的总体参数的估计区间称为置信区间。

（2）对置信区间的理解应注意：

①如果用某种方法构造的所有区间中有95%的区间包含总体参数的真值，5%的区间不包含总体参数的真值，那么，用该方法构造的区间称为置信水平为95%的置信区间。同样，其他置信水平的区间也可以用类似的方式进行表述。

②总体参数的真值是固定的、未知的，而用样本构造的区间则是不固定的。若抽取不同的样本，那么可以得到不同的区间，从这个意义上说，置信区间是一个随机区间，它会因样本的不同而不同，而且不是所有的区间都包含总体参数的真值。一个置信区间就像是为捕获未知参数而撒出去的网，不是所有撒网的地点都能捕获到参数。

③在实际问题中，进行估计时往往只抽取一个样本，此时所构造的是与该样本相联系的一定置信水平下的置信区间。由于用该样本所构造的区间是一个特定的区间，而不再是随机区间，所以无法知道这个样本所产生的区间是否包含总

体参数的真值。我们只能希望这个区间是大量包含总体参数真值的区间中的一个，但它也可能是少数几个不包含参数真值的区间中的一个。

6.什么是置信度？置信度 1-α的含义是什么？

答：（1）置信度：将构造置信区间的步骤重复多次，置信区间中包含总体参数真值的次数所占的比例称为置信度，也叫置信水平或置信系数。

（2）置信度 1-α的含义：在随机抽样中，若重复抽样多次，得到样本 $X_1, X_2, \cdots X_n$ 的多个样本值 $x_1, x_2, \cdots x_n$，对应每个样本值都确定了一个置信区间 $(\hat{\theta}_1,\ \hat{\theta}_2)$，每个这样的区间要么包含了 θ 的真值，要么不包含 θ 的真值。根据伯努利大数定理，当抽样次数充分大时，这些区间中包含 θ 的真值的频率接近于置信度（即概率），即在这些区间中包含真值的区间大约有 $100(1-\partial)\%$ 个，不包含真值的区间大约有 $100\partial\%$ 个。

7.解释 95%的置信区间。

答：在多次抽样中有 95%的样本得到的区间包含总体真值，它的真正意义是如果做了 100 次抽样，大概有 95 次找到的区间包含真值，有 5 次找到的区间不包含真值。95%这个概率不是用来描述某个特定的区间包含总体参数真值的可能性，而是针对随机区间而言的。一个特定的区间“总是包含”或“绝对不包含”参数的真值，不存在“以多大的概率包含总体参数”的问题。但是，用概率可以知道在多次抽样得到的区间中大概有多少个区间包含参数的真值。

8. $z_{\frac{\partial}{2}}\frac{\sigma}{\sqrt{n}}$ 的含义是什么？

答：是样本的估计误差，构成了样本置信区间长度的一半。

9.解释独立样本和匹配样本的含义。

答：（1）独立样本：如果两个样本是从两个总体中独立抽取的，即一个样本中的元素与另一个样本中的元素相互独立，则称为独立样本。

（2）匹配样本：即一个样本中的数据与另一个样本中的数据相对应。匹配样本可以消除由于样本指定的不公平造成的两种方法组装时间上的差异。

10.简述一个总体参数的区间估计。

答：（1）总体均值的区间估计

①结果的四舍五入法则

a.当用原始数据构建置信区间时，置信区间的计算结果应保留的小数点位置要比原始数据中使用的小数点多一位。

b.当不知道原始数据，只使用汇总统计量$(n,\bar{x},s)$时，置信区间的计算结果保留的小数点位数应与样本均值使用的小数点位数相同。

②正态总体、σ^2已知或非正态总体、大样本

a.假定条件：总体服从正态分布，且方差σ^2已知；如果不是正态分布当$n\geq 30$时，可由正态分布来近似。

b.使用正态分布统计量：$z=\dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$。

c.总体均值μ在1-α置信水平下的置信区间为：$\bar{x}\pm z_{\partial/2}\dfrac{\sigma}{\sqrt{n}}$或$\bar{x}\pm z_{\partial/2}\dfrac{s}{\sqrt{n}}(\sigma未知)$。

③正态总体、σ^2未知、小样本

a.假定条件：总体服从正态分布，但方差σ^2未知；且为小样本$(n<30)$。

b.使用t分布统计量：$t=\dfrac{\bar{x}-\mu}{s/\sqrt{n}}\sim t(n-1)$。

c.总体均值μ在1-α置信水平下的置信区间为：$\bar{x}\pm t_{\partial/2}\dfrac{s}{\sqrt{n}}$

（2）总体比例的区间估计

①假定条件：总体服从二项分布；且为大样本。

②使用正态分布统计量z：$z=\dfrac{p-\pi}{\sqrt{\dfrac{\pi(1-\pi)}{n}}}\sim N(0,1)$。

③总体比例p在1-α置信水平下的置信区间为：$p\pm z_{\partial/2}\sqrt{\dfrac{p(1-p)}{n}}$。

（3）总体方差的区间估计

①估计一个总体的方差或标准差。

②假设总体服从正态分布。

③总体方差σ^2的点估计量为s^2，且$\dfrac{(n-1)s^2}{\sigma^2}\sim\chi^2(n-1)$。

④总体方差σ^2在1-α置信水平下的置信区间为

$\frac{(n-1)s^2}{\chi^2_{\partial/2}(n-1)} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\partial/2}(n-1)}$。

11.简述两个总体参数的区间估计。

答：（1）两个总体均值之差的估计

①大样本

a.假定条件：两个总体都服从正态分布，σ_1^2，σ_2^2已知；若两个总体不是正态分布或方差未知，当$n_1 \ge 30, n_2 \ge 30$时，可以用正态分布来近似；两个样本是独立的随机样本。

b.使用正态分布统计量 z，$z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$。

c.两个总体均值之差$\mu_1 - \mu_2$在 1-α置信水平下的置信区间为

$(\overline{x_1} - \overline{x_2}) \pm z_{\partial/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\sigma_1^2, \sigma_2^2已知)$

$(\overline{x_1} - \overline{x_2}) \pm z_{\partial/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\sigma_1^2, \sigma_2^2未知)$。

②小样本$(\sigma_1^2 = \sigma_2^2)$

a.假定条件：两个总体都服从正态分布，方差未知但相等：$\sigma_1^2 = \sigma_2^2$；两个独立的小样本。

b.总体方差的合并估计量：$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$。

c.两个样本均值之差的标准化 t 服从的分布为：$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$。

d.两个总体均值之差$\mu_1 - \mu_2$在 1-α置信水平下的置信区间为

$(\overline{x_1} - \overline{x_2}) \pm t_{\partial/2}(n_1 + n_2 - 2)\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$。

③小样本$(\sigma_1^2 \ne \sigma_2^2)$

a.假定条件：两个总体都服从正态分布，方差未知且不相等；两个独立的小样本。

b.使用统计量 $t=\frac{\left(\overline{x_1}-\overline{x_2}\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}\sim t(v)$。

c.自由度 $v=\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1}+\frac{\left(s_2^2/n_2\right)^2}{n_2-1}}$。

d.两个总体均值之差 $\mu_1-\mu_2$ 在 1-α 置信水平下的置信区间为 $\left(\overline{x_1}-\overline{x_2}\right)\pm t_{\partial/2}(v)\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$。

④匹配大样本

a.假定条件：两个匹配的大样本 $(n_1\geq 30,n_2\geq 30)$。

b.两个总体各观察值的配对差服从正态分布。

c.两个总体均值之差 $\mu_d=\mu_1-\mu_2$ 在 1-α 置信水平下的置信区间为 $\overline{d}(\text{对应差值的均值})\pm z_{\alpha/2}\frac{\sigma_d(\text{对应差值的标准差})}{\sqrt{n}}$。

⑤匹配大样本

a.假定条件：两个匹配的小样本 $(n_1<30,n_2<30)$。

b.两个总体各观察值的配对差服从正态分布。

c.两个总体均值之差 $\mu_d=\mu_1-\mu_2$ 在 1-α 置信水平下的置信区间为 $\overline{d}\pm t_{\partial/2}(n-1)\frac{s_d}{\sqrt{n}}$。

（2）两个总体比例之差的区间估计

①假定条件：两个大样本服从二项分布可以用正态分布来近似；两个样本是独立的。

②两个总体比例之差 $\pi_1-\pi_2$ 在 1-α 置信水平下的置信区间为 $\left(p_1-p_2\right)\pm z_{\partial/2}\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$。

（3）两个总体方差比的区间估计

①比较两个总体的方差比，可用 F 分布。

②用两个样本的方差比来判断：

a.如果 s_1^2/s_2^2 接近于 1，说明两个总体方差很接近；

b.如果 s_1^2/s_2^2 远离 1，说明两个总体方差之间存在差异。

③总体方差比在 1-α 置信水平下的置信区间为 $\frac{s_1^2/s_2^2}{F_{\partial/2}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2/s_2^2}{F_{1-\partial/2}}\left[F_{1-\partial/2}(n_1,n_2)=\frac{1}{F_{\partial/2}(n_2,n_1)}\right]$。

12.简述样本量与置信水平、总体方差、估计误差的关系。

答：（1）估计总体均值时样本量的确定

①单个总体

a.估计总体均值时样本容量 n 为 $n=\frac{z_{\partial/2}^2\sigma^2}{E^2}$,其中$E=z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$。如果$\sigma$的值不知道，可以用以前相同或类似的样本的标准差来代替；也可以用试验调查的办法，选择一个初始样本，以该样本的标准差作为σ的估计值。

b.样本容量 n 与总体方差σ^2、边际误差 E、可靠性系数 z 或 t、置信水平（1-α）的关系为：样本容量与总体方差成正比；与可靠性系数成正比；与置信水平成正比；与边际误差成反比。

c.样本容量的圆整法则：当计算出的样本容量不是整数时，将小数点后面的数值一律进位成整数。

②两个总体

a.设 n_1 和 n_2 为来自两个总体的样本，并假定 $n_1=n_2$。

b.根据均值之差的区间估计公式可得两个样本的容量 n 为 $n_1=n_2=n=\frac{z_{\alpha/2}^2\cdot\left(\sigma_1^2+\sigma_2^2\right)}{E^2}$，其中 $E=z_{\alpha/2}\frac{\sigma_1+\sigma_2}{\sqrt{n}}$。

（2）估计总体比例时样本容量的确定

①单个总体

a.根据比例区间估计公式可得样本容量 n 为

$n=\frac{z_{\partial/2}^2\cdot\pi(1-\pi)}{E^2}$,其中$E=z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}$。在实际应用中，如果$\pi$的值不知道，可以用类似的样本比例来代替；也可以用试验调查的办法，选择一个初试样本，以该样本的比例作为π的估计值。

b.E 的取值一般小于 0.1。

c.π 未知时，可取使$\pi(1-\pi)$最大时的 0.5。

②两个总体

a.设 n_1 和 n_2 为来自两个总体的样本，并假定 $n_1=n_2$。

b.根据比例之差的区间估计公式可得两个样本的容量 n 为 $n_1=n_2=n=\frac{z_{\alpha/2}^2\cdot[\pi_1(1-\pi_1)+\pi_2(1-\pi_2)]}{E^2}$，其中 $E=z_{\alpha/2}\frac{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}{\sqrt{n}}$。

第八章 假设检验

1.假设检验和参数估计有什么相同点和不同点？

答：（1）相同点

①都是根据样本信息推断总体参数；

②都以抽样分布为理论依据建立在概率论基础之上的判断，判断结果都有风险；

③对同一问题的参数进行推断，使用同一样本、同一统计量、同一分布，因而二者可以相互转换。

（2）不同点

①参数估计是以样本资料估计总体参数的可能范围，假设检验是以样本资料检验对总体参数的先验假设是否成立；

②区间估计求得的是以样本估计值为中心的双侧置信区间，假设检验既有双侧检验，也有单侧检验；

③区间估计立足于大概率，通常以较大的把握程度（可信度）1-α去估计总体参数的置信区间；假设检验立足于小概率，通常是给定很小的显著性水平α去检验对总体参数的先验假设是否成立。

2.什么是假设检验中的显著性水平？统计显著是什么意思？

答：（1）显著性水平

①显著性水平α是一个概率值。其含义是当原假设为真时却被拒绝的概率或风险，这也是假设检验中犯弃真错误的概率，它是人们根据检验的要求确定的。通常取α=0.05 或α=0.01，这表明，当做出接受原假设的决定时，其正确的概率为 95%或 99%。

②其意义是，当显著性水平取α，即小概率事件发生了，则认为原假设不成立，α取不同的水平，将直接影响到拒绝域的临界值，并进而影响到判断结果。

（2）统计显著

①显著的（significant）一词的意义在这里并不是“重要的”，而是指“非偶然的”。

②在假设检验中，如果样本提供的证据拒绝原假设，则称样本结果在“统计上是显著的”；同样如果不拒绝原假设，则称样本结果在“统计上是不显著的”。

③一项检验在统计上是“显著的”，意思是指：这样的（样本）结果不是偶然得到的，或者说，不是靠机遇得到的。拒绝原假设，则表示这样的样本结果并不是偶然得到的；同样不拒绝原假设，则表示这样的样本结果只是偶然得到的。

④在“显著”和“不显著”之间没有清楚的界限，只是在 P 值越来越小时，我们就有越来越强的证据。在大样本的情况下，总能把与假设值的任何细微差别都查出来，即使这种差别几乎没有任何实际意义。因此，在实际检验中，不要刻意追求“统计上的”显著性，也不要把统计上的显著性与实际意义上的显著性混同起来。

3.什么是假设检验中的两类错误？如何控制两类错误及其数量关系是什么？

答：（1）弃真错误（α错误、第 I 类错误、显著性水平）

原假设 H_0 为真时却被拒绝了，犯这种错误的概率用α表示。

（2）取伪错误（β错误、第 II 类错误）

原假设 H_0 为伪时却没有被拒绝，犯这种错误的概率用β表示。

（3）两类错误的数量关系

α与β是此消彼长的关系。当样本量 n 固定不变时，若要减少α错误，就会增大犯β错误的机会；若要减少β错误，就会增大犯α错误的机会。若要同时减小两类错误，只有通过增大样本量的方法来实现。

（4）实际中控制α错误或β错误

一般来说，哪一类错误所带来的后果越严重，危害越大，在假设检验中就把哪一类错误作为首要的控制目标。但在假设检验中，有一个原则是：首先控制犯α错误。此原则的原因有二：一是，研究者都遵循一个统一的原则，讨论问题较为方便。二也是主要原因是，从实用的观点看，原假设是什么常常是明确的，而备择假设是什么则常常是模糊的。

4.解释假设检验中的 P 值以及如何利用 P 值进行决策？

答：（1）P 值的含义

P 值是当原假设为真时样本观察结果或更极端结果出现的概率，被称为观察到的（或实测）显著性水平。如果 P 值很小，说明这种情况发生的概率很小，如果这种情况出现了，根据小概率原理，我们就有理由拒绝原假设，P 值越小，拒绝原假设的理由就越充分。

（2）利用 P 值进行决策

若 P 值小于α，则拒绝原假设 H_0；若 P 值大于α，则不拒绝原假设 H_0。

5.显著性水平α与 P 值有何区别？

答：显著性水平α是犯第 I 错误的上限控制值，它只能提供检验结论可靠性的一个大致范围，而对于一个特定的假设检验问题，却无法给出观测结果与原假设之间不一致程度的精确度量。也就是说，仅从显著性水平来比较，如果选择的α值相同，所有检验结论的可靠性都一样。而 P 值可以测量出样本观测结果与原假设中假设的值的偏离程度（P 值越小，说明实际观测到的结果与 H_0 之间的不一致程度就越大），是观测到的显著性水平。

6.什么是小概率？什么是原假设和备择假设？提出假设的原则是什么？

答：（1）小概率

①在一次试验中，一个几乎不可能发生的事件发生的概率；

②在一次试验中小概率事件一旦发生，我们就有理由拒绝原假设；

③小概率由研究者事先确定。

（2）原假设（零假设）

原假设通常是研究者想收集证据予以推翻的假设，是对总体所作出的一个陈述，用 H_0 表示。（总是有符号 $=$、$\leq$ 或 $\geq$）

（3）备择假设（研究假设）

备择假设是研究者想收集证据予以支持的假设，是原假设被拒绝时可供选择的假设，用 H_1 表示。（总是有符号 $\neq$、$<$ 或 $>$）

（4）提出假设的原则

①原假设和备择假设是一个完备事件组，而且相互独立。在一项假设检验中，原假设和备择假设必有一个成立，而且只有一个成立。

②实验中应先确定原假设，再确定备择假设。

③等号“=”总是放在原假设上。

④因研究目的的不同，对同一问题可能提出不同的假设，也可能得出不同的结论。

7.假设检验依据的基本原理和基本步骤是什么？

答：（1）基本原理

①假设检验是先对总体的参数（或分布形式）提出某种假设，然后利用样本信息判断假设是否成立的过程，分为参数检验和非参数检验。

②假设检验在逻辑上运用反证法，统计上依据小概率原理。

（2）基本步骤

①对所考察的总体的分布形式或总体的某些未知参数做出某些假设，称之为原假设，并给出与之对立的备择假设；

②根据检验对象构造合适的检验统计量，并通过数理统计分析确定在原假设成立的条件下该检验统计量的抽样分布；

③在给定的显著性水平下，根据抽样分布得出原假设成立时的临界值，由临界值构造拒绝域和接受域；

④由所抽取的样本资料计算样本统计量的取值，并将其与临界值进行比较，从而对所提出的原假设做出接受还是拒绝的统计判断。

8.举例说明在单侧检验和双侧检验中原假设和备择假设的方向应该如何确定？

答：（1）双侧检验

一般原假设中含有“=”，备择假设中含有“≠”。例如，要确定一台机床生产的零件是否符合标准要求，如果零件的平均直径大于或小于 10 厘米，则表明生产过程不正常，必须进行调整。研究者想收集证据予以支持的假设应该是“生产过程不正常”，因为如果研究者事先认为生产过程正常，也就没有必要去进行检验了。所以建立的原假设和备择假设应为：H_0：$\mu=10$；H_1：$\mu\neq10$。

（2）双侧检验

①左侧检验

备择假设含有符号“<”。例如，某品牌的洗涤剂声称，每瓶的“平均净含量不低于 500 克”。从消费者的利益出发，有关研究人员通过抽检一批产品来验证此声称是否属实。一般来说，研究者抽检的意图是倾向于证明这种洗涤剂的平均净含量并不符合说明书中的陈述，因为这会损害消费者的利益（如果研究者对产品说明丝毫没有质疑，也就没有抽检的必要了）。所以建立的原假设和备择假设应为：H_0：$\mu\geq500$；H_1：$\mu<500$。

②右侧检验

备择假设含有符号“>”。例如，一家机构估计，某城市中家庭拥有汽车的比例超过 30%，为验证这一估计是否正确，该机构随机抽取一个样本进行检验。研究者想收集证据予以支持的假设是“该城市中家庭拥有汽车的比例超过 30%”，因此建立的原假设和备择假设应为：H_0：$\pi\leq30\%$；H_1：$\pi>30\%$。

总结就是研究者总是会支持自己的看法，如果是别人检验该研究者的看法，会提出方向相反的备择假设。不过这样的假设的确定带有一定的主观色彩，因为“研究者推翻的假设”和“研究者支持的假设”最终仍取决于研究者本人的意

向。

9.一个总体参数的检验。

答：（1）总体均值的检验（大样本）

①假定条件：正态总体或非正态总体大样本$(n \geq 30)$。

②检验的 z 统计量：$\begin{matrix} z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)(\sigma^2\text{已知}) \\ z = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}} \sim N(0,1)(\sigma^2\text{未知}) \end{matrix}$。

（2）总体均值的检验（小样本）

①假定条件：总体服从正态分布，且为小样本$(n < 30)$。

②检验统计量：$\begin{matrix} z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)(\sigma^2\text{已知}) \\ t = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}} \sim t(n-1)(\sigma^2\text{未知}) \end{matrix}$。

（3）总体比例的检验

①假定条件：总体服从二项分布且为大样本，可用正态分布来近似。

②检验的 z 统计量：$z = \dfrac{p - \pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} \sim N(0,1)$,其中$\pi_0$为假设的总体比例 。

（4）总体方差或标准差的检验

①假设总体近似服从正态分布。

②检验总体方差或标准差时使用 χ^2 分布。

③检验统计量：$\chi^2 = \dfrac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$。

10.两个总体参数的检验。

答：（1）两个总体均值之差的检验（独立大样本）

①假定条件：两个样本是独立的随机样本；正态总体或非正态总体大样本$(n_1 \geq 30\text{和}n_2 \geq 30)$。

②检验统计量：$z=\dfrac{\left(\overline{x_1}-\overline{x_2}\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}\sim N(0,1)\left(\sigma_1^2、\ \sigma_2^2已知\right)$

$z=\dfrac{\left(\overline{x_1}-\overline{x_2}\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}}}\sim N(0,1)\left(\sigma_1^2、\ \sigma_2^2未知\right)$。

（2）两个总体均值之差的检验（独立小样本、σ_1^2、σ_2^2已知）

①假定条件：两个独立的小样本；两个总体都是正态分布；σ_1^2、σ_2^2已知。

②检验统计量：$z=\dfrac{\left(\overline{x_1}-\overline{x_2}\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}\sim N(0,1)$。

（3）两个总体均值之差的检验（独立小样本、σ_1^2、σ_2^2 未知但相等）

①假定条件：两个独立的小样本；两个总体都是正态分布；σ_1^2、σ_2^2 未知但相等。

②检验统计量：$t=\dfrac{\left(\overline{x_1}-\overline{x_2}\right)-\left(\mu_1-\mu_2\right)}{s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim t\left(n_1+n_2-2\right), s_p^2=\dfrac{\left(n_1-1\right)s_1^2+\left(n_2-1\right)s_2^2}{n_1+n_2-2}$。

（4）两个总体均值之差的检验（独立小样本、σ_1^2、σ_2^2 未知且不相等）

①假定条件：两个独立的小样本；两个总体都是正态分布；σ_1^2、σ_2^2 未知且不相等。

②检验统计量：

$$t=\frac{\left(\overline{x_1}-\overline{x_2}\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}}}=\frac{\left(\overline{x_1}-\overline{x_2}\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\dfrac{s_1^2+s_2^2}{n}}}\sim t\left(n_1+n_2-2\right)=t\left(2\left(n-1\right)\right),\ \ n_1=n_2$$

$$t=\frac{\left(\overline{x_1}-\overline{x_2}\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}}}\sim t(v),\ \ v=\frac{\left(\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(s_1^2/n_1\right)^2}{n_1-1}+\dfrac{\left(s_2^2/n_2\right)^2}{n_2-1}}, n_1\neq n_2$$

（5）两个总体均值之差的检验（匹配样本）

①假定条件：两个总体配对差值构成的总体服从正态分布；配对差是由差值总体中随机抽取的；数据配对或匹配。

②检验统计量：$t=\frac{\bar{d}-d_0}{s_d/\sqrt{n_d}}\sim t(n-1),\bar{d}=\frac{\sum_{i=1}^{n}d_i}{n_d},s_d=\sqrt{\frac{\sum_{i=1}^{n}(d_i-\bar{d})^2}{n_d-1}}$。

（6）两个总体比例之差的检验

①假定条件：两个总体都服从二项分布且均为大样本；可以用正态分布来近似。

②检验统计量：

H_0：$\pi_1-\pi_2=0,z=\frac{p_1-p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}},\left(p=\frac{x_1+x_2}{n_1+n_2}=\frac{p_1n_1+p_2n_2}{n_1+n_2}\right)$

H_0：$\pi_1-\pi_2=d_0,z=\frac{(p_1-p_2)-d_0}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}}$

（7）两个总体方差之比的检验

①假定条件：两个总体都服从正态分布且方差相等；两个独立的随机样本。

②检验统计量：$F=\frac{s_1^2}{s_2^2}\sim F(n_1-1,n_2-1)$。

第九章 分类数据分析

1. 什么是分类数据？如何对分类数据进行分析？

答：（1）分类数据

分类数据是对事物进行分类的结果，其特征是，调查结果虽然用数值表示，但不同数值描述了调查对象的不同特征。分类数据的结果是频数，χ^2 检验是对分类数据的频数进行分析的统计方法。

（2）χ^2 统计量

①应用：χ^2 可以用于测定两个分类变量之间的相关程度；利用 χ^2 统计量可以对分类数据进行拟合优度检验和独立性检验。

②公式：$\chi^2=\sum\frac{(f_0-f_e)^2}{f_e}$，其中 f_e 表示期望值频数；f_0 表示观察值频数。

③特征

a. $\chi^2\geq 0$，因为它是对平方结果的汇总；

b. χ^2 统计量的分布与自由度有关，随着自由度的增加，χ^2 分布的偏斜程

度趋于缓解，逐渐显露出对称性，随着自由度的继续增大，χ^2 分布将趋近于对称的正态分布；

c.χ^2 统计量描述了观察值与期望值的接近程度。两者越接近，即 f_0-f_e 的绝对值越小，计算出的 χ^2 值越小；反之，f_0-f_e 的绝对值越大，计算出的 χ^2 值也越大。χ^2 检验正是通过对 χ^2 的计算结果与 χ^2 分布中的临界值进行比较，做出是否拒绝原假设的统计决策。

2.什么是拟合优度检验？简述对分类数据使用拟合优度检验的步骤，并说明 χ^2 统计量的计算步骤。

答：（1）含义：它是依据总体分布状况，计算出分类变量中各类别的期望频数，与分布的观察频数进行对比，判断期望频数与观察频数是否有显著差异，从而达到对分类变量进行分析的目的。

（2）步骤

①提出假设：H_0：观察频数与期望频数一致

H_1：观察频数与期望频数不一致

②计算 χ^2 统计量：a.计算 (f_0-f_e)；

b.计算 $(f_0-f_e)^2$；

c.计算 $(f_0-f_e)^2/f_e$；

d.计算 $\chi^2=\sum\frac{(f_0-f_e)^2}{f_e}$；

e. 自由度 $df=R-1$，R 为分类变量类型的个数。

③统计决策：a.若 $\chi^2\geq\chi_\alpha^2$，则拒绝 H_0；

b.若 $\chi^2<\chi_\alpha^2$，则不能拒绝 H_0。

3.如何对分类数据的独立性进行检验？简述列联表的构造与列联表的分布。

答：（1）列联表的构造

①列联表是将两个以上的变量进行交叉分类的频数分布表；

②表中横向变量的划分类别视为 R，纵向变量的划分类别视为 C，这样的

表又可以成为$R\times C$列联表。

（2）独立性检验

①对两个分类变量的分析，称为独立性检验，分析过程可以通过列联表的方式呈现，故把这种分析称为列联分析。独立性检验就是分析列联表中行变量与列变量是否相互独立。

②步骤

a.提出假设：H_0：变量 A 与变量 B 是独立的（不存在依赖关系）

H_1：变量 A 与变量 B 不是独立的（存在依赖关系）

b.根据列联表计算每个单元中的期望比例和其相应的频数期望值：

任何一个单元中频数的期望值为$f_e=\frac{RT}{n}\times\frac{CT}{n}\times n=\frac{RT\times CT}{n}$，RT 为给定单元所在行的合计；CT 为给定单元所在列的合计。

c.根据列联表计算χ^2统计量的值，$\chi^2=\sum\frac{(f_0-f_e)^2}{f_e}$，自由度为$(R-1)\times(C-1)$。

d.统计决策：若$\chi^2\geq\chi_\alpha^2$，则拒绝H_0；若$\chi^2<\chi_\alpha^2$，则不能拒绝H_0。

4.简述列联表中的相关度量，并介绍这些系数各自的特点。

答：对两个变量之间相关程度的测定，主要用相关系数表示。

（1）φ相关系数

①适用情况：主要用于描述2×2列联表数据相关程度最常用的一种相关系数，适合 R 或 C 小于等于 2 的情况；

②计算公式：$\varphi=\sqrt{\frac{\chi^2}{n}}=\frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$

③差值$ad-bc$可以反映变量之间相关程度的高低。差值越大，说明两个变量的相关程度越高。当两个变量相互独立，不存在相关关系时，频数间应有下面的关系：$\frac{a}{a+c}=\frac{b}{b+d}$。

④φ系数的特点：φ在$-1\sim1$之间取值，当$\varphi=0$时，两个变量相互独立；当$|\varphi|=1$时，两个变量完全相关；φ的绝对值越大，两变量的相关程度越高。但是，当列联表$R\times C$中的行数 R 或列数 C 大于 2 时，φ系数将随着 R 或 C 的变大而增大，且φ值没有上限。这时用φ系数测定两个变量的相关程度就不够清晰，

可以用列联相关系数。

（2）列联相关系数（列联系数，简称 c 系数）

①适用情况：主要用于列联表大于 2×2 的情况。

②计算公式：$c=\sqrt{\frac{\chi^2}{\chi^2+n}}$

③当列联表中的两个变量相互独立时，系数 $c=0$，并且它不可能大于 1。

④c 系数的特点：其可能的最大值依赖于列联表的行数和列数，且随着 R 和 C 的增大而增大。

⑤c 系数的局限性：根据不同行和列计算的列联系数不便于比较，除非两个列联表中的行数和列数一致。

（3）V 相关系数

①计算公式：$V=\sqrt{\frac{\chi^2}{n\times\min[(R-1),(C-1)]}}$

②V 系数的特点：V 的取值在 $0\sim1$ 之间，当两个变量相互独立时，$V=0$；当两个变量完全相关时，$V=1$。

（4）数值分析

①对于同一个数据，系数 φ、c、v 的结果不同；

②在对不同列联表变量之间的相关程度进行比较时，不同列联表中行与行、列与列的个数要相同，并且采用同一种系数，这样系数值才具有可比性。

5.列联分析中应该注意哪些问题？

答：（1）条件百分表的方向

①一般，在列联表中变量的位置是任意的；

②如果变量 X 与 Y 存在因果关系，令 X 为自变量，Y 为因变量，那么一般做法是将自变量放在列的位置；

③如果因变量在样本内的分布不能代表其在总体内的分布，这时仍按自变量的方向计算百分表就会歪曲实际情况。

（2）χ^2 分布的期望值准则

①用 χ^2 分布进行独立性检验，要求样本量必须足够大，特别是每个单元中的期望频数不能过小；

②如果只有两个单元，则每个单元的期望频数必须是 5 或 5 以上；倘若有两个以上的单元，若有 20%的单元的期望频数小于 5 则不能使用 χ^2 检验。

④处理方法是将较小的期望频数合并，这样便可得到合理的结论。

第十章 方差分析

1.什么是方差分析？它研究的是什么？它的适用情况和目的是什么？

答：（1）概念：方差分析就是通过检验各总体的均值是否相等来判断分类型自变量对数值型因变量是否有显著影响。

（2）研究内容：它所研究的是分类型自变量对数值型因变量的影响。

（3）适用情况：用于两个及两个以上样本均值差别的显著性检验。

（4）目的：通过数据分析找出对该事物有显著影响的因素，各因素之间的交互作用，以及显著影响因素的最佳水平等。

2.要检验多个总体均值是否相等时，为什么不作两两比较，而用方差分析方法？

答：（1）方差分析不仅可以提高检验的效率，同时由于它是将所有的样本信息结合在一起，也增加了分析的可靠性。

（2）检验多个总体均值是否相等时，如果作两两比较，一次只能研究两个样本，检验 n 个总体则需要进行 C_n^2 次的 t 检验，十分繁琐。而且随着个体显著性检验次数的增加，偶然因素导致差别的可能性也会增加（并非均值真的存在差别）。而方差分析方法是同时考虑所有的样本，因此排除了错误累积的概率，从而犯第 I 类错误的概率会小很多。

3.方差分析包括哪些类型？它们有何区别？

答：（1）类型：方差分析可分为单因素方差分析和双因素方差分析。

（2）区别：单因素方差分析研究的是一个分类自变量对一个数值型因变量的影响，而双因素方差分析涉及两个分类型自变量。

4.方差分析中有哪些基本假定？

答：（1）每个总体都应服从正态分布，即对于因素的每一个水平，其观测值是来自正态分布总体的简单随机样本。

（2）各个总体的方差 σ^2 必须相同，即各组观察数据是从具有相同方差的正态总体中抽取的。

（3）观测值是独立的。

5.简述方差分析的基本思想。

答：方差分析是通过对数据误差来源的分析来判断不同总体的均值是否相等，进而判断分类型自变量对数值型因变量是否有显著影响。其基本思想表述如下：

（1）误差分解

在方差分析中，数据的误差是用平方和来表示的，总平方和可以分解为组间平方和与组内平方和。组内误差只包含随机误差，而组间误差既包括随机误差，也包括系统误差。

（2）误差分析

如果组间误差只包含随机误差，而没有系统误差。这时，组间误差与组内误差经过平均后的数值就应该很接近，它们的比值就会接近 1；反之，如果在组间误差中除了包含随机误差外还包含系统误差的话，这时组间误差平均后的数值就会大于组内误差平均后的数值，它们之间的比值就会大于 1。当这个比值大到某种程度时，就认为因素的不同水平之间存在着显著差异，即分类型自变量对数值型因变量有影响。

6.方差分析中，解释因素、水平、观测值、总体的含义。

答：（1）因素（因子）：所要检验的对象。

（2）水平（处理）：因素的不同表现。

（3）观测值：在每个因子水平下得到的样本数据。

（4）总体：因素的每一个水平可以看做是一个总体。

7.解释组内误差和组间误差的含义。

答：（1）组内误差（SSE）

①含义：指每个水平或组的各个样本数据与其组平均值误差平方和，反映了因素的同一水平（总体）下，每个样本各观测值的离散状况；

②组内误差只含有随机误差，而随机误差是由抽样的随机性所造成的。

③反映组内误差大小的平方和称为组内平方和，也称为误差平方和或残差平方和，记为 SSE。

（2）组间误差（SSA）

①含义：指各组平均值与总平均值的误差平方和，反映了各样本均值之间的差异程度即因素的不同水平之间观察值的差异程度。

②组间误差是随机误差和系统误差的总和，而系统误差是由于行业本身的系统性因素造成的。

③反映组间误差的平方和称为组间平方和，也称为因素平方和，记为 SSA。

（3）总误差（SST）

①含义：指各个样本数据与总平均值的误差平方和，反映了全部观测值的

离散状况。

②总误差（SST）=组内误差（SSE）+组间误差（SSA）

③反映全部数据误差大小的平方和称为总平方和，记为 SST。

8.解释组内方差（均方）和组间方差（均方）的含义。

答：（1）组间方差

①含义：指因素的不同水平下各个样本之间的方差。

②SSA 的均方也称为组间均方或组间方差，记为 MSA。

③计算公式：$MSA=\dfrac{\text{组间平方和}}{\text{自由度}}=\dfrac{SSA}{k-1}$。

（2）组内方差

①含义：指因素的同一个水平下样本数据的方差。

②SSE 的均方也称为组内均方或组内方差,记为 MSE。

③计算公式：$MSE=\dfrac{\text{组内平方和}}{\text{自由度}}=\dfrac{SSE}{n-k}$

9.简述单因素方差分析的基本步骤。

答：（1）提出假设

①H_0：$\mu_1=\mu_2=\cdots\mu_k$　　自变量对因变量没有显著影响

②H_1：μ_1，μ_2,…，μ_k不全相等　　自变量对因变量有显著影响

（2）构造检验统计量；

①计算各样本的均值

a.假定从第 i 个总体中抽取一个容量为 n_i 的简单随机样本，第 i 个总体的样本均值为该样本的全部观察值总和除以观察值的个数；

b.计算公式：$\overline{x_i}=\dfrac{\sum_{j=1}^{n_j}x_{ij}}{n_i}(i=1,2,\cdots k)$；

式中，n_i 为第 i 个总体的样本观察值个数；x_{ij} 为第 i 个总体的第 j 个观察值。

②计算全部观察值的总均值

a.全部观察值的总和除以观察值的总个数；

b.计算公式：$\overline{\overline{x}}=\dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}}{n}=\dfrac{\sum_{i=1}^{k}n_i\overline{x_i}}{n}$；

式中，$n=n_1+n_2+\cdots n_k$。

③计算各误差平方和

a. $SST=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\overline{\overline{x}}\right)^2$；

b. $SSE=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\overline{x_i}\right)^2$；

c. $SSA=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\overline{x_i}-\overline{\overline{x}}\right)^2=\sum_{i=1}^{k}n_i\left(\overline{x_i}-\overline{\overline{x}}\right)^2$；

d. $SST=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\overline{\overline{x}}\right)^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\overline{x_i}\right)^2+\sum_{i=1}^{k}n_i\left(\overline{x_i}-\overline{\overline{x}}\right)^2$；

④计算统计量

a.各误差平方和的大小与观察值的多少有关，为消除观察值多少对误差平方和大小的影响，需要将其平均，这就是均方，也称为方差。由误差平方和除以相应的自由度求得；

b.三个平方和对应的自由度；

SST：n-1;SSA：k-1；SSE：n-k；其中 k 为因素水平（总体）的个数，n 为全部观察值的个数；

c.计算 MSA 和 MSE；

d.计算 F 统计量：$F=\frac{MSA}{MSE}\sim F(k-1,n-k)$。

（3）统计决策。

①将统计量的值 F 与给定的显著性水平α的临界值 F_α进行比较，做出对原假设的决策。

②根据给定的显著性水平α，在 F 分布表中查找与第一自由度 $df_1=k-1$，第二自由度 $df_2=n-k$ 相应的临界值 F_α。

③若 $F>F_\alpha$，则拒绝原假设，表明均值之间的差异是显著的，所检验的因素对观察值有显著影响；若 $F<F_\alpha$，则不拒绝原假设，无证据表明所检验的因素对观察值有显著影响。

10.简述双因素无交互作用方差分析的步骤。

答：（1）提出假设

分别对行因素、列因素的均值是否相等提出假设。

（2）计算平方和：

①总误差平方和：$SST=\sum_{i=1}^{k}\sum_{j=1}^{r}\left(x_{ij}-\overline{\overline{x}}\right)^2$

②行因素误差平方和：$SSR=\sum_{i=1}^{k}\sum_{j=1}^{r}\left(\overline{x_{i\cdot}}-\overline{\overline{x}}\right)^2$

③列因素误差平方和：$SSC=\sum_{i=1}^{k}\sum_{j=1}^{r}\left(\overline{x_{\cdot j}}-\overline{\overline{x}}\right)^2$

④随机误差项平方和：$SSE=\sum_{i=1}^{k}\sum_{j=1}^{r}\left(x_{ij}-\overline{x_{i\cdot}}-\overline{x_{\cdot j}}+\overline{\overline{x}}\right)^2$

（3）计算均方

$MSR=\frac{SSR}{k-1}$；$MSC=\frac{SSC}{r-1}$；$MSE=\frac{SSE}{(k-1)(r-1)}$

（4）计算检验统计量

$F_R=\frac{MSR}{MSE}\sim F(k-1,(k-1)(r-1))$

$F_C=\frac{MSC}{MSE}\sim F(r-1,(k-1)(r-1))$

（5）作出统计决策（单侧检验）

11.简述双因素有交互作用方差分析的步骤。

答：（1）提出假设

对行变量、列变量和交互作用分别提出假设。

（2）计算平方和：

①总误差平方和：$.SST=\sum_{i=1}^{k}\sum_{j=1}^{r}\sum_{l=1}^{m}\left(x_{ijl}-\overline{\overline{x}}\right)^2$

②行因素误差平方和：$SSR=\sum_{i=1}^{k}\sum_{j=1}^{r}\sum_{l=1}^{m}\left(\overline{x_{i\cdot}}-\overline{\overline{x}}\right)^2=rm\sum_{i=1}^{k}\left(\overline{x_{i\cdot}}-\overline{\overline{x}}\right)^2$

③列因素误差平方和：$SSC=\sum_{i=1}^{k}\sum_{j=1}^{r}\sum_{l=1}^{m}\left(\overline{x_{\cdot j}}-\overline{\overline{x}}\right)^2=km\sum_{j=1}^{r}\left(\overline{x_{\cdot j}}-\overline{\overline{x}}\right)^2$

④交互作用平方和：$SSRC=\sum_{i=1}^{k}\sum_{j=1}^{r}\sum_{l=1}^{m}\left(\overline{x_{ij}}-\overline{x_{i\cdot}}-\overline{x_{\cdot j}}+\overline{\overline{x}}\right)^2$

⑤随机误差项平方和：$SSE=SST-SSR-SSC-SSRC$

（3）计算均方

$MSR=\frac{SSR}{k-1}$；$MSC=\frac{SSC}{r-1}$；$MSRC=\frac{SSRC}{(k-1)(r-1)}$；$MSE=\frac{SSE}{kr(m-1)}$

（4）计算检验统计量

$$F_R = \frac{MSR}{MSE} \sim F\left(k-1,(k-1)(r-1)\right)$$

$$F_C = \frac{MSC}{MSE} \sim F\left(r-1,(k-1)(r-1)\right)$$

$$F_{RC} = \frac{MSRC}{MSE} \sim F\left((k-1)(r-1),kr(m-1)\right)$$

（5）作出统计决策（单侧检验）

12.方差分析中多重比较（最小显著差异方法 LSD）的作用是什么？并简述其具体步骤。

答：（1）多重比较的作用

通过对总体均值之间的配对比较来进一步检验哪些均值之间存在差异。

（2）多重比较的步骤

①提出假设：H_0：$\mu_i = \mu_j$；H_1：$\mu_i \neq \mu_j$。

②计算检验统计量：$\overline{x_i} - \overline{x_j}$。

③计算 LSD：$LSD = t_{\alpha/2}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$。

④根据显著性水平α作出决策。如果$\left|\overline{x_i} - \overline{x_j}\right| > LSD$，则拒绝原假设；如果$\left|\overline{x_i} - \overline{x_j}\right| < LSD$，则不拒绝原假设。

13.什么是交互作用？

答：如果两个因素搭配在一起会对因变量产生一种新的效应，则这种效应称为交互作用。

14.解释无交互作用和有交互作用的双因素方差分析。

答：无交互作用的双因素方差分析中假定两个因素对因变量的影响是独立的；而有交互作用的双因素方差分析需要考虑交互作用对因变量的影响。

15.解释 R^2 的含义和作用。

答：拒绝原假设表明因素（自变量）与观测值之间有显著关系。组间平方和不等于零就表明两个变量之间有关系（只是是否显著的问题）。

当组间平方和比组内平方和大，而且大到一定程度时，就意味着两个变量之间的关系显著，大得越多，表明它们之间的关系就越强。反之，就意味着两个变量之间的关系不显著，小得越多，表明它们之间的关系越弱。

（1）含义：用组间平方和（SSA）占总平方和（SST）的比例大小来反映变量之间的关系强度，将这一比例记为 R^2，即 $R^2=\frac{SSA(组间SS)}{SST(总SS)}$。

（2）作用：R^2 的平方根可以用来测量自变量与因变量之间的关系强度。

第十一章 一元线性回归

1.解释相关关系的含义，并说明相关关系的特点。

答：（1）含义：变量之间存在的互相依存的不确定的数量关系，称为相关关系。

（2）相关关系的特点

①变量之间确实存在着数量上的依存关系。

②变量之间数量上的关系是不确定、不严格的依存关系，不能用函数关系精确表达。

③一个变量的取值不能由另一个变量唯一确定；当变量 x 取某个值时，变量 y 的取值可能有几个。

④各观测点分布在直线周围。

2.相关分析主要解决哪些问题？

答：相关分析通过对两个变量之间的线性关系的描述与度量，主要解决的问题有：

（1）变量之间是否存在关系？

（2）如果存在关系，它们之间是什么样的关系？

（3）变量之间的关系强度如何？

（4）样本所反映的变量之间的关系能否代表总体变量之间的关系？

3.相关分析中有哪些基本假定？

答：（1）两个变量之间是线性关系；

（2）两个变量都是随机变量。

4.简述相关系数的性质、计算公式及其经验解释。

答：（1）性质

①r 的取值范围是[-1，1]，r 绝对值的大小表示相关程度的高低；$|r|=1$为完全相关，r=1 为完全正相关，r=-1 为完全负相关；r=0，不存在线性相关关系；$-1\le r<0$，为负相关；$0<r\le 1$，为正相关。

②对称性：X 与 Y 的相关系数 r_{xy} 和 Y 与 X 之间的相关系数 r_{yx} 相等；

③相关系数与原点和尺度无关；

④相关系数是线性关联或线性相依的一个度量，它不能用于描述非线性关

系；

⑤相关系数只是两个变量之间线性关联的一个度量，却不一定意味两个变量之间有因果关系；

⑥若 X 与 Y 统计上独立，则它们之间的相关系数为零；但 r=0 不等于说两个变量是独立的。即零相关并不一定意味着独立性。

（2）样本相关系数（线性相关系数）的计算公式

$$r=\frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\sqrt{\sum\left(x-\bar{x}\right)^2\sum\left(y-\bar{y}\right)^2}}=\frac{n\sum xy-\sum x\sum y}{\sqrt{n\sum x^2-\left(\sum x\right)^2}\sqrt{n\sum y^2-\left(\sum y\right)^2}}$$

（3）经验解释（建立在对相关系数显著性检验的基础之上）

①$|r|>0.8$时，可视为高度相关；

②$0.5\le|r|<0.8$，可视为中度相关；

③$0.3\le|r|<0.5$，可视为低度相关；

④$|r|<0.3$，说明两个变量之间的相关程度极弱，可视为不相关。

5.为什么要对相关系数进行显著性检验？

答：（1）在实际的客观现象分析研究中，总体相关系数 ρ 是未知的，相关系数一般都是利用样本数据计算的，它受到抽样波动的影响，因而带有一定的随机性。

（2）样本容量越小，其可信程度就越差，抽取的样本不同，r 的取值也会不同，因此 r 是一个随机变量。能否用样本相关系数来反映总体的相关程度，需要考察样本相关系数的可靠性也就是总体线性相关的存在性检验，因此要进行显著性检验。

6.简述相关系数显著性检验的步骤。

答：（1）提出假设

H_0：$\rho=0$；H_1：$\rho\neq0$

（2）计算检验统计量 t 值

$$t=|r|\sqrt{\frac{n-2}{1-r^2}}\sim t(n-2)$$

（3）进行决策

根据给定的显著性水平α和自由度$df=n-2$，查 t 分布表中相应的临界值$t_{\alpha/2}(n-2)$。若$|t|>t_{\alpha/2}$，则拒绝原假设，表明总体的两个变量之间存在显著的线

性关系。（但这并不意味着变量之间存在重要的相关性）

7.解释回归模型、回归方程、估计的回归方程的含义。

答：（1）回归模型：是对统计关系进行定量描述的一种数学模型。例如：对于具有线性关系的两个变量，可以有一元线性方程来描述它们之间的关系，描述因变量 y 如何依赖自变量 x 和误差项ε的方程称为回归模型。

（2）回归方程：是对变量之间统计关系进行定量描述的一种数学表达式。指具有相关的随机变量和固定变量之间关系的方程。

（3）估计的回归方程：当总体回归系数未知时，必须用样本数据去估计，用样本统计量代替回归方程中的未知参数，就得到了估计的回归方程。

8.一元线性回归模型中有哪些基本假定？

答：（1）因变量 y 与自变量 x 之间具有线性关系；

（2）在重复抽样中，自变量 x 的取值是固定的，即假定 x 是非随机的；

（3）误差项ε是一个期望为零的随机变量，即$E(\varepsilon)=0$；

（4）对于所有的 x 值，误差项ε的方差σ^2都相同；

（5）误差项ε是一个服从正态分布的随机变量，且相互独立，即$\varepsilon \sim N(0,\sigma^2)$。独立性意味着一个特定的 x 值所对应的 ε 与其他 x 值所对应的 ε 不相关，因此，一个特定的 x 值所对应的的 y 值与其他 x 值所对应的 y 值也不相关。

9.简述参数最小二乘法的基本原理。

答：（1）使因变量的观测值与估计值之间的离差平方和达到最小来求得β_0和β_1的方法，即$\sum_{i=1}^{n}(y_i-\hat{y})^2=\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_0x_i)^2$最小。

（2）最小二乘法拟合直线有以下优良的性质，正是由于这些性质，将最小二乘法广泛用于回归模型参数的估计。

①根据最小二乘法得到的回归直线能使离差平方和达到最小，虽然这并不能保证它就是拟合数据的最佳直线，但这毕竟是一条与数据拟合良好的直线应有的性质。

②由最小二乘法求得的回归直线可知β_0和β_1的估计量的抽样分布。

③在某些条件下，β_0和β_1的最小二乘估计量同其他估计量相比，其抽样分布具有较小的标准差。

（3）根据最小二乘法，使 $\sum_{i=1}^{n}(y_i-\hat{y})^2=\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_0x_i\right)^2$ 最小。令 $Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$，在给定样本数据后，Q 是 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的函数，且最小值总是存在。根据微积分的极值定理，对 Q 求相应于 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的偏导数，令其等于 0，便可求出 $\hat{\beta}_0$ 和 $\hat{\beta}_1$，即 $\left.\frac{\partial Q}{\partial\beta_0}\right|_{\beta_0=\hat{\beta}_0}=-2\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)=0$，$\left.\frac{\partial Q}{\partial\beta_1}\right|_{\beta_1=\hat{\beta}_1}=-2\sum_{i=1}^{n}x_i\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)=0$。

解上述方程得 $\hat{\beta}_1=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}$

$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$

注意，回归直线必定通过点 $(\bar{x},\bar{y})$。

10.解释总平方和、回归平方和、残差平方和的含义，并说明它们之间的关系。

答：（1）总平方和：指 n 次观测值的的离差平方和，衡量的是被解释变量 y 波动的程度或不确定性的程度。

（2）回归平方和反映 y 的总变差中由于 x 与 y 之间的线性关系引起的 y 的变化部分，这是可以由回归直线来解释的部分，衡量的是被解释变量 y 不确定性程度中能被解释变量 x 解释的部分。

（3）残差平方和是除了 x 对 y 的线性影响之外的其他因素引起的 y 的变化部分，是不能由回归直线来解释的部分。

（4）它们之间的关系是：总平方和=回归平方和+残差平方和。

11.简述判定系数的含义、计算公式、性质和作用。

答：（1）含义

回归平方和占总平方和的比例称为判定系数，记为 R^2。判定系数是对估计的回归方程拟合优度的度量。

（2）计算公式

$$R^2=\frac{SSR}{SST}=1-\frac{SSE}{SST}=\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}=1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}$$

（3）性质

①R^2 的取值范围为$[0,1]$；

②若所有观测点都落在直线上，SSE=0，则 $R^2=1$，拟合是完全的；如果 y 的变化与 x 无关，x 完全无助于解释 y 的变差，$\hat{y}=\bar{y}$，则 $R^2=0$；

③R^2 越接近于 1，表明回归平方和占总平方和的比例越大，回归直线的拟合程度就越好；反之，R^2 越接近于 0，表明回归平方和占总平方和的比例越小，回归直线的拟合程度就越差。

（4）作用

判定系数测量了回归直线对观测数据的拟合程度。

12. 概述相关分析与回归分析的联系与区别。

答：（1）联系

它们都具有共同的研究对象，都是对变量间相关关系的分析，二者可以相互补充。相关分析可以表明变量间相关关系的性质和程度，只有当变量间存在相当程度的相关关系时，进行回归分析去寻找变量间相关的具体数学形式才有实际的意义。同时，在进行相关分析时，如果要具体确定变量间相关的具体数学形式，又要依赖于回归分析，而且在多个变量的相关分析中相关系数的确定也是建立在回归分析基础上的。

（2）区别

①从研究目的上看，相关分析是用一定的数量指标（相关系数）度量变量间相互关系的方向和程度；回归分析却是要寻找变量间联系的具体数学形式，是要根据自变量的给定值去估计和预测因变量的平均值。

②从对变量的处理看，相关分析对称地对待相互联系的变量，不考虑二者的因果关系，也就是不区分自变量和因变量，相关的变量不一定具有因果关系，均视为随机变量；回归分析是在变量因果关系分析的基础上研究其中的自变量的变动对因变量的具体影响，必须明确划分自变量和因变量，所以回归分析中对变量的处理是不对称的，在回归分析中通常假定自变量在重复抽样中是取固定值的非随机变量，只有因变量是具有一定概率分布的随机变量。

13.请说明一元线性回归中，相关系数和判定系数的关系。

答：在一元线性回归中，相关系数 r 实际上是判定系数 R^2 的平方根，其正负号与回归方程中回归系数的符号相同。

14.在回归分析中，F 检验和 t 检验各有什么作用？

答：（1）在回归分析中，F 检验是为检验自变量和因变量之间的线性关系是否显著，通过均方回归与均方残差之比，构造 F 检验统计量，提出假设，根据显著性水平，作出判断。

（2）t 检验是回归系数的显著性检验，要检验自变量对因变量的影响是否显著，通过构造 t 检验统计量，提出假设，根据显著性水平，作出判断。

15.简述一元线性回归中，线性关系检验和回归系数检验的步骤。

答：（1）线性关系检验的步骤

①提出假设：$H_0:\beta_1=0$（两个变量之间的线性关系不显著）；

②计算检验统计量 F：$F=\dfrac{SSR/1}{SSE/(n-2)}=\dfrac{MSR}{MSE}$；

③作出决策：确定显著性水平α，并根据分子自由度$df_1=1$和分母自由度$df_2=n-2$查 F 分布表，找到相应的临界值 Fα。若$F>F_\alpha$,拒绝原假设，表明两个变量之间的线性关系是显著的；若$F<F_\alpha$，没有证据表明两个变量之间的线性关系显著。

（2）回归系数检验的步骤

①提出假设：$H_0:\beta_1=0;\ \ H_1:\beta_1\neq 0$；

②计算检验统计量 t：$t=\dfrac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$；

③作出决策：确定显著性水平α，并根据自由度$df=n-2$查 t 分布表，找到相应的临界值$t_{\alpha/2}$。若$|t|>t_{\alpha/2}$，则拒绝原假设，回归系数等于 0 的可能性小于α，表明自变量 x 对因变量 y 的影响是显著的，换言之，两个变量之间存在着显著的线性关系；若$|t|<t_{\alpha/2}$，则不拒绝原假设，没有证据表明自变量 x 对因变量 y 的影响显著，或者说，二者之间尚不存在显著的线性关系。

16.回归分析结果的评价包括哪几方面？

答：（1）所估计的回归系数的符号是否与理论或事先预期相一致；

（2）如果理论上认为自变量 x 与因变量 y 之间的关系不仅是正的，而且是统计上显著的，那么所建立的回归方程也应该如此；

（3）用判定系数 R^2 来回答回归模型在多大程度上解释了因变量 y 取值的差异；

（4）考察关于误差项ε的正态性假定是否成立。因为在对线性关系进行 F 检验和对回归系数进行 t 检验时，都要求误差项ε服从正态分布，否则，所有的检验程序将是无效的。检验ε正态性的简单方法是画出残差的散点图或正态概率

图。

17.什么是置信区间估计和预测区间估计？二者有何区别？

答：（1）置信区间估计：是对 x 的一个给定值 x_0，求出 y 的平均值的估计区间。这一区间称为置信区间。

（2）预测区间估计：是对 x 的一个给定值 x_0，求出 y 的一个个别值的估计区间。这一区间称为预测区间。

（3）二者的区别：

①对于同一个 x_0，两个区间的宽度不同，预测区间比置信区间宽一些，也就是说，估计 y 的平均值比预测 y 的一个特定值或个别值更精确。

②置信区间估计是求 y 的平均值的估计区间，而预测区间估计是求 y 的一个个别值的估计区间。

18.残差分析在回归分析中的作用是什么？

答：回归分析是确定两种或两种以上变量间的定量关系的一种统计分析方法。判断回归模型的拟合效果是回归分析的重要内容，在回归分析中，通常用残差分析来判断回归模型的拟合效果，并判定关于误差项的正态假设是否成立。

19.什么是残差分析？并解释估计标准误差。

答：（1）残差分析

①含义：确定有关 ε 的假定是否成立的方法之一。

②残差：是因变量的观测值与根据估计的回归方程求出的预测值之差，用 e 表示，它反应了用估计的回归方程去预测 y_i 而引起的误差。

③残差的计算公式：$e_i = y_i - \hat{y}_i$

④判断误差项 ε 的假定是否成立：

a.利用残差图判断 ε 的方差是否相等：若对所有的 x 值，ε 的方差都相同，而且假定描述变量 x 和 y 之间关系的回归模型是合理的，那么残差图中的所有点都应落在一条水平带中间。

b.利用标准化残差判断 ε 是否服从正态分布：标准化残差是残差除以它的标准差后得到的数值，也称为 Pearson 残差或半学生化残差，用 z_e 表示，计算公式为：$z_{e_i} = \frac{e_i}{s_e} = \frac{y_i - \hat{y}_i}{s_e}$。如果误差项 ε 服从正态分布这一假定成立，那么标准化残差的分布也应服从正态分布。因此，在标准化残差图中，大约有 95%的标准化残差在 $-2 \sim 2$ 之间。

（2）估计标准误差

①残差平方和可以说明实际观测值 y_i 与回归估计值 $\hat{y}_i$ 之间的差异程度。

②定义：度量各实际观测点在直线周围的散布状况的一个统计量，它是均方残差（MSE）的平方根，用 s_e 来表示。

③计算公式：$s_e=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}}=\sqrt{\frac{SSE}{n-2}}=\sqrt{MSE}$

④估计标准误差是对误差项 ε 的估计，它可以看做在排除了 x 对 y 的线性影响后，y 随机波动大小的一个估计量。从估计标准误差的实际意义来看，它反应了用估计的回归方程预测因变量 y 时预测误差的大小。各观测点越靠近直线，s_e 越小，回归直线对各观测点的代表性就越好，根据估计的回归方程进行预测也就越准确。可见 s_e 从另一个角度说明了回归直线的拟合优度。

第十二章 多元线性回归

1.解释多元回归模型、多元回归方程、估计的多元回归方程的含义。

答：（1）多元回归模型：设因变量为 y，k 个自变量分别为 $x_1,x_2,\cdots x_k$，描述因变量 y 如何依赖于自变量 $x_1,x_2,\cdots x_k$ 和误差项 ε 的方程 $y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k+\varepsilon$ 称为多元回归模型。其中，β_0，β_1，…，β_k 是模型的参数；ε 为误差项。

（2）多元回归方程：在多元回归模型的基本假定下，因变量 y 的期望为 $E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k$，该式被称为多元回归方程，它描述了因变量 y 与自变量 $x_1,x_2,\cdots x_k$ 之间的关系。

（3）估计的多元回归方程：回归方程中的参数 β_0，β_1，…，β_k 是未知的，需要利用样本数据去估计它们。当用样本统计量 $\hat{\beta}_0$，$\hat{\beta}_1$，…，$\hat{\beta}_k$ 去估计回归方程中的未知参数 β_0，β_1，…，β_k 时，就得到了估计的多元回归方程 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_kx_k$。式中，$\hat{\beta}_0$，$\hat{\beta}_1$，…，$\hat{\beta}_k$ 是参数 β_0，β_1，…，β_k 的估计值，$\hat{y}$ 是因变量 y 的估计值。$\hat{\beta}_1$，…，$\hat{\beta}_k$ 称为偏回归系数。

2.多元线性回归模型中有哪些基本假定？

答：（1）误差项ε是一个期望值为0的随机变量，即$E(\varepsilon)=0$。这意味着对于给定的$x_1,x_2,\cdots x_k$的值，y的期望值为$E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k$。

（2）对于自变量$x_1,x_2,\cdots x_k$的所有值，ε的方差σ^2都相同。

（3）误差项ε是一个服从正态分布的随机变量，且相互独立，即$\varepsilon\sim N(0,\sigma^2)$。独立性意味着自变量$x_1,x_2,\cdots x_k$的一组特定值所对应的$\varepsilon$与$x_1,x_2,\cdots x_k$任意一组其他值所对应的$\varepsilon$不相关；正态性意味着对于给定的$x_1,x_2,\cdots x_k$的值，因变量y是一个服从正态分布的随机变量。

3.解释多重判定系数和调整的多重判定系数的含义和作用。

答：（1）多重判定系数R^2

①含义：是多元回归中的回归平方和占总平方和的比例，它是度量多元回归方程拟合程度的一个统计量，反映了在因变量的变差中被估计的回归方程所解释的比例，公式为$R^2=\frac{SSR}{SST}=1-\frac{SSE}{SST}$。

②R^2的平方根称为多重相关系数，也称为复相关系数，它度量了因变量同k个自变量的相关程度。

（2）调整的多重判定系数R_a^2

①含义：为避免增加自变量而高估R^2，统计学家提出用样本量n和自变量的个数k去调整R^2，计算出调整的多重判定系数R_a^2。R_a^2表示在用样本量和模型中自变量的个数进行调整后，在因变量的变差中被估计的回归方程所解释的比例，公式为$R_a^2=1-(1-R^2)(\frac{n-1}{n-k-1})$。

②作用：R_a^2同时考虑了样本量n和模型中自变量的个数k的影响，这就使得R_a^2的值永远小于R^2，而且R_a^2的值不会由于模型中自变量个数的增加而越来越接近于1。

4.解释多重共线性的含义。

答：当回归模型中两个或两个以上的自变量彼此相关时，则称回归模型中存在多

重共线性。

5.多重共线性对回归分析有哪些影响？

答：（1）变量之间高度相关时，可能会使回归的结果混乱，甚至会把分析引入歧途；

（2）多重共线性可能对参数估计值的正负号产生影响，特别是β_i的正负号有可能同预期的正负号相反。

6.多重共线性的判别方法主要有哪些？

答：（1）计算模型中各对自变量之间的相关系数，并对各相关系数进行显著性检验。如果有一个或多个相关系数是显著的，就表示模型中所使用的自变量之间相关，因而存在多重共线性问题。

（2）经验判别：出现如下情况暗示存在多重共线性。

①模型中各对自变量之间显著相关。

②当模型的线性关系检验（F 检验）显著时，几乎所有回归系数β_i的t检验却不显著。

③回归系数的正负号与预期的相反。

④容忍度与方差扩大因子（VIF）。某个自变量的容忍度等于 1 减去该自变量为因变量而其他 k-1 个自变量为预测变量时所得到的线性回归模型的判定系数，即$1-R_i^2$。方差扩大因子等于容忍度的倒数，即$VIF=\frac{1}{1-R_i^2}$。容忍度越小，也即方差扩大因子 VIF 越大，多重共线性越严重。通常容忍度小于 0.1，也即 VIF 大于 10 时，存在严重的多重共线性。

7.多重共线性的处理方法有哪些？

答：（1）将一个或多个相关的自变量从模型中剔除，使保留的自变量尽可能不相关。

（2）如果要在模型中保留所有的自变量，那就应该：避免根据t统计量对单个参数β进行检验；对因变量y值的推断（估计或预测）限定在自变量样本值的范围内。

8.在多元线性回归中，选择自变量的方法有哪些？

答：向前选择、向后剔除、逐步回归、最优子集等。

第十三章 时间序列分析和预测

1.简述时间序列的含义及构成要素。

答：（1）含义：时间序列是同一现象在不同时间的相继观察值排列而成的序列。

（2）构成要素：

①趋势（长期趋势）（T）：时间序列在长期内呈现出来的某种持续上升或持续下降的变动。趋势可以是线性的，也可以是非线性的。

②季节性（季节变动）（S）：时间序列在一年内重复出现的周期性波动。

③周期性（循环波动）（C）：时间序列中呈现出来的围绕长期趋势的一种波浪形或振荡式变动。

④随机性（不规则波动）（I）：由偶然性因素对时间序列产生影响，致使时间序列呈现出某种随机波动。时间序列中除去趋势、周期性和季节性之后的偶然性波动。

（3）周期性与趋势和季节性的区别：

①周期性通常是由商业和经济活动引起的，它不同于趋势变动，不是朝着单一方向的持续运动，而是涨落相间的交替波动；

②周期性也不同于季节变动，季节变动有比较固定的规律，且变动周期大多为一年，循环波动则无固定规律，变动周期多在一年以上，且周期长短不一，周期性通常是由经济环境的变化引起的。

2.利用增长率分析时间序列时应该注意哪些问题？

答：（1）当时间序列中的观察值出现 0 或负值时，不宜计算增长率。这是因为对这样的序列计算增长率，要么不符合数学公理，要么无法解释其实际意义。在这种情况下，适宜直接用绝对数进行分析。

（2）在有些情况下，不能单纯就增长率论增长率，要注意将增长率与绝对水平结合起来分析。增长率是一个相对值，它与对比的基期值的大小有很大关系。大的增长率背后，其隐含的绝对值可能很小，小的增长率背后，其隐含的绝对值可能很大。这就是说，由于对比的基点不同，可能会造成增长率数值上的较大差异。在这种情况下，则需要将增长率与绝对水平结合起来进行分析，通常要计算增长 1%的绝对值来克服增长率分析的局限性。增长 1%的绝对值表示增长率每增长一个百分点而增加的绝对数量，其计算公式为：$\text{增长1\%的绝对值}=\frac{\text{前期水平}}{100}$。

3.简述平稳序列和非平稳序列的含义。

答：（1）平稳序列

①含义：基本上不存在趋势的序列。

②特点：这类序列的各观察值基本上在某个固定的水平上波动，虽然在不同的时间段波动的程度不同，但并不存在某种规律，波动可以看成是随机的。

（2）非平稳序列

①含义：是包含趋势、季节性或周期性的序列，它可能只包含其中一种成分，也可能包含有几种成分。

②分类：非平稳序列可以分为有趋势的序列、有趋势和季节性的序列、几种成分混合而成的复合型序列。

4.简述时间序列的预测程序。

答：（1）确定时间序列所包含的成分，也就是确定时间序列的类型。确定趋势成分和季节成分是否存在，可以从绘制时间序列的线图入手。

（2）找出适合此类时间序列的预测方法。如简单平均法、移动平均法、指数平滑法、自回归模型（ARMA）等。

（3）对可能的预测方法进行评估，以确定最佳预测方案。评估的方法就是找出预测值与实际值的差距即预测误差，最优的预测方法也就是预测误差达到最小的方法。

（4）利用最佳预测方案对未来各期的时间序列数值进行预测。

5.简述预测平稳时间序列的几种方法的基本含义。

答：（1）简单平均法

①含义：根据已有的 t 期观察值通过简单平均来预测下一期的数值。

②步骤：

a. 设时间序列已有的 t 期观察值为$Y_1,Y_2,\cdots Y_t$，则 t+1 期的预测值 F_{t+1} 为：

$F_{t+1}=\frac{1}{t}(Y_1+Y_2+\cdots Y_t)=\frac{1}{t}\sum_{i=1}^{t}Y_i$。

b. 到了 t+1 期之后，有了 t+1 的实际值，便可计算出 t+1 期的预测误差 e_{t+1}：

$e_{t+1}=Y_{t+1}-F_{t+1}$。

c. 于是 t+2 期的预测值为：$F_{t+2}=\frac{1}{t}(Y_1+Y_2+\cdots Y_t+Y_{t+1})=\frac{1}{t+1}\sum_{i=1}^{t+1}Y_i$。

d. 依次类推。

③适用情况：简单平均法适合对较为平稳的时间序列进行预测，如果时间序列有趋势或季节成分，该方法的预测则不够准确。

（2）移动平均法

①含义：是通过对时间序列逐期递移求得平均数作为预测值的一种预测方

法。

②分类：简单移动平均法、加权移动平均法

③特点：

a.对原序列有修匀或平滑的作用。时距项数N越大，对数列的修匀作用越强；

b.移动平均项数N为偶数时，需移正平均；

c.平均时距项数N与季节变动长度一致才能消除季节变动；时距项数N和周期一致才能消除周期波动；

d.移动平均会使原序列失去部分信息，平均项数越大，失去的信息越多。

e.N为奇数时：$\frac{N-1}{2}$；N为偶数时：$\frac{N}{2}$。

④简单移动平均法的步骤：

a.将最近的k期数据加以平均，作为下一期的预测值。设移动间隔为$k(1<k<t)$，则t期的移动平均值为：$\overline{Y_t}=\frac{Y_{t-k+1}+Y_{t-k+2}+\cdots+Y_{t-1}+Y_t}{k}$。

b.t+1期的简单移动平均预测值为：$F_{t+1}=\overline{Y_t}=\frac{Y_{t-k+1}+Y_{t-k+2}+\cdots+Y_{t-1}+Y_t}{k}$。

c.同样，t+2期的预测值为：$F_{t+2}=\overline{Y_{t+1}}=\frac{Y_{t-k+2}+Y_{t-k+3}+\cdots+Y_t+Y_{t+1}}{k}$。

d.依次类推。

④适用情况：移动平均法只使用最近k期的数据，在每次计算移动平均值时，移动的间隔都为k。该方法也适合对较为平稳的时间序列进行预测。关键是确定合理的移动间隔k，对于同一个时间序列，采用不同的移动间隔，预测的准确性是不同的。可通过试验选择一个使均方误差最小的移动间隔。

（3）指数平滑法

①含义：是通过对过去的观察值加权平均进行预测的一种方法，该方法使t+1期的预测值等于t期的实际观察值与t期的预测值的加权平均值。

②特点：指数平滑法是加权平均的一种特殊形式，观察值的时间越远，其权数呈现指数下降，因而称为指数平滑。

③分类：指数平滑法有一次指数平滑法、二次指数平滑法、三次指数平滑法等。一次指数平滑法也可用于对时间序列进行修匀，以消除随机波动，找出序列的变化趋势。

④一次指数平滑法（单一指数平滑法）的含义：

a.只有一个平滑系数；

b.观察值离预测时期越久远，权数变得越小；

c.以一段时间的预测值与观察值的线性组合作为第t+1期的预测值，其预测模型为$F_{t+1}=\alpha Y_t+(1-\alpha)F_t$，其中$Y_t$为第t期的实际观察值，$F_t$为第t期的预测

值，α为平滑系数$(0<\alpha<1)$。

⑤一次指数平滑法的步骤：

a. 在开始计算时还没有 1 期的预测值 F_1，通常可以设 F_1 等于 1 期的实际观察值，即 $F_1=Y_1$。

b. 2 期的预测值为：$F_2=\alpha Y_1+(1-\alpha)F_1=\alpha Y_1+(1-\alpha)Y_1$。

c. 3 期的预测值为：$F_3=\alpha Y_2+(1-\alpha)F_2=\alpha Y_2+(1-\alpha)Y_1$。

d. 4 期的预测值为：$F_4=\alpha Y_3+(1-\alpha)F_3=\alpha Y_3+\alpha(1-\alpha)Y_2+(1-\alpha)^2 Y_1$。

e. t+1 期的预测值为：$F_{t+1}=\alpha Y_t+(1-\alpha)F_t$。

⑥预测一次指数平滑法的误差：

a. 对指数平滑法的预测精度，用均方误差来衡量，因此可将 F_{t+1} 改写为：$F_{t+1}=\alpha Y_t+(1-\alpha)F_t=\alpha Y_t+F_t-\alpha F_t=F_t+\alpha(Y_t-F_t)$。

b. 可见，F_{t+1} 是 t 期的预测值 F_t 加上用 α 调整的 t 期的预测误差 (Y_t-F_t)。

⑦一次指数平滑法中α的确定：

a. 不同的α会对预测结果产生不同的影响。当时间序列有较大的随机波动时，宜选较大的α，以便能尽快跟上近期的变化。当时间序列比较平稳时，宜选取较小的α。

b. 选择α时，还应考虑预测误差。用均方误差来衡量预测误差的大小。确定α时，可选择几个进行预测，然后找出预测误差最小的作为最后的值。

6.简述复合型时间序列的预测步骤。

答：复合型时间序列是指含有趋势性、季节性、周期性和随机成分的序列。对这类序列预测方法通常是将时间序列的各个因素依次分解出来，然后再进行预测，分解法预测通常按下面的步骤进行：

（1）测定并分离长期趋势。利用移动平均法、时间回归法等方法来测定出时间序列的长期趋势，并将其从时间序列中分离出去。

（2）确定并分离季节成分。计算季节指数，以确定时间序列中的季节成分。然后将季节成分从时间序列中分离出去，即用每一个时间序列观察值除以相应的季节指数，以消除季节性。

（3）剩余法测定循环变动。再用移动平均法消除（2）中所得的时间序列中的不规则变动，即得到原时间序列中的周期性成分。

（4）建立预测模型并进行预测。对原时间序列建立适当的预测模型，并根

据这一模型进行预测，计算出最后的预测值。

7.简述分解法预测的基本步骤。

答：（1）确定并分离季节成分。计算季节指数，以确定时间序列中的季节成分。然后将季节成分从时间序列中分离出去，即用每一个时间序列观测值除以相应的季节指数，以消除季节因素的影响。

（2）建立预测模型并进行预测。对消除季节成分的时间序列建立适当的预测模型，并根据这一模型进行预测。

（3）计算出最后的预测值。用预测值乘以相应的季节指数，得到最终的预测值。

8.简述季节指数的计算步骤。

答：以移动平均趋势剔除法为例，计算季节指数的基本步骤为：

（1）计算移动平均值（如果是季度数据采用 4 项移动平均，月份数据则采用 12 项移动平均），并将其结果进行“中心化”处理，也就是将移动平均的结果再进行一次 2 项的移动平均，即得出“中心化移动平均值”（CMA）。

（2）计算移动平均的比值，也称为季节比率，即将序列的各观察值除以相应的中心化移动平均值，然后再计算出各比值的季度（或月份）平均值。

（3）季节指数调整。由于各季节指数的平均数应等于 1 或 100%，若根据第（2）步计算的季节比率的平均值不等于 1 时，则需要进行调整。具体方法是：将第（2）步计算的每个季节比率的平均值除以时间序列的总平均值，即得到调整后的季节指数。

9.如何选择时间序列数据的预测方法？

答：（1）首先判断是否存在趋势；

（2）继续判断是否存在季节性；

（3）若不存在趋势但存在季节性或存在趋势也存在季节性，则使用季节性预测法：季节多元回归模型、季节自回归模型、时间序列分解；若不存在趋势也不存在季节性，则使用平滑法预测：简单平均法、移动平均法、指数平滑法；若存在趋势但不存在季节性，则使用趋势预测方法：线性趋势推测、非线性趋势推测、自回归预测模型。

10.评估预测的方法有哪些？

答：（1）平均误差：$ME=\dfrac{\sum_{i=1}^{n}(Y_i-F_i)}{n}$；

（2）平均绝对误差：$MAD=\dfrac{\sum_{i=1}^{n}|Y_i-F_i|}{n}$；

（3）均方误差：$MSE=\dfrac{\sum_{i=1}^{n}(Y_i-F_i)^2}{n}$；

（4）平均百分比误差：$MPE=\dfrac{\sum_{i=1}^{n}\left(\dfrac{Y_i-F_i}{Y_i}\times 100\right)}{n}$；

（5）平均绝对百分比误差：$MAPE=\dfrac{\sum_{i=1}^{n}\left(\dfrac{|Y_i-F_i|}{Y_i}\times 100\right)}{n}$。

第十四章 指数

1.什么是指数？它有哪些性质？

答：（1）指数的概念：指数也称统计指数，是分析社会经济现象数量变化的一种重要统计方法，有狭义和广义之分。狭义的统计指数是指综合反映不能直接相加的社会经济现象总体总动态的相对数；广义的统计指数是指说明同类现象对比的相对数。

具体来说，指数是测定多项内容数量综合变动的相对数。此概念包含两个要点：一是指数的实质是测定多项内容。指数方法论研究如何将多项内容合在一起，从整体上进行反应。二是指数的表现形式为动态相对数，既然是动态相对数，就涉及指标的基期对比，不同要素基期的选择就成为指数方法需要讨论的问题。

（2）性质：

①相对性。指数是总体各变量在不同场合下对比形成的相对数，它可以度量一个变量在不同时间或不同空间的相对变化，如一种商品的价格指数或数量指数。它也可以反映一组变量的综合变动，比如综合物价指数是根据一组商品价格的相对变化并给每种商品的相对数定以不同权数计算出来的，这种指数称为综合指数。另外根据对比两变量所处的是不同时间还是不同空间，它们计算出来的指数分时间性指数和区域性指数。

②综合性。综合性说明指数是一种特殊的相对数，它是由一组变量或项目综合对比形成的。比如，由若干种商品和服务构成的一组消费项目，通过综合后计算价格指数，以反映消费价格的综合变动水平。

③平均性。平均性含义有二：一是指数进行比较的综合数量是作为个别量的一个代表，这本身就具有平均的性质；二是两个综合量对比形成的指数反映了个别量的平均变动水平，比如物价指数反映了多种商品和服务项目价格的平均变动水平。

2.什么是同度量因素？同度量因素在编制加权综合指数中有什么作用？

答：（1）含义：在统计学中，一般把使得不能直接相加的指标过渡到可以直接相加的指标的媒介因素称为同度量因素或同度量系数。如：产品产量指数能够综合地反映多种产品产量的变动情况，然后各种产品的使用价值和计量单位均不相同，其产量无法直接相加和对比。这就需要借助另一个因素即产品出厂价格，使不能直接加总的产量指标过渡到能够相加的价值指标，此处的产品出厂价格就被称为同度量因素。

（2）作用：在编制加权综合指数时，同度量因素起着同度量和权数的双重作用。在本例中，出厂价格越高的产品，其产量与出厂价格的乘积越大，从而它对产量综合指数的影响也越大。

3.拉氏指数与帕氏指数的概念是什么？它们各有什么特点？

答：（1）拉氏指数

①定义：在计算综合指数时将作为权数的同度量因素固定在基期。

②计算公式：

a. 拉式数量指标指数：$I_q=\frac{\sum q_1p_0}{\sum q_0p_0}$；

b. 拉式质量指标指数：$I_p=\frac{\sum q_0p_1}{\sum q_0p_0}$。

③特点：拉式指数是由德国学者拉斯贝尔斯在 1864 年提出来的，它是用基期消费量加权来计算价格指数，这一指数被称为拉式指数。拉氏指数是以基期变量值为权数来计算的指数，它可以消除权数变动对指数的影响，从而使不同时期的指数具有可比性。在实际应用中，拉氏指数方法常用来计算数量指数。因为拉式数量指数是假定价格不变的条件下报告期销售量的综合变动，它不仅可以单纯反映出销售量的综合变动水平，也符合计算销售量指数的实际要求。而拉式价格指数在实际中应用得很少，因为从实际生活角度看，人们更关心在报告期销售量条件下价格变动对实际生活的影响。而拉式价格指数是在假定销售量不变的情况下报告期价格的变动水平，这一指数尽管可以单纯反映价格的变动水平，但不能反映出销售量的变化。

（2）帕氏指数

①定义：在计算综合指数时将作为权数的同度量因素固定在报告期。

②计算公式：

a. 帕式数量指标指数：$I_q=\frac{\sum q_1p_1}{\sum q_0p_1}$；

b. 帕式质量指标指数：$I_p=\frac{\sum q_1p_1}{\sum q_1p_0}$。

③特点：帕式指数是由德国学者帕舍在 1874 年提出来的。帕氏指数是以报告期变量为权数来计算的指数，它不能消除权数变动对指数的影响，因而不同时期的指数缺乏可比性。但帕式指数可以同时反映出价格和消费结构的变化，具有比较明确的经济意义。在实际应用中，常采用帕式公式计算价格、成本等质量指数。而帕式数量指数由于包含了价格的变动，意味着是按调整后的价格来测定物量的综合变动，这本身不符合计算物量指数的目的，因此帕式数量指数在实际中应用得较少。

4.加权平均指数与加权综合指数的概念是什么？它们有何区别与联系？

答：（1）加权综合指数

①拉式指数

②帕式指数

（2）加权平均指数

①定义：以个体指数为基础，通过对个体指数进行加权平均来编制的指数；先计算所研究现象各个项目的个体指数，然后根据所给的价值量指标作为权数对个体指数进行加权平均。

②计算公式：

a. 加权算术平均指数（基期总值加权）：$A_q=\frac{\sum\frac{q_1}{q_0}q_0p_0}{\sum q_0p_0}$；$A_p=\frac{\sum\frac{p_1}{p_0}q_0p_0}{\sum q_0p_0}$；

b. 加权调和平均指数（报告期总值加权）：$H_q=\frac{\sum q_1p_1}{\sum\frac{q_0}{q_1}q_1p_1}$；$H_p=\frac{\sum q_1p_1}{\sum\frac{p_0}{p_1}q_1p_1}$。

（3）区别：

①二者在所使用的权数和计算形式上不同。综合指数是以某一时期的变量值作为权数对另一个变量进行加权，然后采用综合的形式计算出来的；而加权平均指数则是采用某一总量为权数对个体指数加权平均计算出来的；

②二者所依据的计算资料不同。加权综合指数的计算通常需要掌握全面的资料；加权平均指数既可以依据全面资料计算，也可依据非全面资料来计算。

（4）联系：当使用 p_0q_0 为权数时，加权算术平均指数可以变形为加权综合指数；当使用 p_1q_1 为权数时，加权调和平均指数可以变形为加权综合指数。

5.什么是指数体系？它有什么作用？

答：（1）定义：由总量指数及若干个因素指数构成的数量关系式称为指数体系。它一般保持两个对等关系，一是各影响因素指数的连乘积等于总变动指数；二是各因素对总额变动影响差额的总和等于实际发生的总差额。

（2）作用：

①指数体系是进行因素分析的根据。即利用指数体系可以分析复杂经济现象总变动中各因素变动影响方向和程度。

②利用各指数之间的联系进行指数间的相互推算。例如，我国商品销售量总指数往往就是根据商品销售额总指数和价格总指数进行推算的。即商品的销售量指数=销售额指数÷价格指数。

③用综合指数法编制总指数时，指数体系也是确定同度量因素时期的根据之一。因为指数体系是进行因素分析的根据，要求各个指数之间在数量上要保持一定的联系。所以编制产品产量指数时，如用基期价格作同度量因素，那么编制产品价格指数时就必须用报告期的产品产量作为同度量因素；如果编制产品产量指数用报告期价格作为同度量因素，那么编制产品价格指数时就必须用基期的产品产量作为同度量因素。

6.试述平均数指数体系。

答：平均数的变动受两个因素的影响：一个是各组的变量水平（x）；另一个是各组的结构 $\left(\frac{f}{\sum f}\right)$。

（1）总平均水平指数：$I_{xf}=\frac{\overline{x_1}}{\overline{x_0}}=\frac{\sum x_1f_1/\sum f_1}{\sum x_0f_0/\sum f_0}$；

（2）组水平变动指数：$I_x=\frac{\overline{x_1}}{\overline{x_n}}=\frac{\sum x_1f_1/\sum f_1}{\sum x_0f_1/\sum f_1}$；

（3）结构变动指数：$I_f=\frac{\overline{x_n}}{\overline{x_0}}=\frac{\sum x_0f_1/\sum f_1}{\sum x_0f_0/\sum f_0}$。

此时，指数体系的具体表现为：

总平均水平指数=组水平变动指数×结构变动指数，即

$$\frac{\sum x_1 f_1 / \sum f_1}{\sum x_0 f_0 / \sum f_0} = \frac{\sum x_1 f_1 / \sum f_1}{\sum x_0 f_1 / \sum f_1} \times \frac{\sum x_0 f_1 / \sum f_1}{\sum x_0 f_0 / \sum f_0}，\ I_{xf} = I_x \times I_f$$

总平均水平变动额=各组水平变动影响额+结构变动影响额，即

$$\left(\sum x_1 f_1 / \sum f_1 - \sum x_0 f_0 / \sum f_0\right) = \left(\sum x_1 f_1 / \sum f_1 - \sum x_0 f_1 / \sum f_1\right) + \left(\sum x_0 f_1 / \sum f_1 - \sum x_0 f_0 / \sum f_0\right)$$

$$\overline{x_1} - \overline{x_0} = \left(\overline{x_1} - \overline{x_n}\right) + \left(\overline{x_n} - \overline{x_0}\right)$$

我们可以把总体平均数的变动分解为组水平的影响和结构变动的影响。进行分析时，将总体结构看成数量指标，将各组变量值看成质量指标。在研究结构的变动对平均数的影响时，将各组变量值固定在基期，在研究各组变量值的变动对平均数的影响时，将结构固定在报告期。

7.构建综合评价指数时需要考虑哪些方面的问题？

答：（1）进行理论研究，其中包括统计指标理论以及统计指标体系的理论研究，以便为确定所需的评价指标提供一定的理论依据。

（2）建立科学的评价指标体系。所建立的指标体系是否科学与合理，直接关系到评价结果的科学性和准确性。建立指标体系，首先应进行必要的定性研究，对所研究的问题进行深入的分析，尽量选择那些具有一定综合意义的代表性指标；其次，应尽可能运用多元统计方法进行指标的筛选，以提高指标的客观性。

（3）评价方法研究，主要包括综合评价指数的构造方法、指标的赋权方法以及各种评价方法的比较等。

8.构建综合评价指数一般需要哪些步骤？

答：（1）建立综合评价指标体系。

所建立的指标体系是否科学与合理，直接关系到评价结果的科学性和准确性。首先应进行必要的定性研究，对所研究的问题进行深入的分析，尽量选择那些具有一定综合意义的代表性指标；其次，应尽可能运用多元统计方法进行指标的筛选，以提高指标的客观性。

（2）评价指标的无量纲化处理。

由于综合评价需要运用由多个指标组成的指数体系，而这些指标的性质不同，计量单位不同，具有不同的量纲，因此需要对各指标的实际数据进行无量纲化处理，使之具有可比性，在此基础上才有可能进行综合。

（3）确定各项评价指标的权重。

对于不同的指标，不同的人从不同的角度审视，会有不同的看法和评价，所以要在综合评价中确定各项指标的权重。有两类方法：一类是主观确定权数，可以采用多种方式，但共同特征是由有关专家通过研究讨论决定，特点是可以集中专家集体智慧，工作效率比较高，但很难找到客观的评价标准。另一类是客观

确定权数，也有许多不同的实现方法， 但共同特征是权数由实际数据确定。也有一些研究采用专门的统计方法，通过模型或其他计算方式产生权数。这类方法的特点是依据数据，客观性更强，但有时难以反映评价的导向性。

（4）计算综合评价指数。

有了各项指标的无量纲化处理结果，有了各项指标的权重，通过适当的方法，就可以得到综合评价指数。

9.综合评价指数的构建方法是什么？

答：（1）无量纲化处理

①统计标准化：$z_i = \frac{x_i - \bar{x}}{s}$；

②相对标准化：$z_i = \frac{x_i}{x_s}$，x_s为进行标准化确定的对比标准，通常可以选择最优值或平均值作为对比标准。

③功效系数法：

a. 对多目标规划原理中的功效系数加以改进，从而把要评价的指标转化为可以度量的评判分数，公式为：$z_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$；

b. 改进的功效系数法：得到的标准化分数在$60 \sim 100$之间，可以减少极端数值对计算结果的视觉影响，接近人们对分数的一般看法，公式为：

$z_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \times 40 + 60$。

（2）用加权平均方式计算综合评价指数：$I = \frac{\sum_{i=1}^{n} z_i w_i}{\sum_{i=1}^{n} w_i}, 0 \le w_i \le 1, \sum_{i=1}^{n} w_i = 1$。